Utilisation des Divergences entre Mesures en Statistique Inférentielle

Amor Keziou

THÈSE DE DOCTORAT DE l'UNIVERSITÉ PARIS 6

Spécialité : Mathématiques
Option : Statistique

présentée par

Amor KEZIOU

pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PARIS 6

Sujet de la thèse :

Utilisation des Divergences entre Mesures en Statistique Inférentielle

Soutenue le 17 novembre 2003 devant le jury composé de :

Directeur de thèse	M. Michel BRONIATOWSKI
Rapporteurs	M. Patrice BERTAIL
	M. Ya'acov RITOV
Président	M. Denis BOSQ
Examinateurs	Mme Dominique PICARD
	M. Christian P. ROBERT
Invités	M. Michel DELECROIX
	M. Jean-Michel ZAKOIAN

Remerciements

En tout premier lieu, je tiens à remercier très vivement mon directeur de thèse Monsieur Michel Broniatowski pour son soutien constant, ses orientations, ses conseils et ses encouragements qui m'ont permis de mener à terme ce travail. Qu'il trouve ici l'expression de toute ma reconnaissance.

Je remercie très chaleureusement Messieurs Patrice Bertail et Ya'acov Ritov pour avoir accepté d'être rapporteurs de cette thèse. Je leur suis très reconnaissant du temps qu'ils ont consacré à l'évaluation de ce travail. J'apprécie à sa juste valeur leur présence dans le jury.

J'adresse mes sincères remerciements à Monsieur Denis Bosq d'avoir accepté de présider le jury de cette thèse, à Madame Dominique Picard et à Messieurs Christian Robert, Michel Delecroix et Jean-Michel Zakoian pour l'intérêt qu'ils ont accordé à mon travail et pour avoir accepté de participer au jury.

Je souhaite remercier très vivement Monsieur Ksir Brahim, l'un de mes professeurs en maîtrise, de m'avoir appris les notions de base de la théorie des Probabilités et Statistique et de m'avoir orienté vers le Laboratoire de Statistique Théorique et Appliquée de l'Université Paris 6.

Je tiens à remercier également tous les membres du L.S.T.A. et notamment le directeur Paul Deheuvels de m'avoir admis en DEA, Youri A. Koutoyants pour m'avoir encadré en mémoire de DEA, et les membres du Groupe de travail "Entropie-Divergence" qui, de près ou de loin, ont participé à la réalisation de ce travail. Je voudrais rappeler parmi d'autres Emmanuelle Gautherat, Djamel Louani, Emmanuel Guerre, Samuela Leoni avec qui j'ai travaillé le chapitre 2, Rabah Bennis, Fateh Chebana, Alexandre Depire, Jean-Baptiste Aubin et Omar El-Dakkak.

Je voudrais remercier très chaleureusement Louise Lamart et Pascal Epron qui font preuve chaque jour de leur gentillesse et patience. Je voudrais également adresser un salut amical à mes collègues du Laboratoire : Fateh Chebana, Sophie Dabo-Niang, Myriam Maumy, Samuela Leoni, Alexandre Depire, Jean Renaud Pycke, Jean-Baptiste Aubin, Anne Massiani, Pierre Ribereau, Serge Guillas, Rabah Bennis, Salim Bouzebda, Rosaria Ignaccolo, Catia Scricciolo, Fatiha Rachedi, Fama Fall et tant d'autres avec qui j'ai partagé de bons moments.

Je souhaite enfin remercier toute ma famille pour le soutien qu'elle m'a apporté tout au long de la préparation de cette thèse.

Amor KEZIOU.

LSTA-Université Paris 6. Le 10 novembre 2003.

Adresse personnelle :
Monsieur Amor Keziou
5, Villa Saint-Fargeau
75020 Paris.

Adresse professionnelle :
Laboratoire de Statistique Théorique et Appliquée (LSTA),
Université Paris 6, boîte 158,
175, rue du Chevaleret, 75013 Paris.

E-mail : keziou@ccr.jussieu.fr
keziou@math.univ-paris13.fr

Dédicace :

A ma mère et mon père
modeste témoignage de mon infinie tendresse.

Résumé

Nous proposons de nouvelles méthodes d'estimation et de test par optimisation des Divergences entre mesures pour des modèles paramétriques discrets ou continus, pour des modèles à rapport de densités semi-paramétriques et pour des modèles non paramétriques restreints par des contraintes linéaires.

Les méthodes proposées sont basées sur une nouvelle représentation des Divergences entre mesures. Nous montrons que les méthodes du maximum de vraisemblance paramétrique et du maximum de vraisemblance empirique sont des cas particuliers correspondant au choix de la Divergence de Kullback-Leibler modifiée, et que le choix d'autres types de Divergences mène à des estimateurs ayant des propriétés similaires voire meilleurs dans certains cas. De nombreuses perspectives concernant le problème du choix de la Divergence sont notées.

Abstract

We introduce estimation and test procedures through optimization of Divergence for discrete and continuous parametric models, for semiparametric two samples density ratio models and for nonparametric models restricted by linear constraints.

The proposed procedures are based on a new dual representation for Divergences between measures. We show that the maximum parametric likelihood and the maximum empirical likelihood methods are particular cases corresponding to the choice of the modified Kullback-Leibler Divergence, and that the use of other Divergences leads to some estimates having similar properties even better in some cases. Several problems concerning the choice of the Divergence are noted for future investigations.

Table des matières

Introduction Générale **1**

1 Parametric Estimation and Tests through Divergences **19**
1.1 Introduction and notation 19
1.1.1 Examples of ϕ-divergences. 20
1.1.2 Statistical examples and motivations 21
1.2 Duality and ϕ-divergences 26
1.2.1 Application of the duality Lemma 27
1.2.2 Calculation of the Fenchel-Legendre transform of ϕ-divergences 28
1.3 Parametric estimation and tests through minimum ϕ-divergence approach 31
1.3.1 The asymptotic behavior of the DϕDE's and $\widehat{\phi}_n(\alpha, \theta_0)$ for a given α in Θ 33
1.3.2 The asymptotic behavior of the MDϕDE's 36
1.3.3 Composite tests by minimum $\phi-$divergences 37
1.4 Statistical applications 38
1.4.1 Confidence areas based on ϕ-divergences 38
1.4.2 Multiple maxima of the likelihood surface 42
1.5 Proofs 45
1.5.1 Proof of Proposition 1.1 45
1.5.2 Proof of Lemma 1.2 45
1.5.3 Proof of Proposition 1.2 46
1.5.4 Proof of Proposition 1.4 47
1.5.5 Proof of Theorem 1.2 47
1.5.6 Proof of Proposition 1.5 49
1.5.7 Proof of Theorem 1.3 50
1.5.8 Proof of Theorem 1.4 53

2 Case-Control and Semiparametric Density Ratio Models **57**
2.1 Introduction and motivations 57
2.1.1 Comparison of two populations 58

2.1.2 Logistic model and multiplicative-intercept risk model 59
2.2 Semiparametric Estimation and Tests by Divergences 61
2.2.1 ϕ-divergences and dual representation 61
2.2.2 Estimation through the dual representation of ϕ- divergences . 64
2.3 The asymptotic behaviour of the SDϕDE's 66
2.3.1 Consistency. 66
2.3.2 Asymptotic distributions. 67
2.4 Semiparametric ELE and SDϕDE's 70
2.5 Numerical Results 72
2.5.1 Example 1 73
2.5.2 Example 2 74
2.5.3 Example 3 75
2.5.4 Example 4 77
2.6 Concluding remarks 78
2.7 Proofs 82
2.7.1 Proof of Theorem 2.2 82
2.7.2 Proof of Theorem 2.3 82
2.7.3 Proof of Theorem 2.4 86

3 Estimation and Tests for Models satisfying Linear Constraints with Unknown Parameter 87
3.1 Introduction and notation 87
3.1.1 Statistical examples and motivations 88
3.1.2 Minimum divergence estimates 90
3.2 ϕ-Divergences and Projection 96
3.2.1 Existence of ϕ-Projection on general Sets Ω 97
3.2.2 Existence and Characterization of ϕ-Projection on general Sets Ω 98
3.2.3 Existence and Characterization of ϕ-Projection on Sets defined by Linear Constraints 99
3.3 Dual Representation and Estimation of ϕ - Divergences 101
3.4 Estimation for Models satisfying Linear Constraints 103
3.5 Asymptotic properties and Statistical Tests 109
3.5.1 Asymptotic properties of the estimates for a given $\theta \in \Theta$ . . . 110
3.5.2 Asymptotic properties of the estimates $\widehat{\theta}_\phi$ and $\widehat{\phi}(\mathcal{M}, P_0)$ . . . 112
3.5.3 Tests of model 114
3.5.4 Simple tests on the parameter 115
3.5.5 Composite tests on the parameter 116
3.6 Estimates of the distribution function through projected distributions 117
3.7 Empirical likelihood and related methods 119

3.8 Robustness and Efficiency of MEϕD estimates and Simulation Results 122
3.8.1 Example 1.a . . . 125
3.8.2 Example 1.b . . . 126
3.8.3 Example 1.c . . . 128
3.8.4 Example 2.a . . . 132
3.8.5 Example 2.b . . . 133
3.9 Proofs . . . 135
3.9.1 Proof of Theorem 3.1 . . . 135
3.9.2 Proof of Lemma 3.1 . . . 138
3.9.3 Proof of Theorem 3.2 . . . 139
3.9.4 Proof of Theorem 3.3 . . . 140
3.9.5 Proof of Proposition 3.4 . . . 141
3.9.6 Proof of Proposition 3.6 . . . 142
3.9.7 Proof of Theorem 3.4 . . . 143
3.9.8 Proof of Proposition 3.7 . . . 145
3.9.9 Proof of Theorem 3.5 . . . 146
3.9.10 Proof of Theorem 3.6 . . . 151
3.9.11 Proof of Theorem 3.7 . . . 151

4 Annexe : Sur l'estimation de l'entropie 153
4.1 Introduction . . . 153
4.2 Résultats . . . 155
4.3 Démonstrations . . . 156
4.3.1 Démonstration de la Proposition 4.1 . . . 156
4.3.2 Démonstration de la Proposition 4.2 . . . 160
4.3.3 Démonstration de la Proposition 4.3 . . . 160
4.3.4 Démonstration du Théorème 4.1 . . . 160
4.3.5 Démonstration du Théorème 4.2 . . . 161

Bibliographie 162

Liste des tableaux 162

Table des figures 163

Introduction Générale

L'objet de cette thèse est l'étude des divergences entre mesures et l'application de cette étude en statistique inférentielle dans le but de fournir de nouvelles méthodes qui recouvrent et améliorent des méthodes classiques dans certains cas.

Soit $(\mathcal{X}, \mathcal{B})$ un espace mesurable sur lequel seront définies toutes les mesures. Soit $X_1, \ldots, X_n$ un échantillon de loi de probabilité P_0. Nous proposons une nouvelle méthode d'estimation et de test basée sur les divergences entre mesures pour traiter les problèmes statistiques de base suivants :

(**P1**) Problèmes de test de validation des modèles

$$\mathcal{H}_0 \quad : \quad P_0 \ \text{ appartient à } \ \Omega,$$

où Ω est un modèle : une certaine classe de lois de probabilité.

(**P2**) Problèmes d'estimation et de test, sous le modèle Ω, par des méthodes utilisant l'information P_0 appartient à Ω.

(**P3**) Problèmes d'estimation et de test semi-paramétriques à deux échantillons.

Pour les problèmes (**P1**) et (**P2**), nous présentons cette nouvelle approche dans les deux cas suivants

Cas 1 : Le cas classique des modèles paramétriques :

$$\Omega = \{P_\theta, \ \theta \in \Theta\} =: \mathcal{P}, \tag{1}$$

où Θ est un ouvert de $\mathbb{R}^d$.

Cas 2 : Le cas des modèles semi-paramétriques définis par un nombre fini de contraintes linéaires à paramètres inconnus : l'ensemble Ω est la famille de toutes les lois de probabilité vérifiant

$$\int g(x, \theta) \ dQ(x) = 0,$$

où θ, le paramètre d'intérêt, appartient à Θ, un ouvert de $\mathbb{R}^d$ et $g := (g_1, \ldots, g_l)^T$ est une fonction vectorielle définie sur $\mathcal{X} \times \Theta$ à valeurs dans $\mathbb{R}^l$. Les fonctions réelles $\{g_j, \ j = 1, \ldots, l\}$ sont définies sur $\mathcal{X} \times \Theta$.

Notons M^1 l'ensemble de toutes les lois de probabilité et $\mathcal{M}_\theta$ l'ensemble défini par

$$\mathcal{M}_\theta := \left\{ Q \in M^1 \text{ tel que } \int g(x,\theta)\, dQ(x) = 0 \right\}. \tag{2}$$

D'où

$$\Omega = \bigcup_{\theta\in\Theta} \mathcal{M}_\theta = \bigcup_{\theta\in\Theta} \left\{ Q \in M^1 \text{ tel que } \int g(x,\theta)\, dQ(x) = 0 \right\} =: \mathcal{M}. \tag{3}$$

Ces modèles sont fréquemment utilisés en statistique et en économétrie. De nombreux problèmes statistiques peuvent être étudiés dans le cadre de ces modèles ; le Chapitre 3 en présente plusieurs exemples.

Dans les deux cas, 1 et 2, si $P_0 \in \mathcal{P}$ ou $P_0 \in \mathcal{M}$, on note θ_0 la valeur du paramètre θ pour laquelle $P_0 = P_{\theta_0}$ ou $P_0 \in \mathcal{M}_{\theta_0}$, respectivement.

Les méthodes d'estimation et de test que nous développons dans cette thèse sont basées sur les divergences entre mesures. Avant de présenter ces méthodes, nous devons rappeler, brièvement, la notion des divergences, leurs propriétés et leurs domaines d'application en donnant des motivations à l'utilisation de ces critères en statistique.

Les divergences entre lois de probabilité, appelées f-divergences et parfois ϕ-divergences, ont été introduites par Csiszár (1963). (Nous allons utiliser le mot ϕ-divergences ou simplement divergences).

Soit φ une fonction convexe définie sur $[0,+\infty]$ à valeurs dans $[0,+\infty]$ vérifiant $\varphi(1) = 0$. Pour toutes lois de probabilité Q et P telles que Q est absolument continu par rapport à P, la ϕ-divergence entre Q et P est définie par

$$\phi(Q,P) := \int \varphi\left(\frac{dQ}{dP}\right)\, dP. \tag{4}$$

Si Q n'est pas absolument continue par rapport à P, on pose $\phi(Q,P) = +\infty$.

Cette définition, introduite par Rüschendorf (1984), est la version modifiée de la définition originale de Csiszár (1963). Cette dernière ne nécessite pas la continuité absolue de Q par rapport à P, mais elle utilise une mesure dominante commune σ-finie λ pour les lois Q et P. Comme nous allons considérer des ensembles Ω de lois de probabilité toutes absolument continues par rapport à la loi de l'échantillon P_0, il convient d'utiliser la définition 4. Il est à noter que les deux définitions coïncident sur l'ensemble de lois de probabilité absolument continues par rapport à P et dominées par λ.

Pour toute loi de probabilité P, l'application $Q \mapsto \phi(Q,P)$ est convexe et positive. Si $Q = P$, alors $\phi(Q,P) = 0$. De plus, si la fonction $x \mapsto \varphi(x)$ est strictement convexe sur un voisinage de $x = 1$, on a la propriété fondamentale suivante

$$\phi(Q,P) = 0 \text{ si et seulement si } Q = P. \tag{5}$$

Ces propriétés sont présentées et démontrées dans l'article de Csiszár (1963) et dans le livre de Liese and Vajda (1987). Les propriétés de convexité, de positivité et la propriété (5) sont fondamentales pour les problèmes d'estimation et de test que nous allons considérer.

Les ϕ-divergences, contrairement aux critères $\{L_p,\ p \neq 1\}$, sont invariantes par changement de variable. Cependant, elles ne sont en général pas symétriques ; $\phi(Q,P)$ et $\phi(P,Q)$ diffèrent. Elles coïncident si la fonction convexe φ vérifie $\varphi(x) - x\varphi(1/x) = c(x-1)$ où c est une constante (voir Liese and Vajda (1987) Théorème 1.13). Par conséquent, les ϕ-divergences ne sont en général pas des distances.

Largement utilisée en théorie de l'information, la divergence de Kullback-Leibler, notée KL-divergence, est associée à la fonction convexe réelle $\varphi(x) = x\log x - x + 1$; elle est définie comme suit :

$$KL(Q,P) := \int \log\left(\frac{dQ}{dP}\right)\, dQ.$$

La divergence de Kullback-Leibler modifiée, notée KL_m-divergence, est associée à la fonction convexe $\varphi(x) = -\log x + x - 1$, i.e.,

$$KL_m(Q,P) := \int -\log\left(\frac{dQ}{dP}\right)\, dP.$$

D'autres divergences, largement utilisées en statistique inférentielle, sont les divergences de χ^2 et χ^2 modifiée (χ^2_m) :

$$\chi^2(Q,P) := \frac{1}{2}\int \left(\frac{dQ}{dP} - 1\right)^2\, dP$$

et

$$\chi^2_m(Q,P) := \frac{1}{2}\int \frac{\left(\frac{dQ}{dP} - 1\right)^2}{\frac{dQ}{dP}}\, dP$$

qui sont associées aux fonctions convexes $\varphi(x) = \frac{1}{2}(x-1)^2$ et $\varphi(x) = \frac{1}{2}(x-1)^2/x$, respectivement.

Les distances de Hellinger (H) et L_1 sont également des ϕ-divergences ; elles sont associées au fonctions convexes $\varphi(x) = 2(\sqrt{x}-1)^2$ et $\varphi(x) = |x-1|$, respectivement.

Tous les exemples précédents des ϕ-divergences, à l'exception de la distance L_1, font partie de la classe des "divergences de puissance" introduite par Cressie and Read (1984) (voir aussi Liese and Vajda (1987) Chapitre 2) qui est définie par la classe des fonctions convexes

$$x \in \mathbb{R}_+^* \mapsto \varphi_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma - 1)} \tag{6}$$

pour tout $\gamma \in \mathbb{R}\backslash\{0,1\}$, $\varphi_0(x) = -\log x + x - 1$, $\varphi_1(x) = x\log x - x + 1$ et $\varphi_\gamma(0) = \lim_{x\downarrow 0}\varphi_\gamma(x)$, $\varphi_\gamma(+\infty) = \lim_{x\uparrow+\infty}\varphi_\gamma(x)$, pour tout $\gamma \in \mathbb{R}$.

Les divergences de KL, KL_m, χ^2, χ^2_m et H sont ainsi associées au fonctions convexes φ_1, φ_0, φ_2, φ_{-1} et $\varphi_{1/2}$. Parfois, on note ϕ_γ, la divergence associée à φ_γ, pour tout $\gamma \in \mathbb{R}$.

Définition 0.1. *Soit Ω un ensemble de lois de probabilité et P une loi de probabilité. On définit la ϕ-divergence entre l'ensemble Ω et la loi P par*

$$\phi(\Omega, P) := \inf_{Q\in\Omega} \phi(Q, P).$$

Définition 0.2. *Supposons que $\phi(\Omega, P)$ soit finie*[1]*. On appelle ϕ-projection (ou simplement projection) de P sur Ω, notée Q^*, toute loi appartenant à Ω et vérifiant*

$$\phi(Q^*, P) \leq \phi(Q, P), \quad \textit{pour tout} \quad Q \in \Omega.$$

Les divergences entre lois de probabilité sont utilisées dans le problème de reconstruction de lois et dans le problème de moments ; Csiszár *et al.* (1999) considèrent ces problèmes et en présentent une bibliographie.

En ce qui concerne les divergences entre distributions de processus, elles ont été utilisées dans des problèmes de test statistiques et dans des problèmes en relation avec la contiguïté (c.f. Liese and Vajda (1987) Chapitres 3-7).

Pour des modèles paramétriques $\mathcal{P} := \{P_\theta,\ \theta \in \Theta\}$ à support discret fini (que l'on note S), Liese et Vajda introduisent les estimateurs de minimum des ϕ-divergences (c.f. Liese and Vajda (1987) Chapitre 10).

La ϕ-divergence entre P_θ et P_{θ_0}, dans ce cas, s'écrit

$$\phi(P_\theta, P_{\theta_0}) = \sum_{j\in S} \varphi\left(\frac{P_\theta(j)}{P_{\theta_0}(j)}\right) P_{\theta_0}(j).$$

Elle peut être estimée par

$$\widehat{\phi}\,(P_\theta, P_{\theta_0}) := \phi\,(P_\theta, P_n) = \sum_{j\in S} \varphi\left(\frac{P_\theta(j)}{P_n(j)}\right) P_n(j), \tag{7}$$

[1]Ceci est équivalent à : $\exists Q \in \Omega$ tel que $\phi(Q, P) < \infty$.

où P_n est la mesure empirique associée à l'échantillon $X_1, \ldots, X_n$ de loi P_{θ_0}, i.e.,

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

et δ_x désigne la mesure de Dirac au point x, pour tout x.

C'est ainsi que l'on peut définir l'estimateur du minimum de ϕ-divergence comme suit :

$$\widehat{\theta}_\phi := \arg \inf_{\theta \in \Theta} \phi\left(P_\theta, P_n\right) = \arg \inf_{\theta \in \Theta} \sum_{j \in S} \varphi\left(\frac{P_\theta(j)}{P_n(j)}\right) P_n(j). \tag{8}$$

($\widehat{\theta}_\phi$ est la valeur de θ qui minimise la divergence ϕ entre le modèle paramétrique $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ et la mesure empirique P_n).

La classe d'estimateurs $\widehat{\theta}_\phi$ contient l'estimateur du maximum de vraisemblance. Il est obtenu pour la divergence de KL_m.

Lindsay (1994) et Morales *et al.* (1995) étudient les estimateurs $\widehat{\theta}_\phi$ et montrent que tous ces estimateurs, sous des conditions générales de régularité, sont asymptotiquement normaux et asymtotiquement efficaces.

Par ailleurs, Lindsay (1994) et Jiménez and Shao (2001) ont étudié les propriétés d'efficacité et de robustesse ; ils montrent que l'estimateur du minimum de la divergence de Hellinger $\widehat{\theta}_H$ est préférable à l'estimateur du maximum de vraisemblance, $\widehat{\theta}_H$ est le meilleur, parmi tous les estimateurs du minimum des divergences de puissance ϕ_γ, en terme d'"efficacité-robustesse".

Lindsay (1994) et Morales *et al.* (1995) proposent d'utiliser les statistiques

$$\widehat{\phi}\left(\mathcal{P}, P_0\right) := \inf_{\theta \in \Theta} \phi(P_\theta, P_n)$$

pour construire des tests de modèles :

$$\mathcal{H}_0 \ :\ P_0 \ \text{ appartient à } \ \mathcal{P}. \tag{9}$$

En ce qui concerne les tests paramétriques des hypothèses simples ou composées

$$\mathcal{H}_0 \ :\ \theta_0 = \theta_1, \qquad\qquad \mathcal{H}_0 \ :\ \theta_0 \in \Theta_0, \tag{10}$$

où θ_1 est une valeur donnée, Θ_0 un sous-ensemble de Θ, on peut utiliser les statistiques

$$\widehat{\phi}(P_{\theta_1}, P_{\theta_0}) = \phi(P_{\theta_1}, P_n) \tag{11}$$

et

$$\inf_{\theta \in \Theta_0} \widehat{\phi}(P_\theta, P_{\theta_0}) = \inf_{\theta \in \Theta_0} \phi(P_\theta, P_n). \tag{12}$$

Il est à noter que la classe des tests basés sur ces statistiques ne contient pas les tests du rapport des maxima des vraisemblances de Wilks pour les hypothèses simples ou

composées.

Lorsque le support S est continu, les estimateurs dans (8) ne sont pas définis. Supposons que les lois P_θ admettent des densités p_θ par rapport à une mesure dominante σ-finie λ. La ϕ-divergence $\phi(P_\theta, P_{\theta_0})$, dans ce cas, s'écrit

$$\phi(P_\theta, P_{\theta_0}) := \int \varphi\left(\frac{p_\theta(x)}{p_{\theta_0}(x)}\right) p_{\theta_0}(x)\ d\lambda(x).$$

Notons $p_{n,h}$ l'estimateur à noyau de la densité p_{θ_0}, i.e.,

$$p_{n,h}(x) := \frac{1}{h}\int K\left(\frac{x-t}{h}\right)\ dP_n(t).$$

Beran (1977) définit et étudie l'estimateur du minimum de la divergence de Hellinger

$$\widehat{\theta}_H := \arg\inf_{\theta\in\Theta}\int \varphi_{1/2}\left(\frac{p_\theta(x)}{p_{n,h}(x)}\right) p_{n,h}(x)\ dx.$$

Basu and Lindsay (1994) introduisent le principe "minimum disparity estimate" (MDE's). Ils utilisent la version modifiée des densités p_θ définie par

$$p^*_\theta(x) := \frac{1}{h}\int K\left(\frac{x-t}{h}\right) p_\theta(t)\ d\lambda(t).$$

Les estimateurs MDE's sont

$$\widehat{\theta}_\phi := \arg\inf_{\theta\in\Theta}\int \varphi\left(\frac{p^*_\theta(x)}{p_{n,h}(x)}\right) p_{n,h}(x)\ dx. \tag{13}$$

L'estimateur associé à la divergence de KL_m, à la différence du cas discret, ne coïncide pas avec l'estimateur du maximum de vraisemblance à cause du lissage. Pour des modèles réguliers, Basu and Lindsay (1994) montrent que les estimateurs (13) sont asymptotiquement normaux et asymptotiquement efficaces si la fonction φ et le noyau K vérifient certaines conditions. Pour construire des tests de modèles (9), on peut utiliser la statistique (c.f. Beran (1977))

$$\widehat{H}\left(\mathcal{P}, P_0\right) := \inf_{\theta\in\Theta}\int \varphi_{1/2}\left(\frac{p_\theta(x)}{p_{n,h}(x)}\right) p_{n,h}(x)\ dx.$$

On peut aussi utiliser les statistiques

$$\widehat{\phi}\left(\mathcal{P}, P_0\right) := \inf_{\theta\in\Theta}\int \varphi\left(\frac{p^*_\theta(x)}{p_{n,h}(x)}\right) p_{n,h}(x)\ dx.$$

D'autre part, les statistiques

$$\widehat{\phi}\left(P_{\theta_1}, P_{\theta_0}\right) := \int \varphi\left(\frac{p_\theta^*(x)}{p_{n,h}(x)}\right) p_{n,h}(x)\ dx,$$

et

$$\inf_{\theta\in\Theta_0} \int \varphi\left(\frac{p_\theta^*(x)}{p_{n,h}(x)}\right) p_{n,h}(x)\ dx$$

peuvent être utilisées pour tester les hypothèses paramétriques simples ou composées (10). La classe des tests basés sur ces statistiques ne recouvre pas les tests du rapport des maxima des vraisemblaces de Wilks pour les hypothèses simples ou composées.

Chapitre 1

En fait, l'objectif du Chapitre 1 est de donner une nouvelle méthode qui permet de définir les estimateurs de minimum des ϕ-divergences pour des modèles paramétriques dans le cas discret ou continu sans utiliser la technique de lissage. Cette méthode généralise de façon naturelle la méthode du maximum de vraisemblance pour l'estimation ponctuelle et pour les tests statistiques.

Pour estimer la divergence de KL entre certaines classes de lois Ω et une loi de probabilité P sans utiliser la technique de lissage, Broniatowski (2003) propose d'utiliser la représentation "duale" bien connue de la divergence de KL comme la transformée de Fenchel-Legendre de la fonction génératrice des moments.

Dans ce Chapitre, sous des conditions faibles, pour toute divergence ϕ et pour toute loi de probabilité P, nous donnons une représentation "duale" des fonctions des ϕ-divergence $Q \to \phi(Q,P)$ qui permet d'estimer $\phi(Q,P)$ par utilisation directe de la mesure empirique P_n dans le cas continu comme dans le cas discret sans utiliser la technique de lissage.

Comme les fonctions $Q \mapsto \phi(Q,P)$ sont convexes, nous allons utiliser des techniques de dualité : nous allons appliquer le Lemme de dualité suivant (c.f. e.g. Dembo and Zeitouni (1998) Lemma 4.5.8)

Lemme 0.1 (Lemme de dualité). *Soit $\mathcal{S}$ un espace vectoriel topologique Hausdorff localement convexe (e.v.t.h.l.c.). Soit $g : \mathcal{S} \mapsto]-\infty, +\infty]$ une fonction convexe semi-continue inférieurement (s.c.i.). Définissons la transformée de Fenchel-Legendre de g comme suit*

$$g^*(l) := \sup_{x\in\mathcal{S}} \left\{l(x) - g(x)\right\},\ l \in \mathcal{S}^*,$$

où $\mathcal{S}^*$ est le dual topologique de $\mathcal{S}$.

Alors, on a

$$g(x) = \sup_{l\in\mathcal{S}^*} \left\{l(x) - g^*(l)\right\}, \tag{14}$$

i.e., g est la transformée de Fenchel-Legendre de g^*.

Une fois appliqué dans le contexte des ϕ-divergences, le Lemme précédent permet d'obtenir une représentation "duale" pour les ϕ-divergences.

Pour pouvoir appliquer ce Lemme, afin d'obtenir une représentation "duale" des fonctions des divergences $Q \mapsto \phi(Q,P)$, nous devons généraliser la définition des fonctions des divergences $Q \mapsto \phi(Q,P)$ sur un espace vectoriel que nous devons munir d'une topologie appropriée pour laquelle les conditions du Lemme de dualité seront satisfaites.

Notons M l'espace vectoriel de toutes les mesures signées finies et $\mathcal{B}_b$ l'ensemble toutes les fonctions mesurables bornées. On considère également une classe de fonctions $\mathcal{F}$. Définissons l'espace vectoriel

$$M_{\mathcal{F}} := \left\{ Q \in M \ \text{ tel que } \ \int |f| \ d|Q| < \infty, \ \text{ pour tout } f \in \mathcal{F} \right\}.$$

On généralise la définition des fonctions des divergences $Q \mapsto \phi(Q,P)$ sur l'espace vetoriel $M_{\mathcal{F}}$ via l'extention de la définition des fonctions convexes réelles $x \mapsto \varphi(x)$ sur $[-\infty,+\infty]$: dans le Chapitre 1 Section 2, nous présentons comment généraliser la définition des divergences de puissance ϕ_γ via l'extention de la définition des fonctions convexes réelles $x \mapsto \varphi_\gamma(x)$ (voir (6)) sur $[-\infty,+\infty]$.

Dans tout ce qui suit, φ désigne une fonction convexe définie sur $[-\infty,+\infty]$ à valeurs dans $[0,+\infty]$ et vérifie $\varphi(1)=0$. Définissons le domaine de φ, noté D_φ, par

$$D_\varphi := \{ x \in [-\infty,+\infty] \ \text{ tel que } \ \varphi(x) < \infty \}.$$

On munit l'espace vectoriel $M_{\mathcal{F}}$ de la topologie (que l'on note $\tau_{\mathcal{F}}$) la moins fine qui rend les fonctions $Q \mapsto \int f \ dQ$ continues, pour tout $f \in \langle \mathcal{F} \cup \mathcal{B}_b \rangle$, où $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$ désigne l'espace vectoriel engendré par la classe $\mathcal{F} \cup \mathcal{B}_b$.

Le choix de l'espace vectoriel $M_{\mathcal{F}}$ et la topologie $\tau_{\mathcal{F}}$ avec $\mathcal{F}$ une classe de fonctions libre, est le choix le plus convenable pour les problèmes statistiques que nous allons considérer comme nous allons voir dans les Chapitres 1,2 et 3.

Nous montrons que

Proposition 0.1. *L'espace vectoriel $M_{\mathcal{F}}$ muni de la topologie $\tau_{\mathcal{F}}$, ($M_{\mathcal{F}}$,$\tau_{\mathcal{F}}$) est un espace vectoriel topologique Hausdorff localement convexe. De plus, son dual topologique est l'espace vetoriel des fonctions linéaires*

$$\left\{ Q \mapsto \int f \ dQ, \ f \in \langle \mathcal{F} \cup \mathcal{B}_b \rangle \right\}.$$

Proposition 0.2. *Pour toute divergence ϕ et toute loi de probabilité P, les fonctions $Q \mapsto \phi(Q,P)$, étant définies sur $(M_{\mathcal{F}}, \tau_{\mathcal{F}})$ à valeurs dans $[0,+\infty]$, sont semi-continues inférieurement.*

La transformée de Fenchel-Legendre des fonctions $Q \mapsto \phi(Q,P)$ est donnée par

$$T(f,P) = \sup_{Q \in M_{\mathcal{F}}} \left\{ \int f \; dQ - \phi(Q,P) \right\}, \quad \text{pour tout} \quad f \in \langle \mathcal{F} \cup \mathcal{B}_b \rangle.$$

En appliquant le Lemme de dualité, nous obtenons

Proposition 0.3. *Pour tout $Q \in M_{\mathcal{F}}$ et pour toute loi P, on a*

$$\phi(Q,P) = \sup_{f \in \langle \mathcal{F} \cup \mathcal{B}_b \rangle} \left\{ \int f \; dQ - T(f,P) \right\}. \tag{15}$$

Notons ψ la conjuguée convexe (ou la transformée de Fenchel-Legendre) de φ, i.e., la fonction définie par

$$t \in \mathbb{R} \to \psi(t) := \sup_{x \in \mathbb{R}} \{tx - \varphi(t)\},$$

et définissons la fonction φ^* comme suit

$$t \in \text{ Im } \varphi' \;\; \mapsto \;\; \varphi^*(t) := t{\varphi'}^{-1}(t) - \varphi\left({\varphi'}^{-1}(t)\right), \tag{16}$$

où φ' est la dérivée de φ, ${\varphi'}^{-1}$ la fonction inverse de φ' et Im φ' l'ensemble de toutes les valeurs prises par φ'. La fonction φ^* coïncide avec ψ sur l'ensemble Im φ'.

Dans le Théorème ci-dessous, sous des conditions faibles, nous donnons la forme explicite de $T(f,P)$ et une représentation duale similaire à la représentation (15) qui fait intervenir non pas l'espace vectoriel dual en entier $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$ mais seulement une classe $\mathcal{F}$ bien choisie de sorte que l'optimum soit atteint dans la représentation duale. Cela permet de proposer par la suite des estimateurs bien définis dans des modèles paramétriques et semi-paramétriques.

Théorème 0.1. *Supposons que la fonction φ soit strictement convexe et soit de classe $\mathcal{C}^2$ sur l'interieur de son domaine D_φ. Soit Q une mesure signée finie et P une loi de probabilité avec $\phi(Q,P) < \infty$. Soit $\mathcal{F}$ une classe de fonctions telle que*

(i) $\int |f| \; d|Q| < \infty$ *pour tout* $f \in \mathcal{F}$ *;*

(ii) $\varphi'\,(dQ/dP)$ *appartient à* $\mathcal{F}$ *;*

(iii) Im $f \subset$ Im φ', *pour tout* $f \in \mathcal{F}$.

Alors, on a

(1) La divergence $\phi(Q,P)$ admet la "représentation duale"

$$\phi(Q,P) = \sup_{f \in \mathcal{F}} \left\{ \int f \; dQ - \int \varphi^*(f) \; dP \right\} \tag{17}$$

(2) Le supremum dans (17) est unique et est atteint en

$$f(x) = \varphi'\left(\frac{dQ}{dP}(x)\right), \quad \textit{pour tout } x \quad (P - p.s.).$$

A présent, on présente comment appliquer ce Théorème (et en particulier comment choisir la classe $\mathcal{F}$) pour définir les estimateurs du minimum des ϕ-divergences. Supposons que la divergence ϕ vérifie

$$\phi\left(P_\theta, P_\alpha\right) < \infty, \quad \text{pour tout } \theta, \alpha \in \Theta.$$

En appliquant le Théorème 0.1, pour tout $\theta \in \Theta$, la divergence $\phi\left(P_\theta, P_{\theta_0}\right)$ s'écrit sous la forme

$$\phi\left(P_\theta, P_{\theta_0}\right) = \sup_{f \in \mathcal{F}} \left\{ \int f \; dP_\theta - \int \varphi^*(f) \; dP_{\theta_0} \right\}.$$

Comme le supremum est atteint en $f = \varphi'(p_\theta / p_{\theta_0})$ et comme θ_0 est inconnu, nous choisirons la classe $\mathcal{F}$ ainsi

$$\mathcal{F} = \left\{ x \mapsto \varphi'\left(\frac{p_\theta}{p_\alpha}(x)\right), \; \alpha \in \Theta \right\}.$$

On obtient donc

$$\phi\left(P_\theta, P_{\theta_0}\right) = \sup_{\alpha \in \Theta} \left\{ \int \varphi'\left(\frac{p_\theta}{p_\alpha}\right) \; dP_\theta - \int \varphi^*\left(\varphi'\left(\frac{p_\theta}{p_\alpha}\right)\right) \; dP_{\theta_0} \right\}. \tag{18}$$

Par conséquent, nous proposons d'estimer la divergence $\phi\left(P_\theta, P_{\theta_0}\right)$ par

$$\widehat{\phi}\left(P_\theta, P_{\theta_0}\right) := \sup_{\alpha \in \Theta} \left\{ \int \varphi'\left(\frac{p_\theta}{p_\alpha}\right) \; dP_\theta - \int \varphi^*\left(\varphi'\left(\frac{p_\theta}{p_\alpha}\right)\right) \; dP_n \right\}. \tag{19}$$

D'autre part, pour tout $\theta \in \Theta$, le supremum dans (18) est unique et est atteint en $\alpha = \theta_0$ (d'après la partie 2 du Théorème 0.1). On peut estimer θ_0 par

$$\widehat{\alpha}_n(\theta) := \arg\sup_{\alpha \in \Theta} \left\{ \int \varphi'\left(\frac{p_\theta}{p_\alpha}\right) \; dP_\theta - \int \varphi^*\left(\varphi'\left(\frac{p_\theta}{p_\alpha}\right)\right) \; dP_n \right\}.$$

Nous définissons l'estimateur du minimum de la ϕ-divergence de θ_0 par

$$\widehat{\theta}_\phi := \arg\inf_{\theta \in \Theta} \sup_{\alpha \in \Theta} \left\{ \int \varphi'\left(\frac{p_\theta}{p_\alpha}\right) \; dP_\theta - \int \varphi^*\left(\varphi'\left(\frac{p_\theta}{p_\alpha}\right)\right) \; dP_n \right\}. \tag{20}$$

Remarque 0.1. *On constate que la définition des estimateurs $\widehat{\theta}_\phi$ de θ_0 ne suppose pas que les lois P_θ soient discrètes.*

Remarque 0.2. *(Un autre point de vue à l'estimateur du maximum de vraisemblance). La classe d'estimateurs $\widehat{\theta}_\phi$ contient l'esimateur du maximum de vraisemblance ; il est obtenu pour la divergence de Kullback-Leibler modifiée (KL_m), i.e., lorsque $\varphi(x) = \varphi_0(x) = -\log x + x - 1$. L'EMV est celui donc qui minimise l'estimateur de la divergence KL_m entre le modèle paramétrique $\mathcal{P} := \{P_\theta,\ \theta \in \Theta\}$ et la loi de l'échantillon P_{θ_0}.*

Remarque 0.3. *(Estimateur de la divergence de Kullback-Leibler modifiée et test du rapport des maxima des vraisemblances de Wilks). Les statistiques $\widehat{\phi}(P_{\theta_1}, P_{\theta_0})$ (définie par l'expression (19) et $\inf_{\theta\in\Theta_0} \widehat{\phi}(P_\theta, P_{\theta_0})$ peuvent être utilisées pour tester les hypothèses simples et composés (10). Les tests basés sur les statistiques $\widehat{KL_m}(P_{\theta_1}, P_{\theta_0})$ et $\inf_{\theta\in\Theta_0} \widehat{KL_m}(P_\theta, P_{\theta_0})$ coïncident avec les tests du rapport des maxima des vraisemblances de Wilks pour les hypothèses simples et composées.*

Nous étudions le comportement asymptotique des estimateurs introduits des paramètres et nous montrons que les estimateurs $\widehat{\theta}_\phi$, sous des conditions de régularité, sont tous asymptotiquement normaux et asymptotiquement efficaces indépendamment de la divergence considérée. Le comportement asymptotique des estimateurs des divergences est étudié sous le modèle et également sous l'hypothèse de mauvaise spéfication.

Dans la Section 4, nous considérons le problème du choix de la divergence. Nous présentons comment choisir la divergence ϕ qui permet de construire des régions de confiances pour le paramètre θ_0 plus précises que celles obtenues par la méthode du maximum de vraisemblance, et qui utilise la connaissance à priori qu'on peut disposer sur le paramètre θ_0.

Ce chapitre est résumé dans une note au Comptes-Rendus de l'Académie des Sciences sous la référence :

A. Keziou (2003). Dual representation of ϕ-divergences and applications. C. R. Math. Acad. Sci. Paris, Ser. I 336 857-862.

Perspectives : L'étude des problèmes de comparaison des estimateurs $\widehat{\theta}_\phi$ en terme d'efficacité et robustesse nécessite l'étude de l'efficacité aux ordres supérieurs puisque tous les estimateurs sont équivalents au premier ordre. Nous espérons développer certains de ces problèmes dans des travaux futurs.

Chapitre 2

Dans le Chapitre 2, nous considérons le problème d'estimation et de test à deux échantillons pour des modèles à rapport de densités semi-paramétriques. Soient $X_1, \ldots, X_{n_0}$ et $Y_1, \ldots, Y_{n_1}$ deux échantillons de lois G et H, respectivement.

Un modèle à rapport de densités semi-paramétriques s'écrit sous la forme

$$\frac{dH}{dG}(x) = m(\theta_0, x), \tag{21}$$

où θ_0 est le paramètre d'intérêt appartient à un ouvert Θ de $\mathbb{R}^d$ et $m(.,.)$ est une fonction positive.

La Section 1 en présente des exemples et des motivations : tests de comparaison, estimation et tests dans des modèles logistiques.

Le modèle (21) est appelé "multiplicative-intercept-risk model" si $m(.,.)$ est de la forme

$$m(\theta, x) = \exp\{\alpha + r(x, \beta)\}, \tag{22}$$

où $\theta := (\alpha, \beta^T)^T$, $\beta \in \mathbb{R}^{d-1}$ et $r(.,.)$ une fonction spécifiée.

En général, pour comparer deux lois H et G, on estime une mesure de différence (distance, pseudo-distance ou divergence) entre les deux lois.

Dans ce Chapitre, on considère la classe des divergences entre lois de probabilité.

L'estimateur "plug-in" de $\phi(H, G)$, la ϕ-divergence entre H et G, est défini par

$$\widehat{\phi}(H, G) := \phi\left(H_{n_1}^Y, G_{n_0}^X\right) \tag{23}$$

où $H_{n_1}^Y$ et $G_{n_0}^X$ sont les mesures empiriques associées aux deux échantillons $Y_1, \ldots, Y_{n_1}$ et $X_1, \ldots, X_{n_0}$, respectivement. L'estimateur (23) est bien défini seulement lorsque H et G ont un support discret fini commun. Nous proposons donc d'utiliser la représentation duale des ϕ-divergences (voir Théorème 0.1) pour estimer $\phi(H, G)$.

En utilisant la première partie du Théorème 0.1, on obtient

$$\phi(H, G) = \sup_{f \in \mathcal{F}} \left\{ \int f \; dH - \int \varphi^*(f) \; dG \right\}.$$

Comme le supremum est atteint en

$$f = \varphi'\left(\frac{dH}{dG}(x)\right) = \varphi'\left(m(\theta_0, x)\right),$$

et θ_0 est inconnu et est supposé appartenir à Θ, nous allons considérer la classe de fonctions

$$\mathcal{F} = \left\{x \mapsto \varphi'\left(m(\theta, x)\right), \; \theta \in \Theta\right\}.$$

Nous obtenons donc

$$\phi(H, G) = \sup_{\theta \in \Theta} \left\{ \int k(\theta, x) \; dH(x) - \int l(\theta, x) \; dG(x) \right\}, \tag{24}$$

où $k(\theta,x) := \varphi'\left(m(\theta,x)\right)$ et $l(\theta,x) := \varphi'\left(m(\theta,x)\right) m(\theta,x) - \varphi\left(m(\theta,x)\right)$.

Nons proposons d'estimer $\phi(H,G)$ par

$$\widehat{\phi}(H,G) := \sup_{\theta\in\Theta}\left\{\int k(\theta,x)\ dH^Y_{n_1} - \int l(\theta,x)\ dG^X_{n_0}\right\}.$$

Le supremum dans (24) est unique et est atteint en $\theta = \theta_0$; nous proposons donc d'estimer θ_0 par

$$\widehat{\theta}_{n_0,n_1} := \arg\sup_{\theta\in\Theta}\left\{\int k(\theta,x)\ dH^Y_{n_1} - \int l(\theta,x)\ dG^X_{n_0}\right\}.$$

Des résultats de simulations montrent qu'une bonne divergence doit dépendre du rapport $\frac{n_1}{n_0}$. Nous introduisons des divergences dépendant du rapport $\rho := \lim_{n\to\infty}\frac{n_1}{n_0}$ (avec $n := n_1 + n_0$). Notons φ_ρ au lieu de φ et notons ϕ_ρ la divergence associée à φ_ρ. Nous proposons d'estimer la divergence $\phi_\rho(H,G)$, en utilisant le rapport $\rho_n := \frac{n_1}{n_0}$, par

$$\widehat{\phi}_\rho(H,G) := \sup_{\theta\in\Theta}\left\{\int k_{\rho_n}(\theta,x)\ dH^Y_{n_1} - \int l_{\rho_n}(\theta,x)\ dG^X_{n_0}\right\}, \tag{25}$$

avec $k_{\rho_n}(\theta,x) = \varphi'_{\rho_n}\left(m(\theta,x)\right)$ et $l_{\rho_n}(\theta,x) := \varphi'_{\rho_n}\left(m(\theta,x)\right) m(\theta,x) - \varphi_{\rho_n}\left(m(\theta,x)\right)$.

Le paramètre θ_0, peut être estimé par

$$\widehat{\theta}_{\rho_n} := \arg\sup_{\theta\in\Theta}\left\{\int k_{\rho_n}(\theta,x)\ dH^Y_{n_1} - \int l_{\rho_n}(\theta,x)\ dG^X_{n_0}\right\}. \tag{26}$$

Dans la Section 3, on donne les lois limites des estimateurs (25) et (26) dans les deux cas : apparié et non apparié.

La classe d'estimateurs $\widehat{\theta}_{\rho_n}$, dans le cas des modèles (22), recouvre l'estimateur du maximum de vraisemblance semiparamétrique (EMVSP) introduit par Qin (1998). De plus, les estimateurs $\widehat{\theta}_{\rho_n}$, à la différence de l'EMVSP, sont définis dans les deux cas, apparié et non apparié.

Dans le Section 5, nous comparons, par simulations, l'efficacité des différents estimateurs $\widehat{\theta}_{\rho_n}$ et l'EMVSP ; nous comparons également leurs sensibilités dans le cas des données contaminées. Les résultats de simulations montrent que le choix de la divergence dépend du modèle considéré.

Perspectives : Le problème de comparaison des estimateurs $\widehat{\theta}_{\rho_n}$ en terme d'efficacité et robustesse ainsi que le problème de comparaison des tests basés sur les statistiques $\widehat{\phi}_\rho(H,G)$ pourront faire l'objet de travaux ultérieurs.

Chapitre 3

Dans le Chapitre 3, nous considérons les problèmes (**P1**) et (**P2**) dans le cas des modèles non paramétriques définis par un nombre fini de contraintes linéaires à paramètre inconnu. L'ensemble Ω est défini par

$$\Omega := \bigcup_{\theta\in\Theta} \mathcal{M}_\theta := \bigcup_{\theta\in\Theta} \left\{ Q \in M^1 \ / \ \int g(x,\theta)\ dQ(x) = 0 \right\} =: \mathcal{M}.$$

De nombreux problèmes statistiques peuvent être étudiés dans le cadre de ces modèles, la Section 1 en présente plusieurs exemples.

L'approche développée dans ce Chapitre est basée sur des méthodes de projection au sens des ϕ-divergences.

Pour construire des tests de modèles

$$\mathcal{H}_0 \ : \ P_0 \ \text{ appartient à } \ \mathcal{M},$$

on peut utiliser les estimateurs des ϕ-divergences entre le modèle $\mathcal{M}$ et la loi P_0.

L'estimateur "plug-in" de la ϕ-divergence entre l'ensemble $\mathcal{M}_\theta$ est la loi P_0 est défini par

$$\widehat{\phi}\,(\mathcal{M}_\theta, P_0) := \inf_{Q\in\mathcal{M}_\theta} \phi(Q, P_n) = \inf_{Q\in\mathcal{M}_\theta} \int \varphi\left(\frac{dQ}{dP_n}(x)\right)\ dP_n(x). \tag{27}$$

Si l'infimum existe, il est clair qu'il est atteint en une loi absolument continue par rapport à P_n. Définissons donc les ensembles

$$\mathcal{M}_\theta^{(n)} := \left\{ Q \in M^1 \text{ tel que } Q \ll P_n \text{ et } \sum_{i=1}^n g(X_i,\theta) Q(X_i) = 0 \right\},$$

qui peuvent être vus comme des sous-ensembles de $\mathbb{R}^n$.

L'estimateur plug-in (27) s'écrit

$$\widehat{\phi}\,(\mathcal{M}_\theta, P_0) = \inf_{Q\in\mathcal{M}_\theta^{(n)}} \frac{1}{n}\sum_{i=1}^n \varphi\left(nQ(X_i)\right). \tag{28}$$

L'infimum, dans cette expression, peut être atteint en un point appartenant à la frontière de l'ensemble $\mathcal{M}_\theta^{(n)}$. Dans ce cas, on ne peut pas utiliser la méthode de Lagrange pour caractériser l'infimum et calculer $\widehat{\phi}\,(\mathcal{M}_\theta, P_0)$.

Pour remédier à cela, nous proposons de considérer des ensembles de toutes les mesures signées finies Q absolument continues par rapport à P_n et qui vérifient les contraintes

$$\int dQ = 1 \ \text{ et } \ \int g(x,\theta)\ dQ(x) = 0.$$

Dans tout ce qui suit, les ensembles $\mathcal{M}_\theta^{(n)}$ et $\mathcal{M}_\theta$ sont des ensembles de mesures signées finies que l'on définit comme suit

$$\mathcal{M}_\theta^{(n)} := \left\{ Q \in M \ \text{ tel que } \ Q \ll P_n \text{ et } \sum_{i=1}^{n} g(X_i, \theta) Q(X_i) = 0 \right\} \tag{29}$$

et

$$\mathcal{M}_\theta := \left\{ Q \in M \ \text{ tel que } \ \int dQ = 1 \ \text{ et } \ \int g(x, \theta) \, dQ(x) = 0 \right\}, \tag{30}$$

où M désigne l'espace de toutes les mesures signées finies.

Pour tout $\theta \in \Theta$, la divergence $\phi\left(\mathcal{M}_\theta, P_0\right)$ peut être estimée par

$$\widehat{\phi}\left(\mathcal{M}_\theta, P_0\right) := \inf_{Q \in \mathcal{M}_\theta^{(n)}} \frac{1}{n} \sum_{i=1}^{n} \varphi\left(nQ(X_i)\right). \tag{31}$$

La divergence $\phi\left(\mathcal{M}, P_0\right)$ entre le modèle $\mathcal{M}$ et la loi P_0 est estimée par

$$\widehat{\phi}\left(\mathcal{M}, P_0\right) := \inf_{\theta \in \Theta} \inf_{Q \in \mathcal{M}_\theta^{(n)}} \frac{1}{n} \sum_{i=1}^{n} \varphi\left(nQ(X_i)\right). \tag{32}$$

Le paramètre d'intérêt est estimé par

$$\widehat{\theta}_\phi := \arg \inf_{\theta \in \Theta} \inf_{Q \in \mathcal{M}_\theta^{(n)}} \frac{1}{n} \sum_{i=1}^{n} \varphi\left(nQ(X_i)\right). \tag{33}$$

Pour toute divergence ϕ, nous proposons d'appeler $\widehat{\theta}_\phi$ "estimateur du minimum de la ϕ-divergence emirique" (EMϕDE).

La classe d'estimateurs $\widehat{\theta}_\phi$ contient l'estimateur du maximum de vraisemblance empirique (EMVE) (c.f. Owen (1990), Qin and Lawless (1994) et Owen (2001)), il est obtenu pour la divergence de KL_m, i.e., $EMVE = \widehat{\theta}_{KL_m}$.

L'étude du comportement asymptotique des estimateurs (31), (32) et (33), en particulier sous l'alternative de mauvaise spécification, nécessite l'étude des problèmes d'existence et de caractérisation des ϕ-projections de la loi P_0 sur les ensembles $\mathcal{M}_\theta$. Ces problèmes ont fait l'objet des Sections 2 et 3.

Dans la Section 4, en utilisant la représentation duale des ϕ-divergences (voir Théorème 0.1 partie 1), on montre que les estimateurs $\widehat{\phi}\left(\mathcal{M}_\theta, P_0\right)$ peuvent s'écrire sous la forme

$$\widehat{\phi}\left(\mathcal{M}_\theta, P_0\right) = \sup_{t \in \mathcal{C}_\theta^{(n)}} \int m(x, \theta, t) \, dP_n(x), \tag{34}$$

où $\mathcal{C}_\theta^{(n)}$ est un sous-ensemble de $\mathbb{R}^{l+1}$ défini par

$$\mathcal{C}_\theta^{(n)} := \left\{ t \in \mathbb{R}^{(l+1)} \text{ tel que } t_0 + \sum_{j=1}^{l} t_j g_j(X_i, \theta) \in \text{ Im } \varphi', \text{ pour tout } i = 1, \ldots, n \right\},$$

et

$$m(x, \theta, t) \;\; := \;\; t_0 - \varphi^* \left(t_0 + \sum_{j=1}^{l} t_j g_j(x, \theta) \right).$$

L'expression (34) transforme le problème d'optimisation (31) sous contraintes en un problème d'optimisation simple (sans contraintes). D'une part, ceci simplifie en pratique le calcul des estimateurs (31), (32) et (33) et, d'autre part, elle permet d'obtenir les lois limites. En utilisant (34), on peut écrire les estimateurs (32) et (33) sous la forme

$$\widehat{\phi}\,(\mathcal{M}, P_0) = \inf_{\theta \in \Theta} \sup_{t \in \mathcal{C}_\theta^{(n)}} \int m(x, \theta, t)\; dP_n(x), \tag{35}$$

et

$$\widehat{\theta}_\phi = \arg \inf_{\theta \in \Theta} \sup_{t \in \mathcal{C}_\theta^{(n)}} \int m(x, \theta, t)\; dP_n(x). \tag{36}$$

Dans la Section 5, on étudie, sous le modèle et également sous l'alternative de mauvaise spécification, le comportement asymptotique des estimateurs proposés en se servant des représentations (34), (35) et (36). Il est à noter que la méthode du maximum de vraisemblance empirique n'a pas été étudiée sous l'alternative de mauvaise spécification. Nous considérons également le problème de test des modèles, le problème des tests simples et composés relatif au paramètre θ_0 et le problème d'estimation de θ_0 par régions de confiance non paramétriques.

Dans la Section 6, en généralisant le résultat de Qin and Lawless (1994), nous définissons une classe d'estimateurs de la fonction de répartition qui tient compte du fait que la loi P_0 vérifie des contraintes linéaires ; les estimateurs que nous obtenons sont généralement plus efficaces que la fonction de répartition empirique.

Dans la Section 7, nous comparons ces méthodes avec la méthode du maximum de vraisemblance empirique en donnant des éléments de réponse au problème du choix de la divergence.

Dans la Section 8, nous comparons les propriétés d'efficacité et de robustesse des estimateurs $\widehat{\theta}_\phi$. Les résultats de simulations montrent que l'estimateur du minimum de la divergence de Hellinger $\widehat{\theta}_H$ est préférable à l'estimateur du maximum de vraisemblance empirique et possède de bonnes propriétés d'efficacité-robustesse.

Perspectives : Une étude plus profonde du problème de comparaison des estimateurs $\widehat{\theta}_\phi$ en terme d'efficacité et robustesse, et du problème de comparaison des tests basés sur les statistiques $\widehat{\phi}(\mathcal{M}_\theta, P_0)$ et $\widehat{\phi}(\mathcal{M}, P_0)$ pourra faire l'objet de travaux futurs.

Annexe

Dans l'Annexe, nous considérons le problème d'estimation de l'entropie des lois à support dénombrable. Soit P une loi de probabilité discrète sur un espace infini dénombrable $\mathcal{X}$. On étudie la vitesse de convergence presque sûre de l'estimateur "plug-in" de l'entropie de Shannon $H := H(P)$ de la loi de probabilité inconnue P. On démontre aussi la convergence presque sûre de l'estimateur pour des variables aléatoires stationnaires ergodiques, et pour des variables aléatoires stationnaires $\alpha-$mélangeantes sous une condition faible sur la queue de distribution de la loi P. Cette partie est résumée dans une note aux Comptes-Rendus de l'Académie des Sciences sous la référence :

A. Keziou (2002). Sur l'estimation de l'entropie des lois à support dénombrable. C. R. Math. Acad. Sci. Paris, 335 (9), 763-766.

Chapitre 1

Parametric Estimation and Tests through Divergences

Une partie de ce Chapitre est publiée en version réduite sous la référence : Keziou (2003). Dual representation of ϕ-divergences and applications. C. R. Math. Acad. Sci. Paris, Ser. I 336 857-862.

We introduce estimation and test procedures through divergence optimization for discrete and continuous parametric models. This approach is based on a new dual representation for divergences. We treat point estimation and tests for simple and composite hypotheses, extending maximum likelihood techniques. Confidence regions with small area are deduced from this approach, as well as some solution for the problem of multiple maxima of the likelihood surface.

1.1 Introduction and notation

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space. Let φ be a nonnegative convex function defined from $[0, +\infty]$ onto $[0, +\infty]$ and satisfying $\varphi(1) = 0$. Let P be a probability measure (p.m.) defined on $(\mathcal{X}, \mathcal{B})$. For any p.m. Q absolutely continuous (a.c.) with respect to (w.r.t.) P, the ϕ-divergence between Q and P is defined by

$$\phi(Q, P) := \int \varphi\left(\frac{dQ}{dP}\right) \, dP. \tag{1.1}$$

When Q is not a.c. w.r.t. P, we set $\phi(Q, P) := +\infty$. Therefore, $Q \to \phi(Q, P)$ is defined on the whole class of all probability measures on $(\mathcal{X}, \mathcal{B})$. This definition has been introduced by Rüschendorf (1984). A former definition of ϕ-divergences (called f-divergences in Csiszár (1963)) is due to Csiszár (1963) ; his definition does not require absolute continuity of Q with respect to P, but uses a common dominating

σ-finite measure λ for both Q and P; since we will consider a whole set of p.m.'s Q, all a.c. w.r.t. P, it is more convenient to use definition (1.1). Also both definitions coincide on the set of p.m.'s a.c. w.r.t. P and dominated by λ.

For any p.m. P, the mapping $Q \to \phi(Q,P)$ is convex and nonnegative. When $Q = P$, the ϕ-divergence between Q and P is zero. When the function $x \to \varphi(x)$ is a strictly convex function in a neighborhood of $x = 1$, then the following fundamental property holds

$$\phi(Q,P) = 0 \quad \text{if and only if} \quad Q = P.$$

We refer to Csiszár (1963), Csiszár (1967c), Csiszár (1967a) and to Liese and Vajda (1987) Chapter 1, for the proofs of these properties.

Let us conclude these few remarks quoting that in general $\phi(Q,P)$ and $\phi(P,Q)$ are not equal. Indeed, they coincide if and only if there exists some real number c such that for any positive x, it holds $\varphi(x) - x\varphi(1/x) = c(x-1)$ (see Liese and Vajda (1987) Theorem 1.13). Hence, ϕ-divergences usually are not distances, but they merely measure some difference between two p.m.'s. Of course a main feature of divergences between distributions of r.v's X and Y is the invariance property with respect to any common change of variables.

1.1.1 Examples of ϕ-divergences.

Some divergences are widely used in statistics and in Information theory. They are the Kullback-Leibler (KL) and the modified KL (KL_m) divergences, defined as follows

$$KL(Q,P) := \begin{cases} \int \log\left(\frac{dQ}{dP}\right)\, dQ & \text{if} \quad Q \text{ is a.c. w.r.t. } P, \\ +\infty & \text{otherwise,} \end{cases}$$

$$KL_m(Q,P) := \begin{cases} \int -\log\left(\frac{dQ}{dP}\right)\, dP & \text{if} \quad Q \text{ is a.c. w.r.t. } P, \\ +\infty & \text{otherwise.} \end{cases}$$

The first one corresponds to $\varphi(x) = x\log x - x + 1$, and the second one to $\varphi(x) = -\log x + x - 1$. Other well known divergences are the χ^2 and the modified χ^2

$$\chi^2(Q,P) := \begin{cases} \int \frac{1}{2}\left(\frac{dQ}{dP} - 1\right)^2 dP & \text{if} \quad Q \text{ is a.c. w.r.t. } P, \\ +\infty & \text{otherwise,} \end{cases}$$

$$\chi^2_m(Q,P) := \begin{cases} \int \frac{1}{2}\frac{\left(\frac{dQ}{dP}-1\right)^2}{\frac{dQ}{dP}}\, dP & \text{if} \quad Q \text{ is a.c. w.r.t. } P, \\ +\infty & \text{otherwise.} \end{cases}$$

The φ functions are respectively $\varphi(x) = \frac{1}{2}(x-1)^2$ and $\varphi(x) = \frac{1}{2}\frac{(x-1)^2}{x}$.

Both Hellinger and L^1 distances are ϕ-divergences, namely

$$H(Q,P) := \begin{cases} \int 2\left(\sqrt{\frac{dQ}{dP}}-1\right)^2 dP & \text{if} \quad Q \text{ is a.c. w.r.t. } P, \\ +\infty & \text{otherwise,} \end{cases}$$

$$L(Q,P) := \begin{cases} \int \left|\frac{dQ}{dP}-1\right| dP & \text{if} \quad Q \text{ is a.c. w.r.t. } P, \\ +\infty & \text{otherwise,} \end{cases}$$

which correspond to $\varphi(x) = 2\left(\sqrt{x}-1\right)^2$ and $\varphi(x) = |x-1|$. All the above examples except the last one are peculiar cases of the so-called "power divergences", introduced by Cressie and Read (1984) (see also Liese and Vajda (1987) Chapter 2), which are defined through the class of convex real valued functions

$$x \in \mathbb{R}_+^* \to \varphi_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma-1)} \tag{1.2}$$

for γ in $\mathbb{R}\setminus\{0,1\}$, $\varphi_0(x) := -\log x + x - 1$ and $\varphi_1(x) := x\log x - x + 1$. For all $\gamma \in \mathbb{R}$, $\varphi_\gamma(0) := \lim_{x\downarrow 0}\varphi_\gamma(x)$ and $\varphi_\gamma(+\infty) := \lim_{x\uparrow+\infty}\varphi_\gamma(x)$.

The KL-divergence is associated to φ_1, the KL_m to φ_0, the χ^2 to φ_2, the χ^2_m to φ_{-1} and the Hellinger distance to $\varphi_{1/2}$.

For all $\gamma \in \mathbb{R}$, sometimes, we denote ϕ_γ the divergence associated to the convex real valued function φ_γ.

In this Chapter, we are interested in estimation and test using ϕ-divergences. An i.i.d. sample $X_1,\ldots,X_n$ with common unknown distribution P is observed and some p.m. Q is given. We intend to estimate $\phi(Q,P)$ and, more generally, $\inf_{Q\in\Omega}\phi(Q,P)$ where Ω is some set of p.m.'s, as well as the p.m. Q^* achieving the infimum in Ω. In the parametric context, these problems can be well defined and lead to new results in estimation and tests, extending classical notions.

1.1.2 Statistical examples and motivations

1- Tests of fit

Let Q_0 and P be two p.m.'s with same finite discrete support S. It holds

$$\phi(Q_0,P) = \sum_{j\in S}\varphi\left(\frac{Q_0(j)}{P(j)}\right)P(j)$$

which can then be estimated via plug-in, setting

$$\widehat{\phi}_n(Q_0,P) := \phi(Q_0,P_n) = \sum_{j\in S}\varphi\left(\frac{Q_0(j)}{P_n(j)}\right)P_n(j),$$

where P_n is the empirical measure pertaining to the sample $X_1, \ldots, X_n$ with distribution P, namely,

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

in which δ_x is the Dirac measure at point x.

More generally, when Q_0 and P have continuous common support S, consider a partition $A_1, \ldots, A_k$ of S. The divergence $\phi(Q_0, P)$ can be approximated by

$$\phi(Q_0, P) \simeq \sum_{j=1}^{k} \varphi\left(\frac{Q_0(A_j)}{P(A_j)}\right) P(A_j), \tag{1.3}$$

which, in turn is estimated by

$$\widehat{\phi}_n(Q_0, P) = \sum_{j=1}^{k} \varphi\left(\frac{Q_0(A_j)}{P_n(A_j)}\right) P_n(A_j).$$

In this vein, goodness of fit tests have been proposed by Cressie and Read (1984), Landaburu and Pardo (2000) for fixed number of classes, and by Györfi and Vajda (2002) when the number of classes depends on the sample size.

2- Parametric estimation and tests

Let $\{P_\theta, \theta \in \Theta\}$ be some parametric model with Θ an open set in $\mathbb{R}^d$. On the basis of an i.i.d. sample $X_1, \ldots, X_n$ with distribution P_{θ_0}, we want to estimate θ_0, the unknown true value of the parameter and perform statistical tests on the parameter using ϕ-divergences.

When all p.m.'s P_θ share the same finite support S and when the support S does not depend upon θ, we have

$$\phi(P_\theta, P_{\theta_0}) = \sum_{j \in S} \varphi\left(\frac{P_\theta(j)}{P_{\theta_0}(j)}\right) P_{\theta_0}(j).$$

For such models, Liese and Vajda (1987), Lindsay (1994) and Morales *et al.* (1995) introduced the so-called "Minimum ϕ-divergences estimates" (MϕDE's) (Minimum Disparity Estimators in Lindsay (1994)) of the parameter θ_0, defined by

$$\widehat{\theta}_n := \arg \inf_{\theta \in \Theta} \phi(P_\theta, P_n), \tag{1.4}$$

where $\phi(P_\theta, P_n)$ is the plug-in estimate of $\phi(P_\theta, P_{\theta_0})$

$$\phi(P_\theta, P_n) = \sum_{j \in S} \varphi\left(\frac{P_\theta(j)}{P_n(j)}\right) P_n(j).$$

Various parametric tests can be performed based on the previous estimates of the ϕ-divergences; see Lindsay (1994) and Morales *et al.* (1995).

Also in Information theory, estimation of the KL-divergence leads to accurate bounds for the average length of a Shannon code based on a source estimate; see Cover and Thomas (1991) Theorem 5.4.3.

The class of estimates in (1.4) contains the maximum likelihood estimate (MLE). Indeed, when $\varphi(x) = \varphi_0(x) = -\log x + x - 1$, we obtain
$\widehat{\theta}_n := \arg\inf_{\theta\in\Theta} KL_m(P_\theta, P_n) = \arg\inf_{\theta\in\Theta} \sum_{j\in S} -\log(P_\theta(j)) = \text{MLE}$.

When interested in testing hypotheses $\mathcal{H}_0 \ : \ \theta_0 = \alpha_0$ against alternatives $\mathcal{H}_1 \ : \ \theta_0 \neq \alpha_0$, we can use the statistics $\phi(P_{\alpha_0}, P_n)$, the plug-in estimate of the divergence between P_{α_0} and P_{θ_0}, rejecting $\mathcal{H}_0$ for large values of the statistics; see Cressie and Read (1984). In the case when $\varphi(x) = -\log x + x - 1$, this test does not coincide with the maximum likelihood ratio test, which defined through the Wilks likelihood ratio statistic $\lambda_n := 2\log \frac{\sup_{\theta\in\Theta} \prod_{i=1}^n p_\theta(X_i)}{\prod_{i=1}^n p_{\alpha_0}(X_i)}$. The new estimate $\widehat{KL}_m(P_{\alpha_0}, P_{\theta_0})$ of $KL_m(P_{\alpha_0}, P_{\theta_0})$, which is proposed in this Chapter, leads to the maximum likelihood ratio test (see Remark 1.6 below).

When the support S is continuous, the estimates in (1.4) are not defined; Basu and Lindsay (1994) investigate the so-called "minimum disparity estimators" (MDE's) for continuous models; they consider

$$p_{n,h}(x) := \frac{1}{h} \int K\left(\frac{x-t}{h}\right) \, dP_n(t),$$

the kernel estimate of the density p_{θ_0} of the $X_i's$, and the modified version of the p.m.'s p_θ defined by

$$p^*_\theta(x) := \frac{1}{h} \int K\left(\frac{x-t}{h}\right) p_\theta(t) \, dt.$$

The MDE is then defined as

$$\arg\min_{\theta\in\Theta} \phi(p^*_\theta, p_{n,h}) := \arg\min_{\theta\in\Theta} \int \varphi\left(\frac{p^*_\theta(t)}{p_{n,h}(t)}\right) p_{n,h}(t) \, dt.$$

When $\varphi(x) = -\log x + x - 1$, this estimate clearly, due to smoothing, does not coincide with the MLE. Also, the test based on $KL_m(p^*_{\alpha_0}, p_{n,h})$ is different from the likelihood ratio test.

No direct plug-in estimate of $\phi(Q, P)$ can be performed by substitution of P by P_n when Q belongs to some class of p.m.'s a.c. w.r.t. the Lebesgue measure λ. In order to build tests pertaining to the density $p := \frac{dP}{d\lambda}$, Beran (1977) and Berlinet

et al. (1998) proposed to use the smoothed kernel estimate p_n of p. Beran (1977) handles the Hellinger distance, while Berlinet (1999) obtains the limiting distribution of the estimate for the Kullback-Leibler divergence. The extension of their results to other divergences remains an open problem ; see Berlinet (1999), Györfi *et al.* (1998), and Berlinet *et al.* (1998). Also, it seems difficult to use such methods to obtain the limiting distribution of an estimate of $\inf_{Q\in\Omega}\phi(Q,P)$ when Ω is some class of p.m.'s ; this problem will be treated in the present paper when Ω is a parametric class, avoiding the smoothing method.

In any case, an exhaustive study of MϕDE's seems necessary, in a way that would include both the discrete and the continuous support cases. This is precisely the scope of this Chapter.

When the support S is discrete finite, the estimates in (1.4) are well defined for large n, since then the empirical measure gives positive mass to any point of S with probability one. However, when S is infinite or continuous, then the plug-in estimate $\phi(P_\theta, P_n)$ usually takes infinite value when no use is done of some partition-based approximation, as done in (1.3). In Broniatowski (2003), a new estimation procedure is proposed in order to estimate the KL-divergence between some set of p.m.'s Ω and some p.m. P, without making use of any partitioning nor smoothing, but merely making use of the well known dual representation of the KL-divergence as the Fenchel-Legendre transform of the moment generating function.

In this Chapter, we give a new general representation for all the ϕ-divergences and we will use this representation in order to define the minimum ϕ-divergences estimates in both discrete and continuous parametric models. This representation is obtained through an application of the following duality Lemma, whose proof can be found for example in Dembo and Zeitouni (1998) Chapter 4 Lemma 4.5.8 or Azé (1997) Chapter 4.

Lemma 1.1. *Let $\mathcal{S}$ be a locally convex Hausdorff topological linear space, and let $g : \mathcal{S} \to (-\infty, +\infty]$ be some convex lower semi continuous (l.s.c.) function. Define the Fenchel-Legendre transform of g by*

$$g^*(l) := \sup_{x\in\mathcal{S}} \{l(x) - g(x)\}, \quad l \in \mathcal{S}^*,$$

where $\mathcal{S}^*$ is the topological dual space of $\mathcal{S}$. We then have

$$g(x) = \sup_{l\in\mathcal{S}^*} \{l(x) - g^*(l)\},$$

which is to say that g is the Fenchel-Legendre transform of g^*.

When applied in the context of ϕ-divergences, the above Lemma allows us to obtain a general representation for all the ϕ-divergences which we will call "Dual representation of ϕ-divergences". This representation is the starting point for the definition of estimates of the parameter θ_0, which we will call "minimum dual ϕ-divergences estimates" (MDϕDE's). They are defined in parametric models $\{P_\theta, \theta \in \Theta\}$, where the p.m.'s P_θ do not necessarily have finite support ; it can be discrete or continuous, bounded or not. Also the same representation will be applied in order to estimate $\phi(P_\alpha, P_{\theta_0})$ and $\inf_{\alpha\in\Theta_0} \phi(P_\alpha, P_{\theta_0})$ where Θ_0 is a subset of Θ, which leads to various simple and composite tests pertaining to θ_0, the true unknown value of the parameter. When $\varphi(x) = -\log x + x - 1$, the MD$\phi$DE's coincide with the maximum likelihood estimates (see Remark 1.5 below) ; since our approach includes also test procedures, it will be seen that with this peculiar choice for the function φ, we recover the classical likelihood ratio test for simple hypotheses and for composite hypotheses (see Remark 1.6 and Remark 1.7 below).

The interest in divergence techniques is that it provides alternatives to the Maximum Likelihood (ML) method in various contexts. Divergence-based confidence regions for parameters generalize the classical ones, and may be smaller than those obtained via ML. In an other context, a rather common problem in ML estimations arises from multiple maxima of the Likelihood surface, when the model, although identifiable, is close to non identifiability, and the sample size is small. Divergence technique may allow for a reasonable choice among all local maximizers of the likelihood.

This Chapter is organized as follows. In Section 2, we prove that, for all p.m. P, the ϕ-divergences functions $Q \to \phi(Q, P)$ satisfy the conditions of the duality Lemma 1.1, and we will give the dual representation for the ϕ-divergences. Section 3 presents, through the dual representation of ϕ-divergences, various estimates and tests in the parametric framework and deals with their asymptotic properties. Section 4 is devoted to some statistical illustration of our results. We first define a class of divergences enjoying some ordering property, which we call the *symmetric power divergences (s.p.d.'s).* Using s.p.d.'s we can obtain confidence regions with small areas, taking into account the parametric model and the information at hand pertaining to the true unknown parameter. Next, we explore the advantage of divergence approach in the case of multiple maxima of the likelihood surface. We consider the case of a mixture of two normal densities and show, by simulation, that the location of a proper estimate can be obtained through a mixed divergence-likelihood approach.

We sometimes write Pf for $\int f \, dP$ for any measure P and any function f.
All proofs are in Section 5.

1.2 Duality and ϕ-divergences

In this section, for all p.m. P, we prove that the ϕ-divergences functions $Q \to \phi(Q,P)$ satisfy the conditions of the duality Lemma 1.1. Since the space of all p.m.'s is not a linear space, we need to extend the definition in (1.1) of the ϕ-divergences functions $Q \to \phi(Q,P)$ on a linear space. So, let M be the space of all finite signed measures defined on $(\mathcal{X},\mathcal{B})$. We also consider a class $\mathcal{F}$ of measurable real valued functions f defined on $\mathcal{X}$. We denote $\mathcal{B}_b$ the set of all measurable bounded functions defined on $\mathcal{X}$ and $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$ the linear span of $\mathcal{F} \cup \mathcal{B}_b$. Define the linear subspace

$$M_{\mathcal{F}} := \left\{ Q \in M \ \text{ such that } \ \int |f| \, d|Q| < \infty, \ \text{ for all } \ f \text{ in } \mathcal{F} \right\},$$

where $|Q|$ is the total variation of the signed finite measure Q.

We extend the definition in (1.1) of the ϕ-divergences functions $Q \to \phi(Q,P)$ on the whole linear subspace $M_{\mathcal{F}}$ as follows :

For ϕ_γ-divergences : define $\varphi'_\gamma(0)$ by $\varphi'_\gamma(0) := \lim_{x \downarrow 0} \varphi'_\gamma(x)$. When the convex real valued functions φ_γ are not defined on $]-\infty,0[$ or when are defined on whole $\mathbb{R}$ but are not convex function on whole $\mathbb{R}$, we extend the definition of φ_γ through

$$x \in \mathbb{R} \ \mapsto \ \varphi_\gamma(x) \mathbb{1}_{[0,+\infty]}(x) + \left(\varphi'_\gamma(0) x + \varphi(0) \right) \mathbb{1}_{[-\infty,0[}(x). \tag{1.5}$$

Note that for the χ^2-divergence, φ_2 is defined and convex on whole $\mathbb{R}$.
More generally, we can consider any convex function φ defined from $[-\infty,+\infty]$ onto $[0,+\infty]$ satisfying $\varphi(1)=0$. Define the domain of φ through

$$D_\varphi := \{ x \in [-\infty,+\infty] \ \text{ such that } \ \varphi(x) < +\infty \}. \tag{1.6}$$

Since φ is convex function, D_φ is an interval, it may be open or not, bounded or unbounded. Namely, $D_\varphi := (a,b)$ with a and b may be finite or infinite. We assume that $a < 1 < b$, $\varphi(a) := \lim_{x \downarrow a} \varphi(x)$ and $\varphi(b) =: \lim_{x \uparrow b} \varphi(x)$ which may be finite or infinite.

We equip the linear space $M_{\mathcal{F}}$ with the $\tau_{\mathcal{F}}-$topology, which is the weakest topology for which all mappings $Q \to \int f \, dQ$ are continuous for all f in $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$. A base of open neighborhoods for any R in $M_{\mathcal{F}}$ is defined by

$$U(R,\mathcal{A},\varepsilon) := \left\{ Q \in M_{\mathcal{F}} \ \text{ such that } \ \max_{f \in \mathcal{A}} |Qf - Rf| < \varepsilon \right\} \tag{1.7}$$

for $\varepsilon > 0$ and $\mathcal{A}$ a finite collection of functions in $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$.

We refer to Dunford and Schwartz (1962) Chapter 5, for the various topologies induced by classes of functions. Note that the class $\mathcal{B}_b$ induces the so-called τ-topology (see Groeneboom *et al.* (1979) and Gänssler (1971)) and that $M_{\mathcal{B}_b}$ is the whole space M.

Proposition 1.1. *Equip $M_{\mathcal{F}}$ with the $\tau_{\mathcal{F}}$-topology. Then, $M_{\mathcal{F}}$ is a Hausdorff locally convex topological linear space. Further, the topological dual space of $M_{\mathcal{F}}$ is the set of all mappings $Q \to \int f\ dQ$ when f belongs to $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$.*

In view of the above result, we identify the topological dual space of $M_{\mathcal{F}}$ with the linear span of $\mathcal{F} \cup \mathcal{B}_b$, that is with $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$.

We will prove that, for any ϕ-divergence defined as in (1.1), the mapping $Q \to \phi(Q, P)$ defined on $M_{\mathcal{F}}$ (equipped with the $\tau_{\mathcal{F}}$-topology) satisfies the conditions of the duality Lemma 1.1. We establish that the ϕ-divergence mapping $Q \to \phi(Q, P)$ is l.s.c. in the $\tau_{\mathcal{F}}$-topology. We first state

Lemma 1.2. *Let $M_{\mathcal{F}}(P)$ denote the subset of all signed measures in $M_{\mathcal{F}}$ absolutely continuous w.r.t. P. The set $M_{\mathcal{F}}(P)$ is a closed subspace in $(M_{\mathcal{F}}, \tau_{\mathcal{F}})$.*

We can now state the following

Proposition 1.2. *For any ϕ-divergence, the divergence function $Q \to \phi(Q, P)$ from $(M_{\mathcal{F}},\ \tau_{\mathcal{F}})$ onto $[0, +\infty]$ is l.s.c..*

1.2.1 Application of the duality Lemma

We now apply Lemma 1.1 when the space $\mathcal{S}$ is replaced by $M_{\mathcal{F}}$ and when the function g is defined by

$$\begin{array}{ccc} (M_{\mathcal{F}},\ \tau_{\mathcal{F}}) & \to & [0, +\infty] \\ Q & \to & g(Q) = \phi(Q, P), \end{array}$$

in which P is an arbitrary p.m.. Note that the hypotheses in Lemma 1.1 hold and that the topological dual space of $M_{\mathcal{F}}$ is one to one with $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$, following Proposition 1.1 and Proposition 1.2. Hence, the Fenchel-Legendre transform of $Q \to \phi(Q, P)$ is defined for any f in $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$ by

$$T(f, P) := \sup_{Q \in M_{\mathcal{F}}} \left\{ \int f\ dQ - \phi(Q, P) \right\}. \tag{1.8}$$

We thus state

Proposition 1.3. *For any measure Q in $M_{\mathcal{F}}$ and for any p.m. P, it holds*

$$\phi(Q, P) = \sup_{f \in \langle \mathcal{F} \cup \mathcal{B}_b \rangle} \left\{ \int f dQ - T(f, P) \right\}. \tag{1.9}$$

1.2.2 Calculation of the Fenchel-Legendre transform of ϕ-divergences

In this subsection, we give the explicit form of $T(f,P)$ defined in (1.8) under weak assumptions on the convex nonnegative function φ. In the sequel, we will consider only nonnegative strictly convex functions φ defined on $\mathbb{R}$ and which are $\mathcal{C}^2$ on the interior of its domain D_φ (see (1.6) for the definition of D_φ) and satisfying $\varphi(1)=0$; note that all the functions φ_γ (see (1.5)) satisfy these conditions.

Define $\varphi'(a)$, $\varphi''(a)$, $\varphi'(b)$ and $\varphi''(b)$ respectively by $\varphi'(a)=\lim_{x\downarrow a}\varphi'(x)$, $\varphi''(a)=\lim_{x\downarrow a}\varphi''(x)$, $\varphi'(b)=\lim_{x\uparrow b}\varphi'(x)$ and $\varphi''(b)=\lim_{x\uparrow b}\varphi''(x)$. These quantities may be finite or infinite.

Denote ${\varphi'}^{-1}$ the inverse function of φ' and $\text{Im }\varphi'$ the set of all values of φ'.

The convex conjugate (or Legendre-Fenchel transform) of φ will be denoted by ψ, i.e.,

$$t\in\mathbb{R} \;\mapsto\; \psi(t):=\sup_{x\in\mathbb{R}}\{tx-\varphi(x)\}\,. \tag{1.10}$$

Denote φ^* the function defined on $\text{Im }\varphi'$ by

$$t\in \text{ Im }\varphi' \;\mapsto\; \varphi^*(t):=t{\varphi'}^{-1}(t)-\varphi\left({\varphi'}^{-1}(t)\right). \tag{1.11}$$

Clearly, the function φ^* coincides on the set $\text{Im }\varphi'$ with ψ, the convex conjugate of φ.

For concepts about convex functions and properties of convex conjugates see e.g. Rockafellar (1970).

For all f in $\langle\mathcal{F}\cup\mathcal{B}_b\rangle$, define the mapping $G_f : M_\mathcal{F} \to]-\infty,+\infty]$ by

$$G_f(Q):=\phi(Q,P)-\int f\,dQ,$$

from which $T(f,P)=-\inf_{Q\in M_\mathcal{F}}G_f(Q)$. The function $G_f(.)$ is strictly convex. Its domain is

$$Dom(G_f):=\{Q\in M_\mathcal{F} \text{ such that } G_f(Q)<+\infty\}\,.$$

Denote, if it exists, $Q_0:=\arg\inf_{Q\in M_\mathcal{F}}G_f(Q)$, which belongs to $Dom(G_f)$. This implies that Q_0 is a.c. w.r.t. P. By Theorem III.31 in Azé (1997) (see also Luenberger (1969)), since $M_\mathcal{F}$ is convex set, the measure Q_0, if it exists, is the only measure in $Dom(G_f)$ such that for any measure R in $Dom(G_f)$, it holds

$$G'_f(Q_0,R-Q_0)\geq 0,$$

where $G'_f(Q_0, R-Q_0)$ is the derivative of the function G_f at point Q_0 in direction $R-Q_0$. Denote $r = \frac{dR}{dP}$ and $q_0 = \frac{dQ_0}{dP}$. By its very definition, we have

$$\begin{aligned} G'_f(Q_0, R-Q_0) &= \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \{G_f(Q_0 + \epsilon(R-Q_0)) - G_f(Q_0)\} \\ &= \lim_{\epsilon \downarrow 0} \int \frac{1}{\epsilon} \left[\varphi\left(q_0 + \epsilon(r-q_0)\right) - \varphi(q_0)\right] \, dP - \int f \, d(R-Q_0). \end{aligned}$$

Define the function

$$g(\epsilon) := \frac{1}{-\epsilon} \left[\varphi\left(q_0 + \epsilon(r-q_0)\right) - \varphi(q_0)\right].$$

Convexity of φ implies

$$g(\epsilon) \uparrow \varphi'(q_0)(q_0 - r) \quad \text{when} \quad \epsilon \downarrow 0,$$

and for all $0 < \epsilon \leq 1$ and R in $Dom(G_f)$, we have

$$g(\epsilon) \geq g(1) = -\left(\varphi(r) - \varphi(q_0)\right) \in L^1(P).$$

So, by application of the Monotone Convergence Theorem, we obtain

$$G'_f(Q_0, R-Q_0) = \int \left(\varphi'(q_0) - f\ \right) \, d(R-Q_0) \geq 0.$$

Therefore, for any function f in $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$, for which there exists a measure Q_0 in $M_{\mathcal{F}}$ with $\varphi'(dQ_0/dP) = f$ (P-a.s.), it holds $Q_0 = \arg\inf_{Q \in M_{\mathcal{F}}} G_f(Q)$. Note that since f is arbitrary in $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$, it is necessary in order for Q_0 to be properly defined that $\text{Im}\ \varphi' = \mathbb{R}$. It follows that

$$T(f, P) = \int \varphi^*(f) \, dP. \tag{1.12}$$

We see that Q_0 belongs to $M_{\mathcal{F}}$ iff for any g in $\mathcal{F} \cup \mathcal{B}_b$, $\int |g| \ d|Q_0|$ is finite. Further, $dQ_0 = \varphi'^{-1}(f) \, dP$. Hence, (1.12) holds whenever for any g in $\mathcal{F} \cup \mathcal{B}_b$, $\int |g| \left|\varphi'^{-1}(f)\right| \, dP$ is finite. The formula in (1.12) holds for any f in $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$ whenever for any couple of functions f and g in $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$, it holds

$$\int |g| \left|\varphi'^{-1}(f)\right| \, dP \quad \text{is finite.} \tag{1.13}$$

Under (1.13), we have, by Theorem (1.3)

$$\phi(Q, P) = \sup_{f \in \langle \mathcal{F} \cup \mathcal{B}_b \rangle} \left\{ \int f dQ - \int \varphi^*(f) \, dP \right\}. \tag{1.14}$$

Formula (1.12) holds under quite restrictive conditions. Typically, we have requested that $\text{Im}\ \varphi' = \mathbb{R}$ and that (1.13) holds. An important counter-example is the Hellinger divergence for which $\varphi(x) = 2(\sqrt{x}-1)^2$, hence $\text{Im}\ \varphi' = (-\infty, 2)$.

Hopefully, under weak assumptions we can derive a convenient expression for $\phi(Q,P)$, similar as in (1.14), even when (1.13) and the condition $\text{Im}\ \varphi' = \mathbb{R}$ do not hold. This is performed as follows : Let Q be a signed finite measure and P a p.m. with $\phi(Q,P)$ is finite. Assume that the class $\mathcal{F}$ be such that

(i) The mapping $\varphi'(dQ/dP)$ belongs to $\mathcal{F}$;

(ii) for any f in $\mathcal{F}$, $\int |f|\ d|Q|$ is finite;

(iii) for any f in $\mathcal{F}$, $\text{Im}\ f$ is included in $\text{Im}\ \varphi'$.

When (iii) holds, then the same holds, for all f in the convex hull of $\mathcal{F}$ ($\text{Convh}(\mathcal{F})$). Therefore, define

$$f \in \text{Convh}(\mathcal{F}) \to A(f,Q,P) := \int f\ dQ - \int \varphi^*(f)\ dP. \tag{1.15}$$

Introducing $\text{Convh}(\mathcal{F})$ allows for a characterization of the optimizer of the concave function $f \to A(f,Q,P)$ on the convex set $\text{Convh}(\mathcal{F})$, as quoted in Theorem III.31 in Azé (1997) (see also Luenberger (1969)). Furthermore, since φ is strictly convex and is $\mathcal{C}^2$, we can prove that the function $f \in \text{Convh}(\mathcal{F}) \to A(f,Q,P)$ is strictly concave, which implies that the supremum of $f \to A(f,Q,P)$ on $\text{Convh}(\mathcal{F})$, if it exists, is unique (P-a.s.). As above, using directional derivatives, we characterize the supremum and we obtain

$$\arg \sup_{f \in \text{Convh}(\mathcal{F})} A(f,Q,P) = \varphi'\left(\frac{dQ}{dP}\right).$$

Replacing f by $\varphi'(dQ/dP)$ in $A(.,Q,P)$, we obtain $\phi(Q,P)$. We summarize the above arguments as follows.

Theorem 1.1. *Assume that the function φ is strictly convex and is $\mathcal{C}^2$ on the interior of D_φ. Let Q be a signed finite measure and P a p.m. with $\phi(Q,P) < \infty$. Let $\mathcal{F}$ be a class of functions such that (i) Q belongs to $M_{\mathcal{F}}$, (ii) $\varphi'(dQ/dP)$ belongs to $\mathcal{F}$ and (iii) for any f in $\mathcal{F}$, $\text{Im}\ f$ is included in $\text{Im}\ \varphi'$. We then have*

(1) The divergence $\phi(Q,P)$ admits the "dual representation"

$$\phi(Q,P) = \sup_{f \in \mathcal{F}} \left\{ \int f\ dQ - \int \varphi^*(f)\ dP \right\}. \tag{1.16}$$

(2) The supremum in (1.16) is unique $(P-a.s)$ and is reached at $f = \varphi'(dQ/dP)$ $(P-a.s)$.

Remark 1.1. *The difference from (1.14) (or (1.9)) and (1.16) lays in the substitution of $f \in \langle \mathcal{F} \cup \mathcal{B}_b \rangle$ by $f \in \mathcal{F}$ in the sup. This will prove to be an important feature for statistical applications.*

Remark 1.2. *Consider the vector space M_p of all measures Q a.c. w.r.t. P with $\frac{dQ}{dP}$ belongs to $L_p(\mathcal{X}, P)$ $(1 \leq p \leq \infty)$, and the vector space F_q of all functions in $L_q(\mathcal{X}, P)$ with $(1 \leq q \leq \infty)$ and $1/p + 1/q = 1$. The vector spaces M_p and F_q are decomposable in sense of Rockafellar (1968). In this case, we can apply the Corollary to Theorem 2 in Rockafellar (1968) to obtain a representation similar to (1.9) on the spaces M_p and F_q. Furthermore, the explicit form of $T(f, P)$ is given by*

$$T(f, P) = \int \psi(f) \, dP \tag{1.17}$$

where ψ is the convex conjugate of φ.

Remark 1.3. *Various dual representations of convex functionals are given on general vector spaces (see e.g. Rockafellar (1968), Rockafellar (1971)), on L_p spaces $(1 \leq p \leq \infty)$ and on some vector measure spaces (see Borwein and Lewis (1991), Borwein and Lewis (1993), Gamboa and Gassiat (1997), Csiszár* et al. *(1999), Léonard (2001b), Léonard (2001a)). These representations are used in maximum entropy reconstruction, moment problem and some inverse problems.*

1.3 Parametric estimation and tests through minimum ϕ-divergence approach

We consider an identifiable parametric model $\{P_\theta : \theta \in \Theta\}$ defined on some measurable space $(\mathcal{X}, \mathcal{B})$ and Θ is some open set in $\mathbb{R}^d$. For notational clearness we write $\phi(\alpha, \theta)$ for $\phi(P_\alpha, P_\theta)$ for α and θ in Θ. We assume that for any θ in Θ, P_θ has density p_θ with respect to some dominating σ-finite measure λ, which can be either with countable support or not. Assume further that the support S of the measure P_θ does not depend upon θ. On the basis of an i.i.d. sample $X_1, ..., X_n$ with distribution P_{θ_0}, we intend to estimate θ_0 the true value of the parameter. We assume that for any α in Θ, the following condition holds

(C.0)
$$\int \left| \varphi' \left(\frac{p_\alpha(x)}{p_\theta(x)} \right) \right| \, dP_\alpha(x) < \infty, \text{ for any } \theta \in \Theta.$$

This condition is fulfilled if

$$\phi(\alpha, \theta) := \int \varphi \left(\frac{p_\alpha(x)}{p_\theta(x)} \right) \, dP_\theta(x) < \infty, \text{ for any } \theta \in \Theta, \tag{1.18}$$

and φ fulfills the condition of Lemma 8.7 in Liese and Vajda (1987) ; see Liese and Vajda (1987) Lemma 8.9.

For all $\alpha \in \Theta$, consider the class of functions $\mathcal{F}$ defined by

$$\mathcal{F} := \left\{ x \to \varphi' \left(\frac{p_\alpha(x)}{p_\theta(x)} \right),\ \theta \in \Theta \right\}.$$

By Theorem 2.1, when $(C.0)$ holds, we obtain

$$\phi(\alpha, \theta_0) = \sup_{f \in \mathcal{F}} \left\{ \int f\ dP_\alpha - \int \varphi^*(f)\ dP_{\theta_0} \right\},$$

i.e.,

$$\phi(\alpha, \theta_0) = \sup_{\theta \in \Theta} P_{\theta_0} m(\theta, \alpha), \tag{1.19}$$

with

$$m(\theta, \alpha)\ :\ x \to m(\theta, \alpha, x)$$

and

$$m(\theta, \alpha, x) := \int \varphi' \left(\frac{p_\alpha}{p_\theta} \right)\ dP_\alpha - \varphi^* \left(\varphi' \left(\frac{p_\alpha(x)}{p_\theta(x)} \right) \right).$$

Remark 1.4. *The function $\theta \to P_{\theta_0} m(\theta, \alpha)$ has a unique maximizer $\theta = \theta_0$; see Theorem 1.1 part 2.*

For all $\alpha \in \Theta$, define the class of estimates of θ_0, which we call "dual ϕ-divergence estimates" (DϕDE's), by

$$\widehat{\theta}_n(\alpha) := \arg \sup_{\theta \in \Theta} P_n m(\theta, \alpha). \tag{1.20}$$

For any α in Θ, the divergence $\phi(P_\alpha, P_{\theta_0})$ between P_α and P_{θ_0} can be estimated by

$$\widehat{\phi}_n(\alpha, \theta_0) := P_n m(\widehat{\theta}_n(\alpha), \alpha) = \sup_{\theta \in \Theta} P_n m(\theta, \alpha).$$

Further, we have

$$\inf_{\alpha \in \Theta} \phi(\alpha, \theta_0) = \phi(\theta_0, \theta_0) = 0.$$

The infimum in the above display is unique when φ is strictly convex on a neighborhood of 1, and it is achieved at $\alpha = \theta_0$. It follows that a natural definition of estimates of θ_0, which we call "minimum dual ϕ-divergence estimates" (MDϕDE's), is

$$\widehat{\alpha}_n := \arg \inf_{\alpha \in \Theta} \widehat{\phi}_n(\alpha, \theta_0) = \arg \inf_{\alpha \in \Theta} \sup_{\theta \in \Theta} P_n m(\theta, \alpha). \tag{1.21}$$

Remark 1.5. *(An other view at the MLE). The maximum likelihood estimate (MLE) belongs to this class of estimates. Indeed it is obtained when* $\varphi(x) = -\log x + x - 1$, *that is as the dual modified* KL*-divergence estimate or as the minimum dual modified* KL*-divergence estimate, i.e.,* $MLE = DKL_mDE = MDKL_mDE$. *Indeed, we then have* $P_n m(\theta, \alpha) = -\int \log\left(\frac{p_\alpha}{p_\theta}\right)\, dP_n$, *hence by definitions (1.20) and (1.21), we get*

$$\widehat{\theta}_n(\alpha) = \arg\sup_{\theta\in\Theta} -\int \log\left(\frac{p_\alpha}{p_\theta}\right)\, dP_n = \arg\sup_{\theta\in\Theta} \int \log(p_\theta)\, dP_n = MLE$$

independently upon α, *and*

$$\widehat{\alpha}_n = \arg\inf_{\alpha\in\Theta}\sup_{\theta\in\Theta} -\int \log\left(\frac{p_\alpha}{p_\theta}\right)\, dP_n = \arg\sup_{\alpha\in\Theta} \int \log(p_\alpha)\, dP_n = MLE.$$

So, the MLE is the estimate of θ_0 *that minimizes the estimate of the* KL_m*-divergence between the parametric model* $\{P_\theta,\ \theta \in \Theta\}$ *and the p.m.* P_{θ_0}.

1.3.1 The asymptotic behavior of the DϕDE's and $\widehat{\phi}_n(\alpha, \theta_0)$ for a given α in Θ

In this section, we state both weak and strong consistency of the estimates $\widehat{\theta}_n(\alpha)$ of θ_0, the true value of the parameter, as defined in (1.20). We also state their asymptotic normality and evaluate their limiting variance. The hypotheses handled here are similar to those used in van der Vaart (1998) Chapter 5, in the study of M-estimates. Notice that indeed for fixed α, $\widehat{\theta}_n(\alpha)$ are M-estimates.

Denote $\|\ .\ \|$ the Euclidian norm defined on $\mathbb{R}^d$, $m'(\theta, \alpha)$ the d-dimensional vector with entries $\frac{\partial}{\partial\theta_i} m(\theta, \alpha)$ and $m''(\theta, \alpha)$ the $d \times d$-matrix with entries $\frac{\partial^2}{\partial\theta_i \partial\theta_j} m(\theta, \alpha)$. In the sequel, we will assume that condition (C.0) holds, $\phi(\alpha, \theta_0) < \infty$ and that the estimate $\widehat{\theta}_n(\alpha)$ exists.

Consistency

Let Θ_α be the subset of Θ, defined by

$$\Theta_\alpha := \left\{\theta \in \Theta \text{ such that } \int \varphi^*\left(\varphi'\left(\frac{p_\alpha(x)}{p_\theta(x)}\right)\right)\, dP_{\theta_0}(x) < +\infty\right\},$$

and denote Θ_α^c the complementary of the subset Θ_α in the set Θ, i.e.,

$$\Theta_\alpha^c := \left\{\theta \in \Theta \text{ such that } \int \varphi^*\left(\varphi'\left(\frac{p_\alpha(x)}{p_\theta(x)}\right)\right)\, dP_{\theta_0}(x) = +\infty\right\}.$$

Note that Θ_α contains θ_0, if $\phi(\alpha,\theta_0) < \infty$. Define the estimates $\widetilde{\phi}_n(\alpha,\theta_0)$ and $\widetilde{\theta}_n(\alpha)$ by

$$\widetilde{\phi}_n(\alpha,\theta_0) := \sup_{\theta\in\Theta_\alpha} P_n m(\theta,\alpha),$$

$$\widetilde{\theta}_n(\alpha) := \arg\sup_{\theta\in\Theta_\alpha} P_n m(\theta,\alpha).$$

We will consider the following conditions

(C.1) $\sup_{\theta\in\Theta_\alpha} |P_n m(\theta,\alpha) - P_{\theta_0} m(\theta,\alpha)|$ converges to 0 a.s. (resp. in probability) ;

(C.2) for any positive ϵ, $\sup_{\{\theta\in\Theta\,:\,\|\theta-\theta_0\|\geq\epsilon\}} P_{\theta_0} m(\theta,\alpha) < P_{\theta_0} m(\theta_0,\alpha)$;

(C.3) there exist $M < 0$ and $n_0 > 0$ such that, for all $n \geq n_0$, $\sup_{\theta\in\Theta_\alpha^c} P_n m(\theta,\alpha) \leq M$.

Condition (C.2) means that the maximizer θ_0 of the function $\theta \to P_{\theta_0} m(\theta,\alpha)$ is isolated. This condition holds, for example, when the function $\theta \to P_{\theta_0} m(\theta,\alpha)$ is strictly concave. The condition (C.3) makes sense, since for all $\theta \in \Theta_\alpha^c$, $P_{\theta_0} m(\theta,\alpha) = -\infty$.

Proposition 1.4.

(i) Assume that conditions (C.1) and (C.3) hold. Then, the estimate $\widehat{\phi}_n(\alpha,\theta_0)$ converges a.s. (resp. in probability) to $\phi(\alpha,\theta_0)$.

(ii) Assume that (C.1), (C.2) and (C.3) hold. Then the estimate $\widehat{\theta}_n(\alpha)$ converges a.s. (resp. in probability) to θ_0.

Asymptotic distributions

Define the function

$$x \to g(\theta,\alpha,x) := \varphi'\left(\frac{p_\alpha(x)}{p_\theta(x)}\right) p_\alpha(x),$$

and denote I_{θ_0} the information matrix, i.e., the matrix defined by

$$I_{\theta_0} := \int \frac{\dot{p}_{\theta_0}\dot{p}_{\theta_0}^T}{p_{\theta_0}}\, d\lambda.$$

We will consider the following conditions.

(C.4) $\widehat{\theta}_n(\alpha)$ converges in probability to θ_0 ;

(C.5) the function φ is $\mathcal{C}^3$ on $(0,+\infty)$ and there exists a neighborhood $V(\theta_0)$ of θ_0 such that, for all θ in $V(\theta_0)$, the gradient $\dot{p}_\theta$ and the Hessian matrix $\ddot{p}_\theta$ of p_θ exist (λ-a.e.), the partial derivatives of order 1 of p_θ, and the partial derivatives of order 1 and 2 of $\theta \to g(\theta,\alpha,x)$ are dominated (λ-a.e.) by some λ-integrable functions ;

(C.6) the function $\theta \to m(\theta, \alpha)$ is $\mathcal{C}^3$ on a neighborhood $V(\theta_0)$ of θ_0 for all x, and all partial derivatives of order 3 of $\theta \to m(\theta, \alpha)$ are dominated on $V(\theta_0)$ by some $P_{\theta_0}-$integrable function $x \to H(x)$;

(C.7) $P_{\theta_0}\|m'(\theta_0, \alpha)\|^2$ is finite, and the matrix $P_{\theta_0}m''(\theta_0, \alpha, x)$ exists and is invertible.

Theorem 1.2. *Assume that conditions (C.4-7) hold. Then,*

(1) $\sqrt{n}\left(\widehat{\theta}_n(\alpha) - \theta_0\right)$ *converges in distribution to a centered normal variable with covariance matrix*

$$V = \left[-P_{\theta_0}m''(\theta_0, \alpha)\right]^{-1}\left[P_{\theta_0}m'(\theta_0, \alpha)m'(\theta_0, \alpha)^T\right]\left[-P_{\theta_0}m''(\theta_0, \alpha)\right]^{-1}. \quad (1.22)$$

If $\alpha = \theta_0$, *then*

$$-P_{\theta_0}m''(\theta_0, \alpha) = \frac{1}{\varphi''(1)}P_{\theta_0}m'(\theta_0, \alpha)m'(\theta_0, \alpha)^T, \quad (1.23)$$

and

$$V = \varphi''(1)[-P_{\theta_0}m''(\theta_0, \alpha)]^{-1} = I_{\theta_0}^{-1}. \quad (1.24)$$

(2) If $\theta_0 = \alpha$, *then the statistics* $\frac{2n}{\varphi''(1)}\widehat{\phi}_n(\alpha, \theta_0)$ *converge in distribution to a* χ^2 *variable with* d *degrees of freedom.*

(3) If $\theta_0 \neq \alpha$, *then* $\sqrt{n}\left(\widehat{\phi}_n(\alpha, \theta_0) - \phi(\alpha, \theta_0)\right)$ *converges in distribution to a centered normal variable with variance* $\sigma^2 = P_{\theta_0}m(\theta_0, \alpha)^2 - \left(P_{\theta_0}m(\theta_0, \alpha)\right)^2$.

Remark 1.6. *(Maximum likelihood ratio test and* KL_m-*divergence). Using Theorem 1.2, the estimates* $\widehat{\phi}_n(P_{\alpha_0}, P_{\theta_0})$ *can be used to perform tests of the hypothesis* $\mathcal{H}_0 : \theta_0 = \alpha_0$ *against the alternatives* $\mathcal{H}_1 : \theta_0 \neq \alpha_0$ *for some value* α_0. *Since* $\phi(\theta_0, \alpha_0)$ *is nonnegative and takes value* 0 *only when* $\theta_0 = \alpha_0$, *the tests are defined through the critical region*

$$C_\phi := \left\{\frac{2n}{\varphi''(1)}\widehat{\phi}_n(\alpha_0, \theta_0) > q_{(1-\epsilon)}\right\},$$

where $q_{(1-\epsilon)}$ *is the* $(1-\epsilon)$*-quantile of the* χ^2 *distribution with* d *degrees of freedom. Also these tests are all asymptotically of level* ϵ *and asymptotically powerful, since the estimates* $\widehat{\phi}_n(\alpha_0, \theta_0)$ *are* $n-$*consistent estimates of* $\phi(\theta_0, \alpha_0) = 0$ *under* $\mathcal{H}_0$ *and* $\sqrt{n}-$*consistent estimates of* $\phi(\theta_0, \alpha_0)$ *under* $\mathcal{H}_1$.

When $\varphi(x) = -\log x + x - 1$, *we obtain the critical area*

$$C_{KL_m} := \left\{2n\sup_{\theta\in\Theta}P_n\log\left(\frac{p_\theta}{p_{\alpha_0}}\right) > q_{(1-\epsilon)}\right\} = \left\{2\log\frac{\sup_{\theta\in\Theta}\prod_{i=1}^n p_\theta(X_i)}{\prod_{i=1}^n p_{\alpha_0}(X_i)} > q_{(1-\epsilon)}\right\},$$

which is to say that the test is precisely the maximum likelihood ratio test.

1.3.2 The asymptotic behavior of the MDϕDE's

In this Section, we state the strong and weak consistency of the estimates $\widehat{\theta}_n(\widehat{\alpha}_n)$ and the MDϕDE's $\widehat{\alpha}_n$ defined in (1.21). We also state their limiting distributions. We will assume that condition (C.0) is fulfilled, there exists a neighborhood $V(\theta_0)$ of θ_0 such that $\phi(\alpha, \theta_0) < \infty$ for all $\alpha \in V(\theta_0)$, and that both estimates $\widehat{\theta}_n(\widehat{\alpha}_n)$ and $\widehat{\alpha}_n$ exist.

Consistency

Define the estimate $\widetilde{\alpha}_n$ by

$$\widetilde{\alpha}_n := \arg \inf_{\alpha \in \Theta} \sup_{\theta \in \Theta_\alpha} P_n m(\theta, \alpha).$$

We state our results under the following conditions

(C.8) $\sup_{\{\alpha \in \Theta, \theta \in \Theta_\alpha\}} |P_n m(\theta, \alpha) - P_{\theta_0} m(\theta, \alpha)|$ tends to 0 a.s. (resp. in probability) ;

- (a) for any positive ϵ, there exists some positive η, such that for any θ in Θ_α with $\|\theta - \theta_0\| \geq \epsilon$ and for all $\alpha \in \Theta$, it holds $P_{\theta_0} m(\theta, \alpha) < P_{\theta_0} m(\theta_0, \alpha) - \eta$;
- (b) there exists a neighborhood of θ_0, say $V(\theta_0)$, such that for any positive ϵ, there exists some positive η such that for all $\theta \in V(\theta_0)$ and all $\alpha \in \Theta$ satisfying $\|\alpha - \theta_0\| \geq \epsilon$, it holds $P_{\theta_0} m(\theta, \theta_0) < P_{\theta_0} m(\theta, \alpha) - \eta$;

(C.10) there exists some neighborhood $V(\theta_0)$ of θ_0 and a positive function H such that for all θ in $V(\theta_0)$, $|m(\theta, \theta_0, x)| \leq H(x)$ (P_{θ_0}-a.s.) with $P_{\theta_0} H < \infty$;

(C.11) there exist $M < 0$ and $n_0 > 0$ such that, for all $n \geq n_0$, we have $\sup_{\alpha \in \Theta} \sup_{\theta \in \Theta_\alpha^c} P_n m(\theta, \alpha) < M$.

Proposition 1.5. *Assume that conditions (C.8-11) hold. Then,*

(1) $\sup_{\alpha \in \Theta} \|\widehat{\theta}_n(\alpha) - \theta_0\|$ *tends to* 0 *a.s. (resp. in probability).*

(2) The MDϕ estimates $\widehat{\alpha}_n$ *converge to* θ_0 *a.s. (resp. in probability).*

Asymptotic distributions

We will make use of the following conditions.

(C.12) Both estimates $\widehat{\alpha}_n$ and $\widehat{\theta}_n(\widehat{\alpha}_n)$ converge in probability to θ_0 ;

(C.13) the function φ is $\mathcal{C}^3$ on $(0, +\infty)$ and there exists a neighborhood of (θ_0, θ_0), say $V(\theta_0, \theta_0)$ such that, for all (θ, α) in $V(\theta_0, \theta_0)$, the gradient $\dot{p}_\theta$ and the Hessian matrix $\ddot{p}_\theta$ exist (λ-a.e.), the partial derivatives of order 1 of P_θ and the partial derivatives of order 1 and 2 of $(\theta, \alpha) \to g(\theta, \alpha, x)$ are dominated ($\lambda-$a.e.) by some $\lambda-$integrable functions ;

(C.14) the function $(\theta,\alpha)\to m(\theta,\alpha)$ is $\mathcal{C}^3$ on some neighborhood $V(\theta_0,\theta_0)$ of (θ_0,θ_0) for all x, and the partial derivatives of order 3 of $(\theta,\alpha)\to m(\theta,\alpha)$ are all dominated on $V(\theta_0,\theta_0)$ by some $P_{\theta_0}-$integrable function $H(x)$;

(C.15) $P_{\theta_0}\left\|\frac{\partial}{\partial\theta}m(\theta_0,\theta_0)\right\|^2$ and $P_{\theta_0}\left\|\frac{\partial}{\partial\alpha}m(\theta_0,\theta_0)\right\|^2$ are finite, and the Information matrix I_{θ_0} exists and is invertible.

Theorem 1.3. *Assume that conditions (C.12-15) hold. Then both* $\sqrt{n}\,(\widehat{\alpha}_n-\theta_0)$ *and* $\sqrt{n}\left(\widehat{\theta}_n(\widehat{\alpha}_n)-\theta_0\right)$ *converge in distribution to a centered normal variable with covariance matrix* $V=I_{\theta_0}^{-1}$.

1.3.3 Composite tests by minimum $\phi-$divergences

Let $h\ :\ \mathbb{R}^d\to\mathbb{R}^l$ be some function such that the $(d\times l)$-matrix $H(\theta):=\frac{\partial}{\partial\theta}h(\theta)$ exists, is continuous and has rank l with $0<l<d$. Let us define the composite null hypothesis

$$\Theta_0=\{\theta\in\Theta\ \text{ such that }\ h(\theta)=0\}.$$

Let us consider the composite test

$$\mathcal{H}_0\ :\ \theta_0\in\Theta_0\ \text{ versus }\ \mathcal{H}_1\ :\ \theta_0\in\Theta\backslash\Theta_0.$$

This test is equivalent to the following one

$$\mathcal{H}_0\ :\ \theta_0\in g(B_0)\ \text{ versus }\ \mathcal{H}_1\ :\ \theta_0\not\in g(B_0), \tag{1.25}$$

where $g\ :\ \mathbb{R}^{(d-l)}\to\mathbb{R}^d$ is a function such that the matrix $G(\beta):=\frac{\partial}{\partial\beta}g(\beta)$ exists and has rank $(d-l)$, and $B_0:=\left\{\beta\in\mathbb{R}^{(d-l)}\ \text{ such that } g(\beta)\in\Theta_0\right\}$. Therefore $\theta_0\in\Theta_0$ is an equivalent statement for $\theta_0=g(\beta_0),\beta_0\in B_0$.

Under $\mathcal{H}_0$, it holds $\inf_{\beta\in B_0}\phi\left(g(\beta),\theta_0\right)=\phi\left(g(\beta_0),\theta_0\right)=\phi\left(\theta_0,\theta_0\right)=0$ whereas under $\mathcal{H}_1$, $\inf_{\beta\in B_0}\phi\left(g(\beta),\theta_0\right)$ is positive. All ϕ-divergences $\inf_{\beta\in B_0}\phi\left(g(\beta),\theta_0\right)$ between the p.m. P_{θ_0} and the family of distributions $\left\{P_{g(\beta)}\ \text{ such that }\ \beta\in B_0\right\}$ can be estimated by the statistics

$$T_n^\phi:=\inf_{\beta\in B_0}\widehat{\phi}_n(g(\beta),\theta_0):=\inf_{\beta\in B_0}\sup_{\theta\in\Theta}P_nm(\theta,g(\beta)).$$

T_n^ϕ defined above is used in order to perform the tests pertaining to (1.25). Since $\inf_{\beta\in B_0}\phi\left(g(\beta),\theta_0\right)$ is nonnegative under $\mathcal{H}_1$ and takes value 0 only under $\mathcal{H}_0$, we reject $\mathcal{H}_1$ whenever T_n^ϕ takes large values. We obtain in this Section all the limiting distributions for T_n^ϕ under reasonable conditions and we show that all statistics $\frac{2n}{\varphi''(1)}T_n^\phi$ are asymptotically equivalent to the Wilks likelihood ratio statistic for composite hypotheses.

Remark 1.7. *It is to be emphasized that the peculiar choice* $\varphi(x) = -\log x + x - 1$ *yields the celebrated Wilks Likelihood ratio test for the composite null hypothesis, since* $T_n^{KL_m}$ *then writes* $2nT_n^{KL_m} = 2\log \frac{\sup_{\beta\in B_0}\prod p_{g(\beta)}(X_i)}{\sup_{\theta\in\Theta}\prod p_\theta(X_i)}$.

Theorem 1.4. *Let us assume that the conditions in Theorem 1.3 are satisfied. Under* $\mathcal{H}_0$, *the statistics* $\frac{2n}{\varphi''(1)}T_n^\phi$ *converge in distribution to a* χ^2 *r.v. with* $(d-l)$ *degrees of freedom.*

1.4 Statistical applications

In this section, we present two applications of ϕ-divergence techniques. In the first one, we will first define some new divergences based on the power divergences family ; we will see that within this family we are able to determine the best divergence associated to a parametric model and to the amount of information at disposal for the estimation of the parameter.

1.4.1 Confidence areas based on ϕ-divergences

Symmetric power divergences (s.p.d.'s)

For nonnegative γ define the function

$$\psi_\gamma(x) := \varphi_\gamma(x) + \varphi_{1-\gamma}(x), \tag{1.26}$$

with φ_γ defined as in (1.2) (see Figure 1.1). Define next the divergence associated to ψ_γ through

$$\phi_\gamma^s(Q,P) := \int \psi_\gamma\left(\frac{dQ}{dP}\right)\, dP \tag{1.27}$$

whenever Q is a.c. w.r.t. P and $+\infty$ otherwise. Call $\phi_\gamma^s(Q,P)$ the *symmetric power divergence* (s.p.d.) with exponent γ. For example, when $\gamma = 1/2$, then $\phi_\gamma^s(Q,P)$ is twice the Hellinger distance between Q and P ; when $\gamma = 0$, then it is $KL_m(Q,P) + KL(Q,P)$. Easy calculation based on the monotonicity property of the class ψ_γ with respect to γ (see Fig. 1.2) prove that the following basic properties hold.

Proposition 1.6.

(i) For all nonnegative γ *the divergences* $\phi_\gamma^s(Q,P)$ *are symmetric, i.e.,* $\phi_\gamma^s(Q,P) = \phi_\gamma^s(P,Q)$.

(ii) The symmetric power divergences $\phi_\gamma^s(Q,P)$ *enjoy the following monotonicity properties : For* $0 \le \gamma_1 \le \gamma_2 \le 1/2$, *we have* $\phi_{\gamma_1}^s(Q,P) \ge \phi_{\gamma_2}^s(Q,P)$, *and for* $1/2 \le \gamma_1 \le \gamma_2 < \infty$, *we have* $\phi_{\gamma_1}^s(Q,P) \le \phi_{\gamma_2}^s(Q,P)$.

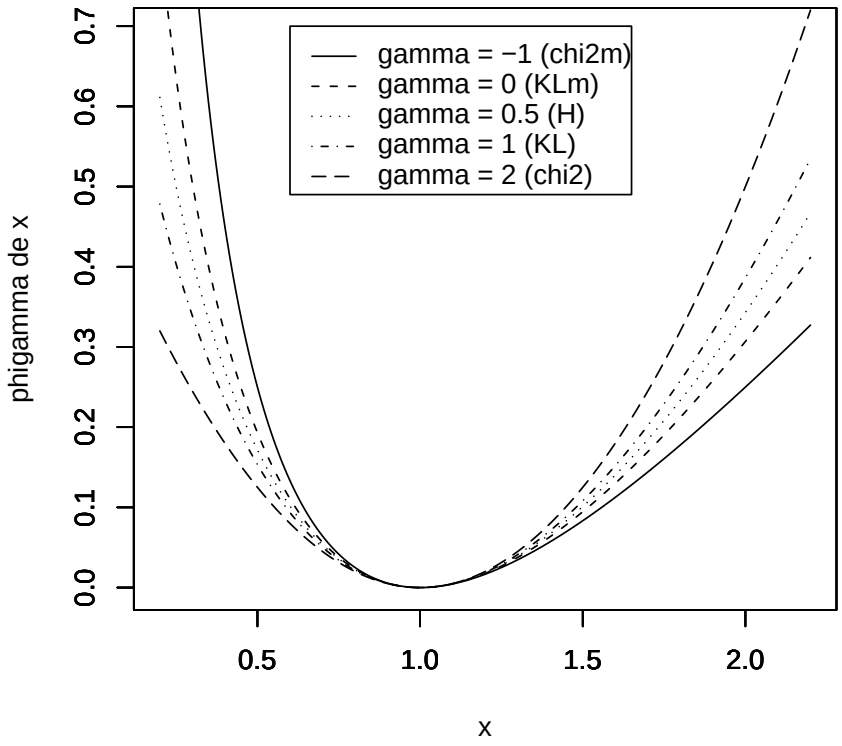

FIG. 1.1 – Divergence functions φ_γ.

Remark 1.8. *An other class of divergences can be defined through the functions* $\varphi_\gamma + \varphi_{-\gamma}$. *This class of divergences also has the monotonicity property and all divergences in this class are bounded below by twice the modified KL divergence. However, they are not generally symmetric in* P *and* Q.

Construction of small confidence regions

Recall from Theorem 1.2 part 2 that for any divergence the statistic $\frac{2n}{\varphi''(1)}\widehat{\phi}_n(\alpha_0, \theta_0)$ is asymptotically distributed as a χ^2 random variable with d degrees of freedom when $\mathcal{H}_0 \ : \ \theta_0 = \alpha_0$ holds. This enables the construction of confidence regions (CR's) for θ_0 with asymptotic level $(1-\epsilon)$. Define

$$I_n^\phi(\epsilon) := \left\{ \alpha_0 \in \Theta \text{ for which } \widehat{\phi}_n(\alpha_0, \theta_0) \leq \frac{\varphi''(1)}{2n} q_{(1-\epsilon)} \right\} \tag{1.28}$$

where $q_{(1-\epsilon)}$ is the $(1-\epsilon)$-quantile of a $\chi^2(d)$ distribution.

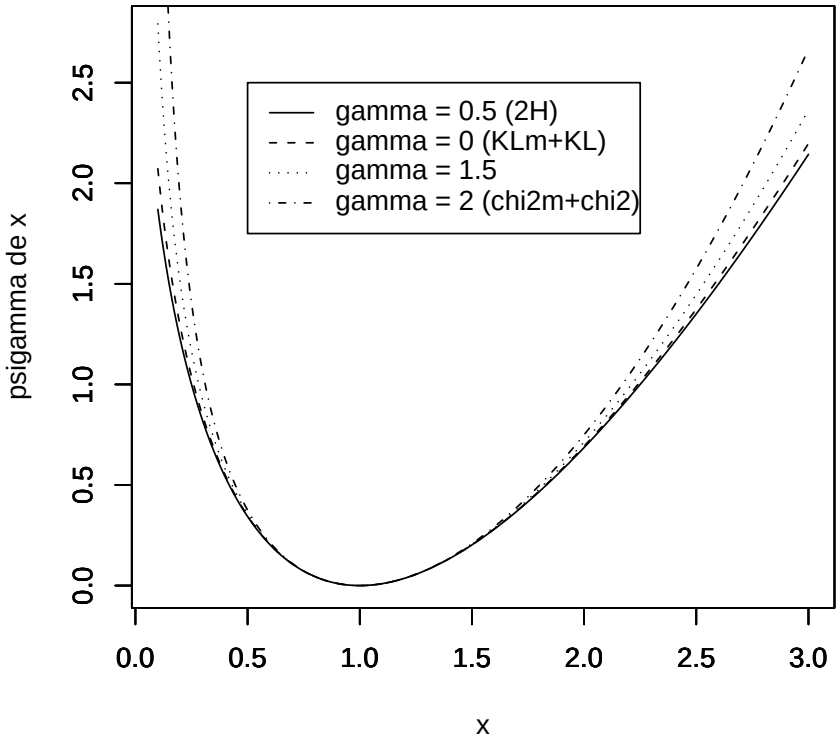

FIG. 1.2 – Divergence functions ψ_γ.

Therefore, $\lim_{n\to\infty} P_{\theta_0}\left\{\theta_0 \in I_n^\phi(\epsilon)\right\} = 1-\epsilon$, where θ_0 stands for the true value of the parameter. So, for any ϕ-divergence, $I_n^\phi(\epsilon)$ is a confidence region for the parameter θ_0 asymptotically of level $(1-\epsilon)$; a good CR should have minimum volume. Hence, screening various CR's according to ϕ allows to obtain some reasonable CR. Since $\widehat{\phi}_n(\alpha_0,\theta_0)$ estimates $\phi(\alpha_0,\theta_0)$, the volume of $I_n^\phi(\epsilon)$ is small when $\phi(\theta,\theta_0)$ is large when θ is close to θ_0. A good divergence should hence have large variations in a neighborhood of θ_0. This in turn can be expressed in terms of the φ function. Monotonicity is a major tool for the choice of the divergence. By Proposition 1.6, all divergences $\phi_\gamma^s(\alpha_0,\theta_0)$ can be compared ; the best one compatible with the model (i.e., $\phi_\gamma^s(\theta,\theta_0)<\infty$ for all θ in Θ) is such that the set of values θ for which $\phi_\gamma^s(\theta,\theta_0)$ is smaller than some bound is the smallest when θ_0 is fixed. The set Θ is obviously dependent upon some a priory knowledge on the unknown parameter θ_0. Not surprisingly we will see that the smallest Θ the smallest volume $I_n^\phi(\epsilon)$ for an adequate choice of the divergence ϕ_γ^s. At the contrary, when no information is available on the para-

meter, the likelihood method prevails, although not in the class of s.p.d.'s. This can be stated in the following way : assume that we know that the parameter θ_0 belongs to some interval $[a, b]$. Then the sup in $\widehat{KL_m}(\alpha, \theta_0) = \sup_{\theta\in[a,b]} - \int \log \frac{p_\alpha}{p_\theta}\, dP_n$ is reached at θ such that $\int \frac{\dot{p}_\theta}{p_\theta}\, dP_n = 0$, which does not make any use of $[a, b]$ for large n. On the other hand, for a symmetric power divergence $\widehat{\phi^s_\gamma}(\alpha, \theta_0) := \sup_{\theta\in[a,b]} P_n m(\theta_0, \alpha)$ depends explicitly upon $[a, b]$. Note that the choice $\gamma = 1/2$ yields to the Hellinger distance based CR, which is the largest among all those based on symmetric power divergences. As for the likelihood criterion, this choice does not make any use of any a priory knowledge on θ_0 since the Hellinger divergence is the lower bound for all s.p.d.'s.

We treat a basic example. The model consists of all exponential distributions on $\mathbb{R}_+$, defined through the density $p_\theta(x) := \theta e^{-\theta x} \mathbb{1}_{\mathbb{R}_+}(x)$. Let us write the conditions linking α, θ and γ in order that $\widehat{\phi^s_\gamma}(\alpha, \theta)$ estimates $\phi^s_\gamma(\alpha, \theta)$. Those are defined by $\phi^s_\gamma(\alpha, \theta) < \infty$; see (1.18).

For the present model, these conditions are

$$\frac{\gamma - 1}{\gamma}\alpha < \theta < \frac{\gamma}{\gamma - 1}\alpha \quad \text{if} \quad \gamma > 1. \tag{1.29}$$

When $0 \le \gamma \le 1$, then condition $\phi^s_\gamma(\alpha, \theta) < \infty$ always holds.

In practice, when some knowledge is available on θ_0, which occurs when the range Θ is not $\mathbb{R}^*_+$ but some strictly smaller set, we may find a small CR through a precise tuning of γ. For example, suppose θ_0 close to 1. Choose a priori interval $[\theta_1, \theta_2]$ centered in 1. For $\gamma > 1$, using the above conditions, we get $\theta_1 = \frac{2\gamma-2}{2\gamma-1}$ and $\theta_2 = \frac{\gamma}{\gamma-1}\theta_1$. Then, evaluate $\widehat{\phi^s_\gamma}(\theta, \theta_0) := \sup_{\theta\in[\theta_1,\theta_2]} P_n m(\theta_0, \theta)$ and define $I^\phi_n(\epsilon)$ as in (1.28). For instance, $\gamma = 2$ (symmetric χ^2) yields the CR with minimal length compatible with a sharp knowledge on θ_0 since $\theta_1 = 2/3$ and $\theta_2 = 4/3$ for the exponential model and $\gamma = 2$ is the largest value of γ for which condition $\phi^s_\gamma(\alpha, \theta) < \infty$ holds for all α and θ in $[\theta_1, \theta_2]$.

We present some empirical results. We have simulated a sample of $n = 100$ i.i.d. r.v.'s with common standard exponential distribution $(\theta_0 = 1)$; the value of the level ϵ is 0.05. The range for γ runs from 0 to 2. For each γ we evaluate θ_1 and θ_2 according to the condition (1.29) and we calculate $I^\phi_n(\epsilon)$, namely $[a_\gamma, b_\gamma]$. As a benchmark we have also calculated $[a_L, b_L] = [0.8729, 1.2923]$. the ML-based CR. The performance of the divergence approach is measured through the ratio $\frac{b_\gamma - a_\gamma}{b_L - a_L}$. Coverage probabilities $P_{\theta_0}[a_\gamma, b_\gamma]$ for those CR's have been calculated on a run of 1000 replications.

We observe that we can gain up to 8% on the ML-based CR (see Table 1.1) when the a priori knowledge on θ_0 is $\theta_0 \in [2/3, 4/3]$. Obviously, when γ gets larger and larger, the interval $[\theta_1, \theta_2]$ shrinks to θ_0 and the CR also tends to θ_0.

γ	0	0.25	0.5	1.25	1.5	2
$[\theta_1, \theta_2]$	$(0, \infty)$	$(0, \infty)$	$(0, \infty)$	$(1/3; 5/3)$	$(0.5; 1.5)$	$(2/3; 4/3)$
a_γ	0.8875	0.8856	0.8849	0.90	0.8961	0.9164
b_γ	1.2945	1.2934	1.2931	1.2962	1.2984	1.3043
$\frac{b_\gamma - a_\gamma}{b_L - a_L}$	0.97	0.9725	0.9732	0.9662	0.9593	0.9250
$P_{\theta_0}[a_\gamma, b_\gamma]$	0.943	0.953	0.951	0.933	0.934	0.926

TAB. 1.1 – Confidence interval for various ϕ^s_γ-divergences

Calculations based on divergences exposed in Remark 1.8 do not provide as good results as the above ones.

As a conclusion we indicate the modus operandi for the obtention of small CR's. Knowing $[\theta_1, \theta_2]$, the parameter set, select through (1.29) the largest value of γ such that $[\theta_1, \theta_2]$ is included in the set of parameters (α, θ) for which $\phi^s_\gamma(\alpha, \theta)$ is finite for all α, θ in $[\theta_1, \theta_2]$.

1.4.2 Multiple maxima of the likelihood surface

Multiple maxima of the likelihood surface arise quite commonly in the study of data generated by mixtures of distributions, when the components play nearly symmetric role. We consider a mixture of two normal distributions, where the only unknown parameters are the means of the components, and the weight of each of them is close to 1/2. So, $\theta := (\mu_1, \mu_2)$ and $p_\theta := p\mathcal{N}(\mu_1, 1) + (1-p)\mathcal{N}(\mu_2, 1)$ and $\mathcal{N}(\mu, \sigma)$ is the normal density with mean μ and standard deviation σ. Consider the likelihood surface $\theta \to L(\theta) := E(\log p_\theta(X))$ where X is a random variable with distribution p_{θ_0}. We consider two examples : **case 1 :** $p = 0.45$ and $\theta_0 = (2, 5)$. **case 2 :** $p = 0.45$ and $\theta_0 = (4, 5)$. In both cases, the likelihood surface has two local maxima, one at θ_0, the global maximum of L, and the second one close to $(5, 2)$ (in case 1) and to $(5, 4)$ (in the case 2). Also in case 2, the values of L at θ_0 and at $(5, 4)$ are quite close to each other. We first exhibit the two contour plots of the surface $L(\theta)$.

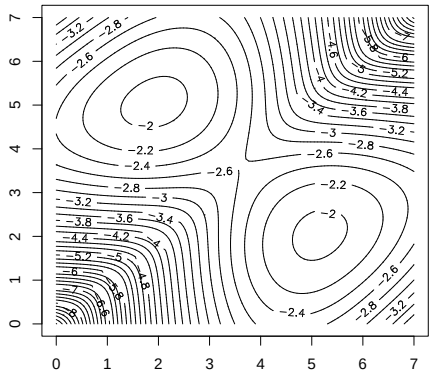

FIG. 1.3 – Contour plot of the Likelihood surface in case 1.

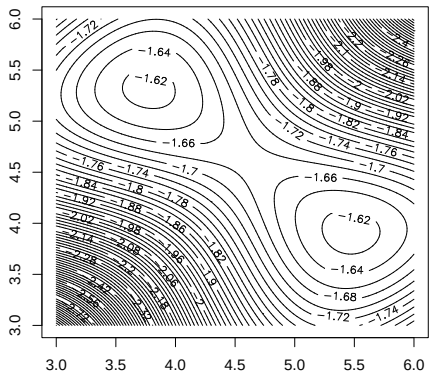

FIG. 1.4 – Contour plot of the Likelihood surface in case 2.

When L is estimated through $L_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(X_i)$ where the X_i' are i.i.d. with distribution p_{θ_0}, fluctuations due to sampling may lead to very small differences in the local levels of the empirical likelihood surface $L_n(\theta)$, as appears in cases 1 and 2, and even to an inversion effect of those maximal values, inducing erroneous estimates ; this appears in case 2. This problem is mentioned in the literature on ML methods (see Small *et al.* (2000) and the references therein), also, for a study of the asymptotic behavior of the sequence of relative maxima of the likelihood, see Fiorin (2000).

For each of cases 1 and 2, $L_n(\theta)$ also has two local maxima. Denote $\theta_n^{(1)}$ the global maximum of $L_n(\theta)$ and $\theta_n^{(2)}$ the second (local) one. Comparison of $p_{\theta_n^{(i)}}$ with respect to the empirical distribution of the sample does not provide a convenient tool for a proper choice of the estimate. In case 1, we obtain $\theta_n^{(1)} = (1.94, 5.11)$, $\theta_n^{(2)} = (5.17, 2.05)$ and $\Delta_L := -L_n(\theta_n^{(1)}) + L_n(\theta_n^{(2)}) = -0.0317$, while for case 2, $\theta_n^{(1)} = (4.86, 4.01)$, $\theta_n^{(2)} = (3.94, 4.76)$ showing the inversion of the estimates, and $\Delta_L = -0.0006$. This indicates that also the likelihood ratio does not allow for a proper choice of the estimate. The values of Δ_L are clearly too small in order to decide in favor of $\theta_n^{(1)}$ or $\theta_n^{(2)}$.

Following Remarks 1.5 and 1.6, we have $\Delta_L = \widehat{KL_m}(\theta_n^{(1)}, \theta_0) - \widehat{KL_m}(\theta_n^{(2)}, \theta_0)$. This suggests to consider some similar tool as Δ_L in order to discriminate between $\theta_n^{(1)}$ and $\theta_n^{(2)}$. Consider the class of power divergences with index γ as defined in 1.2, and define

$$\Delta_{\phi_\gamma} := \widehat{\phi_\gamma}(\theta_n^{(1)}, \theta_0) - \widehat{\phi}_\gamma(\theta_n^{(2)}, \theta_0).$$

Two cases may occur. In the first instance, for all divergence ϕ_γ, Δ_{ϕ_γ} is negative or close to 0. In this case, there is no reason to keep any doubt on the validity of $\theta_n^{(1)}$ as a valid estimate of θ_0. This is precisely our case 1 ; see Table 2. We also note that as γ gets more negative, the gap between $\theta_n^{(1)}$ and $\theta_n^{(2)}$ measured by Δ_{ϕ_γ} increases and argues in favor of $\theta_n^{(1)}$. In our second case, at the contrary, small negative values of Δ_{ϕ_γ} hold for quite large (negative) values of γ, but a change in sign occurs for $\gamma = -50$, and for $\gamma = -60$, the (positive) value of Δ_{ϕ_γ} exceeds $|\Delta_L|$ by a factor of 300 ; see Table 3. This leads us to consider $\theta_n^{(2)}$ as a proper estimate for θ_0. The argument above is obviously quite heuristic in nature. It seems that a good choice in order to check which of the estimates is bona fide is to select γ for which the mapping $\theta \to \phi_\gamma(\theta, \theta_0)$ has large variations in some domain containing $\theta_n^{(1)}$ and $\theta_n^{(2)}$. Furthermore, the rate of convergence of the estimates $\widehat{\phi_\gamma}$ should also play a role in this choice. We do not develop this here.

γ	-40	-30	-25	-20	-10	-5	-2
Δ_{ϕ_γ}	-3.530	-0.6641	-0.3033	-0.1461	-0.0444	-0.0305	-0.0272

γ	$-1(\chi^2)$	$0.5(H)$	$1(KL)$	$2(\chi^2_m)$
Δ_{ϕ_γ}	-0.0269	-0.0281	-0.0297	-0.0447

TAB. 1.2 – Case 1 : $n = 50$, $\theta_n^{(1)} = (1.94, 5.11)$, $\theta_n^{(2)} = (5.17, 2.05)$.

Obviously, a good solution for the estimation in such models turns out to select the estimate associated to the divergence whose variations are the greatest at the neighborhood of the true value of the parameter among the ones compatible with

γ	-60	-50	-40	-30	-20	-10	-5
Δ_{ϕ_γ}	0.1391	0.0051	-0.0057	-0.0047	-0.0632	-0.0009	-0.0002

γ	$-1(\chi^2)$	$0.5(H)$	$1(KL)$
Δ_{ϕ_γ}	-0.0007	-0.0007	-0.0007

TAB. 1.3 – Case 2 : $n = 50$, $\theta_n^{(1)} = (4.86, 4.01)$, $\theta_n^{(2)} = (3.94, 4.76)$.

the model ; this method yields to a calculation similar to the one developed above in order to determine the range of parameters θ and θ' for which $\phi_\gamma(\theta, \theta')$ is finite, where the exponential model is changed into the mixture one. Such calculation is unfortunately very complex. The advantage of the likelihood technique lies precisely in its universality with respect to the model. Divergences may then help to deal with its defaults.

1.5 Proofs

1.5.1 Proof of Proposition 1.1

By Lemma 3 in Dunford and Schwartz (1962) Chapter 5 Section 3, the linear space $M_{\mathcal{F}}$ equipped with the $\tau_{\mathcal{F}}$-topology is a Hausdorff locally convex topological space. The set of all mappings $Q \to \int f \, dQ$ when f belongs to $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$ is a total linear space ; indeed, assume that $\int f \, dQ = 0$ for all f in $< \mathcal{F} \cup \mathcal{B}_b >$, choose $f = \mathbb{1}_{\{B\}}$ for any $B \in \mathcal{B}$ to conclude that $Q = 0$. The proof ends then as a consequence of Theorem 9 in Dunford and Schwartz (1962) Chapter 5 Section 3.

1.5.2 Proof of Lemma 1.2

Let $\overline{M_{\mathcal{F}}(P)}$ denote the closure of $M_{\mathcal{F}}(P)$ in $(M_{\mathcal{F}}, \tau_{\mathcal{F}})$. Assume that there exists R in $\overline{M_{\mathcal{F}}(P)}$ with R not in $M_{\mathcal{F}}(P)$. Then, there exists some B in $\mathcal{B}$ such that $P(B) = 0$ and $R(B) \neq 0$. On the other hand, for all n in $\mathbb{N}$, the set $U := U\left(R, \mathbb{1}_{\{B\}}, 1/n\right)$ is a neighborhood of R (see (1.7)), hence, $U \cap M_{\mathcal{F}}(P)$ is non void. Therefore, we can construct a sequence of measures R_n in $M_{\mathcal{F}}(P)$ such that

$$\left| \int \mathbb{1}_{\{B\}} \, dR - \int \mathbb{1}_{\{B\}} \, dR_n \right| < 1/n.$$

Since $R_n(B) = 0$ for all n in $\mathbb{N}$, we deduce that $R(B) = 0$, a contradiction. This implies that $\overline{M_{\mathcal{F}}(P)} = M_{\mathcal{F}}(P)$, that is $M_{\mathcal{F}}(P)$ is closed in $(M_{\mathcal{F}}, \ \tau_{\mathcal{F}})$. This concludes the proof of Lemma 1.2.

1.5.3 Proof of Proposition 1.2

Let α be a real number. We prove that the set

$$A(\alpha) := \{Q \in M_{\mathcal{F}} \text{ such that } \phi(Q,P) \leq \alpha\}$$

is closed in $(M_{\mathcal{F}}, \tau_{\mathcal{F}})$. By Lemma 1.2, $M_{\mathcal{F}}(P)$ is closed in $(M_{\mathcal{F}}, \tau_{\mathcal{F}})$. Since $A(\alpha)$ is included in $(M_{\mathcal{F}}(P), \tau_{\mathcal{F}})$, we have to prove that $A(\alpha)$ is closed in the subspace $(M_{\mathcal{F}}(P), \tau_{\mathcal{F}})$. Let

$$B(\alpha) := \left\{ f \in L^1(P) \text{ such that } \int \varphi(f(x))\, dP(x) \leq \alpha \right\}.$$

$B(\alpha)$ is a convex set, since φ is a convex function. Furthermore, $B(\alpha)$ is closed in $L^1(P)$. Indeed, let f_n be a sequence in $B(\alpha)$ with $\lim_{n\to\infty} f_n = f^*$, where the limit is intended in $L^1(P)$. Hence, there exists a subsequence f_{n_k} which converges to f^* (P-a.s.). The functions $\varphi(f_{n_k})$ are nonnegative. Further, φ is continuous as a convex function. Therefore, Fatou's Lemma implies

$$\int \varphi(f^*)\, dP \leq \int \liminf_{k\to+\infty} \varphi(f_{n_k})\, dP \leq \liminf_{k\to+\infty} \int \varphi(f_{n_k})\, dP \leq \alpha,$$

which is to say that f^* belongs to $B(\alpha)$. Hence, $B(\alpha)$ is a closed subset in $L^1(P)$. By Theorem 13 in Dunford and Schwartz (1962) Chapter 5 Section 3, $B(\alpha)$ is weakly closed in $L^1(P)$, as a convex set. Denote W the weak topology on $L^1(P)$ and consider the mapping H defined by

$$\begin{array}{rccc} H & : & (M_{\mathcal{F}}(P), \tau_{\mathcal{F}}) & \longrightarrow & (L^1(P), W) \\ & & Q & \longrightarrow & H(Q) = dQ/dP. \end{array}$$

Let us prove that H is weakly continuous, that is $Q \to \int H(Q) g\, dP$ is a continuous mapping for all g in $L^\infty(P)$. Indeed, let g be some function in $L^\infty(P)$. Then, we have

$$\int H(Q) g\, dP = \int (dQ/dP) g\, dP = \int g\, dQ.$$

The mapping $Q \to \int g\, dQ$ is $\tau_{\mathcal{F}}$-continuous; indeed, for all g in $L^\infty(P)$, it holds $P(g > \|g\|_\infty) = 0$, which implies $Q(g > \|g\|_\infty) = 0$, for all Q in $M_{\mathcal{F}}(P)$. Therefore, $\int g\, dQ = \int g \mathbb{1}_{[g \leq \|g\|_\infty]}\, dQ$. Now, the mapping $Q \to \int g \mathbb{1}_{[g \leq \|g\|_\infty]}\, dQ$ is continuous in $\tau_{\mathcal{F}}$-topology since $g 1_{[g \leq \|g\|_\infty]} \in \mathcal{F} \cup \mathcal{B}_b$.
Since $A(\alpha) = \{Q \in M_{\mathcal{F}}(P),\ \phi(Q,P) \leq \alpha\} = H^{-1}(B(\alpha))$, we deduce that $A(\alpha)$ is closed in $(M_{\mathcal{F}}(P), \tau_{\mathcal{F}})$, for any α in $\mathbb{R}$. This proves Proposition 1.2.

1.5.4 Proof of Proposition 1.4

We will prove that, under conditions (C.1) and (C.2), the estimates $\widetilde{\phi}_n(\alpha, \theta_0)$ and $\widetilde{\theta}_n(\alpha)$ are consistent which by condition (C.3), implies that for all n sufficiently large, we have $\widetilde{\phi}_n(\alpha, \theta_0) = \widehat{\phi}_n(\alpha, \theta_0)$ and $\widetilde{\theta}_n(\alpha) = \widehat{\theta}_n(\alpha)$.

We prove the consistency of the estimate $\widetilde{\phi}_n(\alpha, \theta_0)$. For the consistency of $\widetilde{\theta}_n(\alpha)$, we refer to (van der Vaart (1998) Theorem 5.7).

We have

$$\left|\widetilde{\phi}_n(\alpha, \theta_0) - \phi(\alpha, \theta_0)\right| = \left|P_n m(\widetilde{\theta}_n(\alpha), \alpha) - P_{\theta_0} m(\theta_0, \alpha)\right| := |A|,$$

which implies

$$P_n m(\theta_0, \alpha) - P_{\theta_0} m(\theta_0, \alpha) \leq A \leq P_n m(\widetilde{\theta}_n(\alpha), \alpha) - P_{\theta_0} m(\widetilde{\theta}_n(\alpha), \alpha).$$

Both the RHS and the LHS terms in the above display go to 0, under condition (C.1). This implies that A tends to 0, which concludes the proof.

1.5.5 Proof of Theorem 1.2

Proof of (1). Under condition (C.5), derivating under the integral sign, we get

$$P_{\theta_0} m'(\theta_0, \alpha) = 0 \tag{1.30}$$

and

$$P_{\theta_0} m''(\theta_0, \alpha) = -\varphi''(1) \int \frac{\dot{p}_{\theta_0} \dot{p}_{\theta_0}^T}{p_{\theta_0}} \, d\lambda, \tag{1.31}$$

which implies that the matrix $P_{\theta_0} m''(\theta_0, \alpha)$ is symmetrical. By Taylor expansion, there exists $\widetilde{\theta}_n$ inside the segment that links θ_0 and $\widehat{\theta}_n(\alpha)$ with

$$\begin{aligned} 0 &= P_n m'(\widehat{\theta}_n(\alpha), \alpha) \\ &= P_n m'(\theta_0, \alpha) + \left(P_n m''(\theta_0, \alpha)\right)^T (\widehat{\theta}_n(\alpha) - \theta_0) \\ &\quad + \frac{1}{2} (\widehat{\theta}_n(\alpha) - \theta_0)^T P_n m'''(\widetilde{\theta}_n, \alpha)(\widehat{\theta}_n(\alpha) - \theta_0). \end{aligned} \tag{1.32}$$

In (1.32), $P_n m'''(\widetilde{\theta}_n, \alpha)$ is a $d-$vector whose entries are $d \times d$-matrices. By (C.6), we have, for the sup-norm of vectors and matrices

$$\left\| P_n m'''(\widetilde{\theta}_n, \alpha) \right\| = \left\| \frac{1}{n} \sum_{i=1}^{n} m'''(\widetilde{\theta}_n, \alpha)(X_i) \right\| \leq \frac{1}{n} \sum_{i=1}^{n} |H(X_i)|.$$

By the Law of Large Numbers, $P_n m'''(\widetilde{\theta}_n, \alpha) = O_P(1)$. So using (C.4), we can write the last term in the RHS of (1.32) as a $o_P(1)(\widehat{\theta}_n(\alpha) - \theta_0)$. On the other hand by

(C.7), $P_n m''(\theta_0,\alpha) := \frac{1}{n}\sum_{i=1}^n m''(\theta_0,\alpha,X_i)$ converges to the matrix $P_{\theta_0} m''(\theta_0,\alpha)$. Write $P_n m''(\theta_0,\alpha)$ as $P_{\theta_0} m''(\theta_0,\alpha) + o_P(1)$ to obtain from (1.32)

$$-P_n m'(\theta_0,\alpha) = \left(P_{\theta_0} m''(\theta_0,\alpha) + o_P(1)\right)\left(\widehat{\theta}_n(\alpha) - \theta_0\right). \tag{1.33}$$

Under (C.7), by the Central Limit Theorem, we have $\sqrt{n} P_n m'(\theta_0,\alpha) = O_p(1)$, which by (1.33), implies $\sqrt{n}\left(\widehat{\theta}_n(\alpha) - \theta_0\right) = O_p(1)$. Hence, from (1.33), we get

$$\sqrt{n}\left(\widehat{\theta}_n(\alpha) - \theta_0\right) = \left[-P_{\theta_0} m''(\theta_0,\alpha)\right]^{-1}\sqrt{n} P_n m'(\theta_0,\alpha) + o_P(1). \tag{1.34}$$

Under (C.7), the Central Limit Theorem concludes the proof of part (1). In the case when $\alpha = \theta_0$, a simple calculation yields (1.23) and (1.24).

Proof of (2). By Taylor expansion, there exists $\overline{\theta}_n$ inside the segment that links θ_0 and $\widehat{\theta}_n(\alpha)$ with

$$\begin{aligned}
\widehat{\phi}_n(\alpha,\theta_0) &= P_n m(\widehat{\theta}_n(\alpha),\alpha) \\
&= P_n m(\theta_0,\alpha) + \left(P_n m'(\theta_0,\alpha)\right)^T\left(\widehat{\theta}_n(\alpha) - \theta_0\right) \\
&\quad + \frac{1}{2}\left(\widehat{\theta}_n(\alpha) - \theta_0\right)^T P_n m''(\theta_0,\alpha)\left(\widehat{\theta}_n(\alpha) - \theta_0\right) \\
&\quad + \frac{1}{3!}\sum_{1\le i,j,k\le d}\left(\widehat{\theta}_n(\alpha) - \theta_0\right)_i\left(\widehat{\theta}_n(\alpha) - \theta_0\right)_j \times \\
&\quad \times\left(\widehat{\theta}_n(\alpha) - \theta_0\right)_k P_n \frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k} m(\overline{\theta}_n,\alpha).
\end{aligned} \tag{1.35}$$

When $\alpha = \theta_0$, we have $P_n m(\theta_0,\alpha,x) = 0$. Furthermore, by part (1) in Theorem 1.2, it holds $\sqrt{n}(\widehat{\theta}_n(\alpha) - \theta_0) = O_p(1)$. Hence, by (C.4), (C.6) and (C.7), we get

$$\begin{aligned}
\widehat{\phi}_n(\alpha,\theta_0) &= \left(P_n m'(\theta_0,\alpha)\right)^T\left(\widehat{\theta}_n(\alpha) - \theta_0\right) + \\
&\quad \frac{1}{2}\left(\widehat{\theta}_n(\alpha) - \theta_0\right)^T P_{\theta_0} m''(\theta_0,\alpha)\left(\widehat{\theta}_n(\alpha) - \theta_0\right) + o_P(1/n),
\end{aligned} \tag{1.36}$$

which by (1.34), implies

$$\begin{aligned}
\widehat{\phi}_n(\alpha,\theta_0) &= \left[P_n m'(\theta_0,\alpha)\right]^T\left[-P_{\theta_0} m''(\theta_0,\alpha)\right]^{-1}\left[P_n m'(\theta_0,\alpha)\right] + \\
&\quad \frac{1}{2}\left[P_n m'(\theta_0,\alpha)\right]^T\left[P_{\theta_0} m''(\theta_0,\alpha)\right]^{-1}\left[P_n m'(\theta_0,\alpha)\right] + o_P(1/n) \\
&= \frac{1}{2}\left[P_n m'(\theta_0,\alpha)\right]^T\left[-P_{\theta_0} m''(\theta_0,\alpha)\right]^{-1}\left[P_n m'(\theta_0,\alpha)\right] + o_P(1/n).
\end{aligned}$$

This yields to

$$\frac{2n}{\varphi''(1)}\widehat{\phi}_n(\alpha,\theta_0) = \left[\sqrt{n}P_n m'(\theta_0,\alpha)\right]^T [-\varphi''(1)P_{\theta_0} m''(\theta_0,\alpha)]^{-1} \left[\sqrt{n}P_n m'(\theta_0,\alpha)\right] + o_P(1). \tag{1.37}$$

Note that when $\alpha = \theta_0$, calculation yields

$$P_{\theta_0} m'(\theta_0,\alpha) m'(\theta_0,\alpha)^T = -\varphi''(1)P_{\theta_0} m''(\theta_0,\alpha) \text{ and } P_{\theta_0} m'(\theta_0,\alpha) = 0.$$

This implies that the centered r.v's $m'(\theta_0,\alpha,X_i)$ are i.i.d. with variance matrix $(-\varphi''(1)P_{\theta_0} m''(\theta_0,\alpha))$. Combining this with (3.100), we conclude the proof of part (2).

Proof of (3). From (1.36), we can write

$$\begin{aligned} \widehat{\phi}_n(\alpha,\theta_0) &= P_n m(\widehat{\theta}_n(\alpha),\alpha) \\ &= P_n m(\theta_0,\alpha) + (P_n m'(\theta_0,\alpha))^T \left(\widehat{\theta}_n(\alpha) - \theta_0\right) + o_P(\delta_n), \end{aligned} \tag{1.38}$$

where $\delta_n = \left\|\widehat{\theta}_n(\alpha) - \theta_0\right\|$. Using the fact that $\delta_n = O_P(1/\sqrt{n})$ and $P_n m'(\theta_0,\alpha) = P_{\theta_0} m'(\theta_0,\alpha) + o_P(1) = 0 + o_P(1) = o_P(1)$, we obtain from (1.38)

$$\begin{aligned} \sqrt{n}\left(\widehat{\phi}_n(\alpha,\theta_0) - \phi(\alpha,\theta_0)\right) &= \sqrt{n}\left(P_n m(\theta_0,\alpha) - \phi(\alpha,\theta_0)\right) + o_P(1) \\ &= \sqrt{n}\left(P_n m(\theta_0,\alpha) - P_{\theta_0} m(\theta_0,\alpha)\right) + o_P(1), \end{aligned}$$

and the Central Limit Theorem yields to the conclusion of the proof.

1.5.6 Proof of Proposition 1.5

We prove (1). For all $\alpha \in \Theta$, under condition (C.8-10), we prove that $\sup_{\alpha\in\Theta} \|\widetilde{\theta}_n(\alpha) - \theta_0\|$ tends to 0. By the very definition of $\widetilde{\theta}_n(\alpha)$ and the condition (C.8), we have

$$\begin{aligned} P_n m(\widetilde{\theta}_n(\alpha),\alpha) &\geq P_n m(\theta_0,\alpha) \\ &\geq P_{\theta_0} m(\theta_0,\alpha) - o_p(1), \end{aligned}$$

where $o_p(1)$ does not depend upon α (due to condition (C.8)). Hence, we have for all $\alpha \in \Theta$

$$P_{\theta_0} m(\theta_0,\alpha) - P_{\theta_0} m(\widetilde{\theta}_n(\alpha),\alpha) \leq P_n m(\widetilde{\theta}_n(\alpha),\alpha) - P_{\theta_0} m(\widetilde{\theta}_n(\alpha),\alpha) + o_p(1), \tag{1.39}$$

The term in the RHS is less than $\sup_{\{\alpha\in\Theta,\theta\in\Theta_\alpha\}} |P_n m(\theta,\alpha) - P_{\theta_0} m(\theta,\alpha)| + o_p(1)$ which, by (C.8), tends to 0. Let $\epsilon > 0$ be such that $\sup_{\alpha\in\Theta} \|\widetilde{\theta}_n(\alpha) - \theta_0\| > \epsilon$. There

exists some $a_n \in \Theta$ such that $\|\widetilde{\theta}_n(a_n) - \theta_0\| > \epsilon$. Together with $(C.9)(a)$, there exists some $\eta > 0$ such that $P_{\theta_0} m(\theta_0, a_n) - P_{\theta_0} m(\widetilde{\theta}_n(a_n), a_n) > \eta$. We then conclude that

$$P\left\{\sup_{\alpha\in\Theta} \|\widetilde{\theta}_n(\alpha) - \theta_0\| > \epsilon\right\} \leq P\left\{P_{\theta_0} m(\theta_0, a_n) - P_{\theta_0} m(\widetilde{\theta}_n(\alpha), a_n) > \eta\right\},$$

and the RHS term tends to 0 by (1.39). Conditions (C.8) and (C.11) imply that for all n sufficiently large, we have $\widetilde{\theta}_n(\alpha) = \widehat{\theta}_n(\alpha)$, for all α, $\sup_{\{\theta\in\Theta_\alpha\}} P_n m(\theta, \alpha) = \sup_{\{\theta\in\Theta\}} P_n m(\theta, \alpha)$ and $\widetilde{\alpha}_n = \widehat{\alpha}_n$, which concludes the proof of part (1).

We prove (2). From the proof of part (1), it is sufficient to prove that $\widetilde{\alpha}_n$ tends to θ_0. By the very definition of $\widetilde{\alpha}_n$, condition (C.10) and part (1) of Proposition 1.5, we have

$$\begin{aligned} P_n m(\widetilde{\theta}_n(\widetilde{\alpha}_n), \widehat{\alpha}_n) &\leq P_n m(\widetilde{\theta}_n(\theta_0), \theta_0) \\ &\leq P_{\theta_0} m(\widetilde{\theta}_n(\widetilde{\alpha}_n), \theta_0) - o_p(1), \end{aligned}$$

from which

$$\begin{aligned} P_{\theta_0} m(\widetilde{\theta}_n(\widetilde{\alpha}_n), \widetilde{\alpha}_n) - P_{\theta_0} m(\widetilde{\theta}_n(\widetilde{\alpha}_n), \theta_0) &\leq P_{\theta_0} m(\widetilde{\theta}_n(\widetilde{\alpha}_n), \widetilde{\alpha}_n) - P_n m(\widetilde{\theta}_n(\widetilde{\alpha}_n), \widetilde{\alpha}_n) + o_p(1) \\ &\leq \sup_{\{\alpha\in\Theta, \theta\in\Theta_\alpha\}} |P_n m(\theta, \alpha) - P_{\theta_0} m(\theta, \alpha)| + o_p(1). \end{aligned} \tag{1.40}$$

Further, by (1) and condition (C.9)(b), for any positive ϵ, there exists $\eta > 0$ such that

$$P\left\{\|\widetilde{\alpha}_n - \theta_0\| > \epsilon\right\} \leq P\left\{P_{\theta_0} m(\widetilde{\theta}_n(\widetilde{\alpha}_n), \widetilde{\alpha}_n) - P_{\theta_0} m(\widetilde{\theta}_n(\widetilde{\alpha}_n), \theta_0) > \eta\right\},$$

and the RHS term, under condition (C.8), tends to 0 by (1.40). This concludes the proof.

1.5.7 Proof of Theorem 1.3

Derivation under the integral sign using (C.13) and some calculus yield

$$P_{\theta_0} \frac{\partial}{\partial\theta} m(\theta_0, \theta_0) = P_{\theta_0} \frac{\partial}{\partial\alpha} m(\theta_0, \theta_0) = P_{\theta_0} \frac{\partial^2}{\partial\alpha\partial\theta} m(\theta_0, \theta_0) = P_{\theta_0} \frac{\partial^2}{\partial\theta\partial\alpha} m(\theta_0, \theta_0) = 0,$$

$$P_{\theta_0} \frac{\partial^2}{\partial\alpha\partial\alpha} m(\theta_0, \theta_0) = -P_{\theta_0} \frac{\partial^2}{\partial\theta\partial\theta} m(\theta_0, \theta_0) = \varphi''(1) I_{\theta_0}, \tag{1.41}$$

and

$$\begin{aligned} P_{\theta_0}\left[\frac{\partial}{\partial\theta}m(\theta_0,\theta_0)\right]\left[\frac{\partial}{\partial\theta}m(\theta_0,\theta_0)\right]^T &= P_{\theta_0}\left[\frac{\partial}{\partial\alpha}m(\theta_0,\theta_0)\right]\left[\frac{\partial}{\partial\alpha}m(\theta_0,\theta_0)\right]^T \\ &= -P_{\theta_0}\left[\frac{\partial}{\partial\alpha}m(\theta_0,\theta_0)\right]\left[\frac{\partial}{\partial\theta}m(\theta_0,\theta_0)\right]^T \\ &= \varphi''(1)^2 I_{\theta_0}. \end{aligned}$$

By the very definition of $\widehat{\alpha}_n$ and $\widehat{\theta}_n(\widehat{\alpha}_n)$, they both obey

$$\left\{\begin{array}{rcl} P_n\frac{\partial}{\partial\theta}m(\theta,\alpha) &=& 0 \\ P_n\frac{\partial}{\partial\alpha}m(\theta(\alpha),\alpha) &=& 0, \end{array}\right.$$

i.e.,

$$\left\{\begin{array}{rcl} P_n\frac{\partial}{\partial\theta}m(\widehat{\theta}_n(\widehat{\alpha}_n),\widehat{\alpha}_n) &=& 0 \\ P_n\frac{\partial}{\partial\alpha}m(\widehat{\theta}_n(\widehat{\alpha}_n),\widehat{\alpha}_n) + P_n\frac{\partial}{\partial\theta}m(\widehat{\theta}_n(\widehat{\alpha}_n),\widehat{\alpha}_n)\frac{\partial}{\partial\alpha}\widehat{\theta}_n(\widehat{\alpha}_n) &=& 0. \end{array}\right.$$

The second term in the LHS of the second equation above is equal to 0, due to the first equation. Hence, $\widehat{\theta}_n(\widehat{\alpha}_n)$ and $\widehat{\alpha}_n$ are solutions of the somehow simpler system

$$\left\{\begin{array}{rcll} P_n\frac{\partial}{\partial\theta}m(\widehat{\theta}_n(\widehat{\alpha}_n),\widehat{\alpha}_n) &=& 0 & (E1) \\ P_n\frac{\partial}{\partial\alpha}m(\widehat{\theta}_n(\widehat{\alpha}_n),\widehat{\alpha}_n) &=& 0 & (E2). \end{array}\right.$$

Use a Taylor expansion in (E1); there exists $(\widetilde{\theta}_n,\widetilde{\alpha}_n)$ inside the segment that links $(\widehat{\theta}_n(\widehat{\alpha}_n),\widehat{\alpha}_n)$ and (θ_0,θ_0) such that

$$\begin{aligned} 0 &= P_n\frac{\partial}{\partial\theta}m(\theta_0,\theta_0) + \left[\left(P_n\frac{\partial^2}{\partial\theta^2}m(\theta_0,\theta_0)\right)^T,\left(P_n\frac{\partial^2}{\partial\alpha\partial\theta}m(\theta_0,\theta_0)\right)^T\right]a_n \\ &\quad + \frac{1}{2}a_n^T A_n a_n, \end{aligned} \tag{1.42}$$

with $a_n := \left((\widehat{\theta}_n(\widehat{\alpha}_n)-\theta_0)^T,(\widehat{\alpha}_n-\theta_0)^T\right)$ and

$$A_n := \left[\begin{array}{cc} P_n\frac{\partial^3}{\partial\theta^3}m(\widetilde{\theta}_n,\widetilde{\alpha}_n) & P_n\frac{\partial^3}{\partial\theta\partial\alpha\partial\theta}m(\widetilde{\theta}_n,\widetilde{\alpha}_n) \\ P_n\frac{\partial^3}{\partial\alpha\partial\theta^2}m(\widetilde{\theta}_n,\widetilde{\alpha}_n) & P_n\frac{\partial^3}{\partial\alpha^2\partial\theta}m(\widetilde{\theta}_n,\widetilde{\alpha}_n) \end{array}\right].$$

By (C.14), the Law of Large Numbers implies that $A_n{=}O_P(1)$, so using (C.12), we can write the last term in the RHS of (1.42) as $o_P(1)a_n$. On the other hand by (C.15), we can write also $\left[\left(P_n\frac{\partial^2 m(\theta_0,\theta_0)}{\partial\theta^2}\right)^T,\left(P_n\frac{\partial^2 m(\theta_0,\theta_0)}{\partial\alpha\partial\theta}\right)^T\right]$ as

$\left[P_{\theta_0}\frac{\partial^2}{\partial\theta^2}m(\theta_0,\theta_0), P_{\theta_0}\frac{\partial^2}{\partial\alpha\partial\theta}m(\theta_0,\theta_0)\right] + o_P(1)$, to obtain from (1.42)

$$-P_n\frac{\partial}{\partial\theta}m(\theta_0,\theta_0) = \left[P_{\theta_0}\frac{\partial^2}{\partial\theta^2}m(\theta_0,\theta_0) + o_P(1), P_{\theta_0}\frac{\partial^2}{\partial\alpha\partial\theta}m(\theta_0,\theta_0) + o_P(1)\right] a_n. \tag{1.43}$$

In the same way, using a Taylor expansion in (E2) ; there exists $(\overline{\theta}_n, \overline{\alpha}_n)$ inside the segment that links $(\widehat{\theta}_n(\widehat{\alpha}_n), \widehat{\alpha}_n)$ and (θ_0,θ_0) such that

$$\begin{aligned} 0 &= P_n\frac{\partial}{\partial\alpha}m(\theta_0,\theta_0) + \left[\left(P_n\frac{\partial^2}{\partial\theta\partial\alpha}m(\theta_0,\theta_0)\right)^T, \left(P_n\frac{\partial^2}{\partial\alpha^2}m(\theta_0,\theta_0)\right)^T\right] a_n \\ &\quad + \frac{1}{2}a_n^T B_n a_n, \end{aligned} \tag{1.44}$$

with

$$B_n := \begin{bmatrix} P_n\frac{\partial^3}{\partial\theta^2\partial\alpha}m(\overline{\theta}_n,\overline{\alpha}_n) & P_n\frac{\partial^3}{\partial\theta\partial\alpha^2}m(\overline{\theta}_n,\overline{\alpha}_n) \\ P_n\frac{\partial^3}{\partial\alpha\partial\theta\partial\alpha}m(\overline{\theta}_n,\overline{\alpha}_n) & P_n\frac{\partial^3}{\partial\alpha^3}m(\overline{\theta}_n,\overline{\alpha}_n) \end{bmatrix}.$$

As in (1.43), we obtain

$$-P_n\frac{\partial m(\theta_0,\theta_0)}{\partial\alpha} = \left[P_{\theta_0}\frac{\partial^2}{\partial\theta\partial\alpha}m(\theta_0,\theta_0) + o_P(1), P_{\theta_0}\frac{\partial^2}{\partial\alpha^2}m(\theta_0,\theta_0) + o_P(1)\right] a_n. \tag{1.45}$$

Using (1.43), (1.45) and (1.41), we get

$$\begin{aligned} \sqrt{n}a_n &= \sqrt{n}\begin{bmatrix} P_{\theta_0}\frac{\partial^2}{\partial\theta^2}m(\theta_0,\theta_0) & P_{\theta_0}\frac{\partial^2}{\partial\alpha\partial\theta}m(\theta_0,\theta_0) \\ P_{\theta_0}\frac{\partial^2}{\partial\theta\partial\alpha}m(\theta_0,\theta_0) & P_{\theta_0}\frac{\partial^2}{\partial\alpha^2}m(\theta_0,\theta_0) \end{bmatrix}^{-1} \begin{bmatrix} -P_n\frac{\partial}{\partial\theta}m(\theta_0,\theta_0) \\ -P_n\frac{\partial}{\partial\alpha}m(\theta_0,\theta_0) \end{bmatrix} + o_P(1) \\ &= \sqrt{n}\begin{bmatrix} P_{\theta_0}\frac{\partial^2}{\partial\theta^2}m(\theta_0,\theta_0) & 0 \\ 0 & P_{\theta_0}\frac{\partial^2}{\partial\alpha^2}m(\theta_0,\theta_0) \end{bmatrix}^{-1} \begin{bmatrix} -P_n\frac{\partial}{\partial\theta}m(\theta_0,\theta_0) \\ -P_n\frac{\partial}{\partial\alpha}m(\theta_0,\theta_0) \end{bmatrix} + o_P(1) \\ &= \sqrt{n}\begin{bmatrix} \frac{-1}{\varphi''(1)}I_{\theta_0}^{-1} & 0 \\ 0 & \frac{1}{\varphi''(1)}I_{\theta_0}^{-1} \end{bmatrix} \begin{bmatrix} -P_n\frac{\partial}{\partial\theta}m(\theta_0,\theta_0) \\ -P_n\frac{\partial}{\partial\alpha}m(\theta_0,\theta_0) \end{bmatrix} + o_P(1). \end{aligned}$$

We therefore deduce, by CLT, that $\sqrt{n}a_n$ converges in distribution to a centered normal variable with covariance matrix

$$\mathbb{V} = \begin{bmatrix} I_{\theta_0}^{-1} & I_{\theta_0}^{-1} \\ I_{\theta_0}^{-1} & I_{\theta_0}^{-1} \end{bmatrix},$$

which completes the proof of Theorem 1.3.

1.5.8 Proof of Theorem 1.4

We have

$$\begin{aligned} T_n^{\phi} &:= \inf_{\beta\in B_0}\sup_{\theta\in\Theta} P_n m\left(\theta, g(\beta)\right). \\ &= P_n m\left(\theta_n(g(\widehat{\beta}_n)), g(\widehat{\beta}_n)\right), \end{aligned}$$

in which as in the proof of Theorem 1.3, $\theta_n(g(\widehat{\beta}_n))$ et $g(\widehat{\beta}_n)$ are solutions of the system of equations

$$\left\{ \begin{array}{rcl} P_n \frac{\partial}{\partial\theta} m\left(\widehat{\theta}_n(g(\widehat{\beta}_n)), g(\widehat{\beta}_n)\right) &=& 0 \\ P_n \frac{\partial}{\partial\beta} m\left(\widehat{\theta}_n(g(\widehat{\beta}_n)), g(\widehat{\beta}_n)\right) &=& 0. \end{array} \right.$$

In the first equation the partial derivative is intended w.r.t. the first variable θ and in the second one w.r.t. the second variable β in $g(\beta)$. A Taylor expansion $P_n \frac{\partial}{\partial\theta} m\left(\widehat{\theta}_n(g(\widehat{\beta}_n)), g(\widehat{\beta}_n)\right)$ and $P_n \frac{\partial}{\partial\beta} m\left(\widehat{\theta}_n(g(\widehat{\beta}_n)), g(\widehat{\beta}_n)\right)$ in a neighborhood of (θ_0, β_0), similarly as in the proof of Theorem 1.3, we obtain

$$\left[\begin{array}{c} -P_n \frac{\partial}{\partial\theta} m(\theta_0, g(\beta_0)) \\ -P_n \frac{\partial}{\partial\beta} m(\theta_0, g(\beta_0)) \end{array} \right] = \left[\begin{array}{cc} P_{\theta_0} \frac{\partial^2}{\partial\theta^2} m(\theta_0, g(\beta_0)) & P_{\theta_0} \frac{\partial^2}{\partial\beta\partial\theta} m(\theta_0, g(\beta_0)) \\ P_{\theta_0} \frac{\partial^2}{\partial\theta\partial\beta} m(\theta_0, g(\beta_0)) & P_{\theta_0} \frac{\partial^2}{\partial\beta^2} m(\theta_0, g(\beta_0)) \end{array} \right] a_n + o_P(1). \tag{1.46}$$

with $a_n := \left((\widehat{\theta}_n(g(\widehat{\beta}_n)) - \theta_0)^T, (\widehat{\beta}_n - \beta_0)^T\right)$. This implies that $\sqrt{n}a_n = O_P(1)$. So by a Taylor expansion of T_n^{ϕ} in a neighborhood of (θ_0, β_0), we obtain

$$\begin{aligned} T_n^{\phi} &= P_n m\left(\theta_0, g(\beta_0)\right) + \left[\left(P_n \frac{\partial}{\partial\theta} m(\theta_0, g(\beta_0)) \right)^T, \left(P_n \frac{\partial}{\partial\beta} m(\theta_0, g(\beta_0)) \right)^T \right] a_n + \\ &\quad \frac{1}{2} a_n^T C_n a_n + o_P(1/n), \end{aligned} \tag{1.47}$$

with

$$C_n := \left[\begin{array}{cc} P_n \frac{\partial^2}{\partial\theta^2} m(\theta_0, g(\beta_0)) & P_n \frac{\partial^2}{\partial\theta\partial\beta} m(\theta_0, g(\beta_0)) \\ P_n \frac{\partial^2}{\partial\beta\partial\theta} m(\theta_0, g(\beta_0)) & P_n \frac{\partial^2}{\partial\beta^2} m(\theta_0, g(\beta_0)) \end{array} \right].$$

Under $\mathcal{H}_0$ (i.e., $\theta_0 = g(\beta_0)$), we have $P_n m\left(\theta_0, g(\beta_0)\right) = 0$. Write C_n as $C + o_P(1)$ with

$$C := \left[\begin{array}{cc} P_{\theta_0} \frac{\partial^2}{\partial\theta^2} m(\theta_0, g(\beta_0)) & P_{\theta_0} \frac{\partial^2}{\partial\theta\partial\beta} m(\theta_0, g(\beta_0)) \\ P_{\theta_0} \frac{\partial^2}{\partial\beta\partial\theta} m(\theta_0, g(\beta_0)) & P_{\theta_0} \frac{\partial^2}{\partial\beta^2} m(\theta_0, g(\beta_0)) \end{array} \right],$$

to obtain from (1.47)

$$T_n^\phi = \left[\left(P_n\frac{\partial}{\partial\theta}m(\theta_0,g(\beta_0))\right)^T,\left(P_n\frac{\partial}{\partial\beta}m(\theta_0,g(\beta_0))\right)^T\right]a_n + \frac{1}{2}a_n^TCa_n + o_P(1/n). \tag{1.48}$$

From this, using (1.46), we obtain

$$\begin{aligned}2T_n^\phi &= \left[P_n\frac{\partial}{\partial\theta}m(\theta_0,g(\beta_0))\right]^T\left[-P_{\theta_0}\frac{\partial^2}{\partial\theta^2}m(\theta_0,g(\beta_0))\right]^{-1}\left[P_n\frac{\partial}{\partial\theta}m(\theta_0,g(\beta_0))\right]\\ &-\left[P_n\frac{\partial}{\partial\beta}m(\theta_0,g(\beta_0))\right]^T\left[P_{\theta_0}\frac{\partial^2}{\partial\beta^2}m(\theta_0,g(\beta_0))\right]^{-1}\left[P_n\frac{\partial}{\partial\beta}m(\theta_0,g(\beta_0))\right]\\ &+o_p(1/n).\end{aligned}$$

Hence,

$$\frac{2n}{\varphi''(1)}T_n^\phi = U_n^TA^{-1}U_n - V_n^TB^{-1}V_n + o_p(1), \tag{1.49}$$

with

$$\begin{aligned}U_n &:= \frac{\sqrt{n}}{\varphi''(1)}P_n\frac{\partial}{\partial\theta}m(\theta_0,g(\beta_0)),\\ V_n &:= \frac{\sqrt{n}}{\varphi''(1)}P_n\frac{\partial}{\partial\beta}m(\theta_0,g(\beta_0)),\\ A &:= -\frac{1}{\varphi''(1)}P_{\theta_0}\frac{\partial^2}{\partial\theta^2}m(\theta_0,g(\beta_0)),\\ B &:= \frac{1}{\varphi''(1)}P_{\theta_0}\frac{\partial^2}{\partial\beta^2}m(\theta_0,g(\beta_0)).\end{aligned}$$

By (1.41), it holds $A = I_{\theta_0}$. On the other hand

$$\begin{aligned}\frac{\partial}{\partial\beta}m(\theta_0,g(\beta_0)) &= \left[\frac{\partial}{\partial\beta}g(\beta_0)\right]^T\frac{\partial}{\partial g(\beta)}m(\theta_0,g(\beta_0))\\ &= [G(\beta_0)]^T\frac{\partial}{\partial g(\beta)}m(\theta_0,g(\beta_0)).\end{aligned}$$

Moreover, since $\varphi'(1)=0$, then $\frac{\partial}{\partial g(\beta)}m(\theta_0,g(\beta_0)) = -\frac{\partial}{\partial\theta}m(\theta_0,g(\beta_0))$, which implies $P_{\theta_0}\frac{\partial}{\partial\beta}m(\theta_0,g(\beta_0)) = [G(\beta_0)]^T\left[-P_{\theta_0}\frac{\partial}{\partial\theta}m(\theta_0,g(\beta_0))\right]$. In the same way, we can prove that

$$P_{\theta_0}\frac{\partial^2}{\partial\beta^2}m(\theta_0,g(\beta_0)) = [G(\beta_0)]^T\left[-P_{\theta_0}\frac{\partial^2}{\partial\theta^2}m(\theta_0,g(\beta_0))\right][G(\beta_0)].$$

It follows that $V_n = [G(\beta_0)]^T U_n$ and $B = [G(\beta_0)]^T I_{\theta_0} G(\beta_0)$. Combining this result with (1.49), we get

$$\frac{2n}{\varphi''(1)} T_n^{\phi} = U_n^T \left[I_{\theta_0}^{-1} - G(\beta_0) B^{-1} G(\beta_0)^T \right] U_n + o_P(1),$$

which is precisely the asymptotic expression for the Wilks likelihood ratio statistic for composite hypotheses. The proof is completed following therefore the same arguments as for the Wilks likelihood ratio statistic ; see e.g. Sen and Singer (1993) Chapter 5.

Chapitre 2

Estimation and Tests by Divergences for Case-Control and Semiparametric Two-Sample Density Ratio Models

In this Chapter, we introduce, for semiparametric two-samples density ratio models, a new estimation and test method based on estimation of ϕ-divergences between probability measures, using the so-called dual representation of ϕ-divergences. In the particular case of multiplicative-intercept risk model, our method includes the semiparametric maximum empirical likelihood one. Large and small sample size behaviors of some proposed estimates and the semiparametric maximum empirical likelihood estimate are illustrated and their sensibilities in the case of contaminated data are compared.

2.1 Introduction and motivations

In this Chapter, we aim to introduce a new method in order to give new answers for the following problems : tests of comparison of two populations and estimation of the parameters for semiparametric density ratio models.

We dispose of two samples,

$$X_1, \ldots, X_{n_0} \qquad \text{and} \qquad Y_1, \ldots, Y_{n_1}$$

from two unknown distributions, noted G and H respectively. A semiparametric density ratio model is of the form

$$\frac{dH}{dG}(x) = m(\theta_0, x) \tag{2.1}$$

where θ_0 is the parameter of interest, unknown, supposed unique and belongs to some open set $\Theta \subseteq \mathbb{R}^d$. $m(\cdot,\cdot)$ is a nonnegative known function.

In the following, we give some statistical examples and motivations for the model (2.1).

2.1.1 Comparison of two populations

In applications, we often come across with the problem of comparing two samples. Let $X_1,\ldots,X_{n_0}$ and $Y_1,\ldots,Y_{n_1}$ be two samples from unknown distributions G and H, respectively.

The use of the well known $t-$test, requires to assume that both samples are normally distributed with unknown means and common unknown variance. The $t-$test enjoys several optimal properties, for example it is the uniformly most powerful unbiased test (see e.g. Lehmann (1986)). If both H and G are normally distributed with equal variances

$$H = \mathcal{N}(\mu_1,\sigma^2) \text{ and } G = \mathcal{N}(\mu_2,\sigma^2),$$

then, the ratio $\frac{dH}{dG}$ takes the form

$$\frac{dH}{dG}(x) = \exp\{\alpha+\beta x\}, \tag{2.2}$$

where

$$\alpha = \frac{\mu_1^2-\mu_2^2}{2\sigma^2} \qquad \text{and} \qquad \beta = \frac{\mu_2-\mu_1}{\sigma^2}.$$

It follows that testing the hypothesis $\mathcal{H}_0 : H = G$ is equivalent to testing the hypothesis $\mathcal{H}_0 : \beta = 0$. We underline that $\beta = 0$ implies $\alpha = 0$.

Kay and Little (1987) and Fokianos (2002) observed that there are cases in which the choice

$$\frac{dH}{dG}(x) = \exp\{\alpha+\beta r(x)\}, \tag{2.3}$$

where $r(x)$ is an arbitrary but known function of x, is more appropriate.

For such models, Fokianos *et al.* (2001) present a test based on the empirical likelihood approach when the samples $X_1,\ldots,X_{n_0}$ and $Y_1,\ldots,Y_{n_1}$ are independent.

The model (2.1) includes the models (2.2) and (2.3) by taking $m(\theta,x) = \exp\{\alpha+\beta x\}$ and $m(\theta,x) = \exp\{\alpha+\beta r(x)\}$ respectively, and $\theta = (\alpha,\beta)^T$.

In the case when the semiparametric assumption (2.1) fails, the test commonly used is the Mann-Whitney-Wilcoxon test (see for example Randles and Wolfe (1979) and Hollander and Wolfe (1999)).

2.1.2 Logistic model and multiplicative-intercept risk model

Let consider the logistic model which has been widely used in statistical applications for the analysis of binary data (see e.g. Agresti (1990), Hosmer and Lemeshow (1999) and Hosmer and Lemeshow (2000)).

Suppose that y is a binary response variable and that x is the associate covariate vector. The logistic model is of the form

$$\Pr(y = 1|x) = \frac{\exp(\gamma + x^T\beta)}{1 + \exp(\gamma + x^T\beta)}, \quad \gamma \in \mathbb{R}, \quad \beta \in \mathbb{R}^{d-1}. \tag{2.4}$$

Note that the marginal density of x, noted $f(x)$, is left completely unspecified.

One of the major reasons the logistic regression model has seen such wide use, especially in epidemiologic research, is the ease of obtaining adjusted odds ratios from the estimated slope coefficients when sampling is performed conditional on the outcome variables, as in a case-control study.

In a case-control study the binary outcome variable is fixed by stratification. In this type of study design, two random samples of sizes n_0 and n_1 are chosen from the two strata defined by the outcome variable, i.e., from the subsets of the population with $y = 0$ and $y = 1$ respectively. Assume that $x_1, \ldots, x_{n_0}$ are the observed covariates from the control group and let $x_{n_0+1}, \ldots, x_n$ $(n = n_0 + n_1)$ be those from the case group. We aim to estimate the parameters γ and β using the two samples $X_1, \ldots, X_{n_0}$ and $X_{n_0+1}, \ldots, X_n$. We show that the logistic model (2.4) writes in the form of the model (2.1). So, let f denote the density function of the covariates x, and put

$$\pi = \Pr(y = 1) = \int \Pr(y = 1|x) f(x)\, dx$$

and assume that

$$f_i(x) = f(x|y = i) = dF(x|y = i)/dx \qquad i = 0, 1$$

exist and represent the conditional density function of x given $y = i$. It is not difficult to manipulate the case-control likelihood function to obtain a logistic regression model in which the dependent variable is the outcome variable of interest to the investigator. The key step in this development is an application of the Bayes Theorem, that yields

$$\begin{aligned} f_1(x) &= \frac{\exp(\gamma + x^T\beta)}{\pi \cdot (1 + \exp(\gamma + x^T\beta))} f(x) \\ f_0(x) &= \frac{\exp(\gamma + x^T\beta)}{(1 - \pi) \cdot (1 + \exp(\gamma + x^T\beta))} f(x). \end{aligned}$$

So

$$\begin{aligned}\frac{f_1(x)}{f_0(x)} &= \frac{1-\pi}{\pi}\exp(\gamma + x^T\beta) = \\ &= \exp\left\{\gamma + \log\left(\frac{1-\pi}{\pi}\right) + x^T\beta\right\} = \\ &= \exp(\alpha + x^T\beta),\end{aligned}$$

where $\alpha := \gamma + \log\left(\frac{1-\pi}{\pi}\right)$. So the model (2.4) is equivalent to the following two-sample semiparametric model

$$\begin{aligned}x_1, \ldots, x_{n_0} &\sim f(x|y=0) = f_0(x) \\ x_{n_0+1}, \ldots, x_n &\sim f(x|y=1) = f_1(x) = \exp(\alpha + x^T\beta)\cdot f_0(x). \end{aligned} \tag{2.5}$$

More generally, we can consider the following logistic model

$$\Pr(y=1|x) = \frac{\exp(\gamma + r(x,\beta))}{1+\exp(\gamma + r(x,\beta))}, \quad \gamma \in \mathbb{R}, \quad \beta \in \mathbb{R}^{d-1},$$

where $r(x,\beta)$ is a given function of x and β. In this case, we obtain the following two-sample semiparametric model

$$\begin{aligned}x_1, \ldots, x_{n_0} &\sim f_0(x) \\ x_{n_0+1}, \ldots, x_n &\sim f_1(x) = \exp(\alpha + r(x,\beta))\cdot f_0(x), \end{aligned} \tag{2.6}$$

which is called multiplicative-intercept risk model. The models (2.5) and (2.6) are particular cases of the models (2.1) by taking $\frac{dG}{dx} = f_0$, $\frac{dH}{dx} = f_1$ and $\theta = (\alpha, \beta^T)^T$. In the context of the model (2.1), estimate γ and β is equivalent to estimate α and β.

For models (2.1), Qin (1998) presents estimate of θ_0 based on the empirical likelihood approach when the samples $X_1, \ldots, X_{n_0}$ and $Y_1, \ldots, Y_{n_1}$ are independent.

However, an important special case of the case-control study is the matched (or paired) study. In this design, subjects are stratified on the basis of variables believed to be associated with the outcome (an example of stratification variable is the age for each of the individuals in the survey). Within each stratum, samples of cases ($y = 1$) and controls ($y = 0$) are chosen ; the most common matched design includes one case and one control per stratum and is thus referred as 1-1 matched study.

The goal of this Chapter is to present a new approach for estimation of the parameter θ_0 and tests of the hypothesis $\mathcal{H}_0 : H = G$ in two-sample semiparametric models of the form (2.1) with independent or paired samples $X_1, \ldots, X_{n_0}$ and $Y_1, \ldots, Y_{n_1}$.

In general, in order to compare the laws of two variables X and Y, we estimate some measures of difference (distances, pseudo-distances or divergences) between the two laws using the empirical distributions of the two samples. Denote H and G the laws of Y and X, respectively. The empirical measures associated to the samples $Y_1, \ldots, Y_{n_1}$ and $X_1, \ldots, X_{n_0}$ are $H_{n_1}^Y$ and $G_{n_0}^X$ respectively, namely

$$H_{n_1}^Y = \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_{Y_i} \qquad \text{and} \qquad G_{n_0}^X = \frac{1}{n_0} \sum_{i=1}^{n_0} \delta_{X_i},$$

in which δ_x is the Dirac measure at point x.

In this Chapter, we consider divergences, noted $\phi(\cdot, \cdot)$, between probability measures. The choice of the class of ϕ-divergences is motivated by their invariance property w.r.t. change of variables, and by the fact that it covers some classical methods.

The direct "plug-in" estimates $\phi(H_{n_1}^Y, G_{n_0}^X)$ of ϕ-divergences $\phi(H, G)$, as we will see in Section 2.1, is not defined when H and G do not share the same discrete finite support ; in Section 2.2 we will solve this problem using the so-called "dual-representation" of ϕ-divergences (see Keziou (2003)).

This Chapter is organized as follows : in Section 2, we present our estimates, after a brief review of the mathematical instruments necessary to the definition of estimates. In Section 3, we study the asymptotic behavior of the proposed estimates for both cases : two independent samples and two paired samples. In Section 4, we will focus on the multiplicative-intercept-risk model and we compare our method to maximum semiparametric likelihood's ; we will show that the class of the proposed estimates covers the two-sample maximum semiparametric likelihood estimate (SMLE). In Section 5, we illustrate and compare large and small sample size behavior of some estimates and the SMLE and we compare, by numerical results, their sensibility in the case of contaminated data. Concluding remarks and possible developments are presented in Section 6. All proofs are in Section 7.

2.2 Semiparametric Estimation and Tests by Divergences

Before giving the estimates, let us remind the definition and properties of ϕ-divergences between probability measures (p.m.'s).

2.2.1 ϕ-divergences and dual representation

Let $(\mathcal{X}, \mathcal{B})$ some measurable space on which all random variables and all p.m.'s will be defined. Let φ be a convex function defined from $[0, +\infty]$ onto $[0, +\infty]$ with

$\varphi(1) = 0$. Define the domain of φ through

$$D_\varphi := \{x \in [0, +\infty] \text{ such that } \varphi(x) < +\infty\}. \tag{2.7}$$

Since φ is a convex function, then D_φ is an interval in $\mathbb{R}_+$ which may be open or not, bounded or unbounded. Hence, write $D_\varphi := (a, b)$ in which b may be finite or infinite.

We will consider only strictly convex functions φ on D_φ. We assume that D_φ contains $]0, +\infty[$ and that the functions φ are $\mathcal{C}^3$ on $]0, +\infty[$. We define $\varphi(0)$, $\varphi'(0)$, $\varphi''(0)$,$\varphi'''(0)$, $\varphi(+\infty)$, $\varphi'(+\infty)$, $\varphi''(+\infty)$ and $\varphi'''(+\infty)$ respectively by $\lim_{x\downarrow 0} \varphi(x)$, $\lim_{x\downarrow 0} \varphi'(x)$, $\lim_{x\downarrow 0} \varphi''(x)$, $\lim_{x\downarrow 0} \varphi'''(x)$, $\lim_{x\uparrow+\infty} \varphi(x)$, $\lim_{x\uparrow+\infty} \varphi'(x)$,$\lim_{x\uparrow+\infty} \varphi''(x)$ and $\lim_{x\uparrow+\infty} \varphi'''(x)$. These quantities may be finite or infinite.

Definition 2.1. *The ϕ-divergence between two p.m.'s Q and P is defined through*

$$\phi(Q, P) = \int \varphi\left(\frac{dQ}{dP}\right) \, dP \tag{2.8}$$

if Q is absolutely continuous with respect to (a.c. w.r.t.) P, and $\phi(Q, P) = +\infty$ otherwise.

The ϕ-divergences have been introduced by Csiszár (1963), see also Csiszár (1967c), Csiszár (1967a), Csiszár (1967b), Rüschendorf (1984) and Liese and Vajda (1987).

For all probability measure P, the mappings $Q \to \phi(Q, P)$ are convex and are nonnegative. When $Q = P$ then $\phi(Q, P) = 0$. Further, if the function φ is strictly convex on a neighborhood of $x = 1$, then the following fundamental property holds

$$\phi(Q, P) = 0 \text{ if and only if } Q = P. \tag{2.9}$$

For proofs of these properties, we refer to Csiszár (1963), Csiszár (1967c) and Liese and Vajda (1987) Chapter 1.

As a consequence of the property (2.9), we can use estimates of the $\phi-$divergences between H and G to perform tests of the hypothesis $\mathcal{H}_0 : H = G$ rejecting the null hypothesis $\mathcal{H}_0$ when the estimates take large values.

The Kullback-Leibler (KL), modified Kullback-Leibler (KL_m), χ^2, modified χ^2 (χ^2_m), Hellinger (H) and L^1 divergences are respectively associated to the convex functions $\varphi(x) = x \log x - x + 1$, $\varphi(x) = -\log x + x - 1$, $\varphi(x) = \frac{1}{2}(x-1)^2$, $\varphi(x) = \frac{1}{2}(x-1)^2/x$, $\varphi(x) = 2(\sqrt{x} - 1)^2$ and $\varphi(x) = |x - 1|$.

All those divergences except the L^1 one, belong to the class of power divergences introduced in Cressie and Read (1984) (see also Liese and Vajda (1987) Chapter 2). They are defined through the class of convex functions

$$x \in \mathbb{R}^*_+ \mapsto \varphi_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma - 1)} \tag{2.10}$$

if $\gamma \in \mathbb{R} \setminus \{0,1\}$ and by $\varphi_0(x) := -\log x + x - 1$, $\varphi_1(x) := x \log x - x + 1$. For all $\gamma \in \mathbb{R}$, $\varphi_\gamma(0) := \lim_{x \downarrow 0} \varphi_\gamma(x)$ and $\varphi_\gamma(+\infty) := \lim_{x \uparrow +\infty} \varphi_\gamma(x)$.
So, the KL_m divergence is associated to φ_0, the KL to φ_1, the χ^2_m to φ_{-1}, the χ^2 to φ_2, and the Hellinger distance to $\varphi_{1/2}$. For all $\gamma \in \mathbb{R}$, sometimes, we denote ϕ_γ the divergence associated to the convex function φ_γ.

When both H and G share the same finite discrete support S, then the ϕ-divergence $\phi(Q,P)$ writes

$$\phi(H,G) = \sum_{j \in S} \varphi\left(\frac{H(j)}{G(j)}\right) G(j). \tag{2.11}$$

In this case, $\phi(H,G)$ can be estimated by

$$\widehat{\phi}_n(H,G) = \sum_{j \in S} \varphi\left(\frac{H^Y_{n_1}(j)}{G^X_{n_0}(j)}\right) G^X_{n_0}(j). \tag{2.12}$$

In the case when the laws H and G have continuous support or discrete but different or infinite support, then the estimate (2.12) is infinite due to lack of absolute continuity ; Broniatowski (2003) makes use of the well known dual representation of Kullback-Leibler divergence ($KL-$divergence), as the Fenchel-Legendre transform of the moment generating function, in order to estimate the $KL-$divergence between some set of probability measures and some probability measure P, from a sample $X_1, \ldots, X_n$ with common distribution P. Using similar idea, in order to solve the problem of lack of absolute continuity, we make use of the so-called "Dual representation of $\phi-$divergences" (see Keziou (2003)), which has been used in parametric statistics in Keziou (2003), and in Broniatowski and Keziou (2003) and which we recall here.
Denote φ'^{-1} the inverse function of φ' and $\text{Im } \varphi'$ the set of all values of the function φ'.

The convex conjugate (or Fenchel-Legendre transform) of φ will be denoted ψ, i.e.,

$$t \in \mathbb{R} \;\; \mapsto \;\; \psi(t) := \sup_{x \in \mathbb{R}} \{tx - \varphi(x)\}.$$

Denote φ^* the function defined on $\text{Im } \varphi'$ by

$$t \in \text{ Im } \varphi' \;\; \mapsto \;\; \varphi^*(t) := t\varphi'^{-1}(t) - \varphi\left(\varphi'^{-1}(t)\right). \tag{2.13}$$

Clearly, the function φ^* coincides on the set $\text{Im } \varphi'$ with ψ, the convex conjugate of φ.

Theorem 2.1. *Assume that the function φ is strictly convex and is $\mathcal{C}^2$ on the interior of its domain D_φ. Let Q and P be two p.m.'s with $\phi(Q,P) < \infty$. Let $\mathcal{F}$ be a class of functions such that (i) for all f in $\mathcal{F}$, $\int |f|\ dQ$ is finite, (ii) $\varphi'(dQ/dP)$ belongs to $\mathcal{F}$ and (iii) for any f in $\mathcal{F}$, Im f is included in Im φ'. We then have*

(1) The divergence $\phi(Q,P)$ admits the "dual representation"

$$\phi(Q,P) = \sup_{f\in\mathcal{F}} \left\{ \int f\ dQ - \int \varphi^*(f)\ dP \right\}. \tag{2.14}$$

(2) The supremum in (2.14) is unique $(P-a.s)$ and is reached at $f = \varphi'(dQ/dP)$ $(P-a.s)$.

2.2.2 Estimation through the dual representation of ϕ- divergences

Our estimates derive from an application of Theorem 2.1 above.

In order to introduce our estimates, we consider the following notations

$k(\theta)\ :\ x \mapsto k(\theta,x) = \varphi'(m(\theta,x))\,;$

$l(\theta)\ :\ x \mapsto l(\theta,x) = \varphi'(m(\theta,x)) \cdot m(\theta,x) - \varphi(m(\theta,x)).$

We sometimes write Pf for $\int f\ dP$ for any measure P and function f.

We consider also the following conditions

(C.0) $\int |\varphi'(m(\theta,x))|\ dH(x)$ is finite for all $\theta \in \Theta$;

(C.1) $\phi(H,G)$ is finite.

Under conditions (C.0) and (C.1), by application of the above Theorem, for the probability measures H and G, choosing the class of functions

$$\mathcal{F} = \{x \to \varphi'(m(\theta,x))\ /\ \theta \in \Theta\}, \tag{2.15}$$

we obtain, for the divergences $\phi(H,G)$, the representations

$$\phi(H,G) = \sup_{\theta\in\Theta} \{Hk(\theta) - Gl(\theta)\}. \tag{2.16}$$

So, we propose to estimate the divergences $\phi(H,G)$ by

$$\widehat{\phi}_{n_0,n_1}(H,G) := \sup_{\theta\in\Theta} \left\{ H_{n_1}^Y k(\theta) - G_{n_0}^X l(\theta) \right\}. \tag{2.17}$$

Further, the supremum in (2.16) is unique and is reached at $\theta = \theta_0$, that is

$$\theta_0 = \arg\sup_{\theta\in\Theta} \{Hk(\theta) - Gl(\theta)\}. \tag{2.18}$$

So, we propose the following estimates of θ_0

$$\widehat{\theta}_{n_0,n_1} := \arg\sup_{\theta\in\Theta} \widehat{\phi}_{n_0,n_1}(H,G) := \arg\sup_{\theta\in\Theta} \left\{ H_{n_1}^Y k(\theta) - G_{n_0}^X l(\theta) \right\}. \tag{2.19}$$

We underline that the use of the Dual representation of $\phi-$divergences provides estimates for $\phi-$divergences and also estimates for the parameter θ_0, while the plug-in direct approach (2.12) obviously provides only estimates of ϕ-divergences between H and G.

Simulation results show that the choice of a good divergence depends on the ratio n_1/n_0. Hence, we will introduce a class of ϕ-divergences depending on this ratio. In particular, in Section 4, we will consider the divergence associated to the nonnegative convex function

$$\varphi_\rho^\star(x) := x\log x - \frac{1+\rho x}{\rho}\log(1+\rho x) + \frac{1+\rho}{\rho}\log(1+\rho) - \left(\frac{\rho-1}{\rho} - \log(1+\rho)\right)(x-1),$$

which in the multiplicative-intercept risk model yields to an estimate of the parameter θ_0 corresponding to the semiparametric maximum likelihood's.

Denote

$$n := n_0 + n_1 \quad \text{and} \quad \rho := \lim_{n\to\infty} \rho_n := \lim_{n\to\infty} \frac{n_1}{n_0},$$

which we suppose positive and finite. We consider any convex function depending on ρ, noted

$$\begin{array}{rccc} \varphi_\rho & : & \mathbb{R}_+ & \to & [0,\infty] \\ & & x & \mapsto & \varphi_\rho(x) \end{array}$$

satisfying $\varphi_\rho(1) = 0$. Denote $\phi_\rho(H,G)$ the divergence between the probability laws H and G associated to the convex function φ_ρ. Consider also the following notations

$$\begin{aligned} &k_\rho(\theta) \ : \ x \mapsto k_\rho(\theta,x) = \varphi_\rho'(m(\theta,x)); \\ &l_\rho(\theta) \ : \ x \mapsto l_\rho(\theta,x) = \varphi_\rho'(m(\theta,x))\cdot m(\theta,x) - \varphi_\rho(m(\theta,x)). \end{aligned}$$

We introduce estimates for the divergences $\phi_\rho(H,G)$ and for the parameter θ_0 using the ratio $\rho_n := \frac{n_1}{n_0}$ as follows

$$\widehat{\phi}_{\rho_n}(H,G) := \sup_{\theta\in\Theta} \left\{ H_{n_1}^Y k_{\rho_n}(\theta) - G_{n_0}^X l_{\rho_n}(\theta) \right\}, \tag{2.20}$$

and

$$\widehat{\theta}_{\rho_n} := \arg\sup_{\theta\in\Theta} \widehat{\phi}_{\rho_n}(H,G) := \arg\sup_{\theta\in\Theta} \left\{ H_{n_1}^Y k_{\rho_n}(\theta) - G_{n_0}^X l_{\rho_n}(\theta) \right\}, \tag{2.21}$$

which we call "two-sample semiparametric dual $\phi-$divergences estimates" (SDϕDE's).

2.3 The asymptotic behaviour of the SDϕDE's

In this Section, we state both weak and strong consistency of the estimate $\widehat{\theta}_{\rho_n}$ and $\widehat{\phi}_{\rho_n}(H,G)$ defined in (2.21) and (2.20). We also state their limit distributions. The hypotheses handled here are similar to those used in van der Vaart (1998) Chapter 5, in the study of M-estimates. Denote $\|\cdot\|$ the Euclidean norm defined on $\mathbb{R}^d$, $k'_\rho(\theta,x)$, $l'_\rho(\theta,x)$ $d-$dimensional vectors with entries respectively $\frac{\partial}{\partial\theta_i}k_\rho(\theta,x)$ and $\frac{\partial}{\partial\theta_i}l_\rho(\theta,x)$. Let $k''_\rho(\theta,x)$ and $l''_\rho(\theta,x)$ be $d\times d$ matrices with entries respectively $\frac{\partial^2}{\partial\theta_i\partial\theta_j}k_\rho(\theta,x)$ and $\frac{\partial^2}{\partial\theta_i\partial\theta_j}l_\rho(\theta,x)$. In the sequel, we will assume that conditions (C.0) and (C.1) are satisfied, and that the estimates $\widehat{\theta}_{\rho_n}$ exist. Moreover, let α_0 be the value of the parameter θ which satisfies $m(\alpha_0,x)=1$ ($G-$a.s.). Remark that $\theta_0=\alpha_0$ iff $H=G$.

2.3.1 Consistency.

In order to prove consistency of the estimates (2.20) and (2.21), we consider the set

$$\Theta_1 := \{\theta\in\Theta \text{ such that } G\, k_\rho(\theta) < +\infty\},$$

and denote by Θ_1^c the complementary of the subset Θ_1 in the set Θ, i.e.,

$$\Theta_1^c := \{\theta\in\Theta \text{ such that } G\, k_\rho(\theta) = +\infty\}.$$

Note that Θ_1 contains θ_0, if $\phi_\rho(H,G)<\infty$. We also define the estimates $\tilde{\phi}_{\rho_n}(H,G)$ and $\tilde{\theta}_{\rho_n}$ by

$$\tilde{\phi}_{\rho_n}(H,G) := \sup_{\theta\in\Theta_1}\left\{H_{n_1}^Y k_{\rho_n}(\theta) - G_{n_0}^X l_{\rho_n}(\theta)\right\}, \tag{2.22}$$

$$\tilde{\theta}_{\rho_n} := \arg\sup_{\theta\in\Theta_1}\left\{H_{n_1}^Y k_{\rho_n}(\theta) - G_{n_0}^X l_{\rho_n}(\theta)\right\}. \tag{2.23}$$

We give conditions under which we will prove consistency of (2.22) and (2.23), and we will show that these estimates coincide with (2.20) and (2.21) respectively, which implies consistency of the estimates $\widehat{\phi}_{\rho_n}(H,G)$ and $\widehat{\theta}_{\rho_n}$. The conditions are the followings

(C.2) $\sup_{\theta\in\Theta_1}\left|H_{n_1}^Y k_{\rho_n}(\theta) - G_{n_0}^X l_{\rho_n}(\theta) - (Hk_\rho(\theta) - Gl_\rho(\theta))\right| \to 0$ a.s. (resp. in probability);

(C.3) for any positive ε,

$$\sup_{\{\theta\in\Theta:\|\theta-\theta_0\|\geq\varepsilon\}}\{Hk_\rho(\theta) - Gl_\rho(\theta)\} < Hk_\rho(\theta_0) - Gl_\rho(\theta_0),$$

that is the maximizer θ_0 of the function $\theta\mapsto Hk_\rho(\theta) - Gl_\rho(\theta)$ is isolated;

(C.4) there exists $M < 0$ and $\overline{n} > 0$ such that for all $n_0, n_1 \geq \overline{n}$, $\sup_{\theta \in \Theta_1^c} \left\{ H_{n_1}^Y k_{\rho_n}(\theta) - G_{n_0}^X l_{\rho_n}(\theta) \right\} < M$ a.s. (resp. in probability).

Theorem 2.2.

(a) If (C.2), (C.3) and (C.4) hold, then $\widehat{\theta}_{\rho_n}$ converge a.s. (resp. in probability) to θ_0.

(b) If conditions (C.2) and (C.4) hold, then $\widehat{\phi}_{\rho_n}(H, G)$ converge a.s. (resp. in probability) to $\phi_\rho(H, G)$.

Remark 2.1. *If $\Theta_1^c = \varnothing$, condition (C.4) is not necessary for the consistency results in (a) and (b).*

2.3.2 Asymptotic distributions.

We state the limit laws of the proposed estimates for both cases : two independent samples and two paired samples. We will consider the following regularity conditions

(C.5) $\widehat{\theta}_{\rho_n} \to \theta_0$ in probability ;

(C.6) the functions $\theta \mapsto k_\rho(\theta, x)$ and $\theta \mapsto l_\rho(\theta, x)$ are $\mathcal{C}^3$ on a neighborhood $V(\theta_0)$ of θ_0 for all x (G-a.s.), and all partial derivatives of order 3 of $\{\theta \mapsto k_\rho(\theta, x),\ \theta \in V(\theta_0)\}$ and $\{\theta \mapsto l_\rho(\theta, x),\ \theta \in V(\theta_0)\}$ are dominated for all x (G-a.s.) respectively by some functions
$x \mapsto K(x)$, with K an H-integrable function and
$x \mapsto L(x)$, with L a G-integrable function ;

(C.7) the function $\theta \mapsto m(\theta, x)$ is $\mathcal{C}^3$ on $V(\theta_0)$ for all x (G-a.s.), and the partial derivatives of order 1 of $\{\theta \mapsto m(\theta, x),\ \theta \in V(\theta_0)\}$ are dominated for all x (G-a.s.) by some function G-integrable. The integrals $H\left|k'_\rho(\theta_0)\right|$, $H\left|k''_\rho(\theta_0)\right|$, $G\left|l''_\rho(\theta_0)\right|$, and $H\left|\varphi'''(m(\theta_0))\dot{m}(\theta_0)\dot{m(\theta_0)}^T\right|$ are finite.

(C.8) The matrix $\left[Gl''_\rho(\theta_0) - Hk''_\rho(\theta_0)\right]$ is non singular and

$$H\|k'_\rho(\theta_0)\|^2 < \infty, \qquad G\|l'_\rho(\theta_0)\|^2 < \infty;$$

(C.9) Let

$$\rho_{n_1} := \frac{n_1}{n} \qquad \rho_{n_0} := \frac{n_0}{n}.$$

We assume that $\rho_{n_1} \to \rho_1 > 0$ and $\rho_{n_0} \to \rho_0 > 0$, when $n \to \infty$.

Theorem 2.3. *Assume that conditions (C.5-9) hold. Then, if the two samples are independent, we have*

1. $\sqrt{n}(\widehat{\theta}_{\rho_n} - \theta_0)$ *converge to a centered multivariate normal variable with covariance matrix*

$$\begin{aligned} LCM &= \left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right]^{-1} \cdot \Big[\rho_1^{-1} \cdot \Big(Hk'_\rho(\theta_0)\left[k'_\rho(\theta_0)\right]^T - \\ &\quad -Hk'_\rho(\theta_0)\left[Hk'_\rho(\theta_0)^T\right]\Big) + \rho_0^{-1} \cdot \Big(Gl'_\rho(\theta_0)\left[l'_\rho(\theta_0)\right]^T - \\ &\quad - Gl'_\rho(\theta_0)\left[Gl'_\rho(\theta_0)\right]^T\Big)\Big] \cdot \left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right]^{-1}. \end{aligned} \tag{2.24}$$

If $H = G$, then the limit covariance matrix is

$$LCM_0 = \left[\rho_0\rho_1 G(\dot{m}(\alpha_0, x)\left[\dot{m}(\alpha_0, x)\right]^T)\right]^{-1}. \tag{2.25}$$

2. *If $H = G$, then the statistics*

$$\frac{2n\rho_{n_0}\rho_{n_1}}{\varphi''_{\rho_n}(1)}\widehat{\phi}_{\rho_n}(H, G)$$

converge in distribution to a χ^2 variable with d degrees of freedom.

3. *If $H \neq G$, then $\sqrt{n} \cdot \Big(\widehat{\phi}_{\rho_n}(H, G) - \phi_\rho(H, G)\Big)$ converge in distribution to a centered normal variable with variance*

$$W = \rho_1^{-1}\left(Hk_\rho^2(\theta_0) - (Hk_\rho(\theta_0))^2\right) + \rho_0^{-1}\left(Gl_\rho^2(\theta_0) - (Gl_\rho(\theta_0))^2\right). \tag{2.26}$$

Remark 2.2. *All estimates $\widehat{\theta}_{\rho_n}$ are asymptotically normal variables. When H and G differ ($\theta_0 \neq \alpha_0$), the limit variance depends on the choice of the divergence. It would be interesting to obtain the divergence, noted $\phi^\star$, or the explicit form of the convex function, noted $\varphi^\star$, that optimizes the limit variance in the class of SDϕDE's ; denote V_φ the limit variance obtained for the estimate associated to the divergence ϕ : then $\varphi^\star$ is such that the matrix $V_\varphi - V_{\varphi^\star}$ is positive semi-definite for any convex function φ. clearly $\varphi^\star$ depends on the model and in general is not easy to obtain its explicit form. One possible way to simplify this problem is to search the optimal divergence $\phi^\star$ in the class (2.10) of power-divergences, that is to search the number $\gamma^\star$ such that $V_{\varphi_\gamma} - V_{\varphi_{\gamma^\star}}$ is positive semi-definite for all $\gamma \in \mathbb{R}$.*
If $H = G$, that is $\theta_0 = \alpha_0$, all SDϕDE's are asymptotically equivalent and have the same limit variance.

Remark 2.3. *The limit law of the statistics $\frac{2n\rho_{n_0}\rho_{n_1}}{\varphi''_{\rho_n}(1)}\widehat{\phi}_{\rho_n}(H, G)$, when $H = G$, allows to perform tests of the hypothesis $\mathcal{H}_0 : H = G$ against the alternative $\mathcal{H}_1 : H \neq G$.*

Since $\phi(H,G)$ is nonnegative and takes value 0 only when $H=G$, the tests are defined through the critical regions (C_ϕ), obviously depending on ϕ

$$C_\phi := \left\{ \frac{2n\rho_{n_0}\rho_{n_1}}{\varphi''_{\rho_n}(1)} \widehat{\phi}_{\rho_n}(H,G) > q_{(1-\alpha)} \right\}.$$

where $q_{(1-\alpha)}$ is the $(1-\alpha)$ quantile of the χ^2 distribution with d degrees of freedom. Also these tests are all asymptotically of level α and asymptotically powerful, since the estimates $\widehat{\phi}_{\rho_n}$ are n-consistent estimates of $\phi(H,G)=0$ under $\mathcal{H}_0$ and $\sqrt{n}$-consistent estimates of $\phi(H,G)$ under $\mathcal{H}_1$ (see Theorem 2.3 part 3); it would be interesting to obtain the divergence that yields the most powerful test.

In the following Theorem, we give the limit laws of the estimates in the case of paired data. Denote $F_{X,Y}$ the joint law of the variables X and Y. The laws H and G in this case stand for the marginal laws of $F_{X,Y}$.

Theorem 2.4. *Assume that conditions (C.5-9) hold. Then, in the case of paired samples, we have :*

1. *$\sqrt{n}(\widehat{\theta}_{\rho_n} - \theta_0)$ converge to a centered multivariate normal variable with covariance matrix*

$$\begin{aligned} LCM(dep) \;=\; & \left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right]^{-1} \cdot \Big[\rho_1^{-1} \cdot \Big(Hk'_\rho(\theta_0)\left[k'_\rho(\theta_0)\right]^T - \\ & Hk'_\rho(\theta_0)\left[Hk'_\rho(\theta_0)^T\right]\Big) + \rho_0^{-1} \cdot \Big(Gl'_\rho(\theta_0)\left[l'_\rho(\theta_0)\right]^T - \\ & Gl'_\rho(\theta_0)\left[Gl'_\rho(\theta_0)\right]^T\Big) - 2\rho_1^{-1}\min(1,\rho) \cdot \\ & \left(F_{X,Y}k'_\rho(\theta_0)[l'_\rho(\theta_0)]^T - Hk'_\rho(\theta_0)[Gl'_\rho(\theta_0)]^T\right)\Big] \\ & \cdot \left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right]^{-1}. \end{aligned}$$

   If $H=G$, then the limit covariance matrix becomes

$$\begin{aligned} LCM_0(dep) \;=\; & \left[G(\dot{m}(\alpha_0,x)\cdot[\dot{m}(\alpha_0,x)]^T)\right]^{-1} \cdot \left[(\rho_0\rho_1)^{-1}\right. \\ & G(\dot{m}(\alpha_0,x)\cdot[\dot{m}(\alpha_0,x)]^T) - 2\rho_1^{-1}\min(1,\rho)\cdot \\ & \left.\int \dot{m}(\alpha_0,x)\cdot[\dot{m}(\alpha_0,x)]^T\, dF_{X,Y}\right]\cdot \\ & \left[G(\dot{m}(\alpha_0,x)\cdot[\dot{m}(\alpha_0,x)]^T\right]^{-1}. \end{aligned}$$

2. *If $H=G$, then the statistics $\frac{2n}{\varphi''_{\rho_n}(1)}\widehat{\phi}_{\rho_n}(H,G)$ converge in distribution to the variable YY^T, where $Y \sim \mathcal{N}_d(0,M)$, $M := A^{-\frac{1}{2}}V_{dep}A^{-\frac{1}{2}}$,*

$A = \varphi''_\rho(1)\left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right]$ *and*

$$\begin{aligned} V_{dep} &= \frac{\varphi''_\rho(1)}{\rho_0\rho_1}\cdot\left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right] \\ &\quad -2(\varphi''_\rho(1))^2\frac{\min(1,\rho)}{\rho_1}F_{X,Y}\dot{m}(\theta_0,x)(\dot{m}(\theta_0,x))^T. \end{aligned}$$

3. *If* $H \neq G$, *then* $\sqrt{n}\cdot\left(\widehat{\phi}_{\rho_n}(H,G) - \phi_\rho(H,G)\right)$ *converge in distribution to a centered normal variable with variance*

$$\begin{aligned} W_{dep} &= \rho_1^{-1}\left(Hk^2_\rho(\theta_0) - (Hk_\rho(\theta_0))^2\right) + \rho_0^{-1}\left(Gl^2_\rho(\theta_0) - (Gl_\rho(\theta_0))^2\right) - \\ &\quad -2\rho_1^{-1}\min(1,\rho)\left(F_{X,Y}k(\theta_0)l(\theta_0) - Hk(\theta_0)Gl(\theta_0)\right). \end{aligned}$$

2.4 Semiparametric ELE and SDϕDE's

In the present setting, the semiparametric empirical likelihood method for estimation of the parameter θ_0 can be summarized as follows.
For any $\theta \in \Theta$, the likelihood of the two samples $X_1,\ldots,X_{n_0}$ and $Y_1,\ldots,Y_{n_1}$, if they are independent is

$$L = \prod_{i=1}^{n_0} g(X_i)\prod_{j=1}^{n_1} h(Y_j).$$

For simplicity, we write $X_{n_0+1},\ldots,X_n$ the sample $Y_1,\ldots,Y_{n_1}$. Since $h(x) = m(\theta,x)g(x)$, then L writes

$$L = \prod_{i=1}^{n} g(X_i)\prod_{j=n_0+1}^{n} m(\theta,X_j).$$

When G is discrete, $g(X_i)$ stands for $Pr(X = x_i)$. For convenience we write p_i instead of $g(X_i)$. In order to maximize the log-likelihood $l = \log L$, we need to consider distributions with jumps only at each of the observed sample points, so that

$$l = \sum_{i=1}^{n}\log p_i + \sum_{i=n_0+1}^{n}\log m(\theta,X_i)$$

where $p_i \geq 0$, $\sum_{i=1}^n p_i = 1$, $\sum_{i=1}^n p_i\{m(\theta,X_i) - 1\} = 0$.
Along the lines of Qin and Lawless (1994) we have that $p_i = n^{-1}(1 + \lambda\{m(\theta,X_i) - 1\})^{-1}$, where λ is a Lagrange multiplier determined by the equation

$$-\sum_{i=1}^{n}\frac{m(\theta,X_i) - 1}{1 + \lambda\{m(\theta,X_i) - 1\}} = 0.$$

Thus

$$l(\theta,\lambda) = -\sum_{i=1}^{n} \log[1+\lambda\{m(\theta,X_i)-1\}] + \sum_{i=n_0+1}^{n} \log m(\theta,X_i) \tag{2.27}$$

Note that $l(\theta,\lambda)$ is well defined for $\theta \in \Theta$ and $\lambda \in [0,1]$. Taking derivatives with respect to θ and λ, we have the estimating equations

$$\begin{aligned} \frac{\partial l}{\partial \theta} &= -\lambda \sum_{i=1}^{n} \frac{\partial m/\partial\theta}{1+\lambda\{m(\theta,X_i)-1\}} + \sum_{i=n_0+1}^{n} \frac{\partial \log m(\theta,X_i)}{\partial\theta} = 0; \qquad (2.28)\\ \frac{\partial l}{\partial \lambda} &= -\sum_{i=1}^{n} \frac{m(\theta,X_i)-1}{1+\lambda\{m(\theta,X_i)-1\}} = 0. \end{aligned}$$

The maximum semiparametric likelihood estimator of θ_0 is the solution of the above estimating equations.

For the multiplicative-intercept risk model, that is when $\frac{dH}{dG} = m(\theta,x) = \exp\{\alpha + r(\beta,x)\}$ and $\theta = (\alpha,\beta^T)^T$, the Lagrange multiplier λ in (2.28) has the explicit solution $\lambda = \rho_{n_1}$, so the log-likelihood (2.27) can be simplified to

$$l = \sum_{i=n_0+1}^{n} \{\alpha + r(\beta,X_i)\} - \sum_{i=1}^{n} \log[1+\rho_n \exp\{\alpha + r(\beta,X_i)\}].$$

On the other hand, for any measurable $p\times 1$ vector valued function $\eta(x)$ which may depend on $(\alpha,\beta^T)^T$, we have

$$\int \eta(x)\ dH(x) = \int \eta(x) m(\theta,x)\ dG(x).$$

So, we can consider the class of unbiased estimating functions

$$\mathcal{Q} = \left\{ Q_n(\eta,\theta)\ /\ Q_n(\eta,\theta) = \sum_{i=n_0+1}^{n} \eta(\alpha,\beta,X_i) - \sum_{i=1}^{n_0} \rho_n \eta(\alpha,\beta,X_i)\exp\{\alpha + r(\beta,X_i)\} \right\}.$$

The maximum likelihood estimating function $\partial l/\partial\theta$ is equivalent to take

$$\eta(x) = \xi(x) = \frac{1}{1+\rho_n m(\theta,x)} \frac{\partial \log m(\theta,x)}{\partial\theta}.$$

For any measurable function $\eta(x)$, denote $(\widehat{\alpha}_\eta, \widehat{\beta}_\eta)$ the solution of the equation

$$Q_n(\eta,\theta) = 0, \qquad \theta^T = (\alpha,\beta^T).$$

Qin (1998) shows that, under certain regularity conditions,

$$\sqrt{n}\begin{pmatrix} \widehat{\alpha}_\eta - \alpha_0 \\ \widehat{\beta}_\eta - \beta_0 \end{pmatrix} \longrightarrow \mathcal{N}(0, A_\eta)$$

in distribution. He shows also, that the maximum likelihood estimating equation $Q_n(\xi, \theta) = 0$ is optimal in the sense of Godambe (see Godambe (1960) and Godambe and Heyde (1987)) in the class $\mathcal{Q}$, in that $A_\eta - A_\xi$ is positive semidefinite for any measurable function $\eta(x)$ satisfying some conditions (see Qin (1998)).

Now, it is easy to show that in the multiplicative-intercept risk model the semiparametric empirical likelihood estimator is the SD$\phi^\star_\rho$DE corresponding to the choice

$$\varphi^\star_\rho(x) := x\log x - \frac{1+\rho x}{\rho}\log(1+\rho x) + \frac{1+\rho}{\rho}\log(1+\rho) - \left(\frac{\rho-1}{\rho} - \log(1+\rho)\right)(x-1).$$

So, for the multiplicative-intercept risk model, our method covers the two-sample semiparametric empirical likelihood's.

When α is known, in order to obtain the SMLE, we have to solve the system of estimating equations (2.28) and the explicit form of the solution λ is not easy to obtain ; moreover, in general λ is not constant, i.e., depends on the data. So in this case the SMLE does not belong to the class of unbiased estimating functions $\mathcal{Q}$, and hence it is not clear if it is still optimal.

In Section 5, we consider an example for the multiplicative-intercept risk model $m(\theta, x) = \exp\{\alpha + r(\beta, x)\}$ with $\alpha = 0$ known. By simulation, we compare the finite-sample efficiency of various SDϕDE's and SMLE. The finite-sample variance and MSE of SMLE is not optimal for small and medium sample sizes (see figures (2.5) and (2.6)), in that there exist SDϕDE's that have smaller variance and MSE.

2.5 Numerical Results

In this Section, we present some simulation results to illustrate the behavior, and compare the efficiency and the robustness of various SDϕDE's and SMLE. We consider the estimates associated to some power-divergences (2.10), to $\phi^\star_{\rho_n}-$divergence and SMLE.

Various examples of the choices of $m(\theta, x)$ can be founded in the papers by Qin (1998), Kay and Little (1987) and Cox and Ferry (1991).
In order to compare the efficiency of estimates, we present Examples 1,2 and 3 (below). Example 4 concerns comparison of robustness of estimates. Examples 1 and 2 have been considered in Qin (1998).

2.5.1 Example 1

Consider the two-sample length bias problem

$$h(x) = \exp(\theta + \log x)g(x), \qquad x > 0.$$

We take h to be a Gamma distribution with parameters (2,1) and g a standard exponential distribution. This is an example of multiplicative-intercept risk model $m(\theta, x) = \exp\{\alpha + r(\theta, x)\}$ with $r(\theta, x) = \log x$.
We have $\theta_0 = 0$. We compute the SDϕDE's for some power divergences ϕ_γ and for the $\phi_\rho^\star-$divergence using samples of size $n_0 = n_1$ =10, 20, 30, 40, 50, 60, 70, 80, 90, 100. In figures (2.1), (2.2) and (2.3) are reported the averages of the 1000 estimates, the variances and the MSE of the 1000 replications for the SDϕDE's considered.

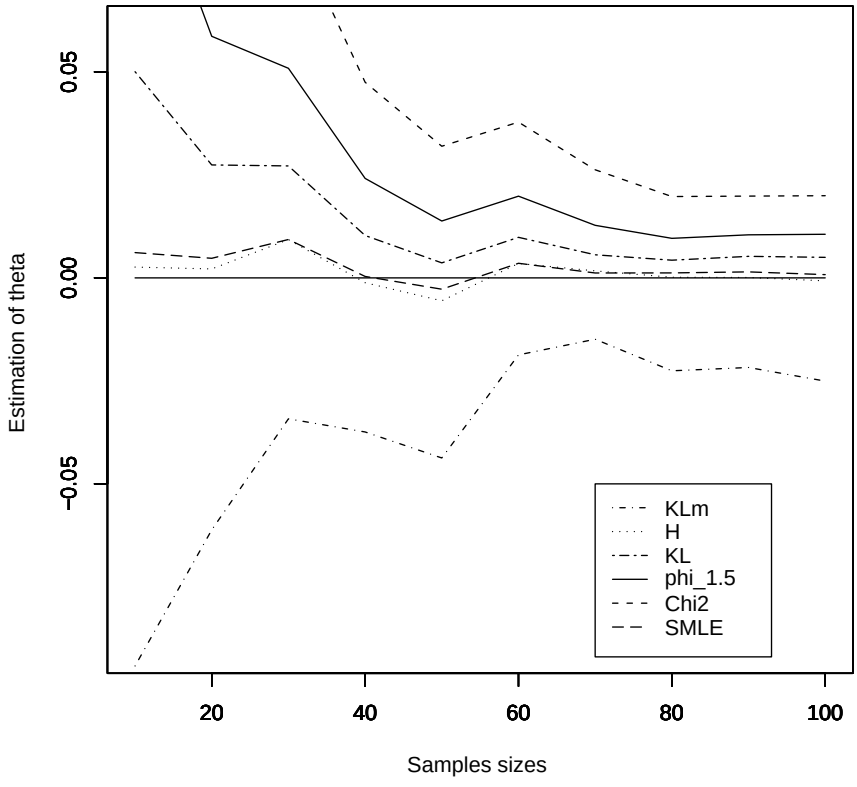

FIG. 2.1 – Estimation of θ in Example 1.

We can see from figure (2.1) that all the estimates converge in a satisfactory way, and in particular the SMLE and the SDHE (the estimate associated to the Hellinger

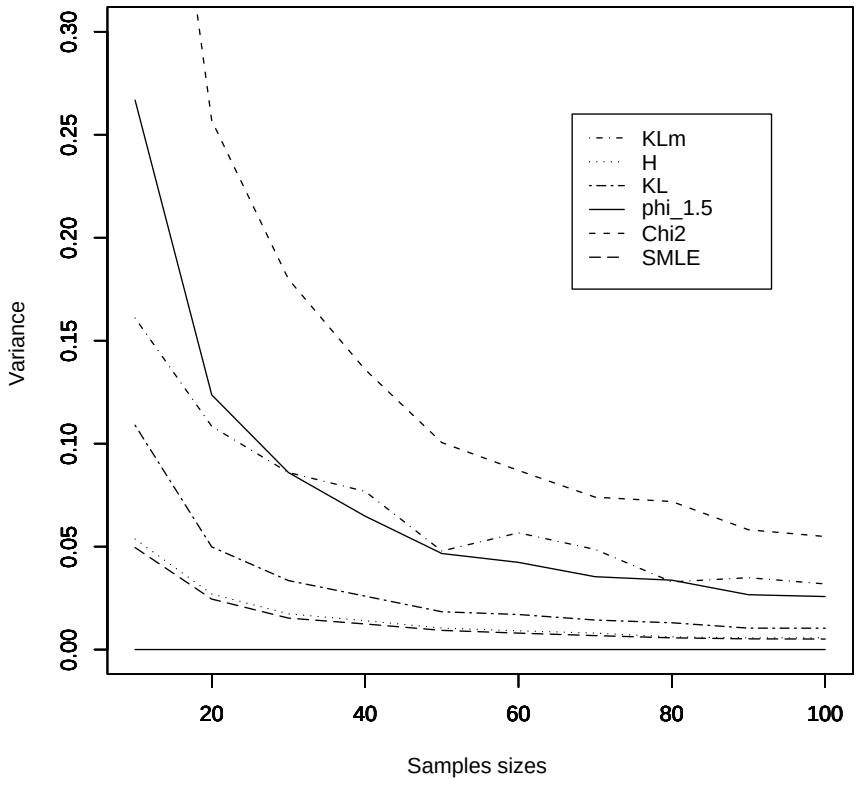

FIG. 2.2 – Variance in Example 1.

divergence). Moreover, they have the smallest finite-sample variance and MSE (see figures (2.2) and (2.3)).

2.5.2 Example 2

Consider the multiplicative-intercept risk model with

$$m(\theta, x) = \exp\{\theta_1 + \theta_2 \log x + \theta_3 \log(1 - x)\}.$$

In the following simulation study, we separately generated uniform $U(0, 1)$ variables for the control group and $Be(2, 2)$ variables for the cases, respectively. Therefore, the true value of $(\theta_1, \theta_2, \theta_3)$ is $(1.792, 1.000, 1.000)$. We compute the SDϕDE's for some power divergences ϕ_γ and for the $\phi_\rho^\star$ divergence using samples of size $n_0 = n_1$ =50, 100, 200, 500. In the table (2.1) are presented MonteCarlo estimates, variances and MSE for the SDϕDE's considered, again with the number of replications equal to

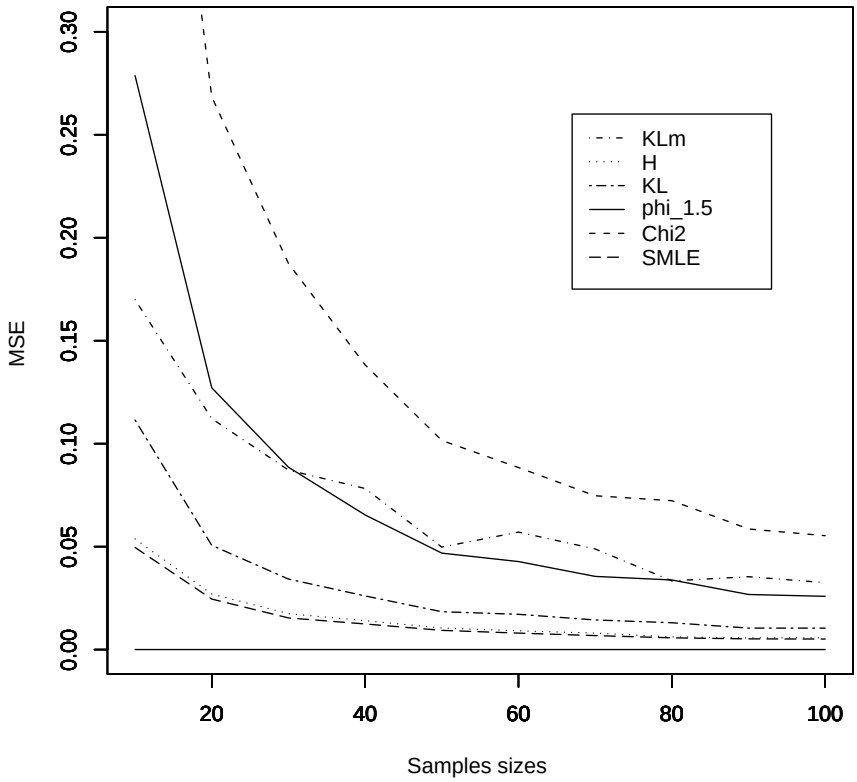

FIG. 2.3 – MSE in Example 1.

1000. We can see that the bias decreases for large sample sizes. For this example the SMLE $= SD\phi^{\star}_{\rho}$DE is clearly the best among the presented estimates.

2.5.3 Example 3

Consider the problem in which

$$h(x) = \frac{1 + (x-\theta)^2}{1 + x^2} g(x). \tag{2.29}$$

We take g to be a standard Cauchy distribution and h a Cauchy distribution with location 20.
Note that this is an example of multiplicative-intercept risk model $m(\theta, x) = \exp\{\alpha + r(\theta, x)\}$ with α known, $\alpha = 0$ and $r(\theta, x) = \log(1 + (x-\theta)^2) - \log(1 + x^2)$.
We remember that for such models it is not clear which is the optimal estimate (see

	$\phi_0 := KL_m$								
n_0, n_1	$\widehat{\theta}_1$	Var	MSE	$\widehat{\theta}_2$	Var	MSE	$\widehat{\theta}_3$	Var	MSE
50	9.4457	2407.26	2465.84	4.7197	529.78	543.61	4.8074	596.32	610.82
100	2.7675	1.9380	2.8896	1.5289	0.6483	0.9280	1.5332	0.5988	0.8831
200	2.3356	0.5862	0.8816	1.2921	0.1893	0.2747	1.3003	0.2051	0.2953
500	2.0754	0.2085	0.2888	1.1518	0.0685	0.0916	1.1563	0.0744	0.0989
	$\phi_{0.5} := H$								
n_0, n_1	$\widehat{\theta}_1$	Var	MSE	$\widehat{\theta}_2$	Var	MSE	$\widehat{\theta}_3$	Var	MSE
50	2.0979	0.8823	0.9759	1.1865	0.3440	0.3788	1.1919	0.3334	0.3702
100	1.9883	0.3709	0.4094	1.1152	0.1438	0.1570	1.1218	0.1383	0.1532
200	1.8888	0.1803	0.1896	1.0550	0.0652	0.0682	1.0631	0.0696	0.0736
500	1.8388	0.0712	0.0734	1.0276	0.0259	0.0267	1.0296	0.0276	0.0285
	$\phi_1 := KL$								
n_0, n_1	$\widehat{\theta}_1$	Var	MSE	$\widehat{\theta}_2$	Var	MSE	$\widehat{\theta}_3$	Var	MSE
50	1.9539	0.7909	0.8171	1.0961	0.3008	0.3101	1.1068	0.2989	0.3103
100	1.9100	0.3355	0.3494	1.0681	0.1302	0.1349	1.0751	0.1244	0.1300
200	1.8384	0.1739	0.1761	1.0239	0.0623	0.0629	1.0344	0.0667	0.0679
500	1.8196	0.0678	0.0686	1.0165	0.0248	0.0251	1.0181	0.0260	0.0263
	$\phi_\rho^\star \equiv SMLE$								
n_0, n_1	$\widehat{\theta}_1$	Var	MSE	$\widehat{\theta}_2$	Var	MSE	$\widehat{\theta}_3$	Var	MSE
50	1.9700	0.7353	0.7670	1.1067	0.2828	0.2942	1.1144	0.2760	0.2891
100	1.9114	0.3194	0.3336	1.0690	0.1234	0.1281	1.0756	0.1188	0.1245
200	1.8422	0.1606	0.1632	1.0265	0.0580	0.0587	1.0358	0.0617	0.0630
500	1.8173	0.0636	0.0642	1.0152	0.0232	0.0235	1.0167	0.0244	0.0247

TAB. 2.1 – Estimation in Example 2.

Qin (1998)).
We have $\theta_0 = 20$. We compute the SDϕDE's for some power divergences ϕ_γ and for the $\phi_\rho^\star$-divergence using samples of size 25, 50, 100, 200. In figures (2.4), (2.5) and (2.6) are presented MonteCarlo estimates, variances and MSE (based on 1000 replications) for the SDϕDE's considered.

We see from figures (2.5) and (2.6), that for finite sample sizes the SDϕ_{-2}E is better than the SMLE and all SDϕDE's considered. It has the smallest variance and MSE.

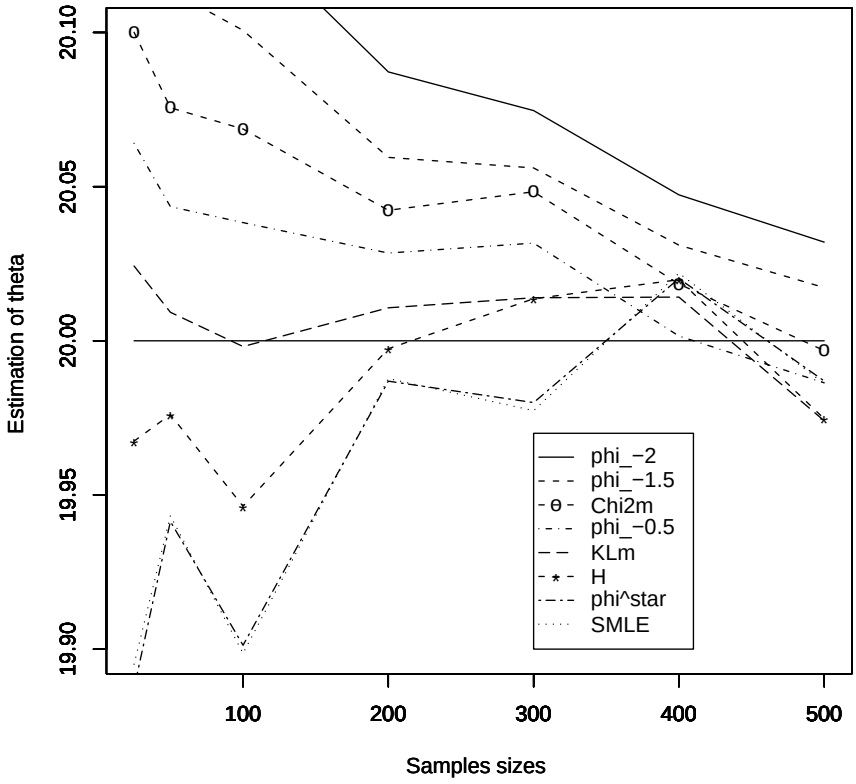

FIG. 2.4 – Estimation of θ in Example 3.

2.5.4 Example 4

Consider the two-sample length bias problem $h(x) = \exp(\theta + \log x)g(x)$, $x > 0$. We take h to be a Gamma distribution with parameters (2,1) and g a standard exponential distribution. We have $\theta_0 = 0$, and we consider $n_0 = n_1 = 100$. We compute the SDϕDE's for some power divergences ϕ_γ and for the $\phi_\rho^\star$-divergence using contaminated data

$$X_i \sim (1-\varepsilon_0)g + \varepsilon_0\delta_{10}, \qquad Y_i \sim (1-\varepsilon_1)h + \varepsilon_1\delta_{15},$$

where $i = 1, \ldots, 200$ and $\varepsilon_0, \varepsilon_1 = \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

In figures (2.7) and (2.8) are presented the contour plots of the MonteCarlo estimates and MSE of the SDϕDE's considered. We can see that the estimate associated to the modified Kullback-Leibler divergence (KL_m) is the most robust when both samples are contaminated. In particular it is almost stable with respect to contamination of the sample X.

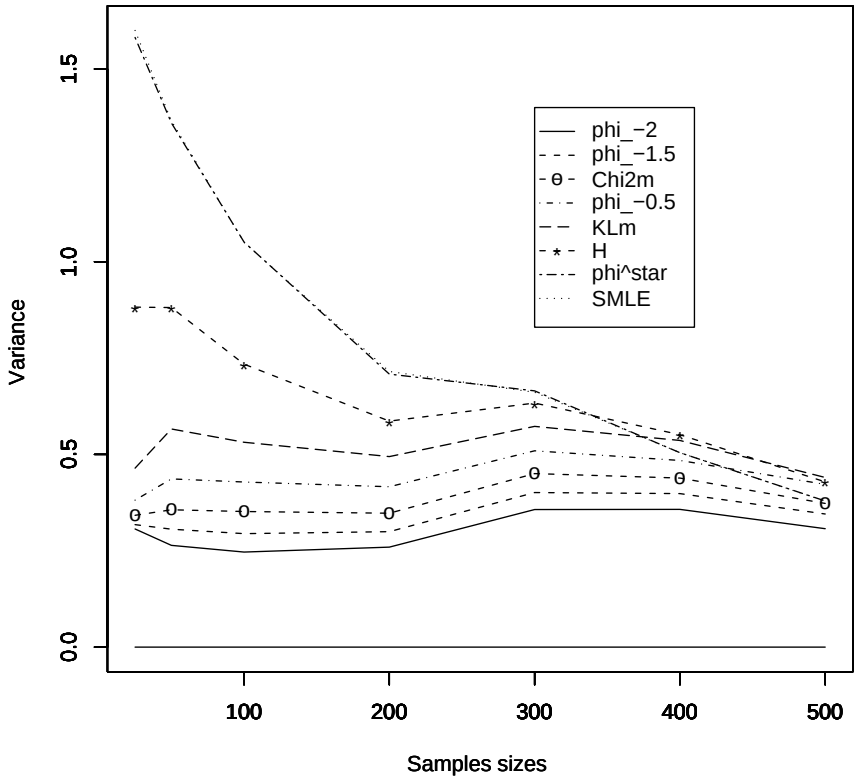

FIG. 2.5 – Variance in Example 3.

2.6 Concluding remarks

We have presented a new method of estimation and tests in semiparametric two-sample density ratio models. This method is based on $\phi-$divergences between probability measures.

In the case of the multiplicative-intercept risk model with intercept unknown, our method covers the semiparametric maximum likelihood one, choosing the $\phi_\rho^\star$-divergence.

For any semiparametric model $m(\theta, \cdot)$, computation of the proposed SDϕDE's is easier to perform then the computation of SMLE which makes use of the Lagrange multiplier. Moreover, our method, contrary to semiparametric maximum likelihood method, applies to the paired-sample case. From our limited simulation results, it seems that the choice of the best estimate in the class of SDϕDE's (in the sense of finite-sample or infinite sample efficiency or robustness) depends on the model, but

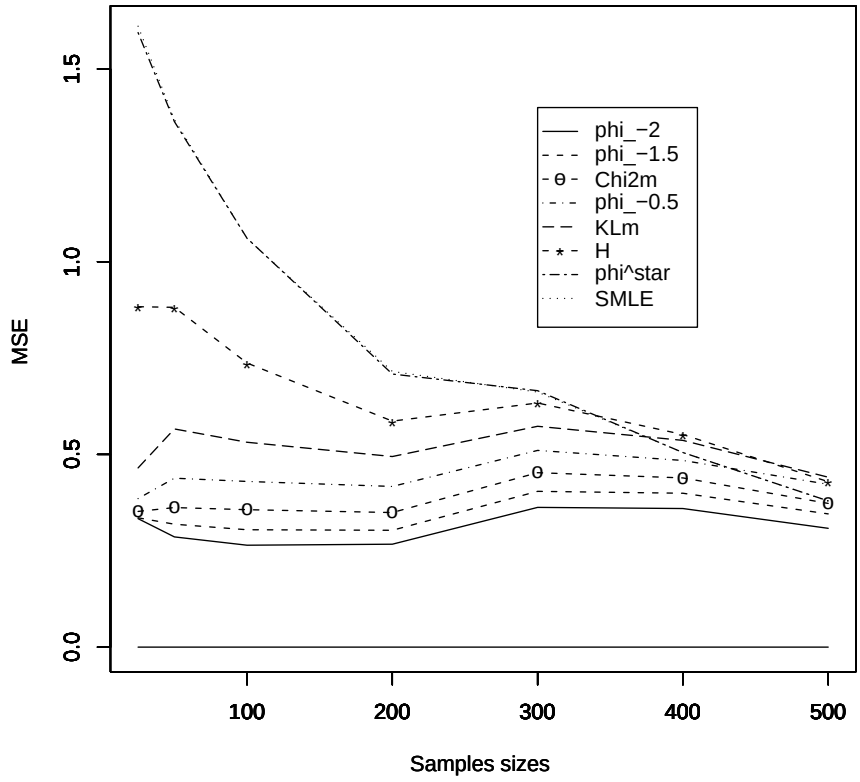

FIG. 2.6 – MSE in Example 3.

we did not give theoretical results about the way of obtaining it.

Also for statistical tests problems it would be interesting to give how to obtain the statistics $\widehat{\phi}_{\rho_n}$ that leads to the most powerful (or robust) test.
Our method can be generalized to corresponding problems involving more than two samples.

These developments will be reported in future communications.

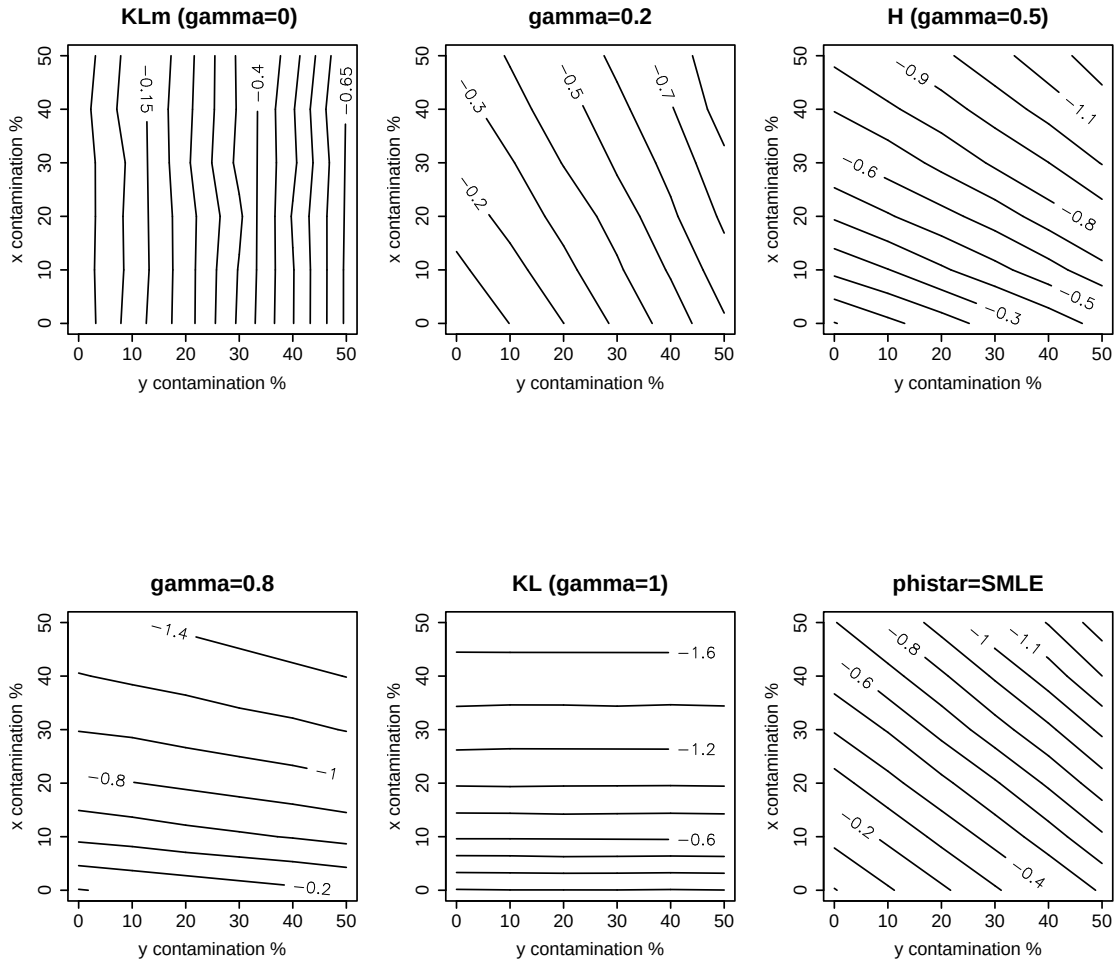

FIG. 2.7 – Contour plots of estimation of θ for contaminated data.

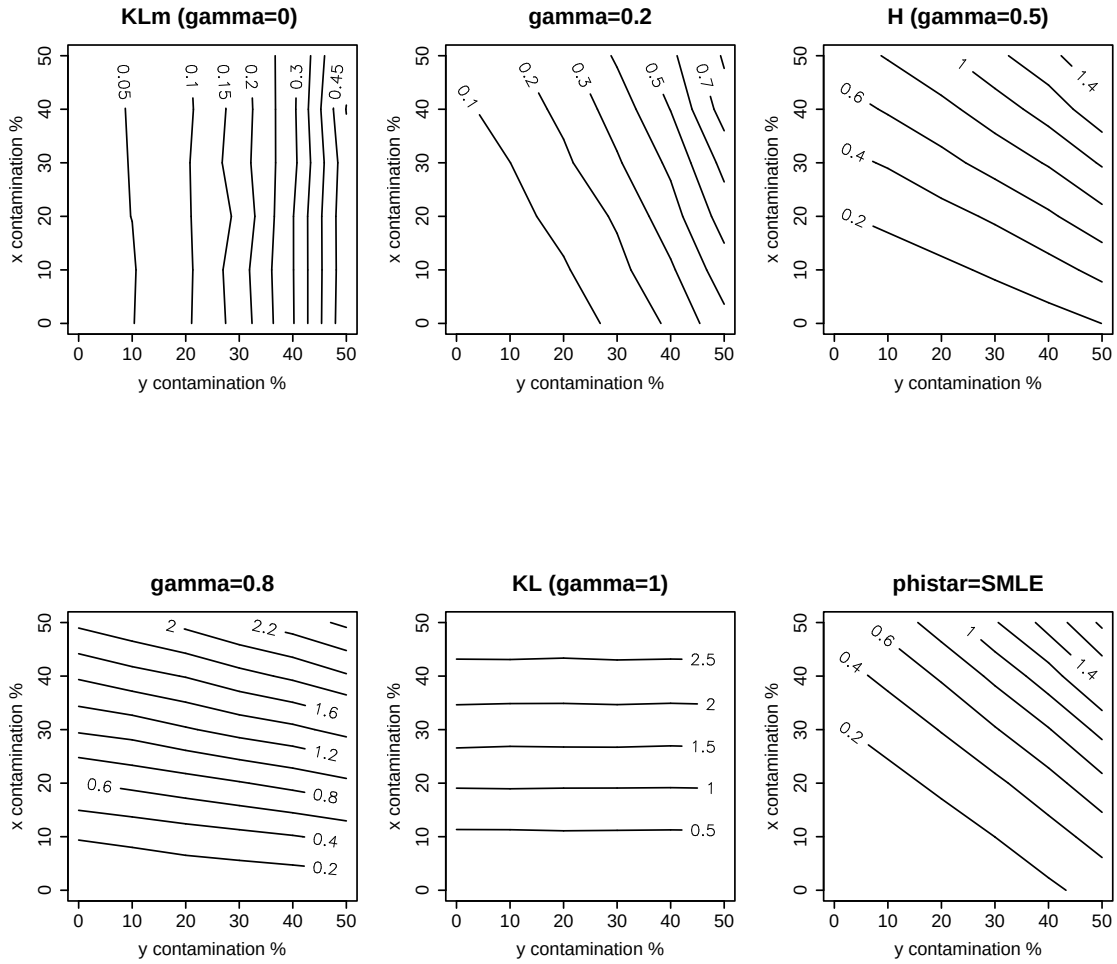

FIG. 2.8 – Contour plots of MSE for contaminated data.

2.7 Proofs

2.7.1 Proof of Theorem 2.2

We will prove that, under conditions (C.2) and (C.3), the estimates $\tilde{\theta}_{\rho_n}$ and $\tilde{\phi}_{\rho_n}(H,G)$ are consistent, which by condition (C.4) implies that for all n_0, n_1 sufficiently large, we have $\tilde{\theta}_{\rho_n} = \widehat{\theta}_{\rho_n}$ and $\tilde{\phi}_{\rho_n}(H,G) = \widehat{\phi}_{\rho_n}(H,G)$.

(a) Since $\tilde{\theta}_{\rho_n}$ is an $M-$estimator, we can apply Theorem 5.7, van der Vaart (1998).

(b) We have

$$\left|\tilde{\phi}_{\rho_n}(H,G) - \phi_\rho(H,G)\right| = \left|H^Y_{n_1} k_{\rho_n}(\tilde{\theta}_{\rho_n}) - G^X_{n_0} l_{\rho_n}(\tilde{\theta}_{\rho_n}) - (Hk_\rho(\theta_0) - Gl_\rho(\theta_0))\right| =: A.$$

By the very definition of $\tilde{\theta}_{\rho_n}$ and (2.18) we obtain

$$\begin{aligned} &H^Y_{n_1} k_{\rho_n}(\theta_0) - G^X_{n_0} l_{\rho_n}(\theta_0) - (Hk_\rho(\theta_0) - Gl_\rho(\theta_0)) \le A \le \\ &\le H^Y_{n_1} k_{\rho_n}(\tilde{\theta}_{\rho_n}) - G^X_{n_0} l_{\rho_n}(\tilde{\theta}_{\rho_n}) - \left(Hk_\rho(\tilde{\theta}_{\rho_n}) - Gl_\rho(\tilde{\theta}_{\rho_n})\right), \end{aligned}$$

and both the RHS and the LHS terms go to 0 under condition (C.2).

2.7.2 Proof of Theorem 2.3

Proof of (1). Under condition (C.7), some calculus yield

$$Hk'_\rho(\theta_0) - Gl'_\rho(\theta_0) = 0 \tag{2.30}$$

and

$$Hk''_\rho(\theta_0) - Gl''_\rho(\theta_0) = -\int \dot{m}(\theta_0, x) \cdot [\dot{m}(\theta_0, x)]^T \cdot \varphi''_\rho(m(\theta_0, x))\, dG(x) \tag{2.31}$$

which implies that the matrix $Hk''_\rho(\theta_0) - Gl''_\rho(\theta_0)$ is symmetric.

By Taylor expansion, using condition (C.6), there exists $\overline{\theta}_{\rho_n}$ inside the segment that links θ_0 and $\widehat{\theta}_{\rho_n}$ with

$$\begin{aligned} 0 &= H^Y_{n_1} k'_{\rho_n}(\widehat{\theta}_{\rho_n}) - G^X_{n_0} l'_{\rho_n}(\widehat{\theta}_{\rho_n}) \\ &= H^Y_{n_1} k'_{\rho_n}(\theta_0) - G^X_{n_0} l'_{\rho_n}(\theta_0) + \left(H^Y_{n_1} k''_{\rho_n}(\theta_0) - G^X_{n_0} l''_{\rho_n}(\theta_0)\right)^T (\widehat{\theta}_{\rho_n} - \theta_0) \\ &\quad + \frac{1}{2}(\widehat{\theta}_{\rho_n} - \theta_0)^T \cdot \left[H^Y_{n_1} k'''_{\rho_n}(\overline{\theta}_{\rho_n}) - G^X_{n_0} l'''_{\rho_n}(\overline{\theta}_{\rho_n})\right] \cdot (\widehat{\theta}_{\rho_n} - \theta_0), \end{aligned} \tag{2.32}$$

where $H^Y_{n_1} k'''_{\rho_n}(\overline{\theta}_{\rho_n})$ and $G^X_{n_0} l'''_{\rho_n}(\overline{\theta}_{\rho_n})$ are $d-$vectors whose entries are $d \times d$ matrices. By (C.6), we have for the sup-norm of vectors and matrices

$$\left\|H^Y_{n_1} k'''_{\rho_n}(\overline{\theta}_{\rho_n}) - G^X_{n_0} l'''_{\rho_n}(\overline{\theta}_{\rho_n})\right\| \le \frac{1}{n_1}\sum_{i=1}^{n_1} \mid K(Y_i) \mid + \frac{1}{n_0}\sum_{i=1}^{n_0} \mid L(X_i) \mid,$$

so by the Law of Large Numbers, we obtain $H^Y_{n_1}k'''_{\rho_n}(\overline{\theta}_{\rho_n}) - G^X_{n_0}l'''_{\rho_n}(\overline{\theta}_{\rho_n}) = O_P(1)$. By (C.5), we have that

$$\frac{1}{2}(\widehat{\theta}_{\rho_n} - \theta_0)^T \cdot \left[H^Y_{n_1}k'''_{\rho_n}(\overline{\theta}_{\rho_n}) - G^X_{n_0}l'''_{\rho_n}(\overline{\theta}_{\rho_n})\right] \cdot (\widehat{\theta}_{\rho_n} - \theta_0) = o_P(1) \cdot (\widehat{\theta}_{\rho_n} - \theta_0)$$

Now

$$H^Y_{n_1}k''_{\rho_n}(\theta_0) - G^X_{n_0}l''_{\rho_n}(\theta_0) = \frac{1}{n_1}\sum_{i=1}^{n_1} k''_{\rho_n}(\theta_0, Y_i) - \frac{1}{n_0}\sum_{i=1}^{n_0} l''_{\rho_n}(\theta_0, X_i)$$

and by the Law of Large Numbers this converges to the matrix $Hk''_\rho(\theta_0) - Gl''_\rho(\theta_0)$, that is

$$H^Y_{n_1}k''_{\rho_n}(\theta_0) - G^X_{n_0}l''_{\rho_n}(\theta_0) = Hk''_\rho(\theta_0) - Gl''_\rho(\theta_0) + o_P(1)$$

so, from (2.32), we get

$$-(H^Y_{n_1}k'_{\rho_n}(\theta_0) - G^X_{n_0}l'_{\rho_n}(\theta_0)) = \left(Hk''_\rho(\theta_0) - Gl''_\rho(\theta_0) + o_P(1)\right)(\widehat{\theta}_{\rho_n} - \theta_0) \qquad (2.33)$$

From this, under (C.8), using the Central Limit Theorem, we obtain $\sqrt{n}(H^Y_{n_1}k'_{\rho_n}(\theta_0) - G^X_{n_0}l'_{\rho_n}(\theta_0)) = O_P(1)$. Using (2.33) we obtain

$$\sqrt{n}(\widehat{\theta}_{\rho_n} - \theta_0) = O_P(1), \qquad (2.34)$$

that is $(\widehat{\theta}_{\rho_n} - \theta_0) = O_P(1/\sqrt{n})$.
Now, from (2.33), using (2.34), it follows

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_{\rho_n} - \theta_0) &= \left[\left(-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right)\right]^{-1} \cdot \left(\frac{1}{\sqrt{\rho_{n_1}}} \cdot \frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1} k'_{\rho_n}(\theta_0, Y_i)\right. \\ &\quad \left. - \frac{1}{\sqrt{\rho_{n_0}}} \cdot \frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0} l'_{\rho_n}(\theta_0, X_i)\right) + o_P(1). \qquad (2.35)\end{aligned}$$

Under (C.8), the Central Limit Theorem concludes the proof of (1), and the Limit Covariance Matrix is :

$$\begin{aligned}LCM &= \left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right]^{-1} \cdot \left\{\rho_1^{-1} \cdot \left(Hk'_\rho(\theta_0)\left[k'_\rho(\theta_0)\right]^T - Hk'_\rho(\theta_0)\left[Hk'_\rho(\theta_0)^T\right]\right) + \right. \\ &\quad \left. + \rho_0^{-1} \cdot \left(Gl'_\rho(\theta_0)\left[l'_\rho(\theta_0)\right]^T - Gl'_\rho(\theta_0)\left[Gl'_\rho(\theta_0)\right]^T\right)\right\} \cdot \left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right]^{-1}.\end{aligned}$$

If $H = G$, we have

$$\left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right]^{-1} = \left[\int \dot{m}(\alpha_0, x) \cdot [\dot{m}(\alpha_0, x)]^T \cdot \varphi''_\rho(1)\, dG(x)\right]^{-1} =: A^{-1}.$$

Now,

$$k'_\rho(\alpha_0) = l'_\rho(\alpha_0) = \varphi''_\rho(1) \cdot \dot{m}(\alpha_0, x)$$

and under (C.7), it holds $Hk'_\rho(\alpha_0) = Gl'_\rho(\alpha_0) = 0$. So, we obtain

$$\rho_1^{-1} \cdot Hk'_\rho(\alpha_0) \left[k'_\rho(\alpha_0)\right]^T + \rho_0^{-1} \cdot Gl'_\rho(\alpha_0) \left[l'_\rho(\alpha_0)\right]^T = (\rho_1\rho_0)^{-1} \cdot \varphi''_\rho(1) \cdot A =: B,$$

since $\rho_0 + \rho_1 = 1$. Finally, under the hypothesis $H = G$, the Covariance Matrix is

$$LCM_0 = A^{-1} \cdot B \cdot A^{-1} = \left[\rho_1\rho_0 \cdot G\left(\dot{m}(\alpha_0, x) \cdot [\dot{m}(\alpha_0, x)]^T\right)\right]^{-1}.$$

Proof of (2). By Taylor expansion, there exists $\overline{\theta}_{\rho_n}$ inside the segment that links θ_0 and $\widehat{\theta}_{\rho_n}$ with

$$\begin{aligned}
\widehat{\phi}_{\rho_n}(H, G) &= H^Y_{n_1} k_{\rho_n}(\widehat{\theta}_{\rho_n}) - G^X_{n_0} l_{\rho_n}(\widehat{\theta}_{\rho_n}) \\
&= H^Y_{n_1} k_{\rho_n}(\theta_0) - G^X_{n_0} l_{\rho_n}(\theta_0) + \left[H^Y_{n_1} k'_{\rho_n}(\theta_0) - G^X_{n_0} l'_{\rho_n}(\theta_0)\right]^T \cdot (\widehat{\theta}_{\rho_n} - \theta_0) \\
&\quad + \frac{1}{2} \cdot (\widehat{\theta}_{\rho_n} - \theta_0)^T \cdot \left[H^Y_{n_1} k''_{\rho_n}(\theta_0) - G^X_{n_0} l''_{\rho_n}(\theta_0)\right] \cdot (\widehat{\theta}_{\rho_n} - \theta_0) \\
&\quad + \frac{1}{3!} \sum_{1 \le i,j,k \le d} (\widehat{\theta}_{\rho_n} - \theta_0)_i \cdot (\widehat{\theta}_{\rho_n} - \theta_0)_j \cdot (\hat{\theta}_{n_0,n_1} - \theta_0)_k \\
&\quad \cdot \left[H^Y_{n_1} \frac{\partial^3}{\partial\theta_i \partial\theta_j \partial\theta_k} k_{\rho_n}(\overline{\theta}_{\rho_n}) - G^X_{n_0} \frac{\partial^3}{\partial\theta_i \partial\theta_j \partial\theta_k} l_{\rho_n}(\overline{\theta}_{\rho_n})\right]. \qquad (2.36)
\end{aligned}$$

When $\theta_0 = \alpha_0$, we have : $H^Y_{n_1} k_{\rho_n}(\theta_0) - G^X_{n_0} l_{\rho_n}(\theta_0) = 0$. We have already shown that $(\widehat{\theta}_{\rho_n} - \theta_0) = O_P\left(\frac{1}{\sqrt{n}}\right)$. Moreover, by condition (C.5) we know that $(\widehat{\theta}_{\rho_n} - \theta_0) = o_P(1)$, so the last term in (2.36) is $o_P\left(\frac{1}{n}\right)$. So, we can write

$$\begin{aligned}
\widehat{\phi}_{\rho_n}(H, G) &= \left[H^Y_{n_1} k'_{\rho_n}(\theta_0) - G^X_{n_0} l'_{\rho_n}(\theta_0)\right]^T \cdot (\widehat{\theta}_{\rho_n} - \theta_0) \\
&\quad + \frac{1}{2} \cdot (\widehat{\theta}_{\rho_n} - \theta_0)^T \cdot \left[H^Y_{n_1} k''_{\rho_n}(\theta_0) - G^X_{n_0} l''_{\rho_n}(\theta_0)\right] \cdot (\widehat{\theta}_{\rho_n} - \theta_0) + o_P\left(\frac{1}{n}\right) \qquad (2.37)
\end{aligned}$$

Now, using formula (2.35) and (2.37), and the fact that

$[H^Y_{n_1}k'_{\rho_n}(\theta_0) - G_{n_0}l'_{\rho_n}(\theta_0)]o_P(1/\sqrt{n}) = o_P(1/n)$, we obtain

$$\begin{aligned}
\widehat{\phi}_{\rho_n}(H,G) &= \left[H^Y_{n_1}k'_{\rho_n}(\theta_0) - G^X_{n_0}l'_{\rho_n}(\theta_0)\right]^T \cdot \left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right]^{-1} \cdot \\
&\quad \left[H^Y_{n_1}k'_{\rho_n}(\theta_0) - G^X_{n_0}l'_{\rho_n}(\theta_0)\right] - \frac{1}{2} \cdot \left[H^Y_{n_1}k'_{\rho_n}(\theta_0) - G^X_{n_0}l'_{\rho_n}(\theta_0)\right]^T \cdot \\
&\quad \left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right]^{-1} \cdot \left[H^Y_{n_1}k'_{\rho_n}(\theta_0) - G^X_{n_0}l'_{\rho_n}(\theta_0)\right] + o_P\left(\frac{1}{n}\right) \\
&= \frac{1}{2} \cdot \left[H^Y_{n_1}k'_{\rho_n}(\theta_0) - G^X_{n_0}l'_{\rho_n}(\theta_0)\right]^T \cdot \left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right]^{-1} \\
&\quad \cdot \left[H^Y_{n_1}k'_{\rho_n}(\theta_0) - G^X_{n_0}l'_{\rho_n}(\theta_0)\right] + o_P\left(\frac{1}{n}\right).
\end{aligned}$$

Then, we have

$$\begin{aligned}
\frac{2n}{\varphi''_{\rho_n}(1)}\widehat{\phi}_{\rho_n}(H,G) &= n \cdot \left[H^Y_{n_1}k'_{\rho_n}(\theta_0) - G^X_{n_0}l'_{\rho_n}(\theta_0)\right]^T \cdot \frac{\left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right]^{-1}}{\varphi''_\rho(1)} \\
&\quad \cdot \left[H^Y_{n_1}k'_{\rho_n}(\theta_0) - G^X_{n_0}l'_{\rho_n}(\theta_0)\right] + o_P(1) = \qquad (2.38) \\
&\quad =: nS^TQ^{-1}S + o_P(1) = \sqrt{n}S^TA^{-\frac{1}{2}}\sqrt{n}A^{-\frac{1}{2}}S + o_P(1).
\end{aligned}$$

When $\theta_0 = \alpha_0$, we have

$$\sqrt{n}\left[H^Y_{n_1}k'_{\rho_n}(\theta_0) - G^X_{n_0}l'_{\rho_n}(\theta_0)\right] \to \mathcal{N}(0,V), \qquad (2.39)$$

where

$$\begin{aligned}
V &= \frac{1}{\rho_1} \cdot Hk'_\rho(\theta_0)\left[k'_\rho(\theta_0)\right]^T + \frac{1}{\rho_0} \cdot Gl'_\rho(\theta_0)\left[l'_\rho(\theta_0)\right]^T \\
&= \left(\frac{1}{\rho_1} + \frac{1}{\rho_0}\right) \cdot \varphi''_\rho(1) \cdot \left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right].
\end{aligned}$$

So we can write $\left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right]^{-1} = (\rho_1\rho_0)^{-1} \cdot \varphi''_\rho(1) \cdot V^{-1}$, and substituting this in (2.38), we get

$$\begin{aligned}
\frac{2n\rho_0\rho_1}{\varphi''_{\rho_n}(1)}\widehat{\phi}_{\rho_n}(H,G) &= n \cdot \left[H^Y_{n_1}k'_{\rho_n}(\theta_0) - G^X_{n_0}l'_{\rho_n}(\theta_0)\right]^T \cdot V^{-1} \\
&\quad \cdot \left[H^Y_{n_1}k'_{\rho_n}(\theta_0) - G^X_{n_0}l'_{\rho_n}(\theta_0)\right] + o_P(1),
\end{aligned}$$

and from (2.39) we obtain (under the hypothesis $H = G$)

$$\frac{2n\rho_{n_0}\rho_{n_1}}{\varphi''_{\rho_n}(1)}\widehat{\phi}_{\rho_n}(H,G) \to \chi^2(d),$$

where $d = dim(\Theta)$.

Proof of (3).

By Taylor expansion, there exists $\overline{\theta}_{\rho_n}$ inside the segment that links θ_0 and $\widehat{\theta}_{n_0,n_1}$ with

$$\begin{aligned}\widehat{\phi}_{\rho_n}(H,G) &= H^Y_{n_1}k_{\rho_n}(\widehat{\theta}_{\rho_n}) - G^X_{n_0}l_{\rho_n}(\widehat{\theta}_{\rho_n}) = \\ &= H^Y_{n_1}k_{\rho_n}(\theta_0) - G^X_{n_0}l_{\rho_n}(\theta_0) + o_P(1)\cdot(\widehat{\theta}_{\rho_n}-\theta_0) \\ &\quad + \frac{1}{2}\cdot(\widehat{\theta}_{\rho_n}-\theta_0)^T\cdot\left[Hk''_\rho(\theta_0) - Gl''_\rho(\theta_0) + o_P(1)\right]\cdot(\widehat{\theta}_{\rho_n}-\theta_0) + o_P\left(\frac{1}{n}\right).\end{aligned}$$

So

$$\begin{aligned}&\sqrt{n}\cdot\left(\widehat{\phi}_{\rho_n}(H,G) - \phi_\rho(H,G)\right) = \\ &= \sqrt{n}\left(H^Y_{n_1}k_{\rho_n}(\theta_0) - G^X_{n_0}l_{\rho_n}(\theta_0) - Hk_\rho(\theta_0) + Gl_\rho(\theta_0)\right) + o_P(1). \qquad (2.40)\end{aligned}$$

Finally, we have

$$\sqrt{n}\cdot\left(\widehat{\phi}_{\rho_n}(H,G) - \phi_\rho(H,G)\right) \to \mathcal{N}(0,W),$$

where

$$W = \rho_1^{-1}\left(Hk^2_\rho(\theta_0) - (Hk_\rho(\theta_0))^2\right) + \rho_0^{-1}\left(Gl^2_\rho(\theta_0) - (Gl_\rho(\theta_0))^2\right).$$

This concludes the proof of Theorem (2.3).

2.7.3 Proof of Theorem 2.4

Proof of (1). Formula (2.35) still holds; under (C.8), applying the Central Limit Theorem, we obtain the Limit Covariance Matrix $LCM(dep)$. Calculations give $LCM_0(dep)$.

Proof of (2). Formula (2.38) still holds. When $\theta_0 = \alpha_0$, we have

$$\sqrt{n}\left[H^Y_{n_1}k'_{\rho_n}(\theta_0) - G^X_{n_0}l'_{\rho_n}(\theta_0)\right] \to \mathcal{N}(0, V_{dep})$$

where

$$V_{dep} = \frac{\varphi''_\rho(1)}{\rho_0\rho_1}\cdot\left[-Hk''_\rho(\theta_0) + Gl''_\rho(\theta_0)\right] - 2(\varphi''_\rho(1))^2\frac{\min(1,\rho)}{\rho_1}F_{H,H}\dot{m}(\theta_0,x)(\dot{m}(\theta_0,x))^T$$

So, we have that $\sqrt{n}S^TA^{-\frac{1}{2}}$ in (2.38) converges to a $\mathcal{N}(0,M)$, where $M = A^{-\frac{1}{2}}V_{dep}A^{-\frac{1}{2}}$.

Proof of (3). Formula (2.40) still holds, and calculations give (3).

Chapitre 3

Estimation and Tests for Models satisfying Linear Constraints with Unknown Parameter

We introduce estimation and test procedures through divergence projections for models satisfying linear constraints with unknown parameter. Several statistical examples and motivations for these models are given. These procedures extend the empirical likelihood method. We treat the problems of existence and characterization of the divergence projections of probability measures on sets of signed finite measures. The asymptotic behavior of the proposed estimates and statistics are studied using the dual representation of divergences and the explicit forms of the divergence projections. We discuss the comparison problem of the procedures including the empirical likelihood one. Efficiency and robustness properties are discussed. A simulation study shows that the Hellinger divergence enjoys good efficiency and robustness properties.

3.1 Introduction and notation

A model satisfying partly specified linear parametric constraints is a family of distributions $\mathcal{M}^1$ all defined on a same measurable space $(\mathcal{X}, \mathcal{B})$, such that, for all Q in $\mathcal{M}^1$, the following condition holds

$$\int g(x, \theta)\ dQ(x) = 0.$$

The unspecified parameter θ belongs to Θ, an open set in $\mathbb{R}^d$. The function $g := (g_1, ..., g_l)^T$ is defined on $\mathcal{X} \times \Theta$ with values in $\mathbb{R}^l$, each of the g_i's being real valued and the functions $g_1, \ldots, g_l, \mathbb{1}_{\mathcal{X}}$ are assumed linearly independent. So $\mathcal{M}^1$ is defined

through l-linear constraints indexed by some $d-$dimensional parameter θ. Denote M^1 the collection of all probability measures on $(\mathcal{X}, \mathcal{B})$, and

$$\mathcal{M}_\theta^1 := \left\{ Q \in M^1 \text{ such that } \int g(x,\theta)\, dQ(x) = 0 \right\}$$

so that

$$\mathcal{M}^1 = \bigcup_{\theta \in \Theta} \mathcal{M}_\theta^1. \tag{3.1}$$

Assume now that we have at hand a sample $X_1, ..., X_n$ of independent random variables (r.v.'s) with common unknown distribution P_0. When P_0 belongs to the model (3.1), we denote θ_0 the value of the parameter θ such that $\mathcal{M}_{\theta_0}$ contains P_0. Obviously, we assume that θ_0 is unique.

The scope of this Chapter is to propose new answers for the following problems

Problem 1 : Does P_0 belong to the model $\mathcal{M}^1$?

Problem 2 : When P_0 is in the model, which is the value θ_0 of the parameter for which $\int g(x,\theta_0)\, dP_0(x) = 0$? Also can we perform simple and composite tests for θ_0 ? Can we construct confidence areas for θ_0 ? Can we give more efficient estimates for the distribution function than the usual cumulative distribution function (c.d.f.) ?

We present some examples and motivations for the models (3.1) and Problems 1 and 2.

3.1.1 Statistical examples and motivations

Example 3.1. *Sometimes we have information relating the first and second moments of a random variable X (see e.g., Godambe and Thompson (1989) and McCullagh and Nelder (1983)). Assume that the second moment of X is $m(\theta_0)$ when its expectation is θ_0, with $m(.)$ some known function. Problem 1 then writes : On the basis of an i.i.d. sample of r.v.'s with same distribution as X, can we assess that some θ_0 exists for which such model holds ? The second problem deals with some estimation and test pertaining to the parameter θ_0 if Problem 1 has a positive issue. Let $\mathcal{M}_\theta^1$ be the set of all p.m.'s Q on $\mathbb{R}$ satisfying $\int (x-\theta)\, dQ(x) = 0$ and $\int (x^2 - m(\theta))\, dQ(x) = 0$, and $\mathcal{M}^1$ be defined as in (3.1), for some subset Θ in the domain of $m(.)$. The function $g(x,\theta)$ equals $(x-\theta, x^2 - m(\theta))^T$. A similar case is when $\mathcal{M}^1$ is the class of all p.m.'s with k moments in some specified subset of $\mathbb{R}^k$.*

Example 3.2. *Consider an i.i.d. bivariate sample (X_i, Y_i), $1 \leq i \leq n$, and assume that $E(X_1) = c$ with c some known number. Suppose that we are interested in the estimation of $\theta_0 = E(Y_1)$. Clearly we are in the same case as previously, with Problem 2 defined by $\mathcal{M}_\theta^1$ the set of all p.m.'s Q on $\mathbb{R}^2$ satisfying $\int g(x,y,\theta)\, dQ(x,y) = 0$ with $g(x,y,\theta) = (x-c, y-\theta)^T$ and $\mathcal{M}^1$ as in (3.1) where Θ is the range of θ's.*

Such problems are common in survey sampling (see e.g., Kuk and Mak (1989) and Chen and Qin (1993)). A similar problem is when $E(X_i) = E(Y_i) = \theta_0$ *by taking* $g((x,y)^T, \theta) = (x-\theta, y-\theta)^T$.

Example 3.3. *Several authors have considered nonparametric estimation of a distribution function* F *when information about certain functionals is available. Haberman (1984) and Sheehy (1987) consider estimation of* $F(x)$ *based on i.i.d. sample* $X_1, \ldots, X_n$ *when it is known that* $\int T(x)\ dF(x) = a$, *for some specified function* $T(\cdot)$. *For this problem, the function* $g(x,\theta)$ *in the model (3.1) is equal to* $T(x) - \theta$ *with* $\theta = a$ *is known.*

The above three examples can be found in Qin and Lawless (1994). Here two cases in connection with classical statistical problems.

Example 3.4. *Suppose that* P_0 *is the distribution of a pair of random variables* (X,Y) *on a product space* $\mathcal{X} \times \mathcal{Y}$ *with known marginal distributions* P_1 *and* P_2. *Bickel* et al. *(1991) study efficient estimation of* $\theta_0 = \int h(x,y)\ dP_0(x,y)$ *for specified function* h. *This problem can be handled in the present context when the spaces* $\mathcal{X}$ *and* $\mathcal{Y}$ *are discrete and finite. Denote* $\mathcal{X} = \{x_1, \ldots, x_k\}$ *and* $\mathcal{Y} = \{y_1, \ldots, y_r\}$. *Consider an i.i.d. bivariate sample* $(X_i, Y_i), 1 \leq i \leq n$ *of the bivariate random variable* (X,Y). *The space* $\mathcal{M}_\theta$ *in this case is the set of all p.m.'s* Q *on* $\mathcal{X} \times \mathcal{Y}$ *satisfying* $\int g(x,y,\theta)\ dQ(x,y) = 0$ *where* $g = (g_1^{(1)}, \ldots, g_k^{(1)}, g_1^{(2)}, \ldots, g_r^{(2)}, g_1)^T$, $g_i^{(1)}(x,y,\theta) = \mathbb{1}_{\{x_i\}\times\mathcal{Y}}(x,y) - P_1(x_i)$, $g_j^{(2)}(x,y,\theta) = \mathbb{1}_{\mathcal{X}\times\{y_j\}}(x,y) - P_2(y_j)$ *for all* $(i,j) \in \{1,\ldots,k\} \times \{1,\ldots,r\}$, *and* $g_1(x,y,\theta) = h(x,y) - \theta$. *Problem 1 turns to be the test for "*P_0 *belongs to* $\bigcup_{\theta\in\Theta} \mathcal{M}_\theta$*", while Problem 2 pertains to the estimation and tests for specific values of* θ. *Motivation and references for this problem are given in Bickel* et al. *(1991) and Bickel* et al. *(1993).*

Example 3.5. *(Generalized linear models). Let* Y *be a random variable and* X *a* l*-dimensional random vector.* Y *and* X *are linked through*

$$Y = m(X, \theta_0) + \varepsilon$$

in which $m(.,.)$ *is some specified real valued function and* θ_0, *the parameter of interest, belongs to some open set* $\Theta \subset \mathbb{R}^d$. ε *is a measurement error. Denote* P_0 *the law of the vector variable* (X,Y) *and suppose that the true value* θ_0 *satisfies the orthogonality condition*

$$\int x(y - m(x,\theta_0))\ dP_0(x,y) = 0.$$

Consider an i.i.d. sample (X_i, Y_i), $1 \leq i \leq n$ *of r.v.'s with same distribution as* (X,Y). *The existence of some* θ *for which the above condition holds is given as*

the solution of Problem 1, while Problem 2 aims to provide its explicit value; here $\mathcal{M}_\theta^1$ *is the set of all p.m.'s* Q *on* $\mathbb{R}^{l+1}$ *satisfying* $\int g(x,y,\theta)\, dQ(x,y) = 0$ *with* $g(x,y,\theta) = x(y - m(x,\theta))$.

Another motivation for our work stems from confidence region (C.R.) estimation techniques. The empirical likelihood method provides such estimation (see Owen (1990)). We will extend this approach providing a wide range of such C.R.'s, each one depending upon a specific criterion, one of those leading to Owen's C.R.

Problems of this kind have been addressed by various authors in the recent literature. The empirical likelihood approach developed by Owen (1988) and Owen (1990) has been adapted in the present setting by Qin and Lawless (1994). Owen (1991) extends the empirical likelihood approach to linear regression models.

The approach which we develop is based on *minimum descrepancy estimates*, which have common features with minimum distance techniques, using merely divergences. We present wide sets of estimates, simple and composite tests and confidence regions for the parameter θ_0 as well as various test statistics for *Problem 1*, all depending on the choice of the divergence. Simulations show that the approach based on Hellinger divergence enjoys good robustness and efficiency properties when handling *Problem 2*. As presented in Section 7, empirical likelihood methods appear to be a special case of the present approach.

3.1.2 Minimum divergence estimates

We first set some general definition and notation. Let P be some probability measure (p.m.). Denote $M^1(P)$ the subset of all p.m.'s which are absolutely continuous (a.c.) with respect to P. Denote M the space of all signed finite measures on $(\mathcal{X}, \mathcal{B})$ and $M(P)$ the subset of all signed finite measures a.c. w.r.t. P. Let φ be a convex function from $[-\infty, +\infty]$ onto $[0, +\infty]$ with $\varphi(1) = 0$. For any signed finite measure Q in $M(P)$, the $\phi-$divergence between Q and the p.m. P is defined through

$$\phi(Q,P) := \int \varphi\left(\frac{dQ}{dP}\right)\, dP. \tag{3.2}$$

When Q is not a.c. w.r.t. P, we set $\phi(Q,P) = +\infty$. This definition extends Rüschendorf (1984)'s one which applies for $\phi-$divergences between p.m.'s ; it also differs from Csiszár (1963)'s one, which requires a common dominating $\sigma-$finite measure, noted λ, for Q and P. Since we will consider subsets of $M^1(P)$ and subsets of $M(P)$, it is more adequate for our sake to use the definition (3.2). Also note that all the just mentioned definitions of $\phi-$divergences coincide on the set of all p.m.'s a.c. w.r.t. P

and dominated by λ.

For all p.m. P, the mappings $Q \in M \to \phi(Q,P)$ are convex and take nonnegative values. When $Q = P$ then $\phi(Q,P) = 0$. Further, if the function $x \to \varphi(x)$ is strictly convex on neighborhood of $x = 1$, then the following basic property holds

$$\phi(Q,P) = 0 \text{ if and only if } Q = P. \tag{3.3}$$

All these properties are presented in Csiszár (1963), Csiszár (1967c) and Liese and Vajda (1987) Chapter 1, for $\phi-$divergences defined on the set of all p.m.'s M^1. When the ϕ-divergences are defined on M, then the same arguments as developed on M^1 hold.

When defined on M^1, the Kullback-Leibler (KL), modified Kullback-Leibler (KL_m), χ^2, modified χ^2 (χ^2_m), Hellinger (H), and L^1 divergences are respectively associated to the convex functions $\varphi(x) = x\log x - x + 1$, $\varphi(x) = -\log x + x - 1$, $\varphi(x) = \frac{1}{2}(x-1)^2$, $\varphi(x) = \frac{1}{2}(x-1)^2/x$, $\varphi(x) = 2(\sqrt{x}-1)^2$ and $\varphi(x) = |x-1|$.

All those divergences except the L^1 one, belong to the class of power divergences introduced in Cressie and Read (1984) (see also Liese and Vajda (1987) Chapter 2). They are defined through the class of convex functions

$$x \in \mathbb{R}_+^* \mapsto \varphi_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma-1)} \tag{3.4}$$

if $\gamma \in \mathbb{R} \setminus \{0,1\}$ and by $\varphi_0(x) := -\log x + x - 1$ and $\varphi_1(x) := x\log x - x + 1$. For all $\gamma \in \mathbb{R}$, $\varphi_\gamma(0) := \lim_{x\downarrow 0}\varphi_\gamma(x)$ and $\varphi_\gamma(+\infty) := \lim_{x\uparrow+\infty}\varphi_\gamma(x)$.

So, the $KL-$divergence is associated to φ_1, the KL_m to φ_0, the χ^2 to φ_2, the χ^2_m to φ_{-1} and the Hellinger distance to $\varphi_{1/2}$.

For all $\gamma \in \mathbb{R}$, sometimes, we denote ϕ_γ the divergence associated to the convex function φ_γ. We define the derivative of φ_γ at 0 by $\varphi'_\gamma(0) := \lim_{x\downarrow 0}\varphi'_\gamma(x)$.

We extend the definition of the power divergences functions $Q \in M^1 \to \phi_\gamma(Q,P)$ onto the whole set of signed finite measures M as follows

When the function $x \to \varphi_\gamma(x)$ is not defined on $(-\infty, 0[$ or when φ_γ is defined on $\mathbb{R}$ but is not a convex function we extend the definition of φ_γ through

$$x \in [-\infty,+\infty] \mapsto \varphi_\gamma(x)\mathbb{1}_{[0,+\infty]}(x) + (\varphi'_\gamma(0)x + \varphi_\gamma(0))\mathbb{1}_{[-\infty,0[}(x). \tag{3.5}$$

For any convex function φ, define the *domain* of φ through

$$D_\varphi = \{x \in [-\infty,+\infty] \text{ such that } \varphi(x) < +\infty\}. \tag{3.6}$$

Since φ is convex, D_φ is an interval which may be open or not, bounded or unbounded. Hence, write $D_\varphi := (a,b)$ in which a and b may be finite or infinite. In this Chapter, we will only consider φ functions defined on $[-\infty,+\infty]$ with values in $[0,+\infty]$

such that $a < 1 < b$, and which satisfy $\varphi(1) = 0$, are strictly convex and are $\mathcal{C}^2$ on the interior of its domain D_φ; we define $\varphi(a)$, $\varphi'(a)$, $\varphi''(a)$, $\varphi(b)$, $\varphi'(b)$ and $\varphi''(b)$ respectively by $\varphi(a) := \lim_{x \downarrow a} \varphi(x)$, $\varphi'(a) := \lim_{x \downarrow a} \varphi'(x)$, $\varphi''(a) := \lim_{x \downarrow a} \varphi''(x)$, $\varphi(b) := \lim_{x \uparrow b} \varphi(x)$, $\varphi'(b) := \lim_{x \uparrow b} \varphi'(x)$ and $\varphi''(b) := \lim_{x \uparrow b} \varphi''(x)$. These quantities may be finite or infinite. All the φ_γ functions (see (3.5)) satisfy these conditions.

Definition 3.1. *Let Ω be some subset in M. The $\phi-$divergence between the set Ω and a p.m. P, noted $\phi(\Omega, P)$, is*

$$\phi(\Omega, P) := \inf_{Q \in \Omega} \phi(Q, P).$$

Definition 3.2. *Assume that $\phi(\Omega, P)$ is finite. A measure $Q^* \in \Omega$ such that*

$$\phi(Q^*, P) \leq \phi(Q, P) \quad \textit{for all} \quad Q \in \Omega$$

is called a $\phi-$projection of P onto Ω. This projection may not exist, or may be not defined uniquely.

We will make use of the concept of $\phi-$divergences in order to perform estimation and tests for the model (3.1).

So, let $X_1, ..., X_n$ denote an i.i.d. sample of r.v.'s with common distribution P_0. Let P_n be the empirical measure pertaining to this sample, namely

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

in which δ_x is the Dirac measure at point x.

When P_0 and all $Q \in \mathcal{M}^1$ share the same discrete finite support S, then the ϕ-divergence $\phi(Q, P_0)$ writes

$$\phi(Q, P_0) = \sum_{j \in S} \varphi\left(\frac{Q(j)}{P_0(j)}\right) P_0(j). \tag{3.7}$$

In this case, $\phi(Q, P_0)$ can be estimated simply through the plug-in of P_n in (3.7), as follows

$$\widehat{\phi}(Q, P_0) := \sum_{j \in S} \varphi\left(\frac{Q(j)}{P_n(j)}\right) P_n(j). \tag{3.8}$$

In the same way, for any θ in Θ, $\phi\left(\mathcal{M}_\theta^1, P_0\right)$ is estimated by

$$\widehat{\phi}\left(\mathcal{M}_\theta^1, P_0\right) := \inf_{Q \in \mathcal{M}_\theta^1} \sum_{j \in S} \varphi\left(\frac{Q(j)}{P_n(j)}\right) P_n(j), \tag{3.9}$$

and $\phi\left(\mathcal{M}^1, P_0\right) = \inf_{\theta\in\Theta} \phi\left(\mathcal{M}_\theta^1, P_0\right)$ can be estimated by

$$\widehat{\phi}\left(\mathcal{M}^1, P_0\right) := \inf_{\theta\in\Theta} \inf_{Q\in\mathcal{M}_\theta^1} \sum_{j\in S} \varphi\left(\frac{Q(j)}{P_n(j)}\right) P_n(j). \tag{3.10}$$

By uniqueness of $\inf_{\theta\in\Theta} \phi\left(\mathcal{M}_\theta^1, P_0\right)$ and since this infimum is reached at $\theta = \theta_0$, we estimate θ_0 through

$$\widehat{\theta}_\phi := \arg\inf_{\theta\in\Theta} \inf_{Q\in\mathcal{M}_\theta^1} \sum_{j\in S} \varphi\left(\frac{Q(j)}{P_n(j)}\right) P_n(j). \tag{3.11}$$

The infimum in (3.9) (i.e., the projection of P_n on $\mathcal{M}_\theta^1$) may be achieved on the frontier of $\mathcal{M}_\theta^1$. In this case the Lagrange method is not valid. We endow our statistical approach in the global context of signed finite measures with total mass 1 satisfying the linear constraints. So, define

$$\mathcal{M}_\theta := \left\{ Q \in M \ \text{ such that } \ \int dQ = 1 \text{ and } \int g(x,\theta)\, dQ(x) = 0 \right\} \tag{3.12}$$

and

$$\mathcal{M} := \bigcup_{\theta\in\Theta} \mathcal{M}_\theta, \tag{3.13}$$

sets of signed finite measures that replace $\mathcal{M}_\theta^1$ and $\mathcal{M}^1$.

As above, we estimate $\phi(\mathcal{M}_\theta, P_0)$, $\phi(\mathcal{M}, P_0)$ and θ_0 respectively by

$$\widehat{\phi}\left(\mathcal{M}_\theta, P_0\right) := \inf_{Q\in\mathcal{M}_\theta} \sum_{j\in S} \varphi\left(\frac{Q(j)}{P_n(j)}\right) P_n(j), \tag{3.14}$$

$$\widehat{\phi}\left(\mathcal{M}, P_0\right) := \inf_{\theta\in\Theta} \inf_{Q\in\mathcal{M}_\theta} \sum_{j\in S} \varphi\left(\frac{Q(j)}{P_n(j)}\right) P_n(j), \tag{3.15}$$

and

$$\widehat{\theta}_\phi := \arg\inf_{\theta\in\Theta} \inf_{Q\in\mathcal{M}_\theta} \sum_{j\in S} \varphi\left(\frac{Q(j)}{P_n(j)}\right) P_n(j). \tag{3.16}$$

Enhancing $\mathcal{M}^1$ to $\mathcal{M}$ is motivated by the following arguments

- For all θ in Θ, denote Q_1^* and Q^* respectively the projection of P_n on $\mathcal{M}_\theta^1$ and on $\mathcal{M}_\theta$, as defined in (3.9) and in (3.14). If Q_1^* is an interior point of $\mathcal{M}_\theta^1$, then, by Proposition 3.37 below, it coincides with Q^*, the projection of P_n on $\mathcal{M}_\theta$, i.e., $Q_1^* = Q^*$. Therefore, in this case, both approaches coincide.

- It may occur that for some θ in Θ, Q_1^*, the projection of P_n on $\mathcal{M}_\theta^1$, is a frontier point of $\mathcal{M}_\theta^1$, which makes a real difficulty for the estimation procedure. We will prove in Theorem 3.5 that $\widehat{\theta}_\phi$, defined in (3.16) and which replaces (3.11), converges to θ_0. This validates the substitution of the sets $\mathcal{M}_\theta^1$ by the sets $\mathcal{M}_\theta$. In the context of test problem, we will prove that the asymptotic distributions of the test statistics pertaining to Problem 1 and 2 are unaffected by this change.

This modification motivates the above extensions in the definitions of the φ functions on $[-\infty, +\infty]$ and of the ϕ-divergences on the whole space of finite signed measures M.

In the case when Q and P_0 share different discrete finite support or share same or different discrete infinite or continuous support, then formula (3.8) is not defined, due to lack of absolute continuity of Q with respect to P_n. Indeed

$$\widehat{\phi}(Q, P_0) := \phi(Q, P_n) = +\infty. \tag{3.17}$$

The plug-in estimate of $\phi(\mathcal{M}_\theta, P_0)$ writes

$$\widehat{\phi}(\mathcal{M}_\theta, P_0) := \inf_{Q \in \mathcal{M}_\theta} \phi(Q, P_n) = \inf_{Q \in \mathcal{M}_\theta} \int \varphi\left(\frac{dQ}{dP_n}(x)\right)\, dP_n(x). \tag{3.18}$$

If the infimum exists, then it is clear that it is reached at a signed finite measure (or probability measure) which is a.c. w.r.t. P_n. So, define the sets

$$\mathcal{M}_\theta^{(n)} := \left\{ Q \in M \ \text{ such that } \ Q \ll P_n,\ \sum_{i=1}^n Q(X_i) = 1 \ \text{ and } \ \sum_{i=1}^n Q(X_i) g(X_i, \theta) = 0 \right\}, \tag{3.19}$$

which may be seen as subsets of $\mathbb{R}^n$. Then, the plug-in estimate (3.18) of $\phi(\mathcal{M}_\theta, P_0)$ writes

$$\widehat{\phi}(\mathcal{M}_\theta, P_0) = \inf_{Q \in \mathcal{M}_\theta^{(n)}} \frac{1}{n} \sum_{i=1}^n \varphi\left(nQ(X_i)\right) \tag{3.20}$$

In the same way, $\phi(\mathcal{M}, P_0) := \inf_{\theta \in \Theta} \inf_{Q \in \mathcal{M}_\theta} \phi(Q, P_0)$ can be estimated by

$$\widehat{\phi}(\mathcal{M}, P_0) = \inf_{\theta \in \Theta} \inf_{Q \in \mathcal{M}_\theta^{(n)}} \frac{1}{n} \sum_{i=1}^n \varphi\left(nQ(X_i)\right). \tag{3.21}$$

By uniqueness of $\inf_{\theta \in \Theta} \phi(\mathcal{M}_\theta, P_0)$ and since this infimum is reached at $\theta = \theta_0$, we estimate θ_0 through

$$\widehat{\theta}_\phi = \arg \inf_{\theta \in \Theta} \inf_{Q \in \mathcal{M}_\theta^{(n)}} \frac{1}{n} \sum_{i=1}^n \varphi\left(nQ(X_i)\right). \tag{3.22}$$

Note that, when P_0 and all $Q \in \mathcal{M}^1$ share the same discrete finite support, then the estimates (3.22), (3.21) and (3.20) coincide respectively with (3.16), (3.15) and (3.14). Hence, in the sequel, we study the estimates $\widehat{\phi}(\mathcal{M}_\theta, P_0)$, $\widehat{\phi}(\mathcal{M}, P_0)$ and $\widehat{\theta}_\phi$ as defined in (3.20), (3.21) and (3.22), respectively.

We propose to call the estimates $\widehat{\theta}_\phi$ defined in (3.22) "Minimum Empirical ϕ-Divergences Estimates" (MEϕDE's).

As will be noticed later on, the "empirical likelihood" paradigm (see Owen (1988) and Owen (1990)), which is based on this plug-in approach, enters as a special case of the statistical issues related to estimation and tests based on $\phi-$divergences with $\varphi(x) = \varphi_0(x) = -\log x + x - 1$, namely on KL_m-divergence. The empirical log-likelihood ratio for the model (3.12), in the context of ϕ-divergences, writes $-n\widehat{KL_m}(\mathcal{M}_\theta, P_0)$. In the case of a single functional, for example when $g(x, \theta) = x-\theta$ with x and θ belong to $\mathbb{R}$, Owen (1988) shows that $2n\widehat{KL_m}(\mathcal{M}_\theta, P_0)$ has an asymptotic $\chi^2_{(1)}$ distribution when P_0 belongs to $\mathcal{M}_\theta$. (see Owen (1988) Theorem 1). This result is a nonparametric version of Wilks's theorem (see Wilks (1938)). In the multivariate case, the same result holds (see Owen (1990) Theorem 1). When we want to extend arguments used in Owen (1988) and Owen (1990) in order to study the limiting behavior of the statistics $\widehat{\phi}(\mathcal{M}_\theta, P_0)$, when $P_0 \in \mathcal{M}_\theta$ (i.e., when $\theta_0 = \theta$) and when $P_0 \notin \mathcal{M}_\theta$ (for example, when $\theta_0 \neq \theta$), most limiting arguments become untractable. We propose to use the so-called "dual representation of $\phi-$divergences"(see Keziou (2003)), a device which is well known for the Kullback-Leibler divergence in the context of large deviations, and which has been used in parametric statistics in Keziou (2003) and in Broniatowski and Keziou (2003). The estimates then turn to be M-estimates whose limiting distributions are obtained through classical methods. On the other hand, the obtention of the limit distributions of the statistics $\widehat{\phi}(\mathcal{M}_\theta, P_0)$ when $P_0 \notin \mathcal{M}_\theta$, requires the study of the existence and the characterization of the projection of the p.m. P_0 on the sets $\mathcal{M}_\theta$.

This Chapter is organized as follows : In Section 2, we focus on existence conditions for the projection of some p.m. P on some subsets Ω of M and we extend some known results which characterize the projection when Ω is defined through linear constraints, that is when $\Omega = \mathcal{M}_\theta$. Section 3 presents the dual representation of the divergences. In Section 4, we give sufficient conditions for existence of ϕ-projections of P_n on the sets $\mathcal{M}_\theta^{(n)}$ (i.e., the existence of the infimum in (3.20)). We also provide sufficient conditions for the existence of ϕ-projections of P_0 on the sets $\mathcal{M}_\theta$ using the dual representation of ϕ-divergences. In Section 5, we study the asymptotic behavior of the proposed estimates (3.20), (3.21) and (3.22) giving solutions to *Problem 2*.

We then address *Problem 1*, namely : does there exist some θ_0 in Θ for which P_0

belongs to $\mathcal{M}_{\theta_0}$?

In Section 6, we propose new estimates for the distribution function using the ϕ-projections of P_n on the model $\mathcal{M}$. We show that the new estimates of the distribution function are generally more efficient than the empirical cumulative distribution function. Section 7 illustrates the concept of empirical likelihood in the context of ϕ-divergences techniques. In Section 8, we focus on robustness and efficiency of the MEϕD estimates. A simulation study aims at emphasizing the specific advantage of the choice of the Hellinger divergence in relation with robustness and efficiency considerations. All proofs are in Section 9.

3.2 ϕ-Divergences and Projection

We now give some notation. Let $\mathcal{F}$ be some class of $\mathcal{B}$-measurable real valued functions f defined on $\mathcal{X}$ and denote $\mathcal{B}_b$ the set of all bounded $\mathcal{B}$-measurable functions defined on $\mathcal{X}$. Define

$$M_{\mathcal{F}}^1 := \left\{ Q \in M^1 \text{ such that } \int |f| \, dQ < \infty, \text{ for all } f \text{ in } \mathcal{F} \right\}$$

and

$$M_{\mathcal{F}} := \left\{ Q \in M \text{ such that } \int |f| \, d|Q| < \infty, \text{ for all } f \text{ in } \mathcal{F} \right\},$$

in which $|Q|$ denotes the total variation of the signed finite measure Q.

Note that when $\mathcal{F} = \mathcal{B}_b$, then $M_{\mathcal{F}}^1 = M^1$ and $M_{\mathcal{F}} = M$.

Denote $\mathcal{D}_\phi$ the *domain of the* $\phi-$divergence, i.e.,

$$\mathcal{D}_\phi := \{ Q \in M \text{ such that } \phi(Q, P) < \infty \}.$$

The set $\mathcal{D}_\phi$ clearly depends on P.

Definition 3.3. *Denote $\tau_{\mathcal{F}}$ the weakest topology on $M_{\mathcal{F}}$ for which all mappings $Q \in M_{\mathcal{F}} \to \int f \, dQ$ are continuous when f belongs to $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$, the linear span of $\mathcal{F} \cup \mathcal{B}_b$.*

A base of open neighborhoods for any $R \in M_{\mathcal{F}}$ is defined by

$$U(R, \mathcal{A}, \epsilon) := \left\{ Q \in M_{\mathcal{F}} \text{ such that } \left| \int f \, dR - \int f \, dQ \right| < \epsilon \text{ for all } f \in \mathcal{A} \right\},$$

where $\mathcal{A}$ is a finite subset of $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$ and ϵ is a positive number.

We refer to Dunford and Schwartz (1962) Chapter 5, for the various topologies induced by classes of functions.

Note that when $\mathcal{F} = \mathcal{B}_b$, then the base of open neighborhoods $U(R, A, \epsilon)$ generates the so-called $\tau-$topology ; see Groeneboom *et al.* (1979) and Gänssler (1971).

For all p.m. P and all ϕ-divergence, the divergence function $Q \to \phi(Q, P)$ defined on the linear space $M_{\mathcal{F}}$ endowed with the $\tau_{\mathcal{F}}$-topology enjoys the following property (see Keziou (2003) Proposition 2.1 and Broniatowski and Keziou (2003) Proposition 2.3).

Proposition 3.1. *The divergence function $Q \to \phi(Q, P)$ defined from $(M_{\mathcal{F}},\ \tau_{\mathcal{F}})$ onto $[0, +\infty]$ is lower semi-continuous (l.s.c.).*

3.2.1 Existence of ϕ-Projection on general Sets Ω

If φ is a strictly convex function, then $Q \to \phi(Q, P)$ is strictly convex and the projection of P on some convex set Ω, say Q^*, is uniquely defined whenever it exists.

By Proposition 3.1, the following result holds

Proposition 3.2. *Let P be some p.m. and Ω some compact subset of $(M_{\mathcal{F}},\ \tau_{\mathcal{F}})$. Then the ϕ-projection of P on Ω exists.*

The above $\tau_{\mathcal{F}}-$topology on $M_{\mathcal{F}}$ is indeed the natural and the most convenient one in order to handle projection properties. It has been introduced in the context of large deviation probabilities by Eichelsbacher and Schmock (1997) for the Kullback-Leibler divergence and it is used in statistics in Broniatowski (2003), Keziou (2003) and Broniatowski and Keziou (2003).

Usually the sets which are to be considered in statistical applications are not compact but merely closed sets ; a typical example is when they are defined by linear constraints

$$\mathcal{M}_\theta := \left\{ Q \in M_{\mathcal{F}} \text{ such that } \int dQ = 1 \text{ and } \int g(x, \theta)\ dQ(x) = 0 \right\} \tag{3.23}$$

which is (3.12) when

$$\mathcal{F} := \{ g_1(., \theta), ..., g_l(., \theta) \} . \tag{3.24}$$

Hence, the set $\mathcal{M}_\theta$ is closed in $M_{\mathcal{F}}$ endowed with $\tau_{\mathcal{F}}$-topology ; this motivates the choice of $\tau_{\mathcal{F}}$-topology.

Using some similar arguments as used in the proof of Proposition 8.5 in Liese and Vajda (1987), we state a general result for the existence of the ϕ-projection of some p.m. P on closed sets Ω.

Theorem 3.1. *Let Ω be some closed set in $M_{\mathcal{F}}$ equipped with the $\tau_{\mathcal{F}}$-topology. Assume that there exists numbers $r > 1$ and $k > 1$ such that $\frac{1}{r} + \frac{1}{k} = 1$,*

$$\lim_{|x|\to\infty} \frac{\varphi(x)}{|x|^r} = +\infty \quad \textit{and} \quad \int |f|^k \; dP < \infty \textit{ for all } \; f \textit{ in } \mathcal{F}.$$

If $\phi(\Omega, P) := \inf_{Q\in\Omega} \phi(Q, P)$ is finite[1], then the projection of P on Ω exists.

Remark 3.1. *Theorem 3.1 remains valid if we substitute $M_{\mathcal{F}}$ by $M^1_{\mathcal{F}}$ or if we substitute $\mathcal{F}$ by $\mathcal{B}_b$.*

Remark 3.2. *The condition in Theorem 3.1 does not hold for the Kullback-Leibler divergence since $\lim_{|x|\to+\infty} \frac{x\log x - x + 1}{|x|^r} = 0$ for all $r > 1$. However, when Ω is closed in the $\tau-$topology, similar arguments as developed in the proof lead to Theorem 3.1 under the weaker assumption $\lim_{|x|\to+\infty} \frac{\varphi(x)}{|x|} = +\infty$.*

Remark 3.3. *By Eichelsbacher and Schmock (1997), whenever for all $\alpha > 0$ and all $f \in \mathcal{F}$, it holds $\int \exp(\alpha|f|) \; dP < \infty$ then the level sets*

$$\left\{Q \in M^1_{\mathcal{F}} \;\; \textit{such that} \;\; KL(Q, P) \le c\right\}$$

are compact in $(M^1_{\mathcal{F}}, \tau_{\mathcal{F}})$ for all real c. Therefore, for any $\tau-$closed set Ω for which $KL(\Omega, P) < \infty$, the projection of P on Ω exists (see Proposition 3.2).

Remark 3.4. *In the case of the Kullback-Leibler divergence and for sets Ω defined through linear constraints, sufficient conditions for the existence of the projection are presented in (Csiszár (1975) Theorem 3.1, Corollary 3.1 and Theorem 3.3).*

Remark 3.5. *Theorem 3.1 remains valid when substituting $M_{\mathcal{F}}$ by $M_{\mathcal{F}} \cap \mathcal{D}_\phi$ endowing $M_{\mathcal{F}} \cap \mathcal{D}_\phi$ with the relative topology.*

3.2.2 Existence and Characterization of ϕ-Projection on general Sets Ω

In this Section, we extend known results pertaining to the projected measure as can be found in Liese and Vajda (1987), Csiszár (1975), Csiszár (1984), Rüschendorf (1984) and Rüschendorf (1987). These authors have characterized the projected measure on subsets of M^1. We expose similar results when considering subsets of $M_{\mathcal{F}}$ and take the occasion to clarify some proofs. Denote

[1] Note that this is equivalent to the following assertion

$$\text{there exists} \;\; Q \in \Omega \;\; \text{such that} \;\; \phi(Q, P) \;\; \text{is finite.}$$

(C.0) There exists a positive δ such that for all c in $[1-\delta, 1+\delta]$, we can find positive numbers c_1, c_2, c_3 such that $\varphi(cx) \leq c_1\varphi(x) + c_2\,|x| + c_3$, for all real x.

Remark 3.6. *Condition (C.0) holds for all power divergences including KL and KL_m divergences.*

We first give two Lemmas, which we will use in the proof of Theorem 3.2 and Theorem 3.3 below.

Lemma 3.1. *Assume that (C.0) holds. Then, for all Q in M and all P in M^1 such that $\phi(Q,P)$ is finite, it holds*

1- for any c in $[1-\delta, 1+\delta]$, $\varphi\left(c\frac{dQ}{dP}\right)$ belongs to $L^1(P)$.

2- $\lim_{c\uparrow 1}\phi(cQ,P) = \phi(Q,P) = \lim_{c\downarrow 1}\phi(cQ,P)$.

Lemma 3.2. *Assume that (C.0) holds. Then, for all Q in $\mathcal{D}_\phi$, $\varphi'(q)q$ belongs to $L^1(P)$, where $q := \frac{dQ}{dP}$.*

Theorem 3.2. *Let Ω be a subset of M. Assume that (C.0) holds. Then*

1- If there exists some Q^* in $\Omega \cap \mathcal{D}_\phi$ such that for all Q in $\Omega \cap \mathcal{D}_\phi$ it holds (i) $\varphi'(q^*)q \in L^1(P)$ and (ii) $\int \varphi'(q^*)\,dQ^* \leq \int \varphi'(q^*)\,dQ$, then Q^* is the projection of P on Ω.

2- Let Ω be convex and P have projection Q^* on Ω. Then, for all Q in $\Omega \cap \mathcal{D}_\phi$, it holds (i) $\varphi'(q^*)q \in L^1(P)$ and (ii) $\int \varphi'(q^*)\,dQ^* \leq \int \varphi'(q^*)\,dQ$.

Remark 3.7. *Theorem 3.2 remains valid for subsets Ω of M^1, the space of all p.m.'s.*

3.2.3 Existence and Characterization of ϕ-Projection on Sets defined by Linear Constraints

In this Subsection, we consider the projection of a p.m. P on a linear set of measures in M defined by an arbitrary family of constraints. So, let $\mathcal{G}$ denote a collection (finite or infinite, countable or not) of real valued functions defined on $(\mathcal{X}, \mathcal{B})$. The class $\mathcal{G}$ is assumed to contain the function $\mathbb{1}_{\mathcal{X}}$. The set Ω is defined through

$$\Omega := \left\{ Q \ \in M \text{ such that } \int dQ = 1 \text{ and } \int g\,dQ = 0 \text{ for all } g \text{ in } \mathcal{G}\setminus\{\mathbb{1}_{\mathcal{X}}\} \right\}. \tag{3.25}$$

The following result states the explicit form of Q^*, the projection of P on Ω, when it exists.

Theorem 3.3. *Assume that (C.0) holds. Then,*

1- P has projection Q^* on Ω if only if Q^* belongs to $\Omega \cap \mathcal{D}_\phi$ and for all Q in $\Omega \cap \mathcal{D}_\phi$, it holds

$$\varphi'(q^*)q \in L^1(P) \quad and \quad \int \varphi'(q^*)\, dQ^* \leq \int \varphi'(q^*)\, dQ.$$

2- If there exists some measure Q^* in $\Omega \cap \mathcal{D}_\phi$ such that $\varphi'(q^*)$ belongs to $\langle \mathcal{G} \rangle$, then Q^* is the projection of P on Ω.

3- If P has projection Q^* on Ω, then $\varphi'(q^*)$ belongs to $\overline{\langle \mathcal{G} \rangle}$, the closure of $\langle \mathcal{G} \rangle$ in $L^1(|Q^*|)$.

Remark 3.8. *Theorem 3.3 remains valid if we substitute the set Ω by the set $\Omega \cap M^1$.*

Remark 3.9. *It should be noticed that the preceding Theorem does not provide a definite description of the projected measure ; indeed, it does not give any information on the support of $|Q^*|$ (see example 3.6 below).*

Remark 3.10. *Versions of this Theorem, for sets of p.m.'s, have been proved for the Kullback-Leibler divergence by Csiszár (1975), and by Rüschendorf (1984) for ϕ-divergences between p.m.'s. We prove it in the present context, that is when the set Ω (see (3.25)) is a subset of the signed finite measures and P is a p.m..*

Example 3.6. *Let $\mathcal{X} := [0,1]$, P the uniform distribution on $[0,1]$, $\mathcal{G} := \left\{ \mathbb{1}_{[0,1]}, I_d \right\}$ where I_d is the identity function. Consider the χ^2-divergence and the set $\mathcal{M}_{1/4}$ defined by*

$$\mathcal{M}_{1/4} := \left\{ Q \in M^1 \text{ such that } \int dQ = 1 \text{ and } \int (x - 1/4)\, dQ(x) = 0 \right\}.$$

Apply the preceding results pertaining to the characterization of the projection of P on $\mathcal{M}_{1/4}$. By Theorem 3.1, P has some projection Q^* on $\mathcal{M}_{1/4}$. By Theorem 3.3, there exists two real numbers c_0 and c_1 such that

$$\frac{dQ^*}{dP}(x) \mathbb{1}_{\{q^*(x)>0\}} = c_0 + c_1 x.$$

The support of Q^* is different from the support of P ; it is strictly included in $[0,1]$. Indeed, if the support of Q^* is $[0,1]$, then by Theorem 3.3 part (3), there exits constants c_0 and c_1 such that

$$dQ^*(x) = (c_0 + c_1 x)\ dP(x) = (c_0 + c_1 x)\, \mathbb{1}_{[0,1]}(x)\, dx. \tag{3.26}$$

Using the fact that Q^* belongs to $\mathcal{M}_{1/4}$, we obtain that $c_0 = 5/2$ and $c_1 = -3$. So, Q^* satisfying $dQ^*(x) = (5/2 - 3x)\ dP(x)$ does not belong to $\mathcal{M}_{1/4}$ (it is not a p.m.), a contradiction with the existence of the projection. This proves that the support of Q^* is strictly included in $[0,1]$. Hence, the support of a projection Q^* of P may be different from the support of P.

3.3 Dual Representation and Estimation of ϕ - Divergences

Here, we introduce the *dual representation* of the ϕ-divergences on the space $(M_{\mathcal{F}}, \tau_{\mathcal{F}})$. The role of this representation is twofold : First, it provides practical tools for the existence and the explicit form of the projection of P_0 on the sets $\mathcal{M}_\theta$; see Proposition 3.5 below. Second, it produces convenient forms of the above estimates (3.20), (3.21) and (3.22). Also their asymptotic properties are easily studied through this approach.

In few words, the dual representation expresses the fact that in a suitable context the convex function $Q \in M_{\mathcal{F}} \to \phi(Q, P)$ is the upper envelope of its support hyperplanes. The first result, which is Proposition 2.1 in Keziou (2003) and in Broniatowski and Keziou (2003), provides the description of the hyperplanes in $M_{\mathcal{F}}$.

Proposition 3.3. *Equip $M_{\mathcal{F}}$ with the $\tau_{\mathcal{F}}$-topology. Then, $M_{\mathcal{F}}$ is a locally convex Hausdorff topological linear space. Further, the topological dual space of $M_{\mathcal{F}}$ is the set of all mappings $Q \to \int f \, dQ$ when f belongs to $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$.*

Having in mind the results in Propositions 3.1 and 3.3, we state the duality Lemma.

Lemma 3.3. *[Duality Lemma] Let $\mathcal{X}$ be a locally convex Hausdorff topological linear space, and $g : \mathcal{X} \to (-\infty, +\infty]$ some convex l.s.c. function. Define*

$$g^*(l) := \sup_{x \in \mathcal{X}} \{l(x) - g(x)\}, \ l \in \mathcal{X}^*$$

where $\mathcal{X}^*$ denotes the topological dual space of $\mathcal{X}$. Then,

$$g(x) = \sup_{l \in \mathcal{X}^*} \{l(x) - g^*(l)\},$$

which is to say that g is the Fenchel-Legendre transform of g^*.

The proof of this Lemma can be found for example in Dembo and Zeitouni (1998) Chapter 4 or Azé (1997) Chapter 4.

For all p.m. P, we apply the above duality Lemma in the context of the space $(M_{\mathcal{F}}, \tau_{\mathcal{F}})$ and the divergences functions $Q \to \phi(Q, P)$ defined from $(M_{\mathcal{F}}, \tau_{\mathcal{F}})$ onto $[0, +\infty]$. According to Proposition 3.3, the Fenchel-Legendre transform of $\phi(., P)$ is defined on $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$ by

$$T(f, P) := \sup_{Q \in M_{\mathcal{F}}} \left\{ \int f \, dQ - \phi(Q, P) \right\}.$$

By Lemma 3.3, we thus have for any Q in $M_{\mathcal{F}}$,

$$\phi(Q,P) := \sup_{f\in\langle\mathcal{F}\cup\mathcal{B}_b\rangle}\left\{\int f\ dQ - T(f,P)\right\}. \tag{3.27}$$

The convex conjugate (or Fenchel-Legendre transform) of φ will be denoted ψ, i.e.,

$$t\in\mathbb{R} \ \mapsto\ \psi(t) := \sup_{x\in\mathbb{R}}\{tx-\varphi(x)\}. \tag{3.28}$$

Denote ${\varphi'}^{-1}$ the inverse function of φ', Im φ' the set of all values of φ' and φ^* the function defined on Im φ' by

$$t\in \text{ Im }\varphi' \ \mapsto\ \varphi^*(t) := t{\varphi'}^{-1}(t) - \varphi\left({\varphi'}^{-1}(t)\right). \tag{3.29}$$

Clearly, the function φ^* coincides on the set Im φ' with ψ, the convex conjugate of φ. For concepts about convex functions and properties of convex conjugates see e.g. Rockafellar (1970).

In Keziou (2003) and in Broniatowski and Keziou (2003), under some conditions, we prove that

$$T(f,P) = \int f{\varphi'}^{-1}(f) - \varphi\left({\varphi'}^{-1}(f)\right)\ dP =: \int \varphi^*(f)\ dP, \tag{3.30}$$

for all $f\in\langle\mathcal{F}\cup\mathcal{B}_b\rangle$.

Under mild conditions on $\mathcal{F}$, formula (3.27) still holds when $\langle\mathcal{F}\cup\mathcal{B}_b\rangle$ is replaced by any class of functions $\mathcal{F}$ that contains $\varphi'(dQ/dP)$ (see Theorem 1.1). Therefore, in such case

$$\phi(Q,P) := \sup_{f\in\mathcal{F}}\left\{\int f\ dQ - \int \varphi^*(f)\ dP\right\}. \tag{3.31}$$

which we write

$$\phi(Q,P) := \sup_{f\in\mathcal{F}}\int m_f(x)\ dP(x), \tag{3.32}$$

with

$$m_f(x) := \int f\ dQ - \varphi^*\left(f(x)\right). \tag{3.33}$$

Remark 3.11. *For any Q in $M_{\mathcal{F}}$ such that $\varphi'(dQ/dP)\in\mathcal{F}$, the maximum in* (3.31) *is unique (P-a.s.) and reached at $f=\varphi'(dQ/dP)$; see Keziou (2003) Theorem 2.1 and Broniatowski and Keziou (2003) Theorem 2.5, (see also Theorem 1.1).*

3.4 Estimation for Models satisfying Linear Constraints

At this point, we must introduce some notational convention for sake of brevity and clearness. For any p.m. P on $\mathcal{X}$ and any measurable real function f on $\mathcal{X}$, Pf denotes $\int f(x)\, dP(x)$. For example, $P_0 g_j(\theta)$ will be used instead of $\int g_j(\theta, x)\, dP_0(x)$. Hence, we are led to define the following functions : denote $\overline{g}$ the function defined on $\mathcal{X} \times \Theta$ with values in $\mathbb{R}^{l+1}$ by

$$\begin{array}{rccl} \overline{g} & : & \mathcal{X} \times \Theta & \to & \mathbb{R}^{(l+1)} \\ & & (x, \theta) & \mapsto & \overline{g}(x, \theta) := (\mathbb{1}_{\mathcal{X}}(x), g_1(x, \theta), \ldots, g_l(x, \theta))^T, \end{array}$$

and for all $\theta \in \Theta$, denote also $\overline{g}(\theta)$, $g(\theta)$, $g_j(\theta)$ the functions defined respectively by

$$\begin{array}{rccl} \overline{g}(\theta) & : & \mathcal{X} & \to & \mathbb{R}^{l+1} \\ & & x & \mapsto & \overline{g}(x, \theta) := (g_0(x, \theta), g_1(x, \theta), \ldots, g_l(x, \theta))^T, \text{ where } g_0(x, \theta) := \mathbb{1}_{\mathcal{X}}(x), \\ g(\theta) & : & \mathcal{X} & \to & \mathbb{R}^l \\ & & x & \mapsto & g(x, \theta) := (g_1(x, \theta), \ldots, g_l(x, \theta))^T \\ \text{and} & & & & \\ g_j(\theta) & : & \mathcal{X} & \to & \mathbb{R} \\ & & x & \mapsto & g_j(x, \theta), \quad \text{for all } j \in \{0, 1, \ldots, l\}. \end{array}$$

We now turn back to the setting defined in the Introduction and consider model (3.12). For fixed θ in Θ, define the class of functions

$$\mathcal{F}_\theta := \{g_0(\theta), g_1(\theta), \ldots, g_l(\theta)\},$$

and consider the set of finite signed measures $\mathcal{M}_\theta$ defined by $(l+1)$ linear constraints as defined in (3.12)

$$\mathcal{M}_\theta := \left\{ Q \in M_{\mathcal{F}_\theta} \text{ such that } \int dQ(x) = 1 \text{ and } \int g(x, \theta)\, dQ(x) = 0 \right\}.$$

We will make use of the dual representation of ϕ-divergences in order to state the explicit form of the estimates (3.20), (3.21) and (3.22). Furthermore, we present explicit tractable conditions for these estimates to be well defined. This will be done in Propositions 3.4 and 3.5 below.

First, we present sufficient conditions which assess the existence of the infimum in (3.20), noted $\widehat{Q}_\theta^*$, the projection of P_n on $\mathcal{M}_\theta$. We also provide conditions under which the Lagrange method can be used to characterize $\widehat{Q}_\theta^*$.

So, define

$$\mathcal{D}_{\phi}^{(n)} := \left\{ Q \in M \text{ such that } Q \ll P_n \text{ and } \frac{1}{n}\sum_{i=1}^{n} \varphi\left(nQ(X_i)\right) < \infty \right\}, \tag{3.34}$$

i.e., the *domain* of the function

$$(Q(X_1), \ldots, Q(X_n))^T \in \mathbb{R}^n \mapsto \frac{1}{n}\sum_{i=1}^{n} \varphi\left(nQ(X_i)\right).$$

We have

Proposition 3.4. *Assume that there exists some measure R in the interior of $\mathcal{D}_{\phi}^{(n)}$ and in $\mathcal{M}_{\theta}^{(n)}$ such that for all Q in $\partial\mathcal{D}_{\phi}^{(n)}$, the frontier of $\mathcal{D}_{\phi}^{(n)}$, we have*

$$\frac{1}{n}\sum_{i=1}^{n} \varphi\left(nR(X_i)\right) < \frac{1}{n}\sum_{i=1}^{n} \varphi\left(nQ(X_i)\right). \tag{3.35}$$

Then the following holds

(i) there exists an unique $\widehat{Q_{\theta}^{*}}$ in $\mathcal{M}_{\theta}^{(n)}$ such that

$$\inf_{Q \in \mathcal{M}_{\theta}^{(n)}} \frac{1}{n}\sum_{i=1}^{n} \varphi\left(nQ(X_i)\right) = \frac{1}{n}\sum_{i=1}^{n} \varphi\left(n\widehat{Q_{\theta}^{*}}(X_i)\right) \tag{3.36}$$

(ii) $\widehat{Q_{\theta}^{*}}$ is an interior point of $\mathcal{D}_{\phi}^{(n)}$ and satisfies for all $i = 1, \ldots, n$

$$\widehat{Q_{\theta}^{*}}(X_i) = \frac{1}{n}\overleftarrow{\varphi'}\left(\sum_{j=0}^{l} \widehat{c}_j g_j(X_i, \theta)\right), \tag{3.37}$$

where $(\widehat{c}_0, \widehat{c}_1, \ldots, \widehat{c}_l)^T := \widehat{c}_{\theta}$ is solution of the system of equations

$$\begin{cases} \int \overleftarrow{\varphi'}\left(\widehat{c}_0 + \sum_{i=1}^{l} \widehat{c}_i g_i(x,\theta)\right) dP_n(x) &= 1 \\ \int g_j(x,\theta)\overleftarrow{\varphi'}\left(\widehat{c}_0 + \sum_{i=1}^{l} \widehat{c}_i g_i(x,\theta)\right) dP_n(x) &= 0, \quad j = 1, \ldots, l. \end{cases} \tag{3.38}$$

Example 3.7. *For the χ^2-divergence, we have $\mathcal{D}_{\chi^2}^{(n)} = \mathbb{R}^n$. Hence condition (3.35) holds whenever $\mathcal{M}_{\theta}^{(n)}$ is not void. Therefore, the above Proposition holds always independently upon the distribution P_0. More generally, the above Proposition holds for any ϕ-divergence which is associated to φ function satisfying $D_{\varphi} = \mathbb{R}$. (See (3.6) for the definition of D_{φ}).*

Example 3.8. *In the case of the modified Kullback-Leibler divergence, which turns to coincide with the empirical likelihood technique (see Section 7), we have* $\mathcal{D}_{KL_m}^{(n)} = (]0,+\infty[)^n$. *For* α *in* Θ, *define the assertion*

$$\begin{array}{rcl} \text{there exists } q & = & (q_1,...,q_n) \text{ in } \mathbb{R}^n \text{ with } 0 < q_i < 1 \text{ for all } i = 1,...,n \\ \text{and } \sum_{i=1}^n q_i g_j(X_i,\alpha) & = & 0 \text{ for all } j = 1,...,l. \end{array} \tag{3.39}$$

A sufficient condition, in order to assess that condition (3.35) in the above Proposition holds, is when (3.39) *holds for* $\alpha = \theta$. *In the case when* $g(x,\theta) = x - \theta$, *this is precisely what is checked in (Owen (1990)), p. 100, when* θ *is an interior point of the convex hull of* $(X_1,...,X_n)$.

Example 3.9. *For the modified* χ^2-*divergence, we have* $\mathcal{D}_{\chi_m^2}^{(n)} = (]0,\infty[)^n$, *and therefore, condition (3.39) for* $\alpha = \theta$ *is sufficient for the condition (3.35) to holds. So, conditions which assess the existence of the projection* $\widehat{Q_\theta^*}$ *are the same for the modified* χ^2-*divergence and the* KL_m*-divergence.*

Remark 3.12. *If there exists* $Q_0 \in \mathcal{M}_\theta^{(n)}$ *such that*

$$a < \inf_i nQ_0(X_i) \leq \sup_i nQ_0(X_i) < b. \tag{3.40}$$

Then, applying Corollary 2.6 in Borwein and Lewis (1991), we get

$$\inf_{Q\in\mathcal{M}_\theta^{(n)}} \frac{1}{n}\sum_{i=1}^n \varphi\left(nQ(X_i)\right) = \sup_{t\in\mathbb{R}^{(l+1)}} \left\{ t_0 - \int \psi\left(t^T\overline{g}(x,\theta)\right)\ dP_n(x) \right\}$$

with dual attainement. Furthermore, if

$$\varphi'(a) < \inf_i \widehat{c}_\theta^T \overline{g}(X_i,\theta) \leq \sup_i \widehat{c}_\theta^T \overline{g}(X_i,\theta) < \varphi'(b),$$

with $\widehat{c}_\theta$ *a dual optimal, then the unique projection* $\widehat{Q_\theta^*}$ *of* P_n *on* $\mathcal{M}_\theta^{(n)}$ *is given by (3.37).*

We will make use of formula (3.31). So, define

$$\mathcal{C}_\theta := \left\{ t \in \mathbb{R}^{l+1} \text{ such that } t^T\overline{g}(.,\theta) \text{ belongs to } \text{Im } \varphi' \ (P_0 - a.s.) \right\}, \tag{3.41}$$

and

$$\mathcal{C}_\theta^{(n)} := \left\{ t \in \mathbb{R}^{l+1} \text{ such } t^T\overline{g}(X_i,\theta) \text{ belongs to } \text{Im } \varphi' \text{ for all } i = 1,\ldots,n \right\}. \tag{3.42}$$

We may omit the subscript θ when unnecessary. Note that both $\mathcal{C}_\theta$ and $\mathcal{C}_\theta^{(n)}$ depend upon the function φ but, for simplicity, we omit the subscript φ.

If P_0 admits a projection Q_θ^* on $\mathcal{M}_\theta$ with the same support as P_0, then using the third part in Theorem 3.3, there exist constants $c_0, \ldots, c_l$, obviously depending on θ, such that

$$\varphi'\left(\frac{dQ_\theta^*}{dP_0}(x)\right) = c_0 + \sum_{j=1}^{l} c_j g_j(x,\theta), \quad \text{for all} \quad x \ (P_0 - a.s.).$$

Since Q_θ^* belongs to $\mathcal{M}_\theta$, the real numbers $c_0, c_1, \ldots, c_l$ are solutions of

$$\begin{cases} \int \varphi'^{-1}\left(c_0 + \sum_{j=1}^{l} c_j g_j(x,\theta)\right)\, dP_0(x) &= 1 \\ \int g_j(x,\theta)\varphi'^{-1}\left(c_0 + \sum_{j=1}^{l} c_j g_j(x,\theta)\right)\, dP_0(x) &= 0,\ j = 1, \ldots, l. \end{cases} \tag{3.43}$$

Since $Q \mapsto \phi(Q, P_0)$ is strictly convex, the projection Q_θ^* of P_0 on the convex set $\mathcal{M}_\theta$ is unique. This implies, by Theorem 3.3 part 2, that the solution

$$c_\theta := (c_0, c_1, \ldots, c_l)^T$$

of the system (3.43) is unique provided that the functions $g_i(\theta)$ are linearly independent.

Further, using the dual representation (3.32), we get

$$\phi(\mathcal{M}_\theta, P_0) := \phi(Q_\theta^*, P_0) = \sup_{f \in \mathcal{F}} \left\{ \int f \ dQ_\theta^* - \int \varphi^*(f) \ dP_0 \right\},$$

and the sup is unique and is reached at $f = \varphi'(dQ_\theta^*/dP_0) = c_0 + \sum_{j=1}^{l} c_j g_j(.,\theta)$, if it belongs to $\mathcal{F}$ (see Remark 3.11). This motivates the choice of the class $\mathcal{F}$ through

$$\mathcal{F} := \left\{ x \to t^T \overline{g}(x,\theta) \quad \text{for} \quad t \quad \text{in} \quad \mathcal{C}_\theta \right\}.$$

It is the smallest class of functions that contains $\varphi'(dQ_\theta^*/dP_0)$ and which does not presume any knowledge on Q_θ^*.

We thus obtain, through (3.32)

$$\phi(\mathcal{M}_\theta, P_0) = \sup_{t \in \mathcal{C}_\theta} \int m(x,\theta,t) \ dP_0(x),$$

where $m(\theta, t)$ is the function defined on $\mathcal{X}$ by

$$\begin{aligned} x \in \mathcal{X} \ \mapsto \ & m(x,\theta,t) := t_0 - \varphi^*\left(t^T \overline{g}(x,\theta)\right) = \\ & t_0 - \left(t^T \overline{g}(x,\theta)\right) \varphi'^{-1}\left(t^T \overline{g}(x,\theta)\right) + \varphi\left(\varphi'^{-1}\left(t^T \overline{g}(x,\theta)\right)\right), \end{aligned}$$

which is (3.33) in the present setting.

With the above notation, we state

$$\phi(\mathcal{M}_\theta, P_0) = \sup_{t \in \mathcal{C}_\theta} P_0 m(\theta, t). \tag{3.44}$$

So, a natural estimate of $\phi(\mathcal{M}_\theta, P_0)$ is

$$\sup_{t \in \mathcal{C}_\theta^{(n)}} P_n m(\theta, t) \tag{3.45}$$

which coincides with the estimate defined in (3.20). Hence, we can write

$$\widehat{\phi}(\mathcal{M}_\theta, P_0) = \sup_{t \in \mathcal{C}_\theta^{(n)}} P_n m(\theta, t). \tag{3.46}$$

which transforms the constrained optimization in (3.20) into the above unconstrained one.

On the other hand, the sup in (3.44) is reached at $t_0 = c_0, \ldots, t_l = c_l$ which are solutions of the system of equations (3.43). i.e.,

$$c_\theta = \arg\sup_{t \in \mathcal{C}_\theta} P_0 m(\theta, t). \tag{3.47}$$

So, a natural estimate of c_θ in (3.47) is therefore defined through

$$\arg \sup_{t \in \mathcal{C}_\theta^{(n)}} P_n m(\theta, t). \tag{3.48}$$

This coincides with $\widehat{c_\theta}$, the solution of the system of equations (3.38). So, we can write

$$\widehat{c_\theta} = \arg \sup_{t \in \mathcal{C}_\theta^{(n)}} P_n m(\theta, t). \tag{3.49}$$

Using (3.46), we obtain the following representations for the estimates $\widehat{\phi}(\mathcal{M}, P_0)$ in (3.21) and $\widehat{\theta}_\phi$ in (3.22)

$$\widehat{\phi}(\mathcal{M}, P_0) = \inf_{\theta \in \Theta} \sup_{t \in \mathcal{C}_\theta^{(n)}} P_n m(\theta, t) \tag{3.50}$$

and

$$\widehat{\theta}_\phi = \arg\inf_{\theta \in \Theta} \sup_{t \in \mathcal{C}_\theta^{(n)}} P_n m(\theta, t), \tag{3.51}$$

respectively.

Formula (3.44) also has the following basic interest : Consider the function

$$t \in \mathcal{C}_\theta \mapsto P_0 m(\theta, t), \tag{3.52}$$

In order for integral (3.52) to be properly defined, we assume that

$$\int |g_i(x,\theta)| \; dP_0(x) < \infty, \quad \text{for all} \quad i \in \{1, \ldots, l\}\,. \tag{3.53}$$

The domain of the function (3.52) is

$$\mathcal{D}_\phi(\theta) := \{t \in \mathcal{C}_\theta \;\text{ such that }\; P_0 m(\theta, t) > -\infty\}\,. \tag{3.54}$$

This function $t \mapsto P_0 m(\theta, t)$ is strictly concave on the convex set $\mathcal{D}_\phi(\theta)$. Whenever it has a maximum t^*, then it is unique, and if it belongs to the interior of $\mathcal{D}_\phi(\theta)$, then it satisfies the first order condition. Therefore t^* satisfies system (3.43). In turn, this implies that the measure Q^* defined through $dQ^* := \varphi'^{-1}\left(t^{*T}\overline{g}(\theta)\right)\; dP_0$ is the projection of P_0 on Ω, by Theorem 3.3 part 2. This implies that Q^* and P_0 share the same support. We summarize the above arguments as follows

Proposition 3.5. *Assume that (3.53) holds and that*

(i) there exists some s in the interior of $\mathcal{D}_\phi(\theta)$ such that for all t in $\partial\mathcal{D}_\phi(\theta)$, the frontier of $\mathcal{D}_\phi(\theta)$, it holds $P_0 m(\theta, t) < P_0 m(\theta, s)$;

(ii) for all t in the interior of $\mathcal{D}_\phi(\theta)$, there exists a neighborhood $V(t)$ of t, such that the classes of functions $\left\{x \to \frac{\partial}{\partial r_i} m(x, \theta, r), \quad r \in V(t)\right\}$ are dominated (P_0-a.s.) by some P_0-integrable function $x \to H(x, \theta)$.

Then P_0 admits an unique projection Q_θ^* on $\mathcal{M}_\theta$ having the same support as P_0 and

$$dQ_\theta^* = \varphi'^{-1}\left({c_\theta}^T \overline{g}(\theta)\right) dP_0, \tag{3.55}$$

where c_θ is the unique solution of the system of equations (3.43).

Remark 3.13. *In the case of KL-divergence, comparing this Proposition with Theorem 3.3 in Csiszár (1975), we observe that the dual formula (3.44) provides weaker conditions on the class of functions $\{\overline{g}(\theta),\ \theta \in \Theta\}$ than the geometric approach.*

Remark 3.14. *The result of Borwein and Lewis (1991), with some additional conditions, provides more practical tools for obtaining the results in Proposition 3.5. Assume that the functions $g_j(\theta)$ belongs to the space $L_p(\mathcal{X}, P_0)$ with $1 \le p \le \infty$ and that the following "constraint qualification" holds*

$$\text{there exists } Q_0 \text{ in } \mathcal{M}_\theta \text{ such that : } \; a < \inf \frac{dQ_0}{dP_0} \le \sup \frac{dQ_0}{dP_0} < b, \tag{3.56}$$

with $\mathcal{M}_\theta$ is the set of all signed measures Q a.c. w.r.t. P_0, satisfying the linear constraints and such that $\frac{dQ}{dP_0}$ belong to $L_q(\mathcal{X}, P_0)$, ($1 \leq q \leq \infty$ and $1/p + 1/q = 1$). In this case, applying Corollary 2 in Borwein and Lewis (1991), we obtain

$$\phi\left(\mathcal{M}_\theta, P_0\right) = \sup_{t \in \mathbb{R}^{(l+1)}} \left\{ t_0 - \int \psi\left(t^T \overline{g}(x,\theta)\right) \, dP_0(x) \right\}$$

(with dual attainement). Furthermore, if for a dual optimal c_θ, it holds

$$\lim_{y \downarrow -\infty} \frac{\varphi(y)}{y} < \inf_x c_\theta^T \overline{g}(x,\theta) \leq \sup_x c_\theta^T \overline{g}(x,\theta) < \lim_{y \uparrow +\infty} \frac{\varphi(y)}{y},$$

then the unique projection Q_θ^* of P_0 on $\mathcal{M}_\theta$ is given by

$$dQ_\theta^* = \psi'\left({c_\theta}^T \overline{g}(\theta)\right) dP_0. \tag{3.57}$$

Note that if ψ is strictly convex, then c_θ is unique and

$$\sup_{t \in \mathbb{R}^{(l+1)}} \left\{ t_0 - \int \psi\left(t^T \overline{g}(x,\theta)\right) \, dP_0(x) \right\} = \sup_{t \in \mathcal{C}_\theta} \left\{ t_0 - \int \varphi^*\left(t^T \overline{g}(x,\theta)\right) \, dP_0(x) \right\},$$

$$\psi'\left({c_\theta}^T \overline{g}(\theta)\right) = {\varphi'}^{-1}\left({c_\theta}^T \overline{g}(\theta)\right).$$

Léonard (2001b) and Léonard (2001a) gives, under minimal conditions, duality theorems of minimum ϕ-divergences and characterization of projections under linear constraints, which generalize the results given by Borwein and Lewis (1991) and Borwein and Lewis (1993). These results are used recently by Bertail (2003) in empirical likelihood.

3.5 Asymptotic properties and Statistical Tests

In the sequel, we assume that the conditions in Proposition 3.4 (or Remark 3.12) and in Proposition 3.5 (or Remark 3.14) hold. This allows to use the representations (3.46), (3.50) and (3.51) in order to study the asymptotic behavior of the proposed estimates (3.20), (3.21) and (3.22). All the results in the present Section are obtained through classical methods of parametric statistics ; see e.g. van der Vaart (1998) and Sen and Singer (1993). We first consider the case when θ is fixed, and we study the asymptotic behavior of the estimate $\widehat{\phi}(\mathcal{M}_\theta, P_0)$ (see (3.20)) of $\phi(\mathcal{M}_\theta, P_0) := \inf_{Q \in \mathcal{M}_\theta} \phi(Q, P_0)$ both when $P_0 \in \mathcal{M}_\theta$ and when $P_0 \notin \mathcal{M}_\theta$. This is done in the first Subsection. In the second Subsection, we study the asymptotic behavior of the EMϕD estimates $\widehat{\theta}_\phi$ and the estimates $\widehat{\phi}\left(\mathcal{M}, P_0\right)$ both in the two cases when P_0 belongs to $\mathcal{M}$ and when P_0 does not belong to $\mathcal{M}$.

The solution of *Problem 1* is given in Subsection 5.3 while *Problem 2* is treated in Subsections 5.1, 5.2, 5.3 and 5.4.

3.5.1 Asymptotic properties of the estimates for a given $\theta \in \Theta$

First we state consistency.

Consistency

We state both weak and strong consistency of the estimates $\widehat{c_\theta}$ and $\widehat{\phi}(\mathcal{M}_\theta, P_0)$ using their representations (3.49) and (3.46), respectively. Denote $\|.\|$ the Euclidian norm defined on $\mathbb{R}^d$ or on $\mathbb{R}^{l+1}$. In order to state consistency, we need to define

$$T_\theta := \{t \in \mathcal{C}_\theta \text{ such that } P_0 m(\theta, t) > -\infty\},$$

and denote T_θ^c the complementary of the set T_θ in the set $\mathcal{C}_\theta$, namely

$$T_\theta^c := \{t \in \mathcal{C}_\theta \text{ such that } P_0 m(\theta, t) = -\infty\}.$$

Note that, by Proposition 3.5, the set T_θ contains c_θ.

We will consider the following condition

(C.1) $\sup_{t \in T_\theta} |P_n m(\theta, t) - P_0 m(\theta, t)|$ converges to 0 a.s. (resp. in probability);

(C.2) there exists $M < 0$ and $n_0 > 0$, such that, for all $n > n_0$, it holds $\sup_{t \in T_\theta^c} P_n m(\theta, t) \leq M$ a.s. (resp. in probability).

The condition (C.2) makes sense, since for all $t \in T_\theta^c$ we have $P_0 m(\theta, t) = -\infty$.

Since the function $t \in T_\theta \mapsto P_0 m(\theta, t)$ is strictly concave, the maximum c_θ is isolated, that is

$$\text{for any positive } \epsilon, \text{ we have } \sup_{\{t \in \mathcal{C}_\theta \,:\, \|t - c_\theta\| \geq \epsilon\}} P_0 m(\theta, t) < P_0 m(\theta, c_\theta). \tag{3.58}$$

Proposition 3.6. *Assume that conditions (C.1) and (C.2) hold. Then*

(i) the estimates $\widehat{\phi}(\mathcal{M}_\theta, P_0)$ converge to $\phi(\mathcal{M}_\theta, P_0)$ a.s. (resp. in probability).

(ii) the estimates $\widehat{c_\theta}$ converge to c_θ a.s. (resp. in probability).

Asymptotic distributions

Denote $m'(\theta, t)$ the $(l+1)$-dimensional vector with entries $\frac{\partial}{\partial t_i} m(\theta, t)$, $m''(\theta, t)$ the $(l+1) \times (l+1)$-matrix with entries $\frac{\partial^2}{\partial t_i \partial t_j} m(\theta, t)$, $\underline{0}_l := (0, \ldots, 0)^T \in \mathbb{R}^l$, $\underline{0}_d := (0, \ldots, 0)^T \in \mathbb{R}^d$, $\underline{c}$ the $(l+1)-$vector defined by $\underline{c} := \left(0, \underline{0}_l^T\right)^T$, and $P_0 g(\theta) g(\theta)^T$ the $l \times l-$matrix defined by

$$P_0 g(\theta) g(\theta)^T := [P_0 g_i(\theta) g_j(\theta)]_{i,j=1,\ldots,l}.$$

We will consider the following assumptions

(A.1) $\widehat{c_\theta}$ converges in probability to c_θ;

(A.2) the function $t \mapsto m(x,\theta,t)$ is $\mathcal{C}^3$ on a neighborhood $V(c_\theta)$ of c_θ for all x (P_0-a.s.), and all partial derivatives of order 3 of the function $\{t \mapsto m(x,\theta,t),\ t \in V(c_\theta)\}$ are dominated by some P_0-integrable function $x \mapsto H(x)$;

(A.3) $P_0\left(\|m'(\theta,c_\theta)\|^2\right)$ is finite, and the matrix $P_0 m''(\theta,c_\theta)$ exists and is invertible.

Theorem 3.4. *Assume that assumptions (A.1-3) hold. Then*

(1) $\sqrt{n}\left(\widehat{c_\theta} - c_\theta\right)$ *converges to a centered normal multivariate variable with covariance matrix*

$$V = \left[-P_0 m''(\theta,c_\theta)\right]^{-1}\left[P_0 m'(\theta,c_\theta) m'(\theta,c_\theta)^T\right]\left[-P_0 m''(\theta,c_\theta)\right]^{-1}. \tag{3.59}$$

In the special case, when P_0 *belongs to* $\mathcal{M}_\theta$*, then* $c_\theta = \underline{c}$ *and*

$$V = \varphi''(1)^2 \begin{bmatrix} 0 & \underline{0}_l^T \\ \underline{0}_l & \left[P_0 g(\theta) g(\theta)^T\right]^{-1} \end{bmatrix}. \tag{3.60}$$

(2) If P_0 *belongs to* $\mathcal{M}_\theta$*, then the statistics*

$$\frac{2n}{\varphi''(1)}\widehat{\phi}\left(\mathcal{M}_\theta, P_0\right)$$

converge in distribution to a χ^2 *variable with* l *degrees of freedom.*

(3) If P_0 *does not belong to* $\mathcal{M}_\theta$*, then*

$$\sqrt{n}\left(\widehat{\phi}\left(\mathcal{M}_\theta, P_0\right) - \phi(\mathcal{M}_\theta, P_0)\right)$$

converges to a centered normal variable with variance

$$\sigma^2 := P_0 m(\theta,c_\theta)^2 - \left(P_0 m(\theta,c_\theta)\right)^2.$$

Remark 3.15.

(a) When specialized to the modified Kullback-Leibler divergence, Theorem 3.4 part (2) gives the limiting distribution of the empirical log-likelihood ratio $2n\widehat{KL_m}(\mathcal{M}_\theta, P_0)$ *which is the result in Owen (1990) Theorem 1.*
Part (3) gives its limiting distribution when P_0 *does not belong to* $\mathcal{M}_\theta$*.*

(b) Nonparametric confidence regions (CR_ϕ) *for* θ_0 *of asymptotic level* $(1-\epsilon)$ *can be constructed using the statistics*

$$\frac{2n}{\varphi''(1)}\widehat{\phi}\,(\mathcal{M}_\theta, P_0)\,,$$

through

$$CR_\phi := \left\{\theta \in \Theta \;\; \textit{such that} \;\; \frac{2n}{\varphi''(1)}\widehat{\phi}\,(\mathcal{M}_\theta, P_0) \leq \; q_{(1-\epsilon)}\right\},$$

where $(1-\epsilon)$ *is the* $(1-\epsilon)$*-quantile of a* $\chi^2(d)$ *distribution. It would be interesting to obtain the divergence leading to optimal confidence regions in the sense of Neyman (1937) (see Takagi (1998)), or the optimal divergence leading to confidence regions with small length (volume, area or diameter) and covering the true value* θ_0 *with large enough probability.*

3.5.2 Asymptotic properties of the estimates $\widehat{\theta}_\phi$ and $\widehat{\phi}(\mathcal{M}, P_0)$

First we state consistency.

Consistency

We assume that when P_0 does not belong to the model $\mathcal{M}$, the minimum, say θ^*, of the function $\theta \in \Theta \mapsto \inf_{Q\in\mathcal{M}_\theta} \phi(Q, P_0)$ exists and is unique. Hence P_0 admits a projection on $\mathcal{M}$ which we denote $Q^*_{\theta^*}$. Obviously when P_0 belongs to the model $\mathcal{M}$, then $\theta^* = \theta_0$ and $Q^*_{\theta^*} = P_0$. We will consider the following conditions

(C.3) $\sup_{\{\theta\in\Theta, t\in T_\theta\}} |P_n m(\theta, t) - P_0 m(\theta, t)|$ tends to 0 a.s. (resp. in probability) ;

(C.4) there exists a neighborhood $V(c_{\theta^*})$ of c_{θ^*} such that

- (a) for any positive ϵ, there exists some positive η such that for all $t \in V(c_{\theta^*})$ and all $\theta \in \Theta$ satisfying $\|\theta - \theta^*\| \geq \epsilon$, it holds $P_0 m(\theta^*, t) < P_0 m(\theta, t) - \eta$;
- (b) there exists some function H such that for all t in $V(c_{\theta^*})$, we have $|m(t, \theta_0)| \leq H(x)$ $(P_0$-a.s.) with $P_0 H < \infty$;

(C.5) there exits $M < 0$ and $n_0 > 0$ such that for all $n \geq n_0$, we have

$$\sup_{\theta\in\Theta} \sup_{t\in T^c_\theta} P_n m(\theta, t) \leq M \;\; \text{a.s. (resp. in probability).} \tag{3.61}$$

Proposition 3.7. *Assume that conditions (C.3-5) hold. Then*

(i) the estimates $\widehat{\phi}(\mathcal{M}, P_0)$ *converge to* $\phi(\mathcal{M}, P_0)$ *a.s. (resp. in probability).*

(ii) $\sup_{\theta\in\Theta} \|\widehat{c_\theta} - c_\theta\|$ *converge to* 0 *a.s. (resp. in probability).*

(iii) The MEϕ*D estimates* $\widehat{\theta}_\phi$ *converge to* θ^* *a.s. (resp. in probability).*

Asymptotic distributions

When $P_0 \in \mathcal{M}$, then by assumption, there exists unique $\theta_0 \in \Theta$ such that $P_0 \in \mathcal{M}_{\theta_0}$. Hence $\theta^* = \theta_0$ and $c_{\theta^*} = c_{\theta_0} = \underline{c}$. We state the limit distributions of the estimates $\widehat{\theta}_\phi$ and $\widehat{c_{\widehat{\theta}_\phi}}$ when $P_0 \in \mathcal{M}$ and when $P_0 \notin \mathcal{M}$. We will make use of the following assumptions

(A.4) Both estimates $\widehat{\theta}_\phi$ and $\widehat{c_{\widehat{\theta}_\phi}}$ converge in probability respectively to θ^* and c_{θ^*};

(A.5) the function $(\theta, t) \mapsto m(x, \theta, t)$ is $\mathcal{C}^3$ on some neighborhood $V(\theta^*, c_{\theta^*})$ for all x (P_0-a.s.), and the partial derivatives of order 3 of the functions $\{(\theta, t) \mapsto m(x, \theta, t),\ (\theta, t) \in V(\theta^*, c_{\theta^*})\}$ are dominated by some P_0-integrable function $H(x)$;

(A.6) $P_0\left(\left\|\frac{\partial}{\partial t} m(\theta^*, c_{\theta^*})\right\|^2\right)$ and $P_0\left(\left\|\frac{\partial}{\partial \theta} m(\theta^*, c_{\theta^*})\right\|^2\right)$ are finite, and the matrix

$$S := \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

with $S_{11} := P_0 \frac{\partial^2}{\partial t^2} m(\theta^*, c_{\theta^*})$, $S_{12} = {S_{21}}^T := P_0 \frac{\partial^2}{\partial t \partial \theta} m(\theta^*, c_{\theta^*})$ and $S_{22} := P_0 \frac{\partial^2}{\partial \theta^2} m(\theta^*, c_{\theta^*})$, exists and is invertible.

Theorem 3.5. *Let P_0 belongs to $\mathcal{M}$ and assumptions (A.4-6) hold. Then, both $\sqrt{n}\left(\widehat{\theta}_\phi - \theta_0\right)$ and $\sqrt{n}\left(\widehat{c_{\widehat{\theta}_\phi}} - \underline{c}\right)$ converge in distribution to a centered multivariate normal variable with covariance matrix, respectively*

$$V = \left\{ \left[P_0 \frac{\partial}{\partial \theta} g(\theta_0) \right] \left[P_0 \left(g(\theta_0) g(\theta_0)^T \right) \right]^{-1} \left[P_0 \frac{\partial}{\partial \theta} g(\theta_0) \right]^T \right\}^{-1}, \tag{3.62}$$

and

$$\begin{aligned} U \;=\; & \varphi''(1)^2 \begin{bmatrix} 0 & \underline{0}_l^T \\ \underline{0}_l & \left[P_0 g(\theta_0) g(\theta_0)^T\right]^{-1} \end{bmatrix} - \varphi''(1)^2 \begin{bmatrix} 0 & \underline{0}_l^T \\ \underline{0}_l & \left[P_0 g(\theta_0) g(\theta_0)^T\right]^{-1} \end{bmatrix} \times \\ & \times \left[\underline{0}_d, P_0 \frac{\partial}{\partial \theta} g(\theta_0) \right]^T V \left[\underline{0}_d, P_0 \frac{\partial}{\partial \theta} g(\theta_0) \right] \begin{bmatrix} 0 & \underline{0}_l^T \\ \underline{0}_l & \left[P_0 g(\theta_0) g(\theta_0)^T\right]^{-1} \end{bmatrix}, \end{aligned}$$

and the estimates $\widehat{\theta}_\phi$ and $\widehat{c_{\widehat{\theta}_\phi}}$ are asymptotically uncorrelated.

Remark 3.16. *When specialized to the modified Kullback-Leibler divergence, then the estimate $\widehat{\theta}_{KL_m}$ is the empirical likelihood estimate (ELE) (noted $\widetilde{\theta}$ in Qin and Lawless (1994)), and the above result gives the limiting distribution of $\sqrt{n}(\widehat{\theta}_{KL_m} - \theta_0)$ which coincides with the result in Theorem 1 in Qin and Lawless (1994). Note also that all MEϕDE's including ELE have the same limiting distribution with the same variance when P_0 belongs to $\mathcal{M}$. Hence they are all equally first order efficient.*

Theorem 3.6. *Assume that P_0 does not belong to $\mathcal{M}$ and that assumptions (A.4-6) hold. Then*

$$\sqrt{n}\begin{pmatrix} \widehat{c_{\widehat{\theta}_\phi}} - c_{\theta^*} \\ \widehat{\theta}_\phi - \theta^* \end{pmatrix}$$

converges in distribution to a centered multivariate normal variable with covariance matrix

$$W = S^{-1}MS^{-1}$$

where

$$M := P_0\left(\begin{pmatrix} \frac{\partial}{\partial t}m\left(\theta^*, c_{\theta^*}\right) \\ \frac{\partial}{\partial \theta}m\left(\theta^*, c_{\theta^*}\right) \end{pmatrix}\begin{pmatrix} \frac{\partial}{\partial t}m\left(\theta^*, c_{\theta^*}\right) \\ \frac{\partial}{\partial \theta}m\left(\theta^*, c_{\theta^*}\right) \end{pmatrix}^T\right).$$

θ^* and c_{θ^*} are characterized by

$$\theta^* := \arg\inf_{\theta\in\Theta}\phi\left(\mathcal{M}_\theta, P_0\right),$$

$$dQ^*_{\theta^*} = \varphi'^{-1}\left(c^T_{\theta^*}\overline{g}(\theta)\right)dP_0 \quad \textit{and} \quad Q^*_{\theta^*} \in \mathcal{M}_{\theta^*}.$$

3.5.3 Tests of model

In order to test the hypothesis $\mathcal{H}_0 : P_0$ belongs to $\mathcal{M}$ against the alternative $\mathcal{H}_1 : P_0$ does not belong to $\mathcal{M}$, we can use the estimates $\widehat{\phi}(\mathcal{M}, P_0)$ of $\phi(\mathcal{M}, P_0)$, the $\phi-$divergences between the model $\mathcal{M}$ and the distribution P_0. Since $\phi(\mathcal{M}, P_0)$ is nonnegative and take value 0 only when P_0 belongs to $\mathcal{M}$ (provided that P_0 admits a projection on $\mathcal{M}$), we reject the hypothesis $\mathcal{H}_0$ when the estimates take large values. In the following Corollary, we give the asymptotic law of the estimates $\widehat{\phi}(\mathcal{M}, P_0)$ both under $\mathcal{H}_0$ and under $\mathcal{H}_1$.

Corollary 3.1.

(i) Assume that the assumptions of Theorem 3.5 hold and that $l > d$. Then, under $\mathcal{H}_0$, the statistics

$$\frac{2n}{\varphi''(1)}\widehat{\phi}(\mathcal{M}, P_0)$$

converge in distribution to a χ^2 variable with $(l-d)$ degrees of freedom.

(ii) Assume that the assumptions of Theorem 3.6 hold. Then, under $\mathcal{H}_1$, we have :

$$\sqrt{n}\left(\widehat{\phi}(\mathcal{M}, P_0) - \phi(\mathcal{M}, P_0)\right) \tag{3.63}$$

converges to centered normal variable with variance

$$\sigma^2 = P_0 m(\theta^*, c_{\theta^*})^2 - \left(P_0 m(\theta^*, c_{\theta^*})\right)^2$$

where θ^* and c_{θ^*} satisfy

$$\theta^* := \arg \inf_{\theta \in \Theta} \phi\left(\mathcal{M}_\theta, P_0\right),$$

$$\varphi'\left(\frac{dQ^*_{\theta^*}}{dP_0}(x)\right) = c^T_{\theta^*} \overline{g}(x, \theta^*) \qquad \textit{and} \qquad Q^*_{\theta^*} \in \mathcal{M}_{\theta^*}.$$

Remark 3.17. *This Theorem allows to perform tests of model of asymptotic level α ; the critical regions are*

$$C_\phi := \left\{ \frac{2n}{\varphi''(1)} \widehat{\phi}(\mathcal{M}, P_0) > q_{(1-\alpha)} \right\}, \tag{3.64}$$

where $q_{(1-\alpha)}$ is the $(1-\alpha)-$quantile of the χ^2 distribution with $(l-d)$ degrees of freedom. Also these tests are all asymptotically powerful, since the estimates $\widehat{\phi}(\mathcal{M}, P_0)$ are $n-$consistent estimates of $\phi(\mathcal{M}, P_0) = 0$ under $\mathcal{H}_0$ and $\sqrt{n}-$consistent estimates of $\phi(\mathcal{M}, P_0)$ under $\mathcal{H}_1$.

We assume now that the p.m. P_0 belongs to $\mathcal{M}$. We will perform simple and composite tests on the parameter θ_0 taking account of the information $P_0 \in \mathcal{M}$.

3.5.4 Simple tests on the parameter

Let

$$\mathcal{H}_0 \quad : \quad \theta_0 = \theta_1 \quad \text{versus} \quad \mathcal{H}_1 \quad : \quad \theta_0 \in \Theta \setminus \{\theta_1\}, \tag{3.65}$$

where θ_1 is a given known value. We can use the following statistics to perform tests pertaining to (3.65)

$$S^\phi_n := \widehat{\phi}(\mathcal{M}_{\theta_1}, P_0) - \inf_{\theta \in \Theta} \widehat{\phi}(\mathcal{M}_\theta, P_0).$$

Since

$$\phi(\mathcal{M}_{\theta_1}, P_0) - \inf_{\theta \in \Theta} \phi(\mathcal{M}_\theta, P_0) = \phi(\mathcal{M}_{\theta_1}, P_0)$$

are nonnegative and take value 0 only when $\theta_0 = \theta_1$, we reject the hypothesis $\mathcal{H}_0$ when the statistics S^ϕ_n take large values.

We give the limit distributions of the statistics S^ϕ_n in the following Corollary which we can prove using some algebra and arguments used in the proof of Theorem 3.5 and Theorem 3.6.

Corollary 3.2.

(i) Assume that assumptions of Theorem 3.5 hold. Then under $\mathcal{H}_0$, the statistics

$$\frac{2n}{\varphi''(1)} S_n^{\phi}$$

converge in distribution to χ^2 variable with d degrees of freedom.

(ii) Assume that assumptions of Theorem 3.5 hold. Then under $\mathcal{H}_1$,

$$\sqrt{n}\left(S_n^{\phi} - \phi\left(\mathcal{M}_{\theta_1}, P_0\right)\right)$$

converges to a centered normal variable with variance

$$\sigma^2 = P_0 m(\theta_1, c_{\theta_1})^2 - \left(P_0 m(\theta_1, c_{\theta_1})\right)^2.$$

Remark 3.18. *When specialized to the KL_m-divergence, the statistic $2nS_n^{KL_m}$ is the empirical likelihood ratio statistic (see Qin and Lawless (1994) Theorem 2).*

3.5.5 Composite tests on the parameter

Let

$$h \ : \ \mathbb{R}^d \ \rightarrow \ \mathbb{R}^k \tag{3.66}$$

be some function such that the $(d \times k)-$matrix $H(\theta) := \frac{\partial}{\partial \theta} h(\theta)$ exists, is continuous and has rank k with $0 < k < d$. Let us define the composite null hypothesis

$$\Theta_0 := \{\theta \in \Theta \ \text{ such that } \ h(\theta) = 0\}. \tag{3.67}$$

We consider the composite test

$$\mathcal{H}_0 \ : \ \theta_0 \in \Theta_0 \quad \text{versus} \quad \mathcal{H}_1 \ : \ \theta_0 \in \Theta \setminus \Theta_0, \tag{3.68}$$

i.e., the test

$$\mathcal{H}_0 \ : \ P_0 \in \bigcup_{\theta \in \Theta_0} \mathcal{M}_\theta \quad \text{versus} \quad \mathcal{H}_1 \ : \ P_0 \in \bigcup_{\theta \in \Theta \setminus \Theta_0} \mathcal{M}_\theta. \tag{3.69}$$

This test is equivalent to the following one

$$\mathcal{H}_0 \ : \ \theta_0 \in f(B_0) \quad \text{versus} \quad \mathcal{H}_1 \ : \ \theta_0 \notin f(B_0), \tag{3.70}$$

where $f \ : \ \mathbb{R}^{(d-k)} \ \rightarrow \ \mathbb{R}^d$ is a function such that the matrix $G(\beta) := \frac{\partial}{\partial \beta} g(\beta)$ exists and has rank $(d-k)$, and $B_0 := \left\{\beta \in \mathbb{R}^{(d-k)} \text{ such that } f(\beta) \in \Theta_0\right\}$. Therefore $\theta_0 \in \Theta_0$ is an equivalent statement for $\theta_0 = f(\beta_0), \beta_0 \in B_0$.

The following statistics are used to perform tests pertaining to (3.70) :

$$T_n^\phi := \inf_{\beta\in B_0} \widehat{\phi}\left(\mathcal{M}_{f(\beta)}, P_0\right) - \inf_{\theta\in\Theta} \widehat{\phi}\left(\mathcal{M}_\theta, P_0\right).$$

Since

$$\inf_{\beta\in B_0} \phi(\mathcal{M}_{f(\beta)}, P_0) - \inf_{\theta\in\Theta} \phi(\mathcal{M}_\theta, P_0) = \inf_{\beta\in B_0} \phi(\mathcal{M}_{f(\beta)}, P_0)$$

are nonnegative and take value 0 only when $\mathcal{H}_0$ holds, we reject the hypothesis $\mathcal{H}_0$ when the statistics T_n^ϕ take large values.

We give the limit distributions of the statistics T_n^ϕ in the following Corollary.

Corollary 3.3.

(i) Assume that assumptions of Theorem 3.5 hold. Under $\mathcal{H}_0$, the statistics T_n^ϕ converge in distribution to a χ^2 variable with $(d-k)$ degrees of freedom.

(ii) Assume that there exists $\beta^* \in B_0$, such that $\beta^* = \arg\inf_{\beta\in B_0} \phi\left(\mathcal{M}_{f(\beta)}, P_0\right)$. If the assumptions of Theorem 3.6 hold for $\theta^* = f(\beta^*)$, then

$$\sqrt{n}\left(T_n^\phi - \phi\left(\mathcal{M}_{\theta^*}, P_0\right)\right)$$

converges to a centered normal variable with variance

$$\sigma^2 = P_0 m(\theta^*, c_{\theta^*})^2 - \left(P_0 m(\theta^*, c_{\theta^*})\right)^2.$$

3.6 Estimates of the distribution function through projected distributions

In this Subsection, the measurable space $(\mathcal{X}, \mathcal{B})$ is $(\mathbb{R}, \mathcal{B}_\mathbb{R})$. For all $\phi-$divergence, by (3.21), we have

$$\widehat{\phi}\left(\mathcal{M}, P_0\right) = \phi\left(\mathcal{M}, P_n\right) = \phi\left(\widehat{Q^*_{\widehat{\theta}_\phi}}, P_n\right).$$

Proposition 3.36 above provides the description of $\widehat{Q^*_{\widehat{\theta}_\phi}}$.

So, for all ϕ-divergence, we estimate the distribution function F using $\widehat{Q^*_{\widehat{\theta}_\phi}}$ the $\phi-$projection of P_n on $\mathcal{M}$, through

$$\begin{aligned}\widehat{F}_n(x) &:= \sum_{i=1}^n \widehat{Q^*_{\widehat{\theta}_\phi}}(X_i)\mathbb{1}_{(-\infty,x]}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^n \overleftarrow{\varphi'}\left(\widehat{c_{\widehat{\theta}_\phi}}^T \overline{g}(X_i, \widehat{\theta}_\phi)\right)\mathbb{1}_{(-\infty,x]}(X_i). \qquad (3.71)\end{aligned}$$

Remark 3.19. *When the estimating equation*

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta) = \underline{0}_d \tag{3.72}$$

admits a solution $\widetilde{\theta}_n$, then P_n belongs to $\mathcal{M}$. If the solution is unique then $\widehat{\theta}_\phi = \widetilde{\theta}_n$. Hence by Proposition 3.4

$$\text{for all } \ i \in \{1,2,\ldots,n\}, \quad \text{we have} \quad \widehat{Q^*_{\widehat{\theta}_\phi}}(X_i) = \frac{1}{n},$$

and $\widehat{F}_n(x)$, in this case, is the empirical cumulative distribution function, i.e.,

$$\widehat{F}_n(x) = F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,x]}(X_i).$$

So, the main interest is in the case where (3.72) does not admit a solution, that is in general when $l > d$.

Remark 3.20. *The ϕ-projections $\widehat{Q^*_{\widehat{\theta}_\phi}}$ of P_n on $\mathcal{M}$ may be signed measures. For all ϕ-divergence satisfying $D_\varphi = \mathbb{R}^*_+$, the ϕ-projection $\widehat{Q^*_{\widehat{\theta}_\phi}}$ is a p.m. if it exists. (for example, KL_m, KL, Hellinger, and χ^2_m divergences all provide p.m.'s).*

We give the limit law of the estimates $\widehat{F}_n$ of the distribution function F in the following Theorem. We will see that the estimate $\widehat{F}_n(x)$ is generally more efficient than the empirical cumulative distribution function $F_n(x)$.

Theorem 3.7. *Under the assumptions of Theorem 3.5, $\sqrt{n}\left(\widehat{F}_n(x) - F(x)\right)$ converges in distribution to a centered normal variable with variance*

$$W(x) = F(x)\left(1-F(x)\right) - \left[P_0\left(g(\theta_0)\mathbb{1}_{(-\infty,x]}\right)\right]^T \Gamma \left[P_0\left(g(\theta_0)\mathbb{1}_{(-\infty,x]}\right)\right], \tag{3.73}$$

with

$$\begin{aligned}\Gamma &= \left[P_0 g(\theta_0) g(\theta_0)^T\right]^{-1} - \left[P_0 g(\theta_0) g(\theta_0)^T\right]^{-1} \left[P_0 \frac{\partial}{\partial\theta} g(\theta_0)\right]^T V \times \\ &\quad \times \left[P_0 \frac{\partial}{\partial\theta} g(\theta_0)\right] \left[P_0 g(\theta_0) g(\theta_0)^T\right]^{-1},\end{aligned}$$

and

$$V = \left\{ \left[P_0 \frac{\partial}{\partial\theta} g(\theta_0)\right] \left[P_0\left(g(\theta_0) g(\theta_0)^T\right)\right]^{-1} \left[P_0 \frac{\partial}{\partial\theta} g(\theta_0)\right]^T \right\}^{-1}.$$

3.7 Empirical likelihood and related methods

In the present setting, the empirical likelihood (EL) approach for the estimation of the parameter θ_0 can be summarized as follows. For any θ in Θ, define the *profile likelihood ratio* of the sample $X := (X_1, ..., X_n)$ through

$$L_n(\theta) := \sup \left\{ \prod_{i=1}^{n} nQ(X_i) \text{ where } Q(X_i) \geq 0,\ \sum_{i=1}^{n} Q(X_i) = 1,\ \sum_{i=1}^{n} g(X_i, \theta) Q(X_i) = 0 \right\}.$$

The estimate of θ_0 through empirical likelihood (EL) approach is then defined by

$$\widehat{\theta}_{EL} := \arg \sup_{\theta \in \Theta} L_n(\theta). \tag{3.74}$$

The paper by Qin and Lawless (1994) introduces $\widehat{\theta}_{EL}$ and presents its properties. In this Section, we show that $\widehat{\theta}_{EL}$ belongs to the family of MEϕD estimates for the specific choice $\varphi(x) = -\log x + x - 1$. We also discuss the problem of the existence of the solution of (3.74) for all n.

When $\varphi(x) = -\log x + x - 1$, formula (3.22) clearly coincides with $\widehat{\theta}_{EL}$. For test of hypotheses given by $\mathcal{H}_0 : P_0 \in \mathcal{M}_\theta$ against $\mathcal{H}_1 : P_0 \notin \mathcal{M}_\theta$ or for construction of nonparametric confidence regions for θ_0, the statistic $2n\widehat{KL_m}(\mathcal{M}_\theta, P_0)$ coincides with the empirical log-likelihood ratio introduced in Owen (1988), Owen (1990) and Qin and Lawless (1994). We state the results of Section 5 in the present context. We will see that the approach of empirical likelihood by divergence minimization, using the dual representation of the KL_m-divergence and the explicit form of the KL_m-projection of P_0, yields to the limit distribution of the statistic $2n\widehat{KL_m}(\mathcal{M}_\theta, P_0)$ under $\mathcal{H}_1$, which can not be achieved using the approach in Owen (1990) and Qin and Lawless (1994). Consider

$$\widehat{\theta}_{KL_m} = \arg \inf_{\theta \in \Theta} \widehat{KL_m}(\mathcal{M}_\theta, P_0)$$

where

$$\widehat{KL_m}(\mathcal{M}_\theta, P_0) = \sup_{t \in \mathcal{C}_\theta} P_n m(\theta, t) \tag{3.75}$$

with $\varphi(x) = \varphi_0(x) = -\log x + x - 1$. The explicit form of $m(\theta, t)$ in this case is

$$\begin{aligned} x \mapsto m(x, \theta, t) &= t_0 - \left(t^T \overline{g}(x, \theta)\right) \frac{1}{1 - t^T \overline{g}(x, \theta)} + \log\left(1 - t^T \overline{g}(x, \theta)\right) + \frac{1}{1 - t^T \overline{g}(x, \theta)} - 1. \\ &= t_0 + \log\left(1 - t^T \overline{g}(x, \theta)\right). \end{aligned} \tag{3.76}$$

For fixed $\theta \in \Theta$, the sup in (3.75), which we have noted $\widehat{c_\theta}$, satisfies the following system

$$\begin{cases} \int \frac{1}{1-c_0-\sum_{j=1}^{l} c_j g_j(x,\theta)}\, dP_n(x) &= 1 \\ \int \frac{g_j(x,\theta)}{1-c_0-\sum_{j=1}^{l} c_j g_j(x,\theta)}\, dP_n(x) &= 0, \quad \text{for all } j = 1, ..., l \end{cases} \tag{3.77}$$

a system of $(l+1)$ equations and $(l+1)$ variables. The projection $\widehat{Q_\theta^*}$ is then obtained using Proposition 3.4 part (ii). We have for all $i \in \{1, \ldots, n\}$

$$\frac{1}{\widehat{Q_\theta^*}(X_i)} = n\left(1 - c_0 - \sum_{j=1}^{n} c_j g_j(X_i, \theta)\right)$$

which, multiplying by $\widehat{Q_\theta^*}(X_i)$ and summing upon i yields $c_0 = 0$. Therefore the system (3.77) reduces to the system (3.3) in Qin and Lawless (1994) replacing $c_1, \ldots, c_l$ by $-t_1, \ldots, -t_l$. Simplify (3.76) plugging $t_0 = 0$. Notice that $2n\widehat{KL_m}(\mathcal{M}_\theta, P_0) = l_E(\theta_0)$ in the notation of Qin and Lawless (1994), and that the function of $t = (0, -\tau_1, \ldots, -\tau_l)$ defined by

$$t \mapsto P_n m(\theta, t)$$

coincide with the function

$$\tau \to P_n \log\left(1 + \tau^T g(., \theta)\right)$$

used in Qin and Lawless (1994). The interest in formula (3.75) lays in the obtention of the limit distributions of $2n\widehat{KL_m}(\mathcal{M}_\theta, P_0)$ under $\mathcal{H}_1$. By Theorem 3.4, we have

$$\sqrt{n}\left(\widehat{KL_m}(\mathcal{M}_\theta, P_0) - KL_m(\mathcal{M}_\theta, P_0)\right)$$

converges to a normal distribution variable, which proves consistency of the test ; this results cannot be obtained by the Qin and Lawless (1994)'s approach.

The choice of φ depends on some a priori knowledge on θ_0. Hopefully, some divergences do not have such an inconvenient. We now clarify this point. For fixed θ in Θ, let $\mathcal{M}_\theta^{(n)}$ and $\mathcal{D}_\phi^{(n)}$ be defined respectively as in (3.19) and in (3.34). Assume that $\mathcal{M}_\theta^{(n)} \cap \mathcal{D}_\phi^{(n)}$ is not void. Then P_n has a projection $\widehat{Q_\theta^*}$ on $\mathcal{M}_\theta^{(n)}$ and $\phi(\widehat{Q_\theta^*}, P_n)$ is finite.
The estimation of θ_0 is achieved minimizing $\widehat{\phi}(\mathcal{M}_\theta, P_0)$ on the sets

$$\Theta_n^\phi := \left\{\theta \in \Theta \text{ such that } \mathcal{M}_\theta^{(n)} \cap \mathcal{D}_\phi^{(n)} \text{ is not void}\right\}.$$

Clearly the description of Θ_n^ϕ depends on the divergence ϕ. Consider the following example, with $n = 2$, $X = (X_1, X_2)$ and $g(x, \theta) = x - \theta$. Then

$$\mathcal{M}_\theta = \left\{(q_1, q_2)^T \text{ such that } q_1 + q_2 = 1 \text{ and } q_1(X_1 - \theta) + q_2(X_2 - \theta) = 0\right\}$$

and

$$\mathcal{D}_{\phi}^{(2)} = \left\{ (q_1, q_2) \ \text{ such that } \ \frac{1}{2} \sum_{i=1}^{2} \varphi(2q_i) < \infty \right\}.$$

When $\phi = KL_m$, then $\mathcal{D}_{KL_m}^{(2)} = \mathbb{R}_+^* \times \mathbb{R}_+^*$. So, according to the value of θ, $\mathcal{M}_{\theta}^{(n)} \cap \mathcal{D}_{KL_m}^{(n)}$ may be void and therefore $\Theta_n^{KL_m}$ has a complex structure. At the opposite, for example when $\phi = \chi^2$, then $\mathcal{D}_{\chi^2}^{(2)} = \mathbb{R}^2$. Hence $\mathcal{M}_{\theta}^{(n)} \cap \mathcal{D}_{\phi}^{(n)} = \mathcal{M}_{\theta}^{(n)}$ which is not void for all θ and hence $\Theta_n^{\chi^2} = \Theta$.

On the other hand, we have for any ϕ-divergence

$$\begin{aligned} \widehat{\theta}_{\phi} &:= \arg \inf_{\theta \in \Theta} \inf_{Q \in \mathcal{M}_{\theta}^{(n)}} \widehat{\phi}(Q, P_0) \\ &= \arg \inf_{\theta \in \Theta} \inf_{Q \in \mathcal{M}_{\theta}^{(n)} \cap \mathcal{D}_{\phi}^{(n)}} \widehat{\phi}(Q, P_0). \end{aligned}$$

When $\mathcal{D}_{\phi}^{(n)} \neq \mathbb{R}^n$, the infimum in θ above should be taken upon Θ_n^{ϕ} which might be quite cumbersome. Owen (2001) indeed mentions such a difficulty.

In relation to this problem, Qin and Lawless (1994) bring some asymptotic arguments in the case of the empirical likelihood. They show that there exists a sequence of neighborhoods

$$V_n(\theta_0) := \left\{ \theta \ \text{ such that } \ \|\theta - \theta_0\| \leq n^{-1/3} \right\}$$

on which, with probability one as n tends to infinity, $L_n(\theta)$ has a maximum. This turns out, in the context of ϕ-divergences, to write that the mapping

$$\theta \mapsto \inf_{Q \in \mathcal{M}_{\theta}^{(n)}} KL_m(Q, P_n)$$

has a minimum when θ belongs to $V_n(\theta_0)$. This interesting result does not solve the problem for fixed n, as θ_0 is unknown. For such problem, the use of ϕ-divergences, satisfying $\mathcal{D}_{\phi}^{(n)} = \mathbb{R}^n$ (for example χ^2-divergence), might give information about θ_0 and localizes it through ϕ-divergence confidence regions (CR_{ϕ}'s).

The choice of the divergence ϕ also depends upon some knowledge on the support of the unknown p.m. P_0. When P_0 has a projection on $\mathcal{M}$ with same support as P_0, Proposition 3.5 yields its description and its explicit calculation. A necessary condition for this is that $\mathcal{C}_{\theta}$, as defined in (3.41), has non void interior in $\mathbb{R}^{(l+1)}$. Consider the case of the empirical likelihood, that is when $\varphi(x) = -\log x + x - 1$; then $\text{Im } \varphi' =]-\infty, 1[$. Consider $g(x, \theta) = x - \theta$, i.e., a constraint on the mean. Assume that the support of P_0 is unbounded. Then

$$\mathcal{C}_{\theta} = \left\{ t \in \mathbb{R}^2 \text{ such that for all } x \ (P_0 - a.s.) \ , \ t_0 + t_1(x - \theta) \in]-\infty, 1[\right\}.$$

Therefore, $t_1 = 0$ and $\mathcal{C}_\theta =]-\infty, 1[\times\{0\}$ which implies that the interior of $\mathcal{C}_\theta$ is void. This results indicates that the support of Q^* is not the same as the support of P_0. Hence in this case we cannot use the dual representation of $KL_m(\mathcal{M}_\theta, P_0)$. The arguments used in Section 5 for the obtention of limiting distributions cannot be used, if the support of P_0 is unbounded, in order to obtain the limiting distribution of the estimates $\widehat{KL_m}(\mathcal{M}_\theta, P_0)$ under $\mathcal{H}_1$ (i.e., when P_0 does not belong to $\mathcal{M}_\theta$). We thus cannot conclude in this case that the tests pertaining to θ_0 are consistent.

3.8 Robustness and Efficiency of MEϕD estimates and Simulation Results

Lindsay (1994) introduced a general instrument for the study of the asymptotic properties of parametric estimates by minimum ϕ-divergences, called Residual Adjustment Function (RAF). We first recall its definition. Let $\{P_\theta \ : \ \theta \in \Theta\}$ be some parametric model defined on a finite set $\mathcal{X}$. Let $X_1, \ldots, X_n$ a sample with distribution P_{θ_0}. A minimum ϕ-divergence estimate (MϕDE) (called also minimum disparity estimator) of θ_0 is given by

$$\widetilde{\theta}_\phi := \arg\inf_{\theta\in\Theta} \sum_{x\in\mathcal{X}} \varphi\left(\frac{P_\theta(x)}{P_n(x)}\right) P_n(x), \tag{3.78}$$

where $P_n(x)$ is the proportion of the sample point that take value x. When the parametric model $\{P_\theta \ : \ \theta \in \Theta\}$ is regular, then $\widetilde{\theta}_\phi$ is solution of the equation

$$\sum_{x\in\mathcal{X}} \varphi'\left(\frac{P_\theta(x)}{P_n(x)}\right) \dot{P}_\theta(x) = 0, \tag{3.79}$$

which writes as

$$\sum_{x\in\mathcal{X}} A_\varphi(\delta(x))\dot{P}_\theta(x) = 0. \tag{3.80}$$

In this display, $A_\varphi(u) := \varphi'\left(\frac{1}{u+1}\right)$ depends only upon the divergence function φ and

$$\delta(x) := \frac{P_n(x)}{P_\theta(x)} - 1$$

is the "Pearson Residual" at x which belongs to $]-1, +\infty[$. The function $A_\varphi(.)$ is the RAF.

The points x for which $\delta(x)$ is close to -1 are called "inliers", whereas points x such that $\delta(x)$ is large are called "outliers". Efficiency properties are linked with

the behavior of $A_\varphi(.)$ in the neighborhood of 0 (see Lindsay (1994) Proposition 3 and Basu and Lindsay (1994)) : the smaller the value of $\left|A''_\varphi(0)\right|$, the more second efficient the estimate $\widetilde{\theta}_\phi$ in the sense of Rao (1961).

It is easy to verify that the RAF's of the power divergences ϕ_γ, defined by the divergence functions in (3.4), have the form

$$A_\gamma(\delta) = \frac{(\delta+1)^{1-\gamma}-1}{(\gamma-1)}. \tag{3.81}$$

In particular, the Mϕ_γDE of (3.79) with the RAF in (3.81) corresponds to the maximum likelihood when $\gamma = 0$, minimum Hellinger distance when $\gamma = 0.5$, minimum χ^2 divergence when $\gamma = 2$, minimum modified χ^2 divergence when $\gamma = -1$ and minimum KL divergence when $\gamma = 1$.

From (3.81), we see that $A''_\gamma(0) = \gamma$. Hence for the maximum likelihood estimate, we have $\left|A''_\gamma(0)\right| = |A''_0(0)| = 0$ which is the smallest value of $\left|A''_\gamma(0)\right|$, $\gamma \in \mathbb{R}$. Therefore, according to Proposition 3 in Lindsay (1994), the maximum likelihood estimate is the most second-order efficient estimate (in the sense of Rao (1961)) among all minimum power divergences estimates.

Robustness features of $\widetilde{\theta}_\phi$ against inliers and outliers are related to the variations of $A_\varphi(u)$ or $\varphi(x)$ when u or x close to -1 and $+\infty$, respectively as seen through the following heuristic arguments. Let ϕ_1 and ϕ_2 two divergences associated to the functions φ_1 and φ_2. If

$$\lim_{x\downarrow 0} \frac{\varphi_1(x)}{\varphi_2(x)} = +\infty,$$

then the estimating equation (3.79) corresponding to φ_1 in not as stable as that corresponding to φ_2, and hence the MEϕ_2DE is more robust than MEϕ_1DE against outliers. If

$$\lim_{x\uparrow+\infty} \frac{\varphi_1(x)}{\varphi_2(x)} = +\infty,$$

then the estimating equation (3.79) corresponding to φ_1 is not as stable as that corresponding to φ_2, and hence the MEϕ_2DE is more robust than MEϕ_1DE against inliers.

In all cases, the divergence associated to the divergence function having the smallest variations on its domain leads to the most robust estimate against both outliers and inliers. Hence, among all power symmetric divergences ϕ^s_γ (see (1.27)) associated to the divergence functions ψ_γ (see (1.26)), the Hellinger divergence leads to the most robust estimate, since it is associated to the divergence function having

the smallest variations (see Figure 1.2).

It is shown also in Jiménez and Shao (2001) that no minimum power divergence estimate (including the maximum likelihood one) is better than the minimum Hellinger divergence in terms of both second-order efficiency and robustness.

In the examples below, we compare by simulations the efficiency and robustness properties of some MEϕDE's for some models satisfying linear constraints. We will see that the minimum empirical Hellinger divergence estimate represents a suitable compromise between efficiency and robustness. A theoretical study of efficiency and robustness properties of MEϕDE's is necessary and should envolve second-order efficiency versus robustness since all MEϕDE's are all equally first-order efficient (see Remark 3.16 and Theorem 3.5).

Numerical Results

We consider for illustration the same model as in Qin and Lawless (1994) Section 5 Example 1. The model $\mathcal{M}_\theta$ (see 3.12) here is the set of all signed finite measures Q satisfying

$$\int dQ = 1 \quad \text{and} \quad \int g(x,\theta)\, dQ(x) = 0, \tag{3.82}$$

with $g(x,\theta) = ((x-\theta),(x^2-2\theta^2-1))^T$ and θ, the parameter of interest, belongs to $\mathbb{R}$.

In Examples 1.a, 1.b, and 1.c below, we compare the efficiency property of various estimates : we generate 1000 pseudorandom samples of sizes 25, 50, 75 and 100 from a normal distribution with mean θ_0 and variance θ_0^2+1 (i.e., $P_0 = \mathcal{N}(\theta_0, \theta_0^2+1)$) for three values of θ_0 : $\theta_0 = 0$ in Example 1.a, $\theta_0 = 1$ in Example 1.b and $\theta_0 = 2$ in Example 1.c. Note that P_0 satisfies (3.82).

For each sample, we consider various estimates of θ_0 : the sample mean estimate (SME), the parametric ML estimate (MLE) based on the normal distribution $\mathcal{N}(\theta, \theta^2+1)$ and MEϕD estimates $\widehat{\theta}_\phi$ associated to the divergences : $\phi = \chi^2_m$, H, KL, χ^2 and KL_m-divergence (which coincides with the MEL one, i.e., MEKL_mE=MELE).

For all divergence ϕ considered, in order to calculate the MEϕDE $\widehat{\theta}_\phi$, we first calculate $\widehat{\phi}(\mathcal{M}_\theta, P_0)$ for all given θ (using the representation (3.46)) by Newton's method, and then minimize it to obtain $\widehat{\theta}_\phi$.

The results of Theorem 3.5 show that for all ϕ-divergence

$$\sqrt{n}\left(\widehat{\theta}_{\phi}-\theta_0\right)\rightarrow\mathcal{N}(0,V)$$

where V is independent of the divergence ϕ; it is given in Theorem 3.5. For the present model, following Qin and Lawless (1994), V writes

$$V=Var(X)-\triangle^{-1}\left[m'(\theta_0)Var(X)+\theta_0 m(\theta_0)-E(X^3)\right])^2 \tag{3.83}$$

where $\triangle = E\left[m'(\theta_0)(X-\theta_0)+m(\theta_0)-X^2\right]^2$ and $m(\theta) := 2\theta^2+1$. Thus $V \leq Var(X)$ which is the variance of $\sqrt{n}\left(\overline{X}_n-\theta_0\right)$ with $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n}X_i$, the sample mean estimate (SME) of θ_0. So, EMϕD estimates are all asymptotically at least as efficient as $\overline{X}_n$.

3.8.1 Example 1.a

In this example the true value of the parameter is $\theta_0 = 0$.

	MEχ^2_mDE		MEKL_mDE=MELE		MEHDE		MEKLDE	
n	mean	var	mean	var	mean	var	mean	var
25	0.0089	0.0314	0.0086	0.0315	0.0084	0.0315	0.0082	0.0314
50	-0.0116	0.0209	-0.0118	0.0210	-0.0119	0.0210	-0.0120	0.0210
75	-0.0025	0.0171	-0.0024	0.0170	-0.0023	0.0170	-0.0022	0.0169
100	-0.0172	0.0112	-0.0174	0.0111	-0.0174	0.0111	-0.0175	0.0112
	MEχ^2DE		PMLE		SME			
n	mean	var	mean	var	mean	var		
25	0.0077	0.0313	0.0026	0.0318	0.0081	0.0394		
50	-0.0125	0.0212	-0.0063	0.0196	-0.0040	0.0200		
75	-0.0019	0.0167	-0.0011	0.0170	0.0013	0.0164		
100	-0.0177	0.0112	-0.0158	0.0108	-0.0149	0.0102		

TAB. 3.1 – Estimated mean and variance of the estimates of θ_0 in Example 1.a.

We can see from Table 3.1 that all the estimates converge in a satisfactory way. The estimated variances are almost the same for all estimates. This is not surprising since the limit variance of all estimates in this Example (when $\theta_0 = 0$) is close to $V(X)$.

3.8.2 Example 1.b

In this example the true value of the parameter is $\theta_0 = 1$.

	MEχ^2_mDE		MEKL_mDE=MELE		MEHDE		MEKLDE	
n	mean	var	mean	var	mean	var	mean	var
25	0.9394	0.0310	0.9387	0.0312	0.9385	0.0313	0.9378	0.0316
50	0.9994	0.0186	0.9967	0.0186	0.9954	0.0186	0.9941	0.0187
75	1.0009	0.0156	0.9988	0.0154	0.9975	0.0154	0.9966	0.0153
100	0.9984	0.0113	0.9959	0.0112	0.9945	0.0112	0.99315	0.0112
	MEχ^2DE		PMLE		SME			
n	mean	var	mean	var	mean	var		
25	0.9350	0.0322	0.9540	0.0325	1.0033	0.0810		
50	0.9909	0.0190	1.0036	0.0174	1.0021	0.0407		
75	0.9940	0.0152	1.0003	0.0149	0.9912	0.0288		
100	0.9900	0.0113	0.9970	0.0107	0.9851	0.0262		

TAB. 3.2 – Estimated mean and variance of the estimates of θ_0 in Example 1.b.

We can see from Table 3.2 and Figure 3.1 that the estimated bias of EϕDE's are all smaller than the SME one for moderate and large sample sizes. Furthermore, from Figure 3.2, we observe that the estimated variances of EϕDE's are all less than the SME one. They lie between that of the sample mean and that of the parametric maximum likelihood estimate. We observe also that the estimated variances of the MELE and MEHDE are equal and are the smallest among the variances of all MEϕDE's considered. It should be emphasized that even for small sample sizes, the MSE of the SM is larger than any of MEϕDE's.

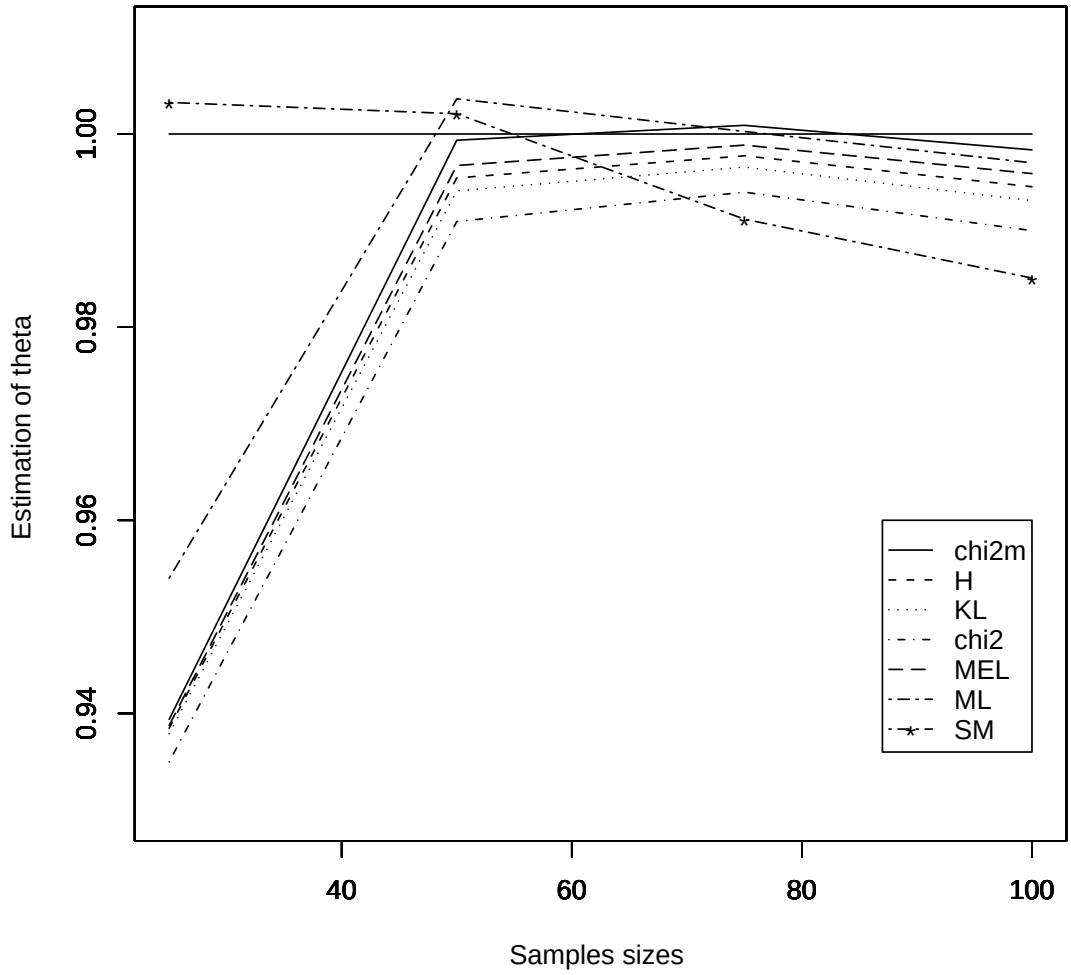

FIG. 3.1 – Estimated mean of the estimates of θ_0 in Example 1.b.

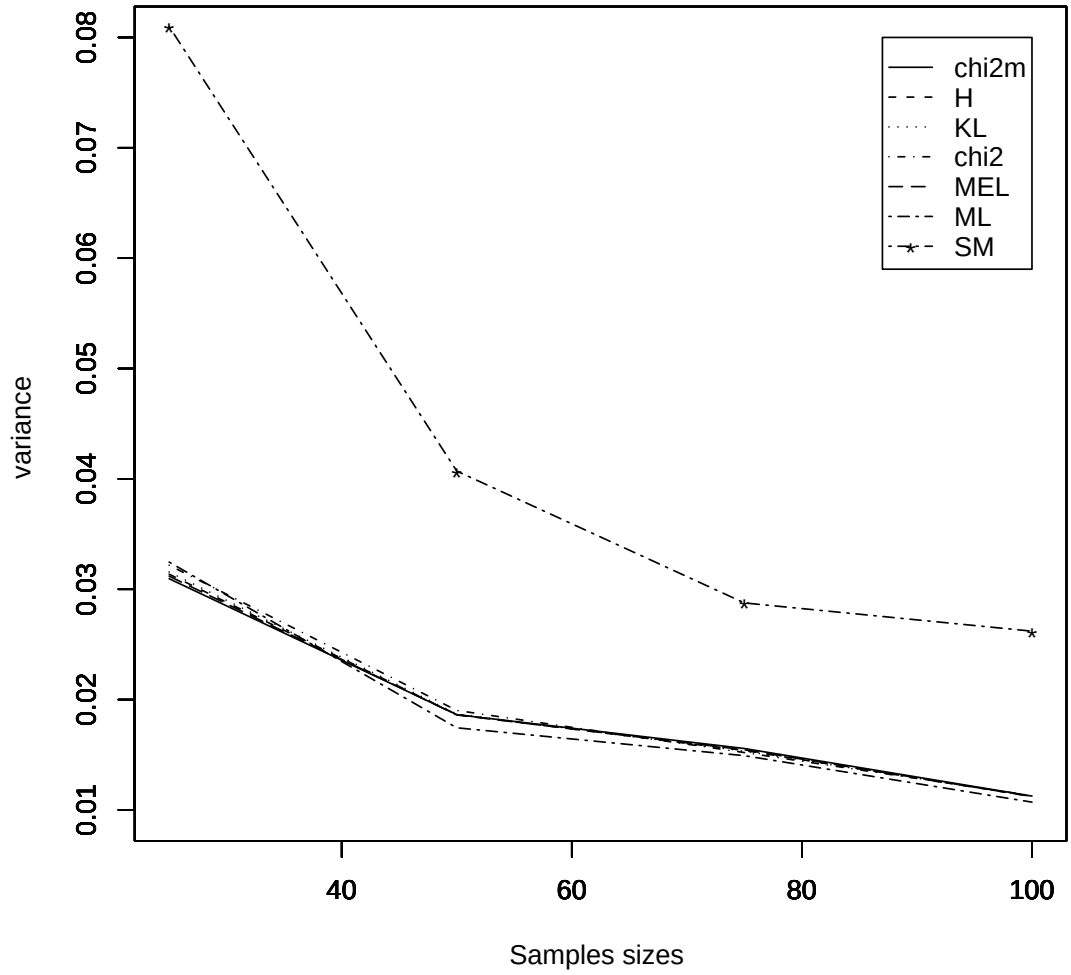

FIG. 3.2 – Estimated variance of the estimates of θ_0 in Example 1.b.

3.8.3 Example 1.c

In this example the true value of the parameter is $\theta_0 = 2$.

In this Example, we can see from Table 3.3 and Figure 3.3 that the estimated bias of EϕDE's are all much smaller than the SME one for moderate and large sample sizes. Furthermore, the estimated variances of MEϕDE's are much smaller than the SME one (see Figure 3.4). The PMLE has the smallest variance. The MEL and MEHD estimates have the smallest variance among all MEϕDE's considered

	MEχ^2_mDE		MEKL_mDE=MELE		MEHDE		MEKLDE	
n	mean	var	mean	var	mean	var	mean	var
25	1.9149	0.0559	1.9092	0.0563	1.9065	0.0567	1.9034	0.0571
50	2.0076	0.0348	2.0012	0.0348	1.9980	0.0349	1.9949	0.0350
75	2.0060	0.0294	2.0015	0.0293	1.9991	0.0292	1.9965	0.0292
100	2.0034	0.0218	1.9982	0.0217	1.9956	0.0217	1.9929	0.0217
	MEχ^2DE		PMLE		SME			
n	mean	var	mean	var	mean	var		
25	1.8953	0.0586	1.9371	0.0591	2.0052	0.2025		
50	1.9876	0.0360	2.0091	0.0326	2.0033	0.1018		
75	1.9911	0.0292	2.0042	0.0286	1.9860	0.0719		
100	1.9873	0.0218	1.9999	0.0204	1.9764	0.0655		

TAB. 3.3 – Estimated mean and variance of the estimates of θ_0 in Example 1.c.

(see Table 3.3). Also in this case, even for small sample sizes, the MSE of the SM is larger than any of MEϕDE's.

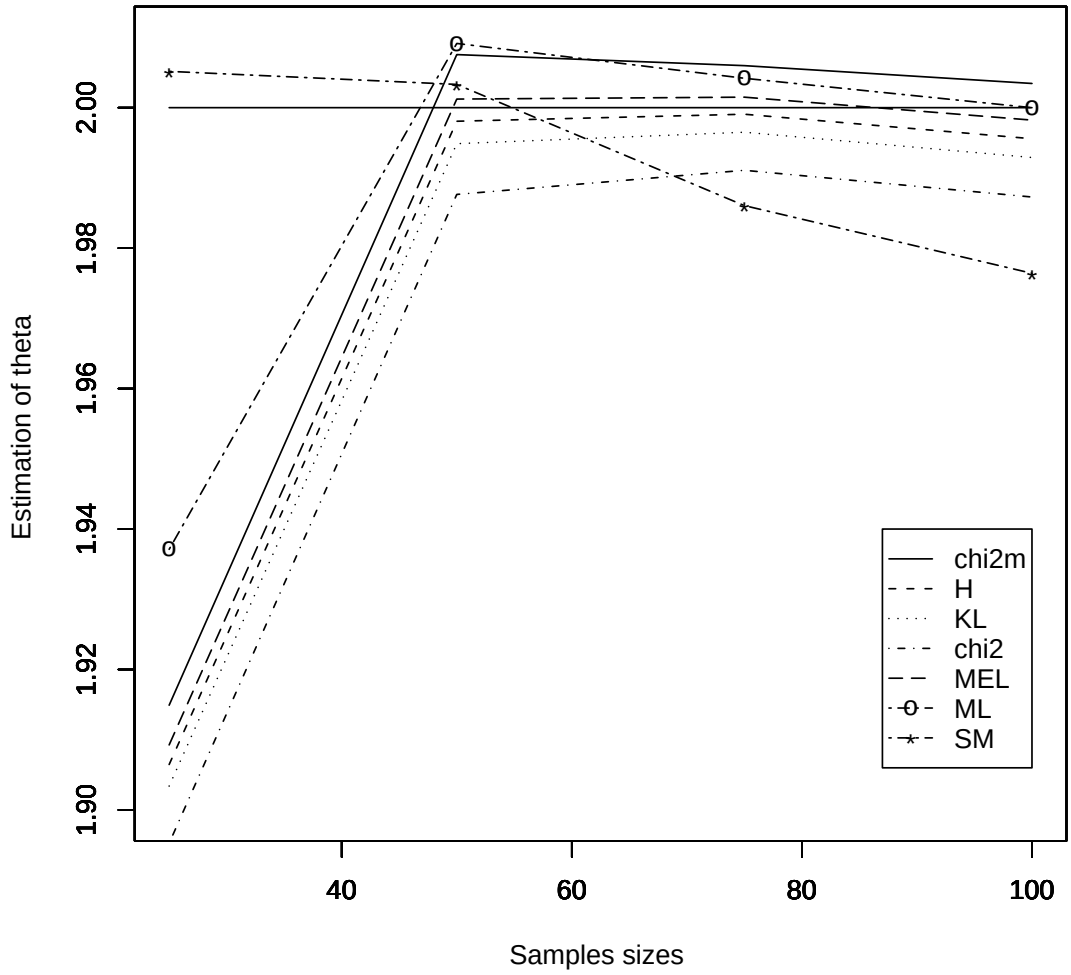

FIG. 3.3 – Estimated mean of the estimates of θ_0 in Example 1.c.

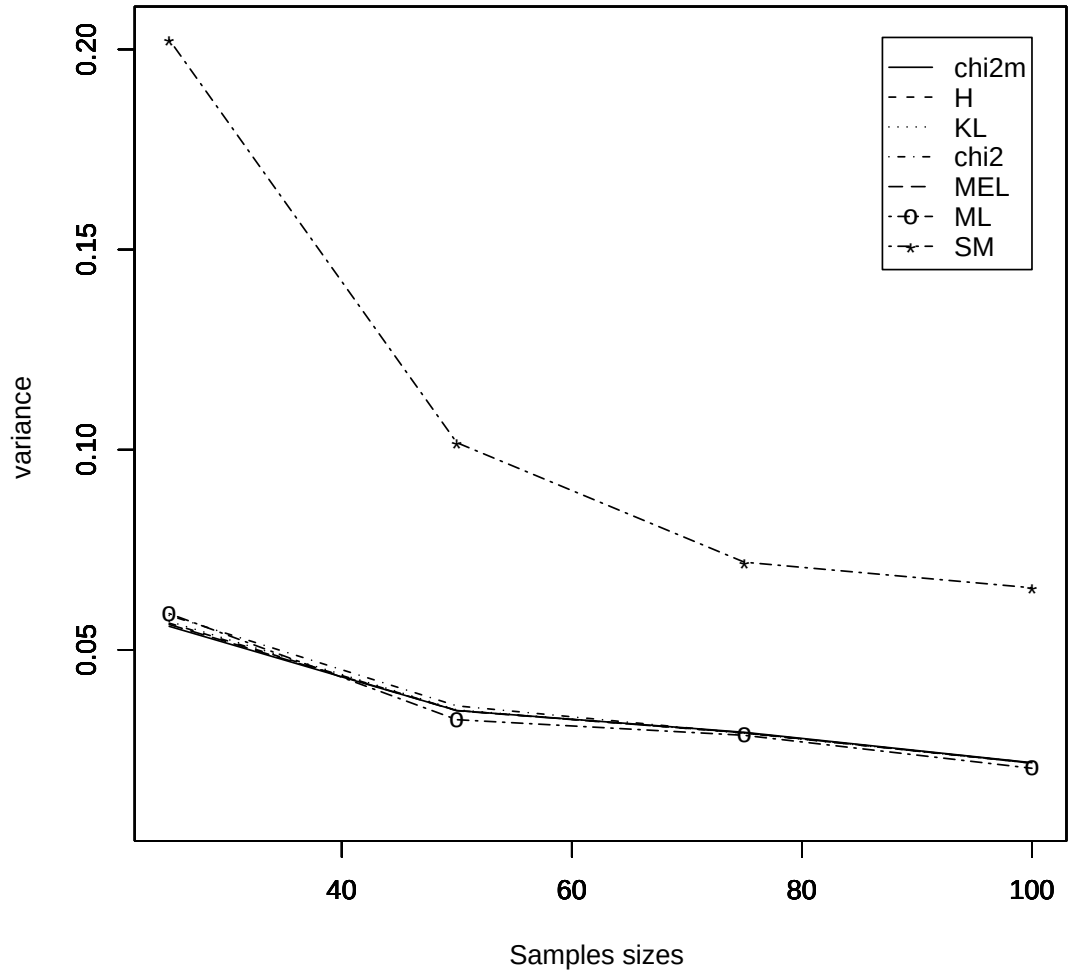

FIG. 3.4 – Estimated variance of the estimates of θ_0 in Example 1.c.

In Examples 2.a and 2.b below, we compare robustness property of the estimates considered above for contaminated data : we consider the same model $\mathcal{M}_\theta$ as in (3.82).

3.8.4 Example 2.a

In this Example, we generate 1000 pseudo-random samples of sizes 25, 50, 75 and 100 from a distribution

$$\widetilde{P_0} = (1-\epsilon)P_0 + \epsilon\delta_5$$

where $P_0 = \mathcal{N}(\theta_0, \theta_0^2 + 1)$, $\epsilon = 0.15$ and $\theta_0 = 2$. We consider the same estimates as in the above examples.

	MEχ^2_mDE		MEKL_mDE=MELE		MEHDE		MEKLDE	
n	mean	var	mean	var	mean	var	mean	var
25	2.1609	0.0654	2.1513	0.0653	2.1453	0.0653	2.1396	0.0652
50	2.2087	0.0303	2.1975	0.0304	2.1912	0.0307	2.1848	0.0309
75	2.2218	0.0214	2.2106	0.0213	2.2046	0.0213	2.1987	0.0215
100	2.2283	0.0151	2.2169	0.0149	2.2110	0.0148	2.2052	0.0149
	MEχ^2DE		PMLE		SME			
n	mean	var	mean	var	mean	var		
25	2.1278	0.0646	2.2088	0.0581	2.4265	0.2178		
50	2.1729	0.0316	2.2296	0.0280	2.4535	0.1076		
75	2.1877	0.0219	2.2337	0.0197	2.4545	0.0721		
100	2.1947	0.0151	2.2352	0.0139	2.4572	0.0543		

TAB. 3.4 – Estimated mean and variance of the estimates of θ_0 in Example 2.a.

In this Example, we can see from Table 3.4 and Figure 3.5 that the MEχ^2D estimate is the most robust and MEχ^2_m estimate is the least robust. We observe also that the MELE which is the MEKL_mDE is less robust than the MEKLDE and that the MEHD estimate is more robust than MEL one.

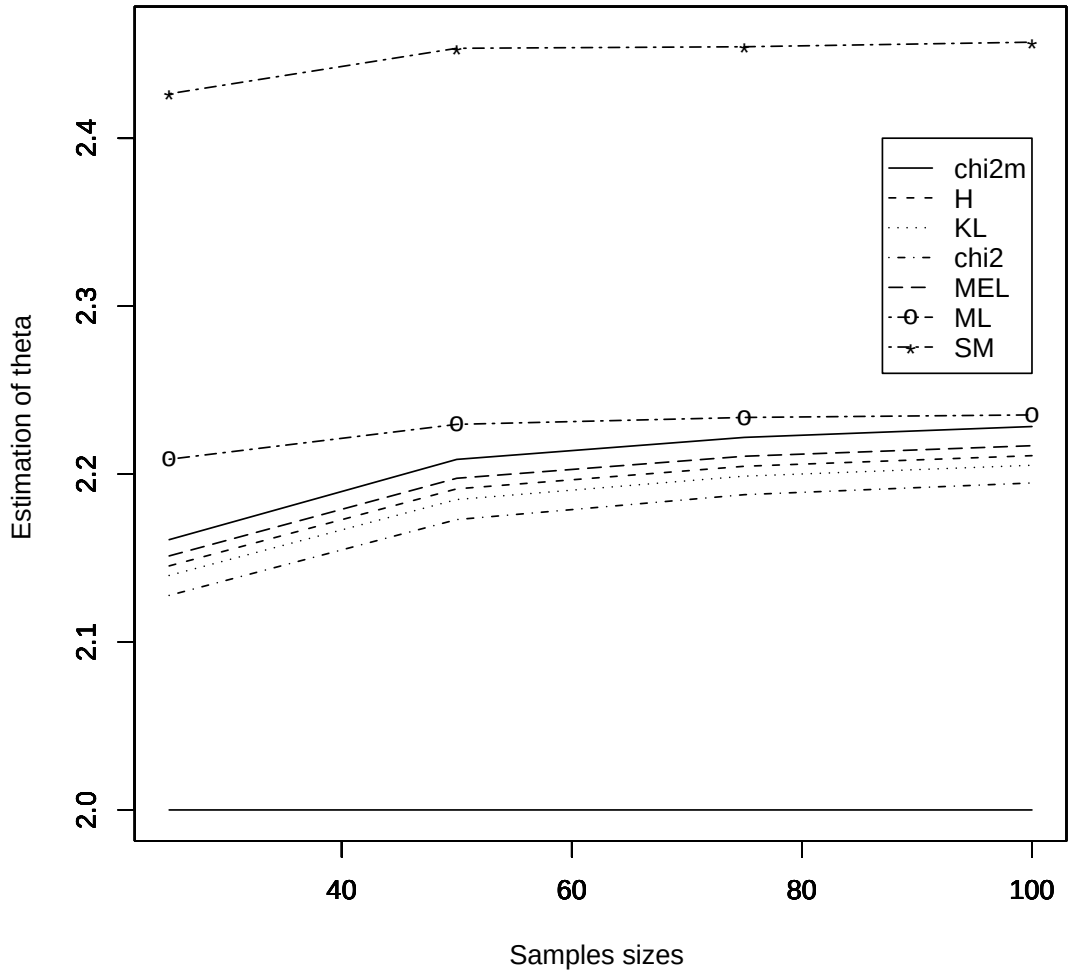

FIG. 3.5 – Estimated mean of the estimates of θ_0 in Example 2.a.

3.8.5 Example 2.b

In this Example, we generate 1000 pseudo-random samples of sizes 50, 100, 150 and 200 from a distribution $P_0 = \mathcal{N}(\theta_0, \theta_0^2 + 1)$ with $\theta_0 = 2$ and we cancel the observations in the interval $[4, 5]$. We consider the same estimates as in the above examples.

In this example, in contrast with Example 2.b, we observe that the MEχ^2_mDE

	MEχ^2_mDE		MEKL_mDE=MELE		MEHDE		MEKLDE	
n	mean	var	mean	var	mean	var	mean	var
50	1.9917	0.0451	1.9784	0.0431	1.9721	0.0426	1.9659	0.0423
100	1.9962	0.0362	1.9844	0.0346	1.9787	0.0341	1.9729	0.0336
150	2.0011	0.0150	1.9903	0.0142	1.9849	0.0139	1.9795	0.0137
200	1.9602	0.0162	1.9516	0.0158	1.9473	0.0157	1.9430	0.0156
	MEχ^2DE		PMLE		SME			
n	mean	var	mean	var	mean	var		
50	1.9522	0.0428	1.9705	0.0358	1.7750	0.1039		
100	1.9590	0.0329	1.9687	0.0298	1.7365	0.0576		
150	1.9671	0.0135	1.9781	0.0121	1.7456	0.0283		
200	1.9325	0.0155	1.9420	0.0146	1.7247	0.0317		

TAB. 3.5 – Estimated mean and variance of the estimates of θ_0 in Example 2.b.

is the most robust, MEχ^2DE is the least robust and MEKLDE is less robust than MEKL_mDE (=MELE).

Generally, if a MEϕDE is more robust than its adjoint[2] (i.e., ME$\phi^\sim$DE) against "outliers", then it is less robust then its adjoint against "inliers" (see Examples 2.a and 2.b). The Hellinger divergence has not this disadvantage since it is self-adjoint (i.e., $H = H^\sim$).

[2]For all divergence ϕ associated to a convex function φ, its adjoint, noted $\phi^\sim$, is by definition the divergence associated to the convex function, noted $\varphi^\sim$ defined by : $\varphi^\sim(x) = x\varphi(1/x)$, for all x.

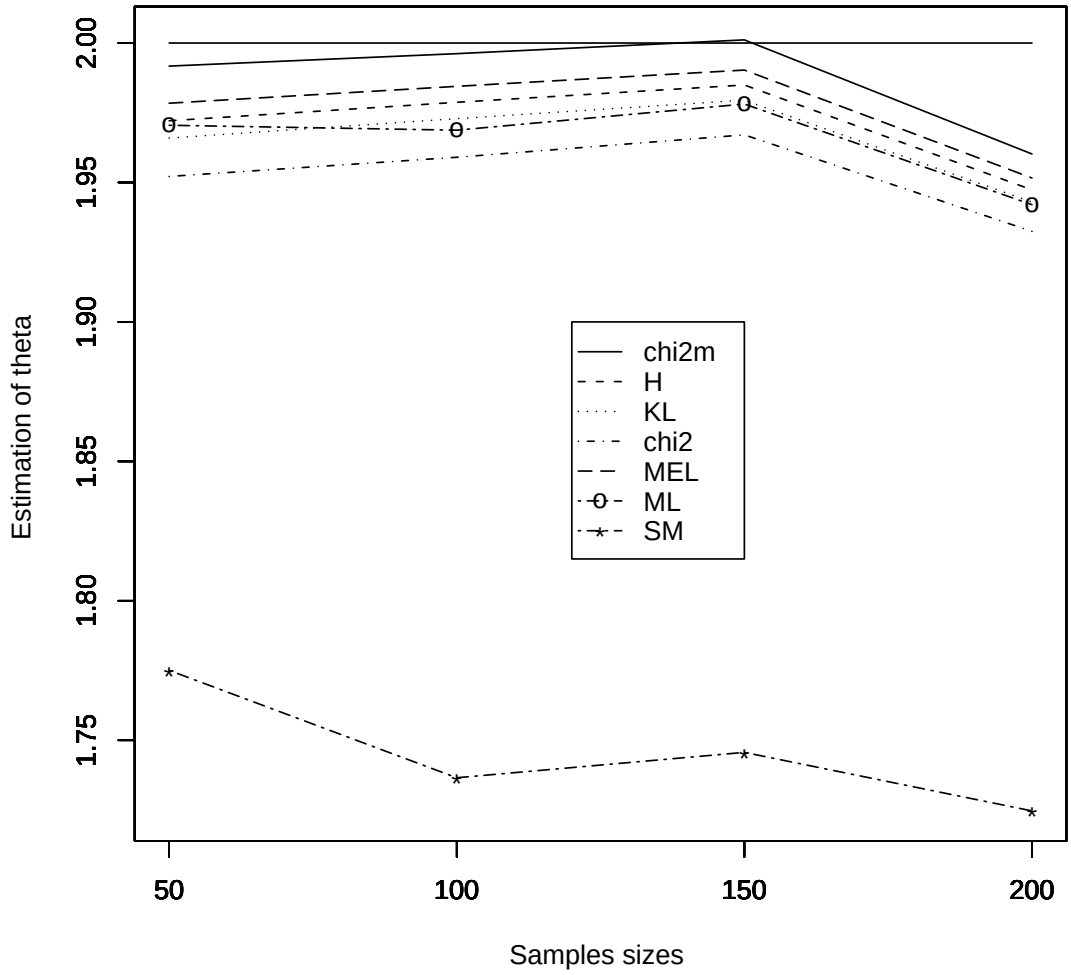

FIG. 3.6 – Estimated mean of the estimates of θ_0 in Example 2.b.

3.9 Proofs

3.9.1 Proof of Theorem 3.1

Denote $m := \phi(\Omega, P)$ and let δ be a positive number. Define the sets

$$\Omega(\delta) := \{Q \in \Omega \text{ such that } \phi(Q, P) \leq m + \delta\}$$

and

$$\Lambda(\delta) := \left\{ q := \frac{dQ}{dP} \text{ such that } Q \in \Omega(\delta) \right\}.$$

The set $\Lambda(\delta)$ is equi-integrable. Indeed, by the assumptions in the Theorem, for all $\epsilon > 0$, there exists $c > 0$, such that for all Q and all x, it holds

$$|q(x)| > c \;\Rightarrow\; \frac{|q(x)|}{\varphi\left(q(x)\right)} < \epsilon \text{ and } \frac{|q(x)|}{\varphi(q(x))^{1/r}} < \epsilon. \tag{3.84}$$

Hence,

$$\begin{aligned} \int |q(x)|\, \mathbb{1}_{\{|q(x)|>c\}}\, dP &= \int \frac{|q(x)|}{\varphi\left(q(x)\right)} \varphi\left(q(x)\right) \mathbb{1}_{\{|q(x)|>c\}}\, dP \\ &\leq \epsilon \int \varphi\left(q(x)\right)\, dP \\ &\leq \epsilon(m+\delta). \end{aligned}$$

This implies that

$$\lim_{c\to+\infty} \sup_{Q\in\Lambda(\delta)} \int |q(x)|\, \mathbb{1}_{\{|q(x)|>c\}}\, dP \leq \epsilon(m+\delta), \quad \text{for all } \epsilon > 0.$$

Hence,

$$\lim_{c\to+\infty} \sup_{Q\in\Lambda(\delta)} \int |q(x)|\, \mathbb{1}_{\{|q(x)|>c\}}\, dP = 0,$$

which is to say that $\Lambda(\delta)$ is equi-integrable. Hence, it is weakly sequentially compact in $L^1(P)$ (see Meyer (1966) p. 39). Consider a sequence Q_n in $\Lambda(\delta)$ such that

$$\lim_{n\to+\infty} \phi(Q_n, P) = \phi(\Omega, P).$$

The sequence $q_n := dQ_n/dP$ belongs to $\Lambda(\delta)$. Therefore, there exists a subsequence $(q_{n_k})_k$ which converges weakly to some function, say q^*, in $L^1(P)$, i.e., the sequence of measures Q_{n_k} converges to Q^* in τ-topology with Q^* defined by $dQ^*/dP = q^*$. We prove now that $Q^* \in M_{\mathcal{F}}$. For all f in $\mathcal{F}$, it holds

$$\begin{aligned} \int |f|\, d|Q^*| &= \int |f||q^*|\, dP \\ &= \int |f||q^*| \mathbb{1}_{\{|q^*|\leq c\}}\, dP + \int |f||q^*| \mathbb{1}_{\{|q^*|>c\}}\, dP \\ &\leq c \int |f|\, dP + \int |f| \frac{|q^*|}{\varphi(q^*)^{1/r}} \varphi(q^*)^{1/r} \mathbb{1}_{\{|q^*|>c\}}\, dP \\ &\leq c \int |f|\, dP + \epsilon \left(\int |f|^k\, dP \right)^{1/k} \left(\int \varphi(q^*)\, dP \right)^{1/r} \\ &= c \int |f|\, dP + \epsilon \left(\int |f|^k\, dP \right)^{1/k} (\phi(Q^*, P))^{1/r} \end{aligned} \tag{3.85}$$

Furthermore, the mapping $Q \in (M, \tau) \to \phi(Q, P)$ is l.s.c.[3] and since Q_{n_k} converges to Q^* in τ-topology, this implies that

$$\phi(Q^*, P) \leq \lim_{k \to +\infty} \phi(Q_{n_k}, P) = \phi(\Omega, P) < \infty.$$

Hence, from (3.85), we deduce $\int |f| \ d|Q^*| < \infty$. We still have to prove that Q^* belongs to Ω. Since Ω is, by assumption, a closed set in $(M_{\mathcal{F}}, \tau_{\mathcal{F}})$, it is enough to show that the sequence $(Q_{n_k})_k$ converges to Q^* in $(M_{\mathcal{F}}, \tau_{\mathcal{F}})$, i.e., $\int f \ dQ_{n_k}$ converges to $\int f \ dQ^*$ for all f in $\mathcal{F}$. So, let f in $\mathcal{F}$, we have

$$\int f \ dQ_{n_k} = \int f \mathbb{1}_{\{|f| \leq b\}} \ dQ_{n_k} + \int f \mathbb{1}_{\{|f| > b\}} \ dQ_{n_k} =: A + B.$$

$$\begin{aligned}
|B| &= \left| \int f \mathbb{1}_{\{|f|>b\}} \ dQ_{n_k} \right| \leq \int |f| \mathbb{1}_{\{|f|>b\}} \ d|Q_{n_k}| \\
&= \int |f| \mathbb{1}_{\{|f|>b\}} |q_{n_k}| \ dP \\
&= \int |f| \mathbb{1}_{\{|f|>b\}} |q_{n_k}| \mathbb{1}_{\left\{|q_{n_k}| \leq c\right\}} \ dP + \int |f| \mathbb{1}_{\{|f|>b\}} |q_{n_k}| \mathbb{1}_{\left\{|q_{n_k}| > c\right\}} \ dP \\
&\leq c \int |f| \mathbb{1}_{\{|f|>b\}} \ dP + \int |f| \mathbb{1}_{\{|f|>b\}} \frac{|q_{n_k}|}{\varphi(q_{n_k})^{1/r}} \varphi(q_{n_k})^{1/r} \mathbb{1}_{\left\{|q_{n_k}| > c\right\}} \ dP \\
&\leq c \int |f| \mathbb{1}_{\{|f|>b\}} \ dP + \epsilon \int |f| \mathbb{1}_{\{|f|>b\}} \varphi(q_{n_k})^{1/r} \ dP \\
&\leq c \int |f| \mathbb{1}_{\{|f|>b\}} \ dP + \epsilon \left(\int |f|^k \mathbb{1}_{\{|f|>b\}} \ dP \right)^{1/k} \left(\int \varphi(q_{n_k}) \ dP \right)^{1/r}.
\end{aligned}$$

We deduce

$$(B1) \leq \int f \ dQ_{n_k} \leq (B2), \tag{3.86}$$

with

$$\begin{aligned}
(B1) \ := \ & \int f \mathbb{1}_{\{|f| \leq b\}} \ dQ_{n_k} - c \int |f| \mathbb{1}_{\{|f|>b\}} \ dP - \\
& \epsilon \left(\int |f|^k \mathbb{1}_{\{|f|>b\}} \ dP \right)^{1/k} \left(\int \varphi(q_{n_k}) \ dP \right)^{1/r}.
\end{aligned}$$

and

$$\begin{aligned}
(B2) \ := \ & \int f \mathbb{1}_{\{|f| \leq b\}} \ dQ_{n_k} + c \int |f| \mathbb{1}_{\{|f|>b\}} \ dP + \\
& \epsilon \left(\int |f|^k \mathbb{1}_{\{|f|>b\}} \ dP \right)^{1/k} \left(\int \varphi(q_{n_k}) \ dP \right)^{1/r}.
\end{aligned}$$

[3] this holds from Proposition 3.1 choosing the class of functions $\mathcal{F} = \mathcal{B}_b$, the class of all bounded measurable real valued functions.

The families of functions $\left\{f_b := |f|\mathbb{1}_{\{|f|>b\}},\ \ b \geq 0\right\}$ and $\left\{f_b^k := |f|^k\mathbb{1}_{\{|f|>b\}},\ \ b \geq 0\right\}$ are dominated respectively by $|f|$ and $|f|^k$. Moreover $\int |f|\ dP$ and $\int |f|\ dP$ are finite by assumption. We thus get by the Dominated Monotone Convergence Theorem

$$\lim_{b\to+\infty}\int |f|\mathbb{1}_{\{|f|>b\}}\ dP = \lim_{b\to+\infty}\int |f|^k\mathbb{1}_{\{|f|>b\}}\ dP = 0.$$

Hence, from (3.86), we get

$$\int f\ dQ^* = \lim_{b\to+\infty}\lim_{k\to+\infty}(B1) \leq \lim_{k\to+\infty}\int f\ dQ_{n_k} \leq \lim_{b\to+\infty}\lim_{k\to+\infty}(B1) = \int f\ dQ^*,$$

which is to say that the sequence $(Q_{n_k})_k$ converges to Q^* in $\tau_{\mathcal{F}}$-topology. Hence, Q^* belongs to Ω which is $\tau_{\mathcal{F}}$- closed. Since $\phi(Q^*, P) \leq \phi(\Omega, P)$, we get $\phi(Q^*, P) = \phi(\Omega, P)$, i.e., the projection of P on Ω exists; it is Q^*, the limit of the subsequence $(Q_{n_k})_k$. This ends the proof of Theorem 3.1.

3.9.2 Proof of Lemma 3.1

1- Under (C.0), for all P in M^1 and all Q in M such that $\phi(Q, P) < \infty$, we have

$$\varphi\left(c\frac{dQ}{dP}\right) \leq c_1\varphi\left(\frac{dQ}{dP}\right) + c_2\left|\frac{dQ}{dP}\right| + c_3.$$

Integrating with respect to P yields

$$\int \varphi\left(c\frac{dQ}{dP}\right)\ dP \leq c_1\phi(Q, P) + c_2\int\left|\frac{dQ}{dP}\right|\ dP + c_3 < \infty.$$

2- For all c in $[1-\delta, 1+\delta]$, define the functions

$$\begin{array}{rcccccl} l_c & : & x \in \mathbb{R} & \mapsto & l_c(x) & := & \varphi(cx)\mathbb{1}_{]-\infty,0[}(cx), \\ g_c & : & x \in \mathbb{R} & \mapsto & g_c(x) & := & \varphi(cx)\mathbb{1}_{[0,1]}(cx), \\ h_c & : & x \in \mathbb{R} & \mapsto & h_c(x) & := & \varphi(cx)\mathbb{1}_{]1,+\infty[}(cx). \end{array}$$

For any c and x, we have $\varphi(cx) = l_c(x) + g_c(x) + h_c(x)$. For all real x, the functions $c \to l_c$ and $c \to h_c$ are increasing, and the function $c \to g_c$ is decreasing. Denote $q := \frac{dQ}{dP}$. Apply the Monotone Convergence Theorem to get

$$\lim_{c\uparrow 1}\int l_c(q)\ dP = \int l_1(q)\ dP \ \text{ and } \ \lim_{c\uparrow 1}\int h_c(q)\ dP = \int h_1(q)\ dP.$$

On the other hand, the class of functions $\{x \to g_c(x),\ c \text{ in } [1-\delta, 1+\delta]\}$ is dominated by the function $x \to g_{1-\delta}(x)$. Furthermore, for all Q in M, $g_{1-\delta}(q)$ belongs to $L^1(P)$ by the condition (C.0). Hence, apply the dominated convergence Theorem to get

$$\lim_{c\uparrow 1}\int g_c(q)\ dP = \int g_1(q)\ dP.$$

Those three limits prove the first part of the claim. The same argument closes the proof of the Lemma.

Proof of Lemma 3.2

Using the convexity of the function φ, we have for all $\epsilon > 0$

$$\frac{\varphi(q) - \varphi\left((1-\epsilon)q\right)}{\epsilon} \leq q\varphi'(q) \leq \frac{\varphi\left((1+\epsilon)q\right) - \varphi(q)}{\epsilon}.$$

By Lemma 3.1, for all ϵ satisfying $0 < \epsilon < \delta$, both the LHS and the RHS terms belong to $L^1(P)$, and hence $\varphi'(q)q \in L^1(P)$.

3.9.3 Proof of Theorem 3.2

Convexity of φ implies, for all positive ϵ,

$$\varphi'(q^*)(q - q^*) \leq \frac{\varphi\left((1-\epsilon)q^* + \epsilon q\right) - \varphi(q^*)}{\epsilon} \leq \varphi(q) - \varphi(q^*). \tag{3.87}$$

The middle term in the above display, by the convexity of φ, decreases to $\varphi'(q^*)(q - q^*)$ when $\epsilon \downarrow 0$. Furthermore, it is dominated by $\varphi(q) - \varphi(q^*)$ which belongs to $L^1(P)$ for all Q in $\mathcal{D}_\phi$. Hence, apply the Monotone Convergence Theorem to get

$$\int \varphi'(q^*)(q - q^*)\, dP = \lim_{\epsilon \downarrow 0} \int \frac{\varphi\left((1-\epsilon)q^* + \epsilon q\right) - \varphi(q^*)}{\epsilon}\, dP, \quad \text{for all } \; Q \in \mathcal{D}_\phi. \tag{3.88}$$

Proof of part 1. Integrating (3.87) with respect to P and using (i) and (ii) in part 1 of the Theorem, we obtain for all Q in $\Omega \cap \mathcal{D}_\phi$

$$\phi(Q, P) - \phi(Q^*, P) \geq \int \varphi'(q^*)(q - q^*)\, dP = \int \varphi'(q^*)\, dQ - \int \varphi'(q^*)\, dQ^* \geq 0. \tag{3.89}$$

Hence, P has projection Q^* on Ω.

Proof of part 2. Convexity of both Ω and $\mathcal{D}_\phi$, implies that for all $Q \in \Omega \cap \mathcal{D}_\phi$, $(1-\epsilon)Q + \epsilon Q^*$ belongs to $\Omega \cap \mathcal{D}_\phi$. Since Q^* is the projection of P on Ω, for all $Q \in \Omega \cap \mathcal{D}_\phi$ and all ϵ satisfying $0 < \epsilon < 1$, we get

$$\phi\left((1-\epsilon)Q + \epsilon Q^*, P\right) - \phi(Q^*, P) \geq 0.$$

Combining this with (3.88) and using the fact that Q^* is the projection of P on Ω, we obtain for all Q in $\Omega \cap \mathcal{D}_\phi$

$$\begin{aligned} \int \varphi'(q^*)(q - q^*)\, dP &= \lim_{\epsilon \downarrow 0} \int \frac{\varphi\left((1-\epsilon)q^* + \epsilon q\right) - \varphi(q^*)}{\epsilon}\, dP \\ &= \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon}\left[\phi\left((1-\epsilon)Q^* + \epsilon Q, P\right) - \phi\left(Q^*, P\right)\right] \geq 0. \end{aligned}$$

On the other hand, integrating (3.87) with respect to P, we obtain for all Q in $\Omega \cap \mathcal{D}_\phi$

$$\int \varphi'(q^*)(q - q^*)\, dP \leq \phi(Q, P) - \phi(Q^*, P) < \infty. \tag{3.90}$$

Hence, (3.90) and (3.90) imply

$$\varphi'(q^*)(q - q^*) \in L^1(P), \quad \text{for all } \ Q \in \Omega \cap \mathcal{D}_\phi. \tag{3.91}$$

By Lemma 3.2, $\varphi'(q^*)q^* \in L^1(P)$. Combining this with (3.91), we obtain that

$$\text{for all } \ Q \in \Omega \cap \mathcal{D}_\phi, \ \text{ we have } \ \varphi'(q^*)q \in L^1(P) \ \text{ and } \ \int \varphi'(q^*)\, dQ^* \leq \int \varphi'(q^*)\, dQ.$$

This ends the proof of Theorem 3.2.

3.9.4 Proof of Theorem 3.3

Proof of part 1. Since Ω is convex, this is a consequence of Theorem 3.2.

Proof of part 2. When $\varphi'(q^*)$ belongs to $\langle \mathcal{G} \rangle$, then for all Q in Ω, $\int \varphi'(q^*)\, dQ^* = \int \varphi'(q^*)\, dQ$ which, by the first part of the present Theorem, proves that P has projection Q^* on Ω. This projection is unique by convexity of Ω.

Proof of part 3. Since Q^* is a signed finite measure, by the Hahn decomposition Theorem, there exists a partition $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ such that $\mathcal{X}_1, \mathcal{X}_2 \in \mathcal{B}$ and satisfying

for all $B \in \mathcal{B}$, such that $B \subset \mathcal{X}_1$ we have $Q^*(B) \geq 0$,

and

for all $B \in \mathcal{B}$, such that $B \subset \mathcal{X}_2$ we have $Q^*(B) \leq 0$.

Denote by Q^*_+ and Q^*_- respectively the nonnegative variation and the nonpositive variation of Q^* which are defined, for all $B \in \mathcal{B}$, by

$$Q^*_+(B) := Q^*(B \cap \mathcal{X}_1) \ \text{ and } \ Q^*_-(B) := -Q^*(B \cap \mathcal{X}_2).$$

So, Q^*_+ and Q^*_- are nonnegative finite measures, $Q = Q^*_+ - Q^*_-$ and the total variation $|Q^*|$ is, by definition, the nonnegative measure $Q^*_+ + Q^*_-$. Denote $\langle \mathcal{G} \rangle^\perp_+$ and $\langle \mathcal{G} \rangle^\perp_-$ respectively the orthogonal of $\langle \mathcal{G} \rangle$ in $L^1(Q^*_+)$ and in $L^1(Q^*_-)$, i.e., the sets defined by

$$\langle \mathcal{G} \rangle^\perp_+ := \left\{ h \in L^\infty(Q^*_+) \ \text{ such that } \ \int fh\, dQ^*_+ = 0, \ \text{ for all } \ f \in \langle \mathcal{G} \rangle \right\}$$

and

$$\langle \mathcal{G} \rangle^\perp_- := \left\{ h \in L^\infty(Q^*_-) \ \text{ such that } \ \int fh\, dQ^*_- = 0, \ \text{ for all } \ f \in \langle \mathcal{G} \rangle \right\}.$$

We will prove that the two following assertions hold

$$\text{for all } \; h \in \langle \mathcal{G} \rangle_+^{\perp}, \;\; \text{we have } \; \int \varphi'(q^*) h \; dQ_+^* = 0, \tag{3.92}$$

and

$$\text{for all } \; h \in \langle \mathcal{G} \rangle_-^{\perp}, \;\; \text{we have } \; \int \varphi'(q^*) h \; dQ_-^* = 0. \tag{3.93}$$

Proof of (3.92). We prove (3.92) by contradiction : Assume that there exists h in $\langle \mathcal{G} \rangle_+^{\perp}$ such that $\int \varphi'(q^*) h \; dQ_+^* \neq 0$. We then have either (a) $\int \varphi'(q^*) h \; dQ_+^* < 0$ or (b) $\int \varphi'(q^*) h \; dQ_+^* > 0$.

Assume (a). For $0 < \epsilon < \delta$,[4] define the measure Q_0 by

$$dQ_0 := \left(1 + \epsilon \frac{h \mathbb{1}_{\mathcal{X}_1}}{|h|_\infty} \right) dQ^*.$$

Then Q_0 belongs to Ω, and, following condition (C.0), Q_0 belongs to $\mathcal{D}_\phi$ by Lemma 3.1. Furthermore,

$$\int \varphi'(q^*) \; dQ_0 = \int \varphi'(q^*) \; dQ^* + \epsilon \frac{1}{|h|_\infty} \int \varphi'(q^*) h \; dQ_+^* < \int \varphi'(q^*) \; dQ^*,$$

which contradicts the fact that Q^* is the projection of P on Ω (see part 2 in Theorem 3.2).

Assume (b). Consider $-h$ instead of h.

We thus have proved (3.92). The same arguments hold for the proof of (3.93). Therefore, $\varphi'(q^*)$ belongs to $\left(\langle \mathcal{G} \rangle_+^{\perp} \right)_+^{\perp}$ and to $\left(\langle \mathcal{G} \rangle_-^{\perp} \right)_-^{\perp}$ respectively the orthogonal of $\langle \mathcal{G} \rangle_+^{\perp}$ in $L^1(Q_+^*)$ and the orthogonal of $\langle \mathcal{G} \rangle_-^{\perp}$ in $L^1(Q_-^*)$. By Hahn-Banach Theorem (see e.g. Brezis (1983) Section 2), we have

$$\left(\langle \mathcal{G} \rangle_+^{\perp} \right)_+^{\perp} = \overline{\langle \mathcal{G} \rangle}_+ \;\; \text{and} \;\; \left(\langle \mathcal{G} \rangle_-^{\perp} \right)_-^{\perp} = \overline{\langle \mathcal{G} \rangle}_-$$

which are respectively the closure of $\langle \mathcal{G} \rangle$ in $L^1(Q_+^*)$ and the closure of $\langle \mathcal{G} \rangle$ in $L^1(Q_-^*)$. This implies that $\varphi'(q^*)$ is in $\overline{\langle \mathcal{G} \rangle}$ the closure of $\langle \mathcal{G} \rangle$ in $L^1(|Q^*|)$. This concludes the proof of Theorem 3.3.

3.9.5 Proof of Proposition 3.4

Proof of part (i). The function

$$(Q(X_1), \ldots, Q(X_n))^T \in \mathbb{R}^n \mapsto \frac{1}{n} \sum_{i=1}^{n} \varphi \left(n Q(X_i) \right)$$

[4] here δ is defined in the condition (C.0).

is continuous and nonnegative on $\mathcal{D}_{\phi}^{(n)}$. Furthermore, the set $\mathcal{M}_{\theta}^{(n)}$ is closed in $\mathbb{R}^n$. Hence, by condition (3.35), the infimum of the function

$$(Q(X_1),\ldots,Q(X_n))^T \in \mathbb{R}^n \mapsto \frac{1}{n}\sum_{i=1}^{n}\varphi\left(nQ(X_i)\right)$$

on the set $\mathcal{D}_{\phi}^{(n)} \cap \mathcal{M}_{\theta}^{(n)}$ exists as an interior point of $\mathcal{D}_{\phi}^{(n)}$. Since the above function is strictly convex and the set $\mathcal{D}_{\phi}^{(n)} \cap \mathcal{M}_{\theta}^{(n)}$ is convex, then this infimum is unique. It is noted $\widehat{Q_{\theta}^{*}}$. This concludes the proof of part (i).
Proof of (ii). Since $(Q(X_1),\ldots,Q(X_n))^T \in \mathbb{R}^n \mapsto \frac{1}{n}\sum_{i=1}^{n}\varphi\left(nQ(X_i)\right)$ is $\mathcal{C}^1$ on the interior of $\mathcal{D}_{\phi}^{(n)}$, and since $\widehat{Q_{\theta}^{*}}$ is in the interior of $\mathcal{D}_{\phi}^{(n)}$, we can use the Lagrange method. This yields the explicit form (3.37) of the projection $\widehat{Q_{\theta}^{*}}$ in which $\widehat{c}_0$ is the Lagrange multiplier associated to the constraint $\sum_{i=1}^{n} Q(X_i) = 1$ and $\widehat{c}_j$ to the constraint $\sum_{i=1}^{l} Q(X_i)g_j(X_i,\theta) = 0$, for all $j = 1,\ldots,l$. This concludes the proof of Proposition 3.4.

3.9.6 Proof of Proposition 3.6

Define the estimates

$$\widetilde{c_{\theta}} = \arg\inf_{t\in T_{\theta}} P_n m(\theta,t) \quad \text{and} \quad \widetilde{\phi}\left(\mathcal{M}_{\theta},P_0\right) = \sup_{t\in T_{\theta}} P_n m(\theta,t).$$

By condition (C.2), for all n sufficiently large, we have

$$\widehat{c_{\theta}} = \widetilde{c_{\theta}} \quad \text{and} \quad \widehat{\phi}\left(\mathcal{M}_{\theta},P_0\right) = \widetilde{\phi}\left(\mathcal{M}_{\theta},P_0\right).$$

We prove that $\widetilde{\phi}\left(\mathcal{M}_{\theta},P_0\right)$ and $\widetilde{c_{\theta}}$ converge to $\phi\left(\mathcal{M}_{\theta},P_0\right)$ and c_{θ} respectively. Since c_{θ} is isolated, then consistency of $\widetilde{c_{\theta}}$ holds as a consequence of Theorem 5.7 in van der Vaart (1998). For the estimate $\widetilde{\phi}\left(\mathcal{M}_{\theta},P_0\right)$, we have

$$\left|\widetilde{\phi}\left(\mathcal{M}_{\theta},P_0\right) - \phi\left(\mathcal{M}_{\theta},P_0\right)\right| = \left|P_n m(\theta,\widetilde{c_{\theta}}) - P_0 m(\theta,c_{\theta})\right| := |A|,$$

which implies

$$P_n m(\theta,c_{\theta}) - P_0 m(\theta,c_{\theta}) < A < P_n m(\theta,\widetilde{c_{\theta}}) - P_0 m(\theta,\widetilde{c_{\theta}}).$$

Both the RHS and the LHS terms in the above display go to 0, under condition (C.1). This implies that A tends to 0, which concludes the proof of Proposition 3.6.

3.9.7 Proof of Theorem 3.4

.

Proof of (1). Some calculus yield

$$P_0 m'(\theta, c_\theta) = P_0 \left(1 - \overleftarrow{\varphi'}\left(c_\theta^T g(\theta)\right), -g_1(\theta)\overleftarrow{\varphi'}\left(c_\theta^T g(\theta)\right), \ldots, -g_l(\theta)\overleftarrow{\varphi'}\left(c_\theta^T g(\theta)\right)\right)^T (\underline{g}^T) \quad (3.94)$$

and

$$P_0 m''(\theta, c_\theta) = P_0 \left[-\frac{g_i g_j}{\varphi''\left(\overleftarrow{\varphi'}\left(c_\theta^T g(\theta)\right)\right)} \right]_{i,j=0,\ldots,l}, \quad (3.95)$$

which implies that the matrix $P_0 m''(\theta, c_\theta)$ is symmetric. Under assumption (A.2), by Taylor expansion, there exists $t_n \in \mathbb{R}^{l+1}$ inside the segment that links c_θ and $\widehat{c}_\theta$ with

$$\begin{aligned} 0 &= P_n m'(\theta, \widehat{c}_\theta) \\ &= P_n m'(\theta, c_\theta) + \left(P_n m''(\theta, c_\theta)\right)^T (\widehat{c}_\theta - c_\theta) \\ &\quad + \tfrac{1}{2} (\widehat{c}_\theta - c_\theta)^T P_n m'''(\theta, t_n) (\widehat{c}_\theta - c_\theta), \end{aligned} \quad (3.96)$$

in which, $P_n m'''(\theta, t_n)$ is a $(l+1)-$vector whose entries are $(l+1)\times(l+1)-$matrices. By (A.2), we have for the sup-norm of vectors and matrices

$$\|P_n m'''(\theta, t_n)\| := \left\| \frac{1}{n} \sum_{i=1}^n m'''(X_i, \theta, t_n) \right\| \leq \frac{1}{n} \sum_{i=1}^n |H(X_i)|.$$

By the Law of Large Numbers (LLN), $P_n m'''(\theta, t_n) = O_P(1)$. So using (A.1), we can write the last term in the right hand side of (3.96) as $o_P(1)(\widehat{c}_\theta - c_\theta)$. On the other hand by (A.3), $P_n m''(\theta, c_\theta) := \frac{1}{n}\sum_{i=1}^n m''(X_i, \theta, c_\theta)$ converges to the matrix $P_0 m''(\theta, c_\theta)$. Write $P_n m''(\theta, c_\theta)$ as $P_0 m''(\theta, c_\theta) + o_P(1)$ to obtain from (3.96)

$$-P_n m'(\theta, c_\theta) = \left(P_0 m''(\theta, c_\theta) + o_P(1)\right)(\widehat{c}_\theta - c_\theta). \quad (3.97)$$

Under (A.3), by the Central Limit Theorem, we have $\sqrt{n} P_n m'(\theta, c_\theta) = O_P(1)$, which by (3.97) implies that $\sqrt{n}(\widehat{c}_\theta - c_\theta) = O_P(1)$. Hence, from (3.97), we get

$$\sqrt{n}(\widehat{c}_\theta - c_\theta) = \left[-P_0 m''(\theta, c_\theta)\right]^{-1} \sqrt{n} P_n m'(\theta, c_\theta) + o_P(1). \quad (3.98)$$

Under (A.3), the Central Limit Theorem concludes the proof of part 1. In the case when P_0 belongs to $\mathcal{M}_\theta$, then $c_\theta^T = (\varphi'(1), \underline{0}^T) := \underline{c}$ and calculation yields

$$P_0 m'(\theta, \underline{c}) m'(\theta, \underline{c})^T = \begin{pmatrix} 0 & \underline{0}_l^T \\ \underline{0}_l & P_0 g(\theta) g(\theta)^T \end{pmatrix} \text{ and } -\varphi''(1) P_0 m''(\theta, \underline{c}) = \begin{pmatrix} 1 & \underline{0}_l^T \\ \underline{0}_l & P_0 g(\theta) g(\theta)^T \end{pmatrix}.$$

A simple calculation yields (3.60).

Proof of (2).

By Taylor expansion, there exists $\bar{t}_n$ inside the segment that links c_θ and $\widehat{c}_\theta$ with

$$\begin{aligned}\widehat{\phi}_n(\mathcal{M}_\theta, P_0) &= P_n m(\theta, \widehat{c}_\theta) \\ &= P_n m(\theta, c_\theta) + \left(P_n m'(\theta, c_\theta)\right)^T (\widehat{c}_\theta - c_\theta) \\ &\quad + \frac{1}{2}(\widehat{c}_\theta - c_\theta)^T \left[P_n m''(\theta, c_\theta)\right] (\widehat{c}_\theta - c_\theta) \\ &\quad + \frac{1}{3!} \sum_{1 \le i,j,k \le d} (\widehat{c}_\theta - c_\theta)_i (\widehat{c}_\theta - c_\theta)_j \times \\ &\quad \times (\widehat{c}_\theta - c_\theta)_k P_n \frac{\partial^3}{\partial t_i \partial t_j \partial t_k} m(\theta, \bar{t}_n). \end{aligned} \tag{3.99}$$

When P_0 belongs to $\mathcal{M}_\theta$, then $c_\theta^T = \underline{c}$. Hence $P_n m(\theta, c_\theta) = P_n m(\theta, \underline{c}) = P_n 0 = 0$.

Furthermore, by part (1) in Theorem 3.4, it holds $\sqrt{n}(\widehat{c}_\theta - c_\theta) = O_p(1)$. Hence, by (A.1), (A.2) and (A.3), we get

$$\begin{aligned}\widehat{\phi}_n(\mathcal{M}_\theta, P_0) &= \left(P_n m'(\theta, c_\theta)\right)^T (\widehat{c}_\theta - c_\theta) + \\ &\quad \frac{1}{2}(\widehat{c}_\theta - c_\theta)^T \left[P_0 m''(\theta, c_\theta)\right] (\widehat{c}_\theta - c_\theta) + o_P(1/n), \end{aligned}$$

which by (3.98), implies

$$\begin{aligned}\widehat{\phi}_n(\mathcal{M}_\theta, P_0) &= \left[P_n m'(\theta, c_\theta)\right]^T \left[-P_0 m''(\theta, c_\theta)\right]^{-1} \left[P_n m'(\theta, c_\theta)\right] + \\ &\quad \frac{1}{2}\left[P_n m'(\theta, c_\theta)\right]^T \left[P_0 m''(\theta, c_\theta)\right]^{-1} \left[P_n m'(\theta, c_\theta)\right] + o_P(1/n) \\ &= \frac{1}{2}\left[P_n m'(\theta, c_\theta)\right]^T \left[-P_0 m''(\theta, c_\theta)\right]^{-1} \left[P_n m'(\theta, c)\right] + o_P(1/n). \end{aligned}$$

This yields to

$$\frac{2n}{\varphi''(1)} \widehat{\phi}_n(\mathcal{M}_\theta, P_0) = \left[\sqrt{n} P_n m'(\theta, c_\theta)\right]^T \left[-\varphi''(1) P_0 m''(\theta, c_\theta)\right]^{-1} \left[\sqrt{n} P_n m'(\theta, c_\theta)\right] + o_P(1). \tag{3.100}$$

Note that when P_0 belongs to $\mathcal{M}_\theta$, then $c_\theta^T = \underline{c}$ and calculation yields

$$P_0 m'(\theta, \underline{c}) m'(\theta, \underline{c})^T = \begin{pmatrix} 0 & \underline{0}_l^T \\ \underline{0}_l & P_0 g(\theta) g(\theta)^T \end{pmatrix} \text{ and } -\varphi''(1) P_0 m''(\theta, \underline{c}) = \begin{pmatrix} 1 & \underline{0}_l^T \\ \underline{0}_l & P_0 g(\theta) g(\theta)^T \end{pmatrix}.$$

Combining this with (3.100), we conclude the proof of part (2).

Proof of part (3). Since $(\widehat{c}_\theta - c_\theta) = O_P(1/\sqrt{n})$ and $P_n m'(\theta, c_\theta) = P_0 m'(\theta, c_\theta) + o_P(1) = 0 + o_P(1) = o_P(1)$, then, using (3.99), we obtain

$$\begin{aligned}\sqrt{n}\left(\widehat{\phi}_n(\mathcal{M}_\theta, P_0) - \phi(\mathcal{M}_\theta, P_0)\right) &= \sqrt{n}\left(\widehat{\phi}_n(\mathcal{M}_\theta, P_0) - P_0 m(\theta, c_\theta)\right) \\ &= \sqrt{n}\left(P_n m(\theta, c_\theta) - P_0 m(\theta, c_\theta)\right) + o_P(1), \end{aligned}$$

and the Central Limit Theorem yields to the conclusion of the proof of Theorem 3.4.

3.9.8 Proof of Proposition 3.7

Define the estimates

$$\widetilde{\theta}_\phi := \arg \inf_{\theta\in\Theta} \sup_{t\in T_\theta} P_n m(\theta, t),$$

$$\widetilde{\phi}\left(\mathcal{M}, P_0\right) := \inf_{\theta\in\Theta} \sup_{t\in T_\theta} P_n m(\theta, t)$$

and for all $\theta \in \Theta$,

$$\widetilde{c_\theta} := \arg \sup_{t\in T_\theta} P_n m(\theta, t).$$

By condition (C.5), for all n sufficiently large, it holds

$$\widehat{\theta}_\phi = \widetilde{\theta}_\phi \text{ and } \widehat{\phi}\left(\mathcal{M}, P_0\right) = \widetilde{\phi}\left(\mathcal{M}, P_0\right).$$

We prove that $\widetilde{\theta}_\phi$ and $\widetilde{\phi}\left(\mathcal{M}, P_0\right)$ are consistent. First, we prove the consistency of $\widetilde{\phi}\left(\mathcal{M}, P_0\right)$. We have

$$\left|\widetilde{\phi}\left(\mathcal{M}, P_0\right) - \phi\left(\mathcal{M}, P_0\right)\right| = \left|P_n m\left(\widetilde{\theta}_\phi, \widetilde{c_{\widetilde{\theta}_\phi}}\right) - P_0 m(\theta^*, c_{\theta^*})\right| =: |A|.$$

This implies

$$P_n m\left(\widetilde{\theta}_\phi, c_{\theta^*}\right) - P_0 m\left(\widetilde{\theta}_\phi, c_{\theta^*}\right) \leq A \leq P_n m\left(\theta^*, \widetilde{c_{\widetilde{\theta}_\phi}}\right) - P_0 m\left(\theta^*, \widetilde{c_{\widetilde{\theta}_\phi}}\right).$$

By condition (C.3), both the RHS and LHS terms in the above display go to 0. This implies that A tends to 0 which concludes the proof of part (i).

Proof of part (ii).

Since for sufficiently large n, by condition (C.5), we have $\widehat{c_\theta} = \widetilde{c_\theta}$ for all $\theta \in \Theta$, the convergence of $\sup_{\theta\in\Theta} \|\widetilde{c_\theta} - c_\theta\|$ to 0 implies (ii). We prove now that $\sup_{\theta\in\Theta} \|\widetilde{c_\theta} - c_\theta\|$ tends to 0. By the very definition of $\widetilde{c_\theta}$ and condition (C.3), we have

$$\begin{aligned} P_n m\left(\theta, \widetilde{c_\theta}\right) &\geq P_n m\left(\theta, c_\theta\right) \\ &\geq P_0 m\left(\theta, c_\theta\right) - o_P(1), \end{aligned} \tag{3.101}$$

where $o_P(1)$ does not depends upon θ (due to condition (C.3)). Hence, we have for all $\theta \in \Theta$,

$$P_0 m\left(\theta, c_\theta\right) - P_0 m\left(\theta, \widetilde{c_\theta}\right) \leq P_n m\left(\theta, \widetilde{c_\theta}\right) - P_0 m\left(\theta, \widetilde{c_\theta}\right) + o_P(1). \tag{3.102}$$

The term in the RHS of the above display is less than

$$\sup_{\theta\in\Theta, t\in T_\theta} |P_n m(\theta,t) - P_0 m(\theta,t)| + o_P(1)$$

which by (C.3), tends to 0. Let $\epsilon > 0$ be such that $\sup_{\theta\in\Theta} \|\widetilde{c_\theta} - c_\theta\| > \epsilon$. There exists some $a_n \in \Theta$ such that $\|\widetilde{c_{a_n}} - c_{a_n}\| > \epsilon$. Together with the strict concavity of the function $t \in T_\theta \to P_0 m(\theta,t)$ for all $\theta \in \Theta$, there exists $\eta > 0$ such that

$$P_0 m\left(a_n, c_{a_n}\right) - P_0 m\left(a_n, \widetilde{c_{a_n}}\right) > \eta.$$

We then conclude that

$$P\left\{\sup_{\theta\in\Theta} \|\widetilde{c_\theta} - c_\theta\| > \epsilon\right\} \leq P\left\{P_0 m\left(a_n, c_{a_n}\right) - P_0 m\left(a_n, \widetilde{c_{a_n}}\right) > \eta\right\},$$

and the RHS term tends to 0 by (3.102). This concludes the proof part (ii).

Proof of part (iii).

We prove that $\widetilde{\theta}_\phi$ converges to θ^*. By the very definition of $\widetilde{\theta}_\phi$, condition (C.4.b) and part (ii), we obtain

$$\begin{aligned} P_n m\left(\widetilde{\theta}_\phi, \widetilde{c_{\widetilde{\theta}_\phi}}\right) &\leq P_n m\left(\theta^*, \widetilde{c_{\theta^*}}\right) \\ &\leq P_0 m\left(\theta^*, \widetilde{c_{\widetilde{\theta}_\phi}}\right) - o_P(1), \end{aligned}$$

from which

$$\begin{aligned} P_0 m\left(\widetilde{\theta}_\phi, \widetilde{c_{\widetilde{\theta}_\phi}}\right) - P_0 m\left(\theta^*, \widetilde{c_{\widetilde{\theta}_\phi}}\right) &\leq P_0 m\left(\widetilde{\theta}_\phi, \widetilde{c_{\widetilde{\theta}_\phi}}\right) - P_n m\left(\widetilde{\theta}_\phi, \widetilde{c_{\widetilde{\theta}_\phi}}\right) + o_P(1) \\ &\leq \sup_{\{\theta\in\Theta, t\in T_\theta\}} |P_n m(\theta,t) - P_0 m(\theta,t)| + o_P(1) \quad (3.103) \end{aligned}$$

Further, by part (ii) and condition (C.4.a), for any positive ϵ, there exists $\eta > 0$ such that

$$P\left\{\left\|\widetilde{\theta}_\phi - \theta^*\right\| > \epsilon\right\} \leq P\left\{P_0 m\left(\widetilde{\theta}_\phi, \widetilde{c_{\widetilde{\theta}_\phi}}\right) - P_0 m\left(\theta^*, \widetilde{c_{\widetilde{\theta}_\phi}}\right) > \eta\right\}.$$

The RHS term, under condition (C.3), tends to 0 by (3.103). This concludes the proof of Proposition 3.7.

3.9.9 Proof of Theorem 3.5

Since $P_0 \in \mathcal{M}$, then $c_\theta = \underline{c}$. Some calculus yield

$$\frac{\partial}{\partial t} m(\theta_0, \underline{c}) = [0, -g_1(\theta_0), \ldots, -g_l(\theta_0)]^T = -\left[0, g(\theta_0)^T\right]^T,$$

$$\frac{\partial}{\partial\theta}m(\theta,t)=-\sum_{j=0}^{l}t_i\overleftarrow{\varphi'}(t^Tg(\theta))\frac{\partial}{\partial\theta}g_j(\theta),\quad \frac{\partial}{\partial\theta}m(\theta_0,\underline{c})=\underline{0}_d, \tag{3.104}$$

$$\frac{\partial^2}{\partial\theta\partial t}m(\theta_0,\underline{c})=\left[\underline{0}_d,-\frac{\partial}{\partial\theta}g_1(\theta_0),\ldots,-\frac{\partial}{\partial\theta}g_l(\theta_0)\right]=-\left[\underline{0}_d,\frac{\partial}{\partial\theta}g(\theta)\right]$$

$$\frac{\partial^2}{\partial t\partial\theta}m(\theta_0,\underline{c})=\left[\frac{\partial^2}{\partial\theta\partial t}m(\theta_0,\underline{c})\right]^T,\ \frac{\partial^2}{\partial\theta^2}m(\theta_0,\underline{c})=\left[\underline{0}_d,\ldots,\underline{0}_d\right],$$

and

$$\frac{\partial^2}{\partial t^2}m(\theta_0,\underline{c})=\frac{1}{\varphi''(1)}[-g_i(\theta_0)g_j(\theta_0)]_{i,j=0,1,\ldots,l}:=\frac{-1}{\varphi''(1)}\left(\overline{g}(\theta_0)\overline{g}(\theta_0)^T\right).$$

Integrating w.r.t. P_0, we obtain

$$P_0\frac{\partial}{\partial t}m(\theta_0,\underline{c})=\underline{0}_l,\quad P_0\frac{\partial}{\partial\theta}m(\theta_0,\underline{c})=\underline{0}_d,\ P_0\frac{\partial^2}{\partial\theta^2}m(\theta_0,\underline{c})=\left[\underline{0}_d,\ldots,\underline{0}_d\right], \tag{3.105}$$

$$P_0\frac{\partial^2}{\partial\theta\partial t}m(\theta_0,\underline{c})=-\left[\underline{0}_d,P_0\frac{\partial}{\partial\theta}g(\theta_0)\right], \tag{3.106}$$

$$P_0\frac{\partial^2}{\partial t\partial\theta}m(\theta_0,\underline{c})=\left[P_0\frac{\partial^2}{\partial\theta\partial t}m(\theta_0,\underline{c})\right]^T=-\left[\underline{0}_d,P_0\frac{\partial}{\partial\theta}g(\theta_0)\right]^T, \tag{3.107}$$

and

$$\begin{aligned}P_0\frac{\partial^2}{\partial t^2}m(\theta_0,\underline{c}) &= \frac{-1}{\varphi''(1)}\left[P_0g_i(\theta_0)g_j(\theta_0)\right]_{i,j=0,1,\ldots,l}.\\ &= \frac{-1}{\varphi''(1)}\begin{pmatrix}1 & \underline{0}_l^T\\ \underline{0}_l & P_0g(\theta_0)g(\theta_0)^T\end{pmatrix}\end{aligned} \tag{3.108}$$

By the very definition of $\widehat{\theta}_\phi$ and $\widehat{c_{\widehat{\theta}_\phi}}$, they both obey

$$\begin{cases}P_n\frac{\partial}{\partial t}m\left(\theta,t\right) &= 0\\ P_n\frac{\partial}{\partial\theta}m\left(\theta,t(\theta)\right) &= 0,\end{cases}$$

i.e.,

$$\begin{cases}P_n\frac{\partial}{\partial t}m\left(\widehat{\theta}_\phi,\widehat{c_{\widehat{\theta}_\phi}}\right) &= 0\\ P_n\frac{\partial}{\partial\theta}m\left(\widehat{\theta}_\phi,\widehat{c_{\widehat{\theta}_\phi}}\right)+P_n\frac{\partial}{\partial t}m\left(\widehat{\theta}_\phi,\widehat{c_{\widehat{\theta}_\phi}}\right)\frac{\partial}{\partial\theta}\widehat{c_{\widehat{\theta}_\phi}} &= 0.\end{cases}$$

The second term in the left hand side of the second equation is equal to 0, due to the first equation. Hence $\widehat{c_{\widehat{\theta}_\phi}}$ and $\widehat{\theta}_\phi$ are solutions of the somehow simpler system

$$\left\{ \begin{array}{rcll} P_n \frac{\partial}{\partial t} m\left(\widehat{\theta}_\phi, \widehat{c_{\widehat{\theta}_\phi}}\right) & = & 0 & (E1) \\ P_n \frac{\partial}{\partial \theta} m\left(\widehat{\theta}_\phi, \widehat{c_{\widehat{\theta}_\phi}}\right) & = & 0 & (E2). \end{array} \right.$$

Use a Taylor expansion in (E1) ; there exists $\left(\widetilde{\theta}_n, \widetilde{t}_n\right)$ inside the segment that links $(\widehat{\theta}_\phi, \widehat{c_{\widehat{\theta}_\phi}})$ and $(\theta_0, \underline{c})$ such that

$$\begin{aligned} 0 &= P_n \frac{\partial}{\partial t} m\left(\theta_0, \underline{c}\right) + \left[\left(P_n \frac{\partial^2}{\partial t^2} m(\theta_0, \underline{c}) \right)^T, \left(P_n \frac{\partial^2}{\partial \theta \partial t} m(\theta_0, \underline{c}) \right)^T \right] a_n \\ &\quad + \frac{1}{2} a_n^T A_n a_n, \end{aligned} \tag{3.109}$$

with

$$a_n := \left(\left(\widehat{c_{\widehat{\theta}_\phi}} - \underline{c} \right)^T, \left(\widehat{\theta}_\phi - \theta_0 \right)^T \right)^T \tag{3.110}$$

and

$$A_n := \left(\begin{array}{cc} P_n \frac{\partial^3}{\partial t^3} m(\widetilde{\theta}, \widetilde{c}_n) & P_n \frac{\partial^3}{\partial t \partial \theta \partial t} m(\widetilde{\theta}, \widetilde{c}_n) \\ P_n \frac{\partial^3}{\partial \theta \partial t^2} m(\widetilde{\theta}, \widetilde{c}_n) & P_n \frac{\partial^3}{\partial \theta^2 \partial t} m(\widetilde{\theta}, \widetilde{c}_n) \end{array} \right). \tag{3.111}$$

By (A.5), the LLN implies that $A_n = O_P(1)$. So using (A.4), we can write the last term in right hand side of (3.109) as $o_P(1)a_n$. On the other hand by (A.6), we can write also
$\left[\left(P_n \frac{\partial^2}{\partial t^2} m(\theta_0, \underline{c}) \right)^T, \left(P_n \frac{\partial^2}{\partial \theta \partial t} m(\theta_0, \underline{c}) \right)^T \right]$ as $\left[P_0 \frac{\partial^2}{\partial t^2} m(\theta_0, \underline{c}), \left(P_0 \frac{\partial^2}{\partial \theta \partial t} m(\theta_0, \underline{c}) \right)^T \right] + o_P(1)$
to obtain from (3.109)

$$-P_n \frac{\partial}{\partial t} m(\theta_0, \underline{c}) = \left[P_0 \frac{\partial^2}{\partial t^2} m(\theta_0, \underline{c}) + o_P(1), \left(P_0 \frac{\partial^2}{\partial \theta \partial t} m(\theta_0, \underline{c}) \right)^T + o_P(1) \right] a_n. \tag{3.112}$$

In the same way, using a Taylor expansion in (E2), there exists $(\overline{\theta}_n, \overline{t}_n)$ inside the segment that links $\left(\widehat{\theta}_\phi, \widehat{c_{\widehat{\theta}_\phi}}\right)$ and $(\theta_0, \underline{c})$ such that

$$\begin{aligned} 0 &= P_n \frac{\partial}{\partial \theta} m(\theta_0, \underline{c}) + \left[\left(P_n \frac{\partial^2}{\partial t \partial \theta} m(\theta_0, \underline{c}) \right)^T, \left(P_n \frac{\partial^2}{\partial \theta^2} m(\theta_0, \underline{c}) \right)^T \right] a_n \\ &\quad + \frac{1}{2} a_n^t B_n a_n, \end{aligned} \tag{3.113}$$

with

$$B_n := \left[\begin{array}{cc} P_n \frac{\partial^3}{\partial t^2 \partial \theta} m(\overline{\theta}_n, \overline{t}_n) & P_n \frac{\partial^3}{\partial t \partial \theta^2} m(\overline{\theta}_n, \overline{t}_n) \\ P_n \frac{\partial^3}{\partial \theta \partial t \partial \theta} m(\overline{\theta}_n, \overline{t}_n) & P_n \frac{\partial^3}{\partial \theta^3} m(\overline{\theta}_n, \overline{t}_n) \end{array} \right].$$

As in (3.112), we obtain

$$-P_n \frac{\partial}{\partial \theta} m(\theta_0, \underline{c}) = \left[\left(P_0 \frac{\partial^2}{\partial t \partial \theta} m(\theta_0, \underline{c}) \right)^T + o_P(1), P_0 \frac{\partial^2}{\partial \theta^2} m(\theta_0, \underline{c}) + o_P(1) \right] a_n. \tag{3.114}$$

From (3.112) and (3.114), we get

$$\begin{aligned} \sqrt{n} a_n &= \sqrt{n} \left(\begin{array}{cc} P_0 \frac{\partial^2}{\partial t^2} m(\theta_0, c(\theta_0)) & \left(P_0 \frac{\partial^2}{\partial \theta \partial t} m(\theta_0, \underline{c}) \right)^T \\ \left(P_0 \frac{\partial^2}{\partial t \partial \theta} m(\theta_0, c(\theta_0)) \right)^T & P_0 \frac{\partial^2}{\partial \theta^2} m(\theta_0, c(\theta_0)) \end{array} \right)^{-1} \times \\ &\quad \times \left(\begin{array}{c} -P_n \frac{\partial}{\partial t} m(\theta_0, \underline{c}) \\ -P_n \frac{\partial}{\partial \theta} m(\theta_0, \underline{c}) \end{array} \right) + o_P(1). \end{aligned} \tag{3.115}$$

Denote S the $(l+1+d) \times (l+1+d)-$matrix defined by

$$S := \left(\begin{array}{cc} S_{11} & S_{12} \\ S_{21} & S_{22} \end{array} \right) := \left(\begin{array}{cc} P_0 \frac{\partial^2}{\partial t^2} m(\theta_0, c(\theta_0)) & \left(P_0 \frac{\partial^2}{\partial \theta \partial t} m(\theta_0, \underline{c}) \right)^T \\ \left(P_0 \frac{\partial^2}{\partial t \partial \theta} m(\theta_0, c(\theta_0)) \right)^T & P_0 \frac{\partial^2}{\partial \theta^2} m(\theta_0, c(\theta_0)) \end{array} \right). \tag{3.116}$$

We have

$$S_{11} = \frac{-1}{\varphi''(1)} \left(\begin{array}{cc} 1 & \underline{0}_l^T \\ \underline{0}_l & P_0 g(\theta_0) g(\theta_0)^T \end{array} \right) \tag{3.117}$$

$$S_{12} = -\left[\underline{0}_d, P_0 \frac{\partial}{\partial \theta} g(\theta_0) \right]^T, \; S_{21} = -\left[\underline{0}_d, P_0 \frac{\partial}{\partial \theta} g(\theta_0) \right] \text{ and} \tag{3.118}$$

$$S_{22} = P_0 \frac{\partial^2}{\partial \theta^2} m(\theta_0, \underline{c}) = [\underline{0}_d, \ldots, \underline{0}_d]. \tag{3.119}$$

The inverse matrix S^{-1} of the matrix S writes

$$S^{-1} = \left(\begin{array}{cc} S_{11}^{-1} + S_{11}^{-1} S_{12} S_{22.1}^{-1} S_{21} S_{11}^{-1} & -S_{11}^{-1} S_{12} S_{22.1}^{-1} \\ -S_{22.1}^{-1} S_{21} S_{11}^{-1} & S_{22.1}^{-1} \end{array} \right), \tag{3.120}$$

where

$$\begin{aligned} S_{22.1} &= -S_{21} S_{11}^{-1} S_{12} \\ &= \left[\underline{0}_d, P_0 \frac{\partial}{\partial \theta} g(\theta_0) \right] [\varphi''(1)] \left[\begin{array}{cc} 1 & \underline{0}_l^T \\ \underline{0}_l & \left[P_0 g(\theta_0) g(\theta_0)^T \right]^{-1} \end{array} \right] \left[\underline{0}_d, P_0 \frac{\partial}{\partial \theta} g(\theta_0) \right]^T \\ &= \varphi''(1) \left[P_0 \frac{\partial}{\partial \theta} g(\theta_0) \right] \left[P_0 g(\theta_0) g(\theta_0)^T \right]^{-1} \left[P_0 \frac{\partial}{\partial \theta} g(\theta_0) \right]^T. \end{aligned} \tag{3.121}$$

From (3.115), using (3.116) and (3.120), we can write

$$\sqrt{n}\begin{pmatrix} \widehat{c_{\widehat{\theta}_\phi}} - \underline{c} \\ \widehat{\theta}_\phi - \theta_0 \end{pmatrix} = \begin{pmatrix} S_{11}^{-1} + S_{11}^{-1}S_{12}S_{22.1}^{-1}S_{21}S_{11}^{-1} & -S_{11}^{-1}S_{12}S_{22.1}^{-1} \\ -S_{22.1}^{-1}S_{21}S_{11}^{-1} & S_{22.1}^{-1} \end{pmatrix} \times \times\sqrt{n}\begin{pmatrix} -P_n\frac{\partial}{\partial t}m(\theta_0,\underline{c}) \\ \underline{0}_d \end{pmatrix} + o_P(1). \tag{3.122}$$

Note that

$$\sqrt{n}\begin{pmatrix} -P_n\frac{\partial}{\partial t}m(\theta_0,\underline{c}) \\ \underline{0}_d \end{pmatrix}, \tag{3.123}$$

under assumption (A.6), by the Central Limit Theorem, converges in distribution to a centered multivariate normal variable with covariance matrix

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \tag{3.124}$$

where

$$M_{11} = \begin{pmatrix} 0 & \underline{0}_l^T \\ \underline{0}_l & P_0 g(\theta_0)g(\theta_0)^T \end{pmatrix},\ M_{12} = \begin{pmatrix} \underline{0}_d^T \\ \vdots \\ \underline{0}_d^T \end{pmatrix},\ M_{21} = \begin{pmatrix} 0 & \underline{0}_l^T \\ \vdots & \vdots \\ 0 & \underline{0}_l^T \end{pmatrix} \text{ and } M_{22} = \begin{pmatrix} \underline{0}_d^T \\ \vdots \\ \underline{0}_d^T \end{pmatrix}. \tag{3.125}$$

Hence, from (3.123), we deduce that

$$\sqrt{n}\begin{pmatrix} \widehat{c_{\widehat{\theta}_\phi}} - \underline{c} \\ \widehat{\theta}_\phi - \theta_0 \end{pmatrix} \tag{3.126}$$

converges in distribution to a centered multivariate normal variable with covariance matrix

$$C = S^{-1}M\left[S^{-1}\right]^T := \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \tag{3.127}$$

and using (3.125) and some algebra, we get

$$C_{11} = \varphi''(1)^2\begin{bmatrix} 0 & \underline{0}_l^T \\ \underline{0}_l & \left[P_0 g(\theta_0)g(\theta_0)^T\right]^{-1} \end{bmatrix} - \varphi''(1)^2\begin{bmatrix} 0 & \underline{0}_l^T \\ \underline{0}_l & \left[P_0 g(\theta_0)g(\theta_0)^T\right]^{-1} \end{bmatrix} \times \times\left[\underline{0}, P_0\frac{\partial}{\partial\theta}g(\theta_0)\right]^T[C_{22}]\left[\underline{0}, P_0\frac{\partial}{\partial\theta}g(\theta_0)\right]\begin{bmatrix} 0 & \underline{0}_l^T \\ \underline{0}_l & \left[P_0 g(\theta_0)g(\theta_0)^T\right]^{-1} \end{bmatrix}, \tag{3.128}$$

$$C_{12} = [\underline{0}_l, \ldots, \underline{0}_l],\ C_{21} = [\underline{0}_d, \ldots, \underline{0}_d] \tag{3.129}$$

and

$$C_{22} = \left\{ \left[P_0 \frac{\partial}{\partial \theta} g(\theta_0) \right] \left[P_0 \left(g(\theta_0) g(\theta_0)^T \right) \right]^{-1} \left[P_0 \frac{\partial}{\partial \theta} g(\theta_0) \right]^T \right\}^{-1}. \tag{3.130}$$

From (3.126), we deduce that C_{11} and C_{22} are respectively the limit covariance matrix of $\sqrt{n}\left(\widehat{c_{\widehat{\theta}_\phi}} - \underline{c}\right)$ and $\sqrt{n}\left(\widehat{\theta}_\phi - \theta_0\right)$, i.e., $U = C_{11}$ and $V = C_{22}$. (3.129) implies that $\sqrt{n}\left(\widehat{c_{\widehat{\theta}_\phi}} - \underline{c}\right)$ and $\sqrt{n}\left(\widehat{\theta}_\phi - \theta_0\right)$ are asymptotically uncorrelated. This concludes the Proof of Theorem 3.5.

3.9.10 Proof of Theorem 3.6

Under assumptions (A.4-6), as in the proof of Theorem 3.5, we obtain

$$\sqrt{n} \begin{pmatrix} \widehat{c_{\widehat{\theta}_\phi}} - c_{\theta^*} \\ \widehat{\theta}_\phi - \theta^* \end{pmatrix} = \sqrt{n} S^{-1} \begin{pmatrix} -P_n \frac{\partial}{\partial t} m(\theta^*, c_{\theta^*}) \\ -P_n \frac{\partial}{\partial \theta} m(\theta^*, c_{\theta^*}) \end{pmatrix} + o_P(1),$$

and the CLT concludes the proof.

3.9.11 Proof of Theorem 3.7

Using Taylor expansion at $(\underline{c}, \theta_0)$, we get

$$\begin{aligned} \widehat{F}_n(x) &:= \sum_{i=1}^n \widehat{Q^*_{\widehat{\theta}_\phi}} \mathbb{1}_{(-\infty,x]}(X_i) := \frac{1}{n} \sum_{i=1}^n \overleftarrow{\varphi'} \left(\widehat{c_{\widehat{\theta}_\phi}}^T \overline{g}(X_i, \widehat{\theta}_\phi) \right) \mathbb{1}_{(-\infty,x]}(X_i) \\ &= F_n(x) + \frac{1}{n} \left[\sum_{i=1}^n \overline{g}(X_i, \theta_0) \mathbb{1}_{(-\infty,x]}(X_i) \right]^T \frac{1}{\varphi''(1)} \left(\widehat{c_{\widehat{\theta}_\phi}} - \underline{c} \right) + o_P(\delta_n), \end{aligned} \tag{3.131}$$

where $\delta_n := \left\| \widehat{c_{\widehat{\theta}_\phi}} - \underline{c} \right\| + \left\| \widehat{\theta}_\phi - \theta_0 \right\|$, which by Theorem 3.5, is equal to $O_P(1/\sqrt{n})$. Hence, (3.131) yields

$$\begin{aligned} \sqrt{n} \left(\widehat{F}_n(x) - F(x) \right) &= \sqrt{n} \left(F_n(x) - F(x) \right) + \\ &\quad + \frac{1}{\varphi''(1)} \left[P_0 \left(\overline{g}(\theta_0) \mathbb{1}_{(-\infty,x]} \right) \right]^T \sqrt{n} \left(\widehat{c_{\widehat{\theta}_\phi}} - \underline{c} \right) + o_P(1). \end{aligned} \tag{3.132}$$

On the other hand, from (3.122), we get

$$\sqrt{n} \left(\widehat{c_{\widehat{\theta}_\phi}} - \underline{c} \right) = H \sqrt{n} \left(-P_n \frac{\partial}{\partial \theta} m(\theta_0, \underline{c}) \right) + o_P(1) \tag{3.133}$$

with

$$H = S_{11}^{-1} + S_{11}^{-1} S_{12} S_{22.1}^{-1} S_{21} S_{11}^{-1}. \tag{3.134}$$

We will use $f(.)$ to denote the function $\mathbb{1}_{(-\infty,x]}(.) - F(x)$, for all $x \in \mathbb{R}$. Substituting (3.133) in (3.132), we get

$$\begin{aligned} \sqrt{n}\left(\widehat{F}_n(x) - F(x)\right) &= \sqrt{n} P_n f + \frac{1}{\varphi''(1)} \left[P_0\left(\overline{g}(\theta_0)\mathbb{1}_{(-\infty,x]}\right)\right]^T H \times \\ &\quad \times \sqrt{n}\left(-P_n \frac{\partial}{\partial\theta} m(\theta_0, \underline{c})\right) + o_P(1). \end{aligned} \tag{3.135}$$

By the Multivariate Central Limit Theorem, the vector $\sqrt{n}\left(P_n f, -\left[P_n \frac{\partial}{\partial\theta} m(\theta_0, \underline{c})\right]^T\right)^T$ converges in distribution to a centered multivariate normal variable which implies that $\sqrt{n}\left(\widehat{F}_n(x) - F(x)\right)$ is asymptotically centered normal variable. We calculate now its limit variance, noted $W(x)$.

$$\begin{aligned} W(x) &= F(x)(1 - F(x)) + \frac{1}{\varphi''(1)^2} \left[P_0\left(\overline{g}(\theta_0)\mathbb{1}_{(-\infty,x]}\right)\right]^T U \left[P_0\left(\overline{g}(\theta_0)\mathbb{1}_{(-\infty,x]}\right)\right] + \\ &\quad + 2\frac{1}{\varphi''(1)} \left[P_0\left(\overline{g}(\theta_0)\mathbb{1}_{(-\infty,x]}\right)\right]^T H \left[P_0\left(-\frac{\partial}{\partial t} m(\theta_0, \underline{c})\mathbb{1}_{(-\infty,x]}\right)\right]. \end{aligned} \tag{3.136}$$

Use the explicit forms of $\frac{\partial}{\partial t} m(\theta_0, \underline{c})$, the matrices U and V and some algebra to obtain

$$W(x) = F(x)\left(1 - F(x)\right) - \left[P_0\left(g(\theta_0)\mathbb{1}_{(-\infty,x]}\right)\right]^T \Gamma \left[P_0\left(g(\theta_0)\mathbb{1}_{(-\infty,x]}\right)\right],$$

with

$$\begin{aligned} \Gamma &= \left[P_0 g(\theta_0) g(\theta_0)^T\right]^{-1} - \left[P_0 g(\theta_0) g(\theta_0)^T\right]^{-1} \left[P_0 \frac{\partial}{\partial\theta} g(\theta_0)\right]^T V \times \\ &\quad \times \left[P_0 \frac{\partial}{\partial\theta} g(\theta_0)\right] \left[P_0 g(\theta_0) g(\theta_0)^T\right]^{-1}. \end{aligned}$$

This concludes the proof of Theorem 3.7.

Chapitre 4

Annexe : Sur l'estimation de l'entropie des lois à support dénombrable

Article publié en version réduite sous la référence : Keziou (2002). Sur l'estimation de l'entropie des lois à support dénombrable. C. R. Math. Acad. Sci. Paris, 335 (9), 763-766.

Soit P une loi de probabilité discrète sur un espace infini dénombrable $\mathcal{X}$. On étudie la vitesse de convergence presque sûre de l'estimateur "plug-in" de l'entropie $H := H(P)$ de la loi de probabilité inconnue P. On démontre aussi la convergence presque sûre de l'estimateur pour des variables aléatoires stationnaires ergodiques, et pour des variables aléatoires stationnaires $\alpha-$mélangeantes sous une condition faible sur la queue de distribution de la loi P.

4.1 Introduction

Soit X une variable aléatoire discrète de loi de probabilité inconnue P, sur un espace infini dénombrable $\mathcal{X}$. Pour tout x appartenant à $\mathcal{X}$, on note $p(x)$ la probabilité $P\{X = x\}$. L' entropie de Shannon de la loi P est définie par

$$H := H(P) := -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) = \mathbf{E}\{-\log_2 p(X)\}$$

où $\log_2$ est le logarithme de base 2, et $\mathbf{E}$ désigne l'espérance.

Soit $X_1, ..., X_n$ un échantillon aléatoire de loi P. L'estimateur "plug-in" $\widehat{H}_n$ de

l'entropie H est défini par

$$\widehat{H}_n := H(\widehat{P}_n) = -\sum_{x\in\mathcal{X}} \widehat{p}_n(x)\log_2 \widehat{p}_n(x) \tag{4.1}$$

où $\widehat{P}_n$ est la mesure empirique définie par $\widehat{p}_n(x) := n^{-1}\sum_{i=1}^n \mathbb{1}_{\{X_i=x\}}$ pour tout $x \in \mathcal{X}$.

Basharin (1959) a étudié les propriétés asymptotiques de l'estimateur plug-in $\widehat{H}_n$ de l'entropie de lois de probabilité discrètes à support fini. Antos et Kontoyiannis (2001) ont montré la convergence presque sûre universelle de $\widehat{H}_n$ dans le cas infini dénombrable. Ils ont montré aussi que pour toute vitesse de convergence $a_n \to 0$, il existe une loi P, telle que pour tout estimateur H_n de $H(P)$ on a

$$\limsup_{n\to\infty} \frac{\mathbf{E}\,|H_n - H(P)|}{a_n} = \infty.$$

Il n'existe donc pas de vitesse de convergence universelle pour les estimateurs de l'entropie ; voir Antos and Kontoyiannis (2001a) et Antos and Kontoyiannis (2001b).

Le but de cette note est d'expliciter des conditions sur de vastes classes $\mathcal{P}$ de lois P, correspondant à des hypothèses naturelles dans le domaine du codage et de la compression universelle de sources avec alphabets grands ou infinis (voir Yang and Jia (2000), Verdú (1998) et les références ci-incluses), pour lesquelles on donnera des vitesses de convergence.

Les références Antos and Kontoyiannis (2001a) et Antos and Kontoyiannis (2001b) fournissent des vitesses de convergence dans $L_1(P)$ et $L_2(P)$ sous des hypothèses assez contraignantes sur P, qui ne sont pas satisfaites par les modèles de Poisson et géométriques utilisés dans Yang and Jia (2000).

Nous démontrons aussi la convergence presque sûre de l'estimateur dans le cas stationnaire ergodique et sationnaire $\alpha-$mélangeant sous une condition faible sur la queue de distribution de P.

L'estimateur plug-in de l'entropie (voir (4.1)) s'écrit sous la forme

$$\widehat{H}_n := -\frac{1}{n}\sum_{i=1}^n \log_2 \widehat{P}_n(X_i). \tag{4.2}$$

Cette écriture facilite l'étude de la vitesse de convergence presque sûre.

De nombreux travaux traitent du cas continu. Les articles de Tsybakov and van der Meulen (1996) et Belzunce *et al.* (2001) fournissent, dans ce cas, des vitesses de convergence de certains estimateurs de l'entropie et présentent une bibliographie.

4.2 Résultats

On donne trois propositions essentielles aux démonstrations des théorèmes à suivre.

Proposition 4.1. *Pour tout* $0 \leq \alpha < 1/2$, $\delta \geq 0$, *on a*

$$\lim_{n\to+\infty} n^{\alpha}(\log_2 n)^{\delta} \sup_{x\in\mathcal{X}} |\widehat{p}_n(x) - p(x)| = 0$$

presque sûrement.

Proposition 4.2. *Pour tout* $0 \leq \alpha < 1/2$, $\delta \geq 0$, *on a*

$$\lim_{n\to+\infty} n^{\alpha}(\log_2 n)^{\delta} \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{p(X_i)<t\}} - P\{p(X) < t\} \right| = 0$$

presque sûrement.

Introduisons l'ensemble

$$A_n = A_n(C,\beta) := \left\{x \in \mathcal{X};\ p(x) \geq Cn^{-\beta}\right\}.$$

Proposition 4.3. *Soient* α, β, γ, C *tels que* $0 < \beta < \alpha < 1/2$, $0 \leq \gamma \leq \alpha - \beta < 1/2$ *et* $C > 0$. *Alors, on a pour tout* $\delta \geq 0$ *:*

$$\lim_{n\to+\infty} n^{\gamma}(\log_2 n)^{\delta} \sup_{x\in A_n} |\log_2 \widehat{p}_n(x) - \log_2 p(x)| = 0$$

presque sûrement.

Dans tout ce qui suit, on suppose que l'entropie H de la loi P est finie. L'hypothèse suivante sur la distribution de la variable X permet de trouver une vitesse de convergence presque sûre de l'estimateur $\widehat{H}_n$.

(H.1) Il existe $\beta' > 0$ tel que

$$\sum_{n=1}^{+\infty} nP\left\{p(X) < n^{-\beta'}\right\} < \infty.$$

Théorème 4.1. *Supposons (H.1). Soient* $0 \leq \gamma < 1/2$ *et* $\delta \geq 0$. *S'il existe* α, β *et* $C > 0$ *vérifiant* $0 < \beta < \alpha < 1/2$, $0 < \gamma < \alpha - \beta < 1/2$ *et*

(H.2)
$$\lim_{n\to+\infty} n^{\gamma}(\log_2 n)^{1+\delta} P\left\{p(X) < Cn^{-\beta}\right\} = 0.$$

Alors, on a presque sûrement

$$\lim_{n\to+\infty} n^{\gamma}(\log_2 n)^{\delta} \left|\widehat{H}_n - H\right| = 0.$$

Remarque 4.1. *L'hypothèse (H.1) est utilisée pour appliquer l'inégalité de Hoeffding et montrer la convergence presque sûre de* $n^{\gamma}(\log_2 n)^{\delta}\left|H_n - H\right|$ *vers* 0. *L'hypothèse (H.2) est utilisée pour montrer que* $n^{\gamma}(\log_2 n)^{\delta}\left|\widehat{H}_n - H_n\right|$ *est négligeable.*

Remarque 4.2. *Les références Antos and Kontoyiannis (2001a) et Antos and Kontoyiannis (2001b) montrent que l'erreur* $L^1(P)$ *de l'estimateur* $\widehat{H}_n$ *est de l'ordre de* $n^{-\frac{q-1}{q}}$ *si la loi* P *vérifie* $c_1/i^q \leq p(i) \leq c_2/i^q$ *pout tout* $i \in \mathbb{N}^*$ *avec* $1 < q < 2$ *et* $c_1, c_2 > 0$ *(voir Antos and Kontoyiannis (2001a) Théorème 7). Cette condition n'est pas vérifiée pour les lois géométriques et les lois de Poisson.*

Remarque 4.3. *Les hypothèses (H.1) et (H.2) contrôlent la probabilité d'observer une modalité correspondant à une petite probabilité, ce qui est plus faible et plus naturelle qu'une hypothèse sur les queues de distribution.*

Soit $(\alpha_n)_{n\in\mathbb{N}}$ la suite de coefficients de mélange fort définie par

$$\alpha_0 := 1/2 \text{ et pour tout } n \in \mathbb{N},\ \alpha_n := \sup_{k\in\mathbb{Z}} \alpha\left(\mathcal{F}_k, \sigma\left(X_{k+n}\right)\right)$$

où $\mathcal{F}_k := \sigma\left(X_i, i \leq k\right)$ est la sigma-algèbre engendrée par $(X_i, i \leq k)$ et α est le coefficient de mélange fort défini pour toutes tribus $\mathcal{A}$, $\mathcal{B}$ par

$$\alpha(\mathcal{A}, \mathcal{B}) := 2\sup\left\{\left|P(A\cap B) - P(A)P(B)\right|,\ (A,B) \in \mathcal{A}\times\mathcal{B}\right\}.$$

On suppose que $X_i = 0$ pour tout $i \leq 0$.

Théorème 4.2. *(a) Soit* $X_1, \ldots, X_n$ *une suite de variables aléatoires stationnaires et ergodiques de loi* P. *Sous l'hypothèse (H.3), l'estimateur* $\widehat{H}_n$ *est presque sûrement convergent.*

(b) Soit $X_1, \ldots, X_n$ *une suite de variables aléatoires stationnaires. Sous l'hypothèse (H.3), si la suite de mélange fort* $(\alpha_n)_{n\in\mathbb{N}}$ *vérifie*

$$\sum_{n\geq 0} \frac{\alpha_n}{n+1} < \infty,$$

alors l'estimateur $\widehat{H}_n$ *est presque sûrement convergent.*

4.3 Démonstrations

4.3.1 Démonstration de la Proposition 4.1

Cette démonstration est basée sur celle de Pollard (1984) de la convergence uniforme de la fonction de répartition empirique. On démontre la Proposition 4.1 en trois étapes.

Etape 1 :

Indépendamment de l'échantillon $X_1, X_2, \ldots, X_n$, nous considérons un autre échantillon X_1^*, X_2^*, ..., X_n^* de même loi P. Pour tout $x \in \mathcal{X}$, définissons la variable aléatoire

$$Y_n(x) := n^{\alpha}(\log_2 n)^{\delta} \left|\widehat{p}_n^*(x) - p(x)\right|,$$

avec

$$\widehat{p}_n^*(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\left\{X_i^* = x\right\}}.$$

On a

$$\mathbf{Var}\ Y_n(x) = n^{(2\alpha-1)}(\log_2 n)^{2\delta} p(x)\left(1 - p(x)\right) \leq \frac{n^{(2\alpha-1)}(\log_2 n)^{2\delta}}{4}.$$

Par l'inégalité de Bienaymé-Tchebitchev on obtient

$$P\left\{n^{\alpha}(\log_2 n)^{\delta}\left|\widehat{p}_n^*(x) - p(x)\right| > \frac{\epsilon}{2}\right\} \leq \frac{4\mathbf{Var}\ Y_n(x)}{\epsilon^2} \leq \frac{n^{(2\alpha-1)}(\log_2 n)^{2\delta}}{\epsilon^2}.$$

D'où

$$P\left\{n^{\alpha}(\log_2 n)^{\delta}\left|\widehat{p}_n^*(x) - p(x)\right| \leq \frac{\epsilon}{2}\right\} \geq 1 - \frac{n^{(2\alpha-1)}(\log_2 n)^{2\delta}}{\epsilon^2}.$$

Par conséquent, il existe $n_0 > 0$, tel que pour tout $n \geq n_0$, on a

$$P\left\{n^{\alpha}(\log_2 n)^{\delta}\left|\widehat{p}_n^*(x) - p(x)\right| \leq \frac{\epsilon}{2}\right\} \geq \frac{1}{2}. \tag{4.3}$$

Nous allons appliquer le Lemme de symetrisation dans Pollard (1984) que nous rappelons ici

Lemme 4.1. *Soient $\{Z(t), t \in T\}$, $\{Z^*(t), t \in T\}$ deux processus indépendants où T est un ensemble d'indices. Supposons qu'il existe deux nombres $\beta > 0$ et $\gamma > 0$ tels que*

$$\forall t \in T: \ P\left\{|Z^*(t)| \leq \gamma\right\} \geq \beta.$$

Alors, on a

$$\forall \epsilon > 0, P\left\{\sup_t |Z(t)| > \epsilon\right\} \leq \beta^{-1} P\left\{\sup_t |Z(t) - Z^*(t)| > \epsilon - \gamma\right\}.$$

Par application du Lemme 4.1 pour $T = \mathcal{X}$, $\gamma = \epsilon/2$, $\beta = 1/2$, $Z(x)$=$n^{\alpha}(\log_2 n)^{\delta}\left(\widehat{p}_n(x) - p(x)\right)$ et $Z^*(x)$=$n^{\alpha}(\log_2 n)^{\delta}\left(\widehat{p}_n^*(x) - p(x)\right)$, on obtient pour tout $n \geq n_0$

$$P\left\{n^{\alpha}(\log_2 n)^{\delta}\sup_{x\in\mathcal{X}}\left|\widehat{p}_n(x) - p(x)\right| > \epsilon\right\} \leq$$

$$2P\left\{n^{\alpha}(\log_2 n)^{\delta}\sup_{x\in\mathcal{X}}\left|\widehat{p}_n(x) - \widehat{p}_n^*(x)\right| > \epsilon/2\right\} \tag{4.4}$$

Etape 2 :

Soient $\sigma_1, \sigma_2, \ldots, \sigma_n$ n variables aléatoires i.i.d. de loi $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2$, et indépendantes de $(X_1, \ldots, X_n,\ X_1^*, \ldots, X_n^*)$. On a

$$P\left\{n^\alpha (\log_2 n)^\delta \sup_{x\in\mathcal{X}} |\widehat{p}_n(x) - \widehat{p}_n^*(x)| > \epsilon/2\right\} =$$

$$P\left\{n^\alpha (\log_2 n)^\delta \sup_{x\in\mathcal{X}} \left| n^{-1}\sum_{i=1}^n \left(\mathbb{1}_{\{X_i=x\}} - \mathbb{1}_{\left\{X_i^*=x\right\}}\right)\right| > \epsilon/2\right\} =$$

$$P\left\{n^\alpha (\log_2 n)^\delta \sup_{x\in\mathcal{X}} \left| n^{-1}\sum_{i=1}^n \sigma_i \left(\mathbb{1}_{\{X_i=x\}} - \mathbb{1}_{\left\{X_i^*=x\right\}}\right)\right| > \epsilon/2\right\} \leq$$

$$P\left\{n^\alpha (\log_2 n)^\delta \sup_{x\in\mathcal{X}} \left| n^{-1}\sum_{i=1}^n \sigma_i \mathbb{1}_{\{X_i=x\}}\right| > \epsilon/4\right\} +$$

$$P\left\{n^\alpha (\log_2 n)^\delta \sup_{x\in\mathcal{X}} \left| n^{-1}\sum_{i=1}^n \sigma_i \mathbb{1}_{\left\{X_i^*=x\right\}}\right| > \epsilon/4\right\} =$$

$$2P\left\{n^\alpha (\log_2 n)^\delta \sup_{x\in\mathcal{X}} \left| n^{-1}\sum_{i=1}^n \sigma_i \mathbb{1}_{\{X_i=x\}}\right| > \epsilon/4\right\}. \tag{4.5}$$

En utilisant (4.4) et (4.5), nous obtenons pour tout $n \geq n_0$

$$P\left\{n^\alpha (\log_2 n)^\delta \sup_{x\in\mathcal{X}} \left|\widehat{P}_n(x) - P(x)\right| > \epsilon\right\} \leq$$

$$4P\left\{n^\alpha (\log_2 n)^\delta \sup_{x\in\mathcal{X}} \left| n^{-1}\sum_{i=1}^n \sigma_i \mathbb{1}_{\{X_i=x\}}\right| > \epsilon/4\right\}. \tag{4.6}$$

Nous allons majorer le deuxième terme de l'inégalité (4.6) conditionnellement aux observations $X_1, \ldots, X_n$. Soient $x_1, \ldots, x_k$ les k valeurs différentes de $X_1, \ldots, X_n$, (évidemment $1 \leq k \leq n$). On obtient

$$P\left\{n^\alpha (\log_2 n)^\delta \sup_{x\in\mathcal{X}} \left| n^{-1}\sum_{i=1}^n \sigma_i \mathbb{1}_{\{X_i=x\}}\right| > \epsilon/4 \,\middle/\, X_1, \ldots, X_n\right\} \leq$$

$$\sum_{j=1}^k P\left\{n^\alpha (\log_2 n)^\delta \left| n^{-1}\sum_{i=1}^n \sigma_i \mathbb{1}_{\{X_i=x_j\}}\right| > \epsilon/4 \,\middle/\, X_1, \ldots, X_n\right\} \leq$$

$$n \max_j P\left\{n^\alpha (\log_2 n)^\delta \left| n^{-1}\sum_{i=1}^n \sigma_i \mathbb{1}_{\{X_i=x_j\}}\right| > \epsilon/4 \,\middle/\, X_1, \ldots, X_n\right\}. \tag{4.7}$$

Etape 3

Pour majorer le terme de droite de l'inégalité (4.7), nous allons appliquer l'inégalité de Hoeffding (voir e.g. Pollard (1984) p.191)

Lemme 4.2 (Inégalité de Hoeffding). *Soient $Y_1, Y_2, \ldots, Y_n$ n variables aléatoires indépendantes, centrées et bornées (pour tout i, il existe a_i, b_i dans $\mathbb{R}$; telsque $a_i \leq Y_i \leq b_i$). Alors on a*

$$P\left\{|Y_1 + \cdots + Y_n| \geq \eta\right\} \leq 2\exp\left[-\frac{2\eta^2}{\sum_{i=1}^n (b_i - a_i)^2}\right]$$

Par application de ce Lemme pour $Y_i = \sigma_i \mathbb{1}_{\{X_i = x\}}$, $a_i = -1$ et $b_i = 1$, nous obtenons

$$P\left\{ n^{\alpha}(\log_2 n)^{\delta} \left| n^{-1} \sum_{i=1}^n \sigma_i \mathbb{1}_{\{X_i = x\}} \right| > \epsilon/4 \,\middle/\, X_1, \ldots, X_n \right\} =$$

$$P\left\{ \left| \sum_{i=1}^n \sigma_i \mathbb{1}_{\{X_i = x\}} \right| > n^{(1-\alpha)} (\log_2 n)^{-\delta} \epsilon/4 \,\middle/\, X_1, \ldots, X_n \right\} \leq$$

$$2\exp\left[-\frac{n^{(1-2\alpha)}(\log_2 n)^{-2\delta}\epsilon^2}{32}\right].$$

D'après (4.7), on obtient donc

$$P\left\{ n^{\alpha}(\log_2 n)^{\delta} \sup_{x \in \mathcal{X}} \left| n^{-1} \sum_{i=1}^n \sigma_i \mathbb{1}_{\{X_i = x\}} \right| > \epsilon/4 \,\middle/\, X_1, \ldots, X_n \right\} \leq$$

$$2n\exp\left[-\frac{\epsilon^2}{32} n^{(1-2\alpha)} (\log_2 n)^{-2\delta}\right]. \tag{4.8}$$

Le dernier terme ne dépend pas du conditionnement. En utilisant (4.6), (4.7) et (4.8), on obtient pour tout $n \geq n_0$

$$P\left\{ n^{\alpha}(\log_2 n)^{\delta} \sup_{x \in \mathcal{X}} |\widehat{p}_n(x) - p(x)| > \epsilon \right\} \leq 8n\exp\left[-\frac{\epsilon^2}{32} n^{(1-2\alpha)} (\log_2 n)^{-2\delta}\right] \tag{4.9}$$

La série de terme général $U_n := 8n\exp\left\{-(\epsilon^2/32) n^{(1-2\alpha)} (\log_2 n)^{-2\delta}\right\}$ est convergente (elle est majorée par une série de Rieman), d'où

$$\sum_{n=1}^{+\infty} P\left\{ n^{\alpha}(\log_2 n)^{\delta} \sup_{x \in \mathcal{X}} |\widehat{p}_n(x) - p(x)| > \epsilon \right\} < \infty.$$

Le lemme de Borel-Cantelli permet de conclure. Ce qui achève la démonstration.

4.3.2 Démonstration de la Proposition 4.2

La démonstration est semblable à celle de la Proposition 4.1.

4.3.3 Démonstration de la Proposition 4.3

Par simple application du Théorème des accroissements finis, on obtient

$$n^{\gamma}(\log_2 n)^{\delta} \sup_{x\in A_n} |\log_2 \widehat{p}_n(x) - \log_2 p(x)| =$$

$$\frac{n^{\gamma}}{\log 2}(\log_2 n)^{\delta} \sup_{x\in A_n} |p(x) + \theta_{n,x}\left(\widehat{p}_n(x) - p(x)\right)|^{-1} |\widehat{p}_n(x) - p(x)| \leq$$

$$\bigcirc\left(n^{\gamma+\beta}\right)(\log_2 n)^{\delta} \sup_{x\in\mathcal{X}} |\widehat{p}_n(x) - p(x)|\,.$$

La Proposition 4.1 permet de conclure.

4.3.4 Démonstration du Théorème 4.1

Définissons la suite $H_n := n^{-1}\sum_{i=1}^{n} -\log_2 p(X_i)$. Par application de l'inégalité de Hoeffding (voir e.g. Pollard (1984) p.191) sous l'hypothèse (H.1), on démontre la convergence preque sûre de $n^{\gamma}(\log_2 n)^{\delta}|H_n - H|$ vers 0. En effet l'hypothèse (H.1) implique $\sum_{n=1}^{\infty} P\left\{\bigcup_{i=1}^{n}\left\{p(X_i) < n^{-\beta'}\right\}\right\} < \infty$. On en déduit par le Lemme de Borel-Cantelli que l'événement $\bigcap_{i=1}^{n}\left\{p(X_i) \geq n^{-\beta'}\right\}$ a lieu presque sûrement pour n assez grand. Les variables aléatoires $Y_i := -\log_2 p(X_i) - H$ sont donc bornées : $-H \leq Y_i \leq \beta' \log_2 n - H$. Par application de l'inégalité de Hoeffding, nous obtenons donc

$$P\left\{n^{\gamma}(\log_2 n)^{\delta}|H_n - H| > \epsilon\right\} := P\left\{|Y_1 + \cdots + Y_n| > n^{1-\gamma}(\log_2 n)^{-\delta}\epsilon\right\}$$

$$\leq 2\exp\left[-\frac{2\epsilon^2 n^{1-2\gamma}}{\beta'^2 \log_2^{2\delta+2} n}\right].$$

La série de terme général $2\exp\left[-2\epsilon^2 n^{1-2\gamma}/\beta'^2 \log_2^{2\delta+2} n\right]$ est convergente pour tout $\epsilon > 0$, et le lemme de Borel-Cantelli implique la convergence presque sûre de $n^{\gamma}(\log_2 n)^{\delta}|H_n - H|$ vers 0. La suite de la démonstration consiste à montrer la convergence presque sûrement de $n^{\gamma}(\log_2 n)^{\delta}\left|\widehat{H}_n - H_n\right|$ vers 0. En utilisant des majorations similaires à celles utilisées dans Guerre (1993) p.86, on obtient

$$n^{\gamma}(\log_2 n)^{\delta}\left|\widehat{H}_n - H_n\right| = n^{\gamma}(\log_2 n)^{\delta}\left|\frac{1}{n}\sum_{i=1}^{n} \log_2 \widehat{p}_n(X_i) - \log_2 p(X_i)\right| \leq$$

$$n^{\gamma}(\log_2 n)^{\delta}\frac{1}{n}\sum_{i=1}^{n}|\log_2 \widehat{p}_n(X_i) - \log_2 p(X_i)| =$$

$$n^{\gamma}(\log_2 n)^{\delta}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i\in A_n}|\log_2 \widehat{p}_n(X_i) - \log_2 p(X_i)| +$$

$$n^{\gamma}(\log_2 n)^{\delta}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i\in A_n^c}|\log_2 \widehat{p}_n(X_i) - \log_2 p(X_i)| := A + B.$$

Le terme A est majoré par $n^{\gamma}(\log_2 n)^{\delta}\sup_{x\in A_n}|\log_2 \widehat{p}_n(x) - \log_2 p(x)|$. D'après la proposition 4.3, ce dernier est négligeable, par conséquent le terme A tend vers 0 presque sûrement. Pour le terme B, on a

$$B \leq n^{\gamma}(\log_2 n)^{\delta}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\left\{p(X_i)<Cn^{-\beta}\right\}}|\log_2 \widehat{p}_n(X_i)| +$$

$$n^{\gamma}(\log_2 n)^{\delta}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\left\{p(X_i)<Cn^{-\beta}\right\}}|\log_2 p(X_i)| := C + D.$$

Or

$$C \leq n^{\gamma}(\log_2 n)^{1+\delta}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\left\{p(X_i)<Cn^{-\beta}\right\}}.$$

D'après la proposition 4.2, sous la condition (H.2), le terme C tend donc vers 0 presque sûrement. Le terme D lui aussi tend vers 0 presque sûrement. En effet

$$\begin{aligned} D &= n^{\gamma}(\log_2 n)^{\delta}\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbb{1}_{\left\{n^{-\beta'}\leq p(X_i)<Cn^{-\beta}\right\}}|\log_2 p(X_i)| \\ &\leq \quad n^{\gamma}\beta'(\log_2 n)^{1+\delta}\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbb{1}_{\left\{p(X_i)<Cn^{-\beta}\right\}}. \end{aligned}$$

D'après la proposition 4.2, sous la condition (H.2), le dernier terme tend vers 0 presque sûrement. Ceci achève la démonstration.

4.3.5 Démonstration du Théorème 4.2

Démonstration

Pour démontrer le théorème 4.2, il suffit de démontrer la convergence presque sûre uniforme de $|\widehat{p}_n(x) - p(x)|$ vers 0 pour les deux cas (a) et (b) (voir la démonstartion du théorème 4.1). Sans perte de généralité on suppose que l'espace $\mathcal{X}$ est l'ensemble $\mathbb{N}$. On a

$$\forall\epsilon > 0, \exists x_0(\epsilon) > 0, \forall x \geq x_0 : P\{X \geq x_0\} < \epsilon/8. \tag{4.10}$$

D'autre part, d'après la loi des grands nombres, on a

$$\exists n_0(\epsilon) > 0, \forall n \geq n_0 : \left| \widehat{P}_n \{X \geq x_0\} - P \{X \geq x_0\} \right| < \epsilon/8,$$

d'où

$$\exists n_0(\epsilon) > 0, \forall n \geq n_0 : \widehat{P}_n \{X \geq x_0\} < \epsilon/4. \tag{4.11}$$

En utilisant l'inégalité

$$\sup_{x \in \mathcal{X}} |\widehat{p}_n(x) - p(x)| \leq \sup_{x < x_0} |\widehat{p}_n(x) - p(x)| + \sup_{x \geq x_0} |\widehat{p}_n(x) - p(x)|,$$

(4.10) et (4.11), nous obtenons

$$\forall n \geq n_0 : \sup_{x \in \mathcal{X}} |\widehat{p}_n(x) - p(x)| \leq \sup_{x < x_0} |\widehat{p}_n(x) - p(x)| + \epsilon/2.$$

Le cas (a) : $\sup_{x<x_0} |\widehat{p}_n(x) - p(x)|$ converge vers 0 presque sûrement car les variables $\delta_{X_1}(x) - p(x), \ldots, \delta_{X_n}(x) - p(x)$ sont stationnaires et ergodiques si $X_1, \ldots, X_n$ le sont.

Le cas (b) : Sous les hypothèses de (b) on a convergence presque sûre de $\sup_{x<x_0} |\widehat{p}_n(x) - p(x)|$ vers 0 (voir Rio (2000) p.55). Ce qui finit la démonstration.

Liste des tableaux

1.1 Confidence interval for various ϕ_γ^s-divergences 42
1.2 Case 1 : $n = 50$, $\theta_n^{(1)} = (1.94, 5.11)$, $\theta_n^{(2)} = (5.17, 2.05)$. 44
1.3 Case 2 : $n = 50$, $\theta_n^{(1)} = (4.86, 4.01)$, $\theta_n^{(2)} = (3.94, 4.76)$. 45

2.1 Estimation in Example 2. 76

3.1 Estimated mean and variance of the estimates of θ_0 in Example 1.a. . 125
3.2 Estimated mean and variance of the estimates of θ_0 in Example 1.b. . 126
3.3 Estimated mean and variance of the estimates of θ_0 in Example 1.c. . 129
3.4 Estimated mean and variance of the estimates of θ_0 in Example 2.a. . 132
3.5 Estimated mean and variance of the estimates of θ_0 in Example 2.b. . 134

Table des figures

1.1 Divergence functions φ_γ. 39
1.2 Divergence functions ψ_γ. 40
1.3 Contour plot of the Likelihood surface in case 1. 43
1.4 Contour plot of the Likelihood surface in case 2. 43

2.1 Estimation of θ in Example 1. 73
2.2 Variance in Example 1. 74
2.3 MSE in Example 1. 75
2.4 Estimation of θ in Example 3. 77
2.5 Variance in Example 3. 78
2.6 MSE in Example 3. 79
2.7 Contour plots of estimation of θ for contaminated data. 80
2.8 Contour plots of MSE for contaminated data. 81

3.1 Estimated mean of the estimates of θ_0 in Example 1.b. 127
3.2 Estimated variance of the estimates of θ_0 in Example 1.b. 128
3.3 Estimated mean of the estimates of θ_0 in Example 1.c. 130
3.4 Estimated variance of the estimates of θ_0 in Example 1.c. 131
3.5 Estimated mean of the estimates of θ_0 in Example 2.a. 133
3.6 Estimated mean of the estimates of θ_0 in Example 2.b. 135

Bibliographie

Agresti, A. (1990). *Categorical data analysis*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

Antos, A. and Kontoyiannis, I. (2001a). Convergence properties of functional estimates for discrete distributions. *Random Structures Algorithms*, **19**(3-4), 163–193. Analysis of algorithms (Krynica Morska, 2000).

Antos, A. and Kontoyiannis, I. (2001b). Estimating the entropy of discrete distributions. *IEEE Internaional Symposium on Information Theory*, **1**, 45–51.

Azé, D. (1997). *Eléments d'analyse convexe et variationnelle*. Ellipses, Paris.

Basharin, G. P. (1959). On a statistical estimate for the entropy of a sequence of independent random variables. *Theor. Probability Appl.*, **4**, 333–336.

Basu, A. and Lindsay, B. G. (1994). Minimum disparity estimation for continuous models : efficiency, distributions and robustness. *Ann. Inst. Statist. Math.*, **46**(4), 683–705.

Belzunce, F., Guillamón, A., Navarro, J., and Ruiz, J. M. (2001). Kernel estimation of residual entropy. *Comm. Statist. Theory Methods*, **30**(7), 1243–1255.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.*, **5**(3), 445–463.

Berlinet, A. (1999). How to get central limit theorems for global errors of estimates. *Appl. Math.*, **44**(2), 81–96.

Berlinet, A., Vajda, I., and van der Meulen, E. C. (1998). About the asymptotic accuracy of Barron density estimates. *IEEE Trans. Inform. Theory*, **44**(3), 999–1009.

Bertail, P. (2003). Empirical likelihood in non parametric and semi-parametric model. *To appear in "Semiparametric model and applications", Birkauser.*

Bickel, P. J., Ritov, Y., and Wellner, J. A. (1991). Efficient estimation of linear functionals of a probability measure P with known marginal distributions. *Ann. Statist.*, **19**(3), 1316–1346.

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.

Borwein, J. M. and Lewis, A. S. (1991). Duality relationships for entropy-like minimization problems. *SIAM J. Control Optim.*, **29**(2), 325–338.

Borwein, J. M. and Lewis, A. S. (1993). Partially-finite programming in L_1 and the existence of maximum entropy estimates. *SIAM J. Optim.*, **3**(2), 248–267.

Brezis, H. (1983). *Analyse fonctionnelle*. Collection Mathématiques Appliquées pour la Maîtrise. Masson, Paris. Théorie et applications.

Broniatowski, M. (2003). Estimation through Kullback-Leibler divergence. *To appear in Mathematical Methods of Statistics*.

Broniatowski, M. and Keziou, A. (2003). Parametric estimation and testing through divergences. *Submitted to Annals of Statistics*.

Chen, J. H. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, **80**(1), 107–116.

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory*. Wiley Series in Telecommunications. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

Cox, T. F. and Ferry, G. (1991). Robust logistic discrimination. *Biometrika*, **78**(4), 841–849.

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B*, **46**(3), 440–464.

Csiszár, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **8**, 85–108.

Csiszár, I. (1967a). Information-type indices of the divergence of distributions . I, II. *Magyar Tud. Akad. Mat. Fiz. Oszt. Közl. 17 (1967), 123-149 ; ibid.*, **17**, 267–291.

Csiszár, I. (1967b). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, **2**, 299–318.

Csiszár, I. (1967c). On topology properties of f-divergences. *Studia Sci. Math. Hungar.*, **2**, 329–339.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, **3**, 146–158.

Csiszár, I. (1984). Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.*, **12**(3), 768–793.

Csiszár, I., Gamboa, F., and Gassiat, E. (1999). MEM pixel correlated solutions for generalized moment and interpolation problems. *IEEE Trans. Inform. Theory*, **45**(7), 2253–2270.

Dembo, A. and Zeitouni, O. (1998). *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics*. Springer-Verlag, New York, second edition.

Dunford, N. and Schwartz, J. (1962). *Linear Operators*. Interscience Publishers.

Eichelsbacher, P. and Schmock, U. (1997). Large deviations of products of empirical measures and U-Empirical measures in strong topologies. Technical report, Bielefeld University.

Fiorin, S. (2000). The strong consistency for maximum likelihood estimates : a proof not based on the likelihood ratio. *C. R. Acad. Sci. Paris Sér. I Math.*, **331**(9), 721–726.

Fokianos, K. (2002). Box-cox transformation for semiparametric comparison of two samples, *to appear in. Shoresh Conference Proceedings*.

Fokianos, K., Kedem, B., Qin, J., and Short, D. A. (2001). A semiparametric approach to the one-way layout. *Technometrics*, **43**(1), 56–65.

Gamboa, F. and Gassiat, E. (1997). Bayesian methods and maximum entropy for ill-posed inverse problems. *Ann. Statist.*, **25**(1), 328–350.

Gänssler, P. (1971). Compactness and sequential compactness in spaces of measures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, **17**, 124–146.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, **31**, 1208–1211.

Godambe, V. P. and Heyde, C. C. (1987). Quasi-likelihood and optimal estimation. *Internat. Statist. Rev.*, **55**(3), 231–244.

Godambe, V. P. and Thompson, M. E. (1989). An extension of quasi-likelihood estimation. *J. Statist. Plann. Inference*, **22**(2), 137–172. With discussion and a reply by the authors.

Groeneboom, P., Oosterhoff, J., and Ruymgaart, F. H. (1979). Large deviation theorems for empirical probability measures. *Ann. Probab.*, **7**(4), 553–586.

Guerre, E. (1993). *Méthodes non paramétriques d'analyse des séries temporelles multivariées : estimation de mesures de dépendance*. Monographs on Statistics and Applied Probability. Doc. d'univ. de Paris 6 : Math.

Györfi, L. and Vajda, I. (2002). Asymptotic distributions for goodness-of-fit statistics in a sequence of multinomial models. *Statist. Probab. Lett.*, **56**(1), 57–67.

Györfi, L., Liese, F., Vajda, I., and van der Meulen, E. C. (1998). Distribution estimates consistent in χ^2-divergence. *Statistics*, **32**(1), 31–57.

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Ann. Statist.*, **12**(3), 971–988.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric statistical methods*. Wiley Series in Probability and Statistics : Texts and References Section. John Wiley & Sons Inc., New York, second edition. A Wiley-Interscience Publication.

Hosmer, Jr., D. W. and Lemeshow, S. (1999). *Applied survival analysis*. Wiley Series in Probability and Statistics : Texts and References Section. John Wiley & Sons Inc., New York. Regression modeling of time to event data, A Wiley-Interscience Publication.

Hosmer, Jr., D. W. and Lemeshow, S. (2000). *Applied logistic regression*. Wiley Series in Probability and Statistics : Texts and References Section. John Wiley & Sons Inc., New York.

Jiménez, R. and Shao, Y. (2001). On robustness and efficiency of minimum divergence estimators. *Test*, **10**(2), 241–248.

Kay, R. and Little, S. (1987). Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika*, **74**(3), 495–501.

Keziou, A. (2002). Sur l'estimation de l'entropie des lois à support dénombrable. *C. R. Math. Acad. Sci. Paris*, **335**(9), 763–766.

Keziou, A. (2003). Dual representation of ϕ-divergences and applications. *C. R. Math. Acad. Sci. Paris*, **336**(10), 857–862.

Kuk, A. Y. C. and Mak, T. K. (1989). Median estimation in the presence of auxiliary information. *J. Roy. Statist. Soc. Ser. B*, **51**(2), 261–269.

Landaburu, E. and Pardo, L. (2000). Goodness of fit tests with weights in the classes based on (h, ϕ)-divergences. *Kybernetika (Prague)*, **36**(5), 589–602.

Lehmann, E. L. (1986). *Testing statistical hypotheses*. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, second edition.

Léonard, C. (2001a). Convex conjugates of integral functionals. *Acta Math. Hungar.*, **93**(4), 253–280.

Léonard, C. (2001b). Minimizers of energy functionals. *Acta Math. Hungar.*, **93**(4), 281–325.

Liese, F. and Vajda, I. (1987). *Convex statistical distances*, volume 95 of *Teubner-Texte zur Mathematik*. BSB B. G. Teubner Verlagsgesellschaft, Leipzig.

Lindsay, B. G. (1994). Efficiency versus robustness : the case for minimum Hellinger distance and related methods. *Ann. Statist.*, **22**(2), 1081–1114.

Luenberger, D. G. (1969). *Optimization by vector space methods*. John Wiley & Sons Inc., New York.

McCullagh, P. and Nelder, J. A. (1983). *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Meyer, P.-A. (1966). *Probabilités et potentiel*. Publications de l'Institut de Mathématique de l'Université de Strasbourg, No. XIV. Actualités Scientifiques et Industrielles, No. 1318. Hermann, Paris.

Morales, D., Pardo, L., and Vajda, I. (1995). Asymptotic divergence of estimates of discrete distributions. *J. Statist. Plann. Inference*, **48**(3), 347–369.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Phil. Trans. Roy. Soc. Ser.*, **A**(236), 333–380.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.*, **18**(1), 90–120.

Owen, A. (1991). Empirical likelihood for linear models. *Ann. Statist.*, **19**(4), 1725–1747.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**(2), 237–249.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall, New York.

Pollard, D. (1984). *Convergence of stochastic processes*. Springer Series in Statistics. Springer-Verlag, New York.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, **85**(3), 619–630.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**(1), 300–325.

Randles, R. H. and Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics*. John Wiley & Sons, New York-Chichester-Brisbane. Wiley Series in Probability and Mathematical Statistics.

Rao, C. R. (1961). Asymptotic efficiency and limiting information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 531–545. Univ. California Press, Berkeley, Calif.

Rio, E. (2000). *Théorie asymptotique des processus aléatoires faiblement dépendants*. Springer-Verlag.

Rockafellar, R. T. (1968). Integrals which are convex functionals. *Pacific J. Math.*, **24**, 525–539.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J.

Rockafellar, R. T. (1971). Integrals which are convex functionals. II. *Pacific J. Math.*, **39**, 439–469.

Rüschendorf, L. (1984). On the minimum discrimination information theorem. *Statist. Decisions*, (suppl. 1), 263–283. Recent results in estimation theory and related topics.

Rüschendorf, L. (1987). Projections of probability measures. *Statistics*, **18**(1), 123–129.

Sen, P. K. and Singer, J. M. (1993). *Large sample methods in statistics*. Chapman & Hall, New York.

Sheehy, A. (1987). Kullback-Leibler constrained estimation of probability measures. *Report, Dept. Statistics, Stanford Univ.*

Small, C. G., Wang, J., and Yang, Z. (2000). Eliminating multiple root problems in estimation. *Statist. Sci.*, **15**(4), 313–341. With comments by John J. Hanfelt, C. C. Heyde and Bing Li, and a rejoinder by the authors.

Takagi, Y. (1998). A new criterion of confidence set estimation : improvement of the Neyman shortness. *J. Statist. Plann. Inference*, **69**(2), 329–338.

Tsybakov, A. B. and van der Meulen, E. C. (1996). Root-n consistent estimators of entropy for densities with unbounded support. *Scand. J. Statist.*, **23**(1), 75–83.

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

Verdú, S. (1998). Fifty years of Shannon theory. *IEEE Trans. Inform. Theory*, **44**(6), 2057–2078. Information theory : 1948–1998.

Wilks, S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, **9**, 60–62.

Yang, E. and Jia, Y. (2000). Universal lossless coding of sources with large or unbounded alphabets. *Numbers, Information and Complexity, (Ingo Althfor, et al, eds.), Kluwer Academic Publishers*, pages 421–442.