Semiparametric M-estimators and their applications to multiple change-point problems

Anouar Abdeldjaoued Ferfache

▶ **To cite this version:**

Anouar Abdeldjaoued Ferfache. Semiparametric M-estimators and their applications to multiple change-point problems. Statistics [math.ST]. Université de Technologie de Compiègne, 2021. English. NNT : 2021COMP2643 . tel-03774522

HAL Id: tel-03774522

https://theses.hal.science/tel-03774522v1

Submitted on 11 Sep 2022

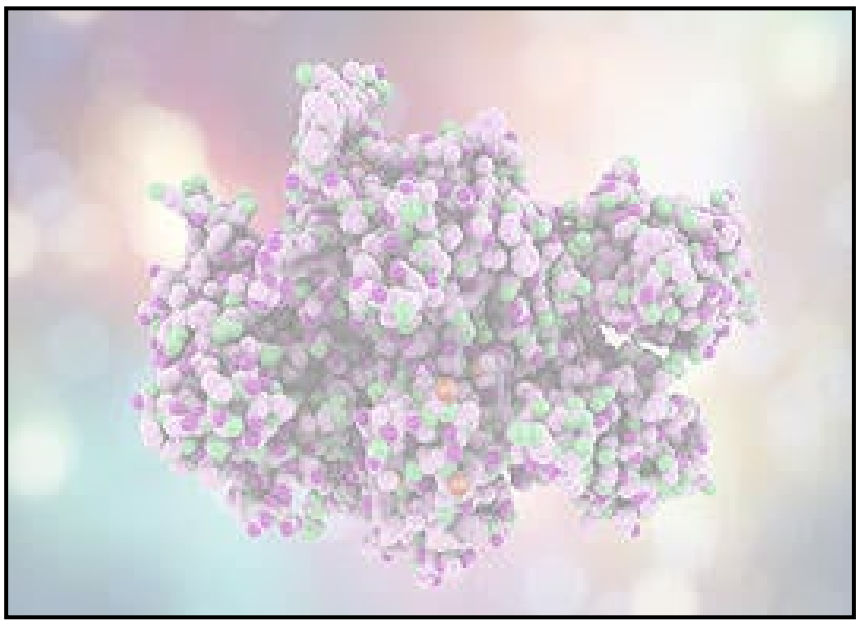
HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Par **Anouar Abdeldjaoued FERFACHE**

Les M-estimateurs semiparamétriques et leurs applications pour les problèmes de ruptures

Thèse présentée
pour l'obtention du grade
de Docteur de l'UTC



Soutenue le 7 décembre 2021
Spécialité : Mathématiques Appliquées et Statistique : Laboratoire de Mathématiques Appliquées de Compiègne (Unité de recherche EA-2222)
D2643

École doctorale des Sciences pour l'ingénieur

Alliance Sorbonne Université

THÈSE

pour obtenir le grade de docteur délivré par

Université de Technologie de Compiègne

Spécialité : Mathématiques Appliquées et Statistique

présentée et soutenue publiquement par

Anouar Abdeldjaoued FERFACHE

le 07/12/2021

Les M-estimateurs semiparamétriques et leurs applications pour les problèmes de ruptures

Directeur de thèse : **Salim BOUZEBDA**

Jury

M. Mokhtar Z. Alaya,	Maître de Conférences	Invité
M. Patrice Bertail,	Professeur	Rapporteur
M. Salim Bouzebda,	Professeur	Directeur
M. Issam Elhattab,	Maître de Conférences	Invité
Mme. Ghislaine Gayraud,	Professeure	Examinatrice
M. Amor Keziou,	Maître de Conférences	Examinateur
M. Nikolaos Limnios,	Professeur	Président du jury
M. Johan Segers,	Professeur	Rapporteur

UTC
Laboratoire de Mathématiques Appliquées de Compiègne (LMAC)
CS 60319 - 57 avenue de Landshut Compiègne Cedex, 60203 France

THESE

Publications

[P.1] *Asymptotic properties of* M*-estimators based on estimating equations and censored data in semi-parametric models with multiple change points*. **Journal of Mathematical Analysis and Applications. 497 (2021)**, no. 2, Paper No. 124883, 44 pp. Joint work with S. Bouzebda.

[P.2] *Asymptotic Properties of Semiparametric* M*-Estimators with Multiple Change Points.* Joint work with S. Bouzebda. Submitted

[P.3] *General* M*-Estimator Processes and their m out of n Bootstrap with Functional Nuisance Parameters*. Joint work with S. Bouzebda. Under revision

[P.4] *Central limit theorems for functional* Z*-estimators with Functional Nuisance Parameters*. Joint work with S. Bouzebda. Submitted

[P.5] *Uniform in bandwidth consistency of conditional* U*-statistics adaptive to intrinsic dimension in presence of censored data*. Joint work with S. Bouzebda and T. El-hadjali. 45 pp. Submitted.

Conferences

[C.1] *Asymptotic Properties of Semiparametric* M*-Estimators with Multiple Change Points.* 14th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2021), London, UK, 18-20th of December 2021.

Résumé de la thèse

Dans cette thèse, nous nous intéressons principalement aux modèles semiparamétriques qui ont reçu beaucoup d'intérêt par leur excellente utilité scientifique et leur complexité théorique intrigante. Dans la première partie, nous considérons le problème de l'estimation d'un paramètre $\boldsymbol{\theta}$, dans un espace de Banach, en maximisant une fonction critère qui dépend d'un paramètre de nuisance inconnu h, éventuellement de dimension infinie. Nous montrons que le bootstrap m out of n, dans ce cadre général, est consistant sous des conditions similaires à celles requises pour la convergence faible des M-estimateurs non-réguliers. Dans ce cadre délicat, des techniques avancées seront nécessaires pour faire face aux estimateurs du paramètre de nuisance à l'intérieur des fonctions critères non régulières. Nous étudions ensuite le bootstrap échangeable pour les Z-estimateurs. L'ingrédient principal est l'utilisation originale d'une identité différentielle qui s'applique lorsque la fonction critère aléatoire est linéaire en termes de mesure empirique. Un grand nombre de schémas de rééchantillonnage bootstrap apparaissent comme des cas particuliers de notre étude. Des exemples d'applications de la littérature sont présentés pour illustrer la généralité et l'utilité de nos résultats. La deuxième partie est consacrée aux modèles statistiques semiparamétriques de ruptures multiples. L'objectif principal de cette partie est d'étudier les propriétés asymptotiques des M-estimateurs semiparamétriques avec des fonctions critères non lisses des paramètres d'un modèle de rupture multiples pour une classe générale de modèles dans lesquels la forme de la distribution peut changer de segment en segment et dans lesquels, éventuellement, il y a des paramètres communs à tous les segments. La consistance des M-estimateurs semi-paramétriques des points de rupture est établie et la vitesse de convergence est déterminée. La normalité asymptotique des M-estimateurs semiparamétriques des paramètres est établie sous des conditions générales. Nous étendons enfin notre étude au cadre des données censurées. Nous étudions les performances de nos méthodologies pour des petits échantillons à travers des études de simulations.

Mots-clés: Processus empirique; M-estimateur; Z-estimateur; Classification; Données censurées; Données manquantes; Convergence faible; Entropie métrique; Échangeable; Rééchantillonnage; Point de ruptures.

Thesis abstract

In this dissertation we are concerned with semiparametric models. These models have success and impact in mathematical statistics due to their excellent scientific utility and intriguing theoretical complexity. In the first part of the thesis, we consider the problem of the estimation of a parameter $\boldsymbol{\theta}$, in Banach spaces, maximizing some criterion function which depends on an unknown nuisance parameter h, possibly infinite-dimensional. We show that the m out of n bootstrap, in a general setting, is weakly consistent under conditions similar to those required for weak convergence of the non smooth M-estimators. In this framework, delicate mathematical derivations will be required to cope with estimators of the nuisance parameters inside non-smooth criterion functions. We then investigate an exchangeable weighted bootstrap for function-valued estimators defined as a zero point of a function-valued random criterion function. The main ingredient is the use of a differential identity that applies when the random criterion function is linear in terms of the empirical measure. A large number of bootstrap resampling schemes emerge as special cases of our settings. Examples of applications from the literature are given to illustrate the generality and the usefulness of our results. The second part of the thesis is devoted to the statistical models with multiple change-points. The main purpose of this part is to investigate the asymptotic properties of semiparametric M-estimators with non-smooth criterion functions of the parameters of multiple change-points model for a general class of models in which the form of the distribution can change from segment to segment and in which, possibly, there are parameters that are common to all segments. Consistency of the semiparametric M-estimators of the change-points is established and the rate of convergence is determined. The asymptotic normality of the semiparametric M-estimators of the parameters of the within-segment distributions is established under quite general conditions. We finally extend our study to the censored data framework. We investigate the performance of our methodologies for small samples through simulation studies.

Key word Empirical processes; M-estimator; Z-estimator; Classification; Censored data; Missing data; Weak convergence; Metric entropy; Exchangeable; Bootstrap; Change-points.

Remerciements

First and foremost, I would like to express my deep gratitude to my research supervisor **Prof. Salim Bouzebda**, for his patient guidance, enthusiastic encouragement and useful critiques of this research work. I am indebted to him for all of his efforts and devotion to make this thesis.
I would like to thank the members of the committee starting by the referees: I would like to thank **Prof. Patrice Bertail** and **Prof. Johan Segers** for devoting time to read and agreeing to evaluate my thesis, also I'm extremely grateful for their insightful and constructive comments. I'm also giving my biggest thanks to the rest of the committee **Dr. Mokhtar Z. Alaya**, **Dr. Issam Elhattab**, **Prof. Ghislaine Gayraud**, **Dr. Amor Keziou** and the president **Prof. Nikolaos Limnios** for being a part of this committee thesis.
My special thanks are extended to the staff of the lab LMAC, especially: **Dr. Faten Jelassi**, who I've worked with as a temporary teaching assistant for her encouragement and great help.
My deepest appreciation is also for the Algerian government for giving me the opportunity to prepare my thesis in France. I would like to thank them for their financial support during these past years. I am particularly grateful to all of my professors of the University of Constantine, Algeria, especially **Prof. Nadia Aib**, **Prof. Fatiha Messaci** and **Prof. Nahima Nemouchi**. I would like to gratefully acknowledge the University of Technology of Compiègne for funding me during these last months of Ph.D. research studies.
I wish furthermore to acknowledge the help provided by **Dr. Issam El Hattab**. His willingness to give his time so generously has been very much appreciated.
I would like to express an exceptional thanks to my parents, my father **Rabah Ferfache** who planted the love of mathematics into my heart and my mother **Moufida Benmerzouk** for their love, understanding, prayers and continuing support during my studies. They are the source of my power and accomplishments. Without forgetting my aunts **Nacera Ferfache**, **Zahia Ferfache** and **Feriel Benmerzouk** for their love and prayers as well.
My second exceptional thanks is to my beloved brothers: **Iheb**, **Djallel** and my brother from another mother **Mohamed Soucha** for their encouragement and always being there for me no matter what.
At last, but never least I would like to thank my wife **Rayane Dorbani** for her love and constant support. Thank you for being my muse and sounding board. But most of all, thank you for being my best friend. I owe you everything.
This thesis is in the memory of my second mother **Fatima Ezohra Ferfache**. May she rest in peace.

Anouar Abdeldjaoued Ferfache

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis we are mainly concerned with semiparametric theory. Semiparametric models are seen in a simple way as sets of probability distributions that cannot be indexed by only a Euclidean parameter, i.e., models that are indexed by an infinite dimensional parameter. Semiparametric models can vary widely in the amount of structure they impose; for example, they can range from nonparametric models for which the model consists of all possible probability distributions, to simple regression models that characterize the regression function parametrically but leave the residual error distribution unspecified. The attention to the semiparametric models is mainly motivated by the problem of misspecification of statistical models. The semiparametric approach to misspecification is to allow the functional form of some components of the model to be unrestricted. We put less restrictions on the probabilistic constraints that our data might have by allowing the space of parameters to be partly infinite dimensional. This approach is an important complement to fully nonparametric models, which may not be very useful with small amounts of data or data of large dimension.

1.1 Introduction

Statistical problems are described using probability models. That is, the data are considered as a realization of a vector of random variables $\mathrm{X}_1,\ldots,\mathrm{X}_n$, where each one of the variables X_i can be a vector of random variables itself, corresponding to the data collected on the i-th individual in a sample of n individuals chosen from some population of interest. Each of these $\mathrm{X}_i, i = 1,\ldots,n$ are measurable functions from some probability space, say $(\Omega,\mathscr{U},\mathbf{P})$ to some measurable space $(\mathscr{X},\mathscr{A})$. We assume that the observations are independent and identically distributed, i.e., we assume an i.i.d. sample. These X_i have a distribution $\mathbb{P}$ on $(\mathscr{X},\mathscr{A})$, so $(\mathscr{X},\mathscr{A},\mathbb{P})$ is a measure space. One then believes that the distribution $\mathbb{P}$ belongs to some statistical or probability model, where a model consists of a class of distributions or densities that we believe the data is generated from. The distributions in a model are identified through a set of parameters.

Recall, a statistical inference aims at learning characteristics of the population from a sample: the population characteristics are parameters and sample characteristics are statistics. The parameters are unknown then to be more informative about them we need to do estimation by

making use of these statistics. The theory of estimation is divided into two main branches; the first one is parametric estimation, which consists in estimating through parametric models, defined as follows.

Definition 1.1.0.1 *A model $\mathscr{P}$ that can be indexed by a Euclidean vector, a vector of a finite number of real values (the parameters), is called a finite-dimensional* ***parametric model****.*

For finite-dimensional parametric models, the class of distributions can be described as

$$\mathscr{P} = \left\{ \mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d \right\}. \tag{1.1.1}$$

The dimension d is some finite positive integer. By considering such a parametric model, we make a lot of assumptions about the data and then the shape of our distribution is fixed and only we consider the estimation of the d true unknown parameters to characterize the distribution of the data. Without making any assumptions on the distribution of the data we obtain the nonparametric model defined as follows.

Definition 1.1.0.2 *A model $\mathscr{P}$ containing all probability distributions on the measurable space $(\mathscr{X}, \mathscr{A})$, is called a* ***nonparametric model****.*

In this case, we do not have a finite-dimensional component of the parameter, it is fully infinite-dimensional. The estimation through this model held the second branch known as a nonparametric estimation. In many practical applications of statistics it is unreasonable to make full finite dimensional parametric assumptions on the probability distributions of the phenomena we observe. On the other hand a nonparametric model might lose "*too much*" of the structure that nevertheless is at hand. So if we want the flexibility of the nonparametric model and want to answer the questions that a parametric model allows us to ask, we will choose a model that is intermediate between them, which is well known as; semiparametric model for which the theory in this thesis is mainly developed. It is defined as follows.

Definition 1.1.0.3 *A model $\mathscr{P}$ containing probability distributions described through a parameter that contains both a finite-dimensional component and an infinite-dimensional component is referred to as a* ***semiparametric model****.*

In this case, the class of distributions is so large that the parameter indexing the model is infinite-dimensional. Thus, a semiparametric model can be seen as an infinite-dimensional model that is essentially smaller than the set of all possible distributions. By allowing the space of parameters to be infinite-dimensional, we are putting less restrictions on the probabilistic constraints that our data might have. A semiparametric model will be denoted by:

$$\mathscr{P} = \left\{ \mathbb{P}_{\boldsymbol{\theta},h} : \boldsymbol{\theta} \in \Theta, h \in \mathscr{H} \right\}. \tag{1.1.2}$$

The main interest will be typically in the finite-dimensional parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, which we call the **parameter of interest** and the infinite dimensional part h belonging to a Banach space $\mathscr{H}$ will be referred to the **nuisance parameter**.

1.1.1 Examples and motivation

We give some important examples in the study of the semiparametric model, the last two examples will be used later in the next chapters of this thesis.

Example 1.1.1.1 *(Cox model) We observe a pair* (T,Z)*, where* T *is a survival time and* Z *is a covariate. The conditional hazard of* T *given* Z *is given by* $\lambda(t \mid z) = \lambda_0(t)e^{\boldsymbol{\theta}^\top z}$*, where* λ_0 *is an unknown baseline hazard function,* $\boldsymbol{\theta}$ *is a parameter of interest that expresses the proportional difference between hazard functions, and the distribution of* Z *is unrestricted.*

Example 1.1.1.2 *(Parametric models) Let* μ *be a fixed* σ *-finite measure on a sample space* $(\mathscr{X},\mathscr{A})$*. We observe* X *with distribution* $\mathbb{P}$ *from the class* $\mathscr{P} = \{\mathbb{P}_{\boldsymbol{\theta}} \ll \mu \mid \boldsymbol{\theta} \in \Theta\}$*, where* $\boldsymbol{\theta}$ *is an open subset of* $\mathbb{R}^d$ *and the parametrization* $\boldsymbol{\theta} \mapsto \mathbb{P}_{\boldsymbol{\theta}}$ *satisfies the following. The* $\operatorname{map} \boldsymbol{\theta} \to \sqrt{\frac{d\mathbb{P}_{\boldsymbol{\theta}}}{d\mu}}$ *from* $\boldsymbol{\theta}$ *to* $L^2(\mu)$ *is Fréchet differentiable with derivative* $s(\boldsymbol{\theta}) \in \mathbb{R}^d$*. The Fisher* $d \times d$ *information matrix for* $\boldsymbol{\theta}$ *given by* $I(\boldsymbol{\theta}) = \int s(\boldsymbol{\theta})^\top s(\boldsymbol{\theta}) d\mu$ *is nonsingular. Finally, the map* $\boldsymbol{\theta} \to s_i(\boldsymbol{\theta})$ *is continuous from* $\boldsymbol{\theta}$ *to* $L^2(\mu)$ *for* $i = 1,\ldots,d$*. Then* $\mathscr{P}$ *is a (finite dimensional) regular parametric model. Such a model is of course a special case of a semiparametric model, see Chapter 2 of Bickel* et al. *[1993].*

Example 1.1.1.3 *(Copula model) We observe* $X = (X_1, X_2)$ *with two-dimensional distribution*

$$F_X(x_1, x_2) = G_{\boldsymbol{\theta}}(G_1(x_1), G_2(x_2)),$$

where $G_{\boldsymbol{\theta}}(\cdot)$ *is a bivariate distribution function known up to the parameter* $\boldsymbol{\theta}$ *and with uniform marginals. The marginal distribution functions* $G_i(\cdot)$ *can both be unknown or one can be known. The purpose of the copula model is to model the covariance structure between* X_1 *and* X_2 *by the parameter* $\boldsymbol{\theta}$ *without affecting the marginal distributions, see Klaassen and Wellner [1997].*

Example 1.1.1.4 *(Regression) Let* Z *and* ϵ *be two independent random vectors and suppose that* $Y = \mu(Z;\boldsymbol{\theta}) + \sigma(Z;\boldsymbol{\theta})\epsilon$ *for known functions* μ *and* σ*. We observe the pair* $X = (Y,Z)$*. If* ϵ *has a parametric distribution and the observed value of* Z *is treated as a constant, then this is just a classical regression model. When the distribution of* ϵ *belongs to an infinite dimensional set, such as all mean zero distributions, we obtain a semiparametric version of the regression model, for instance, see Horowitz [2012].*

Example 1.1.1.5 *(Interval censoring) At the random censoring time* C *we observe whether the "death" time* T *has occurred, i.e. we observe* $X = (C,\Delta)$ *where* Δ *is the indicator of the event* $\{T \le C\}$*. The distribution of* T *and* C *may be as in the previous example, refer to Van Der Laan and Robins [1998].*

Example 1.1.1.6 *(Frailty) Let two survival times* T_1 *and* T_2 *conditional on the random variable* (W,Z) *be independent with conditional hazards of the form* $\lambda(t \mid z) = w\lambda_0(t)e^{\beta^\top z}$*. However, the variable* W *is not observed but independently of* Z *it follows a gamma distribution with mean one and variance* $\boldsymbol{\theta}$*. Thus* W *and* $\boldsymbol{\theta}$ *model the unobserved heterogeneity and we observe* $X = (T_1, T_2, Z)$*, refer to Nielsen* et al. *[1992].*

Example 1.1.1.7 *(Missing at random) Suppose that the second coordinate of* (Y_1, Y_2) *sometimes is missing. If the conditional probability that* Y_2 *is observed depends only on* Y_1*, then we say that* Y_2 *is missing at random (MAR). The interest parameter is typically a function of the distribution of* Y*, for details see van der Vaart [1998].*

Example 1.1.1.8 *(Random censoring) We observe a survival time* T *if it occurs before an independent censoring time* C*, otherwise* C *is observed. If* Δ *is the indicator variable for observing* T*, then the observation is the pair* $X = (T \wedge C, \Delta)$*. The distribution of* T *and* C *may be completely unknown or* T *might follow the Cox model in Example 1.1.1.1 in Andersen* et al. *[1993].*

1.2 M- and Z-estimators

We now give the method that will be used to estimate the parameter of interest, say $\boldsymbol{\theta}_0$. The most important method of constructing statistical estimators is to choose the estimator maximizing some criterion functions. We shall call such estimators M-estimators (from "maximum" or "minimum"). So an M-estimator $\boldsymbol{\theta}_n$ is the approximate maximum of a data-dependent function. To be more precise, let the parameter set be a metric space (Θ, d) and let $X_1, \ldots, X_n$ be i.i.d. observations, the common type of the data-dependent function is:

$$\boldsymbol{\theta} \mapsto M_n(\boldsymbol{\theta}) \equiv \mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{m}(\boldsymbol{\theta}, X_i), \tag{1.2.1}$$

for known objective function $\mathbf{m}(\boldsymbol{\theta}, \cdot)$ on the sample space. By changing the sign of $\mathbf{m}(\boldsymbol{\theta}, \cdot)$ we get the least-squares estimators and by choosing $\mathbf{m}(\boldsymbol{\theta}, \cdot) = \log p_{\boldsymbol{\theta}}(\cdot)$, where $p_{\boldsymbol{\theta}}(\cdot)$ is the density of the observations, we get the corresponding maximum likelihood estimator, these two estimators are the most important examples included by this method, but there are many examples as well. In many situations estimators that maximize a certain map also solve a system of equations, to see this; if the objective function $\mathbf{m}(\boldsymbol{\theta}, \cdot)$ is Férechet differentiable with respect to $\boldsymbol{\theta}$, the maximizing value of the criterion function in (1.2.1) is sought by setting the derivative equal to zero. We shall refer to such kind of estimators as; Z-estimators (from "zero") i.e., estimators that satisfies:

$$\psi_n(\boldsymbol{\theta}) \equiv \mathbb{P}_n \mathbf{m}'_{\boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{m}'(\boldsymbol{\theta}, X_i) = 0, \tag{1.2.2}$$

where $\mathbf{m}'(\boldsymbol{\theta}, X_i) = \frac{\partial \mathbf{m}(\boldsymbol{\theta}, X_i)}{\partial \boldsymbol{\theta}}$. Note that generally in the literature the name M-estimator is also used for what we call Z-estimator and the distinction between the different types of estimators is not always made.

Sometimes the maximum of the criterion function M_n is not taken or the estimating equation ψ_n does not have an exact solution. Then it is natural to use as an estimator a value that almost maximizes the criterion function or is near zero. This yields approximate M-estimators or Z-estimators. Estimators that are sufficiently close to being a point of maximum or a zero often have the same asymptotic behavior.

This class of estimators was first introduced by Huber [1967] for the study of the robustness of has received an important part of the development of modern robust statistics. Huber [1967] and Serfling [1980] studied their asymptotic properties in parametric models. Pakes and Pollard [1989] extended these results by using the modern empirical process theory, which we suggest as in the book of van der Vaart [1998]. Theorems 1.2.0.1 and 1.2.0.2 show the consistency of M− and Z−estimators while Theorem 1.2.0.3 shows their asymptotic normality.
Clearly, the "asymptotic value" of $\hat{\boldsymbol{\theta}}_n$ depends on the asymptotic behavior of the functions M_n and ψ_n. Under suitable normalization there typically exists a deterministic "asymptotic criterion function" $\boldsymbol{\theta} \mapsto \mathrm{M}(\boldsymbol{\theta})$ and $\boldsymbol{\theta} \mapsto \psi(\boldsymbol{\theta})$, which in general they have the following expression;

$$\mathrm{M}(\boldsymbol{\theta}) = \mathbb{P}\mathbf{m}_{\boldsymbol{\theta}} = \int \mathbf{m}_{\boldsymbol{\theta}} d\mathbb{P}, \tag{1.2.3}$$

$$\psi(\boldsymbol{\theta}) = \mathbb{P}\mathbf{m}'_{\boldsymbol{\theta}} = \int \mathbf{m}'_{\boldsymbol{\theta}} d\mathbb{P}. \tag{1.2.4}$$

It seems reasonable to expect that the maximizer (zero point) $\hat{\boldsymbol{\theta}}_n$ of M_n (ψ_n) respectively, converges to the maximizing (zero) value $\boldsymbol{\theta}_0$ of M (ψ) respectively, this is proved in the following theorems, and we say that $\hat{\boldsymbol{\theta}}_n$ is (asymptotically) consistent for $\boldsymbol{\theta}_0$.

Theorem 1.2.0.1 *Let* M_n *be random functions and let* M *be a fixed function of* $\boldsymbol{\theta}$ *such that for every* $\varepsilon > 0$

$$\sup_{\boldsymbol{\theta}\in\Theta} |\mathrm{M}_n(\boldsymbol{\theta}) - \mathrm{M}(\boldsymbol{\theta})| \xrightarrow{\mathbb{P}} 0 \tag{1.2.5}$$

$$\sup_{\boldsymbol{\theta}: d(\boldsymbol{\theta},\boldsymbol{\theta}_0)\geq\varepsilon} \mathrm{M}(\boldsymbol{\theta}) < \mathrm{M}(\boldsymbol{\theta}_0) \tag{1.2.6}$$

Then any sequence of estimators $\hat{\boldsymbol{\theta}}_n$ *with* $\mathrm{M}_n\left(\hat{\boldsymbol{\theta}}_n\right) \geq \mathrm{M}_n(\boldsymbol{\theta}_0) - o_{\mathbb{P}}(1)$ *converges in probability to* $\boldsymbol{\theta}_0$.

Theorem 1.2.0.2 *Let* ψ_n *be random vector-valued functions and let* ψ *be a fixed vector valued function of* $\boldsymbol{\theta}$ *such that for every* $\varepsilon > 0$

$$\sup_{\boldsymbol{\theta}\in\Theta} \left\|\psi_n(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta})\right\| \xrightarrow{\mathbb{P}} 0 \tag{1.2.7}$$

$$\inf_{\boldsymbol{\theta}: d(\boldsymbol{\theta},\boldsymbol{\theta}_0)\geq\varepsilon} \|\psi(\boldsymbol{\theta})\| > 0 = \left\|\psi(\boldsymbol{\theta}_0)\right\| \tag{1.2.8}$$

Then any sequence of estimators $\hat{\boldsymbol{\theta}}_n$ *such that* $\psi_n\left(\hat{\boldsymbol{\theta}}_n\right) = o_{\mathbb{P}}(1)$ *converges in probability to* $\boldsymbol{\theta}_0$.

The natural step after showing the convergence of these estimators concerns the order at which the discrepancy $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}$ converges to zero which depends on the specific situation, but for estimators based on n replications of an experiment the order is often $n^{-1/2}$. Then multiplication with the inverse of this rate creates a proper balance, and the sequence $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)$ converges in distribution, most often a normal distribution. These statements are characterized in the following theorem as given in van der Vaart [1998]. Here we can use a characterization of M-estimators either by maximization or by solving estimating equations.

Theorem 1.2.0.3 *For each $\boldsymbol{\theta}$ in an open subset of Euclidean space, let $x \mapsto \mathbf{m}'_{\boldsymbol{\theta}}(x)$ be a measurable vector-valued function such that, for every $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in a neighborhood of $\boldsymbol{\theta}_0$ and a measurable function $\dot{m}$ with $\mathbb{P}\dot{m}^2 < \infty$*

$$\left\|\mathbf{m}'_{\boldsymbol{\theta}_1}(x) - \mathbf{m}'_{\boldsymbol{\theta}_2}(x)\right\| \leq \dot{m}(x) d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2).$$

Assume that $\mathbb{P}\left\|\mathbf{m}'_{\boldsymbol{\theta}_0}\right\|^2 < \infty$ and that the map $\boldsymbol{\theta} \mapsto \mathbb{P}\mathbf{m}'_{\boldsymbol{\theta}}$ is differentiable at $\boldsymbol{\theta}_0$, with nonsingular derivative matrix $\mathrm{V}_{\boldsymbol{\theta}_0}$. If $\psi_n(\hat{\boldsymbol{\theta}}_n) = \mathbb{P}_n \mathbf{m}'_{\hat{\boldsymbol{\theta}}_n} = o_{\mathbb{P}}\left(n^{-1/2}\right)$, and $\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0$, then

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) = -\mathrm{V}_{\boldsymbol{\theta}_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{m}'_{\boldsymbol{\theta}_0}(\mathrm{X}_i) + o_{\mathbb{P}}(1).$$

In particular, the sequence $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$ is asymptotically normal with mean zero and covariance matrix $\mathrm{V}_{\boldsymbol{\theta}_0}^{-1} \mathrm{P}\mathbf{m}'_{\boldsymbol{\theta}_0} \mathbf{m}'_{\boldsymbol{\theta}_0}{}^{\top} \left(\mathrm{V}_{\boldsymbol{\theta}_0}^{-1}\right)^{\top}$.

In some important cases the objective function $\mathbf{m}(\boldsymbol{\theta}, \cdot)$ is differentiable which yields the rate of convergence to be $\sqrt{n}$ and M-estimators can treated as Z-estimators. There are other interesting situations where the objective function is not smooth with respect to the parameter of interest and then we can't see M-estimators as Z-estimators which implies, in most cases, that the $\sqrt{n}$-consistency replaced by some less rate r_n. The study of this kind of rates was introduced firstly by Chernoff [1964] in the study of the median where he showed that $\hat{\boldsymbol{\theta}}_n$ defined as the maximizer of the function

$$\boldsymbol{\theta} \mapsto \mathbb{P}_n[\boldsymbol{\theta} - 1, \boldsymbol{\theta} + 1],$$

is the M-estimators of the true value of the median $\boldsymbol{\theta}_0$ defined as the maximizer of

$$\boldsymbol{\theta} \mapsto \mathbb{P}[\boldsymbol{\theta} - 1, \boldsymbol{\theta} + 1],$$

converge with rate $n^{1/3}$, the interested reader can check the example 3.2.13 of van der Vaart and Wellner [1996] for more detail. Kim and Pollard [1990] extended this result to more general cases. Before giving theorems for the study of these cases we begin by introducing some notations which will be used after and their definitions are given in chapter 2. Let $\mathscr{F}_\delta$ be a class of functions defined as $\mathscr{F}_\delta = \left\{\mathbf{m}_{\boldsymbol{\theta}} - \mathbf{m}_{\boldsymbol{\theta}_0} : d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq \delta\right\}$ with envelope function F_δ, we require the local bracketing entropy integral to be finite

$$\int_0^\infty \sup_{\delta < \delta_0} \sqrt{\log \mathrm{N}_{[\,]}\left(\epsilon \left\|\mathrm{F}_\delta\right\|_2, \mathscr{F}_\delta, \mathrm{L}_2(\mathbb{P})\right)}\, d\epsilon < \infty. \tag{1.2.9}$$

This condition is needed in order to make all the uniform Lindeberg central limit theorems involved in Theorem 1.2.0.5 work. The next theorem shows the rate of convergence while Theorem 1.2.0.5 treat the asymptotic distribution for M-estimators based on non smooth objective function, we note that the most used method here is the argmax theorem: Theorem 2.2.0.1 given in chapter 2. These theorems are the same as in van der Vaart and Wellner [1996].

Theorem 1.2.0.4 ***(Rate of convergence)***. *Let* $M_n(\boldsymbol{\theta})$ *be stochastic processes indexed by a Euclidean space* $\boldsymbol{\theta}$ *and* $M : \boldsymbol{\theta} \mapsto \mathbb{R}$ *a deterministic function, such that for every* $\boldsymbol{\theta}$ *in a neighborhood of* $\boldsymbol{\theta}_0$ *and some positive constant* $c > 0$,

$$M(\boldsymbol{\theta}) - M(\boldsymbol{\theta}_0) \leq -cd^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0).$$

Furthermore, assume that there exists a function ϕ *such that* $\delta \mapsto \phi(\delta)/\delta^\alpha$ *is decreasing for some* $\alpha < 2$ *and for every* n, *the centered process* $M_n - M$ *satisfies*

$$\mathbb{P} \sup_{d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) < \delta} |(M_n - M)(\boldsymbol{\theta}) - (M_n - M)(\boldsymbol{\theta}_0)| \leq \frac{\phi(\delta)}{\sqrt{n}}. \tag{1.2.10}$$

Let

$$r_n^2 \phi\left(\frac{1}{r_n}\right) \leq \sqrt{n}, \quad \text{for every } n.$$

If the sequence $\hat{\boldsymbol{\theta}}_n$ *satisfies* $M_n(\hat{\boldsymbol{\theta}}_n) \geq M_n(\boldsymbol{\theta}_0) - O_{\mathbb{P}}(r_n^{-2})$ *and converges in probability to* $\boldsymbol{\theta}_0$, *then* $r_n d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_{\mathbb{P}}(1)$. *If the displayed conditions are valid for every* $\boldsymbol{\theta}$ *and* δ, *then the condition that* $\hat{\boldsymbol{\theta}}_n$ *is consistent is unnecessary.*

To derive the limit distribution of $r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ using the argmax theorem, we need to establish the convergence of a multiple of the processes $\gamma \mapsto \mathbb{P}_n(\mathbf{m}_{\boldsymbol{\theta}_0 + \gamma/r_n} - \mathbf{m}_{\boldsymbol{\theta}_0})$ in $\ell^\infty(\gamma : \|\gamma\| \leq K)$ for every K, where $\ell^\infty(A)$ denote the set of bounded functions from A to $\mathbb{R}$, Theorems 2.11.22 and 2.11.23 in van der Vaart and Wellner [1996] give the conditions for their weak convergence.

Theorem 1.2.0.5 ***(Convergence in distribution)*** *For each* $\boldsymbol{\theta}$ *in an open subset of Euclidean space, let* $\mathbf{m}_{\boldsymbol{\theta}}$ *be a measurable function such that* $\boldsymbol{\theta} \mapsto M(\boldsymbol{\theta})$ *is twice continuously differentiable at a point of maximum* $\boldsymbol{\theta}_0$, *with nonsingular second-derivative matrix* V. *Let the entropy condition (1.2.9) hold. Assume that for some continuous function* ϕ, *such that* $\phi^2(\delta) \geq \mathbb{P}F_\delta^2$ *and such that* $\delta \mapsto \phi(\delta)/\delta^\alpha$ *is decreasing for some* $\alpha < 2$, *and for every* $\eta > 0$,

$$\begin{aligned}
\lim_{\delta \downarrow 0} \frac{\mathbb{P}F_\delta^2 \{F_\delta > \eta \delta^{-2} \phi^2(\delta)\}}{\phi^2(\delta)} &= 0, \\
\lim_{\varepsilon \downarrow 0} \limsup_{\delta \downarrow 0} \sup_{\|l-g\| < \varepsilon, \|l\| \vee \|g\| \leq K} \frac{\mathbb{P}\left(\mathbf{m}_{\boldsymbol{\theta}_0 + \delta g} - \mathbf{m}_{\boldsymbol{\theta}_0 + \delta l}\right)^2}{\phi^2(\delta)} &= 0, \\
\lim_{\delta \downarrow 0} \frac{\mathbb{P}\left(\mathbf{m}_{\boldsymbol{\theta}_0 + \delta g} - \mathbf{m}_{\boldsymbol{\theta}_0 + \delta l}\right)^2}{\phi^2(\delta)} &= \mathbb{E}(G(g) - G(h))^2,
\end{aligned}$$

for all K *and some zero-mean Gaussian process* G *such that* $G(g) = G(l)$ *almost surely only if* $l = g$. *Then there exists a version of* G *with bounded, uniformly continuous sample paths on compact. Define* r_n *as the solution of* $r_n^2 \phi(1/r_n) = \sqrt{n}$. *If* $\hat{\boldsymbol{\theta}}_n$ *nearly maximizes the map* $\boldsymbol{\theta} \mapsto \mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta}}$ *for every* n *and converges in probability to* $\boldsymbol{\theta}_0$, *then the sequence* $r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ *converges in distribution to the unique maximizer* $\hat{\gamma}$ *of the process* $\gamma \mapsto G(\gamma) + \frac{1}{2}\gamma' V \gamma$.

This theorem provides a good illustration of the combination of the argmax theorem as in 2.2.0.1 and the rate theorem as in Theorem 1.2.0.4 which it should not be viewed as the only approach for proving the weak convergence.

These results concern parametric models which are extended to nonparametric models by many authors. The maximum likelihood method has been applied for estimating infinitely dimensional parameters, such as cumulative distribution or hazard functions, using entropy methods Wong and Severini [1991] studied its convergence rate and its asymptotic efficiency in estimating smooth functionals of the parameter. They obtained consistency of the maximum likelihood estimator of a nonregular basic parameter at rates of the order n^c, with $0 < c < 1/2$. Gill [1989] showed the efficiency of nonparametric maximum likelihood by the von Mises method and then a revision version was given by Gill and van der Vaart [1993], van der Vaart [1995] extend these results and he showed the efficiency of infinite-dimensional M-estimators. For more details the interested reader can refer to the monographs of van der Vaart and Wellner [1996].
In the preceding paragraphs we have introduced some basic results on M-estimators for parametric and nonparametric models. As previously mentioned, we are mainly concerned with the M-estimators for semiparametric models where there is both a Euclidean parameter of interest $\boldsymbol{\theta}$ and a nuisance parameter h. Obviously, the semiparametric maximum likelihood estimators (MLE) as discussed in Bickel *et al.* [1993], van der Vaart [1998] and Kosorok [2008] are important examples of semiparametric M-estimators, where the objective function is the empirical likelihood one. However, there are numerous other examples of semiparametric M-estimators, including estimators obtained from misspecified semiparametric likelihoods, least-squares and least-absolute deviation.
Let $X_1, \ldots, X_n$ be i.i.d. observations drawn from a semiparametric model $\{\mathbb{P}_{\boldsymbol{\theta},h} : \boldsymbol{\theta} \in \Theta, h \in \mathscr{H}\}$, where Θ is an open subset of $\mathbb{R}^d$ endowed with the Euclidean norm $\|\cdot\|$ and $\mathscr{H}$ is a possibly infinite-dimensional set with a norm $\|\cdot\|_{\mathscr{H}}$. Assume that the true unknown parameter is $(\boldsymbol{\theta}_0, h_0)$. An M-estimator $(\hat{\boldsymbol{\theta}}_n, \hat{h})$ for $(\boldsymbol{\theta}_0, h_0)$ is defined as

$$\left(\hat{\boldsymbol{\theta}}_n, \hat{h}\right) = \operatorname{argmax} M_n(\boldsymbol{\theta}, h), \tag{1.2.11}$$

where

$$(\boldsymbol{\theta}, h) \mapsto M_n(\boldsymbol{\theta}, h) \equiv \mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta},h} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{m}(\boldsymbol{\theta}, h, X_i),$$

and $\mathbf{m}(\cdot,\cdot,\cdot)$ is a known, measurable function defined from $\Theta \times \mathscr{H} \times \mathscr{X}$ to $\mathbb{R}$. We assume the limit criterion function $M(\boldsymbol{\theta}, h) = \mathbb{P}\mathbf{m}_{\boldsymbol{\theta},h}$, has a unique and "well-separated" point of maximum $(\boldsymbol{\theta}_0, h_0)$, i.e., $M(\boldsymbol{\theta}_0, h_0) > \sup_{(\boldsymbol{\theta},h) \notin G} M(\boldsymbol{\theta}, h)$ for every open set G that contains $(\boldsymbol{\theta}_0, h_0)$.
Analysis of the asymptotic behavior of M-estimators can be split into three main steps: (1) establishing consistency; (2) establishing a rate of convergence; and (3) deriving the limiting distribution.
A typical scheme for studying general semiparametric M-estimators is as follows. First, consistency is established with the argmax theorem as in Theorem 2.2.0.1 or a similar method. Second, the rate of convergence for the estimators of all parameters can then be obtained from convergence rate theorem such as Theorem 1.2.0.4. We briefly discuss in the following Section 1.2.1 consistency and rate of convergence results in the semiparametric M-estimation context. The asymptotic behavior of estimators of the Euclidean parameter can be studied with Theorem 1.2.2.1 presented in Section 1.2.2 below.

1.2.1 Consistency and Rate of Convergence

The first steps are to establish consistency and rates of convergence for all parameters of the M-estimator $(\hat{\boldsymbol{\theta}}_n, \hat{h})$. General theory for these aspects is presented in Theorem 1.2.0.4 where the only parameter $\boldsymbol{\theta}$ can replaced by the joint parameter $(\boldsymbol{\theta}, h)$.
Consistency of M-estimators can be achieved by careful application of the argmax theorem, as discussed, for example, in Section 14.2 in Kosorok [2008]. Application of the argmax theorem often involves certain compactness assumptions on the parameter sets along with model identifiability. In this context, it is often sufficient to verify that the class of functions $\{\mathbf{m}_{\boldsymbol{\theta},h} : \boldsymbol{\theta} \in \Theta, h \in \mathscr{H}\}$ is $\mathbb{P}$-Glivenko-Cantelli. Such an approach is used in the proof of consistency for Example 1.2.3.1, given below in Section 1.2.3. More generally, establishing consistency can be quite difficult.
The basic tool in establishing the rate of convergence for an M-estimator is control of the modulus of continuity of the empirical criterion function using entropy integrals over the parameter sets as in (1.2.10). Entropy results in van de Geer [2000] give rate of convergence results for a large variety of models, as we will demonstrate for Examples 1.2.3.1 and 1.2.3.2.

1.2.2 Weak convergence

In this section, we develop theory for establishing $\sqrt{n}$ consistency and asymptotic normality for the Euclidean parameter $\hat{\boldsymbol{\theta}}_n$ obtained from a semiparametric objective function $\mathbf{m}(\cdot)$. In most situations in the literature and along this thesis we consider the M-estimator of $\boldsymbol{\theta}_0$ is defined by:

$$\hat{\boldsymbol{\theta}}_n = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} \mathrm{M}_n\left(\boldsymbol{\theta}, \hat{h}\right) = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \mathbf{m}\left(\boldsymbol{\theta}, \hat{h}, \mathrm{X}_i\right), \tag{1.2.12}$$

where we substitute an estimator $\hat{h}$ for the unknown nuisance parameter h. As mentioned previously, if the objective function is derivable so the M-estimator can be viewed as Z-estimator and $\boldsymbol{\theta}_0$ is then estimated by solving:

$$\psi_n(\boldsymbol{\theta}, \hat{h}) \equiv \mathbb{P}\tilde{\mathbf{m}}_{\boldsymbol{\theta},\hat{h}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{m}}\left(\boldsymbol{\theta}, \hat{h}, \mathrm{X}_i\right) = 0, \tag{1.2.13}$$

where

$$\tilde{\mathbf{m}}(\boldsymbol{\theta}, h) = \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{m}(\boldsymbol{\theta}, h) + \frac{\partial}{\partial t}\bigg|_{t=0} \mathbf{m}(\boldsymbol{\theta}, h(t)).$$

In particular, when $\mathbf{m}(\boldsymbol{\theta}, h) = \mathrm{loglik}(\boldsymbol{\theta}, h)$, (1.2.13) trivially holds and $\tilde{\mathbf{m}}(\boldsymbol{\theta}, h)$ becomes the well studied efficient score function for $\boldsymbol{\theta}$ in semiparametric models, see van der Vaart [1998]. Let $\psi(\boldsymbol{\theta}, h) = \mathbb{P}\tilde{\mathbf{m}}_{\boldsymbol{\theta},h}$ be a deterministic function, which denotes the limit of $\psi_n(\boldsymbol{\theta}, h)$ as $n \to \infty$. In some cases, estimators satisfying (1.2.13) may not exist. Hence (1.2.13) can weakened to the following "nearly-maximizing" condition:

$$\mathbb{P}_n\tilde{\mathbf{m}}\left(\hat{\boldsymbol{\theta}}_n, \hat{h}\right) = o_{\mathbb{P}}\left(n^{-1/2}\right).$$

Theorem 1.2.2.1 *Suppose that* $\hat{\boldsymbol{\theta}}_n$ *satisfying* $\psi_n\left(\hat{\boldsymbol{\theta}}_n, \hat{h}\right) = o_{\mathbb{P}}\left(n^{-1/2}\right)$ *is a consistent estimator of* $\boldsymbol{\theta}_0$ *that is the unique solution to* $\psi(\boldsymbol{\theta}, h_0) = 0$ *in* Θ*, and that* $\hat{h}$ *is an estimator of* h_0 *satisfying* $\left\|\hat{h} - h_0\right\|_{\mathscr{H}} = \mathrm{O}_{\mathbb{P}}\left(n^{-\beta}\right)$ *for some* $\beta > 0$*. Suppose that the following conditions are satisfied.*

(i) (Stochastic equicontinuity.)

$$\frac{\left|n^{1/2}\left(\psi_n - \psi\right)\left(\hat{\boldsymbol{\theta}}_n, \hat{h}\right) - n^{1/2}\left(\psi_n - \psi\right)(\boldsymbol{\theta}_0, h_0)\right|}{1 + n^{1/2}\left|\psi_n\left(\hat{\boldsymbol{\theta}}_n, \hat{h}\right)\right| + n^{1/2}\left|\psi\left(\hat{\boldsymbol{\theta}}_n, \hat{h}\right)\right|} = o_{\mathbb{P}}(1); \tag{1.2.14}$$

(ii) $n^{1/2}\psi_n(\boldsymbol{\theta}_0, h_0) = \mathrm{O}_{\mathbb{P}}(1)$;

(iii) (Smoothness.) (a) If $\beta = 1/2$*, function* $\psi(\boldsymbol{\theta}, h)$ *is Fréchet differentiable at* $(\boldsymbol{\theta}_0, h_0)$*, i.e., there exists a continuous and nonsingular* $d \times d$ *matrix* $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, h_0)$ *and a continuous linear functional* $\dot{\psi}_h(\boldsymbol{\theta}_0, h_0)$ *such that*

$$\begin{aligned}\left|\psi(\boldsymbol{\theta}, h) - \psi(\boldsymbol{\theta}_0, h_0) - \dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \dot{\psi}_h(\boldsymbol{\theta}_0, h_0)[h - h_0]\right| &= o(|\boldsymbol{\theta} - \boldsymbol{\theta}_0|) \\ &\quad + o(\|h - h_0\|);\end{aligned}$$

or (b) if $0 < \beta < 1/2$*, for some* $\alpha > 1$ *satisfying* $\alpha\beta > 1/2$ *we have*

$$\begin{aligned}\left|\psi(\boldsymbol{\theta}, h) - \psi(\boldsymbol{\theta}_0, h_0) - \dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \dot{\psi}_h(\boldsymbol{\theta}_0, h_0)[h - h_0]\right| &= o(|\boldsymbol{\theta} - \boldsymbol{\theta}_0|) \\ &\quad + \mathrm{O}\left(\|h - h_0\|^{\alpha}\right)\end{aligned};$$

(iv) $n^{1/2}\dot{\psi}_h(\boldsymbol{\theta}_0, h_0)\left[\hat{h} - h_0\right] = \mathrm{O}_{\mathbb{P}}(1)$.

Then $\hat{\boldsymbol{\theta}}_n$ *is* $n^{1/2}$*-consistent, and further we have*

$$n^{1/2}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) = \left\{-\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, h_0)\right\}^{-1} n^{1/2}\left\{\left(\psi_n - \psi\right)(\boldsymbol{\theta}_0, h_0) + \dot{\psi}_h(\boldsymbol{\theta}_0, h_0)\left[\hat{h} - h_0\right]\right\} + o_{\mathbb{P}}(1).$$

1.2.3 Examples

Example 1.2.3.1 ***Cox regression model with right censored data.*** *In the Cox regression model, the hazard function of the survival time* T *of a subject with covariate* Z *is modeled as*

$$\lambda(t \mid z) \equiv \lim_{\Delta \to 0} \frac{1}{\Delta} \mathrm{P}(t \le \mathrm{T} < t + \Delta \mid \mathrm{T} \ge t, \mathrm{Z} = z) = \lambda(t) \exp\left(\boldsymbol{\theta}^{\top} z\right),$$

where $\lambda(\cdot)$ *is an unspecified baseline hazard function and* $\boldsymbol{\theta}$ *is a regression vector. In this model, we are usually interested in* $\boldsymbol{\theta}$ *while treating the cumulative hazard function* $h(y) = \int_0^y \lambda(t)dt$ *as the nuisance parameter. The MLE for* $\boldsymbol{\theta}$ *is proven to be semiparametric efficient and widely used in applications. Here we consider the estimation* $\boldsymbol{\theta}_0$*, which corresponds to the study of the log-likelihood as the criterion function* $\mathbf{m}(\boldsymbol{\theta}, h)$*.*

With right censoring of survival time, the data observed is $\mathrm{X} = (\mathrm{Y}, \delta, \mathrm{Z})$*, where* $\mathrm{Y} = \mathrm{T} \wedge \mathrm{C}$*:* C *is a censoring time,* $\delta = 1\!\mathrm{I}\{\mathrm{T} \le \mathrm{C}\}$*, and* Z *is a regression covariate belonging to a compact set* $\mathcal{Z} \subset \mathbb{R}^d$*. We assume that* C *is independent of* T *given* Z*. The log-likelihood is obtained as*

$$\mathbf{m}(\boldsymbol{\theta}, h) = \delta \boldsymbol{\theta}^{\top} z - \exp\left(\boldsymbol{\theta}^{\top} z\right) h(y) + \delta \log h\{y\},$$

where $h\{y\} = h(y) - h(y-)$ is a point mass that denotes the jump of h at point y. The parameter space $\mathscr{H}$ is restricted to a set of non decreasing cadlag functions on the interval $[0,\tau]$ with $h(\tau) \leq \mathfrak{C}$ for some constant $\mathfrak{C}$, see chapter 2 for the definition of cadlag functions. By some algebra, we have

$$\begin{aligned}\tilde{\mathbf{m}}(\boldsymbol{\theta}, h)(x) &= \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{m}(\boldsymbol{\theta}, h) + \frac{\partial}{\partial t}\Big|_{t=0}\mathbf{m}(\boldsymbol{\theta}, h(t)) \\ &= \left[\delta z - z\exp\left(\boldsymbol{\theta}^\top z\right)h(y)\right] \\ &\quad - \left[\delta \mathrm{H}^\dagger(\boldsymbol{\theta}, h)(y) - \exp\left(\boldsymbol{\theta}^\top z\right)\int_0^y \mathrm{H}^\dagger(\boldsymbol{\theta}, h)(u)\,dh(u)\right],\end{aligned}$$

where

$$\mathrm{H}^\dagger(\boldsymbol{\theta}, h)(y) = \frac{\mathbb{P}_{\boldsymbol{\theta},h}\mathrm{Z}\exp\left(\boldsymbol{\theta}^\top \mathrm{Z}\right)1\{\mathrm{Y} \geq y\}}{\mathbb{P}_{\boldsymbol{\theta},h}\exp\left(\boldsymbol{\theta}^\top \mathrm{Z}\right)1\{\mathrm{Y} \geq y\}}.$$

Conditions of Theorem 1.2.2.1 were verified in Cheng [2009]. The convergence rate $\beta = 1/2$ of the estimated nuisance parameter is established in Theorem 3.1 of Murphy and Van Der Vaart [1999], then $n^{1/2}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$ converges in distribution to $\mathcal{N}(0, \mathrm{V})$, where

$$\mathrm{V} = \left[\left\{(\partial/\partial\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\mathbb{P}\tilde{\mathbf{m}}_{\boldsymbol{\theta},h_0}\right\}^{-1}\right]\mathbb{P}\tilde{\mathbf{m}}_{\boldsymbol{\theta}_0,h_0}\tilde{\mathbf{m}}_{\boldsymbol{\theta}_0,h_0}^\top\left[\left\{(\partial/\partial\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\mathbb{P}\tilde{\mathbf{m}}_{\boldsymbol{\theta},h_0}\right\}^{-1}\right]^\top. \tag{1.2.15}$$

Example 1.2.3.2 ***Cox regression model with current status data.*** *Current status data arises when each subject is observed at a single examination time, Y, to determine if an event has occurred. The event time, T, cannot be known exactly. If a vector of covariates, Z, is also available, then the observed data are n i.i.d. realizations of $\mathrm{X} = (\mathrm{Y}, \delta, \mathrm{Z}) \in \mathbb{R}^+ \times \{0,1\} \times \mathbb{R}^d$, where $\delta = 1\!\!1\{\mathrm{T} \leq \mathrm{Y}\}$. The model of the conditional hazard given Z is the same as in the previous example. Throughout the remainder of the discussion, we make the following assumptions:* T *and* Y *are independent given* Z. Z *lies in a compact set almost surely and the covariance of $\mathrm{Z} - \mathbb{E}(\mathrm{Z} \mid \mathrm{Y})$ is positive definite which guarantees the efficient information to be positive definite.* Y *admits a Lebesgue density which is continuous and positive on its support $[\sigma,\tau]$, for which the true nuisance parameter h_0 satisfies $h_0(\sigma-) > 0$ and $h_0(\tau) < \mathfrak{C} < \infty$, and this density is continuously differentiable on $[\sigma,\tau]$ with derivative bounded above and bounded below by zero. Under these assumptions the maximum likelihood estimator of $(\boldsymbol{\theta}, h)$ exists, $\hat{\boldsymbol{\theta}}_n$ is asymptotically efficient and $\left\|\hat{h} - h_0\right\|_{\mathrm{L}_2} = \mathrm{O}_p\left(n^{-1/3}\right)$, where $\|\cdot\|_{\mathrm{L}_2}$ is the norm on $\mathrm{L}_2([\sigma,\tau])$. This is done by defining $d((\boldsymbol{\theta}, h), (\boldsymbol{\theta}_0, h_0)) = \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \|h - h_0\|_{\mathrm{L}_2}$, we use $(\boldsymbol{\theta}, h)$ instead of $\boldsymbol{\theta}$ and taking $\phi(\delta) = \sqrt{\delta}\left(1 + \frac{\sqrt{\delta}}{\delta^2\sqrt{n}}\right)$ in (1.2.10) under entropy conditions as in example 25.11.1 of van der Vaart [1998] we can conclude a convergence rate of $n^{1/3}$ for both $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|$ and $\|\hat{h} - h_0\|_{\mathrm{L}_2}$. The $n^{1/3}$ convergence rate is optimal for the estimation of h, as discussed in Groeneboom and Wellner [1992]. Note that the conditions on the density of* Y *ensure that $\|h - h_0\|_{\mathrm{L}_2}$ is equivalent to $\left(\int_\sigma^\tau \left(h(y) - h_0(y)\right)^2 d\mathrm{F}_\mathrm{Y}(y)\right)^{1/2}$, where $\mathrm{F}_\mathrm{Y}(y)$ is the distribution of the observation time* Y. *The corresponding criterion function, that is, the loglikelihood, is derived as*

$$\mathbf{m}(\boldsymbol{\theta}, h) = \delta\log\left[1 - \exp\left(-h(y)\exp\left(\boldsymbol{\theta}^\top z\right)\right)\right] - (1-\delta)\exp\left(\boldsymbol{\theta}^\top z\right)h(y),$$

then the score function is given, for $a \in L_2([\sigma, \tau])$*, by*

$$\begin{aligned}\tilde{\mathbf{m}}(\boldsymbol{\theta}, h) &= \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{m}(\boldsymbol{\theta}, h) + \frac{\partial}{\partial t}\bigg|_{t=0} \mathbf{m}(\boldsymbol{\theta}, h(t)) \\ &= (zh(c) - a(c))\, Q(x; \boldsymbol{\theta}, h),\end{aligned}$$

where

$$Q(x; \boldsymbol{\theta}, h) = e^{\boldsymbol{\theta}^\top z} \left[\frac{\delta}{\exp\left(e^{\boldsymbol{\theta}^\top z} h(c)\right) - 1} - (1 - \delta) \right]$$

and

$$a(c) = h(y) \frac{\mathbb{P}_{\boldsymbol{\theta},h}\left(ZQ^2(X; \boldsymbol{\theta}, h) \mid Y = y\right)}{\mathbb{P}_{\boldsymbol{\theta},h}\left(Q^2(X; \boldsymbol{\theta}, h) \mid Y = y\right)}.$$

For more detail see Cheng and Kosorok [2008]. The conditions of Theorem 1.2.2.1 were verified in the same reference then asymptotic normality of $\hat{\boldsymbol{\theta}}_n$ *holds with variance has the same form as in (1.2.15).*

It is well known that the asymptotic inferences of semiparametric models often face practical challenges. In particular, the confidence set construction and the asymptotic variance estimation of the estimator for the Euclidean parameter as in (1.2.15), both involve estimating and inverting, hard to estimate, infinite-dimensional operators. One of the most used and it's a powerful method in statistics is the bootstrap method, introduced by Efron [1979] and Efron [1982], who gave a natural way to circumvent the difficulties if its asymptotic validity is established and the feasibility of evaluating M-estimators repeatedly is compatible with available computing resources.

Theoretically, the bootstrap technique is validated for a given problem or a class of problems if we can prove its "consistency": conditional on the observed data, the bootstrap distribution has the same asymptotic behavior, either in probability or almost surely, as the sampling distribution of the original estimator, appropriately centered and scaled see for instance Gill [1989], Giné and Zinn [1990], Præstgaard and Wellner [1993] and Barbe and Bertail [1995]. Principally, we can expect that the asymptotic validity of the nonparametric bootstrap can be proved in a similar way as the asymptotic normality is proved. However, the asymptotic normality of some well-known estimators in many important models has often been explored in the context of those particular models, and it is often not easy to see how to generalize the techniques to obtain asymptotic validity of the nonparametric bootstrap due to the technicalities involved in those models. As discussed before the method of M-estimation provides a general framework that contains many of these estimators. General limit theorems for M-estimators certainly help to envision the common theme behind those examples, not to reinvent the wheel each time a new problem comes up. Note that such a generalisation provides a clean proof for the corresponding bootstrap limit theorems, given the asymptotic results now available for bootstrap empirical processes, for instance, in Giné and Zinn [1990] and Præstgaard and Wellner [1993].

The transfer from a central limit theorem for M-estimators to the corresponding bootstrap limit theorem is based on the fact that the usual stochastic equicontinuity as in (1.2.14) implies

bootstrap equicontinuity under a mild integrability condition which will be discussed in the next chapters. This proof relies on a multiplier inequality similar to that developed by Præstgaard and Wellner [1993]. This type of inequality has different versions in the literature as in Giné and Zinn [1990] and van der Vaart and Wellner [1996] all of them based on a simple formula of summation by parts. For recent reference on the subject, we refer to Cheng [2015] where the general L_p multiplier inequality is developed.

Arcones and Giné [1992] combined the work of Huber [1967] and Pakes and Pollard [1989] on M-estimators with Giné and Zinn [1990] bootstrap central limit theorem for empirical processes to obtain a.s. bootstrap limit theorems with Efron's multinomial weights for finite dimensional M-estimators. Their results make use of stronger conditions on the remainder functions than stochastic equicontinuity. These conditions would be much more difficult to verify in infinite dimensional spaces if a generalization was available. On the other hand, validity of the nonparametric bootstrap in probability is strong enough for practical purposes, which is the reason why we confine ourselves to the "in probability" versions of the bootstrap limit theorems for M-estimators in this thesis. The nonparametric bootstrap technique has been extended to estimating the posterior distribution for some statistics. The idea is to explore the new possibility brought by considering bootstrap weights other than Efron's multinomial weights. This general resampling scheme was first introduced by Rubin [1981], and extensively studied by Barbe and Bertail [1995], who suggested the name "*Weighted bootstrap*" and in Mason and Newton [1992] and Præstgaard and Wellner [1993], who showed that, for a large class of exchangeable weights, the bootstrap empirical processes are asymptotically validated both in probability and almost surely sense. Note that other version of Efron's bootstrap are also studied in Chatterjee and Bose [2005] using the term "*Generalized bootstrap*". Wellner and Zhan [1996] treated the bootstrapped version of Z-estimators, which had given by van der Vaart [1995] in a nonparametric setting, see also Kosorok [2008]. For semiparametric models, Ma and Kosorok [2005] obtained some theoretical results in the case that the bootstrap weights are assumed to be i.i.d. Dixon *et al.* [2005] studied the piggyback bootstrap which is invented solely to draw inferences for the functional parameter h when it is $\sqrt{n}$-consistence. Then Cheng and Huang [2010] gave the general theory for the "*Weighted bootstrap*", and they used examples given in section 1.2.3 to illustrate the applications of their results. As noted before if the objective function is not smooth, the rate of convergence will not be $\sqrt{n}$ and the bootstrap theory is destroyed, this is well known from Knight [1998]'s works. He showed that for the one dimensional median, the "usual" bootstrap is not consistent in non regular situations while the m out of n bootstrap is consistent. This result was generalised to the m out of n bootstrap for nonstandard M-estimators by Bose and Chatterjee [2001]. Lee and Pun [2006] gave the result in the presence of nuisance parameters in parametric models. Then Lee [2010] proved the consistency of such kind of bootstrap in a nonparametric setting. The latest results in this kind of M-estimators are considered by Delsol and Van Keilegom [2020] in semiparametric framework as in (1.2.12), where they showed that under general conditions $r_n\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$ is asymptotically normal. In the last mentioned reference no bootstrap result is investigated.

1.3 Change-points problems

There are many fields of applications where the parameter of interest $\boldsymbol{\theta}$ may change from segment to segment. To be more precise, let us assume that we have a sequence of independent variable $X_1,\ldots,X_n$ and there exist unknown point $n_1,\ldots,n_k$; $0 = n_0 < n_1 < \cdots < n_k < n_{k+1} = n$, such that, for each $j = 1,2,\ldots,k+1$; $X_{n_{j-1}+1},\ldots,X_{n_j}$ are identically distributed with a distribution that depends on j; $F_{n_j}(\cdot)$, this is the well known change-points model. The study of the change-points problem was originally stated by Page [1954], Page [1955] and Page [1957] who first proposed a procedure to detect only one change in a parameter. These models are used in a wide variety of fields, including financial modeling Talih and Hengartner [2005], bioinformatics Muggeo and Adelfio [2011], signal processing Kim *et al.* [2009], climatology Reeves *et al.* [2007], and medical imaging Nam *et al.* [2012]. Many further examples are provided in the monographs Chen and Gupta [2000] and Csörgő and Horváth [1997]. These specific applications may be concerned with changes in the mean, variance, correlation, regression coefficients, or other measures. The parameter of interest $\boldsymbol{\theta}$ can change as in these two cases.

1. Change in mean: the mean of X_i is given by

$$\boldsymbol{\theta}_i = \begin{cases} \boldsymbol{\theta}_1, & \text{if} \quad 1 \le i \le n_1, \\ \boldsymbol{\theta}_2, & \text{if} \quad n_1 + 1 \le i \le n_2, \\ \quad \cdot & \\ \quad \cdot & \\ \quad \cdot & \\ \boldsymbol{\theta}_{k+1}, & \text{if} \quad n_k + 1 \le i \le n, \end{cases}$$

   where $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \neq \cdots \neq \boldsymbol{\theta}_{k+1}$ and the discrete unknown parameter n_i indicates the location of the change-points in the sample.

2. Change in variance: the variance of X_i is given by

$$\boldsymbol{\theta}_i = \begin{cases} \boldsymbol{\theta}_1, & \text{if} \quad 1 \le i \le n_1, \\ \boldsymbol{\theta}_2, & \text{if} \quad n_1 + 1 \le i \le n_2, \\ \quad \cdot & \\ \quad \cdot & \\ \quad \cdot & \\ \boldsymbol{\theta}_{k+1}, & \text{if} \quad n_k + 1 \le i \le n, \end{cases}$$

   where $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \neq \cdots \neq \boldsymbol{\theta}_{k+1}$ and the discrete unknown parameter n_i indicates the location of the change-points in the sample.

Figure 1.1 below illustrates changes in each of these properties on two separate plots.

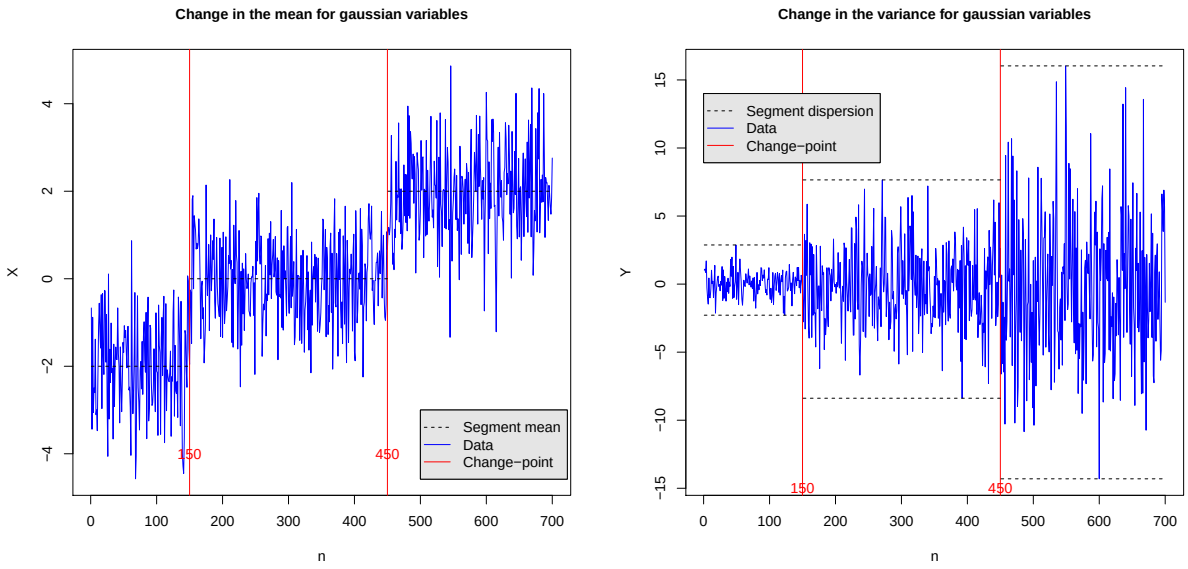


Figure 1.1: Change-points in means **(left)** and variances **(right)** of data generated by samples from a Gaussian distribution.

It is important to study the asymptotic behaviour of a change-point estimator, which includes its consistency, its convergence rate as well as its asymptotic distribution.

Over the years, considerable attention has been devoted to testing and estimation about the change-points. We list methods that could be used in change-point detection tests in literature; Least-square tests, Bayesian analysis tests, maximum likelihood ratio tests, and nonparametric tests are the most widely used among them.

For a single change-point, as in Page [1957], it is assumed that the samples were generated from the same distribution but with different parameters. The estimated location of change-point is the one that maximizes the likelihood function of the hypothesis and the author firstly introduced the CUSUM algorithm in the change-point detection problem. Fisher [1958] is the first to apply the least-squares criterion for a change-point problem to the best of our knowledge. Note that his approach does not come from likelihood maximization but rather from variance minimization. Chernoff and Zacks [1964] estimated the current mean of a normal distribution which was subjected to changes in time. Hinkley [1970] considered the likelihood-based inference to obtain the asymptotic distribution of the maximum likelihood estimator of the change-point under the assumption that the other parameters in the model are known. Hinkley [1972] argued that this asymptotic distribution is also valid when the parameters are unknown. Tang and Gupta [1987] extended the likelihood based approach to the model with a change in variance within normally distributed observations. Yao and Au [1989] proved that the estimated change-point is consistent in probability under mild assumptions, namely the continuity of the cumulative distribution function of the observations and a moment hypothesis. These assumptions are weakened further in Bai and Perron [1998] and the minimax convergence rate of $1/n$

is obtained, here n is the sample size. The least-squares estimation procedure was also shown to be consistent in the case of dependent processes (ARMA) with a single change-point in Bai [1994]. This work extended for weak dependent processes (mixingales) by Bai and Perron [1998]. The technique of using Bayesian inference was applied as a technical device leading to simple robust procedures. A quadratic loss function was used to derive a Bayesian estimator of the current mean for a priori probability distribution on the entire real line, for instance see Chen and Gupta [1997]. Chen [1998] studied the problem of change in the regression coefficients of a linear regression model. Horváth and Rice [2014] studied the change-point problem in the mean of a normal distribution. Dong *et al.* [2015] studied the change-point in the variance of measurement error and explored its convergence rate.

While in multiple change-points; Chen and Gupta [1997] explored testing and locating multiple variance change-points in a sequence of independent Gaussian random variables, assuming known and common mean. Lavielle [1999]; Lavielle and Ludeña [2000] showed the consistency of the least-squares estimate when the number of change-points is known for a large class of dependent processes. He and Severini [2010] showed the rate of convergence of the maximum likelihood estimator for the change-points under a compactness hypothesis and technical assumptions on the behavior of the log-likelihood function assuming the number of changes is known. In the same vein, by using the nonparametric theory of U-statistics Döring [2011] proved the convergence in distribution for the multiple change-points estimators. Hušková and Meintanis [2006] considered a test statistic based on empirical characteristic function, and investigated the probability of type I error and the power of the test by some simulation studies, for the change in distribution. Zou *et al.* [2014] proposed a nonparametric maximum likelihood approach to detect multiple change-points without any parametric assumption on the underlying distributions of the dataset, when the number of changes is unknown. Thus, it is suitable for detection of any change in the distributions.

1.4 Organization of the dissertation

Chapter 2. Mathematical background

This chapter is devoted to the preliminary results for a few specific topics which we will need to be self-contained and better understand the forthcoming chapters. We also review some of the standard facts concerning empirical processes and their weak convergence, with special attention given to the basic tools needed in the treatment of M-estimators and their bootstrap.

Chapter 3. General M-Estimator Processes and their m out of n Bootstrap with Functional Nuisance Parameters

Let us consider the problem of the estimation of some parameter of interest parameter $\boldsymbol{\theta}$, by maximizing some criterion function as follows

$$\hat{\boldsymbol{\theta}}_n = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} \mathrm{M}_n\left(\boldsymbol{\theta}, \hat{h}\right) = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \mathbf{m}\left(\boldsymbol{\theta}, \hat{h}, \mathrm{X}_i\right),$$

where we substitute an estimator $\hat{h}$ for the unknown nuisance parameter h, which belongs to some infinite-dimensional space. Classical estimation methods are mainly based on maximizing the corresponding empirical criterion by substituting the nuisance parameter by some nonparametric estimator. In the context of non smooth objective function Delsol and Van Keilegom [2020] studied the asymptotic properties of the M-estimator of the parameter $\boldsymbol{\theta}$ and they showed that it converges weakly to a maximizers of Gaussian processes in the case of Euclidean parameter with rate slower than $\sqrt{n}$, which is known in this kind of situations. It's well known also that the conventional bootstrap method fails in general to consistently estimate the limit law of this M-estimator. In this chapter; firstly, we extend the work of Delsol and Van Keilegom [2020] by proving the weak convergence of the estimator of the parameter of interest $\boldsymbol{\theta}$, which we suppose it belongs to some Banach space under general conditions by using closed bounded subset instead of compact subset, which is given in the part (i) of Theorem 3.3.3.3. Then we show that the m out of n bootstrap, in this general setting, is weakly consistent under conditions similar to those required for weak convergence of the M-estimators extending of the work of Lee [2010] to semiparametric framework this is in part (ii) of Theorem 3.3.3.3. We do this by first establishing abstract results on the empirical processes in Theorem 3.3.3.2. Non trivial Examples of applications from the literature are given to illustrate the generality and the usefulness of our results. To be more precise, we have considered in detail the single index model with monotone link function in Section 3.4.1, the classification with missing data in Section 3.4.2 and the binary choice model with missing data in Section 3.4.3. Finally, we investigate the performance of the methodology for small samples through a simulation study for the model described in Section 3.4.2. In our simulation we were faced with the delicate problem of the choice of the bootstrap sample size, we refer to Remark 3.3.3.6.

Chapter 4. Central limit theorems for functional Z-estimators with Functional Nuisance Parameters

In this chapter we study ways of bootstrapping the Z-estimators with bootstrap weights different from the multinomial ones which yield the ordinary (or Efron's) bootstrap. More specifically, we consider an exchangeably weighted bootstrap for function-valued estimators defined as a zero point of a function-valued random criterion function. We suppose that the bootstrap weights $\mathbf{W} = \{\mathrm{W}_{ni}, i = 1,2,\ldots,n, n = 1,2,\ldots\}$ are a triangular array defined on the probability space $(\mathcal{Z},\mathcal{E},\widehat{\mathbb{P}})$. Let $\mathrm{W}_n \equiv (\mathrm{W}_{n1},\ldots,\mathrm{W}_{nn})$ be an exchangeable vector of nonnegative weights

which sum to n. Then the exchangeably weighted bootstrap empirical measure is defined by

$$\widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{W}_{ni}\delta_{\mathrm{X}_i}.$$

The bootstrap scores are defined by

$$\psi\left(\boldsymbol{\theta}, \widehat{\mathbb{P}}_n\right)(h) = \widehat{\mathbb{P}}_n \mathrm{B}(\boldsymbol{\theta})(h) \quad \text{for} \quad h \in \mathscr{H}.$$

A bootstrap asymptotic Z-estimator $\widehat{\boldsymbol{\theta}}_n^*$ makes the bootstrap scores or estimating equations $\psi\left(\boldsymbol{\theta}, \widehat{\mathbb{P}}_n\right)$ approximately zero (in probability), i.e.,

$$\left\|\psi\left(\hat{\boldsymbol{\theta}}_n^*, \widehat{\mathbb{P}}_n\right)\right\|_{\mathscr{H}} = o_{\mathbf{P}^*}\left(n^{-1/2}\right),$$

where $\mathbf{P} \equiv \mathbb{P}^{\infty} \times \widehat{\mathbb{P}}$. A large number of bootstrap resampling schemes emerge as special cases of our settings. The main ingredient is the use of a differential identity that applies when the random criterion function is linear in terms of the empirical measure, given in (4.2.14). We have extended this identity to the semiparametric case, which is of independent interest. Our results presented in Theorem 4.2.2.8 are general and do not require linearity of the statistical model in the unknown parameter. The bootstrap limit theorem based on the linearity identity allows the validity of bootstrap to be established with respect to a possibly different norm not equivalent to the one under which the consistency is established. Then we apply these results to justify the bootstrap validity of drawing nonparametric inferences in three complex examples; random right censoring, a simplified frailty model and the double censoring model of nonparametric models. We also consider the semiparametric models and we extend the work of Zhan [2002] to a more delicate framework. The theoretical results established in this chapter, are (or will be) key tools for many further developments in the parametric and the semiparametric models.

Chapter 5. Asymptotic Properties of Semiparametric M-Estimators with Multiple change points

This chapter focuses primarily on the multiple change-points problems in the framework of semiparametric models with smooth objective function. We are interested in models when the distribution of the data is characterized by two parameters of interest, the first one can change from segment to segment and the other is common to all segments where the nuisance parameter may depend on it. Suppose that there exists a random real-valued function $\mathrm{M}_n : \boldsymbol{\Upsilon} \times \prod_{j=1}^{k+1} \Theta_j \times \mathscr{H} \longrightarrow \mathbb{R}$ depending on the data $\mathbf{X}_1, \ldots, \mathbf{X}_n$, such that $\mathrm{M}_n(\boldsymbol{\alpha}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k+1}, \boldsymbol{\lambda}, h_0)$ is an approximation of $\mathrm{M}(\boldsymbol{\alpha}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k+1}, h_0)$. In many situations, we have that

$$\mathrm{M}(\boldsymbol{\alpha}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k+1}, \boldsymbol{\lambda}, h) = \sum_{j=1}^{k+1}\left(\frac{n_j - n_{j-1}}{n}\right)\mathbb{E}[\mathbf{m}_j(\mathbf{X}_{n_j}, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h)],$$

and

$$\mathrm{M}_n(\boldsymbol{\alpha}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k+1}, \boldsymbol{\lambda}, h) = \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}+1}^{n_j} \mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h),$$

where $\mathbf{m}_j(\cdot)$ are a measurable real-valued functions for any $1 \leq j \leq k+1$ such that

$$(\boldsymbol{\alpha}^0, \boldsymbol{\theta}_1^0, \ldots, \boldsymbol{\theta}_{k+1}^0, n_1^0, \ldots, n_k^0) = \underset{0<n_1<n_2<\cdots<n; \boldsymbol{\theta}_j \in \Theta_j, 1 \leq j \leq k+1, \boldsymbol{\alpha} \in \boldsymbol{\Upsilon}}{\operatorname{argmax}} \mathrm{M}(\boldsymbol{\alpha}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k+1}, \boldsymbol{\lambda}, h_0).$$

Statistical models on this setting are used in many fields; however, the theoretical properties of semiparametric M-estimators of such models have received relatively little attention. The main purpose of this chapter is to investigate the asymptotic properties of semiparametric M-estimators with non-smooth criterion functions for a general class of models. These problems form a basically unsolved open problem in the literature. In this general framework, delicate mathematical derivations will be required to cope with estimators of the nuisance parameter inside non-smooth criterion functions, which is not the case in the standard estimation problems with smooth criterion functions. Consistency of the semiparametric M-estimators of the change-points is established and the rate of convergence is determined in Theorems 5.3.1.1 and 5.3.2.1, respectively. The asymptotic normality of the semiparametric M-estimators of the parameters of the within-segment distributions is established under quite general conditions in Theorem 5.3.3.2. These results, together with a generic paradigm for studying semiparametric M-estimators with multiple change-points, provide a valuable extension to previous related research on semiparametric maximum-likelihood estimators. Our theoretical are applied in the classification problem with missing data in the presence of multiple change-points, the details are given in Section 5.3.4. For illustration, we investigate the classification with missing data through a short simulation result.

Chapter 6. Asymptotic properties of M-estimators based on estimating equations and censored data in semi-parametric models with multiple change points

This chapter is devoted to the study of multiple change-points in the general setting of the M-(Z-)estimators where the data are right censored. Survival data in clinical trials or failure time data in reliability studies, for example, are often subject to such censoring. To be more specific, many statistical experiments result in incomplete samples, even under well-controlled conditions. For example, clinical data for surviving most types of disease are usually censored by other competing risks to life which result in death. We assume that the distributions of the random variables and the censored random variables change from segment to segment in the same time; this yields a change in the nuisance parameter which is estimated by the Kaplan-Meier estimator in this case. We assume also that the distribution has a common interest parameter for all segments. This setting is harder than in Chapter 5 due to the change of the Kaplan-Meier estimator. This situation is not studied in literature, and gives the main motivation of the work. More precisely, we estimate the unknown parameters n_j, α and $\boldsymbol{\theta}_j$, $j = 1, \ldots, k+1$ by

maximizing the estimating equations defined by:

$$\boldsymbol{\ell} \equiv \boldsymbol{\ell}(\alpha, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k+1}, n_1, \ldots, n_k) = \sum_{j=1}^{k+1} \frac{(n_j - n_{j-1})}{n} \int_{\mathbb{R}} \mathbf{m}_j(\alpha, \boldsymbol{\theta}_j, x) d\widehat{\mathrm{F}}_{n_j}(x), \tag{1.4.1}$$

where $1 - \widehat{\mathrm{F}}_{n_j}(\cdot)$ is the usual Kaplan-Meier product limit estimator of $1 - \mathrm{F}_{n_j}(\cdot)$ introduced by Kaplan and Meier [1958] and defined by

$$1 - \widehat{\mathrm{F}}_{n_j}(x) = \prod_{i=n_{j-1}+1}^{n_j} \left(1 - \frac{d_i}{n_i}\right)^{1\mathrm{I}_{\{\mathrm{Y}_{(i)} \le x\}}}, \tag{1.4.2}$$

where

$$r_i = \sum_{k=n_{j-1}+1}^{n_j} 1\mathrm{I}_{\{\mathrm{Y}_{(i)} \le \mathrm{Y}_k\}}$$

and

$$d_i = \sum_{k=n_{j-1}+1}^{n_j} 1\mathrm{I}_{\{\mathrm{Y}_{(i)} = \mathrm{Y}_k, \delta_k = 1\}},$$

denoting the number of individuals still at risk at time $\mathrm{Y}_{(i)}$ and the number of deaths at time $\mathrm{Y}_{(i)}$ respectively, and $\mathrm{Y}_{(i)}$ denotes the order statistic of $\mathrm{Y}_{n_{j-1}+1}, \ldots, \mathrm{Y}_{n_j}$ and $1\mathrm{I}_{\mathrm{E}}$ denoting the indicator function of E. For each sample $\mathrm{X}_{n_{j-1}+1}, \ldots, \mathrm{X}_{n_j}$, $j = 1, \ldots, k+1$, and $\mathbf{m}_j(\cdot,\cdot,\cdot)$ is a given measurable function from $\Upsilon \times \Theta_j \times \mathbb{R}$ to $\mathbb{R}$; where Υ and Θ_j are the parameter spaces of α and $\boldsymbol{\theta}_j$ for $j = 1, \ldots, k+1$, respectively. He and Severini [2010] showed the asymptotic properties of the maximum likelihood estimators of the change-points and the parameters of the distribution in parametric case with complete data, here we extend their results to the case of the M-estimators for semiparametric models in the presence of censored data. We investigate the asymptotic properties of M-estimators of the parameters of a multiple change-points model for a general class of models in which the form of the distribution can change from segment to segment and in which, possibly, there is a parameter that is common to all segments, in the setting of a known number of change-points. Consistency of the M-estimators of the change-points is established and the rate of convergence is determined as in Theorems 6.3.0.1 and 6.3.0.3. The asymptotic normality of the M-estimators of the parameters of the within-segment distributions is established via Theorem 6.4.0.4. Since the approaches used in the complete data models are not easily extended to multiple change-points models in the presence of censoring, where we have used some general results of Kaplan-Meier integrals. We investigate the performance of the methodology for samples through a simulation study. We have considered several scenarios to illustrate the performances of the proposed methodology, in particular we have considered the situation of 10 changes in the sample that presents hard problems for the optimization procedures.

1.5 References

Andersen, P. K., Borgan, O. r., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York. 6

Arcones, M. A. and Giné, E. (1992). On the bootstrap of M-estimators and other statistical functionals. In *Exploring the limits of bootstrap (East Lansing, MI, 1990)*, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., pages 13–47. Wiley, New York. 15

Bai, J. (1994). Least squares estimation of a shift in linear processes. *J. Time Ser. Anal.*, **15**(5), 453–472. 18

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, pages 47–78. 17, 18

Barbe, P. and Bertail, P. (1995). *The weighted bootstrap*, volume 98 of *Lecture Notes in Statistics*. Springer-Verlag, New York. 14, 15

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD. 5, 10

Bose, A. and Chatterjee, S. (2001). Generalised bootstrap in non-regular M-estimation problems. *Statist. Probab. Lett.*, **55**(3), 319–328. 15

Chatterjee, S. and Bose, A. (2005). Generalized bootstrap for estimating equations. *Ann. Statist.*, **33**(1), 414–436. 15

Chen, J. (1998). Testing for a change point in linear regression models. *Comm. Statist. Theory Methods*, **27**(10), 2481–2493. 18

Chen, J. and Gupta, A. K. (1997). Testing and locating variance changepoints with application to stock prices. *J. Amer. Statist. Assoc.*, **92**(438), 739–747. 18

Chen, J. and Gupta, A. K. (2000). *Parametric statistical change point analysis*. Birkhäuser Boston, Inc., Boston, MA. 16

Cheng, G. (2009). Semiparametric additive isotonic regression. *J. Statist. Plann. Inference*, **139**(6), 1980–1991. 13

Cheng, G. (2015). Moment consistency of the exchangeably weighted bootstrap for semiparametric M-estimation. *Scand. J. Stat.*, **42**(3), 665–684. 15

Cheng, G. and Huang, J. Z. (2010). Bootstrap consistency for general semiparametric M-estimation. *Ann. Statist.*, **38**(5), 2884–2915. 15

Cheng, G. and Kosorok, M. R. (2008). General frequentist properties of the posterior profile distribution. *Ann. Statist.*, **36**(4), 1819–1853. 14

Chernoff, H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math.*, **16**, 31–41. 8

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, **35**(3), 999–1018. 17

Csörgő, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester. With a foreword by David Kendall. 16

Delsol, L. and Van Keilegom, I. (2020). Semiparametric M-estimation with non-smooth criterion functions. *Ann. Inst. Statist. Math.*, **72**(2), 577–605. 15, 19

Dixon, J. R., Kosorok, M. R., and Lee, B. L. (2005). Functional inference in semiparametric models using the piggyback bootstrap. *Ann. Inst. Statist. Math.*, **57**(2), 255–277. 15

Dong, C., Miao, B., Tan, C., Wei, D., and Wu, Y. (2015). An estimate of a change point in variance of measurement errors and its convergence rate. *Comm. Statist. Theory Methods*, **44**(4), 790–797. 18

Döring, M. (2011). Convergence in distribution of multiple change point estimators. *J. Statist. Plann. Inference*, **141**(7), 2238–2248. 18

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**(1), 1–26. 14

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*, volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa. 14

Fisher, W. D. (1958). On grouping for maximum homogeneity. *J. Amer. Statist. Assoc.*, **53**, 789–798. 17

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method. I. *Scand. J. Statist.*, **16**(2), 97–128. With a discussion by J. A. Wellner and J. Præstgaard and a reply by the author. 10, 14

Gill, R. D. and van der Vaart, A. W. (1993). Non- and semi-parametric maximum likelihood estimators and the von Mises method. II. *Scand. J. Statist.*, **20**(4), 271–288. 10

Giné, E. and Zinn, J. (1990). Bootstrapping general empirical measures. *Ann. Probab.*, **18**(2), 851–869. 14, 15

Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*, volume 19 of *DMV Seminar*. Birkhäuser Verlag, Basel. 13

He, H. and Severini, T. A. (2010). Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli*, **16**(3), 759–779. 18, 22

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, **57**, 1–17. 17

Hinkley, D. V. (1972). Time-ordered classification. *Biometrika*, **59**, 509–523. 17

Horowitz, J. L. (2012). *Semiparametric methods in econometrics*, volume 131. Springer Science & Business Media. 5

Horváth, L. and Rice, G. (2014). Extensions of some classical methods in change point analysis. *TEST*, **23**(2), 219–255. 18

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, pages 221–233. Univ. California Press, Berkeley, Calif. 7, 15

Hušková, M. and Meintanis, S. G. (2006). Change-point analysis based on empirical characteristic functions of ranks. *Sequential Anal.*, **25**(4), 421–436. 18

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **53**, 457–481. 22

Kim, A. Y., Marzban, C., Percival, D. B., and Stuetzle, W. (2009). Using labeled data to evaluate change detectors in a multivariate streaming environment. *Signal Processing*, **89**(12), 2529–2536. 16

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.*, **18**(1), 191–219. 8

Klaassen, C. A. and Wellner, J. A. (1997). Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, pages 55–77. 5

Knight, K. (1998). Bootstrapping sample quantiles in non-regular cases. *Statist. Probab. Lett.*, **37**(3), 259–267. 15

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer Series in Statistics. Springer, New York. 10, 11, 15

Lavielle, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stochastic Process. Appl.*, **83**(1), 79–102. 18

Lavielle, M. and Ludeña, C. (2000). The multiple change-points problem for the spectral distribution. *Bernoulli*, **6**(5), 845–869. 18

Lee, S. M. S. and Pun, M. C. (2006). On m out of n bootstrapping for nonstandard M-estimation with nuisance parameters. *J. Amer. Statist. Assoc.*, **101**(475), 1185–1197. 15

Lee, T.-S. (2010). Change-point problems: bibliography and review. *J. Stat. Theory Pract.*, **4**(4), 643–662. 15, 19

Ma, S. and Kosorok, M. R. (2005). Robust semiparametric m-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, **96**(1), 190–217. 15

Mason, D. M. and Newton, M. A. (1992). A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.*, **20**(3), 1611–1624. 15

Muggeo, V. M. and Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, **27**(2), 161–166. 16

Murphy, S. A. and Van Der Vaart, A. W. (1999). Observed information in semi-parametric models. *Bernoulli*, pages 381–412. 13

Nam, C. F., Aston, J. A., and Johansen, A. M. (2012). Quantifying the uncertainty in change points. *Journal of Time Series Analysis*, **33**(5), 807–823. 16

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sørensen, T. I. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian journal of Statistics*, pages 25–43. 5

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, **41**, 100–115. 16

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, **42**, 523–527. 16

Page, E. S. (1957). On problems in which a change in a parameter occurs at an unknown point. *Biometrika*, **44**(1/2), 248–252. 16, 17

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, **57**(5), 1027–1057. 7, 15

Præstgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, **21**(4), 2053–2086. 14, 15

Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. Q. (2007). A Review and Comparison of Changepoint Detection Techniques for Climate Data. *Journal of Applied Meteorology and Climatology*, **46**(6), 900–915. 16

Rubin, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.*, **9**(1), 130–134. 15

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons, Inc., New York. Wiley Series in Probability and Mathematical Statistics. 7

Talih, M. and Hengartner, N. (2005). Structural learning with time-varying components: tracking the cross-section of financial time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(3), 321–341. 16

Tang, J. and Gupta, A. K. (1987). On testing homogeneity of variances for Gaussian models. *J. Statist. Comput. Simulation*, **27**(2), 155–173. 17

van de Geer, S. A. (2000). *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. 11

Van Der Laan, M. J. and Robins, J. M. (1998). Locally efficient estimation with current status data and time-dependent covariates. *Journal of the American Statistical Association*, **93**(442), 693–701. 5

van der Vaart, A. W. (1995). Efficiency of infinite-dimensional M-estimators. *Statist. Neerlandica*, **49**(1), 9–30. 10, 15

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. 6, 7, 10, 11, 13

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics. 8, 9, 10, 15

Wellner, J. A. and Zhan, Y. (1996). Bootstrapping Z-estimators. *University of Washington Department of Statistics Technical Report*, **308**, 5. 15

Wong, W. H. and Severini, T. A. (1991). On maximum likelihood estimation in infinite-dimensional parameter spaces. *Ann. Statist.*, **19**(2), 603–632. 10

Yao, Y.-C. and Au, S. T. (1989). Least-squares estimation of a step function. *Sankhyā Ser. A*, **51**(3), 370–381. 17

Zhan, Y. (2002). Central limit theorems for functional Z-estimators. *Statist. Sinica*, **12**(2), 609–634. 20

Zou, C., Yin, G., Feng, L., and Wang, Z. (2014). Nonparametric maximum likelihood approach to multiple change-point problems. *Ann. Statist.*, **42**(3), 970–1002. 18

Chapter 2

Mathematical background

2.1 Mathematical Background

In this chapter, we present some of the basic tools and concepts that will be used in the remainder of this thesis. In particular, we make a presentation of the most essential notions and tools concerning semiparametric theory, this being with the most important bibliographical references (for more details, we refer to Bickel *et al.* [1993], van der Vaart and Wellner [1996] van der Vaart [1998] and Kosorok [2008]). More precisely, we review important aspects of semiparametric theory and empirical processes that we need to better understand the main results. We begin by introducing Metric spaces, which are crucial since they provide the descriptive language by which the most important results about stochastic processes are derived and expressed. Outer expectations and outer integrals are crucial to the definition and use of the outer modes of convergence for quantities which are not measurable. Since many statistical quantities of interest are not measurable with respect to the uniform topology, which is often the topology of choice for applications. Linear operators and functional derivatives also play a major role in empirical process methods and are key tools for Z (M)-estimator theory.

2.1.1 Metric Spaces

We introduce some concepts and results for metric spaces. Before giving the definition of metric spaces, we briefly review the topological spaces, the σ-fields, and the measure spaces.

Definition 2.1.1.1 *A collection $\mathcal{O}$ of subsets of a set $\mathcal{X}$ is a topology in $\mathcal{X}$ if:*

(i) $\emptyset \in \mathcal{O}$ *and* $\mathcal{X} \in \mathcal{O}$, *where* $\emptyset$ *is the empty set;*

(ii) *If* $\mathrm{U}_j \in \mathcal{O}$ *for* $j = 1, \ldots, l$, *then* $\bigcap_{j=1}^{l} \mathrm{U}_j \in \mathcal{O}$;

(iii) *If* $\{\mathrm{U}_i\}_{i \in \mathrm{I}}$ *is an arbitrary collection of members of* $\mathcal{O}$ *(finite, countable or uncountable), then* $\bigcup_{i \in \mathrm{I}} \mathrm{U}_i \in \mathcal{O}$.

When $\mathscr{O}$ is a topology in $\mathscr{X}$, then $\mathscr{X}$ (or the pair $(\mathscr{X},\mathscr{O})$) is a topological space, and the members of $\mathscr{O}$ are called the open sets in $\mathscr{X}$. For a subset $A \subset \mathscr{X}$, the relative topology on A consists of the sets $\{A \cap B : B \in \mathscr{O}\}$. A set B in $\mathscr{X}$ is closed if and only if its complement in $\mathscr{X}$, denoted $\mathscr{X} - B$, is open. The closure of an arbitrary set $E \in \mathscr{X}$, denoted $\bar{E}$, is the smallest closed set containing E; while the interior of an arbitrary set $E \in \mathscr{X}$, denoted E°, is the largest open set contained in E. A subset A of a topological space $\mathscr{X}$ is dense if $\bar{A} = \mathscr{X}$. A topological space $\mathscr{X}$ is separable if it has a countable dense subset.

Definition 2.1.1.2 *A collection $\mathscr{A}$ of subsets of a set $\mathscr{X}$ is a σ-field in $\mathscr{X}$ (sometimes called a σ-algebra) if:*

(i) $X \in \mathscr{A}$;

(ii) *If* $U \in \mathscr{A}$, *then* $X - U \in \mathscr{A}$;

(iii) *The countable union* $\bigcup_{j=1}^{\infty} U_j \in \mathscr{A}$ *whenever* $U_j \in \mathscr{A}$ *for all* $j \geq 1$.

When $\mathscr{A}$ is a σ-field in $\mathscr{X}$, then $\mathscr{X}$ (or the pair $(\mathscr{X},\mathscr{A})$) is a measurable space, and the members of $\mathscr{A}$ are called the measurable sets in $\mathscr{X}$.

Definition 2.1.1.3 *For a σ-field $\mathscr{A}$ in a set $\mathscr{X}$, a map $\mu : \mathscr{A} \mapsto \mathbb{R}$ is a measure if:*

(i) $\mu(A) \in [0,\infty]$ *for all* $A \in \mathscr{A}$;

(ii) $\mu(\emptyset) = 0$;

(iii) *For a disjoint sequence* $\{A_j\} \in \mathscr{A}, \mu\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mu\left(A_j\right)$ *(countable additivity).*

If $\mathscr{X} \subset \bigcup_{i \in I} A_i$, where I is finite or countable set of indices with $\mu(A_i) < \infty$ for all $i \in I$, then μ is σ-finite. The triple $(\mathscr{X},\mathscr{A},\mu)$ is called a measure space. If $\mu(\mathscr{X}) = 1$, then μ is a probability measure. For a probability measure P on a set $\mathscr{X}$ with σ-field $\mathscr{A}$, the triple $(\mathscr{X},\mathscr{A},P)$ is called a probability space.

Definition 2.1.1.4 *1. A map $d : \mathbb{D} \times \mathbb{D} \mapsto [0,\infty)$ is a metric or distance if it satisfies;*

(i) $d(x,y) = d(y,x)$;

(ii) $d(x,z) \leq d(x,y) + d(y,z)$ *(the triangle inequality);*

(iii) $d(x,y) = 0$ *if and only if* $x = y$.

2. A metric space is a set $\mathbb{D}$ together with a metric.

If the map $d(\cdot,\cdot)$ satisfies only (i) and (ii) then it is called a semimetric or pseudometric. Technically, a metric space consists of the pair $(\mathbb{D},d)$, but usually only $\mathbb{D}$ is given and the underlying metric d is implied by the context. A map $f : \mathbb{D} \mapsto \mathbb{Y}$ between two semimetric spaces is continuous at *a* point x if and only if $f(x_n) \to f(x)$ for every sequence $x_n \to x$ and it is bounded

if; there exist a constant $M > 0$, such that; for every $x \in \mathbb{D}$, we have $d(f(x), 0) < M$; where the last metric is associate to the space $\mathbb{Y}$. Let $C_b(\mathbb{D})$ denote the set of all continuous and bounded functions $f : \mathbb{D} \mapsto \mathbb{R}$, this set plays an important role in the weak convergence on the metric space $\mathbb{D}$ as we will see in the forthcoming sections.

Definition 2.1.1.5 *A subset* K *is totally bounded if and only if for every* $r > 0$, K *can be covered by finitely many open* r*-balls.*

A very important example of a metric space is a normed space, which are defined below.

Definition 2.1.1.6 *1. A map* $\|\cdot\| : \mathbb{D} \mapsto [0, \infty)$ *is a norm if the following axioms are satisfy; for all* $x, y \in \mathbb{D}$ *and* $\alpha \in \mathbb{R}$

(i) $\|x + y\| \leq \|x\| + \|y\|$ *(triangle inequality);*

(ii) $\|\alpha x\| = |\alpha| \times \|x\|$*;*

(iii) $\|x\| = 0$ *if and only if* $x = 0$.

2. A normed space $\mathbb{D}$ *is a vector space (also called a linear space) equipped with a norm.*

The map $\|\cdot\|$ is a seminorm if it satisfies only (i) and (ii). Note that a normed (respectively seminormed) space is a metric (respectively semimetric) space with $d(x, y) = \|x - y\|$, for all $x, y \in \mathbb{D}$.

Definition 2.1.1.7 *A complete normed space is called a Banach space (completeness being understood with respect to the metric induced by the norm).*

Definition 2.1.1.8 *Let* $\mathbb{D}$ *be a Banach space with a norm* $\|\cdot\|$. *A real valued function* $\langle\cdot,\cdot\rangle : \mathbb{D} \times \mathbb{D} \to \mathbb{R}$ *is called an inner-product (or scalar-product) on* $\mathbb{D}$ *if it has the following properties for any* $x, y, z \in \mathbb{D}$ *and* $\alpha, \beta \in \mathbb{R}$

(i) $\langle x, x\rangle = \|x\|^2 \geq 0$ *with equality iff* $x = 0$*;*

(ii) $\langle x, y\rangle = \langle y, x\rangle$*;*

(iii) $\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$.

The inner product also allows one to talk about orthogonality: We say that " x is orthogonal to y " for $x, y \in \mathbb{D}$ if $\langle x, y\rangle = 0$ and use the notation $x \perp y$ to indicate that x is orthogonal to y.

Definition 2.1.1.9 *A Hilbert space is a Banach space with an inner product.*

A very important result for bounded linear functional in Hilbert spaces is the following:

Theorem 2.1.1.10 ***(Riesz representation theorem)*** *If* $T : \mathbb{H} \mapsto \mathbb{R}$ *is a bounded linear functional on a Hilbert space, then there exists a unique element* $h_0 \in \mathbb{H}$ *such that* $T(h) = \langle h, h_0\rangle$ *for all* $h \in \mathbb{H}$*, and, moreover,* $\|T\| = \|h_0\|$.

The definition of a normed space $\mathbb{D}$ requires that the space is a vector space (and therefore it contains all linear combinations of elements in $\mathbb{D}$). However, it is sometimes of interest to apply norms to subsets $\mathscr{K} \subset \mathbb{D}$ which may not be linear subspaces. In this setting, let $\mathrm{lin}(\mathscr{K})$ denote the linear span of $\mathscr{K}$ (all linear combinations of elements in $\mathscr{K}$), and let $\overline{\mathrm{lin}(\mathscr{K})}$ the closure of $\mathrm{lin}(\mathscr{K})$. Note that both $\mathrm{lin}(\mathscr{K})$ and $\overline{\mathrm{lin}(\mathscr{K})}$ are now vector spaces and that $\overline{\mathrm{lin}(\mathscr{K})}$ equipped with the restriction of the norm to element of $\mathscr{K}$ is also a Banach space.

We present three important examples of metric spaces, the first one is the set of bounded real functions $f : \mathrm{S} \mapsto \mathbb{R}$, where S is an arbitrary set, this space is denoted $\ell^\infty(\mathrm{S})$. The uniform norm $\|x\|_\mathrm{S} \equiv \sup_{t\in\mathrm{S}} |x(t)|$ makes $\ell^\infty(\mathrm{S})$ into a Banach space consisting exactly of all functions $z : \mathrm{S} \mapsto \mathbb{R}$ satisfying $\|z\|_\mathrm{S} < \infty$. Second one is the space $\mathrm{C}[a,b]$, where $a, b \in \overline{\mathbb{R}}$, which consists of continuous functions $z : [a,b] \mapsto \mathbb{R}$. Finally the space $\mathrm{D}[a,b]$, which is the space of cadlag functions which are right-continuous with left-hand limits (cadlag is an abbreviation for continue à droite, limite à gauche). We usually equip these spaces with the uniform norm $\|\cdot\|_{[a,b]}$ inherited from $\ell^\infty([a,b])$. Note that $\mathrm{C}[a,b] \subset \mathrm{D}[a,b] \subset \ell^\infty([a,b])$.

2.1.2 Outer Integrals

Let $(\mathscr{X}, \mathscr{A}, \mathbb{P})$ be an arbitrary probability space and $\mathrm{T} : \mathscr{X} \mapsto \overline{\mathbb{R}}$ an arbitrary map.

Definition 2.1.2.1 *The outer integral (outer expectation) of* T *with respect to* $\mathbb{P}$ *is defined as*

$$\mathbb{E}^*\mathrm{T} = \inf\{\mathbb{E}\mathrm{U} : \mathrm{U} \geq \mathrm{T}, \mathrm{U} : \mathscr{X} \mapsto \overline{\mathbb{R}} \text{ measurable and } \mathbb{E}\mathrm{U} \text{ exists}\},$$

Here, $\mathbb{E}\mathrm{U}$ is understood to exist if at least one of $\mathbb{E}\mathrm{U}^+$or $\mathbb{E}\mathrm{U}^-$is finite, where U^+ and U^- are the envelope functions of U. Analogously, the outer probability of an arbitrary subset B of $\mathscr{X}$ is

$$\mathbb{P}^*(\mathrm{B}) = \inf\{\mathbb{P}(\mathrm{A}) : \mathrm{A} \supset \mathrm{B}, \mathrm{A} \in \mathscr{A}\}.$$

Inner integral and inner probability can be defined in a similar fashion their definition should be obvious. Equivalently, they can be defined by $\mathbb{E}_*\mathrm{T} = -\mathbb{E}^*(-\mathrm{T})$ and $\mathbb{P}_*(\mathrm{B}) = 1 - \mathbb{P}^*(\mathscr{X} - \mathrm{B})$, respectively.

Lemma 2.1.2.2 *For any* $\mathrm{T} : \mathscr{X} \mapsto \overline{\mathbb{R}}$*, there exists a minimal measurable majorant* $\mathrm{T}^* : \mathscr{X} \mapsto \mathbb{R}$ *with*

(i) $\mathrm{T}^* \geq \mathrm{T}$*;*

(ii) For every measurable $\mathrm{U} : \mathscr{X} \mapsto \overline{\mathbb{R}}$ *with* $\mathrm{U} \geq \mathrm{T}$ *a.s.,* $\mathrm{T}^* \leq \mathrm{U}$ *a.s.*

For any T^* satisfying (i) and (ii), $\mathbb{E}^*\mathrm{T} = \mathbb{E}\mathrm{T}^*$, provided $\mathbb{E}\mathbb{T}^*$ exists. The last statement is true if $\mathbb{E}^*\mathrm{T} < \infty$. Note that a maximal measurable minorant is defined by $\mathrm{T}_* = -(-\mathrm{T})^*$ and satisfies the obvious relations in (i) and (ii).

2.1.3 Linear operator

Definition 2.1.3.1 *A linear operator is a map* $\mathrm{T} : \mathbb{D} \mapsto \mathbb{Y}$ *between two normed spaces with the property that* $\mathrm{T}(ax + by) = a\mathrm{T}(x) + b\mathrm{T}(y)$ *for all scalars* a, b *and any* $x, y \in \mathbb{D}$.

When the range space $\mathbb{Y}$ is $\mathbb{R}$, then T is a linear functional. When T is linear, we will often use $\mathrm{T}x$ instead of $\mathrm{T}(x)$.

Definition 2.1.3.2 *Let* $\mathbb{D}, \mathbb{Y}$ *be normed spaces and* $\mathrm{T} : \mathbb{D} \mapsto \mathbb{Y}$ *be a linear operator. The operator* T *is said to be bounded if there exists a* $\mathrm{C} > 0$ *such that*

$$\|\mathrm{T}x\|_{\mathbb{Y}} \leq \mathrm{C}\|x\|_{\mathbb{D}} \quad \text{for all } x \in \mathbb{D}.$$

The norm of the operator is defined as

$$\|\mathrm{T}\| := \sup_{x \in \mathbb{D}, x \neq 0} \frac{\|\mathrm{T}x\|}{\|x\|} = \sup_{x \in \mathbb{D}, \|x\| \leq 1} \|\mathrm{T}x\|. \tag{2.1.1}$$

Here, the norms $\|\cdot\|_{\mathbb{D}}$ and $\|\cdot\|_{\mathbb{Y}}$ are defined by the context.

Let $\mathrm{B}(\mathbb{D}, \mathbb{Y})$ be the space of all bounded linear operators $\mathrm{T} : \mathbb{D} \mapsto \mathbb{Y}$, where $\mathbb{D}$ and $\mathbb{Y}$ are normed spaces. This structure makes the space $\mathrm{B}(\mathbb{D}, \mathbb{E})$ into a normed space with norm $\|\cdot\|$ defined in (2.1.1). When $\mathbb{E}$ is a Banach space, then $\mathrm{B}(\mathbb{D}, \mathbb{E})$ is also a Banach space. When $\mathbb{D}$ is not a Banach space, T has a unique continuous extension to $\overline{\mathbb{D}}$, for instance, see Kosorok [2008].

Definition 2.1.3.3 *For any* $\mathrm{T} \in \mathrm{B}(\mathbb{D}, \mathbb{Y})$*, the null space of* T *is :*

$$\mathscr{N}(\mathrm{T}) \equiv \{x \in \mathbb{D} : \mathrm{T}x = 0\}$$

and its range space is:

$$\mathscr{R}(\mathrm{T}) \equiv \{y \in \mathbb{Y} : \mathrm{T}x = y \text{ for some } x \in \mathbb{D}\}.$$

We have the following two results for inverse operators.

Lemma 2.1.3.4 *Assume* $\mathbb{D}$ *and* $\mathbb{Y}$ *are normed spaces and that* $\mathrm{T} \in \mathrm{B}(\mathbb{D}, \mathbb{Y})$*. Then* T *has a continuous inverse* $\mathrm{T}^{-1} : \mathscr{R}(\mathrm{T}) \mapsto \mathbb{D}$ *if and only if there exists a* $c > 0$ *so that* $\|\mathrm{T}x\| \geq c\|x\|$ *for all* $x \in \mathbb{D}$.

Lemma 2.1.3.5 *Let* $\mathrm{A} = \mathrm{T} + \mathrm{K} : \mathbb{D} \mapsto \mathbb{Y}$ *be a linear operator between Banach spaces, where* T *is both continuously invertible and onto and* K *is compact. Then if* $\mathscr{N}(\mathrm{A}) = \{0\}$, A *is also continuously invertible and onto.*

Theorem 2.1.3.6 ***Banach-Steinhaus Theorem*** *Let* $\mathbb{D}$ *and* $\mathbb{Y}$ *are two Banach spaces and let* $(\mathrm{T}_n)_{n \geq 1}$ *a sequence of* $\mathrm{B}(\mathbb{D}, \mathbb{Y})$*. Then the limit* $\mathrm{T}x = \lim_{n \to \infty} \mathrm{T}_n x$ *exists for every* x *in* $\mathbb{D}$ *if and only if*

(i) the limit $\mathrm{T}x$ *exists for every* x *in a fundamental set, and*

(ii) for each x *in* X *the supremum* $\sup_n |\mathrm{T}_n x| < \infty$.

When the limit $\mathrm{T}x$ exists for each x in $\mathbb{D}$, the operator T is bounded, and

$$|\mathrm{T}| \leq \liminf_{n \to \infty} |\mathrm{T}_n| \leq \sup_n |\mathrm{T}_n| < \infty.$$

2.1.4 Differential of functions

In this section we focus on the concept of differentiation, in the following definitions the two spaces $\mathbb{D}$ and $\mathbb{Y}$ are normed spaces.

Definition 2.1.4.1 *$f : \mathbb{D} \mapsto \mathbb{Y}$ is Gâteaux differentiable at $x \in \mathbb{D}$ if*

$$\forall h \in \mathbb{D}, \exists\, \mathrm{T}_x \in \mathrm{B}(\mathbb{D}, \mathbb{Y})^{1}, \text{ such that, as } t \to 0, \left\| \frac{f(x+th) - f(x)}{t} - \mathrm{T}_x(h) \right\| \to 0.$$

The operator T_x is called the Gâteaux derivative of f at x.

Definition 2.1.4.2 *$f : \mathbb{D} \mapsto \mathbb{Y}$ is Hadamard differentiable at $x \in \mathbb{D}$ if there exist $\mathrm{T}_x \in \mathrm{B}(\mathbb{D}, \mathbb{Y})$ such that, $\forall h \in \mathbb{D}$, if $t \to 0, \|h_t - h\| \to 0$, then*

$$\left\| \frac{f(x+th_t) - f(x)}{t} - \mathrm{T}_x h \right\| \to 0.$$

The operator T_x is called the Hadamard derivative of f at x. The Hadamard differentiability is equivalent to compact differentiability, where compact differentiability satisfies

$$\sup_{h \in \mathrm{K}, x+th \in \mathbb{D}} \left\| \frac{f(x+th) - f(x)}{t} - \mathrm{T}_x h \right\| \to 0, \quad \text{as } t \to 0, \tag{2.1.2}$$

for every compact $\mathrm{K} \subset \mathbb{D}$.

Gâteaux requires the difference quotients to converge to some $\mathrm{T}_x(h)$ for each direction h; Hadamard requires a single $\mathrm{T}_x h$ that works for every direction h. It is equivalent to the convergence in the definition of Gâteaux differentiability being uniform over h in a compact subset of $\mathbb{D}$.

Definition 2.1.4.3 *$f : \mathbb{D} \mapsto \mathbb{Y}$ is Fréchet differentiable at $x \in \mathbb{D}$ if there exist $\mathrm{T}_x \in \mathrm{B}(\mathbb{D}, \mathbb{Y})$ such that, $\forall h \in \mathbb{D}$, if $\|h\| \to 0$, then*

$$\left\| \frac{f(x+th) - f(x) - \mathrm{T}_x h}{\|h\|} \right\| \to 0.$$

This can be viewed as (2.1.2) holds uniformly in h on a bounded subset of $\mathbb{D}$.

Hadamard requires the difference quotients to converge to zero for each direction, possibly with different rates for different directions; Fréchet requires the same rate for each direction. Since compact sets are bounded, Fréchet differentiability implies Hadamard differentiability. They are equivalent for $\mathbb{D} = \mathbb{R}^d$.

Gâteaux differentiability is usually not strong enough for the applications of functional derivatives needed for Z-estimators, while Fréchet differentiability will be needed for Z-estimator theory, while Hadamard differentiability is useful in the delta method.

Much of these materials and discussions are inspired by Section 6 of Kosorok [2008], where there are the proofs of all theorems presented here.

[1] Some authors drop the requirement for linearity here.

2.1.5 Weak convergence

In this section we give some notions for the weak convergence of a stochastic process. It should be noted that the weak convergence of a stochastic process is a generalization of the convergence in law from random vectors to sample paths of the stochastic process. Let $(\mathscr{X}, \mathscr{A}, \mathbb{P})$ be a probability space on which we define the sequence $X_1, \ldots, X_n$ and a collection of random variables $X = \{X(t) = X(t, \omega), \omega \in \mathscr{X}, t \in T\}$, T is an arbitrary index set. Suppose that the set T is equipped with a semi-metric ρ and $(\mathbb{D}, d)$ is a metric space.

Definition 2.1.5.1

- *The collection* $X = \{X(t) = X(t, \omega), \omega \in \mathscr{X}, t \in T\}$, *is a stochastic process.*
- *An empirical process is a stochastic process based on a random observations* $X_1, \ldots, X_n$.
- *For a fixed point* $\omega \in \mathscr{X}$, *the map:*

$$X(\cdot, \omega) : T \mapsto \mathbb{D},$$

is called the sample path of the stochastic process X.

Note that the space $\ell^\infty(T)$ is where most of the action occurs for statistical applications of empirical processes, so in next we will consider $\mathbb{D} = \ell^\infty(T)$, and for $x, y \in \mathbb{D} : d = \sup_{t \in T} |x(t) - y(t)|$ is the uniform distance on $\mathbb{D}$.

Now we say that the process X_n converges weakly to a Borel measurable process X, and we write $X_n \rightsquigarrow X$, if the sample paths of X_n behave in distribution like X when $n \to \infty$. This is reflected in

$$X_n \rightsquigarrow X \iff \forall f \in C_b(\mathbb{D}) : \mathbb{E}^* \left(f(X_n) \right) \longrightarrow \mathbb{E} \left(f(X) \right), \tag{2.1.3}$$

where

$$C_b(\mathbb{D}) := \left\{ f : \mathbb{D} \to \mathbb{R}; \text{with } f \text{ continuous and bounded} \right\}.$$

If $\mathbb{P}$ is the law of X then the last expression can be rewrite as

$$\mathbb{E}^* f(X_n) \to \int f(x) d\mathbb{P}(x), \text{ for every } f \in C_b(\mathbb{D}).$$

However in practice the latter formulation is not easy to handle. An equivalent theorem is given in Theorem 2.1 in Kosorok [2008].

Theorem 2.1.5.2 (Kosorok [2008]) *The stochastic process* X_n *converges weakly to a tight stochastic process* X *in* $\ell^\infty(T)$, *if and only if:*

(i) For all finite $\{t_1, \ldots, t_k\} \subset T$, *the finite-dimensional distribution of* $\{X_n(t_1), \ldots, X_n(t_k)\}$ *converges to that of* $\{X(t_1), \ldots X(t_k)\}$;

(ii) There exists a semi-metric ρ *for which* T *is totally bounded such that for all* $\epsilon > 0$*:*

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}^* \left\{ \sup_{s,t \in T : \rho(s,t) < \delta} |X_n(t) - X_n(s)| > \epsilon \right\} = 0. \tag{2.1.4}$$

Very useful results are the continuous mapping theorem and the Slutsky's Theorem:

Theorem 2.1.5.3 ***(Continuous mapping)*** *Let* $g : \mathbb{D} \mapsto \mathbb{Y}$ *be continuous at all points in* $\mathbb{D}_0 \subset \mathbb{D}$*, where* $\mathbb{D}$ *and* $\mathbb{Y}$ *are metric spaces. Then if* $X_n \rightsquigarrow X$ *in* $\mathbb{D}$*, with* $\mathbb{P}_* (X \in \mathbb{D}_0) = 1$*, then* $g(X_n) \rightsquigarrow g(X)$*.*

Theorem 2.1.5.4 ***(Slutsky's theorem)*** *Suppose* $X_n \rightsquigarrow X$ *and* $Y_n \rightsquigarrow c$*, where* X *is separable and* c *is a fixed constant. Then the following are true:*

(i) $(X_n, Y_n) \rightsquigarrow (X, c)$*.*

(ii) If X_n *and* Y_n *are in the same metric space, then* $X_n + Y_n \rightsquigarrow X + c$*.*

(iii) Assume in addition that the Y_n *are scalars. Then whenever* $c \in \mathbb{R}$*,* $Y_n X_n \rightsquigarrow cX$*. Also, whenever* $c \neq 0, X_n / Y_n \rightsquigarrow X / c$*.*

Generally when dealing with empirical processes the index set $T = \mathscr{F}$ is a class of measurable functions. For this, in the following section we give some definitions and examples concerning these classes.

2.1.6 Classes of functions

This section is devoted to the entropy that is a fundamental tool for the empirical process. The main use of such entropy calculus in this thesis is for establishing rate of convergence M-estimators as discussed in Chapter (1) and evaluating whether the class of functions $\mathscr{F}$ is Glivenko-Cantelli and/or Donsker or neither. There are several additional uses of entropy bounds, we refer the interested reader to the monographs of van der Vaart and Wellner [1996] and Kosorok [2008], see also Pakes and Pollard [1989].

Definition 2.1.6.1 *An envelope function of a class* $\mathscr{F}$ *is any function* $x \mapsto F(x)$ *such that*

$$|f(x)| \leq F(x),$$

for every x *and* f*.*

Definition 2.1.6.2 *A class of subsets* $\mathscr{C}$ *on a set* C *is called a VC-class if there exists a polynomial* $P(\cdot)$ *such that, for every set of* N *points in* C*, the class* $\mathscr{C}$ *picks out at most* $P(N)$ *distinct subsets.*

Definition 2.1.6.3 *The subgraph of a function* $f : \mathscr{X} \mapsto \mathbb{R}$ *is the subset of* $\mathscr{X} \times \mathbb{R}$ *given by*

$$\{(x, t) : t < f(x)\}.$$

Definition 2.1.6.4 *A class of functions $\mathscr{F}$ is called a VC-subgraph class if the collections of all subgraphs of the functions in $\mathscr{F}$ form a VC-class of sets in $\mathscr{X} \times \mathbb{R}$.*

Example 2.1.6.5 *let $\mathscr{C} = \{C \subset \mathscr{X}\}$ and $\mathscr{F}(\mathscr{C}) = \left\{\mathbb{1}_{\{X \in C\}, C \in \mathscr{C}}\right\}$. Then $\mathscr{F}(\mathscr{C})$ is a VC-subgraph class if and only if $\mathscr{C}$ is a VC class of sets.*

Definition 2.1.6.6 *A class $\mathscr{F}$ of measurable functions is $\mathbb{P}$-measurable if the map*

$$(x_1, \ldots, x_2) \mapsto \sup_{f \in \mathscr{F}} \left\| \sum_{i=1}^{n} e_i f(x_i) \right\|$$

is measurable for all $(e_1, \ldots, e_n) \in \mathbb{R}^n$.

A stronger, but easier to verify, measurability assumption is pointwise measurability defined as:

Definition 2.1.6.7 *The class $\mathscr{F}$ is pointwise measurable if there exists a countable subset $\mathscr{G} \subset \mathscr{F}$ such that for every $f \in \mathscr{F}$ there exists a sequence $\{g_l\} \in \mathscr{G}$ with $g_l(x) \to f(x)$ for every x.*

Definition 2.1.6.8 ***(Covering number).*** *Let $(\mathscr{F}, \|\cdot\|)$ be a subset of a normed space of real functions f on some set. The covering number $N(\varepsilon, \mathscr{F}, \|\cdot\|)$ is the minimal number of balls $\{g : \|g - f\| < \varepsilon\}$ of radius ε needed to cover the set $\mathscr{F}$. The entropy (without bracketing) is the logarithm of the covering number. Define*

$$J(\delta, \mathscr{F}) = \sup_{Q} \int_0^{\delta} \sqrt{1 + \log N\left(\varepsilon \|F\|_{Q,2}, \mathscr{F}, L_2(Q)\right)}\, d\varepsilon,$$

where the supremum is taken over all finitely discrete probability measures Q with $\|F\|_{Q,2} > 0$.

Definition 2.1.6.9 ***(Bracketing number).*** *Given two functions l and u, the bracket $[l, u]$ is the set of all functions f with $l \leqslant f \leqslant u$. An ε bracket is a bracket $[l, u]$ with $\|l - u\| < \varepsilon$. The bracketing number $N_{[]}(\varepsilon, \mathscr{F}, \|\cdot\|)$ is the minimum number of ε brackets needed to cover $\mathscr{F}$. The entropy with bracketing is the logarithm of the bracketing number. For a given norm $\|\cdot\|$, define a bracketing integral of a class of functions $\mathscr{F}$ as*

$$J_{[]}(\delta, \mathscr{F}, \|\cdot\|) = \int_0^{\delta} \sqrt{1 + \log N_{[]}(\varepsilon \| |F| \|, \mathscr{F}, \|\cdot\|)}\, d\varepsilon.$$

The next lemma, presents a link between the covering and the packing numbers of a functions class $\mathscr{F}$.

Lemma 2.1.6.10 *For a class of functions $\mathscr{F}$ we have:*

$$N_{[]}(2\epsilon, \mathscr{F}, d) \leq N(\epsilon, \mathscr{F}, d) \leq N_{[]}(\epsilon, \mathscr{F}, d).$$

The following lemma concerns the covering numbers of a VC- type class of functions.

Example 2.1.6.11 *The set $\mathscr{F}$ of all indicator functions $\mathbb{1}_{\{(-\infty,t]\}}$ of cells in $\mathbb{R}$ satisfies :*

$$N(\epsilon, \mathscr{F}, L_2(Q)) \leq \frac{2}{\epsilon^2},$$

for any probability measure Q and $\epsilon \leq 1$. Notice that :

$$\int_0^1 \sqrt{\log\left(\frac{1}{\epsilon}\right)} d\epsilon \leq \int_0^\infty u^{1/2} \exp(-u) du \leq 1.$$

For more details and discussion on this example refer to Example 2.5.4 of van der Vaart and Wellner [1996] and [Kosorok, 2008, p. 157]. The covering numbers of the class of cells $(-\infty, t]$ in higher dimension satisfy a similar bound, but with higher power of $(1/\epsilon)$, see Theorem 9.19 of Kosorok [2008].

Example 2.1.6.12 *(Classes of functions that are Lipschitz in a parameter, Section 2.7.4 in van der Vaart and Wellner [1996]). Let $\mathscr{F}$ be the class of functions $x \mapsto \varphi(t, x)$ that are Lipschitz in the index parameter $t \in T$. Suppose that:*

$$|\varphi(t_1, x) - \varphi(t_2, x)| \leq d(t_1, t_2)\kappa(x)$$

for some metric d on the index set T, the function $\kappa(\cdot)$ defined on the sample space $\mathscr{X}$, and all x. According to Theorem 2.7.11 of van der Vaart and Wellner [1996] and Lemma 9.18 of Kosorok [2008], it follows, for any norm $\|\cdot\|_{\mathscr{F}}$ on $\mathscr{F}$, that :

$$N(\epsilon \|F\|_{\mathscr{F}}, \mathscr{F}, \|\cdot\|_{\mathscr{F}}) \leq N(\epsilon/2, T, d).$$

Hence if (T, d) satisfy $J(\infty, T, d) = \int_0^\infty \sqrt{\log N(\epsilon, T, d)} d\epsilon < \infty$, then the conclusions holds for $\mathscr{F}$.

Example 2.1.6.13 *Let us consider as example the classes of functions that are smooth up to order α defined as follows, see Section 2 of van der Vaart and Wellner [1996]. For $0 < \alpha < \infty$ let $\lfloor \alpha \rfloor$ be the greatest integer strictly smaller than α. For any vector $k = (k_1, \ldots, k_d)$ of d integers define the differential operator :*

$$D^{k_\cdot} := \frac{\partial^{k_\cdot}}{\partial^{k_1} \cdots \partial^{k_d}},$$

where :

$$k_\cdot := \sum_{i=1}^{d} k_i.$$

Then, for a function $\varphi : \mathscr{X} \to \mathbb{R}$, let :

$$\|\varphi\|_\alpha := \max_{k_\cdot \leq \lfloor \alpha \rfloor} \sup_x |D^k \varphi(x)| + \max_{k_\cdot = \lfloor \alpha \rfloor} \sup_{x,y} \frac{D^k \varphi(x) - D^k \varphi(y)}{\|x - y\|^{\alpha - \lfloor \alpha \rfloor}},$$

where the suprema are taken over all x, y in the interior of $\mathscr{X}$ with $x \neq y$. Let $C_M^\alpha(\mathscr{X})$ be the set of all continuous functions $\varphi : \mathscr{X} \to \mathbb{R}$ with :

$$\|\varphi\|_\alpha \leq F.$$

Note that for $\alpha \leq 1$ *this class consists of bounded functions* $\varphi(\cdot)$ *that satisfy a Lipschitz condition. Kolmogorov and Tihomirov [1961] computed the entropy of the classes of* $C_F^\alpha(\mathcal{X})$ *for the uniform norm. As a consequence of their results van der Vaart and Wellner [1996] shows that there exists a constant* K *depending only on* α, d *and the diameter of* $\mathcal{X}$ *such that for every measure* γ *and every* $\epsilon > 0$ *:*

$$\log N_{[\,]}(\epsilon F_\gamma(\mathcal{X}), C_F^\alpha(\mathcal{X}), L_2(\gamma)) \leq K\left(\frac{1}{\epsilon}\right)^{d/\alpha},$$

$N_{[\,]}$ *is the bracketing number, refer to Definition 2.1.6 of van der Vaart and Wellner [1996] and we refer to Theorem 2.7.1 of van der Vaart and Wellner [1996] for a variant of the last inequality. By Lemma 9.18 of Kosorok [2008], we have :*

$$\log N(\epsilon F_\gamma(\mathcal{X}), C_F^\alpha(\mathcal{X}), L_2(\gamma)) \leq K\left(\frac{1}{\epsilon}\right)^{d/\alpha}.$$

2.2 Some useful notes for studying M-estimators

First we begin by introducing some notation needed in this thesis. Let $X_1, \ldots, X_n$ are i.i.d. $\mathbb{P}$ on $\mathcal{X}$. Then the empirical measure $\mathbb{P}_n$ is defined by

$$\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

where δ_x denotes the Dirac measure at x. For each $n \geq 1, \mathbb{P}_n$ denotes the random discrete probability measure which puts mass $1/n$ at each of the n points $X_1, \ldots, X_n$. For a real valued function f on $\mathcal{X}$, we write

$$\mathbb{P}_n(f) := \int f d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

If $\mathcal{F}$ is a class of functions defined on $\mathcal{X}$, then $\{\mathbb{P}_n(f) : f \in \mathcal{F}\}$ is the empirical measure indexed by $\mathcal{F}$. Let us assume that

$$\mathbb{P}f := \int f d\mathbb{P},$$

exists for each $f \in \mathcal{F}$. The empirical process $\mathbb{G}_n$ is defined by

$$\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - P),$$

and the collection of random variables $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ as f varies over $\mathcal{F}$ is called the empirical process indexed by $\mathcal{F}$. We define the following quantity :

$$\|\mathbb{G}_n\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)|.$$

As we discussed in Chapter 1 in the estimating of some parameter of interest $\boldsymbol{\theta}$ by the method of M-estimation we need firstly to prove its consistency and this can hold by using the argmax theorem which is given below :

Theorem 2.2.0.1 ***(Argmax Theorem)*** *Let* M_n, M *be stochastic processes indexed by a metric space* $\mathbb{D}$ *such that* $\mathrm{M}_n \rightsquigarrow \mathrm{M}$ *in* $\ell^\infty(\mathrm{K})$ *for every compact* $\mathrm{K} \subset \mathbb{D}$. *Suppose also that almost all sample paths* $\gamma \mapsto \mathrm{M}(\gamma)$ *are upper semicontinuous and possess a unique maximum at a (random) point* $\hat{\gamma}$, *which as a random map in* $\mathbb{D}$ *is tight. If the sequence* $\hat{\gamma}_n$ *is uniformly tight and satisfies*

$$\mathrm{M}_n\left(\hat{\gamma}_n\right) \geq \sup_{\gamma \in \mathbb{D}} \mathrm{M}_n(\gamma) - o_{\mathrm{P}}(1),$$

then

$$\hat{\gamma}_n \rightsquigarrow \hat{\gamma} \ \ in \ \ \mathbb{D}.$$

The most time we have $\mathrm{M}_n = \mathbb{P}_n$ and $\mathrm{M} = \mathbb{P}$ to derive the weak convergence or the convergence in probability between these quantities indexed by some class of functions $\mathscr{F}$ which is one of the main assumptions to derive the asymptotic for M-(Z)-estimators as in the argmax theorem or Theorem 1.2.0.1 we need that the class $\mathscr{F}$ to be Donsker class or as a restriction Glivenko-Cantelli class, these types of classes is defined below :

Definition 2.2.0.2 *A class* $\mathscr{F}$ *of measurable functions* $f : \mathscr{X} \to \mathbb{R}$ *with* $\mathrm{P}|f| < \infty$ *for every* $f \in \mathscr{F}$ *is called Glivenko-Cantelli (GC) if*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} := \sup_{f \in \mathscr{F}} \left|\mathbb{P}_n f - \mathbb{P} f\right| \to 0, \qquad \textit{in probability (or almost surely).}$$

Definition 2.2.0.3 *A class* $\mathscr{F}$ *of measurable functions* $f : \mathscr{X} \to \mathbb{R}$ *is Donsker if the empirical process* $\left\{\mathbb{G}_n f : f \in \mathscr{F}\right\}$ *indexed by* $\mathscr{F}$ *converges in distribution in the space* $\ell^\infty(\mathscr{F})$ *to a tight random element.*

The next step after proving consistency in the study of estimators constructed from data of size n is at which rate they converge. Generally these rates are function of the size of data n, in our setting of M-estimators in most situation the rate is $\sqrt{n}$, or some less rates r_n in non smooth cases as described before in Chapter 1, these rates can obtained from the modulus of continuity of the criterion function and its limits at the true parameter that is the main problem in this step. A simple, but not necessarily efficient, method is to apply the maximal inequalities given below:

Theorem 2.2.0.4 (van der Vaart and Wellner [1996]) ***(Entropy control with covering number)***. *Let* $\mathscr{F}$ *be a* $\mathbb{P}$*-measurable class of measurable functions with measurable envelope* F. *Then*

$$\mathrm{E}\left[\left\|\mathbb{G}_n(f)\right\|_{\mathscr{F}}^*\right] \leqslant \mathrm{KJ}(1, \mathscr{F})\|\mathrm{F}\|_{\mathbb{P},2},$$

where K *does not depend on* $\mathscr{F}$ *and* F.

Theorem 2.2.0.5 (van der Vaart and Wellner [1996]) ***(Entropy control with bracketing number)*** *Let* $\mathscr{F}$ *be a class of measurable functions with envelop* F. *Then*

$$\mathrm{E}\left[\left\|\mathbb{G}_n(f)\right\|_{\mathscr{F}}^*\right] \leqslant \mathrm{KJ}_{[]}\left(1, \mathbb{F}, \mathrm{L}_2(p)\right)\|\mathrm{F}\|_{\mathbb{P},2},$$

where K *does not depend on* $\mathscr{F}$ *or* F.

2.2.1 Bootstrapped Empirical processes

Let $\mathbb{P}_n$ be the empirical measure of an i.i.d. sample $X_1,\ldots,X_n$ from a probability measure $\mathbb{P}$. Given the sample values, let $X_1^*,\ldots,X_n^*$ be an i.i.d. sample from $\mathbb{P}_n$. The bootstrap[2] empirical measure and process are, respectively, defined by

$$\widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i^*},$$

and

$$\hat{\mathbb{G}}_n = \sqrt{n}\left(\widehat{\mathbb{P}}_n - \mathbb{P}_n\right).$$

Giné and Zinn [1990] proved the following result, for a class of function $\mathscr{F}$ with its envelope F. Here we will consider that the class $\mathscr{F}$ is the collection of indicator functions of sets of the form $[0,c]$, $0<c\leq 1$. Then under measurability restriction on $\mathscr{F}$, we have:

$$\mathbb{G}_n \rightsquigarrow \mathbb{G} \quad \text{and} \quad \mathbb{P}F^2 < \infty,$$

is equivalent to

$$\hat{\mathbb{G}}_n \rightsquigarrow \mathbb{G} \quad \text{for almost all data sequences} \quad X_1, X_2, \ldots,$$

where $\mathbb{G}$ is some tight Brownian bridge and the weak convergence is in $\ell^\infty(\mathscr{F})$. This result is proved "in probability" by the same authors and they settled questions about the validity of Efron's bootstrap in a wide range of situations. We can remark that, the bootstrap empirical measure given before can be expressed as

$$\widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i^*} = \frac{1}{n}\sum_{i=1}^{n}\xi_{ni}\delta_{X_i},$$

where ξ_{ni} is the number of times that X_i is "redrawn" from the original sample. As observed by [Efron, 1982, Section 2.9, pages 17-72], this suggests that there are, in fact, not just one way but several ways to bootstrap; and this is the idea of the exchangeable-weighted bootstrap. Let $\mathbf{W} = \{W_{ni}, i = 1,2,\ldots,n, n = 1,2,\ldots\}$ are a triangular array defined on the probability space $(\mathscr{X},\mathscr{E},\widehat{\mathbb{P}})$. Let $W_n \equiv (W_{n1},\ldots,W_{nn})$ be an exchangeable vector of nonnegative weights which sum to n. Then the exchangeably weighted bootstrap empirical measure is defined by

$$\widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n}W_{ni}\delta_{X_i},$$

with corresponding bootstrap empirical process

$$\hat{\mathbb{G}}_n = \sqrt{n}\left(\widehat{\mathbb{P}}_n - \mathbb{P}_n\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(W_{ni}-1\right)\delta_{X_i}. \tag{2.2.1}$$

The formulation of the weighted bootstrap was originally initiated by Lo [1993]. Mason and Newton [1992] and Præstgaard and Wellner [1993] established sufficient conditions on the weights $\mathbf{W}$ for the exchangeable weighted bootstrap to work asymptotically, where they suggested the following general conditions on $\mathbf{W}$

[2]The bootstrap is the statistical procedure which models sampling from a population by the process of resampling from the sample.

(B.1) The vectors $W_n = (W_{n1}, W_{n2}, \ldots, W_{nn})^T$ are exchangeable for all $n = 1, 2, \ldots$, i.e., for any permutation $\pi = (\pi_1, \ldots, \pi_n)$ of $(1, 2, \ldots, n)$, the joint distribution of

$$\pi(W_n) = \left(W_{n\pi_1}, W_{n\pi_2}, \ldots, W_{n\pi_n}\right)^T$$

is the same as that of W_n.

(B.2) $W_{ni} \geq 0$ for all n, i and $\sum_{i=1}^{n} W_{ni} = n$ for all n.

(B.3) The following $L_{2,1}$ norm of W_{n1} is uniformly bounded:

$$R_n = \int_0^\infty \sqrt{\widehat{\mathbb{P}}(W_{n1} \geq u)}\, du \leq K < \infty$$

(B.4) $\lim_{\lambda \to \infty} \limsup_{n \to \infty} \sup_{t \geq \lambda} t^2 \widehat{\mathbb{P}}\{W_{n1} \geq t\} = 0.$

(B.5) $\frac{1}{n} \sum_{i=1}^{n} (W_{ni} - 1)^2 \to c^2 > 0$ in $\widehat{\mathbb{P}}$-probability.

We note that the Efron's nonparametric bootstrap (or multinomial bootstrap) corresponds to the choice of the weights

$$W_n \sim \text{Multinomial}(n; n^{-1}, \ldots, n^{-1})$$

for which conditions **(B.1)**–**(B.5)** are satisfied. In general, in order to satisfy the conditions **(B.3)**–**(B.5)** we have to impose some moment conditions on W_{ni}, see their Lemma 3.1. The other sampling schemes that satisfy conditions **(B.1)**–**(B.5)**, include *Bayesian bootstrap, Multiplier bootstrap, Double bootstrap and Urn bootstrap*. These examples are sufficient to show that conditions **(B.1)**–**(B.5)** are very general. It is worth noticing that the value of c in **(B.5)** is independent of the sample at hand and depends only on the chosen resampling method, e.g., $c = 1$ for the nonparametric bootstrap and Bayesian bootstrap, whereas $c = \sqrt{2}$ for the double bootstrap. A more precise discussion of this general formulation of the bootstrap and further details can be found in Mason and Newton [1992], Præstgaard and Wellner [1993], Barbe and Bertail [1995], [van der Vaart and Wellner, 1996, §3.6.2., p. 353], [Kosorok, 2008, §10. p. 179], Cheng and Huang [2010]. The interested reader may refer to Billingsley [1968], Aldous [1985] and Kallenberg [2002] for excellent general coverage of the theory of exchangeability. One could claim that general first-order limit theory for the bootstrap was known to Laplace by about 1810 (since Laplace developed one of the earliest general central limit theorems); and that second-order properties were developed by Chebyshev at the end of the 19th Century, as mentioned by Peter Hall in http://www.cms.zju.edu.cn/conference/2005/zlx/peter.pdf. In 1923 Hubback began a series of crop trials, in the Indian states of Bihar and Orissa, in which he developed spatial sampling schemes. In 1927 he published an account of his work in a Bulletin No. 166 of the Indian Agricultural Research Institute. Notice that the idea of bootstrap appeared in different forms in Mahalanobis [1946], Quenouille [1949, 1956], Tukey [1958], [Simon, 1969, Chapters 23-25] and Maritz and Jarrett [1978].

We assume further that the collection $\mathscr{F}$ possesses enough measurability for randomization with i.i.d. multipliers to be possible and the usual Fubini's theorem can be used freely; such a set of conditions is $\mathscr{F} \in \mathrm{NLDM}(\mathbb{P})$ (Nearly Linearly Deviation Measurable), and $\mathscr{F}^2, \mathscr{F}'^2 \in \mathrm{NLSM}(\mathbb{P})$ (Nearly Linearly Supremum Measurable) in the terminology of Giné and Zinn [1990]. Here $\mathscr{F}^2$ and $\mathscr{F}'^2$ denote the classes of squared functions and squared differences of functions from $\mathscr{F}$, respectively. When all of these conditions hold, we write $\mathscr{F} \in \mathrm{M}(\mathbb{P})$. It is known that $\mathscr{F} \in \mathrm{M}(\mathbb{P})$ if $\mathscr{F}$ is countable, or if the empirical processes $\mathbb{G}_n$ are stochastically separable, or if $\mathscr{F}$ is image admissible Suslin (see [Giné and Zinn, 1990, p. 853 and 854]). The following Præstgaard and Wellner [1993]'s result concerns a central limit theorem in probability, for bootstrap empirical process as given in (2.2.1) indexed by the class $\mathscr{F}$.

Theorem 2.2.1.1 *Let $\mathscr{F} \in \mathrm{M}(\mathbb{P})$ be a class of $\mathrm{L}_2(\mathbb{P})$ functions, and let* **W** *be a triangular array of bootstrap weights satisfying assumptions **(B.1)–(B.5)**. Then*

$$\mathscr{F} \quad \textit{is} \quad \mathbb{P}-\textit{Donsker}$$

implies that

$$\hat{\mathbb{G}}_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(\mathrm{W}_{nj} - 1\right) \delta_{\mathrm{X}_j} \rightsquigarrow c\mathbb{G} \quad \textit{in} \quad l^\infty(\mathscr{F}) \quad \textit{in probability,}$$

*where c is given by assumption **(B.5)**.*

If in addition the envelope function F is square integrable then, the result holds almost everywhere.

These results for bootstrapped empirical processes can then be applied to many kinds of bootstrapped estimators since most estimators can be expressed as functionals of empirical processes. Much of the bootstrap results for such estimators will be deferred in Chapters 3 and 4 where we discuss M-estimation and Z-estimation.

2.3 References

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117 of *Lecture Notes in Math.*, pages 1–198. Springer, Berlin. 42

Barbe, P. and Bertail, P. (1995). *The weighted bootstrap*, volume 98 of *Lecture Notes in Statistics*. Springer-Verlag, New York. 42

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD. 29

Billingsley, P. (1968). *Convergence of probability measures*. John Wiley & Sons, Inc., New York-London-Sydney. 42

Cheng, G. and Huang, J. Z. (2010). Bootstrap consistency for general semiparametric M-estimation. *Ann. Statist.*, **38**(5), 2884–2915. 42

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*, volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa. 41

Giné, E. and Zinn, J. (1990). Bootstrapping general empirical measures. *Ann. Probab.*, **18**(2), 851–869. 41, 43

Kallenberg, O. (2002). *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition. 42

Kolmogorov, A. N. and Tihomirov, V. M. (1961). ε-entropy and ε-capacity of sets in functional space. *Amer. Math. Soc. Transl. (2)*, **17**, 277–364. 39

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer Series in Statistics. Springer, New York. 29, 33, 34, 35, 36, 38, 39, 42

Lo, A. Y. (1993). A Bayesian method for weighted sampling. *Ann. Statist.*, **21**(4), 2138–2148. 41

Mahalanobis, P. C. (1946). Sample surveys of crop yields in india. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, **7**(3), 269–280. 42

Maritz, J. S. and Jarrett, R. G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, **73**(361), 194–196. 42

Mason, D. M. and Newton, M. A. (1992). A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.*, **20**(3), 1611–1624. 41, 42

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, **57**(5), 1027–1057. 36

Præstgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, **21**(4), 2053–2086. 41, 42, 43

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *J. Roy. Statist. Soc. Ser. B*, **11**, 68–84. 42

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, **43**, 353–360. 42

Simon, J. (1969). *Basic Research Methods in Social Science*. Random House, New York. 42

Tukey, J. W. (1958). A problem of Berkson, and minimum variance orderly estimators. *Ann. Math. Statist.*, **29**, 614. 42

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. 29

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics. 29, 36, 38, 39, 40, 42

Chapter 3

General M-Estimator Processes and their m out of n Bootstrap with Functional Nuisance Parameters

Ce chapitre développe le contenu d'un article soumis, mis en forme pour être inséré dans le présent manuscrit de thèse.
Title : *General* M*-Estimator Processes and their m out of n Bootstrap with Functional Nuisance Parameters*.

abstract

In the present paper, we consider the problem of the estimation of a parameter $\boldsymbol{\theta}$, in Banach spaces, maximizing some criterion function which depends on an unknown nuisance parameter h, possibly infinite-dimensional. Classical estimation methods are mainly based on maximizing the corresponding empirical criterion by substituting the nuisance parameter by a nonparametric estimator. We show that the M-estimators converge weakly to maximizers of Gaussian processes under rather general conditions. The conventional bootstrap method fails in general to consistently estimate the limit law. We show that the m out of n bootstrap, in this extended setting, is weakly consistent under conditions similar to those required for weak convergence of the M-estimators. The aim of this paper is therefore to extend the existing theory on the bootstrap of the M-estimators. Examples of applications from the literature are given to illustrate the generality and the usefulness of our results. Finally, we investigate the performance of the methodology for small samples through a short simulation study.

Keywords: Gaussian process; M-estimation; Empirical process; m out n of bootstrap; Asymptotic distribution; Nuisance parameter; Semiparametric estimation; non standard distribution; Missing data.
AMS Subject Classifications: Primary : 62G05; 60F17; Secondary : 60F05; 62G09; 62G20; 62H10; 60F15.

3.1 Introduction

The semiparametric modeling has proved to be a flexible tool and provided a powerful statistical modeling framework in a variety of applied and theoretical contexts [refer to Pfanzagl [1990], Bickel *et al.* [1993], van der Vaart and Wellner [1996], van de Geer [2000], and Kosorok [2008]. An important work to be cited is the paper of Pakes and Pollard [1989], where a general central limit theorem is proved for estimators defined by minimization of the length of a vector-valued, random criterion function with no smoothness assumptions. The last reference was extended in different settings, among many others, by Pakes and Olley [1995], Chen *et al.* [2003], Zhan [2002]. Recall that the semiparametric models are statistical models where at least one parameter of interest is not Euclidean. The term "M-estimation" refers to a general method of estimation, where the estimators are obtained by maximizing (or minimizing) certain criterion functions. The most widely used M-estimators include maximum likelihood (MLE), ordinary least-squares (OLS), and least absolute deviation estimators. Notice that the major practical problem of maximum likelihood estimators is the lack of robustness, while many robust estimators achieve robustness at some cost in first-order efficiency. The appeal of the M-estimation method is that in addition to the statistical efficiency of the estimators when the parametric model is correctly specified, these estimators are also robust to contamination when the objective function is appropriately chosen. Throughout the available literature, investigations on the asymptotic properties of the M-estimators, as well as the relevant test statistics, have privileged the parametric case. However, in practice, we need more flexible models that contain both parametric and nonparametric components. This paper concentrates on this specific problem. To formulate the problem that we will treat in this paper, we need the following notation. Let $\mathscr{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be n independent copies of a random element $\mathbf{X}$ in a probability space $(\mathscr{S}, \mathscr{A}, \mathbb{P})$. For a Banach spaces $\mathscr{B}$ and $\mathscr{H}$ equipped with a norm $\|\cdot\|$ and a metric denoted by $d_{\mathscr{H}}(\cdot,\cdot)$ respectively, let $\mathscr{M}_{\Theta,\mathscr{H}}$ be a class of Borel measurable functions $\mathbf{m}_{\boldsymbol{\theta},h} : \mathscr{S} \to \mathbb{R}$, indexed by $\boldsymbol{\theta}$ over some parameter space $\Theta \subset \mathscr{B}$ and $h \in \mathscr{H}$, where $\boldsymbol{\theta}$ is the parameter of interest and h_0 the true value of h consists of nuisance parameter. We define the empirical measure to be

$$\mathbb{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{\mathbf{X}_i},$$

where, for $\mathbf{x} \in \mathscr{S}$, $\delta_{\mathbf{x}}$ is the measure that assigns mass 1 at $\mathbf{x}$ and zero elsewhere. Let $f(\cdot)$ be a real valued measurable function, $f : \mathscr{S} \to \mathbb{R}$. In the modern theory of the empirical it is customary to identify $\mathbb{P}$ and $\mathbb{P}_n$ with the mappings given by

$$f \to \mathbb{P}f = \int f d\mathbb{P}, \quad \text{and} \quad f \to \mathbb{P}_n f = \int f d\mathbb{P}_n = \frac{1}{n} \sum_{k=1}^{n} f(\mathbf{X}_i).$$

The M-estimand of interest $\boldsymbol{\theta}_0$ and its corresponding M-estimator $\boldsymbol{\theta}_n$ are assumed to be well-separated maximizers of the processes $\left\{\mathbb{P}\mathbf{m}_{\boldsymbol{\theta},h_0} : \boldsymbol{\theta} \in \Theta\right\}$ and $\left\{\mathbb{P}_n\mathbf{m}_{\boldsymbol{\theta},\widehat{h}} : \boldsymbol{\theta} \in \Theta\right\}$ for a given consistent sequence of estimators $\widehat{h}$ for h_0, respectively. Under suitable entropy conditions on $\mathscr{M}_{\Theta,\mathscr{H}}$ (defined below) and moment conditions on its envelope, we show that there exist norming sequences $\{\alpha_n\}$ and $\{r_n\}$ such that the random process $\left\{\alpha_n\mathbb{P}_n(\mathbf{m}_{\boldsymbol{\theta}_0+\gamma/r_n,\widehat{h}} - \mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}}) : \gamma \in \mathrm{K}\right\}$

converges weakly, in the sense of Hoffmann-Jørgensen [1991], see van der Vaart and Wellner [1996], in particular their Definition 1.3.3., to the process $\{\mathbb{Z}(\gamma) : \gamma \in K\}$, for each closed bounded subset $K \subset \mathscr{B}$. It follows by an argmax continuous mapping theorem, refer to Kosorok [2008] in particular Chapter 14, that $r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$ converges weakly to $\arg\max_\gamma \mathbb{Z}(\gamma)$. The latter weak limit has a complicated form in general and does not permit explicit computation. It would therefore be of interest to estimate the sampling distribution of $r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$ by the bootstrap for inferencing purposes. Bootstrap samples were introduced and first investigated in Efron [1979]. Since this seminal paper, bootstrap methods have been proposed, discussed, investigated and applied in a huge number of papers in the literature. Being one of the most important ideas in the practice of statistics, the bootstrap also introduced a wealth of innovative probability problems, which in turn formed the basis for the creation of new mathematical theories. The bootstrap can be described briefly as follows. Let $T(\mathbb{P})$ be a functional of an unknown distribution function $\mathbb{P}$, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ a sample from $\mathbb{P}$, and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ an independent and identically distributed [i.i.d.] sample with common distribution given by the empirical distribution $\mathbb{P}_n$ of the original sample. The distribution of $\{T(\mathbb{P}_n) - T(\mathbb{P})\}$ is then approximated by that of $\{T(\widehat{\mathbb{P}}_n^*) - T(\mathbb{P}_n)\}$ conditionally on $\mathbf{X}_1, \ldots, \mathbf{X}_n$, with $\widehat{\mathbb{P}}_n^*$ being the empirical distribution of $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. The key idea behind the bootstrap is that if a sample is representative of the underlying population, then one can make inferences about the population characteristics by resampling from the current sample. The asymptotic theory of the bootstrap with statistical applications has been reviewed in the books among others Efron and Tibshirani [1993] and Shao and Tu [1995]. Chernick [2008], Davison and Hinkley [1997], van der Vaart and Wellner [1996], Hall [1992] and Kosorok [2008]. A major application for an estimator is in the calculation of confidence intervals. By far the most favored confidence interval is the standard confidence interval based on a normal or a Student t-distribution. Such standard intervals are useful tools, but they are based on an approximation that can be quite inaccurate in practice. Bootstrap procedures are an attractive alternative. One way to look at them is as procedures for handling data when one is not willing to make assumptions about the parameters of the populations from which one sampled. The most that one is willing to assume is that the data are a reasonable representation of the population from which they come. One then resamples from the data and draws inferences about the corresponding population and its parameters. The resulting confidence intervals have received the most theoretical study of any topic in the bootstrap analysis. Roughly speaking, it is known that the bootstrap works in the i.i.d. case if and only if the central limit theorem holds for the random variable under consideration. For further discussion we refer the reader to the landmark paper by Giné and Zinn [1989]. It is worth noticing that some special examples reveal that the conventional bootstrap based on resamples of size n breaks down; see, for example, Bose and Chatterjee [2001] and El Bantli [2004]. We focus on a modified form of bootstrap methods, known as the m out of n bootstrap, with a view to remedy the problem of inconsistency. The m out of n scheme modifies the conventional scheme by drawing bootstrap resamples of size $m = o(n)$. See, for example, Bickel *et al.* [1997] for a review of this technique in a variety of contexts. For more recent references on the bootstrap one can refer to Bouzebda [2010], Bouzebda and Limnios [2013], Bouzebda *et al.* [2018], Alvarez-Andrade and Bouzebda [2013, 2015, 2019]

and the reference therein. Denote by $\widehat{\boldsymbol{\theta}}_m^*$ the M-estimator calculated from a bootstrap resample of size m. Weak convergence in probability of the conditional distribution of $r_m\left(\widehat{\boldsymbol{\theta}}_m^* - \boldsymbol{\theta}_n\right)$ to the distribution of $\arg\max_g \mathbb{Z}(g)$ is established under essentially similar conditions for weak convergence of $r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$, provided that $m = o(n)$, $m \to \infty$ and $a_m^2 m^{-1/2} \log n / \log(n/m+1) = o(1)$ for a fixed sequence $\{a_m\}$ depending on the size of the envelope for $\mathscr{M}_{\Theta,\mathscr{H}}$. The asymptotic properties of $\boldsymbol{\theta}_n$ have been established by, among many others, Bose and Chatterjee [2001] and El Bantli [2004], under appropriate concavity or differentiability conditions. Empirical process methods are instrumental tools for evaluating the large sample properties of estimators based on semiparametric models, including consistency, distributional convergence, and validity of the bootstrap. In particular, modern empirical process theory provides a more general approach to theoretical investigation of general M-estimators; see, for example, Dudley [1999], Kim and Pollard [1990], Pollard [1985], van de Geer [2000] and van der Vaart and Wellner [1996]. Most results obtained thus far are, however, restricted to the cases where the Gaussian process $\mathbb{Z}$ has either quadratic mean function or quadratic covariance function. In order to establish stronger results which cover cases where the covariance and mean functions of $\mathbb{Z}$ may take on more general structures, we will use the empirical process approach. Applications of the bootstrap to M-estimation have been investigated deeply in the literature extensively. Relevant theoretical results concern mostly M-estimators with $r_n = n^{1/2}$ and asymptotically Gaussian limits. The most common technique for studying bootstrap M-estimators is the linearization which can not be used in a nonstandard setting. Under standard regularity conditions, the Edgeworth expansions for bootstrap distributions of finite-dimensional M-estimators are Lahiri [1992]. Under a weak form of differentiability condition, Arcones and Giné [1992] investigated bootstrapping finite-dimensional $n^{1/2}$-consistent M-estimators and established an almost sure bootstrap central limit theorem. An in-probability bootstrap central limit theorem for possibly infinite-dimensional Z-estimators is investigated by Wellner and Zhan [1996]. In the setting of the nonregular vector-valued M-estimators obtained from $\mathbf{m}_{\boldsymbol{\theta}}$ concave in $\boldsymbol{\theta}$, Bose and Chatterjee [2001] investigated a weighted form of the bootstrap including the m out of n bootstrap is a special case. The M-estimation for linear models under nonstandard conditions was considered by El Bantli [2004], and proved that the m out of n bootstrap is consistent but the conventional bootstrap is not in general. The Bose and Chatterjee [2001] and El Bantli [2004] results are restricted to the case where $\mathbb{Z}$ has a quadratic covariance function, concavity and differentiability assumptions. Lee and Pun [2006] prove m out of n bootstrap consistency for vector-valued M-estimators under twice-differentiability of the process $\mathbb{P}\mathbf{m}_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ may contain a subvector of nuisance parameters, in which case the process $\mathbb{Z}$ has a quadratic mean function. Lee [2012] gives general result of m out of n bootstrap of M-estimators without the presence of nuisance parameter. Under nonstandard conditions, Lee and Yang [2020] proposed a one-dimensional pivot derived from the criterion function, and prove that its distribution can be consistently estimated by the m out of n bootstrap, or by a modified version of the perturbation bootstrap. They provide a new method for constructing confidence regions which are transformation equivariant and have shapes driven solely by the criterion function.

The main purpose of the present work is to consider a general framework of non-smooth semi-parametric M-estimators extending the setting of Lee [2012] to the $\mathscr{B}$-valued M-estimators in presence of nuisance parameter where the rate of convergence of the nuisance parameter may be different of that of the parameter of interest. More precisely, we consider the m out n bootstrapped version of the M-estimator investigated in Delsol and Van Keilegom [2020], where these authors showed that, their M-estimator converges weakly to some process which is composed on Gaussian process and some deterministic continuous function, which is harder to evaluate for practical use. For that we propose in this paper as a solution of this problem the m out of n bootstrap. We mention at this stage that parameter $\boldsymbol{\theta}$, in the present paper, belongs to some Banach space which is different from the last mentioned work where the parameter of interest is Euclidean. Hence, we restate the results of Delsol and Van Keilegom [2020] under more general conditions. The main aim of the present paper is to provide a first full theoretical justification of the m out of n bootstrap consistency of M-estimators with non-smooth criterion functions of the parameters and gives the consistency rate together with the asymptotic distribution of the parameters of interest $\boldsymbol{\theta}_0$. This requires the effective application of large sample theory techniques, which were developed for the empirical processes. The Lee [2012] results are not directly applicable here since the estimation procedures depend on some nuisance parameters. These results are not only useful in their own right but essential for the derivation of our asymptotic results.

The paper is organized as follows. Section 3.2 introduces the notation and assumptions. Section 3.3 states the main theorems. Though our main objective in the paper is theoretical, we provide in Section 3.5 Monte Carlo simulations of simulations to look at the method's performance in practice. Some concluding remarks are given in Section 7.1. All proofs are gathered in Section 3.6. In the Appendix we apply our theorems and prove as corollaries new m out of n bootstrap consistency results for three examples.

3.2 Notation

We abuse notation slightly by identifying the underlying probability space $(\mathscr{S}, \mathscr{A}, \mathbb{P})$ with the product space $(\mathscr{S}^\infty, \mathscr{A}^\infty, \mathbb{P}^\infty) \times (\mathscr{Z}, \mathscr{C}, \tilde{\mathrm{P}})$. Now $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are equal to the coordinate projections on the first n coordinates. All auxiliary variables, assumed to be independent of the $\mathbf{X}_i$, depend only on the last coordinate. We will use the usual notation of the empirical processes of van der Vaart and Wellner [1996]. Let $\mathbb{Q}$ denote some signed measure on $\mathscr{S}$. Let $\mathscr{F}$ be a class of measurable functions $f : \mathscr{S} \to \mathbb{R}$. Define

$$\|\mathbb{Q}f\|_{\mathscr{F}} = \sup_{f \in \mathscr{F}} |\mathbb{Q}f|.$$

For any $r \geq 1$, denote by $\mathrm{L}^r(\mathbb{Q})$ the class of measurable functions $f : \mathrm{S} \to \mathbb{R}$ with

$$\int |f|^r d\mathbb{Q} < \infty,$$

where $\mathbb{Q}$ is a probability measure. The $L^r(\mathbb{Q})$-norm $\|\cdot\|_{\mathbb{Q},r}$ is defined by

$$\|f\|_{\mathbb{Q},r} = \left(\int |f|^r d\mathbb{Q}\right)^{1/r},$$

for $f \in L^r(\mathbb{Q})$. The essential supremum of $f \in L^\infty(\mathbb{Q})$is denoted by $\|f\|_{\mathbb{Q},\infty}$. The covering number $N(\epsilon, \mathscr{F}, L^r(\mathbb{Q}))$ of a function class $\mathscr{F} \subset L^r(\mathbb{Q})$ is computed with respect to the $L^r(\mathbb{Q})$-norm for radius $\epsilon > 0$. To be more precise, $N(\epsilon, \mathscr{F}, L^r(\mathbb{Q}))$ is the minimum number of balls $\{g : \|g - h\|_{\mathbb{Q},r} < \epsilon\}$ of radius ϵ covering $\mathscr{F}$.
For some random element $\mathbf{Z}$, the probability measure induced by $\mathbf{Z}$ is denoted by $\mathbb{P}_{\mathbf{Z}}$, conditional on all other variables. The empirical process is defined to be

$$\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - \mathbb{P}).$$

The outer and inner probability measures derived from $\mathbb{P}$ are designated by $\mathbb{P}^*$ and $\mathbb{P}_*$, respectively. Outer and inner probability measures to be understood in the sense used in the monograph by van der Vaart and Wellner [1996], in particular their definitions in page 6. Let T be any map from the underlying probability space to the extended real line $\overline{\mathbb{R}}$. The minimal measurable majorant and maximal measurable minorant of T are denoted by T^* and T_*, respectively. For any subset B of the probability space, by similar notation, its indicator function satisfies $\mathbb{1}_{B^*} = \mathbb{1}_B^*$ and $\mathbb{1}_{B_*} = (\mathbb{1}_B)_*$. We draw randomly with replacement from $\mathscr{X}$ independent bootstrap observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$. Let us define

$$\widehat{\mathbb{P}}_m^* = m^{-1} \sum_{i=1}^m \delta_{\mathbf{Y}_i},$$

so that

$$\widehat{\mathbb{P}}_m^* = \sum_{i=1}^m W_i \delta_{\mathbf{X}_i},$$

where $mW = m(W_1, \ldots, W_n)$ is a multinomial vector with m trials and parameters $(n^{-1}, \ldots, n^{-1})$, independent of the $\mathbf{X}_i$. The probability measure induced by bootstrap resampling conditional on $\mathscr{X}$ is denoted by $\mathbb{P}_W$. Let us define the bootstrappped empirical process by

$$\widehat{\mathbb{G}}_m^* = m^{1/2}\left(\widehat{\mathbb{P}}_m^* - \mathbb{P}_n\right).$$

Let T_n denote a sequence of maps. Let $\mathbb{D}$ be a metric space. Let T be a $\mathbb{D}$-valued measurable map from the underlying probability. If T_n is bounded in outer probability, we will write $T_n = O_{\mathbb{P}^*}(1)$, in a similar way, if T_n converges in outer probability to zero, we will write $T_n = o_{\mathbb{P}^*}(1)$. Assume that

$$\lim_{M\to\infty} \liminf_{n\to\infty} \mathbb{P}_W\left\{\|T_n\| < M\right\}_* = 1. \tag{3.2.1}$$

If (3.2.1) holds along almost every sequence $\mathbf{X}_1, \mathbf{X}_2, \ldots$, we write $T_n = O_{\mathbb{P}_W^*}(1)$ a.s. (almost surely). If for any subsequence $\{T_{n'}\}$, there exists a further subsequence $\{T_{n''}\}$ with $T_{n''} = O_{\mathbb{P}_W^*}(1)$ a.s., we write $T_n = O_{\mathbb{P}_W^*}(1)$ i.p. (in probability). We write $T_n = o_{\mathbb{P}_W^*}(1)$ a.s., if, for any $\epsilon > 0$, we have

$$\mathbb{P}_W\left\{\|T_n\| > \epsilon\right\}^* \to 0, \quad \text{as} \quad n \to \infty \tag{3.2.2}$$

almost surely. We write $T_n = o_{\mathbb{P}^*_W}(1)$ i.p., in the case when the convergence (3.2.2) is in probability. The weak convergence of T_n to T, in the sense of Hoffmann-Jørgensen [1991], is denoted by $T_n \rightsquigarrow T$. The space of $\mathbb{D}$-valued functions in $\mathbb{R}$ bounded by 1 in the Lipschitz norm is denoted by $BL_1(\mathbb{D})$. The conditional weak convergence of T_n to a separable T in $\mathbb{D}$ is characterized by the condition

$$\sup_{f \in BL_1(\mathbb{D})} \left| \mathbb{P}^*_W f(T_n) - \mathbb{P} f(T) \right| \to 0. \tag{3.2.3}$$

In the case of the convergence (3.2.3) is in outer probability, we will write write $T_n \rightsquigarrow T$ i.p., in a similar way, if it is outer almost sure, we write $T_n \rightsquigarrow T$ a.s.

Define $\mathscr{M}_{S,\mathscr{H}} = \{\mathbf{m}_{\boldsymbol{\theta},h} : \boldsymbol{\theta} \in S, h \in \mathscr{H}\} \subset \mathscr{M}_{\Theta,\mathscr{H}}$, where $S \subset \Theta$. For any $\delta, \delta_1, \eta > 0$, let us denote by $\mathscr{M}_{\delta,\delta_1}(\eta)$ and $\mathscr{M}_{\delta,\delta_1}$ the class of functions

$$\mathscr{M}_{\delta,\delta_1} = \left\{ \mathbf{m}_{\boldsymbol{\theta},h} - \mathbf{m}_{\boldsymbol{\theta}_0,h} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le \delta, d_{\mathscr{H}}(h, h_0) \le \delta_1, \boldsymbol{\theta} \in \Theta, h \in \mathscr{H} \right\},$$

$$\mathscr{M}_{\delta,\delta_1}(\eta) = \left\{ \mathbf{m}_{\boldsymbol{\theta},h} - \mathbf{m}_{\boldsymbol{\psi},h} : \|\boldsymbol{\theta} - \boldsymbol{\psi}\| < \eta, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \vee \|\boldsymbol{\psi} - \boldsymbol{\theta}_0\| < \delta, d_{\mathscr{H}}(h, h_0) \le \delta_1, \boldsymbol{\theta}, \boldsymbol{\psi} \in \Theta, h \in \mathscr{H} \right\}.$$

The envelope function of $\mathscr{M}_{\delta,\delta_1}$ is denoted by M_{δ,δ_1}. For each $\boldsymbol{\psi} \in \mathscr{B}$ and $h \in \mathscr{H}$ with $\boldsymbol{\theta}_0 + \boldsymbol{\psi} \in \Theta$, define $\tilde{\mathbf{m}}_{\boldsymbol{\psi},h} = \mathbf{m}_{\boldsymbol{\theta}_0+\boldsymbol{\psi},h} - \mathbf{m}_{\boldsymbol{\theta}_0,h}$. For any $\mathscr{T} \subset \mathscr{B}$, the class of bounded functions from $\mathscr{T}$ to $\mathbb{R}$ is denoted by $\ell^\infty(\mathscr{T})$, equipped with the sup norm. In the sequel, for all $x \in \mathscr{S}$ and closed bounded $K \subset \Theta$, assume that

$$\sup_{\boldsymbol{\theta} \in K, h \in \mathscr{H}} |\mathbf{m}_{\boldsymbol{\theta},h}(x) - \mathbb{P}\mathbf{m}_{\boldsymbol{\theta},h}| < \infty.$$

In the sequel, we denote by C a positive constant that may be different from line to line. The choice of the bootstrap sample size m is theoretically governed by **(AB1)** and **(C4)**. The above conditions are typically satisfied by taking $m \propto n^c$, for some sufficiently small $c \in (0,1)$. Empirical determination of m has long been an important problem which has not yet been fully resolved, for more comments see Remark 3.3.3.6 below.

3.3 Main results

In this section, we present four main theorems, each of independent interest, which lead eventually to weak convergence of $r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$ and in-outer-probability m out of n bootstrap consistencies in the context of general M-estimation by applying the argmax theorem in van der Vaart and Wellner [1996] and in Lee [2012] respectively. Let us recall the basic idea. If the argmax functional is continuous with respect to some metric on the space of the criterion functions, then convergence in distribution of the criterion functions will imply the convergence in distribution of their points of maximum, the M-estimators, to the maximum of the limit criterion function. First, we establish consistency of $\boldsymbol{\theta}_n$ and $\widehat{\boldsymbol{\theta}}^*_m$ for $\boldsymbol{\theta}_0$ by the following theorem.

3.3.1 Consistency

In our analysis, we consider the following assumptions. Assume that the sequence of positive constants $r_n \uparrow \infty$, for some fixed $\nu > 1$ and for some function $\ell : (0,\infty) \to [0,\infty)$ which is slowly varying at ∞.

(A1) $\mathbb{P}\left(\widehat{h} \in \mathscr{H}\right) \longrightarrow 1$ as $n \longrightarrow \infty$ and $d_{\mathscr{H}}(\widehat{h}, h_0) \xrightarrow{\mathbb{P}^*} 0$.

(A2) $\mathscr{M}_{\Theta,\mathscr{H}}$ is Glivenko-Cantelli:

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{M}_{\Theta,\mathscr{H}}} = o_{\mathbb{P}^*}(1).$$

(A3) $\lim_{d_{\mathscr{H}}(h,h_0)\to 0} \sup_{\boldsymbol{\theta}\in\Theta} |\mathbb{P}\mathbf{m}_{\boldsymbol{\theta},h} - \mathbb{P}\mathbf{m}_{\boldsymbol{\theta},h_0}| = 0$.

(A4) The parameter of interest $\boldsymbol{\theta}_0$ lies in the interior of Θ and satisfies, for every open $\mathscr{O}$ containing $\boldsymbol{\theta}_0$,

$$\mathbb{P}\mathbf{m}_{\boldsymbol{\theta}_0,h_0} > \sup_{\boldsymbol{\theta}\notin\mathscr{O}} \mathbb{P}\mathbf{m}_{\boldsymbol{\theta},h_0}.$$

(A5) The M-estimator $\boldsymbol{\theta}_n$ satisfies $\mathbb{P}_n\mathbf{m}_{\boldsymbol{\theta}_n,\widehat{h}} \geq \mathbb{P}_n\mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}} - \mathrm{R}_n$, with

$$r_n^{\nu}\ell(r_n)\mathrm{R}_n = o_{\mathbb{P}^*}(1).$$

(AB1) $m = m_n \to \infty$, $m = o(n)$ and $r_m^{\nu}\ell(r_m) = o\left(r_n^{\nu}\ell(r_n)\right)$.

(AB2) $d_{\mathscr{H}}(\widehat{h}_m, h_0) = o_{\mathbb{P}^*_{\mathrm{W}}}(1)$ i.p.

(AB3) The m out of n bootstrap M-estimator $\widehat{\boldsymbol{\theta}}^*_m$ satisfies $\widehat{\mathbb{P}}^*_m\mathbf{m}_{\widehat{\boldsymbol{\theta}}^*_m,\widehat{h}_m} \geq \widehat{\mathbb{P}}^*_m\mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}_m} - \widehat{\mathrm{R}}_n$, with

$$r_m^{\nu}\ell(r_m)\widehat{\mathrm{R}}_n = o_{\mathbb{P}^*_{\mathrm{W}}}(1), \text{ i.p.}$$

Remark 3.3.1.1 **(i)** *Assumption **(A2)** fulfilled under some entropy and moment conditions; see for example, Theorem 2.4.3,* (p.123) *of van der Vaart and Wellner [1996].*

(ii) *Assumption **(A3)** is automatically hold for example if; there exist function $\mathfrak{G}(\cdot)$ and such that for any h in the neighborhood of h_0 and any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we have:*

$$|\mathbf{m}(\mathbf{X}_i,\boldsymbol{\theta},h) - \mathbf{m}(\mathbf{X}_i,\boldsymbol{\theta},h_0)| \leq \mathfrak{G}(\mathbf{X}_i)\, d_{\mathscr{H}}(h,h_0).$$

The function $\mathfrak{G}(\cdot)$ satisfies;

$$\mathbb{P}\mathfrak{G}(\mathbf{X}) < \infty,$$

or the function $h \mapsto \mathbf{m}(x,\boldsymbol{\theta},h)$ is Lipschitz uniformly over x and $\boldsymbol{\theta}$.

(iii) *Assumptions **(A5)** and **(AB3)** are trivially fulfilled when*

$$\mathbb{P}_n\mathbf{m}_{\boldsymbol{\theta}_n,\widehat{h}} \geq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{P}_n\mathbf{m}_{\boldsymbol{\theta},\widehat{h}} - \mathrm{R}_n,$$

and

$$\widehat{\mathbb{P}}^*_m\mathbf{m}_{\widehat{\boldsymbol{\theta}}^*_m,\widehat{h}_m} \geq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \widehat{\mathbb{P}}^*_m\mathbf{m}_{\boldsymbol{\theta},\widehat{h}_m} - \widehat{\mathrm{R}}_n,$$

respectively, which allows to deal with approximations of the value that actually maximizes $\theta \mapsto \mathbb{P}_n\mathbf{m}_{\boldsymbol{\theta},\widehat{h}}$ *and* maximizes $\theta \mapsto \widehat{\mathbb{P}}^*_m\mathbf{m}_{\boldsymbol{\theta},\widehat{h}_m}$ *respectively.*

(iv) *Assumption **(AB2)** poses no difficulty in practice and is met trivially by, for example, setting* $\widehat{h}_m = \widehat{h}$.

(v) *For the finite-dimensional* $\boldsymbol{\theta}$, ***(A5)*** *and* ***(AB3)*** *can be achieved by a global maximization of the processes* $\mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta},\widehat{h}}$ *and* $\widehat{\mathbb{P}}^*_m \mathbf{m}_{\boldsymbol{\theta},\widehat{h}_m}$, *in this situation* $\mathrm{R}_n = \widehat{\mathrm{R}}_n = 0$. *For the infinite-dimensional* $\boldsymbol{\theta}$, *the maximization of the processes may be very complex or not practically feasible. To circumvent this, we need sophisticated algorithms to construct* $\boldsymbol{\theta}_n$ *and* $\widehat{\boldsymbol{\theta}}^*_m$ *fulfilling* ***(A5)*** *and* ***(AB3)***.

(vi) *Finally, it's possible to replace the following assumptions* ***(A2)*** *and* ***(A4)*** *by:*

(A1′) *For every compact* $\mathrm{K} \subset \boldsymbol{\Theta}$, $\mathscr{M}_{\mathrm{K},\mathscr{H}}$ *is Glivenko-Cantelli.*

(A2′) *The map* $\theta \mapsto \mathbb{P}\mathbf{m}_{\boldsymbol{\theta},h_0}$ *is upper semicontinuous with a unique maximum at* $\boldsymbol{\theta}_0$.

(A3′) $\boldsymbol{\theta}_n$ *is uniformly tight.*

(AB1′) $\widehat{\boldsymbol{\theta}}^*_m$ *is uniformly tight i.p.*

Theorem 3.3.1.2 *(i) Assume* ***(A1)-(A5)***. *Then*

$$\boldsymbol{\theta}_n - \boldsymbol{\theta}_0 = o_{\mathbb{P}^*}(1).$$

(ii) Assume ***(A2)***, ***(A3)***, ***(A4)*** *and* ***(AB1)-(AB3)***. *Then*

$$\widehat{\boldsymbol{\theta}}^*_m - \boldsymbol{\theta}_0 = o_{\mathbb{P}^*_{\mathrm{W}}}(1) \quad i.p.$$

Note that, the result of part (i) holds if we replaced ***(A2)*** *and* ***(A4)*** *by* ***((A1′)-(A3′))*** *and the result of part (ii) holds if we replaced* ***(A2)*** *and* ***(A4)*** *by* ***(A1′)***, ***(A2′)*** *and* ***(AB1′)***.

In the sequel, we refer to the sets of assumptions which imply the parts (i) and (ii); **(C)** and **(CP)**; respectively. Next we give the set of assumptions needed to identify rates of convergence of $\boldsymbol{\theta}_n$ and $\widehat{\boldsymbol{\theta}}^*_m$ to $\boldsymbol{\theta}_0$, which is the important step for studying the weak convergence of these estimators.

Remark 3.3.1.3 *We highlight that the parameter of interest* $\boldsymbol{\theta}$ *is not restricted to belong to some Euclidean space as in Delsol and Van Keilegom [2020]. More precisely, we consider the general framework in which* $\boldsymbol{\theta} \in \Theta$, *where* Θ *is a subset of some Banach space* $\mathscr{B}$. *Notice that the result (i) of Theorem 3.3.1.2 is a bit more general than the analogous stated in the last reference, by the fact the conditions imposed are more general in our setting and extend those of Lee [2012] to the semiparametric models.*

3.3.2 Rates of Convergence

Let us introduce the following assumptions:

(B1) $v_n d_{\mathscr{H}}(\widehat{h}, h_0) = \mathrm{O}_{\mathbb{P}^*}(1)$ for some $v_n \longrightarrow \infty$.

(B2) For all $\delta_1 > 0$, there exist $\boldsymbol{\alpha} < \nu$, $K > 0$, $\delta_0 > 0$, for all $n \in \mathbb{N}$ there exist a function φ for which $\delta \mapsto \frac{\varphi(\delta)}{\delta^{\boldsymbol{\alpha}}}$ is decreasing on $(0, \delta_0]$ and $r_n^{\nu}\ell(r_n)n^{-1/2}\varphi(1/r_n) \leq C$ for n sufficiently large and some positive constant C, such that for all $\delta \leq \delta_0$,

$$\mathbb{P}^* \left[\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \leq \delta, d_{\mathscr{H}}(h,h_0) \leq \frac{\delta_1}{v_n}} |\mathbb{G}_n \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h}| \right] \leq K\varphi(\delta).$$

(B3) There exist $\eta_0 > 0$, $C > 0$ and two positive and non-decreasing functions $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ on $(0, \eta_0]$ such that for all $\boldsymbol{\theta} \in \Theta$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta_0$:

$$\mathbb{P}\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,\widehat{h}} \leq W_n \boldsymbol{\psi}_1(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|) - (C + o_{\mathbb{P}^*}(1))\boldsymbol{\psi}_2(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|).$$

Moreover, there exist $\beta_2 > \alpha, \beta_1 < \beta_2, \delta_0 > 0$ such that $\delta \mapsto \boldsymbol{\psi}_1(\delta)\delta^{-\beta_1}$ is non-increasing and $\delta \mapsto \boldsymbol{\psi}_2(\delta)\delta^{-\beta_2}$ is non-decreasing on $(0, \delta_0]$, and such that, for some sequence $r_n \to \infty$,

$$\boldsymbol{\psi}_1\left(r_n^{1-\nu}\ell^{-1}(r_n)\right) W_n = O_{\mathbb{P}^*}\left(\boldsymbol{\psi}_2\left(r_n^{1-\nu}\ell^{-1}(r_n)\right)\right).$$

for definition of $\mathbb{P}$-measurability.

(BB1) $v_m d_{\mathscr{H}}(\widehat{h}_m, h_0) = O_{\mathbb{P}^*_W}(1)$ i.p. for some $v_m \longrightarrow \infty$.

(BB2) With the same notation in assumption **(B2)** we replace r_n (v_n) by r_m (v_m) with assumption **(AB1)** we have;

$$\mathbb{P}^*\mathbb{P}^*_W \left[\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \leq \delta, d_{\mathscr{H}}(h,h_0) \leq \frac{\delta_1}{v_m}} |\widehat{\mathbb{G}}^*_m \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h}| \right] \leq K\varphi(\delta).$$

(BB3) With the same notation in assumption **(B3)** we replace r_n by r_m with assumption **(AB1)** in mind we have;

$$\mathbb{P}\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,\widehat{h}_m} \leq W_m \boldsymbol{\psi}_1(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|) - (C + o_{\mathbb{P}^*}(1))\boldsymbol{\psi}_2(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|),$$

where for some sequence $r_m \to \infty$,

$$\boldsymbol{\psi}_1\left(r_m^{1-\nu}\ell^{-1}(r_m)\right) W_m = O_{\mathbb{P}^*_W}\left(\boldsymbol{\psi}_2\left(r_m^{1-\nu}\ell^{-1}(r_m)\right)\right), \text{ i.p.}$$

Remark 3.3.2.1 **(i)** *Assumption **(B1)** is a high-level assumption. Such condition on the nuisance parameter $\widehat{h}$ could be obtained by many asymptotic results. In the present paper, we are primarily concerned with the cases where the convergence rate of the* M*-estimator of $\boldsymbol{\theta}$ is not affected by the estimation of the nuisance parameter h.*

(ii) *Assumption **(B2)** is fulfilled if we assume that for any $\boldsymbol{x}$ the function $(\boldsymbol{\theta}, h) \to \mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta}, h(\boldsymbol{x}, \boldsymbol{\theta})) - \mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta}_0, h(\boldsymbol{x}, \boldsymbol{\theta}_0))$ is uniformly bounded on an open neighborhood of $(\boldsymbol{\theta}_0, h_0)$, i.e., on*

$$\left\{(\boldsymbol{\theta}, h) : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta_0, d_{\mathscr{H}}(h, h_0) \leq \delta_1'\right\},$$

for some $\delta_0, \delta_1' > 0$. *We consider the class* $\mathscr{M}_{\delta,\delta_1'}$ *for any* $\delta \leq \delta_0$ *and its envelope* $M_{\delta,\delta_1'}$. *For any* δ_1, *we have, for* n *large enough;* $\delta_1 v_n^{-1} \leq \delta_1'$. *After by the entropy conditions on* $\mathscr{M}_{\delta,\delta_1'}$,

$$\int_0^1 \sup_{\delta<\delta_0} \sup_{\mathbb{Q}} \sqrt{1+\log N\left(\epsilon \left\|M_{\delta,\delta_1'}\right\|_{\mathbb{L}_2(\mathbb{Q})}, \mathscr{M}_{\delta,\delta_1'}, \mathbb{L}_2(\mathbb{Q})\right)} d\epsilon < +\infty, \tag{3.3.1}$$

where the second supremum is taken over all finitely discrete probability measures $\mathbb{Q}$ *on* $\mathscr{S}$. *We get;*

$$\mathbb{P}^* \left[\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta, d_{\mathscr{H}}(h,h_0)\leq\frac{\delta_1}{v_n}} |\mathbb{G}_n \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h}| \right] \leq K_1 \sqrt{\mathbb{P}^*\left[M_{\delta,\delta_1'}^2\right]},$$

see Theorems 2.14.1 and 2.14.2 in van der Vaart and Wellner [1996]. Then the last part of ***(B2)*** *holds if* $\varphi(\delta)$ *can be chosen such that*

$$\exists K_0, \forall \delta \leq \delta_0 : \sqrt{\mathbb{P}^*\left[M_{\delta,\delta_1'}^2\right]} \leq K_0 \varphi(\delta). \tag{3.3.2}$$

Note that, all the different rate of convergence r_n *in the literature for smooth or not smooth function satisfied the last term in assumption* ***(B2)***.

(iii) *We choose for simplification* $\boldsymbol{\psi}_1(x) = Id(x) = x$ *and* $\boldsymbol{\psi}_2(x) = x^2$ *in assumption* ***(B3)***, *so it's hold under the following conditions :*

(a) $\Theta \subset \mathscr{B}$, *where* $\mathscr{B}$ *is a Banach space.*

(b) There exists $\delta_2 > 0$ *such that for any* h *satisfying* $d_{\mathscr{H}}(h,h_0) \leq \delta_2$, *the function* $\boldsymbol{\theta} \mapsto \mathbb{P}(\mathbf{m}(\mathbf{X},\boldsymbol{\theta},h))$ *is twice Fréchet differentiable on an open neighborhood of* $\boldsymbol{\theta}_0$,

$$\lim_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\to 0} \sup_{d_{\mathscr{H}}(h,h_0)\leq\delta_2} \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^{-2} \Big| \mathbb{P}\mathbf{m}_{\boldsymbol{\theta},h} - \mathbb{P}\mathbf{m}_{\boldsymbol{\theta}_0,h} - \Gamma(\boldsymbol{\theta}_0,h)(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + \frac{1}{2}\Lambda(\boldsymbol{\theta}_0,h)(\boldsymbol{\theta}-\boldsymbol{\theta}_0,\boldsymbol{\theta}-\boldsymbol{\theta}_0) \Big| = 0.$$

For more detail see [Allaire, 2005, p.306].

(c) $\Gamma(\boldsymbol{\theta}_0,h)(\cdot)$ *is a continuous linear form, with* $\|\Gamma(\boldsymbol{\theta}_0,\widehat{h})\| = O_{\mathbb{P}^*}\left(\frac{1}{r_n^{\nu-1}\ell(r_n)}\right)$ *and* $\Gamma(\boldsymbol{\theta}_0,h_0) = 0$.

(d) $\Lambda(\boldsymbol{\theta}_0,h)(\cdot,\cdot)$ *is bilinear form with* $\Lambda(\boldsymbol{\theta}_0,h_0)$ *is bounded, symmetric, positive definite and elliptic (i.e.* $\Lambda(\boldsymbol{\theta}_0,h_0)(u,u) \geq C\|u\|^2$*) and* $h \mapsto \Lambda(\boldsymbol{\theta}_0,h)$ *is continuous in* h_0, *i.e.,*

$$\lim_{d_{\mathscr{H}}(h,h_0)\to 0} \sup_{u\in\mathbb{R}^k,\|u\|=1} \|(\Lambda(\boldsymbol{\theta}_0,h) - \Lambda(\boldsymbol{\theta}_0,h_0))u\| = 0.$$

These assumptions and the fact that the bilinear form is bounded, it results when $d_{\mathscr{H}}(\widehat{h},h_0) \leq \delta_2$;

$$\begin{aligned}\mathbb{P}\mathbf{m}_{\boldsymbol{\theta},\widehat{h}} - \mathbb{P}\mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}} &= \Gamma(\boldsymbol{\theta}_0,\widehat{h})(\gamma_{\boldsymbol{\theta}}) - \frac{1}{2}\Lambda(\boldsymbol{\theta}_0,h_0)(\gamma_{\boldsymbol{\theta}},\gamma_{\boldsymbol{\theta}}) + o_{\mathbb{P}^*}(\|\gamma_{\boldsymbol{\theta}}\|^2) + o(\|\gamma_{\boldsymbol{\theta}}\|^2) \\ &\leq W_n\|\gamma_{\boldsymbol{\theta}}\| - C\|\gamma_{\boldsymbol{\theta}}\|^2 + \beta_n(\|\gamma_{\boldsymbol{\theta}}\|),\end{aligned}$$

where $\gamma_{\boldsymbol{\theta}} = \boldsymbol{\theta} - \boldsymbol{\theta}_0$. *So* ***(B3)*** *holds with*

$$\mathrm{W}_n = \left\| \Gamma\left(\boldsymbol{\theta}_0, \widehat{h}\right) \right\|.$$

Note that when the space $\Theta \subset \mathbf{E}$ *where* $\mathbf{E}$ *is some Euclidean space, the Fréchet derivatives* $\Gamma(\boldsymbol{\theta}_0, h)$ *and* $\Lambda(\boldsymbol{\theta}_0, h)$ *become the usually derivatives i.e., the Gradient and the Hessian matrix respectively, which is given in* ***Remark 2****(v) of Delsol and Van Keilegom [2020].*

(iv) *Assumption* ***(BB1)*** *poses no difficulty in practice and is met trivially by, for example, setting* $\widehat{h}_m = \widehat{h}$, *like in* ***Remark 3.3.1.1*** *(iv).*

(v) *Assumption* ***(BB2)*** *is a 'high-level' assumption. It serves to control the modulus of continuity of the bootstrapped empirical processes; which is needed to derive the rate of convergence of the bootstrapped estimator* $\widehat{\boldsymbol{\theta}}^*_m$. *Note that when we are in the situation of the n out of n bootstrap this condition is given in Ma and Kosorok [2005] and in Lemma 1 of Cheng and Huang [2010] for more generally in the exchangeable bootstrap weights. In our setting; it's fulfilled under some entropy conditions, for brevity with the same notation in* ***(ii)****, let* $\widetilde{\mathrm{N}}_1, \widetilde{\mathrm{N}}_2, \ldots$ *be i.i.d. symmetrized Poisson variables with parameter* $\frac{1}{2}m/n$ *and* $\varepsilon_1, \varepsilon_2, \ldots$ *are i.i.d. Rademacher variables independent of* $\widetilde{\mathrm{N}}_1, \widetilde{\mathrm{N}}_2, \ldots$ *and* $\mathbf{X}_1, \mathbf{X}_2, \ldots$. *Denote by* $\mathrm{R} = (\mathrm{R}_1, \ldots, \mathrm{R}_n)$ *a random permutation of* $\{1, 2, \ldots, n\}$, *independent of all other variables. Let us introduce*

$$\mathbb{P}^{\mathrm{R}}_k = k^{-1} \sum_{i=1}^{k} \delta_{\mathbf{X}_{\mathrm{R}_i}},$$

for each $k \in \{1, \ldots, n\}$. *By Lemma 3.6.6 of van der Vaart and Wellner [1996] and the paragraph before it (ahead) with sub-Gaussian inequality for Rademacher process we obtain*

$$\mathbb{P}^*_{\mathrm{W}} \left\| \widehat{\mathbb{G}}^*_m \right\|_{\mathscr{M}_{\delta,\delta'_1}} \leq 4\mathbb{P}_{\widetilde{\mathrm{N}}} \left\| \frac{1}{\sqrt{k}} \sum_{i=1}^{n} |\widetilde{\mathrm{N}}_i| \varepsilon_i \delta_{\mathbf{X}_i} \right\|_{\mathscr{M}_{\delta,\delta'_1}}. \tag{3.3.3}$$

Applying now Lemma 3.6.7 of van der Vaart and Wellner [1996] to the right side of (3.3.3) with n_0 *set to* 1 *we get;*

$$\begin{aligned}
\mathbb{P}^*_{\mathrm{W}} \left\| \widehat{\mathbb{G}}^*_m \right\|_{\mathscr{M}_{\delta,\delta'_1}} &\leq 4\mathbb{P}_{\widetilde{\mathrm{N}}} \left\| \frac{1}{\sqrt{k}} \sum_{i=1}^{n} |\widetilde{\mathrm{N}}_i| \varepsilon_i \delta_{\mathbf{X}_i} \right\|_{\mathscr{M}_{\delta,\delta'_1}} \\
&\leq \sqrt{\frac{n}{k}} \left\| \widetilde{\mathrm{N}}_i \right\|_{2,1} \max_{1 \leq k \leq n} \mathbb{P}_{\mathrm{R}} \mathbb{P}_{\varepsilon} \left\| \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \varepsilon_i \delta_{\mathbf{X}_{\mathrm{R}_i}} \right\|^*_{\mathscr{M}_{\delta,\delta'_1}} \\
&\leq \mathrm{C} \max_{1 \leq k \leq n} \mathbb{P}_{\mathrm{R}} \left(\mathbb{P}^{\mathrm{R}}_k \mathrm{M}_{\delta,\delta'_1} \right)^{1/2} \\
&\leq \mathrm{C} \left(\mathbb{P}_n \mathrm{M}_{\delta,\delta'_1} \right)^{1/2},
\end{aligned} \tag{3.3.4}$$

where $\mathrm{C} > \sqrt{\frac{n}{k}} \left\| \widetilde{\mathrm{N}}_i \right\|_{2,1}$ *see Problem 3.6.3 of van der Vaart and Wellner [1996]. By Jensen's inequality the outer expectation of the left side of (3.3.4) is bounded by*

$C\sqrt{\mathbb{P}[M_{\delta,\delta_1'}]^2}$, *for every* $\delta < \delta_1$. *The inequality in assumption* ***(BB2)*** *holds for every* $n \in \mathbb{N}$ *this implied by the fact that the variables we consider are i.i.d.*

(vi) *Finally, for the assumption* ***(BB3)*** *with the same discussion given in* ***(iii)*** *only the choice* $W_n = \left\|\Gamma\left(\boldsymbol{\theta}_0, \widehat{h}\right)\right\|$ *becomes*

$$W_m = \left\|\Gamma\left(\boldsymbol{\theta}_0, \widehat{h}_m\right)\right\|,$$

with $W_m = O_{\mathbb{P}^*_W}\left(\frac{1}{r_m^{\nu-1}\ell(r_m)}\right)$ *i.p.*

Theorem 3.3.2.2 *(i) Assume* ***(C)*** *and* ***(B1)-(B3)****. Then*

$$r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) = O_{\mathbb{P}^*}(1).$$

(ii) Assume ***(CP)*** *and* ***(BB1)-(BB3)****. Then*

$$r_m\left(\widehat{\boldsymbol{\theta}}^*_m - \boldsymbol{\theta}_0\right) = O_{\mathbb{P}^*_W}(1) \quad i.p.$$

Remark 3.3.2.3 *The result (i) of this Theorem still holds for* $\boldsymbol{\theta}$ *belongs to Banach space which is more general of the Theorem 2 of Delsol and Van Keilegom [2020], where the authors are interested in the finite dimensional case. Noting that the choice of* $\nu = 2$ *and* $\ell \equiv 1$ *in assumptions* ***B2*** *and* ***B3****, reduces to the assumptions* ***B2*** *and* ***B3*** *respectively of the last reference.*

3.3.3 Weak Convergence

We start this section by introducing some notation. For any $\boldsymbol{\theta} \in \Theta$ and $h \in \mathscr{H}$, let $\mathscr{K} = \left\{\boldsymbol{\gamma} \in \mathbf{E} : \|\boldsymbol{\gamma}\| \leq K\right\}$ for $K > 0$. Define, for sufficiently large n and for $\gamma \in \mathscr{K}$, the empirical processes

$$\begin{aligned} \mathbb{M}_n(\gamma, h) &= r_n^{\nu}\ell(r_n)(\mathbb{P}_n - \mathbb{P})\widetilde{\mathbf{m}}_{\gamma/r_n,h}, \\ \widehat{\mathbb{M}}_n(\gamma, h) &= r_m^{\nu}\ell(r_m)(\widehat{\mathbb{P}}^*_m - \mathbb{P}_n)\widetilde{\mathbf{m}}_{\gamma/r_m,h}, \end{aligned} \tag{3.3.5}$$

which can be treated as random elements in $\ell^\infty(\mathscr{K})$. Also let for any $\delta > 0$;

$$\begin{aligned} \mathbf{M}_\delta(\cdot) &\geq \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta} |\mathbf{m}(\cdot, \boldsymbol{\theta}, h_0) - \mathbf{m}(\cdot, \boldsymbol{\theta}_0, h_0)|, \\ \mathscr{M}_\delta &= \left\{\mathbf{m}(\cdot, \boldsymbol{\theta}, h_0) - \mathbf{m}(\cdot, \boldsymbol{\theta}_0, h_0) : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta\right\}, \\ \mathscr{M}_\delta(\eta) &= \left\{\mathbf{m}_{\boldsymbol{\theta},h_0} - \mathbf{m}_{\boldsymbol{\psi},h_0} : \|\boldsymbol{\theta} - \boldsymbol{\psi}\| < \eta, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \vee \|\boldsymbol{\psi} - \boldsymbol{\theta}_0\| < \delta, \boldsymbol{\theta}, \boldsymbol{\psi} \in \Theta\right\}. \end{aligned}$$

Finally, for any $p \in \mathbb{N}$ and any $f : \Theta \longrightarrow \mathbb{R}$ and for any $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p) \in \Theta^p$, denote

$$\overline{f}_{\boldsymbol{\gamma}} = (f(\boldsymbol{\gamma}_1), \ldots, f(\boldsymbol{\gamma}_p))^\top.$$

We give the set of assumptions for the asymptotic distribution of the processes given in (3.3.5) and their maximum.

(C1) $r_n\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\| = O_{\mathbb{P}^*}(1)$ and $v_n d_{\mathscr{H}}(\widehat{h}, h_0) = O_{\mathbb{P}^*}(1)$ for some sequences $r_n \longrightarrow \infty$ and $v_n \longrightarrow \infty$, and $r_n^{\nu-2}\ell(r_n) < C$ for some $C > 0$.

(C2) $\boldsymbol{\theta}_0$ lies to the interior of Θ, where $\Theta \subset (\mathscr{B}, \|\cdot\|)$.

(C3) For all $\delta_2, \delta_3 > 0$,

$$\sup_{\substack{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \le \frac{\delta_2}{r_n} \\ d_{\mathscr{H}}(h,h_0) \le \frac{\delta_3}{v_n}}} \frac{|(\mathbb{P}_n - \mathbb{P})\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h} + (\mathbb{P}_n - \mathbb{P})\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h_0}|}{r_n^{-\nu}\ell^{-1}(r_n) + |\mathbb{P}_n\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h}| + |\mathbb{P}_n\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h_0}| + |\mathbb{P}\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h}| + |\mathbb{P}\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h_0}|} = o_{\mathbb{P}^*}(1).$$

(C4) There exists a sequence $\{a_n\}$ with

$$a_m^2 m^{-1/2}\log n/\log(n/m+1) = o(1) \text{ and } a_n^{-1} = \mathrm{O}(1),$$

such that, for all $\mathrm{C}, \eta > 0$ and for every sequence $\{j_n\}$ with $a_n = o(j_n)$,

$$\frac{r_n^{2\nu}\ell^2(r_n)}{n}\mathbb{P}^*\left[\mathbf{M}^2_{\frac{\mathrm{C}}{r_n}}\right] = \mathrm{O}(1) \text{ and } \frac{r_n^{2\nu}\ell^2(r_n)}{n}\mathbb{P}^*\left[\mathbf{M}^2_{\frac{\mathrm{C}}{r_n}}\,\mathrm{1I}_{\left\{\mathbf{M}_{\frac{\mathrm{C}}{r_n}} > \frac{\eta j_n n^{1/2}}{r_n^{\nu}\ell(r_n)}\right\}}\right] = o(1).$$

(C5) For all K and for any $\eta_n \longrightarrow 0$,

$$\sup_{\|\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_2\|<\eta_n, \|\boldsymbol{\gamma}_1\|\vee\|\boldsymbol{\gamma}_2\|\le \mathrm{K}} \frac{r_n^{2\nu}\ell^2(r_n)}{n}\mathbb{P}\left[\mathbf{m}\left(\mathbf{X}, \boldsymbol{\theta}_0 + \frac{\boldsymbol{\gamma}_1}{r_n}, h_0\right) - \mathbf{m}\left(\mathbf{X}, \boldsymbol{\theta}_0 + \frac{\boldsymbol{\gamma}_2}{r_n}, h_0\right)\right]^2 = o(1).$$

(C6) For $\boldsymbol{x}$, the function $\boldsymbol{\theta} \mapsto \mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta}, h_0)$ and almost all paths of the two processes $\boldsymbol{\theta} \mapsto \mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta}, \widehat{h})$ and $\boldsymbol{\theta} \mapsto \mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta}, \widehat{h}_m)$ are uniformly bounded on closed bounded sets (over $\boldsymbol{\theta}$).

(C7) There exist a random and linear function $\mathrm{W}_n : \mathscr{B} \longrightarrow \mathbb{R}$, a deterministic and bilinear function $\mathrm{V} : \mathscr{B} \times \mathscr{B} \longrightarrow \mathbb{R}$ and $\beta_n = o_{\mathbb{P}^*}(1)$; such that for all $\boldsymbol{\theta} \in \Theta$;

$$\mathbb{P}\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,\widehat{h}} = \mathrm{W}_n(\boldsymbol{\gamma}_{\boldsymbol{\theta}}) + \mathrm{V}(\boldsymbol{\gamma}_{\boldsymbol{\theta}}, \boldsymbol{\gamma}_{\boldsymbol{\theta}}) + \beta_n\|\boldsymbol{\gamma}_{\boldsymbol{\theta}}\|^2 + o(\|\boldsymbol{\gamma}_{\boldsymbol{\theta}}\|^2)$$

and

$$\mathbb{P}\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h_0} = \mathrm{V}(\boldsymbol{\gamma}_{\boldsymbol{\theta}}, \boldsymbol{\gamma}_{\boldsymbol{\theta}}) + o(\|\boldsymbol{\gamma}_{\boldsymbol{\theta}}\|^2),$$

where $\boldsymbol{\gamma}_{\boldsymbol{\theta}} = \boldsymbol{\theta} - \boldsymbol{\theta}_0$ and the notation $o(\|\boldsymbol{\gamma}_{\boldsymbol{\theta}}\|^2)$ means

$$\lim_{\|\boldsymbol{\gamma}_{\boldsymbol{\theta}}\| \longrightarrow 0} \frac{o(\|\boldsymbol{\gamma}_{\boldsymbol{\theta}}\|^2)}{\|\boldsymbol{\gamma}_{\boldsymbol{\theta}}\|^2} = 0.$$

Moreover, for any bounded closed set $\mathscr{K} \subset \mathscr{B}$,

$$\exists \tau, \delta_1 > 0, r_n^{\nu-1}\ell(r_n) \sup_{\substack{\boldsymbol{\gamma}\in\mathscr{K}, \delta\le\delta_1 \\ \|\boldsymbol{\gamma}\|\le\delta}} \left|\frac{\mathrm{W}_n(\boldsymbol{\gamma})}{\delta^{\tau}}\right| = \mathrm{O}_{\mathbb{P}}(1) \text{ and } \sup_{\substack{\boldsymbol{\gamma},\boldsymbol{\gamma}'\in\mathscr{K}, \delta\le\delta_1 \\ \|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|\le\delta}} \frac{|\mathrm{V}(\boldsymbol{\gamma},\boldsymbol{\gamma}) - \mathrm{V}(\boldsymbol{\gamma}',\boldsymbol{\gamma}')|}{\delta^{\tau}} < \infty.$$

(C8) There exists a zero-mean Gaussian process $\mathbb{G}$ defined on $\mathscr{B}$ and a continuous function Λ such that for all $p \in \mathbb{N}$ and for all $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p) \in \mathscr{K}^p$,

$$r_n^{\nu-1}\ell(r_n)\overline{\mathrm{W}_n}_{\boldsymbol{\gamma}} + r_n^{\nu}\ell(r_n)\overline{\mathbb{P}_n\tilde{\mathbf{m}}_{\frac{\cdot}{r_n},h_0}}_{\boldsymbol{\gamma}} \rightsquigarrow \overline{\Lambda}_{\boldsymbol{\gamma}} + \overline{\mathbb{G}}_{\boldsymbol{\gamma}}.$$

Moreover, $\mathbb{G}(\boldsymbol{\gamma}) = \mathbb{G}(\boldsymbol{\gamma}')$ a.s. implies that $\boldsymbol{\gamma} = \boldsymbol{\gamma}'$, and

$$\mathbb{P}^*\left(\limsup_{\|\boldsymbol{\gamma}\|\longrightarrow\infty}(\Lambda(\boldsymbol{\gamma}) + \mathbb{G}(\boldsymbol{\gamma})) < \sup_{\boldsymbol{\gamma}\in\mathscr{B}}(\Lambda(\boldsymbol{\gamma}) + \mathbb{G}(\boldsymbol{\gamma}))\right) = 1.$$

(C9) There exists a $\delta_0 > 0$ such that

$$\int_0^{\infty} \sup_{\delta\le\delta_0}\sup_{\mathbb{Q}} \sqrt{\log\left(\mathrm{N}(\epsilon\|\mathbf{M}_\delta\|_{\mathbb{Q},2}, \mathscr{M}_\delta, \mathbb{L}^2(\mathbb{Q}))\right)}d\epsilon < \infty.$$

(C10) For all $\delta, \eta > 0$, the classes $\mathscr{M}_\delta, \mathscr{M}_\delta(\eta)$ and $\mathscr{M}_\delta(\eta)^2$ are $\mathbb{P}$-measurable, see [van der Vaart and Wellner, 1996, p.110] for definition of $\mathbb{P}$-measurability.

(C11) For all $\mathrm{C} > 0$, there exists $n_0 \in \mathbb{N}$ such that for all $n_0 \ge n$,

$$\mathbb{P}_n\mathbf{m}_{\boldsymbol{\theta}_n,\hat{h}} \ge \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\frac{\mathrm{C}}{r_n}} \mathbb{P}_n\mathbf{m}_{\boldsymbol{\theta}_0,\hat{h}} - \mathrm{R}_n,$$

where R_n is given in **(A5)**.

(CB1) $r_m\|\widehat{\boldsymbol{\theta}}_m^* - \boldsymbol{\theta}_0\| = \mathrm{O}_{\mathbb{P}_\mathrm{W}^*}(1)$ i.p. and $v_m d_{\mathscr{H}}(\widehat{h}_m, h_0) = \mathrm{O}_{\mathbb{P}_\mathrm{W}^*}(1)$ i.p. for some sequences $r_m \longrightarrow \infty$ and $v_m \longrightarrow \infty$ and $r_m^{\nu-2}\ell(r_m) \le \mathrm{C}$.

(CB2) For all $\delta_2, \delta_3 > 0$,

$$\sup_{\substack{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\frac{\delta_2}{r_m}\\ d_{\mathscr{H}}(h,h_0)\le\frac{\delta_3}{v_m}}} \frac{|(\widehat{\mathbb{P}}_m^* - \mathbb{P}_n)\tilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h} + (\widehat{\mathbb{P}}_m^* - \mathbb{P}_n)\tilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h_0}|}{r_m^{-\nu}\ell^{-1}(r_m) + |\mathbb{P}_n\tilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h}| + |\mathbb{P}_n\tilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h_0}| + |\widehat{\mathbb{P}}_m^*\tilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h}| + |\widehat{\mathbb{P}}_m^*\tilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h_0}|} = o_{\mathbb{P}^*}(1).$$

(CB3) There exists a random and linear function $\mathrm{W}_m : \mathscr{B} \longrightarrow \mathbb{R}$, and $\beta_m = o_{\mathbb{P}^*}(1)$, such that for all $\boldsymbol{\theta} \in \Theta$;

$$\mathbb{P}\tilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,\widehat{h}_m} = \mathrm{W}_m(\boldsymbol{\gamma}_{\boldsymbol{\theta}}) + \mathrm{V}(\boldsymbol{\gamma}_{\boldsymbol{\theta}}, \boldsymbol{\gamma}_{\boldsymbol{\theta}}) + \beta_n\|\boldsymbol{\gamma}_{\boldsymbol{\theta}}\|^2 + o(\|\boldsymbol{\gamma}_{\boldsymbol{\theta}}\|^2)$$

and

$$\mathbb{P}\tilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h_0} = \mathrm{V}(\boldsymbol{\gamma}_{\boldsymbol{\theta}}, \boldsymbol{\gamma}_{\boldsymbol{\theta}}) + o(\|\boldsymbol{\gamma}_{\boldsymbol{\theta}}\|^2).$$

Moreover, for any closed bounded set $\mathscr{K} \subset \mathbf{E}$,

$$\exists\tau, \delta_1 > 0, r_m^{\nu-1}\ell(r_m) \sup_{\substack{\boldsymbol{\gamma}\in\mathscr{K},\delta\le\delta_1\\ \|\boldsymbol{\gamma}\|\le\delta}} \left|\frac{\mathrm{W}_m(\boldsymbol{\gamma})}{\delta^\tau}\right| = \mathrm{O}_{\mathbb{P}_\mathrm{W}^*}(1) \text{ i.p.}, \sup_{\substack{\boldsymbol{\gamma},\boldsymbol{\gamma}'\in\mathscr{K},\delta\le\delta_1\\ \|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|\le\delta}} \frac{|\mathrm{V}(\boldsymbol{\gamma},\boldsymbol{\gamma}) - \mathrm{V}(\boldsymbol{\gamma}',\boldsymbol{\gamma}')|}{\delta^\tau} < \infty.$$

(CB4)

$$r_m^{\nu-1}\ell(r_m)\overline{\mathrm{W}_{m\boldsymbol{\gamma}}} + r_m^{\nu}\ell(r_m)\overline{\widehat{\mathbb{P}}_m^*\widetilde{\mathbf{m}}_{\frac{\cdot}{r_m},h_0}}_{\boldsymbol{\gamma}} \rightsquigarrow \overline{\Lambda}_{\boldsymbol{\gamma}} + \overline{\mathbb{G}}_{\boldsymbol{\gamma}} \text{ i.p.},$$

where Λ and $\mathbb{G}$ are given in **(C8)** and the weak convergence is conditionally on the sample.

(CB5) For all $C > 0$, there exist $m_0 \in \mathbb{N}$ such that for all $m \geq m_0$,

$$\widehat{\mathbb{P}}_m^*\mathbf{m}_{\widehat{\boldsymbol{\theta}}_m^*,\widehat{h}_m} \geq \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\frac{C}{r_m}} \widehat{\mathbb{P}}_m^*\mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}_m} - \widehat{\mathrm{R}}_n,$$

where $\widehat{\mathrm{R}}_n$ is given in **(AB3)**.

Remark 3.3.3.1 **(i)** *We can obtained the first part of condition **(C1)** by part (i) of Theorem 3.3.2.2.*

(ii) *Assumption **(C3)** holds under the common condition: for all $\delta_2, \delta_3 > 0$,*

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\frac{\delta_2}{r_n}, d_{\mathscr{H}}(h,h_0)\leq\frac{\delta_3}{v_n}} \left|(\mathbb{P}_n-\mathbb{P})\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h} + (\mathbb{P}_n-\mathbb{P})\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h_0}\right| = o_{\mathbb{P}^*}(r_n^{-\nu}\ell^{-1}(r_n)),$$

which is implied by the fact that; there exists a function f and a constant $\delta_0 > 0$ such that for all $\delta_2, \delta_3 < \delta_0$,

$$r_n^{\nu}\ell(r_n)f\left(\frac{\delta_2}{r_n}, \frac{\delta_3}{v_n}\right) = o\left(\sqrt{n}\right),$$

and

$$\begin{aligned}
&\mathbb{P}^*\left[\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\frac{\delta_2}{r_n}, d_{\mathscr{H}}(h,h_0)\leq\frac{\delta_3}{v_n}} \left|(\mathbb{P}_n-\mathbb{P})\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h} + (\mathbb{P}_n-\mathbb{P})\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h_0}\right|\right] \\
&\leq 2\mathbb{P}^*\left[\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\frac{\delta_2}{r_n}, d_{\mathscr{H}}(h,h_0)\leq\frac{\delta_3}{v_n}} \left|\mathbb{G}_n\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h}\right|\right] \\
&\leq \frac{1}{\sqrt{n}}f\left(\frac{\delta_2}{r_n}, \frac{\delta_3}{v_n}\right).
\end{aligned}$$

*Using the same arguments as in **Remark 3.3(ii)**, we get the last inequality.*

(iii) *If we assume that $j_n = \sqrt{n}$, and noting; $\gamma \mapsto \mathbb{M}_n(\gamma, h_0) = \frac{r_n^{\nu}\ell(r_n)}{\sqrt{n}}\mathbb{G}_n\widetilde{\mathbf{m}}_{\gamma/r_n,h_0}$ is the empirical process with indexed class $\frac{r_n^{\nu}\ell(r_n)}{\sqrt{n}}\mathscr{M}_{\frac{C}{r_n}}$ then, under assumption **(B2)**, the assumptions **(C4)** and **(C5)** hold by the following conditions: there exists a $\delta_4 > 0$ such that for all $\delta \leq \delta_4$, $\mathbb{P}^*(\mathbf{M}_\delta^2) \leq K\varphi^2(\delta)$ for some $C > 0$,*

$$\lim_{\delta\longrightarrow 0}\frac{\mathbb{P}^*\left[\mathbf{M}_\delta^2 1\!\!1_{\{\mathbf{M}_\delta > \eta\delta^{-2}\varphi^2(\delta)\}}\right]}{\varphi^2(\delta)} = 0,$$

for all $\eta > 0$ and

$$\lim_{\epsilon\longrightarrow 0}\lim_{\delta\longrightarrow 0}\sup_{\|\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_2\|<\epsilon, \|\boldsymbol{\gamma}_1\|\vee\|\boldsymbol{\gamma}_2\|\leq K}\frac{\mathbb{P}\left[\mathbf{m}\left(\mathbf{X}, \boldsymbol{\theta}_0+\boldsymbol{\gamma}_1\delta, h_0\right) - \mathbf{m}\left(\mathbf{X}, \boldsymbol{\theta}_0+\boldsymbol{\gamma}_2\delta, h_0\right)\right]^2}{\varphi^2(\delta)} = 0,$$

for all $C > 0$, corresponding the case to Theorem 3.2.10 in van der Vaart and Wellner [1996].

(iv) *Let $\mathscr{K}$ be an arbitrary closed bounded subset in $\mathscr{B}$, the first part of condition **(C8)** is used to assume the convergence of the marginals of the process $\gamma \mapsto r_n^{\nu-1}\ell(r_n)\mathrm{W}_n(\gamma)+ r_n^{\nu}\ell(r_n)\mathbb{P}_n\widetilde{\mathbf{m}}_{\frac{\gamma}{r_n},h_0}$ for deriving its weak convergence in $\ell^\infty(\mathscr{K})$ by the fact that it is asymptotically tight; which is fulfilling by using **(C4)**, **(C5)**, **(C9)** and the preceding discussion in **(iii)**. If*

$$r_n^{\nu-1}\ell(r_n)\sup_{\boldsymbol{\gamma}\in\mathscr{K},\boldsymbol{\gamma}\neq 0}\|\mathrm{W}_n(\boldsymbol{\gamma})\|\boldsymbol{\gamma}\|^{-1}\| = o_{\mathbb{P}}(1),$$

we treat the given process as in the parametric case, where its marginals converge provided that

$$\begin{aligned}\lim_{n\to\infty}\frac{r_n^{2\nu}\ell^2(r_n)}{n}\mathbb{P}\left\{\left[\mathbf{m}\left(\mathbf{X},\boldsymbol{\theta}_0+\frac{\gamma_1}{r_n},h_0\right)-\mathbf{m}\left(\mathbf{X},\boldsymbol{\theta}_0+\frac{\gamma_2}{r_n},h_0\right)\right]^2\right\}\\ =\mathbb{P}\left[\left(\mathbb{G}\left(\gamma_1\right)-\mathbb{G}\left(\gamma_2\right)\right)^2\right],\end{aligned}$$

for all γ_1,γ_2 and we lead to a rate of convergence r_n as the solution of

$$r_n^{\nu}\ell(r_n)\varphi(1/r_n)=\sqrt{n},$$

for more detail see Theorem 3.2.10 of van der Vaart and Wellner [1996]. Note that almost all sample paths of the process $\boldsymbol{\gamma}\mapsto\Lambda(\boldsymbol{\gamma})+\mathbb{G}(\boldsymbol{\gamma})$ have a supremum affiliated to their attitude on bounded closed set, which is guaranteed by the last assumption. The dominant term of the deterministic part Λ is usually a negative definite quadratic form and hence exponential inequalities could lead to such result, for example when we are in the situation of the smooth function, one can refer to Lee and Pun [2006], Ma and Kosorok [2005], Kosorok [2008], Kristensen and Salanié [2017] among many others.

(v) *Assumption **(C9)** is a technical assumption, which is the same in the parametric case where the nuisance parameter h_0 is known, needs to show that; the process $\boldsymbol{\gamma}\mapsto r_n^{\nu}\ell(r_n)\mathbb{P}_n\widetilde{\mathbf{m}}_{\frac{\gamma}{r_n},h_0}$ is asymptotically tight, see Theorem 3.2.10 of van der Vaart and Wellner [1996].*

(vi) *First part of **(CB1)** follows by part (ii) of Theorem 3.3.2.2.*

(vii) *Assumption **(CB2)** is automatically hold under the condition : for all $\delta_2,\delta_3>0$,*

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\frac{\delta_2}{r_m},d_{\mathscr{H}}(h,h_0)\leq\frac{\delta_3}{v_m}}\left|(\widehat{\mathbb{P}}_m^*-\mathbb{P}_n)\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h}+(\widehat{\mathbb{P}}_m^*-\mathbb{P}_n)\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h_0}\right|=o_{\mathbb{P}_{\mathrm{W}}^*}(r_m^{-\nu}\ell^{-1}(r_m))\ i.p.$$

This condition is hold if: there exists a function g and a constant $\delta_0>0$ such that for all $\delta_2,\delta_3<\delta_0$,

$$r_m^{\nu}\ell(r_m)g\left(\frac{\delta_2}{r_m},\frac{\delta_3}{v_m}\right)=o\left(\sqrt{m}\right),$$

and

$$\begin{aligned}
&\mathbb{P}^*\mathbb{P}_{\mathrm{W}}^*\left[\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\frac{\delta_2}{r_m},d_{\mathscr{H}}(h,h_0)\leq\frac{\delta_3}{v_m}}\left|(\widehat{\mathbb{P}}_m^*-\mathbb{P}_n)\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h}+(\widehat{\mathbb{P}}_m^*-\mathbb{P}_n)\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h_0}\right|\right]\\
&\leq 2\mathbb{P}^*\mathbb{P}_{\mathrm{W}}^*\left[\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\frac{\delta_2}{r_m},d_{\mathscr{H}}(h,h_0)\leq\frac{\delta_3}{v_m}}\left|\widehat{\mathbb{G}}_m^*\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h}\right|\right]\\
&\leq \frac{1}{\sqrt{m}}g\left(\frac{\delta_2}{r_m},\frac{\delta_3}{v_m}\right).
\end{aligned}$$

*Using the same arguments as in Remark **3.3(v)**, we get the last inequality.*

(viii) *Following similar discussion of the condition **(C7)** provided in **Remark 3(iv)** of Delsol and Van Keilegom [2020], we only change the random function $\mathrm{W}_n(\boldsymbol{\gamma})$ for the bootstrap version to $\mathrm{W}_m(\gamma)=\langle\Gamma(\boldsymbol{\theta}_0,\widehat{h}_m),\boldsymbol{\gamma}\rangle$. If we are in the situation where $\widehat{h}_m$ is calculated from a dataset independently from the bootstrapped sample $(\mathbf{X}_1^*,\ldots,\mathbf{X}_m^*)$, so it is sufficient for assumption **(CB4)** to suppose the conditional weak convergence of each term; $r_m^{\mathrm{v}-1}\ell(r_m)\overline{\mathrm{W}_{m\boldsymbol{\gamma}}}$ and $r_m^{\mathrm{v}}\ell(r_m)\overline{\widehat{\mathbb{P}}_m^*\widetilde{\mathbf{m}}_{\frac{\cdot}{r_n},h_0}}_{\boldsymbol{\gamma}}$ separately. We can get the convergence of the second one as the same in the situation without the nuisance parameter, the interested reader is referred to Lemma 1 of Lee [2012]. Note that if $r_m^{\mathrm{v}-1}\ell(r_m)\Gamma(\boldsymbol{\theta}_0,\widehat{h}_m)\rightarrow\mathrm{W}$ conditionally in distribution, the marginals of the process $\gamma\mapsto\langle r_m^{\mathrm{v}-1}\ell(r_m)\Gamma(\boldsymbol{\theta}_0,\widehat{h}_m),\gamma\rangle$ tend in distribution to the marginals of $\gamma\mapsto\langle\mathrm{W},\gamma\rangle$. Furthermore, if $r_m=\sqrt{m}$ and $\ell\equiv 1$, it is common to assume that*

$$\Gamma(\boldsymbol{\theta}_0,\widehat{h}_m)=m^{-1}\sum_{i=1}^{m}\mathrm{U}_{i,m}+o_{\mathbb{P}_{\mathrm{W}}^*}\left(m^{-1/2}\right),$$

where $\mathrm{U}_{i,m}, i=1,\ldots,m$, are independent and centered random variables. The convergence follows from Lindeberg's condition.

Theorem 3.3.3.2 *(Weak Convergence of Empirical Processes). For all $\mathrm{K}>0$, let $\mathscr{K}=\{\boldsymbol{\gamma}\in\mathbf{E}:\|\boldsymbol{\gamma}\|\leq\mathrm{K}\}$ be a closed bounded subset of $\mathscr{B}$, treating $\boldsymbol{\gamma}\mapsto\mathbb{M}_n(\gamma,\widehat{h})$ and $\boldsymbol{\gamma}\mapsto\widehat{\mathbb{M}}_n(\gamma,\widehat{h}_m)$ as random elements in $\ell(\mathscr{K})$ for sufficiently large n, we have the following results:*

*(i) Assume **(C1)-(C10)**. Then*

$$r_n^{\mathrm{v}}\ell(r_n)\mathbb{P}_n\widetilde{\mathbf{m}}_{\frac{\gamma}{r_n},\widehat{h}}\rightsquigarrow\Lambda(\boldsymbol{\gamma})+\mathbb{G}(\boldsymbol{\gamma}).$$

*(ii) Assume **(A2)**, **(AB1)**, **(B2)**, **(C2)-(C6)**,**(C9)- (C11)** and **(CB1)-(CB4)**. Then*

$$r_m^{\mathrm{v}}\ell(r_m)\widehat{\mathbb{P}}_m^*\widetilde{\mathbf{m}}_{\frac{\gamma}{r_m},\widehat{h}_m}\rightsquigarrow\Lambda(\boldsymbol{\gamma})+\mathbb{G}(\boldsymbol{\gamma})\ \ i.p.$$

Our main results concerning weak convergence of $r_n(\boldsymbol{\theta}_n-\boldsymbol{\theta}_0)$ and m out of n bootstrap consistency are embodied in the following theorem.

Theorem 3.3.3.3 *Assume for any such $\mathscr{K}$ that almost every sample path of the process $\boldsymbol{\gamma} \mapsto \Lambda(\boldsymbol{\gamma}) + \mathbb{G}(\boldsymbol{\gamma})$ achieves its supremum at a unique random point $\gamma_0 = \arg\max_{\gamma \in \mathscr{B}} \Lambda(\boldsymbol{\gamma}) + \mathbb{G}(\boldsymbol{\gamma})$, then;*

*(i) Assume **(C1)-(C11)**. Then*

$$r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \rightsquigarrow \gamma_0.$$

*(ii) Assume **(A2)**, **(AB1)**, **(B2)**, the first part of **(C1)**, **(C2)-(C6)**,**(C9)**- **(C11)** and **(CB1)-(CB5)**. Then*

$$r_m\left(\widehat{\boldsymbol{\theta}}_m^* - \boldsymbol{\theta}_n\right) \rightsquigarrow \gamma_0 \ \ i.p.$$

Remark 3.3.3.4 *The result (i) of the Theorem 3.3.3.2 is the same result of Lemma 1 of Delsol and Van Keilegom [2020] where the parameter of the interest $\boldsymbol{\theta}$ is in a Euclidean space, for the particular case $\nu = 2$ and $\ell \equiv 1$, then by the application of Theorem 3.2.2 of van der Vaart and Wellner [1996] and the uniform tightness of the sequence $r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$, the authors established the weak convergence to some tight random variable γ_0 in $\ell^\infty(\mathscr{K})$ for the compact set $\mathscr{K}$ in their Theorem 3 which is given in the result (i) of the Theorem 3.3.3.3 in this case. In our setting, we provide the weak convergence of the same sequence for the Banach valued parameter by using Theorem of van der Vaart and Wellner [1996] where the compact sets and the uniformed tightness of $r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$ are replaced, respectively, by closed bounded sets with a similar structure as the set $\mathscr{K}$ and $r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) = O_{\mathbb{P}^*}(1)$, as given in Lee [2012] without the nuisance parameter h_0.*

Note that (i) still holds if **(C4)** is replaced by this more weak condition

$$n^{-1} r_n^{2\nu} \ell(r_n)^2 \mathbb{P}^* \mathbf{M}_{c/r_n}^2 \left\{ \mathbf{M}_{c/r_n} > \eta n r_n^{-\nu} \ell^{-1}(r_n) \right\} \to 0.$$

In order to prove the conditional stochastic equicontinuity of the bootstrapped process $\widehat{\mathbb{M}}_n$ we need the condition **(C4)**, that is fulfilled if the uniform integrability condition is imposed for $j_n \geq n^c$, for some $0 < c < 1/4$.

Remark 3.3.3.5 *It is well known that Theorem 3.3.3.3 can be used easily through routine bootstrap sampling, which we describe briefly as follows. More precisely, this can be used, for example, to form confidence bands for the true parameter $\boldsymbol{\theta}$ based* N*, be a large integer, sampled samples $\mathbf{Y}_1^k, \ldots, \mathbf{Y}_m^{(k)}$, $k = 1, \ldots, \mathfrak{N}$. Let $\left(\widehat{\boldsymbol{\theta}}_m^*\right)^{(k)}$ the bootstrapped estimator of $\boldsymbol{\theta}$ based on the sample $\mathbf{Y}_1^{(k)}, \ldots, \mathbf{Y}_m^{(k)}$, $k = 1, \ldots, \mathfrak{N}$. An application of Theorem 3.3.3.3 implies that*

$$\left(\widehat{\boldsymbol{\theta}}_m^*\right)^{(1)}$$

$$\left(r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0), r_m\left(\left(\widehat{\boldsymbol{\theta}}_m^*\right)^{(1)} - \boldsymbol{\theta}_n\right), \ldots, r_m\left(\left(\widehat{\boldsymbol{\theta}}_m^*\right)^{(\mathfrak{N})} - \boldsymbol{\theta}_n\right)\right) \rightsquigarrow \left(\gamma_0, \gamma_0^{(1)}, \ldots, \gamma_0{}^{(\mathfrak{N})}\right) \ \ i.p.,$$

where $\gamma_0^{(1)}, \ldots, \gamma_0^{(\mathfrak{N})}$ are independent copies of γ_0. Notice that we have

$$\lim_{n \to \infty} \mathbb{P}\left(\boldsymbol{\theta}_n - r_n^{-1} c(\alpha) \leq \boldsymbol{\theta}_0 \leq \boldsymbol{\theta}_n + r_n^{-1} c(\alpha)\right) = \mathbb{P}(|\gamma_0| \leq c(\alpha)) = 1 - \alpha.$$

In order to approximate $c(\alpha)$*, one can use the sampling estimator* $\widehat{c}(\alpha)$*, of* $c(\alpha)$*, as the smallest* z *such that*

$$\frac{1}{\mathfrak{N}}\sum_{k=1}^{\mathfrak{N}} \mathbb{1}_{\left\{r_m\left(\left(\widehat{\boldsymbol{\theta}}_m^*\right)^{(k)}-\boldsymbol{\theta}_n\right)\leq z\right\}} \geq 1-\alpha.$$

Remark 3.3.3.6 *In nonregular problems where the conventional* n *out of* n *bootstrap is inconsistent, the* m *out of* n *bootstrap provides a useful remedy to restore consistency. In practice, however, choosing an appropriate* m *needs careful attention. Asymptotically,* $\sqrt{n}, \log n$ *or* $20\log n$ *satisfy the* $o(n)$ *requirement, but in finite sample settings the actual results can vary dramatically depending on the choice. Let* $\mathscr{X}_n = (\mathrm{X}_1, \ldots, \mathrm{X}_n)$ *be a random sample drawn from an unknown distribution* F*, and* $\mathrm{T}_n(\mathscr{X}_n, \mathrm{F})$ *be a statistical functional of interest. Under mild conditions the* m *out of* n *bootstrap distribution* $\mathscr{L}^*_{m,n}$ *provides a consistent estimator of the distribution* $\mathscr{L}_n$ *of* $\mathrm{T}_n(\mathscr{X}_n, \mathrm{F})$*, provided that the bootstrap sample size* m *is properly chosen, refer to Götze and Račkauskas [2001] and Bickel* et al. *[1997]. Empirical selection of* m *has long been an important problem, which has been discussed by, for example, Datta and McCormick [1995], Hall* et al. *[1995] and Politis* et al. *[1999a] in different contexts. The prevailing idea is to estimate a theoretically optimal sample size* m*, defined in a frequentist sense to be the value of* m *which minimises the expected value of some metric measure* $d\left(\mathscr{L}_n, \mathscr{L}^*_{m,n}\right)$ *between* $\mathscr{L}_n$ *and* $\mathscr{L}^*_{m,n}$*. The problem can be solved using bootstrap samples of size* m*, where* $m \to \infty$ *and* $m/n \to 0$*. Bickel and Sakov [2008] proposed an adaptive rule to select a value* $\widehat{m}$ *and discuss its properties. The authors show, under some conditions, that* $\widehat{m}/n \xrightarrow{\mathrm{P}} 1$ *when the* n *bootstrap works, but* $\widehat{m} \to \infty$ *and* $\widehat{m}/n \to 0$ *when the* n*-bootstrap does not work. More precisely, the authors suggested the following rule for choosing* m*:*

1. *Consider a sequence of* m*'s of the form*

$$m_j = \left\lfloor q^j n \right\rfloor, \quad \text{for } j = 0, 1, 2, \ldots, \quad 0 < q < 1,$$

   where $\lfloor\alpha\rfloor$ *denotes the smallest integer* $\geq \alpha$*.*

2. *For each* m_j*, find* $\mathrm{L}^*_{m_j,n}$ *(in practice this is done by Monte-Carlo).*

3. *Let* d *be some metric consistent with convergence in law, and set*

$$\hat{m} = \underset{m_j}{\operatorname{argmin}}\, d\left(\mathrm{L}^*_{m_j,n}, \mathrm{L}^*_{m_{j+1},n}\right).$$

If there is more than one value of m *which minimizes the difference, then we pick the largest one. These results mean that the rule behaves well under both situations. Swanepoel [1986] proposed* $m = (2/3)n$ *to obtain the desired coverage probability of a confidence interval. Alin* et al. *[2017] have considered* $m = n^j$ *where the value* j *satisfies* $n^j = \frac{2}{3}n$*. Solving this equation for* j*, this expression leads to the choice*

$$m = n^j \quad \text{for } j = 1 + \frac{\log\left(\frac{2}{3}\right)}{\log(n)},$$

for which we note $0 < j < 1$, *so that* $m = o(n)$. *Götze and Račkauskas [2001] have suggested the estimation* m *by minimising* $d\left(\mathscr{L}^*_{m,n}, \mathscr{L}^*_{m/2,n}\right)$, *yielding an optimal bootstrap sample size in the sense of Wei* et al. *[2016], provided that the latter has order* $o_{\mathbb{P}}(n)$. *Wei* et al. *[2016] have investigated stochastic version of the optimal bootstrap sample size, defined as the minimiser of an error measure calculated directly from the observed sample. The authors have developed procedures for calculating the stochastically optimal value of* m. *The performance of their methodology is illustrated in the special forms of Edgeworth-type expansions which are typically satisfied by statistics of the shrinkage type.*

Remark 3.3.3.7 *An alternative approach, known as subsampling, uses without-replacement subsamples instead of with-replacement bootstrap samples to estimate the limiting distribution. Unlike the* m *out of* n *bootstrap, the consistency of which derives from a notion of local uniform continuity on the space of distribution functions, validity of subsampling follows from the asymptotics of a U-statistic of degree* m, *and consistency can be proved under minimal conditions (see, e.g., Politis and Romano [1994] and Politis, Politis* et al. *[1999b] for a general exposition of subsampling), Thus, subsampling is more general than the m-bootstrap since fewer assumptions are required. However, the m-bootstrap has the advantage that it allows for the choice of* $m = n$. *In particular, if the n-bootstrap works and is known to be second order correct for some pivotal roots, the selection rule for* m *includes the particular case* $m/n \to 1$. *In that case, unlike subsampling, the m-bootstrap enjoys the second order properties of the n-bootstrap. We mention that the higher-order asymptotic results are clearly essential for a detailed comparison of the two approaches under conditions when they are both consistent. Results so far are only sporadic for either approach, however. Under regularity conditions, both approaches suffer from a loss of efficiency, which can be recovered to some extent by extrapolation (Bickel* et al. *[1997], Politis* et al. *[1999b]). Bickel and Sakov [2008] studied the effects of extrapolation also for nonregular cases that admit Edgeworth expansions of a particular form. We note that extrapolation is easier to implement on the* m *out of* n *bootstrap than on subsampling, for which a finite-population correction factor is explicitly required. Since in all situations of interest, so far, the conditions for consistency of the m-bootstrap are satisfied, we consider only the sampling with replacement case. It is easily seen that if* $m = o(n^{1/2})$, *then ties in bootstrap samples are asymptotically negligible, and the two approaches are equivalent to first order. For more details, for instance, we refer Lee and Pun [2006], Romano and Shaikh [2012], Bertail [1997], Bertail* et al. *[1999] and Politis* et al. *[2001].*

3.4 Applications

We present in this section some examples which can not handled with the classical theory of semiparametric estimators and their m out of n bootstrap version cannot be applied while theory of the paper can be applied. This illustrates the usefulness of our results. Delsol and Van Keilegom [2020] provided some examples of situations in which the existing theory on semiparametric estimators cannot be applied, whereas their result could be applied. It is worth

noticing that the aim of this section is to verify the bootstrap conditions that are different from those used for the non bootstrapped estimators checked in the last mentioned reference. Although only three examples will be given here, they stand as archetypes for a variety of models that can be investigated by the methodology of the present paper.

3.4.1 Single index model with monotone link function

The single index regression models are typical examples which are given

$$\mathrm{Y} = g\left(\mathbf{X}^{\top}\boldsymbol{\beta}\right) + \varepsilon \tag{3.4.1}$$

where $\mathbb{P}(\varepsilon|\mathbf{X}) = 0, \mathrm{Var}(\varepsilon|\mathbf{X}) < \infty$ and we assume that the unknown function $g(\cdot)$ is monotone, we refer to Ichimura [1993] for more details. On the basis of the sample $(\mathbf{X}_1, \mathrm{Y}_1), \ldots, (\mathbf{X}_n, \mathrm{Y}_n)$ coming from the model (3.4.1), we make use of the the pool-adjacent-violators algorithm to construct and estimator of the function $g(\cdot)$. This gives a non-smooth estimator $\widehat{g}_{\boldsymbol{\beta}}(\cdot)$ of $g_{\boldsymbol{\beta}}(\mathbf{z}) = \mathbb{E}\left[\mathrm{Y}|\mathbf{X}^{\top}\boldsymbol{\beta} = \mathbf{z}\right]$. Next, by using the least-squares estimation method we estimate $\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left[-n^{-1} \sum_{i=1}^{n} \left(\mathrm{Y}_i - \widehat{g}_{\boldsymbol{\beta}}\left(\mathbf{X}_i^{\top}\boldsymbol{\beta}\right)\right)^2 \right].$$

By the fact that $\widehat{g}_{\boldsymbol{\beta}}(\cdot)$ is of non-smooth nature implies that the criterion function is not smooth in $\boldsymbol{\beta}$. This is a situation where the theory of the present paper can be applied.

3.4.2 Classification with missing data

Let $\mathbf{X}_1 = (\mathbf{X}_{11}, \mathbf{X}_{12}), \ldots, \mathbf{X}_n = (\mathbf{X}_{n1}, \mathbf{X}_{n2})$ be independent and identically distributed random copies of the random vector $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, coming from two underlying populations. For $j = 0, 1$, let $\mathbf{Y}_i = j$ when the $\mathbf{X}_i$ comes from the population j. Let us denote by $\mathbf{Y}$ the population indicator associated with the vector $\mathbf{X}$. Using the information of available data, we seek to find a classification method for novel observations with unknown true population.

The classification is performed by regressing $\mathbf{X}_2$ on $\mathbf{X}_1$ making use of the parametric criterion function $f_{\boldsymbol{\theta}}(\cdot)$, and choosing $\boldsymbol{\theta}$ that maximize the following

$$\mathbb{P}\,1\!\!1_{\{\mathbf{Y}=1, \mathbf{X}_2 \geq f_{\boldsymbol{\theta}}(\mathbf{X}_1)\}} + \mathbb{P}\,1\!\!1_{\{\mathbf{Y}=0, \mathbf{X}_2 < f_{\boldsymbol{\theta}}(\mathbf{X}_1)\}}. \tag{3.4.2}$$

Let $\boldsymbol{\theta}_0$ denote the maximizer of (3.4.2) with respect to all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, here $\boldsymbol{\Theta}$ is assumed to be a compact subset of $\mathbb{R}^k$ containing as an interior point $\boldsymbol{\theta}_0$. Now assume that $\mathbf{Y}_i$'s are subject to some missing mechanism. Let $\Delta_i$be a random variable (respectively Δ) equals to 1 when we observe the random variable $\mathbf{Y}_i$ (respectively $\mathbf{Y}$), and 0 otherwise. Let $\mathbf{Z}_1 = (\mathbf{X}_1, \mathbf{Y}_1\Delta_1, \Delta_1), \ldots, \mathbf{Z}_n = (\mathbf{X}_n, \mathbf{Y}_n\Delta_n, \Delta_n)$ be the observations at hand. The missing at random mechanism in considered in the following sense

$$\mathbb{P}\left(1\!\!1_{\{\Delta=1\}}|\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}\right) = \mathbb{P}\left(1\!\!1_{\{\Delta=1\}}|\mathbf{X}_1\right) := p_0\left(\mathbf{X}_1\right).$$

Note that the relation (3.4.2) can be written

$$\mathbb{E}\left[\frac{1\!\!1_{\{\Delta=1\}}}{p_0(\mathbf{X}_1)}\left\{1\!\!1_{\{\mathbf{Y}=1,\mathbf{X}_2\geq f_{\boldsymbol{\theta}}(\mathbf{X}_1)\}}+1\!\!1_{\{\mathbf{Y}=0,\mathbf{X}_2<f_{\boldsymbol{\theta}}(\mathbf{X}_1)\}}\right\}\right].$$

We define

$$\mathbf{m}_{\boldsymbol{\theta},p}(\mathbf{Z})=\frac{1\!\!1_{\{\Delta=1\}}}{p(\mathbf{X}_1)}\left\{1\!\!1_{\{\mathbf{Y}=1,\mathbf{X}_2\geq f_{\boldsymbol{\theta}}(\mathbf{X}_1)\}}+1\!\!1_{\{\mathbf{Y}=0,\mathbf{X}_2<f_{\boldsymbol{\theta}}(\mathbf{X}_1)\}}\right\},$$

here the infinite dimensional nuisance parameter $p(\cdot)$ belonging to some functional space $\mathscr{P}$ to be specified later. Consequently, the estimator $\boldsymbol{\theta}_n$ of $\boldsymbol{\theta}_0$ is given by

$$\boldsymbol{\theta}_n=\arg\max_{\boldsymbol{\theta}\in\theta}\mathbb{P}_n\mathbf{m}_{\boldsymbol{\theta},\widehat{p}},$$

where, for any x and a bandwidth sequence $h=h_n$,

$$\widehat{p}(x)=\sum_{i=1}^{n}\frac{\mathrm{K}_h(x-\mathbf{X}_{i1})}{\sum_{j=1}^{n}\mathrm{K}_h(x-\mathbf{X}_{j1})}1\!\!1_{\{\Delta_i=1\}},$$

where the kernel function $\mathrm{K}(\cdot)$ is assumed to be a density function having support $[-1,1]$, $\mathrm{K}_h(u)=\frac{\mathrm{K}(\frac{u}{h})}{h}$. Nonparametric regressions with missing have long attracted a great deal of attention, for good sources of references to research literature in this area along with statistical applications consult Müller [2009], Pérez-González *et al.* [2009] and Koul *et al.* [2012] among many others.

3.4.3 Binary choice model with missing data

Let us define the binary choice model, in the linear regression function framework, by

$$\begin{cases}\mathrm{U} &= \mathbf{X}^\top\boldsymbol{\beta}-\varepsilon,\\ \mathrm{Y} &= 1\!\!1(\mathrm{U}\geq 0),\end{cases}$$

where we assume that ε is zero median conditionally on $\mathbf{X}$. The random variable Y is missing at random with the probability, to observe Y, depending on $\mathbf{X}$ via the following relation

$$\mathbb{P}(1\!\!1_{\{\Delta=1\}}|\mathbf{X},\mathrm{Y})=\mathbb{P}\left(1\!\!1_{\{\Delta=1\}}|\mathbf{X}^\top\gamma\right):=p\left(\mathbf{X}^\top\gamma\right),$$

where $\Delta=1$ when we observe Y and 0 elsewhere. The observed data for the preceding model are given by of i.i.d. triplets $(\mathbf{X}_1,\mathrm{Y}_1\Delta 1,\Delta_n),\ldots,(\mathbf{X}_n,\mathrm{Y}_n\Delta_n,\Delta_n)$. To estimate $p_\gamma(z)=\mathbb{P}\left(1\!\!1_{\{\Delta=1\}}|\mathbf{X}^\top\boldsymbol{\gamma}=z\right)$, we use the following

$$\widehat{p}_{\boldsymbol{\gamma}}(z)=\sum_{i=1}^{n}\frac{\mathrm{K}_h\left(\mathbf{X}_i^\top\boldsymbol{\gamma}-z\right)}{\sum_{j=1}^{n}\mathrm{K}_h\left(\mathbf{X}_j^\top\boldsymbol{\gamma}-z\right)}1\!\!1_{\{\Delta_i=1\}}.$$

The parameter estimate is given by

$$(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})=\arg\max_{\boldsymbol{\beta},\boldsymbol{\gamma}}\mathbb{P}_n\mathbf{m}_{\boldsymbol{\beta},\boldsymbol{\gamma},\widehat{p}_{\boldsymbol{\gamma}}},$$

where

$$\mathbf{m}_{\boldsymbol{\beta},\boldsymbol{\gamma},p} = \frac{1\!\!1_{\{\Delta_i=1\}}}{p\left(\mathbf{X}_i^\top \boldsymbol{\gamma}\right)}\left[2 1\!\!1_{\{Y_i=1\}} - 1\right] 1\!\!1_{\{\mathbf{X}_i^\top \boldsymbol{\beta}\geq 0\}}.$$

The existing theory cannot be applied here by the fact that the function $\mathbf{m}_{\boldsymbol{\beta},\boldsymbol{\gamma},p}$ is smooth in $\boldsymbol{\gamma}$ but non-smooth in $\boldsymbol{\beta}$.

Now we will study in full detail the example in 3.4.2 and we work out the verification of the conditions of Theorems 3.3.1.2, 3.3.2.2, 3.3.3.2 and 3.3.3.3 the most of this conditions verified in Section 7 of Delsol and Van Keilegom [2020] by noting that $\nu = 2$ and $\ell \equiv 1$, so our focuses is to verify the conditions needed for the m out of n bootstrapped version. In the beginning we give some information about the nuisance function and her space and some notation. Let $\mathscr{P}$ be the space of functions $p : \mathbf{R}_{\mathbf{X}_1} \to \mathbb{R}$ that are continuously differentiable, for which

$$\sup_{\mathbf{x}_1 \in \mathbf{R}_{\mathbf{X}_1}} p(\mathbf{x}_1) \leq M < \infty, \ \sup_{\mathbf{x}_1 \in \mathbf{R}_{\mathbf{X}_1}} |p^{'}(\mathbf{x}_1)| \leq M \ \text{ and } \ \inf_{\mathbf{x}_1 \in \mathbf{R}_{\mathbf{X}_1}} p(\mathbf{x}_1) > \eta/2,$$

where

$$\eta = \inf_{\mathbf{x}_1 \in \mathbf{R}_{\mathbf{X}_1}} p_0(\mathbf{x}_1) > 0$$

and $\mathbf{R}_{\mathbf{X}_1}$ is the support of $\mathbf{X}_1$, where we suppose it is a compact subspace of $\mathbb{R}$. We equip the space $\mathscr{P}$ with the supremum norm:

$$d_{\mathscr{P}}(p_1, p_2) = \sup_{\mathbf{x}_1 \in \mathbf{R}_{\mathbf{X}_1}} |p_1(\mathbf{x}_1) - p_2(\mathbf{x}_1)| \ \text{ for any } \ p_1, p_2 \in \mathscr{P}.$$

After, the conditions of the consistency are verified as follows, **(A1)** holds true provided the functions $p_0(\cdot)$ and $K(\cdot)$ are continuously differentiable. For assumption **(A2)** we can showing that the bracketing number of the class $\mathscr{F} = \{\mathbf{m}_{\boldsymbol{\theta},p}, \boldsymbol{\theta} \in \Theta, p \in \mathscr{P}\}$; $N_{[\,]}(\epsilon, \mathscr{F}, \mathbb{L}_{\mathbb{P}})$ is finite for all $\epsilon > 0$, by using Corollary 2.7.2 of van der Vaart and Wellner [1996], we get

$$N_{[\,]}(\epsilon, \mathscr{P}, \mathbb{L}_{\mathbb{P}}) \leq \exp\{\mathfrak{K}\epsilon^{-1}\}, \tag{3.4.3}$$

and

$$N_{[\,]}\left(\epsilon, \{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}, \mathbb{L}_{\mathbb{P}}\right) \leq \exp\{\mathfrak{K}\epsilon^{-1}\},$$

by the properties of the set $\mathscr{P}$ and the fact that $\mathbf{x} \mapsto f_{\boldsymbol{\theta}}(\mathbf{x})$ is continuously differentiable over $\boldsymbol{\theta}$ with bounded derivative and as a consequence it's easily to show that

$$N_{[\,]}(\epsilon, \mathscr{T}, \mathbb{L}_{\mathbb{P}}) \leq \exp\{\mathfrak{K}\epsilon^{-1}\}, \tag{3.4.4}$$

for the class $\mathscr{T} = \left\{(\mathbf{x}_1, \mathbf{x}_2) \to 1\!\!1_{\{x_2 \geq f_{\boldsymbol{\theta}}(\mathbf{x}_1)\}} : \boldsymbol{\theta} \in \Theta\right\}$. From (3.4.3) and (3.4.4) we get;

$$N_{[\,]}(\epsilon, \mathscr{F}, \mathbb{L}_{\mathbb{P}}) \leq \exp\{\mathfrak{K}\epsilon^{-1}\}.$$

Then assumption **(A3)** is straightforward. Assumption **(A4)** is an identifiability condition to ensure the uniqueness of $\boldsymbol{\theta}_0$ and **(A5)** is verified by construction of the estimator $\boldsymbol{\theta}_n$. The consistency of $\boldsymbol{\theta}_n$ is then follows. For the conditions of the bootstrap version they are verified as follows; fist part of assumption **(AB1)** is satisfied by definition of the m out of n bootstrap,

where the second part in this situation follows directly by noting that if $r_n = n^{\kappa}$, we get $r_m = m^{\kappa}$ for some $\kappa > 0$, by consequent we have $r_m^2 = o(r_n^2)$. For **(AB2)** as mentioned in **remark 3.1(v)** we take $\widehat{p}_m(\cdot) = \widehat{p}(\cdot)$ where we replace the variables $\mathbf{X}_{1i}$ and Δ_i by $\mathbf{X}_{1i}^*$ and Δ_i^* respectively in $\widehat{p}(\cdot)$; i.e.,

$$\widehat{p}_m(x_1) = \sum_{i=1}^{m} \frac{\mathrm{K}_h(x_1 - \mathbf{X}_{i1}^*)}{\sum_{j=1}^{m} \mathrm{K}_h(x_1 - \mathbf{X}_{j1}^*)} \mathbb{1}_{\{\Delta_i^*=1\}},$$

we remark that

$$\mathbb{P}_{\mathrm{W}}\left(\frac{1}{m}\sum_{i=1}^{m} \mathrm{K}_h(x_1 - \mathbf{X}_{i1}^*)\mathbb{1}_{\{\Delta_i^*=1\}}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{K}_h(x_1 - \mathbf{X}_{i1})\mathbb{1}_{\{\Delta_i=1\}},$$

which implies $d_{\mathscr{H}}\left(\widehat{p}_m, \widehat{p}\right) = o_{\mathbb{P}_{\mathrm{W}}^*}(1)$ i.p. By the triangular inequality we get

$$d_{\mathscr{H}}\left(\widehat{p}_m, p_0\right) \leq d_{\mathscr{H}}\left(\widehat{p}_m, \widehat{p}\right) + d_{\mathscr{H}}\left(\widehat{p}, p_0\right) = o_{\mathbb{P}_{\mathrm{W}}^*}(1), \text{ i.p.}$$

(AB3) is verified by construction of the estimator $\widehat{\boldsymbol{\theta}}_m^*$. Which implies the consistency of $\widehat{\boldsymbol{\theta}}_m^*$. Next for the rate of convergence we show only conditions **(B2)** and **(B3)**. For **(B2)**, it suffices by **remark 3.3(ii)** to show (3.3.1) and (3.3.2). For that by uses of the relation between covering and bracketing numbers and Corollary 2.7.2 of van der Vaart and Wellner [1996] we get that

$$\log \mathrm{N}\left(\epsilon \left\| \mathrm{M}_{\delta,\delta_1'} \right\|_{\mathbb{L}_2(\mathbb{Q})}, \mathscr{M}_{\delta,\delta_1'}, \mathbb{L}_2(\mathbb{Q})\right) \leq \exp\{\mathfrak{K}\epsilon^{-1}\},$$

for every probability measure $\mathbb{Q}$ on $\mathbb{R}^4$, which implies our relation in (3.3.1), (3.3.2) is verified by the choice $\varphi(\delta) = \sqrt{\delta}$ as consequence we get **(B2)**. For **(B3)**, it follows directly like in section7 of the same reference which described this example and by the choice of the two functions $\boldsymbol{\psi}_1(\cdot)$ and $\boldsymbol{\psi}_2(\cdot)$ given in **Remark 3.3(iii)**, which implies **(B3)**. By their discussion for the rates r_n, v_n and the bandwidth h of the kernel; it follows

$$\boldsymbol{\theta}_n - \boldsymbol{\theta}_0 = \mathrm{O}_{\mathbb{P}^*}\left(n^{-1/3}\right).$$

We verify the assumption **(BB1)** as in the verification of condition **(AB2)** by choosing $\widehat{p}_m(\cdot) = \widehat{p}(\cdot)$ we get $v_m^{-1} = \sqrt{\frac{\log m}{mh}} + h$, where $h = h_m$. Assumption **(BB2)** holds by the same argument given for **(B2)**. For assumption **(BB3)**, we check conditions (b)-(d) of **Remark 3.3(iii)**. We obtain

$$\Gamma\left(\boldsymbol{\theta}_0, p\right) = \mathbb{P}\left[\frac{p_0\left(\mathbf{X}_1\right)}{p\left(\mathbf{X}_1\right)}\left\{1 - 2\mathbb{P}\left(\mathbb{1}_{\{\mathbf{Y}=1\}}|\mathbf{X}_1, \mathbf{X}_2\right)\right\} f_{\mathbf{X}_2|\mathbf{X}_1}\left(f_{\boldsymbol{\theta}_0}\left(\mathbf{X}_1\right)\right) \frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}\left(\mathbf{X}_1\right)\right], \tag{3.4.5}$$

and

$$\begin{aligned}\Lambda\left(\boldsymbol{\theta}_0, p\right) =& \mathbb{P}\left[\frac{p_0\left(\mathbf{X}_1\right)}{p\left(\mathbf{X}_1\right)}\left\{1 - 2\mathbb{P}\left(\mathbb{1}_{\{\mathbf{Y}=1\}}|\mathbf{X}_1, \mathbf{X}_2\right)\right\}\left\{f'_{\mathbf{X}_2|\mathbf{X}_1}\left(f_{\boldsymbol{\theta}_0}\left(\mathbf{X}_1\right)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}\left(\mathbf{X}_1\right)\right)^2\right.\right. \\ &\left.\left. + f_{\mathbf{X}_2|\mathbf{X}_1}\left(f_{\boldsymbol{\theta}_0}\left(\mathbf{X}_1\right)\right)\frac{\partial^2}{\partial\boldsymbol{\theta}^2} f_{\boldsymbol{\theta}_0}\left(\mathbf{X}_1\right)\right\}\right],\end{aligned}$$

provided the derivatives in $\Lambda\left(\boldsymbol{\theta}_0, p\right)$ all exist. By the definition of maximum it follows that $\Gamma\left(\boldsymbol{\theta}_0, p_0\right) = 0$ and $\Lambda\left(\boldsymbol{\theta}_0, p_0\right)$ is negative. Noting that

$$\|\Gamma\left(\boldsymbol{\theta}_0, \widehat{p}_m\right)\| = \mathrm{O}_{\mathbb{P}_{\mathrm{W}}}(r_m^{-1}) \text{ i.p.}$$

if r_m satisfies

$$r_m\left(m^{-1/2} + h_m + \frac{\log m}{m h_m}\right) = \mathrm{O}(1),$$

by noting that the expectation in (3.4.5) is taken with respect to $\mathbf{Z}$ and $\mathbf{W}$ when we are working with $\widehat{p}_m$, since our function are measurable, we obtain such result by applying Fubini's Theorem. This condition on r_m and the other given in **(BB2)** which is satisfied for $r_m = \mathrm{O}(m^{1/3})$ are reconcilable provided

$$m h_m^3 = \mathrm{O}(1) \quad \text{and} \quad \frac{(\log m)^{3/2}}{m h_m^{3/2}} = \mathrm{O}(1).$$

Note that if we assume that $p_0(\cdot)$ is twice continuously differentiable we can weaken the first condition to $m h_m^6 = \mathrm{O}(1)$, as a consequence we get the v_m^{-1} of $\widehat{p}_m$ would be $\mathrm{O}\left(\sqrt{\frac{\log m}{m h_m}} + h_m^2\right)$, which is faster than $r_m^{-1} = m^{-1/3}$ of $\widehat{\boldsymbol{\theta}}_m^*$ provided $m h_m^3 \longrightarrow \infty$. The level of complexity of the latter case is less than the case where p_0 is only once differentiable, And we do not discuss it any further, therefore. We conclude that,

$$\widehat{\boldsymbol{\theta}}_m^* - \boldsymbol{\theta}_0 = \mathrm{O}_{\mathbb{P}_{\mathrm{W}}^*}(m^{-1/3}) \text{ i.p.}$$

Finally, for the weak convergence of $\boldsymbol{\theta}_n$, we note that our assumptions **(C4)** is satisfied for $j_n = \sqrt{n}$ like in **Remark3.5 (iii)** and **(C9)** hold similarly to **(B2)**. By consequence $n^{1/3}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$ converges weakly. Where assumption **(CB1)** follows from part (ii) of Theorem (3.3.2.2) and condition **(BB1)**, by similar proof of condition **(BB2)** we get **(CB2)**. We get from **Remark 3.3 (iii)**, (**vi**) and **Remark 3.5 (viii)** that assumption **(CB3)** holds, provided that

$$|\Lambda(\boldsymbol{\theta}_0, p_0)| < \infty.$$

Clearly we have for some positive constant $c > 0$ that $m^{-1/3} < \mathrm{C}$. For assumption **(CB4)**, we have

$$r_m \mathrm{W}_m(\gamma) = r_m \Gamma(\boldsymbol{\theta}_0, \widehat{p}_m)\gamma = o_{\mathbb{P}_{\mathrm{W}}}(1) \text{ i.p.},$$

provided $m h_m^3 = o(1)$ and $\frac{\log^{3/2} m}{m h_m^{3/2}} = o(m^{-1/2})$, by using what we discuss already for **(BB3)**. Next, by the result given to the process in (3.6.7) i.e., the process $\gamma \mapsto \mathbb{G}_n \frac{r_m^2}{\sqrt{m}} \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0}$ converges weakly to the process $\mathbb{G}(\gamma)$ and condition **(AB1)**, we get

$$\begin{aligned} r_m^2 \widehat{\mathbb{P}}_m^* \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, p_0} &= r_m^2 \left[\left(\widehat{\mathbb{P}}_m^* - \mathbb{P}_n\right) \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, p_0} + \sqrt{\frac{m}{n}} \mathbb{G}_n \frac{\widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, p_0}}{\sqrt{m}} + \mathbb{P} \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, p_0} \right] \\ &= r_m^2 \left(\widehat{\mathbb{P}}_m^* - \mathbb{P}_n\right) \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, p_0} + \frac{1}{2} \Lambda(\boldsymbol{\theta}_0, p_0) \gamma^2 + o_{\mathbb{P}}(1), \end{aligned}$$

with $\Gamma(\boldsymbol{\theta}_0, p_0) = 0$ and

$$\Lambda(\gamma) = \frac{1}{2} \Lambda(\boldsymbol{\theta}_0, p_0) \gamma^2.$$

The process $\gamma \to r_m^2 \left(\widehat{\mathbb{P}}_m^* - \mathbb{P}_n\right) \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, p_0}$ are the same given in Lee [2012] where there is no presence of nuisance parameter. Hence, we can follow the same steps given in Lemma 1 of Lee [2012] and get the convergence of the marginals using Lindeberg's condition and some regularity assumption on $f_{\mathbf{X}_1/\mathbf{X}_2}$ and $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}$. By construction of the estimator $\widehat{\boldsymbol{\theta}}_m^*$, condition **(CB5)** follows. Then we get the asymptotic distribution of $r_m \left(\widehat{\boldsymbol{\theta}}_m^* - \boldsymbol{\theta}_n\right)$ from part (ii) of Theorem 3.3.3.3.

3.5 Numerical results

We provide numerical illustrations regarding the asymptotic distribution of estimators in the classification with missing data, details are provided in Section 3.4.2. The computing program codes were implemented in R. In our simulation, we will show resampling bootstrap samples of size n fails while resampling with size m satisfies the conditions given in previous sections for the consistency of the bootstrap. Let us describe the model, define

$$\mathrm{X}_2 = \max(\min(\mathrm{U} + \epsilon, 1), 0),$$

where $\mathrm{U} \sim \mathscr{U}[0,1]$, $\epsilon \sim \mathscr{U}[-.1,.1]$ and $\mathrm{X}_1 \sim \mathscr{U}[0,1]$, with X_1, ϵ and U are independent. Let

$$\mathrm{Y} = \mathbb{1}\{\mathrm{U} \geq f_\theta(\mathrm{X}_1)\}, \tag{3.5.1}$$

were $f_\theta(x_1) = \theta x_1$, for some θ, we define

$$p(x_1) = \mathbb{P}(\Delta = 1 \mid \mathrm{X}_1 = x_1) = \alpha_0 + (x_1 - 0.5)^2.$$

The data is composed of $\mathrm{Z}_i = (\mathrm{X}_{i1}, \mathrm{X}_{i2}, \mathrm{Y}_i \Delta_i, \Delta_i)$ $i = 1, 2, \ldots, n$ from the described model. For the bandwidth, we use $h_n = \frac{c_h}{\sqrt{n}}$ ($h_m = \frac{c_h}{\sqrt{m}}$), which satisfies the requirements of regularity conditions of the asymptotic theory. In this simulation, we use the quadratic kernel defined by

$$\mathrm{K}(u) = \frac{15}{16}\left(1 - u^2\right)^2 \mathbb{1}\{|u| \leq 1\},$$

which is a density function having support $[-1,1]$. The results given below are based on three different value of n, we took $n = 250$, $n = 1000$ and $n = 2000$ and the true value to be $\theta^0 = 1$, we choose $c_h = 3.5$ and $\alpha_0 = 0.5$, this choice is not restrictive, we can obtain the same desired result with different value of c_h and α_0 for example $c_h = 2$ or 5 and $\alpha_0 = .25$ or .75. The bootstrap procedure is as follows, for each value of m we generate B independent bootstrap samples $\left\{\mathrm{Z}_{ib}^* : i \leq m\right\}$ for $b = 1, \ldots, \mathrm{B}$, using some method of bootstrapping, and for each given value of m, we compute an estimator $\theta_m^{(b)}$ based on the b-th bootstrapped sample. Our main objective is to give a comparison between the distribution of $n^{1/3}\left(\theta_n - \theta^0\right)$ with the m out of n bootstrap distribution of $m^{1/3}\left(\widehat{\theta}_m^* - \theta_n\right)$. To achieve this goal, we have used the Kolmogorov distance between the distributions of $n^{1/3}\left(\theta_n - \theta^0\right)$ and $m^{1/3}\left(\widehat{\theta}_m^* - \theta_n\right)$ by averaging over 1000 and 1500 m out of n bootstrap sample drawn from one chosen arbitrarily random sample. Table 3.1 displays the results for $n = 250$, $n = 1000$ and $n = 2000$ which show that the most

accurate estimates are given for the choices of $m = 50$, $m = 60$ and $m = 110$ respectively. Deviations from these choices in either direction result in deteriorating accuracy. In Figures 3.1-3.3, we give the empirical distribution of the true distribution and the empirical distribution of the bootstrapped one for some values of m given in Table 3.1, which each figure compares the estimated bootstrap empirical distribution with those of $n^{1/3}\left(\theta_n - \theta^0\right)$ for the different values of n. All these figures show that the classical bootstraps (n out n bootstrap) fail while the m out n bootstraps are consistent. Figures 3.4-3.6 show the root mean squared error (RMSE) of the estimator $\widehat{\theta}^*_m$ for several values of m given in Table 3.1, for each value of n. In figures 3.7-3.9 we draw the graph of the criterion function for some given values of m for each value of n. Figure 3.10 presents the concentration of the different estimator values of θ^0 based on 1000 replica random sample with the corresponding graph of the criterion function as a function of θ for a different value of n.

One can see as in any other inferential context, the greater the sample size, the better.

$n = 250$		$n = 1000$		$n = 2000$	
m	KD	m	KD	m	KD
10	0,1733	50	0,0880	50	0,0730
20	0,1267	60	**0,0843**	100	0,0780
30	0,0853	70	0,0870	110	**0,0717**
40	0,0687	80	0,1267	120	0,1150
50	**0,0527**	90	0,1197	130	0,0950
60	0,0793	100	0,1040	140	0,1103
70	0,0780	110	0,1180	150	0,0997
80	0,1213	120	0,1133	160	0,1437
90	0,0953	130	0,1080	170	0,1557
100	0,1183	140	0,1283	180	0,1337
125	0,1383	150	0,1073	190	0,1523
150	0,1453	200	0,1553	200	0,1480
175	0,1773	275	0,2187	300	0,2073
200	0,1757	350	0,2543	400	0,2530
225	0,2057	425	0,2917	500	0,2833
250	0,1993	500	0,2990	750	0,3533
		750	0,3787	1000	0,3953
		1000	0,4187	1250	0,4310
				1500	0,4537
				2000	0,5240

Table 3.1: Kolmogorov Distance (KD) Between Distributions of $n^{1/3}\left(\theta_n - \theta^0\right)$ and $m^{1/3}\left(\widehat{\theta}^*_m - \theta_n\right)$, for $n = 250$, $n = 1000$ and $n = 2000$.

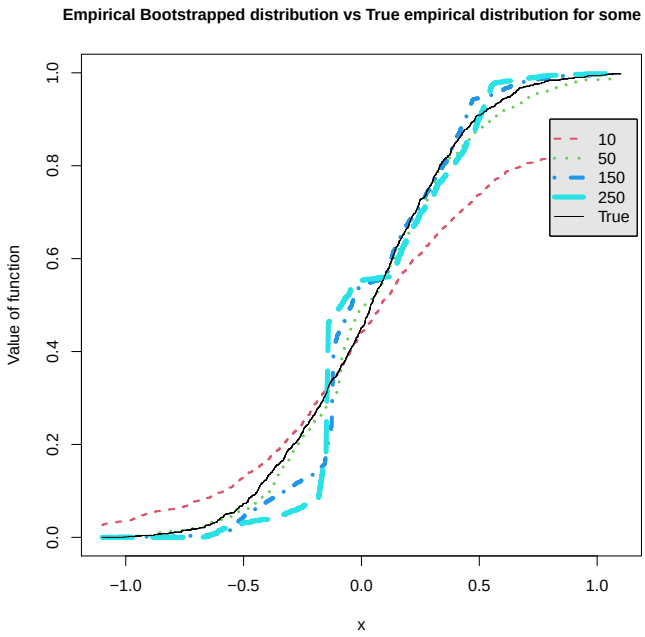


Figure 3.1: Empirical distribution of $n^{1/3}\left(\theta_n - \theta^0\right)$ compared with those of $m^{1/3}\left(\widehat{\theta}^*_m - \theta_n\right)$, $m = 50$, $m = 110$, $m = 200$, $m = 250$ and $n = 250$.

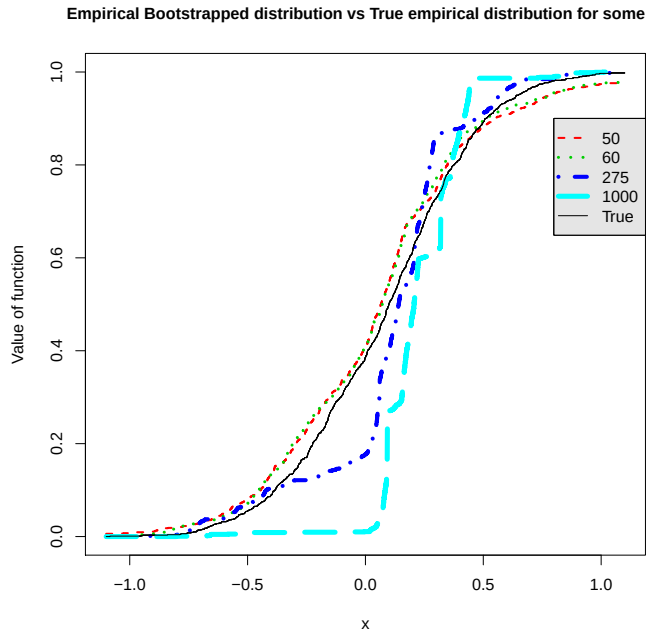


Figure 3.2: Empirical distribution of $n^{1/3}\left(\theta_n - \theta^0\right)$ compared with those of $m^{1/3}\left(\widehat{\theta}^*_m - \theta_n\right)$, $m = 50$, $m = 60$, $m = 275$, $m = 1000$ and $n = 1000$.

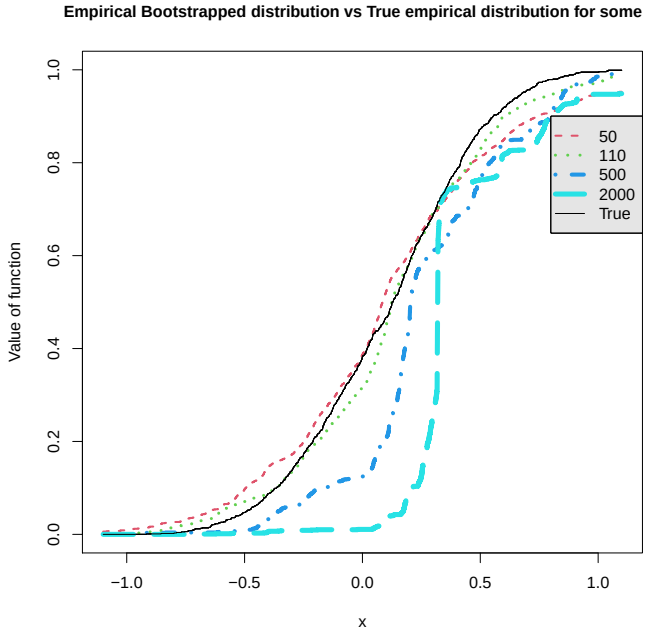


Figure 3.3: Empirical distribution of $n^{1/3}\left(\theta_n - \theta^0\right)$ compared with those of $m^{1/3}\left(\widehat{\theta}^*_m - \theta_n\right)$, $m = 50$, $m = 110$, $m = 500$, $m = 2000$ and $n = 2000$.

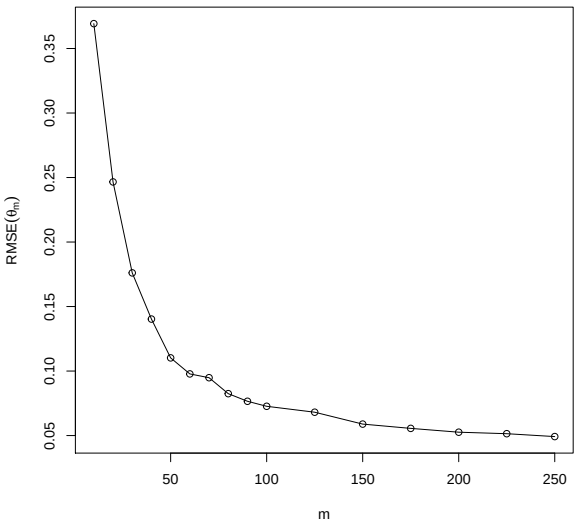


Figure 3.4: The RMSE of $\widehat{\theta}^*_m$ in function of m, for $n = 250$.

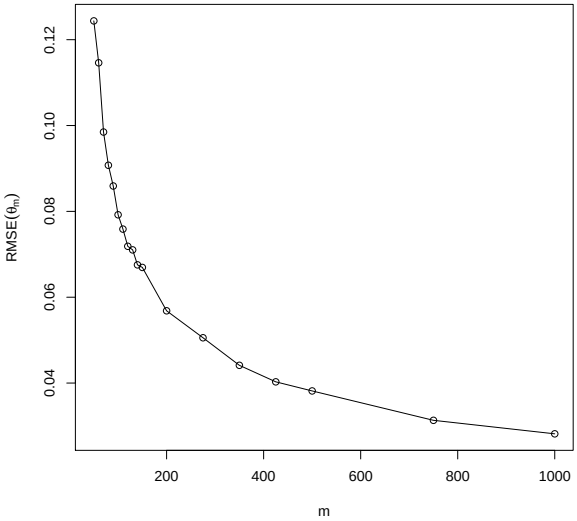


Figure 3.5: The RMSE of $\widehat{\theta}^*_m$ in function of m, for $n = 1000$.

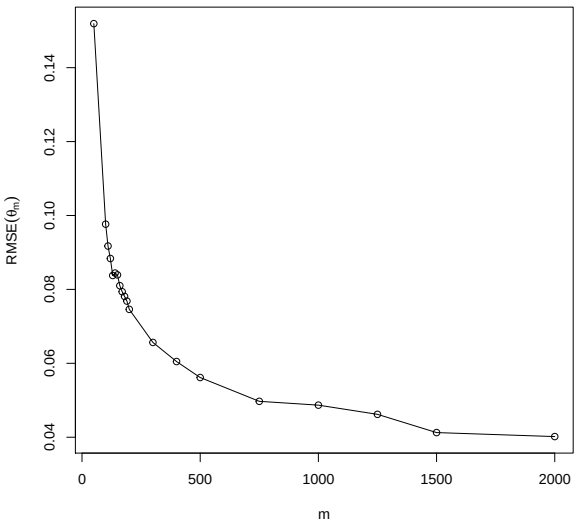


Figure 3.6: The RMSE of $\widehat{\theta}^*_m$ in function of m, for $n = 2000$.

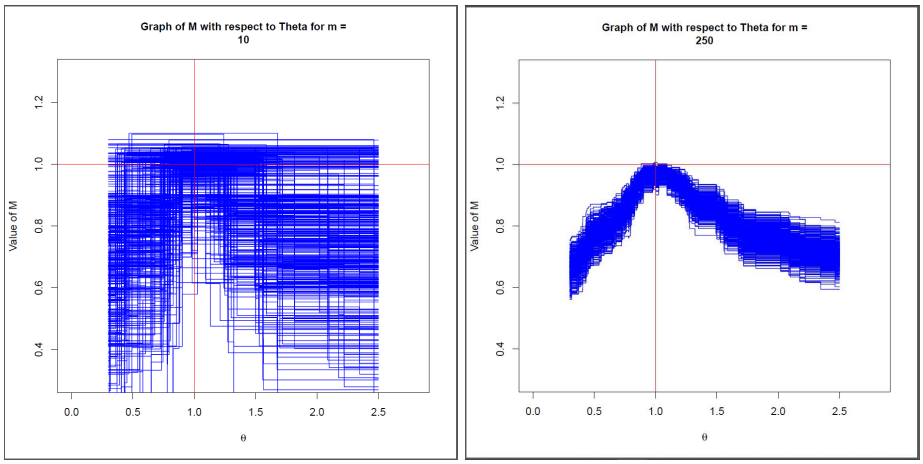


Figure 3.7: Graph of the criterion function based on bootstrapped sample for $n = 250$.

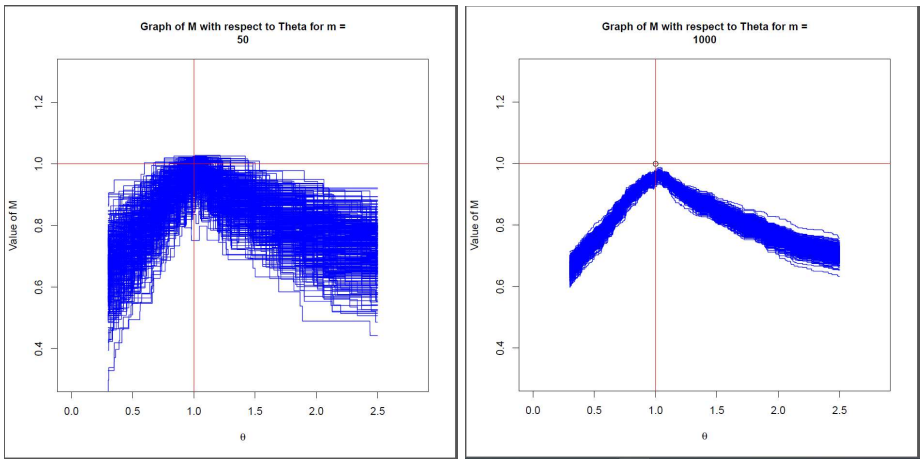


Figure 3.8: Graph of the criterion function based on bootstrapped sample for $n = 1000$.

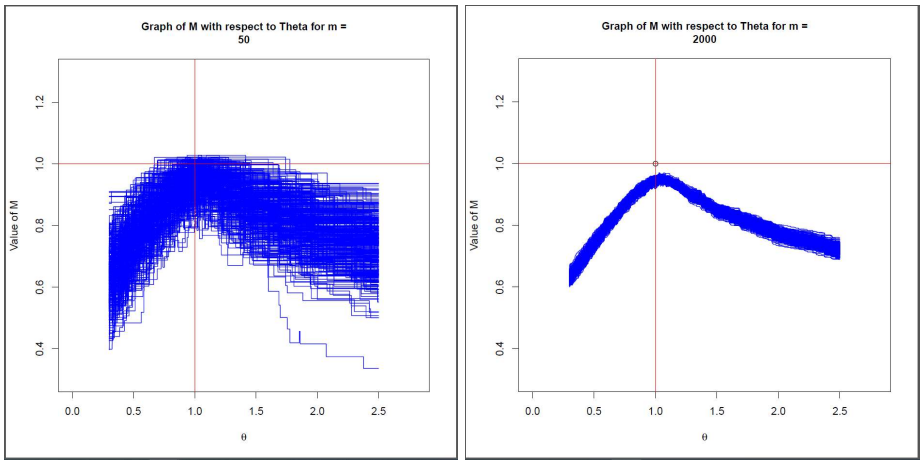


Figure 3.9: Graph of the criterion function based on bootstrapped sample for $n = 2000$.

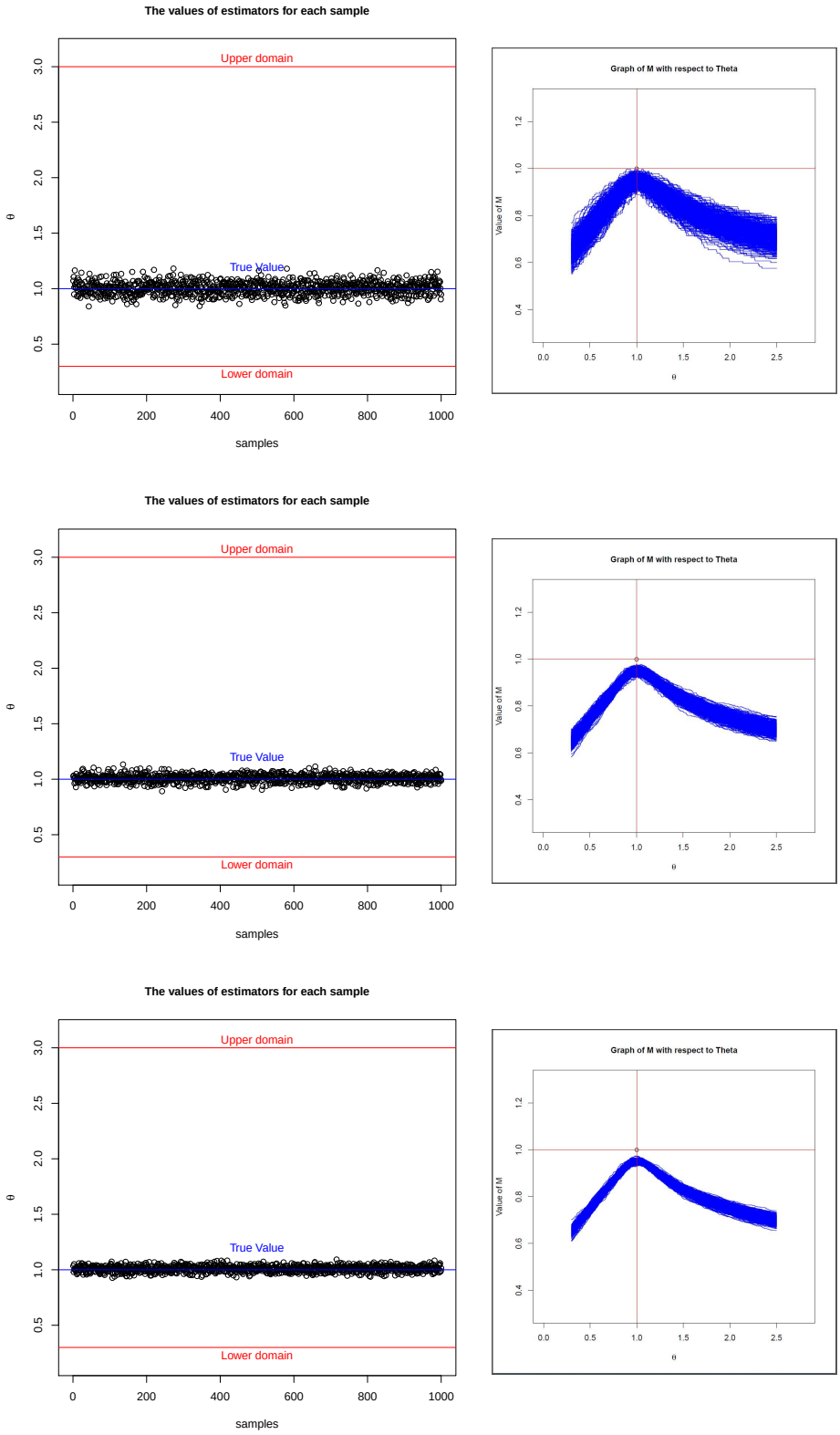


Figure 3.10: Concentration of estimators and the graph of each corresponding criterion function for different values of $n = 250$, $n = 1000$ and $n = 2000$.

3.6 Mathematical developments

In this section, we give the proofs of the asymptotic results of our M-estimator $\boldsymbol{\theta}_n$ and its bootstrap version.

Proof of Theorem 3.3.1.2

Part (i) follows directly from Theorem 1 of Delsol and Van Keilegom [2020]. For (ii), note that **(AB1)** and **(A2)** imply that

$$\|\widehat{\mathbb{P}}_m^* - \mathbb{P}\|_{\mathscr{M}_{\Theta,\mathscr{H}}} = o_{\mathbb{P}_W^*}(1) \ \text{ a.s.} \tag{3.6.1}$$

By using the result in Lemma 3.6.16 of van der Vaart and Wellner [1996]. We have; for every $\eta > 0$ there is $\delta > 0$, such that

$$\begin{aligned}
&\mathbb{P}_W^*\left(\|\widehat{\boldsymbol{\theta}}_m^* - \boldsymbol{\theta}_0\| > \eta\right) \\
&\leq \mathbb{P}_W^*\left(\mathbb{P}\mathbf{m}_{\boldsymbol{\theta}_0,h_0} - \mathbb{P}\mathbf{m}_{\widehat{\boldsymbol{\theta}}_m^*,h_0} > \delta\right) \\
&\leq \mathbb{P}_W^*\left(2\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\mathbb{P}\mathbf{m}_{\boldsymbol{\theta},\widehat{h}_m} - \mathbb{P}\mathbf{m}_{\boldsymbol{\theta},h_0}| + \mathbb{P}\mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}_m} - \mathbb{P}\mathbf{m}_{\widehat{\boldsymbol{\theta}}_m^*,\widehat{h}_m} > \delta\right) \\
&\leq \mathbb{P}_W^*\left(2\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\mathbb{P}\mathbf{m}_{\boldsymbol{\theta},\widehat{h}_m} - \mathbb{P}\mathbf{m}_{\boldsymbol{\theta},h_0}| + 2\|\widehat{\mathbb{P}}_m^* - \mathbb{P}\|_{\mathscr{M}_{\boldsymbol{\Theta},\mathscr{H}}} > \delta - \widehat{\mathrm{R}}_n\right).
\end{aligned}$$

Making use of the assumption **(AB3)**, there is $n_0 \in \mathbb{N}$, such that for every $n \geq n_0$, we obtain the existence of $\delta' > 0$, such that $\delta - \widehat{\mathrm{R}}_n \geq 4\delta'$ i.p., and the last expression is bounded by:

$$\begin{aligned}
&\mathbb{P}_W^*(\|\widehat{\boldsymbol{\theta}}_m^* - \boldsymbol{\theta}_0\| > \eta) \\
&\leq \mathbb{P}_W^*\left(2\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\mathbb{P}\mathbf{m}_{\boldsymbol{\theta},\widehat{h}_m} - \mathbb{P}\mathbf{m}_{\boldsymbol{\theta},h_0}| + 2\|\widehat{\mathbb{P}}_m^* - \mathbb{P}\|_{\mathscr{M}_{\boldsymbol{\Theta},\mathscr{H}}} > 4\delta'\right) \\
&\leq \mathbb{P}_W^*\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\mathbb{P}\mathbf{m}_{\boldsymbol{\theta},\widehat{h}_m} - \mathbb{P}\mathbf{m}_{\boldsymbol{\theta},h_0}| > \delta'\right) + \mathbb{P}_W^*\left(\|\widehat{\mathbb{P}}_m^* - \mathbb{P}\|_{\mathscr{M}_{\boldsymbol{\Theta},\mathscr{H}}} > \delta'\right).
\end{aligned}$$

By using the assumptions **(AB1)**, **(A3)**, **(AB3)** in combination with (3.6.1), we obtain the desired result. □

Proof of Theorem 3.3.2.2

Firstly note that, we will give the proof of this theorem for the particular choice of function

$$\boldsymbol{\psi}_1(x) = \mathrm{I}d(x) = x \text{ and } \boldsymbol{\psi}_2(x) = \frac{x^{\nu}}{\ell(1/x)} \text{ for every } x \neq 0.$$

It worth noticing that this condition is in agreement with those used in Lee [2012] in the parametric setting. Let β_n be the $o_{\mathbb{P}^*}(1)$ in assumption **(B3)** and we define the sets

$$\mathrm{S}_{j,n} = \left\{\boldsymbol{\theta} \in \Theta : 2^{j-1} < r_n\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq 2^j\right\},$$

we observe

$$\Theta \backslash \boldsymbol{\theta}_0 = \bigcup_{j=1}^{\infty} \mathrm{S}_{j,n}.$$

Our objective is to show that; for any $\epsilon > 0$, there exists $\tau_\epsilon > 0$ such that

$$\mathbb{P}^* \left(r_n \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\| > \tau_\epsilon \right) < \epsilon, \tag{3.6.2}$$

for any n sufficiently large. In the sequel, we work with arbitrary fixed $\epsilon > 0$. For any $\delta, \delta_1, \mathrm{M}, \mathrm{K}, \mathrm{K}^{'} > 0$, by using the condition **(A5)**, we readily obtain

$$\begin{aligned}
&\mathbb{P}^* \left(r_n \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\| > 2^{\mathrm{M}} \right) \\
&\leq \sum_{\mathrm{M} \leq j, 2^j \leq \delta r_n} \mathbb{P}^* \left(\sup_{\boldsymbol{\theta} \in \mathrm{S}_{j,n}} [\mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta}, \widehat{h}} - \mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta}_0, \widehat{h}}] \geq -\mathrm{K} r_n^{-\nu} \ell(r_n)^{-1}, \mathrm{A}_n \right) \\
&\quad + \mathbb{P}^* \left(2\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\| \geq \delta \right) + \mathbb{P}^* \left(r_n^{\nu} \ell(r_n) |\mathrm{R}_n| > \mathrm{K} \right) + \mathbb{P}^* \left(r_n^{\nu-1} \ell(r_n) |\mathrm{W}_n| > \mathrm{K}^{'} \right) \\
&\quad + \mathbb{P}^* \left(|\beta_n| > \frac{\mathrm{C}}{2} \right) + \mathbb{P}^* \left(d_{\mathscr{H}} \left(\widehat{h}, h_0 \right) > \frac{\delta_1}{\nu_n} \right),
\end{aligned}$$

where

$$\mathrm{A}_n = \left\{ r_n^{\nu-1} \ell(r_n) |\mathrm{W}_n| \leq \mathrm{K}^{'}, |\beta_n| \leq \frac{\mathrm{C}}{2}, d_{\mathscr{H}} \left(\widehat{h}, h_0 \right) \leq \frac{\delta_1}{\nu_n} \right\}.$$

Indeed, we can write

$$\begin{aligned}
&\mathbb{P}^* \left(r_n \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\| > 2^{\mathrm{M}}, 2\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\| < \delta, r_n^{\nu} \ell(r_n) |\mathrm{R}_n| \leq \mathrm{K}, \mathrm{A}_n \right) \\
&\quad \leq \sum_{j \geq \mathrm{M}, 2^j \leq \delta r_n} \mathbb{P}^* \left(\boldsymbol{\theta}_n \in \mathrm{S}_{j,n}, r_n^{\nu} \ell(r_n) |\mathrm{R}_n| \leq \mathrm{K}, \mathrm{A}_n \right) \\
&\quad \leq \sum_{j \geq \mathrm{M}, 2^j \leq \delta r_n} \mathbb{P}^* \left(\sup_{\theta \in \mathrm{S}_{j,n}} \left[\mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta}, \widehat{h}} - \mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta}_0, \widehat{h}} \right] \geq -\mathrm{R}_n, r_n^{\nu} \ell(r_n) |\mathrm{R}_n| \leq \mathrm{K}, \mathrm{A}_n \right) \\
&\quad \leq \sum_{j \geq \mathrm{M}, 2^j \leq \delta r_n} \mathbb{P}^* \left(\sup_{\theta \in \mathrm{S}_{j,n}} \left[\mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta}, \widehat{h}} - \mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta}_0, \widehat{h}} \right] \geq -\mathrm{K} r_n^{-\nu} \ell(r_n)^{-1}, \mathrm{A}_n \right).
\end{aligned}$$

Condition **(C)** implies, for all $\delta > 0$, that there exists n_ϵ, such that, for $n > n_\epsilon$, we have

$$\mathbb{P}^* \left(2\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\| \geq \delta \right) < \frac{\epsilon}{6}.$$

By the definitions of R_n, W_n and under condition **(B1)**, there exist $\delta_1, \mathrm{K}_\epsilon, \mathrm{K}_\epsilon^{'}$ and $\mathrm{K}_{2,\epsilon}$ such that we have

$$\begin{aligned}
&\mathbb{P}^* \left(r_n^{\nu} \ell(r_n) |\mathrm{R}_n| > \mathrm{K}_\epsilon \right) < \tfrac{\epsilon}{6}, \quad \mathbb{P}^* \left(r_n^{\nu-1} \ell(r_n) |\mathrm{W}_n| > \mathrm{K}_\epsilon^{'} \right), \\
&\mathbb{P}^* \left(|\beta_n| > \tfrac{\mathrm{C}}{2} \right) < \tfrac{\epsilon}{6}, \quad \mathbb{P}^* \left(d_{\mathscr{H}}(\widehat{h}, h_0) > \tfrac{\delta_1}{\nu_n} \right) < \tfrac{\epsilon}{6}.
\end{aligned} \tag{3.6.3}$$

For n large than some n_1. We fix $\delta < \delta_0$ and suppose $n \geq \max(n_0, n_1, n_\epsilon)$, for $2^j \leq \delta r_n$, we have the assumptions **(B2)** and **(B3)** are fulfilled on all $\mathrm{S}_{j,n}$. For each fixed j such that $2^j \leq \delta r_n$,

under assumption **(B3)**, for all $\boldsymbol{\theta} \in \mathrm{S}_{j,n}$, we then have

$$\begin{aligned}
&\mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta},\widehat{h}} - \mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}} \\
&\leq \mathbb{P}\mathbf{m}_{\boldsymbol{\theta},\widehat{h}} - \mathbb{P}\mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}} + \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \leq \frac{2^j}{r_n}} \left| \mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta},\widehat{h}} - \mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}} - \mathbb{P}\mathbf{m}_{\boldsymbol{\theta},\widehat{h}} + \mathbb{P}\mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}} \right| \\
&\leq |\mathrm{W}_n| \frac{2^j}{r_n} - (\mathrm{C} - \beta_n) \frac{2^{\nu(j-1)}}{r_n^{\nu} \ell(2^{-(j-1)} r_n)} + \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \leq \frac{2^j}{r_n}} \left| \mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta},\widehat{h}} - \mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}} - \mathbb{P}\mathbf{m}_{\boldsymbol{\theta},\widehat{h}} + \mathbb{P}\mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}} \right| \\
&\leq |\mathrm{W}_n| \frac{2^j}{r_n} - (\mathrm{C} - \beta_n) \frac{1}{2^{-j\nu} r_n^{\nu} \ell(2^{-j} r_n)} + n^{-1/2} \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \leq \frac{2^j}{r_n}} \left| \mathbb{G}_n \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,\widehat{h}} \right|.
\end{aligned}$$

Consequently, we obtain the following inequalities;

$$\begin{aligned}
&\mathbb{P}^* \left(\sup_{\theta \in \mathrm{S}_{j,n}} \left[\mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta},\widehat{h}} - \mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}} \right] \geq -\mathrm{K}_\epsilon r_n^{-\nu} \ell(r_n)^{-1}, \mathrm{A}_n \right) \\
&\leq \mathbb{P}^* \left(n^{-1/2} \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \leq \frac{2^j}{r_n}, d_{\mathscr{H}}(h,h_0) \leq \frac{\delta_1}{\nu_n}} \left| \mathbb{G}_n \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h} \right| \geq \frac{\mathrm{C}}{2} \frac{2^{j\nu}}{r_n^{\nu} \ell(2^{-j} r_n)} - \mathrm{K}'_\epsilon \frac{2^j}{r_n^{\nu} \ell(r_n)} - \mathrm{K}_\epsilon \frac{1}{r_n^{\nu} \ell(r_n)} \right) \\
&\leq \mathbb{P}^* \left(n^{-1/2} \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \leq \frac{2^j}{r_n}, d_{\mathscr{H}}(h,h_0) \leq \frac{\delta_1}{\nu_n}} \left| \mathbb{G}_n \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h} \right| \right. \\
&\qquad \left. \geq \frac{2^{j\nu}}{r_n^{\nu} \ell(2^{-j} r_n)} \left(\frac{\mathrm{C}}{2} - \mathrm{K}'_\epsilon \frac{\ell(2^{-j} r_n)}{2^{j(\nu-1)} \ell(r_n)} - \mathrm{K}_\epsilon \frac{2^{-j\nu} \ell(2^{-j} r_n)}{\ell(r_n)} \right) \right).
\end{aligned}$$

For any $\lambda > 0$, we can find a non-decreasing function ξ such that

$$x^{\lambda} \ell(x) \sim \xi(x) \text{ as } x \to \infty.$$

It follows that $\frac{2^{-j\lambda} \ell(2^{-j} r_n)}{\ell(r_n)}$ is uniformly bounded for $\mathrm{M} \leq j \leq \log_2 \delta r_n$ and for all n. Making use of the condition **(B2)** in combination with the Chebyshev's inequality and the fact that $\varphi(c\delta) \leq c^{\alpha} \delta$ for all $c \geq 1$, there exists a positive constant C' and for any $\lambda > 0$, we have

$$\begin{aligned}
&\mathbb{P}^* \left(\sup_{\theta \in \mathrm{S}_{j,n}} \left[\mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta},\widehat{h}} - \mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}} \right] \geq -\mathrm{K}_\epsilon r_n^{-\nu} \ell(r_n)^{-1}, \mathrm{A}_n \right) \\
&\leq \mathrm{C}' 2^{-j\nu} r_n^{\nu} \ell\left(2^{-j} r_n\right) n^{-1/2} \varphi\left(\frac{2^j}{r_n}\right) \\
&\leq \mathrm{C}' 2^{-j(\nu-\lambda)} r_n^{\nu} \ell\left(r_n\right) n^{-1/2} \varphi\left(\frac{2^j}{r_n}\right) \\
&\leq \mathrm{C}' 2^{-j(\nu'-\alpha)} r_n^{\nu} \ell(r_n) n^{-1/2} \varphi\left(\frac{1}{r_n}\right),
\end{aligned}$$

where $\nu' = \nu - \lambda > \alpha$. By choosing small value of λ and by using the proprieties of the function $\varphi(\cdot)$, we infer that

$$\sum_{M\leq j, 2^j\leq \delta r_n} \mathbb{P}^*\left(\sup_{\boldsymbol{\theta}\in S_{j,n}} [\mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta},\widehat{h}} - \mathbb{P}_n \mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}}] \geq -K r_n^{-\nu}\ell(r_n)^{-1}, A_n\right) \leq \sum_{M\leq j} 2^{-j(\nu'-\alpha)},$$

the last expression tends to 0 as $M\to\infty$, so we obtain the result (i) of our theorem for sufficiently large value of M and n.
For (ii) we have :

$$\begin{aligned}
&\mathbb{P}^*_W\left(r_m\|\widehat{\boldsymbol{\theta}}^*_m - \boldsymbol{\theta}_0\| > 2^M\right)\\
&\leq \sum_{M\leq j, 2^j\leq \delta r_m} \mathbb{P}^*_W\left(\sup_{\boldsymbol{\theta}\in S_{j,n}} [\widehat{\mathbb{P}}^*_m \mathbf{m}_{\boldsymbol{\theta},\widehat{h}_m} - \widehat{\mathbb{P}}^*_m \mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}_m}] \geq -K r_m^{-\nu}\ell(r_m)^{-1}, A_m\right)\\
&\quad + \mathbb{P}^*_W\left(2\|\widehat{\boldsymbol{\theta}}^*_m - \boldsymbol{\theta}_0\| \geq \delta\right) + \mathbb{P}^*_W\left(r_m^{\nu}\ell(r_m)|\widehat{R}_n| > K\right) + \mathbb{P}^*_W\left(r_m^{\nu-1}\ell(r_m)|W_m| > K'\right)\\
&\quad + \mathbb{P}^*_W\left(|\beta_n| > \frac{C}{2}\right) + \mathbb{P}^*_W\left(d_{\mathscr{H}}\left(\widehat{h}_m, h_0\right) > \frac{\delta_1}{v_m}\right). \qquad (3.6.4)
\end{aligned}$$

We obtain from assumption **(BB3)**, for each fixed j such that $2^j < \delta r_m$ and for all $\boldsymbol{\theta}\in S_{m,j}$

$$\begin{aligned}
&\widehat{\mathbb{P}}^*_m \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,\widehat{h}_m}\\
&\leq \mathbb{P}\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h} + \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq \frac{2^j}{r_m}} \left|\widehat{\mathbb{P}}^*_m \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,\widehat{h}_m} - \mathbb{P}_n \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h} + \mathbb{P}_n \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h} - \mathbb{P}\widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h}\right|\\
&\leq |W_m|\frac{2^j}{r_m} - (C-\beta_n)\frac{1}{2^{-j\nu} r_m^{\nu}\ell(2^{-j}r_m)} + m^{-1/2}\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq \frac{2^j}{r_m}} \left|\widehat{\mathbb{G}}^*_m \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,\widehat{h}_m}\right|\\
&\quad + n^{-1/2}\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq \frac{2^j}{r_m}} \left|\mathbb{G}_n \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,\widehat{h}_m}\right|.
\end{aligned}$$

This gives us, by using Chebyshev's inequality, for some $C' > 0$

$$\begin{aligned}
&\mathbb{P}^*_W\left(\sup_{\boldsymbol{\theta}\in S_{j,n}} [\widehat{\mathbb{P}}^*_m \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,\widehat{h}_m}] \geq -K r_m^{-\nu}\ell(r_m)^{-1}\right)\\
&\leq \mathbb{P}^*_W\left(n^{-1/2}\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq \frac{2^j}{r_m}, d_{\mathscr{H}}(h,h_0)\leq \frac{\delta_1}{v_m}} \left|\mathbb{G}_n \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h}\right| + m^{-1/2}\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq \frac{2^j}{r_m}, d_{\mathscr{H}}(h,h_0)\leq \frac{\delta_1}{v_m}} \left|\widehat{\mathbb{G}}^*_m \widetilde{\mathbf{m}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_0,h}\right|\right.\\
&\qquad \left. \geq \frac{2^{j\nu}}{r_m^{\nu}\ell(2^{-j}r_m)}\left(\frac{C}{2} - K'_\epsilon \frac{\ell(2^{-j}r_m)}{2^{j(\nu-1)}\ell(r_m)} - K_\epsilon \frac{2^{-j\nu}\ell(2^{-j}r_m)}{\ell(r_m)}\right)\right)\\
&\leq C' 2^{-j\nu'} r_m^{\nu}\ell(r_m) m^{-1/2}\left\{\mathbb{P}^*_W\left\|\widehat{\mathbb{G}}^*_m\right\|_{\mathscr{M}_{2^j/r_m,\delta_1/v_m}} + m^{1/2}n^{-1/2}\|\mathbb{G}_n\|_{\mathscr{M}_{2^j/r_m,\delta_1/v_m}}\right\}.
\end{aligned}$$

From assumptions **(B2)** and **(BB2)** the outer expectation of the first term in right of (3.6.4) is

bounded by

$$\begin{aligned}
&\sum_{M\le j,2^j\le\delta r_m} \mathbb{P}\mathbb{P}^*_W\left(\sup_{\boldsymbol{\theta}\in S_{j,n}}[\widehat{\mathbb{P}}^*_m\mathbf{m}_{\boldsymbol{\theta},\widehat{h}_m}-\widehat{\mathbb{P}}^*_m\mathbf{m}_{\boldsymbol{\theta}_0,\widehat{h}_m}]\ge -Kr_m^{-\nu}\ell(r_m)^{-1},A_m\right)\\
&\le \sum_{M\le j,2^j\le\delta r_m} C'2^{-j\nu'}r_m^{\nu}\ell(r_m)m^{-1/2}\left\{\mathbb{P}\mathbb{P}^*_W\left\|\widehat{\mathbb{G}}^*_m\right\|_{\mathscr{M}_{2^j/r_m,\delta_1/v_m}}+m^{1/2}n^{-1/2}\mathbb{P}\left\|\mathbb{G}_n\right\|_{\mathscr{M}_{2^j/r_m,\delta_1/v_m}}\right\}\\
&\le \sum_{M\le j,2^j\le\delta r_m} C'2^{-j\nu'}r_m^{\nu}\ell(r_m)m^{-1/2}\varphi\left(\frac{2^j}{r_m}\right)\\
&\quad+\sum_{M\le j,2^j\le\delta r_m} C'm^{1/2}n^{-1/2}2^{-j\nu'}r_m^{\nu}\ell(r_m)m^{-1/2}\varphi\left(\frac{2^j}{r_m}\right)\\
&\le C'\sum_{M\le j}2^{-j(\nu'-\alpha)}+C'm^{1/2}n^{-1/2}\sum_{M\le j}2^{-j(\nu'-\alpha)},
\end{aligned}$$

with assumption **(AB1)** in mind the last two terms converge to 0 as $M, n\to\infty$, the outer expectation of the others terms in (3.6.4) are $o_{\mathbb{P}^*_W}(1)$ i.p., by Lemma 3 of Cheng and Huang [2010], which completes the proof of Theorem 3.3.2.2. □

Proof of Theorem 3.3.3.2

The proof of the first part **(i)** of Theorem 3.3.3.2 is given in Lemmas 1, 2 and 3 of Delsol and Van Keilegom [2020], where in our setting we use bounded closed subsets in the place of compact subsets. We note by their Lemma 2, we obtain the existence of $\xi_{1,n},\xi_{2,n},\xi_{3,n}$ such that

$$\sup_{\gamma\in\mathscr{K}}|\xi_{l,n}|=o_{\mathbb{P}}(1),\ \text{for } l=1,2,3,$$

and the following decomposition

$$r_n^{\nu}\ell(r_n)\mathbb{P}_n\widetilde{\mathbf{m}}_{\frac{\gamma}{r_n},\widehat{h}}(1+\xi_{1,n})=\left[r_n^{\nu-1}\ell(r_n)W_n+r_n^{\nu}\ell(r_n)\mathbb{P}_n\widetilde{\mathbf{m}}_{\frac{\gamma}{r_n},h_0}\right](1+\xi_{2,n})+\xi_{3,n}.$$

By their Lemma 3, the properties of the function $\gamma\mapsto W_n(\gamma)$ and the assumptions of Theorem 3.3.3.2; we obtain the weak convergence of the process

$$\gamma\mapsto r_n^{\nu-1}\ell(r_n)W_n(\gamma)+r_n^{\nu}\ell(r_n)\mathbb{P}_n\widetilde{\mathbf{m}}_{\frac{\gamma}{r_n},h_0}.$$

Briefly, we have the following decomposition;

$$T_n(\gamma)=r_n^{\nu-1}\ell(r_n)W_n(\gamma)+r_n^{\nu}\ell(r_n)\mathbb{P}_n\widetilde{\mathbf{m}}_{\frac{\gamma}{r_n},h_0}=T_{1,n}(\gamma)+T_{2,n}(\gamma),$$

where

$$T_{1,n}(\gamma)=\mathbb{M}_n(\gamma,h_0)$$

and

$$T_{2,n}=r_n^{\nu}\ell(r_n)\mathbb{P}\widetilde{\mathbf{m}}_{\frac{\gamma}{r_n},h_0}+r_n^{\nu-1}\ell(r_n)W_n(\gamma).$$

The process $\gamma\mapsto T_{1,n}(\gamma)$ does not depend on the estimation of nuisance parameter, so it can be studied in a similar way as in the parametric model, by Theorem 2.11.1 of van der Vaart and

Wellner [1996] and the use of assumptions **(C4)**, **(C5)**, **(C9)** and **(C10)**, we obtain its uniformly asymptotic equicontinuity. For the process $\gamma \mapsto T_{2,n}(\gamma)$, we can show that it is asymptotically uniformly equicontinuous by the same method given in the proof of their Lemma 3. By Theorem 1.5.7 and 1.5.4 of van der Vaart and Wellner [1996], we obtain the asymptotic tightness and the weak convergence of T_n to $\Lambda + \mathbb{G}$ in $\ell(\mathcal{K})$ and using Addendum 1.5.8 in the same reference; the almost all paths of the limiting process on $\mathcal{K}$ are uniformly continuous with respect to $\|\cdot\|$. Finally by Slutsky's theorem we obtain the desired result. □

For part (ii) we are in the situation to show the weak convergence of the bootstrapped process, which follows directly from Slutsky's theorem and Lemmas 3.6.0.1 and 3.6.0.2 given bellow.

Lemma 3.6.0.1 *Let $\mathcal{K} = \{\boldsymbol{\gamma} \in \mathbf{E} : \|\boldsymbol{\gamma}\| \leq K\}$. Then under assumptions of part **(ii)** of Theorem 3.3.3.2, for all $\gamma \in \mathcal{K}$, there exist $z_{0,m}$, $z_{1,m}$, $z_{2,m}$, such that*

$$\sup_{\gamma \in \mathcal{K}} |z_{j,m}| = o_{\mathbb{P}^*_W}(1), \ i.p., \ j = 0,1,2,$$

and

$$\begin{aligned} & r_m^{\nu} \ell(r_m) \widehat{\mathbb{P}}^*_m \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}_m} \left(1 + z_{0,m}\right) \\ = & \left[r_m^{\nu} \ell(r_m) \widehat{\mathbb{P}}^*_m \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0} + r_m^{\nu-1} \ell(r_m) W_m(\gamma) \right] \left(1 + z_{1,m}\right) + z_{2,m}. \end{aligned}$$

Proof of Lemma 3.6.0.1

We need to introduce the following notation

$$\begin{aligned} \alpha_{0,n}(\gamma) &= \frac{\mathbb{P}_n \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}} - \mathbb{P} \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}} - \mathbb{P}_n \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0} + \mathbb{P} \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0}}{r_n^{-\nu} \ell^{-1}(r_n) + \left| \mathbb{P}_n \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}} \right| + \left| \mathbb{P}_n \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0} \right| + |\mathbb{P} \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}}| + \left| \mathbb{P} \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0} \right|}, \\ \alpha_{0,m}(\gamma) &= \frac{\widehat{\mathbb{P}}^*_m \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}} - \mathbb{P}_n \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}} - \widehat{\mathbb{P}}^*_m \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0} + \mathbb{P}_n \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0}}{r_m^{-\nu} \ell^{-1}(r_m) + \left| \widehat{\mathbb{P}}^*_m \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}} \right| + \left| \widehat{\mathbb{P}}^*_m \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0} \right| + |\mathbb{P}_n \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}}| + \left| \mathbb{P}_n \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0} \right|}, \\ s_{n,h}(\gamma) &= \operatorname{sign}\left[\mathbb{P}_n \widetilde{\mathbf{m}}_{\frac{\gamma}{r_n}, h} \right], \\ s_h(\gamma) &= \operatorname{sign}\left[\mathbb{P} \widetilde{\mathbf{m}}_{\frac{\gamma}{r_n}, h} \right], \\ s_{m,h}(\gamma) &= \operatorname{sign}\left[\widehat{\mathbb{P}}^*_m \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h} \right]. \end{aligned}$$

The set $\mathcal{K}$ is bounded and $\boldsymbol{\theta}_0$ belongs to the interior of Θ, there exist $m_{\mathcal{K}}$ such that for all $m \geq m_{\mathcal{K}}$ and for all $\gamma \in \mathcal{K}$, the quantity $\boldsymbol{\theta}_0 + \frac{\gamma}{r_m}$ is in Θ. Then for all $\gamma \in \mathcal{K}$ we have;

$$\begin{aligned} \widehat{\mathbb{P}}^*_m \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}} &= \widehat{\mathbb{P}}^*_m \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0} + \mathbb{P} \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}} - \mathbb{P} \widetilde{\mathbf{m}}_{\frac{\gamma}{r_n}, h_0} \\ &\quad + \alpha_{0,m}(\gamma) \left(r_m^{-\nu} \ell^{-1}(r_m) + \left| \widehat{\mathbb{P}}^*_m \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}} \right| \right. \\ &\quad \left. + \left| \widehat{\mathbb{P}}^*_m \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0} \right| + |\mathbb{P}_n \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}}| + \left| \mathbb{P}_n \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0} \right| \right) \\ &\quad + \alpha_{0,n}(\gamma) \left(r_n^{-\nu} \ell^{-1}(r_n) + \left| \mathbb{P}_n \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}} \right| + \left| \mathbb{P}_n \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0} \right| + |\mathbb{P} \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, \widehat{h}}| + \left| \mathbb{P} \widetilde{\mathbf{m}}_{\frac{\gamma}{r_m}, h_0} \right| \right). \end{aligned}$$

This can be rewritten as follows

$$\begin{aligned} & r_m^{\nu}\ell(r_m)\widehat{\mathbb{P}}^*_m\widetilde{\mathbf{m}}_{\frac{\gamma}{r_m},\widehat{h}_m}\left(1-\alpha_{0,m}(\gamma)s_{m,\widehat{h}}(\gamma)\right) \\ = \; & r_m^{\nu}\ell(r_m)\widehat{\mathbb{P}}^*_m\widetilde{\mathbf{m}}_{\frac{\gamma}{r_m},h_0}\left(1+\alpha_{0,m}(\gamma)s_{m,h_0}(\gamma)\right) \\ & +r_m^{\nu}\ell(r_m)\mathbb{P}\widetilde{\mathbf{m}}_{\frac{\gamma}{r_m},\widehat{h}_m}\left(1+\alpha_{0,n}(\gamma)s_{\widehat{h}}(\gamma)\right) \\ & -r_m^{\nu}\ell(r_m)\mathbb{P}\widetilde{\mathbf{m}}_{\frac{\gamma}{r_m},h_0}\left(1-\alpha_{0,n}(\gamma)s_{h_0}(\gamma)\right)+z'_{2,m}, \end{aligned} \tag{3.6.5}$$

where;

$$\begin{aligned} z'_{2,m} \; = \; & \alpha_{0,m}(\gamma)+r_m^{\nu}\ell(r_m)\left(\alpha_{0,m}(\gamma)+\alpha_{0,n}(\gamma)\right)\left|\mathbb{P}_n\widetilde{\mathbf{m}}_{\frac{\gamma}{r_m},\widehat{h}}\right| \\ & +r_m^{\nu}\ell(r_m)\left(\alpha_{0,m}(\gamma)+\alpha_{0,n}(\gamma)\right)\left|\mathbb{P}_n\widetilde{\mathbf{m}}_{\frac{\gamma}{r_m},h_0}\right|+\frac{r_m^{\nu}\ell(r_m)}{r_n^{\nu}\ell(r_n)}\alpha_{0,n}(\gamma). \end{aligned}$$

We get from the assumption **(CB1)** and **(CB3)** that

$$\begin{aligned} r_m^{\nu}\ell(r_m)\left[\mathbb{P}\widetilde{\mathbf{m}}_{\frac{\gamma}{r_m},\widehat{h}}-\mathbb{P}\widetilde{\mathbf{m}}_{\frac{\gamma}{r_m},h_0}\right] \; = \; & r_m^{\nu-1}\ell(r_m)\mathrm{W}_m(\gamma)+r_m^{\nu-2}\ell(r_m)\beta_n\|\gamma\|^2 \\ & +r_m^{\nu-2}\ell(r_m)o(\|\gamma\|^2) \\ := \; & r_m^{\nu-1}\ell(r_m)\mathrm{W}_m(\gamma)+\alpha_{1,n}(\gamma). \end{aligned} \tag{3.6.6}$$

By combining 3.6.5 and 3.6.6, we infer that

$$\begin{aligned} & r_m^{\nu}\ell(r_m)\widehat{\mathbb{P}}^*_m\widetilde{\mathbf{m}}_{\frac{\gamma}{r_m},\widehat{h}_m}\left(1+z_{0,m}\right) \\ = \; & \left[r_m^{\nu}\ell(r_m)\widehat{\mathbb{P}}^*_m\widetilde{\mathbf{m}}_{\frac{\gamma}{r_m},h_0}+r_m^{\nu-1}\ell(r_m)\mathrm{W}_m(\gamma)\right]\left(1+z_{1,m}\right)+z_{2,m}, \end{aligned}$$

where

$$\begin{aligned} z_{0,m}(\gamma) &= -\alpha_{0,m}(\gamma)s_{m,\widehat{h}}(\gamma) \\ z_{1,m}(\gamma) &= \alpha_{0,m}(\gamma)s_{m,h_0}(\gamma) \\ z_{2,m}(\gamma) &= z'_{2,m}(\gamma)+z''_{2,m}(\gamma), \end{aligned}$$

and

$$\begin{aligned} z''_{2,m}(\gamma) \; = \; & \alpha_{0,n}(\gamma)\left[1+\left(\mathrm{V}(\gamma,\gamma)+r_m^{\nu-2}\ell(r_m)o(\|\gamma\|^2)\right)\left(s_{\widehat{h}}+s_{h_0}\right)(\gamma)\right. \\ & \left.+\left(r_m^{\nu-1}\ell(r_m)\mathrm{W}_m(\gamma)+\alpha_{1,n}(\gamma)\right)\left(s_{\widehat{h}}-s_{n,h_0}\right)(\gamma)\right] \\ & +\alpha_{1,n}(\gamma)\left(1+z_{1,m}(\gamma)\right). \end{aligned}$$

It is easily to show that

$$\sup_{\gamma\in\mathscr{K}}\left|z_{j,m}\right|=o_{\mathbb{P}^*_{\mathrm{W}}}(1) \;\text{ i.p., for }\; j=0,1,2,$$

by using assumptions **(A2)**, **(AB1)** **(C3)**, **(CB2)**, **(CB3)** and Lemma 3 of Cheng and Huang [2010]. □

Lemma 3.6.0.2 *Under the assumptions of Lemma 3.6.0.1, the process*

$$\gamma\mapsto r_m^{\nu}\ell(r_m)\widehat{\mathbb{P}}^*_m\widetilde{\mathbf{m}}_{\frac{\gamma}{r_m},h_0}+r_m^{\nu-1}\ell(r_m)\mathrm{W}_m(\gamma)$$

converges weakly conditionally in probability to the process

$$\gamma\mapsto\Lambda(\boldsymbol{\gamma})+\mathbb{G}(\boldsymbol{\gamma}) \text{ in } \ell^{\infty}(\mathscr{K}).$$

Proof of Lemma 3.6.0.2

Making use of the assumption **(CB4)**, we need only to show the equicontinuity of the process

$$\mathrm{T}_m : \gamma \mapsto r_m^{\nu}\ell(r_m)\widehat{\mathbb{P}}_m^*\tilde{\mathbf{m}}_{\frac{\gamma}{r_m},h_0} + r_m^{\nu-1}\ell(r_m)\mathrm{W}_m(\gamma).$$

One can see that the process T_m can be decomposed into the sum three processes in the following way

$$\mathrm{T}_m = \sum_{i=1}^{3}\mathrm{T}_{i,m},$$

where

$$\begin{aligned}
\mathrm{T}_{1,m} &: \gamma \mapsto r_m^{\nu}\ell(r_m)\left((\widehat{\mathbb{P}}_m^* - \mathbb{P}_n)\tilde{\mathbf{m}}_{\frac{\gamma}{r_m},h_0}\right),\\
\mathrm{T}_{2,m} &: \gamma \mapsto r_m^{\nu}\ell(r_m)\left((\mathbb{P}_n - \mathbb{P})\tilde{\mathbf{m}}_{\frac{\gamma}{r_m},h_0}\right),\\
\mathrm{T}_{3,m} &: \gamma \mapsto r_m^{\nu}\ell(r_m)\mathbb{P}\tilde{\mathbf{m}}_{\frac{\gamma}{r_m},h_0} + r_m^{\nu-1}\ell(r_m)\mathrm{W}_m(\gamma).
\end{aligned}$$

We shall study separately the properties of each process. Firstly, we note that by assumption **(C6)**, **(CB3)** and **(AB1)**, for sufficiently large m, we have $\boldsymbol{\theta}_0 + \frac{\mathscr{K}}{r_m} \subset \Theta$ and then the processes $\mathrm{T}_{1,m}$, $\mathrm{T}_{2,m}$ and $\mathrm{T}_{3,m}$ take values in $\ell^\infty(\mathscr{K})$. The process $\mathrm{T}_{2,m}$ can be treated as in the proof of part **(i)** by reformulating it to this form

$$\begin{aligned}
\mathrm{T}_{2,m}(\gamma) &= r_m^{\nu}\ell(r_m)\left((\mathbb{P}_n - \mathbb{P})\tilde{\mathbf{m}}_{\frac{\gamma}{r_m},h_0}\right)\\
&= \sqrt{\frac{m}{n}}\mathbb{G}_n\frac{r_m^{\nu}\ell(r_m)}{\sqrt{m}}\tilde{\mathbf{m}}_{\frac{\gamma}{r_m},h_0},
\end{aligned} \tag{3.6.7}$$

as in the proof of **(i)** apply Theorem 2.11.22 of van der Vaart and Wellner [1996] to the process

$$\gamma \mapsto \mathbb{G}_n\frac{r_m^{\nu}\ell(r_m)}{\sqrt{m}}\tilde{\mathbf{m}}_{\frac{\gamma}{r_m},h_0},$$

by assumptions **(C4)**, **(C5)**, **(C9)** and **(C10)** we get its uniform equicontinuity with respect to $\|\cdot\|$ on $\mathscr{K}$ and by the use of assumption **(AB1)**, we obtain our main result for the process $\mathrm{T}_{2,m}$. Then the process $\mathrm{T}_{1,m}$ also does not depend to the estimation of the nuisance parameter, it can be treated in the same way as in part **(ii)** of Lemma 2 in Lee [2012]. Briefly, we want to show that

$$\Delta_n \overset{\text{def}}{=} \mathrm{P}_{\mathrm{W}}^*\sup\left\{\left|\mathrm{T}_{1,m}(\gamma) - \mathrm{T}_{1,m}(\gamma')\right| : \|\gamma - \gamma'\| \le \delta_n, \gamma, \gamma' \in \mathscr{K}\right\} \to 0 \ \text{ i.p.}$$

Define the class

$$\mathscr{H}_n = r_m^{\nu}\ell\left(r_m\right)m^{-1/2}\mathscr{M}_{d/r_m}\left(\delta_n/r_m\right),$$

and let

$$\tilde{\mathrm{M}}_n = r_m^{\nu}\ell\left(r_m\right)m^{-1/2}\mathrm{M}_{d/r_m}$$

its envelope function. Making use of the condition **(B2)**, we readily infer that

$$\mathbb{P}^*\sup\left\{|f| : f \in \mathscr{H}_n\right\} \le \mathrm{C}\mathbb{P}^*\tilde{\mathrm{M}}_n \le \mathrm{C} < \infty.$$

It follows by Lemmas 2.9.1 and 3.6.6 of van der Vaart and Wellner [1996] that, for $1 \le n_0 \le n$;

$$\begin{aligned}\mathbb{P}^*\Delta_n &\le C n_0 m^{-1/2}\left(\mathbb{E}\max_{1\le i\le n}\tilde{\mathrm{N}}_i\right)\mathbb{P}^*\tilde{\mathrm{M}}_n \\ &\quad + C n^{1/2} m^{-1/2}\int_0^\infty \left(\mathbb{P}\left\{\left|\tilde{\mathrm{N}}_1\right| > x\right\}\right)^{1/2} dx \max_{n_0\le j\le n}\mathbb{P}^*\left\| j^{-1/2}\sum_{i=n_0}^{j}\epsilon_i\delta_{\mathbf{X}_i}\right\|_{\mathscr{H}_n},\end{aligned}$$

where $\tilde{\mathrm{N}} = (\tilde{\mathrm{N}}_1, \tilde{\mathrm{N}}_2, \ldots)$ and $\epsilon = (\epsilon_1, \epsilon_2, \ldots)$ are independent sequences of i.i.d. symmetrized Poisson variables with parameter $m/(2n)$ and i.i.d. Rademacher variables, respectively and both being independent of $\mathbf{X}_1, \ldots, \mathbf{X}_n$. By Jensen's inequality, problem 3.6.3 of van der Vaart and Wellner [1996] and **(B2)**, we readily get

$$\mathbb{P}^*\Delta_n \le C n_0 m^{-1/2}\log n/\log(n/m+1) + C\max_{n_0\le j\le n}\mathbb{P}^*\left\| j^{-1/2}\sum_{i=n_0}^{j}\epsilon_i\delta_{\mathbf{X}_i}\right\|_{\mathscr{H}_n}. \tag{3.6.8}$$

By taking

$$n_0 = n_{0,n} = a_m m^{1/4}\{\log(n/m+1)/\log n\}^{1/2} \in [1, n],$$

it follows, by condition **(C4)**, that

$$n_0 m^{-1/2}\log n/\log(n/m+1) \to 0 \quad \text{as } n\to\infty. \tag{3.6.9}$$

We refer to the integrand in **(C9)** by $\mathscr{N}(\epsilon)$ for $\epsilon > 0$. By using the triangular inequality, the properties of sub-Gaussianity of Rademacher processes, under **(C10)** and using the Cauchy-Schwarz inequality with **(B2)**, we obtain

$$\begin{aligned}\max_{n_0\le j\le n}\mathbb{P}^*\left\| j^{-1/2}\sum_{i=n_0}^{j}\epsilon_i\delta_{\mathbf{X}_i}\right\|_{\mathscr{H}_n} &= \max_{n_0\le j\le n}\mathbb{P}_{\mathrm{X}}\mathbb{P}_\epsilon\left\| j^{-1/2}\sum_{i=n_0}^{j}\epsilon_i\delta_{\mathbf{X}_i}\right\|_{\mathscr{H}_n} \\ &\le 2\max_{n_0\le j\le n}\mathbb{P}_{\mathrm{X}}\mathbb{P}_\epsilon\left\| j^{-1/2}\sum_{i=1}^{j}\epsilon_i\delta_{\mathbf{X}_i}\right\|_{\mathscr{H}_n} \\ &\le C\max_{n_0\le j\le n}\left\{\mathbb{P}^*\left(\int_0^{\Psi_{n,j}}\mathscr{N}(\epsilon)d\epsilon\right)^2\right\}^{1/2}\left(\mathbb{P}^*\tilde{\mathrm{M}}_n^2\right)^{1/2} \\ &\le C\max_{n_0\le j\le n}\left\{\mathbb{P}^*\left(\int_0^{\Psi_{n,j}}\mathscr{N}(\epsilon)d\epsilon\right)^2\right\}^{1/2}, \end{aligned}\tag{3.6.10}$$

where

$$\Psi_{n,j} = \sup\left\{\|f\|_{\mathbb{P}_j,2} : f\in\mathscr{H}_n\right\}.$$

Our aim is to show that

$$\Psi_{n_k,j_k} = o_{\mathbb{P}^*}(1) \quad \text{as } k\to\infty, \tag{3.6.11}$$

for an arbitrary subsequence $\{n_k : k = 1,2,\ldots\}$ of $\{n\}$, and any arbitrary sequence $\{j_k\}$ such that $n_{0,n_k} \le j_k \le n_k$ for all $k = 1,2,\ldots$. Write $m_k^* = m_{n_k}$. Define, for any $\gamma\in\mathscr{K}$,

$$\mathrm{Z}_{ki}(\gamma) = r_{m_k^*}^{\nu}\ell\left(r_{m_k^*}\right)m_k^{*-1/2}j_k^{-1/2}\tilde{m}_{\gamma/r_{m_k^*}}(\mathbf{X}_i).$$

As in the proof of part **(ii)** of Lemma 2 of Lee [2012]; he showed these variables satisfy the condition of Theorem 2.11.1 of van der Vaart and Wellner [1996], which is implies our result in (3.6.11) for arbitrary subsequence n_k and $j_k \in [n_{0,n_k}, n]$, by arguing as in the proof of this Theorem. It then follows by the dominated convergence theorem that the bound in (3.6.10) has limsup equal to 0 as $n \to \infty$. Substituting this and (3.6.9) into (3.6.8) to obtain the desired result. Finally, for the process $T_{3,m}$, for large value of m, we have $\boldsymbol{\theta}_0 + \frac{\gamma}{r_m} \in \Theta$ by using the assumption **(CB3)**, we get, for all $0 < \delta < \delta_1$,

$$
\begin{aligned}
&\sup_{\gamma,\gamma' \in \mathcal{K}, \|\gamma-\gamma'\| \le \delta} \left| T_{3,m}(\gamma) - T_{3,m}(\gamma') \right| \\
&= \sup_{\gamma,\gamma' \in \mathcal{K}, \|\gamma-\gamma'\| \le \delta} \left| r_m^{\nu-1} \ell(r_m) W_m(\gamma - \gamma') + r_m^{\nu-2} \ell(r_m) \left(V(\gamma,\gamma) - V(\gamma',\gamma') \right) \right. \\
&\quad \left. + r_m^{\nu} \ell(r_m) \left(o\left(\frac{\|\gamma\|^2}{r_m^2} \right) + o\left(\frac{\|\gamma'\|^2}{r_m^2} \right) \right) \right| \\
&\le \delta^{\tau} \left(r_m^{\nu-1} \ell(r_m) \sup_{\gamma \in \mathcal{K}, \delta \le \delta_1, \|\gamma\| \le \delta} \left| \frac{W_m(\gamma)}{\delta^{\tau}} \right| + r_m^{\nu-2} \ell(r_m) \sup_{\gamma,\gamma' \in \mathcal{K}, \delta \le \delta_1, \|\gamma-\gamma'\| \le \delta} \frac{\left| V(\gamma,\gamma) - V(\gamma',\gamma') \right|}{\delta^{\tau}} \right) \\
&\quad + b_m \\
&:= \delta^{\tau} \alpha_m + b_m, \qquad (3.6.12)
\end{aligned}
$$

where

$$
b_m \le \sup_{\gamma,\gamma' \in \mathcal{K}} r_m^{\nu} \ell(r_m) \left(o\left(\frac{\|\gamma\|^2}{r_m^2} \right) + o\left(\frac{\|\gamma'\|^2}{r_m^2} \right) \right) \to 0, \text{ as } m \to \infty,
$$

and $\alpha_m = O_{\mathbb{P}^*_W}(1)$ i.p. uniformly over $\delta \le \delta_1$. From this, we obtain, for any $\epsilon > 0$ and $\eta > 0$,

$$
\begin{aligned}
&\mathbb{P}^*_W \left(\sup_{\gamma,\gamma' \in \mathcal{K}, \|\gamma-\gamma'\| \le \delta} \left| T_{3,m}(\gamma) - T_{3,m}(\gamma') \right| > \epsilon \right) \\
&\le \mathbb{P}^*_W \left(\delta^{\tau} \alpha_m + b_m > \epsilon, \alpha_m \le C, |b_m| < \frac{\epsilon}{2} \right) + \mathbb{P}^*_W (\alpha_m > C) \\
&\le \mathbb{P}^*_W \left(\delta^{\tau} > \frac{\epsilon}{2C} \right) + \mathbb{P}^*_W (\alpha_m > C).
\end{aligned}
$$

By choosing C_η such that the last term is bounded by η for large value of m, and taking $\delta \le \delta_1 \wedge \left(\frac{\epsilon}{2C_\eta} \right)^{\frac{1}{\tau}}$, which implies the main result for the process $T_{3,m}$. Finally by using the fact that

$$
T_m = T_{1,m} + T_{2,m} + T_{3,m},
$$

we obtain the desired result on the process T_m. □

Proof of Theorem 3.3.3.3

Making use of the result **(i)** in Theorem 3.3.3.2 in connection with the assumption **(C8)**, we infer that we have almost all paths of the process $\gamma \mapsto \mathbb{G}(\gamma) + \Lambda(\gamma)$ are uniformly continuous on

every $\mathscr{K} \subset \mathscr{B}$, and reaching the supremum at an unique point γ_0. For part **(i)**, an application of **(i)** in Theorem 3.3.3.2, for any closed bounded $\mathscr{K} \subset \mathscr{B}$, gives

$$\mathbb{T}_n(\gamma) = r_n^{\nu}\ell(r_n)\mathbb{P}_n\tilde{\mathbf{m}}_{\frac{\gamma}{r_n},\widehat{h}} \rightsquigarrow \Lambda(\boldsymbol{\gamma}) + \mathbb{G}(\boldsymbol{\gamma}), \text{ in } \ell^{\infty}(\mathscr{K}).$$

We get from the assumption **(C11)** that

$$\mathbb{T}_n\left(r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)\right) \geq \sup_{\gamma \leq K} \mathbb{T}_n(\gamma) - o_{\mathbb{P}}(1).$$

Noting that γ_0 is the unique, well-separated, maximizer of $\mathbb{G}(\gamma) + \Lambda(\gamma)$, then part **(i)** follows by Theorem 3.2.2 of van der Vaart and Wellner [1996], where compact sets and uniform tightness of $r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$ are replaced respectively by closed bounded sets with similar structure as the set $\mathscr{K}$ and

$$r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) = \mathrm{O}_{\mathbb{P}^*}(1).$$

For part **(ii)**, we infer that

$$\widehat{\mathbb{T}}_m(\gamma) = r_m^{\nu}\ell(r_m)\widehat{\mathbb{P}}_m^*\tilde{\mathbf{m}}_{\frac{\gamma}{r_m},\widehat{h}_m} \rightsquigarrow \Lambda(\boldsymbol{\gamma}) + \mathbb{G}(\boldsymbol{\gamma}) \text{ i.p. in } \ell(\mathscr{K}).$$

By combining the assumption **(CB5)** with the first part of **(CB1)**, we have respectively

$$\widehat{\mathbb{T}}_m\left(r_m(\widehat{\boldsymbol{\theta}}_m^* - \boldsymbol{\theta}_0)\right) \geq \sup_{\gamma \leq K} \widehat{\mathbb{T}}_m(\gamma) - o_{\mathbb{P}_W^*}(1)$$

and

$$r_m(\widehat{\boldsymbol{\theta}}_m^* - \boldsymbol{\theta}_0) = \mathrm{O}_{\mathbb{P}_W^*}(1), \text{ i.p.}$$

An application of Lemma 4(ii) of Lee [2012] gives

$$r_m(\widehat{\boldsymbol{\theta}}_m^* - \boldsymbol{\theta}_0) \rightsquigarrow \gamma_0, \text{ i.p.}$$

It follows from the first part of the assumptions **(C1)**, **(AB1)** and Slutsky's theorem that

$$r_m(\widehat{\boldsymbol{\theta}}_m^* - \boldsymbol{\theta}_n) = r_m(\widehat{\boldsymbol{\theta}}_m^* - \boldsymbol{\theta}_0) - \frac{r_m}{r_n} r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \rightsquigarrow \gamma_0 \text{ i.p.}$$

Hence the proof of the statement **(ii)** is complete. □

3.7 References

Alin, A., Martin, M. A., Beyaztas, U., and Pathak, P. K. (2017). Sufficient m-out-of-n (m/n) bootstrap. *J. Stat. Comput. Simul.*, **87**(9), 1742–1753. 66

Allaire, G. (2005). *Analyse numérique et optimisation: une introduction à la modélisation mathématique et à la simulation numérique*. Editions Ecole Polytechnique. 57

Alvarez-Andrade, S. and Bouzebda, S. (2013). Strong approximations for weighted bootstrap of empirical and quantile processes with applications. *Stat. Methodol.*, **11**, 36–52. 49

Alvarez-Andrade, S. and Bouzebda, S. (2015). On the local time of the weighted bootstrap and compound empirical processes. *Stoch. Anal. Appl.*, **33**(4), 609–629. 49

Alvarez-Andrade, S. and Bouzebda, S. (2019). Some selected topics for the bootstrap of the empirical and quantile processes. *Theory Stoch. Process.*, **24**(1), 19–48. 49

Arcones, M. A. and Giné, E. (1992). On the bootstrap of M-estimators and other statistical functionals. In *Exploring the limits of bootstrap (East Lansing, MI, 1990)*, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., pages 13–47. Wiley, New York. 50

Bertail, P. (1997). Second-order properties of an extrapolated bootstrap without replacement under weak assumptions. *Bernoulli*, **3**(2), 149–179. 67

Bertail, P., Politis, D. N., and Romano, J. P. (1999). On subsampling estimators with unknown rate of convergence. *J. Amer. Statist. Assoc.*, **94**(446), 569–579. 67

Bickel, P. J. and Sakov, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statist. Sinica*, **18**(3), 967–985. 66, 67

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD. 48

Bickel, P. J., Götze, F., and van Zwet, W. R. (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statist. Sinica*, **7**(1), 1–31. Empirical Bayes, sequential analysis and related topics in statistics and probability (New Brunswick, NJ, 1995). 49, 66, 67

Bose, A. and Chatterjee, S. (2001). Generalised bootstrap in non-regular M-estimation problems. *Statist. Probab. Lett.*, **55**(3), 319–328. 49, 50

Bouzebda, S. (2010). Bootstrap de l'estimateur de Hill: théorèmes limites. *Ann. I.S.U.P.*, **54**(1-2), 61–72. 49

Bouzebda, S. and Limnios, N. (2013). On general bootstrap of empirical estimator of a semi-Markov kernel with applications. *J. Multivariate Anal.*, **116**, 52–62. 49

Bouzebda, S., Papamichail, C., and Limnios, N. (2018). On a multidimensional general bootstrap for empirical estimator of continuous-time semi-Markov kernels with applications. *J. Nonparametr. Stat.*, **30**(1), 49–86. 49

Chen, X., Linton, O., and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, **71**(5), 1591–1608. 48

Cheng, G. and Huang, J. Z. (2010). Bootstrap consistency for general semiparametric M-estimation. *Ann. Statist.*, **38**(5), 2884–2915. 58, 85, 87

Chernick, M. R. (2008). *Bootstrap methods: a guide for practitioners and researchers*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition. 49

Datta, S. and McCormick, W. P. (1995). Bootstrap inference for a first-order autoregression with positive innovations. *J. Amer. Statist. Assoc.*, **90**(432), 1289–1300. 66

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*, volume 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. With 1 IBM-PC floppy disk (3.5 inch; HD). 49

Delsol, L. and Van Keilegom, I. (2020). Semiparametric M-estimation with non-smooth criterion functions. *Ann. Inst. Statist. Math.*, **72**(2), 577–605. 51, 55, 58, 59, 64, 65, 67, 70, 81, 85

Dudley, R. M. (1999). *Uniform central limit theorems*, volume 63 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge. 50

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**(1), 1–26. 49

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York. 49

El Bantli, F. (2004). M-estimation in linear models under nonstandard conditions. *J. Statist. Plann. Inference*, **121**(2), 231–248. 49, 50

Giné, E. and Zinn, J. (1989). Necessary conditions for the bootstrap of the mean. *Ann. Statist.*, **17**(2), 684–691. 49

Götze, F. and Račkauskas, A. (2001). Adaptive choice of bootstrap sample sizes. In *State of the art in probability and statistics (Leiden, 1999)*, volume 36 of *IMS Lecture Notes Monogr. Ser.*, pages 286–309. Inst. Math. Statist., Beachwood, OH. 66, 67

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York. 49

Hall, P., Horowitz, J. L., and Jing, B.-Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, **82**(3), 561–574. 66

Hoffmann-Jørgensen, J. (1991). *Stochastic processes on Polish spaces*. Various publications series. Aarhus Universitet. Matematisk Institut. 49, 53

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics*, **58**(1-2), 71–120. 68

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.*, **18**(1), 191–219. 50

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer Series in Statistics. Springer, New York. 48, 49, 63

Koul, H. L., Müller, U. U., and Schick, A. (2012). The transfer principle: a tool for complete case analysis. *Ann. Statist.*, **40**(6), 3031–3049. 69

Kristensen, D. and Salanié, B. (2017). Higher-order properties of approximate estimators. *J. Econometrics*, **198**(2), 189–208. 63

Lahiri, S. N. (1992). On bootstrapping M-estimators. *Sankhyā Ser. A*, **54**(2), 157–170. 50

Lee, S. M. S. (2012). General M-estimation and its bootstrap. *J. Korean Statist. Soc.*, **41**(4), 471–490. 50, 51, 53, 55, 64, 65, 73, 81, 88, 90, 91

Lee, S. M. S. and Pun, M. C. (2006). On m out of n bootstrapping for nonstandard M-estimation with nuisance parameters. *J. Amer. Statist. Assoc.*, **101**(475), 1185–1197. 50, 63, 67

Lee, S. M. S. and Yang, P. (2020). Bootstrap confidence regions based on M-estimators under nonstandard conditions. *Ann. Statist.*, **48**(1), 274–299. 50

Ma, S. and Kosorok, M. R. (2005). Robust semiparametric m-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, **96**(1), 190–217. 58, 63

Müller, U. U. (2009). Estimating linear functionals in nonlinear regression with responses missing at random. *Ann. Statist.*, **37**(5A), 2245–2277. 69

Pakes, A. and Olley, S. (1995). A limit theorem for a smooth class of semiparametric estimators. *J. Econometrics*, **65**(1), 295–332. 48

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, **57**(5), 1027–1057. 48

Pérez-González, A., Vilar-Fernández, J. M., and González-Manteiga, W. (2009). Asymptotic properties of local polynomial regression with missing data and correlated errors. *Ann. Inst. Statist. Math.*, **61**(1), 85–109. 69

Pfanzagl, J. (1990). *Estimation in semiparametric models*, volume 63 of *Lecture Notes in Statistics*. Springer-Verlag, New York. Some recent developments. 48

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.*, **22**(4), 2031–2050. 67

Politis, D. N., Romano, J. P., and Wolf, M. (1999a). *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York. 66

Politis, D. N., Romano, J. P., and Wolf, M. (1999b). *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York. 67

Politis, D. N., Romano, J. P., and Wolf, M. (2001). On the asymptotic theory of subsampling. *Statist. Sinica*, **11**(4), 1105–1124. 67

Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory*, **1**(3), 295–313. 50

Romano, J. P. and Shaikh, A. M. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *Ann. Statist.*, **40**(6), 2798–2822. 67

Shao, J. and Tu, D. S. (1995). *The jackknife and bootstrap*. Springer Series in Statistics. Springer-Verlag, New York. 49

Swanepoel, J. W. H. (1986). A note on proving that the (modified) bootstrap works. *Comm. Statist. A—Theory Methods*, **15**(11), 3193–3203. 66

van de Geer, S. A. (2000). *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. 48, 50

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics. 48, 49, 50, 51, 52, 53, 54, 57, 58, 61, 62, 63, 65, 70, 71, 81, 85, 86, 88, 89, 90, 91

Wei, B., Lee, S. M. S., and Wu, X. (2016). Stochastically optimal bootstrap sample size for shrinkage-type statistics. *Stat. Comput.*, **26**(1-2), 249–262. 67

Wellner, J. A. and Zhan, Y. (1996). Bootstrapping Z-estimators. *Preprint.* 50

Zhan, Y. (2002). Central limit theorems for functional Z-estimators. *Statist. Sinica*, **12**(2), 609–634. 48

Chapter 4

Central limit theorems for functional Z-estimators with Functional Nuisance Parameters

Ce chapitre développe le contenu d'un article soumis, mis en forme pour être inséré dans le présent manuscrit de thèse.
Title : *Central limit theorems for functional Z-estimators with Functional Nuisance Parameters*.

Abstract

We consider an exchangeably weighted bootstrap for function-valued estimators defined as a zero point of a function-valued random criterion function. A large number of bootstrap resampling schemes emerge as special cases of our settings. The main ingredient is the use of a differential identity that applies when the random criterion function is linear in terms of the empirical measure. Our results are general and do not require linearity of the statistical model in terms of the unknown parameter. We consider also the semiparametric models extending the work of Zhan [2002] to a more delicate framework. The theoretical results established in this paper, are (or will be) key tools for many further developments in the parametric and the semiparametric models.

Key words : Bootstrap; Z-estimators; estimating equations; exchangeable; infinite-dimensional; nonparametric maximum likelihood; score equation; self-consistency; weak convergence; Semiparametric inference; M-estimators.
Mathematics Subject Classification : 62F03, 62F10, 62F12, 62H12, 62H15.

4.1 Introduction and motivations

Parametric estimation has been the subject of intense investigation for many years and this has led to the development of a large variety of methods. Because of numerous applications and their important role in mathematical statistics, the problem of estimating the parametric models

has been the subject of considerable interest during the last decades. For good sources of references to research literature in this area along with statistical Pfanzagl [1994], Lindsey [1996], Bickel *et al.* [1998], Lehmann and Casella [1998], van der Vaart [1998], Lehmann and Romano [2005] and Cheng [2017]. Assume that the model can be parameterized as $\boldsymbol{\theta} \mapsto \mathbb{P}_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta}$ is the parameter that we are interested in. Various parametric methods of estimation have been extensively investigated, among others, including the method of moments, Least Square Estimator (LSE), Maximum of Likelihood (ML) and Delta method. Attention was confined to parametric models and much effort has been expended in constructing efficient estimators. This paper is devoted to the investigation of "Z-estimators" in a general setting, in other words, estimators that are the solutions of estimating equations. Note that the limiting distributions of the resulting Z-estimators are rather complicated, which does not permit explicit computation in practice. To overcome that difficulty, we shall propose a general bootstrap and study some of its asymptotic properties by means of the modern theory of the empirical processes. The interest in considering general bootstrap instead of particular cases lies in the fact that we need, in general, a more flexible modeling to handle the problems in practice. In a variety of statistical problems, the bootstrap provides a simple method for circumventing technical difficulties due to intractable distribution theory and has become a powerful tool for setting confidence intervals and critical values of tests for composite hypotheses. Bootstrap samples were introduced and first investigated in Efron [1979]. Since this seminal paper, bootstrap methods have been proposed, discussed, investigated and applied in a huge number of papers in the literature. Being one of the most important ideas in the practice of statistics, the bootstrap also introduced a wealth of innovative probability problems, which in turn formed the basis for the creation of new mathematical theories. The asymptotic theory of the bootstrap with statistical applications has been reviewed in the books among others Chernick [2008a], Manly [2007], Good [2005], Chernick [2008b], Davison and Hinkley [1997], van der Vaart and Wellner [1996], Hall [1992] and Kosorok [2008]. A substantial body of literature, reviewed in Beran [2003], gives conditions for the bootstrap to be satisfied in order to provide desirable distributional approximations. In Bickel *et al.* [1997], the performance of different kinds of bootstrap procedures is investigated through asymptotic results and small sample simulation studies. Note that the bootstrap, according to Efron's original formulation (see Efron [1979]), presents some drawbacks. Namely, some observations may be used more than once while others are not sampled at all. To overcome that problem, a more general formulation of the bootstrap has been introduced, the *weighted* (or *smooth*) bootstrap, which has also been shown to be computationally more efficient in several applications. For a survey of further results on weighted bootstrap the reader is referred to Barbe and Bertail [1995]. Another resampling scheme was proposed in Rubin [1981] and was extensively studied by Bickel and Freedman [1981], who suggested the name "weighted bootstrap", e.g., Bayesian Bootstrap when the vector of weights

$$(\mathrm{W}_{n1}, \ldots, \mathrm{W}_{nn}) = (\mathrm{D}_{n1}, \ldots, \mathrm{D}_{nn}),$$

is equal in distribution to the vector of n spacings of $n-1$ ordered uniform $(0,1)$ random variables, that is,

$$(\mathrm{D}_{n1},\ldots,\mathrm{D}_{nn}) \sim \text{ Dirichlet}(n;1,\ldots,1).$$

The interested reader may refer to Lo [1993]. The case

$$(\mathrm{D}_{n1},\ldots,\mathrm{D}_{nn}) \sim \text{ Dirichlet}(n;4,\ldots,4),$$

was considered in [Weng, 1989, Remark 2.3] and [Zheng and Tu, 1988, Remark 5]. These resampling plans lead to the interest of a unified approach, generically designated as general weighted resampling, was first proposed by Mason and Newton [1992] and amongst others extended by Præstgaard and Wellner [1993].
The main purpose of the present work is to consider a general framework of the bootstrap of the Z-estimators completing the work of Zhan [2002]. More precisely, we consider the exchangeable bootstrapped version of the Z-estimators investigated in Zhan [2002]. Zhan [2002] showed that the Z-estimators converge weakly to some process which is hard to evaluate for practical use. To overcome this problem, we propose in this paper the exchangeable bootstrap. The main aim of the present paper is to provide a first full theoretical justification of the exchangeable bootstrap consistency of Z-estimators with the same spirit of Zhan [2002]. This requires the effective application of large sample theory techniques, which were developed for the empirical processes. The Zhan [2002] results are not directly applicable here since we are considering the bootstrapped versions. These results are not only useful in their own right but essential for the derivation of our asymptotic results. At this point, it is worth noting that the approaches adopted in the present paper are different from those used in Zhan [1996], where the traditional arguments are used. The second aim of this work is to consider the semiparametric Z-estimators. Semiparametric models are statistical models where at least one parameter of interest is not Euclidean. The most basic scenario is one in which the finite-dimensional parameters are the parameters of interest, the unknown functions, also called infinite-dimensional parameters, are nuisance parameters. The success of semiparametric methods is due to both; their excellent scientific intriguing theoretical and flexibility of modeling framework for complex data, and proven to be useful in a variety of contexts Banerjee *et al.* [2009], Cheng [2009], Huang [1999], Zeng and Lin [2007], Zhang and Yu [2008], Cheng and Huang [2010] and Ma and Kosorok [2005]. To highlight the importance of the semiparametric models, Kosorok [2009] expanded the scope of the review of Wellner *et al.* [2006] into new domains, including scientific philosophy and graduate education, as well as to touch on a few additional theoretical aspects not discussed previously. The second aim of the present paper is to extend the work of Zhan [2002] to the delicate semiparametric setting. In particular, we have extended the key tool of Zhan [2002], Lemma 4.2.2.1, that will be instrumental for the generalisation to the semiparametric framework.

The paper is organized as follows. Section 4.2 introduces the notation and the framework of the exchangeable bootstraps. Section 4.2.2 states the main theorems, Theorem 4.2.2.5 and 4.2.2.8 for the limiting distributions. In Section 4.3 we apply our theorems to some non-trivial examples

including the random right censoring, the simplified frailty model and the double censoring model. Section 4.4 provides the results for the semiparametric setting where the main results are presented in Theorems 4.5.3.1 and 4.5.4.2. All proofs are gathered in Section 4.6.

4.2 Bootstrapped Z-estimators

In the sequel, we use a notation similar to that used in Zhan [2002] including some changes absolutely necessary for our setting. Let $X_i, i = 1,2,\ldots$ be independent and identically distributed observations from a distribution $\mathbb{P} \in \mathscr{P}$ on a probability space $(\mathscr{X}, \mathscr{A})$, where $\mathscr{P}$ denotes the set of all probability measures on $(\mathscr{X}, \mathscr{A})$. For definiteness and for ease of dealing with measurability issues, we view X_i as the i th coordinate projection from the canonical probability space $(\mathscr{X}^\infty, \mathscr{A}^\infty, \mathbb{P}^\infty)$ into the i th copy of $\mathscr{X}$. Suppose that the collection $\mathscr{P}$ is parametrized by $\boldsymbol{\theta} \in \Theta$, where Θ is assumed to be a smooth surface in a Banach space $(\mathbf{B}, \|\cdot\|)$ with a norm $\|\cdot\|$. We are interested in estimating a functional parameter $\boldsymbol{\theta}_0 \in \Theta$, the true parameter. Let $\ell^\infty(\mathscr{H})$ denote the set of bounded functions from $\mathscr{H}$ to $\mathbb{R}$, for some set $\mathscr{H}$, and let $\|\cdot\|_{\mathscr{H}}$ denote the uniform norm on $\ell^\infty(\mathscr{H})$. Suppose that ψ is a sequence of random maps (functions of the data $X_1, \ldots, X_n$) from Θ to $\ell^\infty(\mathscr{H})$. Thus $\psi(\boldsymbol{\theta})(h)$ is a real-valued random variable for each $\boldsymbol{\theta} \in \Theta$ and $h \in \mathscr{H}$, and $\|\psi(\boldsymbol{\theta})\|_{\mathscr{H}} < \infty$ for each $\boldsymbol{\theta} \in \Theta$. These are often given by

$$\psi(\boldsymbol{\theta}, \mathbb{P}_n)(h) = \mathbb{P}_n \mathrm{B}(\boldsymbol{\theta})(h) \quad \text{for} \quad h \in \mathscr{H},$$

where $\mathrm{B}(\boldsymbol{\theta})$ is a map (the score operator for Θ) from $\mathscr{H}$ to some subset $\mathscr{F}(\boldsymbol{\theta})$ of $L_2(\mathbb{P}_{\boldsymbol{\theta}})$ for each $\boldsymbol{\theta} \in \Theta$. Define the set $\mathscr{F}(\Theta) = \bigcup_{\boldsymbol{\theta} \in \Theta} \mathscr{F}(\boldsymbol{\theta})$. For simplicity of notation, we omit Θ in $\mathscr{F}(\Theta)$ and simply write $\mathscr{F}$, for instance, see van der Vaart [1995] for a similar formulation. The empirical measure $\mathbb{P}_n$ of the first n observations is defined by

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

and the empirical process is

$$\mathbb{G}_n = \sqrt{n}\,(\mathbb{P}_n - \mathbb{P})$$

as usual we will use linear functional notation, and write $\mathbb{P}f = \int f d\mathbb{P}$ for $f \in \mathscr{F} \subset L_2(\mathbb{P})$, and we will consider $\mathbb{G}_n$ as indexed by some collection of functions $\mathscr{F}$ and $\mathbb{P} \in \mathscr{P}$. Suppose that ψ is a deterministic map from Θ to $\ell^\infty(\mathscr{H})$; this can be viewed as the "population version" of the maps ψ. When ψ is given in terms of $\mathrm{B}(\boldsymbol{\theta})$ as above,

$$\psi(\boldsymbol{\theta}, \mathbb{P})(h) = \mathbb{P}\mathrm{B}(\boldsymbol{\theta})(h) \quad \text{for} \quad h \in \mathscr{H}.$$

Suppose that $\mathrm{B}(\boldsymbol{\theta})$ is bounded in the sense that $\|\psi(\boldsymbol{\theta}, \mathbb{P})\|_{\mathscr{H}} = \|\mathbb{P}\mathrm{B}(\boldsymbol{\theta})\|_{\mathscr{H}} < \infty$ for all $\mathbb{P} \in \mathscr{P}$. Then $\psi(\boldsymbol{\theta}, \mathbb{P}) \in \ell^\infty(\mathscr{H})$ for each fixed $\boldsymbol{\theta} \in \Theta$. The empirical process $\mathbb{G}_n \mathrm{B}(\boldsymbol{\theta})$ acting on $\mathrm{B}(\boldsymbol{\theta})$ is also a function in $\ell^\infty(\mathscr{H})$ for fixed $\boldsymbol{\theta} \in \Theta$. An asymptotic functional Z-estimator is a sequence of estimators $\{\hat{\boldsymbol{\theta}}_n\}$ of $\boldsymbol{\theta}_0$ which makes the "scores" $\mathbb{P}_n \mathrm{B}(\boldsymbol{\theta})(h)$ approximately zero: that is

$$\left\|\psi\left(\hat{\boldsymbol{\theta}}_n, \mathbb{P}_n\right)\right\|_{\mathscr{H}} = o_{\mathbb{P}^*}\left(n^{-1/2}\right).$$

In this section we will focus on the asymptotic validity of the exchangeably weighted bootstrap of infinite-dimensional Z-estimators in the following sense: if the weak convergence of this estimator demonstrated by Zhan [2002], then the asymptotic consistency of the bootstrapped version is guaranteed by some additional assumptions as we will describe in the following subsections. Let us recall the main idea of his work. First to prove the weak convergence of a Z-estimators and its bootstrapped version, the common technique used is to write it as a linear approximation of some process which converges to some Brownian bridge, i.e.,

$$\psi\left(\hat{\boldsymbol{\theta}}_n, \mathbb{P}\right) - \psi\left(\boldsymbol{\theta}_0, \mathbb{P}\right) = \dot{\psi}\left(\boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) + o_{\mathbb{P}^*}\left(\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|\right), \tag{4.2.1}$$

and

$$\psi\left(\hat{\boldsymbol{\theta}}_n^*, \mathbb{P}\right) - \psi\left(\boldsymbol{\theta}_0, \mathbb{P}\right) = \dot{\psi}\left(\boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0\right) + o_{\mathbb{P}^*}\left(\left\|\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0\right\|\right), \tag{4.2.2}$$

where $\hat{\boldsymbol{\theta}}_n^*$ is the bootstrapped version of $\hat{\boldsymbol{\theta}}_n$ defined in the next section. This linearisation is validated by using the Taylor series applied to $\psi\left(\boldsymbol{\theta}, \mathbb{P}\right)$, which is Fréchet differentiable with respect to the norm $\|\cdot\|$ of the parameter space Θ. The theory of empirical process can be used to rewrite the last equations respectively as;

$$\dot{\psi}\left(\boldsymbol{\theta}_0\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right) = -\mathbb{G}_n \mathrm{B}\left(\boldsymbol{\theta}_0\right) + o_{\mathbb{P}^*}\left(\sqrt{n}\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|\right) + o_{\mathbb{P}^*}(1), \tag{4.2.3}$$

and

$$\dot{\psi}\left(\boldsymbol{\theta}_0\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_n\right)\right) = -\hat{\mathbb{G}}_n \mathrm{B}\left(\boldsymbol{\theta}_0\right) + o_{\mathbb{P}^*}\left(\sqrt{n}\left\|\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0\right\|\right) + o_{\mathbb{P}^*}\left(\sqrt{n}\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|\right) + o_{\mathbb{P}^*}(1), \tag{4.2.4}$$

were $\hat{\mathbb{G}}_n$ is defined in (4.2.9), then by the boundedness and the invertibility of the derivative $\dot{\psi}\left(\boldsymbol{\theta}_0\right)$ with respect to the same norm $\|\cdot\|$, we can prove that $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\theta}}_n^*$ are $\sqrt{n}-$consistency. By assuming that; $\mathbb{G}_n \mathrm{B}\left(\boldsymbol{\theta}_0\right) \rightsquigarrow \mathbb{Z}_0$ and $\hat{\mathbb{G}}_n \mathrm{B}\left(\boldsymbol{\theta}_0\right) \rightsquigarrow \hat{\mathbb{Z}}_0$ it follows that $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \rightsquigarrow -\dot{\psi}^{-1}\left(\boldsymbol{\theta}_0\right)\left(\mathbb{Z}_0\right)$ and $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_n\right) \rightsquigarrow -\dot{\psi}^{-1}\left(\boldsymbol{\theta}_0\right)\left(\hat{\mathbb{Z}}_0\right)$ are asymptotically normal, by applying the continuous mapping theorem, as a consequence the desired result hold, for more detail we refer the reader to van der Vaart [1995] and Wellner and Zhan [1996]. The problem which can occur is that the invertibility of the derivative operator $\dot{\psi}\left(\boldsymbol{\theta}_0\right)$ and the differentiability of the function $\psi(\boldsymbol{\theta}, \mathbb{P})$ cannot hold with respect to the same norm $\|\cdot\|$. Let us clarify this point and we focus in estimating the distribution function in double censoring modal where the natural parameter space Θ in this case is the set of all distribution functions on $[0, \infty)$ as described in more details in example (4.3.3). In such case we remark that; the derivative $\dot{\psi}\left(\boldsymbol{\theta}_0\right)$ is only invertible with respect to the weaker norm $\|\cdot\|_{\mathrm{K}}$ and the differentiability of $\psi(\boldsymbol{\theta}, \mathbb{P})$ can hold only for the strong norm $\|\cdot\|$, consequently, the preceding arguments fail to demonstrate the weak convergence of the Z-estimators and of its bootstrapped version. To overcome this problem, Zhan [2002] developed an identity in his Lemma 2.1 which connect the Fréchet differentiability of the function $\psi(\boldsymbol{\theta}, \mathbb{P})$ and the Fréchet differentiability of the function $\boldsymbol{\theta} \mapsto \mathbb{P}_{\boldsymbol{\theta}}$, as a consequence he obtained a connection between $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$ and $\mathbb{G}_n \mathrm{B}\left(\boldsymbol{\theta}_0\right)$:

$$\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right) = -\dot{\mathbb{P}}_{\hat{\boldsymbol{\theta}}_n}\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right) \mathrm{B}\left(\hat{\boldsymbol{\theta}}_n\right).$$

This identity was derived before in ad hoc manner, in the multiplicative censoring model Vardi and Zhang [1992], in the double censoring model Gu and Zhang [1993] and in the interval truncation model Tsai and Zhang [1995]. To examine the efficiency of the MLE van der Laan [1995] also derived an identity. Note that the common feature in these problems is that the probability measures $\mathbb{P}_{\boldsymbol{\theta}}$ are convex linearly indexed by $\boldsymbol{\theta}$ as in the multiplicative and double censoring model, or nearly so up to a normalizing constant as in the interval truncation model. For a general class of models in which convex linearity can be boundedly extended to the linear span of the parameter space, such linearity identity can be proved by Fréchet differentiability of the likelihood equations $\psi(\boldsymbol{\theta}, \mathbb{P}_{\boldsymbol{\theta}}) = 0$. For more details see Section 2.1 of Zhan [2002] also see Wellner and Zhan [1996]. The idea is to allow a linearization applied to $\mathbb{P}_{\boldsymbol{\theta}}$ instead of $\psi(\boldsymbol{\theta}, \mathbb{P})$ through its derivative operator $\dot{\mathbb{P}}_{\boldsymbol{\vartheta}}(\cdot)$. Models where $\mathbb{P}_{\boldsymbol{\theta}}$ is bounded convex linearity, the differential $\dot{\mathbb{P}}_{\hat{\boldsymbol{\theta}}_n}\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right) \mathrm{B}\left(\hat{\boldsymbol{\theta}}_n\right)$ exactly equals the difference $\sqrt{n}\left(\psi\left(\hat{\boldsymbol{\theta}}_n, \mathbb{P}_{\hat{\boldsymbol{\theta}}_n}\right) - \psi\left(\hat{\boldsymbol{\theta}}_n, \mathbb{P}_{\boldsymbol{\theta}_0}\right)\right)$. Consequently, we have

$$\begin{aligned} \dot{\psi}\left(\hat{\boldsymbol{\theta}}_n\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right) &= -\sqrt{n}\psi\left(\hat{\boldsymbol{\theta}}_n, \mathbb{P}_{\boldsymbol{\theta}_0}\right) \\ &= -\mathbb{G}_n \mathrm{B}\left(\boldsymbol{\theta}_0\right) + o_{\mathbb{P}^*}(1). \end{aligned} \tag{4.2.5}$$

One can see the difference between the last expression and the other given in (4.2.3), where $\hat{\boldsymbol{\theta}}_n$ must converge with $n^{-1/2}$ rate with respect to the norm $\|\cdot\|$ to validate the linearization, then the asymptotic normality, such a condition is not needed in (4.2.5) because the linearization is perfect. The sequence of Z-estimators $\hat{\boldsymbol{\theta}}_n$ still converge at the $n^{-1/2}$ rate in any norm as long as the derivative operators $\dot{\psi}(\hat{\boldsymbol{\theta}}_n)$ is boundedly invertible with respect to it. This circumvents the problem described before by obtaining a weak convergence as well as the rate control in one step. Intuitively the same problem can occur in the study of the bootstrapped version $\hat{\boldsymbol{\theta}}_n^*$ in such case, so we will use the same arguments given by Zhan [2002] with some additional arguments needed for the bootstrap study, which allow us to show the weak convergence in probability of the bootstrapped Z-estimators $\hat{\boldsymbol{\theta}}_n^*$.

4.2.1 Definitions and notation

We begin by some definitions which is needed in the following results described in Section 4.2.2. The function $\psi(\boldsymbol{\theta}, \mathbb{P})$, as a map from Θ to $\ell^\infty(\mathscr{H})$, is Fréchet differentiable with respect to the norm $\|\cdot\|$ at a point $\boldsymbol{\vartheta} \in \Theta$ if there is a bounded linear operator $\dot{\psi}\left(\boldsymbol{\vartheta}, \mathbb{P}_{\boldsymbol{\vartheta}}\right)(\cdot)$ mapping from $(\mathrm{lin}(\Theta), \|\cdot\|)$ to $(\ell^\infty(\mathscr{H}), \|\cdot\|_{\mathscr{H}})$ such that

$$\left\|\psi\left(\boldsymbol{\theta}, \mathbb{P}_{\boldsymbol{\vartheta}}\right) - \psi\left(\boldsymbol{\vartheta}, \mathbb{P}_{\boldsymbol{\vartheta}}\right) - \dot{\psi}\left(\boldsymbol{\vartheta}, \mathbb{P}_{\boldsymbol{\vartheta}}\right)\left(\boldsymbol{\theta} - \boldsymbol{\vartheta}\right)\right\|_{\mathscr{H}} = o(\|\boldsymbol{\theta} - \boldsymbol{\vartheta}\|).$$

Denote the operator $\dot{\psi}\left(\boldsymbol{\theta}, \mathbb{P}_{\boldsymbol{\theta}}\right)$ by $\dot{\psi}(\boldsymbol{\theta}) : \dot{\psi}(\boldsymbol{\theta}) \equiv \dot{\psi}\left(\boldsymbol{\theta}, \mathbb{P}_{\boldsymbol{\theta}}\right)$. The operator $\dot{\psi}(\boldsymbol{\theta})$ is continuous as a function of $\boldsymbol{\theta}$ at $\boldsymbol{\vartheta}$ if

$$\|\dot{\psi}(\boldsymbol{\theta}) - \dot{\psi}(\boldsymbol{\vartheta})\| \equiv \sup_{\|a\| \leq 1} \|\dot{\psi}(\boldsymbol{\theta})(a) - \dot{\psi}(\boldsymbol{\vartheta})(a)\|_{\mathscr{H}} \longrightarrow 0 \tag{4.2.6}$$

as $\|\boldsymbol{\theta} - \boldsymbol{\vartheta}\| \to 0$. Recall that for a fixed $\boldsymbol{\vartheta} \in \Theta$, the operator $\mathrm{B}(\boldsymbol{\vartheta})$ is bounded in the sense that $\|\mathbb{P}\mathrm{B}(\boldsymbol{\vartheta})\|_{\mathscr{H}} < \infty$ for all $\mathbb{P} \in \mathscr{P}$. Thus for a fixed $\boldsymbol{\vartheta} \in \Theta$, the probability measure $\mathbb{P}_{\boldsymbol{\theta}}$ induces

a mapping $\boldsymbol{\theta} \mapsto \mathbb{P}_{\boldsymbol{\theta}}\mathrm{B}(\boldsymbol{\vartheta})$ from Θ to $\ell^\infty(\mathscr{H})$. The map $\mathbb{P}_{\boldsymbol{\theta}}\mathrm{B}(\boldsymbol{\vartheta})$, as a function of $\boldsymbol{\theta}$, is Fréchet differentiable with respect to the norm $\|\cdot\|$ at a point $\boldsymbol{\vartheta} \in \Theta$ if there is a linear operator $\dot{\mathbb{P}}_{\boldsymbol{\vartheta}}(\cdot)$ such that $\dot{\mathbb{P}}_{\boldsymbol{\vartheta}}(\cdot)\mathrm{B}(\boldsymbol{\vartheta})$ is bounded and

$$\left\|\mathbb{P}_{\boldsymbol{\theta}}\mathrm{B}(\boldsymbol{\vartheta}) - \mathbb{P}_{\boldsymbol{\vartheta}}\mathrm{B}(\boldsymbol{\vartheta}) - \dot{\mathbb{P}}_{\boldsymbol{\vartheta}}(\boldsymbol{\theta}-\boldsymbol{\vartheta})\mathrm{B}(\boldsymbol{\vartheta})\right\|_{\mathscr{H}} = o(\|\boldsymbol{\theta}-\boldsymbol{\vartheta}\|).$$

Given that Θ is a subset in a Banach space $(\mathbf{B}, \|\cdot\|)$, the closure $\overline{\mathrm{lin}(\Theta)}$ is a Banach space with the same norm $\|\cdot\|$ (Lemma II.1.3 on page 50 , Dunford and Schwartz [1958], Part I). Because $(\ell^\infty(\mathscr{H}), \|\cdot\|_{\mathscr{H}})$ is also a Banach space, the bounded operators $\dot{\psi}^{-1}(\boldsymbol{\theta})$ and $\dot{\psi}(\boldsymbol{\theta})$ can be uniquely extended to the closures of their domains by continuity (see, e.g., Lemma I.6.16 on page 23 of Dunford and Schwartz [1958], Part I). The unique continuous extensions of $\dot{\psi}^{-1}(\boldsymbol{\theta})$ and $\dot{\psi}(\boldsymbol{\theta})$ on the closures of their domains are also denoted by $\dot{\psi}^{-1}(\boldsymbol{\theta})$ and $\dot{\psi}(\boldsymbol{\theta})$. The extension $\dot{\psi}^{-1}(\boldsymbol{\theta})$ on $\overline{\mathscr{R}(\dot{\psi})}$ is also the inverse of the extension $\dot{\psi}(\boldsymbol{\theta})$ on $\mathrm{lin}(\Theta)$. For the examples we deal with in Section 4.3, and for other examples, it is true that $\overline{\mathscr{R}(\dot{\psi})}$ does not depend on $\boldsymbol{\theta}$. We use $\overline{\mathscr{R}(\dot{\psi})}$ instead of $\overline{\mathscr{R}(\dot{\psi}(\boldsymbol{\theta}))}$ to denote the common subspace on which every $\dot{\psi}^{-1}(\boldsymbol{\theta})$ reside. Now consider bootstrapping the functional Z-estimators described before. We shall consider a wide class of bootstrap procedures as possible. We will use the notation and results of Præstgaard and Wellner [1993] for "exchangeably weighted" bootstraps. We suppose that the bootstrap weights $\mathbf{W} = \{\mathrm{W}_{ni}, i = 1,2,\ldots,n, n = 1,2,\ldots\}$ are a triangular array defined on the probability space $(\mathscr{Z}, \mathscr{E}, \widehat{\mathbb{P}})$. Let $\mathrm{W}_n \equiv (\mathrm{W}_{n1},\ldots,\mathrm{W}_{nn})$ be an exchangeable vector of nonnegative weights which sum to n. Then the exchangeably weighted bootstrap empirical measure is defined by

$$\widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{W}_{ni}\delta_{\mathrm{X}_i}.$$

The bootstrap scores are defined by

$$\psi\left(\boldsymbol{\theta}, \widehat{\mathbb{P}}_n\right)(h) = \widehat{\mathbb{P}}_n\mathrm{B}(\boldsymbol{\theta})(h) \quad \text{for} \quad h \in \mathscr{H}. \tag{4.2.7}$$

A bootstrap asymptotic Z-estimator $\widehat{\boldsymbol{\theta}}_n^*$ makes the bootstrap scores or estimating equations $\psi\left(\boldsymbol{\theta}, \widehat{\mathbb{P}}_n\right)$ approximately zero (in probability):

$$\left\|\psi\left(\hat{\boldsymbol{\theta}}_n^*, \widehat{\mathbb{P}}_n\right)\right\|_{\mathscr{H}} = o_{\mathbf{P}^*}\left(n^{-1/2}\right), \tag{4.2.8}$$

where $\mathbf{P} \equiv \mathbb{P}^\infty \times \widehat{\mathbb{P}}$. Moreover, define the product probability space

$$\left(\mathscr{X}^\infty, \mathscr{B}^\infty, \mathbb{P}^\infty\right) \times (\mathscr{Z}, \mathscr{E}, \widehat{\mathbb{P}}) = \left(\mathscr{X}^\infty \times \mathscr{Z}, \mathscr{B}^\infty \times \mathscr{E}, \mathbf{P}\right),$$

for the joint randomness involved. The notation superscript "*" or subscript "$_*$" denotes outer or an inner probability respectively: e.g., $\mathbb{P}^*$ indicates outer probability corresponding to $\mathbb{P}^\infty$. Here are the hypotheses which will be imposed on the bootstrap weights W_n:

(B.1) The vectors $\mathrm{W}_n = (\mathrm{W}_{n1}, \mathrm{W}_{n2}, \ldots, \mathrm{W}_{nn})^{\mathrm{T}}$ are exchangeable for all $n = 1,2,\ldots$, i.e., for any permutation $\pi = (\pi_1,\ldots,\pi_n)$ of $(1,2,\ldots,n)$, the joint distribution of

$$\pi(\mathrm{W}_n) = \left(\mathrm{W}_{n\pi_1}, \mathrm{W}_{n\pi_2}, \ldots, \mathrm{W}_{n\pi_n}\right)^{\mathrm{T}}$$

is the same as that of W_n.

(B.2) $W_{ni} \geq 0$ for all n, i and $\sum_{i=1}^{n} W_{ni} = n$ for all n.

(B.3) The following $L_{2,1}$ norm of W_{n1} is uniformly bounded:

$$R_n = \int_0^\infty \sqrt{\widehat{\mathbb{P}}\left(W_{n1} \geq u\right)} du \leq K < \infty$$

(B.4) $\lim_{\lambda\to\infty} \limsup_{n\to\infty} \sup_{t\geq\lambda} t^2 \widehat{\mathbb{P}}\{W_{n1} \geq t\} = 0$.

(B.5) $(1/n)\sum_{i=1}^{n} (W_{ni} - 1)^2 \to c^2 > 0$ in $\widehat{\mathbb{P}}$-probability.

The bootstrap empirical process $\hat{\mathbb{G}}_n$ is defined by

$$\hat{\mathbb{G}}_n = \sqrt{n}\left(\widehat{\mathbb{P}}_n - \mathbb{P}_n\right). \tag{4.2.9}$$

Discussion about these conditions and weights is given in Section 2.2.1.

4.2.2 Main results

In this subsection we introduce the conditions needed to obtain the results of Zhan [2002] for the weak convergence of $\hat{\boldsymbol{\theta}}_n$, after we give the key assumption for our main theorems which treat the weak convergence in probability of the bootstrapped version $\hat{\boldsymbol{\theta}}_n^*$.

(C.1) For all $\boldsymbol{\theta} \in \Theta, \psi(\boldsymbol{\theta}, \mathbb{P}_{\boldsymbol{\theta}}) = \mathbb{P}_{\boldsymbol{\theta}} B(\boldsymbol{\theta}) \equiv 0$ in $\ell^\infty(\mathscr{H})$.

(C.2) As $n \to \infty$, for any decreasing $\delta_n \downarrow 0$, the stochastic equicontinuity condition

$$\sup\left\{\|\mathbb{G}_n(B(\boldsymbol{\theta}) - B(\boldsymbol{\theta}_0))\|_{\mathscr{H}} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta_n\right\} = o_{\mathbb{P}^*}(1)$$

holds at the point $\boldsymbol{\theta}_0$.

(C.3) At the point $\boldsymbol{\theta}_0, \mathbb{G}_n B(\boldsymbol{\theta}_0) \rightsquigarrow \mathbb{Z}_0$ in $\ell^\infty(\mathscr{H})$, where $\rightsquigarrow$ indicates weak convergence in $\ell^\infty(\mathscr{H})$ to a tight Borel measurable random element $\mathbb{Z}_0$.

(C.4) For a fixed $\boldsymbol{\vartheta} \in \Theta$, the operator $\mathbb{P}_{\boldsymbol{\theta}} B(\boldsymbol{\vartheta})$ as a function of $\boldsymbol{\theta}$ is Fréchet differentiable with respect to the norm $\|\cdot\|$ at $\boldsymbol{\vartheta}$. Furthermore, the function $\boldsymbol{\theta} \mapsto \psi(\boldsymbol{\theta}, \mathbb{P})$ from Θ to $\ell^\infty(\mathscr{H})$ is Fréchet differentiable with respect to the norm $\|\cdot\|$. The operator $\dot{\psi}(\boldsymbol{\theta})$ is continuous as a function of $\boldsymbol{\theta}$ in the sense of (4.2.6).

(C.5) For every fixed $\boldsymbol{\theta} \in \Theta$, the operator $\dot{\psi}(\boldsymbol{\theta})$ from $(\overline{\text{lin}(\Theta)}, \|\cdot\|)$ to $(\ell^\infty(\mathscr{H}), \|\cdot\|_{\mathscr{H}})$ has a bounded inverse $\dot{\psi}^{-1}(\boldsymbol{\theta})$ on a fixed subspace $\overline{\mathscr{R}(\dot{\psi})} \subset \ell^\infty(\mathscr{H})$. Furthermore $\dot{\psi}^{-1}(\boldsymbol{\theta})$ as an operator sequence converges to $\dot{\psi}^{-1}(\boldsymbol{\theta}_0)$ as $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \to 0$:

$$\left\|\dot{\psi}^{-1}(\boldsymbol{\theta})(f) - \dot{\psi}^{-1}(\boldsymbol{\theta}_0)(f)\right\| \longrightarrow 0$$

for all $f \in \overline{\mathscr{R}(\dot{\psi})}$.

Let us define the following notation;

$$\mathscr{D} \equiv \mathscr{D}(\mathrm{R}) = \{\mathrm{B}(\boldsymbol{\theta})(h) : h \in \mathscr{H}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \mathrm{R}\}. \tag{4.2.10}$$

The envelope function of the class $\mathscr{D}(\mathrm{R})$ is:

$$\mathrm{D}(\mathrm{R}) \equiv \sup\{|\mathrm{B}(\boldsymbol{\theta})(h)(x)| : h \in \mathscr{H}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \mathrm{R}\}. \tag{4.2.11}$$

For a sequence of positive numbers δ_n as in condition **(C.2)**, let

$$\begin{aligned}\mathscr{D}_n &\equiv \{\mathrm{B}(\boldsymbol{\theta})(h) - \mathrm{B}(\boldsymbol{\theta}_0)(h) : h \in \mathscr{H}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta_n\} \\ &\equiv \{\mathrm{B}_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)(h) : h \in \mathscr{H}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta_n\},\end{aligned} \tag{4.2.12}$$

with envelope function defined as:

$$\mathrm{D}_n(x) \equiv \sup\{|\mathrm{B}_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)(h)(x)| : h \in \mathscr{H}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta_n\}, \tag{4.2.13}$$

with this notation, condition **(C.2)** can be rewritten as

$$\|\mathbb{G}_n\|_{\mathscr{D}_n} = o_{\mathbb{P}^*}(1).$$

In the following Lemma we recall the identity which is the key idea for solving the problem introduced in Section 4.2, which is a direct result of Lemma 4.5.1.1 without the nuisance parameter.

Lemma 4.2.2.1 (Zhan [2002]) *Assume **(C.1)**. For any $\boldsymbol{\vartheta} \in \Theta$, suppose that $\psi(\boldsymbol{\theta}, \mathbb{P})$ is Fréchet differentiable with respect to the norm $\|\cdot\|$ in a neighborhood of $\boldsymbol{\vartheta}$, and the operator $\dot{\psi}(\boldsymbol{\theta})$ is continuous as a function of $\boldsymbol{\theta}$ at $\boldsymbol{\vartheta}$ as in (4.2.6). If $\mathbb{P}_{\boldsymbol{\theta}}\mathrm{B}(\boldsymbol{\vartheta})$ is Fréchet differentiable with respect to the norm $\|\cdot\|$ at $\boldsymbol{\vartheta} \in \Theta$, then the operator $\psi(\boldsymbol{\theta}, \mathbb{P}_{\boldsymbol{\theta}})$ as a function of $\boldsymbol{\theta}$ is Fréchet differentiable with respect to the norm $\|\cdot\|$ at $\boldsymbol{\vartheta} \in \Theta$ and the following identity holds for all $a \in \mathrm{lin}(\Theta)$;*

$$\dot{\psi}(\boldsymbol{\vartheta})(a) + \dot{\mathbb{P}}_{\boldsymbol{\vartheta}}(a)\mathrm{B}(\boldsymbol{\vartheta}) = 0 \quad \textit{in} \quad \ell^\infty(\mathscr{H}). \tag{4.2.14}$$

Theorem 4.2.2.2 *[Zhan [2002]] Let $\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\| \to_{\mathbb{P}^*} 0$ be a sequence of consistent Z-estimators. Assume **(C.1)** through **(C.5)**. Then*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \rightsquigarrow -\dot{\psi}^{-1}(\boldsymbol{\theta}_0)(\mathbb{Z}_0) \quad \textit{in} \quad (\overline{\mathrm{lin}(\Theta)}, \|\cdot\|).$$

In this theorem the weak convergence is proved by the author for models where $\mathbb{P}_{\boldsymbol{\theta}}$ is not linearly parametrized, which implies that the linearity identity given in (4.2.14) is not perfect which implies that

$$\begin{aligned}\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right) &= \dot{\mathbb{P}}_{\hat{\boldsymbol{\theta}}_n}\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right)\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n\right) \\ &= -\mathbb{G}_n\mathrm{B}(\boldsymbol{\theta}_0) + o_{\mathbb{P}^*}\left(\sqrt{n}\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|\right) + o_{\mathbb{P}^*}(1).\end{aligned}$$

In this case, we must show that $\hat{\boldsymbol{\theta}}_n$ is $\sqrt{n}-$consistence, then by **(C.5)** $\dot{\psi}^{-1}(\boldsymbol{\theta})$ converges to $\dot{\psi}^{-1}(\boldsymbol{\theta}_0)$ as $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \rightarrowtail 0$ which implies the operator $\dot{\psi}^{-1}\left(\hat{\boldsymbol{\theta}}_n\right)$ is uniformly bounded in $\mathbb{P}^*$-probability by the Banach-Steinhaus theorem, by consequence the desired result hold by applying the continuous mapping theorem.
Now we will focus in models where $\mathbb{P}_{\boldsymbol{\theta}}$ is linearly convex, this means that; if $\boldsymbol{\theta} = \lambda_1\boldsymbol{\theta}_1 + \lambda_2\boldsymbol{\theta}_2 \in \operatorname{lin}(\Theta)$ implies $\mathbb{P}_{\boldsymbol{\theta}} = \lambda_1\mathbb{P}_{\boldsymbol{\theta}_1} + \lambda_2\mathbb{P}_{\boldsymbol{\theta}_2} \in \mathscr{P}$ for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$ and any real numbers λ_1 and λ_2 such that $\lambda_1 \geq 0, \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 = 1$. Convex linearity is referred to as bounded with respect to a norm $\|\cdot\|$ on $\operatorname{lin}(\Theta)$ if:

(C.6) For any $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ in Θ, and any real numbers $\lambda_1, \ldots, \lambda_k, k \geq 1$, there is a constant $\mathfrak{C} < \infty$ such that

$$\left\| \sum_{i=1}^{k} \lambda_i \mathbb{P}_{\boldsymbol{\theta}_i} \mathrm{B}(\boldsymbol{\vartheta}) \right\|_{\mathscr{H}} \leq \mathfrak{C} \left\| \sum_{i=1}^{k} \lambda_i \boldsymbol{\theta}_i \right\|$$

holds for every fixed $\boldsymbol{\vartheta} \in \Theta$, where we recall $\mathrm{B}(\boldsymbol{\vartheta})$ is the score operator mapping from $\mathscr{H}$ to $\mathscr{F}$.

Lemma 4.2.2.3 *If the parametrization $\boldsymbol{\theta} \rightarrowtail \mathbb{P}_{\boldsymbol{\theta}}$ is boundedly convex linear, then the mapping $\mathbb{P}_{\boldsymbol{\theta}}\mathrm{B}(\boldsymbol{\vartheta})$ is Fréchet differentiable with respect to the norm $\|\cdot\|$ at all $\boldsymbol{\vartheta} \in \Theta$ and the derivative operator $\dot{\mathbb{P}}_{\boldsymbol{\vartheta}}(\cdot)\mathrm{B}(\boldsymbol{\vartheta})$ is given by $\dot{\mathbb{P}}_{\boldsymbol{\vartheta}}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\mathrm{B}(\boldsymbol{\vartheta}) = \mathbb{P}_{\boldsymbol{\theta}_1}\mathrm{B}(\boldsymbol{\vartheta}) - \mathbb{P}_{\boldsymbol{\theta}_2}\mathrm{B}(\boldsymbol{\vartheta})$ for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and $\boldsymbol{\vartheta}$ in Θ.*

This lemma follows directly as a special case from Lemma 4.5.4.1 without the nuisance parameter. One can remark by the identity developed in (4.2.14) and the Lemma 4.2.2.3 we get a perfect linearization, i.e., we have for all $a = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \in \operatorname{lin}(\Theta)$ and $\boldsymbol{\vartheta} \in \Theta$ that;

$$\begin{aligned} \dot{\psi}(\boldsymbol{\vartheta})(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) &= -\dot{\mathbb{P}}_{\boldsymbol{\vartheta}}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\mathrm{B}(\boldsymbol{\vartheta}) \\ &= \mathbb{P}_{\boldsymbol{\theta}_2}\mathrm{B}(\boldsymbol{\vartheta}) - \mathbb{P}_{\boldsymbol{\theta}_1}\mathrm{B}(\boldsymbol{\vartheta}). \end{aligned} \tag{4.2.15}$$

For such model Zhan [2002] use the following assumptions for his main theorem:

(C.4′) The function $\psi(\boldsymbol{\theta}, \mathbb{P})$ as a map from Θ to $\ell^\infty(\mathscr{H})$ is Fréchet differentiable with respect to the norm $\|\cdot\|$. The operator $\dot{\psi}(\boldsymbol{\theta})$ is continuous as a function of $\boldsymbol{\theta}$ in the sense of (4.2.6).

(C.5′) For every fixed $\boldsymbol{\theta} \in \Theta$, the operator $\dot{\psi}(\boldsymbol{\theta})$ from $(\overline{\operatorname{lin}(\Theta)}, \|\cdot\|_{\mathrm{K}})$ to $(\ell^\infty(\mathscr{H}), \|\cdot\|_{\mathscr{H}})$ has a bounded inverse $\dot{\psi}^{-1}(\boldsymbol{\theta})$ on a fixed subspace $\overline{\mathscr{R}(\dot{\psi})} \subset \ell^\infty(\mathscr{H})$. Furthermore $\dot{\psi}^{-1}(\boldsymbol{\theta})$ as an operator sequence converges to $\dot{\psi}^{-1}(\boldsymbol{\theta}_0)$ as $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \rightarrow 0$:

$$\left\| \dot{\psi}^{-1}(\boldsymbol{\theta})(f) - \dot{\psi}^{-1}(\boldsymbol{\theta}_0)(f) \right\|_{\mathrm{K}} \longrightarrow 0$$

for all $f \in \overline{\mathscr{R}(\dot{\psi})}$.

In condition **(C.4′)** we have not to assume that the function $\boldsymbol{\theta} \rightarrowtail \mathbb{P}_{\boldsymbol{\theta}}$ is Fréchet differentiable as in **(C.4)**, because it holds by the definition of the convex linearity of the model, while **(C.5′)** differs from **(C.5)** by the use of the norm $\|\cdot\|_{\mathrm{K}}$, where the derivative operator $\dot{\psi}(\boldsymbol{\theta})$ is invertible with respect to it instead the norm $\|\cdot\|$ and this is the main ingredient used for solving the main problem as given in the following theorem.

Theorem 4.2.2.4 *[Zhan [2002]] For a model with bounded convex linearity specified in **(C.6)**, assume **(C.1)** through **(C.3)**, **(C.4′)** and **(C.5′)**. Then a sequence of consistent Z-estimators $\hat{\boldsymbol{\theta}}_n$ such that $\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\| \to_{\mathbb{P}^*} 0$ is actually asymptotically normal:*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \rightsquigarrow -\dot{\psi}^{-1}\left(\boldsymbol{\theta}_0\right)\left(\mathbb{Z}_0\right) \quad in \quad \left(\overline{\mathrm{lin}(\Theta)}, \|\cdot\|_{\mathrm{K}}\right).$$

In models where the parametrization $\boldsymbol{\theta} \mapsto \mathbb{P}_{\boldsymbol{\theta}}$ is boundedly convex linear as in **(C.6)** and by choosing $\boldsymbol{\theta}_1 = \boldsymbol{\vartheta} = \hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_0$ in (4.2.15), we get by **(C.1)**;

$$\begin{aligned}\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right) &= -\sqrt{n}\mathbb{P}_{\boldsymbol{\theta}_0}\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n\right) \\ &= -\mathbb{G}_n\mathrm{B}\left(\boldsymbol{\theta}_0\right) + o_{\mathbb{P}^*}(1).\end{aligned}$$

This identity is perfect which give us a great flexibility for choosing any norm as long as that the operator derivative $\dot{\psi}(\boldsymbol{\theta})$ is boundedly invertible with respect to it, then the rate of convergence and the weak convergence obtained in one step.
In the rest of this section we will give the condition needed for the asymptotic limit of the bootstrapped version of Z-estimator $\hat{\boldsymbol{\theta}}_n^*$ satisfies (4.2.8) and our main Theorems. The following condition on the envelope function $\mathrm{D}_n(x)$ is used for validity of the bootstrap consistency as given in Wellner and Zhan [1996]:

(CB) For each sequence $\delta_n \to 0$ the envelope functions D_n of the classes $\mathscr{D}_n$ satisfy

$$\lim_{\lambda\to\infty}\limsup_{n\to\infty}\sup_{t\geq\lambda} t^2\mathbb{P}^*\left(\mathrm{D}_n\left(\mathrm{X}_1\right) > t\right) = 0.$$

For our bootstrap results, we further assume that the collection $\mathscr{D}$ or $\mathscr{D}(\mathrm{R})$ possesses enough measurability for randomization with i.i.d. multipliers to be possible and the usual Fubini's theorem can be used freely; such a set of conditions is $\mathscr{D} \in \mathrm{NLDM}(\mathbb{P})$ (Nearly Linearly Deviation Measurable), and $\mathscr{D}^2, \mathscr{D}'^2 \in \mathrm{NLSM}\ (\mathbb{P})$ (Nearly Linearly Supremum Measurable) in the terminology of Giné and Zinn [1990]. Here $\mathscr{D}^2$ and $\mathscr{D}'^2$ denote the classes of squared functions and squared differences of functions from $\mathscr{D}$, respectively. When all of these conditions hold, we write $\mathscr{D} \in \mathrm{M}(\mathbb{P})$. It is known that $\mathscr{D} \in \mathrm{M}(\mathbb{P})$ if $\mathscr{D}$ is countable, or if the empirical processes $\mathbb{G}_n$ are stochastically separable, or if $\mathscr{D}$ is image admissible Suslin (see Giné and Zinn [1990], page 853 and 854).

Theorem 4.2.2.5 *Let $\hat{\boldsymbol{\theta}}_n$ be a sequence of consistent asymptotic Z-estimators and $\hat{\boldsymbol{\theta}}_n^*$ be a sequence of consistent bootstrap asymptotic Z-estimators with exchangeable bootstrap weights satisfying **(B.1)** through **(B.5)**: $\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\| \to_{\mathbb{P}^*} 0$ and $\left\|\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0\right\| \to_{\mathbf{P}^*} 0$ in $\mathbb{P}^*$-probability. If assumptions **(C.1)** through **(C.5)** and **(CB)** hold, then*

(i)

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \rightsquigarrow -\dot{\psi}^{-1}\left(\boldsymbol{\theta}_0\right)\left(\mathbb{Z}_0\right) \quad in \quad \left(\overline{\mathrm{lin}(\Theta)}, \|\cdot\|\right).$$

(ii)

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n\right) &= -\dot{\psi}^{-1}\left(\boldsymbol{\theta}_0\right)\left(\hat{\mathbb{G}}_n\mathrm{B}\left(\boldsymbol{\theta}_0\right)\right) + o_{\hat{\mathbb{P}}}(1) \\ &\rightsquigarrow -\dot{\psi}^{-1}\left(\boldsymbol{\theta}_0\right)\left(c\cdot\hat{\mathbb{Z}}_0\right) \quad in \quad \left(\overline{\mathrm{lin}(\Theta)}, \|\cdot\|\right),\end{aligned}$$

in $\mathbb{P}^*$ -probability, where $\hat{\mathbb{Z}}_0 \overset{d}{=} \mathbb{Z}_0$, and c is the constant given in ***(B.5)****.*

To prove the last theorem, we need the following two lemmas, in the first we establish the bootstrap equicontinuity condition which implies by the stochastic equicontinuity condition **(C.2)**, then by making use of this result, it holds the $\sqrt{n}-$consistency of $\hat{\boldsymbol{\theta}}_n^*$ in the second lemma.

Lemma 4.2.2.6 *Under* ***(C.2)***, ***(CB)***, ***(B1)-(B5)*** *and assume that $\mathscr{D} \in \mathrm{M}(\mathbb{P})$ for some $\mathrm{R} > 0$. Then for any positive sequence $\delta_n \to 0$, it holds that*

$$\Delta_n \equiv \sup\left\{\left\|\hat{\mathbb{G}}_n\left(\mathrm{B}(\boldsymbol{\theta}) - \mathrm{B}\left(\boldsymbol{\theta}_0\right)\right)\right\|_{\mathscr{H}} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta_n\right\} = \|\hat{\mathbb{G}}_n\|_{\mathscr{D}_n} = o_{\mathbf{P}^*}(1)$$

That is $\Delta_n = o_{\hat{\mathbb{P}}}(1)$ in $\mathbb{P}^*-$probability.

This result holds directly by applying the multiplier inequality as in Lemma 4.1 of Wellner and Zhan [1996] to the empirical process indexed by $\mathscr{D}_n$.
In the following, we use $\mathrm{O}_{\mathbb{P}^*}(1)$ to denote maps whose norm is of order $\mathrm{O}_{\mathbb{P}^*}(1)$. The same rule will be applied to terms of order $o_{\mathbb{P}^*}(1)$, and $o_{\mathbf{P}^*}(1)$ respectively. In the proof of Theorem 4.2.2.5, the stochastic equicontinuity assumption **(C.2)** and the Fréchet differentiability assumption **(C.4)** and the convergence of the operator $\dot{\psi}^{-1}(\boldsymbol{\theta})$ to $\dot{\psi}^{-1}(\boldsymbol{\theta}_0)$ as $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \to 0$. The condition **(C.5)** is used to deduce $\sqrt{n}-$consistency of $\widehat{\boldsymbol{\theta}}_n$ from consistency. Similarly, bootstrap equicontinuity given in the preceding lemma together with the differentiability assumption **(C.4)** and the convergence of the operator $\dot{\psi}(\boldsymbol{\theta})$ **(C.5)** allows us to prove $\sqrt{n}-$consistency of the bootstrap starting from consistency of the bootstrap estimator.

Lemma 4.2.2.7 *Assume conditions of Theorem 4.2.2.5 hold and $\hat{\boldsymbol{\theta}}_n^*$ is consistent: $\|\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0\| = o_{\hat{\mathbb{P}}}(1)$ in $\mathbb{P}^*-$probability, then in $\mathbb{P}^*-$probability, we have*

$$\sqrt{n}\|\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0\| = o_{\hat{\mathbb{P}}}(1).$$

One can remark that for proving the asymptotic normality of the sequence $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n\right)$ in the previous theorem, we began by proving the $\sqrt{n}-$consistency of $\hat{\boldsymbol{\theta}}_n^*$ with respect to the norm $\|\cdot\|$, which is both the operator $\psi(\boldsymbol{\theta}, \mathbb{P})$ is Fréchet differentiable and the operator $\dot{\psi}(\boldsymbol{\theta})$ is invertible with respect to it. As we noticed in the previous section that there are some cases where these cannot hold, as given in Zhan [2002] and examples in Section 4.3. Now we are in position to give a solution for this problem where the weak convergence and the $\sqrt{n}-$consistency of $\hat{\boldsymbol{\theta}}_n^*$ are proved in one step and this is hold with respect to any chosen norm which is the operator $\dot{\psi}(\boldsymbol{\theta})$ is invertible with respect to it. The following theorem is one of our main results of this work.

Theorem 4.2.2.8 *Let $\hat{\boldsymbol{\theta}}_n$ be a sequence of consistent asymptotic Z-estimators and $\hat{\boldsymbol{\theta}}_n^*$ be a sequence of consistent bootstrap asymptotic Z-estimators with exchangeable bootstrap weights satisfying* ***(B.1)*** *through* ***(B.5)****: $\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\| \to_{\mathbb{P}^*} 0$ and $\left\|\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0\right\| \to_{\mathbf{P}^*} 0$ in $\mathbb{P}^*$ -probability. For a model with bounded convex linearity specified in* ***(C.6)****, assume* ***(C.1)*** *through* ***(C.3)****,* ***(C.4′)*** *and* ***(C.5′)*** *then*

(i′)

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \rightsquigarrow -\dot{\psi}^{-1}\left(\boldsymbol{\theta}_0\right)\left(\mathbb{Z}_0\right) \quad \textit{in} \quad \left(\overline{\mathrm{lin}(\Theta)}, \|\cdot\|_K\right).$$

(ii′)

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n\right) &= -\dot{\psi}^{-1}\left(\boldsymbol{\theta}_0\right)\left(\hat{\mathbb{G}}_n \mathrm{B}\left(\boldsymbol{\theta}_0\right)\right) + o_{\hat{\mathbb{P}}}(1) \\ &\rightsquigarrow -\dot{\psi}^{-1}\left(\boldsymbol{\theta}_0\right)\left(c \cdot \hat{\mathbb{Z}}_0\right) \quad \textit{in} \quad \left(\overline{\mathrm{lin}(\Theta)}, \|\cdot\|_K\right),\end{aligned}$$

in $\mathbb{P}^*$ *-probability, where* $\hat{\mathbb{Z}}_0 \overset{d}{=} \mathbb{Z}_0$*, and* c *is the constant given in* ***(B.5)****.*

4.3 Examples

In this section we give three examples to illustrate the usefulness of Theorem 4.2.2.5 and Theorem 4.2.2.8. The first two example concern models without convex linearly parametrization, the first one is the classical model of random right censorship and the second is a simplified frailty model, these models have a nice analytical properties such that all derivatives and their inverses are quite explicit and easily calculable, so the asymptotic normality of the bootstrapped version of the MLE can established by traditional arguments and the same conclusion can also be obtained by Theorem 4.2.2.5. These examples are given in Zhan [2002], for which our bootstrap are validated.

4.3.1 Random right censoring

Here we have $X \sim F$ and $Y \sim G$ on $\mathbb{R}^+$, and we observe (a random sample of data) with the distribution of $(Z, \Delta) \equiv \left(X \wedge Y, \mathbb{1}_{[X \leq Y]}\right)$. Following Gill [1989], we parametrize the model in terms of the cumulative hazard function $\Lambda(\cdot)$ corresponding to $F(\cdot)$ given by

$$\Lambda(x) = \int_0^x \frac{1}{1 - F(u-)} dF(u).$$

We are interested in estimating the unknown cumulative hazard function $\Lambda_0(\cdot)$ on $[0, \infty)$ on the basis of n of i.i.d. observations (Z_i, Δ_i) from

$$P_\Lambda^{(0)}(z) = \mathbb{P}_\Lambda\{Z \leq z, \Delta = 0\} = \int_0^z \overline{F}(y) dG(y), \tag{4.3.1}$$

$$P_\Lambda^{(1)}(z) = \mathbb{P}_\Lambda\{Z \leq z, \Delta = 1\} = \int_0^z \overline{G}(x-) dF(x), \tag{4.3.2}$$

where $\overline{F}(\cdot) = 1 - F(\cdot)$ and $\overline{G}(\cdot) = 1 - G(\cdot)$. Let Θ be the set of all cumulative hazard functions on the positive real line $[0, \infty[$, equipped with the uniform norm $\|\cdot\|$. Then assuming that $\Lambda(\cdot)$ and $G(\cdot)$ are absolutely continuous with densities $\lambda(\cdot)$ and $g(\cdot)$ respectively, the joint density density of (Z, Δ) on $\mathbb{R}^+ \times \{0, 1\}$ is given by

$$\mathbb{P}_{\Lambda, G}(z, \Delta) = \lambda(z)^\Delta \exp(-\Lambda(z)) g(z)^{1-\Delta} (1 - G(z))^\Delta, \quad \text{for} \quad z \in \mathbb{R}^+, \Delta \in \{0, 1\}.$$

Following Gill [1989], consider one-dimensional parametric submodels of the full nonparametric model by defining $\{\Lambda_{\boldsymbol{\eta}} : |\boldsymbol{\eta}| \leq \boldsymbol{\eta}_0\}$ by

$$\frac{d\Lambda_{\boldsymbol{\eta}}}{d\Lambda}(x) = 1 + \boldsymbol{\eta} h(x),$$

for $h(\cdot)$ bounded. The score for this submodel (for $n = 1$) is:

$$\frac{\partial}{\partial \boldsymbol{\eta}} \log \mathbb{P}_{\Lambda_{\boldsymbol{\eta}}, G}(z, \Delta)\Big|_{\boldsymbol{\eta}=0} = \Delta h(z) - \int_0^z h d\Lambda \equiv \mathrm{B}(\Lambda) h(z, \Delta).$$

In this model, we choose $\mathscr{H} = \{h_t = \mathbb{1}_{[0,t]} : 0 \leq t \leq \tau\}$ where $\tau < \tau_H \equiv \inf\{x : H(x) = 1\}$ and $1 - H(x) = (1 - F_0(x))(1 - G_0(x)) = \mathbb{P}_0(Z > x)$. ∞. The parameter space Θ can thus be identified with all cumulative hazard functions restricted to the interval $[0, \tau]$. Then we have

$$\begin{aligned} \mathrm{B}(\Lambda)(h_t)(z, \Delta) &= \Delta \mathbb{1}_{[0,t]}(z) - \int_0^z \mathbb{1}_{[0,t]}(u) d\Lambda(u) \\ &= \Delta \mathbb{1}_{[0,t]}(z) - \Lambda(z \wedge t), \end{aligned}$$

$$\psi(\Lambda, \mathbb{P}_n)(h_t) = \mathbb{P}_n \mathrm{B}(\Lambda)(h_t) = \frac{1}{n} \sum_{i=1}^n \left\{ \Delta_i \mathbb{1}_{[0,t]}(Z_i) - \int_0^t \mathbb{1}_{[Z_i \geq u]} d\Lambda(u) \right\}, \tag{4.3.3}$$

and

$$\begin{aligned} \psi(\Lambda, \mathbb{P}_{\Lambda_0}) = \mathbb{P}_{\Lambda_0} \mathrm{B}(\Lambda)(h_t) &= \int_0^t \overline{G}(u-) dF_0(u) - \int_0^t \overline{H}(u-) d\Lambda(u) \\ &= \int_0^t \overline{H}(u-) d\Lambda_0(u) - \int_0^t \overline{H}(u-) d\Lambda(u). \end{aligned} \tag{4.3.4}$$

From (4.3.3) it follows that the score equations for the maximum likelihood estimator $\widehat{\Lambda}_n(\cdot)$ of $\Lambda(\cdot)$ are

$$\begin{aligned} 0 = \psi(\widehat{\Lambda}_n, \mathbb{P}_n)(h_t) &= \mathbb{P}_n \mathrm{B}(\widehat{\Lambda}_n)(h_t) \\ &= \frac{1}{n} \sum_{i=1}^n \Delta_i \mathbb{1}_{[0,t]}(Z_i) - \frac{1}{n} \sum_{i=1}^n \int_0^t \mathbb{1}_{[Z_i \geq u]} d\widehat{\Lambda}_n(u) \\ &\equiv H_n^{(1)}(t) - \int_0^t \overline{H}_n(u-) d\widehat{\Lambda}_n(u) \quad \text{for} \quad 0 \leq t \leq \tau. \end{aligned}$$

In this case the score equations have an explicit solution: from (4.3.3) it follows that

$$\widehat{\Lambda}_n(t) = \int_0^t \frac{1}{\overline{H}_n(u-)} dH_n^{(1)}(u), \quad 0 \leq t \leq \tau, \tag{4.3.5}$$

the well-known Nelson-Aalen estimator of $\Lambda(\cdot)$. From here the maximum likelihood estimator of $F(\cdot)$ is given by the product integral:

$$1 - \widehat{F}_n(t) = \prod_{s \leq t} \left(1 - \Delta \widehat{\Lambda}_n(s)\right), \quad 0 \leq t \leq \tau$$

this is the Kaplan and Meier [1958] "product-limit" estimator of $F(\cdot)$. Note that we choose this example to well clarify for the reader the difference between using the traditional arguments as in Wellner and Zhan [1996] and the result in Theorem 4.2.2.5, for our Theorem we need to show that the inverse operator of the derivative is uniformly continuous, let us clarify more. Condition **(C.1)** holds directly by substitute $\mathbb{P}_{\Lambda_0} \equiv \mathbb{P}$ with $\mathbb{P}_\Lambda$ for any $\Lambda \in \Theta$ in (4.3.4), **(C.2)** and **(C.3)** here are the same as **(A.2)** and **(A.3)** in Wellner and Zhan [1996], so we don't check them here, for **(C.4)** we observe that; the operator

$$\Lambda \mapsto \mathbb{P}_\Lambda B(\mathfrak{d}) = \int_0^t \overline{H}(u-) d\,(\Lambda - \mathfrak{d})\,(u),$$

is linear in Λ and the second term is independent of Λ which implies its Fréchet differentiability at all $\Lambda \in \Theta$ and its operator derivative is given by for any $\mathfrak{d} \in \Theta$;

$$\dot{\mathbb{P}}_\Lambda(\Lambda - \Lambda_0)B(\mathfrak{d}) = \int_0^t \overline{H}(u-) d\,(\Lambda - \Lambda_0)\,(u). \tag{4.3.6}$$

With the same arguments for $\psi(\Lambda, \mathbb{P})$ it follows by using the integration by parts for any $\Lambda \in \Theta$;

$$\begin{aligned}\dot{\psi}(\Lambda)(\Lambda - \Lambda_0)(h_t) &= -\int_0^t \overline{H}(u-) d\,(\Lambda - \Lambda_0)\,(u) \\ &= -\overline{H}(t)\,(\Lambda - \Lambda_0)\,(t) + \int_0^t (\Lambda - \Lambda_0)\,(u) d\overline{H}(u).\end{aligned} \tag{4.3.7}$$

The operators $\dot{\psi}(\Lambda)(\cdot)$ and $\dot{\mathbb{P}}_\Lambda(\cdot)B(\mathfrak{d})$ are bounded linear from $(\overline{\text{lin}(\Theta)}, \|\cdot\|)$ to $(\ell^\infty(\mathscr{H}), \|\cdot\|_{\mathscr{H}})$, we observe that by combining (4.3.6) and (4.3.7), the identity in (4.2.14) holds. From (4.3.7) we remark that the operator $\dot{\psi}(\Lambda)$ is uniformly continuous as function of Λ, which can be rewritten as; $\dot{\psi}(\Lambda)(J)(t) = K(t)$, for a given $K \in \overline{\mathscr{R}(\dot{\psi})} \subset \ell^\infty(\mathscr{H})$. It follows that;

$$\dot{\psi}^{-1}(\Lambda)(K)(t) = -\int_0^t \frac{1}{\overline{H}(u-)} dK(u) = J(t).$$

One can remark that for all $\Lambda \in \Theta$ we have; $\dot{\psi}(\Lambda)(\cdot) = \dot{\psi}(\Lambda_0)(\cdot)$ as an operator independent of Λ, which implies the condition **(C.5)**. From **(C.1)** to **(C.5)** it follows from Theorem 4.2.2.5 that the Maximum Likelihood Estimator $\hat{\Lambda}_n$, i.e., the Nelson-Aalen estimator, satisfies

$$\sqrt{n}\left(\hat{\Lambda}_n - \Lambda_0\right) \rightsquigarrow -\dot{\psi}^{-1}(\Lambda_0)\,(\mathbb{Z}_0) \quad \text{in} \quad (\overline{\text{lin}(\Theta)}, \|\cdot\|).$$

To prove that the bootstrap Theorem 4.2.2.5, it suffices to verify the envelope integrability assumption **(CB)**, which is the same as assumption **(A.5)** in Wellner and Zhan [1996], then it follows for any exchangeable weighted bootstrap with weights satisfying **(B.1)**-**(B.5)** the validity of the bootstrapped version Λ_n^* in probability as in Theorem 4.2.2.8;

$$\sqrt{n}\left(\hat{\Lambda}_n^* - \hat{\Lambda}_n\right) \rightsquigarrow -\dot{\psi}_0^{-1}(\Lambda_0)\left(c \cdot \hat{\mathbb{Z}}_0\right) \quad \text{in} \quad (\overline{\text{lin}(\Theta)}, \|\cdot\|),$$

in $\mathbb{P}^*$-probability, where $\hat{\mathbb{Z}}_0 \overset{d}{=} \mathbb{Z}_0$. The Efron's nonparametric bootstrap of this estimator is studied by Akritas [1986] and Gill [1989], and for the Bayesian bootstrap see Lo [1993]. For further examples of such weights, we refer the reader to the monograph Præstgaard and Wellner [1993].

4.3.2 A simplified frailty model

Let $Z \sim \text{Gamma}(\nu_0, 1)$ be a known gamma frailty. Conditional on $Z = z$, we observe independent random variables (X, Y) with a common, absolutely continuous hazard function $z\Lambda_0$. Based on n i.i.d. observations (X_i, Y_i) with distribution

$$\mathbb{P}\{X > x, Y > y\} = 1/\left[1 + \Lambda_0(x) + \Lambda_0(y)\right]^{\nu_0},$$

we are interested in estimating Λ_0 on $[0, \tau]$, where $\tau < \infty$ is a real number such that $\Lambda_0(\tau) < \infty$. Let $\Theta \subset \ell^\infty(\mathscr{H}_p)$ be the parameter space, where $\mathscr{H}_p$ is a set of real functions $h(\cdot)$ defined on $[0, \infty)$ with bounded variation $\|h\|_v < p$ on $[0, \tau]$ and identical to zero on (τ, ∞). The set $\mathscr{H}_p$ is considered as a space equipped with the variation norm $\|\cdot\|_v$ defined by

$$\|h\|_v \equiv |h(0)| + \vee_0^\tau(h).$$

A bounded linear functional $\Lambda(h) \in \ell^\infty(\mathscr{H}_p)$ is given by

$$\Lambda(h) = \int_{[0,\infty)} h(x) d\Lambda(x),$$

with

$$\|\Lambda\|_{\mathscr{H}_p} = \sup_{h \in \mathscr{H}_p} \left| \int_{[0,\infty)} h(x) d\Lambda(x) \right| < \infty.$$

The parameter space Θ can thus be identified with all absolutely continuous integrated hazard functions Λ restricted to the interval $[0, \tau]$, such that $\Lambda(u) \equiv \Lambda_0(u)$ for $u > \tau$. We will not distinguish between a functional $\Lambda \in \Theta$ and a hazard function $\Lambda(u)$. The score operator $B(\Lambda)$ is obtained by differentiating the log-likelihood along a curve passing through $\Lambda \in \Theta$. It is a function of Λ mapping from $\mathscr{H}_p$ to a set $\mathscr{F}$ of $L_2(\mathbb{P})$ functions defined on the sample space:

$$B(\Lambda)(h)(x, y) = h(x) + h(y) - (\nu_0 + 2) \frac{\int_{[0,x]} h(u) d\Lambda(u) + \int_{[0,y]} h(u) d\Lambda(u)}{1 + \Lambda(x) + \Lambda(y)}.$$

This example was considered by Murphy [1995] in the context of counting process, generalized by Parner [1998] to the case with covariates, van der Vaart [1995] used it as an example to motivate the Central Limit Theorem for functional parameters and Wellner and Zhan [1996] for studying the consistency of the bootstrap in these models, where they used the traditional argument, we show that the same asymptotic results are obtainable from Theorem 4.2.2.5. Note that the most conditions were verified in Zhan [2002], hence we need only to verify the envelope integrability condition **(CB)**. The envelope function D_n defined in (4.2.13) in this case is bounded by

$$\begin{aligned} D_n(x) &= \sup\left\{ |B_n(\Lambda, \Lambda_0)(h)(x)| : h \in \mathscr{H}_p, \|\Lambda - \Lambda_0\| \le \delta_n \right\} \\ &\le 2(\nu_0 + 2) \sup\left\{ \left| \frac{\int_{[0,x]} h(u) d\Lambda(u) + \int_{[0,y]} h(u) d\Lambda(u)}{1 + \Lambda(x) + \Lambda(y)} \right| : h \in \mathscr{H}_p, \|\Lambda - \Lambda_0\| \le R \right\} \\ &\le 2p(\nu_0 + 2). \end{aligned}$$

Hence condition **(CB)** holds directly and then we get from Theorem 4.2.2.5;

$$\sqrt{n}\left(\hat{\Lambda}_n - \Lambda_0\right) \rightsquigarrow -\dot{\psi}^{-1}(\Lambda_0)\left(\mathbb{Z}_0\right) \quad \text{in} \quad (\overline{\text{lin}(\Theta)}, \|\cdot\|_{\mathscr{H}_p}),$$

and

$$\sqrt{n}\left(\hat{\Lambda}_n^* - \hat{\Lambda}_n\right) \rightsquigarrow -\dot{\psi}^{-1}(\Lambda_0)\left(c\cdot\hat{\mathbb{Z}}_0\right) \quad \text{in} \quad (\overline{\text{lin}(\Theta)}, \|\cdot\|_{\mathscr{H}_p}),$$

in $\mathbb{P}^*$-probability, where $\hat{\mathbb{Z}}_0 \stackrel{d}{=} \mathbb{Z}_0$.

4.3.3 The double censoring model

Let $X \sim F_0$ be a non-negative random variable. Let (Y,Z) be a pair of nonnegative random censoring times independent of the random variable X that satisfies $\mathbb{P}(Y \le Z) = 1$. We observe a pair (W,Δ) of random variables, defined by

$$(W,\Delta) = \begin{cases} (X,1) & \text{if} \quad Y < X \le Z; \\ (Z,2) & \text{if} \quad X > Z; \\ (Y,3) & \text{if} \quad X \le Y, \end{cases}$$

where $(W,\Delta) \sim \mathbb{P}$. We are interested in estimating the distribution function $F_0(\cdot)$ from i.i.d. pairs $(W_i,\Delta_i) \sim \mathbb{P}, i = 1,\ldots,n$. Let $K(t) = G_Y(t) - G_Z(t)$, where $G_Y(t) = \mathbb{P}(Y \le t)$ and $G_Z(t) = \mathbb{P}(Z \le t)$ are the marginal distribution function of Y and Z, respectively. It follows from the censoring mechanism that the distribution $\mathbb{P}_{F_0}$ is equivalent to the following three marginals for $\Delta = 1,2,3$,

$$\mathbb{P}_F^{(1)}(t) \equiv \mathbb{P}_F\{W \le t, \Delta = 1\} = \int_{[0,t]} K(u-)dF(u), \tag{4.3.8}$$

$$\mathbb{P}_F^{(2)}(t) \equiv \mathbb{P}_F\{W \le t, \Delta = 2\} = \int_{[0,t]} (1-F(u))dG_Z(u), \tag{4.3.9}$$

$$\mathbb{P}_F^{(3)}(t) \equiv \mathbb{P}_F\{W \le t, \Delta = 3\} = \int_{[0,t]} F(u)dG_Y(u). \tag{4.3.10}$$

The marginal distribution function for $W(\cdot)$ under the true $F_0(\cdot)$ is

$$H_{\mathbb{P}}(t) = \sum_{j=1}^{3} \mathbb{P}_{F_0}^{(j)}(t).$$

Let Θ be the set of all distribution functions defined on $[0,\infty[$. For parametric submodels of the full model of the form $\left\{\mathbb{P}_{F_{\eta,G}} : |\eta| \le \eta_0\right\}$ with $F_\eta(x)$ given by

$$\frac{dF_\eta}{dF} = 1 + \eta\left(h - \int_{[0,\infty)} h dF\right),$$

where $h(\cdot)$ is any given bounded measurable function on $\mathbb{R}^+$, it is straightforward to compute the score operator $B(F)(h)$:

$$\begin{aligned} B(F)(h)(w,\delta) = {} & \mathbb{1}_{[\delta=1]}\left(h(w) - \int_{[0,\infty[} h dF\right) - \mathbb{1}_{[\delta=2]} \frac{\int_{[0,W]}\left(h - \int_{[0,\infty)} h dF\right) dF}{(1-F(w))} \\ & + \mathbb{1}_{[\delta=3]} \frac{\int_{[0,w]}\left(h - \int_{[0,\infty[} h dF\right) dF}{F(w)}. \end{aligned}$$

When $h = h_t \in \mathscr{H} = \{h_t = 1_{[0,t]}(\cdot) : t \in [0,\infty)\}$, the last equation can be rewritten as

$$B(F)(h_t)(w,\delta) = \left(1_{[0,t]}(w) - F(t)\right) - \mathbb{1}_{[\delta=2, w\le t]}\frac{1-F(t)}{1-F(w)} + \mathbb{1}_{[\delta=3, w>t]}\frac{F(t)}{F(w)}. \tag{4.3.11}$$

Integrating with respect to $\mathbb{P}$ we get the operator ψ, for $t \in [0,\infty[$

$$\psi(F,\mathbb{P})(h_t) = H_{\mathbb{P}}(t) - F(t) - \int_{[0,t]}\frac{1-F(t)}{1-F(u)}d\mathbb{P}^{(2)}_{F_0}(u) + \int_{(t,\infty)}\frac{F(t)}{F(u)}d\mathbb{P}^{(3)}_{F_0}(u). \tag{4.3.12}$$

The set of all Z -estimators $\hat{F}_n$ in this model contains the set of all self-consistent estimators defined by $\psi\left(\hat{F}_n, \mathbb{P}_n\right)(h_t) \equiv 0$ for all $t \ge 0$. It is well known that $\hat{F}_n$ is consistent in the uniform norm, see Gu and Zhang [1993] and Wellner and Zhan [1996]. Before verifying the conditions, we will discuss some results and clarify the situation here, for that we define some notation which will be used for what follows. Let $\tau_0 = \sup\{t : F_0(t) = 0\}$ and $\tau_1 = \inf\{t : F_0(t) = 1\}$. Let $D_0[\tau_0,\tau_1]$ be the Banach space of all real-valued functions defined on $[\tau_0,\tau_1]$ which are right-continuous and have left-limits:

$$D_0[\tau_0,\tau_1] = \{a : F_0(t) = 0 \rightsquigarrow a(t) = 0, F_0(t-) = 1 \rightsquigarrow a(t-) = 0, F_0(t) = 1 \rightsquigarrow a(t) = 0\}.$$

Let $(D_K[\tau_0,\tau_1], \|\cdot\|_K)$ denote the completion of $D_0[\tau_0,\tau_1]$ under the K-norm $\|a\|_K = \|Ka\|$. Further restrict Θ to be all distribution functions on $[0,\infty)$ such that $F \in \Theta$ implies $F - F_0 \in D_0[\tau_0,\tau_1]$. Note that the operator given in (4.3.12) is a Fréchet derivative operator with respect to the uniform norm $\|\cdot\|$ with derivative given by

$$-\dot{\psi}(F)(a)(h_t) = (K + A)(a)(t), \tag{4.3.13}$$

where

$$(Ka)(h_t) = K(t)a(t)$$
$$A \equiv A(F, G_Y, G_Z)(a)(h_t) = \int_{[0,t]}\frac{1-F(t)}{1-F(u)}a(u)dG_Z(u) + \int_{(t,\infty)}\frac{F(t)}{F(u)}a(u)dG_Y(u).$$

We must assume that $\inf_{\tau_0\le t\le\tau_1} K(t) > 0$; to show the invertibility of the operator $\dot{\psi}(F)$ with respect to the uniform norm $\|\cdot\|$. This is done if we remark that the operator A is a compact operator following from the proof of Lemma 2 in Gu and Zhang [1993] and the range of $K(t)$ is not closed without the above condition, then it results that the operator $\dot{\psi}(F)$ is not invertible with respect to $\|\cdot\|$. Under the following conditions

(DC1)

$$K(t-) > 0 \text{ on } \{t : F_0(t) > 0 \text{ or } F_0(t-) < 1\}. \tag{4.3.14}$$

(DC2) For any $0 < \eta < 1$,

$$\int_{0<F_0(u)<1-\eta}\frac{dG_Z(u)}{G_Y(u) - G_Z(u)} + \int_{\eta<F_0(u)<1}\frac{dG_Y(u)}{G_Y(u) - G_Z(u)} < \infty. \tag{4.3.15}$$

Gu and Zhang [1993] showed the weak convergence of the self-consistent estimator in the entire support i.e.,

$$\sqrt{n}\left(\hat{F}_n - F_0\right) \rightsquigarrow -\dot{\psi}^{-1}(F_0)(\mathbb{Z}_0) \quad \text{in} \quad (D_K[\tau_0,\tau_1], \|\cdot\|_K),$$

where they are shown that the operator $\dot{\psi}(F)$ is indeed invertible with respect to the norm $\|\cdot\|_K$, without assuming that $\inf_{\tau_0 \le t \le \tau_1} K(t) > 0$. Leaving the case of self-consistent estimator to the general case of Z-estimators, intuitively the question is: can we apply the traditional arguments to get such results of Gu and Zhang [1993] on the $\hat{F}_n$ and the bootstrapped version F_n^* ? Unfortunately, this cannot hold because the operator $\psi(F,\mathbb{P})$ given in (4.3.12) is not Fréchet differentiable with respect to the norm $\|\cdot\|_K$, which allows to Wellner and Zhan [1996] to obtain the weak convergence in some restrict interval, i.e., they showed that for some c, d satisfy $0 \le \tau_0 < c < d < \tau_1$ that

$$\sqrt{n}\left(\hat{F}_n - F_0\right) \rightsquigarrow -\dot{\psi}^{-1}(F_0)(\mathbb{Z}_0) \quad \text{in } (D[c,d], \|\cdot\|),$$

and

$$\sqrt{n}\left(\hat{F}_n^* - \hat{F}_n\right) \rightsquigarrow -\dot{\psi}^{-1}(F_0)\left(c\cdot\hat{\mathbb{Z}}_0\right) \quad \text{in } (D[c,d], \|\cdot\|),$$

in $\mathbb{P}^*$-probability, where $\hat{\mathbb{Z}}_0 \overset{d}{=} \mathbb{Z}_0$, and c is the constant given in the definition **(B.5)** for the exchangeable weights. Now our theory takes place and let us obtain the weak convergence on the entire support as described below: for the verification of the conditions we need only to verify the assumption **(CB)**, because the rest of them are the same as verified by Zhan [2002]. For that we remark for some $R > 0$, the class of function

$$\{|B(F)(h_t)(w,\delta)| : t \in [0,\infty[, \|F - F_0\| \le R\}$$

has a constant envelope 1, as a consequence the class of functions $\mathscr{D}_n$ in (4.2.12) has an envelope D_n in (4.2.13) which is bounded by 2, hence **(CB)** holds.

Note that the operator $\dot{\psi}(F)(\cdot)$ in (4.3.13) can be regarded as a mapping from $D_K[\tau_0,\tau_1]$ into $D_0[\tau_0,\tau_1]$, under the condition **(DC1)** and **(DC2)** it has a bounded inverse on $D_0[\tau_0,\tau_1]$: $\dot{\psi}^{-1}(F) : D_0[\tau_0,\tau_1] \mapsto D_K[\tau_0,\tau_1]$. Furthermore $\dot{\psi}^{-1}(F)$ is continuous in F :

$$\left\|\dot{\psi}^{-1}(F)(f) - \dot{\psi}^{-1}(F_0)(f)\right\|_K \to 0,$$

for any $f \in \mathscr{R}(\dot{\psi}) = D_0[\tau_0,\tau_1]$ and F such that $\|F - F_0\| \to 0$ and $F - F_0 \in D_0[\tau_0,\tau_1]$. For more detail see Lemma 2 of Gu and Zhang [1993]. The conditions **(C.1)**-**(C.3)**, **(C.4′)**, **(C.5′)**, **(C.6)** and **(CB)** are verified, then we have the following theorem.

Theorem 4.3.3.1 *Suppose that **(CD1)** and **(CD2)** hold. Then all asymptotic Z -estimators $\hat{F}_n$ are asymptotically normal*

$$\sqrt{n}\left(\hat{F}_n - F_0\right) \rightsquigarrow -\dot{\psi}^{-1}(F_0)(\mathbb{Z}_0) \quad \textit{in } (D_K[\tau_0,\tau_1], \|\cdot\|_K),$$

*where $\mathbb{Z}_0$ is a Gaussian random element in $D_0[\tau_0,\tau_1]$. For any bootstrap weights satisfying **(B.1)** through **(B.5)**, all bootstrap asymptotic Z-estimators $\hat{F}_n^*$ are also asymptotically normal:*

$$\sqrt{n}\left(\hat{F}_n^* - \hat{F}_n\right) \rightsquigarrow -\dot{\psi}^{-1}(F_0)\left(c\cdot\hat{\mathbb{Z}}_0\right) \quad \textit{in } (D_K[\tau_0,\tau_1], \|\cdot\|_K)$$

*in $\mathbb{P}^*$ -probability, where $\hat{\mathbb{Z}}_0 \overset{d}{=} \mathbb{Z}_0$, and c is the constant given in the definition **(B.5)** for the exchangeable weights.*

Bootstrap weights

Let us present some examples of the bootstrap weights satisfying the conditions **(B.1)–(B.5)**, we can refer to Præstgaard and Wellner [1993] and Cheng [2015] for further details. More precisely, the following examples are provided in this compressed form Cheng [2015], we have included some minor changes necessary for our setting.

Example 4.3.3.2 (i.i.d.-Weighted Bootstraps) *In this example, the bootstrap weights are defined as* $W_{ni} = \omega_i/\overline{\omega}_n$*, where* $\omega_1, \omega_2, \ldots, \omega_n$ *are i.i.d. positive r.v.s. with* $\|\omega_1\|_{2,1} < \infty$*, where*

$$\|W_{n1}\|_{2,1} = \int_0^\infty \sqrt{\mathbb{P}_W(W_{n1} \geq u)}\, du,$$

$$\overline{\omega}_n = \sum_{i=1}^n \omega_i.$$

Thus, we can choose $\omega_i \sim Exponential(1)$ *or* $\omega_i \sim Gamma(4,1)$*. The former corresponds to the Bayesian bootstrap. The multiplier bootstrap is often thought to be a smooth alternative to the nonparametric bootstrap; see Lo [1993]. The value of* c^2 *is calculated as*

$$Var(\omega_1)/(E\omega_1)^2.$$

Example 4.3.3.3 (Efron's bootstrap) *As already mentioned, the weights for the Efron bootstrap satisfy the conditions* **(B.1)–(B.5)** *with* $c^2 = 1$ *and are* $W_n \sim Multinomial(n; n^{-1}, \ldots, n^{-1})$.

Example 4.3.3.4 (The delete-h Jackknife) *In the delete-h jackknife, see Shao and Wu [1987], the bootstrap weights are generated by permuting the deterministic weights*

$$w_n = \left\{\frac{n}{n-h}, \ldots, \frac{n}{n-h}, 0, \ldots, 0\right\} \quad with \sum_{i=1}^n w_{ni} = n.$$

Specifically, we have $W_{nj} = w_{nR_n(j)}$ *where* $R_n(\cdot)$ *is a random permutation uniformly distributed over* $\{1, \ldots, n\}$*. In Condition* **(B.5)**, $c^2 = h/(n-h)$*. Thus, we need to choose* $h/n \to \alpha \in (0,1)$ *such that* $c > 0$*. Therefore, the usual jackknife with* $h = 1$ *is inconsistent for estimating the distribution.*

Let us recall some examples from Janssen [2005].

Example 4.3.3.5 *The* $m(n)$ *out of* n*-bootstrap weights*

$$W_{ni} = m(n)^{1/2}\left(\frac{1}{m(n)}M_{ni} - \frac{1}{n}\right)$$

are given by a multinomial distributed random variable $(M_{n1}, \ldots, M_{n,n})$ *with sample size*

$$m(n) = \sum_{i=1}^n M_{ni}$$

and equal success probability. In this case, the conditions **(B.1)–(B.5)** *are valid, (details of the proof are given in [Janssen and Pauls, 2003, (8.37)-(8.46)]).*

Example 4.3.3.6 *The $m(n)$-double bootstrap can be described by the weights*

$$\mathrm{W}_{ni} = \frac{m(n)^{1/2}}{\sqrt{2}}\left(\frac{1}{m(n)}\mathrm{M}'_{ni} - \frac{1}{n}\right)$$

Here $(\mathrm{M}'_{n1},\ldots,\mathrm{M}'_{nn})$ denotes a conditional multinomial distributed variable with sample size $m(n) = \sum_{i=1}^{n}\mathrm{M}_{ni}$ and success probability $\mathrm{M}_{ni}/m(n)$ for the i-th cell given by the first example, (details of this example are discussed in Lemma 6.2 of Janssen [2005]).

Remark 4.3.3.7 *As was pointed out in Præstgaard and Wellner [1993], the preceding mentioned bootstraps are "smoother" in some sense than the multinomial bootstrap since they put some (random) weight at all elements in the sample, whereas the multinomial bootstrap puts positive weight at about $1-(1-n^{-1})^n \to 1-e^{-1} = 0.6322$ proportion of each element of the sample, on the average. Notice that when $\omega_i \sim Gamma(4,1)$ so that the W_{ni}/n are equivalent to four-spacings from a sample of $4n-1$ Uniform $(0,1)$ random variables. In Weng [1989] and van Zwet [1979], it was noticed that, in addition to being four times more expensive to implement, the choice of four-spacings depends on the functional of interest and is not universal.*

Remark 4.3.3.8 *It is worth noticing that an appropriate choice of the bootstrap weights W_{ni}'s implies a smaller limit variance, that is, c^2 is smaller than 1. For instance, typical example is the multivariate hypergeometric bootstrap, refer to [Præstgaard and Wellner, 1993, Example 3.4] and the Subsample Bootstrap, [Pauly, 2012, Remark 2.2-(3)]. A detailed discussion about the choice of the weights is certainly out of the scope of the present paper, we refer for review to Barbe and Bertail [1995] and Shao and Tu [1995].*

4.4 Semiparametric framework

The context for a central limit theorem for Z-estimators includes an empirical measure $\mathbb{P}_n$ for n i.i.d. observations and a score operator $\mathrm{B}(\boldsymbol{\theta},\boldsymbol{\eta})$ depending on a parameter $\boldsymbol{\theta}$ of interest and a nuisance parameter $\boldsymbol{\eta}$. Let us give some clarification of this context; assume that the model $\mathscr{P}$ can be parametrized as $(\boldsymbol{\theta},\boldsymbol{\eta}) \mapsto \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}$, where both $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ belongs to an infinite-dimensional sets. Denote $\mathbb{P}$ as the expectation under the true distribution. More generally, consider a statistical model $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}(\mathrm{X})$, with n i.i.d. observations $\mathrm{X}_1,\ldots,\mathrm{X}_n$ drawn from $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}$, where $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\eta} \in \mathfrak{F}$. Assume that the two spaces Θ to be a smooth surface in a Banach space $(\mathbf{B}, \|\cdot\|_\Theta)$ with a norm $\|\cdot\|_\Theta$ and $(\mathfrak{F}, \|\cdot\|_\mathfrak{F})$ possibly but not necessarily subset of Banach space with a norm $\|\cdot\|_\mathfrak{F}$, respectively and the true unknown parameter is $(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)$. An M-estimator $(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n)$ of $(\boldsymbol{\theta},\boldsymbol{\eta})$ has the form

$$(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n) = \operatorname{argmax}\left\{\frac{1}{n}\sum_{i=1}^{n} m(\boldsymbol{\theta},\boldsymbol{\eta},\mathrm{X}_i)\right\}, \tag{4.4.1}$$

where $m(\cdot)$ is a known deterministic function. If we assume that $m(\cdot)$ is Fréchet differentiable with respect to both parameters, so often the maximizing value in (4.4.1) is sought by setting derivatives equal zero, which is given by:

$$\psi\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n,\mathbb{P}_n\right) \equiv \mathbb{P}_n\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right) = 0,$$

where $\mathrm{B}(\cdot,\cdot) : \mathscr{H} \times \mathscr{L} \mapsto \mathscr{F}$ is the score operator acting on the two linear spans of the partial derivative of $m(\cdot)$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ respectively, i.e.,

$$\mathrm{B}(\boldsymbol{\theta},\boldsymbol{\eta})(h,l) = \frac{\partial}{\partial\boldsymbol{\theta}} m(\boldsymbol{\theta},\boldsymbol{\eta})h - \frac{\partial}{\partial\boldsymbol{\eta}} m(\boldsymbol{\theta},\boldsymbol{\eta})l,$$

and $\mathscr{F}$ is some subset of $\mathrm{L}_2(\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}})$ for each $(\boldsymbol{\theta},\boldsymbol{\eta}) \in \Theta \times \Im$, for that we define the set

$$\mathscr{F}(\Theta,\Im) = \bigcup_{\boldsymbol{\theta}\in\Theta,\boldsymbol{\eta}\in\Im} \mathscr{F}(\boldsymbol{\theta},\boldsymbol{\eta}).$$

For notational convenience, we omit Θ and $\Im$ in $\mathscr{F}(\Theta,\Im)$ and write $\mathscr{F}$. We are interested in proving a central limit theorem for Z-estimators $\{\hat{\boldsymbol{\theta}}_n\}$ which is estimated by solving

$$\psi\left(\boldsymbol{\theta},\hat{\boldsymbol{\eta}}_n,\mathbb{P}_n\right) = \mathbb{P}_n \mathrm{B}\left(\boldsymbol{\theta},\hat{\boldsymbol{\eta}}_n\right) = 0, \tag{4.4.2}$$

where we substitute an estimator $\hat{\boldsymbol{\eta}}_n$ for the unknown nuisance parameter.

Remark 4.4.0.1 *1. In some cases, the estimators satisfying (4.4.2) may not exist. We give its weaken version which is known by "nearly maximizing" condition:*

$$\psi\left(\boldsymbol{\theta},\hat{\boldsymbol{\eta}}_n,\mathbb{P}_n\right) = \mathbb{P}_n \mathrm{B}\left(\boldsymbol{\theta},\hat{\boldsymbol{\eta}}_n\right) = o_{\mathbb{P}}(n^{-1/2}), \tag{4.4.3}$$

2. If Θ is included in $\mathbb{R}^k$ and

$$m(\boldsymbol{\theta},\boldsymbol{\eta},x) = \ell(\boldsymbol{\theta},\boldsymbol{\eta},x) = \log\mathrm{lik}(\boldsymbol{\theta},\boldsymbol{\eta})(x)$$

the likelihood function of $(\boldsymbol{\theta},\boldsymbol{\eta})$, then

$$\mathrm{B}(\boldsymbol{\theta},\boldsymbol{\eta})(a,l) = a^\top \frac{\partial}{\partial\boldsymbol{\theta}} \ell(\boldsymbol{\theta},\boldsymbol{\eta}) - \frac{\partial}{\partial\boldsymbol{\eta}} \ell(\boldsymbol{\theta},\boldsymbol{\eta})l,$$

where $a^\top$ is the transpose of the vector $a \in \mathbb{R}^k$. One way of estimating $\boldsymbol{\theta}$ is by solving the efficient score equations

$$\mathbb{P}_n\left(a^\top \frac{\partial}{\partial\boldsymbol{\theta}} \ell(\boldsymbol{\theta},\boldsymbol{\eta}) - \frac{\partial}{\partial\boldsymbol{\eta}} \ell(\boldsymbol{\theta},\boldsymbol{\eta})l\right) = 0.$$

For more details see, for instance, van der Vaart [1998].

Let $\mathbb{P}$ denote the true probability. To prove the central limit theorem, traditional argument assumes that the operator $\psi(\boldsymbol{\theta},\boldsymbol{\eta},\mathbb{P})$ is Fréchet differentiable in $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ with respect to the norm $\|\cdot\|$ on the product space, we can take it as $\|\cdot\| = \|\cdot\|_\Theta + \|\cdot\|_\Im$, with derivative $\dot{\psi}_{\boldsymbol{\theta}}$ and $\dot{\psi}_{\boldsymbol{\eta}}$, respectively. One expands $\psi(\boldsymbol{\theta},\boldsymbol{\eta},\mathbb{P})$ at the true $(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)$ and evaluates the linear approximation at $(\boldsymbol{\theta},\boldsymbol{\eta})$ in some neighborhood of $(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)$, i.e.,

$$\begin{aligned}\psi\left(\boldsymbol{\theta},\boldsymbol{\eta},\mathbb{P}\right) - \psi\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0,\mathbb{P}\right) &= \dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0;\boldsymbol{\eta}_0\right)(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + \dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)([\boldsymbol{\eta}-\boldsymbol{\eta}_0]) \\ &\quad + o_{\mathbb{P}^*}\left(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_\Theta\right) + o_{\mathbb{P}^*}\left(\left\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\right\|_\Im\right).\end{aligned} \tag{4.4.4}$$

Suppose

$$\psi\left(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0, \mathbb{P}\right) = \mathbb{P}\mathrm{B}\left(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0\right) = 0,$$

and that the theory of empirical process can be used to show

$$\mathbb{G}_n \mathrm{B}\left(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_n\right) = \mathbb{G}_n \mathrm{B}\left(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0\right) + o_{\mathbb{P}^*}(1),$$

where $\mathbb{G}_n$ is the empirical process and $(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_n)$ is some sequence converging to $(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)$. By (4.4.3) and some algebra, the difference

$$\sqrt{n}\left(\psi\left(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_n, \mathbb{P}\right) - \psi\left(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0, \mathbb{P}\right)\right) = -\mathbb{G}_n \mathrm{B}\left(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0\right) + o_{\mathbb{P}^*}(1),$$

(see Lemma 4.5.3.2 for details), so that the linearization in (4.4.4) implies

$$\begin{aligned} \dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right) \quad = \quad & -\mathbb{G}_n \mathrm{B}\left(\boldsymbol{\theta}_0\right) + \sqrt{n}\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)([\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0]) \\ & + o_{\mathbb{P}^*}\left(\sqrt{n}\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|_{\Theta}\right) + o_{\mathbb{P}^*}\left(\sqrt{n}\|\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0\|_{\mathfrak{I}}\right) + o_{\mathbb{P}^*}(1) \quad (4.4.5) \end{aligned}$$

The asymptotic distribution of $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$ is determined by the asymptotic joint distribution of the random variables $\mathbb{G}_n \mathrm{B}\left(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0\right)$ and $\sqrt{n}\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)([\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0])$. By assuming the bounded invertibility of the operator $\dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0\right)$ with respect to the same norm $\|\cdot\|_{\Theta}$ used in (4.4.5), we can improve on the consistency of $\hat{\boldsymbol{\theta}}_n$ and prove that $\hat{\boldsymbol{\theta}}_n$ actually converges with a $n^{-1/2}$ rate, i.e.,

$$\sqrt{n}\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|_{\Theta} = \mathrm{O}_{\mathbb{P}^*}(1),$$

(see Lemma 4.5.3.3). With this boundedness and the $\sqrt{n}-$consistency of $\hat{\boldsymbol{\eta}}_n$ the dominant error term $o_{\mathbb{P}^*}\left(\sqrt{n}\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|_{\Theta}\right)$ and $o_{\mathbb{P}^*}\left(\sqrt{n}\|\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0\|_{\mathfrak{I}}\right)$ in (4.4.5) vanishes as n goes to infinity. Hence, by the continuous mapping theorem,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \rightsquigarrow -\dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0\right)(\mathbf{Z}_0),$$

where $\mathbf{Z}_0$ is the limit low of the process

$$-\mathbb{G}_n \mathrm{B}\left(\boldsymbol{\theta}_0\right) + \sqrt{n}\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)([\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0]).$$

In the rest of the paper we refer to the last process as

$$\mathbf{Z}_n = -\mathbb{G}_n \mathrm{B}\left(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0\right) + \sqrt{n}\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)([\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0]).$$

The same problem described before can occur in the setting of semi-parametric framework when the parameter of interest is lying to some infinite dimensional space, where the main difficulty with the classical arguments is that the boundedness invertibility of the derivative operator $\dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0\right)$ with respect to the norm $\|\cdot\|$ used in linearization (4.4.4). For example, without the nuisance parameter, we can see it clearly in the double censoring model given before in (4.3.3). To prove that the estimator $\hat{\boldsymbol{\theta}}_n$ converge to $\boldsymbol{\theta}_0$ with rate $n^{-1/2}$ and thereby validate the linearization and prove a central limit theorem by this argument, however, both the invertibility of $\dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0\right)$ and the differentiability of $\psi(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P})$ with respect to $\boldsymbol{\theta}$ have to be

established with respect to the same norm $\|\cdot\|_\Theta$. At this point one may wonder if the weaker norm $\|\cdot\|_K$ should be used in the place of $\|\cdot\|_\Theta$ in linearization given that the derivative operator is invertible with respect to it. The answer is no; in the double censoring model $\psi(\boldsymbol{\theta}, \boldsymbol{\eta}_0, \mathbb{P})$ is not differentiable with respect to the $\|\cdot\|_K$ norm, as we saw this before.
In an interesting class of models, there is an identity that connects $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$ to the weakly convergent quantity $\mathbf{Z}_n$:

$$\begin{aligned}\dot{\psi}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right) = &-\sqrt{n}\dot{\psi}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0)(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0) \\ &- \dot{\mathbb{P}}_{\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0}\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right), \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)\right) \mathrm{B}\left(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_n\right)\end{aligned}$$

(with $\mathfrak{d} = \hat{\boldsymbol{\theta}}_n$, $\nu = \eta_0$, $a = \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ and $b = \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)$ in Lemma 4.5.1.1). To the best of our knowledge there is not such an identity in the semiparametric framework. A common feature in these problems is that the probability measures $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}$ are convex linearly indexed by $(\boldsymbol{\theta}, \boldsymbol{\eta})$, or we can weaken it to only be convex linearly indexed by $(\boldsymbol{\theta}, \boldsymbol{\eta})$. For a general class of models in which convex linearity can be boundedly extended to the linear span of the parameter space, this linearity identity can be established via Fréchet differentiability of the likelihood equations $\psi\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\right) = 0$. See Section 4.5.4 for more details, also see Section 4.2.2 for functional parameters. This identity allows a linearization applied to $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}$ instead of $\psi(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P})$ through its derivative operator $\dot{\mathbb{P}}_{\mathfrak{d},\nu}(\cdot,\cdot)$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. For models $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}$ with bounded convex linearity, the differential $\dot{\mathbb{P}}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right) \mathrm{B}\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0\right)$ exactly equals the difference $\sqrt{n}\left(\psi\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0, \mathbb{P}_{\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0}\right) - \psi\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0, \mathbb{P}_{\boldsymbol{\theta}_0, \boldsymbol{\eta}_0}\right)\right)$ and the differential $\dot{\mathbb{P}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0)\left(\sqrt{n}\left(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0\right)\right) \mathrm{B}\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0\right)$ exactly equals the difference $\sqrt{n}\left(\psi\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0, \mathbb{P}_{\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_n}\right) - \psi\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0, \mathbb{P}_{\boldsymbol{\theta}_0, \boldsymbol{\eta}_0}\right)\right)$ as in Lemma 4.5.4.1. Consequently, we have

$$\dot{\psi}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right) = -\mathbf{Z}_n + o_{\mathbb{P}^*}(1). \tag{4.4.6}$$

Unlike (4.4.5) where $\hat{\boldsymbol{\theta}}_n$ must converge with an $n^{-1/2}$ rate with respect to $\|\cdot\|_\Theta$ to validate the linearization, there is no need to require this condition in (4.4.6) because the linearization is perfect. The Z-estimators $\{\hat{\boldsymbol{\theta}}_n\}$ still have to converge at the $n^{-1/2}$ rate, but they may converge in any norm as long as the derivative operator is invertible with respect to it. Theorem 4.5.4.2 is a rigorous statement of this argument. For a model $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}$ that is not linearly parameterized, the linearity identity leads to

$$\dot{\psi}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right) = -\mathbf{Z}_n \mathrm{B}\left(\boldsymbol{\theta}_0\right) + o_{\mathbb{P}^*}\left(\sqrt{n}\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|_\Theta\right) + o_{\mathbb{P}^*}\left(\sqrt{n}\left\|\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0\right\|_{\mathfrak{F}}\right) + o_{\mathbb{P}^*}(1).$$

The term $o_{\mathbb{P}^*}\left(\sqrt{n}\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|\right)$ comes from approximating $\dot{\mathbb{P}}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right) \mathrm{B}\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0\right)$ by the difference $\sqrt{n}\left(\psi\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0, \mathbb{P}_{\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0}\right) - \psi\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0, \mathbb{P}_{\boldsymbol{\theta}_0, \boldsymbol{\eta}_0}\right)\right)$. In this case, the uniform boundedness of $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\eta})(\cdot)$ with respect to the norm $\|\cdot\|_\Theta$ is required to improve the rate of convergence for $\hat{\boldsymbol{\theta}}_n$ to make the linearization valid, and a central limit theorem follows. Theorem 4.5.3.1 formulates this argument precisely.

4.5 Central limit theorem

For product space $\ell^\infty(\mathscr{H} \times \mathscr{L})$ we define its norm as

$$\|\cdot\|_{\mathscr{H}\times\mathscr{L}} = \|\cdot\|_{\mathscr{H}} + \|\cdot\|_{\mathscr{L}}.$$

For each fixed $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\eta} \in \Im$ and $\mathbb{P} \in \mathscr{P}$, define the operator $\psi(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P}) = \mathbb{P}\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\eta})$ from $\mathscr{H} \times \mathscr{L}$ to the real line $\mathbb{R}$. Suppose that $\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\eta})$ is bounded in the sense that

$$\|\psi(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P})\|_{\mathscr{H}\times\mathscr{L}} = \|\mathbb{P}\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\eta})\|_{\mathscr{H}\times\mathscr{L}} < \infty,$$

for all $\mathbb{P} \in \mathscr{P}$. Then $\psi(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P}) \in \ell^\infty(\mathscr{H} \times \mathscr{L})$ for each fixed $(\boldsymbol{\theta}, \boldsymbol{\eta}) \in \Theta \times \Im$. The empirical process $\mathbb{G}_n\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\eta})$ acting on $\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\eta})$ is also a function in $\ell^\infty(\mathscr{H} \times \mathscr{L})$ for fixed $(\boldsymbol{\theta}, \boldsymbol{\eta}) \in \Theta \times \Im$. A functional Z-estimator for $\boldsymbol{\theta}_0$ is a sequence of estimates $\{\hat{\boldsymbol{\theta}}_n\} \in \Theta$ which makes the "scores" $\mathbb{P}_n\mathrm{B}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}_n)(h, l), (h, l) \in \mathscr{H} \times \mathscr{L}$, approximately zero:

$$\left\|\psi\left(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_n, \mathbb{P}_n\right)\right\|_{\mathscr{H}\times\mathscr{L}} = o_{\mathbb{P}^*}\left(n^{-1/2}\right),$$

where $\mathbb{P}^*$ denotes the outer probability of $\mathbb{P}^\infty$.

4.5.1 A differential identity

The function $\psi(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P})$ as a map from $\Theta \times \Im$ to $\ell^\infty(\mathscr{H} \times \mathscr{L})$, is Fréchet differentiable with respect to $(\boldsymbol{\theta}, \boldsymbol{\eta})$ and to the norm $\|\cdot\|$ at a point $(\boldsymbol{\vartheta}, \boldsymbol{\nu}) \in \Theta \times \Im$ if there is two bounded linear operators $\dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}, \boldsymbol{\nu}, \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right)(\cdot)$ and $\dot{\psi}_{\boldsymbol{\eta}}\left(\boldsymbol{\vartheta}, \boldsymbol{\nu}, \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right)(\cdot)$ the partial Fréchet differential; mapping from $(\mathrm{lin}(\Theta), \|\cdot\|_\Theta)$ to $(\ell^\infty(\mathscr{H} \times \mathscr{H}), \|\cdot\|_{\mathscr{H}\times\mathscr{L}})$ and from $(\mathrm{lin}(\Im), \|\cdot\|_\Im)$ to $(\ell^\infty(\mathscr{H} \times \mathscr{L}), \|\cdot\|_{\mathscr{H}\times\mathscr{L}})$ such that

$$\begin{aligned}\left\|\psi\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right) - \psi\left(\boldsymbol{\vartheta}, \boldsymbol{\nu}, \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right) - \dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}, \boldsymbol{\nu}, \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right)(\boldsymbol{\theta} - \boldsymbol{\vartheta}) - \dot{\psi}_{\boldsymbol{\eta}}\left(\boldsymbol{\vartheta}, \boldsymbol{\nu}, \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right)(\boldsymbol{\eta} - \boldsymbol{\nu})\right\|_{\mathscr{H}\times\mathscr{L}} \\ = o(\|\boldsymbol{\theta} - \boldsymbol{\vartheta}\|_\Theta) + o(\|\boldsymbol{\eta} - \boldsymbol{\nu}\|_\Im).\end{aligned}$$

For notational convenience, we denote the operators

$$\dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\right) \text{ by } \dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\eta}) : \dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\eta}) \equiv \dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\right),$$

and

$$\dot{\psi}_{\boldsymbol{\eta}}\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\right) \text{ by } \dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\eta}) : \dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\eta}) \equiv \dot{\psi}_{\boldsymbol{\eta}}\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\right).$$

Recall that for a fixed $\boldsymbol{\vartheta} \in \Theta$ and $\boldsymbol{\nu} \in \Im$ the operator $\mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu})$ is bounded in the sense that

$$\|\mathbb{P}\mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu})\|_{\mathscr{H}\times\mathscr{L}} < \infty \text{ for all } \mathbb{P} \in \mathscr{P}.$$

Thus for a fixed $(\boldsymbol{\vartheta}, \boldsymbol{\nu}) \in \Theta \times \Im$ the probability measure $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}$ induces a mapping $(\boldsymbol{\theta}, \boldsymbol{\eta}) \mapsto \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu})$ from $\Theta \times \Im$ to $\ell^\infty(\mathscr{H} \times \mathscr{L})$. The map $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu})$, as a function of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is Fréchet differentiable with respect to the norm $\|\cdot\|$ at a point $(\boldsymbol{\vartheta}, \boldsymbol{\nu}) \in \Theta \times \Im$ if there is two linear operators $\dot{\mathbb{P}}_{\boldsymbol{\theta}}(\boldsymbol{\vartheta}, \boldsymbol{\nu})(\cdot)$

and $\dot{\mathbb{P}}_{\boldsymbol{\eta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(\cdot)$ the partial Fréchet differential; such that $\dot{\mathbb{P}}_{\boldsymbol{\theta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(\cdot)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ and $\dot{\mathbb{P}}_{\boldsymbol{\eta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(\cdot)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ are bounded and

$$\begin{aligned}\left\|\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})-\mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})-\dot{\mathbb{P}}_{\boldsymbol{\theta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(\boldsymbol{\theta}-\boldsymbol{\vartheta})\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})-\dot{\mathbb{P}}_{\boldsymbol{\eta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(\boldsymbol{\eta}-\boldsymbol{\nu})\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})\right\|_{\mathscr{H}\times\mathscr{L}}\\ =o(\|\boldsymbol{\theta}-\boldsymbol{\vartheta}\|_{\Theta})+o(\|\boldsymbol{\eta}-\boldsymbol{\nu}\|_{\Im}).\end{aligned}$$

Lemma 4.5.1.1 *Assume that* $\psi\left(\boldsymbol{\theta},\boldsymbol{\eta},\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\right)\equiv 0$ *for all* $(\boldsymbol{\theta},\boldsymbol{\eta})\in\Theta\times\Im$. *For any* $(\boldsymbol{\vartheta},\boldsymbol{\nu})\in\Theta\times\Im$, *suppose that* $\psi(\boldsymbol{\theta},\boldsymbol{\eta},\mathbb{P})$ *is Fréchet differentiable with respect to the norm* $\|\cdot\|$ *in a neighborhood of* $(\boldsymbol{\vartheta},\boldsymbol{\nu})$, *and the operators* $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ *and* $\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ *are continuous as functions of* $(\boldsymbol{\theta},\boldsymbol{\eta})$ *at* $(\boldsymbol{\vartheta},\boldsymbol{\nu})$, *i.e.;*

$$\|\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})-\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})\|\equiv\sup_{\|a\|\leq 1}\|\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})(a)-\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(a)\|_{\mathscr{H}\times\mathscr{L}}\longrightarrow 0 \tag{4.5.1}$$

$$\|\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta},\boldsymbol{\eta})-\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})\|\equiv\sup_{\|b\|\leq 1}\|\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta},\boldsymbol{\eta})(b)-\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(b)\|_{\mathscr{H}\times\mathscr{L}}\longrightarrow 0 \tag{4.5.2}$$

as $\|\boldsymbol{\theta}-\boldsymbol{\vartheta}\|_{\Theta}\to 0$ *and* $\|\boldsymbol{\eta}-\boldsymbol{\nu}\|_{\Im}\to 0$, *respectively. If* $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ *is Fréchet differentiable with respect to the norm* $\|\cdot\|$ *at* $(\boldsymbol{\vartheta},\boldsymbol{\nu})\in\Theta\times\Im$, *then the operator* $\psi\left(\boldsymbol{\theta},\boldsymbol{\eta},\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\right)$ *as a function of* $\boldsymbol{\theta}$ *and* $\boldsymbol{\eta}$ *is Fréchet differentiable with respect to the norm* $\|\cdot\|$ *at* $(\boldsymbol{\vartheta},\boldsymbol{\nu})\in\Theta\times\Im$ *and the following identity holds for all* $(a,b)\in\mathrm{lin}(\Theta)\times\mathrm{lin}(\Im)$:

$$\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(a)+\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(b)+\dot{\mathbb{P}}_{\boldsymbol{\theta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(a)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})+\dot{\mathbb{P}}_{\boldsymbol{\eta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(b)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})=0. \tag{4.5.3}$$

4.5.2 A condition of uniform boundedness

The uniform boundedness of the operators $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)$ is needed to establish the rate of convergence for a sequence of Z-estimators $\{\hat{\boldsymbol{\theta}}_n\}$. This property is also needed to asymptotically replace $\dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)(-\mathbf{Z}_n)$ by $\dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)(-\mathbf{Z}_n)$ for a consistent estimator $\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)$ and thus allows us to apply the continuous mapping theorem on $\dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)(\mathbf{Z}_n)$ to obtain a central limit theorem. As mentioned in the preceding paragraph our parameter of interest is $\boldsymbol{\theta}\in\Theta$, as a consequence we focus on the partial Fréchet derivative the operator $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ and its inverse $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})$, rather than the partial Fréchet derivative the operator $\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta},\boldsymbol{\eta})$, where Θ is a subset in a Banach space $(\mathbf{B},\|\cdot\|_{\Theta})$, the closure $\overline{\mathrm{lin}(\Theta)}$ is a Banach space with the same norm $\|\cdot\|_{\Theta}$ (Lemma II.1.3 on page 50, Dunford and Schwartz [1958] Part I). Because $(\ell^{\infty}(\mathscr{H}\times\mathscr{L}),\|\cdot\|_{\mathscr{H}\times\mathscr{L}})$ is also a Banach space, the bounded operators $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})$ and $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ can be uniquely extended to the closures of their domains by continuity (see, e.g., Lemma I.6.16 on page 23 of Dunford and Schwartz [1958], Part I).

The unique continuous extensions of $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})$ and $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ on the closures of their domains are also denoted by $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})$ and $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$. The extension $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})$ on $\overline{\mathscr{R}(\dot{\psi}_{\boldsymbol{\theta}})}$ is also the inverse of the extension $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ on $\mathrm{lin}(\Theta)$. We use $\mathscr{R}(\dot{\psi}_{\boldsymbol{\theta}})$ instead of $\mathscr{R}(\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta}))$ to denote the common subspace on which every $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})$ resides.

Lemma 4.5.2.1 *Suppose that, for every fixed* $(\boldsymbol{\theta},\boldsymbol{\eta}) \in \Theta \times \Im$, *the operator* $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ *mapping from* $\left(\overline{\mathrm{lin}(\Theta)}, \|\cdot\|_{\mathrm{K}}\right)$ *to* $(\ell^\infty(\mathscr{H}\times\mathscr{L}), \|\cdot\|_{\mathscr{H}\times\mathscr{L}})$ *has a bounded inverse* $\dot{\psi}^{-1}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ *on a fixed subspace* $\overline{\mathscr{R}(\dot{\psi})} \subset \ell^\infty(\mathscr{H}\times\mathscr{L})$. *Further assume that* $\dot{\psi}^{-1}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ *converges on* $\overline{\mathscr{R}(\dot{\psi})}$ *to* $\dot{\psi}^{-1}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)$ *with respect to a norm* $\|\cdot\|_{\mathrm{K}}$: *for any* $f \in \overline{\mathscr{R}(\dot{\psi})}$

$$\left\|\dot{\psi}^{-1}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})(f) - \dot{\psi}^{-1}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)(f)\right\|_{\mathrm{K}} \longrightarrow 0 \tag{4.5.4}$$

as $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_\Theta \to 0$ *and* $\left\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\right\|_\Im \to 0$. *Assume that* $\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|_\Theta \to_{\mathbb{P}^*} 0$, $\left\|\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\right\|_\Im \to_{\mathbb{P}^*} 0$ *and that* $\mathbb{Z}_n \rightsquigarrow \mathbb{Z}_0$ *in* $\ell^\infty(\mathscr{H}\times\mathscr{L})$ *as* $n\to\infty$. *Then:*

$$\left\|\left(\dot{\psi}^{-1}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right) - \dot{\psi}^{-1}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)\right)(\mathbf{Z}_n)\right\|_{\mathrm{K}} = o_{\mathbb{P}^*}(1).$$

4.5.3 A central limit theorem

We need the following assumptions for a central limit theorem.

(H.1) For all $(\boldsymbol{\theta},\boldsymbol{\eta}) \in \Theta\times\Im$; $\psi\left(\boldsymbol{\theta},\boldsymbol{\eta},\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\right) = \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\theta},\boldsymbol{\eta}) \equiv 0$ in $\ell^\infty(\mathscr{H}\times\mathscr{L})$.

(H.2) As $n\to\infty$, for any decreasing $\delta_n \downarrow 0$, the stochastic equicontinuity condition

$$\sup\left\{\left\|\mathbb{G}_n\left(\mathrm{B}(\boldsymbol{\theta},\boldsymbol{\eta}) - \mathrm{B}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)\right)\right\|_{\mathscr{H}\times\mathscr{L}} : \left(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_\Theta \vee \left\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\right\|_\Im\right) \le \delta_n\right\} = o_{\mathbb{P}^*}(1),$$

holds at the point $(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)$.

(H.3) $\sqrt{n}\|\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\|_\Im = \mathrm{O}_{\mathbb{P}^*}(1)$.

(H.4) The process $\mathbf{Z}_n = -\mathbb{G}_n\mathrm{B}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right) + \sqrt{n}\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)([\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0]) \rightsquigarrow \mathbf{Z}_0$ in $\ell^\infty(\mathscr{H}\times\mathscr{L})$, where $\rightsquigarrow$ indicates weak convergence in $\ell^\infty(\mathscr{H}\times\mathscr{L})$ to a tight Borel measurable random element $\mathbf{Z}_0$.

(H.5) For a fixed $(\boldsymbol{\vartheta},\boldsymbol{\nu}) \in \Theta\times\Im$, the operator $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ as a function of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is Fréchet differentiable with respect to the norm $\|\cdot\|_{\Theta\times\Im} = \|\cdot\|_\Theta + \|\cdot\|_\Im$ at $(\boldsymbol{\vartheta},\boldsymbol{\nu})$. Furthermore, the function $(\boldsymbol{\theta},\boldsymbol{\eta}) \mapsto \psi(\boldsymbol{\theta},\boldsymbol{\eta},\mathbb{P})$ from $\Theta\times\Im$ to $\ell^\infty(\mathscr{H}\times\mathscr{L})$ is Fréchet differentiable with respect to the norm $\|\cdot\|$ at $(\boldsymbol{\vartheta},\boldsymbol{\nu})$, i.e., the following equality hold respectively:

$$\begin{aligned}\left\|\psi\left(\boldsymbol{\theta},\boldsymbol{\eta},\mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right) - \psi\left(\boldsymbol{\vartheta},\boldsymbol{\nu},\mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right) - \dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta},\boldsymbol{\nu}\right)\left(\boldsymbol{\theta}-\boldsymbol{\vartheta}\right) - \dot{\psi}_{\boldsymbol{\eta}}\left(\boldsymbol{\vartheta},\boldsymbol{\nu}\right)\left(\boldsymbol{\eta}-\boldsymbol{\nu}\right)\right\|_{\mathscr{H}\times\mathscr{L}} \\ = o(\|\boldsymbol{\theta}-\boldsymbol{\vartheta}\|_\Theta) + o(\|\boldsymbol{\eta}-\boldsymbol{\nu}\|_\Im),\end{aligned}$$

$$\begin{aligned}\left\|\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}) - \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}) - \dot{\mathbb{P}}_{\boldsymbol{\theta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(\boldsymbol{\theta}-\boldsymbol{\vartheta})\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}) - \dot{\mathbb{P}}_{\boldsymbol{\eta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(\boldsymbol{\eta}-\boldsymbol{\nu})\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})\right\|_{\mathscr{H}\times\mathscr{L}} \\ = o(\|\boldsymbol{\theta}-\boldsymbol{\vartheta}\|_\Theta) + o(\|\boldsymbol{\eta}-\boldsymbol{\nu}\|_\Im).\end{aligned}$$

The operators $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ and $\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ are continuous as a function of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ respectively in the sense of (4.5.1) and (4.5.2).

(H.6) For every fixed $(\boldsymbol{\theta},\boldsymbol{\eta}) \in \Theta\times\Im$ the operator $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ from $(\overline{\text{lin}(\Theta)}, \|\cdot\|_\Theta)$ to $(\ell^\infty(\mathscr{H}\times\mathscr{L}), \|\cdot\|_{\mathscr{H}\times\mathscr{L}})$ has a bounded inverse $\dot{\psi}^{-1}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ on a fixed subspace $\overline{\mathscr{R}(\dot{\psi})} \subset \ell^\infty(\mathscr{H}\times\mathscr{L})$. Furthermore $\dot{\psi}^{-1}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ as an operator sequence converges to $\dot{\psi}^{-1}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)$ as $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_\Theta \to 0$ and $\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|_\Im \to 0$ we have

$$\left\|\dot{\psi}^{-1}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})(f) - \dot{\psi}^{-1}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)(f)\right\|_\Theta \longrightarrow 0.$$

Theorem 4.5.3.1 *Let $\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|_\Theta \to_{\mathbb{P}^*} 0$ be a sequence of consistent Z-estimators. Assume **(H.1)** through **(H.6)**. Then*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right) \rightsquigarrow -\dot{\psi}^{-1}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)(\mathbf{Z}_0) \quad \textit{in} \quad (\overline{\text{lin}(\Theta)}, \|\cdot\|_\Theta).$$

We begin to prove Theorem 4.5.3.1 with the following lemma. It asserts that the standardized estimating equations behave asymptotically as $\mathbb{G}_n\mathrm{B}(\boldsymbol{\theta}_0)$ under our assumptions.

Lemma 4.5.3.2 *Let **(H.1)** and **(H.2)** hold. Then*

$$\sqrt{n}\psi\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n,\mathbb{P}\right) = -\mathbb{G}_n\mathrm{B}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right) + o_{\mathbb{P}^*}(1).$$

The next lemma shows that $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)$ is actually $\mathrm{O}_{\mathbb{P}^*}(1)$ under the mentioned assumptions.

Lemma 4.5.3.3 *Assume **(H.1)** through **(H.6)** and that $\hat{\boldsymbol{\theta}}_n$ is consistent: $\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\| \to_{\mathbb{P}^*} 0$. Then*

$$\sqrt{n}\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\| = \mathrm{O}_{\mathbb{P}^*}(1).$$

4.5.4 Bounded convex linearity

The parametrization $(\boldsymbol{\theta},\boldsymbol{\eta}) \mapsto \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}$ is said to be convex linear if $(\boldsymbol{\theta},\boldsymbol{\eta}) = (\lambda_1\boldsymbol{\theta}_1+\lambda_2\boldsymbol{\theta}_2, \lambda_1'\boldsymbol{\eta}_1 + \lambda_2'\boldsymbol{\eta}_2) \in \text{lin}(\Theta)\times\text{lin}(\Im)$ implies

$$\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}} = \lambda_1\lambda_1'\mathbb{P}_{\boldsymbol{\theta}_1,\boldsymbol{\eta}_1} + \lambda_2\lambda_1'\mathbb{P}_{\boldsymbol{\theta}_2,\boldsymbol{\eta}_1} + \lambda_1\lambda_2'\mathbb{P}_{\boldsymbol{\theta}_1,\boldsymbol{\eta}_2} + \lambda_2\lambda_2'\mathbb{P}_{\boldsymbol{\theta}_2,\boldsymbol{\eta}_2} \in \mathscr{P},$$

for any $(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1),(\boldsymbol{\theta}_2,\boldsymbol{\eta}_2) \in \Theta\times\Im$ and any real numbers λ_1, λ_1', λ_2 and λ_2' such that $\lambda_1,\lambda_1' \geq 0, \lambda_2,\lambda_2' \geq 0$ and $\lambda_1+\lambda_2 = \lambda_1'+\lambda_2' = 1$. Convex linearity is referred to as bounded with respect to a norm $\|\cdot\|$ on $\text{lin}(\Theta)\times\text{lin}(\Im)$ if

(H.7) For any $(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1),\ldots,(\boldsymbol{\theta}_k,\boldsymbol{\eta}_k)$ in $\Theta\times\Im$, and any real numbers $\lambda_1,\ldots,\lambda_k,\lambda_1',\ldots,\lambda_k', k \geq 1$, there is a constant $\mathbf{C} < \infty$ and $\alpha > 1$ such that

$$\left\|\sum_{i=1}^k\sum_{j=1}^k \lambda_i\lambda_j'\mathbb{P}_{\boldsymbol{\theta}_i,\boldsymbol{\eta}_j}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})\right\|_{\mathscr{H}\times\mathscr{L}} \leq \mathbf{C}\left\|\sum_{i=1}^k\lambda_i\boldsymbol{\theta}_i\right\|_\Theta + \mathbf{C}\left\|\sum_{j=1}^k\lambda_j'\boldsymbol{\eta}_j\right\|_\Im^\alpha, \tag{4.5.5}$$

holds for every fixed $(\boldsymbol{\vartheta},\boldsymbol{\nu}) \in \Theta\times\Im$ where $\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ is the score operator mapping from $\mathscr{H}\times\mathscr{L}$ to $\mathscr{F}$.

Lemma 4.5.4.1 *If the parametrization* $(\boldsymbol{\theta},\boldsymbol{\eta}) \mapsto \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}$ *is boundedly convex linear, then the mapping* $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ *is Fréchet differentiable with respect to the norm* $\|\cdot\|$ *at all* $(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1) \in \Theta \times \Im$ *and the partial derivative operator* $\dot{\mathbb{P}}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1)(\cdot)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ *and* $\dot{\mathbb{P}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1)(\cdot)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ *are given by*

$$\dot{\mathbb{P}}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1)(\boldsymbol{\theta}-\boldsymbol{\theta}_1)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}) = \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}_1}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}) - \mathbb{P}_{\boldsymbol{\theta}_1,\boldsymbol{\eta}_1}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}),$$
$$\dot{\mathbb{P}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1)(\boldsymbol{\eta}-\boldsymbol{\eta}_1)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}) = \mathbb{P}_{\boldsymbol{\theta}_1,\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}) - \mathbb{P}_{\boldsymbol{\theta}_1,\boldsymbol{\eta}_1}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}),$$

for any $(\boldsymbol{\theta},\boldsymbol{\eta})$, $(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1)$ *and* $(\boldsymbol{\vartheta},\boldsymbol{\nu})$ *in* $\Theta \times \Im$.

In view of Lemma (4.5.1.1) the differential identity (4.5.3) for models with bounded convex linearity can be improved to

$$\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1)(\boldsymbol{\theta}-\boldsymbol{\theta}_1) + \dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1)(\boldsymbol{\eta}-\boldsymbol{\eta}_1) = -\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}_1}\mathrm{B}(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1) - \mathbb{P}_{\boldsymbol{\theta}_1,\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1), \tag{4.5.6}$$

for any $(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1),(\boldsymbol{\theta},\boldsymbol{\eta}) \in \Theta \times \Im$. We choose $(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1) = (\boldsymbol{\theta}_n,\boldsymbol{\eta}_n)$ and $(\boldsymbol{\theta},\boldsymbol{\eta}) = (\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)$, we get;

$$\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n,\boldsymbol{\eta}_n)(\boldsymbol{\theta}_0-\boldsymbol{\theta}_n) + \dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_n,\boldsymbol{\eta}_n)(\boldsymbol{\eta}_0-\boldsymbol{\eta}_n) = -\mathbb{P}_{\boldsymbol{\theta}_0,\boldsymbol{\eta}_n}\mathrm{B}(\boldsymbol{\theta}_n,\boldsymbol{\eta}_n) - \mathbb{P}_{\boldsymbol{\theta}_n,\boldsymbol{\eta}_0}\mathrm{B}(\boldsymbol{\theta}_n,\boldsymbol{\eta}_n).$$

For these models, a strong enough norm $\|\cdot\|_\Theta$ may be used to obtain the differentiability of $\psi\left(\boldsymbol{\theta},\boldsymbol{\eta},\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\right)$ and condition **(H.6)** and therefore the identity (4.5.6). Then a weaker norm $\|\cdot\|_{\mathrm{K}}$ applied to the space Θ can be used to establish the invertibility of $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})$ and the pointwise convergence in (4.5.4). The difference on the right of (4.5.6) also implies that no rate control, such as that in Lemma 4.5.3.2 is needed. This is the reason for which we can actually obtain asymptotic normality with the weaker norm. This usually improves the applicability of the central limit theorem. To be more specific, the assumptions replacing **(H.5)** and **(H.6)** are the following.

(H.5′) The function $\psi(\boldsymbol{\theta},\boldsymbol{\eta},\mathbb{P})$ as a map from $\Theta \times \Im$ to $\ell^\infty(\mathscr{H} \times \mathscr{L})$ is Fréchet differentiable with respect to the norm $\|\cdot\|_{\Theta\times\Im} = \|\cdot\|_\Theta + \|\cdot\|_\Im$. The operators $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ and $\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ are continuous as a function of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ respectively in the sense of (4.5.1) and (4.5.2).

(H.6′) For every fixed $(\boldsymbol{\theta},\boldsymbol{\eta})$ the operator $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ from $(\overline{\mathrm{lin}(\Theta)},\|\cdot\|_{\mathrm{K}})$ to $(\ell^\infty(\mathscr{H}\times\mathscr{L}),\|\cdot\|_{\mathscr{H}\times\mathscr{L}})$ has a bounded inverse $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})$ on a fixed subspace $\overline{\mathscr{R}(\dot{\psi})} \subset \ell^\infty(\mathscr{H}\times\mathscr{L})$. Furthermore $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})$ as an operator sequence converges to $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)$ as $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_\Theta \to 0$ and $\left\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\right\|_\Im \to 0$ we have

$$\left\|\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})(f) - \dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)(f)\right\|_{\mathrm{K}} \longrightarrow 0.$$

Theorem 4.5.4.2 *For a model with bounded convex linearity specified in* ***(H.7)*** *assume* ***(H.1)*** *through* ***(H.4)***, ***(H.5′)*** *and* ***(H.6′)***, *for a sequence of consistent* Z*-estimators* $(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n)$, *we have* $\hat{\boldsymbol{\theta}}_n$ *is asymptotically normal and,*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right) \rightsquigarrow -\dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)(\mathbf{Z}_0) \quad \textit{in} \quad (\overline{\mathrm{lin}(\Theta)},\|\cdot\|_{\mathrm{K}}).$$

Theorem 4.5.4.2 is mainly motivated by Vardi and Zhang [1992] on the multiplicative censoring model and Gu and Zhang [1993] on the double censoring model. The key assumptions are formulated in **(H.5′)** and **(H.6′)** which are not explicit in these two papers. Other assumptions such as **(H.1)**, **(H.2)** and **(H.3)** are mainly from the traditional arguments, see Huber [1964], Huber [1967], Pakes and Pollard [1989], Pollard [1989], Pollard [1985], van der Vaart [1994] and van der Vaart [1995].

4.6 Mathematical developments

This section is devoted to the proofs of our main result. The previously presented notation continues to be used in the following.

Proof of Lemma 4.2.2.7

By **(C.5)**, as $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \to 0$ the sequence of continuous linear operators $\dot{\psi}^{-1}(\boldsymbol{\theta})$ converge on $\overline{\mathscr{R}(\dot{\psi})}$ to $\dot{\psi}^{-1}(\boldsymbol{\theta}_0)$ as mapping from the Banach space $\overline{\mathscr{R}(\dot{\psi})}$ to the Banach space $\mathrm{lin}(\Theta)$. So by the Banach-Steinhaus theorem (for example, Theorem II.3.6 on page 60 of Dunford and Schwartz [1988]) the norm of the operators $\dot{\psi}^{-1}(\boldsymbol{\theta})$ is uniformly bounded:

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\beta} \|\dot{\psi}^{-1}(\boldsymbol{\theta})\| \leq 1/\gamma < \infty,$$

for some positive numbers $\beta, \gamma > 0$. Thus for any $a \in \mathrm{lin}(\Theta)$, we have

$$\|a\| = \|\dot{\psi}^{-1}(\boldsymbol{\theta})\dot{\psi}(\boldsymbol{\theta})(a)\| \leq \frac{\|\dot{\psi}(\boldsymbol{\theta})(a)\|_{\mathscr{H}}}{\gamma}, \tag{4.6.1}$$

for all $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \leq \beta$. In identity (4.2.14), we take $a = \sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0\right)$ and $\boldsymbol{\vartheta} = \hat{\boldsymbol{\theta}}_n^*$, it follows by the linearity of $\dot{\mathbb{P}}_{\boldsymbol{\vartheta}}(a)\mathrm{B}(\boldsymbol{\vartheta})$ in a, the definition of Féchet differentiability of $\boldsymbol{\theta} \mapsto \mathbb{P}_{\boldsymbol{\theta}}\mathrm{B}(\boldsymbol{\vartheta})$ and **(C.1)** that;

$$\begin{aligned}
\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0\right)\right) &= -\dot{\mathbb{P}}_{\hat{\boldsymbol{\theta}}_n^*}\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0\right)\right)\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n^*\right) \\
&= \sqrt{n}\left(\mathbb{P}_{\boldsymbol{\theta}_0}\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n^*\right) - \mathbb{P}_{\hat{\boldsymbol{\theta}}_n^*}\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n^*\right)\right) + o_{\mathbf{P}^*}(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0)\|) \\
&= \sqrt{n}\left(\mathbb{P}_{\boldsymbol{\theta}_0} - \mathbb{P}_n\right)\left(\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n^*\right) - \mathrm{B}\left(\boldsymbol{\theta}_0\right)\right) \\
&\quad + \sqrt{n}\left(\mathbb{P}_n - \hat{\mathbb{P}}_n\right)\left(\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n^*\right) - \mathrm{B}\left(\boldsymbol{\theta}_0\right)\right) \\
&\quad - \sqrt{n}\hat{\mathbb{P}}_n\left(\mathrm{B}\left(\boldsymbol{\theta}_0\right)\right) + o_{\mathbf{P}^*}(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0)\|).
\end{aligned} \tag{4.6.2}$$

The last one holds by the definition of $\hat{\boldsymbol{\theta}}_n^*$ given in (4.2.8). Therefore by the triangular inequality we have for a consistent $\hat{\boldsymbol{\theta}}_n^*$ that;

$$\begin{aligned}
\left\|\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0\right)\right)\right\|_{\mathscr{H}} - \left\|\hat{\mathbb{G}}_n\left(\mathrm{B}\left(\boldsymbol{\theta}_0\right)\right)\right\|_{\mathscr{H}} &\leq \|\mathbb{G}_n\|_{\mathscr{D}_n} + \|\hat{\mathbb{G}}_n\|_{\mathscr{D}_n} + o_{\mathbf{P}^*}(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0)\|) \\
&= o_{\mathbf{P}^*}(1) + o_{\mathbf{P}^*}(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0)\|).
\end{aligned}$$

Then by this result and the boundedness in (4.2.8) we obtain in $\mathbb{P}^*$ for n sufficiently large:

$$\begin{aligned}\gamma\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^*-\boldsymbol{\theta}_0\right) &\leq \left\|\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^*-\boldsymbol{\theta}_0\right)\right)\right\|_{\mathscr{H}} \\ &\leq \left\|\hat{\mathbb{G}}_n\left(\mathrm{B}\left(\boldsymbol{\theta}_0\right)\right)\right\|_{\mathscr{H}}+o_{\mathbf{P}^*}(1)+o_{\mathbf{P}^*}(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n^*-\boldsymbol{\theta}_0)\|).\end{aligned}$$

Since **(C.3)** and Theorem 2.2 of Præstgaard and Wellner [1993] imply that $\left\|\hat{\mathbb{G}}_n\left(\mathrm{B}\left(\boldsymbol{\theta}_0\right)\right)\right\|_{\mathscr{H}} = o_{\widehat{\mathbb{P}}}(1)$ in $\mathbb{P}^*-$probability, consequently the desired result follows. □

Proof of Theorem 4.2.2.5

By assumptions **(C.1)**, **(C.4)** and the identity (4.2.14) we get that;

$$\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^*-\boldsymbol{\theta}_0\right)\right)=\sqrt{n}\mathbb{P}_{\boldsymbol{\theta}_0}\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n^*\right)+o_{\mathbf{P}^*}\left(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n^*-\boldsymbol{\theta}_0)\|\right) \tag{4.6.3}$$

and

$$\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\right)=\sqrt{n}\mathbb{P}_{\boldsymbol{\theta}_0}\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n\right)+o_{\mathbf{P}^*}\left(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)\|\right). \tag{4.6.4}$$

Subtracting (4.6.3) from (4.6.4), we obtain

$$\begin{aligned}&\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^*-\boldsymbol{\theta}_n\right)\right)-\left[\left(\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n\right)-\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n^*\right)\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\right)\right] \\ &= \sqrt{n}\mathbb{P}_{\boldsymbol{\theta}_0}\left(\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n^*\right)-\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n\right)\right)+o_{\widehat{\mathbb{P}}}\left(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n^*-\boldsymbol{\theta}_0)\|\right)+o_{\mathbb{P}^*}\left(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)\|\right) \\ &= -\sqrt{n}\hat{\mathbb{P}}_n\mathrm{B}\left(\boldsymbol{\theta}_n\right)+\hat{\mathbb{G}}_n\left(\mathrm{B}\left(\boldsymbol{\theta}_n\right)-\mathrm{B}(\hat{\boldsymbol{\theta}}_n^*)\right)+\mathbb{G}_n\left(\mathrm{B}\left(\boldsymbol{\theta}_n\right)-\mathrm{B}(\hat{\boldsymbol{\theta}}_n^*)\right)+o_{\widehat{\mathbb{P}}}\left(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n^*-\boldsymbol{\theta}_0)\|\right) \\ &\quad +o_{\mathbb{P}^*}\left(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)\|\right) \\ &= -\hat{\mathbb{G}}_n\mathrm{B}\left(\boldsymbol{\theta}_0\right)+\hat{\mathbb{G}}_n\left(\mathrm{B}\left(\boldsymbol{\theta}_0\right)-\mathrm{B}(\hat{\boldsymbol{\theta}}_n^*)\right)+\mathbb{G}_n\left(\mathrm{B}\left(\boldsymbol{\theta}_n\right)-\mathrm{B}(\hat{\boldsymbol{\theta}}_n^*)\right)+o_{\widehat{\mathbb{P}}}\left(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n^*-\boldsymbol{\theta}_0)\|\right) \\ &\quad +o_{\mathbb{P}^*}\left(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)\|\right).\end{aligned} \tag{4.6.5}$$

Note that the operator $\dot{\psi}(\cdot)$ is continuous as in (4.2.6) and the sequence $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)$ is asymptotically tight, making use of **(C.2)**, the result in Lemmas 4.2.2.6 and 4.2.2.7 and the definition of $\hat{\boldsymbol{\theta}}_n$; then the triangular inequality with (4.6.5) leads to:

$$\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^*-\boldsymbol{\theta}_n\right)\right)=-\hat{\mathbb{G}}_n\left(\mathrm{B}\left(\boldsymbol{\theta}_0\right)\right)+o_{\mathbf{P}^*}(1). \tag{4.6.6}$$

By applying the Banach-Steinhaus theorem to the convergent sequence of operators $\dot{\psi}^{-1}(\boldsymbol{\theta})$ by **(C.5)**, then the consistency of $\hat{\boldsymbol{\theta}}_n^*$ imply that the operator norm of $\dot{\psi}^{-1}(\hat{\boldsymbol{\theta}}_n^*)$ is uniformly bounded in $\mathbb{P}^*-$probability when n is sufficiently large. It maps a term of $o_{\mathbf{P}^*}(1)$ in the $\|\cdot\|_{\mathscr{H}}$-norm into a term of $o_{\mathbf{P}^*}(1)$ in $\|\cdot\|$-norm: $\dot{\psi}^{-1}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(o_{\mathbf{P}^*}(1)\right)=o_{\mathbf{P}^*}(1)$. This means that

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^*-\boldsymbol{\theta}_n\right) &= \dot{\psi}^{-1}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(\hat{\mathbb{G}}_n\mathrm{B}\left(\boldsymbol{\theta}_0\right)+o_{\mathbf{P}^*}(1)\right) \\ &= \dot{\psi}^{-1}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(\hat{\mathbb{G}}_n\mathrm{B}\left(\boldsymbol{\theta}_0\right)\right)+o_{\mathbf{P}^*}(1).\end{aligned} \tag{4.6.7}$$

Since $\hat{\mathbb{G}}_n \mathrm{B}(\boldsymbol{\theta}_0) \rightsquigarrow c \cdot \hat{\mathbb{Z}}_0$ in $\ell^\infty(\mathscr{H})$ in $\mathbb{P}^*-$probability by **(C.3)** and Theorem 2.2 of Præstgaard and Wellner [1993], then by the triangular inequality and Lemma 2.2 in Zhan [2002] (applied with the K-norm replaced by $\|\cdot\|$) we obtain

$$\dot{\psi}^{-1}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(\hat{\mathbb{G}}_n \mathrm{B}(\boldsymbol{\theta}_0)\right) = \dot{\psi}^{-1}(\boldsymbol{\theta}_0)\left(\hat{\mathbb{G}}_n \mathrm{B}(\boldsymbol{\theta}_0)\right) + o_{\mathbf{P}^*}(1),$$

in $\mathbb{P}^*-$probability. Noting that a term of order $o_{\mathbf{P}^*}(1)$ is also a term of an order $o_{\mathbb{P}^*}(1)$ in $\mathbb{P}^*-$probability. Hence it follows

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_n\right) \rightsquigarrow -\dot{\psi}^{-1}(\boldsymbol{\theta}_0)\left(c \cdot \hat{\mathbb{Z}}_0\right)$$

in $\mathbb{P}^*-$probability, in $(\overline{\mathrm{lin}(\Theta)}, \|\cdot\|)$ as $n \to \infty$ by Slutsky's theorem and the continuous mapping theorem. □

Proof of Theorem 4.2.2.8

By Lemma 4.2.2.3, take $\mathfrak{d} = \boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_n^*$, and $\boldsymbol{\theta}_2 = \hat{\boldsymbol{\theta}}$ in (4.2.15), **(C.1)** and use $\mathbb{P}$ to denote $\mathbb{P}_{\boldsymbol{\theta}_0}$, we obtain by the linearity of the parametrization $\boldsymbol{\theta} \mapsto \mathbb{P}_{\boldsymbol{\theta}}$

$$\begin{aligned}
\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_n\right)\right) &= \sqrt{n}\mathbb{P}_{\boldsymbol{\theta}_n - \boldsymbol{\theta}_0}\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n^*\right) + \sqrt{n}\mathbb{P}\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n^*\right) \\
&= \dot{\psi}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(\sqrt{n}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_n)\right) + \sqrt{n}\mathbb{P}\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n^*\right) \\
&= \sqrt{n}\mathbb{P}\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n^*\right) + \left[\left(\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n^*\right) - \dot{\psi}\left(\hat{\boldsymbol{\theta}}_n\right)\right)\left(\sqrt{n}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_n)\right)\right] \\
&\quad + \dot{\psi}\left(\hat{\boldsymbol{\theta}}_n\right)\left(\sqrt{n}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_n)\right) \\
&= \sqrt{n}\mathbb{P}\left(\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n^*\right) - \mathrm{B}\left(\hat{\boldsymbol{\theta}}_n\right)\right) + \left[\left(\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n^*\right) - \dot{\psi}\left(\hat{\boldsymbol{\theta}}_n\right)\right)\left(\sqrt{n}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_n)\right)\right] \\
&= -\hat{\mathbb{G}}_n \mathrm{B}(\boldsymbol{\theta}_0) + \hat{\mathbb{G}}_n\left(\mathrm{B}(\boldsymbol{\theta}_0) - \mathrm{B}(\hat{\boldsymbol{\theta}}_n^*)\right) + \mathbb{G}_n\left(\mathrm{B}(\boldsymbol{\theta}_n) - \mathrm{B}(\hat{\boldsymbol{\theta}}_n^*)\right) \\
&\quad + \left[\left(\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n^*\right) - \dot{\psi}\left(\hat{\boldsymbol{\theta}}_n\right)\right)\left(\sqrt{n}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_n)\right)\right].
\end{aligned}$$

Now the asymptotic tightness of the sequence $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$ with respect to $\|\cdot\|_{\mathrm{K}}$ holds from $\mathbf{i}'$, by assumption **(C.4′)**, **(C.2)**, Lemma 4.2.2.6 and the consistency of $\hat{\boldsymbol{\theta}}_n^*$, the last equality is written as;

$$\dot{\psi}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_n\right)\right) = -\hat{\mathbb{G}}_n \mathrm{B}(\boldsymbol{\theta}_0) + o_{\mathbf{P}^*}(1),$$

where the term $o_{\mathbf{P}^*}(1)$ denotes a term whose $\|\cdot\|_{\mathscr{H}}-$norm is of order $o_{\mathbf{P}^*}(1)$. Since $\dot{\psi}^{-1}(\boldsymbol{\theta})$ converges to $\dot{\psi}^{-1}(\boldsymbol{\theta}_0)$ on $\overline{\mathscr{R}(\dot{\psi})}$, the Banach-Steinhaus theorem implies that the operator norm of $\dot{\psi}^{-1}\left(\hat{\boldsymbol{\theta}}_n^*\right)$ is uniformly bounded in $\mathbb{P}^*-$probability when n is sufficiently large. It then maps a term of $o_{\mathbf{P}^*}(1)$ in the $\|\cdot\|_{\mathscr{H}}-$norm into a term of $o_{\mathbf{P}^*}(1)$ in K$-$norm: $\dot{\psi}^{-1}\left(\hat{\boldsymbol{\theta}}_n^*\right)(o_{\mathbf{P}^*}(1)) = o_{\mathbf{P}^*}(1)$. This means that

$$\begin{aligned}
\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_n\right) &= \dot{\psi}^{-1}\left(\hat{\boldsymbol{\theta}}_n\right)\left(-\hat{\mathbb{G}}_n \mathrm{B}(\boldsymbol{\theta}_0) + o_{\mathbf{P}^*}(1)\right) \\
&= -\dot{\psi}^{-1}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(\hat{\mathbb{G}}_n \mathrm{B}(\boldsymbol{\theta}_0)\right) + o_{\mathbf{P}^*}(1).
\end{aligned}$$

Since $\hat{\mathbb{G}}_n \mathrm{B}(\boldsymbol{\theta}_0) \rightsquigarrow c \cdot \hat{\mathbb{Z}}_0$ in $\ell^\infty(\mathscr{H})$ in $\mathbb{P}^*-$probability by **(C.3)** and Theorem 2.2 of Præstgaard and Wellner [1993], then by the triangular inequality and Lemma 2.2 in Zhan [2002], we obtain

$$\dot{\psi}^{-1}\left(\hat{\boldsymbol{\theta}}_n^*\right)\left(\hat{\mathbb{G}}_n \mathrm{B}\left(\boldsymbol{\theta}_0\right)\right) = \dot{\psi}^{-1}\left(\boldsymbol{\theta}_0\right)\left(\hat{\mathbb{G}}_n \mathrm{B}\left(\boldsymbol{\theta}_0\right)\right) + o_{\mathbf{P}^*}(1).$$

Hence

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_n\right) \rightsquigarrow -\dot{\psi}^{-1}\left(\boldsymbol{\theta}_0\right)\left(c \cdot \hat{\mathbb{Z}}_0\right) \text{ in } \left(\overline{\mathrm{lin}(\Theta)}, \|\cdot\|_{\mathrm{K}}\right)$$

in $\mathbb{P}^*-$probability as $n \to \infty$ by Slutsky's theorem and the continuous mapping theorem. □

Proof of Lemma 4.5.1

For any $(\boldsymbol{\vartheta}, \boldsymbol{\nu}) \in \Theta \times \Im$, we have

$$\begin{aligned}
\psi\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\right) - \psi\left(\boldsymbol{\vartheta}, \boldsymbol{\nu}, \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right) &= \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}} \mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu}) \\
&= \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}(\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu})) + \left(\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}} - \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right) \mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu}) \\
&\quad + \left(\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}} - \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right)\left(\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu})\right). \qquad (4.6.8)
\end{aligned}$$

Since $\psi(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P})$ is Fréchet differentiable at $(\boldsymbol{\vartheta}, \boldsymbol{\nu})$ and the map $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu})$ as a function of $(\boldsymbol{\theta}, \boldsymbol{\eta})$ is Fréchet differentiable with respect to the norm $\|\cdot\|$ at $(\boldsymbol{\vartheta}, \boldsymbol{\nu})$ the two first term of (4.6.8) can be written respectively as

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}(\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu})) &= \psi\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right) - \psi\left(\boldsymbol{\vartheta}, \boldsymbol{\nu}, \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right) \\
&= \dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}, \boldsymbol{\nu}, \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right)(\boldsymbol{\theta} - \boldsymbol{\vartheta}) + \dot{\psi}_{\boldsymbol{\eta}}\left(\boldsymbol{\vartheta}, \boldsymbol{\nu}, \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right)(\boldsymbol{\eta} - \boldsymbol{\nu}) \\
&\quad + o(\|\boldsymbol{\theta} - \boldsymbol{\vartheta}\|_\Theta) + o(\|\boldsymbol{\eta} - \boldsymbol{\nu}\|_\Im),
\end{aligned}$$

and

$$\begin{aligned}
\left(\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}} - \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right) \mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu}) &= \dot{\mathbb{P}}_{\boldsymbol{\theta}}(\boldsymbol{\vartheta}, \boldsymbol{\nu})(\boldsymbol{\theta} - \boldsymbol{\vartheta})\mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu}) + \dot{\mathbb{P}}_{\boldsymbol{\eta}}(\boldsymbol{\vartheta}, \boldsymbol{\nu})(\boldsymbol{\eta} - \boldsymbol{\nu})\mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu}) \\
&\quad + o(\|\boldsymbol{\theta} - \boldsymbol{\vartheta}\|_\Theta) + o(\|\boldsymbol{\eta} - \boldsymbol{\nu}\|_\Im).
\end{aligned}$$

The operator $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}$ acts on $\mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu})$ linearly, the rest term on the right hand of (4.6.8) can be handled as

$$\begin{aligned}
&\left(\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}} - \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right)\left(\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu})\right) \\
&= \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\left(\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu})\right) - \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\left(\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu})\right) \\
&= \dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}, \boldsymbol{\nu}\right)(\boldsymbol{\theta} - \boldsymbol{\vartheta}) + \dot{\psi}_{\boldsymbol{\eta}}\left(\boldsymbol{\vartheta}, \boldsymbol{\nu}\right)(\boldsymbol{\eta} - \boldsymbol{\nu}) - \dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}, \boldsymbol{\eta}\right)(\boldsymbol{\theta} - \boldsymbol{\vartheta}) - \dot{\psi}_{\boldsymbol{\eta}}\left(\boldsymbol{\theta}, \boldsymbol{\eta}\right)(\boldsymbol{\eta} - \boldsymbol{\nu}) \\
&\quad + o(\|\boldsymbol{\theta} - \boldsymbol{\vartheta}\|_\Theta) + o(\|\boldsymbol{\eta} - \boldsymbol{\nu}\|_\Im) \\
&= \left(\dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}, \boldsymbol{\nu}\right) - \dot{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}, \boldsymbol{\eta}\right)\right)(\boldsymbol{\theta} - \boldsymbol{\vartheta}) + \left(\dot{\psi}_{\boldsymbol{\eta}}\left(\boldsymbol{\vartheta}, \boldsymbol{\nu}\right) - \dot{\psi}_{\boldsymbol{\eta}}\left(\boldsymbol{\theta}, \boldsymbol{\eta}\right)\right)(\boldsymbol{\eta} - \boldsymbol{\nu}) \\
&\quad + o(\|\boldsymbol{\theta} - \boldsymbol{\vartheta}\|_\Theta) + o(\|\boldsymbol{\eta} - \boldsymbol{\nu}\|_\Im).
\end{aligned}$$

We obtained the first term in the last equality by applying the Fréchet differentiability of $\psi\left(\boldsymbol{\vartheta}, \boldsymbol{\nu}, \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\right) = \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu})$ at $(\boldsymbol{\theta}, \boldsymbol{\eta})$. Applying the triangle inequality and the conditions of continuity of $\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\eta})$ and $\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\eta})$, we get

$$\| \left(\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}} - \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\right) \left(\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu})\right) \|_{\mathscr{H}\times\mathscr{L}} = o(\|\boldsymbol{\theta} - \boldsymbol{\vartheta}\|_{\Theta}) + o(\|\boldsymbol{\eta} - \boldsymbol{\nu}\|_{\mathfrak{I}}).$$

Which implies the Fréchet differentiability of $\psi\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\right)$ as a function of $(\boldsymbol{\theta}, \boldsymbol{\eta})$ with respect to the norm $\|\cdot\|$ at $(\boldsymbol{\vartheta}, \boldsymbol{\nu})$ and its Fréchet derivative is given by

$$\dot{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\vartheta}, \boldsymbol{\nu})(a) + \dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\vartheta}, \boldsymbol{\nu})(b) + \dot{\mathbb{P}}_{\boldsymbol{\theta}}(\boldsymbol{\vartheta}, \boldsymbol{\nu})(a)\mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu}) + \dot{\mathbb{P}}_{\boldsymbol{\eta}}(\boldsymbol{\vartheta}, \boldsymbol{\nu})(b)\mathrm{B}(\boldsymbol{\vartheta}, \boldsymbol{\nu}).$$

By the uniqueness of the Fréchet derivative and the fact that $\psi\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\right) \equiv 0$ the identity in (4.5.3) holds. □

Proof of Lemma 4.5.2.1

For any compact set $\mathbb{C} \subset \overline{\mathscr{R}(\dot{\psi})} \subset \ell^{\infty}(\mathscr{H} \times \mathscr{L})$, let $\mathbb{C}(\delta)$ be the δ-enlargement of $\mathbb{C}$ defined by

$$\mathbb{C}(\delta) = \left\{ f \in \overline{\mathscr{R}(\dot{\psi})} : \left\| f - f' \right\|_{\mathscr{H}\times\mathscr{L}} \leq \delta \text{ for some } f' \in \mathbb{C} \right\}.$$

We show that

$$\sup \left\{ \left\| \left(\dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_n\right) - \dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0\right)\right)(f) \right\|_{\mathrm{K}} : f \in \mathbb{C}(\delta) \right\} \longrightarrow 0 \tag{4.6.9}$$

as $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\Theta} \to 0$ and then $\delta \to 0+$. Indeed, by (4.5.4) and the Banach-Steinhaus theorem, the operator norm of $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}, \boldsymbol{\eta})$ is uniformly bounded:

$$\sup_{(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\Theta} \vee \|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|_{\mathfrak{I}}) \leq \beta} \left\| \dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}, \boldsymbol{\eta}) \right\| \leq \mathrm{M} < \infty,$$

for some positive numbers $\beta > 0$ and $\mathrm{M} > 0$. The uniform boundedness of the operators $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}, \boldsymbol{\eta})$ is equivalent to their uniform continuity as mappings in Banach spaces, so that the pointwise convergence in (4.5.4) directly implies the uniform convergence in the norm $\|\cdot\|_{\mathrm{K}}$. Now since $\mathbf{Z}_n \in \mathscr{R}(\dot{\psi})$ converges weakly to $\mathbf{Z}_0$ in $\left(\overline{\mathscr{R}(\dot{\psi})}, \|\cdot\|_{\mathscr{H}\times\mathscr{L}}\right)$, by its asymptotically tightness: for every $\epsilon > 0$ there exists a compact set $\mathbb{C} \subset \overline{\mathscr{R}(\dot{\psi})}$ such that

$$\liminf_{n\to\infty} \mathbb{P}_* \{\mathbf{Z}_n \in \mathbb{C}(\delta)\} \geq 1 - \epsilon,$$

for every $\delta > 0$; see van der Vaart and Wellner [1996] Section 1.3. Making use of the equation (4.6.9), we have

$$\left\| \left(\dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_n\right) - \dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0\right)\right)(\mathbf{Z}_n) \right\|_{\mathrm{K}} = o_{\mathbb{P}^*}(1),$$

as $n \to \infty$ and then $\delta \to 0+$. □

Proof of Lemma 4.5.3.2

Since $\psi\left(\boldsymbol{\theta},\boldsymbol{\eta},\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\right)\equiv 0$ for all $(\boldsymbol{\theta},\boldsymbol{\eta})\in\Theta\times\Im$, we have by the definitions of $\psi(\boldsymbol{\theta},\boldsymbol{\eta},\mathbb{P})$ and the Z-estimator

$$\left\|\sqrt{n}\mathbb{P}_n\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\right\|_{\mathscr{H}\times\mathscr{L}}=o_{\mathbb{P}^*}(1),$$

and

$$\begin{aligned}\sqrt{n}\left(\psi\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n,\mathbb{P}\right)-\psi\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0,\mathbb{P}\right)\right)&=-\mathbb{G}_n\left(\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\right)+o_{\mathbb{P}^*}(1)\\&=-\mathbb{G}_n\mathrm{B}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)-\mathbb{G}_n\left(\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)-\mathrm{B}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)\right)+o_{\mathbb{P}^*}(1).\end{aligned}$$

By **(H.2)**, the consistency of $(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n)$ and the fact that;

$$\begin{aligned}&\mathbb{P}^*\left\{\left\|\mathbb{G}_n\left(\mathrm{B}(\boldsymbol{\theta},\boldsymbol{\eta})-\mathrm{B}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)\right)\right\|_{\mathscr{H}\times\mathscr{L}}\geq\epsilon\right\}\\&\leq\ \mathbb{P}^*\left\{\sup_{(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_\Theta\vee\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|_\Im)\leq\delta_n}\left\|\mathbb{G}_n\left(\mathrm{B}(\boldsymbol{\theta},\boldsymbol{\eta})-\mathrm{B}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)\right)\right\|_{\mathscr{H}\times\mathscr{L}}\geq\epsilon\right\}\\&\quad+\mathbb{P}^*\left\{\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|_\Theta>\delta_n\right\}+\mathbb{P}^*\left\{\left\|\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\right\|_\Im>\delta_n\right\},\end{aligned}$$

it follows that

$$\left\|\mathbb{G}_n\left(\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)-\mathrm{B}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)\right)\right\|_{\mathscr{H}\times\mathscr{L}}=o_{\mathbb{P}^*}(1).$$

Hence:

$$\begin{aligned}&\left\|\mathbb{G}_n\mathrm{B}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)+\sqrt{n}\left(\psi\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n,\mathrm{P}\right)-\psi\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0,\mathrm{P}\right)\right)\right\|_{\mathscr{H}\times\mathscr{L}}\\&\leq\ \left\|\mathbb{G}_n\left(\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)-\mathrm{B}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)\right)\right\|_{\mathscr{H}\times\mathscr{L}}+o_{\mathbb{P}^*}(1)\\&=\ o_{\mathbb{P}^*}(1).\end{aligned}$$

By the last inequality and **(H.1)** the desired result follows. □

Proof of Lemma 4.5.3.3

Mapping from the Banach space $\overline{\mathscr{R}(\dot{\psi})}$ to the Banach space $\overline{\mathrm{lin}(\Theta)}$, the sequence of continuous linear operators $\dot{\psi}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})$ converges on $\overline{\mathscr{R}(\dot{\psi})}$ to $\dot{\psi}^{-1}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)$ as $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_\Theta\to 0$ and $\left\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\right\|_\Im\to 0$ by **(H.6)**. Hence, by the Banach-Steinhaus theorem (for example, Theorem II.3.6 on page 60 of Dunford and Schwartz [1988]), the norm of the operators $\dot{\psi}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})$ is uniformly bounded:

$$\sup_{(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_\Theta\vee\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|_\Im)\leq\beta}\left\|\dot{\psi}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})\right\|\leq 1/\alpha<\infty,$$

for some positive numbers

$$0<\alpha<\infty\text{ and }\beta>0.$$

Thus for any $a\in\overline{\mathrm{lin}(\Theta)}$, we have

$$\begin{aligned}\|a\|_\Theta&=\left\|\dot{\psi}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})(\dot{\psi}(\boldsymbol{\theta},\boldsymbol{\eta})(a))\right\|_\Theta\\&\leq\left\|\dot{\psi}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})\right\|\times\|\dot{\psi}(\boldsymbol{\theta},\boldsymbol{\eta})(a)\|_{\mathscr{H}\times\mathscr{L}}\\&\leq(1/\alpha)\|\dot{\psi}(\boldsymbol{\theta},\boldsymbol{\eta})(a)\|_{\mathscr{H}\times\mathscr{L}}.\end{aligned}$$

Hence

$$\alpha\|a\| \leq \|\dot{\psi}(\boldsymbol{\theta},\boldsymbol{\eta})(a)\|_{\mathscr{H}\times\mathscr{L}}, \tag{4.6.10}$$

for all $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_\Theta \leq \beta$. Take $a = \sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$, $b = \sqrt{n}\left(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0\right)$, $\boldsymbol{\vartheta} = \hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\nu} = \hat{\boldsymbol{\eta}}_n$ in identity (4.5.3). By the linearity of the operators $\dot{\mathbb{P}}_{\boldsymbol{\vartheta}}(a)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$, $\dot{\psi}_{\boldsymbol{\eta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(b)$ and $\dot{\mathbb{P}}_{\boldsymbol{\nu}}(b)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ in a and b respectively and the definition of Fréchet differentiability of $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ as function of $(\boldsymbol{\theta},\boldsymbol{\eta})$, we have;

$$\begin{aligned}
&\dot{\psi}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\right)\\
&= -\dot{\psi}_{\boldsymbol{\eta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\right)\right) - \dot{\mathbb{P}}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\right)\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\\
&\quad -\dot{\mathbb{P}}_{\boldsymbol{\eta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\right)\right)\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\\
&= -\sqrt{n}\dot{\psi}_{\boldsymbol{\eta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\right) - \sqrt{n}\left(\dot{\mathbb{P}}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\right.\\
&\quad \left.+\dot{\mathbb{P}}_{\boldsymbol{\eta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\right)\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\right)\\
&= -\sqrt{n}\dot{\psi}_{\boldsymbol{\eta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\right) - \sqrt{n}\left(\mathbb{P}_{\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n}\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right) + \mathbb{P}_{\boldsymbol{\theta}_0,\boldsymbol{\eta}_0}\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\right.\\
&\quad \left.+ o_{\mathbb{P}^*}\left(\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|_\Theta\right) + \mathrm{O}_{\mathbb{P}^*}\left(\|\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\|_{\mathfrak{F}}^{\alpha}\right)\right)\\
&= -\sqrt{n}\dot{\psi}_{\boldsymbol{\eta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\right) - \sqrt{n}\psi\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n,\mathbb{P}\right) + o_{\mathbb{P}^*}(\sqrt{n}\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|_\Theta)\\
&\quad +\mathrm{O}_{\mathbb{P}^*}(\sqrt{n}\|\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\|_{\mathfrak{F}}^{\alpha})\\
&= -\mathbf{Z}_n + o_{\mathbb{P}^*}\left(\sqrt{n}\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|_\Theta\right) + o_{\mathbb{P}^*}(1).
\end{aligned}$$

The last equality holds by Lemmas 4.5.3.2 and **(C.5)**. Therefore, by the boundedness (4.6.10) we obtain

$$\begin{aligned}
\alpha\sqrt{n}\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|_\Theta &\leq \left\|\dot{\psi}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\right)\right\|_{\mathscr{H}\times\mathscr{L}}\\
&\leq \|\mathbf{Z}_n\|_{\mathscr{H}\times\mathscr{L}} + o_{\mathbb{P}^*}(1)\cdot\sqrt{n}\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|_\Theta + o_{\mathbb{P}^*}(1),
\end{aligned}$$

in $\mathbb{P}^*$ -probability when n is sufficiently large. The conclusion of the lemma follows from **(H.4)** which assert that the term $\|\mathbf{Z}_n\|_{\mathscr{H}\times\mathscr{L}}$ is of an order of $\mathrm{O}_{\mathbb{P}^*}(1)$. □

Proof of Theorem 4.5.3.1

By the Fréchet differentiability of $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ at $(\boldsymbol{\vartheta},\boldsymbol{\nu})$ we have

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}) - \mathbb{P}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}) - \dot{\mathbb{P}}_{\boldsymbol{\theta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(\boldsymbol{\theta}-\boldsymbol{\vartheta})\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}) - \dot{\mathbb{P}}_{\boldsymbol{\eta}}(\boldsymbol{\vartheta},\boldsymbol{\nu})(\boldsymbol{\eta}-\boldsymbol{\nu})\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})\\
= o(\|\boldsymbol{\theta}-\boldsymbol{\vartheta}\|_\Theta) + o(\|\boldsymbol{\eta}-\boldsymbol{\nu}\|_{\mathfrak{F}}).
\end{aligned}$$

Substituting $\hat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\vartheta}$, $\boldsymbol{\theta}_0$ for $\boldsymbol{\theta}$, $\hat{\boldsymbol{\eta}}_n$ for $\boldsymbol{\nu}$ and $\boldsymbol{\eta}_0$ for $\boldsymbol{\eta}$ and using $\mathbb{P}$ to denote $\mathbb{P}_{\boldsymbol{\theta}_0,\boldsymbol{\eta}_0}$, we obtain

$$\begin{aligned}
&\dot{\mathbb{P}}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\\
&\quad +\dot{\mathbb{P}}_{\boldsymbol{\eta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\right)\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\\
&= \mathbb{P}_{\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n}\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right) - \mathbb{P}_{\boldsymbol{\theta}_0,\boldsymbol{\eta}_0}\mathrm{B}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\\
&\quad +o_{\mathbb{P}^*}\left(\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|\right) + o_{\mathbb{P}^*}(\|\boldsymbol{\eta}_0-\hat{\boldsymbol{\eta}}_n\|_{\mathfrak{F}})\\
&= \psi\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n,\mathrm{P}\right) + o_{\mathbb{P}^*}\left(\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|\right)\\
&\quad +o_{\mathbb{P}^*}(\|\boldsymbol{\eta}_0-\hat{\boldsymbol{\eta}}_n\|_{\mathfrak{F}}).
\end{aligned}$$

Note that by the identity (4.5.3) we have

$$\begin{aligned}\dot{\psi}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\right) &= -\sqrt{n}\psi\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n,\mathbb{P}\right)-\sqrt{n}\dot{\psi}_{\boldsymbol{\eta}}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\right)\\ &\quad +o_{\mathbb{P}^*}\left(\sqrt{n}\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|_\Theta\right)+o_{\mathbb{P}^*}\left(\sqrt{n}\|\boldsymbol{\eta}_0-\hat{\boldsymbol{\eta}}_n\|_\Im\right)\\ &= \mathbb{G}_n\mathrm{B}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)-\sqrt{n}\dot{\psi}_{\boldsymbol{\eta}}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)\left(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\right)+o_{\mathbb{P}^*}(1).\end{aligned}$$

The last equality follows from the consistency of $(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n)$, **(H.1)** through **(H.6)**, Lemma 4.5.3.2 and Lemma 4.5.3.3. Note that by **(H.6)** the operator sequence $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})$ converges to $\dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)$ on $\overline{\mathscr{R}(\dot{\psi})}$ as $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_\Theta$ and $\left\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\right\|_\Im$ both converge to 0. Hence the Banach-Steinhaus theorem and the consistency of $(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n)$ imply that the operator norm of $\dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)$ is uniformly bounded in $\mathbb{P}^*$-probability for n is sufficiently large. It maps a term of $o_{\mathbb{P}^*}(1)$ in the $\|\cdot\|_{\mathscr{H}\times\mathscr{L}}$-norm into a term of $o_{\mathbb{P}^*}(1)$ in $\|\cdot\|_\Theta$-norm, that is

$$\dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(o_{\mathbb{P}^*}(1)\right)=o_{\mathbb{P}^*}(1).$$

This means that

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right) &= \dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)\left(\mathbb{G}_n\mathrm{B}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)-\sqrt{n}\dot{\psi}_{\boldsymbol{\eta}}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)\left(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\right)+o_{\mathbb{P}^*}(1)\right)\\ &= -\dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)(\mathbf{Z}_n)+o_{\mathbb{P}^*}(1).\end{aligned}$$

By the triangle inequality and Lemma 4.5.2.1 (applied with the K-norm $\|\cdot\|_{\mathrm{K}}$ replaced by $\|\cdot\|_\Theta$) we obtain $\dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n\right)(\mathbf{Z}_n)=\dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)(\mathbb{Z}_n)+o_{\mathbb{P}^*}(1)$. Hence we have

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\rightsquigarrow-\dot{\psi}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0\right)(\mathbf{Z}_0)$$

in $(\overline{\mathrm{lin}(\Theta)},\|\cdot\|)$ as $n\to\infty$ by the continuous mapping theorem. □

Proof of Lemma 4.5.4.1

Let $(\boldsymbol{\theta},\boldsymbol{\eta})=\sum_{i=1}^k\sum_{j=1}^k(\lambda_i\boldsymbol{\theta}_i,\lambda_j'\boldsymbol{\eta}_j)\in\mathrm{lin}(\Theta)\times\mathrm{lin}(\Im)$ be a linear combination of the $(\boldsymbol{\theta}_i,\boldsymbol{\eta}_i)$'s. We want to prove that

$$\mathrm{L}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})=\sum_{i=1}^k\sum_{j=1}^k\lambda_i\lambda_j'\mathbb{P}_{\boldsymbol{\theta}_i,\boldsymbol{\eta}_j}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}),\tag{4.6.11}$$

is a bounded bilinear extension of $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ to $\mathrm{lin}(\Theta)\times\mathrm{lin}(\Im)$. First by (4.5.5), if a linear combination of the elements $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k$ and $\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_k$ is equal to the zero element

$$\sum_{i=1}^k\sum_{j=1}^k(\lambda_i\boldsymbol{\theta}_i,\lambda_j'\boldsymbol{\eta}_j)=0,$$

then

$$\sum_{i=1}^k\sum_{j=1}^k\lambda_i\lambda_j'\mathbb{P}_{\boldsymbol{\theta}_i,\boldsymbol{\eta}_j}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})=0$$

as well. From this observation, the value of the mapping $\mathrm{L}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ is uniquely determined by $(\boldsymbol{\theta},\boldsymbol{\eta})\in\mathrm{lin}(\Theta)\times\mathrm{lin}(\Im)$. It is not hard to verify that $\mathrm{L}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ is a bilinear mapping from

$\mathrm{lin}(\Theta)\times\mathrm{lin}(\Im)$ to $\ell^\infty(\mathscr{H}\times\mathscr{L})$. The boundedness $\left\|\mathrm{L}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})\right\|_{\mathscr{H}\times\mathscr{L}}\le \mathbf{C}\|(\boldsymbol{\theta},\boldsymbol{\eta})\|$ of $\mathrm{L}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ follows from (4.5.5). And it is easy to verify that $\mathrm{L}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ is an extension of $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ to $\mathrm{lin}(\Theta)\times\mathrm{lin}(\Im)$ with $\mathrm{L}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})\equiv\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ for all $(\boldsymbol{\theta},\boldsymbol{\eta})\in\Theta\times\Im$ by (4.6.11).

For any bounded bilinear mapping $\mathrm{A}:\mathrm{lin}(\Theta)\times\mathrm{lin}(\Im)\mapsto\ell^\infty(\mathscr{H}\times\mathscr{L})$, the Fréchet derivative of A at $(\boldsymbol{\vartheta},\boldsymbol{\nu})$ is simply given by

$$d\mathrm{A}_{\boldsymbol{\vartheta},\boldsymbol{\nu}}(a,b)=\mathrm{A}(a,\boldsymbol{\nu})+\mathrm{A}(\boldsymbol{\vartheta},b).$$

Now the mapping $\mathrm{L}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}):\mathrm{lin}(\Theta)\times\mathrm{lin}(\Im)\mapsto\ell^\infty(\mathscr{H}\times\mathscr{L})$ is bounded and bilinear, hence it is Fréchet differentiable at $(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1)\in\mathrm{lin}(\Theta)\times\mathrm{lin}(\Im)$, and its derivative operator is given by for $(a,b)\in\mathrm{lin}(\Theta)\times\mathrm{lin}(\Im)$

$$\mathrm{L}_{a,\boldsymbol{\eta}_1}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})+\mathrm{L}_{\boldsymbol{\theta}_1,b}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}).$$

Since $\mathrm{L}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})=\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})$ for any $(\boldsymbol{\theta},\boldsymbol{\eta})\in\Theta\times\Im$, we have

$$\begin{aligned}\dot{\mathbb{P}}_{\boldsymbol{\theta}_1,\boldsymbol{\eta}_1}(a,b)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})&=\dot{\mathbb{P}}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1)(a)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})+\dot{\mathbb{P}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1)(b)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})\\&=\mathrm{L}_{a,\boldsymbol{\eta}_1}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})+\mathrm{L}_{\boldsymbol{\theta}_1,b}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}),\end{aligned}$$

by the uniqueness of the Fréchet derivative. Therefore, for $a=\boldsymbol{\theta}-\boldsymbol{\theta}_1$ with $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_1$ belonging to Θ and $b=\boldsymbol{\eta}-\boldsymbol{\eta}_1$ with $\boldsymbol{\eta}$ and $\boldsymbol{\eta}_1$ belonging to $\Im$, we have

$$\begin{aligned}\dot{\mathbb{P}}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1)(\boldsymbol{\theta}-\boldsymbol{\theta}_1)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})&=\mathrm{L}_{(\boldsymbol{\theta}-\boldsymbol{\theta}_1,\boldsymbol{\eta}_1)}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})\\&=\mathrm{L}_{(\boldsymbol{\theta},\boldsymbol{\eta}_1)}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})-\mathrm{L}_{(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1)}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})\\&=\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}_1}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})-\mathbb{P}_{\boldsymbol{\theta}_1,\boldsymbol{\eta}_1}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}),\\\dot{\mathbb{P}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_1,\boldsymbol{\eta}_1)(\boldsymbol{\eta}-\boldsymbol{\eta}_1)\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})&=\mathrm{L}_{(\boldsymbol{\theta}_1,\boldsymbol{\eta}-\boldsymbol{\eta}_1)}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})\\&=\mathrm{L}_{\boldsymbol{\theta}_1,\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})-\mathrm{L}_{\boldsymbol{\theta}_1,\boldsymbol{\eta}_1}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})\\&=\mathbb{P}_{\boldsymbol{\theta}_1,\boldsymbol{\eta}}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu})-\mathbb{P}_{\boldsymbol{\theta}_1,\boldsymbol{\eta}_1}\mathrm{B}(\boldsymbol{\vartheta},\boldsymbol{\nu}),\end{aligned}$$

which completes the proof of the lemma. □

Proof of Theorem 4.5.4.2

By Lemma 4.5.4.1, we take $\boldsymbol{\theta}_1=\hat{\boldsymbol{\theta}}_n$, $\boldsymbol{\eta}_1=\boldsymbol{\eta}_0$, $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ and $\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}_n$ in 4.5.6 and use $\mathbb{P}$ to denote $\mathbb{P}_{\boldsymbol{\theta}_0,\boldsymbol{\eta}_0}$, we obtain the following three equality, by using **(H.1)**, **(H.7)** and the boundedness of the score operator

$$\begin{aligned}\dot{\psi}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{\eta}_0)(\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}_n)&=-\dot{\psi}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{\eta}_0)(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0)-\mathbb{P}_{\boldsymbol{\theta}_0,\hat{\boldsymbol{\eta}}_n}\mathrm{B}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{\eta}_0)-\mathbb{P}_{\hat{\boldsymbol{\theta}}_n,\boldsymbol{\eta}_0}\mathrm{B}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{\eta}_0)\\&=-\dot{\psi}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{\eta}_0)(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0)-\mathbb{P}_{\boldsymbol{\theta}_0,\boldsymbol{\eta}_0}\mathrm{B}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{\eta}_0)+\mathrm{O}_{\mathbb{P}}(\|\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\|_{\Im}^{\alpha})\\&=-\dot{\psi}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{\eta}_0)(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0)+(\mathbb{P}_n-\mathbb{P})\left(\mathrm{B}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{\eta}_0)-\mathrm{B}(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n)\right)\\&\quad-\mathbb{P}_n\left(\mathrm{B}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{\eta}_0)-\mathrm{B}(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n)\right)-\mathbb{P}\left(\mathrm{B}(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\eta}}_n)\right)+o_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)\\&\quad+\mathrm{O}_{\mathbb{P}}(\|\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\|_{\Im}^{\alpha}).\end{aligned}$$

Hence we have

$$\begin{aligned}\dot{\psi}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0)\left(\sqrt{n}\left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_n\right)\right) &= \mathbf{Z}_n - \mathbb{G}_n\left(\mathrm{B}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0) - \mathrm{B}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_n)\right) \\ &\quad + \mathrm{O}_{\mathbb{P}}\left(n^{1/2-\alpha/2}\right) + o_{\mathbb{P}}(1) \\ &= \mathbf{Z}_n + o_{\mathbb{P}}(1).\end{aligned}$$

While the last two equalities follow from the continuity of the operator $\dot{\psi}_{\boldsymbol{\eta}}(\cdot,\cdot)$ by **(H.5)**, Lemma 4.5.3.2 and **(H.2)**. The above term $o_{\mathbb{P}^*}(1)$ denotes a term whose $\|\cdot\|_{\mathscr{H}\times\mathscr{L}}$-norm is of order $o_{\mathbb{P}^*}(1)$ since $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}, \boldsymbol{\eta})$ converges to $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)$ on $\overline{\mathscr{R}(\dot{\psi})}$, the Banach-Steinhaus theorem implies that the operator norm of $\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0)$ is uniformly bounded in $\mathbb{P}^*$-probability when n is sufficiently large. It then maps a term of $o_{\mathbb{P}^*}(1)$ in the $\|\cdot\|_{\mathscr{H}\times\mathscr{L}}$-norm into a term of $o_{\mathbb{P}^*}(1)$ in K-norm

$$\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0)\left(o_{\mathbb{P}^*}(1)\right) = o_{\mathbb{P}^*}(1).$$

This means that

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) &= -\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0)\left(\mathbf{Z}_n + o_{\mathbb{P}^*}(1)\right) \\ &= -\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0)(\mathbf{Z}_n) + o_{\mathbb{P}^*}(1).\end{aligned}$$

By Lemma 4.5.2.1

$$\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_0)(\mathbf{Z}_n) = \dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)(\mathbf{Z}_n) + o_{\mathbb{P}^*}(1).$$

Hence, by the continuous mapping theorem, we have as $n \to \infty$

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \rightsquigarrow -\dot{\psi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)(\mathbf{Z}_0) \quad \text{in} \quad \left(\overline{\mathrm{lin}(\Theta)}, \|\cdot\|_{\mathrm{K}}\right).$$

Thus the proof Theorem 4.5.4.2 is complete. □

4.7 References

Akritas, M. G. (1986). Bootstrapping the Kaplan-Meier estimator. *J. Amer. Statist. Assoc.*, **81**(396), 1032–1038. 111

Banerjee, M., Mukherjee, D., and Mishra, S. (2009). Semiparametric binary regression models under shape constraints with an application to Indian schooling data. *J. Econometrics*, **149**(2), 101–117. 99

Barbe, P. and Bertail, P. (1995). *The weighted bootstrap*, volume 98 of *Lecture Notes in Statistics*. Springer-Verlag, New York. 98, 117

Beran, R. (2003). The impact of the bootstrap on statistical algorithms and theory. *Statist. Sci.*, **18**(2), 175–184. Silver anniversary of the bootstrap. 98

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, **9**(6), 1196–1217. 98

Bickel, P. J., Götze, F., and van Zwet, W. R. (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statist. Sinica*, **7**(1), 1–31. Empirical Bayes, sequential analysis and related topics in statistics and probability (New Brunswick, NJ, 1995). 98

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1998). *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag, New York. Reprint of the 1993 original. 98

Cheng, G. (2009). Semiparametric additive isotonic regression. *J. Statist. Plann. Inference*, **139**(6), 1980–1991. 99

Cheng, G. (2015). Moment consistency of the exchangeably weighted bootstrap for semiparametric M-estimation. *Scand. J. Stat.*, **42**(3), 665–684. 116

Cheng, G. and Huang, J. Z. (2010). Bootstrap consistency for general semiparametric M-estimation. *Ann. Statist.*, **38**(5), 2884–2915. 99

Cheng, R. (2017). *Non-standard parametric statistical inference*. Oxford University Press, Oxford. 98

Chernick, M. R. (2008a). *Bootstrap methods: a guide for practitioners and researchers*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition. 98

Chernick, M. R. (2008b). *Bootstrap methods: a guide for practitioners and researchers*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition. 98

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*, volume 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. With 1 IBM-PC floppy disk (3.5 inch; HD). 98

Dunford, N. and Schwartz, J. T. (1958). *Linear operators part I: general theory*, volume 243. Interscience publishers New York. 103, 122

Dunford, N. and Schwartz, J. T. (1988). *Linear operators. Part I*. Wiley Classics Library. John Wiley & Sons, Inc., New York. General theory, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1958 original, A Wiley-Interscience Publication. 126, 131

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**(1), 1–26. 98

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method. I. *Scand. J. Statist.*, **16**(2), 97–128. With a discussion by J. A. Wellner and J. Præstgaard and a reply by the author. 109, 110, 111

Giné, E. and Zinn, J. (1990). Bootstrapping general empirical measures. *Ann. Probab.*, **18**(2), 851–869. 107

Good, P. (2005). *Permutation, parametric and bootstrap tests of hypotheses*. Springer Series in Statistics. Springer-Verlag, New York, third edition. 98

Gu, M. and Zhang, C.-H. (1993). Asymptotic properties of self-consistent estimators based on doubly censored data. *The Annals of Statistics*, pages 611–624. 102, 114, 115, 126

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York. 98

Huang, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.*, **27**(5), 1536–1563. 99

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101. 126

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, pages 221–233. Univ. California Press, Berkeley, Calif. 126

Janssen, A. (2005). Resampling Student's t-type statistics. *Ann. Inst. Statist. Math.*, **57**(3), 507–529. 116, 117

Janssen, A. and Pauls, T. (2003). How do bootstrap and permutation tests work? *Ann. Statist.*, **31**(3), 768–806. 116

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **53**, 457–481. 111

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer Series in Statistics. Springer, New York. 98

Kosorok, M. R. (2009). What's so special about semiparametric methods? *Sankhyā*, **71**(2, Ser. A), 331–353. 99

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition. 98

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition. 98

Lindsey, J. K. (1996). *Parametric statistical inference*. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York. 98

Lo, A. Y. (1993). A Bayesian method for weighted sampling. *Ann. Statist.*, **21**(4), 2138–2148. 99, 111, 116

Ma, S. and Kosorok, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *J. Multivariate Anal.*, **96**(1), 190–217. 99

Manly, B. F. J. (2007). *Randomization, bootstrap and Monte Carlo methods in biology*. Chapman & Hall/CRC Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, third edition. 98

Mason, D. M. and Newton, M. A. (1992). A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.*, **20**(3), 1611–1624. 99

Murphy, S. A. (1995). Asymptotic theory for the frailty model. *The annals of statistics*, pages 182–198. 112

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, **57**(5), 1027–1057. 126

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics*, **26**(1), 183–214. 112

Pauly, M. (2012). Consistency of the subsample bootstrap empirical process. *Statistics*, **46**(5), 621–626. 117

Pfanzagl, J. (1994). *Parametric statistical theory*. De Gruyter Textbook. Walter de Gruyter & Co., Berlin. With the assistance of R. Hamböker. 98

Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory*, pages 295–313. 126

Pollard, D. (1989). Asymptotics via empirical processes. *Statist. Sci.*, **4**(4), 341–366. With comments and a rejoinder by the author. 126

Præstgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, **21**(4), 2053–2086. 99, 103, 111, 116, 117, 127, 128, 129

Rubin, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.*, **9**(1), 130–134. 98

Shao, J. and Tu, D. S. (1995). *The jackknife and bootstrap*. Springer Series in Statistics. Springer-Verlag, New York. 117

Shao, J. and Wu, C. (1987). Heteroscedasticity-robustness of jackknife variance estimators in linear models. *The Annals of Statistics*, pages 1563–1579. 116

Tsai, W.-Y. and Zhang, C.-H. (1995). Asymptotic properties of nonparametric maximum likelihood estimator for interval-truncated data. *Scandinavian journal of statistics*, pages 361–370. 102

van der Laan, M. J. (1995). *Efficient and inefficient estimation in semiparametric models*, volume 114 of *CWI Tract*. Stichting Mathematisch Centrum, Centrum voor Wiskunde en Informatica, Amsterdam. 102

van der Vaart, A. (1994). Maximum likelihood estimation with partially censored data. *The Annals of Statistics*, pages 1896–1916. 126

van der Vaart, A. W. (1995). Efficiency of infinite-dimensional M-estimators. *Statist. Neerlandica*, **49**(1), 9–30. 100, 101, 112, 126

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. 98, 118

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics. 98, 130

van Zwet, W. R. (1979). The Edgeworth expansion for linear combinations of uniform order statistics. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics (Hradec Králové, 1978)*, pages 93–101. North-Holland, Amsterdam-New York. 117

Vardi, Y. and Zhang, C.-H. (1992). Large sample study of empirical distributions in a random-multiplicative censoring model. *The Annals of Statistics*, pages 1022–1039. 102, 126

Wellner, J. A. and Zhan, Y. (1996). Bootstrapping Z-estimators. *University of Washington Department of Statistics Technical Report*, **308**, 5. 101, 102, 107, 108, 111, 112, 114, 115

Wellner, J. A., Klaassen, C. A. J., and Ritov, Y. (2006). Semiparametric models: a review of progress since BKRW (1993). In *Frontiers in statistics*, pages 25–44. Imp. Coll. Press, London. 99

Weng, C.-S. (1989). On a second-order asymptotic property of the Bayesian bootstrap mean. *Ann. Statist.*, **17**(2), 705–710. 99, 117

Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **69**(4), 507–564. With discussion and a reply by the authors. 99

Zhan, Y. (1996). *Bootstrapping functional M-estimators*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–University of Washington. 99

Zhan, Y. (2002). Central limit theorems for functional Z-estimators. *Statist. Sinica*, **12**(2), 609–634. 97, 99, 100, 101, 102, 104, 105, 106, 107, 108, 109, 112, 115, 128, 129

Zhang, C. and Yu, T. (2008). Semiparametric detection of significant activation for brain fMRI. *Ann. Statist.*, **36**(4), 1693–1725. 99

Zheng, Z. G. and Tu, D. S. (1988). Random weighting methods in regression models. *Sci. Sinica Ser. A*, **31**(12), 1442–1459. 99

Chapter 5

Asymptotic Properties of Semiparametric M-Estimators with Multiple Change Points

Ce chapitre développe le contenu d'un article soumis, mis en forme pour être inséré dans le présent manuscrit de thèse.

Title : *Asymptotic Properties of Semiparametric* M*-Estimators with Multiple Change Points*.

abstract

Statistical models with multiple change-points are used in many fields; however, the theoretical properties of semiparametric M-estimators of such models have received relatively little attention. The main purpose of the present work is to investigate the asymptotic properties of semiparametric M-estimators with non-smooth criterion functions of the parameters of multiple change-points model for a general class of models in which the form of the distribution can change from segment to segment and in which, possibly, there are parameters that are common to all segments. Consistency of the semiparametric M-estimators of the change points is established and the rate of convergence is determined. The asymptotic normality of the semiparametric M-estimators of the parameters of the within-segment distributions is established under quite general conditions. These results, together with a generic paradigm for studying semiparametric M-estimators with multiple change-points, provide a valuable extension to previous related research on (semi)parametric maximum-likelihood estimators. For illustration, the classification with missing data in the model is investigated in detail and a short simulation result is provided.

Key words : Semiparametric inference; multiple change-points; change-point fraction; common parameter; consistency; convergence rate; M-estimators; Empirical processes; bracketing numbers.

Mathematics Subject Classification : 62F03, 62F10, 62F12, 62H12, 62H15.

5.1 Introduction and motivations

Change-point detection has become a popular tool for identifying locations in a data sequence when a stochastic system is subject to sudden external influences and is encountered in almost every field of science. Reasons behind these changes might be different and their detection helps to investigate them and properly react to them. The problem of detecting breaks in a sequence of random variables has a long history. Early work on this problem can be found in Page [1954, 1955, 1957] who investigated quality of control problems and proposed a sequential scheme for identifying changes in the mean of a sequence of independent random variables, numerous authors have worked on this problem. Despite a relatively long tradition in statistics, change point analysis is a very active field and has become increasingly popular in the last years due to its importance in many areas where data is collected over time. More precisely, methods in change point analysis have been developed to address data analytic questions in a lot of fields for example bioinformatics (recombination detection, Minin *et al.* [2005]), prediction of transmembrane helix locations Lio and Vannucci [2000], segmentation of microarray data Erdman and Emerson [2008], detection of changes in the DNA copy number Olshen *et al.* [2004], Fu and Curnow [1990] and Braun *et al.* [2000], medicine (estimation of phase transitions in pain symptoms Desmond *et al.* [2002]), climate (analysis of tropical cyclone activity Chu and Zhao [2004] and Reeves *et al.* [2007] for review), security applications (monitoring for denial-of-service attacks Wang *et al.* [2004]), and other intrusions in computer networks Tartakovsky *et al.* [2006], bio-informatics Liu *et al.* [2018], linguistics (text segmentation Choi [2000]), audio and video processing (audio segmentation Lu *et al.* [2002]), speech segmentation Shriberg *et al.* [2000], network traffic data analysis Lung-Yut-Fong *et al.* [2012], temporal video segmentation Koprinska and Carrato [2001], quality control (calibration for aircraft testing Mahmoud *et al.* [2007]), or economics and finance (identifying and dating change-points in stock market volatility Aggarwal *et al.* [1999]), in modeling and forecasting of changes in financial data Lavielle and Teyssière [2006], Spokoiny [2009], in the evolution of macroeconomic variables Bai and Perron [2003], change point detection in comparative genomics for early cancer diagnosis Lai *et al.* [2005] and in many cases such methodology has become standard. The change point is of prime importance in many learning tasks such as signal segmentation Abou-Elailah *et al.* [2015] and Kim *et al.* [2009]. Mazhar *et al.* [2018] investigated change point detection and clustering for sequences of data points. Building upon recent theoretical advances characterizing the limiting distribution-free behavior of the Wasserstein two-sample test, Cheng *et al.* [2019] proposed a novel unsupervised algorithm for distribution-free change point detection. Chen [2019] proposed an approach based on nearest neighbor information for change-point with interesting application in detecting global structural changes in social networks. The statistical community now enjoys a vast literature on change point analysis where many of the most natural and common questions have received at least some attention. For a broader presentation of the field of change-point analysis along with statistical applications and machine learning, we refer the reader to the monographs by Brodsky and Darkhovsky [1993], Carlstein *et al.* [1994], Chen and Gupta [2000], Wu [2005], Pons [2018], Tartakovsky *et al.* [2015] and Truong *et al.* [2020].

There are numerous approaches to investigate the change point problem, the reader is referred to the monograph Csörgő and Horváth [1997] for an in-depth treatment of these approaches. The problem of detecting abrupt changes has been discussed intensively in a time series context, we may refer to Jandhyala *et al.* [2013], Aue and Horváth [2013], Alvarez-Andrade and Bouzebda [2014] and Horváth and Rice [2014] for a review of the literature. The problem of detecting change-points in a sequence of random variables can be stated as follows: a sequence of random variables has a set of characteristics, such as the mean and/or the variance, that follow a piecewise constant structure. Then, the goal is to detect the number of times that these characteristics change from a set of values to another, as well as the location of the changes. Additionally, it is of interest to estimate the characteristics in each constant period. Compared to single change-point detection, multiple change-points detection is a much more challenging problem. Change point problems have been mainly focused on changes in the mean and/or the variance of univariate sequences and in the mean and/or the covariance matrix of multivariate sequences. The problem of semiparametric change-point problems have attracted considerable attention in the literature. For example, Guan [2007] considered semiparametric tests for one change-point and one epidemic alternatives models by maximum empirical likelihood method. In the semiparametric change-point regression model, Xing and Ying [2012] have developed an estimation procedure that relies on recent advances in semiparametric analysis based on counting process argument and multiple change-points inference. In Bouzebda and Keziou [2013] and Bouzebda [2014], a semiparametric maximum-likelihood-type test statistic is proposed and proved to have the same limit null distribution as the classical parametric likelihood one. Under some mild conditions, the limiting law of the proposed test statistic, suitably normalized and centralized, is shown to be double exponential, under the null hypothesis of no change in the parameter of copula models In Bouzebda and Keziou [2013], the asymptotic distribution of the proposed statistic under specified alternatives is shown to be normal, and an approximation to the power function is given. Zhang and Tian [2020] suggested the semiparametric test for the multiple change-points problems, by using the maximum empirical likelihood to get the estimations of change-points.

However, the case of general semiparametric M-estimation has been much less explored. It is worth noticing that semiparametric M-estimation was investigated in the case where the criterion function satisfies certain smoothness properties, which are not satisfied in some applications. To overcome this problem, Delsol and Van Keilegom [2020] investigated the semiparametric M-estimation in a general setting, in order to cover non-smooth M-estimators as well. The main purpose of the present work is to consider a general framework of non-smooth semiparametric M-estimators in multiple change-points models. Our paper is to provide a first full theoretical justification of the consistency of M-estimators with non-smooth criterion functions of the parameters of a general class of multiple change-points models and gives the asymptotic distribution of the parameters of the within-segment distributions by using the abstract theory of the empirical processes. This requires the effective application of large sample theory techniques, which were developed for the empirical processes.

Although the idea of our estimation approach follows that in He and Severini [2010], we allow for infinite-dimensional nuisance parameters in our estimation procedures as in Delsol and Van Keilegom [2020]. He and Severini [2010] have established asymptotic properties of the likelihood estimates for parametric models. Their results are not directly applicable here since the two-step estimation of the semiparametric model depends on some nuisance parameters, yielding to the use of different arguments in our proofs to cope with the general framework of non-smooth semi-parametric M-estimators. These results are not only useful in their own right but essential to the investigation of the present paper. In this sense, we extend the work of He and Severini [2010] to the multiple change points in the semiparametric model. The addition of the multiple change points in the model adds more extra complexity in the proofs compared to the paper of Delsol and Van Keilegom [2020].

The layout of the paper is organized as follows. Section 5.2 introduces the proposed estimation procedure, notation and definition needed to state our main results. Section 5.3 derives the asymptotic properties of the non-smooth semi-parametric M-estimators including the consistency with rate and the asymptotic distribution. The finite sample performance of the proposed procedure is illustrated by means of Monte Carlo simulations in Section 5.4. Finally, Section 7.3 provides some conclusions. To avoid interrupting the flow of the presentation, all mathematical developments are relegated to Section 5.5.

5.2 Notation and definitions

During the whole of the paper, we suppose that the data $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent random vectors. The set $\boldsymbol{\Upsilon} \times \boldsymbol{\Theta}$ denotes a parameter set (usually but not necessarily of finite dimension) and $\mathscr{H}$ denotes an infinite-dimensional parameter set. Suppose that there exists a non-random measurable real-valued function $\mathrm{M} : \boldsymbol{\Upsilon} \times \boldsymbol{\Theta} \times \mathscr{H} \longrightarrow \mathbb{R}$, such that

$$(\boldsymbol{\alpha}^0, \boldsymbol{\theta}^0) = \underset{(\boldsymbol{\alpha},\boldsymbol{\theta}) \in \boldsymbol{\Upsilon} \times \boldsymbol{\Theta}}{\operatorname{argmax}} \mathrm{M}(\cdot, \boldsymbol{\alpha}, \boldsymbol{\theta}, h(\cdot, \boldsymbol{\alpha})),$$

and suppose $(\boldsymbol{\alpha}^0, \boldsymbol{\theta}^0)$ is unique and belongs to the interior of $\boldsymbol{\Upsilon} \times \boldsymbol{\Theta}$. Let $(\boldsymbol{\alpha}^0, \boldsymbol{\theta}^0)$ and $h_0 \in \mathscr{H}$ be the true unknown finite- and infinite-dimensional parameters. We allow that the functions $h \in \mathscr{H}$ depend on the parameters $\boldsymbol{\alpha}$ and the vector $\mathbf{X}$, But we will always suppress this dependency for notational convenience when no misunderstanding is possible. We also use, for example , the following abbreviated notation:

$$(\boldsymbol{\alpha}, \boldsymbol{\theta}, h) := (\boldsymbol{\alpha}, \boldsymbol{\theta}, h(\cdot, \boldsymbol{\alpha})), (\boldsymbol{\alpha}, \boldsymbol{\theta}, h_0) := (\boldsymbol{\alpha}, \boldsymbol{\theta}, h_0(\cdot, \boldsymbol{\alpha})), \text{ and } (\boldsymbol{\alpha}^0, \boldsymbol{\theta}^0, h_0) := (\boldsymbol{\alpha}^0, \boldsymbol{\theta}^0, h_0(\cdot, \boldsymbol{\alpha}^0)).$$

We suppose the sets $\boldsymbol{\Theta}$, $\boldsymbol{\Upsilon}$ and $\mathscr{H}$ are metric spaces and we denote their metrics by d_1, d_2 and $d_{\mathscr{H}}$, respectively. Since the nuisance parameter is permitted to depend on $\boldsymbol{\alpha}$, by implication we define $d_{\mathscr{H}}(h, h_0)$ uniformly over $\boldsymbol{\alpha}$, i.e.,

$$d_{\mathscr{H}}(h, h_0) := \sup_{\boldsymbol{\alpha} \in \boldsymbol{\Upsilon}} d^1_{\mathscr{H}}(h(\cdot, \boldsymbol{\alpha}), h_0(\cdot, \boldsymbol{\alpha})),$$

for some metric $d^1_{\mathscr{H}}(\cdot,\cdot)$. Assume that there exists unknown change points $n_1,\ldots,n_k$,

$$0 = n_0 < n_1 < n_2 < \cdots < n_k < n_{k+1} = n,$$

such that, for each $j = 1,\ldots,k+1$, $\mathbf{X}_{n_{j-1}+1},\ldots,\mathbf{X}_{n_j}$ are identically distributed with a distribution that depends on j. The following notation will be used

$$\begin{aligned}
\lambda_j &= \frac{n_j}{n}, \text{ for any } j = 1,\ldots,k,\\
\lambda^0_j &= \frac{n^0_j}{n}, \text{ for any } j = 1,\ldots,k,\\
\boldsymbol{\lambda} &= (\lambda_1,\lambda_2,\ldots,\lambda_k),\\
\boldsymbol{\lambda}^0 &= (\lambda^0_1,\lambda^0_2,\ldots,\lambda^0_k),
\end{aligned}$$

where $n^0_j, j = 1,\ldots,k$ are the true change-points locations. Note that $\boldsymbol{\lambda}^0$ is taken to be a constant vector as n goes to infinity, which is a common assumption in the literature, see for example He and Severini [2010] and Zou *et al.* [2014].

Suppose that there exists a random real-valued function

$\mathrm{M}_n : \boldsymbol{\Upsilon} \times \prod_{j=1}^{k+1}\boldsymbol{\Theta}_j \times \mathscr{H} \longrightarrow \mathbb{R}$ depending on the data $\mathbf{X}_1,\ldots,\mathbf{X}_n$, such that $\mathrm{M}_n(\boldsymbol{\alpha},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{k+1},\boldsymbol{\lambda},h_0)$ is an approximation of $\mathrm{M}(\boldsymbol{\alpha},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{k+1},h_0)$. In many situations, we have that

$$\mathrm{M}(\boldsymbol{\alpha},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{k+1},\boldsymbol{\lambda},h) = \sum_{j=1}^{k+1}\left(\frac{n_j - n_{j-1}}{n}\right)\mathbb{E}[\mathbf{m}_j(\mathbf{X}_{n_j},\boldsymbol{\alpha},\boldsymbol{\theta}_j,h)],$$

and

$$\mathrm{M}_n(\boldsymbol{\alpha},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{k+1},\boldsymbol{\lambda},h) = \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}+1}^{n_j}\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h),$$

where $\mathbf{m}_j(\cdot)$ are a measurable real-valued functions for any $1 \le j \le k+1$ such that

$$(\boldsymbol{\alpha}^0,\boldsymbol{\theta}^0_1,\ldots,\boldsymbol{\theta}^0_{k+1},n^0_1,\ldots,n^0_k) = \underset{0<n_1<n_2<\cdots<n;\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j,1\le j\le k+1,\boldsymbol{\alpha}\in\boldsymbol{\Upsilon}}{\operatorname{argmax}} \mathrm{M}(\boldsymbol{\alpha},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{k+1},\boldsymbol{\lambda},h_0).$$

Suppose that for each $\boldsymbol{\alpha}$ there is an initial non-parametric estimator $\widehat{h}(\cdot,\boldsymbol{\alpha})$ for $h_0(\cdot,\boldsymbol{\alpha})$. This nonparametric estimator depends on the model in question and can be based on, e.g., kernels, splines or neural networks. Again, for notational simplicity, we let $(\boldsymbol{\alpha},\widehat{h}) = (\boldsymbol{\alpha},\widehat{h}(\cdot,\boldsymbol{\alpha}))$. We have to estimate the unknown parameter $(\boldsymbol{\alpha}^0,\boldsymbol{\theta}^0_1,\ldots,\boldsymbol{\theta}^0_{k+1},n^0_1,\ldots,n^0_k)$ by any $\left(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\theta}}_1,\ldots,\widehat{\boldsymbol{\theta}}_{k+1},\widehat{n}_1,\ldots,\widehat{n}_k\right)$ that "approximately solves" the following sample maximization problem

$$(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\theta}}_1,\ldots,\widehat{\boldsymbol{\theta}}_{k+1},\widehat{n}_1,\ldots,\widehat{n}_k) = \underset{0<n_1<n_2<\cdots<n;\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j,1\le j\le k+1,\boldsymbol{\alpha}\in\boldsymbol{\Upsilon}}{\operatorname{argmax}} \mathrm{M}_n(\boldsymbol{\alpha},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{k+1},\boldsymbol{\lambda},\widehat{h}).$$

In the set of conditions given in the next sections, we will formalize what we mean with "approximate solution". Let us introduce the following notation

$$\begin{aligned}
\boldsymbol{\phi}_j &= (\boldsymbol{\alpha},\boldsymbol{\theta}_j) \text{ for any } j = 1,\ldots,k,\\
\boldsymbol{\phi}^0_j &= (\boldsymbol{\alpha}^0,\boldsymbol{\theta}^0_j) \text{ for any } j = 1,\ldots,k,\\
\boldsymbol{\phi} &= (\boldsymbol{\alpha},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_{k+1}),\\
\boldsymbol{\phi}^0 &= (\boldsymbol{\alpha}^0,\boldsymbol{\theta}^0_1,\boldsymbol{\theta}^0_2,\ldots,\boldsymbol{\theta}^0_{k+1}).
\end{aligned}$$

Define

$$\mathbf{M}_n(\boldsymbol{\phi}_j, \lambda_j^0, h) = \frac{1}{n} \sum_{i=n_{j-1}^0+1}^{n_j^0} \mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h),$$
$$\mathbf{M}(\boldsymbol{\phi}_j, \lambda_j^0, h) = (\lambda_j^0 - \lambda_{j-1}^0)\mathbb{E}[\mathbf{m}_j(\mathbf{X}_{n_j^0}, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h)].$$

The proof of the consistency of our estimators is based in the following approach

$$\begin{aligned}
\mathfrak{W} &= \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}+1}^{n_j}\left[\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, \widehat{h}) - \mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0)\right] \\
&\quad + \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}+1}^{n_j}\left[\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0))\right] \\
&\quad - \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}^0+1}^{n_j^0}\left[\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}^0, \boldsymbol{\theta}_j^0, h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}^0, \boldsymbol{\theta}_j^0, h_0))\right] \\
&\quad + \sum_{j=1}^{k+1}(\lambda_j - \lambda_{j-1})\mathbb{E}[\mathbf{m}_j(\mathbf{X}_{n_j}, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0)] \\
&\quad - \sum_{j=1}^{k+1}(\lambda_j^0 - \lambda_{j-1}^0)\mathbb{E}[\mathbf{m}_j(\mathbf{X}_{n_j^0}, \boldsymbol{\alpha}^0, \boldsymbol{\theta}_j^0, h_0)] \\
&= \mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}, \widehat{h}) - \mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}, h_0) + \mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}, h_0) - \mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}, h_0) - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h_0) \\
&\quad + \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h_0) + \mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}, h_0) - \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h_0). \qquad (5.2.1)
\end{aligned}$$

We obviously have that

$$\underset{0<n_1<n_2<\cdots<n;\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j,1\le j\le k+1,\boldsymbol{\alpha}\in\boldsymbol{\Upsilon}}{\operatorname{argmax}} \mathfrak{W} = \underset{0<n_1<n_2<\cdots<n;\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j,1\le j\le k+1,\boldsymbol{\alpha}\in\boldsymbol{\Upsilon}}{\operatorname{argmax}} \mathbf{M}_n(\boldsymbol{\alpha}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k+1}, \boldsymbol{\lambda}, \widehat{h}).$$

Let us introduce

$$\begin{aligned}
\mathbf{U} &= \mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}, \widehat{h}) - \mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}, h_0) \\
&= \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}+1}^{n_j}\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, \widehat{h}) - \mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0). \qquad (5.2.2) \\
\mathbf{L} &= \mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}, h_0) - \mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}, h_0) - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h_0) + \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h_0) \\
&= \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}+1}^{n_j}\left[\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0))\right] \\
&\quad - \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}^0+1}^{n_j^0}\left[\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}^0, \boldsymbol{\theta}_j^0, h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}^0, \boldsymbol{\theta}_j^0, h_0))\right].
\end{aligned}$$

Alternatively, we may write

$$\mathbf{L} = \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=1}^{k+1}\Bigg\{\sum_{t\in\widetilde{n}_{ji}}\left[\mathbf{m}_j(\mathbf{X}_t,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0)-\mathbb{E}(\mathbf{m}_j(\mathbf{X}_t,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0))\right] - \sum_{t\in\widetilde{n}_{ji}}\left[m_i(\mathbf{X}_t,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_i^0,h_0)-\mathbb{E}(m_i(\mathbf{X}_t,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_i^0,h_0))\right]\Bigg\}, \tag{5.2.3}$$

where

$$\widetilde{n}_{ji} = [n_{j-1}+1, n_j]\cap[n_{i-1}^0+1, n_i^0] \text{ for } i,j = 1,2,\ldots,k+1.$$

Note that when the families

$$\mathscr{F}_j = \{\mathbf{m}_j(\cdot,\boldsymbol{\phi}_j,h), \boldsymbol{\phi}_j\in\boldsymbol{\Phi}_j, h\in\mathscr{H}\}$$

are Glivenko-Cantelli for each $j = 1,2,\ldots,k+1$, we have L approaches 0 as $n\longrightarrow 0$ and

$$\mathbf{M}(\boldsymbol{\phi},\boldsymbol{\lambda},h_0)-\mathbf{M}(\boldsymbol{\phi}^0,\boldsymbol{\lambda}^0,h_0) = \sum_{j=1}^{k+1}\sum_{i=1}^{k+1}\frac{n_{ji}}{n}\int_{\mathbb{R}}[\mathbf{m}_j(x,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0)-m_i(x,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_i^0,h_0)]\,d\mathrm{F}_{n_i^0}(x). \tag{5.2.4}$$

where n_{ji} is the number of observations of the interested variables in the set $[n_{j-1}+1,n_j]\cap[n_{i-1}^0+1,n_i^0]$, for $i,j = 1,\ldots,k+1$, and $\mathrm{F}_{n_i^0}(\cdot)$ is the true function of distribution for each subsample $\mathbf{X}_{n_{i-1}^0+1},\ldots,\mathbf{X}_{n_i^0}$. All through of the paper, we use $\mathrm{O}_{\mathbb{P}}(1)$ and $o_{\mathbb{P}}(1)$ notation of Mann and Wald [1943], as exposed in Chernoff [1956], where $\mathbb{P}$ is the joint probability defined on product spaces. When applied to vectors and matrices, the symbols should be interpreted entry by entry.

5.3 Main results

5.3.1 Consistency

In this section, we consider the consistency of the M-estimators that can be achieved by the argmax theorem in van der Vaart and Wellner [1996]. Let us recall the basic idea. If the argmax functional is continuous with respect to some metric on the space of the criterion functions, then convergence in distribution of the criterion functions will imply the convergence in distribution of their points of maximum, the M-estimators, to the maximum of the limit criterion function. So in this section we will introduce the set of sufficient assumptions which guarantee the weak consistency of the estimators $\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\theta}}_1,\ldots,\widehat{\boldsymbol{\theta}}_{k+1},\widehat{\lambda}_1,\ldots,\widehat{\lambda}_k$, which it will be considered as an initial step for the next subsequent sections, where we will treat the rate of convergence and the asymptotic distribution of the estimators $\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\theta}}_1,\ldots,\widehat{\boldsymbol{\theta}}_{k+1}$. The proof of the asymptotic distribution of $\widehat{\lambda}_1,\ldots,\widehat{\lambda}_k$, should require a complex methodology, and we leave this problem open for future research.

Without loss of generality and unlike to the work of Delsol and Van Keilegom [2020], we assume our functions and estimators are measurable so we don't use the terminology of outer expectation and probability, see Pakes and Pollard [1989].

In our analysis, we consider the following assumptions.

(A1) $\widehat{\boldsymbol{\phi}} \in \boldsymbol{\Phi}$, $\widehat{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}$ and $\mathbf{M}_n(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\lambda}}, \widehat{h}) > \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h}) + o_{\mathbb{P}}(1)$.

(A2) For all $\epsilon > 0$, there exist a $\delta > 0$ such that $d(\boldsymbol{\phi}, \boldsymbol{\phi}^0) > \epsilon$ or $\left\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\right\|_\infty > \epsilon$ implies

$$\mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h_0) > \mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}, h_0) + \delta.$$

(A3) It is assumed that for $j = 1, \ldots, k+1$,

$$m_{j+1}(\cdot, \boldsymbol{\alpha}^0, \boldsymbol{\theta}^0_{j+1}, h_0) \neq \mathbf{m}_j(\cdot, \boldsymbol{\alpha}^0, \boldsymbol{\theta}^0_j, h_0)$$

on a set of non-zero measure.

(A4) $\mathbb{P}(\widehat{h} \in \mathscr{H}) \longrightarrow 1$ as $n \longrightarrow \infty$ and $d_{\mathscr{H}}(\widehat{h}, h_0) \xrightarrow{\mathbb{P}} 0$.

(A5) For any $j = 1, 2, \ldots, k+1$ and any integers s, t satisfying $0 \leq s < t \leq n$,

$$\mathbb{E}\left[\max_{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j, \boldsymbol{\alpha} \in \boldsymbol{\Upsilon}} \left(\sum_{i=s+1}^{t} \mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0) - \mathbb{E}[\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0)]\right)^2\right] \leq \mathrm{A}(t-s)^r,$$

where $r < 2$ and A is a constant.

(A6) There exist function $\mathfrak{G}(\cdot)$ and such that for any h in the neighborhood of h_0, any $j = 1, 2, \ldots, k+1$ and any $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j$, $\boldsymbol{\alpha} \in \boldsymbol{\Upsilon}$ we have:

$$|\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h) - \mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0)| \leq \mathfrak{G}(\mathbf{X}_i) d_{\mathscr{H}}(h, h_0).$$

The function $\mathfrak{G}(\cdot)$ satisfies for any $i = 1, 2, \ldots, k+1$,

$$\int \mathfrak{G}^2(x) d\mathrm{F}_{n_i^0}(x) < \infty.$$

We state now our fist result.

Theorem 5.3.1.1 (Consistency) *Under assumptions (A1)-(A6), we have*

$$\widehat{\lambda}_i \xrightarrow{\mathbb{P}} \lambda_i^0, \widehat{\boldsymbol{\theta}}_j \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_j^0 \text{ and } \widehat{\boldsymbol{\alpha}} \xrightarrow{\mathbb{P}} \boldsymbol{\alpha}^0,$$

where $\widehat{\lambda}_i = \frac{\widehat{n}_i}{n}$ *for* $i = 1, \ldots, k$ *and* $j = 1, \ldots, k+1$.

Note that if we are in the situation of the estimator of maximum likelihood in parametric models, i.e.

$$\mathbf{m}_j(x, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0) = \log f_j(x, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0),$$

where $f_j(\cdot, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0)$ is the density function with known true function h_0, Theorem 5.3.1.1 reduces to the Theorem 2.1 of He and Severini [2010].

Remark 5.3.1.2 *Let us discuss in this remark the imposed conditions to highlight their generality.*

(i) *Assumption* ***(A1)*** *is trivially fulfilled if*

$$\mathbf{M}_n(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\lambda}}, \widehat{h}) \geq \sup_{\boldsymbol{\phi}\in\boldsymbol{\Phi}, \lambda\in\boldsymbol{\Lambda}} \mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}, \widehat{h}) + o_{\mathbb{P}}(1).$$

(ii) *Assumption* ***(A3)*** *guarantees that the distributions in two neighboring segments are different; clearly, this is required for the change points to be well defined.*

(iii) *Assumption* ***(A5)*** *is technical requirements on the behavior of the function* $\mathbf{m}_j(\cdot)$ *between and within segments, respectively. This condition is used to ensure that the information regarding the within- and between-segment parameters grows quickly enough to establish the consistency and the rate of convergence of the parameters estimators. Note that where*

$$\mathbf{m}_j(\cdot, \boldsymbol{\phi}_j, h_0) = \log f_j(\cdot, \boldsymbol{\phi}_j, h_0)$$

these conditions are relatively weak; it is easy to check that they are satisfied by at least all distributions in the exponential family, for detail see He and Severini [2010], Lavielle [1999] and Lavielle and Ludeña [2000].

(iv) *Assumption* ***(A6)*** *is automatically fulfilled when : for any* $j = 1, 2, \ldots, k+1$ *the function* $\mathbf{m}_j(\cdot)$ *is continuously differentiable in* $h = h(\cdot)$*, we note its derivative by*

$$\begin{aligned} \mathrm{G}_j(x, \boldsymbol{\alpha}, h) &= \frac{\partial \mathbf{m}_j(x, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h)}{\partial h}|_{h=h(x,\boldsymbol{\alpha})}, \\ \mathrm{G}_j(x, \boldsymbol{\alpha}) &= \mathrm{G}_j(x, \boldsymbol{\alpha}, h_0), \end{aligned}$$

and assume that the function $\mathrm{G}_j(\cdot)$ *exists a.e. Also assume that there is envelopes* $\mathfrak{G}_j(\cdot)$ *with the property that*

$$|\mathrm{G}_j(x, \boldsymbol{\alpha}, h)| \leq \mathfrak{G}_j(x),$$

and

$$\int \mathfrak{G}_j^2(x) d\mathrm{F}_{n_i^0}(x) \leq \mathscr{K} \text{ for some } \mathscr{K} < \infty \text{ for any } j, i = 1, \ldots, k+1.$$

So we have for any h *in the neighborhood of* h_0*:*

$$\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, \widehat{h}) - \mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0) = \mathrm{G}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \overline{h}(\mathbf{X}_i, \boldsymbol{\alpha}))(\widehat{h}(\mathbf{X}_i, \boldsymbol{\alpha}) - h_0(\mathbf{X}_i, \boldsymbol{\alpha})),$$

where $\overline{h}(x, \boldsymbol{\alpha}) \in [\widehat{h}(x, \boldsymbol{\alpha}), h_0(x, \boldsymbol{\alpha})]$*. Under this conditions we obtain that:*

$$|\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, \widehat{h}) - \mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0)| \leq \mathfrak{G}_j(\mathbf{X}_i) d_{\mathscr{H}}(\widehat{h}, h_0).$$

We can choose for example

$$\mathfrak{G}(\cdot) = \max_{1\leq j\leq k+1} \mathfrak{G}_j(\cdot).$$

5.3.2 Rate of convergence

In the present section, we will consider the rate of convergence of the considered estimators. Generally speaking, the basic tool in establishing the rate of convergence for an M-estimator is control of the modulus of continuity of the empirical criterion function using entropy integrals over the parameter sets, one can refer at this point to the books of van der Vaart and Wellner [1996], van de Geer [2000] and Kosorok [2008]. To establish the convergence rate, we need to assume their consistency, such a result can be held by using the Theorem 5.3.1.1. Theorem 5.3.2.1 provides the rate of convergence for the estimators of change points. Theorem 5.3.2.3 gives the rate of convergence of the parameters of the within-segment distributions. The fact that the rate of convergence of $\widehat{\boldsymbol{\lambda}}$ to $\boldsymbol{\lambda}^0$ is faster than the rate convergence given in Theorem 5.3.2.3, will be instrumental for the results in the following section.
We introduce the following assumptions.

(B1) For any $j = 1, \ldots, k+1$, any $\boldsymbol{\alpha}$, $\boldsymbol{\theta}_j$; for $i = 1, \ldots, k+1$,

$$\int_{\mathbb{R}} \mathbf{m}_j(x, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0) d\mathrm{F}_{n_i^0}(x) \leq \int_{\mathbb{R}} \mathbf{m}_i(x, \boldsymbol{\alpha}^0, \boldsymbol{\theta}_i^0, h_0) d\mathrm{F}_{n_i^0}(x).$$

We can check under condition **(B1)**, there exist $\mathrm{C}_1 > 0$ for any $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ such that

$$\mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}, h_0) - \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h_0) \leq -\mathrm{C}_1 \left\| \boldsymbol{\lambda} - \boldsymbol{\lambda}^0 \right\|_\infty, \tag{5.3.1}$$

where

$$\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\|_\infty = \max_{1 \leq j \leq k} |\lambda_j - \lambda_j^0|.$$

Theorem 5.3.2.1 *Under assumptions **(A3)-(A6)** and **(B1)**, we have*

$$\lim_{\eta \longrightarrow \infty} \lim_{n \longrightarrow \infty} \mathbb{P}\left(n \left\| \widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 \right\|_\infty \geq \eta\right) = 0,$$

where

$$\widehat{\lambda} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_k), \left\| \widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 \right\|_\infty = \max_{1 \leq j \leq k} \left| \widehat{\lambda}_j - \lambda_j^0 \right|.$$

That is, for $i = 1, 2, \ldots, k$,

$$\widehat{\lambda}_i - \lambda_i^0 = \mathrm{O}_{\mathbb{P}}\left(n^{-1}\right).$$

Once more, we stress the fact Theorem 5.3.2.1 extends and complements Theorem 2.2 of He and Severini [2010], by including the estimator of maximum likelihood, in parametric models, as a particular case.

Remark 5.3.2.2 *Assumption **(B1)** is to ensure that the expectation of the function associates with the true parameters is the maximum in the true sample, when we consider the particular case $\mathbf{m}_j(\cdot, \boldsymbol{\phi}_j, h) = \log f_j(\cdot, \boldsymbol{\phi}_j, h)$, this assumption comes directly from the distance of Kullback-Leibler, for further details, we refer to He and Severini [2010], or when the function $m(\cdot, \cdot, \cdot)$*

is independent of the index j, i.e., the same function of all segments for example when the variables are assumed to be from normal distribution and there is a change in variances and having the same mean, or conversely, so we have all parameters are in the same set, i.e., $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}$ for any $j = 1, 2, \ldots, k+1$. Another example is that the variables are assumed to follow the Weibull's distribution. In the M*-estimation theory, this condition is required to ensure that the true parameters are the points that maximize the criterion function. For more details, see also van der Vaart and Wellner [1996].*

In the following theorem, we give the rate of convergence of the parameters of the within-segment, we give the general conditions extending those in Delsol and Van Keilegom [2020] to cope with the general setting of multiple change point problems.

(B2) $d(\widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}^0) \xrightarrow{\mathbb{P}} 0$ and $v_n d_{\mathscr{H}}(\widehat{h}, h_0) = \mathrm{O}_{\mathbb{P}}(1)$ for some $v_n \longrightarrow \infty$.

(B3) For all $\delta_1 > 0$, there exist $\boldsymbol{\alpha} < 2$, $\mathrm{K} > 0$, $\delta_0 > 0$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ there exists a function ψ_n for which $\delta \mapsto \frac{\psi_n(\delta)}{\delta^{\boldsymbol{\alpha}}}$ is decreasing on $(0, \delta_0]$ and for all $\delta \leq \delta_0$,

$$\mathbb{E}\left[\sup_{d(\boldsymbol{\phi},\boldsymbol{\phi}^0)\leq\delta, d_{\mathscr{H}}(h,h_0)\leq\frac{\delta_1}{\sqrt{n}}} |\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, h) - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h) - \mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, h) + \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h)|\right] \leq \mathrm{K}\frac{\psi_n(\delta)}{\sqrt{n}}.$$

(B4) There exists a constant $\mathrm{C} > 0$, a sequence $r_n \longrightarrow \infty$, and variables $\mathrm{W}_n = \mathrm{O}_{\mathbb{P}}(r_n^{-1})$ and $\beta_n = o_{\mathbb{P}}(1)$, such that for all $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ satisfying $d(\boldsymbol{\phi}, \boldsymbol{\phi}^0) \leq \delta_0$:

$$\mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) - \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h}) \leq \mathrm{W}_n d(\boldsymbol{\phi}, \boldsymbol{\phi}^0) - \mathrm{C} d(\boldsymbol{\phi}, \boldsymbol{\phi}^0)^2 + \beta_n d(\boldsymbol{\phi}, \boldsymbol{\phi}^0)^2.$$

(B5) We have

$$\mathbf{M}_n(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\lambda}}, \widehat{h}) \geq \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h}) + \mathrm{O}_{\mathbb{P}}(r_n^{-2}),$$

$$r_n^2 \psi_n(r_n^{-1}) \leq \sqrt{n} \quad \text{and} \quad r_n^2 = o(n).$$

Under these conditions and after giving the rate of convergence of the estimators of change points fractions, we will prove the r_n^{-1}-consistent of the estimator $\widehat{\boldsymbol{\phi}}$ like in the i.i.d. case. Hence, the sequence r_n plays an important role in the above assumptions and should be chosen in the sharpest possible way. Before giving the theorem and its proof, we discuss these assumptions in more detail and clarify that they have the ability to hold even when there is a change in the distribution.

In the following theorem we provide the rate of convergence of $\widehat{\boldsymbol{\phi}}$ to $\boldsymbol{\phi}^0$.

Theorem 5.3.2.3 *Under conditions **(B1)-(B5)**, we have, as $n \to \infty$,*

$$r_n d(\widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}^0) = \mathrm{O}_{\mathbb{P}}(1).$$

Remark 5.3.2.4 *(i) Assumption* ***(B2)*** *is a "high-level" assumption. Many asymptotic results make it possible to get those conditions on both, the* M*-estimator* $\widehat{\boldsymbol{\phi}}$ *and the nuisance estimator* $\widehat{h}$*.In general, the nuisance estimator 's convergence rate is slower than the best convergence rate of the* M*-estimator. We are interested in researching instances where the convergence rate of the* M*-estimator is not influenced by the fact that the nuisance parameter needs to be calculated.*

(ii) Assumption ***(B3)*** *is a "high-level" assumption, and it fulfilled if we impose this condition in each of the true sub-sample* $\mathbf{X}_{n_{j-1}^0+1},\ldots,\mathbf{X}_{n_j^0}$ *for each* $j=1,\ldots,k+1$*. For this end, we assume for each* j *that for any* z *the function* $(\boldsymbol{\alpha},\boldsymbol{\theta}_j,h)\mapsto \mathbf{m}_j(z,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h(z,\boldsymbol{\alpha}))-\mathbf{m}_j(z,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_j^0,h(z,\boldsymbol{\alpha}^0))$ *is bounded on open neighborhood of* $(\boldsymbol{\phi}_j^0,h_0)$*, i.e., on*

$$\{(\boldsymbol{\phi}_j,h): d(\boldsymbol{\phi}_j,\boldsymbol{\phi}_j^0)\leq \delta_0, d_{\mathscr{H}}(h,h_0)\leq \delta_1^{'}\}$$

for some $\delta_0,\delta_1^{'}>0$*. Let us consider for each* j *the class*

$$\mathscr{F}^j_{\delta,\delta_1^{'}}=\{\mathbf{m}_j(\cdot,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h(\cdot,\boldsymbol{\alpha}))-\mathbf{m}_j(\cdot,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_j^0,h(\cdot,\boldsymbol{\alpha}^0)), d(\boldsymbol{\phi}_j,\boldsymbol{\phi}_j^0)\leq\delta, d_{\mathscr{H}}(h,h_0)\leq\delta_1^{'}\}$$

for any $\delta\leq\delta_0$ *and denote its envelope by* $\mathscr{M}^j_{\delta,\delta_1^{'}}$*. For any* δ_1*, we have* $\frac{\delta_1}{\upsilon_n}\leq\delta_1^{'}$ *for* n *large enough. Let us recall the definition of the bracketing numbers. For any borel measurable function* $f:\mathscr{S}\mapsto\mathbb{R}$*, we define the bracket* $[\![f^-,f^+]\!]$*, between two Borel functions* f^- *and* f^+*, to be the set of Borel functions* f *fulfilling* $f^-\prec f\prec f^+$*, the symbol* $\prec$ *standing for the everywhere pointwise comparison between real functions on* $\mathscr{S}$*. Denoting by* $\mathrm{N}_{[\,]}(\epsilon,\mathscr{F},\|\cdot\|_{\mathbb{Q},2})$ *the minimal number of brackets with* $\|\cdot\|_{\mathbb{Q},2}$ *diameter less than* ϵ *needed to cover* $\mathscr{F}$*, refer to Definition 2.1.6 of van der Vaart and Wellner [1996].) Then under entropy conditions on* $\mathscr{F}^j_{\delta,\delta_1^{'}}$*; we get,*

$$\sup_{\delta\leq\delta_0}\int_0^1\sqrt{1+\log \mathrm{N}_{[\,]}\left(\epsilon\left\|\mathrm{M}^j_{\delta,\delta_1^{'}}\right\|\mathbb{L}_2(\mathbb{P}^*),\mathscr{F}^j_{\delta,\delta_1^{'}},\mathbb{L}_2(\mathbb{P})\right)}d\epsilon<+\infty, \tag{5.3.2}$$

there exist $\mathrm{K}_1>0$ *independent of* δ *such a way that for all* $\delta\leq\delta_0$*, we have*

$$\mathbb{E}\left[\sup_{d(\boldsymbol{\phi},\boldsymbol{\phi}^0)\leq\delta, d_{\mathscr{H}}(h,h_0)\leq\delta_1^{'}}|\mathbf{M}_n(\boldsymbol{\phi}_j,\lambda_j^0,h)-\mathbf{M}_n(\boldsymbol{\phi}_j^0,\lambda_j^0,h)-\mathbf{M}(\boldsymbol{\phi}_j,\lambda_j^0,h)+\mathbf{M}(\boldsymbol{\phi}_j^0,\lambda_j^0,h)|\right]$$
$$\leq \mathrm{K}_1\frac{\sqrt{\mathbb{E}[\mathscr{M}^j_{\delta,\delta_1^{'}}]^2}}{\sqrt{n}},$$

see Theorems 2.14.1 and 2.14.2 in van der Vaart and Wellner [1996]. Then the last part of ***(B3)*** *holds if* $\psi_n(\delta)$ *can be chosen such that*

$$\exists \mathrm{K}_0,\forall\delta\leq\delta_0:\sqrt{\mathbb{E}[\mathscr{M}^j_{\delta,\delta_1^{'}}]^2}\leq \mathrm{K}_0\psi_n(\delta). \tag{5.3.3}$$

*For more details of this assumption, entropy condition and the case of the function $\psi_n(\cdot)$ whose give us the different expression of the rate r_n, we refer the reader to Delsol and Van Keilegom [2020], all the different rate of convergence r_n in the literature for smooth or not smooth function satisfied the last term in assumption **(B3)**. Note that the function $\psi_n(\cdot)$ can not be the same as all true sub-sample for this in the last expression 5.3.3 we will have k functions $\psi_n^j(\cdot)$ for this case we can take*

$$\psi_n(\cdot) = \sum_{j=1}^{k} \psi_n^j(\cdot), \quad or \quad \psi_n(\cdot) = \max_{1\le j\le k} \psi_n^j(\cdot),$$

the same think with the rate r_n maybe there are a k rates of convergence r_n^j, for showing this theorem and the weak convergence in the next section we need to ensure that: all this different rates are equivalent sequences, or more generally

$$\max_{1\le j\le k} \frac{r_n^j}{r_n^i} \longrightarrow a_i,$$

where $a_i \in]0,1]$ for any $i = 1,\dots,k$. So in this case we can take $r_n = \max_{1\le j\le k} r_n^j$, or one of these rates.

*(iii) With the same argument in the previous remark assumption **(B4)** is implied when the following conditions hold for every true sub-sample:*

(a) $\boldsymbol{\Upsilon} \times \boldsymbol{\Theta}_j \subset \mathbb{R}^{d+d_j}$ for some integers d, d_j for $j = 1,\dots,k+1$ and

$$d(\boldsymbol{\phi}_j, \boldsymbol{\phi}_j^0) = \|\boldsymbol{\phi}_j - \boldsymbol{\phi}_j^0\|_\infty,$$

this usual norm is chosen for technical calculation for giving the result to our sample, note that the usual norms on $\mathbb{R}^m$ are equivalent, consequently, our choice is not restrictive.

(b) There exists $\delta_2 > 0$ such that for any h satisfying $d_{\mathscr{H}}(h, h_0) \le \delta_2$, for any $j = 1,\dots,k+1$ the function $\boldsymbol{\phi}_j \mapsto \mathbb{E}(\mathbf{m}_j(\mathbf{X}, \boldsymbol{\phi}_j, h))$ is twice continuously differentiable on an open neighborhood of $\boldsymbol{\phi}^0$, we have:

$$\lim_{\|\boldsymbol{\phi}-\boldsymbol{\phi}^0\|\longrightarrow 0} \sup_{d_{\mathscr{H}}(h,h_0)\le\delta_2} \|\boldsymbol{\phi}-\boldsymbol{\phi}^0\|^{-2} \Big| \mathbf{M}(\boldsymbol{\phi}, \lambda_j^0, h) - \mathbf{M}(\boldsymbol{\phi}^0, \lambda_j^0, h) - \Gamma_j(\boldsymbol{\phi}^0, h)(\boldsymbol{\phi}-\boldsymbol{\phi}^0) - \frac{1}{2}(\boldsymbol{\phi}-\boldsymbol{\phi}^0)^\top \Omega_j(\boldsymbol{\phi}^0, h)(\boldsymbol{\phi}-\boldsymbol{\phi}^0) \Big| = 0.$$

(c) $\|\Gamma_j(\boldsymbol{\phi}^0, \widehat{h})\| = O_{\mathbb{P}}(r_n^{-1})$ and $\Gamma_j(\boldsymbol{\phi}^0, h_0) = 0$.

(d) $\Omega_j(\boldsymbol{\phi}^0, h_0)$ is negative define, and $h \mapsto \Gamma_j(\boldsymbol{\phi}^0, h)$ is continuous in h_0. These conditions imply :

$$\begin{aligned} \mathbb{E}(\mathbf{m}_j(\mathbf{X}, \boldsymbol{\phi}_j, \widehat{h})) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}, \boldsymbol{\phi}_j^0, \widehat{h})) &= \langle \Gamma_j(\theta_0, \hat{h}), \gamma_{\phi_j} \rangle + \frac{1}{2} \left(\gamma_{\phi_j}\right)^{\mathrm{T}} \Omega_j(\theta_0, h_0) \left(\gamma_{\phi_j}\right) \\ &\quad + \left\|\gamma_{\phi_j}\right\|^2 o_{\mathrm{P}^*}(1) + o\left(\left\|\gamma_{\phi_j}\right\|^2\right) \\ &\le \mathrm{W}_n d(\boldsymbol{\phi}_j, \boldsymbol{\phi}_j^0) - \mathrm{C} d(\boldsymbol{\phi}_j, \boldsymbol{\phi}_j^0)^2 + \beta_n d(\boldsymbol{\phi}_j, \boldsymbol{\phi}_j^0)^2, \end{aligned}$$

where $\gamma_{\phi_j} = \phi_j - \phi_j^0$, *then we can get*

$$d(\boldsymbol{\phi}, \boldsymbol{\phi}^0) \leq \sum_{j=1}^{k+1} d(\boldsymbol{\phi}_j, \boldsymbol{\phi}_j^0) \leq k d(\boldsymbol{\phi}, \boldsymbol{\phi}^0),$$

it holds also when we replace this norm by $d(\cdot,\cdot)^2$.

(iv) Condition ***(B5)*** *holds automatically under the following classical assumption:*

$$\mathbf{M}_n(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\lambda}}, \widehat{h}) \geq \sup_{\lambda \in \boldsymbol{\Lambda}, \boldsymbol{\phi} \in \boldsymbol{\Phi}} \mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}, \widehat{h}) + \mathrm{O}_{\mathbb{P}}(r_n^{-2}).$$

Under the condition $r_n^2 = o(n)$, *we obtain the rate* $\sqrt{n}$ *and this is needed for the result of Lemma 5.5.0.5, where we give the weak convergence of the parameters of the within-segment distributions. Note that when all the points of change are known or when there isn't a change in the distribution we can drop this condition and we add the possibility that* r_n *reaches* $\sqrt{n}$. *Kim* et al. *[1990] seminal paper, dealing with the estimation in parametric models, is to be mentioned here. In a neighborhood of a fixed parameter point, an* $r_n = n^{1/3}$ *rescaling of the parameter is compensated for by an* $n^{2/3}$ *rescaling of the empirical measure, resulting in a limiting Gaussian process. The authors arguments rely on a simple new sufficient condition for a Gaussian process to achieve its maximum almost surely at a unique point. More precisely, the authors have deduced limit theorems for several statistics defined by maximization or constrained minimization of a process derived from the empirical measure by introducing a modified continuous mapping theorem for the location of the maximizing value. In particular the authors have established a new functional central limit theorem for empirical processes indexed by classes of functions. An extension to the setting* $r_n = n^{\alpha}$ *for some* $\alpha > 0$ *may be found in van der Vaart and Wellner [1996] and Kosorok [2008], where the interested reader may found more details on the subject.*

5.3.3 Asymptotic distribution

In the preceding results, we have obtained

$$\widehat{\lambda} - \lambda^0 = \mathrm{O}_{\mathbb{P}}\left(\frac{1}{n}\right),$$

and

$$r_n d(\widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}^0) = \mathrm{O}_{\mathbb{P}}(1).$$

Our aim now is to study the asymptotic distribution of $r_n(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0)$. In this section we will assume that the parameter space $\boldsymbol{\Phi}$ is equipped with the Euclidean norm $\|\cdot\|$. Let us start by giving some notation, for any $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ and $h \in \mathscr{H}$, let

$$\begin{aligned}
\mathbf{B}_n(\boldsymbol{\phi},\boldsymbol{\lambda},h) &= \mathbf{M}_n(\boldsymbol{\phi},\boldsymbol{\lambda},h)-\mathbf{M}_n(\boldsymbol{\phi}^0,\boldsymbol{\lambda},h),\\
\mathbf{B}(\boldsymbol{\phi},\boldsymbol{\lambda},h) &= \mathbf{M}(\boldsymbol{\phi},\boldsymbol{\lambda},h)-\mathbf{M}(\boldsymbol{\phi}^0,\boldsymbol{\lambda},h),\\
\mathbf{M}_\delta(\cdot) &\geq \sup_{\|\boldsymbol{\phi}-\boldsymbol{\phi}^0\|\leq\delta}\left|\sum_{j=1}^{k+1}[\mathbf{m}_j(\cdot,\boldsymbol{\phi}_j,h_0)-\mathbf{m}_j(\cdot,\boldsymbol{\phi}_j^0,h_0)]\right|,\\
\mathbf{M}_{j,\delta}(\cdot) &\geq \sup_{\|\boldsymbol{\phi}_j-\boldsymbol{\phi}_j^0\|\leq\delta}\left|\mathbf{m}_j(\cdot,\boldsymbol{\phi}_j,h_0)-\mathbf{m}_j(\cdot,\boldsymbol{\phi}_j^0,h_0)\right|,
\end{aligned}$$

for any $\delta>0$. Also, let

$$\mathscr{M}_{j,\delta}=\{\mathbf{m}_j(\cdot,\boldsymbol{\phi}_j,h_0)-\mathbf{m}_j(\cdot,\boldsymbol{\phi}_j^0,h_0),\|\boldsymbol{\phi}_j-\boldsymbol{\phi}_j^0\|\leq\delta\},$$

and

$$\mathscr{M}_\delta=\left\{\sum_{j=1}^{k+1}\mathbf{m}_j(\cdot,\boldsymbol{\phi}_j,h_0)-\mathbf{m}_j(\cdot,\boldsymbol{\phi}_j^0,h_0),\|\boldsymbol{\phi}-\boldsymbol{\phi}^0\|\leq\delta\right\}.$$

Finally, for any $p\in\mathbb{N}$, for any $f:\boldsymbol{\Phi}\longrightarrow\mathbb{R}$ and for any $\boldsymbol{\gamma}=(\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_p)\in\boldsymbol{\Phi}^p$, denote

$$\overline{f}_{\boldsymbol{\gamma}}=(f(\boldsymbol{\gamma}_1),\ldots,f(\boldsymbol{\gamma}_p))^\top.$$

We give the assumptions to investigate the weak convergence without change in the distribution followed by their adaptation for each true sub-sample.

(C1) $r_n\|\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^0\|=\mathrm{O}_{\mathbb{P}}(1)$ and $v_n d_{\mathscr{H}}(\widehat{h},h_0)=\mathrm{O}_{\mathbb{P}}(1)$ for some sequences $r_n\longrightarrow\infty$ and $v_n\longrightarrow\infty$.

(C2) $\boldsymbol{\phi}^0$ belongs to the interior of $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}\subset(\mathbf{E},\|\cdot\|)$, where $\mathbf{E}$ is a finite dimensional Euclidean (i.e., $\mathbf{E}=\mathbb{R}^m$ for some m).

(C3) For all $\delta_2,\delta_3>0$,

$$\sup_{\|\boldsymbol{\phi}-\boldsymbol{\phi}^0\|\leq\frac{\delta_2}{r_n},d_{\mathscr{H}}(\widehat{h},h_0)\leq\frac{\delta_3}{v_n}}|\mathbf{B}_n(\boldsymbol{\phi},\boldsymbol{\lambda}^0,h)-\mathbf{B}(\boldsymbol{\phi},\boldsymbol{\lambda}^0,h)-\mathbf{B}_n(\boldsymbol{\phi},\boldsymbol{\lambda}^0,h_0)+\mathbf{B}(\boldsymbol{\phi},\boldsymbol{\lambda}^0,h_0)|=o_{\mathbb{P}}(r_n^{-2}).$$

(C4) For all $\mathrm{K},\eta>0$ and for any $j=1,\ldots,k+1$

$$\frac{r_n^4}{n}\mathbb{E}\left[\mathbf{M}^2_{j,\frac{\mathrm{K}}{r_n}}\right]=\mathrm{O}(1)\ \ and\ \ \frac{r_n^4}{n}\mathbb{E}\left[\mathbf{M}^2_{j,\frac{\mathrm{K}}{r_n}}\mathbb{1}_{\{r_n^2\mathbf{M}_{j,\frac{\mathrm{K}}{r_n}}>\eta n\}}\right]=o(1).$$

(C5) For all $\mathrm{K}>0$, for any $j=1,\ldots,k+1$ and for any $\eta_n\longrightarrow0$,

$$\sup_{\|\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_2\|<\eta_n,\|\boldsymbol{\gamma}_1\|\vee\|\boldsymbol{\gamma}_2\|\leq\mathrm{K}}\frac{r_n^4}{n}\mathbb{E}\left[\mathbf{m}_j\left(\mathbf{X},\boldsymbol{\phi}_j^0+\frac{\boldsymbol{\gamma}_1}{r_n},h_0\right)-\mathbf{m}_j\left(\mathbf{X},\boldsymbol{\phi}_j^0+\frac{\boldsymbol{\gamma}_2}{r_n},h_0\right)\right]^2=o(1).$$

(C6) For $z\in\mathbf{F}$, fol all $j=1,\ldots,k+1$, the function $\boldsymbol{\phi}_j\mapsto\mathbf{m}_j(z,\boldsymbol{\phi}_j,h_0)$ and almost all paths of the process $\boldsymbol{\phi}_j\mapsto\mathbf{m}_j(z,\boldsymbol{\phi}_j,\widehat{h})$ are uniformly (over $\boldsymbol{\phi}_j$) bounded on compact sets.

(C7) There exists $\beta_n = o_{\mathbb{P}}(1)$, a random and linear function $W_n : \mathbb{E} \longrightarrow \mathbb{R}$, and a deterministic and bilinear function $V : \mathbb{E} \times \mathbb{E} \longrightarrow \mathbb{R}$ such that for all $\boldsymbol{\phi} \in \boldsymbol{\Phi}$.

$$\mathbf{B}(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) = W_n(\boldsymbol{\gamma}_{\boldsymbol{\phi}}) + V(\boldsymbol{\gamma}_{\boldsymbol{\phi}}, \boldsymbol{\gamma}_{\boldsymbol{\phi}}) + \beta_n \|\boldsymbol{\gamma}_{\boldsymbol{\phi}}\|^2 + o(\|\boldsymbol{\gamma}_{\boldsymbol{\phi}}\|^2),$$

and

$$\mathbf{B}(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, h_0) = V(\boldsymbol{\gamma}_{\boldsymbol{\phi}}, \boldsymbol{\gamma}_{\boldsymbol{\phi}}) + o(\|\boldsymbol{\gamma}_{\boldsymbol{\phi}}\|^2),$$

where $\boldsymbol{\gamma}_{\boldsymbol{\phi}} = \boldsymbol{\phi} - \boldsymbol{\phi}^0$. Moreover, for any compact set $\mathscr{K} \subset \mathbf{E}$,

$$\exists \tau, \delta_1 > 0, r_n \sup_{\substack{\boldsymbol{\gamma} \in \mathscr{K}, \delta \leq \delta_1 \\ \|\boldsymbol{\gamma}\| \leq \delta}} \left|\frac{W_n(\boldsymbol{\gamma})}{\delta^\tau}\right| = O_{\mathbb{P}}(1) \quad and \quad \sup_{\substack{\boldsymbol{\gamma}, \boldsymbol{\gamma}' \in \mathscr{K}, \delta \leq \delta_1 \\ \|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\| \leq \delta}} \frac{|V(\boldsymbol{\gamma}, \boldsymbol{\gamma}) - V(\boldsymbol{\gamma}', \boldsymbol{\gamma}')|}{\delta^\tau} < \infty.$$

(C8) For all $K > 0$, there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,

$$\mathbf{M}_n(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\lambda}}, \widehat{h}) \geq \sup_{\|\boldsymbol{\phi} - \boldsymbol{\phi}^0\| \leq \frac{K}{r_n}} \mathbf{M}_n(\boldsymbol{\phi}, \widehat{\boldsymbol{\lambda}}, \widehat{h}) + o_{\mathbb{P}}(r_n^{-2}).$$

(C9) There exists a deterministic continuous function Γ and a zero-mean Gaussian process $\mathbb{G}$ defined on $\mathbf{E}$ such that for all $p \in \mathbb{N}$ and for all $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p) \in \mathbf{E}^p$,

$$r_n \overline{W_n}_{\boldsymbol{\gamma}} + r_n^2 \overline{B_n\left(\boldsymbol{\phi}^0 + \frac{\cdot}{r_n}, \boldsymbol{\lambda}^0, h_0\right)}_{\boldsymbol{\gamma}} \Rightarrow \overline{\Gamma}_{\boldsymbol{\gamma}} + \overline{\mathbb{G}}_{\boldsymbol{\gamma}},$$

where "$\Rightarrow$" denotes the weak convergence. Moreover, $\mathbb{G}(\boldsymbol{\gamma}) = \mathbb{G}(\boldsymbol{\gamma}')$ a.s. implies that $\boldsymbol{\gamma} = \boldsymbol{\gamma}'$, and

$$\mathbb{P}\left(\limsup_{\|\boldsymbol{\gamma}\| \longrightarrow \infty} (\Gamma(\boldsymbol{\gamma}) + \mathbb{G}(\boldsymbol{\gamma})) < \sup_{\boldsymbol{\gamma} \in \mathbf{E}} (\Gamma(\boldsymbol{\gamma}) + \mathbb{G}(\boldsymbol{\gamma}))\right) = 1.$$

(C10) There exists a $\delta_0 > 0$ such that

$$\int_0^\infty \sup_{\delta \leq \delta_0} \sqrt{\log\left(N_{[\,]}(\epsilon \|\mathbf{M}_\delta\|_{\mathbb{P},2}, \mathscr{M}_\delta, \mathbb{L}^2(\mathbb{P}))\right)} d\epsilon < \infty.$$

We will show that $r_n(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0)$ converges to the unique maximizer of the process $\boldsymbol{\gamma} \mapsto \Gamma(\boldsymbol{\gamma}) + \mathbb{G}(\boldsymbol{\gamma})$, where $\Gamma(\cdot)$ and $\mathbb{G}(\cdot)$ are defined in **(C9)**. At first we discuss the above assumptions in more detail.

Remark 5.3.3.1

*(i) From Theorem 5.3.2.3 we can obtain the first part of assumption **(C1)**.*

*(ii) Assumption **(C3)** is automatically fulfilled if we have for each true sub-sample : for all* $\delta_2, \delta_3 > 0$

$$\sup_{\|\boldsymbol{\phi}_j - \boldsymbol{\phi}_j^0\| \leq \frac{\delta_2}{r_n}, d_{\mathscr{H}}(h, h_0) \leq \frac{\delta_3}{v_n}} |\mathbf{B}_n(\boldsymbol{\phi}_j, \lambda_j^0, h) - \mathbf{B}(\boldsymbol{\phi}_j, \lambda_j^0, h) - \mathbf{B}_n(\boldsymbol{\phi}_j, \lambda_j^0, h_0) + \mathbf{B}(\boldsymbol{\phi}, \lambda_j^0, h_0)| = o_{\mathbb{P}}(r_n^{-2}),$$

where

$$\mathbf{B}_n(\boldsymbol{\phi}_j, \lambda_j^0, h) = \mathbf{M}_n(\boldsymbol{\phi}_j, \lambda_j^0, h) - \mathbf{M}_n(\boldsymbol{\phi}_j^0, \lambda_j^0, h),$$

and

$$\mathbf{B}(\boldsymbol{\phi}_j, \lambda_j^0, h) = \mathbf{M}(\boldsymbol{\phi}_j, \lambda_j^0, h) - \mathbf{M}(\boldsymbol{\phi}_j^0, \lambda_j^0, h).$$

This condition is satisfied if: there exists a function f_j and a constant $\delta_0 > 0$ such that for all $\delta_2, \delta_3 < \delta_0$,

$$r_n^2 f_j\left(\frac{\delta_2}{r_n}, \frac{\delta_3}{v_n}\right) = o(\sqrt{n}),$$

and

$$\mathbb{E}\left[\sup_{\|\boldsymbol{\phi}_j - \boldsymbol{\phi}_j^0\| \leq \frac{\delta_2}{r_n}, d_{\mathscr{H}}(\widehat{h}, h_0) \leq \frac{\delta_3}{v_n}} \left|\mathbf{B}_n(\boldsymbol{\phi}_j, \lambda_j^0, h) - \mathbf{B}(\boldsymbol{\phi}_j, \lambda_j^0, h) - \mathbf{B}_n(\boldsymbol{\phi}_j, \lambda_j^0, h_0) + \mathbf{B}(\boldsymbol{\phi}, \lambda_j^0, h_0)\right|\right] \leq \frac{1}{\sqrt{n}} f_j\left(\frac{\delta_2}{r_n}, \frac{\delta_3}{v_n}\right).$$

This last bound may be obtained using the same arguments as in ***Remark 3.6****(ii).*

(iii) We assume that assumption ***(B3)*** *holds with $\psi_n \equiv \psi$ not depending on n and continuous. Let us mention that in the particular case when $\psi_n(\cdot) = \sum_{j=1}^{k} \psi_n^j(\cdot)$, where $\psi_n^j(\cdot)$ is calculated for each true sub-sample $\mathbf{X}_{n_{j-1}^0+1}, \ldots, \mathbf{X}_{n_j^0}$ for each $j = 1, \ldots, k+1$, which is formed by i.i.d. random vectors, leads to the functions $\psi_n^j(\cdot) \equiv \psi^j(\cdot)$ independent of n for each $j = 1, \ldots, k+1$, one can refer to van der Vaart and Wellner [1996] and Kosorok [2008]) for more discussion. If we consider the situation when $r_n \longrightarrow \infty$ such that $r_n^2 \psi(r_n^{-1}) = \sqrt{n}$, then assumption* ***(C4)*** *and* ***(C5)*** *are implied by the following ones: there exists a $\delta_4 > 0$ such that for all $\delta \leq \delta_4$, for all $j = 1, \ldots, k+1$, $\mathbb{E}(\mathbf{M}_{j,\delta}^2) \leq \mathrm{K}\psi^2(\delta)$ for some $\mathrm{K} > 0$,*

$$\lim_{\delta \longrightarrow 0} \frac{\mathbb{E}\left[\mathbf{M}_{j,\delta}^2 \mathbb{1}_{\{\mathbf{M}_{j,\delta} > \eta\delta^{-2}\psi^2(\delta)\}}\right]}{\psi^2(\delta)} = 0$$

for all $\eta > 0$, and

$$\lim_{\epsilon \longrightarrow 0} \lim_{\delta \longrightarrow 0} \sup_{\|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\| < \epsilon, \|\boldsymbol{\gamma}_1\| \vee \boldsymbol{\gamma}_2 \leq \mathrm{K}} \frac{\mathbb{E}\left[\mathbf{m}_j\left(\mathbf{X}, \boldsymbol{\phi}_j^0 + \boldsymbol{\gamma}_1\delta, h_0\right) - \mathbf{m}_j\left(\mathbf{X}, \boldsymbol{\phi}_j^0 + \boldsymbol{\gamma}_2\delta, h_0\right)\right]^2}{\psi^2(\delta)} = 0$$

for all $\mathrm{K} > 0$, using the same arguments as in the proof of Theorem 3.2.10 in van der Vaart and Wellner [1996]. These assumptions are used for the investigation of the variance of each process

$$\boldsymbol{\gamma} \mapsto r_n^2\left(\mathbf{B}_n\left(\boldsymbol{\phi}_j^0 + \frac{\boldsymbol{\gamma}}{r_n}, \lambda_j^0, h_0\right) - \mathbf{B}\left(\boldsymbol{\phi}_j^0 + \frac{\boldsymbol{\gamma}}{r_n}, \lambda_j^0, h_0\right)\right)$$

and in the proof of their weak convergence.

(iv) *Assumption* ***(C6)*** *ensures that for any compact* $\mathscr{K} \subset \mathbf{E}$ *each process* $\boldsymbol{\gamma} \mapsto r_n^2 \mathbf{B}_n\left(\boldsymbol{\phi}_j^0 + \frac{\boldsymbol{\gamma}}{r_n}, \lambda_j^0, \widehat{h}\right)$ *and* $\boldsymbol{\gamma} \mapsto r_n^2 \mathbf{B}_n\left(\boldsymbol{\phi}_j^0 + \frac{\boldsymbol{\gamma}}{r_n}, \lambda_j^0, h_0\right) + r_n \mathrm{W}_n(\boldsymbol{\gamma}),\ j = 1, \ldots, k+1$ *take value in* $\ell^\infty(\mathscr{K})$*, for the second process we used also assumption* ***(C7)****, which gives us that*

$$\boldsymbol{\gamma} \mapsto r_n^2 \mathbf{B}_n\left(\boldsymbol{\phi}^0 + \frac{\boldsymbol{\gamma}}{r_n}, \boldsymbol{\lambda}^0, \widehat{h}\right)$$

and

$$\boldsymbol{\gamma} \mapsto r_n^2 \mathbf{B}_n\left(\boldsymbol{\phi}^0 + \frac{\boldsymbol{\gamma}}{r_n}, \boldsymbol{\lambda}^0, h_0\right) + r_n \mathrm{W}_n(\boldsymbol{\gamma})$$

are in $\ell^\infty(\mathscr{K})$ *like sum of process taking their values in this set.*

(v) *We assume the assumptions (a)-(d) from* ***Remark 3.6****(iii) hold for each* $j = 1, \ldots, k+1$*. Following the same ideas as in this remark, it's easy to show* ***(C7)*** *is fulfilled if we have for each* $j = 1, \ldots, k+1$

$$\mathbf{B}(\boldsymbol{\phi}, \lambda_j^0, \widehat{h}) = \mathrm{W}_{j,n}(\boldsymbol{\gamma}_{\boldsymbol{\phi}}) + \mathrm{V}_j(\boldsymbol{\gamma}_{\boldsymbol{\phi}}, \boldsymbol{\gamma}_{\boldsymbol{\phi}}) + \beta_n \|\boldsymbol{\gamma}_{\boldsymbol{\phi}}\|^2 + o(\|\boldsymbol{\gamma}_{\boldsymbol{\phi}}\|^2)$$

and

$$\mathbf{B}(\boldsymbol{\phi}, \lambda_j^0, h_0) = \mathrm{V}_j(\boldsymbol{\gamma}_{\boldsymbol{\phi}}, \boldsymbol{\gamma}_{\boldsymbol{\phi}}) + o(\|\boldsymbol{\gamma}_{\boldsymbol{\phi}}\|^2),$$

with $\mathbf{E} = \mathbb{R}^m$*,*

$$\mathrm{W}_n(\boldsymbol{\gamma}) = \sum_{j=1}^{k+1} \mathrm{W}_{j,n}(\boldsymbol{\gamma}) = \sum_{j=1}^{k+1} \langle \Omega_j(\boldsymbol{\phi}^0, \widehat{h}), \boldsymbol{\gamma} \rangle$$

and

$$\mathrm{V}(\boldsymbol{\gamma}, \boldsymbol{\gamma}) = \sum_{j=1}^{k+1} \mathrm{V}_j(\boldsymbol{\gamma}, \boldsymbol{\gamma}) = \frac{1}{2} \boldsymbol{\gamma}^\top \Gamma(\boldsymbol{\phi}^0, h_0) \boldsymbol{\gamma};$$

where $\Gamma(\cdot, \cdot) = \sum_{j=1}^{k+1} \Gamma_j(\cdot, \cdot)$ *whenever*

$$\sup_{u \in \mathbb{R}^m, \|u\|=1} \|\Gamma(\boldsymbol{\phi}^0, h_0) u\| < \infty.$$

We assume also the two last expressions in ***(C7)*** *hold when we replace* $\mathrm{W}_n(\cdot)$ *and* $\mathrm{V}(\cdot, \cdot)$ *by* $\mathrm{W}_{j,n}(\cdot)$ *and* $\mathrm{V}_j(\cdot, \cdot)$ *respectively.*

(vi) *Assumption* ***(C8)*** *allows to consider estimators* $\widehat{\boldsymbol{\phi}}$ *that are approximations of the value that actually maximizes the map* $\boldsymbol{\phi} \mapsto \mathbf{M}_n(\boldsymbol{\phi}, \widehat{\boldsymbol{\lambda}}, \widehat{h})$*.*

(vii) *If we assume that assumption* ***(C9)*** *is fulfilled for every process*

$$\boldsymbol{\gamma} \mapsto r_n^2 \mathbf{B}_n\left(\boldsymbol{\phi}_j^0 + \frac{\boldsymbol{\gamma}}{r_n}, \lambda_j^0, h_0\right) + r_n \mathrm{W}_{j,n}(\boldsymbol{\gamma}), j = 1, \ldots, k+1;$$

so this condition is used for showing their weak convergence (in the $\ell^\infty(\mathscr{K})$*) to the processes* $\boldsymbol{\gamma} \mapsto \Gamma_j(\boldsymbol{\gamma}) + \mathbb{G}_j(\boldsymbol{\gamma}),\ j = 1, \ldots, k+1$ *from the fact that they are asymptotically tight. If*

$$r_n \sup_{\boldsymbol{\gamma} \in \mathscr{K}, \boldsymbol{\gamma} \neq 0} \|\mathrm{W}_{j,n}(\boldsymbol{\gamma}) \|\boldsymbol{\gamma}\|^{-1}\| = o_{\mathbb{P}}(1), j = 1, \ldots, k+1,$$

we are in the same situation as in the parametric case and we obtain the convergence of the marginals of each process of the true sub-sample and with this we can obtain the convergence of the marginals of the sum of these processes like in the assumption ***(C9)*** *because the variables are independent. The last part of this assumption on the process* $\boldsymbol{\gamma} \mapsto \Gamma(\boldsymbol{\gamma}) + \mathbb{G}(\boldsymbol{\gamma})$ *is used to show that almost all sample paths have a supremum which is only related to their behavior on compact sets. The dominant term of the deterministic part* Γ *is usually a negative definite quadratic form and hence exponential inequalities could lead to such a result, see Remark 3(vi) of Delsol and Van Keilegom [2020].*

(viii) We used assumption ***(C10)*** *to show the asymptotically tightness of the process*

$$\boldsymbol{\gamma} \mapsto r_n^2 \mathbf{B}_n\left(\boldsymbol{\phi}^0 + \frac{\boldsymbol{\gamma}}{r_n}, \boldsymbol{\lambda}^0, h_0\right),$$

it's automatically fulfilled if we assume for each $j = 1, \ldots, k+1$

$$\int_0^\infty \sup_{\delta \leq \tilde{\delta}_0} \sqrt{\log\left(N_{[\,]}(\epsilon \|\mathbf{M}_{j,\delta}\|_{\mathbb{P},2}, \mathscr{M}_{j,\delta}, \mathbb{L}^2(\mathbb{P}))\right)} d\epsilon < \infty,$$

these are used to show that the processes

$$\boldsymbol{\gamma} \mapsto r_n^2 \mathbf{B}_n\left(\boldsymbol{\phi}_j^0 + \frac{\boldsymbol{\gamma}}{r_n}, \lambda_j^0, h_0\right), \text{ for } j = 1, \ldots, k+1$$

are asymptotically tight which, in turn, implies the result for the sum of these processes. For weaker conditions of these assumptions based on converging numbers we refer the reader to Theorems 2.11.22, 2.11.23 and 3.2.10 of van der Vaart and Wellner [1996], those are the same as in the parametric case, where h_0 *is known (see Theorem 3.2.10 in van der Vaart and Wellner [1996]). The same holds true with* ***(C4)-(C5)****.*

After giving the assumptions for the asymptotic distribution of $\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0$ and their clarification, we give now the main result of this paper. Notice that in the model without change in the distribution of the data, our theorem reduces to Theorem 3 in Delsol and Van Keilegom [2020].

Theorem 5.3.3.2 *Under conditions* ***(C1)-(C10)*** *we have; for all* $K > 0$ *the process*

$$\boldsymbol{\gamma} \mapsto r_n^2 \mathbf{B}_n\left(\boldsymbol{\phi}^0 + \frac{\boldsymbol{\gamma}}{r_n}, \widehat{\boldsymbol{\lambda}}, \widehat{h}\right)$$

converges weakly to

$$\boldsymbol{\gamma} \mapsto \Gamma(\boldsymbol{\gamma}) + \mathbb{G}(\boldsymbol{\gamma}) \text{ in } \ell^\infty(\mathscr{K})$$

where $\mathscr{K} = \{\boldsymbol{\gamma} \in \mathbf{E} : \|\boldsymbol{\gamma}\| \leq K\}$*. Moreover for any such* $\mathscr{K}$ *almost all paths of the limiting process have a unique maximizer* $\boldsymbol{\gamma}^0$ *on* $\mathscr{K}$*. We assume now that* $\boldsymbol{\gamma}^0$ *is measurable. Then,* $r_n(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0)$ *converges in distribution to* $\boldsymbol{\gamma}^0$*.*

Remark 5.3.3.3 *In the present work we have assumed that the number of changes in the sample is known, which is not the case in the real applications. To circumvent this, we can use the binary segmentation method proposed in Vostrikova [1981], which is a "top down" procedure, in the sense that one tests all the data to determine if there is at least one change-point and iterates the procedure in the intervals immediately to the "left" and "right" of the most recently detected change-point.*

Remark 5.3.3.4 *For notational convenience, we have considered that the nuisance parameter $h(\cdot)$ depends only on the common parameter α and not on the within-segment parameters. This situation is fulfilled when we study the change point for the copula semiparametric models. In this situation it commonly assumed that the nuisance parameters (the nonparametric margins) are not subject to changes within-segment, only the dependence parameters vary from segment to segment, we refer to Bouzebda and Keziou [2013], Bouzebda [2012, 2014] and the references therein.*

5.3.4 Example : classification with missing data in model with change point

We will give an example of classification with missing data and we keep the same notation as in Delsol and Van Keilegom [2020] where we add in this example the case of many but known changes in distribution. We recall the example without a change point. Let us consider i.i.d. data $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2})$ $(i = 1, 2, \ldots, n)$ having the same distribution as $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$. We suppose that these data come in reality from two underlying populations. Let $\mathbf{Y}_i$ be j if observation i belongs to population j $(j = 0, 1)$, and let $\mathbf{Y}$ be the population indicator for the vector $\mathbf{X}$. Based on these data, we wish to establish a classification rule for new observations, for which it will be unknown to which population they belong. The classification consists in regressing $\mathbf{X}_2$ on $\mathbf{X}_1$ via a parametric regression function $f_{\boldsymbol{\theta}}(\cdot)$, and choosing $\boldsymbol{\theta}$ by maximising the criterion

$$\mathbb{P}(\mathbf{Y} = 1, \mathbf{X}_2 \geq f_{\boldsymbol{\theta}}(\mathbf{X}_1)) + \mathbb{P}(\mathbf{Y} = 0, \mathbf{X}_2 < f_{\boldsymbol{\theta}}(\mathbf{X}_1)). \tag{5.3.4}$$

Let $\boldsymbol{\theta}^0$ be the value of $\boldsymbol{\theta}$ that maximizes (5.3.4) with respect to all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a compact subset of $\mathbb{R}^k$, whose interior contains $\boldsymbol{\theta}^0$. We suppose now that some of the $\mathbf{Y}_i$'s are missing. Let Δ_i (respectively Δ) be 1 if $\mathbf{Y}_i$ (respectively $\mathbf{Y}$) is observed, and 0 otherwise. Hence, our data consist of i.i.d. vectors $\mathbf{Z}_i = (\mathbf{X}_i, \mathbf{X}_i\Delta_i, \Delta_i)$ $(i = 1, 2 \ldots, n)$. We assume that the missing at random mechanism holds true, in the sense that

$$\mathbb{P}(\Delta = 1 | \mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}) = \mathbb{P}(\Delta = 1 | \mathbf{X}_1) := p^0(\mathbf{X}_1).$$

Note that the expression 5.3.4 can be written as follows

$$\mathbb{E}\left[\frac{\mathbb{1}\{\Delta = 1\}}{p^0(\mathbf{X}_1)}\left\{\mathbb{1}\{\mathbf{Y} = 1, \mathbf{X}_2 \geq f_{\boldsymbol{\theta}}(\mathbf{X}_1)\} + \mathbb{1}\{\mathbf{Y} = 0, \mathbf{X}_2 < f_{\boldsymbol{\theta}}(\mathbf{X}_1)\}\right\}\right],$$

where $\mathbb{1}\{\mathrm{A}\}$ denotes the indicator function of A. The reader could find the expression of $m(\mathbf{Z}, \boldsymbol{\theta}, p)$, $\mathbf{M}(\boldsymbol{\theta}, p)$, $\mathbf{M}_n(\boldsymbol{\theta}, p)$ and the non-parametric estimator $\widehat{p}(\cdot)$ of $p^0(\cdot)$ in the same reference. We will

adapt the preceding problem formulation to the context of change point, i.e., we consider independent data $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2})$ $(i = 1, 2, \ldots, n)$ where they are i.i.d. if $n_{j-1}^0 < i \leq n_j^0$ $j = 1, 2 \ldots, k+1$ where n_j^0 is the true point of change in distribution and k is the known number of change and we assume that the data $(\mathbf{X}_{i1})$ $(i = 1, 2, \ldots, n)$ are i.i.d. with the same distribution as $\mathbf{X}_1$, under this model we obtain $k+1$ parametric functions $f_{\boldsymbol{\theta}_j}(\cdot)$, and choosing $\boldsymbol{\theta}_j$ by maximizing the criterion

$$\mathbb{P}\left(\mathbf{Y} = 1, \mathbf{X}_{n_j^0 2} \geq f_{\boldsymbol{\theta}_j}(\mathbf{X}_{n_j^0 1})\right) + \mathbb{P}\left(\mathbf{Y} = 0, \mathbf{X}_{n_j^0 2} < f_{\boldsymbol{\theta}_j}(\mathbf{X}_{n_j^0 1})\right). \tag{5.3.5}$$

Let $\boldsymbol{\theta}_j^0$ be the value of $\boldsymbol{\theta}_j$ that maximizes (5.3.5) with respect to all $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j$, where $\boldsymbol{\Theta}_j$ is a compact subset of $\mathbb{R}^{d_j}$, whose interior contains $\boldsymbol{\theta}_j^0$. We suppose that some of the $\mathbf{Y}_i$'s are missing. in this case our data consist of independent vectors $\mathbf{Z}_i = (\mathbf{X}_i, \mathbf{Y}_i \Delta_i, \Delta_i)$ $(i = 1, 2 \ldots, n)$ where they are i.i.d. if $n_{j-1}^0 < i \leq n_j^0$ $j = 1, 2 \ldots, k+1$. We assume like in model without a change in distribution that the missing at random mechanism holds true, in the following sense

$$\mathbb{P}(\Delta = 1 | \mathbf{X}_{i1}, \mathbf{X}_{i2}, \mathbf{Y}) = \mathbb{P}(\Delta = 1 | \mathbf{X}_{i1}) := p^0(\mathbf{X}_1).$$

We introduce our statistics, for $j = 1, \ldots, k+1$,

$$\begin{aligned} \mathbf{m}_j(\mathbf{Z}_i, \boldsymbol{\theta}_j, p) &= \frac{\mathbb{1}\{\Delta_i = 1\}}{p(\mathbf{X}_{i1})} \Big\{ \mathbb{1}\{\mathbf{Y}_i = 1, \mathbf{X}_{i2} \geq f_{\boldsymbol{\theta}_j}(\mathbf{X}_{i1})\} \\ &\quad + \mathbb{1}\{\mathbf{Y}_i = 0, \mathbf{X}_{i2} < f_{\boldsymbol{\theta}_j}(\mathbf{X}_{i1})\} \Big\}. \end{aligned} \tag{5.3.6}$$

where the nuisance function $p(\cdot)$ belongs to a space $\mathscr{P}$ to be defined later. Also, let

$$\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}, p) = \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}+1}^{n_j} \mathbf{m}_j(\mathbf{Z}_i, \boldsymbol{\theta}_j, p) \tag{5.3.7}$$

$$\mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}, p) = \sum_{j=1}^{k+1} (\lambda_j - \lambda_{j-1}) \mathbb{E}[\mathbf{m}_j(\mathbf{Z}_i, \boldsymbol{\theta}_j, p)], \tag{5.3.8}$$

where $\boldsymbol{\phi} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k+1})$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$ and $\boldsymbol{\Phi} = \mathbb{R}^{d_1 + \cdots + d_{k+1}}$. Consequently the estimators $\widehat{\boldsymbol{\phi}}$ and $\widehat{\boldsymbol{\lambda}}$ of $\boldsymbol{\phi}^0$ and $\boldsymbol{\lambda}^0$ respectively are given by

$$(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\lambda}}) = \underset{0 < n_1 < \ldots < n_k < n; \boldsymbol{\phi} \in \boldsymbol{\Phi}}{\operatorname{argmax}} \mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}, p),$$

where for any x_1,

$$\widehat{p}(x_1) = \sum_{i=1}^{n} \frac{\mathbb{K}_h(x_1 - \mathbf{X}_{i1})}{\sum_{j=1}^{n} \mathbb{K}_h(x_1 - \mathbf{X}_{j1})} \mathbb{1}\{\Delta_i = 1\},$$

where $\mathbb{K}(\cdot)$ is a density function with support $[-1, 1]$, $\mathbb{K}_h(u) = \frac{\mathbb{K}(\frac{u}{h})}{h}$ and $h = h_n$ is a bandwidth sequence. Non parametric regression with missing data has been studied very extensively in the literature, see, e.g., Müller [2009], Pérez-González *et al.* [2009], Koul *et al.* [2012], among many others. This is one of examples where we can apply our theory in semiparametric estimators with model of change point and the usefulness of the asymptotic result of this paper

in such cases where the criterion function is not-differentiable with respect to $\boldsymbol{\theta}_j$ for every $j = 1, 2, \ldots, k+1$. Now we will study in full detail this example, and we work out the verification of conditions of Theorems 5.3.1.1, 5.3.2.1, 5.3.2.3 and 5.3.3.2 the most of these conditions verified in Section 7 of Delsol and Van Keilegom [2020] for each true sub sample or when our data are i.i.d. In the beginning we give some information about the nuisance function, their appropriate space and some notation. Suppose $d(\boldsymbol{\phi}, \boldsymbol{\phi}^0)$ is the euclidean distance $\|\cdot\|$. Let $\mathscr{P}$ be the space of functions $p : \mathbf{R}_{\mathbf{X}_1} \to \mathbb{R}$ that are continuously differentiable, and for which

$$\sup_{x_1 \in \mathbf{R}_{\mathbf{X}_1}} p(x_1) \leq \mathrm{M} < \infty, \quad \sup_{x_1 \in \mathbf{R}_{\mathbf{X}_1}} |p^{'}(x_1)| \leq \mathrm{M} \text{ and } \inf_{x_1 \in \mathbf{R}_{\mathbf{X}_1}} p(x_1) > \eta/2,$$

where $\eta = \inf_{x_1 \in \mathbf{R}_{\mathbf{X}_1}} p^0(x_1) > 0$, and where $\mathbf{R}_{\mathbf{X}_1}$ is the support of $\mathbf{X}_1$ which is supposed to be a compact subspace of $\mathbb{R}$. We equip the space $\mathscr{P}$ with the supremum norm:

$$d_{\mathscr{P}}(p_1, p_2) = \sup_{x_1 \in \mathbf{R}_{\mathbf{X}_1}} |p_1(x_1) - p_2(x_1)|,$$

for any $p_1, p_2 \in \mathscr{P}$. After, the conditions of the consistency are verified as follows, **(A1)** is verified by construction of the estimators $\widehat{\boldsymbol{\phi}}$ and $\widehat{\boldsymbol{\lambda}}$. Condition **(A2)** is an identifiability condition, needed to ensure the uniqueness of $\boldsymbol{\phi}^0$ and $\boldsymbol{\lambda}^0$, also **(A3)** is to ensure that there is a change in distribution, **(A4)** holds true provided the functions $p_0(\cdot)$ and $\mathbb{K}(\cdot)$ are continuously differentiable. Concerning the condition **(A5)**, we have the functions $\mathbf{m}_j(\cdot, \boldsymbol{\theta}_j, p)$ and $[\mathbf{m}_j(\cdot, \boldsymbol{\theta}_j, p)]^2$ are bounded for all $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j$, $j = 1, 2, \ldots, k+1$, note that for each $j = 1, 2, \ldots, k+1$; the function $\boldsymbol{\theta}_j \to \mathbf{m}_j(\cdot, \boldsymbol{\theta}_j, p) - \mathbb{E}\mathbf{m}_j(\cdot, \boldsymbol{\theta}_j, p)$ take value in $[-\frac{1}{\eta}, \frac{1}{\eta}]$ for all $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j$ which implies the existence of $\boldsymbol{\theta}_j^*$ such that

$$\max_{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j} |\mathbf{m}_j(\cdot, \boldsymbol{\theta}_j, p) - \mathbb{E}\mathbf{m}_j(\cdot, \boldsymbol{\theta}_j, p)| \leq \mathbf{m}_j(\cdot, \boldsymbol{\theta}_j^*, p) - \mathbb{E}\mathbf{m}_j(\cdot, \boldsymbol{\theta}_j^*, p).$$

So the assumption **(A5)** is satisfied with $r = 1$; for any $j = 1, 2, \ldots, k+1$, any $0 \leq s < t \leq n$, we have

$$\begin{aligned}
&\max_{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j} \left(\sum_{i=s+1}^{t} \mathbf{m}_j(\mathbf{Z}_i, \boldsymbol{\theta}_j, p^0) - \mathbb{E}\mathbf{m}_j(\mathbf{Z}_i, \boldsymbol{\theta}_j, p^0) \right)^2 \\
&\leq \max_{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j} \left\{ \sum_{i=s+1}^{t} \left(\mathbf{m}_j(\mathbf{Z}_i, \boldsymbol{\theta}_j, p^0) - \mathbb{E}\mathbf{m}_j(\mathbf{Z}_i, \boldsymbol{\theta}_j, p^0) \right)^2 \right. \\
&\quad \left. + 2 \sum_{s+1 \leq i < k \leq t} \left| \left(\mathbf{m}_j(\mathbf{Z}_i, \boldsymbol{\theta}_j, p^0) - \mathbb{E}\mathbf{m}_j(\mathbf{Z}_i, \boldsymbol{\theta}_j, p^0) \right) \left(\mathbf{m}_j(\mathbf{Z}_k, \boldsymbol{\theta}_j, p^0) - \mathbb{E}\mathbf{m}_j(\mathbf{Z}_k, \boldsymbol{\theta}_j, p^0) \right) \right| \right\} \\
&\leq \sum_{i=s+1}^{t} \max_{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j} \left(\mathbf{m}_j(\mathbf{Z}_i, \boldsymbol{\theta}_j, p^0) - \mathbb{E}\mathbf{m}_j(\mathbf{Z}_i, \boldsymbol{\theta}_j, p^0) \right)^2 \\
&\quad + 2 \sum_{s+1 \leq i < k \leq t} \left(\mathbf{m}_j(\mathbf{Z}_i, \boldsymbol{\theta}_j^*, p^0) - \mathbb{E}\mathbf{m}_j(\mathbf{Z}_i, \boldsymbol{\theta}_j^*, p^0) \right) \left(\mathbf{m}_j(\mathbf{Z}_k, \boldsymbol{\theta}_j^*, p^0) - \mathbb{E}\mathbf{m}_j(\mathbf{Z}_k, \boldsymbol{\theta}_j^*, p^0) \right).
\end{aligned}$$

The result follows from the fact that the variables are independent. The condition **(A6)** is verified directly for this model, for each $j = 1, 2, \ldots, k+1$ we have;

$$|\mathbf{m}_j(\mathbf{Z_i}, \boldsymbol{\theta}_j, p) - \mathbf{m}_j(\mathbf{Z_i}, \boldsymbol{\theta}_j, p^0)| \leq \frac{\mathbb{1}\{\Delta_i = 1\}}{2\eta^2} d_{\mathscr{P}}(p, p^0),$$

and hence the consistency of $\widehat{\boldsymbol{\theta}}_j, \widehat{\lambda}_j \;\; j = 1,2,\ldots,k+1$ follows. Next, we verify the B-conditions. Condition **(B1)** is implied by the definition of $\boldsymbol{\theta}_i^0$ that maximizes the probability in (5.3.5) in each true sub-sample $n_{i-1}^0+1,\ldots,n_i^0$ where $i = 1,2,\ldots,k+1$; to clarify this let $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j$ we obtain

$$\begin{aligned}\int_{\mathbb{R}} \mathbf{m}_j(z,\boldsymbol{\theta}_j,p^0)d\mathrm{F}_{n_i^0}(z) &= \mathbb{P}\left(\mathbf{Y}=1,\mathbf{X}_{n_i^0 2} \geq f_{\boldsymbol{\theta}_j}(\mathbf{X}_{n_i^0 1})\right) + \mathbb{P}\left(\mathbf{Y}=0,\mathbf{X}_{n_i^0 2} < f_{\boldsymbol{\theta}_j}(\mathbf{X}_{n_i^0 1})\right) \\ &\leq \mathbb{P}\left(\mathbf{Y}=1,\mathbf{X}_{n_i^0 2} \geq f_{\boldsymbol{\theta}_i^0}(\mathbf{X}_{n_i^0 1})\right) + \mathbb{P}\left(\mathbf{Y}=0,\mathbf{X}_{n_i^0 2} < f_{\boldsymbol{\theta}_i^0}(\mathbf{X}_{n_i^0 1})\right),\end{aligned}$$

this holds true for each $i, j = 1,2,\ldots,k+1$, which implies the rate of convergence of $\widehat{\boldsymbol{\lambda}}_j \;\; j = 1,2,\ldots,k$. The condition **(B2)** holds with

$$v_n^{-1} = \mathrm{K}[(nh)^{-1/2}(\log n)^{1/2} + h].$$

Conditions **(B3)** and **(B4)** hold by the Remark 3.6(ii) and (iii) respectively if they are satisfied for each true sub-sample which holds for this example. We conclude that

$$\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0 = \mathrm{O}_{\mathbb{P}}\left(n^{\frac{1}{3}}\right),$$

note that this rate verifies the last part of **(B5)**. Finally, we check the conditions needed for establishing the asymptotic distribution of $\widehat{\boldsymbol{\phi}}$. Condition **(C1)** follows from Theorem 5.3.2.3 and condition **(B2)**, whereas **(C2)** is immediately satisfied. Condition **(C3)** holds with the same method given for condition **(B3)** for each true sub-sample. For **(C4)** and **(C5)**, we remark the function $\psi_n(\delta) = \mathrm{K}\delta^{1/2}$ in condition **(B3)** is independent of n and continuous. **(C4)** and **(C5)** are therefore verified, provided the conditions set out in **Remarks 3.7**(iii) are verified. Next, condition **(C6)** easily follows from the fact that our functions $\mathbf{m}_j(z,\cdot,p) \;\; j = 1,2,\ldots,k+1$ are sums of indicator functions for fixed z and $p(\cdot)$. After for condition **(C7)**, it's satisfied provided that

$$|\Gamma(\boldsymbol{\phi}^0, h_0| < \infty,$$

following **Remarks 3.6**(iii) and **3.7**(v). Condition **(C8)** holds true by construction of the estimator $\widehat{\boldsymbol{\phi}}$. For condition **(C9)**, we note that for each $j = 1,2,\ldots,k+1$,

$$r_n \mathrm{W}_{j,n}(\boldsymbol{\gamma}) = r_n \Omega_j(\boldsymbol{\phi}^0, \widehat{p})\boldsymbol{\gamma} = o_{\mathbb{P}}(1),$$

provided $nh^3 = o(1)$ and $\frac{(\log n)^{3/2}}{nh^{3/2}} = o(1)$, with the same argument as in Delsol and Van Keilegom [2020] for their condition **(C9)** we obtain the result. **(C10)** can be demonstrated in a similar manner as **(B3)**. The asymptotic distribution of $r_n(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0)$ now follows from Theorem 5.3.3.2.

5.4 Numerical results

We provide numerical illustrations regarding the bias and the root mean-squared error (RMSE). The computing program codes were implemented in R. In our simulation, the scenario of two change-points is considered, i.e.,

$$\mathrm{X}_{i2} = \max(\min(\mathrm{U}+\epsilon, 1), 0),$$

where $U \sim \mathcal{U}[0,1]$, $\epsilon \sim \mathcal{U}[-r,r]$ for small value of $r > 0$, and $X_1 \sim \mathcal{U}[0,1]$, with X_1, ϵ and U are independent. Let

$$Y_i = \mathbb{1}\{U_i \geq f_{\theta_j}(X_{i1})\}, \qquad n_{j-1}+1 \leq i \leq n_j, \quad j \in \{1,2,3\}; i = 1,\ldots,n,$$

where $X_{11},\ldots,X_{n1}$ $(U_1,\ldots,U_n)$ are i.i.d. sample of X_1 (U) and $f_{\theta_j}(x_1) = \theta_j x_1$, for some $\boldsymbol{\theta} = (\theta_1,\theta_2,\theta_3)$, we define

$$p(x_1) = \mathbb{P}(\Delta = 1 \mid X_1 = x_1) = \alpha_0 + \alpha_1 (x_1 - 0.5)^2.$$

The data is composed of $(X_{i1}, X_{i2}, Y_i\Delta_i, \Delta_i)$ $i = 1,2,\ldots,n$ from the described model. For the bandwidth, we work with $h = \frac{c_h}{\sqrt{n}}$, which satisfies the requirements of regularity derived from the asymptotic theory and for kernel we worked with quadratic kernel; $K(u) = \frac{15}{16}\left(1-u^2\right)^2 \mathbb{1}\{|u| \leq 1\}$. In Table 5.1 and 5.2, we show the bias and the RMSE of the estimator $(\widehat{\boldsymbol{\theta}}, \widehat{n}_1, \widehat{n}_2)$ for two values of the size of the sample $n = 500$ and $n = 1000$, where we consider for two different values of n; $r = 0.1$ and $r = 0.2$, $\alpha_0 = 0.25, \alpha_0 = 0.5$ and $\alpha_0 = 0.75$, $\alpha_1 = 1$ and $c_h = 2, c_h = 3.5$ and $c_h = 5$. For the sample 500, we consider the true value (TV) of the parameter to be estimated is $(\boldsymbol{\theta}^0, n_1^0, n_2^0) = (0.75, 1, 1.5, 150, 350)$ and for the sample size 1000 is $(\boldsymbol{\theta}^0, n_1^0, n_2^0) = (0.75, 1, 1.5, 350, 650)$. The results are based on 1000 Monte Carlo runs. In the estimation of the parameter of the change point model as in any other inferential context, the greater the sample size, the better. From the following two tables we observe that both the bias and the RMSE are quite small, for moderate sample size. The results are better when α_0 increases or r decreases. From the results reported in tables, one can see that the estimation of $(\boldsymbol{\theta}, n_1, n_2)$ is not very sensitive to the choice of the bandwidth h. In order to extract methodological recommendations for the use of the procedures proposed in this work, it will be interesting to conduct extensive Monte Carlo experiments to compare our procedures with other scenarios presented in the literature, but this would go well beyond the scope of the present paper.

Table 5.1: M-estimators of the parameters of within segments and change-points, with sample size 500.

			$ch = 2.0$		$ch = 3.5$		$ch = 5.0$	
	α_0	**T.V**	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
r = 0.1	0.25	0.75	0.05	0.125	0.057	0.12	0.063	0.118
		1	0.016	0.145	0.002	0.146	-0.017	0.156
		1.5	0.113	0.199	0.107	0.188	0.116	0.196
		150	-2.197	9.818	-2.283	9.819	-3.507	10.459
		350	1.589	9.28	2.53	9.299	3.011	9.944
r = 0.1	0.5	0.75	0.075	0.116	0.079	0.118	0.086	0.121
		1	-0.027	0.155	-0.044	0.162	-0.055	0.17
		1.5	0.121	0.186	0.131	0.193	0.135	0.191
		150	-4.31	10.152	-5.265	10.843	-5.982	10.942
		350	3.85	10.019	4.954	10.345	5.374	10.705
r = 0.1	0.75	0.75	0.083	0.113	0.094	0.119	0.096	0.122
		1	-0.041	0.159	-0.058	0.172	-0.073	0.178
		1.5	0.124	0.178	0.146	0.187	0.149	0.19
		150	-5.093	10.35	-6.318	10.793	-6.825	11.101
		350	4.789	9.937	5.525	10.615	6.155	11.11
r = 0.2	0.25	0.75	0.051	0.147	0.053	0.141	0.058	0.144
		1	0.025	0.166	0.009	0.171	0	0.17
		1.5	0.113	0.214	0.117	0.209	0.129	0.214
		150	-2.332	10.424	-2.8	10.247	-3.312	10.452
		350	1.161	9.968	2.324	10.237	2.346	10.12
r = 0.2	0.5	0.75	0.066	0.128	0.072	0.128	0.078	0.127
		1	-0.016	0.166	-0.033	0.169	-0.048	0.174
		1.5	0.128	0.201	0.134	0.204	0.15	0.205
		150	-4.221	10.336	-4.899	10.858	-5.705	10.934
		350	3.64	10.113	4.635	10.481	5.45	10.607
r = 0.2	0.75	0.75	0.081	0.123	0.09	0.125	0.09	0.125
		1	-0.035	0.165	-0.054	0.178	-0.061	0.177
		1.5	0.134	0.192	0.156	0.201	0.158	0.203
		150	-4.682	10.504	-6.111	10.879	-6.472	11.009
		350	4.348	10.072	5.449	10.631	5.743	10.81

Table 5.2: M-estimators of the parameters of within segments and change-points, with sample size 1000.

			$ch = 2.0$		$ch = 3.5$		$ch = 5.0$	
	α_0	**T.V**	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
r = 0.1	0.25	0.75	0.072	0.106	0.068	0.107	0.072	0.107
		1	-0.005	0.149	-0.01	0.147	-0.016	0.15
		1.5	0.111	0.172	0.108	0.171	0.119	0.175
		350	-3.447	10.037	-3.83	10.159	-4.156	10.127
		350	2.474	9.567	2.753	9.632	3.157	9.757
r = 0.1	0.5	0.75	0.062	0.1	0.067	0.104	0.074	0.106
		1	0.005	0.132	-0.012	0.143	-0.024	0.149
		1.5	0.096	0.152	0.107	0.161	0.113	0.163
		350	-3.066	9.672	-3.678	10.087	-4.37	10.242
		350	1.748	9.272	2.478	9.906	3.102	10.146
r = 0.1	0.75	0.75	0.077	0.106	0.086	0.112	0.093	0.115
		1	-0.028	0.152	-0.049	0.164	-0.06	0.171
		1.5	0.119	0.166	0.132	0.176	0.138	0.179
		350	-3.882	10.105	-5.549	10.728	-5.923	10.878
		350	3.866	9.956	4.867	10.46	5.717	10.746
r = 0.2	0.25	0.75	0.069	0.121	0.066	0.119	0.066	0.117
		1	0.002	0.162	-0.004	0.161	-0.003	0.162
		1.5	0.126	0.196	0.123	0.189	0.121	0.192
		350	-3.678	10.385	-3.325	10.338	-3.768	10.293
		350	2.318	9.975	2.412	10.143	2.879	10.127
r = 0.2	0.5	0.75	0.062	0.11	0.064	0.112	0.07	0.115
		1	0.019	0.144	-0.003	0.151	-0.021	0.162
		1.5	0.114	0.164	0.118	0.172	0.134	0.182
		350	-2.85	9.24	-3.72	9.948	-4.647	10.189
		350	1.706	9.129	2.631	9.496	3.512	10.135
r = 0.2	0.75	0.75	0.078	0.111	0.085	0.116	0.088	0.118
		1	-0.018	0.156	-0.038	0.166	-0.048	0.173
		1.5	0.132	0.176	0.14	0.179	0.143	0.182
		350	-4.332	10.03	-5.271	10.329	-5.585	10.813
		350	3.473	9.887	4.675	10.273	4.881	10.457

5.5 Mathematical developments

This section is devoted to the proofs of our results. The previously defined notation continues to be used below. Before giving the proof of the Theorem 5.3.1.1, we start with two lemmas needed to establish the weak convergence and their convergence rate.

Lemma 5.5.0.1 *Under assumption (A4) and (A6), we have for any* $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ *and any* $\boldsymbol{\phi} \in \boldsymbol{\Phi}$

$$\mathbf{U} = o_{\mathbb{P}}(1).$$

This lemma is the basic idea for giving the results like consistency and the rate of convergence for our M-estimators, it shown that the variable $\mathbf{U}$ is defined in (5.2.2) is near zero when $\widehat{h}$ is close to h_0 for any $\boldsymbol{\phi}_j$ $j = 1,2,\ldots,k+1$ and any n_i $i = 1,2,\ldots,k$, when n grows, under general condition and even there is a change in the distribution.

Proof of Lemma 5.5.0.1

For any $\delta > 0$, under assumption **(A6)**, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}} |\mathbf{U}| > \delta\right) \\
&= \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}} \left|\frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}+1}^{n_j} \mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,\widehat{h}) - \mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0)\right| > \delta\right) \\
&\leq \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}} \left\{\frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}+1}^{n_j} |\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,\widehat{h}) - \mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0)|\right\} > \delta\right) \\
&\leq \mathbb{P}\left(\left\{\frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}^0+1}^{n_j^0} \mathfrak{G}(\mathbf{X}_i)\right\} d_{\mathscr{H}}(\widehat{h},h_0) > \frac{\delta}{2}\right).
\end{aligned}$$

We take the result from **(A4)** and the law of large numbers for i.i.d. variables since by assumptions, for any $i = 1,2,\ldots,k+1$, we have

$$\int \mathfrak{G}^2(x)\, d\mathrm{F}_{n_i^0}(x) < \infty.$$

Hence the proof is complete. □

Lemma 5.5.0.2 *Under assumption (A5) it follows: for any* $j = 1,2,\ldots,k+1$, *any* $0 \leq m_1 < m_2 \leq n$ *and any positive number* $\epsilon > 0$, *there exists a constant* A_j, *independent of* ϵ, *and a constant* $r > 2$, *such that*

$$\begin{aligned}
&\mathbb{P}\left(\max_{m_1\leq s<t\leq m_2,\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j,\boldsymbol{\alpha}\in\boldsymbol{\Upsilon}} \left|\sum_{i=s+1}^{t} \mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0) - \mathbb{E}[\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0)]\right| > \epsilon\right) \\
&\leq \mathrm{A}_j \frac{(m_2-m_1)^r}{\epsilon^2}.
\end{aligned} \tag{5.5.1}$$

Proof of Lemma 5.5.0.2

By the fact that the all variables are independent so with Assumption **(A4)** in mind, equation (5.5.1) can be achieved by induction with respect to m_2. The induction method is similar to the one used in Móricz *et al.* [1982], so its proof is omitted here. □

Proof of Theorem 5.3.1.1

Let us introduce

$$\begin{aligned}
\boldsymbol{\Lambda}_\eta &= \{\boldsymbol{\lambda} \in \boldsymbol{\Lambda} : \left\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\right\|_\infty > \eta\}, \\
\boldsymbol{\Phi}_\eta &= \{\boldsymbol{\phi} \in \boldsymbol{\Phi} : d(\boldsymbol{\phi}, \boldsymbol{\phi}^0) > \eta\}, \\
\boldsymbol{\Phi} &= \boldsymbol{\Upsilon} \times \boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2 \times \cdots \times \boldsymbol{\Theta}_{k+1}, \\
\boldsymbol{\Lambda} &= \{(\lambda_1, \lambda_2, \ldots, \lambda_k) | \lambda_j = \frac{n_j}{n} : j = 1, \ldots, k; 0 < n_1 < \cdots < n_k < n\}.
\end{aligned}$$

We have the following chain of inequalities

$$\begin{aligned}
&\mathbb{P}(\left\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0\right\|_\infty > \eta) \\
&\leq \mathbb{P}\left(\max_{\lambda \in \boldsymbol{\Lambda}_\eta, \boldsymbol{\phi} \in \boldsymbol{\Phi}} \mathfrak{W} > 0\right) \\
&= \mathbb{P}\left(\max_{\lambda \in \boldsymbol{\Lambda}_\eta, \boldsymbol{\phi} \in \boldsymbol{\Phi}} \{\mathbf{U} + \mathbf{L} + \mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}, h_0) - \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h_0)\} > 0\right) \\
&\leq \mathbb{P}\left(\max_{\lambda \in \boldsymbol{\Lambda}_\eta, \boldsymbol{\phi} \in \boldsymbol{\Phi}} |\mathbf{U} + \mathbf{L}| > \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h_0) - \max_{\lambda \in \boldsymbol{\Lambda}_\eta, \boldsymbol{\phi} \in \boldsymbol{\Phi}} \mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}, h_0)\right) \\
&\leq \mathbb{P}\left(\max_{\lambda \in \boldsymbol{\Lambda}_\eta, \boldsymbol{\phi} \in \boldsymbol{\Phi}} |\mathbf{U}| > \frac{\delta}{2}\right) + \mathbb{P}\left(\max_{\lambda \in \boldsymbol{\Lambda}_\eta, \boldsymbol{\phi} \in \boldsymbol{\Phi}} |\mathbf{L}| > \frac{\delta}{2}\right) \\
&\leq \mathbb{P}\left(\max_{\lambda \in \boldsymbol{\Lambda}_\eta, \boldsymbol{\phi} \in \boldsymbol{\Phi}} |\mathbf{U}| > \frac{\delta}{2}\right) \\
&\quad + \mathbb{P}\left(\max_{\lambda \in \boldsymbol{\Lambda}_\eta, \boldsymbol{\phi} \in \boldsymbol{\Phi}} \left| \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}+1}^{n_j} \left[\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0))\right] \right.\right. \\
&\quad \left.\left. - \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}^0+1}^{n_j^0} \left[\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}^0, \boldsymbol{\theta}_j^0, h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}^0, \boldsymbol{\theta}_j^0, h_0))\right] \right| > \frac{\delta}{2}\right) \\
&\leq \mathbb{P}\left(\max_{\lambda \in \boldsymbol{\Lambda}_\eta, \boldsymbol{\phi} \in \boldsymbol{\Phi}} |\mathbf{U}| > \frac{\delta}{2}\right) \\
&\quad + \mathbb{P}\left(\max_{\lambda \in \boldsymbol{\Lambda}_\eta, \boldsymbol{\phi} \in \boldsymbol{\Phi}} \left| \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}+1}^{n_j} \left[\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0))\right] \right| > \frac{\delta}{4}\right) \\
&\quad + \mathbb{P}\left(\left| \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}^0+1}^{n_j^0} \left[\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}^0, \boldsymbol{\theta}_j^0, h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}^0, \boldsymbol{\theta}_j^0, h_0))\right] \right| > \frac{\delta}{4}\right)
\end{aligned}$$

$$\leq \sum_{j=1}^{k+1} \mathbb{P}\Bigg(\max_{0\leq n_{j-1}<n_j\leq n, \boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j, \boldsymbol{\alpha}\in\boldsymbol{\Upsilon}} \frac{1}{n} \times \Bigg| \sum_{i=n_{j-1}+1}^{n_j} \left[\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0))\right] \Bigg| > \frac{\delta}{4(k+1)} \Bigg)$$

$$+ \sum_{j=1}^{k+1} \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=n_{j-1}^0+1}^{n_j^0} \left[\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_j^0,h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_j^0,h_0))\right] \right| > \frac{\delta}{4(k+1)} \right)$$

$$+ \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}} |\mathbf{U}| > \frac{\delta}{2} \right).$$

It follows from Lemma 5.5.0.1 and Lemma 5.5.0.2 the result. For $\widehat{\boldsymbol{\phi}}$, we similarly obtain

$$\begin{aligned}
&\mathbb{P}(d(\widehat{\boldsymbol{\phi}},\boldsymbol{\phi}^0) > \eta) \\
&\leq \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}_\eta} \mathfrak{W} > 0 \right) \\
&= \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}_\eta} \{\mathbf{U}+\mathbf{L}+\mathbf{M}(\boldsymbol{\phi},\boldsymbol{\lambda},h_0) - \mathbf{M}(\boldsymbol{\phi}^0,\boldsymbol{\lambda}^0,h_0)\} > 0 \right) \\
&\leq \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}_\eta} |\mathbf{U}+\mathbf{L}| > \mathbf{M}(\boldsymbol{\phi}^0,\boldsymbol{\lambda}^0,h_0) - \max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}_\eta} \mathbf{M}(\boldsymbol{\phi},\boldsymbol{\lambda},h_0) \right) \\
&\leq \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}_\eta} |\mathbf{U}| > \frac{\delta}{2} \right) + \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}_\eta} |\mathbf{L}| > \frac{\delta}{2} \right) \\
&\leq \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}_\eta} |\mathbf{U}| > \frac{\delta}{2} \right) \\
&\quad + \mathbb{P}\Bigg(\max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}_\eta} \Bigg| \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}+1}^{n_j} \left[\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0))\right] \\
&\qquad\qquad + \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}^0+1}^{n_j^0} \left[\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_j^0,h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_j^0,h_0))\right] \Bigg| > \frac{\delta}{2} \Bigg) \\
&\leq \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}_\eta} |\mathbf{U}| > \frac{\delta}{2} \right) \\
&\quad + \mathbb{P}\Bigg(\max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}_\eta} \Bigg| \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}+1}^{n_j} \left[\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0))\right] \Bigg| > \frac{\delta}{4} \Bigg) \\
&\quad + \mathbb{P}\left(\left| \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}^0+1}^{n_j^0} \left[\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_j^0,h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_j^0,h_0))\right] \right| > \frac{\delta}{4} \right) \\
&\leq \sum_{j=1}^{k+1} \mathbb{P}\Bigg(\max_{0\leq n_{j-1}<n_j\leq n, \boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j, \boldsymbol{\alpha}\in\boldsymbol{\Upsilon}} \frac{1}{n} \times \Bigg| \sum_{i=n_{j-1}+1}^{n_j} \left[\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0))\right] \Bigg| > \frac{\delta}{4(k+1)} \Bigg)
\end{aligned}$$

$$+\sum_{j=1}^{k+1}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=n_{j-1}^0+1}^{n_j^0}\left[\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_j^0,h_0)-\mathbb{E}(\mathbf{m}_j(\mathbf{X}_i,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_j^0,h_0))\right]\right|>\frac{\delta}{4(k+1)}\right)$$
$$+\mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda},\boldsymbol{\phi}\in\boldsymbol{\Phi}_\eta}|\mathbf{U}|>\frac{\delta}{2}\right).$$

It follows from Lemma 5.5.0.1 and Lemma 5.5.0.2 the desired result. □

Proof of Theorem 5.3.2.1

Let us first define, for any $\eta>0$,

$$\boldsymbol{\Lambda}_{\eta,n}=\{\lambda\in\boldsymbol{\Lambda}:n\left\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^0\right\|_\infty\geq\eta\}.$$

Because of the consistency of $\widehat{\lambda}$, we need to consider only those terms observations are in $\widetilde{n}_{j,j-1}$, $\widetilde{n}_{j,j}$ and $\widetilde{n}_{j,j+1}$ for all j in equation (5.2.3). Therefore we have

$$\begin{aligned}
&\mathbb{P}(n\left\|\widehat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}^0\right\|_\infty\geq\eta)\\
&\leq\ \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda}_{\eta,n},\boldsymbol{\phi}\in\boldsymbol{\Phi}}\mathfrak{W}>0\right)\\
&\leq\ \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda}_{\eta,n},\boldsymbol{\phi}\in\boldsymbol{\Phi}}\{\mathbf{U}+\mathbf{L}+\mathbf{M}(\boldsymbol{\phi},\boldsymbol{\lambda},h_0)-\mathbf{M}(\boldsymbol{\phi}^0,\boldsymbol{\lambda}^0,h_0)\}>0\right)\\
&\leq\ \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda}_{\eta,n},\boldsymbol{\phi}\in\boldsymbol{\Phi}}\left[\mathbf{U}+\frac{\mathbf{M}(\boldsymbol{\phi},\boldsymbol{\lambda},h_0)-\mathbf{M}(\boldsymbol{\phi}^0,\boldsymbol{\lambda}^0,h_0)}{2}\right]>0\right)\\
&\quad+\mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda}_{\eta,n},\boldsymbol{\phi}\in\boldsymbol{\Phi}}\left[\mathbf{L}+\frac{\mathbf{M}(\boldsymbol{\phi},\boldsymbol{\lambda},h_0)-\mathbf{M}(\boldsymbol{\phi}^0,\boldsymbol{\lambda}^0,h_0)}{2}\right]>0\right).
\end{aligned}$$

The second term is bounded by

$$\begin{aligned}
&\mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda}_{\eta,n},\boldsymbol{\phi}\in\boldsymbol{\Phi}}\left[\mathbf{L}+\frac{\mathbf{M}(\boldsymbol{\phi},\boldsymbol{\lambda},h_0)-\mathbf{M}(\boldsymbol{\phi}^0,\boldsymbol{\lambda}^0,h_0)}{2}\right]>0\right)\\
&\leq\ \sum_{j=1}^{k+1}\mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda}_{\eta,n},\boldsymbol{\phi}\in\boldsymbol{\Phi}}\left\{\frac{1}{n}\sum_{t\in\widetilde{n}_{jj}}\left[\mathbf{m}_j(\mathbf{X}_t,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0)-\mathbb{E}(\mathbf{m}_j(\mathbf{X}_t,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0))\right]\right.\right.\\
&\quad\left.\left.-\frac{1}{n}\sum_{t\in\widetilde{n}_{jj}}\left[m_i(\mathbf{X}_t,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_i^0,h_0)-\mathbb{E}(m_i(\mathbf{X}_t,\boldsymbol{\alpha}^0,\boldsymbol{\theta}_i^0,h_0))\right]+\frac{\mathbf{M}(\boldsymbol{\phi},\boldsymbol{\lambda},h_0)-\mathbf{M}(\boldsymbol{\phi}^0,\boldsymbol{\lambda}^0,h_0)}{6(k+1)}\right\}>0\right)\\
&\quad+\sum_{j=2}^{k+1}\mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda}_{\eta,n},\boldsymbol{\phi}\in\boldsymbol{\Phi}}\left\{\frac{1}{n}\sum_{t\in\widetilde{n}_{j,j-1}}\left[\mathbf{m}_j(\mathbf{X}_t,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0)-\mathbb{E}(\mathbf{m}_j(\mathbf{X}_t,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0))\right]\right.\right.
\end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{n}\sum_{t\in\widetilde{n}_{j,j-1}}\Big[m_{j-1}(\mathbf{X}_t,\boldsymbol{\alpha}^0,\boldsymbol{\theta}^0_{j-1},h_0)-\mathbb{E}(m_{j-1}(\mathbf{X}_t,\boldsymbol{\alpha}^0,\boldsymbol{\theta}^0_{j-1},h_0))\Big]+\frac{\mathbf{M}(\boldsymbol{\phi},\boldsymbol{\lambda},h_0)-\mathbf{M}(\boldsymbol{\phi}^0,\boldsymbol{\lambda}^0,h_0)}{6k}\Bigg\}>0\Bigg)\\
&+\sum_{j=1}^{k}\mathbb{P}\Bigg(\max_{\lambda\in\boldsymbol{\Lambda}_{\eta,n},\boldsymbol{\phi}\in\boldsymbol{\Phi}}\Bigg\{\frac{1}{n}\sum_{t\in\widetilde{n}_{j,j+1}}\big[\mathbf{m}_j(\mathbf{X}_t,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0)-\mathbb{E}(\mathbf{m}_j(\mathbf{X}_t,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0))\big]\\
&-\frac{1}{n}\sum_{t\in\widetilde{n}_{j,j+1}}\Big[m_{j+1}(\mathbf{X}_t,\boldsymbol{\alpha}^0,\boldsymbol{\theta}^0_{j+1},h_0)-\mathbb{E}(m_{j+1}(\mathbf{X}_t,\boldsymbol{\alpha}^0,\boldsymbol{\theta}^0_{j+1},h_0))\Big]+\frac{\mathbf{M}(\boldsymbol{\phi},\boldsymbol{\lambda},h_0)-\mathbf{M}(\boldsymbol{\phi}^0,\boldsymbol{\lambda}^0,h_0)}{6k}\Bigg\}>0\Bigg)\\
\equiv\;&\sum_{j=1}^{k+1}\mathrm{I}_{1j}+\sum_{j=2}^{k+1}\mathrm{I}_{2j}+\sum_{j=1}^{k}\mathrm{I}_{3j}. \qquad (5.5.2)
\end{aligned}
$$

First, consider the probability formula I_{1j} in the above equation for any $j=1,2,\ldots,k+1$. The consistency of $\widehat{\boldsymbol{\lambda}}$ allows us to restrict our attention to the case $n_{jj}>\frac{1}{2}(n^0_j-n^0_{j-1})$. For this case, there exists a constant $\mathfrak{C}>0$ such that

$$
\begin{aligned}
\mathbf{M}(\boldsymbol{\phi},\boldsymbol{\lambda},h_0)-\mathbf{M}(\boldsymbol{\phi}^0,\boldsymbol{\lambda}^0,h_0)&=\sum_{j=1}^{k+1}\sum_{i=1}^{k+1}\frac{n_{ji}}{n}\int_{\mathbb{R}}[\mathbf{m}_j(x,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0)-m_i(x,\boldsymbol{\alpha}^0,\boldsymbol{\theta}^0_i,h_0)]d\mathrm{F}_{n^0_i}(x)\\
&\le\frac{n^0_j-n^0_{j-1}}{2n}\mathbb{E}(\mathbf{m}_j(\mathbf{X},\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0)-\mathbf{m}_j(\mathbf{X},\boldsymbol{\alpha}^0,\boldsymbol{\theta}^0_j,h_0))\\
&\le-\mathfrak{C}\frac{n^0_j-n^0_{j-1}}{2n}. \qquad (5.5.3)
\end{aligned}
$$

Therefore the probability I_{1j} is upper bounded by

$$
\begin{aligned}
\mathrm{I}_{1j}\;\le\;&\mathbb{P}\Bigg(\max_{n^0_{j-1}\le s<t\le n^0_j,\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j,\boldsymbol{\alpha}\in\boldsymbol{\Upsilon}}\sum_{i=s+1}^{t}\big[\mathbf{m}_j(\mathbf{X}_t,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0)-\mathbb{E}(\mathbf{m}_j(\mathbf{X}_t,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0))\big]>\mathfrak{C}\frac{n^0_j-n^0_{j-1}}{12(k+1)}\Bigg)\\
&+\mathbb{P}\Bigg(\sum_{i=s+1}^{t}\big[m_i(\mathbf{X}_t,\boldsymbol{\alpha}^0,\boldsymbol{\theta}^0_i,h_0)-\mathbb{E}(m_i(\mathbf{X}_t,\boldsymbol{\alpha}^0,\boldsymbol{\theta}^0_i,h_0))\big]>\mathfrak{C}\frac{n^0_j-n^0_{j-1}}{12(k+1)}\Bigg).
\end{aligned}
$$

The result in Lemma 5.5.0.2 shows that $\mathrm{I}_{1j}\longrightarrow 0$ as $n,\eta\longrightarrow\infty$. Next, we consider I_{2j} in (5.5.2), for any $j=2,\ldots,k+1$. For this case we have $\lambda_{j-1}<\lambda^0_{j-1}$, and I_{2j} is bounded by

$$
\begin{aligned}
\mathrm{I}_{2j}\;\le\;&\mathbb{P}\Bigg(\max_{\lambda\in\boldsymbol{\Lambda}_{\eta,n},\boldsymbol{\phi}\in\boldsymbol{\Phi}}\Bigg\{\frac{1}{n}\sum_{t\in\widetilde{n}_{j,j-1}}\big[\mathbf{m}_j(\mathbf{X}_t,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0)-\mathbb{E}(\mathbf{m}_j(\mathbf{X}_t,\boldsymbol{\alpha},\boldsymbol{\theta}_j,h_0))\big]\\
&+\frac{\mathbf{M}(\boldsymbol{\phi},\boldsymbol{\lambda},h_0)-\mathbf{M}(\boldsymbol{\phi}^0,\boldsymbol{\lambda}^0,h_0)}{12k}\Bigg\}>0\Bigg)\\
&+\mathbb{P}\Bigg(\max_{\lambda\in\boldsymbol{\Lambda}_{\eta,n},\boldsymbol{\phi}\in\boldsymbol{\Phi}}\Bigg\{\frac{1}{n}\sum_{t\in\widetilde{n}_{j,j-1}}\big[m_{j-1}(\mathbf{X}_t,\boldsymbol{\alpha},\boldsymbol{\theta}_{j-1},h_0)-\mathbb{E}(m_{j-1}(\mathbf{X}_t,\boldsymbol{\alpha},\boldsymbol{\theta}_{j-1},h_0))\big]\\
&+\frac{\mathbf{M}(\boldsymbol{\phi},\boldsymbol{\lambda},h_0)-\mathbf{M}(\boldsymbol{\phi}^0,\boldsymbol{\lambda}^0,h_0)}{12k}\Bigg\}>0\Bigg)\\
\equiv\;&\mathrm{I}^{(1)}_{2j}+\mathrm{I}^{(2)}_{2j}.
\end{aligned}
$$

With the same method we show that $\mathrm{I}^{(1)}_{2j}$ and $\mathrm{I}^{(2)}_{2j}$ are negligible for n and η grow to infinity, so we just treat the first term. Only two cases have to be considered.

If $n^0_{j-1} - n_{j-1} \leq \eta$, then

$$\begin{aligned} \mathrm{I}^{(1)}_{2j} &\leq \mathbb{P}\left(\max_{n_{j-1}\leq s<t\leq n^0_{j-1}, \boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j, \boldsymbol{\alpha}\in\boldsymbol{\Upsilon}} \left| \sum_{i=s+1}^{t} \left[\mathbf{m}_j(\mathbf{X}_t, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_t, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0)) \right] \right| > \frac{\mathrm{C}_1\eta}{12k} \right) \\ &\leq \frac{(n^0_{j-1} - n_{j-1})^r}{(\mathrm{C}_1\eta)^2}(12k)^2 \leq \eta^{r-2}\left(\frac{12k}{\mathrm{C}_1}\right)^2. \end{aligned}$$

If $n^0_{j-1} - n_{j-1} > \eta$, for the other case, then

$$\mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}, h_0) - \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h_0) \leq -\mathrm{C}_1 \frac{(n^0_{j-1} - n_{j-1})}{n}.$$

Therefore, we obtain that

$$\begin{aligned} &\mathrm{I}^{(1)}_{2j} \\ &\leq \mathbb{P}\left(\max_{n_{j-1}\leq s<t\leq n^0_{j-1}, \boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j, \boldsymbol{\alpha}\in\boldsymbol{\Upsilon}} \left| \sum_{i=s+1}^{t} \left[\mathbf{m}_j(\mathbf{X}_t, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0) - \mathbb{E}(\mathbf{m}_j(\mathbf{X}_t, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0)) \right] \right| > \frac{\mathrm{C}_1(n^0_{j-1} - n_{j-1})}{12k} \right) \\ &\leq (n^0_{j-1} - n_{j-1})^{r-2}\left(\frac{12k}{\mathrm{C}_1}\right)^2. \end{aligned}$$

The result is a direct consequence of Lemma 5.5.0.2. In the same way we can prove the same result for I_{3j}. For the first term we obtain from 5.5.3 and assumption **(B1)** that

$$\begin{aligned} &\mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda}_{\eta,n}, \boldsymbol{\phi}\in\boldsymbol{\Phi}} \left[\mathbf{U} + \frac{\mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}, h_0) - \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h_0)}{2} \right] > 0 \right) \\ &\leq \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda}_{\eta,n}, \boldsymbol{\phi}\in\boldsymbol{\Phi}} \left[\frac{1}{n}\sum_{j=1}^{k+1} \sum_{i=n_{j-1}+1}^{n_j} \mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, \widehat{h}) - \mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0) \right] > \mathfrak{C}\frac{\lambda^0_j - \lambda^0_{j-1}}{2} \right) \\ &\leq \mathbb{P}\left(\max_{\lambda\in\boldsymbol{\Lambda}, \boldsymbol{\phi}\in\boldsymbol{\Phi}} \left[\frac{1}{n}\sum_{j=1}^{k+1} \sum_{i=n_{j-1}+1}^{n_j} \mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, \widehat{h}) - \mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h_0) \right] > \mathfrak{C}' \right), \end{aligned}$$

the last term converges to zero as $n \longrightarrow \infty$ by applying Lemma 5.5.0.1. Therefore Theorem 5.3.2.1 is proved. □

Proof of Theorem 5.3.2.3

Recall that ξ_n is the $\mathrm{O}_{\mathbb{P}}(r_n^{-2})$-quantity involved in assumption **(B5)**. We introduce the sets

$$\mathrm{S}_{j,n} = \{\boldsymbol{\phi} \in \boldsymbol{\Phi} : 2^{j-1} < r_n d(\boldsymbol{\phi}, \boldsymbol{\phi}^0) \leq 2^j\},$$

we observe

$$\boldsymbol{\Phi}\backslash\boldsymbol{\phi}^0 = \cup_{j=1}^{\infty}\mathrm{S}_{j,n}.$$

Our objective is to show that for any $\epsilon > 0$ there exist $\tau_\epsilon > 0$ such that

$$\mathbb{P}(r_n d(\widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}^0) > \tau_\epsilon) < \epsilon, \tag{5.5.4}$$

for any n sufficiently large. In the next we work with arbitrary fixed ϵ. For any $\delta, \delta_1, K, K^{'}, K^{''}$, $K_1, K_2 > 0$, we obtain the following bound using condition **(B5)** and the result in Theorem 5.3.2.1

$$\begin{aligned}
&\mathbb{P}(r_n d(\widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}^0) > 2^M) \\
&\leq \sum_{M \leq j, 2^j \leq \delta r_n} \mathbb{P}\left(\sup_{\boldsymbol{\phi} \in S_{j,n}} [\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h})] \geq -K r_n^{-2}, A_n \right) \\
&\quad + \mathbb{P}\left(2 d(\widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}^0) \geq \delta\right) + \mathbb{P}\left(r_n^2 |\xi_n| > K^{'}\right) + \mathbb{P}\left(r_n |W_n| > K^{''}\right) + \mathbb{P}\left(|\beta_n| > \frac{C}{2}\right) \\
&\quad + \mathbb{P}\left(d_{\mathscr{H}}(\widehat{h}, h_0) > \frac{\delta_1}{v_n}\right) + \mathbb{P}\left(n \left\| \widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 \right\| > K_2\right),
\end{aligned}$$

where

$$A_n = \left\{ r_n |W_n| \leq K^{''}, |\beta_n| \leq \frac{C}{2}, d_{\mathscr{H}}(\widehat{h}, h_0) \leq \frac{\delta_1}{v_n}, n \left\| \widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 \right\| \leq K_2 \right\}.$$

Indeed, we can write

$$\begin{aligned}
&\mathbb{P}\left(r_n d(\widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}^0) > 2^M, 2 d(\widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}^0) < \delta, r_n^2 |\xi_n| \leq K^{'}, n \left\| \widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 \right\| \leq K_2, A_n \right) \\
&\leq \sum_{M \leq j, 2^j \leq \delta r_n} \mathbb{P}\left(\widehat{\boldsymbol{\phi}} \in S_{j,n}, r_n^2 |\xi_n| \leq K^{'}, n \left\| \widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 \right\| \leq K_2, A_n \right) \\
&\leq \sum_{M \leq j, 2^j \leq \delta r_n} \mathbb{P}\left(\sup_{\boldsymbol{\phi} \in S_{j,n}, \lambda \in \boldsymbol{\Lambda}} [\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}, \widehat{h}) - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h})] \geq -K^{'} r_n^{-2}, n \left\| \widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 \right\| \leq K_2, A_n \right) \\
&= \sum_{M \leq j, 2^j \leq \delta r_n} \mathbb{P}\left(\sup_{\boldsymbol{\phi} \in S_{j,n}, \lambda \in \boldsymbol{\Lambda}} [\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}, \widehat{h}) - \mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) + \mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) \right. \\
&\qquad \left. - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h})] \geq -K^{'} r_n^{-2}, n \left\| \widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 \right\| \leq K_2, A_n \right) \\
&\leq \sum_{M \leq j, 2^j \leq \delta r_n} \mathbb{P}\left(\sup_{\boldsymbol{\phi} \in \boldsymbol{\Phi}, \lambda \in \boldsymbol{\Lambda}} [\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}, \widehat{h}) - \mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h})] + \sup_{\boldsymbol{\phi} \in S_{j,n}} [\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) \right. \\
&\qquad \left. - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h})] \geq -K^{'} r_n^{-2}, n \left\| \widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 \right\| \leq K_2, A_n \right) \\
&\leq \sum_{M \leq j, 2^j \leq \delta r_n} \mathbb{P}\left(\sup_{\boldsymbol{\phi} \in S_{j,n}} [\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h})] \geq -K r_n^{-2}, A_n \right).
\end{aligned}$$

Note that the passage from the before last expression to the last is justified by the result in lemma (5.5.0.3). Assumption **(B2)** implies for all $\delta > 0$ there exists n_ϵ, such that, for $n > n_\epsilon$,

$$\mathbb{P}(2 d(\widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}^0) \geq \delta) < \frac{\epsilon}{7}.$$

By definition of ξ_n, W_n, under assumption **(B2)** and the result of theorem 5.3.2.1, there exist $\delta_1, k_\epsilon^{'}, k_\epsilon^{''}$ and $K_{2,\epsilon}$ such that

$$\begin{aligned}
&\mathbb{P}\left(r_n^2 |\xi_n| > K_\epsilon^{'}\right) < \tfrac{\epsilon}{7}, \quad \mathbb{P}\left(r_n |W_n| > K_\epsilon^{''}\right), \quad \mathbb{P}\left(|\beta_n| > \tfrac{C}{2}\right) < \tfrac{\epsilon}{7}, \\
&\mathbb{P}\left(d_{\mathscr{H}}(\widehat{h}, h_0) > \tfrac{\delta_1}{v_n}\right) < \tfrac{\epsilon}{7} \quad \text{and} \quad \mathbb{P}\left(n \left\| \widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 \right\| > K_{2,\epsilon}\right) < \tfrac{\epsilon}{7}.
\end{aligned} \tag{5.5.5}$$

For n large than some n_1. We fix $\delta < \delta_0$ and suppose $n \geq \max(n_0, n_1, n_\epsilon)$, for $2^j \leq \delta r_n$ we have the assumption **(B3)** and **(B4)** are fulfilled on all $S_{j,n}$. For each fixed j such that $2^j \leq \delta r_n$, we

have under assumption **(B4)**, for all $\boldsymbol{\phi} \in \mathrm{S}_{j,n}$:

$$\begin{aligned}
&\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h}) \\
&\quad \leq \quad \mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) - \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h}) \\
&\qquad + \sup_{d(\boldsymbol{\phi}, \boldsymbol{\phi}^0) \leq \frac{2^j}{r_n}} |\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h}) - \mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) + \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h})| \\
&\quad \leq \quad |\mathrm{W}_n| \frac{2^j}{r_n} - (\mathrm{C} - \beta_n) \frac{2^{2j-2}}{r_n^2} \\
&\qquad + \sup_{d(\boldsymbol{\phi}, \boldsymbol{\phi}^0) \leq \frac{2^j}{r_n}} |\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h}) - \mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) + \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h})|.
\end{aligned}$$

Consequently, we obtain the following inequality:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{\phi} \in \mathrm{S}_{j,n}} [\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h})] \geq -\mathrm{K} r_n^{-2}, \mathrm{A}_n \right) \\
&\quad \leq \quad \mathbb{P}\left(\sup_{d(\boldsymbol{\phi}, \boldsymbol{\phi}^0) \leq \frac{2^j}{r_n}, d_{\mathscr{H}}(h, h_0) \leq \frac{\delta_1}{v_n}} |\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, h) - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h) - \mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, h) + \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h)| \right. \\
&\qquad \left. \geq \frac{2^{2j-2}}{r_n^2} \left(\frac{\mathrm{C}}{2} - \mathrm{K}_\epsilon'' 2^{2-j} - \mathrm{K}_\epsilon' 2^{2-2j} \right) \right).
\end{aligned}$$

Now, there exists M_ϵ such that for all $j \geq \mathrm{M}_\epsilon$, we have

$$\frac{\mathrm{C}}{2} - \mathrm{K}_\epsilon'' 2^{2-j} - \mathrm{K}_\epsilon' 2^{2-2j} \geq \frac{\mathrm{C}}{4}.$$

By consequent, if $\mathrm{M} \geq \mathrm{M}_\epsilon$, using assumption **(B3)** in combination with Chebyshev's inequality we readily obtain

$$\begin{aligned}
&\sum_{\mathrm{M} \leq j, 2^j \leq \delta r_n} \mathbb{P}\left(\sup_{\boldsymbol{\phi} \in \mathrm{S}_{j,n}} [\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, \widehat{h}) - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, \widehat{h})] \geq -\mathrm{K} r_n^{-2}, \mathrm{A}_n \right) \\
&\quad \leq \quad \sum_{\mathrm{M} \leq j, 2^j \leq \delta r_n} \mathbb{P}\left(\sup_{d(\boldsymbol{\phi}, \boldsymbol{\phi}^0) \leq \frac{2^j}{r_n}, d_{\mathscr{H}}(h, h_0) \leq \frac{\delta_1}{v_n}} |\mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, h) - \mathbf{M}_n(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h) \right. \\
&\qquad \left. -\mathbf{M}(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, h) + \mathbf{M}(\boldsymbol{\phi}^0, \boldsymbol{\lambda}^0, h)| \geq \frac{\mathrm{C} 2^{2j-2}}{4 r_n^2} \right) \\
&\quad \leq \quad \sum_{\mathrm{M} \leq j, 2^j \leq \delta r_n} \frac{4\mathrm{K} r_n^2}{\mathrm{C} 2^{2j-2}} \frac{\psi_n(\frac{2^j}{r_n})}{\sqrt{n}} \\
&\quad \leq \quad \frac{4\mathrm{K} r_n^2}{\mathrm{C}\sqrt{n}} \sum_{\mathrm{M} \leq j, 2^j \leq \delta r_n} \frac{2^{j\boldsymbol{\alpha}} \psi_n(\frac{1}{r_n})}{2^{2j-2}} \\
&\quad \leq \quad \frac{16\mathrm{K}}{\mathrm{C}} \sum_{\mathrm{M} \leq j} 2^{j(\boldsymbol{\alpha}-2)}.
\end{aligned}$$

By the fact that $\boldsymbol{\alpha} < 2$, the series $\sum_{M \leq j} 2^{j(\boldsymbol{\alpha}-2)}$ converges, so there exists $M_\epsilon^{'} \geq M_\epsilon$, such that

$$\frac{16K}{C} \sum_{M \leq j} 2^{j(\boldsymbol{\alpha}-2)} \leq \frac{\epsilon}{7}.$$

The theorem is proved by choosing $\tau_\epsilon = 2^{M_\epsilon^{'}}$ in (5.5.4). □

Lemma 5.5.0.3 *Under conditions **(A3)-(A6)** and **(B1)** we obtain : for any $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ and $h \in \mathscr{H}$*

$$\mathbf{M}_n(\boldsymbol{\phi}, \widehat{\boldsymbol{\lambda}}, h) = \mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, h) + O_{\mathbb{P}}\left(\frac{1}{n}\right).$$

Proof of Lemma 5.5.0.3

We have the following decomposition

$$\begin{aligned}
&\mathbf{M}_n(\boldsymbol{\phi}, \widehat{\boldsymbol{\lambda}}, h) - \mathbf{M}_n(\boldsymbol{\phi}, \boldsymbol{\lambda}^0, h) \\
&= \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=\widehat{n}_{j-1}+1}^{\widehat{n}_j} \mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h) - \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}^0+1}^{n_j^0} \mathbf{m}_j(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h) \\
&= \frac{1}{n} \sum_{j=1}^{k+1} \Bigg\{ \mathbb{1}_{\{n_j^0 \leq \widehat{n}_j, n_{j-1}^0 \leq \widehat{n}_{j-1}\}} \left[\sum_{i=n_j^0+1}^{\widehat{n}_j} \mathbf{m}_j(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h) - \sum_{i=n_{j-1}^0+1}^{\widehat{n}_{j-1}} \mathbf{m}_j(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h) \right] \\
&\quad + \mathbb{1}_{\{n_j^0 < \widehat{n}_j, \widehat{n}_{j-1} < n_{j-1}^0\}} \left[\sum_{i=n_j^0+1}^{\widehat{n}_j} \mathbf{m}_j(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h) + \sum_{i=\widehat{n}_{j-1}+1}^{n_{j-1}^0} \mathbf{m}_j(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h) \right] \\
&\quad + \mathbb{1}_{\{\widehat{n}_j < n_j^0, n_{j-1}^0 \leq \widehat{n}_{j-1}\}} \left[- \sum_{i=\widehat{n}_j+1}^{n_j^0} \mathbf{m}_j(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h) - \sum_{i=n_{j-1}^0+1}^{\widehat{n}_{j-1}} \mathbf{m}_j(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h) \right] \\
&\quad + \mathbb{1}_{\{\widehat{n}_j \leq n_j^0, \widehat{n}_{j-1} < n_{j-1}^0\}} \left[- \sum_{i=\widehat{n}_j+1}^{n_j^0} \mathbf{m}_j(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h) + \sum_{i=\widehat{n}_{j-1}+1}^{n_{j-1}^0} \mathbf{m}_j(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\theta}_j, h) \right] \Bigg\}.
\end{aligned}$$

It follows from Theorem 5.3.2.1 the desired result. □

Proof of Theorem 5.3.3.2

In the first we prove the weak convergence of the process

$$\boldsymbol{\gamma} \mapsto r_n^2 \mathbf{B}_n\left(\boldsymbol{\phi}^0 + \frac{\boldsymbol{\gamma}}{r_n}, \widehat{\boldsymbol{\lambda}}, \widehat{h}\right),$$

which is proved in Lemma 5.5.0.4. The rest of the proof is based on somewhat similar arguments as those used to state the Argmax theorem in van der Vaart and Wellner [1996] without a change points, where the weak convergence of the empirical process implies the convergence in

distribution of its point of maximum, the M-estimators. Note that the set $\mathbf{E}$ is σ-compact metric space i.e.,

$$\mathbf{E} = \cup_{j=1}^{\infty} \mathcal{K}_j,$$

where, for any positive sequence $(a_j)_{j\in\mathbb{N}^*}$,

$$\mathcal{K}_j = \{\boldsymbol{\gamma} \in \mathbf{E} : \|\boldsymbol{\gamma}\| \leq a_j\}.$$

After we deduce from assumption **(C9)**, Lemma 5.5.0.5 and Lemma 5.5.0.6 together that almost all paths of the limiting process $\boldsymbol{\gamma} \mapsto \Gamma(\boldsymbol{\gamma}) + \mathbb{G}(\boldsymbol{\gamma})$ attain their supreme at the unique point $\boldsymbol{\gamma}^0$, following the same ideas in the parametric case without change points (see Theorem 3.2.10 in van der Vaart and Wellner [1996]). We assume now that $\boldsymbol{\gamma}^0$ is measurable. The weak convergence of $r_n(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0)$ to $\boldsymbol{\gamma}^0$ is equivalent to the statement (Portmanteau's Theorem) :

$$\limsup_{n\longrightarrow\infty} \mathbb{P}\left(r_n(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0) \in \mathrm{C}\right) \leq \mathbb{P}\left(\boldsymbol{\gamma}^0 \in \mathrm{C}\right), \quad \text{for every closed set } \mathrm{C}.$$

Let C be an arbitrary closed subset of $\mathbf{E}$ and fix $\epsilon > 0$. The set $\mathbf{E}$ is σ-compact and which implies that the random $\boldsymbol{\gamma}^0$ is tight combining this with the assumption **(C1)** we can find $\mathrm{K}_\epsilon > 0$ and a compact set

$$\mathcal{K}_\epsilon = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma}\| \leq \mathrm{K}_\epsilon\},$$

such that

$$\mathbb{P}(\boldsymbol{\gamma}^0 \notin \mathrm{K}_\epsilon) \leq \epsilon/2 \quad \text{and} \quad \mathbb{P}(r_n(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0) \notin \mathrm{K}_\epsilon) \leq \epsilon/2.$$

It follows from these last expressions

$$\begin{aligned}
&\limsup_{n\longrightarrow\infty} \mathbb{P}\left(r_n(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0) \in \mathrm{C}\right) \\
&\quad\leq \quad \mathbb{P}\left(r_n(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0) \in \mathrm{C} \cap \mathrm{K}_\epsilon, \boldsymbol{\gamma}^0 \in \mathrm{K}_\epsilon\right) \\
&\qquad\quad + \limsup_{n\longrightarrow\infty} \mathbb{P}\left(\{r_n(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0) \notin \mathrm{K}_\epsilon\} \cup \{\boldsymbol{\gamma}^0 \notin \mathrm{K}_\epsilon\}\right) \\
&\quad\leq \quad \mathbb{P}\left(r_n(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0) \in \mathrm{C} \cap \mathrm{K}_\epsilon, \boldsymbol{\gamma}^0 \in \mathrm{K}_\epsilon\right) + \epsilon. \qquad (5.5.6)
\end{aligned}$$

Now we use Lemma 5.5.0.4 and assumption **(C8)** we obtain that

$$\begin{aligned}
&\limsup_{n\longrightarrow\infty} \mathbb{P}\left(r_n(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0) \in \mathrm{C} \cap \mathrm{K}_\epsilon, \boldsymbol{\gamma}^0 \in \mathrm{K}_\epsilon\right) \\
&\quad\leq \quad \limsup_{n\longrightarrow\infty} \mathbb{P}\left(\sup_{\boldsymbol{\gamma}\in\mathrm{K}_\epsilon\cap\mathrm{C}} r_n^2 \mathbf{B}_n\left(\boldsymbol{\phi}^0 + \frac{\boldsymbol{\gamma}}{r_n}, \widehat{\boldsymbol{\lambda}}, \widehat{h}\right) \geq \sup_{\boldsymbol{\gamma}\in\mathrm{K}_\epsilon} r_n^2 \mathbf{B}_n\left(\boldsymbol{\phi}^0 + \frac{\boldsymbol{\gamma}}{r_n}, \widehat{\boldsymbol{\lambda}}, \widehat{h}\right) + o_{\mathbb{P}}(1), \boldsymbol{\gamma}^0 \in \mathrm{K}_\epsilon\right) \\
&\quad\leq \quad \mathbb{P}\left(\sup_{\boldsymbol{\gamma}\in\mathrm{K}_\epsilon\cap\mathrm{C}} (\Gamma + \mathbb{G})(\boldsymbol{\gamma}) \geq \sup_{\boldsymbol{\gamma}\in\mathrm{K}_\epsilon} (\Gamma + \mathbb{G})(\boldsymbol{\gamma}), \boldsymbol{\gamma}^0 \in \mathrm{K}_\epsilon\right), \qquad (5.5.7)
\end{aligned}$$

by Slutskys lemma and Portmanteaus theorem. On the other hand, for every open set G containing $\boldsymbol{\gamma}^0$, we have

$$(\Gamma + \mathbb{G})(\boldsymbol{\gamma}^0) > \sup_{\boldsymbol{\gamma}\in\mathrm{G}^\mathrm{C}\cap\mathrm{K}_\epsilon} (\Gamma + \mathbb{G})(\boldsymbol{\gamma}).$$

This together with (5.5.7) imply

$$\limsup_{n\longrightarrow\infty}\mathbb{P}\left(r_n(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^0)\in \mathrm{C}\cap \mathrm{K}_\epsilon, \boldsymbol{\gamma}^0\in \mathrm{K}_\epsilon\right)\leq \mathbb{P}\left(\boldsymbol{\gamma}^0\in \mathrm{C}\right).$$

Consequently, it follows from (5.5.6) that for all $\epsilon>0$,

$$\limsup_{n\longrightarrow\infty}\mathbb{P}\left(r_n(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^0)\in \mathrm{C}\right)\leq \mathbb{P}\left(\boldsymbol{\gamma}^0\in \mathrm{C}\right)+\epsilon.$$

The last inequality hold for every $\epsilon>0$, so it also holds for $\epsilon=0$. Consequently, the result holds from Portmanteau theorem. □

Lemma 5.5.0.4 *For all* $\mathrm{K}>0$*, let* $\mathscr{K}=\{\boldsymbol{\gamma}\in\mathbf{E}:\|\boldsymbol{\gamma}\|\leq \mathrm{K}\}$ *be a compact subset of* $\mathbf{E}$*. Then, under assumptions of Theorem 5.3.3.2, the process*

$$\boldsymbol{\gamma}\mapsto r_n^2\mathbf{B}_n\left(\boldsymbol{\phi}^0+\frac{\boldsymbol{\gamma}}{r_n},\widehat{\boldsymbol{\lambda}},\widehat{h}\right)$$

converges weakly to the process

$$\boldsymbol{\gamma}\mapsto \Gamma(\boldsymbol{\gamma})+\mathbb{G}(\boldsymbol{\gamma}) \text{ in } \ell^\infty(\mathscr{K}).$$

Moreover, almost all paths of the limiting process are continuous (uniformly on every compact $\mathscr{K}$*) with respect to* $\|\cdot\|$*.*

Proof of Lemma 5.5.0.4

The result of this lemma follows directly from Slutsky's theorem, Lemma 5.5.0.5 and Lemma 5.5.0.6. On the other hand, $\|\cdot\|$ makes $\mathscr{K}$ totally bounded (since it is compact) and

$$\boldsymbol{\gamma}\mapsto r_n^2\mathbf{B}_n\left(\boldsymbol{\phi}^0+\frac{\boldsymbol{\gamma}}{r_n},\boldsymbol{\lambda}^0,h_0\right)+r_n\mathrm{W}_n(\boldsymbol{\gamma})$$

is asymptotically uniformly $\|\cdot\|$-equicontinuous in probability, asymptotically tight and converges weakly to

$$\boldsymbol{\gamma}\mapsto \Gamma(\boldsymbol{\gamma})+\mathbb{G}(\boldsymbol{\gamma})$$

in $\ell^\infty(\mathscr{K})$ (see proof of Lemma 5.5.0.6). Thus, almost all paths of the limiting process are uniformly $\|\cdot\|$-continuous on $\mathscr{K}$ (see Theorem 1.5.7 in van der Vaart and Wellner [1996]). Moreover, because $\mathbf{E}$ may be covered by a countable sequence of such compact sets, almost all paths of the limiting process are $\|\cdot\|$-continuous on $\mathbf{E}$. □

Lemma 5.5.0.5 *Let* $\mathscr{K}=\{\boldsymbol{\gamma}\in\mathbf{E}:\|\boldsymbol{\gamma}\|\leq \mathrm{K}\}$*. Then under assumptions of Theorem 5.3.2.1 and assumptions of Theorem 5.3.3.2 we have respectively :*

1. $\mathbf{B}_n\left(\boldsymbol{\phi}^0+\frac{\boldsymbol{\gamma}}{r_n},\widehat{\boldsymbol{\lambda}},\widehat{h}\right)=\mathbf{B}_n\left(\boldsymbol{\phi}^0+\frac{\boldsymbol{\gamma}}{r_n},\boldsymbol{\lambda}^0,\widehat{h}\right)+\mathrm{O}_{\mathbb{P}}(n^{-1})$.
2. *There exist* $\xi_{1,n},\xi_{2,n},\xi_{3,n}$ *such that* $\sup_{\boldsymbol{\gamma}\in\mathscr{K}}|\xi_{l,n}|=o_{\mathbb{P}}(1)$*,* $l=1,2,3$*, and*

$$r_n^2\mathbf{B}_n\left(\boldsymbol{\phi}^0+\frac{\boldsymbol{\gamma}}{r_n},\boldsymbol{\lambda}^0,\widehat{h}\right)(1+\xi_{1,n})=\left[r_n^2\mathbf{B}_n\left(\boldsymbol{\phi}^0+\frac{\boldsymbol{\gamma}}{r_n},\boldsymbol{\lambda}^0,h_0\right)+r_n\mathrm{W}_n(\boldsymbol{\gamma})\right](1+\xi_{2,n})+\xi_{3,n}.$$

Proof of Lemma 5.5.0.5

The first assertion is a direct application of Lemma 5.5.0.3. We have

$$\begin{aligned}\mathbf{B}_n\left(\boldsymbol{\phi}^0+\frac{\boldsymbol{\gamma}}{r_n},\widehat{\boldsymbol{\lambda}},\widehat{h}\right) &= \mathbf{M}_n\left(\boldsymbol{\phi}^0+\frac{\boldsymbol{\gamma}}{r_n},\widehat{\boldsymbol{\lambda}},\widehat{h}\right)-\mathbf{M}_n(\boldsymbol{\phi}^0,\widehat{\boldsymbol{\lambda}},\widehat{h})\\ &= \mathbf{M}_n\left(\boldsymbol{\phi}^0+\frac{\boldsymbol{\gamma}}{r_n},\boldsymbol{\lambda}^0,\widehat{h}\right)-\mathbf{M}_n(\boldsymbol{\phi}^0,\boldsymbol{\lambda}^0,\widehat{h})+\mathrm{O}_{\mathbb{P}}(n^{-1})\\ &= \mathbf{B}_n\left(\boldsymbol{\phi}^0+\frac{\boldsymbol{\gamma}}{r_n},\boldsymbol{\lambda}^0,\widehat{h}\right)+\mathrm{O}_{\mathbb{P}}(n^{-1}).\end{aligned}$$

We can use the same proof of Lemma 2 in Delsol and Van Keilegom [2020] for showing the last assertions where we have the true change points in the expression of our process and it satisfies their conditions, hence we obtain the result. □

Lemma 5.5.0.6 *Let $\mathscr{K}=\{\boldsymbol{\gamma}\in\mathbf{E}:\|\boldsymbol{\gamma}\|\leq \mathrm{K}\}$. Then, under the assumptions of Theorem 5.3.3.2 the process*

$$\boldsymbol{\gamma}\mapsto r_n^2\mathbf{B}_n\left(\boldsymbol{\phi}^0+\frac{\boldsymbol{\gamma}}{r_n},\boldsymbol{\lambda}^0,h_0\right)+r_n\mathrm{W}_n(\boldsymbol{\gamma})$$

is asymptotically tight, asymptotically uniformly equicontinuous with respect to $\|\cdot\|$ on $\mathscr{K}$, and it converges weakly to the process

$$\boldsymbol{\gamma}\mapsto\Gamma(\boldsymbol{\gamma})+\mathbb{G}(\boldsymbol{\gamma}) \text{ in } \ell^\infty(\mathscr{K}).$$

Proof of Lemma 5.5.0.6

The proof of this lemma is the same as Lemma 3 in Delsol and Van Keilegom [2020], we can write the process

$$\mathrm{T}_{j,n}:\boldsymbol{\gamma}\mapsto r_n^2\mathbf{B}_n\left(\boldsymbol{\phi}_j^0+\frac{\boldsymbol{\gamma}}{r_n},\lambda_j^0,h_0\right)+r_n\mathrm{W}_{j,n}(\boldsymbol{\gamma}),$$

for each $j=1,\dots,k+1$ as the sum of two process

$$\mathrm{Y}_{j,n}=r_n^2\left(\mathbf{B}_n\left(\boldsymbol{\phi}_j^0+\frac{\boldsymbol{\gamma}}{r_n},\lambda_j^0,h_0\right)-\mathbf{B}\left(\boldsymbol{\phi}_j^0+\frac{\boldsymbol{\gamma}}{r_n},\lambda_j^0,h_0\right)\right)$$

and

$$\mathrm{Z}_{j,n}=r_n^2\mathbf{B}\left(\boldsymbol{\phi}_j^0+\frac{\boldsymbol{\gamma}}{r_n},\lambda_j^0,h_0\right)+r_n\mathrm{W}_{j,n}.$$

Here we have k process satisfy the assumptions of the Theorem 5.3.3.2 converging weakly by application of the Lemma 3 in the same reference, so we obtain the result by summing these process i.e.,

$$r_n^2\mathbf{B}_n\left(\boldsymbol{\phi}^0+\frac{\boldsymbol{\gamma}}{r_n},\boldsymbol{\lambda}^0,h_0\right)+r_n\mathrm{W}_n(\boldsymbol{\gamma})=\sum_{j=1}^{k+1}\mathrm{T}_{j,n}(\boldsymbol{\gamma}).$$

Almost all paths of the limiting process $\boldsymbol{\gamma}\mapsto\Gamma(\boldsymbol{\gamma})+\mathbb{G}(\boldsymbol{\gamma})$ on $\mathscr{K}$ are uniformly continuous with respect to $\|\cdot\|$ by the Addendum 1.5.8 in van der Vaart and Wellner [1996]. □

5.6 References

Abou-Elailah, A., Gouet-Brunet, V., and Bloch, I. (2015). Detection of abrupt changes in spatial relationships in video sequences. In A. Fred, M. De Marsico, and M. Figueiredo, editors, *Pattern Recognition: Applications and Methods*, pages 89–106, Cham. Springer International Publishing. 142

Aggarwal, R., Inclan, C., and Leal, R. (1999). Volatility in emerging stock markets. *Journal of financial and Quantitative Analysis*, **34**(1), 33–55. 142

Alvarez-Andrade, S. and Bouzebda, S. (2014). Some nonparametric tests for change-point detection based on the $\mathbb{P}$-$\mathbb{P}$ and $\mathbb{Q}$-$\mathbb{Q}$ plot processes. *Sequential Anal.*, **33**(3), 360–399. 143

Aue, A. and Horváth, L. (2013). Structural breaks in time series. *J. Time Series Anal.*, **34**(1), 1–16. 143

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of applied econometrics*, **18**(1), 1–22. 142

Bouzebda, S. (2012). On the strong approximation of bootstrapped empirical copula processes with applications. *Math. Methods Statist.*, **21**(3), 153–188. 160

Bouzebda, S. (2014). Asymptotic properties of pseudo maximum likelihood estimators and test in semi-parametric copula models with multiple change points. *Math. Methods Statist.*, **23**(1), 38–65. 143, 160

Bouzebda, S. and Keziou, A. (2013). A semiparametric maximum likelihood ratio test for the change point in copula models. *Stat. Methodol.*, **14**, 39–61. 143, 160

Braun, J. V., Braun, R. K., and Müller, H.-G. (2000). Multiple changepoint fitting via quasi-likelihood, with application to DNA sequence segmentation. *Biometrika*, **87**(2), 301–314. 142

Brodsky, B. E. and Darkhovsky, B. S. (1993). *Nonparametric methods in change-point problems*, volume 243 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht. 142

Carlstein, E., Müller, H.-G., and Siegmund, D., editors (1994). *Change-point problems*, volume 23 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA. Papers from the AMS-IMS-SIAM Summer Research Conference held at Mt. Holyoke College, South Hadley, MA, July 11–16, 1992. 142

Chen, H. (2019). Sequential change-point detection based on nearest neighbors. *Ann. Statist.*, **47**(3), 1381–1407. 142

Chen, J. and Gupta, A. K. (2000). *Parametric statistical change point analysis*. Birkhäuser Boston, Inc., Boston, MA. 142

Cheng, K. C., Aeron, S., Hughes, M. C., Hussey, E., and Miller, E. L. (2019). Optimal transport based change point detection and time series segment clustering. 142

Chernoff, H. (1956). Large-sample theory: Parametric case. *The Annals of Mathematical Statistics*, **27**(1), 1–22. 147

Choi, F. Y. (2000). Advances in domain independent linear text segmentation. *arXiv preprint cs/0003083*. 142

Chu, P.-S. and Zhao, X. (2004). Bayesian change-point analysis of tropical cyclone activity: The central north pacific case. *Journal of Climate*, **17**(24), 4893–4901. 142

Csörgő, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester. With a foreword by David Kendall. 143

Delsol, L. and Van Keilegom, I. (2020). Semiparametric M-estimation with non-smooth criterion functions. *Ann. Inst. Statist. Math.*, **72**(2), 577–605. 143, 144, 147, 151, 153, 159, 160, 162, 163, 178

Desmond, R. A., Weiss, H. L., Arani, R. B., Soong, S.-j., Wood, M. J., Fiddian, P. A., Gnann, J. W., and Whitley, R. J. (2002). Clinical applications for change-point analysis of herpes zoster pain. *Journal of pain and symptom management*, **23**(6), 510–516. 142

Erdman, C. and Emerson, J. W. (2008). A fast bayesian change point analysis for the segmentation of microarray data. *Bioinformatics*, **24**(19), 2143–2148. 142

Fu, Y.-X. and Curnow, R. N. (1990). Maximum likelihood estimation of multiple change points. *Biometrika*, **77**(3), 563–573. 142

Guan, Z. (2007). Semiparametric tests for change-points with epidemic alternatives. *Journal of statistical planning and inference*, **137**(6), 1748–1764. 143

He, H. and Severini, T. A. (2010). Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli*, **16**(3), 759–779. 144, 145, 148, 149, 150

Horváth, L. and Rice, G. (2014). Extensions of some classical methods in change point analysis. *TEST*, **23**(2), 219–255. 143

Jandhyala, V., Fotopoulos, S., MacNeill, I., and Liu, P. (2013). Inference for single and multiple change-points in time series. *J. Time Series Anal.*, **34**(4), 423–446. 143

Kim, A. Y., Marzban, C., Percival, D. B., and Stuetzle, W. (2009). Using labeled data to evaluate change detectors in a multivariate streaming environment. *Signal Processing*, **89**(12), 2529–2536. 142

Kim, J., Pollard, D., *et al.* (1990). Cube root asymptotics. *The Annals of Statistics*, **18**(1), 191–219. 154

Koprinska, I. and Carrato, S. (2001). Temporal video segmentation: A survey. *Signal processing: Image communication*, **16**(5), 477–500. 142

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer Series in Statistics. Springer, New York. 150, 154, 157

Koul, H. L., Müller, U. U., and Schick, A. (2012). The transfer principle: a tool for complete case analysis. *Ann. Statist.*, **40**(6), 3031–3049. 161

Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**(19), 3763–3770. 142

Lavielle, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stochastic Process. Appl.*, **83**(1), 79–102. 149

Lavielle, M. and Ludeña, C. (2000). The multiple change-points problem for the spectral distribution. *Bernoulli*, **6**(5), 845–869. 149

Lavielle, M. and Teyssière, G. (2006). Détection de ruptures multiples dans des séries temporelles multivariées. *Liet. Mat. Rink.*, **46**(3), 351–376. 142

Lio, P. and Vannucci, M. (2000). Wavelet change-point prediction of transmembrane proteins. *Bioinformatics*, **16**(4), 376–382. 142

Liu, S., Wright, A., and Hauskrecht, M. (2018). Change-point detection method for clinical decision support system rule monitoring. *Artificial Intelligence in Medicine*, **91**, 49 – 56. 142

Lu, L., Zhang, H.-J., and Jiang, H. (2002). Content analysis for audio classification and segmentation. *IEEE Transactions on speech and audio processing*, **10**(7), 504–516. 142

Lung-Yut-Fong, A., Lévy-Leduc, C., and Cappé, O. (2012). Distributed detection/localization of change-points in high-dimensional network traffic data. *Stat. Comput.*, **22**(2), 485–496. 142

Mahmoud, M. A., Parker, P. A., Woodall, W. H., and Hawkins, D. M. (2007). A change point method for linear profile data. *Quality and Reliability Engineering International*, **23**(2), 247–268. 142

Mann, H. B. and Wald, A. (1943). On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, **14**(3), 217–226. 147

Mazhar, O., Rojas, C., Fischione, C., and edit Mohammad Reza Hesamzadeh (2018). Bayesian model selection for change point detection and clustering. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3433–3442, Stockholmsmässan, Stockholm Sweden. PMLR. 142

Minin, V. N., Dorman, K. S., Fang, F., and Suchard, M. A. (2005). Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, **21**(13), 3034–3042. 142

Móricz, F. A., Serfling, R. J., and Stout, W. F. (1982). Moment and probability bounds with quasisuperadditive structure for the maximum partial sum. *Ann. Probab.*, **10**(4), 1032–1040. 168

Müller, U. U. (2009). Estimating linear functionals in nonlinear regression with responses missing at random. *Ann. Statist.*, **37**(5A), 2245–2277. 161

Olshen, A. B., Venkatraman, E., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, **5**(4), 557–572. 142

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, **41**, 100–115. 142

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, **42**, 523–527. 142

Page, E. S. (1957). On problems in which a change in a parameter occurs at an unknown point. *Biometrika*, **44**(1/2), 248–252. 142

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, **57**(5), 1027–1057. 147

Pérez-González, A., Vilar-Fernández, J. M., and González-Manteiga, W. (2009). Asymptotic properties of local polynomial regression with missing data and correlated errors. *Ann. Inst. Statist. Math.*, **61**(1), 85–109. 161

Pons, O. (2018). *Estimations and tests in change-point models*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ. 142

Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. Q. (2007). A Review and Comparison of Changepoint Detection Techniques for Climate Data. *Journal of Applied Meteorology and Climatology*, **46**(6), 900–915. 142

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech communication*, **32**(1-2), 127–154. 142

Spokoiny, V. (2009). Multiscale local change point detection with applications to value-at-risk. *Ann. Statist.*, **37**(3), 1405–1436. 142

Tartakovsky, A., Nikiforov, I., and Basseville, M. (2015). *Sequential analysis*, volume 136 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL. Hypothesis testing and changepoint detection. 142

Tartakovsky, A. G., Rozovskii, B. L., Blazek, R. B., and Kim, H. (2006). A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE transactions on signal processing*, **54**(9), 3372–3382. 142

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, **167**, 107299. 142

van de Geer, S. A. (2000). *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. 150

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics. 147, 150, 151, 152, 154, 157, 159, 175, 176, 177, 178

Vostrikova, L. J. (1981). Discovery of "discord" in multidimensional random processes. *Dokl. Akad. Nauk SSSR*, **259**(2), 270–274. 160

Wang, H., Zhang, D., and Shin, K. G. (2004). Change-point monitoring for the detection of dos attacks. *IEEE Transactions on dependable and secure computing*, **1**(4), 193–208. 142

Wu, Y. (2005). *Inference for change-point and post-change means after a CUSUM test*, volume 180 of *Lecture Notes in Statistics*. Springer, New York. 142

Xing, H. and Ying, Z. (2012). A semiparametric change-point regression model for longitudinal observations. *J. Amer. Statist. Assoc.*, **107**(500), 1625–1637. 143

Zhang, S. and Tian, B. (2020). Semiparametric method for identifying multiple change-points in financial market. *Communications in Statistics-Simulation and Computation*, pages 1–16. 143

Zou, C., Yin, G., Feng, L., and Wang, Z. (2014). Nonparametric maximum likelihood approach to multiple change-point problems. *Ann. Statist.*, **42**(3), 970–1002. 145

Chapter 6

Asymptotic properties of M-estimators based on estimating equations and censored data in semi-parametric models with multiple change points

Ce chapitre développe le contenu d'un article publié dans le journal **"Journal of Mathematical Analysis and Applications."**497 (2021), no. 2, 124883, mis en forme pour être inséré dans le présent manuscrit de thèse.
Asymptotic properties of M*-estimators based on estimating equations and censored data in semi-parametric models with multiple change points*. **JMAA.** 497 (2021), no. 2, 124883,

Abstract

Statistical models with multiple change points in presence of censored data are used in many fields; however, the theoretical properties of M-estimators of such models have received relatively little attention. The main purpose of the present work is to investigate the asymptotic properties of M-estimators of the parameters of a multiple change-point model for a general class of models in which the form of the distribution can change from segment to segment and in which, possibly, there are parameters that are common to all segments, in the setting of a known number of change points. Consistency of the M-estimators of the change points is established and the rate of convergence is determined. The asymptotic normality of the M-estimators of the parameters of the within-segment distributions is established. Since the approaches used in the complete data models are not easily extended to multiple change-point models in the presence of censoring, we have used some general results of Kaplan-Meier integrals. We investigate the performance of the methodology for small samples through a simulation study.

Key words: Semiparametric inference; multiple change-points; change-point fraction; common parameter; consistency; convergence rate; M-estimators; Z-estimators; censored data; Kaplan

Meier integrals; Argmax theorem; Central limit theorem.
Mathematics Subject Classification :62F12; 62F03; 62G20; 60F05; 62N02; 62E20; 62P20.

6.1 Introduction and motivations

In major real data investigation, the stationarity assumption has been frequently used. However, in practice, time series entail in their dependence structure and therefore modelling non-stationary processes using stationary methods to capture their time-evolving dependence aspects most likely result in a crude approximation. Change-point detection plays a critical role in such situations. Notice that the problem of change-points in a sequence of random variables has a long history. Early work on this problem can be found in Page [1954, 1955, 1957] who investigated quality control problems and proposed a sequential scheme for identifying changes in the mean of a sequence of independent random variables. Over time, methods in change point analysis have been developed to address data analytic questions in fields ranging from biology to finance, and in many cases such methodology has become standard. The statistical community now enjoys a vast literature on change point analysis where many of the most natural and common questions have received at least some attention. For a broader presentation of the field of change-point analysis along with statistical applications, we refer the reader to the monographs by Brodsky and Darkhovsky [1993], Csörgő and Horváth [1997], Chen and Gupta [2000], Wu [2005] and Pons [2018], just to cite a few. We refer to the paper of Lee [2010] for a list of comprehensive bibliography of books and research papers on this topic. The problem of detecting abrupt changes has been discussed intensively in a time series context, we may refer to Jandhyala *et al.* [2013] and Aue and Horváth [2013] for a review of the literature. Recent references on the subject include Chen [2019], Chu and Chen [2019], Garreau and Arlot [2018], Tan and Zhang [2019], Nkurunziza and Fu [2019], Qian *et al.* [2019] and El Ktaibi and Ivanoff [2019]. Compared to single change-point detection, multiple change-points detection is a much more challenging problem. Work on detection for multiple change-points began in the 1980s (e.g., Vostrikova [1981], Yin [1988], Yao [1988]). There exists a rich literature devoted to this field, we refer to Truong *et al.* [2020] for review of change-point and some extensions. For the censored setting, there are only a few papers dealing with detection of changes, for single change-point, we refer to Stute [1996] who provided an estimator of the change point based on the U-statistics. Gombay and Liu [2000], Hušková and Neuhaus [2004], Al-Awadhi and Aly [2005], Wang and Zheng [2012] have considered test procedures for change-point. He [2017, 2015] considered the multiple change-points for particular distributions. To our best knowledge there the case where the change occurs for the two variables, i.e., the censored variable and the censorship variable in general setting was not investigated in the literature up to present. Notice that multiple change-points problem occurs for the survival function due to hazard change according to evolving time. For example, a cancer survival function can change abruptly or smoothly at a few time points. For example, Kim *et al.* [2020] applied their method to find the change-points for leukemia survival data and identified the change-points. However

multiple change-points problems are not much considered due to its computational complexity and theoretical difficulty. Hušková and Neuhaus [2004] have investigated the problem of single change when the variables are assumed to be independent but not necessarily identically distributed. While the body of work about the change-point constitutes a rich literature, it mainly deals with the inference of a single change in a short or moderate sized sequence. Detecting multiple change-points in a very long sequence has emerged as an important problem that has attracted more and more attention recently, we refer to Niu *et al.* [2016]. There is a literature on the change-point problem and their applications and it is not the purpose of the present paper to survey this extensive literature.

The main purpose of the present work is to consider a general framework and the characterization of the asymptotic properties of semi-parametric M-estimators based on censored data in models with multiple change-points, this generalization is far from being trivial and harder to control the estimator of Kaplan-Meier of each sample, which form a basically unsolved open problem in the literature. We aim at filling this gap in the literature by combining results He and Severini [2010] with techniques handling the Kaplan Meier integrals. However, as will be seen later, the problem requires much more than "simply" combining ideas from the existing results. In fact, delicate mathematical derivations will be required to cope with Kaplan Meier integrals in our context.

We start by giving some notations and definitions that are needed for the forthcoming sections. Let $X_1,\dots,X_n$ be n independent random variables censoring by n independent random variables $C_1,\dots,C_n$ respectively, where X_i and C_i are independent for all i, so we observe

$$\left(Y_i = X_i \wedge C_i, \delta_i = 1\!\!1_{\{X_i \le C_i\}}\right), \quad \text{for} \quad 1 \le i \le n.$$

Survival data in clinical trials or failure time data in reliability studies, for example, are often subject to such censoring. To be more specific, many statistical experiments result in incomplete samples, even under well-controlled conditions. For example, clinical data for surviving most types of disease are usually censored by other competing risks to life which result in death. We suppose that there exists unknown change points $n_1,\dots,n_k$, such that

$$0 = n_0 < n_1 < \cdots < n_k < n_{k+1} = n,$$

where for each $j = 1,\dots,k+1$, $(X_{n_{j-1}+1}, C_{n_{j-1}+1}),\dots,(X_{n_j}, C_{n_j})$, are i.i.d. with distribution function depending on j. Here, we consider semi-parametric change-points models in which the distribution function of $X_{n_{j-1}+1},\dots,X_{n_j}$ is parametric. We suppose that the theoretical distribution $F_{n_j^0}(\cdot) =: F(\alpha^0, \theta_j^0, \cdot)$ of X_i, $i = 1,\dots,n$, depends on the real common parameter α^0 for all $j = 1,\dots,k+1$ and the real within-segment θ_j^0, for each $j = 1,\dots,k+1$ which are assumed to be unknown. In this model, there are k real change points $n_1^0,\dots,n_k^0$ but unknown, where the number of change point k is assumed to be known. We estimate the unknown parameters n_j, α

and θ_j, $j = 1, \ldots, k+1$ by maximizing the estimating equations defined by:

$$\boldsymbol{\ell} \equiv \boldsymbol{\ell}(\alpha, \theta_1, \ldots, \theta_{k+1}, n_1, \ldots, n_k) = \sum_{j=1}^{k+1} \frac{(n_j - n_{j-1})}{n} \int_{\mathbb{R}} \mathbf{m}_j(\alpha, \theta_j, x) d\widehat{\mathrm{F}}_{n_j}(x), \tag{6.1.1}$$

where $1 - \widehat{\mathrm{F}}_{n_j}(\cdot)$ is the usual Kaplan-Meier product limit estimator of $1 - \mathrm{F}_{n_j}(\cdot)$ introduced by Kaplan and Meier [1958] and defined by

$$1 - \widehat{\mathrm{F}}_{n_j}(x) = \prod_{i=n_{j-1}+1}^{n_j} \left(1 - \frac{d_i}{n_i}\right)^{\mathrm{1I}_{\{\mathrm{Y}_{(i)} \le x\}}}, \tag{6.1.2}$$

where

$$r_i = \sum_{k=n_{j-1}+1}^{n_j} \mathrm{1I}_{\{\mathrm{Y}_{(i)} \le \mathrm{Y}_k\}}$$

and

$$d_i = \sum_{k=n_{j-1}+1}^{n_j} \mathrm{1I}_{\{\mathrm{Y}_{(i)} = \mathrm{Y}_k, \delta_k = 1\}},$$

denoting the number of individuals still at risk at time $\mathrm{Y}_{(i)}$ and the number of deaths at time $\mathrm{Y}_{(i)}$ respectively, and $\mathrm{Y}_{(i)}$ denotes the order statistic of $\mathrm{Y}_{n_{j-1}+1}, \ldots, \mathrm{Y}_{n_j}$ and $\mathrm{1I}_{\mathrm{E}}$ denoting the indicator function of E. For each sample $\mathrm{X}_{n_{j-1}+1}, \ldots, \mathrm{X}_{n_j}$, $j = 1, \ldots, k+1$, and $\mathbf{m}_j(\cdot)$ is a given measurable function from $\Upsilon \times \Theta_j \times \mathbb{R}$ to $\mathbb{R}$; Υ and Θ_j are the parameter spaces of α, θ_j for $j = 1, \ldots, k+1$, respectively. Simple calculation gives

$$\boldsymbol{\ell}(\alpha, \theta_1, \ldots, \theta_{k+1}, n_1, \ldots, n_k) = \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}+1}^{n_j} \frac{\mathbf{m}_j(\alpha, \theta_j, \mathrm{Y}_i)\delta_i}{\mathrm{S}_{\mathrm{C}}^{n_j}(\mathrm{Y}_i^-)}, \tag{6.1.3}$$

where $\mathrm{S}_{\mathrm{C}}^{n_j}(\cdot)$ is the Kaplan-Meier product limit estimator of $1 - \mathrm{G}_{n_j}(\cdot)$, for each sample $\mathrm{C}_{n_{j-1}+1}, \ldots, \mathrm{C}_{n_j}$, $j = 1, \ldots, k+1$.

Our result is a generalization for the work of He and Severini [2010] in the sense that we consider the M-estimation in the censored data setting. He and Severini [2010] investigated statistical models with multiple change-points and established the theoretical properties of the maximum likelihood estimators. Their results are not directly applicable here since we consider a more general framework. These results are not only useful in their own right but essential to establish the theoretical properties of our estimators. Under no censoring, there are a number of results available on the asymptotic properties of parameter estimators in change-point models with $\mathbf{m}_j(\alpha, \theta_j, x) = \log f_j(\alpha, \theta_j, x)$. See, for example, Hinkley [1970, 1972], Hinkley and Hinkley [1970], Bhattacharya [1987], Fu and Curnow [1990a,b], Jandhyala and Fotopoulos [1999, 2001] and Hawkins [2001]; the two monographs Chen and Gupta [2000] and Csörgő and Horváth [1997], and for the M-estimators we refer to Hušková [1996]. In Gombay and Horváth [1994], a maximum-likelihood-type statistic is proposed for testing a sequence

of observations for no change in the parameter against a possible change, this work is extended to the semi-parametric setting in Bouzebda and Keziou [2013] and Bouzebda [2014]. It is worth noticing that M-estimators include the least squares estimators, several robust version of means and notably their predecessor, the maximum likelihood estimate (MLE) with $\mathbf{m}_j(\alpha,\theta_j,\cdot) = \log f_j(\alpha,\theta_j,\cdot)$, $f(\cdot)$ being the probability density function. Strong consistency of M-estimators can be verified as that of the MLEs, and it is possible to avoid the differentiability condition of the density function $f_j(\alpha,\theta_j,x)$ in the MLE case. This approach was first employed by Wald [1949] and later extended, for example, by LeCam [1953], Kiefer and Wolfowitz [1956], Bahadur [1967], Huber [1967], Pfanzagl [1969] and Perlman [1972] among others. Asymptotic properties of Huber's M-estimators based on complete data are well understood nowadays and can be found, for example, in Huber [1981] and van der Vaart [1998], among others.

In the presence of censoring very little is known about the general large sample properties of M-estimators. Reid [1981] derived the influence function and the asymptotic normality of a truncated type M-estimator. (Some modifications are required in Reid's arguments, cf. Andersen *et al.* [2012]. Oakes [1986] considered M-estimators with $\mathbf{m}_j(\alpha,\theta_j,\cdot) = \log f_j(\alpha,\theta_j,\cdot)$ and called them approximate MLEs since the corresponding M-estimators are no longer the MLEs. Borgan [1984] studied the asymptotic properties of the MLE. Another type of M-estimator, based on the cumulative hazard function and aiming at inclusion of the MLEs under censoring, is discussed in Hjort [1985]. Wang [1995] has established the strong consistency of this type of estimators under general conditions which can be applied to parametric, semi-and non-parametric models.

The main objective of our paper is to provide a full theoretical justification of the consistency of M-estimators of the parameters of a general class of multiple change-points models and gives the asymptotic distribution of the parameters of the within-segment distributions. This requires the effective application of large sample theory techniques, which were developed for the empirical processes, refer to Section 6.4 where we have used results from the work of Pakes and Pollard [1989].

The article is structured as follows. Section 6.2 is devoted to the statement of our notations and assumptions. In Section 6.3, the asymptotic properties of our estimators are derived. The general theory of the Z-estimators is considered in Section 6.4. In Section 6.5, we specify the estimation procedure for the maximum likelihood. The finite sample performance of the latter is illustrated by means of Monte Carlo simulations in Section 6.6. Some concluding remarks are given in Section 7.4. To avoid interrupting the flow of the presentation, all mathematical developments are relegated to Section 6.7. Section 6.8 gives some basic definitions and preliminaries needed to state our results.

6.2 Notation and assumptions

In this section, we introduce the notation needed to state the asymptotic results in Section 6.3. The parameter spaces Υ and Θ_j are the subset of $\mathbb{R}^d$ and $\mathbb{R}^{d_j}$ respectively. Let

$$\begin{aligned}
\lambda_j &= \frac{n_j}{n}, \text{ for any } j = 1,\ldots,k,\\
\lambda_j^0 &= \frac{n_j^0}{n}, \text{ for any } j = 1,\ldots,k,\\
\boldsymbol{\lambda} &= (\lambda_1,\lambda_2,\ldots,\lambda_k),\\
\boldsymbol{\lambda}^0 &= (\lambda_1^0,\lambda_2^0,\ldots,\lambda_k^0),\\
\boldsymbol{\theta} &= (\theta_1,\theta_2,\ldots,\theta_{k+1}),\\
\boldsymbol{\theta}^0 &= (\theta_1^0,\theta_2^0,\ldots,\theta_{k+1}^0),\\
\boldsymbol{\phi} &= (\alpha,\theta_1,\theta_2,\ldots,\theta_{k+1}),\\
\boldsymbol{\phi}^0 &= (\alpha^0,\theta_1^0,\theta_2^0,\ldots,\theta_{k+1}^0),\\
S_{F_{n_j^0}}(\cdot) &= 1 - F_{n_j^0}(\cdot),\\
S_{G_{n_j^0}}(\cdot) &= 1 - G_{n_j^0}(\cdot).
\end{aligned}$$

We have for each $j = 1,\ldots,k$,

$$1 - H_{n_j^0}(\cdot) = (1 - G_{n_j^0}(\cdot))(1 - F_{n_j^0}(\cdot)).$$

Let $\tau_{F_{n_j^0}}(\cdot)$ (resp. $\tau_{G_{n_j^0}}(\cdot)$) be the upper bound of the support of $F_{n_j^0}(\cdot)$ (resp. $G_{n_j^0}(\cdot)$). Note that $\boldsymbol{\lambda}^0$ is taken to be a constant vector as n goes to infinity. Let Λ be the set of the configurations of change-points and Φ the set of parameters,

$$\begin{aligned}
\Lambda &= \{(\lambda_1,\lambda_2,\ldots,\lambda_k) : \lambda_j = \frac{n_j}{n},\ \ j = 1,\ldots,k,\ \ 0 < n_1 < \cdots < n_k < n\},\\
\Phi &= \Theta_1 \times \Theta_2 \times \cdots \times \Theta_{k+1} \times \Upsilon.
\end{aligned}$$

The criterion function computed over the segment j of $\boldsymbol{\lambda}$ is defined by

$$\mathscr{G}_n(\Upsilon_j,\theta_j,\alpha) = \frac{(n_j - n_{j-1})}{n}\int_{\mathbb{R}} \mathbf{m}_j(\alpha,\theta_j,x)\, d\widehat{F}_{n_j}(x).$$

Consequently, we can rewrite the function $\boldsymbol{\ell}$ given in (6.1.1) as

$$\boldsymbol{\ell} = \sum_{j=1}^{k+1} \mathscr{G}_n(\Upsilon_j,\theta_j,\alpha).$$

Estimators of all change-points, all within-segment parameters and the common parameter are defined by maximization of the function $\boldsymbol{\ell}$ in $\Lambda \times \Phi$, i.e.,

$$(\widehat{\alpha},\widehat{\theta}_1,\ldots,\widehat{\theta}_{k+1},\widehat{n}_1,\ldots,\widehat{n}_k) = \underset{0<n_1<n_2<\cdots<n;\theta_j\in\Theta_j,1\le j\le k+1,\alpha\in\Upsilon}{\operatorname{argmax}} \boldsymbol{\ell}. \tag{6.2.1}$$

For a given configuration $\boldsymbol{\lambda}$, $\left(\widehat{\theta}_j(\lambda_j), \widehat{\alpha}(\lambda_j)\right)$

maximizes $\mathcal{G}_n(Y_j, \theta_j, \alpha)$. We can remark that, when $\boldsymbol{\lambda} = \boldsymbol{\lambda}^0$, the estimate of $(\boldsymbol{\theta}^0, \alpha^0)$ obtained by maximizing $\boldsymbol{\ell}(\alpha, \theta_1, \ldots, \theta_{k+1}, n_1^0, \ldots, n_k^0)$ converge to $(\boldsymbol{\theta}^0, \alpha^0)$

under the Assumptions 6.2.0.1 and the first part of the Assumption 6.2.0.2 for complete data, by the result of van der Vaart [1998] and by add the first part of Assumption 6.2.0.5, we get the convergence for censored data by the result of Wang [1995]. In the case where the change point fraction $\boldsymbol{\lambda}^0$ is unknown, the M-estimators

$(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}}, \widehat{\alpha})$ is the value of $(\boldsymbol{\lambda}, \boldsymbol{\theta}, \alpha)$ that maximizes $\boldsymbol{\ell}(\alpha, \theta_1, \ldots, \theta_{k+1}, n_1^0, \ldots, n_k^0)$ in $\Lambda \times \Phi$. Thus $(\widehat{\theta}_j, \widehat{\alpha}) \overset{\text{def}}{=} \left(\widehat{\theta}_j(\widehat{\lambda}_j), \widehat{\alpha}(\widehat{\lambda}_j)\right)$

is the M-estimator of (θ_j^0, α_j^0) computed in the segment j of the estimated configuration of change-points $\widehat{n}_j$, refer for similar arguments to Lavielle and Ludeña [2000]. Let us introduce

$$L^0(\alpha, \theta_1, \ldots, \theta_{k+1}) = \sum_{j=1}^{k+1} \frac{(n_j^0 - n_{j-1}^0)}{n} \int_{\mathbb{R}} \mathbf{m}_j(\alpha, \theta_j, x) dF_{n_j^0}(x), \tag{6.2.2}$$

where $F_{n_j^0}(\cdot)$ (respectivement $G_{n_j^0}(\cdot)$) is the true function of distribution for the sample $X_{n_{j-1}^0+1}, \ldots, X_{n_j^0}$ (resp. $C_{n_{j-1}^0+1}, \ldots, C_{n_j^0}$), $j = 1, \ldots, k+1$. The following decomposition will play an instrumental role in the proofs of Theorem 6.3.0.1 and Theorem 6.3.0.3. Define a function $\mathbf{W}'$ by

$$\begin{aligned}
\mathbf{W}' = & \sum_{j=1}^{k+1} \sum_{i=1}^{k+1} \frac{n_{ji}}{n} \left\{ \int [\mathbf{m}_j(\alpha, \theta_j, x) - \mathbf{m}_i(\alpha^0, \theta_i^0, x)] dF_{n_i^0} \right\} \\
& + \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}+1}^{n_j} \left\{ \frac{\mathbf{m}_j(\alpha, \theta_j, Y_i)\delta_i}{S_C^{n_j}(Y_i^-)} - \mathbb{E}(\mathbf{m}_j(\alpha, \theta_j, X_i)) \right\} \\
& - \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}^0+1}^{n_j^0} \left\{ \frac{\mathbf{m}_j(\alpha^0, \theta_j^0, Y_i)\delta_i}{S_C^{n_j^0}(Y_i^-)} - \mathbb{E}(\mathbf{m}_j(\alpha^0, \theta_j^0, X_i)) \right\},
\end{aligned} \tag{6.2.3}$$

where n_{ji} is the number of observations of the interested variables in the set

$$[n_{j-1} + 1, n_j] \cap [n_{i-1}^0 + 1, n_i^0],$$

for $i, j = 1, \ldots, k+1$. We obviously have that

$$\underset{0 < n_1 < n_2 < \cdots < n; \theta_j \in \Theta_j, 1 \le j \le k+1, \alpha \in \Upsilon}{\operatorname{argmax}} \boldsymbol{\ell} = \underset{0 < n_1 < n_2 < \cdots < n; \theta_j \in \Theta_j, 1 \le j \le k+1, \alpha \in \Upsilon}{\operatorname{argmax}} \mathbf{W}';$$

thus, the M-estimators may be defined as the maximizers of $\mathbf{W}'$ rather than as the maximizers of $\boldsymbol{\ell}$. Our idea is to replace EKM $S_C^{n_j}(\cdot)$ in (6.2.3) by the theoretical survival function $S_{G_{n_j^0}}(\cdot)$ and to proof the difference between the EKM based on the estimated survival function and the EKM based on the theoretical survival function is negligible, in probability, as n goes to infinity, see (6.7.0.3). Notice that $S_C^{n_j^0}(\cdot)$ converges to $S_{G_{n_j^0}}(\cdot)$, so we can replace the EKM, at the price of

some complicated calculations. Let $b(\alpha,\theta_j,\alpha^0,\theta_i^0)$ be defined by

$$\begin{aligned} b(\alpha,\theta_j,\alpha^0,\theta_i^0) &= \mathbb{E}(\mathbf{m}_j(\alpha,\theta_j,\mathrm{X}_i)) - \mathbb{E}(\mathbf{m}_i(\alpha^0,\theta_i^0,\mathrm{X}_i)) \\ &= \int_{\mathbb{R}} [\mathbf{m}_j(\alpha,\theta_j,x) - \mathbf{m}_i(\alpha^0,\theta_i^0,x)]\, d\mathrm{F}_{n_i^0}(x), \end{aligned} \tag{6.2.4}$$

for $i,j = 1,\ldots,k+1$. We substitute $\mathbf{W}^{'}$ by $\mathbf{W}$ after replacing the EKM by its true survival function and we define

$$\mathbf{W} = \mathbf{W}_1 + \mathbf{W}_2,$$

where

$$\mathbf{W}_1 = \sum_{j=1}^{k+1}\sum_{i=1}^{k+1} \frac{n_{ji}}{n} b(\alpha,\theta_j,\alpha^0,\theta_i^0) \tag{6.2.5}$$

and

$$\begin{aligned} \mathbf{W}_2 =& \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}+1}^{n_j} \left\{ \frac{\mathbf{m}_j(\alpha,\theta_j,\mathrm{Y}_i)\delta_i}{\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_i^-)} - \mathbb{E}(\mathbf{m}_j(\alpha,\theta_j,\mathrm{X}_i)) \right\} \\ & - \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}^0+1}^{n_j^0} \left\{ \frac{\mathbf{m}_j(\alpha^0,\theta_j^0,\mathrm{Y}_i)\delta_i}{\mathrm{S}_{\mathrm{G}_{n_z^0}}(\mathrm{Y}_i^-)} - \mathbb{E}(\mathbf{m}_j(\alpha^0,\theta_j^0,\mathrm{X}_i)) \right\}. \end{aligned}$$

Alternatively, we may write

$$\begin{aligned} \mathbf{W}_2 =& \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=1}^{k+1} \left\{ \sum_{t\in\tilde{n}_{ji}} \left[\frac{\mathbf{m}_j(\alpha,\theta_j,\mathrm{Y}_t)\delta_t}{\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_t^-)} - \mathbb{E}(\mathbf{m}_j(\alpha,\theta_j,\mathrm{X}_t)) \right] \right. \\ & \left. - \sum_{t\in\tilde{n}_{ji}} \left[\frac{\mathbf{m}_i(\alpha^0,\theta_i^0,\mathrm{Y}_t)\delta_t}{\mathrm{S}_{\mathrm{G}_{n_i^0}}(\mathrm{Y}_t^-)} - \mathbb{E}(\mathbf{m}_i(\alpha^0,\theta_i^0,\mathrm{X}_t)) \right] \right\}, \end{aligned} \tag{6.2.6}$$

where

$$\tilde{n}_{ji} = [n_{j-1}+1, n_j] \cap [n_{i-1}^0+1, n_i^0].$$

We note that in the particular case where $\mathbf{m}_j(\cdot) = \log f_j(\cdot)$, we get $\mathbf{W}_1$ is a weighted sum of the negative Kullback-Leibler distances, and $\mathbf{W}_2 \to 0$ as $n \to 0$, by applying Proposition 6.8.1.1. In our analysis, the following assumptions will be needed.

Assumption 6.2.0.1 *1. Assume that for* $j = 1,\ldots,k+1$,

$$\mathbf{m}_{j+1}(\alpha^0,\theta_{j+1}^0,x) \neq \mathbf{m}_j(\alpha^0,\theta_j^0,x)$$

on a set of non-zero measure.

2. For any $j = 1,\ldots,k+1$, *any* α, θ_j; *for* $i = 1,\ldots,k+1$,

$$\int_{\mathbb{R}} (\mathbf{m}_j(\alpha,\theta_j,x))\, d\mathrm{F}_{n_i^0}(x) \leq \int_{\mathbb{R}} (\mathbf{m}_i(\alpha^0,\theta_i^0,x))\, d\mathrm{F}_{n_i^0}(x).$$

The first part of this assumption guarantees that the distributions in two neighboring segments are different. Clearly, this is required for the change-points to be well defined, and the second part is to ensure that the expectation of the function associates with the true parameters is the maximum in the true sample, when we consider the particular case $\mathbf{m}_j(\cdot) = \log f_j(\cdot)$, this assumption comes directly from the distance of Kullback-Leibler, for further details, we refer to He and Severini [2010], or when the function $g(\cdot)$ is independent of the index j, i.e., the same function of all segments for example when the variables are assumed to be from normal distribution and there is a change in variances and having the same mean, or conversely, so we have all parameters are in the same set, i.e., $\theta_j \in \Theta$ for any $j = 1, 2, \ldots, k+1$, for the uncensored case, another example if the variables are assumed to follow the Weibull's distribution. In the M-estimation theory, this condition is required to ensure that the true parameters are the points that maximize the criterion function. For more details see also van der Vaart and Wellner [1996].

Assumption 6.2.0.2 *Assume that*

1. *for $j = 1, \ldots, k+1$, θ_j and θ_j^0 are contained in Θ_j, where Θ_j is a compact subset of $\mathbb{R}^{d_j}$; α and α^0 are contained in Υ, where Υ is a compact subset of $\mathbb{R}^d$; here $d, d_1, \ldots, d_{k+1}$ are non-negative integers.*

2. *$\boldsymbol{\ell}(\alpha, \boldsymbol{\theta})$ is second-order continuously differentiable with respect to α, $\boldsymbol{\theta}$, and there is an interchangeability of integration and differentiation in (6.2.2).*

Compactness of the parameter space is used to insure that the maximum is achievable and to establish the consistency of the M-estimators of

$$\frac{n_1}{n}, \ldots, \frac{n_k}{n}, \theta_1, \ldots, \theta_{k+1}, \alpha,$$

for discussions and details on this condition and its necessity in general model, the reader can refer to Huber [1981] for complete data and Wang [1995] for censored data. Differentiability of the given function is used to justify some Taylor series expansions, interchangeability of integration and differentiation is a technical assumption used for the variance expression in (6.4.5). The second part of the Assumption 6.2.0.2 ensures the existence of the variance of the M-estimates. Both parts of Assumption 6.2.0.2 are relatively weak and are essentially the same as conditions used in parametric models for censored data without change-points, see Wang [1999].

Assumption 6.2.0.3 *Assume that*

1. *for any $j = 1, \ldots, k+1$ and any integers s, t satisfying $0 \le s < t \le n$,*

$$\mathbb{E}\left(\max_{\theta_j \in \Theta_j, \alpha \in \Upsilon} \left[\sum_{i=s+1}^{t} \left(\sum_{z=1}^{k+1} \frac{\mathbf{m}_j(\alpha, \theta_j, Y_i)\delta_i}{S_{G_{n_z^0}}(Y_i^-)} \mathrm{II}_{\{n_{z-1}+1 \le i \le n_z\}} - \mathbb{E}(\mathbf{m}_j(\alpha, \theta_j, X_i)) \right) \right]^2 \right) \le C(t-s)^r,$$

where $r < 2$ and C is a constant.

2. *for any* $j = 1, \ldots, k+1$ *and any integers* s, t *satisfying* $n_{j-1}^0 \leq s < t \leq n_j^0$,

$$\mathbb{E}\left(\max_{\theta_j \in \Theta_j, \alpha \in \Upsilon} \left[\sum_{i=s+1}^{t} \left(\sum_{z=1}^{k+1} \frac{\mathbf{m}_j(\alpha, \theta_j, \mathrm{Y}_i)\delta_i}{\mathrm{S}_{\mathrm{G}_{n_z^0}}(\mathrm{Y}_i^-)} \mathrm{II}_{\{n_{z-1}+1 \leq i \leq n_z\}} - \frac{\mathbf{m}_j(\alpha^0, \theta_j^0, \mathrm{Y}_i)\delta_i}{\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_i^-)} - b(\alpha, \theta_j, \alpha^0, \theta_j^0) \right) \right]^2 \right) \leq \mathrm{D}(t-s)^r,$$

where $b(\alpha, \theta_j, \alpha^0, \theta_j^0)$ *is introduced in equation (6.2.4),* $r < 2$ *and* D *is a constant.*

Parts 1 and 2 of Assumption 6.2.0.3 are technical requirements on the behavior of the function $\mathbf{m}_j(\cdot)$ between and within segments, respectively. This condition is used to ensure that the information regarding the within- and between-segment parameters grows quickly enough to establish consistency and asymptotic normality of the parameter estimators. Note that where $\mathbf{m}_j(\cdot) = \log f_j(\cdot)$ these conditions are relatively weak; it is easy to check that they are satisfied by at least all distributions in the exponential family, for more details refer to He and Severini [2010].

Assumption 6.2.0.4 1. *The parameter* $\boldsymbol{\phi}^0$ *is the unique root of* $\rho(\boldsymbol{\phi}) = 0$.

2. *The matrix* $\mathrm{C}(\boldsymbol{\phi}^0)$ *defined in (6.4.3) is finite.*

Assumption 6.2.0.5 1. *Assume that (R1), in the appendix, hold for* $\tau_{\mathrm{F}_{n_j^0}}$ *and* $\tau_{\mathrm{G}_{n_j^0}}$ *for any* $j = 1, 2, \ldots, k+1$.

2. *Assume that (R2) and (R3), in the appendix, hold for any* $j = 1, 2, \ldots, k+1$ *when we replace* φ *by* $\psi_{j(l)}$, $1 \leq l \leq d + d_1 + \cdots + d_{k+1}$, $\gamma_0(\cdot)$ *by* $\gamma_{j0}(\cdot)$, $\mathrm{H}_1(\cdot)$ *by* $\mathrm{H}_{j1}(\cdot)$, $\mathrm{C}(x)$ *by* $\mathrm{C}_j(x)$ *and* $\mathrm{F}(\cdot)$ *by* $\mathrm{F}_{n_j^0}(\cdot)$.

Assumption 6.2.0.6 *Assume that for every* $j = 1, \ldots, k$ *and for* $t > 0$; $\mathrm{S}_{\mathrm{C}}^{n_j}(t) > 0$ *and* $\mathrm{S}_{\mathrm{G}_{n_j^0}}(t) > 0$.

The first part of the Assumption 6.2.0.4 is quite classical condition in the Z-estimation theory. The second part is used to justify the existence of variance-covariance expression. We use the Assumption 6.2.0.5 for the SLLN and CLT of each true sub-sample in the presence of censoring. Assumption 6.2.0.6 is imposed to justify the finiteness of some expressions when we have $\mathrm{S}_{\mathrm{C}}^{n_j}(\cdot)$ and $\mathrm{S}_{\mathrm{G}_{n_j^0}}(\cdot)$ in the denominator for each j.

6.3 Asymptotic results

In this section, we establish the consistency of the M-estimators by using the argmax theorem in van der Vaart and Wellner [1996]. For reader convenience, let us recall the basic idea. If the argmax functional is continuous with respect to some metric on the space of the criterion functions, then convergence in distribution of the criterion functions will imply the convergence

in distribution of their points of maximum, the M-estimators, to the maximum of the limit criterion function. So in this section we will give our first main result; the weak consistency of the estimators $\widehat{\alpha}, \widehat{\theta}_1, \ldots, \widehat{\theta}_{k+1}, \widehat{\lambda}_1, \ldots, \widehat{\lambda}_k$, which it will be considered as an initial step for the next results, where we will treat the rate of convergence and the asymptotic distribution of the estimators $\widehat{\alpha}, \widehat{\theta}_1, \ldots, \widehat{\theta}_{k+1}$. The results presented in this section extends and complements the theory of He and Severini [2010] in several ways. On the first hand, when all the data are observed and the criterion function is replaced by the probability density function, i.e., $\mathbf{m}_j(\cdot) = \log f_j(\cdot)$, our Theorem 6.3.0.1 becomes their Theorem 2.1 and our Theorem 6.3.0.3 becomes their Theorem 2.2. On the other hand, we consider the censored data setting in semi-parametric models that is quite different from the framework of the last mentioned reference. Let us recall that the estimators $(\widehat{\alpha}, \widehat{\theta}_1, \ldots, \widehat{\theta}_{k+1}, \widehat{n}_1, \ldots, \widehat{n}_k)$ are defined in equation (6.2.1). The following theorem gives the consistency of the model's parameters estimators

$(\widehat{\alpha}, \widehat{\theta}_1, \ldots, \widehat{\theta}_{k+1}, \widehat{n}_1, \ldots, \widehat{n}_k)$.

Theorem 6.3.0.1 (Consistency) *Under Assumption 3.1, part 1 of Assumption 3.2, part 1 of Assumption 3.3 and Assumption 3.6, we have, as* $n \to \infty$,

$$\widehat{\lambda}_i \xrightarrow{\mathbb{P}} \lambda_i^0, \ \widehat{\theta}_j \xrightarrow{\mathbb{P}} \theta_j^0 \ \text{and} \ \widehat{\alpha} \xrightarrow{\mathbb{P}} \alpha^0,$$

where

$$\widehat{\lambda}_i = \frac{\widehat{n}_i}{n}$$

for $i = 1, \ldots, k$ *and* $j = 1, \ldots, k+1$.

Remark 6.3.0.2 *It is worth noting that* $\widehat{n}_i$, $i = 1, \ldots, k$ *are not consistent. Here we consider the consistency of the change point fractions* $\widehat{\lambda}_i$, $i = 1, \ldots, k$, *in a similar spirit as in Hinkley [1970]. The weak consistency of the parameters* $\widehat{\alpha}$ *and* $\widehat{\theta}_j$, $j = 1, \ldots, k+1$ *is based on the classical* M-*estimators techniques for the censored data in the complex setting of the multiple change-points models.*

The proof of this theorem is based on the proof of Theorem 6.3.0.1 in He and Severini [2010]. The proof of Theorem 6.3.0.1 is captured in the forthcoming Sect. 6.7.

The following theorem give the convergence rate of the estimator $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_k$ the change-points coefficients $\lambda_1, \ldots, \lambda_k$.

Theorem 6.3.0.3 (Convergence rate) *Under Assumption 3.1, part 1 of Assumption 3.2, Assumption 3.3 and Assumption 3.6, we have*

$$\lim_{\eta \to \infty} \lim_{n \to \infty} \mathbb{P}\left(n \left\| \widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 \right\|_\infty \geq \eta\right) = 0,$$

where

$$\widehat{\boldsymbol{\lambda}} = \left(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_k\right), \quad \left\| \widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 \right\|_\infty = \max_{1 \leq j \leq k} \left| \widehat{\lambda}_j - \lambda_j^0 \right|.$$

That is, for $i = 1, 2, \ldots, k$,

$$\widehat{\lambda}_i - \lambda_i^0 = \mathrm{O}_{\mathbb{P}}\left(n^{-1}\right).$$

The proof of this theorem is based on the proof of Theorem 2.2 in He and Severini [2010]. The proof of Theorem 6.3.0.3 is captured in the forthcoming Sect. 6.7.

Remark 6.3.0.4 *The proof of the asymptotic distribution of* $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_k$*, should require a complex methodology, and we leave this problem open for future research.*

Remark 6.3.0.5 *In the comparison of the nonparametric regression estimators, Korostelëv and Tsybakov [1993] argued that the minimax approach is one of the correct ways. Raimondo [1998] considered the sharp change-point problem as an extension of earlier problems in change-point analysis related to the nonparametric regression. Raimondo [1998] proposed a test function for the local regularity of a signal that characterizes such a point as a global maximum and developed a suboptimal wavelet estimator. Goldenshluger* et al. *[2008] considered the problem of nonparametric estimation of signal change-points from indirect and noisy observations, where the estimation problem is analyzed in a general minimax framework. The authors provide lower bounds for minimax risks and propose rate-optimal estimation procedures, one can refer to the last reference for more details on the subject. Shiryaev [2016] considered the change-point quickest detection problem for Brownian motion. The minimax test proposed by Lorden [1971], is used to solve this problem. An original complete and remarkable proof of the CUSUM statistics optimality is constructed and given in detail. Pergamenchtchikov and Tartakovsky [2019] established very general conditions for some models under which the weighted Shiryaev-Roberts procedure is asymptotically optimal, in the minimax sense. In the setting of the multiple change-points when the number of change-points in known, Bai and Perron [1998] obtained the rate* $1/n$ *in the multiple linear regression setting, even the least-squares estimator is consistent with the optimal rate* $1/n$*; see Hao* et al. *[2013] and the references therein. Using the maximum likelihood estimators, He and Severini [2010], obtained the same rate, while in the nonparametric maximum likelihood approach Dumbgen* et al. *[1991] showed that the optimal rate is* $1/n$ *in the single change-point setting, which is generalized by Zou* et al. *[2014a] when they fixed the number of change-points. Notice that the rate* $1/n$ *obtained in Theorem 6.3.0.3 is the minimax rate when the number of change-points is known. The rate convergence* $1/n$ *of the estimated change-points fractions plays a crucial role to obtain standard root-n asymptotic normality of the estimated parameter* $\hat{\boldsymbol{\phi}}$*.*

6.4 Z-estimators

In this section, we give the Z-estimators of $\boldsymbol{\phi}$ when the functions $\mathbf{m}_j(\cdot)$ are differentiable in $\boldsymbol{\phi}$, in two step the first step is maximizing the equation (6.1.1) in n_j, $j = 1, 2 \ldots, k$, and in the second step, we find the solution to the estimating equation given by

$$\begin{aligned} \rho_n(\alpha, \theta_1, \ldots, \theta_{k+1}) = \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\phi}} &= \sum_{j=1}^{k+1} \frac{(\widehat{n}_j - \widehat{n}_{j-1})}{n} \int_{\mathbb{R}} \psi_j(\alpha, \theta_j, x) d\widehat{\mathrm{F}}_{\widehat{n}_j}(x), \\ &= \sum_{j=1}^{k+1} (\widehat{\lambda}_j - \widehat{\lambda}_{j-1}) \rho_{\widehat{n}_j}(\alpha, \theta_j), \end{aligned} \tag{6.4.1}$$

where $\widehat{n}_j$ is the maximizers of n_j and $\psi_j(\alpha,\theta_j,x)=\frac{\partial \mathbf{m}_j(\alpha,\theta_j,x)}{\partial \boldsymbol{\phi}_i}$, $i=1,\ldots,k+2$, from $\Upsilon\times\Theta_j\times\mathbb{R}$ to $\mathbb{R}^{d+d_1+\cdots+d_{k+1}}$; satisfies

$$\rho(\alpha^0,\theta_1^0,\ldots,\theta_{k+1}^0)=\sum_{j=1}^{k+1}\frac{(n_j^0-n_{j-1}^0)}{n}\int_{\mathbb{R}}\psi_j(\alpha^0,\theta_j^0,x)d\mathrm{F}_{n_j^0}(x)=0,$$

and, for each $j=1,2,\ldots,k+1$,

$$\rho_{n_j^0}(\alpha^0,\theta_j^0)=\int_{\mathbb{R}}\psi_j(\alpha^0,\theta_j^0,x)d\mathrm{F}_{n_j^0}(x)=0.$$

Let

$$\begin{aligned}\rho_n^0(\alpha,\theta_1,\ldots,\theta_{k+1})=\frac{\partial \boldsymbol{\ell^0}}{\partial\boldsymbol{\phi}}&=\sum_{j=1}^{k+1}\frac{(n_j^0-n_{j-1}^0)}{n}\int_{\mathbb{R}}\psi_j(\alpha,\theta_j,x)d\widehat{\mathrm{F}}_{n_j^0}(x).\\&=\sum_{j=1}^{k+1}(\lambda_j^0-\lambda_{j-1}^0)\rho_{n_j^0}^0(\alpha,\theta_j).\end{aligned}$$

Notice that Z-estimators include the maximum likelihood estimators, when

$$\psi_j(\boldsymbol{\phi},x)=\frac{\partial\log f_j(\boldsymbol{\phi},x)}{\partial\boldsymbol{\phi}},$$

where $f(\cdot)$ is the density function, generalized method of moment estimators when

$$\psi_j(\boldsymbol{\phi},x)=h(x)-\mathbb{E}_{\boldsymbol{\phi}}h(x),$$

for some function $h(\cdot)$, asymptotic properties are given in Huber [1981], Serfling [1980], van der Vaart and Wellner [1996] and van der Vaart [1998] among others. For the censored data, the case

$$\psi_j(\boldsymbol{\phi},x)=\frac{\partial\log f_j(\boldsymbol{\phi},x)}{\partial\boldsymbol{\phi}},$$

no longer correspond to the maximum likelihood estimators. Oakes [1986] referred to this particular type of Z-estimator as the approximate maximum likelihood estimators and points out its computational and potential robustness advantages over the classical maximum likelihood estimators. Wang [1999] has established the strong consistency of this type of estimators. The asymptotic normality is obtained, under restrictive conditions, by Reid [1981]. Wang [1999] established general asymptotic normality results, which are comparable to those in Cramér [1946], Huber [1967] and subsequent work, he provided the influence curves of a Z-estimator. In this section, we give the asymptotic results and the rate of convergence of Z-estimators under censored data in models with multiple change-points, after approximating the points of change and giving the general conditions for the asymptotic normality, similar to those considered in Wang [1999]. The main hurdle for the full development of the asymptotic properties of Z-estimators is the work of Stute [1995] obtained the most general CLT for $\int\varphi d\widehat{\mathrm{F}}_n$ with an arbitrary function $\varphi(\cdot)$. For any $j=1,2,\ldots,k+1$, let $\psi_{j(l)}(\alpha,\theta_j,\cdot)$ denote the l-th component of

$\psi_j(\alpha,\theta_j,\cdot)$. Replace $\varphi(\cdot)$ by $\psi_{j(l)}(\alpha,\theta_j,\cdot)$ in (6.8.3) and (6.8.5), $H_0(\cdot)$ (resp. $H_1(\cdot)$, $H_{pn}(\cdot)$) by $H_{j0}(\cdot)$ (resp. $H_{j1}(\cdot)$, $H_{j,pn}(\cdot)$) in (6.8.1) and (6.4.2) where

$$H_{j,pn}(y) = \frac{1}{n_j^0 - n_{j-1}^0} \sum_{i=n_{j-1}^0}^{n_j^0} 1\!\!1_{\{Y_i \le y, \delta_i = p\}}, \quad \text{for } p = 0,1, \tag{6.4.2}$$

$H(\cdot)$ (resp. $F(\cdot)$, $G(\cdot)$) by $H_{n_j^0}(\cdot)$ (resp. $F_{n_j^0}(\cdot)$, $G_{n_j^0}(\cdot)$), $C(\cdot)$ by $C_j(\cdot)$ in (6.8.3) and (6.8.4) and denote the corresponding $\gamma_i(\cdot)$'s and U by $\gamma_{ji(l)}(\cdot), i = 0,1,2$ and $U(\psi_{j(l)})$ respectively. It now follows from Proposition 6.8.1.2, and the multivariate central limit theorem that,

$$\sqrt{n}\int_{\mathbb{R}} \psi_j(\alpha,\theta_j,x)\, d(\widehat{F}_{n_j^0} - F_{n_j^0})(x)$$

converges in distribution to a multivariate normal distribution with zero mean and covariance matrix $C_j(\psi_j,\alpha,\theta_j,F_{n_j^0},G_{n_j^0})$, whose (i,l)-entry is

$$\begin{aligned} C_{j(il)}(\psi_j,\alpha,\theta_j,F_{n_j^0},G_{n_j^0}) &= \mathbb{E}(U(\psi_{j(i)})U(\psi_{j(l)})) \\ &= \mathbb{E}\Big\{[\psi_{j(i)}(\alpha,\theta_j,Y)\gamma_{j0(i)}(Y)\delta + \gamma_{j1(i)}(Y)(1-\delta) - \gamma_{j2(i)}(Y)\int_{\mathbb{R}} \psi_{j(i)}(\alpha,\theta_j,x)\,dF(x)] \\ &\quad [\psi_{j(l)}(\alpha,\theta_j,Y)\gamma_{j0(l)}(Y)\delta + \gamma_{j1(l)}(Y)(1-\delta) - \gamma_{j2(l)}(Y) - \int_{\mathbb{R}} \psi_{j(l)}(\alpha,\theta_j,x)\,dF(x)]\Big\}. \end{aligned}$$

Let

$$C(\boldsymbol{\phi}) = \sum_{j=1}^{k+1} (\lambda_j^0 - \lambda_j^0) C_j(\psi_j,\alpha,\theta_j,F_{n_j^0},G_{n_j^0}), \tag{6.4.3}$$

and

$$\frac{\partial}{\partial \boldsymbol{\phi}}\psi_j(\alpha,\theta_j,x) = \left(\frac{\partial}{\partial \phi_l}\psi_{j(i)}(\alpha,\theta_j,x)\right)_{il},$$

denote the $(d + d_1 + \cdots + d_{k+1}) \times (d + d_1 + \cdots + d_{k+1})$ derivative matrix of ψ with respect to $\boldsymbol{\phi}$, let $\Gamma_{F_{n_j^0}}(t)$ and $\Gamma(t)$ denote the $(d + d_1 + \cdots + d_{k+1}) \times (d + d_1 + \cdots + d_{k+1})$ matrix with

$$\Gamma_{F_{n_j^0}}(t) = \int \frac{\partial}{\partial \boldsymbol{\phi}}\psi_j(\alpha,\theta_j,x)\,|_{\boldsymbol{\phi}=t}\, dF_{n_j^0}(x),$$

$$\Gamma(t) = \sum_{j=1}^{k+1} \frac{n_j^0 - n_{j-1}^0}{n} \Gamma_{F_{n_j^0}}(t), \tag{6.4.4}$$

$$\Sigma = \left[\Gamma(\boldsymbol{\phi}^0)\right]^{-1} C(\boldsymbol{\phi}^0) \left[\Gamma(\boldsymbol{\phi}^0)^\top\right]^{-1}, \tag{6.4.5}$$

where $A^\top$ denotes the transpose of a matrix A.
The following theorem gives the consistency of $\widehat{\boldsymbol{\phi}}$.

Theorem 6.4.0.1 *Under the Assumptions of Theorem 6.3.0.3, the function $\rho(\cdot)$ is continuous and for every $\epsilon > 0$, for $n \to \infty$,*

$$\sup_{\boldsymbol{\phi}\in\Phi} \left\|\rho_n^0(\boldsymbol{\phi}) - \rho(\boldsymbol{\phi})\right\| \xrightarrow{\mathbb{P}} 0,$$

$$\inf_{\boldsymbol{\phi}:\|\boldsymbol{\phi}^0\|\ge\epsilon} \left\|\rho(\boldsymbol{\phi})\right\| > 0 = \left\|\rho\left(\boldsymbol{\phi}^0\right)\right\|.$$

Then any sequence of estimators $\widehat{\boldsymbol{\phi}}$ such that $\rho_n\left(\widehat{\boldsymbol{\phi}}\right) = o_{\mathbb{P}}(1)$ converges in probability to $\boldsymbol{\phi}^0$.

The proof of Theorem 6.4.0.1 is captured in the forthcoming Sect. 6.7.
The conditions of the last theorem are given in van der Vaart [1998] when the data are complete and without change in distribution, here we give the conditions under the presence of censoring where we use the Kaplan-Meier integral, the first condition of this theorem is satisfies when the families

$$\mathscr{F}_j = \{\psi_j(\alpha, \theta_j, \cdot), \alpha \in \Upsilon, \theta_j \in \Theta_j\}$$

are Glivenko-Cantelli and the functions $F_{n_j^0}(\cdot)$ are continuous for each $j = 1, 2, \ldots, k+1$ for more detail see Stute [1995] and Bae and Kim [2003], compactness of the set Φ and the continuity of $\psi_j(\cdot)$ for any $j = 1, 2, \ldots, k+1$ with the first part of Assumption 6.2.0.4 implies the condition 2 of Theorem 6.4.0.1.
In the next theorem, we will give weaker conditions than those in the previous theorem, these conditions are introduced in Pakes and Pollard [1989]. Note that the first condition is to insure the estimator $\widehat{\boldsymbol{\phi}}$ is taken as any value that comes close enough to provide a global minimum for $\|\rho_n(\cdot)\|$, since $\boldsymbol{\phi}^0$ is included in the set over which the minimum is taken, $\|\rho_n(\widehat{\boldsymbol{\phi}})\|$ cannot be much bigger than $\|\rho_n(\boldsymbol{\phi}^0)\|$. If the quantity $\rho_n(\boldsymbol{\phi}^0)$ is eventually close to zero, the second assumption on $\rho(\boldsymbol{\phi}^0)$ implies that $\rho_n(\widehat{\boldsymbol{\phi}})$ must also get close to zero. If small values of $\|\rho_n(\boldsymbol{\phi})\|$ can occur only near $\boldsymbol{\phi}^0$, this forces $\widehat{\boldsymbol{\phi}}$ to be close to $\boldsymbol{\phi}^0$ by the third condition.

Theorem 6.4.0.2 *Under the following conditions*

(i)

$$\left\|\rho_n\left(\widehat{\boldsymbol{\phi}}\right)\right\| \leq o_{\mathbb{P}}(1) + \inf_{\boldsymbol{\phi}\in\Phi} \|\rho_n(\boldsymbol{\phi})\|;$$

(ii)

$$\rho_n(\boldsymbol{\phi}^0) = o_{\mathbb{P}}(1);$$

(iii)

$$\sup_{\|\boldsymbol{\phi}^0\|>\eta} \|\rho_n(\boldsymbol{\phi})\|^{-1} = O_{\mathbb{P}}(1) \quad \textit{for each} \quad \eta > 0.$$

Then any sequence of estimators $\widehat{\boldsymbol{\phi}}$ *such that* $\rho_n\left(\widehat{\boldsymbol{\phi}}\right) = o_{\mathbb{P}}(1)$ *converges in probability to* $\boldsymbol{\phi}^0$.

The proof of Theorem 6.4.0.2 is captured in the forthcoming Sect. 6.7.

The next theorem gives conditions under which $\widehat{\boldsymbol{\phi}}$, which is now assumed to converge in probability to $\boldsymbol{\phi}^0$, satisfies a central limit theorem like a Z-estimator. The argument breaks naturally into two steps. First we establish $\sqrt{n}$-consistency by means of a comparison between $\|\rho_n^0(\widehat{\boldsymbol{\phi}})\|$ and $\|\rho_n^0(\boldsymbol{\phi}^0)\|$. Informally stated, the new equicontinuity condition (iii) implies that

$$\|\rho(\boldsymbol{\phi})\| \leq O_{\mathbb{P}}(\|\rho_n(\boldsymbol{\phi})\|) + O_{\mathbb{P}}\left(\left\|\rho_n\left(\boldsymbol{\phi}^0\right)\right\|\right) + o_{\mathbb{P}}\left(n^{-1/2}\right)$$

uniformly near $\boldsymbol{\phi}^0$. Since $\widehat{\boldsymbol{\phi}}$ comes close to minimizing $\|\rho_n(\cdot)\|$, the quantity $\|\rho_n(\widehat{\boldsymbol{\phi}})\|$ cannot be much larger than $\|\rho_n(\boldsymbol{\phi}^0)\|$, which is of order $O_{\mathbb{P}}(n^{-1/2})$. Approximate linearity of $\rho(\cdot)$ in a neighborhood of $\boldsymbol{\phi}^0$ transfers the same rate of convergence to $\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0$. The argument for the second step need only values of $\boldsymbol{\phi}$ in a $O_{\mathbb{P}}\left(n^{-1/2}\right)$ neighborhood of $\boldsymbol{\phi}^0$ (see page 1040 in Pakes

and Pollard [1989]). The combination of conditions (ii) and (iii) shows that $\rho_n^0(\cdot)$ is uniformly well approximated by a linear function $L_n(\cdot)$. The $\boldsymbol{\phi}_n^*$ that minimizes $\|L_n(\cdot)\|$ has an explicit form, from which asymptotic normality of $\sqrt{n}(\boldsymbol{\phi}_n^* - \boldsymbol{\phi}^0)$ is easily established. A comparison between $\|\rho_n(\boldsymbol{\phi}_n^*)\|$ and $\|\rho_n^0(\widehat{\boldsymbol{\phi}})\|$ shows that $\widehat{\boldsymbol{\phi}}$ must lie within $O_{\mathbb{P}}(n^{-1/2})$ of $\boldsymbol{\phi}_n^*$, which implies the desired central limit theorem.
The following theorem provides the central limit theorem for the estimator $\widehat{\boldsymbol{\phi}}$.

Theorem 6.4.0.3 *Let $\widehat{\boldsymbol{\phi}}$ be a consistent estimator of $\boldsymbol{\phi}^0$, under the Assumptions of Theorem 6.3.0.3, Assumption 6.2.0.4 and*

(i) $\|\rho_n^0(\widehat{\boldsymbol{\phi}})\| \leq o_{\mathbb{P}}(n^{-1/2})$;

(ii) $\rho(\cdot)$ *is differentiable at* $\boldsymbol{\phi}^0$ *with a derivative matrix* Ω *of full rank;*

(iii) for every sequence η_n *of positive numbers that converges to zero,*

$$\sup_{\|\boldsymbol{\phi}-\boldsymbol{\phi}^0\|<\eta_n} \frac{\|\rho_n^0(\boldsymbol{\phi}) - \rho(\boldsymbol{\phi}) - \rho_n^0(\boldsymbol{\phi}^0)\|}{n^{-1/2} + \|\rho_n^0(\boldsymbol{\phi})\| + \|\rho(\boldsymbol{\phi})\|} = o_{\mathbb{P}}(1);$$

(iv) $\boldsymbol{\phi}^0$ *is an interior point of* Φ,

then we have, as $n \to \infty$,

$$\sqrt{n}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0) \rightsquigarrow N\left(0, (\Omega^{-1})C(\boldsymbol{\phi}^0)(\Omega^{-1})^\top\right).$$

The proof of Theorem 6.4.0.3 is captured in the forthcoming Sect. 6.7.
From Proposition 6.8.1.2 the central limit theorem follows. Note that if we can interchange between the integration and differentiation in (6.4.5), we take

$$\Omega = \Gamma(\boldsymbol{\phi}^0).$$

The proof of Theorem 6.4.0.3 is similar to the proof in Pakes and Pollard [1989] but in our case, $\rho_n^0(\cdot)$ (resp $\rho_{n_j^0}^0(\cdot), j = 1, 2, \ldots, k+1$) is not available, we have only $\rho_n(\cdot)$ (respectively $\rho_{\widehat{n}_j}(\cdot), j = 1, 2, \ldots, k+1$), the result expression (6.7.4) in Lemma 6.7.0.4 gives us the asymptotic equivalence when n is large enough. The condition (i) and (iii) are automatically fulfilled when

(i)$'$
$$\left\|\rho_{n_j^0}^0(\alpha, \theta_j)\right\| \leq o_{\mathbb{P}}(n^{-1/2}), j = 1, 2, \ldots, k+1;$$

(iii)$'$
$$\sup_{\|(\alpha,\theta_j)-(\alpha^0,\theta_j^0)\|<\eta_n} \left\|\rho_{n_j^0}^0(\alpha, \theta_j) - \rho_{n_j^0}(\alpha, \theta_j) - \rho_{n_j^0}^0(\alpha^0, \theta_j^0)\right\| = o_{\mathbb{P}}(n^{-1/2}), j = 1, 2, \ldots, k+1.$$

Note that for the conditions (i) (resp (i)$'$) and (iii) (resp (iii)$'$) which they are assumed for $\rho_n^0(\cdot)$ (respectively $\rho_{n_j^0}^0(\cdot)$) the same under result in Lemma 6.7.0.4, we can show this conditions are required also for $\rho_n(\cdot)$ (respectively $\rho_{\widehat{n}_j}(\cdot)$) and conversely. In the next theorem, we give the asymptotic normality of $\sqrt{n}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0)$ for $\widehat{\boldsymbol{\phi}}$ as an M-estimator or Z-estimator the proof is much similar.

Theorem 6.4.0.4 (Asymptotic normality) *Under part 2 of Assumption 6.2.0.2 for $\boldsymbol{\phi}$ in a neighborhood of $\boldsymbol{\phi}^0$, and let $\Gamma(\boldsymbol{\phi}^0)$ defined in (6.4.4) be a finite and non-singular $(d + d_1 + \cdots + d_{k+1}) \times (d + d_1 + \cdots + d_{k+1})$ matrix. Assume that the assumptions of Lemma 6.7.0.5 with part 1 of Assumption 6.2.0.5 hold for*

$$s(\boldsymbol{\phi}, x) = \left(\frac{\partial}{\partial \phi_l} \psi_{j(i)}(\alpha, \theta_j, x) \right)_{il}, \quad 1 \leq i, l \leq d + d_1 + \cdots + d_{k+1},$$

for any j, and part 2 of Assumption 6.2.0.5. Under Assumption 6.2.0.3 and Assumption 6.2.0.4, any sequence of Z-estimates $\widehat{\boldsymbol{\phi}}$ satisfying

$$\widehat{\boldsymbol{\phi}} \xrightarrow{\mathbb{P}} \boldsymbol{\phi}^0$$

is asymptotically normal with

$$\sqrt{n}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0) \rightsquigarrow \mathrm{N}(0, \Sigma),$$

where Σ is defined in (6.4.5).

The proof of Theorem 6.4.0.4 is captured in the forthcoming Sect. 6.7.

Remark 6.4.0.5 *Change-point detection has received enormous attention due to the emergence of an increasing amount of temporal data. In the present work, we are mainly concerned with the estimation of the model parameters. We have assumed that the number of changes in the sample is known, which is not the case in real application. Without the need to know the number of change-points in advance, Zou et al. [2014b] proposed a nonparametric maximum likelihood approach to detecting multiple change-points. It is worth noting that the determination of the number of change-points k in a dataset has been crucial to multiple change-points analysis for a long time. It is often approached as a model selection problem, since k drives the model dimension. we can use the binary segmentation (BinSeg) method proposed in Vostrikova [1981], which is a "top down" procedure, in the sense that one tests all the data to determine if there is at least one change-point and iterates the procedure in the intervals immediately to the "left" and "right" of the most recently detected change-point. This procedure is widely used motivated by the low computational complexity and the is conceptually easy to implement compared to the Exhaustive Search as described by Niu et al. [2016] in Section 3.1. Each stage of BinSeg involves search for a single change-point, which means that if a given segment contains multiple change-points in certain unfavourable configurations, BinSeg may fail to perform adequately on it, as it attempts to fit the "wrong" model. Fryzlewicz [2014] shows that relatively restrictive theoretical assumptions are needed for BinSeg to offer near-optimal performance*

in terms of the accuracy of estimation of the change-point locations, refer to Korkas and Fryzlewicz [2017] and Fryzlewicz [2018]. In the last reference a new solution is proposed giving a 'tail-greedy', bottom-up transform for one-dimensional data, which results in a nonlinear but conditionally orthonormal, multiscale decomposition of the data with respect to an adaptatively chosen unbalanced Haar wavelet basis, which avoids the disadvantages of the classical divisive BinSeg. When the number of changes is unknown, Lavielle [1999], Lavielle and Ludeña [2000] proposed its estimation by minimizing a penalized contrast function. Very recently, Zou et al. *[2020] proposed a data-driven selection criterion that is applicable to most kinds of popular change-point detection methods, including in particular the binary segmentation and the optimal partitioning algorithms. The main idea is to select the number of change-points that minimizes the squared prediction error, which measures the fit of a specified model for a new sample. The authors investigated a unified parametric framework which includes classical univariate or multivariate location and scale problems, ordinary least-squares, generalized linear models, and many others as special cases, provided that the corresponding objective (likelihood or loss) function can be recast into their asymptotically equivalent least-squares problems. In Zou* et al. *[2014c], the number of change-points is determined by the Bayesian information criterion and the locations of the change-points can be estimated via the dynamic programming algorithm and the use of the intrinsic order structure of the likelihood function. Under some general conditions, Zou* et al. *[2014c] showed that the new method provides consistent estimation with an optimal rate. We refer to the last reference for more discussions. For more details, we refer to Truong* et al. *[2020], where the authors presented a selective survey of algorithms for the offline detection of multiple change-points.*

6.5 Maximum likelihood estimators

In this section we will consider the maximum likelihood estimators in models with multiple change points in the censored data framework. To unburden our notation a bit, we assume that the censoring variables C are independent and identically distributed with distribution function $G(\cdot)$ and density function $g(\cdot)$, with respect to the Lebesgue measure λ. Let the lifetime X and the censoring time C be positive continuous random variables assumed to be independent. Recall that, the distribution function of the lifetime X is $F(\alpha, \theta, \cdot)$ with density function $f(\alpha, \theta, \cdot)$, with respect to the Lebesgue measure λ, where α and θ are the unknown parameters to be estimated. In the random censorship from the right model, one observes the pairs (Y, δ), where $Y = \min(X, C)$ and $\delta = 1\!\!1\{X \leqslant C\}$. Let $(Y_i, \delta_i), 1 \leqslant i \leqslant n$, denote a random sample of (Y, δ) that one observes, and $Y_{(1)} < \cdots < Y_{(m)}$ denote the m distinct ordered values of Y's. When there are ties among the Y's, we have $m < n$. The likelihood function for this sample is given by

$$\mathscr{L}(\alpha, \theta) = \prod_{i=1}^{n} f_{Y,\Delta}\left(\alpha, \theta, \delta_i, y_i\right),$$

where $f_{Y,\Delta}(\cdot)$ is the density function of the couple (Y, Δ) with respect to the product measure $\lambda \otimes \mu$ with λ is the measure of Lebesgue and μ is the counting measure on the set $\{0, 1\}$. The

likelihood function can be rewritten as follows

$$\mathscr{L}(\alpha,\theta)=\prod_{i=1}^{n} f_{Y,\Delta}\left(\alpha,\theta,\delta_i,y_i\right)=\prod_{i=1}^{n}\left(f\left(\alpha,\theta,\delta_i,y_i\right)\mathrm{G}\left(y_i\right)\right)^{\delta_i}\left(\mathrm{g}\left(y_i\right)\left(1-\mathrm{F}\left(\alpha,\theta,\delta_i,y_i\right)\right)\right)^{1-\delta_i}. \tag{6.5.1}$$

By the hypothesis that the distribution of the censored data is independent of the unknown parameters α and θ so the maximization of $(\alpha,\theta)\mapsto\mathscr{L}(\alpha,\theta)$ is equivalent to the maximization of the pseudo-likelihood given by

$$\mathrm{L}(\alpha,\theta)=\prod_{i=1}^{n}\left(f\left(\alpha,\theta,y_i\right)\right)^{\delta_i}\left(1-\mathrm{F}\left(\alpha,\theta,y_i\right)\right)^{1-\delta_i}. \tag{6.5.2}$$

Now, we consider model with known k change in the distribution, i.e.,

$$\mathrm{X}_i\sim\mathrm{F}(\alpha,\theta,x),\quad n_{j-1}+1\le i\le n_j, j=1,\ldots,k+1; i=1,\ldots,n.$$

In this case, the likelihood function given in (6.5.2), can be written as follows

$$\mathrm{L}(\alpha,\theta_1,\ldots,\theta_{k+1},n_1,\ldots,n_k)=\prod_{j=1}^{k+1}\prod_{i=n_{j-1}+1}^{n_j}\left(f\left(\alpha,\theta_j,y_i\right)\right)^{\delta_i}\left(1-\mathrm{F}\left(\alpha,\theta_j,y_i\right)\right)^{1-\delta_i},$$

which implies that the log-likelihood function is given by

$$\begin{aligned}\ell &\equiv \ell(\alpha,\theta_1,\ldots,\theta_{k+1},n_1,\ldots,n_k)\\ &= \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}+1}^{n_j}\left\{\delta_i\log f\left(\alpha,\theta_j,y_i\right)+(1-\delta_i)\log(1-\mathrm{F}\left(\alpha,\theta_j,y_i\right))\right\},\end{aligned} \tag{6.5.3}$$

where $\mathrm{F}\left(\alpha,\theta_j,y\right)>0$ for all $j=1,\ldots,k+1$. The maximization is taken with respect to the vector $(\alpha,\theta_1,\ldots,\theta_{k+1},n_1,\ldots,n_k)$, so the multiplication by the factor $1/n$ does not affect the optimization problem, which is needed for asymptotic results.

Although only two examples will be given here, they stand as archetypes for a variety of parametric families that can be investigated in a similar way. Let us specify the log-likelihood function for the exponential and Gaussian random variables.

Exponential distribution

We consider the following model

$$\begin{aligned}&\mathrm{X}_i\sim\mathrm{Exp}(\theta_j),\quad n_{j-1}+1\le i\le n_j, j=1,\ldots,k+1; i=1,\ldots,n.\\ &\mathrm{C}_i\sim\mathrm{Exp}(\beta_j),\quad n_{j-1}+1\le i\le n_j, j=1,\ldots,k+1; i=1,\ldots,n,\end{aligned} \tag{6.5.4}$$

where $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_{k+1})$ is assumed to be known. The log-likelihood function is given by

$$\begin{aligned}\ell(\theta_1,\ldots,\theta_{k+1},n_1,\ldots,n_k) &= \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}+1}^{n_j}\left\{\delta_i\log\left(\theta_j e^{-\theta_j y_i}\right)+(1-\delta_i)\log\left(e^{-\theta_j y_i}\right)\right\} \qquad (6.5.5)\\ &= \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}+1}^{n_j}\left\{\delta_i\log(\theta_j)-\delta_i\theta_j y_i-(1-\delta_i)\theta_j y_i\right\}\\ &= \frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}+1}^{n_j}\left\{\delta_i\log(\theta_j)-\theta_j y_i\right\},\end{aligned}$$

where y_i are the observed values.

Normal distribution

We now consider the uncensored case, where the variables are normal with change only in mean from segment to segment and fixed variance, this means that the change occurs only in θ_j and $\alpha \equiv 1$, i.e.,

$$X_i \sim \mathcal{N}(\theta_j, 1), \quad n_{j-1}+1 \le i \le n_j, j = 1,\ldots,k+1; i = 1,\ldots,n. \tag{6.5.6}$$

The log-likelihood function in this case is given by

$$\ell(\theta_1,\ldots,\theta_{k+1}, n_1,\ldots,n_k) = -\frac{1}{n}\sum_{j=1}^{k+1}\sum_{i=n_{j-1}+1}^{n_j}\frac{(x_i-\theta_j)^2}{2}.$$

6.6 Numerical results

This section is concerned with the evaluation of the finite sample performance of the proposed estimation procedure using the the maximum likelihood in (6.5.2) with samples of different sizes and different censoring rate. We provide numerical illustrations regarding the bias, the variance and the root mean-squared error RMSE. The computing program codes were implemented in R. In our simulation, we choose one sample of $n = 1000$ observations with 10 change-points, i.e., $k = 10$ with true location;

$$n\boldsymbol{\lambda}^0 = (50, 150, 240, 330, 410, 520, 610, 710, 820, 930)$$

and we consider two cases of true within-parameter

a. The first case is:

$$\boldsymbol{\theta}^0 = (5, 3, 1, 6, 2, 7, 3, 1, 8, 2, 7). \tag{6.6.1}$$

b. The second case is:

$$\boldsymbol{\theta}^0 = (0.5, 0.3, 1, 1.6, 0.2, 0.75, 0.35, 1, 0.5, 2, 1.5). \tag{6.6.2}$$

We will consider different intensities of censoring in the sample. The censoring random variables $C_1,\ldots,C_n$ are generated from distribution depending on some parameter $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_{11})$ calibrated to attain the desired censoring rate (5%, 10% or 30%). The three scenarios of the censoring rate (proportion) (cr) are given for the first case of true within-parameter as follows.

(i) cr = 5%, with censoring random variables

$$C_i \sim \text{Exp}(\beta_j), \qquad n_{j-1}+1 \le i \le n_j, \quad j = 1,\ldots,k+1; \quad i = 1,\ldots,n,$$

where $\boldsymbol{\beta} = (0.26, 0.16, 0.05, 0.32, 0.11, 0.37, 0.16, 0.05, 0.42, 0.11, 0.37)$;

(ii) cr = 10%, with censoring random variables

$$C_i \sim \text{Exp}(\beta_j), \qquad n_{j-1}+1 \le i \le n_j, \quad j = 1,\dots,k+1; i = 1,\dots,n,$$

where $\boldsymbol{\beta} = (0.56, 0.33, 0.11, 0.67, 0.22, 0.78, 0.33, 0.11, 0.89, 0.22, 0.78)$;

(iii) cr = 30%, with censoring random variables

$$C_i \sim \text{Exp}(\beta_j), \qquad n_{j-1}+1 \le i \le n_j, \quad j = 1,\dots,k+1; \quad i = 1,\dots,n,$$

where $\boldsymbol{\beta} = (2.14, 1.29, 0.43, 2.57, 0.86, 3, 1.29, 0.43, 3.43, 0.86, 3)$.

The following figures display the simulated data.

The simulation results are reported in the following Tables 6.1-6.8.

cr=5%					
Parameter	True value	Mean	BIAS	SD	RMSE
n_1	50	41.889	-8.111	8.17	11.512
n_2	150	149.757	-0.243	0.938	0.969
n_3	240	240.086	0.086	0.5	0.507
n_4	330	329.256	-0.744	1.794	1.942
n_5	410	413.289	3.289	1.137	3.48
n_6	520	524.486	4.486	1.591	4.76
n_7	610	607.943	-2.057	0.248	2.071
n_8	710	707.961	-2.039	1.783	2.708
n_9	820	819.895	-0.105	0.531	0.541
n_{10}	930	930.058	0.058	0.246	0.252
θ_1	5	5.018	0.018	0.278	0.279
θ_2	3	3.153	0.153	0.093	0.179
θ_3	1	0.952	-0.048	0.024	0.053
θ_4	6	5.665	-0.335	0.143	0.363
θ_5	2	2.104	0.104	0.055	0.118
θ_6	7	6.747	-0.253	0.167	0.302
θ_7	3	2.943	-0.057	0.083	0.101
θ_8	1	0.83	-0.17	0.019	0.17
θ_9	8	6.217	-1.783	0.181	1.791
θ_{10}	2	2.257	0.257	0.05	0.262
θ_{11}	7	7.398	0.398	0.206	0.448

Table 6.1: Maximum likelihood estimator for censored case sample size 1000, Exponential distribution, cr=5%.

cr=10%					
Parameter	True value	Mean	BIAS	SD	RMSE
n_1	50	44.402	-5.598	10.202	11.637
n_2	150	149.192	-0.808	1.794	1.967
n_3	240	240.285	0.285	1.536	1.562
n_4	330	328.762	-1.238	2.248	2.567
n_5	410	413.502	3.502	1.594	3.847
n_6	520	526.493	6.493	2.786	7.065
n_7	610	607.891	-2.109	0.35	2.137
n_8	710	708.551	-1.449	2.156	2.598
n_9	820	819.698	-0.302	0.931	0.978
n_{10}	930	930.252	0.252	1.409	1.431
θ_1	5	4.712	-0.288	0.324	0.433
θ_2	3	2.978	-0.022	0.114	0.116
θ_3	1	0.903	-0.097	0.034	0.102
θ_4	6	5.408	-0.592	0.211	0.628
θ_5	2	2	0	0.075	0.075
θ_6	7	6.349	-0.651	0.218	0.686
θ_7	3	2.758	-0.242	0.11	0.265
θ_8	1	0.79	-0.21	0.027	0.211
θ_9	8	5.934	-2.066	0.223	2.077
θ_{10}	2	2.136	0.136	0.068	0.152
θ_{11}	7	7.029	0.029	0.276	0.278

Table 6.2: Maximum likelihood estimator for censored case sample size 1000, Exponential distribution, cr=10%.

cr=30%					
Parameter	True value	Mean	BIAS	SD	RMSE
n_1	50	46.634	-3.366	11.468	11.952
n_2	150	147.882	-2.118	3.559	4.142
n_3	240	240.709	0.709	3.765	3.832
n_4	330	327.203	-2.797	3.874	4.778
n_5	410	415.482	5.482	4.573	7.139
n_6	520	526.44	6.44	5.595	8.531
n_7	610	607.617	-2.383	0.808	2.516
n_8	710	709.498	-0.502	2.976	3.018
n_9	820	818.8	-1.2	3.079	3.304
n_{10}	930	930.896	0.896	3.987	4.086
θ_1	5	3.689	-1.311	0.412	1.374
θ_2	3	2.318	-0.682	0.161	0.7
θ_3	1	0.699	-0.301	0.051	0.304
θ_4	6	4.284	-1.716	0.346	1.75
θ_5	2	1.56	-0.44	0.119	0.455
θ_6	7	5.014	-1.986	0.342	2.014
θ_7	3	2.121	-0.879	0.169	0.894
θ_8	1	0.613	-0.387	0.042	0.388
θ_9	8	4.683	-3.317	0.321	3.332
θ_{10}	2	1.658	-0.342	0.105	0.357
θ_{11}	7	5.481	-1.519	0.456	1.585

Table 6.3: Maximum likelihood estimator for censored case sample size 1000, Exponential distribution, cr=30%.

The uncensored case with exponential distribution					
Parameter	True value	Mean	BIAS	SD	RMSE
n_1	50	51.99	1.99	7.333	7.598
n_2	150	151.035	1.035	4.612	4.727
n_3	240	239.578	-0.422	1.711	1.762
n_4	330	330.995	0.995	4.8	4.902
n_5	410	409.522	-0.478	3.634	3.665
n_6	520	521.002	1.002	7.534	7.601
n_7	610	611.214	1.214	4.83	4.98
n_8	710	709.687	-0.313	1.23	1.269
n_9	820	820.649	0.649	2.977	3.047
n_{10}	930	929.143	-0.857	3.341	3.449
θ_1	5	5.327	0.327	0.793	0.858
θ_2	3	2.992	-0.008	0.309	0.309
θ_3	1	0.994	-0.006	0.109	0.109
θ_4	6	6.162	0.162	0.705	0.724
θ_5	2	1.975	-0.025	0.231	0.232
θ_6	7	7.222	0.222	0.754	0.786
θ_7	3	3.015	0.015	0.347	0.347
θ_8	1	0.995	-0.005	0.102	0.102
θ_9	8	8.16	0.16	0.799	0.814
θ_{10}	2	1.995	-0.005	0.196	0.196
θ_{11}	7	7.177	0.177	0.913	0.93

Table 6.4: Maximum likelihood estimator for uncensored case sample size 1000, Exponential distribution.

cr=5%					
Parameter	True value	Mean	BIAS	SD	RMSE
n_1	50	51.806	1.806	1.101	2.115
n_2	150	150.306	0.306	0.813	0.869
n_3	240	240.098	0.098	0.669	0.676
n_4	330	328.851	-1.149	0.486	1.247
n_5	410	410.159	0.159	0.722	0.739
n_6	520	516.455	-3.545	4.788	5.958
n_7	610	610.061	0.061	0.247	0.254
n_8	710	709.884	-0.116	0.429	0.444
n_9	820	820.073	0.073	0.325	0.333
n_{10}	930	962.175	32.175	6.143	32.756
θ_1	0.5	0.452	-0.048	0.014	0.049
θ_2	0.3	0.274	-0.026	0.006	0.026
θ_3	1	0.901	-0.099	0.021	0.1
θ_4	1.6	1.608	0.008	0.037	0.038
θ_5	0.2	0.199	-0.001	0.005	0.005
θ_6	0.75	0.708	-0.042	0.019	0.046
θ_7	0.35	0.331	-0.019	0.009	0.02
θ_8	1	1.09	0.09	0.025	0.094
θ_9	0.5	0.496	-0.004	0.01	0.011
θ_{10}	2	1.852	-0.148	0.035	0.151
θ_{11}	1.5	1.315	-0.185	0.053	0.192

Table 6.5: Maximum likelihood estimator for censored case sample size 1000, Exponential distribution, cr=5%.

cr=10%					
Parameter	True value	Mean	BIAS	SD	RMSE
n_1	50	51.267	1.267	2.348	2.668
n_2	150	150.536	0.536	1.206	1.32
n_3	240	240.3	0.3	1.574	1.602
n_4	330	329.284	-0.716	1.276	1.463
n_5	410	410.441	0.441	1.492	1.556
n_6	520	517.234	-2.766	5.228	5.915
n_7	610	610.114	0.114	0.347	0.366
n_8	710	709.815	-0.185	0.568	0.597
n_9	820	820.118	0.118	0.387	0.404
n_{10}	930	959.984	29.984	8.658	31.209
θ_1	0.5	0.43	-0.07	0.019	0.072
θ_2	0.3	0.26	-0.04	0.008	0.04
θ_3	1	0.854	-0.146	0.03	0.148
θ_4	1.6	1.512	-0.088	0.07	0.111
θ_5	0.2	0.188	-0.012	0.007	0.014
θ_6	0.75	0.671	-0.079	0.026	0.082
θ_7	0.35	0.312	-0.038	0.012	0.039
θ_8	1	1.035	0.035	0.034	0.049
θ_9	0.5	0.471	-0.029	0.014	0.032
θ_{10}	2	1.759	-0.241	0.05	0.246
θ_{11}	1.5	1.251	-0.249	0.072	0.258

Table 6.6: Maximum likelihood estimator for censored case sample size 1000, Exponential distribution, cr=10%.

cr=30%					
Parameter	True value	Mean	BIAS	SD	RMSE
n_1	50	50	0	4.247	4.247
n_2	150	151.532	1.532	2.612	3.028
n_3	240	242.759	2.759	6.564	7.12
n_4	330	329.113	-0.887	1.923	2.117
n_5	410	410.806	0.806	3.192	3.292
n_6	520	517.033	-2.967	7.82	8.364
n_7	610	610.573	0.573	1.508	1.613
n_8	710	710.099	0.099	3.118	3.119
n_9	820	819.151	-0.849	1.898	2.079
n_{10}	930	953.294	23.294	15.092	27.756
θ_1	0.5	0.34	-0.16	0.028	0.162
θ_2	0.3	0.202	-0.098	0.013	0.098
θ_3	1	0.67	-0.33	0.048	0.332
θ_4	1.6	1.19	-0.41	0.084	0.417
θ_5	0.2	0.144	-0.056	0.01	0.056
θ_6	0.75	0.526	-0.224	0.036	0.226
θ_7	0.35	0.24	-0.11	0.018	0.11
θ_8	1	0.805	-0.195	0.055	0.202
θ_9	0.5	0.36	-0.14	0.022	0.14
θ_{10}	2	1.362	-0.638	0.095	0.644
θ_{11}	1.5	0.984	-0.516	0.137	0.533

Table 6.7: Maximum likelihood estimator for censored case sample size 1000, Exponential distribution, cr=30%.

The uncensored case with exponential distribution					
Parameter	True value	Mean	BIAS	SD	RMSE
n_1	50	52.22	2.22	7.01	7.353
n_2	150	149.176	-0.824	4.311	4.389
n_3	240	239.614	-0.386	11.84	11.847
n_4	330	330.306	0.306	1.468	1.5
n_5	410	409.383	-0.617	3.368	3.424
n_6	520	521.288	1.288	7.961	8.065
n_7	610	609.221	-0.779	5.202	5.26
n_8	710	711.774	1.774	9.263	9.431
n_9	820	819.397	-0.603	2.822	2.886
n_{10}	930	935.153	5.153	17.891	18.618
θ_1	0.5	0.53	0.03	0.079	0.085
θ_2	0.3	0.296	-0.004	0.029	0.03
θ_3	1	0.999	-0.001	0.115	0.115
θ_4	1.6	1.673	0.073	0.184	0.199
θ_5	0.2	0.2	0	0.022	0.022
θ_6	0.75	0.766	0.016	0.075	0.076
θ_7	0.35	0.344	-0.006	0.041	0.041
θ_8	1	1.038	0.038	0.112	0.119
θ_9	0.5	0.493	-0.007	0.046	0.047
θ_{10}	2	2.077	0.077	0.23	0.243
θ_{11}	1.5	1.462	-0.038	0.246	0.249

Table 6.8: Maximum likelihood estimator for uncensored case sample size 1000, Exponential distribution.

After we consider the case of complete data, i.e., $Y_i = X_i$ and $\delta_i = 1$ for all $i = 1, \ldots, n$ in the same model given in (6.5.4), the log-likelihood in (6.5.5) is written in this form

$$\ell(\theta_1, \ldots, \theta_{k+1}, n_1, \ldots, n_k) = \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}+1}^{n_j} \{\log(\theta_j) - \theta_j y_i\},$$

with the same true location $\boldsymbol{\lambda}^0$ and the same true within-parameters $\boldsymbol{\theta}^0$, we have the following results in Table 6.4.
Finally, consider the case of normal distribution for complete data model given in (6.5.6), with sample size $n = 1000$ with 10 change-points, with true location given by

$$n\boldsymbol{\lambda} = (70, 160, 250, 340, 440, 540, 630, 730, 820, 920)$$

and the true within-parameter is given

$$\boldsymbol{\theta} = (-5, 3, 0, 4, -1, 3, -3, 10, 4, -2, 0).$$

The results are reported in Table 6.9.

The uncensored case with normal distribution					
Parameter	True value	Mean	BIAS	SD	RMSE
n_1	70	70	0	0	0
n_2	160	160.003	0.003	0.55	0.55
n_3	250	249.998	-0.002	0.24	0.24
n_4	340	340.001	0.001	0.104	0.104
n_5	440	440.002	0.002	0.268	0.268
n_6	540	540.002	0.002	0.044	0.044
n_7	630	630	0	0	0
n_8	730	730.001	0.001	0.07	0.07
n_9	820	820	0	0.044	0.044
n_{10}	920	920.048	0.048	1.301	1.302
θ_1	-5	-5.004	-0.004	0.123	0.123
θ_2	3	2.996	-0.004	0.105	0.105
θ_3	0	-0.003	-0.003	0.104	0.104
θ_4	4	4.004	0.004	0.106	0.106
θ_5	-1	-1.005	-0.005	0.097	0.097
θ_6	3	2.999	-0.001	0.099	0.099
θ_7	-3	-2.997	0.003	0.103	0.103
θ_8	10	9.995	-0.005	0.095	0.095
θ_9	4	3.997	-0.003	0.103	0.103
θ_{10}	-2	-2.004	-0.004	0.098	0.099
θ_{11}	0	0.003	0.003	0.112	0.112

Table 6.9: Maximum likelihood estimator for uncensored case sample size 1000, normal distribution.

From tables and figures, the best results are obtained when the data is complete, and the results in the censoring case are satisfactory when the censoring rate is moderate 5%, 10% and 30% and the performance are deteriorated when the censoring rate increase. The following figures are computed for the three rates of censoring and for complete data for model given in (6.5.4) with 1000 replicate from samples with sizes from 1000 to 10000 i.e., size = $(70,90,90,90,100,100,90,100,90,100,80) * k$; $k = 1,\ldots,10,$. By inspecting Figures 6.1-6.6 for the first case (6.6.1) and Figures 6.7-6.12 for the second one (6.6.2), one can see that as in any other inferential context, the greater the sample size, the better. In the literature, it is commonly used two or three changes in the sample for the finite sample experiments. In the present simulations, we have optimized the likelihood criterion with respect to 21 parameters $(n_1,\ldots,n_{10},\theta_1,\ldots,\theta_{11})$ simultaneously, including 10 changes in the sample, which has a computational cost. This can be circumvented by using the penalized likelihood criterion. In order to extract methodological recommendations for the use of the procedures proposed in this work, it will be interesting to conduct extensive Monte Carlo experiments to compare our procedures

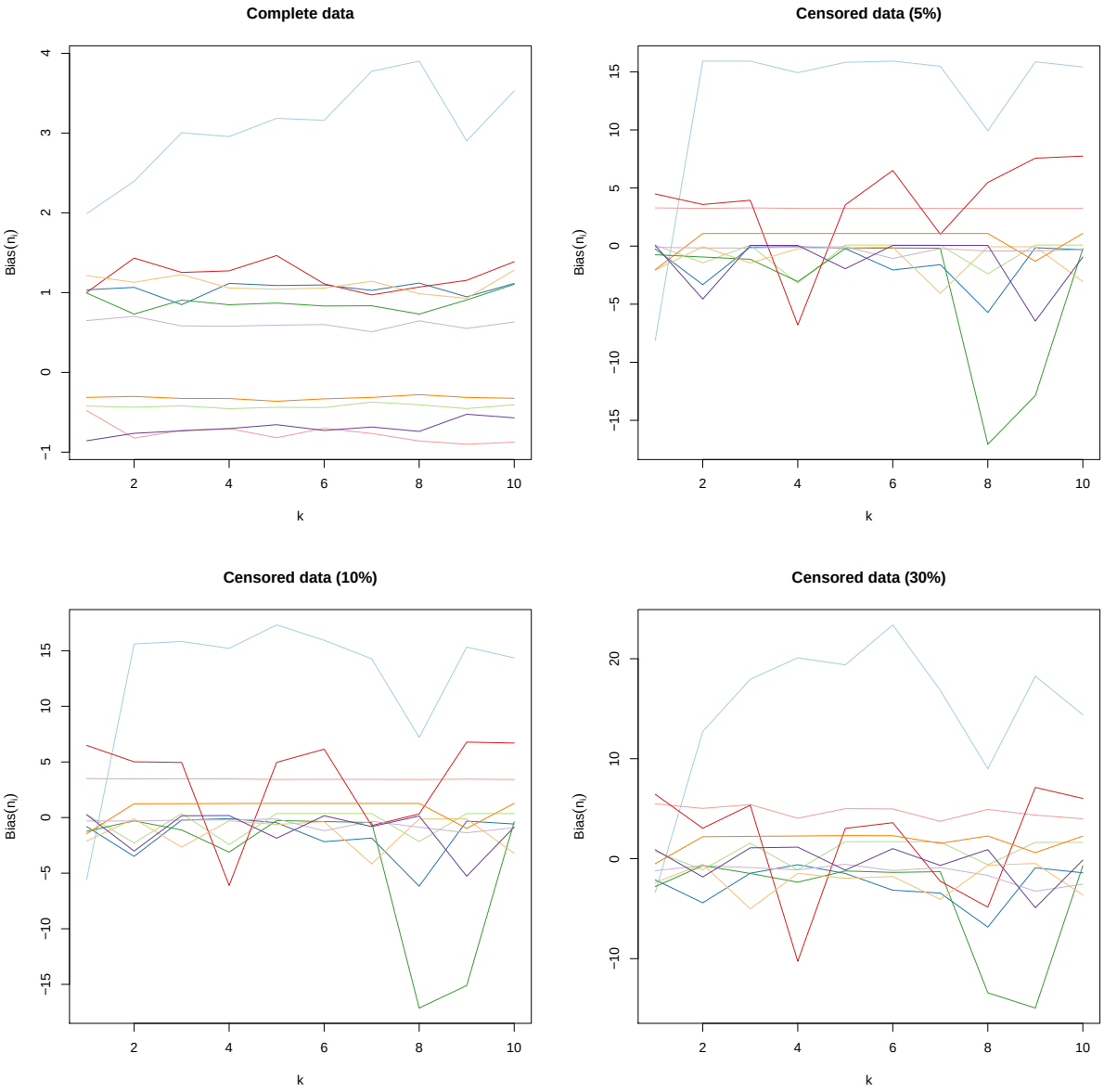


Figure 6.1: Bias of $\widehat{n}_j$, $j = 1, \ldots, 10$.

with other scenarios presented in the literature, but this would go well beyond the scope of the present paper.

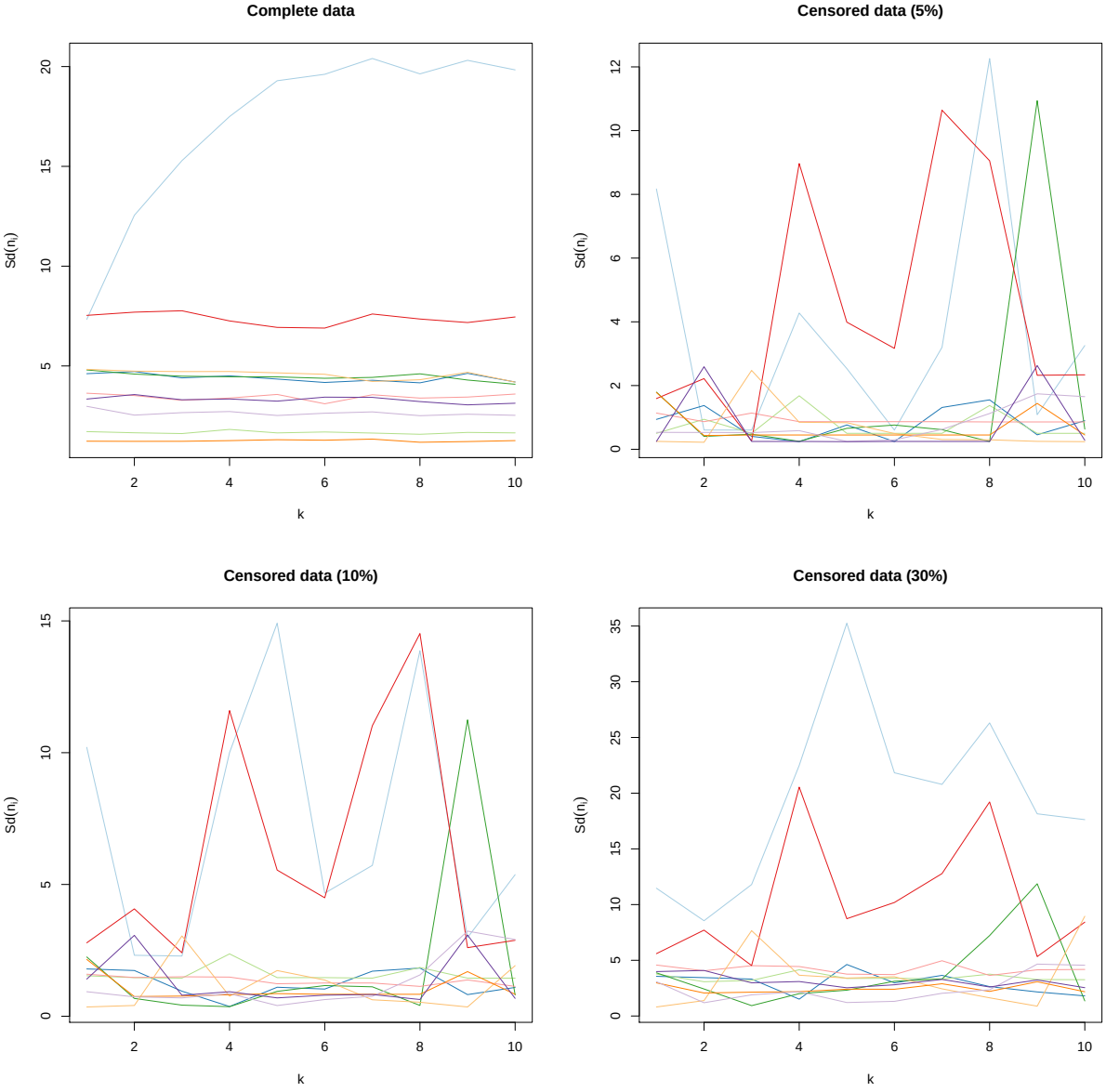


Figure 6.2: Standard deviation of $\widehat{n}_j$, $j = 1, \ldots, 10$.

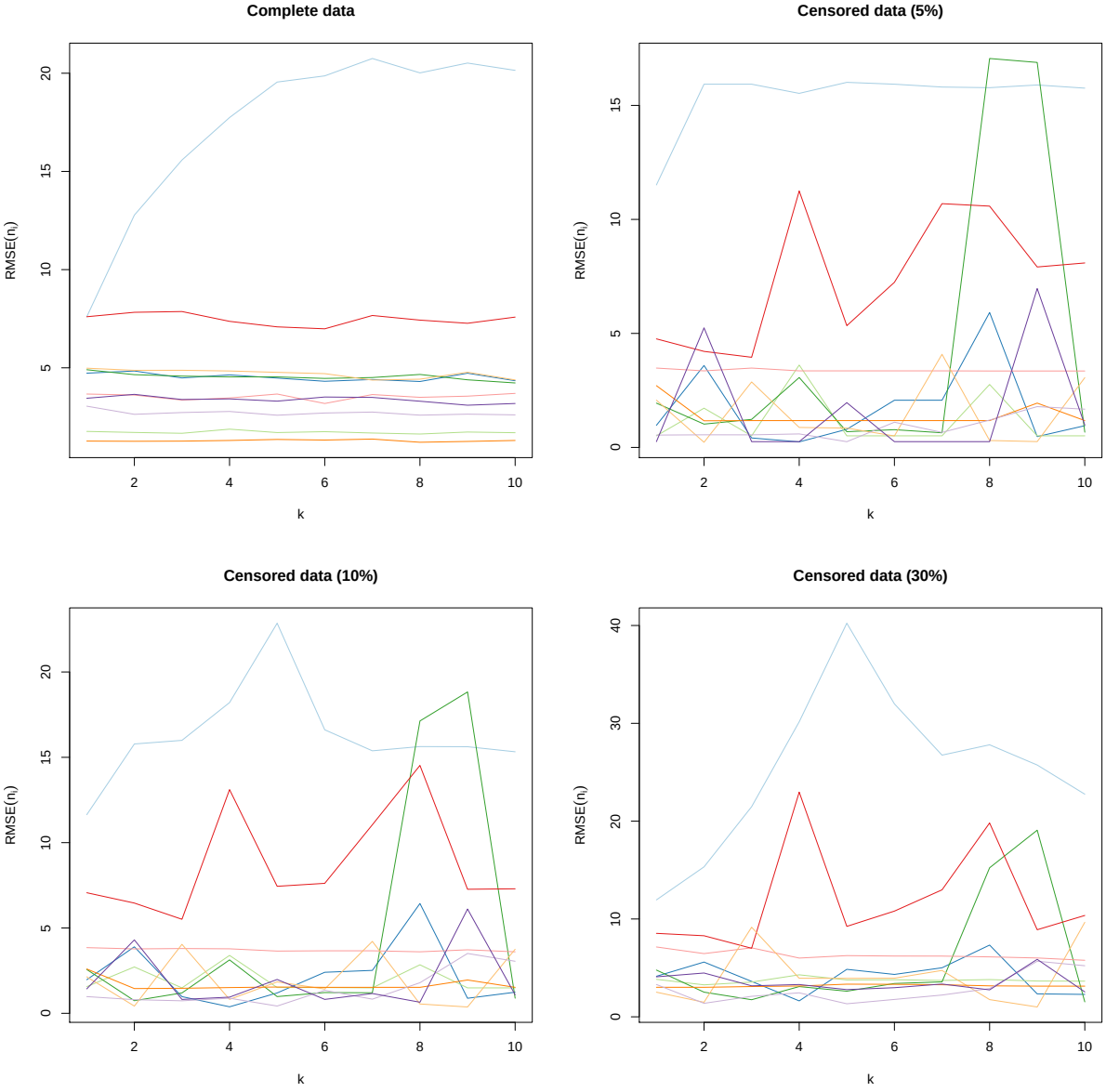


Figure 6.3: Root of MSE of $\widehat{n}_j$, $j = 1, \ldots, 10$.

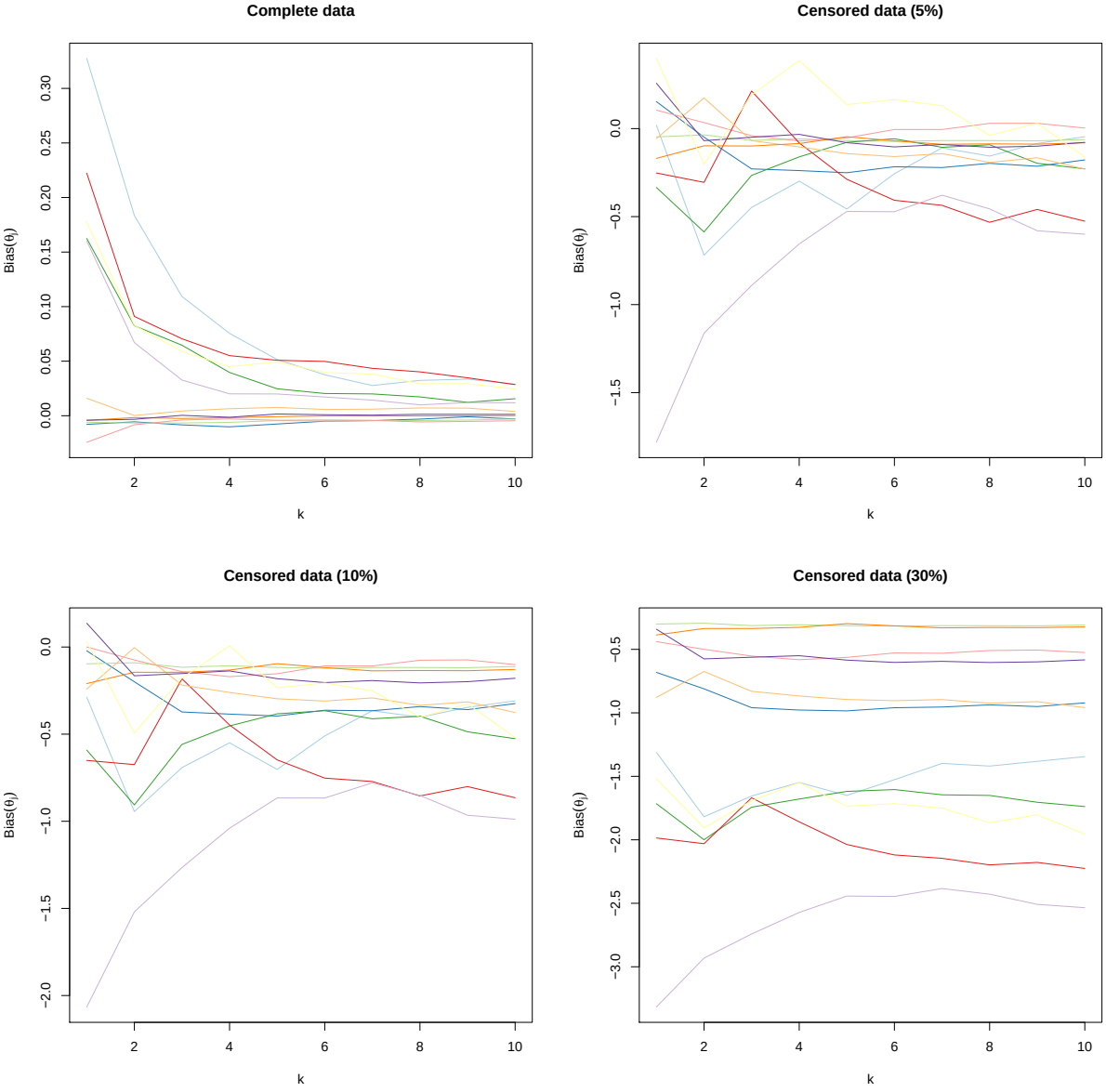


Figure 6.4: Bias of $\widehat{\theta}_j$, $j = 1,\ldots,11$.

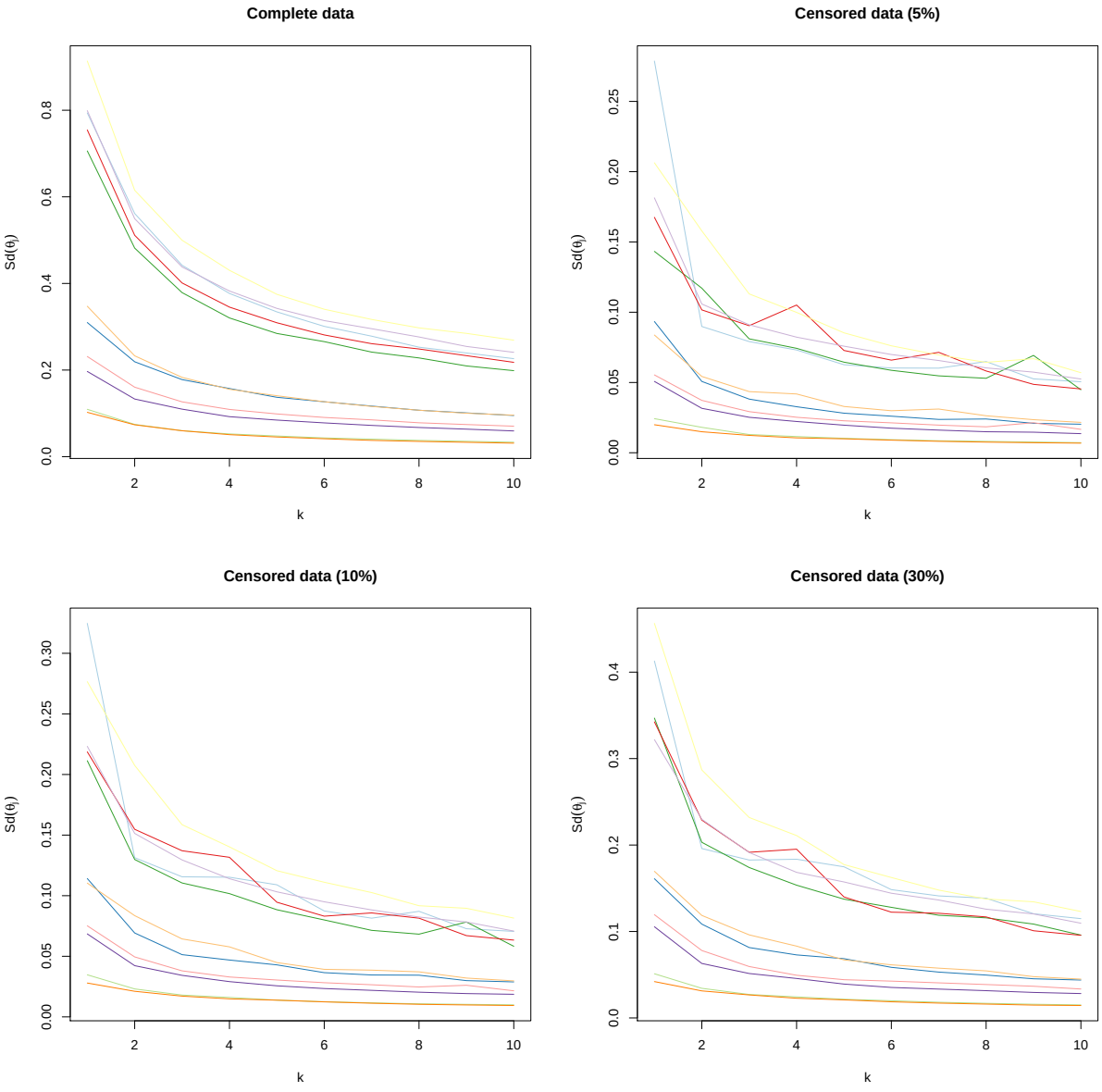


Figure 6.5: Standard deviation of $\widehat{\theta}_j$, $j = 1,\ldots,11$.

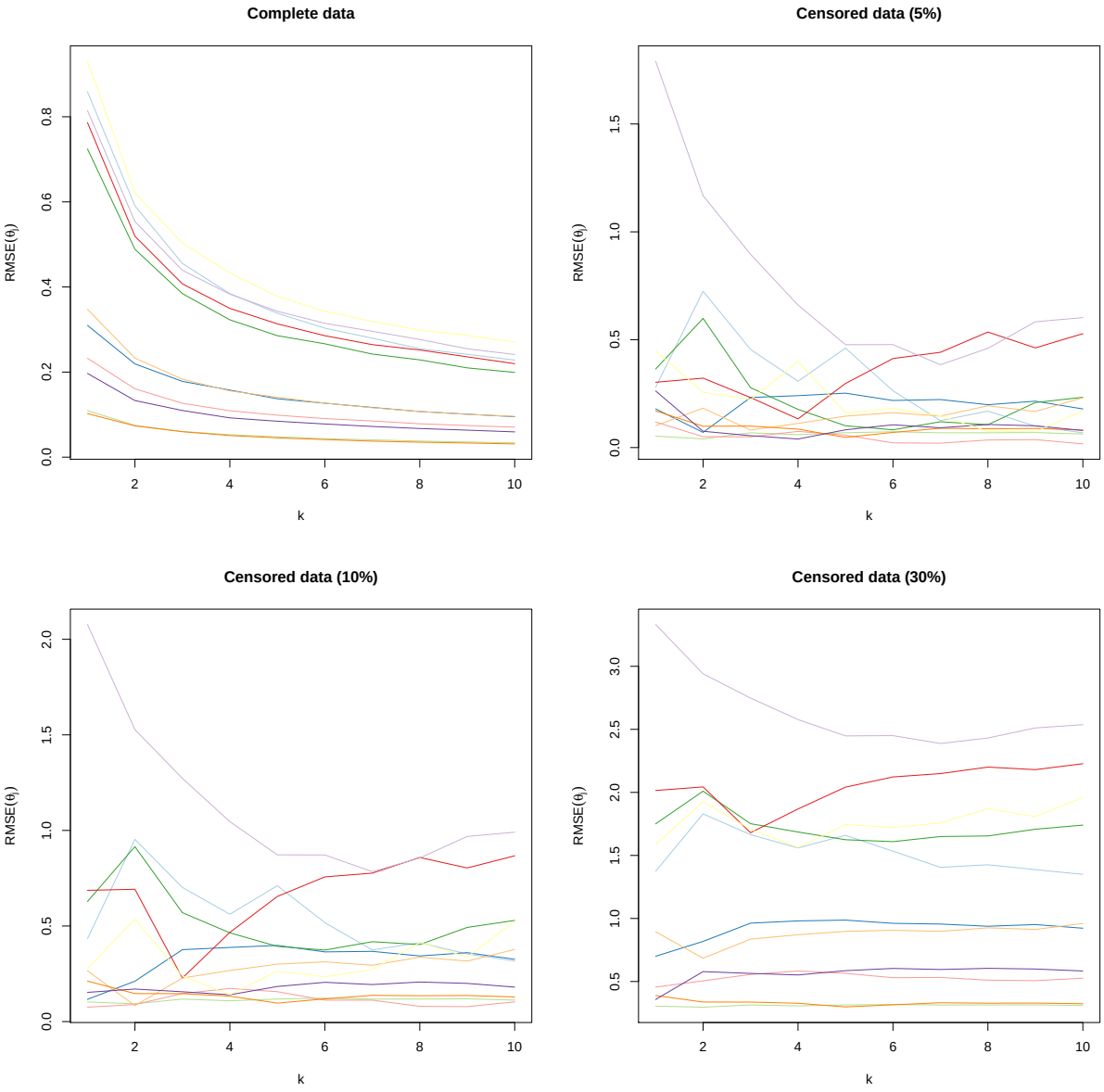


Figure 6.6: Root of MSE of $\widehat{\theta}_j$, $j = 1, \ldots, 11$.

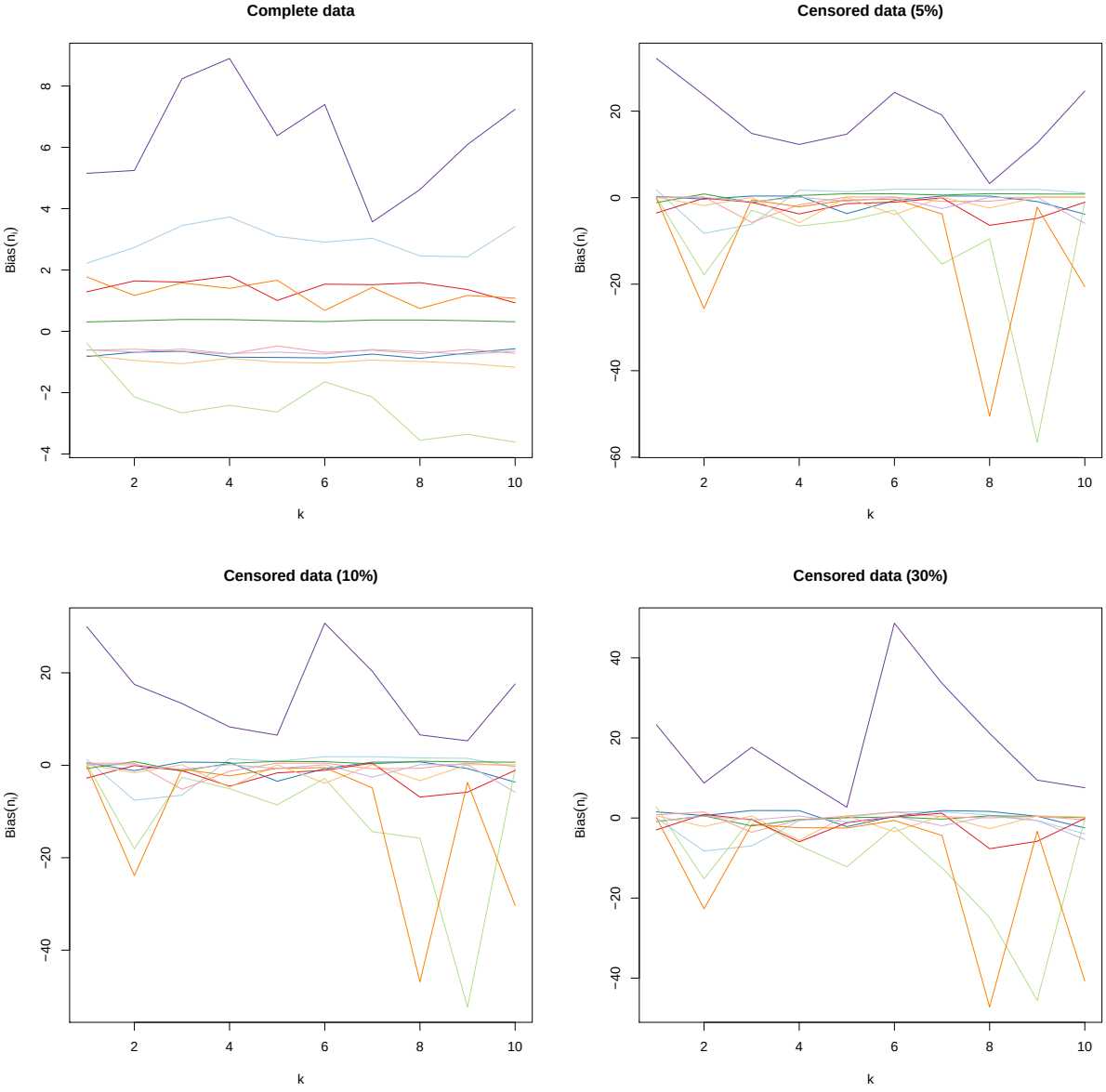


Figure 6.7: Bias of $\widehat{n}_j$, $j = 1, \ldots, 10$.

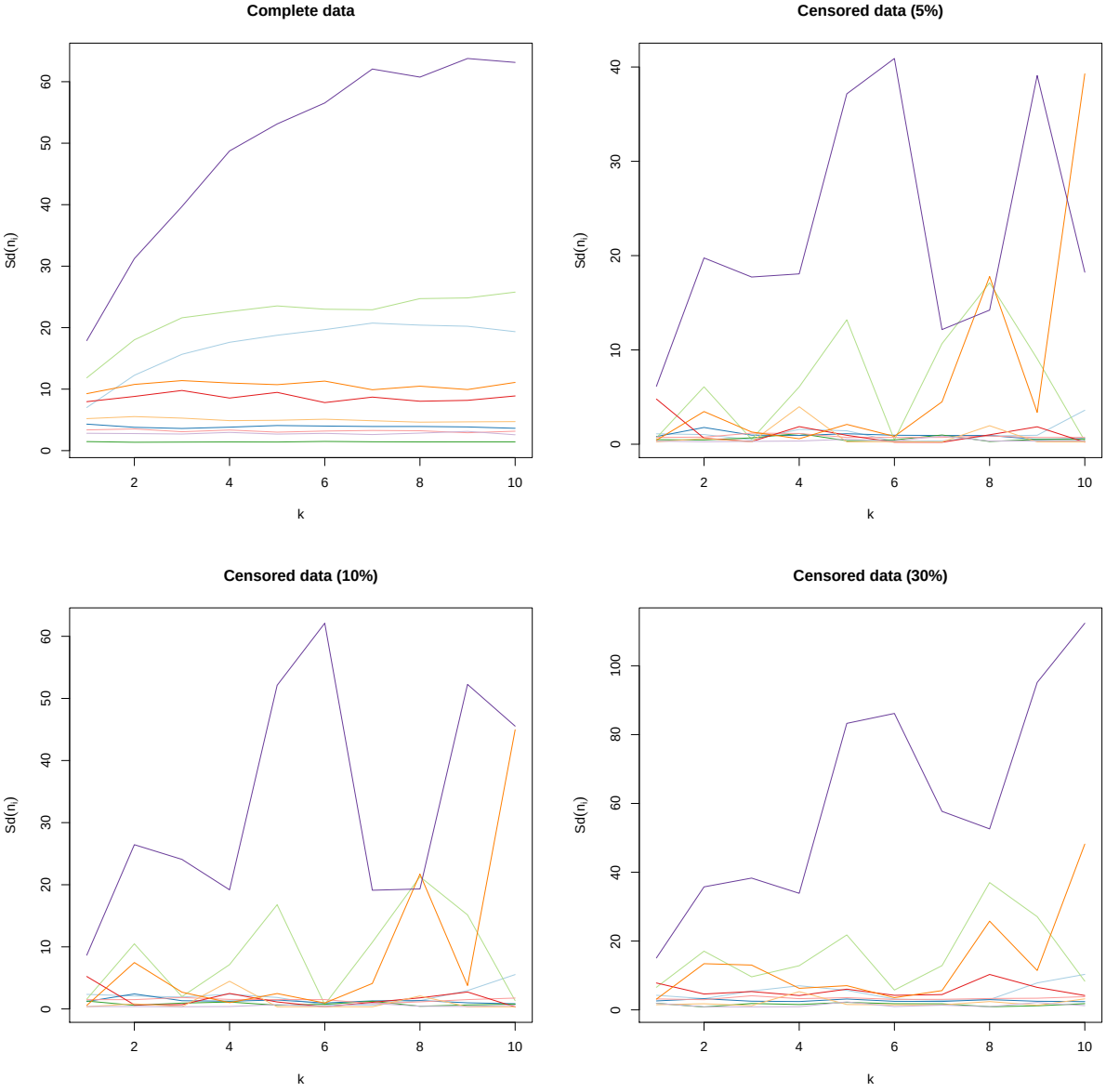


Figure 6.8: Standard deviation of $\widehat{n}_j$, $j = 1, \ldots, 10$.

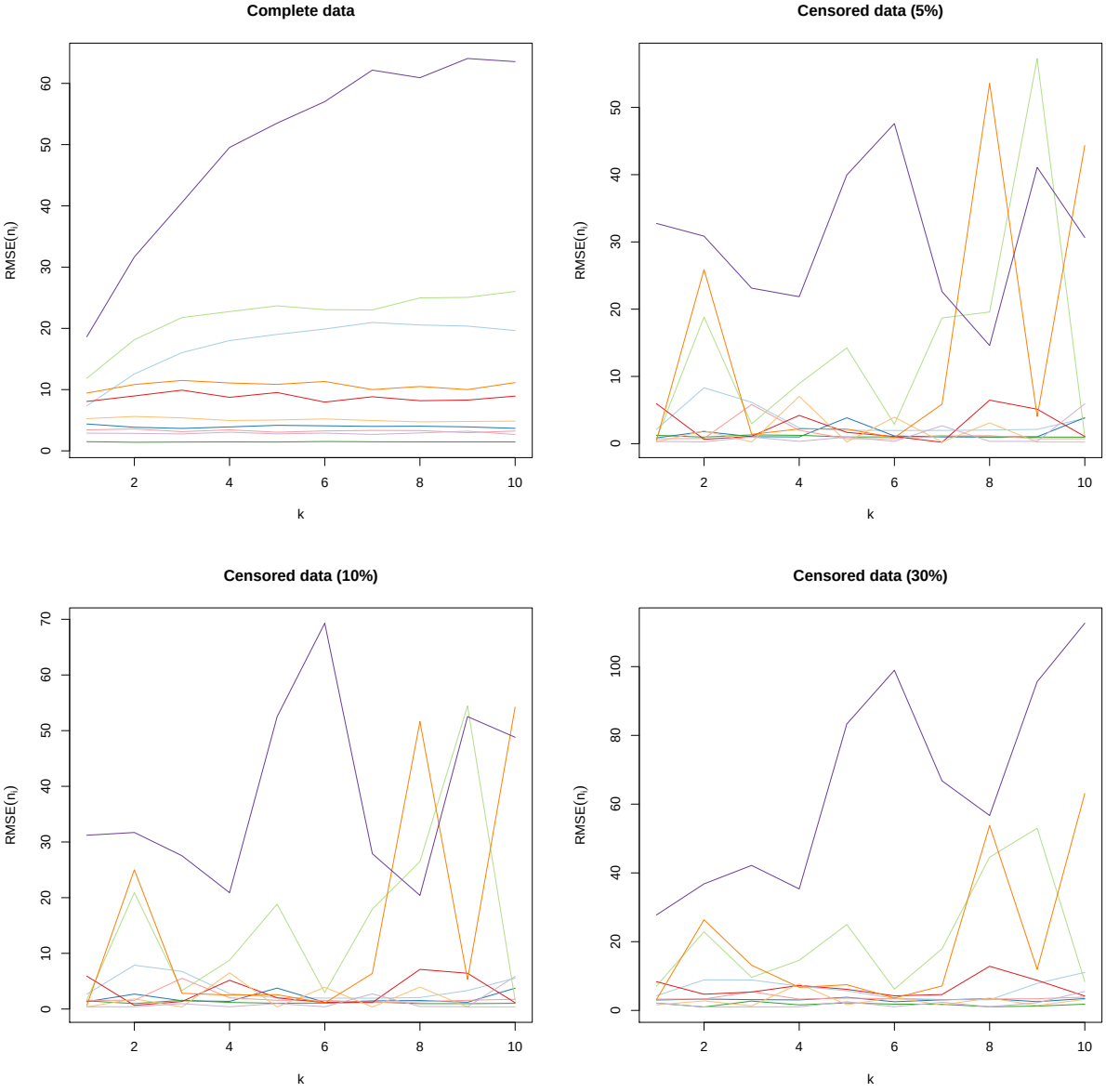


Figure 6.9: Root of MSE of $\widehat{n}_j$, $j = 1, \ldots, 10$.

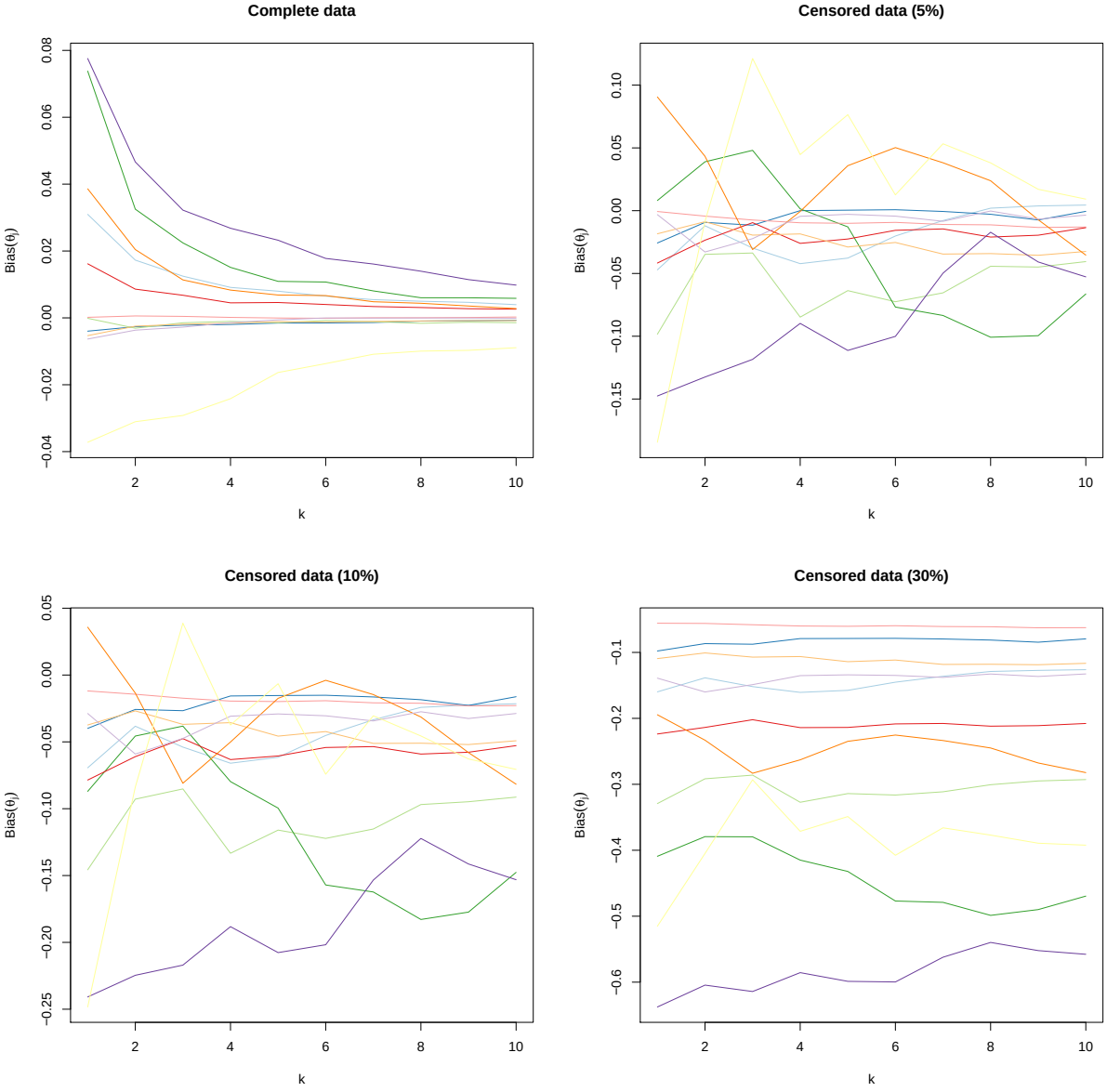


Figure 6.10: Bias of $\widehat{\theta}_j$, $j = 1, \ldots, 11$.

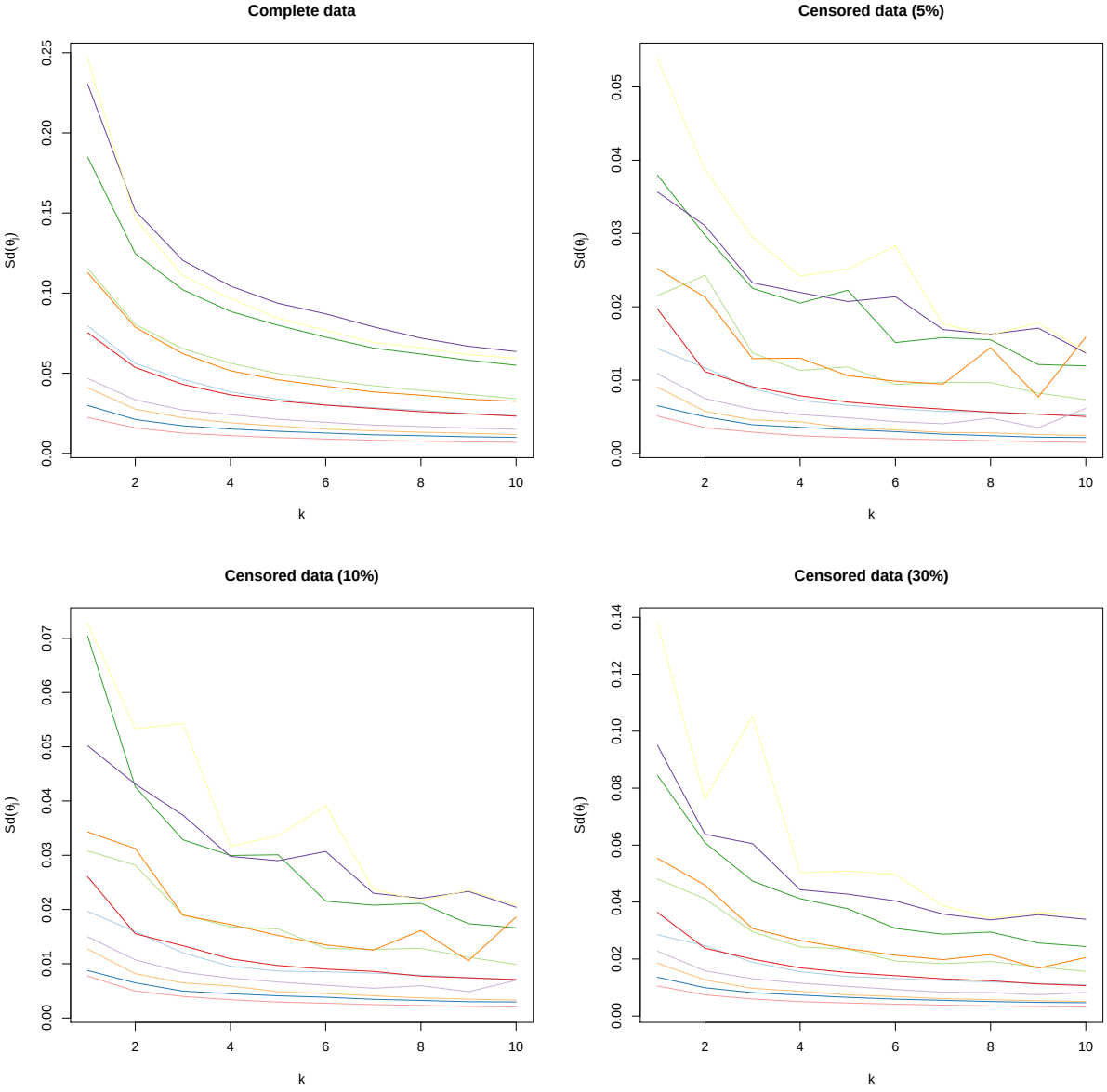


Figure 6.11: Standard deviation of $\widehat{\theta}_j$, $j = 1, \ldots, 11$.

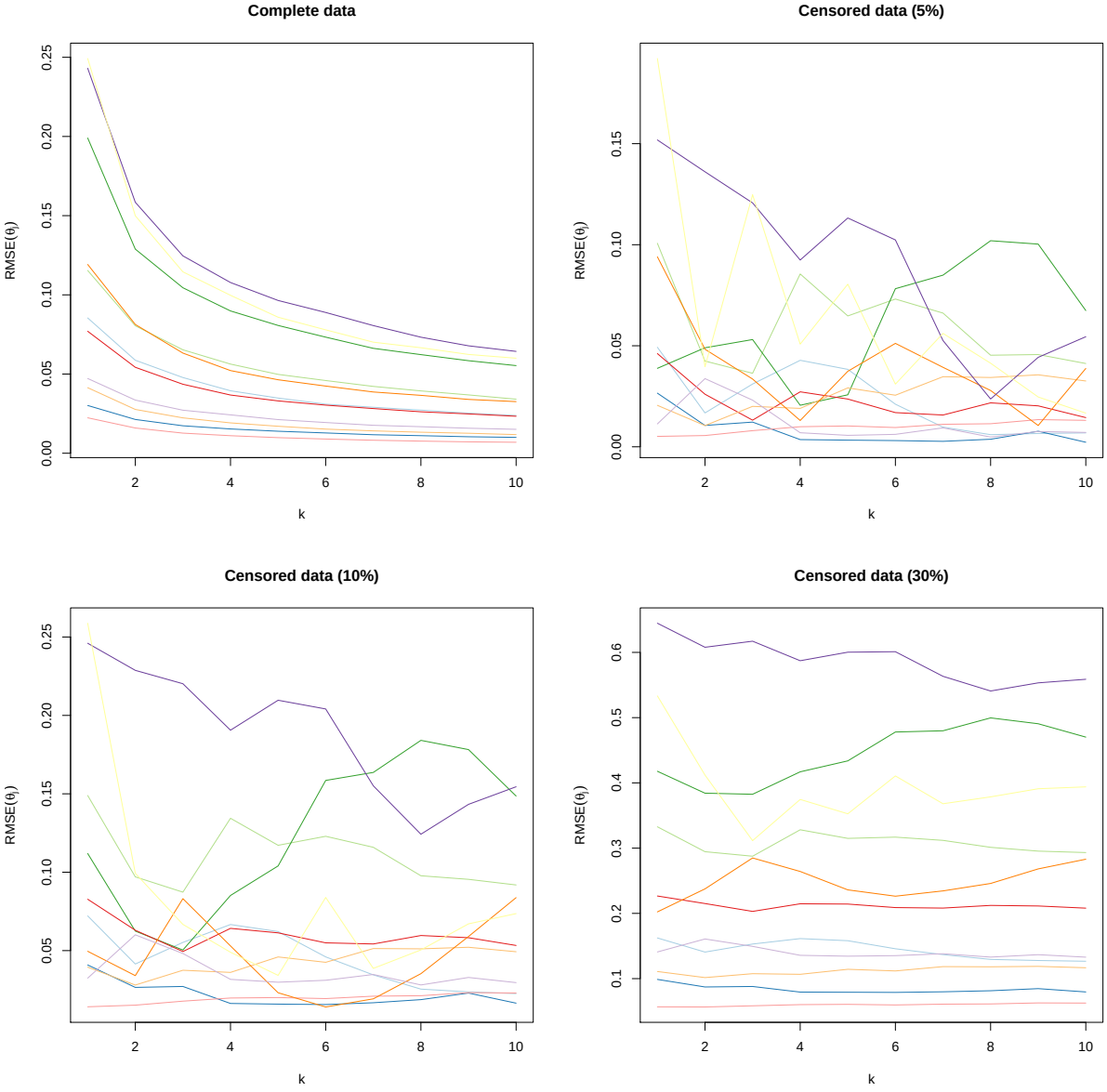


Figure 6.12: Root of MSE of $\widehat{\theta}_j$, $j = 1, \ldots, 11$.

6.7 Mathematical developments

This section is devoted to the proofs of our results. The previously defined notation continues to be used below.
The proof of Theorem 6.3.0.1 will based on the Lemma 6.7.0.1 and Lemma 6.7.0.2. The following lemma gives a bound for the term $\mathbf{W}_1$ given in equation (6.2.5).

Lemma 6.7.0.1 *Under the Assumption 3.1 and the first part of Assumption 3.2, there exist two positive constants* $C_1 > 0$ *and* $C_2 > 0$ *such that, for any* λ *and* ϕ*, we have*

$$\mathbf{W}_1 \leq -\max\{C_1\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^0\|_\infty, C_2\varrho(\phi,\boldsymbol{\phi}^0)\},$$

where

$$\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^0\|_\infty = \max_j|\lambda_j-\lambda_j^0| \ \text{ and } \ \varrho(\boldsymbol{\phi},\boldsymbol{\phi}^0) = \max_j|b(\alpha,\theta_j,\alpha^0,\theta_j^0)|.$$

Proof of Lemma 6.7.0.1

The proof of this lemme follows the similar arguments used in the proof of Lemma 3.1 in He and Severini [2010]. Recall that

$$\begin{aligned} b(\alpha,\theta_j,\alpha^0,\theta_i^0) &= \mathbb{E}(\mathbf{m}_j(\alpha,\theta_j,X_i)) - \mathbb{E}(\mathbf{m}_i(\alpha^0,\theta_i^0,X_i)) \\ &= \int_{\mathbb{R}}[\mathbf{m}_j(\alpha,\theta_j,x)-\mathbf{m}_i(\alpha^0,\theta_i^0,x)]\,dF_{n_i^0}(x). \end{aligned}$$

Let us define, for $i = 1,2,\ldots,k$,

$$h_i(\beta,\boldsymbol{\phi}^0) = \sup_{1\leq j\leq k}\sup_{\theta_j\in\Theta_j}\sup_{\alpha\in\Upsilon}[\beta b(\alpha,\theta_j,\alpha^0,\theta_{i+1}^0)+(1-\beta)b(\alpha,\theta_j,\alpha^0,\theta_i^0)],$$

where $\beta\in[0,1]$. We have

$$h_i(0,\boldsymbol{\phi}^0) = h_i(1,\boldsymbol{\phi}^0) = 0 \text{ for } i = 1,2,\ldots,k.$$

One can check that $h_i(\beta,\boldsymbol{\phi}^0)$ is a convex function with respect to β for any $i = 1,2,\ldots,k$. Let

$$H_i(\boldsymbol{\phi}^0) = 2h_i(1/2,\boldsymbol{\phi}^0).$$

It follows from the Assumption 6.2.0.1 that $H_i(\boldsymbol{\phi}^0) < 0$. If we let

$$\overline{H}(\boldsymbol{\phi}^0) = \max_{1\leq i\leq k} H_i(\boldsymbol{\phi}^0),$$

then we have $\overline{H}(\boldsymbol{\phi}^0) < 0$. Let

$$\Delta_{\boldsymbol{\lambda}}^0 = \min_{1\leq j\leq k-1}|\lambda_{j+1}^0-\lambda_j^0|.$$

Consider the change-point configuration λ in such a way that

$$\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^0\|_\infty \leq \Delta_{\boldsymbol{\lambda}}^0/4.$$

For any $j = 1, 2, \ldots, k$, there are two cases: a candidate change-point fraction λ_j may be on the left or on the right of the true change-point fraction λ_j^0. For any j with λ_j on the right of λ_j^0, we have that $\lambda_{j-1} \le \lambda_j^0 \le \lambda_j$. Then

$$\mathbf{W}_1 \le \frac{n_{j,j+1}}{n} b(\alpha, \theta_j, \alpha^0, \theta_{j+1}^0) + \frac{n_{jj}}{n} b(\alpha, \theta_j, \alpha^0, \theta_j^0).$$

If we define

$$\beta_{j,j+1} = \frac{n_{j,j+1}}{n_{j,j+1} + n_{jj}},$$

the case $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\|_\infty \le \Delta_{\boldsymbol{\lambda}}^0/4$ gives that $\beta_{j,j+1} \le 1/2$ and

$$\mathbf{W}_1 \le (\lambda_j - \lambda_j^0)\overline{\mathrm{H}}(\boldsymbol{\phi}^0).$$

For any j with λ_j on the left of λ_j^0, we have that $\lambda_j \le \lambda_j^0 \le \lambda_{j+1}$. Similarly, we define

$$\beta_{j,j-1} = \frac{n_{j,j-1}}{n_{j,j-1} + n_{jj}},$$

we get $\beta_{j,j-1} \le 1/2$ and

$$\mathbf{W}_1 \le (\lambda_j^0 - \lambda_j)\overline{\mathrm{H}}(\boldsymbol{\phi}^0).$$

Therefore, if $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\|_\infty \le \Delta_{\boldsymbol{\lambda}}^0/4$, we readily obtain that

$$\mathbf{W}_1 \le \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\|_\infty \overline{\mathrm{H}}(\boldsymbol{\phi}^0).$$

On the other hand, we have

$$\mathbf{W}_1 \le \min_{1 \le j \le k+1} b(\alpha, \theta_j, \alpha^0, \theta_j^0) \frac{n_{jj}}{n} = -\max_{1 \le j \le k+1} |b(\alpha, \theta_j, \alpha^0, \theta_j^0)| \frac{n_{jj}}{n}.$$

For any j, we have $\frac{n_{jj}}{n} \ge \Delta_{\boldsymbol{\lambda}}^0/2$, so we infer that

$$\mathbf{W}_1 \le -\frac{1}{2}\Delta_{\boldsymbol{\lambda}}^0 \varrho(\boldsymbol{\phi}, \boldsymbol{\phi}^0).$$

Now, consider the other case of change-point fraction configuration $\boldsymbol{\lambda}$, where

$$\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\|_\infty > \Delta_{\boldsymbol{\lambda}}^0/4.$$

It is obvious that there exists a pair of integers (i, j) such that $n_{ij} \ge n\Delta_{\boldsymbol{\lambda}}^0/4$, $n_{i,j+1} \ge n\Delta_{\boldsymbol{\lambda}}^0/4$ and $n_{ij} \ge n_{i,j+1}$. Let

$$\beta_{i,j+1} = \frac{n_{i,j+1}}{n_{i,j+1} + n_{ij}}.$$

For any $\boldsymbol{\phi}$, we have

$$\begin{aligned}\mathbf{W}_1 &\le \frac{n_{i,j+1} + n_{ij}}{n} [\beta_{i,j+1} b(\alpha, \theta_i, \alpha^0, \theta_{j+1}^0) + (1 - \beta_{i,j+1}) b(\alpha, \theta_i, \alpha^0, \theta_j^0)] \\ &\le \frac{1}{2}\left(\frac{\Delta_{\boldsymbol{\lambda}}^0}{2}\right)^2 \overline{\mathrm{H}}(\boldsymbol{\phi}^0).\end{aligned}$$

Combining the results from the two cases of $\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^0\|_\infty \le \Delta^0_{\boldsymbol{\lambda}}/4$ and $\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^0\|_\infty > \Delta^0_{\boldsymbol{\lambda}}/4$, it follows that

$$\mathbf{W}_1 \le \frac{1}{2}\left(\frac{\Delta^0_{\boldsymbol{\lambda}}}{2}\right)^2 \overline{\mathrm{H}}(\boldsymbol{\phi}^0)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^0\|_\infty,$$

and

$$\mathbf{W}_1 \le -\frac{\Delta^0_{\boldsymbol{\lambda}}}{2}\min\left[\rho(\boldsymbol{\phi},\boldsymbol{\phi}^0), -\frac{\Delta^0_{\boldsymbol{\lambda}}}{4}\overline{\mathrm{H}}(\boldsymbol{\phi}^0)\right]. \tag{6.7.1}$$

Note that (6.7.1) can be simplified. Let us define

$$\rho(\boldsymbol{\phi},\boldsymbol{\phi}^0) = \max_{1\le j\le k+1}\sup_{\theta_j\in\Theta_j}\sup_{\alpha\in\Upsilon}|b(\alpha,\theta_j,\alpha^0,\theta_j^0)|.$$

It follows from the inequality (6.7.1) that we have

$$\mathbf{W}_1 \le -\frac{\Delta^0_{\boldsymbol{\lambda}}}{2}\rho(\boldsymbol{\phi},\boldsymbol{\phi}^0)\min\left[\frac{\rho(\boldsymbol{\phi},\boldsymbol{\phi}^0)}{\rho(\boldsymbol{\phi},\boldsymbol{\phi}^0)}, -\frac{\Delta^0_{\boldsymbol{\lambda}}}{4}\overline{\mathrm{H}}(\boldsymbol{\phi}^0)/\rho(\boldsymbol{\phi},\boldsymbol{\phi}^0)\right].$$

If $-\frac{\Delta^0_{\boldsymbol{\lambda}}}{4}\overline{\mathrm{H}}(\boldsymbol{\phi}^0)/\rho(\boldsymbol{\phi},\boldsymbol{\phi}^0) \le 1$, then we infer that

$$\mathbf{W}_1 \le (\Delta^0_{\boldsymbol{\lambda}}/2)^2(\rho(\boldsymbol{\phi},\boldsymbol{\phi}^0)/\rho(\boldsymbol{\phi},\boldsymbol{\phi}^0))(\overline{\mathrm{H}}(\boldsymbol{\phi}^0)/2).$$

If $-\frac{\Delta^0_{\boldsymbol{\lambda}}}{4}\overline{\mathrm{H}}(\boldsymbol{\phi}^0)/\rho(\boldsymbol{\phi},\boldsymbol{\phi}^0) > 1$, we readily obtain

$$\mathbf{W}_1 \le -(\Delta^0_{\boldsymbol{\lambda}}/2)\rho(\boldsymbol{\phi},\boldsymbol{\phi}^0).$$

Letting

$$\mathrm{C}_2 = \min\{(\Delta^0_{\boldsymbol{\lambda}}/2)^2|\overline{\mathrm{H}}(\boldsymbol{\phi}^0)|/(2\rho(\boldsymbol{\phi},\boldsymbol{\phi}^0)), \Delta^0_{\boldsymbol{\lambda}}/2\},$$

inequality (6.7.1) implies that

$$\mathbf{W}_1 \le -\mathrm{C}_2\rho(\boldsymbol{\phi},\boldsymbol{\phi}^0).$$

Setting

$$\mathrm{C}_1 = (\Delta^0_{\boldsymbol{\lambda}}/2)^2|\overline{\mathrm{H}}(\boldsymbol{\phi}^0)|/2,$$

we finally have the desired result. □

The following lemma describes between-segment properties and within-segment properties of the model.

Lemma 6.7.0.2 *Under the Assumption 6.2.0.6, part 1 and 2 of the Assumption 6.2.0.3 respectively, it follows that*

(I) For any $j = 1,2,\ldots,k+1$, any $0 \le m_1 < m_2 \le n$ and any positive number $\epsilon > 0$, there exists a constant A_j, independent of ϵ, and a constant $r > 2$, such that

$$\mathbb{P}\left(\max_{m_1\le s<t\le m_2,\theta_j\in\Theta_j,\alpha\in\Upsilon}\left|\sum_{i=s+1}^{t}\left(\sum_{z=1}^{k+1}\frac{\mathbf{m}_j(\alpha,\theta_j,\mathrm{Y}_i)\delta_i}{\mathrm{S}_{\mathrm{G}_{n_z^0}}(\mathrm{Y}_i^-)}\mathrm{1I}_{\{n_{z-1}+1\le i\le n_z\}} - \mathbb{E}(\mathbf{m}_j(\alpha,\theta_j,\mathrm{X}i))\right)\right| > \epsilon\right)$$
$$\le \mathrm{A}_j\frac{(m_2-m_1)^r}{\epsilon^2}. \tag{6.7.2}$$

(II) For any $j = 1, 2, \ldots, k+1$ *and any positive number* $\epsilon > 0$*, there exist a constant* B_j*, independent of* ϵ*, and a constant* $r > 2$*, such that*

$$\mathbb{P}\left(\max_{n_{j-1}^0 \le s < t \le n_j^0, \theta_j \in \Theta_j, \alpha \in \Upsilon} \left[\sum_{i=s+1}^{t} \left(\sum_{z=1}^{k+1} \frac{\mathbf{m}_j(\alpha, \theta_j, Y_i)\delta_i}{S_{G_{n_z^0}}(Y_i^-)} \mathrm{II}_{\{n_{z-1}+1 \le i \le n_z\}} - \frac{\mathbf{m}_j(\alpha^0, \theta_j^0, Y_i)\delta_i}{S_{G_{n_j^0}}(Y_i^-)} - b(\alpha, \theta_j, \alpha^0, \theta_j^0) \right) \right] > \epsilon \right) \le B_j \frac{(n_j^0 - n_{j-1}^0)^r}{\epsilon^2}. \tag{6.7.3}$$

Proof of Lemma 6.7.0.2

By the fact that all variables at hand are independent and keeping the part 1 of the Assumption 6.2.0.1 in mind, equation (6.7.2) can be achieved by induction with respect to m_2. The induction method is similar to the one used in Móricz *et al.* [1982], so its proof is omitted here. Using part 2 of the Assumption 6.2.0.1, equation (6.7.3) can be proved similarly by the same induction method. For more details, we can refer to He and Severini [2010]. □

Proof of Theorem 6.3.0.1

Let us introduce the following notation

$$\begin{aligned} \Lambda &= \{(\lambda_1, \lambda_2, \ldots, \lambda_k) : \lambda_j = \frac{n_j}{n}, j = 1, \ldots, k; 0 < n_1 < \cdots < n_k < n\}, \\ \Lambda_\eta &= \{\boldsymbol{\lambda} \in \Lambda : \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\|_\infty > \eta\}, \\ \Phi &= \Theta_1 \times \Theta_2 \times \cdots \times \Theta_{k+1} \times \Upsilon, \\ \Phi_\eta &= \{\boldsymbol{\phi} \in \Phi : \varrho(\boldsymbol{\phi}, \boldsymbol{\phi}^0) > \eta\}. \end{aligned}$$

Then, for any $\eta > 0$, it follows from an application of Lemma 6.7.0.1 that

$$-\max_{\boldsymbol{\lambda} \in \Lambda_\eta, \boldsymbol{\phi} \in \Phi} \mathbf{W}_1 \ge C_1 \eta \quad \text{and} \quad -\max_{\boldsymbol{\lambda} \in \Lambda, \boldsymbol{\phi} \in \Phi_\eta} \mathbf{W}_1 \ge C_2 \eta.$$

Therefore, we readily obtain that

$$
\begin{aligned}
&\mathbb{P}(\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^0\|_\infty>\eta)\\
&\le \mathbb{P}\left(\max_{\boldsymbol{\lambda}\in\Lambda_\eta,\boldsymbol{\phi}\in\Phi}\mathbf{W}>0\right)\le \mathbb{P}\left(\max_{\boldsymbol{\lambda}\in\Lambda_\eta,\boldsymbol{\phi}\in\Phi}\mathbf{W}_2>-\max_{\boldsymbol{\lambda}\in\Lambda_\eta,\boldsymbol{\phi}\in\Phi}\mathbf{W}_1\right)\le \mathbb{P}\left(\max_{\boldsymbol{\lambda}\in\Lambda_\eta,\boldsymbol{\phi}\in\Phi}|\mathbf{W}_2|>C_1\eta\right)\\
&\le \mathbb{P}\left(\max_{\boldsymbol{\lambda}\in\Lambda_\eta,\boldsymbol{\phi}\in\Phi}\sum_{j=1}^{k+1}\frac{1}{n}\left|\sum_{i=n_{j-1}+1}^{n_j}\left\{\frac{\mathbf{m}_j(\alpha,\theta_j,Y_i)\delta_i}{S_{G_{n_j^0}}(Y_i^-)}-\mathbb{E}(\mathbf{m}_j(\alpha,\theta_j,X_i))\right\}\right|>\frac{C_1\eta}{2}\right)\\
&+\mathbb{P}\left(\sum_{j=1}^{k+1}\frac{1}{n}\left|\sum_{i=n_{j-1}^0+1}^{n_j^0}\left\{\frac{\mathbf{m}_j(\alpha^0,\theta_j^0,Y_i)\delta_i}{S_{G_{n_j^0}}(Y_i^-)}-\mathbb{E}(\mathbf{m}_j(\alpha^0,\theta_j^0,X_i))\right\}\right|>\frac{C_1\eta}{2}\right)\\
&\le\sum_{j=1}^{k+1}\mathbb{P}\left(\max_{0\le n_{j-1}<n_j\le n,\theta_j\in\Theta_j,\alpha\in\Upsilon}\frac{1}{n}\left|\sum_{i=n_{j-1}+1}^{n_j}\left\{\frac{\mathbf{m}_j(\alpha,\theta_j,Y_i)\delta_i}{S_{G_{n_j^0}}(Y_i^-)}-\mathbb{E}(\mathbf{m}_j(\alpha,\theta_j,X_i))\right\}\right|>\frac{C_1\eta}{2(k+1)}\right)\\
&+\sum_{j=1}^{k+1}\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=n_{j-1}^0+1}^{n_j^0}\left\{\frac{\mathbf{m}_j(\alpha^0,\theta_j^0,Y_i)\delta_i}{S_{G_{n_j^0}}(Y_i^-)}-\mathbb{E}(\mathbf{m}_j(\alpha^0,\theta_j^0,X_i))\right\}\right|>\frac{C_1\eta}{2(k+1)}\right).
\end{aligned}
$$

It follows from Lemma 6.7.0.2 that, as $n\longrightarrow+\infty$,

$$
\mathbb{P}(\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^0\|_\infty>\eta)\le 2\left[\frac{2(k+1)}{C_1\eta}\right]^2\left(\sum_{j=1}^{k+1}A_j\right)n^{r-2}\longrightarrow 0.
$$

For the estimator $\widehat{\boldsymbol{\phi}}$, we obtain in a similar way that

$$
\begin{aligned}
&\mathbb{P}(\varrho(\widehat{\boldsymbol{\phi}},\boldsymbol{\phi}^0)>\eta)\le\mathbb{P}\left(\max_{\boldsymbol{\lambda}\in\Lambda,\boldsymbol{\phi}\in\Phi_\eta}\mathbf{W}>0\right)\\
&\le\sum_{j=1}^{k+1}\mathbb{P}\left(\max_{0\le n_{j-1}<n_j\le n,\theta_j\in\Theta_j,\alpha\in\Upsilon}\frac{1}{n}\left|\sum_{i=n_{j-1}+1}^{n_j}\left\{\frac{\mathbf{m}_j(\alpha,\theta_j,Y_i)\delta_i}{S_{G_{n_j^0}}(Y_i^-)}-\mathbb{E}(\mathbf{m}_j(\alpha,\theta_j,X_i))\right\}\right|>\frac{C_2\eta}{2(k+1)}\right)\\
&+\sum_{j=1}^{k+1}\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=n_{j-1}^0+1}^{n_j^0}\left\{\frac{\mathbf{m}_j(\alpha^0,\theta_j^0,Y_i)\delta_i}{S_{G_{n_j^0}}(Y_i^-)}-\mathbb{E}(\mathbf{m}_j(\alpha^0,\theta_j^0,X_i))\right\}\right|>\frac{C_2\eta}{2(k+1)}\right).
\end{aligned}
$$

Once more, an application of Lemma 6.7.0.2 shows, as $n\to+\infty$, that

$$
\mathbb{P}\left(\varrho(\widehat{\boldsymbol{\phi}},\boldsymbol{\phi}^0)>\eta\right)\longrightarrow 0.
$$

Noting the fact that $b(\alpha,\theta_j,\alpha^0,\theta_j^0)=0$ if and only if $\alpha=\alpha^0$ and $\theta_j=\theta_j^0$, for $j=1,\ldots,k+1$, completes the proof of Theorem 6.3.0.1. □

Proof of Theorem 6.3.0.3

Let us first define, for any $\eta>0$,

$$
\Lambda_{\eta,n}=\left\{\boldsymbol{\lambda}\in\Lambda: n\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^0\|_\infty\ge\eta\right\}.
$$

Making use of the consistency of the change point fraction $\widehat{\boldsymbol{\lambda}}$, we need to consider only the observations in $\tilde{n}_{j,j-1}$, $\tilde{n}_{j,j}$ and $\tilde{n}_{j,j+1}$ for all j in equation (6.2.6). Therefore, we have

$$
\begin{aligned}
&\mathbb{P}\left(n\|\widehat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}^0\|_\infty \geq \eta\right)\\
&\leq \sum_{j=1}^{k+1}\mathbb{P}\left(\max_{\boldsymbol{\lambda}\in\Lambda_{\eta,n},\boldsymbol{\phi}\in\Phi}\left\{\frac{1}{n}\sum_{t\in\tilde{n}_{jj}}\left[\frac{\mathbf{m}_j(\alpha,\theta_j,\mathrm{Y}_t)\delta_t}{\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_t^-)}-\mathbb{E}(\mathbf{m}_j(\alpha,\theta_j,\mathrm{X}_t))\right]\right.\right.\\
&\left.\left.-\frac{1}{n}\sum_{t\in\tilde{n}_{jj}}\left[\frac{\mathbf{m}_j(\alpha^0,\theta_j^0,\mathrm{Y}_t)\delta_t}{\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_t^-)}-\mathbb{E}(\mathbf{m}_j(\alpha^0,\theta_j^0,\mathrm{X}_t))\right]+\frac{1}{3(k+1)}\mathbf{W}_1\right\}>0\right)\\
&+\sum_{j=2}^{k+1}\mathbb{P}\left(\max_{\boldsymbol{\lambda}\in\Lambda_{\eta,n},\boldsymbol{\phi}\in\Phi}\left\{\frac{1}{n}\sum_{t\in\tilde{n}_{j,j-1}}\left[\frac{\mathbf{m}_j(\alpha,\theta_j,\mathrm{Y}_t)\delta_t}{\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_t^-)}-\mathbb{E}(\mathbf{m}_j(\alpha,\theta_j,\mathrm{X}_t))\right]\right.\right.\\
&\left.\left.-\frac{1}{n}\sum_{t\in\tilde{n}_{j,j-1}}\left[\frac{\mathbf{m}_{j-1}(\alpha^0,\theta_{j-1}^0,\mathrm{Y}_t)\delta_t}{\mathrm{S}_{\mathrm{G}_{n_{j-1}^0}}(\mathrm{Y}_t^-)}-\mathbb{E}(\mathbf{m}_{j-1}(\alpha^0,\theta_{j-1}^0,\mathrm{X}_t))\right]+\frac{1}{3k}\mathbf{W}_1\right\}>0\right)\\
&+\sum_{j=1}^{k}\mathbb{P}\left(\max_{\boldsymbol{\lambda}\in\Lambda_{\eta,n},\boldsymbol{\phi}\in\Phi}\left\{\frac{1}{n}\sum_{t\in\tilde{n}_{j,j+1}}\left[\frac{\mathbf{m}_j(\alpha,\theta_j,\mathrm{Y}_t)\delta_t}{\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_t^-)}-\mathbb{E}(\mathbf{m}_j(\alpha,\theta_j,\mathrm{X}_t))\right]\right.\right.\\
&\left.\left.-\frac{1}{n}\sum_{t\in\tilde{n}_{j,j+1}}\left[\frac{\mathbf{m}_{j+1}(\alpha^0,\theta_{j+1}^0,\mathrm{Y}_t)\delta_t}{\mathrm{S}_{\mathrm{G}_{n_{j+1}^0}}(\mathrm{Y}_t^-)}-\mathbb{E}(\mathbf{m}_{j+1}(\alpha^0,\theta_{j+1}^0,\mathrm{X}_t))\right]+\frac{1}{3k}\mathbf{W}_1\right\}>0\right)\\
&\equiv\sum_{j=1}^{k+1}\mathrm{I}_{1j}+\sum_{j=2}^{k+1}\mathrm{I}_{2j}+\sum_{j=1}^{k}\mathrm{I}_{3j}.
\end{aligned}
$$

First, consider the probability formulas I_{1j} in the above equation for any $j=1,2,\ldots,k+1$. The consistency of $\widehat{\boldsymbol{\lambda}}$ allows us to restrict our attention to the case $n_{jj}>\frac{1}{2}(n_j^0-n_{j-1}^0)$. For this case, we have that

$$
\mathbf{W}_1\leq\frac{n_j^0-n_{j-1}^0}{2n}b(\alpha,\theta_j,\alpha_0,\theta_j^0).
$$

Therefore, we readily obtain that

$$
\begin{aligned}
\mathrm{I}_{1j}&\leq\mathbb{P}\left(\max_{n_{j-1}^0\leq s<t\leq n_j^0,\theta_j\in\Theta_j,\alpha\in\Upsilon}\left[\sum_{i=s+1}^{t}\left(\sum_{z=j-1}^{j+1}\frac{\mathbf{m}_j(\alpha,\theta_j,\mathrm{Y}_i)\delta_i}{\mathrm{S}_{\mathrm{G}_{n_z^0}}(\mathrm{Y}_i^-)}\mathrm{II}_{\{n_{z-1}+1\leq i\leq n_z\}}\right.\right.\right.\\
&\left.\left.\left.-\frac{\mathbf{m}_j(\alpha^0,\theta_j^0,\mathrm{Y}_i)\delta_i}{\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_i^-)}-b(\alpha,\theta_j,\alpha^0,\theta_j^0)\right)\right]>\frac{a(n_j^0-n_{j-1}^0)}{6(k+1)}\right)\\
&\leq\frac{(n_j^0-n_{j-1}^0)^r}{(n_j^0-n_{j-1}^0)^2}(6(k+1))^2=n^{r-2}(\lambda_j^0-\lambda_{j-1}^0)^{r-2}(6(k+1))^2,
\end{aligned}
$$

where

$$
a=\max_{\theta_j\in\Theta,\alpha\in\Upsilon}|b(\alpha,\theta_j,\alpha^0,\theta_j^0)|.
$$

Equation (6.7.3) can then be applied to show that $\mathrm{I}_{1j} \to 0$ as $n, \eta \to \infty$. Next, we consider the probability formula I_{2j} for any $j = 2, \ldots, k+1$. In this case, we can see that

$$\lambda_{j-1} < \lambda_{j-1}^0.$$

We infer readily

$$\begin{aligned}
\mathrm{I}_{2j} \leq & \mathbb{P}\left(\max_{\boldsymbol{\lambda} \in \Lambda_{\eta,n}, \boldsymbol{\phi} \in \Phi} \left\{ \frac{1}{n} \sum_{t \in \tilde{n}_{j,j-1}} \left[\frac{\mathbf{m}_j(\alpha, \theta_j, \mathrm{Y}_t)\delta_t}{\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_t^-)} - \mathbb{E}(\mathbf{m}_j(\alpha, \theta_j, \mathrm{X}_t)) \right] + \frac{1}{6k}\mathbf{W}_1 \right\} > 0 \right) \\
& + \mathbb{P}\left(\max_{\boldsymbol{\lambda} \in \Lambda_{\eta,n}, \boldsymbol{\phi} \in \Phi} \left\{ -\frac{1}{n} \sum_{t \in \tilde{n}_{j,j-1}} \left[\frac{\mathbf{m}_{j-1}(\alpha, \theta_{j-1}, \mathrm{Y}_t)\delta_t}{\mathrm{S}_{\mathrm{G}_{n_{j-1}^0}}(\mathrm{Y}_t^-)} - \mathbb{E}(\mathbf{m}_{j-1}(\alpha, \theta_{j-1}, \mathrm{X}_t)) \right] + \frac{1}{6k}\mathbf{W}_1 \right\} > 0 \right) \\
\equiv & \mathrm{I}_{2j}^{(1)} + \mathrm{I}_{2j}^{(2)}.
\end{aligned}$$

Notice that $\mathrm{I}_{2j}^{(1)}$ and $\mathrm{I}_{2j}^{(2)}$ can be handled in the same way, so we just show how to handle $\mathrm{I}_{2j}^{(1)}$. Only two cases have to be considered.
If $n_{j-1}^0 - n_{j-1} \leq \eta$, then

$$\begin{aligned}
\mathrm{I}_{2j}^{(1)} \leq & \mathbb{P}\left(\max_{n_{j-1} \leq s < t \leq n_{j-1}^0, \theta_j \in \Theta_j, \alpha \in \Upsilon} \left| \sum_{i=s+1}^{t} \left[\frac{\mathbf{m}_j(\alpha, \theta_j, \mathrm{Y}_t)\delta_t}{\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_t^-)} - \mathbb{E}(\mathbf{m}_j(\alpha, \theta_j, \mathrm{X}_t)) \right] \right| > \frac{\mathrm{C}_1 \eta}{6k} \right) \\
& \leq \frac{(n_{j-1}^0 - n_{j-1})^r}{(\mathrm{C}_1 \eta)^2} (6k)^2 \\
& \leq \eta^{r-2} \left(\frac{6k}{\mathrm{C}_1} \right)^2.
\end{aligned}$$

Equation (6.7.2) of Lemma 6.7.0.2 gives that $\mathrm{I}_{2j}^1 \to 0$, as $n, \eta \to \infty$. If $n_{j-1}^0 - n_{j-1} > \eta$, for the other case, then we have

$$\mathbf{W}_1 \leq -\mathrm{C}_1 \frac{(n_{j-1}^0 - n_{j-1})}{n}.$$

Therefore, we infer that

$$\begin{aligned}
\mathrm{I}_{2j}^{(1)} \leq & \mathbb{P}\left(\max_{n_{j-1} \leq s < t \leq n_{j-1}^0, \theta_j \in \Theta_j, \alpha \in \Upsilon} \left| \sum_{i=s+1}^{t} \left[\frac{\mathbf{m}_j(\alpha, \theta_j, \mathrm{Y}_t)\delta_t}{\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_t^-)} - \mathbb{E}(\mathbf{m}_j(\alpha, \theta_j, \mathrm{X}_t)) \right] \right| > \frac{\mathrm{C}_1 (n_{j-1}^0 - n_{j-1})}{6k} \right) \\
& \leq (n_{j-1}^0 - n_{j-1})^{r-2} \left(\frac{6k}{\mathrm{C}_1} \right)^2,
\end{aligned}$$

which converges to zero as $n, \eta \to \infty$, by equation (6.7.2) of Lemma (6.7.0.2). I_{3j} can be handled in a similar way as I_{2j}. Therefore the proof of Theorem 6.3.0.3 is complete. □
The following lemma establishes that the difference between the Kaplan Meier based on estimated proportion of the sample and the true one is asymptotically negligible.

Lemma 6.7.0.3 *Assume that, for* $i = 1, 2, \ldots, k$,

$$\widehat{\lambda}_i - \lambda_i^0 = o_{\mathbb{P}}(1).$$

We have for each $i = 1, 2, \ldots, k$

$$\left(1 - \widehat{\mathrm{F}}_{\widehat{n}_i}(x)\right) = \left(1 - \mathrm{F}_{n_i^0}(x)\right) + o_{\mathbb{P}}(1).$$

Proof of Lemma 6.7.0.3

For every $\epsilon > 0$ there exist $\eta^{'} > 0$ and $\eta^{''} > 0$ such that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x\leq\tau_{\mathrm{F}_{n_j^0}}}|\widehat{\mathrm{F}}_{\widehat{n}_j}(x)-\widehat{\mathrm{F}}_{n_j^0}(x)|>\epsilon\right)\\
&= \mathbb{P}\left(\sup_{x\leq\tau_{\mathrm{F}_{n_j^0}}}|\widehat{\mathrm{F}}_{\widehat{n}_j}(x)-\widehat{\mathrm{F}}_{n_j^0}(x)|>\epsilon,\widehat{n}_{j-1}=n_{j-1}^0,\widehat{n}_j=n_j^0\right)\\
&\quad+\mathbb{P}\left(\sup_{x\leq\tau_{\mathrm{F}_{n_j^0}}}|\widehat{\mathrm{F}}_{\widehat{n}_j}(x)-\widehat{\mathrm{F}}_{n_j^0}(x)|>\epsilon,\widehat{n}_{j-1}\neq n_{j-1}^0,\widehat{n}_j\neq n_j^0\right)\\
&\quad+\mathbb{P}\left(\sup_{x\leq\tau_{\mathrm{F}_{n_j^0}}}|\widehat{\mathrm{F}}_{\widehat{n}_j}(x)-\widehat{\mathrm{F}}_{n_j^0}(x)|>\epsilon,\widehat{n}_{j-1}\neq n_{j-1}^0,\widehat{n}_j=n_j^0\right)\\
&\quad+\mathbb{P}\left(\sup_{x\leq\tau_{\mathrm{F}_{n_j^0}}}|\widehat{\mathrm{F}}_{\widehat{n}_j}(x)-\widehat{\mathrm{F}}_{n_j^0}(x)|>\epsilon,\widehat{n}_{j-1}=n_{j-1}^0,\widehat{n}_j\neq n_j^0\right)\\
&\leq 2\mathbb{P}(\widehat{n}_{j-1}\neq n_{j-1}^0)+2\mathbb{P}(\widehat{n}_j\neq n_j^0)\\
&\leq 2\mathbb{P}(|\widehat{\lambda}_{j-1}-\lambda_{j-1}^0|>\eta^{'})+2\mathbb{P}(|\widehat{\lambda}_j-\lambda_j^0|>\eta^{''})\xrightarrow[n\to\infty]{\mathbb{P}}0.
\end{aligned}$$

Hence the proof is complete. □

The following lemma gives the approximation of the Kaplan Meier integral based on the estimated proportion of the sample.

Lemma 6.7.0.4 *For any $j = 1,\ldots,k+1$, under the conditions of Theorem 6.3.0.3 and the result of Lemma 6.7.0.3 we have*

$$\int_{\mathbb{R}}\psi_j(\alpha,\theta_j,x)d\widehat{\mathrm{F}}_{\widehat{n}_j}(x)-\int_{\mathbb{R}}\psi_j(\alpha,\theta_j,x)d\widehat{\mathrm{F}}_{n_j^0}(x)=\mathrm{O}_{\mathbb{P}}\left(\frac{1}{n}\right).$$

As a consequence of this lemma, for every $\boldsymbol{\phi}\in\Phi$, we have that

$$\begin{aligned}
\rho_n(\alpha,\theta_1,\ldots,\theta_{k+1}) &= \sum_{j=1}^{k+1}(\widehat{\lambda}_j-\widehat{\lambda}_{j-1})\int_{\mathbb{R}}\psi_j(\alpha,\theta_j,x)d\widehat{\mathrm{F}}_{\widehat{n}_j}(x)\\
&= \sum_{j=1}^{k+1}\left(\lambda_j^0-\lambda_{j-1}^0+\mathrm{O}_{\mathbb{P}}\left(n^{-1}\right)\right)\left(\int_{\mathbb{R}}\psi_j(\alpha,\theta_j,x)d\widehat{\mathrm{F}}_{n_j^0}(x)+\mathrm{O}_{\mathbb{P}}(n^{-1})\right)\\
&= \rho_n^0(\alpha,\theta_1,\ldots,\theta_{k+1})+\mathrm{O}_{\mathbb{P}}\left(n^{-1}\right). \qquad (6.7.4)
\end{aligned}$$

Proof of Lemma 6.7.0.4

We have

$$
\begin{aligned}
&\int_{\mathbb{R}} \psi_j(\alpha,\theta_j,x)\,d\widehat{\mathrm{F}}_{\widehat{n}_j}(x) - \int_{\mathbb{R}} \psi_j(\alpha,\theta_j,x)\,d\widehat{\mathrm{F}}_{n_j^0}(x) \\
&= \frac{\sum_{i=\widehat{n}_{j-1}+1}^{\widehat{n}_j} (n_j^0 - n_{j-1}^0)\frac{\psi_j(\alpha,\theta_j,\mathrm{Y}_i)\Delta_i}{\mathrm{S}_{\mathrm{C}}^{\widehat{n}_j}(\mathrm{Y}_i^-)} - \sum_{i=n_{j-1}^0+1}^{n_j^0} (\widehat{n}_j - \widehat{n}_{j-1})\frac{\psi_j(\alpha,\theta_j,\mathrm{Y}_i)\Delta_i}{\mathrm{S}_{\mathrm{C}}^{n_j^0}(\mathrm{Y}_i^-)}}{(n_j^0 - n_{j-1}^0)(\widehat{n}_j - \widehat{n}_{j-1})} \\
&= \mathbb{1}_{\{\widehat{n}_j \le n_j^0,\, n_{j-1}^0 \le \widehat{n}_{j-1}\}} \left[\sum_{i=n_{j-1}^0+1}^{\widehat{n}_{j-1}} \frac{-\psi_j(\alpha,\theta_j,\mathrm{Y}_i)\Delta_i}{(n_j^0 - n_{j-1}^0)\left(\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_i^-) + o_{\mathbb{P}}(1)\right)} \right. \\
&\quad + \sum_{i=\widehat{n}_{j-1}+1}^{\widehat{n}_j} \frac{\left((n_j^0 - n_{j-1}^0) - (\widehat{n}_j - \widehat{n}_{j-1})\right)\psi_j(\alpha,\theta_j,\mathrm{Y}_i)\Delta_i}{(\widehat{n}_j - \widehat{n}_{j-1})(n_j^0 - n_{j-1}^0)\left(\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_i^-) + o_{\mathbb{P}}(1)\right)} \\
&\quad \left. + \sum_{i=\widehat{n}_j+1}^{n_j^0} \frac{-\psi_j(\alpha,\theta_j,\mathrm{Y}_i)\Delta_i}{(n_j^0 - n_{j-1}^0)\left(\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_i^-) + o_{\mathbb{P}}(1)\right)} \right] \\
&\quad + \mathbb{1}_{\{\widehat{n}_j \le n_j^0,\, \widehat{n}_{j-1} < n_{j-1}^0\}} \left[\sum_{i=\widehat{n}_{j-1}+1}^{n_{j-1}^0} \frac{\psi_j(\alpha,\theta_j,\mathrm{Y}_i)\Delta_i}{(\widehat{n}_j - \widehat{n}_{j-1})\left(\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_i^-) + o_{\mathbb{P}}(1)\right)} \right. \\
&\quad + \sum_{i=n_{j-1}^0+1}^{\widehat{n}_j} \frac{\left((n_j^0 - n_{j-1}^0) - (\widehat{n}_j - \widehat{n}_{j-1})\right)\psi_j(\alpha,\theta_j,\mathrm{Y}_i)\Delta_i}{(\widehat{n}_j - \widehat{n}_{j-1})(n_j^0 - n_{j-1}^0)\left(\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_i^-) + o_{\mathbb{P}}(1)\right)} \\
&\quad \left. + \sum_{i=\widehat{n}_j+1}^{n_j^0} \frac{-\psi_j(\alpha,\theta_j,\mathrm{Y}_i)\Delta_i}{(n_j^0 - n_{j-1}^0)\left(\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_i^-) + o_{\mathbb{P}}(1)\right)} \right] \\
&\quad + \mathbb{1}_{\{n_j^0 < \widehat{n}_j,\, n_{j-1}^0 \le \widehat{n}_{j-1}\}} \left[\sum_{i=n_{j-1}^0+1}^{\widehat{n}_{j-1}} \frac{-\psi_j(\alpha,\theta_j,\mathrm{Y}_i)\Delta_i}{(n_j^0 - n_{j-1}^0)\left(\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_i^-) + o_{\mathbb{P}}(1)\right)} \right. \\
&\quad + \sum_{i=\widehat{n}_{j-1}+1}^{n_j^0} \frac{\left((n_j^0 - n_{j-1}^0) - (\widehat{n}_j - \widehat{n}_{j-1})\right)\psi_j(\alpha,\theta_j,\mathrm{Y}_i)\Delta_i}{(\widehat{n}_j - \widehat{n}_{j-1})(n_j^0 - n_{j-1}^0)\left(\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_i^-) + o_{\mathbb{P}}(1)\right)} \\
&\quad \left. + \sum_{i=n_j^0+1}^{\widehat{n}_j} \frac{\psi_j(\alpha,\theta_j,\mathrm{Y}_i)\Delta_i}{(\widehat{n}_j - \widehat{n}_{j-1})\left(\mathrm{S}_{\mathrm{G}_{n_j^0}}(\mathrm{Y}_i^-) + o_{\mathbb{P}}(1)\right)} \right]
\end{aligned}
$$

$$+1\!\!1_{\{n_j^0<\widehat{n}_j,\widehat{n}_{j-1}<n_{j-1}^0\}}\left[\sum_{i=\widehat{n}_{j-1}+1}^{n_{j-1}^0}\frac{\psi_j(\alpha,\theta_j,Y_i)\Delta_i}{(\widehat{n}_j-\widehat{n}_{j-1})\left(S_{G_{n_j^0}}(Y_i^-)+o_{\mathbb{P}}(1)\right)}\right.$$
$$+\sum_{i=n_{j-1}^0+1}^{n_j^0}\frac{\left((n_j^0-n_{j-1}^0)-(\widehat{n}_j-\widehat{n}_{j-1})\right)\psi_j(\alpha,\theta_j,Y_i)\Delta_i}{(\widehat{n}_j-\widehat{n}_{j-1})(n_j^0-n_{j-1}^0)\left(S_{G_{n_j^0}}(Y_i^-)+o_{\mathbb{P}}(1)\right)}$$
$$\left.+\sum_{i=n_j^0+1}^{\widehat{n}_j}\frac{\psi_j(\alpha,\theta_j,Y_i)\Delta_i}{(\widehat{n}_j-\widehat{n}_{j-1})\left(S_{G_{n_j^0}}(Y_i^-)+o_{\mathbb{P}}(1)\right)}\right].$$

An application of Theorem 6.3.0.3 gives the desired result. □

Proof of Theorem 6.4.0.1

For every $\epsilon>0$ there exists $\eta>0$, such that we have

$$\begin{aligned}
&\mathbb{P}\left(\|\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^0\|>\epsilon\right)\\
&\leq \mathbb{P}\left(\|\rho(\widehat{\boldsymbol{\phi}})-\rho(\boldsymbol{\phi}^0)\|>\eta\right)\\
&\leq \mathbb{P}\left(\|\rho(\widehat{\boldsymbol{\phi}})-\rho_n^0(\widehat{\boldsymbol{\phi}})+\rho_n(\widehat{\boldsymbol{\phi}})-\rho_n^0(\widehat{\boldsymbol{\phi}})+\rho_n(\widehat{\boldsymbol{\phi}})-\rho(\boldsymbol{\phi}^0)\|>\eta\right)\\
&\leq \mathbb{P}\left(\sup_{\boldsymbol{\phi}\in\Phi}\|\rho_n^0(\boldsymbol{\phi})-\rho(\boldsymbol{\phi})\|>\frac{\eta}{3}\right)+\mathbb{P}\left(\|\rho_n(\widehat{\boldsymbol{\phi}})-\rho_n^0(\widehat{\boldsymbol{\phi}})\|>\frac{\eta}{3}\right)\\
&\quad+\mathbb{P}\left(\|\rho_n(\widehat{\boldsymbol{\phi}})-\rho(\boldsymbol{\phi}^0)\|>\frac{\eta}{3}\right),
\end{aligned}$$

the assumptions of Theorem 6.4.0.1 combined with the relation (6.7.4) show that the last term converges in probability to zero as n converges to infinity. □

Proof of Theorem 6.4.0.2

Let us first take $\epsilon>0$ and $\eta>0$ fixed constants. Condition (ii) implies that there exists a finite $\mathfrak{M}$, such that for large value of n, we have

$$\mathbb{P}\left(\sup_{\|\boldsymbol{\phi}-\boldsymbol{\phi}_0\|>\eta}\|\rho_n(\boldsymbol{\phi})\|^{-1}>\mathfrak{M}\right)<\epsilon.$$

Notice that the parameter $\widehat{\boldsymbol{\phi}}$ satisfies

$$\rho_n(\widehat{\boldsymbol{\phi}})=O_{\mathbb{P}}(1),$$

so we readily obtain

$$\mathbb{P}\left(\|\rho_n(\widehat{\boldsymbol{\phi}})\|^{-1}>\mathfrak{M}\right)\longrightarrow 1.$$

It follows that, with probability of at least $1-\epsilon$ for all n large enough,

$$\|\rho_n(\widehat{\boldsymbol{\phi}})\|^{-1} > \mathfrak{M} \geq \sup_{\|\boldsymbol{\phi}-\boldsymbol{\phi}_0\|>\eta} \|\rho_n(\boldsymbol{\phi})\|^{-1}.$$

These inequalities force $\widehat{\boldsymbol{\phi}}$ to lie within a distance η of $\boldsymbol{\phi}^0$, that is,

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^0\| > \eta\right) \leq \epsilon.$$

Since ϵ and η can be chosen arbitrarily close to zero, the asserted convergence in probability is established. □

Proof of Theorem 6.4.0.3

We will follow the proof of Pakes and Pollard [1989]. First we prove $\sqrt{n}$-consistency. The assumed consistency allows us to choose a sequence η_n that converge to zero slowly enough to ensure that

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^0\| > \eta_n\right) \longrightarrow 0.$$

With probability tending to one for this sequence, the supremum in the condition (iii) runs over a range that includes the random value $\widehat{\boldsymbol{\phi}}$. Thus we have

$$\|\rho_n^0(\widehat{\boldsymbol{\phi}}) - \rho(\widehat{\boldsymbol{\phi}}) - \rho_n^0(\boldsymbol{\phi}^0)\| \leq o_{\mathbb{P}}(n^{-1/2}) + o_{\mathbb{P}}(\|\rho_n^0(\widehat{\boldsymbol{\phi}})\|) + o_{\mathbb{P}}(\|\rho(\widehat{\boldsymbol{\phi}})\|).$$

By the triangle inequality, the left-hand side is larger than

$$\|\rho(\widehat{\boldsymbol{\phi}})\| - \|\rho_n^0(\widehat{\boldsymbol{\phi}})\| - \|\rho_n^0(\boldsymbol{\phi}^0)\|.$$

Thus we obtain

$$\|\rho(\widehat{\boldsymbol{\phi}})\|[1-o_{\mathbb{P}}(1)] \leq o_{\mathbb{P}}(n^{-1/2}) + \|\rho_n^0(\widehat{\boldsymbol{\phi}})\|[1+o_{\mathbb{P}}(1)] + \|\rho_n^0(\boldsymbol{\phi}^0)\|.$$

From conditions (i) and the asymptotic normality of $\sqrt{n}\rho_n^0(\boldsymbol{\phi}^0)$ it follows that

$$\|\rho(\widehat{\boldsymbol{\phi}})\| = O_{\mathbb{P}}(n^{-1/2}).$$

The differentiability condition (ii) implies the existence of a positive constant C for which, near $\boldsymbol{\phi}^0$, (recall that $\rho(\boldsymbol{\phi}^0)=0$), we have

$$\|\rho(\boldsymbol{\phi})\| \geq C\|\boldsymbol{\phi}-\boldsymbol{\phi}^0\|.$$

In particular, we infer that

$$\left\|\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^0\right\| = O_{\mathbb{P}}\left(\|\rho(\widehat{\boldsymbol{\phi}})\|\right) = O_{\mathbb{P}}\left(n^{-1/2}\right).$$

Next, we establish asymptotic normality of $\sqrt{n}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^0)$, by arguing that $\rho_n^0(\boldsymbol{\phi})$ is very well approximated by the linear function

$$L_n(\boldsymbol{\phi}) = \Omega\left(\boldsymbol{\phi}-\boldsymbol{\phi}^0\right) + \rho_n^0\left(\boldsymbol{\phi}^0\right)$$

within a $O_{\mathbb{P}}\left(n^{-1/2}\right)$ neighborhood of $\boldsymbol{\phi}^0$. More precisely, we need the approximation error to be of order $o_{\mathbb{P}}\left(n^{-1/2}\right)$ at $\widehat{\boldsymbol{\phi}}$ and at the $\boldsymbol{\phi}_n^*$ that maximizes $\|L_n(\cdot)\|$ globally. This follows directly from (ii) and (iii) together with the $\sqrt{n}$-consistency results already established

$$\begin{aligned}\left\|\rho_n^0\left(\widehat{\boldsymbol{\phi}}\right)-L_n\left(\widehat{\boldsymbol{\phi}}\right)\right\| &\leq \left\|\rho_n^0\left(\widehat{\boldsymbol{\phi}}\right)-\rho\left(\widehat{\boldsymbol{\phi}}\right)-\rho_n^0\left(\boldsymbol{\phi}^0\right)\right\| \\ &\quad+\left\|\rho\left(\boldsymbol{\phi}\right)-\Omega\left(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^0\right)\right\| \\ &\leq o_{\mathbb{P}}\left(n^{-1/2}\right)+o_{\mathbb{P}}\left(\left\|\rho_n^0\left(\widehat{\boldsymbol{\phi}}\right)\right\|\right)+o_{\mathbb{P}}\left(\left\|\rho\left(\widehat{\boldsymbol{\phi}}\right)\right\|\right) \\ &\quad+o_{\mathbb{P}}\left(\left\|\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^0\right\|\right) \\ &= o_{\mathbb{P}}\left(n^{-1/2}\right).\end{aligned}$$

To correspond to a minimum of $\|L_n(\cdot)\|$, the vector $\Omega(\boldsymbol{\phi}_n^*-\boldsymbol{\phi}^0)$ must be equal to the projection of $-\rho_n^0(\boldsymbol{\phi}^0)$ onto the column space of Ω. Hence, we obtain

$$\sqrt{n}(\boldsymbol{\phi}_n^*-\boldsymbol{\phi}^0)=-\sqrt{n}(\Omega^\top\Omega)^{-1}\Omega^\top\rho_n^0(\boldsymbol{\phi}^0).$$

The right-hand side has the asymptotic normal distribution specified in the statement of the theorem. Consequently

$$\boldsymbol{\phi}_n^*=\boldsymbol{\phi}^0+O_{\mathbb{P}}(n^{-1/2}).$$

Because $\boldsymbol{\phi}^0$ is in the interior point of Φ this implies that $\boldsymbol{\phi}_n^*$ lies in Φ with probability tending to one. From the differentiability condition (ii) and condition (iii), we readily obtain that

$$\|\rho_n^0(\boldsymbol{\phi}_n^*)\|=O_{\mathbb{P}}(n^{-1/2}).$$

Then we can argue as for $\widehat{\boldsymbol{\phi}}$ to deduce that

$$\|\rho_n^0(\boldsymbol{\phi}_n^*)-L_n(\boldsymbol{\phi}_n^*)\|=o_{\mathbb{P}}(n^{-1/2}).$$

We now know that ρ_n^0 and L_n are close at both $\widehat{\boldsymbol{\phi}}$, which almost minimizes $\|\rho_n^0\|$, and $\boldsymbol{\phi}_n^*$, which minimizes $\|L_n\|$. This forces $\widehat{\boldsymbol{\phi}}$ to come close to minimizing $\|L_n\|$. That is,

$$\|L_n(\boldsymbol{\phi}_n^*)\|=\|L_n(\widehat{\boldsymbol{\phi}})\|+o_{\mathbb{P}}(n^{-1/2}).$$

So we have

$$\|L_n(\boldsymbol{\phi}_n^*)\|^2=\|L_n(\widehat{\boldsymbol{\phi}})\|^2+o_{\mathbb{P}}(n^{-1}),$$

the across product term being absorbed into $o_{\mathbb{P}}(n^{-1})$ because $\|L_n(\boldsymbol{\phi}_n^*)\|$ is of order $O_{\mathbb{P}}(n^{-1/2})$. The quadratic form of $\|L_n(\boldsymbol{\phi})\|^2$ has the simple expansion

$$\|L_n(\boldsymbol{\phi})\|^2=\|L_n(\boldsymbol{\phi}_n^*)\|^2+\|\Omega(\boldsymbol{\phi}-\boldsymbol{\phi}_n^*)\|^2,$$

about its global minimum. Put $\boldsymbol{\phi}$ equal to $\widehat{\boldsymbol{\phi}}$, then equate the two expressions for $\|L_n(\widehat{\boldsymbol{\phi}})\|^2$ to deduce that

$$\|\Omega(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}_n^*)\|^2=o_{\mathbb{P}}\left(n^{-1/2}\right).$$

Since the matrix Ω has full rank, this is equivalent to

$$\sqrt{n}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0) = \sqrt{n}(\boldsymbol{\phi}_n^* - \boldsymbol{\phi}^0) + o_{\mathbb{P}}(1),$$

from which the asserted central limit theorem follows. □

If we replace conditions (i) by (i)$'$ and (iii) by (iii)$'$ in Theorem 6.4.0.3 we will obtain the same result of Theorem (3.3) in Pakes and Pollard [1989] under each true sub sample, we get $L_n(\boldsymbol{\phi})$ is sum of $k+1$ linear function given by

$$L_{n_j^0}(\boldsymbol{\phi}) = \Gamma_{F_{n_j^0}}(\boldsymbol{\phi}^0)(\boldsymbol{\phi} - \boldsymbol{\phi}^0) + \rho^0_{n_j^0}(\boldsymbol{\phi}^0), j = 1, 2, \ldots, k+1.$$

For notation ease, we put $\boldsymbol{\phi}$ in function for each subsample because there is no influence for other parameters to the ones we are working on.

The following lemma gives the convergence of the Kaplan Meier integrals.

Lemma 6.7.0.5 *Let $s(\boldsymbol{\phi}, x)$ be any real function with, for any $j = 1, 2, \ldots, k+1$,*

$$\int_{\mathbb{R}} |s(\boldsymbol{\phi}^0, x)| dF_{n_j^0}(x) < \infty.$$

Assume that the condition (R1) (in the appendix) with replacement of the functions $H(\cdot)$, $F(\cdot)$ and $G(\cdot)$ by the functions $H_{n_j^0}(\cdot)$, $F_{n_j^0}(\cdot)$ and $G_{n_j^0}(\cdot)$ respectively for each $j = 1, 2, \ldots, k+1$, holds for

$$\varphi(x) = s(\boldsymbol{\phi}^0, x).$$

For any sequence $\widehat{\boldsymbol{\phi}} \xrightarrow{\mathbb{P}} \boldsymbol{\phi}^0$, it follows that, for any $j = 1, 2, \ldots, k+1$,

$$\int_{\mathbb{R}} s(\widehat{\boldsymbol{\phi}}, x) d\widehat{F}_{\widehat{n}_j}(x) \xrightarrow{\mathbb{P}} \int_{\mathbb{R}} s(\boldsymbol{\phi}^0, x) dF_{n_j^0}(x),$$

provided that any one of the following conditions holds, for any $j = 1, 2, \ldots, k+1$,

(i) $s(\boldsymbol{\phi}, x)$ is continuous at $\boldsymbol{\phi}^0$ uniformly in x.

(ii)

$$\int_{\mathbb{R}} \sup_{\{\boldsymbol{\phi}: |\boldsymbol{\phi} - \boldsymbol{\phi}^0| \leq \beta\}} |s(\boldsymbol{\phi}, x) - s(\boldsymbol{\phi}^0, x)| dF_{n_j^0}(x) = h_\beta \to 0 \ \text{ as } \ \beta \to 0.$$

(iii) $s(\cdot, \cdot)$ is continuous in x for $\boldsymbol{\phi}$ in a neighborhood of $\boldsymbol{\phi}^0$, and

$$\lim_{\boldsymbol{\phi} \to \boldsymbol{\phi}^0} \| s(\boldsymbol{\phi}, \cdot) - s(\boldsymbol{\phi}^0, \cdot) \|_V = 0.$$

(iv) $\int_{\mathbb{R}} s(\boldsymbol{\phi}, x) dF_{n_j^0}(x)$ is continuous at $\boldsymbol{\phi} = \boldsymbol{\phi}^0$, and s is continuous in x for $\boldsymbol{\phi}$ in a neighborhood of $\boldsymbol{\phi}^0$, and

$$\lim_{\boldsymbol{\phi} \to \boldsymbol{\phi}^0} \| s(\boldsymbol{\phi}, \cdot) - s(\boldsymbol{\phi}^0, \cdot) \|_V < \infty.$$

(v) $\int_{\mathbb{R}} s(\boldsymbol{\phi}, x) dF_{n_j^0}(x)$ is continuous at $\boldsymbol{\phi} = \boldsymbol{\phi}^0$, and

$$\int_{\mathbb{R}} s(\boldsymbol{\phi}, x) d\widehat{F}_{n_j^0}(x) \xrightarrow{\mathbb{P}} \int_{\mathbb{R}} s(\boldsymbol{\phi}, x) dF_{n_j^0}(x) < \infty,$$

uniformly for $\boldsymbol{\phi}$ in a neighborhood of $\boldsymbol{\phi}^0$.

Proof of Lemma 6.7.0.5

The proof of this lemma is based on the Lemma 6.7.0.4 and Lemma 1 in Wang [1999]. □

Proof of Theorem 6.4.0.4

Note that $\rho_n(\boldsymbol{\phi})$ is differentiable in $\boldsymbol{\phi}$ by the conditions imposed on $\psi_j(\cdot)$. The multivariate mean value theorem thus implies that

$$\rho_n(\widehat{\boldsymbol{\phi}}) = \rho_n(\boldsymbol{\phi}^0) + \left(\sum_{j=1}^{k+1}(\widehat{\lambda}_j - \widehat{\lambda}_{j-1})\Gamma_{\widehat{F}_{\widehat{n}_j}}(\xi_n)\right)(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0),$$

where

$$\|\xi_n - \boldsymbol{\phi}^0\| \leq \|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0\|$$

and recall that $\|\cdot\|$ is the Euclidean norm. By using the fact that

$$\rho_n(\widehat{\boldsymbol{\phi}}) = 0, \rho_{n_j^0}(\alpha^0, \theta_j^0) = 0$$

in combination with Lemma 6.7.0.4, we infer that

$$\sqrt{n}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0) = -\left(\sum_{j=1}^{k+1}(\widehat{\lambda}_j - \widehat{\lambda}_{j-1})\Gamma_{\widehat{F}_{\widehat{n}_j}}(\xi_n)\right)^{-1}\left[\sum_{j=1}^{k+1}(\widehat{\lambda}_j - \widehat{\lambda}_{j-1})\left(\sqrt{n}\left\{\int \psi_j(\alpha^0, \theta_j^0, x)d\widehat{F}_{n_j^0}(x) - \int \psi_j(\alpha^0, \theta_j^0, x)dF_{n_j^0}(x)\right\} + O_{\mathbb{P}}(n^{-1/2})\right)\right].$$

Once more, Lemma 6.7.0.4 implies that we have

$$\begin{aligned}\rho_n(\boldsymbol{\phi}^0) &= \sum_{j=1}^{k+1}(\widehat{\lambda}_j - \widehat{\lambda}_{j-1})\int \psi_j(\alpha^0, \theta_j^0, x)d\widehat{F}_{\widehat{n}_j}(x) \\ &= \sum_{j=1}^{k+1}(\widehat{\lambda}_j - \widehat{\lambda}_{j-1})\left[\int \psi_j(\alpha^0, \theta_j^0, x)d\widehat{F}_{n_j^0}(x) + O_{\mathbb{P}}(n^{-1})\right].\end{aligned}$$

By Theorem 6.3.0.3, we have entries of

$$\sum_{j=1}^{k+1}(\widehat{\lambda}_j - \widehat{\lambda}_{j-1})\Gamma_{\widehat{F}_{\widehat{n}_j}}(\xi_n)$$

converges in probability to the entries of $\Gamma(\boldsymbol{\phi}^0)$. The theorem now follows from combining Proposition 6.8.1.2, Theorem 6.3.0.3 and Slutsky's theorem. □

6.8 Appendix

In the sequel of this section, we use a notation similar to that used in Wang [1999] including some changes absolutely necessary for our setting. We present, for the convenience of the reader, the random censorship model in Section 6.1 without change points. Let $F(\cdot)$ denote the lifetime distribution of X and $G(\cdot)$ the censoring distribution of C. Assume the independence of X and C, which implies that the distribution $H(\cdot)$ of the observation $Y = \min(X, C)$ satisfies

$$1 - H(\cdot) = (1 - F(\cdot))(1 - G(\cdot)).$$

6.8.1 SLLN and CLT for Kaplan Meier integrals

Let us begin by introducing some results on the Kaplan Meier integrals playing a central role in this study. For any specified real function $\varphi(\cdot)$, we state in this section the strong law of large numbers (SLLN) and the central limit theorem (CLT) for the Kaplan Meier integral

$$\int_{\mathbb{R}} \varphi(x) d\widehat{\mathrm{F}}_n(x).$$

Such results constitute the main tools to study the limiting behavior of M-(Z)-estimates in the next sections. For any distribution function $\mathrm{L}(\cdot)$, let

$$\tau_{\mathrm{L}} = \sup\{x : \mathrm{L}(x) < 1\}$$

denote the upper bound of the support of $\mathrm{L}(\cdot)$. Let

$$\triangle \mathrm{L}(x) = \mathrm{L}(x) - \mathrm{L}(x^{-})$$

denote the probability mass of $\mathrm{L}(\cdot)$ at x. Since one can only observe data in the range of $[0, \tau_{\mathrm{H}}]$, it is possible to estimate $\int_{\mathbb{R}} \varphi(x) d\mathrm{F}(x)$ consistency only if $\tau_{\mathrm{F}} = \tau_{\mathrm{H}}$ or if $\varphi(x)$ is zero for $x \geq \tau_{\mathrm{H}}$. The specific requirement for strong consistency is formulated in the following condition :

(R1) at least one of (i) or (ii) below holds:

(i) For some $u < \tau_{\mathrm{H}}$, $\varphi(x) = 0$ for $u < x \leq \tau_{\mathrm{H}}$.

(ii) $\tau_{\mathrm{F}} \leq \tau_{\mathrm{G}}$, where equality may hold except when $\mathrm{G}(\cdot)$ is continuous at τ_{F} and

$$\triangle \mathrm{F}(\tau_{\mathrm{F}}) > 0.$$

Note that (R1) (ii) implies $\tau_{\mathrm{F}} = \tau_{\mathrm{H}}$, and is the necessary and sufficient condition so that $\mathrm{F}(\cdot)$ can be estimated consistently on its entire support. Such a requirement can be dispensed with only the fact that the function $\varphi(\cdot)$ satisfies the requirement (R1) (i) which then results in a truncated Kaplan-Meier integral. Note that only one of the two, but not both, conditions in (i) and (ii) need to hold for (R1). We state in the next proposition the strong consistency of

$$\int_{\mathbb{R}} \varphi(x) d\widehat{\mathrm{F}}_n(x),$$

which follows from the condition (R1), Theorem 1.1 and Corollary 1.2 of Stute and Wang [1993]. Note that the original strong law in Stute and Wang [1993] requires further that $\mathrm{F}(\cdot)$ and $\mathrm{G}(\cdot)$ have no common point of discontinuity. Such a restriction was later discovered to be dispensable, see Stute [1995] for details.

Proposition 6.8.1.1 (Strong law of Large Numbers) *Under the condition (R1) and for any function* $\varphi(\cdot)$ *fulfilling*

$$\int_{\mathbb{R}} |\varphi(x)| d\mathrm{F}(x) < \infty,$$

it follows, with probability one, that

$$\int_{\mathbb{R}} \varphi(x) d\widehat{\mathrm{F}}_n(x) \to \int_{\mathbb{R}} \varphi(x) d\mathrm{F}(x).$$

Moreover, under (R1) (ii), it follows, with probability one, that

$$\sup_{-\infty<x\leq\tau_{\mathrm{H}}} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}(x)| \to 0.$$

Proposition 6.8.1.1 essentially implies that the law of large numbers for censored data hold under the same condition, namely the integrability of $\varphi(\cdot)$, as for the uncensored case. The CLT however requires a little more than the uncensored case. Denote $m(y) = \mathbb{P}(\delta = 1|\mathrm{Y} = y)$ and denote the subdistribution functions for the censored and uncensored observations, respectively, by

$$\begin{aligned} \mathrm{H}_0(y) &= \mathbb{P}(\mathrm{Y} \leq y, \delta = 0) = \int_{-\infty}^{y} (1 - m(t)) d\mathrm{H}(t) = \int_{-\infty}^{y} (1 - \mathrm{F}(t)) d\mathrm{G}(t), \\ \mathrm{H}_1(y) &= \mathbb{P}(\mathrm{Y} \leq y, \delta = 1) = \int_{-\infty}^{y} m(t) d\mathrm{H}(t) = \int_{-\infty}^{y} (1 - \mathrm{G}(t^-)) d\mathrm{F}(t), \end{aligned} \tag{6.8.1}$$

and let the corresponding empirical estimates be denoted by

$$\mathrm{H}_{pn}(y) = \frac{1}{n} \sum_{i=1}^{n} 1\!\mathrm{I}_{\{\mathrm{Y}_i \leq y, \delta_i = p\}}, \quad \text{for} \quad p = 0, 1. \tag{6.8.2}$$

Note that

$$\mathrm{H}_0(\cdot) + \mathrm{H}_1(\cdot) = \mathrm{H}(\cdot).$$

The asymptotic representation of $\int_{\mathbb{R}} \varphi(x) d\widehat{\mathrm{F}}_n(x)$ as a sum of i.i.d. variables defined in (6.8.5) and (6.8.7), is based upon the following expressions

$$\begin{aligned} \gamma_0(x) &= \exp\left\{ \int_{\mathbb{R}} \frac{1\!\mathrm{I}_{\{y<x\}} d\mathrm{H}_0(y)}{1 - \mathrm{H}(y)} \right\}, \\ \gamma_1(x) &= [1 - \mathrm{H}(x)]^{-1} \int_{\mathbb{R}} 1\!\mathrm{I}_{\{y<x\}} \varphi(x) \gamma_0(y) d\mathrm{H}_1(y), \\ \gamma_2(x) &= \int_{\mathbb{R}} \varphi(z) \gamma_0(z) \mathrm{C}(x \wedge z) d\mathrm{H}_1(z), \end{aligned} \tag{6.8.3}$$

where

$$\mathrm{C}(x) = \int_{\mathbb{R}} \frac{1\!\mathrm{I}_{\{y<x\}} d\mathrm{H}_0(y)}{[1 - \mathrm{H}(y)]^2} = \int_{\mathbb{R}} \frac{1\!\mathrm{I}_{\{y<x\}} d\mathrm{G}(y)}{[1 - \mathrm{F}(y)][1 - \mathrm{G}(y)]^2}, \tag{6.8.4}$$

refer to Stute [1995] for more details. Let U denote the random variable defined by

$$\mathrm{U} = \varphi(\mathrm{Y}) \gamma_0(\mathrm{Y}) \delta + \gamma_1(\mathrm{Y})(1 - \delta) - \gamma_2(\mathrm{Y}) - \int_{\mathbb{R}} \varphi(x) d\mathrm{F}(x). \tag{6.8.5}$$

It turns out that $\mathbb{E}(U) = 0$. The variance of U depends on $\varphi(\cdot)$, $F(\cdot)$ and $G(\cdot)$ and is given by

$$\begin{aligned}\sigma^2(\varphi, F, G) &= \mathrm{Var}(U) \\ &= \int_{\mathbb{R}} \varphi^2(y)\gamma_0^2(y)dH_1(y) - \int_{\mathbb{R}} \gamma_1^2(y)dH_0(y) \\ &\quad - \left(\int_{\mathbb{R}} \varphi(x)dF(x)\right)^2 + \int_{\mathbb{R}} \frac{\gamma_1^2(y)[1-m(y)]^2}{1-H(y)} \triangle H(y)dH(y). \end{aligned} \tag{6.8.6}$$

Clearly, the last integral vanishes for a continuous $H(\cdot)$. The additional requirements for the asymptotic normality of $\int_{\mathbb{R}} \varphi(x)d(\widehat{F}_n(x) - F(x))$ are

(R2)
$$\mathbb{E}[\varphi(Y)\gamma_0(Y)\delta]^2 = \int_{\mathbb{R}} \varphi^2(y)\gamma_0^2(y)dH_1(y) < \infty,$$

(R3)
$$\int_{\mathbb{R}} |\varphi(x)|C^{1/2}(x)dF(x) < \infty.$$

For more discussion of these conditions see Wang [1999]. We now present the asymptotic normality results of

$$\int \varphi(x)d(\widehat{F}_n(x) - F(x)),$$

which follow from Theorem 1 of Stute [1995] and (R1).

Proposition 6.8.1.2 (Central limit theorem) *Assume that the conditions (R1)-(R3) are satisfied. Then we have the following representation*

$$\int_{\mathbb{R}} \varphi(x)d(\widehat{F}_n - F)(x) = n^{-1}\sum_{i=1}^{n} U_i + o_{\mathbb{P}}\left(n^{-1/2}\right), \tag{6.8.7}$$

where the U_is are i.i.d. copies of the variable U by replacing the Y and δ in (6.8.5) by Y_i and δ_i, respectively. Thus, for $\sigma^2(\varphi, F, G)$ defined in (6.8.6), we have the following convergence in distribution, as $n \to \infty$,

$$n^{1/2}\int_{\mathbb{R}} \varphi(x)d(\widehat{F}_n - F)(x) \rightsquigarrow N(0, \sigma^2(\varphi, F, G)). \tag{6.8.8}$$

For continuous distribution function $H(\cdot)$, the asymptotic variance in (6.8.8) becomes

$$\sigma^2(\varphi, F, G) = \int_{-\infty}^{\infty} \frac{\left(\int_x^{\infty} \varphi'(t)[1-F(t)]dt\right)^2}{[1-H(x)]^2} dH_1(x). \tag{6.8.9}$$

The last equality in (6.8.9) follows from (6.8.1). A variance estimate can be obtained by replacing $F(\cdot)$, $H_1(\cdot)$ and $H(\cdot)$ respectively by their empirical estimates, for more details we refer the reader to Wang [1999].

6.9 References

Al-Awadhi, F. and Aly, E.-E. A. A. (2005). On the performance of logrank tests in change point problems for randomly censored data. *J. Stat. Theory Appl.*, **4**(3), 292–302. 186

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2012). *Statistical models based on counting processes*. Springer Science & Business Media. 189

Aue, A. and Horváth, L. (2013). Structural breaks in time series. *J. Time Series Anal.*, **34**(1), 1–16. 186

Bae, J. and Kim, S. (2003). The uniform law of large numbers for the Kaplan-Meier integral process. *Bull. Austral. Math. Soc.*, **67**(3), 459–465. 199

Bahadur, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Statist.*, **38**, 303–324. 189

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, pages 47–78. 196

Bhattacharya, P. K. (1987). Maximum likelihood estimation of a change-point in the distribution of independent random variables: general multiparameter case. *J. Multivariate Anal.*, **23**(2), 183–208. 188

Borgan, Ø. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scandinavian Journal of Statistics*, pages 1–16. 189

Bouzebda, S. (2014). Asymptotic properties of pseudo maximum likelihood estimators and test in semi-parametric copula models with multiple change points. *Math. Methods Statist.*, **23**(1), 38–65. 189

Bouzebda, S. and Keziou, A. (2013). A semiparametric maximum likelihood ratio test for the change point in copula models. *Stat. Methodol.*, **14**, 39–61. 189

Brodsky, B. E. and Darkhovsky, B. S. (1993). *Nonparametric methods in change-point problems*, volume 243 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht. 186

Chen, H. (2019). Sequential change-point detection based on nearest neighbors. *Ann. Statist.*, **47**(3), 1381–1407. 186

Chen, J. and Gupta, A. K. (2000). *Parametric statistical change point analysis*. Birkhäuser Boston, Inc., Boston, MA. 186, 188

Chu, L. and Chen, H. (2019). Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data. *Ann. Statist.*, **47**(1), 382–414. 186

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton Mathematical Series, vol. 9. Princeton University Press, Princeton, N. J. 197

Csörgő, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester. With a foreword by David Kendall. 186, 188

Dumbgen, L. *et al.* (1991). The asymptotic behavior of some nonparametric change-point estimators. *The Annals of Statistics*, **19**(3), 1471–1495. 196

El Ktaibi, F. and Ivanoff, B. G. (2019). Bootstrapping the empirical distribution of a stationary process with change-point. *Electron. J. Stat.*, **13**(2), 3572–3612. 186

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, **42**(6), 2243–2281. 201

Fryzlewicz, P. (2018). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *Ann. Statist.*, **46**(6B), 3390–3421. 202

Fu, Y.-X. and Curnow, R. N. (1990a). Locating a changed segment in a sequence of Bernoulli variables. *Biometrika*, **77**(2), 295–304. 188

Fu, Y.-X. and Curnow, R. N. (1990b). Maximum likelihood estimation of multiple change points. *Biometrika*, **77**(3), 563–573. 188

Garreau, D. and Arlot, S. (2018). Consistent change-point detection with kernels. *Electron. J. Stat.*, **12**(2), 4440–4486. 186

Goldenshluger, A., Juditsky, A., Tsybakov, A. B., and Zeevi, A. (2008). Change-point estimation from indirect observations. I. Minimax complexity. *Ann. Inst. Henri Poincaré Probab. Stat.*, **44**(5), 787–818. 196

Gombay, E. and Horváth, L. (1994). An application of the maximum likelihood test to the change-point problem. *Stochastic Process. Appl.*, **50**(1), 161–171. 188

Gombay, E. and Liu, S. (2000). A nonparametric test for change in randomly censored data. *Canad. J. Statist.*, **28**(1), 113–121. 186

Hao, N., Niu, Y. S., and Zhang, H. (2013). Multiple change-point detection via a screening and ranking algorithm. *Statist. Sinica*, **23**(4), 1553–1572. 196

Hawkins, D. M. (2001). Fitting multiple change-point models to data. *Comput. Statist. Data Anal.*, **37**(3), 323–341. 188

He, C. (2017). Bayesian multiple change-point estimation for exponential distribution with truncated and censored data. *Comm. Statist. Theory Methods*, **46**(12), 5827–5839. 186

He, C. B. (2015). Parameter estimation of Weibull distribution with multiple change points for truncated and censored data. *Appl. Math. J. Chinese Univ. Ser. A*, **30**(2), 127–138. 186

He, H. and Severini, T. A. (2010). Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli*, **16**(3), 759–779. 187, 188, 193, 194, 195, 196, 226, 229

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, **57**, 1–17. 188, 195

Hinkley, D. V. (1972). Time-ordered classification. *Biometrika*, **59**, 509–523. 188

Hinkley, D. V. and Hinkley, E. A. (1970). Inference about the change-point in a sequence of binomial variables. *Biometrika*, **57**, 477–488. 188

Hjort, N. L. (1985). Discussion of the paper by andersen, p. k. and borgan, ø. *Scand. J. Statist.*, **12**(2), 97–158. With discussion and a reply by the authors. 189

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, pages 221–233. Univ. California Press, Berkeley, Calif. 189, 197

Huber, P. J. (1981). *Robust statistics*. John Wiley & Sons, Inc., New York. Wiley Series in Probability and Mathematical Statistics. 189, 193, 197

Hušková, M. (1996). Tests and estimators for the change point problem based on M-statistics. *Statist. Decisions*, **14**(2), 115–136. 188

Hušková, M. and Neuhaus, G. (2004). Change point analysis for censored data. *J. Statist. Plann. Inference*, **126**(1), 207–223. 186, 187

Jandhyala, V., Fotopoulos, S., MacNeill, I., and Liu, P. (2013). Inference for single and multiple change-points in time series. *J. Time Series Anal.*, **34**(4), 423–446. 186

Jandhyala, V. K. and Fotopoulos, S. B. (1999). Capturing the distributional behaviour of the maximum likelihood estimator of a changepoint. *Biometrika*, **86**(1), 129–140. 188

Jandhyala, V. K. and Fotopoulos, S. B. (2001). Rate of convergence of the maximum likelihood estimate of a change-point. *Sankhyā Ser. A*, **63**(2), 277–285. 188

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **53**, 457–481. 188

Kiefer, J. and Wolfowitz, J. (1956). Sequential tests of hypotheses about the mean occurrence time of a continuous parameter Poisson process. *Naval Res. Logist. Quart.*, **3**, 205–219 (1957). 189

Kim, J., Cheon, S., and Jin, Z. (2020). Bayesian multiple change-points estimation for hazard with censored survival data from exponential distributions. *J. Korean Statist. Soc.*, **49**(1), 15–31. 186

Korkas, K. K. and Fryzlewicz, P. (2017). Multiple change-point detection for non-stationary time series using wild binary segmentation. *Statist. Sinica*, **27**(1), 287–311. 202

Korostelëv, A. P. and Tsybakov, A. B. (1993). *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York. 196

Lavielle, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stochastic Process. Appl.*, **83**(1), 79–102. 202

Lavielle, M. and Ludeña, C. (2000). The multiple change-points problem for the spectral distribution. *Bernoulli*, **6**(5), 845–869. 191, 202

LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Publ. Statist.*, **1**, 277–329. 189

Lee, T.-S. (2010). Change-point problems: bibliography and review. *J. Stat. Theory Pract.*, **4**(4), 643–662. 186

Lorden, G. (1971). Procedures for reacting to a change in distribution. *Ann. Math. Statist.*, **42**, 1897–1908. 196

Móricz, F. A., Serfling, R. J., and Stout, W. F. (1982). Moment and probability bounds with quasisuperadditive structure for the maximum partial sum. *Ann. Probab.*, **10**(4), 1032–1040. 229

Niu, Y. S., Hao, N., and Zhang, H. (2016). Multiple change-point detection: a selective overview. *Statist. Sci.*, **31**(4), 611–623. 187, 201

Nkurunziza, S. and Fu, K. (2019). Improved inference in generalized mean-reverting processes with multiple change-points. *Electron. J. Stat.*, **13**(1), 1400–1442. 186

Oakes, D. (1986). An approximate likelihood procedure for censored data. *Biometrics*, **42**(1), 177–182. 189, 197

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, **41**, 100–115. 186

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, **42**, 523–527. 186

Page, E. S. (1957). On problems in which a change in a parameter occurs at an unknown point. *Biometrika*, **44**(1/2), 248–252. 186

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, **57**(5), 1027–1057. 189, 199, 200, 236, 238

Pergamenchtchikov, S. and Tartakovsky, A. G. (2019). Asymptotically optimal pointwise and minimax change-point detection for general stochastic models with a composite post-change hypothesis. *J. Multivariate Anal.*, **174**, 104541, 20. 196

Perlman, M. D. (1972). On the strong consistency of approximate maximum likelihood estimators. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. I: Theory of statistics*, pages 263–281. 189

Pfanzagl, J. (1969). Consistent estimation of a location parameter in the presence of an incidental scale parameter. *Ann. Math. Statist.*, **40**, 1353–1357. 189

Pons, O. (2018). *Estimations and tests in change-point models*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ. 186

Qian, G., Wu, Y., and Xu, M. (2019). Multiple change-points detection by empirical Bayesian information criteria and Gibbs sampling induced stochastic search. *Appl. Math. Model.*, **72**, 202–216. 186

Raimondo, M. (1998). Minimax estimation of sharp change points. *Ann. Statist.*, **26**(4), 1379–1397. 196

Reid, N. (1981). Influence functions for censored data. *Ann. Statist.*, **9**(1), 78–92. 189, 197

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons, Inc., New York. Wiley Series in Probability and Mathematical Statistics. 197

Shiryaev, A. N. (2016). On the minimax optimality of CUSUM statistics in change point problems for Brownian motion. *Teor. Veroyatn. Primen.*, **61**(4), 837–844. 196

Stute, W. (1995). The statistical analysis of Kaplan-Meier integrals. In *Analysis of censored data (Pune, 1994/1995)*, volume 27 of *IMS Lecture Notes Monogr. Ser.*, pages 231–254. Inst. Math. Statist., Hayward, CA. 197, 199, 240, 241, 242

Stute, W. (1996). Changepoint problems under random censorship. *Statistics*, **27**(3-4), 255–266. 186

Stute, W. and Wang, J.-L. (1993). The strong law under random censorship. *Ann. Statist.*, **21**(3), 1591–1607. 240

Tan, L. and Zhang, Y. (2019). M-estimators of U-processes with a change-point due to a covariate threshold. *J. Bus. Econom. Statist.*, **37**(2), 248–259. 186

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, **167**, 107299. 186, 202

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. 189, 191, 197, 199

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics. 193, 194, 197

Vostrikova, L. J. (1981). Discovery of "discord" in multidimensional random processes. *Dokl. Akad. Nauk SSSR*, **259**(2), 270–274. 186, 201

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statistics*, **20**, 595–601. 189

Wang, J. and Zheng, M. (2012). Wavelet detection of change points in hazard rate models with censored dependent data. *J. Nonparametr. Stat.*, **24**(3), 765–781. 186

Wang, J.-L. (1995). M-estimators for censored data: strong consistency. *Scand. J. Statist.*, **22**(2), 197–205. 189, 191, 193

Wang, J.-L. (1999). Asymptotic properties of M-estimators based on estimating equations and censored data. *Scand. J. Statist.*, **26**(2), 297–318. 193, 197, 239, 242

Wu, Y. (2005). *Inference for change-point and post-change means after a CUSUM test*, volume 180 of *Lecture Notes in Statistics*. Springer, New York. 186

Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.*, **6**(3), 181–189. 186

Yin, Y. Q. (1988). Detection of the number, locations and magnitudes of jumps. *Comm. Statist. Stochastic Models*, **4**(3), 445–455. 186

Zou, C., Yin, G., Feng, L., and Wang, Z. (2014a). Nonparametric maximum likelihood approach to multiple change-point problems. *Ann. Statist.*, **42**(3), 970–1002. 196

Zou, C., Yin, G., Feng, L., and Wang, Z. (2014b). Nonparametric maximum likelihood approach to multiple change-point problems. *Ann. Statist.*, **42**(3), 970–1002. 201

Zou, C., Yin, G., Feng, L., and Wang, Z. (2014c). Nonparametric maximum likelihood approach to multiple change-point problems. *Ann. Statist.*, **42**(3), 970–1002. 202

Zou, C., Wang, G., and Li, R. (2020). Consistent selection of the number of change-points via sample-splitting. *Ann. Statist.*, **48**(1), 413–439. 202

Chapter 7

Conclusions and perspectives

7.1 Concluding remarks : Chapter 3

In Chapter 3, we have considered the estimation of a parameter $\boldsymbol{\theta}$ that maximizes a certain criterion function depending on an unknown, possibly infinite-dimensional nuisance parameter h. We have followed the common estimation procedure by maximizing the corresponding empirical criterion, in which the nuisance parameter is replaced by some nonparametric estimator. We show that the M-estimators converge weakly to maximizers of Gaussian processes in an abstract setting permitting a great flexibility for applications. We have established that the m out of n bootstrap, in this extended setting, is weakly consistent under conditions similar to those required for weak convergence of the M-estimators in the general framework of Lee [2012], when an additional difficulty comes from the nuisance parameters. The goal of this paper is therefore to extend the existing theory on the bootstrap of the M-estimators, this generalization is far from being trivial and harder to control the nuisance parameter in non-standard framework, which form a basically unsolved open problem in the literature. This requires the effective application of large sample theory techniques, which were developed for the empirical processes. Examples of applications are given to illustrate the generality and the usefulness of our results. It would be interesting to extend the results to dependent framework, this would require further theory which are out of the scope of the present article. An important question is how to extend our findings to the setting of incomplete data (consored data, missing data, etc). This will be a subject of investigation for future work.

7.2 Concluding remarks : Chapter 4

In Chapter 4, we are primarily interested in the exchangeably weighted bootstrap for function-valued estimators defined as a zero point of a function-valued random criterion function. The motivation of considering general bootstrap it to permit a unified treatment for resampling methods and provides a more flexible framework to handle practical problems. We have used a differential identity that applies when the random criterion function is linear in terms of the empirical

measure, that is crucial to establish our main results for the bootstrap. In particular, we do not require linearity of the statistical model in the unknown parameter. The second part of this work is devoted to the semiparametric models extending the results of Zhan [2002] to a more delicate framework. It will be of interest to develop a non-asymptotic Gaussian approximation theory for distributions of Z-estimators together with a Gaussian multiplier bootstrap approximation method. The proof of such a statement, however, should require a different methodology than that used in the present paper, and we leave this problem open for future research.

7.3 Concluding remarks : Chapter 5

In Chapter 5, we investigate the asymptotic properties of semiparametric M-estimators with non-smooth criterion functions of the parameters of a multiple change-point model for a general class of models in which the form of the distribution can change from segment to segment and in which, possibly, there are parameters that are common to all segments. The simulation results show the good performance of the procedure. More precisely, we derive the consistency with rate together with the asymptotic distribution by using the modern theory of the empirical processes. It would be of interest to establish the asymptotic distribution of estimators of the change point coefficient $\boldsymbol{\lambda}$. A future research direction would be to study the problem of estimation in semi-parametric models as such investigated in this work in the setting of serially dependent observations, which requires nontrivial mathematics, that goes well beyond the scope of the present paper.

7.4 Concluding remarks : Chapter 6

In Chapter 6 some important problems in the analysis of multiple change-point models were not considered. One is that the asymptotic distribution of the M-estimator of the vector of change points was not considered, see for example Hinkley [1970] for a treatment of this problem in a single change-point model and Döring [2011] for multiple change points. Thus, this is essentially a separate research topic. However, the asymptotic properties obtained in this paper are necessary for the establishment of the asymptotic distribution of the M-estimator of the vector of change points in this model. This will be a subject of investigation for future work.

Another important problem is to extend the results of this paper to the case in which the number of change points is not known and must be determined from the data. Another direction of research is that the methods and arguments in this paper can be extended to other types of incomplete data (e.g. truncation, double censoring, interval censoring etc.) or data subject to sampling bias, where the Kaplan-Meier product-limit estimate $\widehat{\mathrm{F}}_{n_j}(\cdot)$ will be replaced by an appropriate estimate, usually the non-parametric maximum likelihood estimate of the true lifetime distribution function. Such an extension is straightforward whenever, for the suitable choice of $\widehat{\mathrm{F}}_{n_j}(\cdot)$, the CLT of $\int_{\mathbb{R}} \varphi(x) d\widehat{\mathrm{F}}_{n_j}(x)$ have been established for an arbitrary function $\varphi(\cdot)$. It would be interesting to cleanly extend the results to this, but this would require further theory which

are out of the scope of the present article. Change point estimation is a classical problem in mathematical statistics which, with its broad range of applications in learning problems, has started to gain attention in the machine learning community. An important question is how to apply our findings in such problems. Finally, the optimization problems become computationally complex when the number of parameters is large, it will be interesting to consider the penalized version of the likelihood function to alleviate such difficulties.

7.5 References

Döring, M. (2011). Convergence in distribution of multiple change point estimators. *J. Statist. Plann. Inference*, **141**(7), 2238–2248. 250

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, **57**, 1–17. 250

Lee, S. M. S. (2012). General M-estimation and its bootstrap. *J. Korean Statist. Soc.*, **41**(4), 471–490. 249

Zhan, Y. (2002). Central limit theorems for functional Z-estimators. *Statist. Sinica*, **12**(2), 609–634. 250