Dimension reduction in regression

François Portier

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de
DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Mathématiques et applications
Ecole doctorale Matisse

présentée par

François Portier

préparée à l'IRMAR - UMR CNRS 6625
Institut de recherche mathématiques de Rennes
U.F.R. de mathématiques

Réduction de la dimension en régression

Thèse soutenue à Rennes
le 2 juillet 2013

devant le jury composé de :

Bernard BERCU
Professeur à l'université de Bordeaux 1 / examinateur

Patrice BERTAIL
Professeur à l'université de Paris 10 / examinateur

Bernard DELYON
Professeur à l'université de Rennes 1 / directeur de thèse

Stéphane GIRARD
Chargé de recherche à l'INRIA Rhône-Alpes/ rapporteur

Ingrid VAN KEILEGOM
Professeur à l'université catholique de Louvain / examinateur

Céline VIAL
Maitre de conférence à l'université de Lyon / examinateur

après décision de :

Stéphane GIRARD
Chargé de recherche à l'INRIA Rhône-Alpes / rapporteur

Vladimir SPOKOINY
Professeur à l'université Humboldt de Berlin / rapporteur

Remerciements :

Je tiens tout d'abord à exprimer ma plus profonde reconnaissance à Bernard Delyon qui a été un excellent directeur de thèse. Je le remercie pour la confiance qu'il a bien voulu m'accorder durant ces trois ans, pour la patience dont il a fait preuve, et pour le soutien intellectuel qu'il m'a toujours apporté. Au travers de discussions variées, de nombreux calculs et dessins aux tableaux, il a su me transmettre sa passion pour la statistique mathématique. Travailler à ses côtés a été un grand plaisir.

Je remercie vivement Stéphane Girard et Vladimir Spokoiny de m'avoir fait l'honneur de rapporter sur cette thèse ainsi que pour l'intérêt qu'ils ont porté à mon travail.

Mes remerciements s'adressent également à Bernard Bercu, Patrice Bertail, Ingrid Van Keilegom, et Céline Vial pour avoir accepté de prendre part à mon jury. Je remercie particulièrement Ingrid Van Keilegom avec qui je suis très heureux de travailler l'année prochaine.

J'aimerais adresser ma reconnaissance à tous les membres des équipes de recherche de probabilités et de statistique. Je remercie en particulier Lionel Truquet pour ses conseils (concernant notamment mon travail sur le bootstrap) et pour son accompagnement dans mes missions d'enseignement à ses côtés. Je remercie les doctorants Samuel, Mathieu, Quentin, Julie et Damien avec qui il a toujours été un plaisir d'échanger sur différents thèmes de la statistique.

Je souhaite également remercier certains doctorants qui ont été à mes côtés. Les anciens, Jean-Louis, Yoann, Arnaud, Basile, Mathieu et Gaël qui m'ont transmis le savoir, ainsi que les plus jeunes, Renan et Pierre-Yves qui sont venus faire leurs premiers pas dans mon bureau. Merci à Cyrille pour sa présence très agréable le midi. Merci à Charles pour sa gentillesse. Merci à Baptiste pour son aide, sa sympathie, son amitié.

Par ailleurs, j'adresse mes vifs remerciements à l'ensemble des équipes administratives de l'IRMAR, de l'UFR de mathématiques et de l'école doctorale MATISSE pour leur aide constante durant la préparation de ma thèse. Je remercie également l'équipe informatique et le personnel de la bibliothèque.

Je tiens maintenant à remercier mes amis en commençant par Rémi et Nadine qui m'ont toujours soutenu. Je remercie très chaleureusement Benjamin, Thibaut, Dany, Bruno, Clarisse, Jean-Maxime, Anne-Claire, Aurore, Julien, Lise, Nicolas, Tchang, Marion, Romain, David, Corentin, et Marie notamment pour leur présence aujourd'hui, même si, pour certains, ce n'est que par la pensée (cela ne fait aucun doute).

Je remercie Franck et Elisabeth pour leurs encouragements continuels ainsi que Roger

et Marie-André.

Je remercie mes parents, Cécile et Philippe, pour m'avoir offert un cadre idéal à la réalisation de mes études, leurs conseils bienveillants et leur bonne humeur. Je remercie mon frère, Jean-Julien, pour les nombreuses distractions qu'il a su me proposer.

Enfin, je remercie Emeline pour toute son aide et pour sa présence chaque jour.

Résumé : Dans cette thèse, nous étudions le problème de réduction de la dimension dans le cadre du modèle de régression suivant

$$Y = g(\beta^T X, e),$$

où $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, la fonction g est inconnue et le bruit e est indépendant de X. Nous nous intéressons à l'estimation de la matrice $\beta \in \mathbb{R}^{p\times d}$ où $d \leq p$ (dont la connaissance permet d'obtenir de bonnes vitesses de convergence pour l'estimation de g). Ce problème est traité en utilisant deux approches distinctes. La première, appelée *régression inverse* nécessite la *condition de linéarité* sur X. La seconde, appelée *semi-paramétrique* ne requiert pas une telle condition mais seulement que X possède une densité lisse.

Dans le cadre de la régression inverse, nous étudions deux familles de méthodes respectivement basées sur $\mathbb{E}[X\psi(Y)]$ et $\mathbb{E}[XX^T\psi(Y)]$. Pour chacune de ces familles, nous obtenons les conditions sur ψ permettant une estimation exhaustive de β, aussi nous calculons la fonction ψ optimale par minimisation de la variance asymptotique. De plus, pour l'estimation de $\mathbb{E}[X\mathbb{1}_{\{Y\leq\cdot\}}]$, nous démontrons la convergence de notre estimateur à vitesse $\sqrt{n}$ vers un processus Gaussien.

Dans le cadre de l'approche semi-paramétrique, nous proposons une méthode basée sur le gradient de la fonction de régression. Sous des hypothèses semi-paramétriques classiques, nous montrons la convergence faible de notre estimateur à vitesse $\sqrt{n}$ dans l'espace des fonctions continues, nous donnons aussi les conditions d'exhaustivité de l'estimation de β.

Enfin, quel que soit l'approche considérée, une question fondamentale est soulevée : comment choisir la dimension de β ? Pour cela, nous proposons une méthode d'estimation du rang d'une matrice par test d'hypothèse bootstrap.

Mots-clés : Réduction de la dimension en régression ; Régression inverse ; Modèle à directions révélatrices ; Estimation du gradient de la régression ; Estimation de rang de matrice ; Test d'hypothèse par bootstrap.

Abstract : In this thesis, we study the problem of dimension reduction through the following regression model

$$Y = g(\beta^T X, e),$$

where $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, the function g is unknown and the noise e is independent of X. We are interested in the estimation of the matrix $\beta \in \mathbb{R}^{p\times d}$ where $d \leq p$ (whose knowledge provides good convergence rates for the estimation of g). This problem is processed according to two different approaches. The first one, called the *inverse regression*, needs the *linearity condition* on X. The second one, called *semiparametric*, do not require such an assumption but only that X has a smooth density.

In the context of inverse regression, we focus on two families of methods respectively based on $\mathbb{E}[X\psi(Y)]$ and $\mathbb{E}[XX^T\psi(Y)]$. For both families, we provide conditions on ψ

that allow an exhaustive estimation of β, and also we compute the better function ψ by minimizing the asymptotic variance. Otherwise, for the estimation of $\mathbb{E}[X\mathbb{1}_{\{Y\leq\cdot\}}]$, we provide an estimator that converges to a Gaussian process with rate $\sqrt{n}$.

In the semiparametric context, we study a method for the estimation of the gradient of the regression function. Under some classical semiparametric assumptions, we show the weak convergence of our estimator with rate $\sqrt{n}$ in the space of continuous functions, and we give conditions for the exhaustivity of the estimation of β.

Finally, within each point, an important question is raised : how to choose the dimension of β ? For this we propose a method that estimates of the rank of a matrix by bootstrap hypothesis testing.

Keywords : Sufficient dimension reduction ; Inverse regression ; Multiple index model ; Average derivative estimator ; Rank estimation ; Bootstrap hypothesis testing.

Contents

Introduction générale

Le sujet d'étude de cette thèse se place dans le cadre général de la régression statistique. Les modèles de régression consistent en l'étude de l'influence d'un jeu de variables $X \in \mathbb{R}^p$, appelées variables explicatives, sur une variable $Y \in \mathbb{R}$, dite variable à expliquer. Quitte à augmenter le nombre p de variables explicatives, on peut supposer l'existence d'une fonction $g : \mathbb{R}^p \to \mathbb{R}$ telle que $Y = g(X)$. Dans ce cas, Y est complètement déterminé par X et ce type de modélisation est qualifiée de déterministe. La modélisation probabiliste est plus ambitieuse et se résume suivant le modèle de régression additif

$$Y = g(X) + e, \tag{1}$$

où (X, Y) est désormais aléatoire, $e \in \mathbb{R}$, appelé le bruit, est une variable aléatoire indépendante de X, et $g : \mathbb{R}^p \to \mathbb{R}$, dénommée fonction de lien, est inconnue. Contrairement à la modélisation déterministe, les variables explicatives du modèle (1) ne portent pas toute l'information concernant Y, il demeure une partie inconnue e indépendante de X qui influence Y. Si Y a un moment d'ordre 2 fini, l'hypothèse d'indépendance entre l'erreur et les variables explicatives implique que la fonction de lien est égale à l'espérance conditionnelle de Y sachant X, i.e.

$$g(x) = \mathbb{E}[Y|X = x],$$

ce qui d'une part rend g unique et d'autre part identifie g à la meilleure approximation dans L^2 de Y.

L'estimation de la fonction de lien est un thème important en statistique et une approche remarquable pour y parvenir est l'estimation non-paramétrique. Cette dernière se caractérise par l'absence d'hypothèses "fortes" sur la loi des variables (X, Y) ou sur la fonction g. Elle fait ainsi opposition à une approche appelée paramétrique dans laquelle on suppose traditionnellement que g appartient à une classe de fonctions de dimension finie, par exemple le modèle linéaire $g(x) = \beta^T x$, où $\beta \in \mathbb{R}^p$. La statistique non-paramétrique quant à elle s'illustre notamment grâce aux méthodes dites à noyau, lesquelles, très largement étudiées, jouissent de belles propriétés quant à l'estimation de g. Un estimateur non-paramétrique précurseur est l'estimateur de Nadaraya-Watson (1949) que nous noterons $\widehat{g}$. Les premiers résultats obtenus, comme par exemple la convergence ponctuelle,

rassurent quant à son utilisation. En particulier, sous certaines conditions, on obtient

$$\mathbb{E}[(g(x) - \widehat{g}(x))^2] \leq \frac{C_1(x)}{nh^p} + C_2(x)h^4,$$

où h est un paramètre de l'estimateur de Nadaraya-Watson appelé la fenêtre. Déjà on entrevoit le problème du choix de la fenêtre puisque, pour assurer la convergence ponctuelle de l'erreur quadratique moyenne (MSE) de l'estimateur, il faut d'une part que $h \to 0$ et d'autre part que $nh^p \to +\infty$. Ainsi quand p est grand on impose à h de converger très lentement, ce qui dégrade les vitesses induites par la majoration précédente [1]. Cette perte en vitesse de convergence lorsque la dimension augmente est inhérente aux méthodes à noyau. Couramment dénommé le *fléau de la dimension*, ce problème a fait l'objet d'une recherche active ces dernières années. Au sein des solutions proposées, on distingue deux approches différentes : la première consiste en la sélection de variables. On citera par exemple les estimateurs de type LASSO ou encore les tests de significativité. La seconde, qui correspond au cadre de travail de cette thèse, consiste à imposer une certaine forme à la fonction g, c'est notamment le cas des modèles additifs, des modèles partiellement linéaires, ou encore des modèles à directions révélatrices. En particulier, le sujet de notre étude est résumé par le modèle

$$Y = g(\beta^T X) + e, \tag{2}$$

où $\beta \in \mathbb{R}^{p \times d_0}$ est de rang plein, $g : \mathbb{R}^{d_0} \to \mathbb{R}$ est inconnue et $e \perp\!\!\!\perp X$. En d'autres termes, le modèle (2) est équivalent au modèle (1) lorsque l'on spécifie la fonction de lien $g \circ \beta^T$. Notons que la nouvelle fonction de lien obtenue vit toujours dans un espace de dimension infinie, c'est pourquoi cette approche est souvent qualifiée de semi-paramétrique. Une telle modélisation permet d'agréger l'effet sur Y d'un groupe de p variables explicatives en le réduisant à un plus petit groupe de d_0 variables, composé de combinaisons linéaires des p variables. En d'autre termes, le modèle (2) effectue une réduction linéaire du nombre de variables tout en conservant une fonction de lien non nécessairement linéaire. Cette modélisation apparait ainsi plus flexible que la sélection de variables. Dans la littérature statistique, deux noms différents sont attribués aux modèles de type (2) : *modèle à directions révélatrices* (MIM pour *multiple index model*) ou *sufficient dimension reduction* (SDR). Dans le premier cas span(β) est appelé *l'espace index*, dans le second il est dénommé *espace central moyen*. Nous employons pour l'instant le second type de dénomination. L'espace central moyen est noté E_m. Dès à présent, on introduit une généralisation du modèle (2) dans lequel le bruit n'est pas nécessairement additif

$$Y = g(\beta^T X, e), \tag{3}$$

où $\beta \in \mathbb{R}^{p \times d_0}$ est de rang plein, $g : \mathbb{R}^{d_0+1} \to \mathbb{R}$ est inconnue, et $e \perp\!\!\!\perp X$. L'espace de réduction de la dimension associé span(β) est appelé espace central et est noté E_c [2]. Dans les

1. Plus formellement les bornes minimax présentées pour l'estimation de densité dans [58] nous montrent que le MSE ne peut converger plus vite que $n^{-r/(r+p)}$, si la densité f est r fois dérivable et $f^{(r)}$ est Lipschitzienne.

2. Comme E_c et E_m seront étudié indépendamment, la notation d_0 est utilisée pour les deux sans que leurs dimensions soient nécessairement égales.

deux modèles présentés, des vecteurs $(\beta_1, \cdots, \beta_{d_0}) = \beta$ sont appelé directions révélatrices. Les modélisations (2) et (3) sont guidées par deux enjeux majeurs :

- l'estimation de l'espace de réduction de la dimension,
- l'estimation de la fonction g.

Il est naturel d'estimer E_c ou E_m dans un premier temps car leurs estimations peuvent intervenir dans l'estimation de g. En effet sous certaines conditions, estimer g à l'aide d'une estimation préalable de E_c permet de récupérer des vitesses de convergences convenables, peu endommagées par le fléau de la dimension. Dans cette thèse ce problème ne sera pas abordé et nous nous pencherons uniquement sur l'estimation des espaces de réduction de la dimension : E_c et E_m.

Problématique et plan de la thèse

Le sujet d'étude de cette thèse s'articule autour des modèles (2) et (3) chacun autorisant une réduction de la dimension. L'objectif principal est l'estimation des espaces de réduction de la dimension. Il est tout d'abord intéressant d'effectuer une dichotomie entre le jeu d'hypothèses suffisant à l'estimation de E_c et celui suffisant à l'estimation de E_m. Étant donné la complexité des deux modèles, l'estimation de E_m demande des hypothèses plus faibles ou similaires à celles nécessaires à l'estimation de E_c. Ainsi, on peut se demander s'il est possible d'estimer ces deux espaces de façon consistante sans hypothèse "forte" sur la loi des variables explicatives. Dans le cas du modèle (3), nous répondons négativement en supposant que X est elliptique. Une telle hypothèse nous permet à la fois de travailler avec un modèle aussi général que (3) mais aussi d'obtenir des vitesses de convergence paramétrique. Dans le cas du modèle (2), on répond par l'affirmative. Néanmoins l'approche proposée soulève une nouvelle question, à savoir : parvient-on à conserver les vitesses paramétriques obtenues dans le cadre précédent ? Au sein même des deux approches décrites, la dimension d_0 des espaces de réductions de la dimension (aussi appelé dimension du modèle) est inconnue. L'estimation de ce paramètre qui, nous le verrons, se fait indépendamment de l'estimation de l'espace lui même, est un point délicat de la réduction de la dimension. En effet si ce dernier est sous-estimé alors l'estimation de l'espérance conditionnelle est relativement rapide mais le modèle n'est plus valide et certains effets ne sont pas pris en compte. Dans le cas contraire, les estimateurs convergent lentement à cause du fléau de la dimension.

Dans une première partie (chapitres 1 et 2), nous étudions le modèle (3) en choisissant de restreindre la classe des variables explicatives considérées par une hypothèse sur la loi de X. En cela nous nous plaçons dans la lignée *régression inverse* (RIV) instituée par l'article fondateur de Li en 1991 [66]. L'étude proposée met en avant les propriétés théoriques de nouvelles méthodes permettant l'estimation de E_c. Une attention particulière est accordée à l'exhaustivité de l'estimation de E_c et à la minimisation de la variance asymptotique de l'estimation. Enfin, nous proposons quelques simulations qui illustrent le bon comportements des méthodes en pratique.

La deuxième partie de la thèse (Chapitre 3) a pour objet l'estimation de la dimension dans les modèles (2) et (3). Tout comme dans [66] et l'article plus récent de Bura et Yang

de 2011 [12], nous reformulons le problème précédent comme un problème d'estimation de rang de matrice par test d'hypothèse. Pour une large classe de statistiques, nous proposons une procédure bootstrap de calcul des quantiles pour le test d'appartenance à une variété.

Enfin dans une dernière partie (Chapitre 4), nous étudions le modèle (2) sans imposer de restrictions sur les variables explicatives. Ainsi, le travail présenté est une approche semi-paramétrique pour l'estimation de l'espace central moyen dont l'articles de Härdle et Stoker de 1989 [53] et celui de Hristache, Juditsky et Spokoiny de 2001 [56] sont des références. L'estimateur proposé conserve les vitesses de convergence paramétrique quelle que soit la dimension du modèle.

Présentation des résultats

1 La Régression inverse

1.1 Existence et unicité de l'espace central et de l'espace central moyen

L'espace central E_c et l'espace central moyen E_m ont été introduits dans l'introduction de manière informelle. En particulier, ils ne sont pas uniques, par exemple tout espace contenant E_c vérifie (3). Afin de définir formellement E_c et E_m, on définit les espaces de réduction de la dimension (DRS) et les espaces moyens de réduction de la dimension (MDRS). Un DRS est un espace vectoriel tel que chacune de ses bases vérifie le modèle (3). Un MDRS est un espace vectoriel tel que chacune de ses bases vérifie le modèle (2). Notons que notre définition des MDRS diffère de celle employée habituellement dans la littérature (voir par exemple [21] et [83]). La proposition suivante donne deux caractérisations des DRS, ces dernières sont souvent utilisées à titre de définition des DRS (voir par exemple l'ouvrage de Cook de 1998 [19]).

Proposition 1. *Soit β une base de E. Les affirmations suivantes sont équivalentes :*

(i) *E est un DRS*

(ii) *Pour tout A mesurable, $\mathbb{P}(Y \in A|X) = \mathbb{P}(Y \in A|\beta^T X)$*

(iii) $Y \perp\!\!\!\perp X \mid \beta^T X$

Le dernier point signifie que Y et X sont indépendants conditionnellement à $\beta^T X$. L'équivalence entre (iii) et (i) se montre en utilisant une propriété classique de chaine de Markov (voir par exemple le livre de Benaïm et El Karoui [4] Théorème 2.4.3 page 92). L'équivalence entre (iii) et (ii) est une propriété de base de l'indépendance conditionnelle. Notons par ailleurs, qu'il existe toujours un DRS : $\mathbb{R}^p$. On peut alors définir le plus petit espace, au sens de l'inclusion, qui soit un DRS. Hélas, rien ne garantit son unicité. Par exemple, si $X = (X_1, X_2)$ et $X_1 = X_2$ p.s. alors le modèle $Y = g(X_1, e)$ possède deux DRS de dimension minimale. Une condition suffisante pour obtenir l'unicité est que l'intersection de deux DRS soit un DRS.

Définition 2. *Lorsqu'il existe un unique DRS de dimension minimale, celui-ci est appelé espace central et est noté E_c. Dans ce cas on a*

$$E_c = \bigcap E$$

où l'intersection est prise sur tous les DRS.

Contrairement au cas des DRS, il n'existe pas nécessairement de MDRS. On se place sous le modèle additif (1) afin de réaliser une démarche analogue à la précédente. Nous verrons par la suite, qu'imposer une telle structure unifie les définitions des espaces de réduction de la dimension.

Proposition 3. *Soit β une base de E. Supposons (1), les affirmations suivantes sont équivalentes :*

(i) E *est un MDRS*

(ii) $\mathbb{E}(Y|X) = \mathbb{E}(Y|\beta^T X)$

(iii) $Y \perp\!\!\!\perp \mathbb{E}[Y|X]|\beta^T X$

Notons que le point (iii) est la définition instaurée par Cook [21], communément employée dans la littérature. Comme (1) est supposé, l'équivalence entre (i) et (ii) est évidente. Pour démontrer que (iii) implique (ii), on peut vérifier que $\text{var}(\mathbb{E}[Y|X]|\beta^T X) = 0$ (voir [21] pour une preuve complète). Ainsi sous le modèle (1), il existe toujours un MDRS : $\mathbb{R}^p$. On peut aussi définir le plus petit espace, au sens de l'inclusion, qui soit un MDRS.

Définition 4. *Lorsqu'il existe un unique MDRS de dimension minimale, celui-ci est appelé espace central moyen et est noté E_m. Dans ce cas on a*

$$E_m = \bigcap E$$

où l'intersection est prise sur tous les MDRS.

Au regard des notions introduites précédemment, l'existence de E_c (resp. E_m) est équivalente à l'unicité du DRS (resp. MDRS) de dimension minimale. A propos de l'existence de E_c plusieurs résultats se trouvent dans les articles et livres de Cook [18] et [19]. Ces derniers sont résumés dans le théorème suivant.

Théorème 5 (Cook (1998) [19])**.** *Si X possède une densité dont le support est convexe, alors E_c existe.*

A notre connaissance, il n'existe pas de résultats concernant l'existence de E_m, ni de résultats à propos de l'existence de E_c lorsque le vecteur X a une distribution discrète. Dans le Chapitre 1, nous démontrons que l'hypothèse de convexité du support, supposée dans le Théorème 5, n'est pas nécessaire. Le résultat est le suivant, il est énoncé page 50.

Théorème 6 (Chapitre 1, Théorème 1.1, page 50)**.** *Si X possède une densité telle que la frontière de son support est de mesure de Lebesgue nulle, alors E_c existe. Si de plus (2) est vrai, alors le E_m existe.*

Afin de terminer la présentation des espaces de réduction de la dimension, notons que le modèle (2) est un cas particulier de (3). Ainsi un MDRS est un DRS et

$$E_c \subset E_m,$$

sous les conditions d'existence de E_c et E_m. Par ailleurs, si le modèle additif (1) est vrai, comme (iii) de la Proposition 1 implique (iii) de la Proposition 3, un DRS est un MDRS. On a donc démontré l'inclusion inverse. Ainsi si le modèle additif (1) est vrai, alors

$$E_c = E_m,$$

sous les conditions d'existence de E_c et E_m.

Dans la suite, nous supposons que E_c et E_m existent.

1.2 Caractérisation de l'espace central

On introduit la variable standardisée $Z = \Sigma^{-1/2}(X - \mathbb{E}[X])$ où $\Sigma = \text{var}(X)$ peut être supposée inversible car E_c est inclus dans son image. Le modèle (3) se réécrit avec la variable Z de la façon suivante

$$Y = g(\eta^T Z, e),$$

où $\eta = \Sigma^{1/2}\beta$, $g : \mathbb{R}^{p+1} \to \mathbb{R}$ et $e \perp\!\!\!\perp Z$. On définit donc l'espace central standardisé

$$E_c' = \text{span}(\eta) = \Sigma^{1/2} E_c.$$

On note

$$d_0 \text{ la dimension de } E_c,$$

$P_c = \eta\eta^T$, où les vecteurs $(\eta_1, \cdots, \eta_{d_0}) = \eta$ sont appelés directions révélatrices standardisées.

Les Propositions 1 et 3 sont toujours valides lorsque l'on remplace X par Z.

1.2.1 La méthode *Sliced Inverse Regression* (SIR)

L'auteur Li propose le premier d'utiliser la fonction de régression inverse $y \mapsto \mathbb{E}[Z|Y = y]$ pour caractériser E_c [66]. En cela l'article fait figure de précurseur dans la littérature RIV. On introduit maintenant une condition importante dans l'approche RIV, appelée condition de linéarité.

Condition de linéarité (LC) .

$$\mathbb{E}[Z|P_c Z] = P_c Z$$

Notons que LC ne peut être testée car cette condition dépend du modèle. On a déjà vu, dans l'introduction, que LC est vérifée pour les variables Gaussiennes. Plus généralement, les variables de loi sphérique vérifient LC (voir l'Annexe A pour plus de détails sur les lois sphériques et LC). La condition de sphéricité est atténuée dans [49] où l'on démontre que lorsque p est grand, le vecteur Z est presque sphérique. Notons que l'avantage de LC par rapport à l'hypothèse de sphéricité est que cette dernière permet l'ajout dans le vecteur Z de variables indépendantes de Y et de Z. Par exemple, le vecteur (Z, ϵ), où $Z \perp\!\!\!\perp \epsilon$ suit une loi exponentielle de moyenne 1, n'est pas sphérique alors que LC est toujours vérifiée.

Le théorème suivant, dû à Li est le fondement de la méthode SIR [66]. Une démonstration est proposée dans l'Annexe A.

Théorème 7 (Li (1991) [66])**.** *Supposons que le couple (X, Y) vérifie (3), LC et que $\mathbb{E}[\|X\|^2] < \infty$, alors on a $\mathbb{E}[Z|Y] \in E'_c$.*

On définit la matrice

$$M_{\text{SIR}} = \mathbb{E}[\mathbb{E}[Z|Y]\mathbb{E}[Z|Y]^T]. \tag{4}$$

On remarque que, d'après le théorème 7, $\text{span}(M_{\text{SIR}}) \subset E'_c$ [1]. Ainsi les vecteurs propres de M_{SIR} associés aux valeurs propres non-nulles appartiennent à E'_c.

La méthode SIR se résume à estimer M_{SIR} et à extraire les vecteurs propres de l'estimateur associés aux d plus grandes valeurs propres. Ces vecteurs propres sont des estimés des directions révélatrices standardisées.

Soient $(X_i, Y_i)_{1 \leq i \leq n}$ une suite de variables aléatoires i.i.d., définissons $\widehat{Z}_i = \widehat{\Sigma}^{-1/2}(X_i - \overline{X})$, où $\widehat{\Sigma} = \overline{(X - \overline{X})(X - \overline{X})}$ et $\overline{\cdot}$ signifie la moyenne empirique. L'estimateur de M_{SIR} proposé dans [66] est

$$\widehat{M}_{\text{SIR}} = \sum_{h=1}^{H} \widehat{p}_h \widehat{m}_h \widehat{m}_h^T,$$

1. L'espace engendré par M_{SIR} est égal à l'espace engendré par la courbe de régression inverse $\text{span}(\mathbb{E}[Z|Y = y],\ y \in \text{support}(Y))$.

où $\widehat{m}_h = \frac{\widehat{C}_h}{\widehat{p}_h}$, $\widehat{C}_h = n^{-1}\sum_{i=1}^n \widehat{Z}_i \mathbb{1}_{\{Y_i \in I(h)\}}$ et $\widehat{p}_h = n^{-1}\sum_{i=1}^n \mathbb{1}_{\{Y_i \in I(h)\}}$, et $I = \{I(h),\ h = 1, \cdots, H\}$ est une partition du support de Y.

Notons que l'on peut approcher $\mathbb{E}[Z|Y]$ par $\mathbb{E}[Z|\widetilde{Y}] = \sum_{h=1}^H m_h \mathbb{1}_{\{Y \in I(h)\}}$, où $m_h = \frac{C_h}{p_h}$, $C_h = \mathbb{E}[Z\mathbb{1}_{\{Y \in I(h)\}}]$, et $p_h = \mathbb{P}(Y \in I(h))$, et $\widetilde{Y}$ est la variable Y discrétisée selon la partition I. En définissant

$$\widetilde{M}_{\mathrm{SIR}} = \mathbb{E}[\mathbb{E}[Z|\widetilde{Y}]\mathbb{E}[Z|\widetilde{Y}]^T] = \sum_{h=1}^H p_h m_h m_h^T,$$

on remarque que $\widehat{M}_{\mathrm{SIR}}$ est un estimateur de $\widetilde{M}_{\mathrm{SIR}}$, qui lui même est une approximation de M_{SIR} [2]. On peut donc comprendre l'estimateur $\widehat{M}_{\mathrm{SIR}}$ comme une estimation par Nadaraya-Watson de la régression inverse, sans moyenne mobile.

1.2.2 Faut-il que H tende vers l'infini ?

Quelques auteurs ont étudié les propriétés de convergence de différents estimateurs de M_{SIR}. Par exemple, Zhu et Ng démontrent en 1995, sous certaines hypothèses, que $\sqrt{n}(\widehat{M}_{\mathrm{SIR}} - M_{\mathrm{SIR}})$ converge en loi à condition que $\sqrt{n} \leq H \leq n/2$ [86] (en particulier on peut avoir uniquement deux observations dans chaque tranche). Des méthodes moins naïves que l'estimateur initialement proposé dans [66] sont envisagées. La méthode *Kernel Inverse Regression* (KIR) [85] estime M_{SIR} par une méthode à noyau. Néanmoins, les travaux de [86] et [85] impliquent dans certains cas, une détérioration des vitesses en $\sqrt{n}$ causée par l'estimation non-paramétrique de $\mathbb{E}[Z|Y]$. De plus, ce type d'approches pose un nouveau problème qui est le choix du paramètre de lissage (taille des tranches pour $\widehat{M}_{\mathrm{SIR}}$, longueur de la fenêtre pour l'estimation à noyau).

Lorsque H est fixe, la matrice $\widetilde{M}_{\mathrm{SIR}}$ peut être estimée sans difficulté à vitesse $\sqrt{n}$. De plus on a la propriété

$$\mathrm{span}(\widetilde{M}_{\mathrm{SIR}}) \subset E_c'.$$

Ainsi on peut se demander dans quels sens sont les inclusions au sein des espaces $\mathrm{span}(M_{\mathrm{SIR}})$ et $\mathrm{span}(\widetilde{M}_{\mathrm{SIR}})$, où plutôt si une estimation non-paramétrique est vraiment nécessaire. Dans un second temps on peut s'interroger sur la façon d'obtenir une estimation exhaustive de E_c sans utiliser d'estimateur à noyau. C'est au sein de cette problématique que se place le travail de la première partie.

1.2.3 *Order 1 test function* (TF1)

La famille TF1 est étudiée dans le Chapitre 1. Son intérêt principal est d'offrir à l'approche RIV une famille de méthodes qui d'une part, ne requièrent pas d'estimation non-paramétrique et d'autre part, sont exhaustives au sens où elles récupèrent la totalité des espaces engendrés par la courbe de régression inverse.

2. SIR est une ACP de réalisations de la variable $\mathbb{E}[Z|\widetilde{Y}]$, et non de $\mathbb{E}[Z|Y]$ inobservable si Y possède une densité.

Les méthodes de TF1 s'intéressent aux vecteurs

$$\mathbb{E}[Z\psi(Y)],$$

où $\psi : \mathbb{R} \to \mathbb{R}$, qui sont dans E_c d'après le Théorème 7. En remarquant que $M_{\text{SIR}} = \mathbb{E}[Z\mathbb{E}[Z|Y]^T]$, on peut comprendre TF1 comme une généralisation de SIR. Un membre de la famille TF1 définit un espace

$$E^1_{\Psi_H} = \text{span}(\mathbb{E}[Z\psi_h(Y)],\ \psi_h \in \Psi_H),$$

où Ψ_H est un ensemble de fonctions de cardinal fini H. Nous obtenons le résultat suivant.

Théorème 8 (Chapitre 1, Théorème 1.3, page 52)**.** *Supposons que le couple (X, Y) vérifie le modèle (3), LC, la condition d'exhaustivité*[3] *de l'ordre 1, et que $\mathbb{E}[\|X\|^2] < \infty$. Si de plus Ψ est une famille totale dans $L_1(\|Z\|)$*[4]*, alors il existe Ψ_H un sous-ensemble fini de Ψ tel que $E^1_{\Psi_H} = E_c$.*

Une version mesurable du théorème de Weierstrass est énoncée dans le Théorème 1.B, page 83. Ce dernier nous informe qu'une famille qui sépare les points est totale dans les espaces L_p. Le résultat précédent répond à la question soulevée dans la section 1.2.2, puisque l'espace E_c peut-être retrouvé avec un nombre fini de fonctions.

1.2.4 La méthode *Sliced Average Variance Estimation* (SAVE)

La méthode SAVE est introduite afin de pallier un défaut des méthodes telles que SIR ou MD, qui en particulier, reposent sur une caractérisation de E_c par la fonction de régression inverse. Le défaut en question est souvent présenté sous le nom de *pathologie de SIR*. Un exemple qui peut lui être associé est le modèle

$$Y = Z^{(1)} + (Z^{(2)})^2 + e,$$

où $e \perp\!\!\!\perp Z = (Z^{(1)}, \cdots, Z^{(p)})$ et $(Z^{(1)}, Z^{(2)}) = (Z^{(1)}, -Z^{(2)})$. En supposant LC, on obtient que $\text{span}(M_{\text{SIR}}) \subset \text{span}(e_1)$ alors que $E'_c = \text{span}(e_1, e_2)$, où les e_i's sont les vecteurs de la base canonique de $\mathbb{R}^p$. Ainsi, la condition d'exhaustivité de l'ordre 1 supposée dans le Théorème 8 n'est plus vérifiée.

Pour remédier à ce problème, on suggère dans [66] l'utilisation de la fonction de variance inverse définie par $y \mapsto \text{var}(Z|Y = y)$. La méthode SAVE, basée sur cette même idée, est introduite et étudiée par Cook et Weisberg en 1991 [23]. Elle nécessite une condition supplémentaire appelée la condition de variance conditionnelle constante.

Condition de variance conditionnelle constante (CCV) .

$$\text{var}(Z|P_cZ) = I - P_c.$$

3. Cette condition est énoncée page 51, Condition 3. Elle est équivalente à $\text{span}(M_{\text{SIR}}) = E_c$.
4. L'espace $L_1(\|Z\|)$ est introduit page 51.

Sous CCV et LC, on a un énoncé similaire au Théorème 7 pour la fonction de variance inverse.

Théorème 9 (Cook et Weisberg (1991)[23])**.** *Supposons que LC et CCV soient vérifiées et que X possède un moment d'ordre 2, alors* $\text{span}(I - \text{var}(Z|Y)) \subset E_c'$.

La méthode SAVE caractérise l'espace central standardisé par $\text{span}(M_{\text{SAVE}})$ [5] où

$$M_{\text{SAVE}} = (I - \text{var}(Z|Y))^2.$$

L'estimation de M_{SAVE} se fait d'une manière similaire à celle de M_{SIR}. On définit $\widehat{M}_{\text{SAVE}}$ l'estimateur qui réalise un tranchage de la variance conditionnelle suivant le partition I. Sa limite, lorsque H est fixe, est notée $\widetilde{M}_{\text{SAVE}}$.

On peut soulever une problématique similaire à celle de SIR dans le choix de la limite de l'estimateur, à savoir M_{SAVE} ou $\widetilde{M}_{\text{SAVE}}$. Les auteurs Li et Zhu analysent en 2007 les conditions de convergence en loi de $\sqrt{n}(\widehat{M}_{\text{SAVE}} - M_{\text{SAVE}})$ [68]. Contrairement à SIR, le choix de H est déterminant pour la convergence à vitesse $\sqrt{n}$.

1.2.5 *Order 2 test function* (TF2)

Dans une problématique similaire à l'introduction de TF1 (i.e. faut-il estimer M_{SAVE} plutôt que $\widetilde{M}_{\text{SAVE}}$?), on définit la famille TF2. Son introduction est d'autant plus légitime que la convergence à vitesse $\sqrt{n}$ de $\widehat{M}_{\text{SAVE}}$ est sensible au choix du paramètre H. Les méthodes au sein de TF2 ont pour objet d'intérêt les matrices

$$\mathbb{E}[ZZ^T\psi(Y)],$$

où $\psi : \mathbb{R} \to \mathbb{R}$. Dans le développement de TF2, CCV est remplacé par la condition de variance conditionnelle diagonale, plus générale.

Condition de variance conditionnelle diagonale (DCV) .

$$\text{var}(Z|P_cZ) = \lambda_\omega^* Q_c$$

où λ_ω^* est une variable aléatoire.

Nous démontrons, Théorème 1.8 page 57, sous DCV et LC, que $\text{span}(\mathbb{E}[(ZZ^T - \lambda_\omega^* I)\psi(Y)]) \subset E_c'$. Ainsi un membre de la famille TF2 définit un espace

$$E^2_{\Psi_H} = \text{span}(\mathbb{E}[(ZZ^T - \lambda_\omega^* I)\psi_h(Y)],\ \psi_h \in \Psi_H).$$

Pour TF2 le résultat est un peu différent de celui présenté pour TF1, il est exprimé dans le théorème suivant.

5. La matrice M_{SAVE} estime de façon plus complète l'espace E_c que SIR. On réfère à la partie 1.4 pour plus d'informations sur ce point.

Théorème 10 (Chapitre 1, Théorème 1.9, page 58)**.** *Supposons que le couple* (X, Y) *vérifie le modèle (3), LC, DVC, la condition d'exhaustivité*[6] *de l'ordre* 2*, et que* $\mathbb{E}[\|X\|^2] < \infty$*. Si de plus* Ψ *est une famille totale dans* $L_1(\|Z\|^2)$[7]*, alors il existe* ψ *une combinaison linéaire d'un nombre fini de fonctions dans* Ψ *tel que* $E^2_\psi = E'_c$*.*

Remarquons que les méthodes de TF2 sont valides sous un jeu d'hypothèses plus léger que les méthodes telles que SAVE qui requièrent CCV[8]. De plus, nous obtenons la même conclusion que pour TF1, à savoir qu'il n'est pas nécessaire de faire tendre H vers 0 afin d'approcher M_{SAVE}. En effet, un conséquence du Théorème 10 est que, pour un nombre fini H, $E_{\Psi_H} = E'_c$.

1.2.6 Optimalité de l'estimation

Pour chacune des familles TF1 et TF2, on calcule la fonction optimale par rapport au critère[9]

$$\text{MSE} = \mathbb{E}[\text{dist}(\widehat{E}, E_c)^2],$$

où $\widehat{E}$ est l'estimé de E_c. Les fonctions optimales obtenues pour TF1 et TF2 sont données dans le Chapitre 1, respectivement page 54 et 60. Chacune des optimisations requiert le calcul de la variance asymptotique de la variable $\text{dist}(\widehat{E}, E_c)$. Pour TF1, plusieurs fonctions ψ sont nécessaires pour estimer E_c (Théorème 1.9). L'optimisation est donc conduite par rapport à la fonction $(\psi_1, \cdots, \psi_{d_0})$, sous la contrainte que $(\mathbb{E}[Z\psi_1(Y), \cdots, \mathbb{E}[Z\psi_{d_0}(Y)])$ appartienne à la variété de Stieltjes. Pour TF2, on optimise selon une seule fonction (Théorème 10). Les méthodes OF1 et OF2 associées, requièrent une estimation non-paramétrique car les fonctions optimales sont inconnues. Elles sont testées par simulation à la fin du Chapitre .

Les simulations présentées à la fin du Chapitre 1 soulignent une similitude entre SIR et OF1. En revanche, OF2 ne ressemble à aucune autre méthode considérée. On notera sa grande précision.

1.2.7 Autres méthodes

Tout comme les membres optimaux 0F1 et 0F2, la méthode *minimum discrepancy* (MD) est optimale au sein d'une famille de méthodes, nommée *inverse regression* (IR), dont SIR est un représentant. Néanmoins le critère choisi est différent.

6. Cette condition est énoncée page 58.

7. L'espace $L_1(\|Z\|)$ est introduit page 51.

8. De plus nous verrons dans le Chapitre 1, page 83, que CCV et la sphéricité des variables Z équivaut à l'hypothèse de normalité de Z. Ceci est assez contraignant puisque sous de telles hypothèses on préfère utiliser d'autres approches telles que l'approche de type vraisemblance développée dans [20].

9. La distance d'espaces vectoriels $\text{dist}(\cdot, \cdot)$ est définie par $\text{dist}(E, E') = \|P_E - P_{E'}\|_F$, où P_E (resp. $P_{E'}$) est le projecteur orthogonal sur E (resp. E') et $\|\cdot\|_F$ est la norme de Frobenius.

Les méthodes de la famille IR estiment toujours E_c à l'aide des vecteurs $\widehat{C}_{\text{SIR}} = (\widehat{p}_1^{-1/2}\widehat{C}_1, \cdots, \widehat{p}_H^{-1/2}\widehat{C}_H)$. La différence avec SIR réside dans la façon de recueillir les directions révélatrices. La méthode SIR extrait les vecteurs singuliers à gauche, associés aux plus grandes valeurs singulières de la matrice $\widehat{C}_{\text{SIR}}$. En utilisant un lemme classique sur la décomposition spectrale [10], SIR est équivalent à minimiser la distance de la matrice $\widehat{C}_{\text{SIR}}$ à la variété des matrice de rang d_0, i.e. à résoudre

$$\underset{\text{rank}(M)=d_0, M\in\mathbb{R}^{p\times H}}{\text{argmin}} \|\widehat{C}_{\text{SIR}} - M\|_F^2.$$

Toute matrice solution du problème précédent est de rang d_0 et a pour espace image $\text{span}(\widehat{M}_{\text{SIR}})$. La famille IR autorise un changement dans la norme choisie pour évaluer la distance de l'estimateur à la variété $\{\text{rank}(M) = d_0\}$. Elle est définie par la classe d'estimateurs

$$\underset{\text{rank}(M)=d_0, M\in\mathbb{R}^{p\times H}}{\text{argmin}} \text{vec}(\widehat{C}_{\text{IR}} - M)^T \widehat{V}\, \text{vec}(\widehat{C}_{\text{IR}} - M),$$

où $\widehat{C}_{\text{IR}} = \widehat{\Sigma}^{-1}(\widehat{C}_1, \cdots, \widehat{C}_H)$, $\widehat{V} \in \mathbb{R}^{pH\times pH}$. Au sein de cette famille d'estimateurs, $\widehat{C}_{\text{MD}}$ correspond au choix $\widehat{V} = \widehat{\Gamma}$, où $\widehat{\Gamma}$ est un estimateur consistant de la variance asymptotique de $\sqrt{n}\widehat{C}_{IR}$. Les auteurs démontrent que $\sqrt{n}\widehat{C}_{\text{MD}}$, défini par l'équation précédente, possède la plus petite variance au sein de la famille IR.

Citons d'autres méthodes basées sur la fonction de variance inverse qui présentent de bons résultats. La méthode SIR-II ou plus généralement SIR-α étudiée par Gannoun et Saracco en 2003 [44], consiste en un mélange des méthodes SIR et SAVE et se réalise par une pondération des deux matrices concernées. Plus récemment, les méthodes *cummulative regression* (CR) [65], et *Directinal regression* (DR) [64] ont produit de bons résultats en simulation.

1.2.8 Estimation de l'espace central moyen quand la dimension est connue

Un autre espace de réduction de la dimension est d'un intérêt particulier : l'espace moyen E_m. Les méthodes de type RIV ayant pour objectif son estimation ([67], [21]) supposent le point (iii) de la Proposition 3 : $\mathbb{E}[Y|X] = \mathbb{E}[Y|PX]$. Cette dernière hypothèse, initialement à l'origine de la définition des MDRS est particulièrement intéressante pour les modèles de type additif (1). Or dans ce cas $E_c = E_m$. En conclusion, si un modèle additif est présumé, ce qui est le cadre naturel induit par l'espace moyen, alors les méthodes d'estimation pour l'espace central s'appliquent puisque $E_c = E_m$.

Par ailleurs, nous verrons dans le Chapitre 4 comment estimer E_m dans un cadre plus général que celui imposé par les hypothèses LC et CCV. Ainsi l'approche de type RIV pour l'estimation de E_m nous parait moins justifiée que les précédentes.

10. Ce lemme est énoncé dans le Chapitre 3, Lemme 3.8, page 110.

1.3 Estimation de la dimension

L'estimation de la dimension de E_c ou E_m est fondamentale au sein des méthodes de type RIV car elle permet d'estimer le bon nombre de vecteurs singuliers à conserver lors de la décomposition spectrale de la matrice d'intérêt (par exemple C_{SIR}, C_{MD}, M_{SAVE}). Ici nous ne traitons pas cette question de manière exhaustive car le sujet sera plus largement abordé dans la section 2, introduction au Chapitre 3. Nous donnons néanmoins la procédure d'estimation la plus populaire. Détaillée par Li en 1991 [66], cette dernière est l'objet de tests de rang de la matrice d'intérêt notée dans cette partie $M \in \mathbb{R}^{p\times H}$. Comme $\text{span}(M) \subset E$, où $E = E_c$ ou E_m, il y a un léger abus de langage à parler d'estimation de la dimension. Pour tester un rang donné, on peut considérer le jeu d'hypothèses

$$H_0 : \ \text{rank}(M) = m \qquad \text{contre} \qquad H_1 : \ \text{rank}(M) > m.$$

Pour tester le rang de M, on effectue un test séquentiel qui consiste à mener plusieurs fois le test précédent jusqu'à acceptation. Plus précisément, on commence par tester $m = 0$, si ce test est rejeté on incrémente $m := m + 1$. Le test est reconduit jusqu'à la première acceptation. La statistique initialement proposée est

$$\widehat{\Lambda}_1 = n \sum_{k=m+1}^{p} \widehat{\lambda}_k^2 \tag{5}$$

où $(\widehat{\lambda}_1, \cdots, \widehat{\lambda}_p)$ sont les valeurs singulières de la matrice $\widehat{M} \in \mathbb{R}^{p\times H}$, elle-même estimateur de M. Pour beaucoup de méthodes, la statistique $\widehat{\Lambda}_1$ produit un test consistant (voir la section 2 pour plus de détails). Nous donnons le résultat associé à SIR. Ce dernier, d'abord valable dans le cas de variables explicatives gaussiennes [66] fut étendu au résultat suivant dans [15].

Proposition 11 (Li (1991) [66], Bura et Cook (2001)[15])**.** *Supposons que le couple* (X, Y) *vérifie LC et CCV et que* $\mathbb{E}[\|X\|^2] < \infty$*, alors si* $\text{rank}(M) = m$*, on a*

$$\widehat{\Lambda}_1 \xrightarrow{\text{d}} \chi^2_{(p-m)(p-H-1)}.$$

Dans le Chapitre 1, nous donnons le même type de résultat pour TF1 (Théorème 1.13, page 64). Ce dernier est une généralisation de la Proposition 11 à une famille quelconque de fonctions, avec l'hypothèse DCV.

1.4 La méthode continuous inverse regression pour l'exhaustivité de l'estimation

L'exhaustivité de l'estimation est un thème plus récemment abordé dans la littérature. On parle d'exhaustivité de la méthode A lorsque $E_A = E_c$, où E_A est l'espace à estimer. Par exemple la condition d'exhaustivité de la méthode SIR est la suivante :

$$\text{Pour tout vecteur } \eta \in E_c', \ \mathbb{E}[\eta^T Z|Y] \text{ est de variance non-nulle.} \tag{6}$$

On obtient la même condition por la méthode MD puisqu'il est facile de voir que $E_{\text{SIR}} = E_{\text{MD}}$. Néanmoins, la pathologie SIR (voir section 1.2.4), met en avant une certaine classe de modèles pour lesquels les méthodes basées sur la fonction de régression inverse (SIR, KIR, MD) ne sont pas exhaustives.

En conséquence les méthodes qui s'appuient sur la fonction de variance inverse (SAVE, SIR-II, SIR-α, CR, DR) retrouvent vraisemblablement plus de directions révélatrices que les précédentes. Dans ce sens, on remarque le résultat de Ye et Weiss en 2003 qui nous indique que $E_{\text{SIR}} \subset E_{\text{SAVE}}$ [11] [82]. En plus du résultat concernant SAVE, on peut montrer de la même manière que $E_{\text{SAVE}} = E_{\text{SIR-II}} = E_{\text{DR}}$ (voir [64]). Dans les articles [64] et [63], les auteurs s'intéressent aux conditions d'exhaustivités des méthodes CR et DR. Par exemple pour DR, ils obtiennent le résultat suivant.

Proposition 12 (Li et Wang (2007) [64])**.** *Supposons que pour tout vecteur $\eta \in E'_c$, $\mathbb{E}[\eta^T Z|Y]$ ou $\mathbb{E}[(\eta^T Z)^2|Y]$ est de variance non-nulle, alors la méthode DR est exhaustive.*

Concernant TF1 et TF2, nous obtenons le même type de conditions, adaptées à l'hypothèse DCV (voir pages 51 et 58).

Les conditions énoncées précédemment assurent que la matrice à estimer engendre l'espace central tout entier. Par exemple pour SIR (mais cela reste vrai pour la plupart des autres méthodes), la condition (6) implique que $\text{span}(M_{\text{SIR}}) = E'_c$. Malheureusement, pour estimer cette matrice, un partitionnement de la variable Y est demandé et la matrice à estimer devient $\widetilde{M}_{\text{SIR}}$ qui est une approximation de M_{SIR}. Ainsi on peut reformuler la définition d'exhaustivité de la méthode SIR par

$$\text{span}(\widetilde{M}_{\text{SIR}}) = E'_c,$$

ce qui met en jeu le partitionnement de Y utilisé. Le Théorème 8 nous indique que si H est suffisamment grand, la condition précédente est vérifiée. Néanmoins ce résultat ne nous fournit pas d'indication sur le choix de la partition en pratique. Pour y remédier, on introduit dans le chapitre 2 la méthode continuous inverse regression (CIR) qui estime l'espace

$$\text{span}(\mathbb{E}[Z\mathbb{1}_{\{Y\leq y\}}],\ y \in \mathbb{R}),$$

à l'aide de la matrice

$$M_\Phi = \int \mathbb{E}[Z\mathbb{1}_{\{Y\leq y\}}]\mathbb{E}[Z\mathbb{1}_{\{Y\leq y\}}]^T \mathrm{d}\Phi(y),$$

11. Pour le démontrer il suffit de remarquer que $M_{\text{SAVE}} = M^2_{\text{SIR}} + A$ avec A positive, ainsi $\text{span}(M_{\text{SIR}}) = \text{span}(M^2_{\text{SIR}}) \subset \text{span}(M^2_{\text{SIR}} + A) = \text{span}(M_{\text{SAVE}})$.

où Φ est une distribution de probabilité. L'estimateur considéré est

$$\widehat{M}_\Phi = n^{-2} \sum_{i,j}^{n} \widehat{Z}_i \widehat{Z}_j^T (1 - \max(\Phi(Y_i), \Phi(Y_i))).$$

Le résultat principal du chapitre 2 concerne l'estimation de M_Φ, d'une part lorsque Φ est connue, et d'autre part lorsque $\Phi = F$, la distribution de probabilité de Y (inconnue). Dans ce dernier cas, $\widehat{M}_F = n^{-2} \sum_{i,j}^{n} \widehat{Z}_i \widehat{Z}_j^T (1 - \max(\mathbb{F}(Y_i), \mathbb{F}(Y_i)))$ où $\mathbb{F}$ est la distribution de probabilité empirique de Y.

Théorème 13 (Chapitre 2, Théorème 2.8, page 95)**.** *Supposons que* $\mathbb{E}[\|X\|^2] < +\infty$*, et que* Y *possède une densité continue, alors*

$$n^{1/2}(\widehat{M}_F - M_F) \text{ converge vers un vecteur Gaussien.}$$

Si de plus, Φ *est une distribution de probabilité strictement croissante, alors*

$$n^{1/2}(\widehat{M}_\Phi - M_\Phi) \text{ converge vers un vecteur Gaussien.}$$

Le théorème précédent est une conséquence de l'étude de la convergence faible des processus définis par

$$\mathbb{W}_\Phi : u \mapsto n^{-1/2} \sum_{i=1}^{n} X_i \mathbb{1}_{\{Y_i \leq \Phi^-(u)\}} - \mathbb{E}[X \mathbb{1}_{\{Y \leq \Phi^-(u)\}}]$$

et

$$\mathbb{W}_F : u \mapsto n^{-1/2} \sum_{i=1}^{n} X_i \mathbb{1}_{\{Y_i \leq \mathbb{F}^-(u)\}} - \mathbb{E}[X \mathbb{1}_{\{Y \leq F^-(u)\}}],$$

où Φ^- est l'inverse généralisée de la fonction Φ, et les deux processus sont des éléments de l'espace des fonctions continues à gauche avec limite à droite (càd-làg) noté $D[0,1]$. L'étude de chacune de leurs convergences résulte de considérations différentes. Pour $\mathbb{W}_\Phi$, nous démontrons la convergence faible dans l'espace $D[0,1]$ muni de la norme de Skorohod. Une telle approche est décrite dans le livre de Billingsley de 1979 [8]. Pour $\mathbb{W}_F$, la technique de la Delta method employée dans la preuve nous impose de travailler avec une autre notion de convergence faible. Cette dernière est définie à l'aide de l'intégrale extérieure et fait l'objet du livre de van der Vaart et Wellner paru en 1996 [77]. Les résultats obtenus sont résumés dans le théorème suivant où nous ne différencions pas les différents types de convergence.

Théorème 14 (Chapitre 2, théorèmes 2.2 et 2.6, pages 88 et 93)**.** *Supposons que* $\mathbb{E}[\|X\|^2] < +\infty$*, et que* Y *possède une densité continue, alors*

$$\mathbb{W}_F \text{ converge vers un processus Gaussien.}$$

Si de plus, Φ *est une distribution de probabilité strictement croissante, alors*

$$\mathbb{W}_\Phi \text{ converge vers un processus Gaussien.}$$

En plus d'impliquer le Théorème 13, utile à l'estimation de l'espace central, les résultats précédents nous permettent de proposer des tests de type Cramér-von Mises ou Kolmogorov. Ces derniers font l'objet de la section 2.5.

2 Estimation de la dimension par test d'hypothèse et bootstrap

Dans un premier temps, on reformule dans un cadre plus général le problème d'estimation de la dimension de E_c qui, on l'a vu dans la partie 1.3, s'apparente à un problème d'estimation de rang de matrice. Après une revue des résultats existants en estimation de rang, nous expliquons les modalités et intérêts de l'emploi du bootstrap pour les tests d'hypothèses. Enfin nous présentons les contributions apportées à ce domaine dans la troisième partie.

2.1 Estimation du rang d'une matrice par test d'hypothèse

L'estimation du rang d'une matrice est primordiale au sein de l'approche RIV mais pas seulement. En effet beaucoup d'autres domaines nécessitent ce type de connaissance. On peut citer par exemple l'analyse en composante principale (ACP) où le nombre de facteurs du modèle égale la dimension de la matrice de covariance. Un autre exemple est le modèle auto-régressif à moyenne mobile (ARMA). En effet, les ordres d'un modèle ARMA sont égaux aux rangs de certaines matrices de Toeplitz. Pour plus d'exemple et de détails sur les applications de l'estimation de rang, on pourra lire Gill et Lewbel [45].

2.1.1 Cadre de travail de l'estimation de rang

On peut tester le rang de M_0 noté d_0 à l'aide des deux hypothèses suivantes, il existe un estimateur $\widehat{M}$ telle que

$$n^{1/2}(\widehat{M} - M_0) \xrightarrow{\mathrm{d}} W \tag{7}$$

où vec(W) est un vecteur Gaussien de moyenne nulle et de variance asymptotique Γ, et de plus il existe une matrice $\widehat{\Gamma}$ telle que

$$\widehat{\Gamma} \xrightarrow{\mathbb{P}} \Gamma. \tag{8}$$

Aussi, certaines méthodes supposent que Γ est inversible. Sous ces hypothèses, l'estimation du rang de M_0 peut se faire par test d'hypothèse, et plus précisément par plusieurs tests séquentiels, dont chacun teste un rang donné :

$$H_0 : \ \mathrm{rank}(M) = m \qquad \text{contre} \qquad H_1 : \ \mathrm{rank}(M) > m. \tag{9}$$

Ainsi, on commence par tester $d_0 = 0$, puis si le test est rejeté on test $d_0 = 1$ et ainsi de suite, jusqu'au premier test non-rejeté [12]. Bien entendu, le niveau global du test n'est pas égal au niveau de chacun des tests. Ce dernier est difficile à calculer. Dans notre étude, on se concentre sur le test (9) où m est fixé.

12. Un test différent est proposé par Barrios et Velilla en 2007 [3], il est évoqué plus en détail dans la section 2.4.

2.1.2 Quelques statistiques de test

Gill et Lewbel en 1992 [45] sont les premiers à considérer le problème d'estimation du rang de matrice dans le cadre général donné par les hypothèses (7) et (8). Leur approche consiste à examiner la décomposition $\widehat{L}\widehat{D}\widehat{U}$ de la matrice $\widehat{M}$, où $\widehat{L}$ (resp. $\widehat{U}$) est une matrice triangulaire inférieure (resp. supérieure) et $\widehat{D}$ est diagonale. La matrice $\widehat{D}$ a deux comportements distincts selon H_0 et H_1, ce qui permet aux auteurs, sous les hypothèses (7) et (8), de démontrer la consistance du test (9). Le test proposé est corrigé par Cragg et Donald en 1996 [25] et nous référons à ce dernier article pour plus d'information [13]. Au sein de cette littérature, on remarque le travail de Cragg et Donald de 1997 [26] où il est proposé une statistique de type *minimum distance*

$$\widehat{\Lambda}_3 = \min_{\text{rank}(M)=m} \text{vec}(\widehat{M} - M)^T \widehat{\Gamma}^{-1} \text{vec}(\widehat{M} - M). \tag{10}$$

Les auteurs démontrent sous les conditions (7), (8), H_0 et Γ inversible, que cette dernière converge vers une loi du χ^2 à $(p-m)(H-m)$ degrés de liberté. Ils montrent aussi que sous H_1 la statistique tend vers l'infini en probabilité. Sous les mêmes hypothèses, les articles de Robin et Smith de 2000 [72] et de Kleibergen et Paap de 2006 [61] s'intéressent au comportement de différentes transformations des valeurs singulières.

Dans la littérature RIV, l'estimation du rang a été traitée de différentes façons, souvent en lien avec les hypothèses LC et CCV, afin de préciser les variances asymptotiques. Néanmoins les tests qui y sont développés peuvent être mis en relation avec le cadre précédent. Récemment Bura et Yang adaptent la Proposition 11 au cadre précédent. Ils obtiennent le résultat suivant pour $\widehat{\Lambda}_1$, définie par (5).

Proposition 15 (Cook et Bura (2001) [15], Bura et Yang (2011) [12])**.** *Supposons (7) et H_0, alors on a*

$$\widehat{\Lambda}_1 \xrightarrow{\text{d}} \sum \nu_k W_k^2$$

où les ν_k's sont les valeurs propres de la matrice $(Q_2 \otimes Q_1)\Gamma(Q_2 \otimes Q_1)$ et les W_k's sont i.i.d. gaussiens centrés-réduits.

Ainsi $\widehat{\Lambda}_1$ s'intègre dans la littérature de l'estimation de rang parmi les statistiques basées sur la décomposition en valeurs singulières. Toujours dans [12], les auteurs proposent une version re-normalisée de $\widehat{\Lambda}_1$, définie par

$$\widehat{\Lambda}_2 = n \,\text{vec}(\widehat{Q}_1 \widehat{M} \widehat{Q}_2)^T [(\widehat{Q}_2 \otimes \widehat{Q}_1)\widehat{\Gamma}(\widehat{Q}_2 \otimes \widehat{Q}_1)]^+ \text{vec}(\widehat{Q}_1 \widehat{M} \widehat{Q}_2), \tag{11}$$

où M^+ est l'inverse de Moore-Penrose de M, $\widehat{Q}_1$ et $\widehat{Q}_2$ sont les projecteurs orthogonaux sur les sous-espaces singuliers, à gauche et à droite, de M associés aux $p-m$ plus petites valeurs singulières. Ils obtienent le résultat suivant.

13. Dans l'article en question [45], les auteurs commettent une erreur quant à la distribution asymptotique de $\widehat{D}$. Cette erreur est révélée dans [25].

Proposition 16 (Bura et Yang (2011) [12]). *Supposons (7), (8) et H_0, alors on a*

$$\widehat{\Lambda}_2 \xrightarrow{\mathrm{d}} \chi^2_s,$$

avec $s = \min(\mathrm{rank}(\Gamma), (p-d)(H-d))$.

Le test proposé par Cook et Ni en 2005 [22] pour la méthode MD (voir section 1.2.7) se base sur la statistique $\widehat{\Lambda}_3$ définie par (10) et étudiée, aussi, dans [26] [14].

Proposition 17 (Cragg et Donald (1997) [26], Cook et Ni (2005) [22]). *Supposons (7), (8),* Γ *inversible et H_0, alors on a* [15]

$$\widehat{\Lambda}_3 \xrightarrow{\mathrm{d}} \chi^2_{(p-m)(H-m)}.$$

En utilisant les propositions 15, 16 et 17 on peut montrer que chacune des statistiques $\widehat{\Lambda}_1$, $\widehat{\Lambda}_2$ et $\widehat{\Lambda}_3$ produit un test (9) consistant (14).

Les statistiques $\widehat{\Lambda}_1$, $\widehat{\Lambda}_2$ et $\widehat{\Lambda}_3$ jouent un rôle important dans la suite puisqu'elles constituent les principaux exemples d'applications de la procédure bootstrap étudiée dans le Chapitre 3.

2.1.3 Comportement des statistiques en pratique

Remarquons que pour réaliser le test avec $\widehat{\Lambda}_1$, il faut approcher sa loi limite car cette dernière est composée de quantités inconnues. Les auteurs Bentler et Xie [5] et Wood [79] proposent certaines approximations de la loi du chi2 pondérée. Les auteurs étudient trois manières d'y parvenir (voir page 115 pour plus d'information). Ces approximations de la loi asymptotique peuvent causer une perte dans la précision du test. Remarquons par ailleurs que les statistiques $\widehat{\Lambda}_2$, et $\widehat{\Lambda}_3$ n'ont pas ce problème puisque leurs lois asymptotiques sont des chi2. Néanmoins leurs calculs nécessitent l'inversion d'une matrice souvent grande ce qui produit, à faible nombre d'échantillons, de mauvaises estimations.

Considérons les modèles

$$Y = X^{(1)} + 0.5e \tag{12}$$

$$Y = \cos(2X^{(1)} - 1) + 0.2e, \tag{13}$$

où $e \perp\!\!\!\perp X \in \mathbb{R}^6$, $X \overset{\mathrm{d}}{=} \mathcal{N}(0, I)$ et $e \overset{\mathrm{d}}{=} \mathcal{N}(0, 1)$. Pour ces deux modèles, la matrice $C_{\mathrm{SIR}} = (C_1, \cdots, C_H)$ est de rang 1. On utilise son estimateur par tranchage $\widehat{C}_{\mathrm{SIR}}$, défini dans la

14. Les résultats de [22] sont similaires à ceux de [26] bien que les outils utilisés dans les preuves soient différents.

15. Notons que le degré de liberté du chi2 limite n'est pas en $(p-m)(H-m-1)$ comme énoncé dans [22], ceci est du au fait que Γ est inversible dans notre énoncé.

section 1.2.7, pour calculer les statistiques $\widehat{\Lambda}_1$, $\widehat{\Lambda}_2$, et $\widehat{\Lambda}_3$ et mener le test (9). Notons que les propositions 15, 16 et 17 s'appliquent. Pour un nombre de tranches $H = 5$ et différentes valeurs de n, on calcule les niveaux estimés avec 2000 échantillons, pour un niveau nominal de 5%. On utilise 5 statistiques : les trois approximations de [79] et [5] pour $\widehat{\Lambda}_1$, et $\widehat{\Lambda}_2$ et $\widehat{\Lambda}_3$. Les résultats des simulations sont présentés Figure 1 pour le modèle (12) et Figure 2 pour le modèle (13).

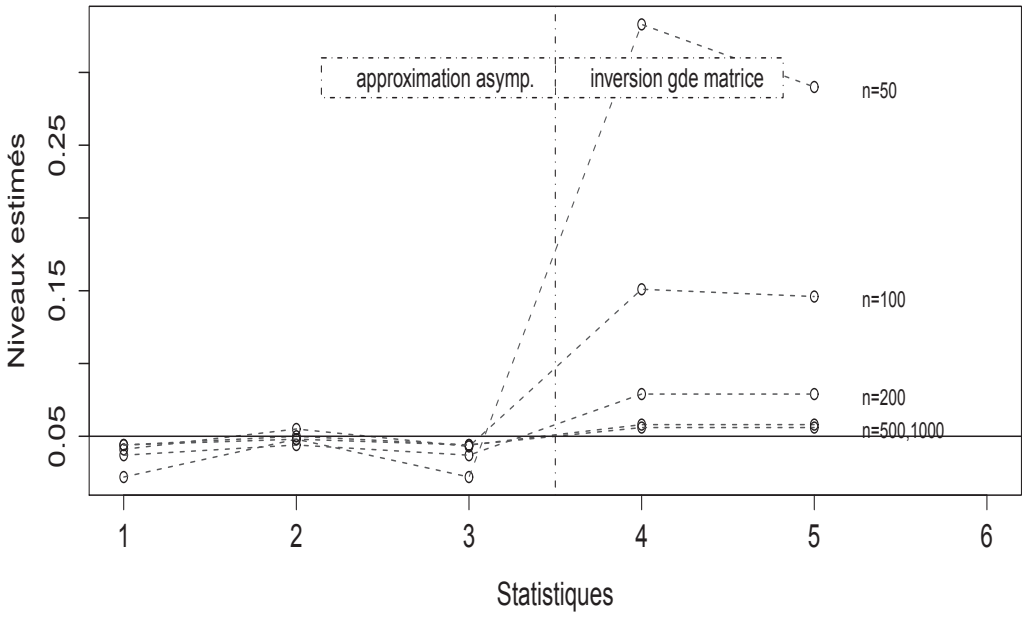

Figure 1 – Modèle (12). Niveaux estimés lors du test (9) avec $\widehat{\Lambda}_1$ (les approximations de [79] et [5] sont en abscisse 1,2,3), $\widehat{\Lambda}_2$ et $\widehat{\Lambda}_3$ (abscisse 4 et 5).

Les approximations [79] et [5] se comportent bien dans le modèle (12) qui est linéaire. Lorsque la dépendance se complique dans le modèle (13), l'approximation n'est plus valide puisque même à $n = 1000$, le niveau nominal n'est pas atteint. A faible nombre d'échantillons ($n = 50$), pour les deux modèles étudiés, aucune des statistiques $\widehat{\Lambda}_2$, $\widehat{\Lambda}_3$ ne produit un test satisfaisant. De plus la convergence est relativement lente dans le modèle linéaire (12).

Nous remarquons donc la difficulté d'estimer la dimension dans les modèles de type RIV. Ceci est d'autant plus marqué que les modèles présentés sont de petite dimension $d_0 = 1$.

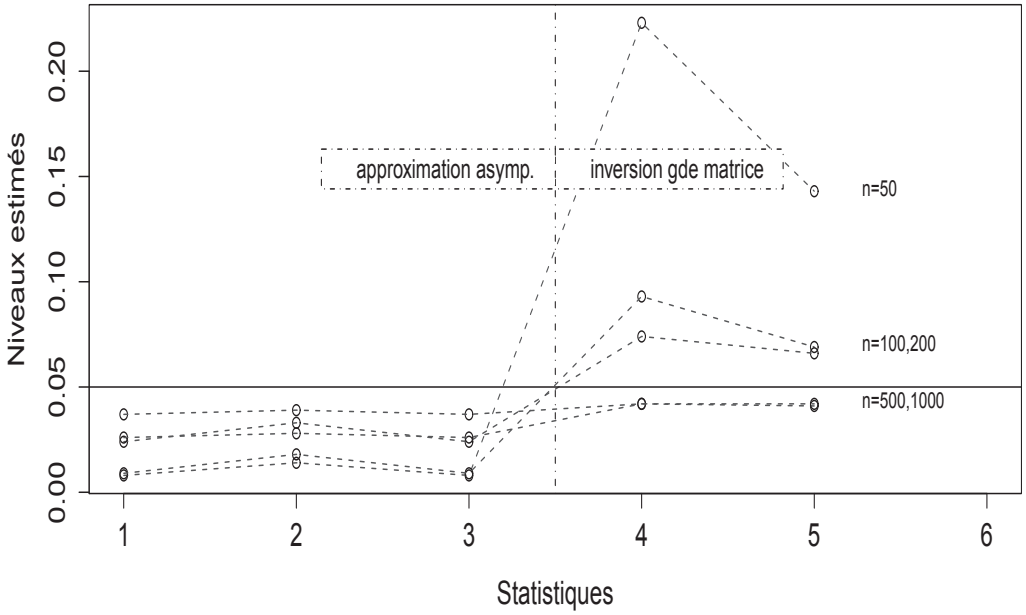

FIGURE 2 – Modèle (13). Niveaux estimés lors du test (9) avec $\widehat{\Lambda}_1$ (les approximations de [79] et [5] sont en abscisse 1,2,3), $\widehat{\Lambda}_2$ et $\widehat{\Lambda}_3$ (abscisse 4 et 5).

2.2 Test d'hypothèse traditionnel et bootstrap

Le bootstrap est une technique de ré-échantillonnage instaurée par Efron en 1982 [38]. Cette approche propose une alternative à la comparaison asymptotique, qui à faible taille d'échantillon, est une approximation souvent grossière.

Soit $(X_1, \cdots, X_n)$ un échantillon i.i.d. et $T(X_1, \cdots, X_n)$ une statistique. Efron propose de tirer un nouvel échantillon $(X_1^*, \cdots, X_n^*)$ selon une loi uniforme au sein du premier échantillon $(X_1, \cdots, X_n)$, et ensuite, d'inférer sur le comportement de la statistique $T(X_1, \cdots, X_n)$ à l'aide de la loi de la statistique bootstrap $T(X_1^*, \cdots, X_n^*)$, dénommée loi bootstrap. Nous verrons que cette loi, si elle est inconnue, peut être approchée avec une grande précision. Aussi, sous certaines conditions, la loi de la statistique est plus proche de la loi bootstrap que de la loi asymptotique. D'autres procédures de ré-échantillonnage sont proposées par Barbe et Bertail en 1995 [2]. L'un des avantages du bootstrap est donc la précision qu'il procure par rapport à la comparaison asymptotique. Le bootstrap est notamment utilisé pour estimer les moments d'une statistique, pour dé-biaiser un estimateur ou encore pour l'estimation de quantiles, lors de tests d'hypothèses et de constructions d'intervalles de confiance.

2.2.1 Le calcul des quantiles par bootstrap pour un test d'hypothèse

Afin d'expliquer comment utiliser le bootstrap lors de tests d'hypothèses, nous prenons l'exemple simple du test d'égalité de moyennes dans le cadre d'un échantillon de n variables réelles $X_1, \cdots, X_n$ i.i.d. telles que $\theta_0 = \mathbb{E}[X_1]$ et $0 < \gamma = \text{var}(X) < \infty$. Les hypothèses du test sont

$$H_0 : \ \theta_0 = \mu \qquad \text{contre} \qquad H_1 : \ \theta_0 \neq \mu,$$

et une statistique de test raisonnable est $\widehat{\Lambda} = n\|\widehat{\theta} - \mu\|^2$, où $\widehat{\theta} = \overline{X}$ puisqu'elle vérifie les deux propositions suivantes. Si H_0 est vrai, alors $\widehat{\Lambda} \xrightarrow{\text{d}} \gamma\chi^2_1$ et si H_1 est vérifiée alors $\widehat{\Lambda}$ diverge vers l'infini en probabilité. Un test d'hypothèse traditionnel de niveau α est la comparaison de $\widehat{\Lambda}$ avec le quantile d'ordre α de sa loi limite sous H_0, noté $q_\infty(\alpha)$. On dit qu'un test est consistant si

$$\mathbb{P}_{H_0}\left(\widehat{\Lambda} > q_\infty(\alpha)\right) \longrightarrow 1 - \alpha \qquad \text{et} \qquad \mathbb{P}_{H_1}\left(\widehat{\Lambda} > q_\infty(\alpha)\right) \longrightarrow 1, \tag{14}$$

ce qui est, dans notre cas, une conséquence directe des deux convergences évoquées précédemment.

Le test d'hypothèse bootstrap se différencie du test d'hypothèse traditionnel par sa méthode de calcul du quantile utilisé pour le test. Alors que le test traditionnel utilise un quantile de la loi asymptotique sous H_0, le test bootstrap utilise un quantile calculé par bootstrap.

Pour cela, on introduit ici un bootstrap par poids indépendants. Prenons $(w_i)_{1\leq i\leq n}$ une suite de variables indépendantes de moyenne 0 et de variance 1 et définissons l'estimateur bootstrap

$$W^* = n^{-1}\sum_{i=1}^{n} w_i(X_i - \overline{X}).$$

Remarquons que par le TCL de Lindeberg, conditionnellement aux X_i's, $\sqrt{n}W^*$ a la même limite en loi que $\sqrt{n}(\widehat{\theta}-\theta_0)$. On dira que $\sqrt{n}W^*$ « bootstrap » $\sqrt{n}(\widehat{\theta}-\theta_0)$. Plus précisément, $\sqrt{n}W^*$ bootstrap la loi sous H_0 de $\sqrt{n}(\widehat{\theta}-\mu)$. Définissons $\Lambda^* = n\|W^*\|^2$. Le test bootstrap est basé sur le résultat

$$\Lambda^* \text{ bootstrap la loi sous } H_0 \text{ de } \widehat{\Lambda}, \tag{15}$$

qui est une conséquence de ce qui précède. Cette convergence est équivalente à la convergence $\widehat{F}(x) \xrightarrow{\text{p.s.}} F_\infty(x)$ pour tout $x \in \mathbb{R}$, où $\widehat{F}$ et F_∞ sont respectivement les fonctions

de répartitions de Λ^*, conditionnellement aux X_i's, et de $\gamma\chi_1^2$. Comme la fonction q_∞ est continue sur $]0,1[$, on peut montrer en utilisant [76], Lemme 21.2, que pour tout $\alpha \in]0,1[$, $\widehat{q}(\alpha) \overset{\text{p.s.}}{\to} q_\infty(\alpha)$, où $\widehat{q}$ est l'inverse généralisée de $\widehat{F}$. Dès lors on obtient la consistance du test bootstrap c.a.d.

$$\mathbb{P}_{H_0}\left(\widehat{\Lambda} > \widehat{q}(\alpha)\right) \longrightarrow 1-\alpha \qquad \text{and} \qquad \mathbb{P}_{H_1}\left(\widehat{\Lambda} > \widehat{q}(\alpha)\right) \longrightarrow 1, \tag{16}$$

où la première convergence est une conséquence du théorème de Slutsky, la deuxième étant évidente. On vient de démontrer la consistance du bootstrap par poids indépendants. Pour plus d'informations, notamment pratiques, concernant le test d'égalité de moyennes par bootstrap on réfère à l'article de Hall et Wilson de 1991 [51]. Pour bien comprendre la méthodologie précédente, remarquons que la clé du test d'hypothèse par bootstrap est que la valeur de $\widehat{q}(\alpha)$ est indépendante de la réalisation de H_0. En particulier, que H_0 ou H_1 soit vraie, on a toujours (15). On rejoint maintenant un point de vue observé dans [51].

Conseil 1. *Que l'hypothèse H_0 ou H_1 soit vraie, la statistique bootstrap reproduit le comportement de la statistique sous H_0.*

En revanche, d'un point de vue théorique, (16) n'est pas suffisant pour préférer le test d'hypothèse bootstrap au test d'hypothèse traditionnel. La partie suivante présente un résultat plus précis.

2.2.2 La précision du bootstrap lorsque la statistique est pivotale

Pour démontrer (16), l'outil principal est un TCL pour variables non-identiquement distribuées. Des résultats similaires nous permettent de démontrer la consistance du test traditionnel. Ainsi, pour permettre de distinguer les deux approches, il est d'usage d'utiliser une généralisation du TCL telle que le théorème de Berry-Essein. L'ouvrage de Hall de 1992 [48] étudie le bootstrap à l'aide de développements d'Edgeworth. Ces derniers sont des développements asymptotiques de distributions de statistiques. On propose le résultat suivant, dans lequel $\theta^* = \overline{X^*}$, $\gamma^* = \overline{(X^* - \overline{X^*})^2}$ et $(X_i^*)_{1\leq i\leq n}$ est le ré-échantillonnage d'Efron.

Théorème 18 (Hall (1992) [48])**.** *Supposons que la distribution de X_1 est non-portée par un réseau et que $\mathbb{E}[X_1^2] < \infty$, alors pour tout $x \in \mathbb{R}$, on a*

$$|\mathbb{P}(n^{1/2}(\theta^* - \widehat{\theta})/\gamma^* \leq x | X_1, \cdots, X_n) - \mathbb{P}(n^{1/2}(\widehat{\theta} - \theta_0)/\widehat{\gamma} \leq x)| = O_{\mathbb{P}}(n^{-1}).$$

Rappelons que pour la distribution asymptotique, un développement d'Edgeworth nous donne

$$|\mathbb{P}(n^{1/2}(\widehat{\theta} - \theta_0)/\widehat{\gamma} \leq x) - \Phi(x)| = O(n^{-1/2}),$$

où Φ est la fonction de répartition de la loi Gaussienne centrée réduite. Nous appelons statistique pivotale une statistique dont la loi asymptotique ne dépend pas de quantités

inconnues. Par exemple la statistique $\widehat{\Lambda}$ n'est pas pivotale puisque sa loi asymptotique dépend de γ. Dans [48] on remarque que dès qu'une statistique n'est pas pivotale, alors le Théorème 18 n'a plus raison d'être. Fort de cette remarque nous énonçons le second conseil de [51].

Conseil 2. *La statistique de test est pivotale.*

Quelques résultats concernant les tests d'hypothèses bootstrap se trouvent dans l'article de Hall et Presnell de 1999 [50]. Les auteurs considèrent le test d'égalité de moyennes décrit plus haut. Ils montrent que

$$\mathbb{P}(n^{1/2}(\widehat{\theta} - \theta_0)/\widehat{\gamma} \leq \widehat{q}(\alpha)) = \alpha + O(n^{-1}),$$

où $\widehat{q}(\alpha)$ est calculé par *biased bootstrap* (voir page 105, paragraphe (ii) pour plus de details), ou par le bootstrap d'Efron.

2.2.3 Le test d'hypothèse bootstrap en pratique

On dénombre deux possibilités pour le calcul de $\widehat{q}(\alpha)$. Soit on connait la loi conditionnelle de Λ^* et donc sa fonction quantile $\widehat{q}$. Par exemple dans le cas du test précédent, si on prend les w_i's Gaussiens alors $\Lambda^* \overset{\mathrm{d}}{=} \widehat{\gamma}\chi^2_1$, où $\widehat{\gamma} = (X - \overline{X})^2$. Soit on ne connait pas a priori la loi de Λ^*, alors on peut tirer B suites de variables $(w_i)_{1\leq i\leq n}$ respectant les conditions énoncées et obtenir B versions de Λ^* indépendantes, notées Λ^*_1,...,Λ^*_B. Cet échantillon nous permet d'estimer le quantile $\widehat{q}$. De plus, comme B est arbitraire, on peut obtenir la précision que l'on souhaite lors de l'estimation de $\widehat{q}$.

2.3 Le bootstrap contraint

Dans le chapitre 3, nous proposons une méthode dénommée *bootstrap contraint* (bootstrap CS) permettant la réalisation d'un test d'hypothèse bootstrap d'appartenance à une variété. La méthodologie développée a pour application directe le test de rang (9) avec les statistiques $\widehat{\Lambda}_1$, $\widehat{\Lambda}_2$ et $\widehat{\Lambda}_3$. Nous présentons tout d'abord le problème de manière générale.

Soit $\mathcal{M}$ une variété localement lisse de codimension q, les hypothèses du test sont

$$H_0 : \ \theta_0 \in \mathcal{M} \qquad \text{contre} \qquad H_1 : \ \theta_0 \notin \mathcal{M}, \tag{17}$$

où $\theta_0 \in \mathbb{R}^p$. On note J_g la jacobienne, si elle existe, de la fonction g. Sous H_0, on définit la fonction $g : \mathbb{R}^p \to \mathbb{R}^q$, $\mathcal{C}^\infty$, telle que $V \cap \mathcal{M} = \{g = 0\}$, où V est un voisinage de θ_0, et $J_g(\theta_0)$ est de rang plein.

2.3.1 Un famille de statistiques adaptée au test d'appartenance à une variété

Soit $\widehat{\theta} \in \mathbb{R}^p$, un estimateur de θ_0. On introduit le famille $\mathcal{D}$ composée de statistiques qui évaluent de différentes manières la distance de $\widehat{\theta}$ à la variété $\mathcal{M}$. Formellement, on définit

l'estimateur contraint $\widehat{\theta}_c$ comme le point le plus proche de $\widehat{\theta}$ selon une certaine distance, qui appartient à $\mathcal{M}$. Plus précisément, posons

$$\widehat{\theta}_c = \underset{\theta \in \mathcal{M}}{\operatorname{argmin}} \ (\widehat{\theta} - \theta)^T \widehat{A} (\widehat{\theta} - \theta), \tag{18}$$

où $\widehat{A} \in \mathbb{R}^{p \times p}$. Ensuite, on définit $\widehat{\Lambda}$ qui évalue une autre distance entre $\widehat{\theta}_c$ à $\widehat{\theta}$. Prenons

$$\widehat{\Lambda} = n(\widehat{\theta} - \widehat{\theta}_c)^T \widehat{B} (\widehat{\theta} - \widehat{\theta}_c), \tag{19}$$

où $\widehat{B} \in \mathbb{R}^{p \times p}$. La famille D est l'ensemble des statistiques $\widehat{\Lambda}$. Sous certaines conditions les membres de $\mathcal{D}$ produisent un test consistant, ce qui nous donne une première indication quant au choix de $\widehat{A}$ et $\widehat{B}$.

Proposition 19 (Chapitre 3, conséquence de la Proposition 3.4, page 108)**.** *Supposons* H_0, $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Delta)$, $\widehat{B} \xrightarrow{\mathbb{P}} B$ *symétrique et* $\widehat{A} \xrightarrow{\mathbb{P}} A$ *est inversible, alors on a*

$$\widehat{\Lambda} \xrightarrow{\mathrm{d}} \sum_{k=1}^{p} \nu_k W_k^2,$$

où les W_k*'s sont des variables Gaussiennes i.i.d. et les* ν_k*'s sont les valeurs propres de la matrice*

$$\Delta^{1/2} J_g(\theta_0)^T (J_g(\theta_0) A^{-1} J_g(\theta_0)^T)^{-T} J_g(\theta_0) A^{-T} B A^{-1} \\ J_g(\theta_0)^T (J_g(\theta_0) A^{-1} J_g(\theta_0)^T)^{-1} J_g(\theta_0) \Delta^{1/2}.$$

En montrant que sous H_1, la statistique $\widehat{\Lambda}$ tend vers l'infini en probabilité, on obtient le résultat suivant.

Théorème 20. *Sous les conditions de la Proposition 19, la famille* $\mathcal{D}$ *teste (17) de façon consistante.*

La réalisation d'un tel test n'est pas simple pour autant. Listons quelques problèmes liés à sa mise en œuvre, et donnons plus d'indications concernant le choix de $\widehat{A}$ et $\widehat{B}$.

(1) En général, la loi limite n'est pas pivotale. Une première possibilité est d'estimer toutes les quantités inconnues de la loi limite et de simuler cette dernière afin d'obtenir une estimation des quantiles. Pour simplifier une telle asymptotique, on peut se placer dans la sous-classe $\widehat{A} = \widehat{B}$ symétrique et de rang plein [16]. Et afin de résoudre définitivement ce problème, on peut prendre $\widehat{A} = \Delta^{-1}$. Dans ce cas $\widehat{\Lambda}$ est pivotale et converge vers un chi2. Une autre possibilité réside dans l'approximation proposée dans [5] et [79] qui consiste à approcher la loi asymptotique de la statistique par une loi du chi2 de même moyenne ou de même variance.

16. Dans cette classe, la loi asymptotique de $\widehat{\Lambda}$ s'exprime avec la matrice $\Delta^{1/2} J_g(\theta_0)^T (J_g(\theta_0) A^{-1} J_g(\theta_0)^T)^{-1} J_g(\theta_0) \Delta^{1/2}$

(2) En prenant $\widehat{A} = \widehat{B} = \Delta^{-1}$, on évite le problème (1), mais on se trouve confronté à une deuxième difficulté. En effet, la dimension p peut être grande ce qui rend l'inversion de A difficile, et donc qui écarte la statistique de sa loi asymptotique. Dans une telle configuration, on peut par exemple, prendre $\widehat{A} = \widehat{B} = I$ afin d'éviter l'inversion d'une grande matrice.

(3) Enfin si la matrice $J_g(\theta_0)$ est inconnue, on ne peut tout simplement pas réaliser le test avec la Proposition 19.

En conséquence, l'utilisation du bootstrap est encouragée par deux arguments principaux. Tout d'abord, d'après la Section 2.2.2, le bootstrap jouit d'une grande précision lorsque la statistique est pivotale. Par exemple, si $\widehat{A} = \widehat{B} = \Delta^{-1}$, le bootstrap récupère des vitesses convenables, endommagées par l'inversion matricielle (problème (2) et Figure 1). Par ailleurs, lorsque la loi asymptotique est inconnue (problème (1) et figure 2), le bootstrap permet d'éviter certaines approximations. En particulier si on ne peut estimer la loi limite, (problème (3)) le bootstrap réalise un test consistant, alors que l'approche traditionnelle échoue.

2.3.2 Le bootstrap CS pour la famille $\mathcal{D}$

Le bootstrap CS est une procédure composée de deux étapes. Tout d'abord, soucieux de respecter le Conseil (1), on crée une première version bootstrap de $\widehat{\theta}$ qui ressemble à $\widehat{\theta}$ sous H_0. On définit

$$\theta_0^* = \widehat{\theta}_c + n^{-1/2} W^*, \qquad \text{avec} \quad \mathcal{L}_\infty(W^*|\widehat{P}) = \mathcal{L}_\infty(n^{1/2}(\widehat{\theta} - \theta_0)) \quad \text{a. s.},$$

où $\mathcal{L}_\infty$ signifie la loi asymptotique. Ainsi θ_0^* se trouve proche de la variété $\mathcal{M}$. Notons même que la distance de θ_0^* à $\mathcal{M}$ est en $O_{\mathbb{P}}(n^{-1/2})$, tout comme la distance entre $\widehat{\theta}$ et $\mathcal{M}$ lorsque H_0 est vérifiée, sous les hypothèses de la Proposition 19. Ensuite, on applique à θ_0^* les mêmes opérations qu'à $\widehat{\theta}$ afin de définir les versions bootstrap

$$\theta_c^* = \underset{\theta \in \mathcal{M}}{\operatorname{argmin}} \ (\theta_0^* - \theta)^T A^* (\theta_0^* - \theta) \qquad \text{et} \qquad \Lambda^* = n(\theta_0^* - \theta_c^*)^T B^* (\theta_0^* - \theta_c^*),$$

où $A^* \in \mathbb{R}^{p \times p}$ et $B^* \in \mathbb{R}^{p \times p}$ représentent les versions bootstrap de $\widehat{A}$ et $\widehat{B}$. En particulier on pourra prendre $A^* = \widehat{A}$ et $B^* = \widehat{B}$. Notre résultat principal concernant le bootstrap CS est le suivant. boottht est

Théorème 21 (Chapitre 3, Théorème 3.6, page 109)**.** *Supposons que* $\widehat{\theta} \overset{\text{a.s.}}{\to} \theta_0$, $\widehat{A} \overset{\mathbb{P}}{\to} A$ *est inversible,* $\widehat{B} \overset{\mathbb{P}}{\to} B$. *Si de plus,* $\mathcal{L}_\infty(\sqrt{n}(\theta_0^* - \widehat{\theta}_c)|\widehat{P}) = \mathcal{L}_\infty(\sqrt{n}(\widehat{\theta} - \theta_0))$ *p.s. possède une densité, et conditionnellement p.s.* $A^* \overset{\mathbb{P}}{\to} A$, $B^* \overset{\mathbb{P}}{\to} B$, *alors on a*

$$\mathbb{P}_{H_0}(\widehat{\Lambda} > \widehat{q}(\alpha)) \longrightarrow 1 - \alpha, \quad \textit{and} \quad \mathbb{P}_{H_1}(\widehat{\Lambda} > \widehat{q}(\alpha)) \longrightarrow 1.$$

En d'autres termes, le test (17) avec la famille $\mathcal{D}$ et l'évaluation des quantiles par le bootstrap CS est consistant.

2.3.3 Application

Le test d'appartenance à une variété (17) a pour application les tests de rang évoqués précédemment. Plus précisément, on démontre, dans le Chapitre 3, que pour tout $k \in \{1, 2, 3\}$, $\widehat{\Lambda}_k \in \mathcal{D}$. Ainsi on peut utiliser le bootstrap CS pour l'estimation du rang d'une matrice et donc pour l'estimation de la dimension dans les modèles de type RIV.

Les simulations du Chapitre 3 nous montrent le bon comportement du bootstrap CS en pratique. Dans toute les situations rencontrées, le bootstrap procure un test plus précis que le test traditionnel.

Nous utilisons le bootstrap CS pour tester la dimension du modèle (13), rencontré à la section 2.1.3. Les quantiles sont calculés avec un échantillon bootstrap de taille $B = 1000$ [17]. Les résultats des niveaux estimés sont présentées Figure 3. Cette dernière est constituée du graphique de la Figure 2 auquel on a ajouté les résultats du test bootstrap.

Pour chacune des configurations présentées, le test bootstrap l'emporte sur son homologue traditionnel. De plus, à partir de $n = 100$, n'importe quel test bootstrap est meilleur que tous les autres tests traditionnels.

2.4 Autres approches

Une autre façon d'inférer le rang d'une matrice est envisagée par Daudin, Duby et Trécourt en 1988 [28]. Cette dernière repose sur l'écart entre $\ker(M)$ et une estimation de cet espace. Suivant cette idée, un critère basé sur la distance de Frobénius est proposé par Besse en 1991 [6],

$$\widehat{R} = n \operatorname{tr}(\widehat{P}(I - P)),$$

où P et $\widehat{P}$ sont respectivement les projecteurs orthogonaux sur la somme directe des espaces propres de M_0 et $\widehat{M}$ associés au $p - m$ plus petites valeurs propres. Remarquons que si H_0 de (17) est vérifiée, alors $\widehat{R} = n \operatorname{tr}((\widehat{P} - P)(I - P)(\widehat{P} - P))$ et en utilisant les résultats de convergence des projecteurs propres (voir Annexe B, Théorème B.10), on a la convergence en loi de $\widehat{R}$ sous H_0. Sous H_1, $\widehat{R}$ n'est pas définie. Par ailleurs, la quantité $\widehat{R}$ est inconnue. En 1998, Ferré propose une estimation de l'espérance de $\widehat{R}$ [43]. Dans [6], une version bootstrap de $\widehat{R}$ est considérée. On définit

$$R^* = n \operatorname{tr}(P^*(I - \widehat{P})),$$

17. En pratique, à partir de $B = 200$, le nombre d'échantillon bootstrap n'influence pas la précision du test.

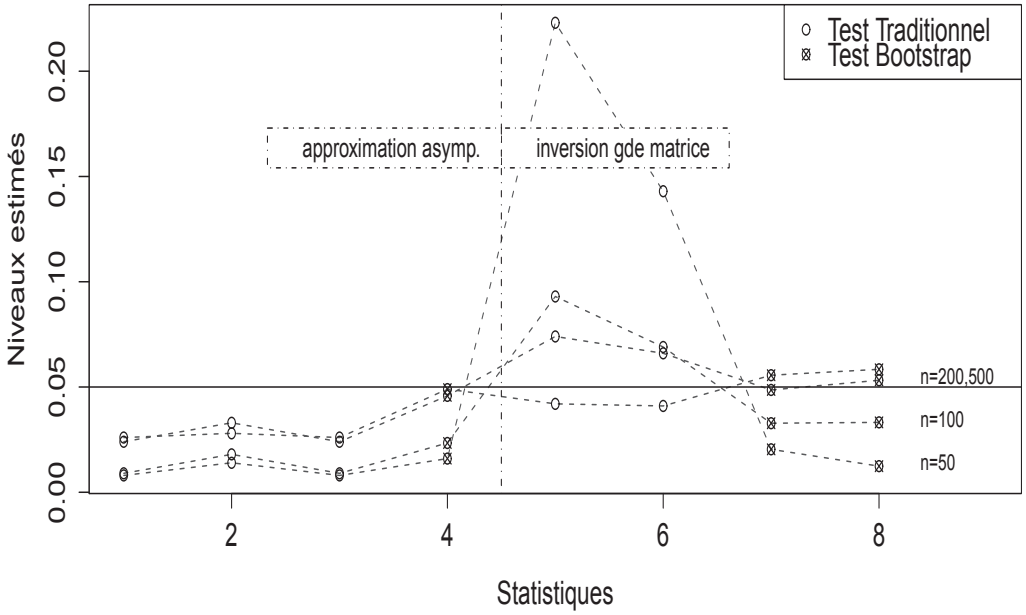

Figure 3 – Modèle (13). Niveaux estimés lors du test (9) avec $\widehat{\Lambda}_1$ (les approximations de [79] et [5] sont en abscisse 1,2,3, et le bootstrap CS en abscisse 4), $\widehat{\Lambda}_2$ et $\widehat{\Lambda}_3$ (abscisse 5 et 6 pour les tests traditionnels et abscisse 7 et 8 pour les tests bootstrap).

où P^* est une version bootstrap de $\widehat{P}$. Il est possible de démontrer la convergence de R^* sous H_0. Sous H_1, les projecteurs considérés ne convergent pas et donc on s'attend à avoir une statistique élevée. Les simulations conduites dans chacun des articles montrent le bon comportement de cette approche. Notons que contrairement au test d'hypothèse bootstrap, une telle procédure utilise le bootstrap afin de calculer des versions de la statistique de test. Un exemple de ce type d'approche se trouve dans Barrios et Velila (2007), les auteurs considèrent les statistiques de type $\sqrt{n}(\widehat{\lambda} - \lambda)$ où λ (resp. λ) est une valeur propre de la matrice M (resp. $\widehat{M}$).

3 Estimation semi-paramétrique de l'espace central moyen

Nous avons présenté précédemment la littérature RIV qui constitue une des approches pour l'estimation de l'espace central et de l'espace central moyen. En particulier, nous avons souligné le choix de cette dernière de restreindre la classe des variables explicatives considérées dans les modèles de régression (2) et (3). Les variables explicatives doivent en effet vérifier la condition de linéarité pour que les quantités estimées vivent dans les espaces de réduction de la dimension. Une telle restriction permet d'éviter une estimation non-paramétrique parfois couteuse en terme de vitesse de convergence. L'approche semi-paramétrique qui s'est développée dans le cadre des modèles de réduction de la dimension tels que (2) et (3) est en opposition avec la littérature RIV dans le sens où cette dernière allège considérablement les hypothèses portant sur la loi des variables explicatives.

La plupart des méthodes semi-paramétriques connues dans ce domaine requièrent tout d'abord l'estimation non-paramétrique d'une certaine fonction, disons h. En conséquence, les vitesses de convergence de l'estimateur non-paramétrique $\widehat{h}$ de h sont bien inférieures aux vitesses paramétriques. Dans un second temps, $\widehat{h}$ est utilisé pour estimer les espaces de réduction de la dimension considérés. L'enjeu principal est de récupérer des vitesses paramétriques lors de cette seconde étape.

Dans la suite nous nous intéressons à l'estimation semi-paramétrique de l'espace central moyen [18]. Nous supposons que $((X,Y),(X_i,Y_i)_{1\leq i\leq n})$ est une suite de variables aléatoires vérifiant le modèle (3). De plus, l'espace central moyen E_m est supposé unique. Nous rappelons que $E_m = \mathrm{span}(\beta)$ où β est appelé l'index.

3.1 La M-estimation semi-paramétrique.

La M-estimation semi-paramétrique dans les modèles de réduction de la dimension fût originalement étudiée par Ichimura en 1993 [59]. Afin de présenter cette approche, on remarque tout d'abord que l'index du modèle (3) est donné par

$$\beta = \underset{\beta\in\mathbb{R}^{p\times d_0}}{\mathrm{argmin}}\ \mathbb{E}[(Y-\mathbb{E}[Y|\beta^T X])^2]. \tag{20}$$

Cette formule est une conséquence du développement

$$\mathbb{E}[(Y-\mathbb{E}[Y|\beta^T X])^2] = \mathbb{E}[(\mathbb{E}[Y|X]-\mathbb{E}[Y|\beta^T X])^2] + \mathbb{E}[(Y-\mathbb{E}[Y|X])^2],$$

celui-ci implique que β de (20) minimise $\mathbb{E}[(\mathbb{E}[Y|X]-\mathbb{E}[Y|\beta^T X])^2]$, qui vaut 0 si et seulement si $\mathrm{span}(\beta) = E_m$ d'après la Proposition 3. Si la fonction g_0 définie par $g = g_0 \circ \beta$ était connue, alors une procédure paramétrique classique serait d'estimer β par

$$\underset{\beta\in\mathbb{R}^{p\times d_0}}{\mathrm{argmin}}\ \sum_{i=1}^{n}(Y_i - g_0(\beta^T X_i))^2,$$

18. L'estimation semi-paramétrique de l'espace central a fait l'objet de récentes études. Quelques méthodes ont été proposées par Xia en 2007 [80] et par Zeng et Zhu en 2010 [84].

néanmoins g_0 est inconnue dans notre cadre. Ainsi on introduit l'estimateur de Nadaraya-Watson de la fonction de régression $\mathbb{E}[Y|\beta^T X]$ définie par

$$\widehat{g}_\beta(x) = \frac{\sum_{i=1}^n Y_i K(h^{-1}\beta^T(X_i - x))}{\sum_{i=1}^n K(h^{-1}\beta^T(X_i - x))},$$

pour tout $x \in \mathbb{R}^p$. L'idée est de reproduire les principes généraux de la M-estimation paramétrique, où g_0 est connue, au cas où la fonction g_0 est estimée comme précédemment. L'estimateur de l'index proposé est le suivant

$$\widehat{\beta}_{\mathrm{M}} = \underset{\beta \in \mathbb{R}^{p\times d_0}}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - \widehat{g}_\beta(X_i))^2 I_i, \tag{21}$$

où $I_i = \mathbb{1}_{\{X_i \in Q\}}$ et $Q \subset \mathbb{R}^p$. Le term I_i est appelé *trimming term*, il est introduit pour empêcher les trop petites valeurs du dénominateur de $\widehat{g}_\beta(X_i)$ de mal influencer l'estimation. Ce terme assure à la fois le bon comportement théorique de l'estimateur mais aussi sa stabilité algorithmique. Il est démontré dans [59] que l'estimateur obtenu est asymptotiquement normal.

Une généralisation des travaux de [59] est proposée par Delecroix, Hristache et Patilea en 2003 [30], lesquels étudient le comportement de M-estimateurs semi-paramétriques basés sur d'autres fonctions de contraste que la fonction carrée utilisée dans (21). De plus les auteurs utilisent une procédure où l'optimisation (21) permet aussi de choisir la fenêtre h de façon optimale (voir aussi [52] sur ce point). La normalité asymptotique de l'estimateur de β est démontrée sous certaines hypothèses.

La M-estimation semi-paramétrique produit donc une estimation à vitesse paramétrique. Néanmoins, le principal défaut de cette approche est lié au problème d'optimisation de type (21), nécessaire au calcul de l'estimateur. Ce dernier n'est pas convexe et sa résolution en dimension multiple est difficile. Une solution à ce problème est apportée par Xia en 2002 [81] qui propose la méthode *minimum average variance estimation* (MAVE) qui s'inspire de la M-estimation semi-paramétrique de la façon suivante. Au lieu de minimiser (21) à l'aide de l'estimateur $\widehat{g}_\beta$ évalué en chacun des points X_i's (comme c'esst le cas pour la M-estimation), l'auteur propose de mener cette minimisation à l'aide de valeurs approchées des points $\widehat{g}_\beta(X_i)$'s. Pour cela, on utilise l'idée de l'estimation par lissage localement linéaire (voir [40]) , ce qui nous permet d'approcher la fonction $x \mapsto \mathbb{E}[(Y - g_0(\beta^T X))^2 | X = x]$ au point X_i par

$$\min_{a\in\mathbb{R}, b\in\mathbb{R}^{d_0}} \sum_{j=1}^n (Y_j - (a + b^T\beta^T(X_j - X_i)))^2 K(h^{-1}(X_i - X_j)).$$

Revenant au problème initial (20), on obtient naturellement l'estimateur de la méthode MAVE donné par

$$\widehat{\beta}_{\mathrm{MAVE}} = \underset{\beta\in\mathbb{R}^{p\times d_0},\ a_i\in\mathbb{R},\ b_i\in\mathbb{R}^{d_0}}{\operatorname{argmin}} \sum_{i=1}^n \sum_{j=1}^n (Y_j - (a_i + b_i^T\beta^T(X_j - X_i)))^2 K(h^{-1}(X_i - X_j)),$$

sous la contrainte $\beta^T\beta = I$. Le précédent problème lié à l'optimisation (21) est résolu puisque ce dernier est désormais quadratique sous contrainte. Sa solution est exprimée dans l'article en question. De plus il est démontré, sous certaines conditions, que

$$\mathrm{dist}(\beta, \widehat{\beta}) = O_{\mathbb{P}}\left(h^2 + \frac{\log(n)}{nh^{p-1}}\right),$$

où $\mathrm{dist}(\cdot,\cdot)$ est une distance sur $\mathbb{R}^{p\times d_0}$. A notre connaissance, la normalité asymptotique de la méthode MAVE n'a pas été obtenue.

3.2 Utilisation du gradient de la régression

Afin de remédier aux problèmes causés par l'optimisation de la fonction de contraste en M-estimation, une seconde approche basée sur le gradient de la fonction de régression s'est développée. En effet, pour tout x dans le support de X, on a $\nabla g(x) \in E_m$, et on peut démontrer que [19]

$$\mathrm{span}(\nabla g(x),\ x \text{ dans le support de } X) = E_m.$$

Afin d'estimer l'espace engendré par le gradient de la régression dans le cas du single index, il est naturel d'estimer $\mathbb{E}[\nabla g(X)]$. Dans le cas où E_m est de dimension plus grande que 1, on peut estimer

$$\eta_\phi = \mathbb{E}[\nabla g(X)\phi(X)], \tag{22}$$

et ensuite faire varier $\phi : \mathbb{R}^p \to \mathbb{R}$ afin de retrouver plusieurs directions. La littérature se décompose selon deux méthodes d'estimation de (22). La première estime directement le gradient par un estimateur de g localement linéaire. La seconde utilise la formule d'intégration par partie (IPP) afin d'estimer une quantité du type $\mathbb{E}[Y\Phi(X)]$, où $\Phi : \mathbb{R}^p \to \mathbb{R}^p$ est une fonction inconnue que l'on doit estimer.

3.2.1 Estimation direct du gradient

L'estimation du gradient de la régression est effectuée simultanément avec l'estimation de la régression. Plus précisément, on utilise l'estimateur localement linéaire de la régression (voir par exemple [40]) qui est défini au point X_i par

$$\begin{pmatrix} \widehat{g}(X_i) \\ \widehat{\nabla g}(X_i) \end{pmatrix} = \underset{a\in\mathbb{R}, b\in\mathbb{R}^p}{\mathrm{argmin}} \sum_{j=1}^{n} (Y_j - a - b^T(X_j - X_i))^2 K(h^{-1}(X_i - X_j)).$$

Cette approche est l'objet de l'articles de Hristache, Juditsky, et Spokoiny de 2001 [56] et de l'article de Hristache, Juditsky, Polzehl et Spokoiny de 2001 [55]. Dans ces articles, les auteurs estiment (22) tout d'abord par

$$\widehat{\eta}_\phi = n^{-1}\sum_{i=1}^{n} \widehat{\nabla g}(X_i)\phi(X_i),$$

19. Voir par exemple la preuve du Lemme 4.5 page 142.

où la fonction ϕ vit dans une certaine base de fonctions. Tout d'abord, sous certaines conditions, notamment de design non-aléatoire, les auteurs fournissent des vitesses en $n^{-1/p}$ pour l'estimation de η_ϕ par $\widehat{\eta}_\phi$. En conséquence, le premier estimateur proposé n'est pas $\sqrt{n}$ consistent. Afin d'y remédier, les auteurs définissent un second estimateur qui adapte le choix de la fenêtre à la structure de l'échantillon. En effet, la première estimation nous fournit une indication quant à la direction de l'espace E_m, on utilise cette indication afin d'élargir la fenêtre dans les directions orthogonales aux vecteurs $\widehat{\eta}_\phi$'s (où g varie peu) et de la limiter dans les directions $\widehat{\eta}_\phi$'s (où g varie beaucoup). En itérant ce processus, les auteurs obtiennent la vitesse $\sqrt{n}$ pour l'estimateur final. Néanmoins, leurs résultats se limitent au cas où la dimension est inférieur où égale à 4, en prenant en compte l'extension proposée par Dalalyan, Juditsky et Spokoiny en 2008 [27].

3.2.2 Estimation du gradient par IPP

Partons de l'équation (22) et posons $\phi(X) = \frac{\psi(X)}{f(X)}$, où f est la densité du vecteur X. Définissons

$$\eta_\psi = \int \nabla g(x)\psi(x)\mathrm{d}x.$$

Sous certaines conditions de régularité nécessaire à l'IPP (voir Lemme 4.4 page 141) , on obtient

$$\eta_\psi = \mathbb{E}\left[\frac{Y\nabla\psi(X)}{f(X)}\right].$$

Ainsi l'estimation des vecteurs η_ψ peut se faire par l'intermédiaire de l'estimation de f. Une approche similaire est développée dans les articles de Powell, Stock et Stoker de 1989 [71], de Härdle et Stoker de 1989 [53] et plus récemment de Zeng et Zhu de 2010 [84] (voir page 129 pour plus de précisions sur les méthodes en question).

Suivant cette approche, dans le chapitre 4, nous proposons une nouvelle méthode d'estimation semi-paramétrique de E_m. La méthode proposée se base sur l'estimation de la matrice

$$M = \int \eta_\psi \eta_\psi^T,$$

où la somme est prise sur une certaine famille de fonctions $\mathcal{F}$. D'une part, notre méthode est motivée par ses propriétés d'exhaustivité, i.e. $\mathrm{span}(M) = E_m$ (voir Lemme 4.5). D'autre part, elle permet d'obtenir une vitesse paramétrique pour l'estimation de l'espace moyen. L'estimateur proposé est

$$\widehat{M} = \int_{\mathcal{F}} \widehat{\eta}_\psi \widehat{\eta}_\psi^T,$$

avec

$$\widehat{\eta}_\psi = n^{-1}\sum_{i=1}^n \frac{Y_i\nabla\psi(X_i)}{\widehat{f}(X_i)} \quad \text{et} \quad \widehat{f}(x) = (nh^p)^{-1}\sum_{j=1}^n K(h^{-1}(X_j - x)),$$

et le résultat principal est le suivant. Les hypothèses qui y figurent sont exposées page 132.

Théorème 22 (Chapitre 4, Théorème 4.3, page 139)**.** *Supposons que (A1-A6) et (A7') soient satisfaites, alors*

$$\sqrt{n}(\widehat{M} - M) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma_2),$$

où Σ_2 *est la variance de la variable aléatoire* $Z_2 + Z_2^T$ *avec*

$$Z_2 = \frac{Y_1 - g(X_1)}{f(X_1)} \int_{\mathcal{F}} \nabla\psi(X_1)\eta_\psi^T.$$

De plus, nous obtenons aussi des conditions garantissant la convergence faible dans l'espace des fonctions indexées sur la classe de fonction $\mathcal{F}$. Ce résultat est énoncé dans le Théorème 4.6 page 143.

3.3 Approximation d'intégrales par lissage par noyau

Un des lemmes utilisé pour la démonstration des résultats énoncés précédemment est d'un interêt particulier puisque son utilité dépasse le cadre de la réduction de la dimension. En effet ce dernier lemme concerne l'estimation de la valeur d'une intégrale et vient directement concurrencer la traditionnelle méthode dite de Monte-carlo. Le résultat est le suivant, c'est une conséquence du Lemme 4.1 page 133.

Lemme 23 (Chapitre 4, Lemme 4.1, page 133)**.** *Supposons que* φ *est une fonction Hölderienne sur son support compact. Alors si* K *et* f *vérifient quelques conditions de régularité ((A1-A4) et (B1)) on a*

$$n^{1/2}\left(n^{-1}\sum_{i=1}^{n}\frac{\varphi(X_i)}{\widehat{f}(X_i)} - \int \varphi(x)\mathrm{d}x\right) \xrightarrow{\mathbb{P}} 0.$$

Dans le Lemme 4.1, les vitesses de convergence du lemme précédent sont explicitées lorsque la fonction φ est Hölderienne sur $\mathbb{R}^p$ à support compact. Ce résultat est intéressant car il donne de meilleures vitesses que la méthode de Monte-Carlo traditionnelle. En effet, si dans le lemme précédent, on remplace $\widehat{f}$ par la vrai densité f, alors sous certaines conditions, on obtient par le TCL que

$$n^{1/2}\left(n^{-1}\sum_{i=1}^{n}\frac{\varphi(X_i)}{f(X_i)} - \int \varphi(x)\mathrm{d}x\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \mathrm{var}\left(\frac{\varphi(X)}{f(X)}\right)\right).$$

Chapter 1

Test function : A new approach for covering the central subspace

Abstract : This paper studies a general family of methods for sufficient dimension reduction (SDR) called the test function (TF), based on the introduction of a nonlinear transformation of the response. By considering order 1 and 2 conditional moments of the predictors given the response, we distinguish two classes of methods. The optimal members of each class are calculated with respect to the asymptotic mean squared error between the central subspace (CS) and its estimate. Moreover the theoretical background of TF is developed under weaker conditions than the existing methods. Accordingly, simulations confirm that the resulting methods are highly accurate.

Key words : Inverse regression ; Slicing estimation ; Sufficient dimension reduction ; Central subspace.

This chapter was published in
Journal of multivariate Analysis,
volume 115, March 2013, Pages 84-107,
under the title :
Optimal transformation : A new approach for covering the central subspace.
It has been written in collaboration with Bernard Delyon.

1.1 Introduction

Dimension reduction in regression aims at improving poor convergence rates derived from the nonparametric estimation of the regression function in large dimension. It attempts to provide methods that challenge the curse of dimensionality by reducing the number of predictors. A specific dimension reduction framework, called the *sufficient dimension reduction* has drawn attention in the last few years. Let Y be a random variable and X a p-dimensional random vector. To reduce the number of predictors, it is proposed to replace X by a number smaller than p of linear combinations of the predictors. The new covariate vector has the form PX, where P can be chosen as an orthogonal projector on a subspace E of $\mathbb{R}^p$. Clearly, this kind of methods relies on an alchemy between the dimension of E, which needs to be as small as possible, and the conservation of the information carried by X about Y through the projection on E. In the SDR literature, mainly two kind of spaces have been studied. First a *dimension reduction subspace* (DRS) ([66]) is defined by the conditional independence property

$$Y \perp\!\!\!\perp X \mid P_c X, \tag{1.1}$$

where P_c is the orthogonal projector on a DRS. With words, it means that knowing $P_c X$, there is no more information carried by X about Y. It is possible to show that (1.1) is equivalent to

$$\mathbb{P}(Y \in A|X) = \mathbb{P}(Y \in A|P_c X) \quad \text{a.s.}, \tag{1.2}$$

for any measurable set A, or there exists a noise e and a function f such that Y has the representation

$$Y = f(P_c X, e) \quad \text{with} \quad e \perp\!\!\!\perp X.$$

Moreover under some additional conditions (see for instance [19]), the intersection of all the DRS is itself a DRS. Consequently, there exists a unique DRS with minimum dimension and we call it the *central subspace* (CS). In this article the CS is noted E_c. Secondly, another space called a *mean dimension reduction subspace* (MDRS) has been defined in [21] with the property

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|P_m X] \quad \text{a.s.}, \tag{1.3}$$

where P_m is the orthogonal projector on a MDRS. Clearly, the existence of a MDRS requires a weaker assumption than the existence of a DRS and therefore it seems to be more appropriate to the context of regression. Because of the equivalent formulation of equation (1.3),

$$Y \perp\!\!\!\perp \mathbb{E}[Y|X] \mid P_m X,$$

the definition of a MDRS imposes that all the dependence between Y and its regression function on X is carried by $P_m X$. If the intersection of all the MDRS is itself a MDRS, then it is called the *central mean subspace* (CMS) ([21]). In the following the CMS is noted E_m. Finally, notice that because a DRS is a MDRS, the CS contains the CMS.

There exist many methods for estimating the CS and the CMS and these methods can be divided into two groups, those who require some assumptions on the distribution of

the covariates and those who do not. The second group includes *structure adaptive method* (SAM) ([55]), *minimum average variance estimation* (MAVE) ([81]), *structural adaptation via maximum minimization* (SAMM) ([27]). Those methods are free from conditions on the predictors but require a nonparametric estimation of the regression function $\mathbb{E}[Y|X = x]$. More recently, the *central solution space* (CSS) ([63]) has also been introduced to alleviate some common assumptions on the distribution of the predictors. In this article we are concerned only with methods of the first group. They are presented in the following paragraph.

For the sake of clarity, from now on we work in terms of standardized covariate $Z = \Sigma^{-\frac{1}{2}}(X - \mathbb{E}[X])$ with $\Sigma = \text{var}(X)$ is a full rank matrix. Hence we define the standardized CS as $\Sigma^{\frac{1}{2}}E_c$. Since there is no ambiguity, we still note it E_c and we still denote by P_c the orthogonal projector on this subspace. Define d as the dimension of E_c. For any matrix M, we note $\text{span}(M)$ the space generated by the column vectors of M, and $\text{vec}(M)$ the vector of columns of M. The usual Kronecker product will be noted $\otimes$ and we denote by $Z^{(k)}$ the k-th component of the vector Z.

All the methods of the first group derive from the principle of inverse regression : instead of studying the regression curve which implies high dimensional estimation problems, the study is based on the inverse regression curve $\mathbb{E}[Z|Y = y]$ or the inverse variance curve $\text{var}(Z|Y = y)$. We will respectively refer to the order 1 and order 2 approaches. To infer about the CS, order 1 methods require that

Assumption 1. *(Linearity condition)*

$$Q_c\mathbb{E}[Z|P_cZ] = 0 \quad a.s.,$$

where $Q_c = I - P_c$. Under the linearity condition and the existence of the CS, it follows that $\mathbb{E}[Z|Y] \in E_c$ a.s. and then if we divide the range of Y into H slices $I(h)$, we have for every h,

$$m_h = \mathbb{E}[Z|Y \in I(h)] \in E_c, \tag{1.4}$$

and clearly, the space generated by some estimators of the m_h's estimates the CS, or more precisely a subspace of the CS. To obtain a basis of this subspace, [66] proposed a principal component analysis and this led to an eigendecomposition of the matrix

$$\widetilde{M}_{\text{SIR}} = \sum_h p_h m_h m_h^T, \tag{1.5}$$

where $p_h = \mathbb{P}(Y \in I(h))$. Many methods relying on the inverse regression curve such as *sliced inverse regression* (SIR) ([66]) have been developed. Other ways to estimate the inverse regression curve are investigated in *kernel inverse regression* (KIR) ([85]) and *parametric inverse regression* (PIR) ([14]). Instead of a principal component analysis, the minimization of a discrepancy function is studied in *inverse regression estimator* (IRE) ([22]) to obtain a basis of the CS. In [83], some polynomial transformations of the response are considered to estimate some subspaces of the CS. For a complete background about order 1 methods, we refer to [22].

By considering regression models like $Y = |Z^{(1)}| + e$, with Z having a symmetric distribution and $e \perp\!\!\!\perp Z$, some authors (for instance [66]) noticed that sometimes, $\mathbb{E}[Z|Y] = 0$ a.s. and refer to the SIR pathology when it occurs. Order 2 methods have been introduced to handle such a situation. In addition to the linearity condition order 2 methods require that

Assumption 2. *(Constant conditional variance (CCV))*

$$\mathrm{var}(Z|P_cZ) = Q_c \quad a.s.,$$

then under the linearity condition, CCV and the existence of the CS, it follows that $\mathrm{span}(var(Z|Y) - I) \in E_c$ a.s. and by considering a slicing of the response, we have

$$\mathrm{span}(v_h - I) \subset E_c, \tag{1.6}$$

where $v_h = \mathrm{var}(Z|Y \in I(h))$. Since the spaces generated by the matrices $(v_h - I)$'s are included in the CS, *sliced average variance estimation* (SAVE) in [23] proposed to make an eigendecomposition of the matrix

$$\widetilde{M}_{\mathrm{SAVE}} = \sum_h p_h(v_h - I)^2,$$

to derive a basis of the CS. Another combination of matrices based on the inverse variance curve is *sliced inverse regression-II* (SIR-II) ([66]). More recently, *contour regression* (CR) ([65]), and *directional regression* (DR) ([64]) investigate a new kind of estimator based on empirical directions. Besides, methods for estimating the CMS also require Assumptions 1 and 2. They include *principal Hessian direction* (pHd) ([67]), and *iterative Hessian transformation* (IHT) ([21]). In order to clear the failure of certain methods when facing pathological models and to keep their efficiency in other cases, some combinations of the previous methods as SIR and SIR-II, SIR and pHd or SIR and SAVE have been studied in [44] and [82].

As we have just highlighted, Assumptions 1 and 2 are needed to respectively characterize the CS with the inverse regression curve and the inverse variance curve. A first point is that the linearity condition and CCV assumed together are really close to an assumption of normality on the predictors. Moreover for each quoted method, these assumptions guarantee only that the estimated CS is asymptotically included in the true CS. A crucial point in SDR and a recent new challenge is to propose some methods that allow a comprehensive estimation of the CS under mild conditions. Recent researches are concerned with this problem, [65] and [64] proposed a new set of assumptions that guarantees the exhaustiveness of the estimation, i.e. the whole CS is estimated.

In this paper, we propose a general point of view about SDR by introducing the test function method (TF). The original basic idea of TF is to investigate the dependence between Z and Y by introducing nonlinear transformations of Y, and inferring about the CS through their covariances with Z or ZZ^T. Actually, an important difference between

TF and other methods is that neither the inverse regression curve and nor the inverse variance curve are estimated as it is suggested by equations (1.4) and (1.6). In this paper, these two curves are some working tools but the inference about the CS is obtained through some covariances. More precisely, the CS is obtained either by an inspection of the range of

$$\mathbb{E}[Z\psi(Y)],$$

when ψ varies in a well chosen finite family of function or either by an eigendecomposition of

$$\mathbb{E}[ZZ^T\psi(Y)],$$

where ψ is a well chosen function. Hence two kinds of methods can be distinguished, the order 1 test function methods (TF1) and the order 2 test function methods (TF2). Notice that $\widetilde{M}_{\text{SIR}}$ is an estimate of $\mathbb{E}[Z\mathbb{E}[Z|Y]^T]$, hence SIR may be seen as a particular case of TF1. For similar reasons, TF1 also extends results of [83] who considered polynomial transformations. Besides, the results regarding TF2 are somewhat more interesting because just a single transformation $\psi : \mathbb{R} \to \mathbb{R}$ is sufficient to have an accurate estimate. As a consequence, there are few connections between TF2 and the order 2 existing methods, for instance SAVE and DR involve transformations of the form $\mathbb{E}[ZZ^TA(Y)]$ where $A(Y)$ is a matrix.

This paper has two principal objectives : to provide a general theoretical study of TF1 and TF2 linked with the background of the existing methods, and to derive the optimal members of each methodology through an asymptotic variance minimization. The optimal members are respectively called order 1 optimal function (OF1) and order 2 optimal function (OF2), they correspond to two distinguish methods for the estimation of the CS. As a result, a significant improvement in accuracy is targeted by OF1 and OF2. We show that TF allows to relax some hypotheses commonly assumed in the literature, especially we alleviate the CCV hypothesis for TF2. Moreover for both methodology TF1 and TF2, we provide mild conditions ensuring an exhaustive characterization of the CS. The present work is divided into the three following principal parts :

- Existence of the CS and the CMS
- Exhaustiveness of TF
- Optimality for TF

More precisely, it is organized as follows. In section 1.2, we propose some new conditions ensuring the existence of the CS and the CMS. Section 1.3 is devoted to TF1 : we present some conditions that guarantee the exhaustiveness of the method and then we calculate the optimal transformation of the response for TF1 to minimize the estimation error. By following the same path, we study TF2 in section 1.4. Accordingly, we propose two plug-in methods derived from the minimization of the mean squared error : OF1 and OF2. The estimation of the dimension of the CS is addressed in section 1.5. Finally, in section 1.6 we compare both methods to existing ones through simulations.

1.2 Existence of the central subspace and the central mean subspace

Conditions on the uniqueness of subspaces that allow a dimension reduction are investigated in this section. This problem has drawn attention early in the literature but it seems not to be the case any more. As a consequence of the definition of the CS (resp. CMS), its existence is equivalent to the uniqueness of a DRS (resp. MDRS) with minimal dimension. In [19], Proposition 6.4 p.108, it is shown that the existence of the CS can be obtained by constraining the distribution of X to have a convex density support. Moreover, in [21], the existence of the CMS is ensured under the same condition than the CS. We prove in Theorem 1.1 below that the convexity assumption can be significantly weakened.

Theorem 1.1. *Under (1.1), if X has a density such that the Lebesgue measure of the boundary of its support is equal to* 0*, then the CS and the CMS exist.*

The proof is postponed to Appendix 1.8.1. Since TF is only concerned about the CS estimation, we assume from now on its existence.

1.3 Order 1 test function

A way to introduce TF1 is to consider some relevant facts about the SIR estimation. As explained in the Introduction, SIR consists in estimating the matrix

$$M_{\mathrm{SIR}} = \mathbb{E}\left[Z\mathbb{E}[Z|Y]^T\right],$$

whose column space is included in the CS. To make that possible, a slicing approximation of the conditional expectation $\mathbb{E}[Z|Y]$ is conducted and it leads to $\widetilde{M}_{\mathrm{SIR}}$ of equation (1.5). Because $p_h > 0$, it is clear that

$$\mathrm{span}(\widetilde{M}_{\mathrm{SIR}}) = \mathrm{span}\left(\mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}],\ h = 1, ..., H\right),$$

and it follows that SIR estimates a subspace spanned by the covariances between Z and a family of Y-measurable functions. The first goal of TF1 is to extend SIR to some other families of functions Ψ_H, in order to estimate E_c with $\mathrm{span}\left(\mathbb{E}[Z\psi(Y)],\ \psi \in \Psi_H\right)$. Besides, notice that

$$\widetilde{M}_{\mathrm{SIR}} = \mathbb{E}\left[Z\left(\phi_1(Y), ..., \phi_p(Y)\right)\right],$$

with $\phi_k(y) = \sum_h \alpha_{k,h}\mathbb{1}_{\{y\in I(h)\}}$ and $\alpha_{k,h} = \mathbb{E}[Z^{(k)}|Y \in I(h)]$. It follows that

$$\mathrm{span}\left(\mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}],\ h = 1, ..., H\right) = \mathrm{span}\left(\mathbb{E}[Z\phi_k(Y)],\ k = 1, ..., p\right),$$

and clearly SIR synthesizes the information contained in a set of H vectors into a set of p vectors. Although each of these spaces are equal, it is not the case for their respective estimators with finite sample. Accordingly, another issue for TF1 is to choose the p functions ϕ_k's in order to minimize the variance of the estimation.

The following theorem is not new at all. Yet, it makes a simple link between TF1 and the CS. We introduce the function space $L_p(r(\omega))$ defined as

$$L_p(r(\omega)) = \{\psi : \mathbb{R} \to \mathbb{R} \; ; \quad \mathbb{E}[|\psi(Y)|^p r(\omega)] < +\infty\},$$

where $r : \mathbb{R} \to \mathbb{R}_+$ is a measurable function and ω a random variable.

Theorem 1.2. *Assume that Z satisfies Assumption 1 and has a finite first moment. Then, for every measurable function $\psi \in L_1(\|Z\|)$, we have*

$$\mathbb{E}[Z\psi(Y)] \in E_c.$$

The linearity condition is often equated with an assumption of sphericity on the distribution of the predictors. It is well known that if Z is spherical then it satisfies the linearity condition but the converse is false. Actually, linearity condition and sphericity are not so closely related : in [34], it is shown that a random variable Z is spherical if and only if $\mathbb{E}[QZ|PZ] = 0$ for every rank 1 projector P and $Q = I - P$. Clearly, at this stage, the sphericity seems to be a too large restriction to obtain the linearity condition. However unlike the sphericity, since we do not know P_c, the linearity condition could not be checked on the data. For instance, an assumption close to the linearity condition is to ask the distribution of Z to be invariant by the orthogonal symmetry to the space E_c, i.e. $Z \overset{\mathrm{d}}{=} (2P_c - I)Z$. Then for any measurable function f,

$$\mathbb{E}[Q_c Z f(P_c Z)] = -\mathbb{E}[Q_c Z f(P_c Z)],$$

which implies the linearity condition. Recalling that sphericity means invariance in distribution by every orthogonal transformation, we have just shown that an invariance in distribution by a particular one suffices to get the linearity condition. Moreover, the assumption of sphericity suffers from the fact that if we add to Z some independent components, the resulting vector is no longer spherical whereas the linearity condition is still satisfied.

1.3.1 Exhaustiveness for TF1

As a consequence of Theorem 1.2, spaces generated by $(\mathbb{E}[Z\psi_1], ..., \mathbb{E}[Z\psi_H])$ are included in E_c. Our goal is to obtain the converse inclusion. Because TF1 is an extension of SIR, this one has a central place in the following argumentation. We start by giving a necessary and sufficient condition for covering the entire CS with SIR. Then under the same condition we extend SIR to a new class of methods.

Assumption 3. *(Order 1 coverage condition) For every nonzero vector $\eta \in E_c$, $\mathbb{E}[\eta^T Z|Y]$ has a nonzero variance.*

The previous assumption is clearly equivalent to $\mathrm{span}(M_{\mathrm{SIR}}) = E_c$. Moreover, it is always true that for H large enough $\mathrm{span}(M_{\mathrm{SIR}}) = \mathrm{span}(\widetilde{M}_{\mathrm{SIR}})$. Then we have the equivalent form

$$\mathrm{span}(\widetilde{M}_{\mathrm{SIR}}) = E_c$$

which was called the coverage condition in [22]. Nevertheless we use the former to make a link with some assumptions developed in [64] (see below Assumption 5 for more details). The aim is to shed light on some coverage-type result replacing the conditional expectation $\mathbb{E}[Z|Y]$ in M_{SIR} by some known and finite family of functions. Particularly, the previous equation provides such a result but only for the family of indicator functions.

Theorem 1.3. *Assume that Z and Y satisfy assumptions 1 and 3. Assume also that Z has a finite second moment. If Ψ is a total countable family in the space $L_1(\|Z\|)$, then one can extract a finite subset Ψ_H of Ψ such that*

$$\text{span}\left(\mathbb{E}[Z\psi(Y)],\ \psi \in \Psi_H\right) = E_c.$$

Remark 1.4. According to Theorem 1.B, quoted in Appendix 1.8.2, we can apply Theorem 1.3 with any family of functions that separates the points, for example polynomials, complex exponentials or indicators. Especially for polynomials, we extend a result stated in [83], Proposition 4, whose purpose is that E_c can be covered with the family $\Psi_H = \{Y^h, h = 1, ..., H\}$ if H goes to infinity.

To make possible a simple use of this theorem we need to recall this result. If $u = (u_1, ..., u_H)$ is a family of vectors in $\mathbb{R}^p$, then $\text{span}(uu^T) = \text{span}(u)$. Thus, if we denote by $\psi_1, ..., \psi_H$ some elements of a family that separates the points, then the CS can be obtained by making an eigendecomposition of the order 1 test function matrix associated with the functions $\psi_1, ..., \psi_H$ defined as

$$M_{\text{TF1}} = \sum_{h=1}^{H} \mathbb{E}[Z\psi_h(Y)]\mathbb{E}[Z\psi_h(Y)]^T.$$

Especially, under the conditions of Theorem 1.3, the eigenvectors associated with a nonzero eigenvalue of M_{TF1} generate E_c. Moreover, as pointed out before, for H large enough $\text{span}(\widetilde{M}_{\text{SIR}}) = \text{span}(M_{\text{SIR}})$. A proof of this result is cleared up by Theorem 1.3. By applying it with the family of indicator functions, it gives that

$$\text{span}\left(\mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}],\ h = 1, ..., H\right) = \text{span}(\widetilde{M}_{\text{SIR}}) = \text{span}(M_{\text{SIR}}) = E_c,$$

for H is sufficiently large. Moreover, SIR can be understood as a particular TF1. Expression (1.5) implies that

$$\widetilde{M}_{\text{SIR}} = \sum_{h=1}^{H} p_h^{-1}\mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}]\mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}]^T,$$

then, SIR is equivalent to TF1 realized with the weighted family of indicator functions $\left(\frac{\mathbb{1}_{\{Y\in I(h)\}}}{\sqrt{p_h}}\right)$. Besides, for any family of functions, the space spanned by M_{TF1} is invariant by positive weighting of the functions. Nevertheless with a finite sample, it is no longer the case for the estimated space and intuitively it seems that such a weighting could influence the convergence rate and improve the accuracy of TF1. The choice of the weights for the family of indicators is debated is section 1.3.2.

1.3.2 Optimality for TF1 : OF1

In this section, we develop a plug-in method based on the minimization of the variance estimation in the case of the family of indicator functions for Ψ_H. Theorem 1.3 and Remark 1.4 imply that the whole subspace E_c can be covered by the family of vectors $\{\mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}],\ h=1,...,H\}$ for a suitable partition $I(h)$. To provide a basis of E_c, it suffices to extract d orthogonal vectors living in this space. This procedure is realized by SIR. Nevertheless, the issue here is somewhat more complicated, we want to find d orthogonal vectors that have the smallest asymptotic mean squared error for the estimation of E_c. Let $(Z_1,Y_1),\dots,(Z_n,Y_n)$, with $Z_i=\Sigma^{-1/2}(X_i-\mathbb{E}[X])$, be an i.i.d. sample from model (1.1). To measure the estimation error, we define the quantity

$$\mathrm{MSE}=\mathbb{E}\left[\|P_c-\widehat{P}_c\|_{\mathrm{F}}^2\right], \tag{1.7}$$

where $\|\cdot\|_{\mathrm{F}}$ stands for the Frobenius norm and $\widehat{P}_c$ is derived from the family of vector $\widehat{\eta}=(\widehat{\eta}_1,...,\widehat{\eta}_d)$ defined as

$$\widehat{\eta}_k=\frac{1}{n}\sum_{i=1}^{n}Z_i\psi_k(Y_i),\quad \text{with}\quad \psi_k(Y)=(\mathbb{1}_{\{Y\in I(1)\}},...,\mathbb{1}_{\{Y\in I(H)\}})\alpha_k=\mathbb{1}_Y^T\alpha_k,$$

and $\alpha_k\in\mathbb{R}^H$. Besides, we introduce $\eta=(\eta_1,...,\eta_d)$ with $\eta_k=\mathbb{E}[Z\psi_k(Y)]$. Consequently, we aim at minimizing MSE according to the family $(\psi_k)_{1\le k\le d}$, or equivalently according to the matrix $\alpha=(\alpha_1,...,\alpha_d)\in\mathbb{R}^{H\times d}$. Moreover, since we have

$$\mathrm{MSE}\ =\ \mathbb{E}[d-\widehat{d}]+2\mathbb{E}[\mathrm{tr}(Q_c\widehat{P}_c)], \tag{1.8}$$

and we suppose that d is known, the minimization of MSE relies only on the minimization of the second term in the previous equality. Hence, this naturally leads to the minimization problem

$$\min_{\alpha}\ \lim_{n\to+\infty} n\mathbb{E}[\mathrm{tr}(Q_c\widehat{P}_c)],$$

under the constraint of orthogonality of the family $(\eta_k)_{1\le k\le d}$. For the sake of clarity, we prefer to minimize the expectation of the limit in distribution, instead of the limit of the expectation when n goes to infinity, of the sequence $n\,\mathrm{tr}(Q_c\widehat{P}_c)$. To set out clearly the next proposition, let us introduce some notations. Define the matrices $C\in\mathbb{R}^{p\times H}$, $D\in\mathbb{R}^{H\times H}$, such that

$$\begin{aligned} C=(C_1,...,C_H) \qquad &\text{with}\quad C_h=\mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}],\\ D=\mathrm{diag}_h\,d_h \qquad &\text{with}\quad d_h=\left(\mathbb{E}[\|Q_cZ\|^2\mathbb{1}_{\{Y\in I(h)\}}]\right),\end{aligned}$$

and

$$G=D^{-\frac{1}{2}}C^TCD^{-\frac{1}{2}}.$$

The matrix G is the Gram matrix of the vector family $(C_h/\sqrt{d_h})_{1\leq h\leq H}$, Theorem 1.3 and Remark 1.4 ensure that its rank is equal to d. Besides, G is diagonalizable and so we define $V=(V_1,V_2)\in\mathbb{R}^{p\times(d+(p-d))}$ such that

$$V^TGV=\begin{pmatrix}D_0 & 0\\ 0 & 0\end{pmatrix},$$

where $D_0\in\mathbb{R}^{d\times d}$.

Proposition 1.5. *If Z has a finite second order moment, then the random variable $n\,\mathrm{tr}(Q_c\widehat{P}_c)$ has a limit in law W_α as $n\to+\infty$. The minimization problem*

$$\min_{\alpha}\ \mathbb{E}\,[W_\alpha]\quad u.c.\quad \eta^T\eta=Id,$$

has a unique solution, up to orthogonal transformations, given by

$$\alpha=D^{-\frac{1}{2}}V_1D_0^{-\frac{1}{2}}.$$

To make a link with other methods and facilitate the programming of OF1, let us express the solution in another way. Instead of expressing the solution in terms of weights α_k's assigned to the indicator functions, we express it in terms of the vectors η_k's associated with these weights. Since the set of functions associated with OF1 is invariant by orthogonal transformations, we choose $\alpha=D^{-\frac{1}{2}}V_1D_0^{-\frac{1}{2}}$ to simplify the next calculation. We have

$$D^{-\frac{1}{2}}C^TCD^{-\frac{1}{2}}V_1=V_1D_0,$$

multiplying by $CD^{-\frac{1}{2}}$ on the left and by $D_0^{-\frac{1}{2}}$ on the right, this gives

$$CD^{-1}C^TCD^{-\frac{1}{2}}V_1D_0^{-\frac{1}{2}}=CD^{-\frac{1}{2}}V_1D_0^{-\frac{1}{2}}D_0.$$

Defining the particular order 1 test function matrix $\widetilde{M}_{\mathrm{OF1}}=CD^{-1}C^T$, and noticing that $\eta=CD^{-\frac{1}{2}}V_1D_0^{-\frac{1}{2}}$, the previous equation is equivalent to

$$\widetilde{M}_{\mathrm{OF1}}\eta=\eta D_0.$$

Thus, since $\widetilde{M}_{\mathrm{OF1}}$ has the same rank as G, we have shown that the vectors η_k's deriving from the optimal weighted family, are the eigenvectors of $\widetilde{M}_{\mathrm{OF1}}$ associated with the nonzero eigenvalues. Besides, it is easy to verify that the previous development is still true when each quantity is replaced by its estimate. Therefore, OF1 relies on the eigendecomposition of an estimator of the matrix $\widetilde{M}_{\mathrm{OF1}}$, whereas SIR is obtained through an eigendecomposition of the matrix $\widetilde{M}_{\mathrm{SIR}}$. To compare both methods, we write their expressions as follows

$$\widetilde{M}_{\mathrm{SIR}}=\sum_{h=1}^{H}\frac{C_hC_h^T}{p_h},\qquad \widetilde{M}_{\mathrm{OF1}}=\sum_{h=1}^{H}\frac{C_hC_h^T}{d_h}. \tag{1.9}$$

Hence, SIR and OF1 are closely related because both methods try to obtain the space generated by the C_h's through some PCA. This information seems to be collected more rapidly with OF1 because it minimizes the criterion (1.7), and as a consequence the convergence rate would be better. This idea is supported by the expression of $\widetilde{M}_{\mathrm{OF1}}$ in which bad slices are less weighted. While $\widetilde{M}_{\mathrm{SIR}} \underset{H\to+\infty}{\to} M_{\mathrm{SIR}}$, $\widetilde{M}_{\mathrm{OF1}}$ converges to

$$M_{\mathrm{OF1}} = \mathbb{E}\left[Z\frac{\mathbb{E}[Z|Y]}{\mathbb{E}[\|Q_cZ\|^2|Y]}\right].$$

As a consequence, OF1 requires the knowledge of Q_c. Therefore we set out a plug-in method to compute Q_c.

OF1 algorithm :

0. Standardization of X into Z. Initialize $\widehat{Q}_c = I$.

1. Compute

$$\widehat{d}_h = \frac{1}{n}\sum_{i=1}^{n}\|\widehat{Q}_cZ_i\|^2\mathbb{1}_{\{Y_i\in I(h)\}}, \quad \widehat{C}_h = \frac{1}{n}\sum_{i=1}^{n}Z_i\mathbb{1}_{\{Y_i\in I(h)\}}$$

$$\text{and} \quad \widehat{M} = \sum_{h=1}^{H}\frac{\widehat{C}_h\widehat{C}_h^T}{\widehat{d}_h}.$$

2. Extract $\widehat{\eta} = (\widehat{\eta}_1, ..., \widehat{\eta}_d)$: the d eigenvectors of $\widehat{M}$ with largest eigenvalues.

3. $\widehat{Q}_c = I - \widehat{\eta}\widehat{\eta}^T$.

Steps 1 to 3 are repeated until convergence is achieved and then $\widehat{\eta}$ is the estimated basis of the standardized CS derived from OF1. The estimated directions of the CS are $\Sigma^{-\frac{1}{2}}\widehat{\eta}$. At the end of the paper, OF1 is compared to SIR through some simulations.

Naturally the previous development can be carried out with some other total family of functions than indicators, say $\Psi_H = (\psi_1, ..., \psi_H)$. The calculation is quite similar, assuming that each ψ_h belong to $L_2(\|Z\|^2)$, the optimization leads to an analogous solution than previously replacing $\widetilde{M}_{OF1}$ by the matrix $CD_{\Psi_H}C^T$, with $D_{\Psi_H} = \mathbb{E}[\|Q_cZ\|^2\Psi_H\Psi_H^T]$.

1.4 Order 2 test function.

Basically, TF2 relies on the same approach as TF1 with the difference that it involves higher conditional moments of Z knowing Y. Indeed, we are interested in the space generated by the column vectors of the matrix $\mathbb{E}[ZZ^T\psi(Y)]$ where ψ denotes a measurable function. The following issues are addressed : we first investigate the exhaustiveness of TF2, especially we propose some conditions on ψ that guarantee a comprehensive estimate of the CS, then we look for optimality by introducing OF2.

Let us start with a known fact often presented as the SIR pathology. Consider the regression model

$$Y = g(Z^{(1)}, Z^{(2)}, e), \tag{1.10}$$

where $e \perp\!\!\!\perp Z \in \mathbb{R}^p$ and g is symmetric with respect to its first coordinate. Assume also that $(Z^{(1)}, Z^{(2)}) \overset{\mathrm{d}}{=} (-Z^{(1)}, Z^{(2)})$. Then thanks to the linearity condition we have $Q_c\mathbb{E}[Z\psi(Y)] = 0$ whereas the previous considerations clearly imply that $\mathbb{E}[Z^{(1)}\psi(Y)] = \mathbb{E}[-Z^{(1)}\psi(Y)]$. Therefore for any measurable function ψ, we have that $\mathbb{E}[Z\psi(Y)] = \mathbb{E}[(0, Z^{(2)}, 0, ..., 0)^T\psi(Y)]$ and consequently the first direction $(1, 0, ..., 0)^T$ cannot be reached by any method based on the inverse regression curve. Clearly, TF1 is sensitive to the SIR pathology. Facing this difficulty an idea developed first in [66] and [23] is to explore some higher conditional moments of Z given Y. Thus methods as SIR-II, SAVE, CR, or DR are interested in some properties of the matrix $\mathbb{E}[ZZ^T|Y]$. It is also the case for TF2. Nevertheless we do not follow the same path as other order 2 methods, especially regarding the assumptions required to explore this second order moment. Order 2 methods usually assume that Z has a spherical distribution or at least satisfies the linearity condition, and secondly that $\mathrm{var}(Z|P_cZ)$ is constant, i.e. CCV. In [11], Proposition 1.A, stated in Appendix B, shows how strong are the last two assumptions. Accordingly, the assumptions required for order 2 methods are really close to the assumption of normality on the distribution of the predictors. TF2 works under weaker conditions. Actually, the CCV condition is no longer needed and we substitute it with the following one.

Assumption 4. *(Diagonal conditional variance (DCV))*

$$\mathrm{var}(Z|P_cZ) = \lambda_\omega^* Q_c \quad a.s.,$$

with λ_ω^* a real random variable.

To facilitate future proofs and to clear up such a condition we provide an equivalent form in the following lemma.

Lemma 1.6. *Assume that Z has a finite second moment. Then the following assertions are equivalent,*

1. *for any orthogonal transformation H such that $HP_c = P_c$, we have*

$$\mathrm{var}(Z|P_cZ) = \mathrm{var}(HZ|P_cZ),$$

2. *there exists λ_ω^* a real random variable such that* $\mathrm{var}(Z|P_cZ) = \lambda_\omega^* Q_c$.

Moreover, under the linearity condition, we have $\lambda_\omega^* = \frac{1}{p-d}\mathbb{E}\left[\|Q_cZ\|^2|P_cZ\right]$.

Remark 1.7. Proposition 1.A indicates that coupling CCV and the spherical assumption is equivalent to the normality assumption for Z, which is quite restrictive. In our framework, since sphericity implies DCV, we alleviate this strong link between order 2 methods and the Gaussian assumption. Indeed, if Z is spherical, then its distribution is invariant by

any orthogonal transformation, and we have for any measurable function f and for any orthogonal matrix H,

$$\mathbb{E}[ZZ^T f(P_c Z)] = \mathbb{E}[HZZ^T H^T f(P_c HZ)].$$

In particular, the previous equation is true for any H which leaves invariant vectors of E_c and we obtain (1) of Lemma 1.6 which is equivalent to DCV. Thus, we have just proved that the spherical assumption implies DCV.

The following theorem is the analogous of Theorem 1.2 for TF2. We define

$$M_\psi = \mathbb{E}[ZZ^T \psi(Y)] \qquad \text{and} \qquad \lambda_\psi^* = \frac{1}{p-d}\mathbb{E}\left[\|Q_c Z\|^2 \psi(Y)\right].$$

Theorem 1.8. *Assume that Z satisfies assumptions 1 and 4 and has a finite second moment. Then, for every measurable function $\psi \in L_1(\|Z\|^2)$, we have*

$$\operatorname{span}(M_\psi - \lambda_\psi^* I) \subset E_c.$$

In practice, because λ_ψ^* is unknown, it seems difficult to use Theorem 1.8. Nevertheless, we do not need to know this particular eigenvalue, this issue is addressed in Remark 1.12. Besides, a consequence of Theorem 1.8 is that $E_c^\perp$ is included in the eigenspace of the matrix M_ψ associated with the eigenvalue λ_ψ^*. Therefore, if all the other eigenvalues are different from λ_ψ^*, the eigenspace associated with λ_ψ^* is equal to $E_c^\perp$. If this is true, the inclusion in Theorem 1.8 becomes an equality, i.e. all the directions of E_c could be recovered. This idea has a central place in the next section where this eigenvalue problem is addressed.

1.4.1 Exhaustiveness for TF2

An important tool in this section is the eigendecomposition of the matrix M_ψ, therefore we try to be more clear in introducing the following notations. Let λ_ψ and λ_Y be two functions $\mathbb{R}^p \to \mathbb{R}$ respectively defined by

$$\lambda_\psi(\eta) = \mathbb{E}[(\eta^T Z)^2 \psi(Y)] \quad \text{and} \quad \lambda_Y(\eta) = \mathbb{E}[(\eta^T Z)^2 | Y],$$

for every $\eta \in \mathbb{R}^p$. Notice that if η is a unit eigenvector of M_ψ (resp. $\mathbb{E}[ZZ^T|Y]$), then $\lambda_\psi(\eta)$ (resp. $\lambda_Y(\eta)$) is equal to the eigenvalue of the matrix M_ψ (resp. $\mathbb{E}[ZZ^T|Y]$) associated with η. However, recalling that $E_c^\perp$ is included in an eigenspace of M_ψ and $\mathbb{E}[ZZ^T|Y]$, the functions λ_ψ and λ_Y are both constant on the centered spheres of $E_c^\perp$. Their respective values on the unit sphere of $E_c^\perp$ are noted λ_ψ^* and λ_Y^*.

Definition. Let ψ be a measurable function, we call ψ-space the vector space of $\mathbb{R}^p$

$$E_\psi = \operatorname{span}(M_\psi - \lambda_\psi^*) = \operatorname{span}\left(\eta \in B(0,1) \subset \mathbb{R}^p,\ M_\psi \eta = \lambda_\psi^* \eta\right)^\perp.$$

Theorem 1.8 indicates that any ψ-space is included in E_c. However, there is no guarantee of the existence of a ψ-space equal to E_c. We follow the same path as for TF1, i.e. we consider some transformations of Y belonging to a dense family. Nevertheless, the results are quite different because we provide the existence of a single function ψ such that $E_\psi = E_c$. A unique additional assumption is needed.

Assumption 5. *(Order 2 coverage condition)*

$$\forall \eta \in E_c,\ \|\eta\| = 1 \quad \mathbb{P}\left(\mathbb{E}\left[(\eta^T Z)^2 | Y\right] = \mathbb{E}\left[\left.\frac{\|Q_c Z\|^2}{p-d}\right| Y\right]\right) < 1.$$

Assumption 5 reflects some similarities with other work such as [65] and [64]. As highlighted in Remark 1.7, our set of assumptions is weaker than their because DCV has replaced CCV. To match their context, assume that CCV is satisfied. Then, Assumption 5 becomes "$\mathbb{E}[(\eta^T Z)^2|Y]$ is nondegenerate", i.e. is not a constant almost surely. Otherwise, TF1 allows an exhaustive estimation of the CS provided that $\mathbb{E}[(\eta^T Z)|Y]$ is nondegenerate. Thus the exhaustiveness condition of TF is the union of the two previous, i.e.

$$\mathbb{E}[(\eta^T Z)^2|Y] \quad \text{or} \quad \mathbb{E}[(\eta^T Z)|Y] \quad \text{is nondegenerate,}$$

which is the same than the one proposed for DR in [64]. Accordingly, TF evolves in a more general context given by DCV but the assumptions ensuring its exhaustiveness are similar. These assumptions can be understood as theoretical ones because they are difficult to check in practice.

Theorem 1.9. *Assume that Z and Y satisfy assumptions 1, 4 and 5. Assume also that Z has a finite second moment, then if Ψ is a total countable family in the space $L_1(\|Z\|^2)$, there exists ψ a finite linear combination of functions in Ψ such that*

$$E_\psi = E_c.$$

Theorem 1.9 highlights some relevant facts about TF2. In addition to providing the existence of a ψ-space equal to E_c, it gives some information about the function ψ to be used. Indeed, Theorem 1.B indicates that the relevant families of functions for TF2 are those that separate the points. Hence, as for TF1, this suggests the use of TF2 with any of these families. For each such family, there exists a function ψ such that $E_\psi = E_c$, yet it does not provide an explicit form of such a ψ. Hence, we set out the following corollary which is the counterpart of Theorem 1.3 for TF2.

Corollary 1.10. *Assume that Z and Y satisfy assumptions 1, 4 and 5. Assume also that Z has a finite second moment then, if Ψ is a total countable family in the space $L_1(\|Z\|^2)$, we have*

$$\bigoplus_{\Psi_H} E_\psi = E_c,$$

where Ψ_H is a finite subset of Ψ.

1.4.2 Optimality for TF2 : OF2

For TF1 we needed at least d functions to recover the CS entirely. For this reason, it was convenient to develop a framework with weighted indicators because it led to a matrix optimization problem. In other words we fixed the class of functions for TF1 to solve a

finite dimensional optimization problem. Actually, for TF2 we follow a different path : we choose to optimize over all the measurable functions thanks to Gâteaux derivatives.

We have already highlighted that the eigenvectors of the matrix M_ψ can be decomposed into two blocks : the ones associated with the eigenvalue λ_ψ^* and the others which necessarily belong to E_c. Theorem 1.9 goes further by arguing that for some ψ, the eigenvectors associated with different eigenvalues than λ_ψ^* generate E_c. Therefore, P_c can be derived from this set of eigenvectors. A natural way to proceed is to estimate each quantity by its empirical version. Recall that $(Z_1, Y_1), \dots, (Z_n, Y_n)$, with $Z_i = \Sigma^{-1/2}(X_i - \mathbb{E}[X])$, is an i.i.d. sample from model (1.1). We define

$$\widehat{M}_\psi = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T \psi(Y_i)$$

and the function $\widehat{\lambda}_\psi : \mathbb{R}^p \to \mathbb{R}$ as $\widehat{\lambda}_\psi(\eta) = \eta^T \widehat{M}_\psi \eta$ for every $\eta \in \mathbb{R}^p$. Since d is assumed to be known, we define the projector $\widehat{P}_c = \widehat{\eta}_\psi \widehat{\eta}_\psi^T$ where $\widehat{\eta}_\psi \in \mathbb{R}^{p\times d}$ are the d eigenvectors of $\widehat{M}_\psi$ associated with the eigenvalues the farthest from λ_ψ^*. Because of the symmetry of the matrix $\widehat{M}_\psi$ and M_ψ, the convergence $\widehat{M}_\psi \xrightarrow{\mathbb{P}} M_\psi$ implies the convergence in probability of the associated eigenvalues (see [36] for some details). As a consequence, one can express the projectors with the Riesz formula. Let $\mathcal{C}$ be a contour of the complex plan which encloses the eigenvalues different from λ_ψ^*. We prefer to work with P_c and its estimator $\widehat{P}_c$ expressed as

$$P_c = \oint_{\mathcal{C}} (Iz - M_\psi)^{-1} \mathrm{d}z \quad \text{and} \quad \widehat{P}_c = \oint_{\mathcal{C}} (Iz - \widehat{M}_\psi)^{-1} \mathrm{d}z.$$

Because of equation (1.8), we minimize MSE through the quantity $\mathbb{E}[\mathrm{tr}(Q_c \widehat{P}_c)]$. As we did for OF1, we first calculate the limit in law of the random variable $n\,\mathrm{tr}(Q_c\widehat{P}_c)$, as n goes to infinity and then we derive its expectation. The next proposition is dedicated to this calculation.

Proposition 1.11. *Let $\psi \in L_2(\|Z\|^4)$ such that $E_\psi = E_c$. Then $n\,\mathrm{tr}(Q_c\widehat{P}_c)$ has a limit in law W_ψ and*

$$\mathbb{E}[W_\psi] = \mathrm{tr}\left(\mathbb{E}\left[ZZ^T\|Q_cZ\|^2\psi(Y)^2\right] P_c (P_c M_\psi - I\lambda_\psi^*)^{-2}\right).$$

The above proposition provides the expression of the quantity to minimize with respect to the function ψ. The next lines are attached to find a minimizer of $\mathbb{E}[W_\psi]$. This informal calculation leads to a fixed point equation whose solution is expected to be a global minimum of $\mathbb{E}[W_\psi]$. Thanks to Proposition 1.11 the quantity to minimize can be written as

$$\mathbb{E}[W_\psi] = \mathrm{tr}(\mathbb{E}[ZZ^T P_c\|Q_cZ\|^2\psi(Y)^2](P_cM_\psi - I\lambda_\psi^*)^{-2}),$$

or introducing the notations $A = ZZ^T P_c\|Q_cZ\|^2$ and $B = P_cZZ^T - \frac{\|Q_cZ\|^2}{p-d} I$,

$$\mathbb{E}[W_\psi] = \mathrm{tr}\left(\mathbb{E}[A\psi(Y)^2]\ \mathbb{E}[B\psi(Y)]^{-2}\right).$$

Thus we are looking for ψ such that

$$\frac{\partial}{\partial t}\mathbb{E}[W_{\psi+t\delta}]\Big|_{t=0} = 0,$$

for every bounded measurable function δ, or equivalently,

$$\mathbb{E}\Big[2\,\mathrm{tr}\left(A\delta\psi\mathbb{E}[B\psi]^{-2}\right) - \mathrm{tr}\left(\mathbb{E}[A\psi^2]\ \mathbb{E}[B\psi]^{-1}\{B\delta\mathbb{E}[B\psi]^{-1} + \mathbb{E}[B\psi]^{-1}B\delta\}\mathbb{E}[B\psi]^{-1}\right)\Big] = 0,$$

where δ and ψ stand for $\delta(Y)$ and $\psi(Y)$. Define the functions $A(Y) = \mathbb{E}[A|Y]$ and $B(Y) = \mathbb{E}[B|Y]$. Since the previous equation is true for any Y-measurable random variable $\delta(Y)$, we obtain

$$2\,\mathrm{tr}\left(A(Y)\psi(Y)\mathbb{E}[B\psi]^{-2}\right) - \mathrm{tr}\left(\mathbb{E}[A\psi^2]\mathbb{E}[B\psi]^{-1}\{B(Y)\mathbb{E}[B\psi]^{-1} + \mathbb{E}[B\psi]^{-1}B(Y)\}\mathbb{E}[B\psi]^{-1}\right) = 0 \quad \text{a.s.},$$

which leads to the implicit equation

$$\psi(y) = \frac{\mathrm{tr}\left(\mathbb{E}[B\psi]^{-1}\mathbb{E}[A\psi^2]\mathbb{E}[B\psi]^{-1}\{\mathbb{E}[B\psi]^{-1}B(y) + B(y)\mathbb{E}[B\psi]^{-1}\}\right)}{2\,\mathrm{tr}(A(y)\mathbb{E}[B\psi]^{-2})}.$$

Since $P_c = \eta_\psi\eta_\psi^T$, we have

$$\mathbb{E}[B\psi]^{-1}\eta_\psi = \eta_\psi D_\psi,$$

where $D_\psi = \mathrm{diag}_k(\lambda_\psi(\eta_k) - \lambda_\psi^*)^{-1}$ and η_k is the k-th column vector of η_ψ. Besides, a simple use of the linearity condition provides that $\mathbb{E}[\eta^T ZZ^T|Y] = \mathbb{E}[\eta^T ZZ^T P_c|Y]$ for every $\eta \in E_c$. Consequently, we have

$$\eta_\psi^T B(y) = \eta_\psi^T B(y) P_c.$$

and then, we obtain

$$\psi(y) = \frac{\mathrm{tr}\left(D_\psi A_\psi D_\psi\{D_\psi\widetilde{B}(y) + \widetilde{B}(y)D_\psi\}\right)}{2\,\mathrm{tr}\left(\widetilde{A}(y)D_\psi^2\right)},$$

where

$$A_\psi = \mathbb{E}\left[\eta_\psi^T ZZ^T\eta_\psi\|Q_c Z\|^2\psi(Y)^2\right],\ \widetilde{A}(y) = \eta_\psi^T A(y)\eta_\psi,\ \widetilde{B}(y) = \eta_\psi^T B(y)\eta_\psi,$$

are $d\times d$ matrices. Using the symmetry of the matrices A_ψ and $\widetilde{B}(y)$, and some well-known properties of the trace, we obtain

$$\psi(y) = \frac{\mathrm{tr}\left(D_\psi A_\psi D_\psi\widetilde{B}(y)D_\psi\right)}{\mathrm{tr}\left(\widetilde{A}(y)D_\psi^2\right)}. \tag{1.11}$$

A solution of Equation (1.11) is noted ψ_{OF2}, it is an optimal function inside the TF2 framework with respect to criterion (1.7). Hence, we define the OF2 matrix as

$$M_{\mathrm{OF2}} = \mathbb{E}[ZZ^T\psi_{\mathrm{OF2}}(Y)].$$

To calculate ψ_{OF2}, we propose an iteration of the fixed point equation (1.11). Before we state a more accurate algorithm to compute OF2, in particular to estimate the matrix M_{OF2}, we need to approximate ψ_{OF2}. Indeed, since $\widetilde{A}$ and $\widetilde{B}$ are unknown functions, one can use a slicing approximation and define $\widetilde{\psi}_{\mathrm{OF2}}$ as a solution of

$$\psi(y) = \sum_h \frac{\mathrm{tr}\left(D_\psi A_\psi D_\psi \widetilde{B}_h D_\psi\right)}{\mathrm{tr}\left(\widetilde{A}_h D_\psi^2\right)} \mathbb{1}_{\{y\in I(h)\}},$$

where $\widetilde{A}_h = \mathbb{E}[\widetilde{A}(Y)\mathbb{1}_{\{y\in I(h)\}}]$ and $\widetilde{B}_h = \mathbb{E}[\widetilde{B}(Y)\mathbb{1}_{\{y\in I(h)\}}]$. Now we set out the OF2 method based on the family of indicator functions. The following algorithm describes the iterations needed to implement our method. For a better understanding, we based the algorithm on the weights α_h's instead of the function $\psi(y) = \sum_h \alpha_h \mathbb{1}_{\{y\in I(h)\}}$. Besides $\widehat{A}_{\widehat{\psi}}$ and $\widehat{D}_{\widehat{\psi}}$ are noted $\widehat{A}$ and $\widehat{D}$, and we will need

$$M_h = \mathbb{E}[ZZ^T\mathbb{1}_{\{Y\in I(h)\}}] \quad \text{and} \quad \lambda_h = \mathbb{E}\left[\frac{\|Q_c Z\|^2}{p-d}\mathbb{1}_{\{Y\in I(h)\}}\right].$$

OF2 algorithm :

0. Standardization of X into Z. Compute

$$\widehat{M}_h = \frac{1}{n}\sum_{i=1}^n Z_i Z_i^T \mathbb{1}_{\{Y_i\in I(h)\}}, \quad \widehat{\lambda}_h = \mathrm{median}(\lambda \in \mathrm{spectrum}(\widehat{M}_h)),$$

   and initialize $\widehat{\alpha}_h \overset{\mathrm{d}}{=} \mathcal{U}[0,1]$ for every $h = 1, ..., H$.

1. Identify [1] the eigenvectors $\widehat{\eta} = (\widehat{\eta}_1, ..., \widehat{\eta}_d) \in E_c$ of $\widehat{M} = \sum_h \widehat{\alpha}_h \widehat{M}_h$.

2. Derive $\widehat{D} = \mathrm{diag}_k(\widehat{\lambda}_{\widehat{\psi}}(\widehat{\eta}_k) - \widehat{\lambda}^*_{\widehat{\psi}})^{-1}$ with $\widehat{\psi}(y) = \sum_h \widehat{\alpha}_h \mathbb{1}_{\{y\in I(h)\}}$, $\widehat{Q}_c = I - \widehat{\eta}\widehat{\eta}^T$ and

$$\widehat{A} = \sum_h \widehat{\alpha}_h \widehat{\eta}^T \widehat{A}_h \widehat{\eta}, \quad \text{with } \widehat{A}_h = \frac{1}{n}\sum_{i=1}^n Z_i Z_i^T \|\widehat{Q}_c Z_i\|^2 \mathbb{1}_{\{Y_i\in I(h)\}}.$$

3. Compute

$$\widehat{\alpha}_h = \frac{\mathrm{tr}\left(\widehat{D}^2\widehat{A}\widehat{D} \;\; (\widehat{\eta}^T\widehat{M}_h\widehat{\eta} - \widehat{\lambda}_h I)\right)}{\mathrm{tr}\left(\widehat{\eta}^T\widehat{A}_h\widehat{\eta}\;\widehat{D}^2\right)}.$$

1. See Remark 1.12 for some details about this point.

Repeat the last three steps until the convergence is achieved. The resulting function $\widehat{\psi}_{\mathrm{OF2}}$ is an estimate of the function ψ_{OF2}. Finally the set of vectors $\widehat{\eta}$ forms an estimated basis of the standardized CS. The space generated by $\Sigma^{-\frac{1}{2}}\widehat{\eta}$ provides an estimate of E_c by OF2.

Remark 1.12. An important practical issue for TF2 and in particular for OF2 is the way we identify the eigenvectors of $\widehat{M_\psi}$ that converge to some vectors of E_c or equivalently the way we identify their associated eigenvalues. This intervenes at each iteration of our algorithm to estimate D_ψ and η_ψ. Although λ^*_ψ is unknown, the theoretical background of TF2 advocates for an identification process based on the eigenvalues. Indeed, as it is pointed out by Theorem 1.8, the eigenvalues of M_ψ associated with eigenvectors included in $E_c^\perp$ are equal. We built an algorithm based on this fact but it was not sufficiently robust to small samples. We thus prefer to develop another one which takes into account the eigenvectors of $\widehat{M_\psi}$. Let η be an eigenvector of $\widehat{M_\psi}$, the identification process is based on a measure of the dependence between $\eta^T Z$ and Y. More precisely, we consider the Pearson's chi-squared statistic of the test of independence between $\eta^T Z$ and Y. Therefore, for each eigenvector we divide the range of $\eta^T Z$ into H slices noted $J(h)$ and we calculate

$$S(\eta) = \sum_{h,h'} \frac{\left(p_{hh'} - \overline{p_{hh'}}^h \, \overline{p_{hh'}}^{h'}\right)^2}{\overline{p_{hh'}}^h \, \overline{p_{hh'}}^{h'}}$$

where $p_{h,h'} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Y_i \in I(h)\}} \mathbb{1}_{\{(\eta^T Z_i) \in J(h')\}}$ and $\overline{\cdot}^h$ is the mean over h. Then the d eigenvectors of $\widehat{M_\psi}$ having the highest values of S are identified as converging in E_c. As a consequence, at step 2 of the OF2 algorithm, the $\widehat{\lambda}_{\widehat{\psi}}(\widehat{\eta}_k)$'s are the eigenvalues of $\widehat{M}$ associated with the eigenvectors $\widehat{\eta}_k$'s with the d highest values of S, $\lambda^*_{\widehat{\psi}}$ is the median over the other eigenvalues. In section 1.6, we performed OF2 with this algorithm.

1.5 Estimation of the dimension

All along the article, the dimension of the CS was assumed to be known. Its estimation is a crucial point in SDR since it corresponds to the number of explicative variables we keep in the regression. Clearly if the dimension is underestimated, then we loose some information about the response, and on the contrary we cannot get the suitable nonparametric convergence rates for the estimation of the regression function. We raise this issue for TF1 and TF2. The estimation of d can be reasonably conducted after the estimation of the matrix of interest, say M, in the following way. As we pointed out before, under some conditions, one can get

$$\mathrm{span}(M) = E_c,$$

and clearly, the estimation of d amounts to estimate the rank of M. Actually, to estimate the rank of such matrix, one can use the hypothesis testing methodology proposed by [66] whose null hypothesis is

$$H_0 \; : \quad d = m \quad \text{against} \quad H_1 \; : \quad d > m,$$

where d stands for the true dimension. Then we start by testing $d = 0$ against $d > 0$ which can be seen as a test for the existence of a DRS. If it is rejected we go a step further $m := m + 1$ until the first acceptance. If $d = m$ is accepted, then m is an estimate of the dimension of E_c. The usual statistic employed in SDR is

$$\widehat{\Lambda} = n \sum_{k=1}^{p-m} \widehat{\lambda}_k^2$$

where $(\widehat{\lambda}_1, ..., \widehat{\lambda}_p)$ are the singular values of an estimator of M arranged in ascending order. Roughly speaking, the statistic goes to infinity under H_1 because at least one of the eigenvalues goes to a positive constant. Under H_0 and some mild conditions $\widehat{\Lambda}$ converges in law. This is the issue raised by Theorem 1 in [12], stated in Appendix 1.8.2 as Theorem 1.C. Thanks to this theorem, most of the SDR methods can provide an estimate of the dimension of E_c. For SIR, because $\widetilde{M}_{\text{SIR}} = C \operatorname{diag}_h(p_h^{-1}) C^T$, it is preferable to apply Theorem 1.C directly with the matrix $\operatorname{diag}_h(p_h)^{-1/2} C$, then we define $\widehat{\Lambda}_{\text{SIR}}$ as $\widehat{\Lambda}$ with $M = C \operatorname{diag}_h(p_h^{-1/2})$. Because of the unknown asymptotic distribution of $\widehat{\Lambda}$ under H_0 in general, it is interesting to study the behavior of the statistic $\widehat{\Lambda}$ under some usual SDR assumptions in order to take advantage of the substantial simplifications they involve. For instance, [15] show that under the linearity condition and CCV, $\widehat{\Lambda}_{\text{SIR}}$ is asymptotically chi-squared. Hence in the following, we provide the asymptotic distribution of $\widehat{\Lambda}$ in a general TF1 context without specifying the family of function $\Psi_H = (\psi_1, ..., \psi_H)^T$. Moreover our study involves both sets of assumptions : CCV and DCV (see Remark 1.7 for details about such assumptions). We use a parametrization quite similar to that of section 1.3.2 by defining the matrix

$$M_\alpha = C\alpha(C\alpha)^T,$$

with $\alpha \in \mathbb{R}^{H \times H}$ could be unknown, $C = (C_1, \ldots, C_H)$, and $C_h = \mathbb{E}[Z\psi_h(Y)]$. Define also U_0 and V_0 as the respective basis of the left and right singular spaces of the matrix $C\alpha$ associated with the singular value 0. Assume that $(X_1, Y_1), \ldots, (X_n, Y_n)$ is an i.i.d. sample from model (1.1) and define

$$\widehat{Z}_i = \widehat{\Sigma}^{-1/2}(X_i - \overline{X}),$$

with $\overline{\cdot}$ the empirical mean. Then we can define the estimator

$$\widehat{M_\alpha} = \widehat{C}\widehat{\alpha}(\widehat{C}\widehat{\alpha})^T,$$

where $\widehat{C} = (\widehat{C}_1, \ldots, \widehat{C}_H)$, $\widehat{C}_h = \frac{1}{n}\sum_{i=1}^n \widehat{Z}_i \psi_h(Y_i)$, and $\widehat{\alpha} \in \mathbb{R}^{H \times H}$ is an estimator of α. The next theorem studies the asymptotic distribution of

$$\widehat{\Lambda}_{\text{TF1}} = n \sum_{k=1}^{p-m} \widehat{\lambda}_k^2,$$

where $(\widehat{\lambda}_1, ..., \widehat{\lambda}_p)$ are the singular values of $\widehat{C}\widehat{\alpha}$ arranged in ascending order.

Theorem 1.13. *Under H_0, assume that Z satisfies assumptions 1 and 4 (resp. 1 and 2) and has a finite second moment, then if $\psi_h \in L_2(\|Z\|^2)$ and $\sqrt{n}(\widehat{C}\widehat{\alpha} - C\alpha)$ has an asymptotic Gaussian distribution, we have*

$$\widehat{\Lambda}_{\mathrm{TF1}} \xrightarrow{\mathrm{d}} \sum_{k=1}^{H-d} \omega_k \xi_k,$$

where the ξ_k's are i.i.d. chi-squared variables with $p-d$ degrees of freedom and the ω_k's are the eigenvalues of the matrix $V_0^T \alpha^T \Delta \alpha V_0$ where

$$\Delta = \mathbb{E}\left[(p-d)^{-1}\|Q_c Z\|^2 (\Psi_H(Y) - \mathbb{E}[\Psi_H(Y)])(\Psi_H(Y) - \mathbb{E}[\Psi_H(Y)])^T\right]$$
$$(\textit{resp. } \Delta = \mathrm{var}(\Psi_H(Y))).$$

The above theorem is a general statement about the estimation of the dimension of E_c for TF1. Notice that the framework employed contains SIR and OF1 as special cases. We highlight in the following some relevant applications. Under CCV, considering the indicator functions and taking $\alpha = \mathrm{diag}_h\, p_h^{-1}$, we obtain the same result as [15], Corollary 1, regarding M_{SIR}. Besides, it is easy to show that CCV implies that $d_h = p_h(p-d)$, then if $\alpha = \mathrm{diag}_h\, d_h^{-1}$, we provide the asymptotic law of $\widehat{\Lambda}_{\mathrm{TF1}}$ for OF1, i.e.

$$\widehat{\Lambda}_{\mathrm{OF1}} \xrightarrow{\mathrm{d}} (p-d)^{-1} \chi^2_{(p-d)(H-d-1)}.$$

The above convergence highlights that, as SIR, OF1 provides a pivotal test for the considered statistic under CCV.

In general the asymptotic distribution of $\widehat{\Lambda}_{\mathrm{TF1}}$ is no longer chi-squared and the weights ω_k's need to be estimated. Theorem 1.13 emphasizes a pivotal version of such a test for any family of functions thanks to a good specification of the matrix α. For clarity assume that Δ is a full rank matrix, one can take $\alpha = \Delta^{-1/2}$ in Theorem 1.13 under DCV or CCV. We get for both

$$\widehat{\Lambda}_{\mathrm{TF1}} \xrightarrow{\mathrm{d}} \chi^2_{(p-d)(H-d)},$$

where α can be respectively estimated by

$$\frac{1}{n(p-d)} \sum_{i=1}^{n} \|\widehat{Q}_c \widehat{Z}_i\|^2 (\Psi_H(Y_i) - \overline{\Psi_H})(\Psi_H(Y_i) - \overline{\Psi_H})^T$$
$$\text{and} \quad \frac{1}{n} \sum_{i=1}^{n} (\Psi_H(Y_i) - \overline{\Psi_H})(\Psi_H(Y_i) - \overline{\Psi_H})^T,$$

where $\widehat{Q}_c$ is estimated from the considered TF1 method. Taking advantage of the SDR context, this kind of approach goes in the sense of the Wald-type pivotal statistic studied for instance in [12].

Using the same approach, it is possible to obtain the asymptotic distribution of such statistic for TF2. Nevertheless, such matrices are not positive and then the test needs to

be based on the sum of squares of the eigenvalues of M_{TF2}. In this case, the eigenvalues ω_k's in Theorem 1.C are more complicated than for TF1 even if we assume DCV or CCV. As a consequence it seems less attractive to follow the same path as previously. However one could follow [12], Theorem 1, to provide a consistent test, assuming sufficient finite moments for Z in order to ensure the convergence of $\widehat{\Lambda}$ on the one hand, and in order to estimate consistently the weights ω_k's on the other hand.

1.6 Simulations

In this section, we evaluate OF1, OF2 and some other SDR methods through different regression models. We first compare OF1 with SIR and IRE and then, we compare OF2 to some order 2 methods through pathological models for order 1 methods (see Example 1.10). To measure the performance of a method we evaluate the estimation error with the following distance : for two subspace E_1 and E_2, if P_1 and P_2 are their respective orthogonal projectors, the distance between E_1 and E_2 is

$$\mathrm{Dist}(E_1, E_2) = \|P_1 - P_2\|_{\mathrm{F}}. \tag{1.12}$$

In the following study, each method is evaluated for a single model. Each boxplot is based on 100 runs of the considered model. All along the simulation study, in order to appreciate the real intrinsic quality of each method, we assume that the variance and the mean of the predictors are known. As a consequence we do not take into account the bias introduced by poor estimates of the variance and the mean. Besides, we compare the distance (1.12) between the estimated standardized directions and the standardized CS.

For each method, when the response is continuous, we discretize its range into H slices, each containing the same number of observations. Both methods OF1 and OF2 require the iteration of the so called OF1 and OF2 algorithms (see section 1.3.2 and 1.4.2). In each case, the number of iterations equals 5. Finally, this simulation study is organized according to four examples that combine different distributions for the predictors.

1.6.1 OF1 and order 1 methods

The order 1 methods we computed include SIR and IRE. Let us consider the case where the predictors have a Gaussian distribution. Clearly $P_c Z$ and $Q_c Z$ are two independent random vectors and then $\mathbb{E}[\|Q_c Z\|^2|Y] = \mathbb{E}[\mathbb{E}[\|Q_c Z\|^2|P_c Z]|Y] = p - d$. Therefore $\mathrm{span}(M_{\mathrm{OF1}}) = \mathrm{span}(M_{\mathrm{SIR}})$ and OF1 is similar to SIR. Simulations made in this case highlight the similarity between the selected methods and are not presented here. Besides, to reach a point of view developed in the simulation study of [22], we are interested in the link between the variations of $\mathrm{var}(Z|Y)$ and the performance of the presented methods. Clearly, according to equation (1.9), the variations of the random variable $\mathbb{E}[\|Q_c Z\|^2|Y]$ emphasize the differences between SIR and OF1. Indeed if this one is a constant, then $d_h = \mathbb{E}[\|Q_c Z\|^2 \mathbb{1}_{\{Y \in I(h)\}}] = (p-d)p_h$ and OF1 is the same method as SIR. Consequently, SIR estimates are near optimal with respect to criterion (1.7) when the variations of

$\mathbb{E}[\|Q_c Z\|^2|Y]$ are near 0. Besides, if this random variable is nonconstant then also the d_h's and the differences between both methods are emphasized. Moreover, the random variables $\mathbb{E}[\|Q_c Z\|^2|Y]$ and $\mathrm{var}(Z|Y)$ are strongly linked, and as it was the case to distinguish IRE from SIR, the variations of $\mathrm{var}(Z|Y)$ play an important role to differentiate OF1 from SIR. Consequently, to point out the differences between these methods, we generate non-Gaussian predictors in the following two examples.

Example 1.14. Let $N_1 \in \mathbb{R}^p$, $N_2 \in \mathbb{R}^p$ be two independent standard Gaussian vectors, let ϵ be a Bernoulli random variable with mean 1/2. The predictor vector $X = (X^{(1)}, \ldots, X^{(p)})$ is generated as a Gaussian mixture through the equation

$$X = (\mu_1 + \sigma_1 N_1)\epsilon + N_2(1-\epsilon),$$

and it would be interesting to consider different values of $\sigma_1 \in \mathbb{R}$ and $\mu_1 \in \mathbb{R}^p$. We introduce the following models

$$\begin{aligned} &\text{Model I}: \qquad Y = \tanh(X^{(1)}/3) + 0.1e \\ &\text{Model II}: \qquad Y = X^{(2)}|1 + X^{(1)}/3| + e \end{aligned}$$

where $e \overset{\mathrm{d}}{=} \mathcal{N}(0,1)$. For Model I, an interesting parametrization is

$$\mu_1 = (a, 0, \ldots, 0)^T,$$

and then we can consider different values for a and σ_1. Such a distribution for the predictors induces two regimes. To highlight differences between both regimes respectively determined by $\epsilon = 1$ and $\epsilon = 0$, one can take the parameter a far from 0 and $\sigma_1 \neq 1$, say $a = 6$ and $\sigma_1 = .5$. Clearly the C_h's corresponding to small Y, have more chances to come from the second regime $\epsilon = 0$, which induces a poor estimate for such C_h's. On the contrary, the other C_h's tend to be well estimated. In this case the error of the SIR method is to uniformly weight these slices whereas OF1 does not. To be more comprehensive, we compute the methods with different parametrizations. The boxplots and the averages of the distance (1.12) between the standardized CS and its estimates over 100 simulated samples are given in figure 1.1. With the same model, in this figure, we also provide a graph to describe the effect of an increase of p.

For Model II, E_c has dimension 2 and then is more difficult to estimate. We consider

$$\mu_1 = (6, 2, \ldots, 0)^T,$$

and essentially, Model II provides similar graphs and interpretations as Model I. As a result, we analyze through this model the impact of an increase of n. The corresponding graph has been included in figure 1.1.

For each model and in all the parameter configurations, OF1 performs better than SIR. Between OF1 and IRE, the conclusion is quite a lot more mitigated. The chosen configurations reflect different kinds of difficulties. The situation presented in the first

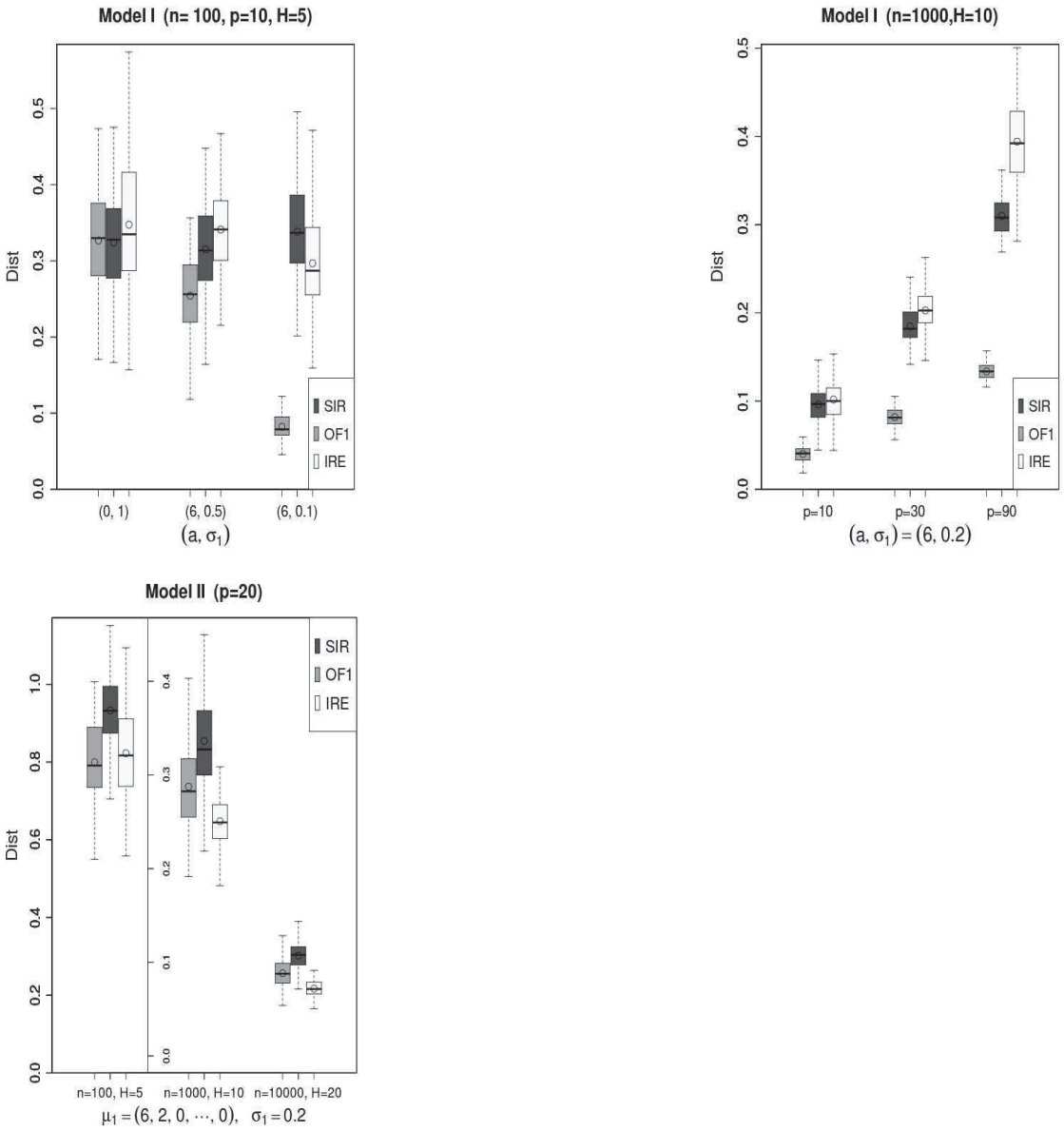

Figure 1.1 – Plot of the distance error for OF1, SIR and IRE in Example 1.14.

graph reflects a too small sample number $n = 100$ with respect to $p = 10$ to provide a good estimate. When $(\mu_1, \sigma_1) = (0, 1)$, the predictors are normally distributed and there are no significant differences between the methods. By increasing μ_1 and reducing σ_1, we move away from the Gaussian assumption and OF1 is the only one to improve its accuracy. Indeed, OF1 performs better than SIR and IRE around 86% of the time when $(\mu_1, \sigma_1) = (6, 0.5)$ and 100% of the time when $(\mu_1, \sigma_1) = (6, 0.1)$. Besides the second graph shows that OF1 is more robust to a high dimensional set-up. The most sensitive method to the increase of p is IRE because it requires the estimation of a large matrix. Finally, the last graph emphasizes that IRE is the most accurate when n is large.

Example 1.15. This example is interesting because it includes logistic models in the SDR framework. It is inspired from [22], Model A. We generalize their model by introducing some noise as described in the following. Let ϵ be a real random variable uniformly distributed on $\{1, 2, 3\}$, and let $N_1 \in \mathbb{R}^p$, $N_2 \in \mathbb{R}^p$, $N_3 \in \mathbb{R}^p$ be independent Gaussian vectors with respective moments $(\mu_1 \mathbf{1}, \sigma_1^2 I)$, $(\mu_2 \mathbf{1}, \sigma_2^2 I)$ and $(\mu_3 \mathbf{1}, \sigma_3^2 I)$ where $\mathbf{1} = (1, ..., 1)^T$. The vector X is generated as a Gaussian mixture through the equation

$$X = N_1 \mathbb{1}_{\{\epsilon=1\}} + N_2 \mathbb{1}_{\{\epsilon=2\}} + N_3 \mathbb{1}_{\{\epsilon=3\}},$$

and Y with the proportional-odds model defined by

$$\text{Model III :} \qquad Y = \sum_{j=1}^{3} j \mathbb{1}_{\{\pi_{j-1} \leq U \leq \pi_j\}},$$

with $U \overset{\mathrm{d}}{=} \mathcal{U}([0, 1])$ and the cumulative probability functions

$$\pi_0 = 0, \quad \pi_1 = \frac{\exp(\theta_1 - \mathbf{1}^T X)}{1 + \exp(\theta_1 - \mathbf{1}^T X)}, \quad \pi_2 = \frac{\exp(\theta_2 - \mathbf{1}^T X)}{1 + \exp(\theta_2 - \mathbf{1}^T X)}, \quad \pi_3 = 1.$$

First note that Model III implies that $Y = f(\mathbf{1}^T X, U)$ and as a consequence the CS exists for this kind of models. In our case, the CS is generated by the vector $\mathbf{1}$ and the CS is equal to the standardized CS. For clarity, we prefer to work with the mean and the standard error of the predictors divided respectively by p and $\sqrt{p}$ so that the mean and the standard error of $\mathbf{1}^T X$ do not depend on p. Working with the new scaled parameters, we fix $\mu_2 = 5$, $\mu_3 = 8$, $\sigma_2 = 0.5$, and $\sigma_3 = 0.5$. Then we can specify the cumulative probability functions by taking $\theta_1 = 3.5$ and $\theta_2 = 6.5$, so that it is realistic with respect to the means μ_2 and μ_3. To visualize such a model, one could draw in the same plot the cumulative probability functions π_1, $\pi_2 - \pi_1$, and $1 - \pi_2$, and the density of $\mathbf{1}^T X$. Each state of the response tends to correspond to some regime of the Gaussian mixture. The parameter H is fixed to 3, the number of states of the response. In figure 1.2, we test the accuracy of OF1, SIR and IRE facing Model III for different configurations of the parameters μ_1 and σ_1. The dimension p and the sample number n have been taken to provide neither a simple situation, nor a too difficult one.

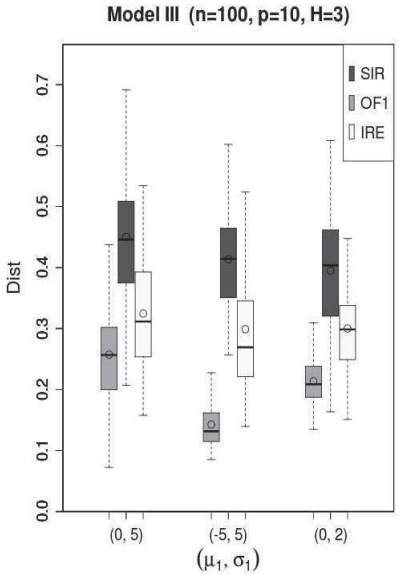

Figure 1.2 – Plot of the distance error for OF1 and SIR in Example 1.15.

In figure 1.2, the presented graph starts by a model with a lot of noise. The second and third situation reflects respectively a shift of the mean μ_1 and a shift of the variance σ_1. In each case, this reduces the noise and the estimation of the CS is more accurate for all the methods. Again when some estimated $\widehat{C}_h$'s have a small variance, OF1 manages to take advantage of the situation.

1.6.2 OF2 and order 2 methods

We compare several well-known order 2 dimension reduction methods with OF2. Order 2 methods we computed include SAVE, pHd, SIR-II and DR. For the considered models, pHd does not work as well as the others. Therefore we focus on a comparison between SAVE, DR, SIR-II and OF2. We computed the OF2 algorithm detailed in section 1.4.2 and the simulations we made truly argued in favor of its convergence : after 5 iterations the resulting matrix is nearly stable. It was also interesting to compare criterion (1.12) between the first iteration matrix and the final one. The difference between both was highly significant. Another important point is that OF2 is not as close to DR, SAVE and SIR-II as OF1 is close to SIR. The following simulations highlight this fact and we expect to have a large scope by providing many kinds of models with different parameter settings. We begin this section by providing the results obtained with Gaussian predictors.

Example 1.16. We consider the three following regression models

$$\text{Model IV}: \quad Y = \tanh\left(\frac{|X^{(1)}|}{2}\right) + 0.1e$$

$$\text{Model V}: \quad Y = 0.4(X^{(1)})^2 + \sqrt{|X^{(2)}|} + 0.2e$$

$$\text{Model VI}: \quad Y = 1.5X^{(1)}X^{(2)}\ e$$

with $e \overset{\text{d}}{=} \mathcal{N}(0,1)$ and $X \overset{\text{d}}{=} \mathcal{N}(0, I_p)$. The standardized CS and the CS of these models are equal. For model IV, the CS is spanned by $(1, 0, ..., 0)$, whereas in Model V and VI, it is a two dimensional subspace generated by $(1, 0, ..., 0)$ and $(0, 1, 0, ..., 0)$. We consider different parameter configurations for which every presented method is in a convenient situation. We compute SAVE, DR, SIR-II and OF2 with (n, p, H) equal to $(100, 6, 5)$, $(500, 10, 5)$ and $(1000, 20, 10)$. For each configuration, 100 simulated random samples have been generated and the resulting boxplots with their averages are presented in figure 1.3.

For all the selected models, OF2 performs better than all other methods. The most significant improvement happens for Model IV in which our method performs better than the others around 99% of the time in the setting $(100, 5, 6)$. When n increased, Of2 was never worse than the others. Note that for $n = 100$, 500, the average error of OF2 is two times smaller than the average error of DR, SAVE or SIR-II. For $n = 1000$ this factor goes to three. The results of the simulations for model V are really close to model IV. Model VI is a more complicated one for each method, we have to wait $n = 500$ to notice substantial differences in the distribution of the criterion. In every model, as n increases the improvement of OF2 is substantial. As a consequence and according to the plots in Figure 1.1 it seems clear that the asymptotic distribution of the distance error of OF2 has a smaller mean and variance than the other methods. Besides, for the selected models SAVE, DR and SIR-II perform in a similar way and are asymptotically equivalent.

Example 1.17. To conclude we present the results obtained with non-Gaussian but spherical predictors. Define $X = \rho U$ with U a uniformly distributed vector on the unit sphere of $\mathbb{R}^p$, independent of ρ, a real random variable. Clearly, X has a spherical distribution. Moreover, by taking

$$\rho = \epsilon\ |10 + 0.5N_1|\ +\ (1-\epsilon)\ |30 + 0.5N_2|,$$

with $N_1 \overset{\text{d}}{=} \mathcal{N}(0,1)$, $N_2 \overset{\text{d}}{=} \mathcal{N}(0,1)$ and $\epsilon \overset{\text{d}}{=} \mathcal{B}(\frac{1}{2})$, the distribution of X is far from a normal distribution. We study again Model VI but also the following ones,

$$\text{Model VII}: \quad Y = |X^{(1)}| + \left(\frac{X^{(2)}}{4}\right)^2 + 0.5e$$

$$\text{Model VIb}: \quad Y = X^{(1)}X^{(2)}\ e$$

where $e \overset{\text{d}}{=} \mathcal{N}(0,1)$. Model VI has been modified to reduce the signal to noise ratio. The directions to estimate, the parameter configurations and the number of simulated random

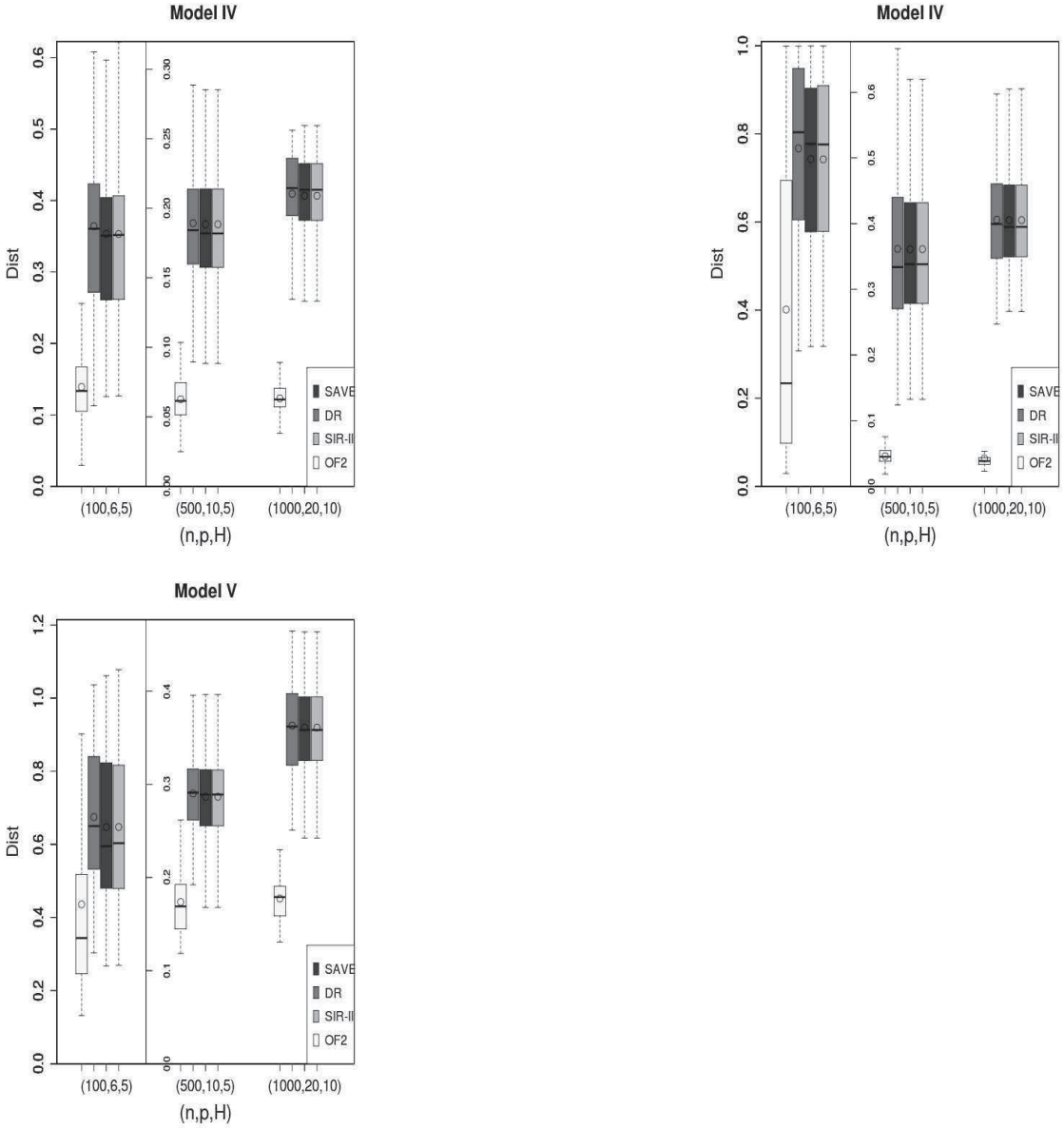

Figure 1.3 – Plot of the distance error for OF2, DR, SAVE and SIR-II in Example 1.16.

samples are the same as in the Gaussian case studied previously. Boxplots with their associated averages are presented in figure 1.4.

A general remark regarding Figure 1.4 is that the transition from normal to spherical predictors went well for OF2 comparing to other methods. Model IV still reflects the most important improvement of OF2. When n is large, it performs around eight times better than the others. In Model VIb, the accuracy of OF2 deteriorates by changing the distribution of the predictors from Gaussian into spherical. Finally, Model VII provides a standard new situation where the improvement of OF2 is highly significant.

In the development of OF2, model VI was of particular interest. Whether predictors are normal or spherical, OF2 is highly sensitive to the identification of the CS directions. For $n = 100$ the mean is less than the median, and it is no longer the case for n larger than 100. This marked change in the boxplots is explained by the presence of small outliers in the first situation and large outliers in the second one. Indeed as n is getting larger, OF2 performs better but however the mean is shifted by the presence of outliers that reflects uncommon difficult situations. This results from the eigenvector identification process described in Remark 1.12. Clearly OF2 relies on the way we identify eigenvectors of M_ψ that belong to E_c. To make that possible, a test of independence between the response and the projected predictors is conducted. Outliers of model VI for n equal to 500 and 1000 are the consequence of a bad eigenvector choice realized by this test. When n is sufficiently large this no longer occurs. When the OF2 algorithm is iterated more than 5 times, it happens only very few times.

1.7 Concluding remarks

The article introduces the basis of a new methodology for SDR. The introduction of some transformations of the response and the optimization with respect to these transformations were the original ideas of this work and have led us to some new methods of investigation in SDR. A surprising point was the high degree of similarity between SIR and OF1. As the simulations pointed out, it could be better to use OF1 when the intra-slice variance is nonconstant. IRE also behaves well in such situations but it has some problems when p is large because of the estimation of a large matrix. Our main contribution relates to order 2 methods, in particular we propose a new class of methods, TF2, that no longer needs the CCV assumption. Moreover, the simulation study sheds light on the high accuracy of OF2 over other order 2 methods. However, one can propose some lines of research that could improve the TF framework.

Regarding the estimation of the dimension, some prospects can be found in the Pearson's chi-squared statistic used in the OF2 algorithm (see Remark 1.12) to select the eigenvectors that belong to the CS. Clearly, this approach tries to take full advantage of the regression context offered by SDR. Work along this line to estimate the dimension of the CS is in progress and up until now simulations in this sense have provided good results.

Besides, both optimizations OF1 and OF2 do not take into account the estimation error on the variance and the mean of the predictors in the asymptotic decomposition of

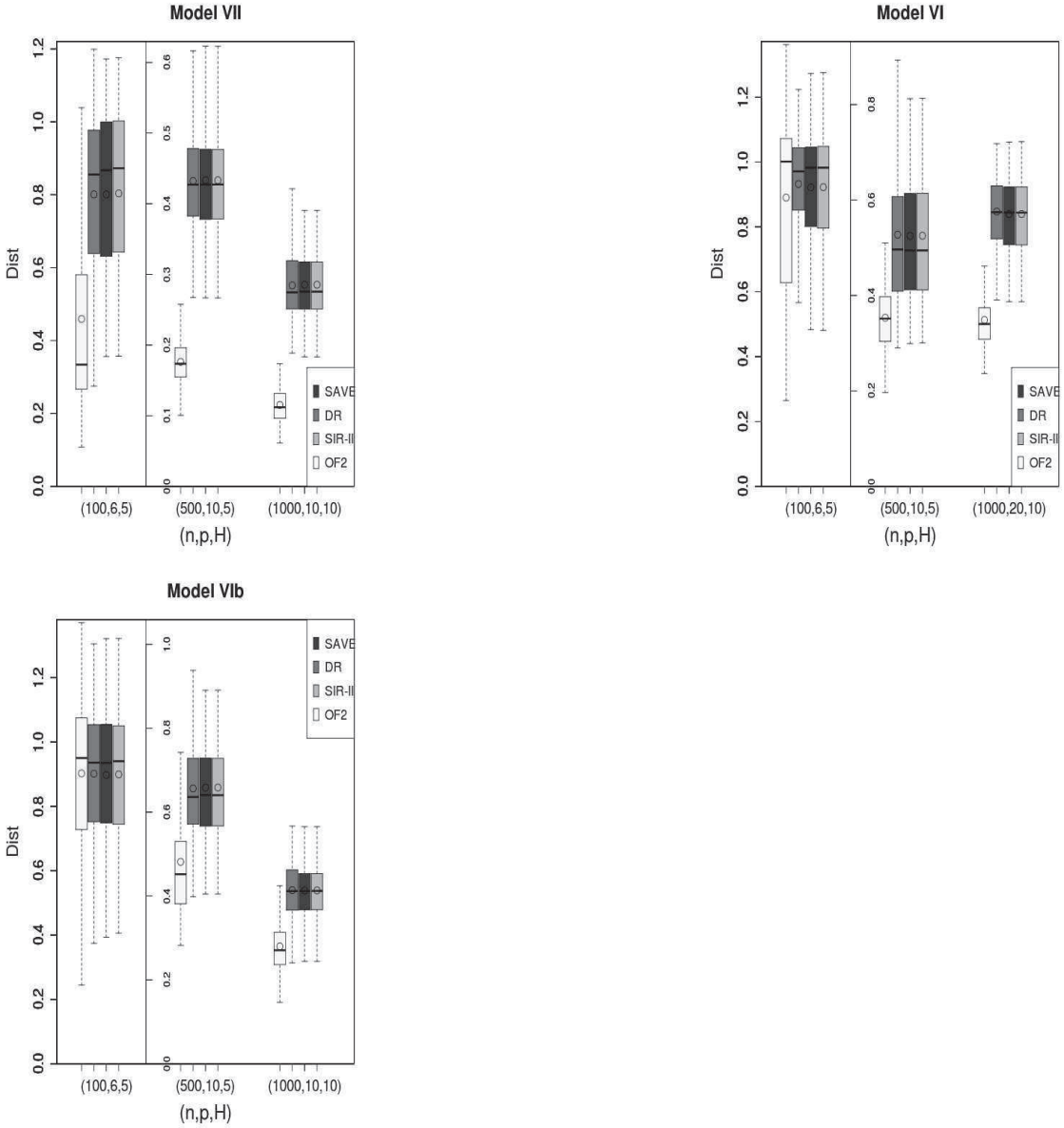

Figure 1.4 – Plot of the distance error for OF2, SAVE and DR in Example 1.17.

the criterion (1.7). This optimization leads to more complicated results that should be validated by simulations.

Finally, in many cases the regression function has different kinds of components, in particular there can be some pathological components for order 1 methods (see equation (1.10)). To handle such cases, one can calculate

$$M = \alpha M_1 + (1-\alpha) M_2,$$

where M_1 and M_2 are matrices of two different SDR methods. A spectral decomposition of M gives a hybrid estimate of the CS. Such ideas were recommended by [44], and [82] proposed a bootstrap method to select the parameter α. This includes the combinations of SIR and SAVE, SIR and pHd, SIR and SIR-II. Besides, it is commonly known that

$$M_{\mathrm{SAVE}} = \mathbb{E}[\mathrm{var}(Z|Y)^2] + M_{\mathrm{SIR}} - I,$$

and that

$$M_{\mathrm{DR}} = \mathbb{E}[\mathbb{E}[(ZZ^T|Y] - I)^2] + M_{\mathrm{SIR}}^2 + \mathrm{tr}(M_{\mathrm{SIR}}) M_{\mathrm{SIR}},$$

making SAVE and DR some combinations of SIR and order 2 methods. Therefore SAVE and DR do not only involve order 2 moments of Z, unlike TF2. Moreover TF1 only involves order 1 moments of Z. As a consequence, it seems more realistic to develop hybrid methods based on TF1 and TF2. Especially, the choice of the parameter α could be realized by the optimization of a well chosen criterion as has been done independently to derive OF1 and OF2.

1.8 Proofs and related results

1.8.1 Proofs of the stated results

Proof of Theorem 1.1. The standardization of the predictors does not change the presentation of this result, hence we present it for X. The proof is divided into three principal parts : we first give a lemma about the intersection of two MDRS, then we apply it to prove the statement of the theorem about the CMS, finally using this last result we conclude the proof for the CS.

Lemma 1.18. *If the restriction of X to the ball of $\mathbb{R}^p$ with radius r and center x_0 has a strictly positive density, then the intersection of two MDRS is a MDRS on this ball, i.e.*

$$(\mathbb{E}[Y|X] - \mathbb{E}[Y|RX]) \mathbb{1}_{\{X \in B(x_0, r)\}} = 0 \quad \text{a.s.},$$

where R denotes the orthogonal projector onto their intersection.

Proof. We first make the proof for a ball centered at 0, and then we apply it to $X - x_0$. Let E and E' be two MDRS, P and P' their respective orthogonal projectors, and R the orthogonal projector onto $E \cap E'$. Using the definition of a MDRS,

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|PX] = \mathbb{E}[Y|P'X] \quad \text{a.s.}$$

Let $g(PX)$ and $h(P'X)$ denote the last two random variables in the preceding equation. Using that X has a strictly positive density on the unit sphere, we can write

$$g(Px) = h(P'x) \quad \text{a.e. on } B(0,r). \tag{1.13}$$

Let $\varepsilon > 0$, and φ_k be a unit approximation with compact support $B(0,\varepsilon)$, we define the function $f_k : B(0,r) \to \mathbb{R}$ such that

$$f_k(x) = (g \circ P) * \varphi_k\ (x).$$

Then, we have for all x,

$$f_k(x) = \int g(P(x-y))\varphi_k(y)\mathrm{d}y = f_k(Px).$$

Moreover, for all $x \in B(0, r-\varepsilon)$, since in the above integral $x - y \in B(0,r)$, using (1.13) we derive

$$f_k(x) = (h \circ P') * \varphi_k\ (x),$$

and similarly we obtain $f_k(x) = f_k(P'x)$. Since $f_k(x) = f_k(Px) = f_k(P'x)$, a simple iteration process provides for all $x \in B(0, r-\varepsilon)$,

$$f_k(x) = f_k((PP')^n x).$$

Since f_k is a continuous function and $\lim\limits_{n\to+\infty}(PP')^n = R$, we have

$$f_k(x) = f_k(Rx), \quad x \in B(0, r-\varepsilon).$$

To conclude, the unit approximation theorem gives us the convergence

$$f_k \circ R \xrightarrow{\mathrm{L}_1} g \circ P.$$

Thus, from $f_k(RX)$ we can derive a subsequence $f_{n_k}(RX)$ that converges almost surely to $g(PX)$, proving that $\mathbb{E}[Y|X]$ is a function of RX. This completes the first part of the proof.

Now suppose that X has a strictly positive density onto the ball of radius r and center x_0. Define $\widetilde{X} = X - x_0$, it is clear that a MDRS for X is also a MDRS for $\widetilde{X}$ and conversely. Then, since $\widetilde{X}$ is centered in 0, the intersection of two MDRS is still a MDRS for $\widetilde{X}$ and obviously for X. □

Existence of the CMS. Denote by $F \subset \mathbb{R}^p$ the support of the density of X. A first step consists of showing that its interior $\mathring{F}$ can be covered by a countable number of balls included in $\mathring{F}$. Secondly, we apply Lemma 1.18 to each of this balls to obtain that the intersection of two MDRS on $\mathring{F}$ is a MDRS on $\mathring{F}$. Finally, the uniqueness is shown.

Let $x \in \mathring{F}$, then there exists $r > 0$ such that $B(x,r) \subset \mathring{F}$. It is possible to find a ball, with rational center and radius, included in $B(x,r)$ and containing x. Thus any x of $\mathring{F}$ is contained in a ball with center and radius rational that is included in $\mathring{F}$. In other

words, the set A formed by all the balls $B(x_q, r_q) \subset \mathring{F}$, with x_q and r_q rationals, covers $\mathring{F}$. Therefore, by applying Lemma 1.18, we have for all $B(x_q, r_q) \in A$,

$$|\mathbb{E}[Y|X] - \mathbb{E}[Y|RX]|\mathbb{1}_{\{X \in B(x_q, r_q)\}} = 0 \quad \text{a.s.},$$

since A is a countable set,

$$\sum_{(x_q, r_q) \in A} |\mathbb{E}[Y|X] - \mathbb{E}[Y|RX]|\mathbb{1}_{\{X \in B(x_q, r_q)\}} = 0 \quad \text{a.s.},$$

then,

$$|\mathbb{E}[Y|X] - \mathbb{E}[Y|RX]| \sum_{(x_q, r_q) \in A} \mathbb{1}_{\{X \in B(x_q, r_q)\}} = 0 \quad \text{a.s.}$$

By assumption $\mathbb{P}(X \in \mathring{F}) = 1$, then the right-hand side is almost surely strictly positive, and thus

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|RX] \quad \text{a.s.}$$

Consequently, the intersection of two MDRS is a MDRS. To complete the proof, assume that two MDRS have minimum dimension. Their intersection has at least minimum dimension because it is a MDRS. So they are equal.

Existence of the CS. Using similar arguments about the dimension of vector spaces, we only need to show that the intersection of two DRS is a DRS. Let E and E' be two DRS. By equations (1.2) and (1.3), E and E' are also MDRS for the random variables $\mathbb{1}_{Y \in A}$ and X. We have just showed that the intersection of two MDRS is a MDRS. Then for all measurable sets A, $E \cap E'$ is a MDRS for $\mathbb{1}_{Y \in A}$ and X. Equivalently, $E \cap E'$ is a DRS. □

Proof of Theorem 1.3. Assumption 3 implies that $\{\mathbb{E}[Z\mathbb{E}[Z^{(k)}|Y]],\ k = 1, ..., p\}$ generates E_c. First, let us show that any vector of this family can be approximated by $\mathbb{E}[Z\phi(Y)]$, where ϕ is a linear combination of functions in Ψ. Let $\varepsilon > 0$ and $k \in \{1, ..., p\}$, since Ψ is a total family in $L_1(\|Z\|)$, there exists ϕ_k a finite linear combination of functions in Ψ such that

$$\mathbb{E}\left[\|Z\|\ |\phi_k(Y) - \mathbb{E}[Z^{(k)}|Y]|\right] \leq \varepsilon,$$

besides, we have

$$\|\mathbb{E}[Z\phi_k(Y)] - \mathbb{E}[Z\mathbb{E}[Z^{(k)}|Y]]\| \quad \leq \quad \mathbb{E}\left[\|Z\|\ \left|\phi_k(Y) - \mathbb{E}[Z^{(k)}|Y]\right|\ \right],$$

and therefore,

$$\|\mathbb{E}[Z\phi_k(Y)] - \mathbb{E}[Z\mathbb{E}[Z^{(k)}|Y]]\| \leq \varepsilon. \tag{1.14}$$

Here an important point is that $\mathbb{E}[Z\phi_k(Y)] \in E_c$, it implies that

$$\text{span}\left(\mathbb{E}[Z\phi_k(Y)],\ k = 1, ..., p\right) \subset \text{span}(M_{\text{SIR}}), \tag{1.15}$$

Moreover, (1.14) and the continuity of the determinant involve that the rank of the set of vectors $\mathbb{E}[Z\phi_k(Y)]$'s is equal to d if ε is small enough. Then, instead of an inclusion

(1.15) becomes an equality and we complete the proof by recalling that each ϕ_k is a linear combination of a finite number of functions in Ψ. □

Proof of Proposition 1.5. We first calculate the expectation of the limit in law of the sequence $n\,\mathrm{tr}(Q_c\widehat{P}_c)$ and then we solve the optimization problem. Since

$$n\,\mathrm{tr}(Q_c\widehat{P}_c) = n\,\mathrm{tr}(\widehat{\eta}^T Q_c\widehat{\eta}\ (\widehat{\eta}^T\widehat{\eta})^{-1}) = \mathrm{tr}(\sqrt{n}(\widehat{\eta}^T-\eta^T)Q_c\sqrt{n}(\widehat{\eta}-\eta)(\widehat{\eta}^T\widehat{\eta})^{-1}),$$

Slutsky's theorem and the continuity of the operator $\mathrm{tr}(\cdot)$ provide that $n\,\mathrm{tr}(Q_c\widehat{P}_c)$ converges to $\mathrm{tr}(\delta^T Q_c\delta)$ in distribution, where $\delta\in\mathbb{R}^{p\times d}$ is the limit in law of the sequence $\sqrt{n}(\widehat{\eta}-\eta)$, i.e. a normal vector with mean 0 (we can get ride of the quantity $(\widehat{\eta}^T\widehat{\eta})^{-1}$ because of the constraint and $\widehat{\eta}^T\widehat{\eta}\xrightarrow{\mathbb{P}}\eta^T\eta$). Thus it remains to calculate the expectation of this limit, notice that

$$\mathbb{E}\left[W_\alpha\right] = \mathbb{E}\left[\mathrm{tr}(\delta^T Q_c\delta)\right] \quad = \quad \sum_{k=1}^{d}\mathrm{tr}\left(Q_c\mathbb{E}[\delta_k\delta_k^T]\right),$$

where δ_k stands for the limit in law of the sequence $\sqrt{n}(\widehat{\eta}_k-\eta_k)$. Finally, since its variance is equal to $\mathrm{var}(Z\psi_k(Y))$ and using the linearity condition, we find that

$$\mathbb{E}\left[W_\alpha\right] = \sum_{k=1}^{d}\mathbb{E}\left[\|Q_cZ\|^2\psi_k(Y)^2\right].$$

Now let us formulate the minimization problem with respect to the matrix α. Using that the $I(h)$ are pairwise disjoint, we have

$$\mathbb{E}\left[W_\alpha\right] = \sum_{k=1}^{d}\alpha_k^T\mathbb{E}[\|Q_cZ\|^2\mathbb{1}_Y\mathbb{1}_Y^T]\alpha_k = \mathrm{tr}(\alpha^T D\alpha),$$

and also,

$$\eta^T\eta = \alpha^T C^T C\alpha = (D^{\frac{1}{2}}\alpha)^T G D^{\frac{1}{2}}\alpha.$$

From both previous equations, we set out the equivalent minimization problem

$$\min_{\alpha}\ \mathrm{tr}\left(\alpha^T D\alpha\right) \quad \text{u.c.} \quad (D^{\frac{1}{2}}\alpha)^T G D^{\frac{1}{2}}\alpha = Id,$$

then, from the variable change $U = V^T D^{\frac{1}{2}}\alpha$ we derive

$$\min_{U}\ \mathrm{tr}(U^TU) \quad \text{u.c.} \quad U^T\begin{pmatrix} D_0 & 0\\ 0 & 0\end{pmatrix}U = Id.$$

By writing $U^T = (U_1^T, U_2^T)$ we notice that there is no constraint on U_2, which implies that $U_2 = 0$. Consequently, it remains to solve

$$\min_{U_1}\ \mathrm{tr}(U_1U_1^T) \quad \text{u.c.} \quad U_1U_1^T = D_0^{-1},$$

where $U_1 \in \mathbb{R}^{d\times d}$, and where the quantity to minimize is fixed by the constraint. Then, a solution is $U_1 = D_0^{-\frac{1}{2}} H$ where H is any orthogonal matrix. Hence, the solution of the minimization problem is

$$\alpha = D^{-\frac{1}{2}} VU = D^{-\frac{1}{2}} V_1 D_0^{-\frac{1}{2}} H.$$

□

Proof of Lemma 1.6. Let us begin in the easiest way : (2) ⇒ (1). Let H be any orthonormal matrix as described in (1). Because $HQ_cH^T = I - HP_cH^T = Q_c$, by multiplying (2) on the left side by H and on the right side by H^T, we find that

$$\operatorname{var}(HZ|P_cZ) = \lambda_\omega^* Q_c = \operatorname{var}(Z|P_cZ).$$

The other way is based on a good choice of the matrix H. Let γ be a unit vector of $E_c^\perp$, and define $H = I - 2\gamma\gamma^T$. Clearly, H is symmetric and satisfies to the requirement of (1). So that, we have

$$\operatorname{var}(Z|P_cZ) = (I - 2\gamma\gamma^T) \operatorname{var}(Z|P_cZ)(I - 2\gamma\gamma^T),$$

developing the right hand side, it follows that

$$\operatorname{var}(Z|P_cZ)\gamma\gamma^T = 2\operatorname{var}(\gamma^T Z|P_cZ)\gamma\gamma^T - \gamma\gamma^T \operatorname{var}(Z|P_cZ),$$

and finally, multiplying by γ on the right, we find

$$\operatorname{var}(Z|P_cZ)\gamma = \operatorname{var}(\gamma^T Z|P_cZ)\gamma. \tag{1.16}$$

Therefore, any $\gamma \in E_c^\perp$ is an eigenvector of the matrix $\operatorname{var}(Z|P_cZ)$ and thus, $E_c^\perp$ is included in an eigenspace of this matrix. Denote by λ_ω^* the eigenvalue associated with $E_c^\perp$. Since the columns of Q_c are vectors of $E_c^\perp$, we have

$$\operatorname{var}(Z|P_cZ)Q_c = \lambda_\omega^* Q_c,$$

which implies that

$$\operatorname{var}(Z|P_cZ) = \operatorname{var}(Q_cZ|P_cZ) = \lambda_\omega^* Q_c,$$

and (1) ⇒ (2) is completed.

The value of λ_ω^* can be given by equation (1.16). Under the linearity condition we have for every unit vector $\gamma \in E_c^\perp$,

$$\lambda_\omega^* = \operatorname{var}(\gamma^T Z|P_cZ) = \mathbb{E}[(\gamma^T Z)^2|P_cZ],$$

and hence it suffices to take $\gamma = \frac{1}{\sqrt{p-d}} \sum_{k=1}^{p-d} \gamma_k$ where $(\gamma_1, ..., \gamma_{p-d})$ is an orthonormal basis of $E_c^\perp$, to obtain

$$\lambda_\omega^* = \frac{1}{p-d} \mathbb{E}\left[\|Q_cZ\|^2|P_cZ\right].$$

□

Proof of Theorem 1.8. To make a complete proof, we need to show that all the vectors in $E_c^\perp$ are eigenvectors of the symmetric matrix $M_\psi - \lambda_\psi^* I$ associated with the eigenvalue 0. The existence of the CS ensures that

$$M_\psi - \lambda_\psi^* I = \mathbb{E}[(\mathbb{E}[ZZ^T|P_c Z] - \lambda_\omega^* I)\psi(Y)],$$

besides, thanks to the linearity condition and DCV, we have

$$\mathbb{E}[ZZ^T|P_c Z] = \lambda_\omega^* Q_c + P_c Z Z^T P_c.$$

Thus, for any $\gamma \in E_c^\perp$ we have $(M_\psi - \lambda_\psi^* I)\gamma = 0$ and the proof is completed. □

Proof of Theorem 1.9. The proof relies on Lemma 1.D and Lemma 1.E. Both are results about vector spaces of non-invertible matrices. For clarity and since it does not deal directly with the subject of the paper, we state and prove these lemmas in Appendix B.

Let Ψ be a total countable family in $L_1(\|Z\|^2)$, Theorem 1.8 indicates that $E_c^\perp \subset E_\psi^\perp$ for any $\psi \in \Psi$. Then it suffices to show that there exists ψ a finite linear combination of functions in Ψ such that $\dim(E_\psi) = \operatorname{rank}(M_\psi - \lambda_\psi^* I) = d$. In the basis (P_1, P_2), where P_1 and P_2 are respectively bases of E_c and $E_c^\perp$, the matrix $M_\psi - \lambda_\psi^* I$ can be written as

$$\begin{pmatrix} N_\psi & 0 \\ 0 & 0 \end{pmatrix},$$

with $N_\psi = P_1^T(M_\psi - \lambda_\psi^*)P_1$. Notice that the space

$$\mathcal{M} = \{N_\psi,\ \psi = \sum_h \alpha_h \psi_h\},$$

is a vector space of symmetric matrices with dimension $d \times d$. In the basis (P_1, P_2), Assumption 5 becomes

$$\forall \eta \in \mathbb{R}^d, \quad \mathbb{P}(\eta^T N_Y \eta = 0) < 1,$$

with $N_Y = P_1^T(M_Y - \lambda_Y^*)P_1$. Clearly, this implies that

$$\forall \eta \in \mathbb{R}^d, \quad \exists \psi, \quad \eta^T N_\psi \eta \neq 0, \tag{1.17}$$

and because Ψ is a total family in $L_1(\|Z\|^2)$, the function ψ in the previous equation could be a finite linear combination of functions in Ψ and then $N_\psi \in \mathcal{M}$. Thus to conclude the proof, one can notice that given a vector subspace $\mathcal{M} \subset \mathbb{R}^{d\times d}$ of symmetric matrices, if (1.17) holds, then there exists an invertible matrix in $\mathcal{M}$. This assertion is true because it is the contrapositive of the statement of Lemma 1.E. □

Proof of Corollary 1.10. From Theorem 1.9 we have $E_\psi = E_c$ where $\psi = \sum_{h=1}^H \alpha_h \psi_h$. Hence, we need to show that $E_\psi \subset \oplus E_{\psi_h}$ since the other inclusion is trivial. Suppose

that there exists $\eta \in E_\psi$ with norm 1 such that $\eta \perp \oplus E_{\psi_h}$. Then by definition, for every $h = 1, ..., H$, we have

$$M_{\psi_h}\eta = \lambda^*_{\psi_h}\eta,$$

and we can obtain

$$M_\psi\eta = \sum_{h=1}^{H} \alpha_h \lambda^*_{\psi_h}\eta = \lambda^*_\psi\eta,$$

which is impossible because $\eta \in E_\psi$. □

Proof of Proposition 1.11. We have

$$\begin{aligned} Q_c\widehat{P}_c &= Q_c(\widehat{P}_c - P_c) \\ &= Q_c \oint_{\mathcal{C}} (Iz - \widehat{M}_\psi)^{-1} - (Iz - M_\psi)^{-1}\mathrm{d}z \\ &= Q_c \oint_{\mathcal{C}} (Iz - \widehat{M}_\psi)^{-1}(M_\psi - \widehat{M}_\psi)(Iz - M_\psi)^{-1}\mathrm{d}z, \end{aligned}$$

and then, we can obtain

$$\begin{aligned} Q_c\widehat{P}_c = Q_c \oint_{\mathcal{C}} &(Iz - M_\psi)^{-1}(M_\psi - \widehat{M}_\psi)(Iz - M_\psi)^{-1}\mathrm{d}z \\ &+ Q_c \oint_{\mathcal{C}} (Iz - \widehat{M}_\psi)^{-1}(M_\psi - \widehat{M}_\psi)(Iz - M_\psi)^{-1}(M_\psi - \widehat{M}_\psi)(Iz - M_\psi)^{-1}\mathrm{d}z. \end{aligned}$$

Consider the trace of the first term in the above equation, since Q_c and $(Iz - M_\psi)^{-1}$ commute we have

$$\begin{aligned} \mathrm{tr}\left(Q_c \oint_{\mathcal{C}} (Iz - M_\psi)^{-1}(M_\psi - \widehat{M}_\psi)\ (Iz - M_\psi)^{-1}\mathrm{d}z \right) \\ = \mathrm{tr}\left((M_\psi - \widehat{M}_\psi) \oint_{\mathcal{C}} Q_c(Iz - M_\psi)^{-2}\mathrm{d}z \right). \end{aligned}$$

Besides, it is clear that

$$Q_c(Iz - M_\psi)^{-1} = \frac{Q_c}{(z - \lambda^*_\psi)}, \tag{1.18}$$

and recalling that λ^*_ψ is outside $\mathcal{C}$, we have $\oint_{\mathcal{C}} \frac{1}{(z-\lambda^*_\psi)^{-2}}\mathrm{d}z = 0$ and we get

$$\begin{aligned} \mathrm{tr}\left(Q_c\widehat{P}_c\right) = \mathrm{tr}\Bigg(Q_c \oint_{\mathcal{C}} &(Iz - \widehat{M}_\psi)^{-1}(M_\psi - \widehat{M}_\psi) \\ &(Iz - M_\psi)^{-1}(M_\psi - \widehat{M}_\psi)(Iz - M_\psi)^{-1}\mathrm{d}z \Bigg). \end{aligned}$$

Denote by Δ the limit in law of $\sqrt{n}(\widehat{M}_\psi - M_\psi)$, since $\widehat{M}$ goes to M in probability, Slutsky's Theorem implies the convergence $n\,\mathrm{tr}(Q_c\widehat{P}_c) \xrightarrow{\mathrm{d}} W_\psi$ with

$$W_\psi = \mathrm{tr}\left(Q_c \oint_{\mathcal{C}} (Iz - M_\psi)^{-1}\Delta(Iz - M_\psi)^{-1}\Delta(Iz - M_\psi)^{-1}\mathrm{d}z\right).$$

Now we use equation (1.18) to obtain

$$W_\psi = \mathrm{tr}\left(Q_c\Delta \oint_{\mathcal{C}} \frac{(Iz - M_\psi)^{-1}}{(z - \lambda_\psi^*)^2}\mathrm{d}z\Delta Q_c\right), \tag{1.19}$$

where the above integral can be calculated in the following way. Splitting it into two terms and using (1.18), we have

$$\begin{aligned}\oint_{\mathcal{C}} \frac{(Iz - M_\psi)^{-1}}{(z - \lambda_\psi^*)^2}\mathrm{d}z &= \oint_{\mathcal{C}} \frac{P_c(Iz - M_\psi)^{-1}}{(z - \lambda_\psi^*)^2}\mathrm{d}z + \oint_{\mathcal{C}} \frac{Q_c(Iz - M_\psi)^{-1}}{(z - \lambda_\psi^*)^2}\mathrm{d}z \\ &= \oint_{\mathcal{C}} \frac{P_c(Iz - M_\psi)^{-1}}{(z - \lambda_\psi^*)^2}\mathrm{d}z + Q_c \oint_{\mathcal{C}} \frac{1}{(z - \lambda_\psi^*)^3}\mathrm{d}z.\end{aligned}$$

It is not difficult to show that the last term in the previous equation equals 0. Regarding the first term, since for every $k \in \{1, ..., d\}$ we have

$$\begin{aligned}P_c \oint_{\mathcal{C}} \frac{(Iz - M_\psi)^{-1}}{(z - \lambda_\psi^*)^2}\mathrm{d}z\eta_k &= \eta_k \oint_{\mathcal{C}} \frac{(z - \lambda_\psi(\eta_k))^{-1}}{(z - \lambda_\psi^*)^2}\mathrm{d}z \\ &= \frac{\eta_k}{(\lambda_\psi(\eta_k) - \lambda_\psi^*)^2} = P_c(P_cM_\psi - I\lambda_\psi^*)^{-2}\eta_k,\end{aligned}$$

and since all the vectors in $E_c^\perp$ belong to the kernel of this matrix, we get

$$P_c \oint_{\mathcal{C}} \frac{(Iz - M_\psi)^{-1}}{(z - \lambda_\psi^*)^2}\mathrm{d}z = P_c(P_cM_\psi - I\lambda_\psi^*)^{-2}.$$

Injecting it in (1.19), we obtain

$$W_\psi = \mathrm{tr}\left(\Delta Q_c\Delta P_c(P_cM_\psi - I\lambda_\psi^*)^{-2}\right),$$

and it remains to calculate its expectation. The linearity condition implies that $Q_cM_\psi P_c = 0$, and we have

$$\mathbb{E}[\Delta Q_c\Delta P_c] = \lim_{n\to+\infty} n\mathbb{E}\left[\widehat{M}_\psi Q_c\widehat{M}_\psi P_c\right] = \mathbb{E}[ZZ^T P_c\|Q_cZ\|^2\psi(Y)^2],$$

which completes the proof. □

Proof of Theorem 1.13. The proof involves a result in [12], stated in Appendix 1.8.2 as Theorem 1.C.

By applying Theorem 1.C to the matrix $\widehat{C}\widehat{\alpha}$, one can notice that the asymptotic distribution of $\widehat{\Lambda}_{\mathrm{TF1}}$ depends only on the variance of the asymptotic law of

$$\sqrt{n}\,\mathrm{vec}(U_0^T(\widehat{C}\widehat{\alpha} - C\alpha)V_0).$$

Let W be a random vector following this distribution. By the linearity condition, we have

$$U_0^T(\widehat{C}\widehat{\alpha} - C\alpha)V_0 = U_0^T\widehat{C}\widehat{\alpha}V_0 = U_0^T\widehat{\Sigma}^{-1/2}(\widehat{\Sigma}^{-\frac{1}{2}}\widehat{C} - \widehat{\Sigma}^{-\frac{1}{2}}C)\widehat{\alpha}V_0.$$

Since $C\alpha V_0 = 0$, $\widehat{\alpha} \xrightarrow{\mathbb{P}} \alpha$ and $\widehat{\Sigma} \xrightarrow{\mathbb{P}} \Sigma$, by Slutsky's theorem W has the same law as the asymptotic distribution of

$$\sqrt{n}\,\mathrm{vec}(U_0^T\Sigma^{-1/2}\widehat{\Sigma}^{-\frac{1}{2}}\widehat{C}\alpha V_0),$$

By the linearity condition $U_0\Sigma^{-1/2}X_i = U_0\Sigma^{-1/2}(X_i - \mathbb{E}[X]) = U_0Z_i$, and one can obtain

$$U_0^T\Sigma^{-1/2}\widehat{\Sigma}^{-\frac{1}{2}}\widehat{C}\alpha V_0 = U_0^T(\overline{Z\Psi_H^T(Y)} - \overline{Z}\ \overline{\Psi_H^T(Y)})\alpha V_0.$$

We notice that

$$\sqrt{n}(\overline{Z\Psi_H^T(Y)} - \overline{Z}\ \overline{\Psi_H^T(Y)}) = \sqrt{n}\left(\overline{Z(\Psi_H^T(Y) - \mathbb{E}[\Psi_H^T(Y)])}\right) + o_{\mathbb{P}}(1),$$

and we provide the decomposition

$$\sqrt{n}\,\mathrm{vec}(U_0^T(\widehat{C}\widehat{\alpha} - C\alpha)V_0) = (V_0^T\alpha^T \otimes U_0^T)\sqrt{n}\,\mathrm{vec}\left(\overline{Z\Phi(Y)}\right) + o_{\mathbb{P}}(1),$$

with the notation $\Phi(Y) = \Psi_H(Y) - \mathbb{E}[\Psi_H(Y)]$. By the central limit theorem, we get

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{vec}(Z_i\Phi(Y_i)^T) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \mathrm{var}(\Phi(Y)\otimes Z)\right).$$

Clearly, using the linearity condition we have

$$\mathrm{var}(W) = (V_0^T\alpha^T \otimes I)\mathbb{E}[\Phi(Y)\Phi(Y)^T \otimes (U_0^TZZ^TU_0)](\alpha V_0 \otimes I).$$

Under DCV one can get

$$\mathbb{E}[\Phi(Y)\Phi(Y)^T \otimes (U_0^TZZ^TU_0)] = \mathbb{E}[(p-d)^{-1}\|Q_cZ\|^2\Phi(Y)\Phi(Y)^T \otimes I_{p-d}],$$

under CCV one can obtain

$$\mathbb{E}[\Phi(Y)\Phi(Y)^T \otimes (U_0^TZZ^TU_0)] = \mathbb{E}[\Phi(Y)\Phi(Y)^T \otimes I_{p-d}],$$

and the conclusion follows.

□

1.8.2 Few results

Theorem 1.A. (Bryc (1995) [11], Theorem 4.1.4, p.48) *Let Z be a random vector of $\mathbb{R}^p$ ($p \geq 2$) with a finite second order moment. If Z is spherical and if $\mathrm{var}(Z|PZ) = const.$ for some orthogonal projector P, then Z is normal and conversely.*

Theorem 1.B. (Coudène (2002) [24]) *Let $p \in [0, +\infty[$, μ a Borel probability measure on $[0,1]$, and $f_n : [0,1] \to \mathbb{R}$ a family of bounded measurable functions that separates the points :*

$$\forall x, y \in [0,1],\ x \neq y,\ \exists n \in \mathbb{N} \quad \text{such that} \quad f_n(x) \neq f_n(y).$$

Then the algebra spanned by the functions f_n's and the constants is dense in $L_p([0,1], \mu)$.

Theorem 1.C. (Bura and Yang (2011) [12]) *Assume* $\mathrm{rank}(M) = d$ *and that* $\sqrt{n}\,\mathrm{vec}(\widehat{M} - M) \xrightarrow{d} \mathcal{N}(0, \Gamma)$. *Then*

$$\widehat{\Lambda} \xrightarrow{\mathrm{d}} \sum_{k=1}^{s} \omega_k X_k^2,$$

where the X_k's are independent standard normal random variables and the ω_k's are the ordered eigenvalues of $(V^T \otimes U^T)\Gamma(V \otimes U)$, with $s = \min(\mathrm{rank}(\Gamma),\ (p-d)(H-d))$ and U and V are respectively basis of the left and right singular spaces of M associated with the singular value 0.

The following lemma deals with vector space structure and rank-deficient matrices. We refer to [33], Proposition 3 for a more general approach. In particular, this lemma implies Lemma 1.E which has a central place in the proof of Theorem 1.9.

Lemma 1.D. *Let M, $N \in \mathbb{R}^{d \times d}$ and $\alpha_0 > 0$. If* $\mathrm{rank}(N + \alpha M) \leq \mathrm{rank}(N)$ *for all $\alpha \leq \alpha_0$, then we have*

$$M \ker(N) \subset \mathrm{Im}(N).$$

Proof. Denote by P_α the characteristic polynomial of $N + \alpha M$ and define $r_\alpha = \mathrm{rank}(N + \alpha M)$ and $k_\alpha = \dim(\ker(N + \alpha M)) = d - r_\alpha$. Because of the continuity of the determinant, the coefficients of P_α converge to the coefficients of P_0, then P_α converges uniformly to P_0 on every compact. By the definition of k_0, P_0 is such that

$$P_0(x) = x^{k_0} Q_0(x) \quad \text{with } Q_0(0) \neq 0.$$

Now we use the uniform convergence. For α small enough we have $P_\alpha^{(k_0)}(0) \neq 0$, and this gives the upper bound $k_\alpha \leq k_0$. Using the assumption we obtain $k_0 = k_\alpha$. Therefore, for some α_0, we have

$$Q_\alpha(0) \neq 0, \quad \alpha \leq \alpha_0.$$

Clearly, there exists a contour $\mathcal{C}$ such that none of the nonzero eigenvalues of $N + \alpha M$ belong to $\mathcal{C}$, $\alpha \leq \alpha_0$. Using the residue theorem, we can express the orthogonal projectors Π_0 and Π_α on the kernel of the matrices N and $N + \alpha M$ as follows,

$$\Pi_0 = \oint_{\mathcal{C}} (N - zI)^{-1} \mathrm{d}z, \quad \text{and} \quad \Pi_\alpha = \oint_{\mathcal{C}} (N + \alpha M - zI)^{-1} \mathrm{d}z,$$

and one can get

$$\Pi_0 - \Pi_\alpha = \alpha \oint_{\mathcal{C}} (N - zI)^{-1} M (N + \alpha M - zI)^{-1} \mathrm{d}z.$$

Because as α goes to 0, none of the eigenvalues of N and $N + \alpha M$ crosses $\mathcal{C}$, the integral converges and then we derive that $\lim_{\alpha \to 0} \Pi_\alpha = \Pi_0$. Besides, we have

$$(N + \alpha M)\Pi_\alpha = 0, \quad \text{and} \quad N\Pi_0 = 0,$$

then we get $N(\Pi_0 - \Pi_\alpha) = \alpha M \Pi_\alpha$, and we obtain

$$\mathrm{Im}(M\Pi_\alpha) \subset \mathrm{Im}(N).$$

We conclude the proof using the continuity of Π_α. □

Lemma 1.E. *Let $\mathcal{M} \subset \mathbb{R}^{d \times d}$ be a vector space of non-invertible symmetric matrices. We have*

$$\exists u \in \mathbb{R}^d, \quad \forall M \in \mathcal{M}, \quad u^T M u = 0.$$

Proof. Since $\mathcal{M}$ is a vector space, we can apply Lemma 1.D with N a matrix of maximal rank in $\mathcal{M}$ and any $M \in \mathcal{M}$. Then, for every $u \in \ker(N)$, there exists $y \in \mathbb{R}^d$ such that

$$Mu = Ny.$$

Because N is symmetric, by multiplying the left-hand side by u^T, we obtain $u^T M u = 0$. □

Chapter 2

Continuous inverse regression with application to Cramér-von Mises testing

Abstract : In this chapter, we propose a new method called continuous inverse regression (CIR) for the estimation of the central subspace. Most of the existing methods that estimate the central subspace involve a slicing of the response (e.g. SIR, IRE, SAVE, DR,...) but in general, we do not know how to slice the response in such a way that the whole central subspace is estimated. Our method solves this problem because the slicing of the response is no longer required, while it keeps the exhaustivity of the estimation. The method CIR is based on the fact that $\mathbb{E}[X\mathbb{1}_{\{Y\leq y\}}]$ belongs to the central subspace for every $y \in \mathbb{R}$. More precisely, CIR inspects the range of the matrix

$$M_\Phi = \int_{\mathbb{R}} \mathbb{E}[X\mathbb{1}_{\{Y\leq y\}}]\mathbb{E}[X\mathbb{1}_{\{Y\leq y\}}]^T \mathrm{d}\Phi(y),$$

where Φ can be a known distribution function or the estimated distribution function of Y. In this chapter, given an i.i.d. sample with unknown distribution, we define an empirical estimator of M_{CIR} which is easy to compute and we obtain its asymptotic normality. From this result, we derive some Cramér-von Mises tests to select the dimension of the central subspace and to test if whether a predictor has an effect on the explanatory variables.

Key words : Dimension reduction ; Sliced inverse regression ; Weak convergence in $D[0, 1]$.

2.1 Introduction

Let $(X_i, Y_i)_{1\leq i\leq n}$ be an i.i.d. sample drawn from the model

$$Y_i = g(\beta^T X_i, e_i), \tag{2.1}$$

where $X_i \in \mathbb{R}^p$ is independent from $e_i \in \mathbb{R}$, $Y_i \in \mathbb{R}$, $\beta \in \mathbb{R}^{p\times d_0}$ and $g : \mathbb{R}^{d_0+1} \to \mathbb{R}$. Moreover, throughout the chapter, we will assume that the variable Y has a strictly positive density and that the central subspace $\text{span}(\beta)$ is unique, we denote it by E_c.

One of the most popular method to estimate E_c is the *sliced inverse regression* (SIR) [66] which has been introduced in section 1.2.1. We recall that the space estimated by SIR is equal to

$$E_{\text{SIR}} = \Sigma^{-1}\,\text{span}(C_1, \cdots, C_H),$$

where $C_h = \mathbb{E}[X_1 \mathbb{1}_{\{Y_1\in I(h)\}}]$, and the $I(h)$'s form a partition of the range of the variable Y. Theorem 1.3 implies that for H sufficiently large, and under some condition (the order 1 coverage condition page 51), we have $E_{\text{SIR}} = E_c$. This result is important because it ensures that when H increases, SIR eventually estimates the whole subspace. Nevertheless, this is not sufficient to guarantee a complete estimation of E_c in practice since we do not know how to choose H.

In this work, we consider the family of vectors $\{\mathbb{E}[X_1 \mathbb{1}_{\{Y_1\leq y\}}],\ y \in \mathbb{R}\}$. Since it separates the points, we have the same result than previously, that is a finite number of this family spanned E_c. Then, there exists a set of real number s_H such that

$$E^H_{\text{CIR}} = \text{span}(\mathbb{E}[X_1 \mathbb{1}_{\{Y_1\leq y\}}],\ y \in s_H) = E_c.$$

As stated in the following proposition, considering the family of functions $\mathbb{1}_{\{.\leq y\}}$ leads to the same estimator than SIR if we consider a finite number of functions.

Proposition 2.1. *There exist some $I(h)$'s and a set s_H such that $E_{SIR} = E^H_{CIR}$.*

To prove this result, it suffices to notice that $C_h = \mathbb{E}[X\mathbb{1}_{\{Y\leq y_h\}}] - \mathbb{E}[X\mathbb{1}_{\{Y\leq y_{h-1}\}}]$ for some y_h, y_{h-1}. As a consequence, both methods SIR and CIR at H finite, are the same. The main advantage of considering E^H_{CIR} with respect to E_{SIR} is that we can estimate the limiting space when $H \to \infty$ without the need of nonparametric estimation (kernel smoother, wavelet,etc...). This is made possible by defining the matrix

$$M_\Phi = \int \mathbb{E}[X_1 \mathbb{1}_{\{Y_1\leq y\}}]\mathbb{E}[X_1 \mathbb{1}_{\{Y_1\leq y\}}]^T \mathrm{d}\Phi(y),$$

where Φ is a probability measure covering the whole support of Y [1]. As a consequence, we have that

$$E_{\text{CIR}} = \text{span}(\mathbb{E}[X_1 \mathbb{1}_{\{Y_1\leq y\}}],\ y \in \mathbb{R}) = \text{span}(M_\Phi).$$

1. If Φ is a discrete probability measure, we get back to earlier considerations with the space E^H_{CIR}.

Because the sum is taken over all the functions in the family $(\mathbb{1}_{\{\cdot\leq y\}})_{y\in\mathbb{R}}$, we have that

$$E^H_{\mathrm{CIR}} \subset E_{\mathrm{CIR}} = E_c,$$

provided that the order 1 coverage condition holds. In particular, we avoid the choice of s_H while keeping all the directions offer by the vectors $\mathbb{E}[X_1\mathbb{1}_{\{Y_1\leq y\}}]$. For the estimation, let us define the processes $\mathbb{G}$ and G by

$$\mathbb{G}(y) = \frac{1}{n}\sum_{i=1}^{n} X_i\mathbb{1}_{\{Y_i\leq y\}} \qquad \text{and} \qquad G(y) = \mathbb{E}[X_1\mathbb{1}_{\{Y_1\leq y\}}].$$

One has the formula

$$M_\Phi = \int_{[0,1]} G\circ\Phi^-(u)G\circ\Phi^-(u)^T\mathrm{d}u.$$

To show this, we notice that the variable $\Phi^-(U)$ has distribution Φ , whenever U is uniformly distributed on $[0,1]$ (see for instance [76] page 305). The function Φ^- is the generalized inverse of Φ and is given by

$$\Phi^-(u) = \inf\{y\in\mathbb{R} \ : \ \Phi(y)\geq u\}.$$

As a consequence, we define $\widehat{M}_\Phi$ as the estimator of M_Φ given by

$$\widehat{M}_\Phi = \int_{[0,1]} \mathbb{G}\circ\Phi^-(u)\mathbb{G}\circ\Phi^-(u)^T\mathrm{d}u.$$

Our estimator is easy to compute since a quick calculation gives that

$$\widehat{M}_\Phi = n^{-2}\sum_{i,j}^{n} X_iX_j^T(1-\max(\Phi(Y_i),\Phi(Y_i))). \tag{2.2}$$

Note that the processes $\mathbb{G}\circ\Phi^-$ and $G\circ\Phi^-$ are elements of $D[0,1]$, the space of càd-làg functions on $[0,1]$. Moreover, if Φ^- is continuous, then $G\circ\Phi^-$ belongs to $C[0,1]$, the space of continuous functions $[0,1]$.

The study of the asymptotic behavior of $\widehat{M}_\Phi$ is the main topic of the chapter. To obtain the convergence in law of $\sqrt{n}(\widehat{M}_\Phi - M_\Phi)$, our approach consists in the two following steps :

(A) Weak convergence of the process $\sqrt{n}(\mathbb{G}-G)$ to a limit in $C[0,1]$.

(B) Continuous mapping theorem to obtain the convergence of $\sqrt{n}(\widehat{M}_\Phi - M_\Phi)$.

When Φ is a known function, we can show the weak convergence of the process $\sqrt{n}(\mathbb{G}-G)$ in the classical sense, i.e. in the space $D[0,1]$ equipped with the Skorohod distance as studied for instance in [8]. This is the approach employed in section 2.2.

A natural choice for Φ is the function $F : y \mapsto \mathbb{P}(Y_1 \leq y)$. We study the asymptotic behavior of the associated estimator in section 2.3. In this case, we can not follow the same path than previously because F is unknown and needs to be estimated. Its estimation induces some random effects that influence the weak convergence. To show the convergence in this case, we also employed the theory of convergence in metric space by following (A) and (B) [2]. Nevertheless, we utilize another notion of weak convergence, which is the convergence in the sense of the outer integral, studied for instance in [77].

In section 2.4, we apply the result obtained in the sections 2.2 and 2.3 to derive the root n consistency of the estimator of M_Φ when Φ is known, and M_F. Some tests are proposed in section 2.5.

2.2 Asymptotic behavior when Φ is known

In the following theorem, we provide the weak convergence of the process

$$\sqrt{n}(\mathbb{G} \circ \Phi^- - G \circ \Phi^-)$$

in $D[0,1]$ endowed with the Skorohod metric. We denote by "$\Rightarrow$" the weak convergence in the Skorohod space as defined in [8]. We define the functions

$$\Sigma(y) = \mathbb{E}[X_1 X_1^T \mathbb{1}_{\{Y_1 \leq y\}}] \qquad \text{and} \qquad \Gamma_1(y,z) = \Sigma(\min(y,z)) - G(y)G(z)^T.$$

Theorem 2.2. *Assume that $\mathbb{E}[\|X_1\|^2]$ is finite, Φ is an increasing function [3] and Y_1 has a continuous density, then we have*

$$\sqrt{n}(\mathbb{G} \circ \Phi^- - G \circ \Phi^-) \Rightarrow W_1,$$

where W_1 is a Gaussian process with covariance function $\Gamma_1(\Phi^-, \Phi^-)$.

Proof. Let us define $\mathbb{W}_1 = \sqrt{n}(\mathbb{G} \circ \Phi^- - G \circ \Phi^-)$. We show the statement of the theorem by applying Theorem 13.5 in [8] page 142. This reduces to show the convergence of the finite dimensional law and the tightness of each coordinate, since it is equivalent to the tightness in the product space. For the tightness, we check the conditions (13.12) and (13.13) by showing the following much restrictive conditions [4] : Let $1 \leq k \leq p$, we provide that

$$\mathbb{E}[(W_{1,k}(1) - W_{1,k}(1-\delta))^2] \longrightarrow 0 \quad \text{when } \delta \to 0, \tag{2.3}$$

and, for any $r \leq s \leq t$,

$$\mathbb{E}[(\mathbb{W}_{1,k}(s) - \mathbb{W}_{1,k}(r))^2 (\mathbb{W}_{1,k}(t) - \mathbb{W}_{1,k}(s))^2] \leq 3(c(t) - c(r))^2, \tag{2.4}$$

2. It seems even more as a necessary path here since unlike the previous case, the link with the theory of U-statistic, provided by equation (2.2) is not clear here.

3. This is also true if Φ covers the whole support of Y, but this implies some technicalities.

4. They are much restrictive because of the Markov inequality and equation (13.14) in [8].

where $c(t) = \mathbb{E}[X_{1,k}^2 \mathbb{1}_{\{\Phi(Y_1)\leq t\}}]$ is a continuous and non-decreasing function.
• The tightness : We define $a_i(u) = X_{i,k}\mathbb{1}_{\{\Phi(Y_i)\leq u\}} - G_k(u)$ where $X_{i,k}$ and $G_k(u)$ stand for the k-th coordinate of X_i and $G(u)$. For any $0 \leq r < s < t \leq 1$, we note $a_i[r,s] = a_i(s) - a_i(r)$. We have

$$\begin{aligned} n^2\mathbb{E}[(\mathbb{W}_{1,k}(s) - \mathbb{W}_{1,k}(r))^2(\mathbb{W}_{1,k}(t) - \mathbb{W}_{1,k}(s))^2] &= \mathbb{E}[(\sum_{i=1}^n a_i[r,s])^2(\sum_{i=1}^n a_i[s,t])^2] \\ &= n\mathbb{E}[a_i[r,s]^2 a_i[s,t]^2] + n(n-1)\mathbb{E}[a_i[r,s]^2]\mathbb{E}[a_i[s,t]^2] + 2n(n-1)\mathbb{E}[a_i[r,s]a_i[s,t]]^2 \\ &\leq n\mathbb{E}[a_i[r,s]^2 a_i[s,t]^2] + 3n(n-1)\mathbb{E}[a_i[r,s]^2]\mathbb{E}[a_i[s,t]^2] \\ &\leq 3n^2\mathbb{E}[a_i[r,s]^2 a_i[s,t]^2], \end{aligned}$$

now, using that $[r,s]$ and $[s,t]$ are disjoint, by a little calculus we obtain that

$$\mathbb{E}[a_i[r,s]^2 a_i[s,t]^2] \leq 2\mathbb{E}[X_{1,k}^2 \mathbb{1}_{\{\Phi(Y_1)\in[r,s]\}}]\mathbb{E}[X_{1,k}^2 \mathbb{1}_{\{\Phi(Y_1)\in[s,t]\}}],$$

which leads to

$$\mathbb{E}[(\mathbb{W}_{1,k}(s) - \mathbb{W}_{1,k}(r))^2(\mathbb{W}_{1,k}(t) - \mathbb{W}_{1,k}(s))^2] \leq 6\mathbb{E}[X_{1,k}^2 \mathbb{1}_{\{\Phi(Y_1)\in[r,t]\}}]^2.$$

The function c is continuous because F and Φ^- are continuous functions (because Φ is increasing). It remains to show (2.3) by noting that $\mathbb{E}[(W_{1,k}(1) - W_{1,k}(1-\delta))^2] = c(1) - c(1-\delta)$ and by using again the continuity of c.
• Convergence of the finite dimensionnal laws : We study the asymptotic law of the $p \times K$-dimensional matrix $(\mathbb{W}_1(u_1), ..., \mathbb{W}_1(u_K))$. By applying the CLT, we obtain that

$$(\mathbb{W}_1(u_1), ..., \mathbb{W}_1(u_K)) \xrightarrow{\mathrm{d}} (W_1(u_1), ..., W_1(u_K)),$$

where $\mathrm{vec}(W_1(u_1), ..., W_1(u_K))$ is a Gaussian vector with mean 0 and covariance matrix having the block decomposition $(\Gamma_1(\Phi^-(u_k), \Phi^-(u_l)))_{1\leq k,l\leq K}$.

□

2.3 Asymptotic behavior when Φ is the unknown distribution function of Y.

In this section we consider the case where $\Phi = F$. Our goal is to estimate the matrix

$$M_F = \int_{[0,1]} G \circ F^-(u) G \circ F^-(u)^T \mathrm{d}u.$$

Because F is unknown, we can not use the estimator of the previous section. As a consequence, we introduce the empirical processes

$$\mathbb{F}(y) = n^{-1}\sum_{i=1}^n \mathbb{1}_{\{Y_i\leq y\}} \qquad \text{and} \qquad \mathbb{H}(u) = n^{-1}\sum_{i=1}^n X_i \mathbb{1}_{\{\mathbb{F}(Y_i)\leq u\}},$$

respectively defined for each $y \in \mathbb{R}$ and $u \in [0,1]$. Our estimator of M_F is the matrix

$$\widehat{M}_F = \int_{[0,1]} \mathbb{H}(u)\mathbb{H}(u)^T \mathrm{d}u.$$

An important remark for the following is that, without loss of generality, the variables Y_i can be assumed uniformly distributed on $[0,1]$. To show this, we write $\mathbb{F}(Y_i) = n^{-1}\mathrm{card}\{Y_j \leq Y_i, j = 1, \cdots, n\}$, and we notice that the rank statistics are invariant by non decreasing transformation. Then we have that

$$\mathbb{F}(Y_i) = n^{-1}\mathrm{card}\{F(Y_j) \leq F(Y_i), j = 1, \cdots, n\} \qquad \text{a.s.},$$

where $F(Y_i)$ is uniformly distributed on $[0,1]$, because of the continuity of F. As a result we can put $F = \mathrm{id}$.

To obtain the convergence of M_F, we follow the two steps (A) and (B) stated in the introduction. Concerning the point (A), that is $\sqrt{n}(\mathbb{H} - H)$ converges in the space $D[0,1]$, we follow an approach employed in [77], page 389, and also in [42], both in the context of the weak convergence of the empirical copula process. Standard methods will give the weak convergence of the empirical process $\sqrt{n}((\mathbb{F},\mathbb{G}) - (F,G))$ in the space $D[0,1]$. Since

$$\mathbb{H} = \psi(\mathbb{F},\mathbb{G}) = \mathbb{G} \circ \mathbb{F}^-$$

where $\mathbb{F}^-$ is the generalized inverse of $\mathbb{F}$ and $\psi : D[0,1] \times D[0,1] \to D[0,1]$. The Delta method would be a great tool to derive the weak limit of $\mathbb{H}$. Note that because the Delta method requires the Hadamard differentiability, weak convergence in the space $D[0,1]$ endowed with the Skorohod metric is not adapted because it is not a topological vector space [5]. To handle such problems, principally related to the poor structure of the space $D[0,1]$ equipped with the Skorohod metric, one is led to consider the space $D([0,1], \|\cdot\|_\infty)$. The drawback of this approach is that many functions of interest are not measurable with respect to the Borel σ-field [6], in particular, weak convergence in the classical sense can not be defined. For this reason, the author Hoffman Jorgensen in 1991 introduced another notion of weak convergence, defined with the outer integralc [54]. This one authorizes some elements of the considered sequences of being non measurable, provided that their limit is measurable. As a result, one can equip the space $D[0,1]$ with a stronger metric such as the distance associated with the supremum norm $\|\cdot\|_\infty$.

5. The Skorohod distance is not a norm and the addition is not continuous.

6. With respect to the Borel σ-field induces by the supremum norm, some random elements in $D[0,1]$ are not measurable. The classical example is the map $\Omega = [0,1] \to D[0,1]$ defined by $\omega \mapsto \mathbb{1}_{[\omega,1]}(\cdot)$, $\Omega = [0,1]$ (see for instance the book of Pollard [69]). This measurability problem was initially neglected by Donsker in 1952 [32] and then noticed by Chibisov in 1965 [17]. The introduction of the Skorohod metric was the first proposed solution. This one permit to work in a separable space and this approach is explained in [8]. Other solutions are studied in [69] but they are not considered here.

The outer integral of W, a random element in $D[0,1]$, is defined by

$$\mathbb{E}^*[W] = \inf\{\mathbb{E}[U] : U \geq W, U \text{ measurable and } \mathbb{E}[U] \text{ is finite}\}.$$

A sequence of random elements $\mathbb{W}$ in $(D[0,1], \|\cdot\|_\infty)$ endowed with the supremum norm $\|\cdot\|_\infty$, converges weakly in the sense of the outer integral to a measurable element $W \in D[0,1]$ if

$$\mathbb{E}^*[f(\mathbb{W})] \longrightarrow \mathbb{E}[f(W)],$$

for every f bounded continuous real function on $(D[0,1], \|\cdot\|_\infty)$. A complete study of this notion of weak convergence is proposed in [77]. If such a convergence is realized for $\mathbb{W} \in D[0,1]$, with limit W, we will say that $\mathbb{W}$ converges weakly to W in the space $(D[0,1], \|\cdot\|_\infty)$ and we will note $\mathbb{W} \Rightarrow^* W$. We say that a class of measurable functions $\mathcal{F} \subset D[0,1]$ is Donsker if the process $\mathbb{W}$ indexed on $\mathcal{F}$ and defined by

$$\mathbb{W}(f) = n^{-1/2} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i)],$$

is such that

$$\mathbb{W} \Rightarrow^* W,$$

where W is a tight measurable element.

Convergence results are obtained following this scheme :

- Use characterization of Donsker class to obtain $\sqrt{n}((\mathbb{F}, \mathbb{G}) - (F, G)) \Rightarrow^* W_2$.
- Use the Delta method to get $\mathbb{H} \Rightarrow^* W_3$.

The processes W_2 and W_3 are some tight measurable elements in the space $C[0,1]$.

We need the following Lemma, which is very similar to Theorem 2.2 with the difference that the weak convergence is provided in the space $(D[0,1], \|\cdot\|_\infty)$. As a result its proof involves different and somewhat more technical considerations related to the entropy of the class of function $\mathcal{G} = \{(x,y) \mapsto x\mathbb{1}_{[0,t]}(y),\ t \in [0,1]\}$.

Lemma 2.3. *Assume that $\mathbb{E}[\|X_1\|^2]$ is finite, then $\mathcal{G}$ is Donsker. If moreover, Y_1 is uniformly distributed, then we have*

$$\sqrt{n}(\mathbb{G} - G) \Rightarrow^* \widetilde{W}_1.$$

where $\widetilde{W}_1$ is a Gaussian process with covariance function Γ_1.

Proof. Firstly, it is classical that $\mathcal{F} = \{\mathbb{1}_{[0,t]},\ t \in [0,1]\}$ is Donsker (see for instance [77], Example 2.5.4, page 129) and that

$$n^{-1/2}\sum_{i=1}^{n}(\mathbb{1}_{\{Y_i \le t\}} - F(t)) \Rightarrow^* B,$$

where B is the Brownian bridge. In particular, from the same reference, the covering number of $\mathcal{F}$ is such that

$$N(\epsilon, \mathcal{F}, L_2(Q)) \le \frac{2}{\epsilon^2}. \tag{2.5}$$

Hence it remains to show that $\mathcal{G}$ is still a Donsker class. Functions of $\mathcal{G}$ have the form $g = \phi(\mathrm{id}, f)$ for some $f \in \mathcal{F}$ and $\phi(x,y) = xy$. We write

$$\mathcal{G} = \phi(\{\mathrm{id}\}, \mathcal{F}).$$

Let f_1 and f_2 be functions in $\mathcal{F}$, since we have

$$|\phi \circ (\mathrm{id}, f_1)(x,y) - \phi \circ (\mathrm{id}, f_2)(x,y)|^2 = x^2(f_1(y) - f_2(y))^2,$$

one can apply Theorem 2.10.20 page 199 in [77] (the condition above corresponds (2.10.19)). In view of the bound for the covering number of $\mathcal{F}$ given in (2.5), and the fact that the covering number of a single element is 1, the uniform entropy condition is checked and the class $\mathcal{G}$ is Donsker. As in Theorem 2.2, if Y_1 is uniformly distributed, the limit is identified with the limit of the finite dimensional laws.

□

Now it is easy to obtain the weak convergence of the process $\sqrt{n}((\mathbb{F}, \mathbb{G}) - (F, G))$. To state this result, we define the function $\Gamma_3 : [0,1] \times [0,1] \to \mathbb{R}^{p \times p}$ by

$$\Gamma_3(u,v) = (\partial\widetilde{G}(u), -I)\ \Gamma_2((u,u),(v,v))\ (\partial\widetilde{G}(v), -I)^T$$

where Γ_2 is the function $\Gamma_2 : [0,1]^2 \times [0,1]^2 \to \mathbb{R}^{(p+1)\times(p+1)}$ defined by

$$\Gamma_2((u,u'),(v,v')) = \begin{pmatrix} \min(u,v) - uv & \widetilde{G}(\min(u,v')) - u\widetilde{G}(v')^T \\ \widetilde{G}(\min(u',v)) - \widetilde{G}(u')v & \widetilde{\Sigma}(\min(u',v')) - \widetilde{G}(u')\widetilde{G}(v')^T \end{pmatrix},$$

with $\widetilde{G}(u) = \mathbb{E}[X_1 \mathbb{1}_{\{F(Y_1) \le u\}}]$ and $\widetilde{\Sigma}(u) = \mathbb{E}[X_1 X_1^T \mathbb{1}_{\{F(Y_1) \le u\}}]$.

Lemma 2.4. *Assume that $\mathbb{E}[\|X_1\|^2]$ is finite, and Y_1 is uniformly distributed, then we have*

$$\sqrt{n}((\mathbb{F}, \mathbb{G}) - (F, G)) \Rightarrow^* W_2,$$

where W_2 is a Gaussian process with covariance function Γ_2.

Proof. By Lemma 2.3, we have that each process $\sqrt{n}(\mathbb{F} - F)$ and $\sqrt{n}(\mathbb{G} - G)$ converges weakly. As a consequence, because tightness is equivalent to tightness of each coordinates, we have that $\sqrt{n}((\mathbb{F}, \mathbb{G}) - (F, G))$ is tight [7]. Then, one can conclude the first step by providing the convergence in law of the finite dimensional laws of $\mathbb{W}_2 = \sqrt{n}((\mathbb{F}, \mathbb{G}) - (F, G))$. By applying the CLT, we obtain that

$$(\mathbb{W}_2(u_1, v_1), ..., \mathbb{W}_2(u_K, v_K)) \xrightarrow{\mathrm{d}} (W_2(u_1, v_1), ..., W_2(u_K, v_K)),$$

where $\mathrm{vec}(W_2(u_1, v_1), ..., W_2(u_K, v_K))$ is a Gaussian vector with mean 0 and covariance matrix having the block decomposition $(\Gamma_2((u_k, v_k), (u_l, v_l)))_{1 \leq k,l \leq K}$. □

We are now able to state the main result of the chapter, which derives the asymptotic law of the process

$$n^{1/2}(\mathbb{H} - H).$$

The proof is based on the Delta method in metric spaces stated in [77], Theorem 3.9.4, page 374. We state here a restricted version adapted to the present context. We recall that a function $\psi : D \to E$ is Hadamard differentiable at the point f if there exists a continuous linear map $\psi'_f : D \to E$ such that

$$t_n^{-1}(\psi(f + t_n h_n) - \psi(f)) \longrightarrow \psi'_f(h),$$

for every sequence $(t_n, h_n) \to (0, h)$. We say that f is Hadamard differentiable at f tangentially to $C \subset D$ if $h \in C$ in the previous definition.

Lemma 2.5 (van der Vaart and Wellner (1996) [77], Theorem 3.9.4, page 374)**.** *Assume that $\psi : D[0,1] \times D[0,1] \to D[0,1]$ is Hadamard differentiable at f tangentially to $C[0,1] \times C[0,1]$, then if $\sqrt{n}(\mathbb{L} - L) \Rightarrow^* W$, $W \in C[0,1] \times C[0,1]$ and is tight, we have*

$$\sqrt{n}(\psi(\mathbb{L}) - \psi(L)) \Rightarrow^* \psi'_L(W).$$

Theorem 2.6. *Assume that $\mathbb{E}[\|X_1\|^2]$ is finite and Y_1 has a continuous density, then we have*

$$n^{1/2}(\mathbb{H} - H) \Rightarrow^* W_3,$$

where W_3 is a Gaussian process with covariance function Γ_3.

Proof. In the whole proof $F = \mathrm{id}$. Firstly we introduce the function $\psi : D[0,1] \times D[0,1] \to D[0,1]$ defined by

$$\psi(f_1, f_2) = f_2 \circ f_1^-,$$

7. The tightness here is defined with respect to the outer integral. We keep this implicit because we still have similar conclusions as in the classical case. The useful ones here are provided by Lemma 1.3.8 and Theorem 1.5.4 in [77].

and we note that we have the decomposition

$$\psi : (f_1, f_2) \mapsto (f_1^-, f_2) \mapsto f_2 \circ f_1^-.$$

Using Lemma 3.9.23, assertion (ii) page 386 in [77], the first map above reduced to $f \mapsto f^-$ is Hadamard differentiable at the function F tangentially to $C[0,1]$. Moreover its derivative at F is given by $h_1 \mapsto -h_1$. The second map is clearly Hadamard differentiable and its derivative at (F, f_2) is given by $(h_1, h_2) \mapsto -h_1 \times \partial f_2 + h_2$. By the chain rule, the function ψ is Hadamard differentiable on $\{F\} \times D[0,1]$ tangentially to $C[0,1] \times C[0,1]$. At the point (F, f_2), its derivative is given by $(h_1, h_2) \mapsto -h_1 \times \partial f_2 + h_2$. Then we can use the Delta method stated in Lemma 2.5 to conclude that the limit W_3 has the representation

$$W_3 = -W + \partial G \times B.$$

where (B, W) is a Gaussian process with covariance function Γ_2. This is the statement of the Theorem. □

Essentially because the Skorohod metric is less restrictive than the uniform metric, we have the following corollary that implies the weak convergence in the space $D[0,1]$ endowed with the uniform metric.

Corollary 2.7. *Assume that $\mathbb{E}[\|X_1\|^2]$ is finite, and Y_1 has a continuous density, then we have*

$$n^{1/2}(\mathbb{H} - H) \Rightarrow W_3,$$

where W_3 is defined in the statement of Theorem 2.6.

Proof. Since the set of all continuous functions with respect to the Skorohod metric is included in the set of all continuous functions with respect to the uniform metric, we have by Theorem 2.6 that

$$\mathbb{E}^*[f(n^{1/2}(\mathbb{H} - H))] \longrightarrow \mathbb{E}[f(W_3)],$$

for every f bounded continuous real function on $D[0,1]$ with respect to the Skorohod metric. It remains to note, that $\mathbb{E}^*$ can be replaced by $\mathbb{E}$ in the previous equation because the random variable is $f(n^{1/2}(\mathbb{H} - H))$ is now measurable.

□

2.4 Estimation of the central subspace

In the previous sections, we obtain that both processes $\sqrt{n}(\mathbb{G} - G)$ and $\sqrt{n}(\mathbb{H} - H)$ converge in $D[0,1]$ endowed with the Skorohod distance. Thanks to the continuous mapping theorem, the convergence of $\sqrt{n}(\widehat{M}_\Phi - M_\Phi)$ and $\sqrt{n}(\widehat{M}_F - M_F)$ are implied, respectively by the weak convergences of $\sqrt{n}(\mathbb{G} - G)$ and $\sqrt{n}(\mathbb{H} - H)$ when the associated transformation is continuous. If Y has a density, each limit belongs to $C[0,1]$ on which the Skorohod norm is equivalent to the uniform norm. With respect to the uniform norm, the application $f \mapsto \int f$ is clearly continuous. As a consequence, we obtain the following Theorem.

Theorem 2.8. *Assume that* $\mathbb{E}[\|X_1\|^2]$ *is finite and* Y_1 *has a continuous density, then we have*

$$n^{1/2}(\widehat{M}_F - M_F) \text{ converges to a Gaussian vector.}$$

If moreover, Φ *is an increasing distribution function, we have*

$$n^{1/2}(\widehat{M}_\Phi - M_\Phi) \text{ converges to a Gaussian vector.}$$

Proof. We make the proof for the first assertion, the second one can be treated similarly. We have

$$(\mathbb{H}\mathbb{H}^T - HH^T) = (\mathbb{H} - H)H^T + H(\mathbb{H} - H)^T + (\mathbb{H} - H)(\mathbb{H} - H)^T.$$

Then by Corollary 2.7 and the Delta-method, $\sqrt{n}(\mathbb{H}\mathbb{H}^T - HH^T)$ converges weakly to $W_3H^T + HW_3^T$. By the continuous mapping theorem in [8], Theorem 2.7, page 21, we obtain

$$n^{1/2}(\widehat{M}_F - M_F) \xrightarrow{\mathrm{d}} \int_{[0,1]} W_3(u)H(u)^T + H(u)W_3(u)^T \mathrm{d}u.$$

□

2.5 Cramér-von Mises tests

2.5.1 Testing the dimensionnality

Theorem 2.8 is sufficient to build a test for the dimension of the central subspace. By Theorem 1 and 2 in [12], we can consider statistics based on the eigendecomposition of the matrices M_F of M_Φ. This statistics lead to consistent tests provided their limit in law are known. This is not the case here, because the variance of the limiting laws, stated in Theorem 2.8, depends on unknown quantities. Hence a first possibility is to estimate these unknown quantities. As highlighted in the next chapter, this may cause some loss in the accuracy of the test. This is even more true here, since the asymptotic variance is hard to estimate.

Another more realistic possibility is to use the bootstrap. We introduce the bootstrap sample $(X_1^*, Y_1^*), \cdots, (X_n^*, Y_n^*)$ drawn from the original one $((X_1, Y_1), \cdots, (X_n, Y_n))$. We define the bootstrap quantities $\mathbb{C}^*$, $\mathbb{H}^*$, M_Φ^* and M_F^* exactly the same way as they have been defined but with the bootstrap sample. To provide the consistency of the bootstrap, an intermediary result deals with the process $\mathbb{C}^*$ and $\mathbb{H}^*$. Using Theorem 3.6.2 in [77], we obtain the weak convergence of

$$n^{1/2}(\mathbb{C}^* - \mathbb{C}) \quad \text{and} \quad n^{1/2}(\mathbb{H}^* - \mathbb{H}),$$

to the same limit as $n^{1/2}(\mathbb{C} - C)$ and $n^{1/2}(\mathbb{H} - H)$. As a consequence and using Theorem 3.9.11 in [77], we get that the bootstrap of the matrix $\widehat{M}_\Phi$ and $\widehat{M}_F$ works. We refer to the following chapter for more details about the use of the bootstrap in rank estimation.

2.5.2 Testing the non effect of a predictor or a group of predictors

Let $X_\eta \in \mathbb{R}$ be a predictor. As a result, there exists $\eta \in \mathbb{R}^p$ such that $X_\eta = \eta^T X$. We have the decomposition $\eta = \beta + \gamma$ where $\beta \in E_c$ and $\gamma \in E_c^\perp$. Clearly, the predictor X_η is used to predict Y if and only if $\beta \neq 0$. On the contrary, X_γ has no effect on Y if $\eta \in E_c^\perp$. As a consequence, we introduce the hypotheses

$$H_0 : \ \eta \in E_c^\perp \quad \text{against} \quad H_1 : \ \eta \notin E_c^\perp. \tag{2.6}$$

Under the order 1 coverage condition page 51, which basically says that E_c is spanned by the inverse regression curve, the previous hypotheses are equivalent to

$$H_0 : \ \eta^T M_F \eta = 0 \quad \text{against} \quad H_1 : \ \eta^T M_F \eta \neq 0,$$

where the matrix M_F can also be replaced by M_Φ. For M_F, because of the formula (2.2), a good statistic for the test is

$$\widehat{\Lambda}_F = n^{-3/2} \sum_{i,j}^{n} (\eta^T X_i)(\eta^T X_j)(1 - \max(\widehat{F}(Y_i), \widehat{F}(Y_j))).$$

Since the asymptotic law of $\widehat{\Lambda}_F$ is complicated, we also recommend to use the bootstrap to derive the asymptotic quantile.

The testing procedure we just described can also be done for groups of variables. It consist in the same argumentation, with the difference that η equals a set of several vectors.

2.5.3 Testing an estimated subspace

In this section we focus on the method OF2 introduced in the previous chapter, even if this works for every method that estimates E_c. To select the good eigenvectors of the estimated matrix M_{OF2}, we proposed a test in section 1.12. This test is an independence test between Y and $\widehat{\eta}^T X$ where $\widehat{\eta}$ is an estimated vector by the method OF2. As a consequence, such a procedure is different from the usual test of rank. Nevertheless the test we developped was too strong with respect to the context. Indeed, the null hypothesis $\eta^T X \perp\!\!\!\perp Y$ implies that $\eta \in E_c^\perp$ but the converse is not true in general. As a result, we over estimate the number of vector that belong to E_c.

Assume that $\widehat{Q}$ is the orthogonal projector associated to some directions of the method OF2. Under H_0, we have that $\sqrt{n}(\widehat{Q} - Q)$ converges in law. As in the previous section for testing (2.6), we consider the statistic

$$\widehat{\Lambda}_F = n^{-3/2} \sum_{i,j}^{n} (\widehat{Q}X_i)(\widehat{Q}X_j)^T (1 - \max(\widehat{F}(Y_i), \widehat{F}(Y_j))).$$

To show that this statistic has a weak limit under H_0, one can write

$$\widehat{Q}\mathbb{H} = (\widehat{Q} - Q)H + Q(\mathbb{H} - H) + (\widehat{Q} - Q)(\mathbb{H} - H),$$

then by Theorem 2.6, the term of the write is a $o_{\mathbb{P}}(n^{-1/2})$. Then it remains to derive the asymptotic law of the vector $((\widehat{Q}-Q)H, Q(\mathbb{H}-H))$ which can be obtained by the finite dimensional laws because the tightness is a direct consequence of Theorem 2.6. Under H_1, the term $QH \neq 0$, and the statistic goes to infinity in probability.

2.6 Conclusion

There is three main points that could be the subject of further researches. Firstly, one can look for optimality in the choice of the distribution Φ. Secondly, the approach developped in this chapter can be extended to other classes of functions than indicators. Indeed, one can consider the vectors

$$\mathbb{E}[X\psi_t(Y)],$$

with ψ_t a family that separates the points (e.g. Fourrier, wavelet, kernel, ...). Thirdly, such an approach needs to be developped for order 2 methods such as TF2 because of their exhaustiveness in the estimation.

Chapter 3

Bootstrap testing of the rank of a matrix via least squared constrained estimation

Abstract : In order to test if an unknown matrix M_0 has a given rank (null hypothesis), we consider a statistic that is a squared distance between an estimator $\widehat{M}$ and the manifold of fixed-rank matrix. Under the null hypothesis, this statistic converges to a weighted chi-squared distribution. In this paper, we introduce the constrained bootstrap in order to build bootstrap estimates of the law under the null hypothesis of this statistic. As a result, the constrained bootstrap is employed to estimate the quantile for testing the rank. We provide the consistency of the procedure and the simulations shed light one the accuracy of the bootstrap method with respect to the traditional asymptotic comparison. More generally, the results are extended to test whether an unknown parameter belongs to a sub-manifold locally smooth. Finally, the constrained bootstrap is easy to compute, it handles a large family of tests and it works under mild assumptions.

Keywords. Rank estimation ; Least squared constrained estimation ; Bootstrap ; Hypothesis testing.

This chapter is under review in the
Journal of the american statistical association,
it has been written in collaboration with Bernard Delyon.

3.1 Introduction

Let $M_0 \in \mathbb{R}^{p\times H}$ be an unknown matrix. To infer about the rank of M_0 with hypothesis testing, the general framework usually considered is the following : there exists an estimator $\widehat{M} \in \mathbb{R}^{p\times H}$ of M_0 such that

$$n^{1/2}(\widehat{M} - M_0) \xrightarrow{\mathrm{d}} W, \qquad \text{with} \qquad \mathrm{vec}\,(W) = \mathcal{N}(0, \Gamma) \tag{3.1}$$

where $\mathrm{vec}(\cdot)$ vectorizes a matrix by stacking its columns. In the whole paper the hatted quantities are random sequences that depend on the sample number n, all the limits are taken with respect to n. Moreover there exists an estimator $\widehat{\Gamma}$ such that

$$\widehat{\Gamma} \xrightarrow{\mathbb{P}} \Gamma, \tag{3.2}$$

and in some cases, one may ask that

$$\Gamma \text{ is full rank.} \tag{3.3}$$

Let d_0 be the rank of M_0 and $m \in \{1, ..., p\}$, we consider the set of hypotheses

$$H_0 \ : \quad d_0 = m \qquad \text{against} \qquad H_1 \ : \quad d_0 > m. \tag{3.4}$$

Thus d_0 would be estimated the following way : we start by testing $m = 0$, if H_0 is rejected we go a step further $m := m + 1$, if not we stop the procedure and the estimated rank is $\widehat{d} = m$. In this paper, by considering the hypotheses (3.4) we focus on each step of this procedure.

Many different statistical tests appeared in the literature for this purpose. For instance Cragg and Donald [25] introduced a statistic based on the LU decomposition of $\widehat{M}$, Kleibergen and Paap [61] studied the asymptotic behaviour of some transformation of the singular values of $\widehat{M}$, and Cragg and Donald [26] considered the minimum of a squared distance under rank constraint. In some other fields with similar issues, close ideas have been developed : Bura and Yang [13] examined a Wald type statistic depending on the singular decomposition of $\widehat{M}$ and Cook and Ni [22] also considered the minimum of a squared distance under rank constraint. Although based on different considerations, each of the previous work relies on the test described by (3.4). For comprehensiveness, in this paper we consider the following three statistics. The first one is introduced by Li [66] as

$$\widehat{\Lambda}_1 = n \sum_{k=m+1}^{p} \widehat{\lambda}_k^2 \tag{3.5}$$

where $(\widehat{\lambda}_1, ..., \widehat{\lambda}_p)$ are the singular values of $\widehat{M}$ arranged in descending order. Under H_0 and (3.1), this statistic converges in law to a weighted chi-squared distribution [13]. The main drawback of such a test is that $\widehat{\Lambda}_1$ is not pivotal, i.e. its asymptotic law depends on

unknown quantities that are M_0 and Γ. Accordingly the consistency of the associated test requires assumptions (3.1) and (3.2). In [13] a standardized version of $\widehat{\Lambda}_1$ is studied with

$$\widehat{\Lambda}_2 = n \operatorname{vec}(\widehat{Q}_1 \widehat{M} \widehat{Q}_2)^T [(\widehat{Q}_2 \otimes \widehat{Q}_1) \widehat{\Gamma} (\widehat{Q}_2 \otimes \widehat{Q}_1)]^+ \operatorname{vec}(\widehat{Q}_1 \widehat{M} \widehat{Q}_2) \tag{3.6}$$

where M^+ stands for the Moore-Penrose inverse of M and $\widehat{Q}_1$ and $\widehat{Q}_2$ are respectively the orthogonal projectors on the left and right singular spaces of $\widehat{M}$ associated with the $p-m$ smallest singular values. The authors proved that under H_0, if (3.1) and (3.2) hold, the Wald-type statistic $\widehat{\Lambda}_2$ is asymptotically chi-squared distributed. Besides, [26] and [22] proposed a constrained estimator by minimizing a squared distance under a fixed-rank constraint as

$$\widehat{\Lambda}_3 = n \min_{\operatorname{rank}(M)=m} \operatorname{vec}(\widehat{M} - M)^T \widehat{\Gamma}^{-1} \operatorname{vec}(\widehat{M} - M), \tag{3.7}$$

which is also asymptotically chi-squared distributed under H_0, assuming (3.1), (3.2) and (3.3). We will refer the minimum discrepancy approach. Although the statistics $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$ have the convenience of being pivotal, they both require the inversion of a large matrix and this may cause robustness problems when the sample number is not large enough. For $\alpha \in (0,1)$ and under the relevant assumptions, each of these statistics $\widehat{\Lambda}_1$, $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$, is consistent at level α in testing (3.4), i.e. the level goes to $1-\alpha$ and the power goes to 1 as n goes to ∞.

Nevertheless the estimation of the quantile is difficult because either the asymptotic distribution depends on the data (non pivotality represented by $\widehat{\Lambda}_1$), or the true distribution may be quite different than the asymptotic one (slow rates of convergence represented by $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$). The objective of the paper is to propose a bootstrap method for quantile estimation in this context.

An important remark which instigates the sketch of the paper is that all the previous statistics share the form

$$\widehat{\Lambda} = n \| \widehat{B}^{1/2} \operatorname{vec}(\widehat{M} - \widehat{M}_c) \|^2 \tag{3.8}$$

with

$$\widehat{M}_c = \operatorname*{argmin}_{\operatorname{rank}(M)=m} \| \widehat{A}^{1/2} \operatorname{vec}(\widehat{M} - M) \|^2 \tag{3.9}$$

where $\| \cdot \|$ is the Euclidean norm, $\widehat{A} \in \mathbb{R}^{pH \times pH}$, $\widehat{B} \in \mathbb{R}^{pH \times pH}$. The values of $\widehat{A}$ and $\widehat{B}$ corresponding to the statistics $\widehat{\Lambda}_1$, $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$ are summarized in the Table 3.1 (See Section 3.2 for the details).

We refer to traditional testing (resp. bootstrap testing) when the statistic is compared to its asymptotic quantile (resp. bootstrap quantile). The bootstrap test is said to be consistent at level α if

$$\mathbb{P}_{H_0}\left(\widehat{\Lambda} > \widehat{q}(\alpha)\right) \longrightarrow 1-\alpha \qquad \text{and} \qquad \mathbb{P}_{H_1}\left(\widehat{\Lambda} > \widehat{q}(\alpha)\right) \longrightarrow 1, \tag{3.10}$$

	$\widehat{\Lambda}_1$	$\widehat{\Lambda}_2$	$\widehat{\Lambda}_3$
$\widehat{A}$	I	I	$\widehat{\Gamma}^{-1}$
$\widehat{B}$	I	$[(\widehat{Q}_2 \otimes \widehat{Q}_1)\widehat{\Gamma}(\widehat{Q}_2 \otimes \widehat{Q}_1)]^+$	$\widehat{\Gamma}^{-1}$

TABLE 3.1 – Values of $\widehat{A}$ and $\widehat{B}$ in (3.8) and (3.9) for computing $\widehat{\Lambda}_1$, $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$.

where $\widehat{q}(\alpha)$ is the quantile of level α calculated by bootstrap. The advantage of bootstrap testing is its high level of accuracy under H_0 with respect to traditional testing. This fact is emphasized by considering the two possibilities : when the statistic is pivotal and when the asymptotic law of the statistic depends on unknown quantities. Firstly, as highlighted by Hall [48], when the statistic is pivotal, under some conditions the gap between the distribution of the statistic and its bootstrap distribution is $O_{\mathbb{P}}(n^{-1})$. Since the normal approximation leads to a difference $O(n^{-1/2})$, the bootstrap enjoys a better level of accuracy. Secondly if the asymptotic law of the statistic is unknown, the bootstrap appears even more as a convenient alternative because it avoids its estimation. In [51], Hall and Wilson gave two advices for the use of the bootstrap testing :

A) Whatever the sample is under H_0 or H_1, the bootstrap estimates the law of the statistic under H_0.

B) The statistic is pivotal.

The first guideline is the most crucial because if it fails it may lead to inconsistency of the test. The second guideline aims at improving the accuracy of the test by taking full advantage of the accuracy of the bootstrap. In this paper we propose a new procedure for bootstrap testing in least squared constraint estimation (LSCE) (estimators as (3.8) are particular cases), called constrained bootstrap (CS bootstrap). More precisely, the CS bootstrap aims at testing whether a parameter belongs or not to a submanifold and so generalised the test (3.4). Our main result is the consistency of the CS bootstrap under mild conditions. As a consequence we provide a consistent bootstrap testing procedure for testing (3.4) with the statistic $\widehat{\Lambda}_1$, $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$. For the sake of clarity, we address the CS bootstrap in the next section. Section 3 is dedicated to rank estimation with special interest to the bootstrap of the statistic $\widehat{\Lambda}_1$, $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$. Finally, the last section emphasizes the accuracy of the bootstrap in rank estimation by providing a simulation study in sufficient dimension reduction (SDR). Accordingly, the sketch of the paper is as follows :

- The CS bootstrap in LSCE
- Bootstrap testing procedure for $\widehat{\Lambda}_1$, $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$
- Application to SDR

3.2 The constrained bootstrap for LSCE and hypothesis testing

Because of (3.8) LSCE has a central place in the paper. Moreover since LSCE intervenes in many statistical fields as M-estimation or hypothesis testing, this section is independent from the rest of the paper.

3.2.1 LSCE

Let $\theta_0 \in \mathbb{R}^p$ be called the parameter of interest, and let $\widehat{\theta} \in \mathbb{R}^p$ be an estimator of θ_0. We define the constrained estimator of θ_0 as

$$\widehat{\theta}_c = \underset{\theta \in \mathcal{M}}{\operatorname{argmin}} \ (\widehat{\theta} - \theta)^T \widehat{A} (\widehat{\theta} - \theta), \tag{3.11}$$

where $\mathcal{M}$ is a submanifold of $\mathbb{R}^p$ with co-dimension q, and $\widehat{A} \in \mathbb{R}^{p \times p}$. The constrained statistic is defined as

$$\widehat{\Lambda} = n(\widehat{\theta} - \widehat{\theta}_c)^T \widehat{B} (\widehat{\theta} - \widehat{\theta}_c). \tag{3.12}$$

where $\widehat{B} \in \mathbb{R}^{p \times p}$. Note that if $\widehat{A}$ is full rank, the unique minimizer of (3.11) without constraint is $\widehat{\theta}$, hence it could be understood as the unconstrained estimator. We introduce now the notion of nonsingular point in $\mathcal{M}$. This one is needed to express the Lagrangian first order condition of the optimization (3.11). For any function $g = (g_1, \ldots, g_p) : \mathbb{R}^p \to \mathbb{R}^q$, define its Jacobian as $J_g = (\nabla g_1, ..., \nabla g_q)$, where ∇ stands for the gradient operator.

Definition 3.1. *We say that θ is $\mathcal{M}$-nonsingular if $\theta \in \mathcal{M}$ and if there exists a neighbourhood V and a function $g : \mathbb{R}^p \to \mathbb{R}^q$ continuously differentiable on V with $J_g(\theta)$ full rank such that*

$$V \cap \mathcal{M} = \{g = 0\}.$$

As a consequence any point of a locally smooth submanifold is nonsingular, e.g. any matrix with rank m is a nonsingular point in the submanifold $\operatorname{rank}(M) = m$. We prove in Proposition 3.4 that if θ_0 is $\mathcal{M}$-nonsingular, $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Delta)$ and $\widehat{B} = \widehat{A} \xrightarrow{\mathbb{P}} A$ is full rank, then we have

$$\widehat{\Lambda} \xrightarrow{\mathrm{d}} \sum_{k=1}^{p} \nu_k W_k^2, \tag{3.13}$$

where the W_k's are i.i.d. Gaussian random variables and the ν_k's are the eigenvalues of the matrix $\Delta^{1/2} J_g(\theta_0)^T (J_g(\theta_0) A^{-1} J_g(\theta_0)^T)^{-1} J_g(\theta_0) \Delta^{1/2}$. Especially, the case $A = \Delta^{-1}$ is interesting because $\widehat{\Lambda}$ is asymptotically chi-squared distributed with q degrees of freedom. Otherwise, if $\theta_0 \notin \mathcal{M}$, $\widehat{\Lambda}$ goes to infinity in probability. Those facts shed light on a consistent testing procedure based on LSCE with the hypotheses

$$H_0 \ : \quad \theta_0 \in \mathcal{M} \qquad \text{against} \qquad H_1 \ : \quad \theta_0 \notin \mathcal{M} \tag{3.14}$$

and the decision rule to reject H_0 if $\widehat{\Lambda}$ is larger than a quantile of its asymptotic law. Accordingly the previous framework can be seen as an extension of the Wald test statistic which handles the simple hypothesis $\theta_0 = \theta$ with the statistic $(\widehat{\theta} - \theta)^T \Delta^{-1} (\widehat{\theta} - \theta)$.

3.2.2 The bootstrap in LSCE

Since LSCE is a particular case of estimating equation, we review the bootstrap literature with two principal directions : estimating equation and hypothesis testing. For clarity we alleviate the framework in this section : let $X_1, \cdots, X_n$ be an i.i.d. sequence of real random variables with law P, define $\gamma = \mathrm{var}(X_1)$, $\widehat{\gamma} = \overline{(X - \overline{X})^2}$, we put $\theta_0 = \mathbb{E}[X_1]$, $\widehat{\theta} = \overline{X}$, and $A = B = \gamma^{-1}$ where $\overline{\cdot}$ stands for the empirical mean.

The original bootstrap was introduced in [37] in the following way. Let $X_1^*, \ldots, X_n^*$ be an i.i.d. sequence of real random variables with law $\widehat{P} = n^{-1} \sum_{i=1}^n \delta_{X_i}$, define $\theta^* = \overline{X^*}$, the distribution of $\sqrt{n}(\theta^* - \widehat{\theta})$ conditionally on the sample, that we call the bootstrap distribution, is "close" to the distribution of $\sqrt{n}(\widehat{\theta} - \theta_0)$, that we call the true distribution (in the rest of the paper we just say "conditionally" instead of "conditionally on the sample"). For instance, it is shown in [76] that the bootstrap distribution converges weakly to the true distribution almost surely. One says that $\sqrt{n}(\theta^* - \widehat{\theta})$ bootstraps $\sqrt{n}(\widehat{\theta} - \theta_0)$ and we will write

$$\mathcal{L}_\infty(n^{1/2}(\theta^* - \widehat{\theta}) | \widehat{P}) = \mathcal{L}_\infty(n^{1/2}(\widehat{\theta} - \theta_0)) \quad \text{a. s.},$$

where $\mathcal{L}_\infty(\cdot)$ and $\mathcal{L}_\infty(\cdot|\widehat{P})$ both mean the asymptotic laws with the difference that the later is conditional on the sample. Equivalently, one has for every $x \in \mathbb{R}$, $\mathbb{P}(\sqrt{n}(\theta^* - \widehat{\theta}) \leq x | \widehat{P}) \overset{\text{a.s.}}{\to} \mathbb{P}(\sqrt{n}(\widehat{\theta} - \theta_0) \leq x)$, but the use of the bootstrap is legitimate by a more general results stated in [48], which says that

$$|\mathbb{P}(n^{1/2}(\theta^* - \widehat{\theta})/\gamma^* \leq x | \widehat{P}) - \mathbb{P}(n^{1/2}(\widehat{\theta} - \theta_0)/\widehat{\gamma} \leq x)| = O_{\mathbb{P}}(n^{-1}) \tag{3.15}$$

with $\gamma^* = \overline{(X^* - \overline{X^*})^2}$, provided that P is non-lattice. Besides, one has

$$|\mathbb{P}(n^{1/2}(\widehat{\theta} - \theta_0)/\widehat{\gamma} \leq x) - \Phi(x)| = O_{\mathbb{P}}(n^{-1/2}),$$

where Φ is the cumulative distribution function (c.d.f.) of the standard normal law. Variations of Efron's resampling plan are proposed in [2] under the name of weighted bootstrap. For a complete introduction about the bootstrap we refer to [48]. We now present three different bootstrap techniques related to LSCE [1].

(i) The classical bootstrap (C bootstrap)

The literature about the bootstrap in Z and M-estimation, see respectively [16] and [1], is based on the following principle : if $\theta_M = \underset{\theta \in \Theta}{\mathrm{argmin}}\ \mathbb{E}[\phi(X, \theta)]$ is estimated

1. A bootstrap with a Delta-method approach (see [76], chapter 23, Theorem 5) fails because $x \to \min_{\|\theta\|=1} \|x - \theta\|$ is not continuously differentiable on the unit circle.

by $\widehat{\theta}_M = \underset{\theta\in\Theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \phi(X_i,\theta)$ where Θ is an open set, then the bootstrap of $\sqrt{n}(\widehat{\theta}_M - \theta_M)$ is carried out by the quantity $\sqrt{n}(\theta^*_M - \widehat{\theta}_M)$ with

$$\theta^*_M = \underset{\theta\in\Theta}{\operatorname{argmin}}\ n^{-1}\sum_{i=1}^n w_i\phi(X_i,\theta), \tag{3.16}$$

where (w_i) is a sequence of random variables. The particular case where the vector $(w_1,\ldots,w_n)$ is distributed as $\text{mult}(n,(n^{-1},\ldots,n^{-1}))$ leads to a direct application of original Efron's bootstrap to M-estimation. Since such a bootstrap has been extensively studied, we refer to the C bootstrap. To the knowledge of the authors, the C bootstrap when Θ has empty interior has not been studied yet. Nevertheless one may sight its bad behaviour for the test of equal mean $H_0:\ \theta_0 = \mu$. The least squared constrained statistic

$$n\widehat{\gamma}^{-1}(\widehat{\theta}-\mu)^2,$$

is indeed the score statistic associated to the M-estimator with $\phi(x,\theta) = \widehat{\gamma}^{-1}(x-\theta)^2$ and $\Theta = \{\mu\}$. Clearly the C bootstrap through $n\gamma^{*-1}(\theta^*-\mu)^2$ does not work because of its bad behaviour under H_1 for instance. In this case it is better to use

$$n\gamma^{*-1}(\theta^*-\widehat{\theta})^2,$$

but it cannot handle the cases of more involved hypotheses [2]. Whereas the C bootstrap is not really connected with hypothesis testing, the two following bootstrap procedures are more related to the present work.

(ii) The biased bootstrap (B bootstrap)

The B bootstrap is introduced in [50] and is directly motivated by hypothesis testing. The original idea of their work is to re-sample with respect to the distribution $\widehat{P}_b = \sum_{i=1}^n \omega_i\delta_{X_i}$, where the ω_i's maximize

$$\sum_{i=1}^n \log(\omega_i) \quad \text{under the constraints} \quad \begin{array}{r} \frac{1}{n}\sum_{i=1}^n \omega_i X_i = \mu \\ \sum_{i=1}^n \omega_i = 1 \end{array}. \tag{3.17}$$

Since the ω_i's minimize the Kulback-Leibler distance between $\widehat{P}$ and $\widehat{P}_b$, one can see the resulting distribution as the closest to the original one satisfying the mean constraint. The authors presented interesting results for the test of equal mean $\theta_0 = \mu$, essentially the bootstrap statistic $n\gamma^{*-1}(\theta^*_b-\mu)^2$, with $\theta^*_b = \overline{X^*_b}$, $X^*_{b,i}$ sampled from $\widehat{P}_b$, has a chi-squared limiting distribution either H_0 or H_1 is assumed. As a result both guidelines (A) and (B) are checked. They go further by showing that the B bootstrap outclasses the asymptotic normal approximation for quantile estimation in the sense that $|\widehat{q}(\alpha) - q_n(\alpha)| = O_{\mathbb{P}}(n^{-1})$ whereas $|q_n(\alpha) - q_\infty(\alpha)| = O(n^{-1/2})$, where q_∞, q_n

2. We refer to [51] for a study of this bootstrap in order to test $\theta_0 = \mu$.

and $\widehat{q}_n$ are the quantile functions of the standard normal distribution, the statistic $n\widehat{\gamma}^{-1}(\widehat{\theta}-\mu)^2$ under H_0 and the bootstrapped statistic, respectively. Although the B bootstrap matches the context of hypothesis testing, it has been designed to handle the particular test of equal mean. To the knowledge of the authors the study of the B bootstrap has not been extended to other tests. Facing (3.17), the main drawback of the B bootstrap deals with algorithmic difficulties. Indeed when the constraint becomes more involved, solving (3.17) is more difficult. As a result it is not sure that this method could handle other situations such as fixed-rank constraints.

(iii) The estimating function bootstrap (EF bootstrap)

Now $X_i \in \mathbb{R}^p$. Some other ideas about the bootstrap of the Z-estimators can be found in [62] and [57], and can be summarized as follows. Considering the score statistic $\widehat{S} = \sqrt{n}\sum_{i=1}^n \frac{\partial\phi}{\partial\theta}(X_i,\theta_0)$, [57] showed that it could be bootstrapped by

$$S^* = n^{-1/2}\sum_{i=1}^n w_i \frac{\partial\phi}{\partial\theta}(X_i,\widehat{\theta}),$$

where (w_i) is a sequence of random variables. This bootstrap is called the EF bootstrap and revealed nice computational properties. Moreover the authors argued for its use in quantile estimation in order to test if $g(\theta_0)=0$, where $g:\mathbb{R}^p\to\mathbb{R}^q$ is the constraint function, by recommending essentially to use $S^{*T}J_g(\widehat{\theta})^T\left(J_g(\widehat{\theta})\gamma^*J_g(\widehat{\theta})^T\right)^{-1}J_g(\widehat{\theta})S^*$. Applying it to the least squared context $\phi(x,\theta)=\|\widehat{\gamma}^{-1/2}(x-\theta)\|^2$, the EF bootstrap is carried out by

$$n(\theta^*-\widehat{\theta})^TJ_g(\widehat{\theta})^T\left(J_g(\widehat{\theta})\gamma^*J_g(\widehat{\theta})^T\right)^{-1}J_g(\widehat{\theta})(\theta^*-\widehat{\theta}).$$

Although it verifies both guidelines (A) and (B) (see the article for details), one can see that the good behaviour of such an approach is more based on the rank deficiency of $J_g(\widehat{\theta})$ than on the bootstrap of $\sqrt{n}(\widehat{\theta}-\widehat{\theta}_c)$. Indeed $\sqrt{n}(\theta^*-\widehat{\theta})$ bootstraps the non constrained estimator $\sqrt{n}(\widehat{\theta}-\theta_0)$. Then as the authors noticed, it is first of all a bootstrap of the Wald-type statistic $n\widehat{S}^TJ_g(\theta_0)^T\left(J_g(\theta_0)\widehat{\gamma}J_g(\theta_0)^T\right)^{-1}J_g(\theta_0)\widehat{S}$ which has fortunately the same asymptotic law than the targeted one. This may induce some loss in accuracy. Moreover, it requires the knowledge of the function J_g which is not the case for fixed rank constraints where the g depends on the limit M_0 (see Remark 3.2 for some details).

Essentially both (i) and (ii) provide a bootstrap for testing simple hypotheses. The EF bootstrap proposed in (iii) extends this limited scope by including tests of the form $g(\theta_0)=0$ where g is known. Nevertheless it does not handle the test (3.4) as it is highlighted by the following remark.

Remark 3.2. Testing (3.4) with $\widehat{\Lambda}_3$ results in an optimization with the constraint $\operatorname{rank}(M)=m$. Since the subspace of fixed rank matrices is a submanifold locally smooth with codimension $(p-d)(H-d)$, at every point M, there exists a neighbourhood V and a $\mathcal{C}^\infty$

function $g : V \to \mathbb{R}^{(p-d)(H-d)}$ such that $V \cap \{\text{rank}(M) = m\} = \{g = 0\}$ and $J_g(M)$ has full rank. Moreover, we have

$$\|\Gamma^{-1/2}\,\text{vec}(\widehat{M}_c - M_0)\| \leq 2\|\Gamma^{-1/2}\,\text{vec}(\widehat{M} - M_0)\|.$$

If now (3.1) holds, the right-hand side term goes to 0 in probability and $\widehat{M}_c \xrightarrow{\mathbb{P}} M_0$. As a consequence, if Γ is invertible, for any neighbourhood of M_0, from a certain rank, $\widehat{M}_c$ belongs to it with probability 1. Then under H_0 since M_0 has rank m the constrained estimator has the expression

$$\widehat{M}_c = \underset{g(M)=0}{\text{argmin}} \|\Gamma^{-1/2}\,\text{vec}(\widehat{M}_c - M)\|,$$

with g depending on M_0. Unfortunately we do not know neither g nor $J_g(M_0)$. This entails some problems relating to the later approach.

3.2.3 The constrained bootstrap

The CS bootstrap is introduced in order to solve all the issues we have raised through the previous little review which are essentially : computational difficulties and small scope of the existing methods. The CS bootstrap targets an estimation $\widehat{q}(\alpha)$ of the quantile under H_0 of $\widehat{\Lambda}$. The consistency of the procedure, i.e. (3.10), forms the main result about the CS bootstrap. Another important issue which occurs beforehand in the section is the bootstrap of the law of

$$n^{1/2}(\widehat{\theta}_c - \theta_0) \quad \text{under } H_0.$$

Basically, we show that a bootstrap of the unconstrained estimator $\sqrt{n}(\widehat{\theta} - \theta_0)$ allows a bootstrap of the constrained estimator $\sqrt{n}(\widehat{\theta}_c - \theta_0)$ under H_0. We point out that the heuristic in CS bootstrap is rather different than the C and EF bootstrap. Otherwise it shares the idea to "reproduce" H_0 even if H_1 is realized with the B bootstrap. Assuming that we can bootstrap $\sqrt{n}(\widehat{\theta} - \theta_0)$, the CS bootstrap calculation of the statistic is realized as follows :

The CS bootstrap procedure

Compute

$$\theta_0^* = \widehat{\theta}_c + n^{-1/2}W^*, \qquad \text{with} \quad \mathcal{L}_\infty(W^*|\widehat{P}) = \mathcal{L}_\infty(n^{1/2}(\widehat{\theta} - \theta_0)) \quad \text{a. s.}, \tag{3.18}$$

where the simulation of W^* can be done by a standard bootstrap procedure [3]. Calculate

$$\theta_c^* = \underset{\theta \in \mathcal{M}}{\text{argmin}}\ (\theta_0^* - \theta)^T A^*(\theta_0^* - \theta), \qquad \text{and} \qquad \Lambda^* = n(\theta_0^* - \theta_c^*)^T B^*(\theta_0^* - \theta_c^*), \tag{3.19}$$

where $A^* \in \mathbb{R}^{p\times p}$ and $B^* \in \mathbb{R}^{p\times p}$ [4].

3. The bootstrap procedure to get W^* is not specified because it depends on $\widehat{\theta}$. For instance, if $\widehat{\theta}$ is a mean over some i.i.d. random variables, one can use the Efron's traditional bootstrap and if $\widehat{\theta}$ is a M-estimator, one should use a bootstrap as detailed by equation (3.16).

4. Assumptions about A^* and B^* are provided further in the statements of the propositions.

Intuitively, this choice appears natural because θ_0^* equals $\widehat{\theta}_c$ plus a small perturbation going to 0. Accordingly θ_0^* is somewhat reproducing the behaviour of $\widehat{\theta}$ under H_0, especially because W^* has the right asymptotic variance. As we should notice, A^* and B^* could be chosen as $\widehat{A}$ and $\widehat{B}$ but this is not the best choice in practice. As it is highlighted in (3.15), we should normalize by the associated bootstrap quantities (e.g. the variance computed on the bootstrap sample). The following lemma gives a first order decomposition of the bootstrap law $\sqrt{n}(\theta_c^* - \widehat{\theta}_c)$ under mild conditions. The following lemma is proved in the Appendix.

Lemma 3.3. *Let $\mathcal{M}$ be a submanifold. Assume there exists $\widehat{\theta}_c \in \mathcal{M}$ and θ_c a $\mathcal{M}$-nonsingular point such that $\widehat{\theta}_c \overset{\text{a.s.}}{\to} \theta_c$. If moreover $\mathcal{L}_\infty(\sqrt{n}(\theta_0^* - \widehat{\theta}_c)|\widehat{P})$ exists a.s. and conditionally a.s. $A^* \overset{\mathbb{P}}{\to} A$ is full rank, then we have conditionally a.s.*

$$n^{1/2}(\theta_c^* - \widehat{\theta}_c) = (I - P)n^{1/2}(\theta_0^* - \widehat{\theta}_c) + o_{\mathbb{P}}(1),$$

with $P = A^{-1}J_g^T(\theta_c)(J_g(\theta_c)A^{-1}J_g^T(\theta_c))^{-1}J_g(\theta_c)$.

Note that if θ_0 is $\mathcal{M}$-nonsingular and $\mathcal{L}_\infty(\sqrt{n}(\widehat{\theta} - \theta_0)|\widehat{P})$ exists, we can apply Lemma 3.3 with $\widehat{\theta}_c = \theta_c = \theta_0$. This gives the following proposition :

Proposition 3.4. *Let $\mathcal{M}$ be a submanifold. Assume that $\mathcal{L}_\infty(\sqrt{n}(\widehat{\theta} - \theta_0)|\widehat{P})$ exists with θ_0 $\mathcal{M}$-nonsingular. Assume also that $\widehat{A} \overset{\mathbb{P}}{\to} A$ is full rank, then we have*

$$n^{1/2}(\widehat{\theta}_c - \theta_0) = (I - P)n^{1/2}(\widehat{\theta} - \theta_0) + o_{\mathbb{P}}(1),$$

with $P = A^{-1}J_g^T(\theta_0)(J_g(\theta_0)A^{-1}J_g^T(\theta_0))^{-1}J_g(\theta_0)$.

Proposition 3.4 leads easily to (3.13) and extends classical results [9] about constrained estimators with constraint $\{g = 0\}$ to manifold type constraints. Besides statements of Lemma 3.3 and Proposition 3.4 together explain the preceding definition of θ_0^* in (3.18). They also lead to the following theorem.

Theorem 3.5. *Let $\mathcal{M}$ be a submanifold. Assume that $\widehat{\theta} \overset{\text{a.s.}}{\to} \theta_0$ with θ_0 $\mathcal{M}$-nonsingular and $\widehat{A} \overset{\mathbb{P}}{\to} A$ hold. If moreover (3.18) holds and conditionally a.s. $A^* \overset{\mathbb{P}}{\to} A$ is full rank, then we have*

$$\mathcal{L}_\infty(n^{1/2}(\theta_c^* - \widehat{\theta}_c)|\widehat{P}) = \mathcal{L}_\infty(n^{1/2}(\widehat{\theta}_c - \theta_0)) \quad \text{a. s.}.$$

Essentially, Theorem 3.5 is an application of Lemma 3.3 under H_0, indeed as we saw in the proof of Lemma 3.3, equation (3.25), the assumption $\widehat{\theta} \overset{\text{a.s.}}{\to} \theta_0 \in \mathcal{M}$ implies that $\widehat{\theta}_c \overset{\text{a.s.}}{\to} \theta_c$. Nevertheless under H_1 nothing guarantee such a convergence (see Example 3.7 below). Roughly speaking, asking for an equality in law under H_1 as in Theorem 3.5 may be too much to ask. However as stated in the following theorem we do not require that $\widehat{\theta}_c$ converges a.s. to a constant to provide that the power of the corresponding test goes to 1.

This leads to the consistency of the CS bootstrap for hypothesis testing. For the statement of the consistency theorem, we need to define the quantile function of the bootstrap statistic

$$\widehat{q}(\alpha) = \inf\{x \ : \ \widehat{F}(x) \geq 1-\alpha\},$$

where $\widehat{F}$ is the c.d.f. of Λ^* conditionally on the sample.

Theorem 3.6. *Let $\mathcal{M}$ be a manifold. Assume that $\widehat{\theta} \overset{\text{a.s.}}{\to} \theta_0$ with θ_0 $\mathcal{M}$-nonsingular under H_0. We assume also that $\widehat{A} \overset{\mathbb{P}}{\to} A$ is full rank, $\widehat{B} \overset{\mathbb{P}}{\to} B$. If moreover $\mathcal{L}_\infty(\sqrt{n}(\theta_0^* - \widehat{\theta}_c)|\widehat{P}) = \mathcal{L}_\infty(\sqrt{n}(\widehat{\theta} - \theta_0))$ a.s. has a density, and conditionally a.s. $A^* \overset{\mathbb{P}}{\to} A$, $B^* \overset{\mathbb{P}}{\to} B$, then we have*

$$\mathbb{P}_{H_0}(\widehat{\Lambda} > \widehat{q}(\alpha)) \longrightarrow 1-\alpha, \quad \textit{and} \quad \mathbb{P}_{H_1}(\widehat{\Lambda} > \widehat{q}(\alpha)) \longrightarrow 1.$$

In other words, the test described in (3.14) with statistic $\widehat{\Lambda}$ and CS bootstrap calculation of quantile is consistent.

We provide the following example under H_1, where $\widehat{\theta}_c$ does not converge to a constant in probability. Although we cannot get the conclusion of Theorem 3.5, the least squared constrained statistic still converges in distribution.

Example 3.7. Let $(X_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence such that $X_1 \overset{\text{d}}{=} \mathcal{N}(0,1)$. Define $\widehat{\theta} = \overline{X}$, and $H_0 : \ \theta_0^2 = 1$. Clearly H_0 does not hold and naturally the statistic $n \min_{\theta^2=1} \|\widehat{\theta} - \theta\|^2$ goes to infinity in probability. One can find that $\widehat{\theta}_c = \text{sign}(\overline{X})$ which does not converge. Since

$$\theta_c^* = \underset{\theta^2=1}{\text{argmin}} \|\theta_0^* - \theta\|^2 \quad \text{and} \quad \theta_0^* = \widehat{\theta}_c + n^{-1/2} W^*,$$

we get that $\theta_c^* = \widehat{\theta}_c$ a.s. and naturally, we do not have the asymptotic given by Theorem 3.5. Besides, the convergence to a chi-squared distribution holds for the quantity $n \min_{\theta^2=1} \|\theta_0^* - \theta\|^2$.

3.3 Rank estimation with hypothesis testing

In this section through a review of the literature about rank estimation, we apply the results obtained in section 3.2.1 to provide a consistent bootstrap procedure for the test described by (3.4) associated with the statistics $\widehat{\Lambda}_1$, $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$. We define $q_0 = p - d_0$ the dimension of the kernel of M_0^T. We denote by $(\lambda_1, ..., \lambda_p)$ the singular values of M_0 arranged in descending order and we write the SVD of M_0 as

$$M_0 = (U_1 \ U_0) \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_0^T \end{pmatrix},$$

with $U_1 \in \mathbb{R}^{p\times d_0}$, $U_0 \in \mathbb{R}^{p\times q_0}$, $V_1 \in \mathbb{R}^{H\times d_0}$, $V_0 \in \mathbb{R}^{H\times q_0}$, and $D_1 = \text{diag}(\lambda_1, ..., \lambda_{d_0})$. For $m \in \{1, \cdots, p\}$, we note $q = p - m$ and we write the SVD of $\widehat{M}$ as

$$\widehat{M} = (\widehat{U}_1 \ \widehat{U}_0) \begin{pmatrix} \widehat{D}_1 & 0 \\ 0 & \widehat{D}_0 \end{pmatrix} \begin{pmatrix} \widehat{V}_1^T \\ \widehat{V}_0^T \end{pmatrix},$$

with $\widehat{U}_1 \in \mathbb{R}^{p\times m}$, $\widehat{U}_0 \in \mathbb{R}^{p\times q}$, $\widehat{V}_1 \in \mathbb{R}^{H\times m}$, $\widehat{V}_0 \in \mathbb{R}^{H\times q}$, $\widehat{D}_1 = \mathrm{diag}(\widehat{\lambda}_1, ..., \widehat{\lambda}_m)$ and $\widehat{D}_0 = \mathrm{diag}(\widehat{\lambda}_{m+1}, ..., \widehat{\lambda}_p)$. We also introduce the orthogonal projectors

$$Q_1 = I - P_1 = U_0 U_0^T, \qquad Q_2 = I - P_2 = V_0 V_0^T,$$
$$\widehat{Q}_1 = I - \widehat{P}_1 = \widehat{U}_0 \widehat{U}_0^T, \qquad \widehat{Q}_2 = I - \widehat{P}_2 = \widehat{V}_0 \widehat{V}_0^T.$$

Whereas the link between $\widehat{\Lambda}_3$ and LSCE is evident, the one conecting $\widehat{\Lambda}_1$ and $\widehat{\Lambda}_2$ to LSCE relies on the following classical lemma, whose proof is avoided.

Lemma 3.8. *Let $\widehat{M} \in \mathbb{R}^{p\times H}$, it holds that*

$$\underset{\mathrm{rank}(M)=m}{\mathrm{argmin}} \ \|\widehat{M} - M\|_F^2 = \widehat{P}_1 \widehat{M} \widehat{P}_2, \quad \textit{and} \quad \|\widehat{M} - \widehat{P}_1 \widehat{M} \widehat{P}_2\|_F^2 = \sum_{k=m+1}^{p} \widehat{\lambda}_k^2,$$

where $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p$ are the singular values of $\widehat{M}$ arranged in descending order, and $\widehat{P}_1$ and $\widehat{P}_2$ are orthogonal right and left singular projectors of $\widehat{M}$ associated with $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_m$.

Note that in the previous lemma, $\widehat{P}_1$ and $\widehat{P}_2$ are uniquely determined if and only if $\widehat{\lambda}_m \neq \widehat{\lambda}_{m+1}$.

3.3.1 Nonpivotal statistic

As stated in the introduction, the statistic $\widehat{\Lambda}_1 = n \sum_{k=m+1}^{p} \widehat{\lambda}_k^2$ can be used to arbitrate between the hypotheses of (3.4). Basically, if $H_0 : \ d_0 = m$ is realized, all the eigenvalues of the sum goes to 0 and $\widehat{\Lambda}_1$ has a weighted chi-squared limiting distribution. Otherwise, at least one eigenvalue converges in probability to a positive number and for any $A > 0$, $\mathbb{P}(\widehat{\Lambda}_1 > A) \longrightarrow 1$. The following proposition describes the asymptotic behaviour of $\widehat{\Lambda}_1$ [5]. It was stated in [15] and some recent extension can be found in [13]. Our statement goes further because we are also concerned about the estimation of the asymptotic law of $\widehat{\Lambda}_1$, i.e. the estimation of the weights that intervenes in the weighted chi-squared asymptotic law. Besides, the proof we give in the Appendix is quite simple [6].

Proposition 3.9. *Under H_0, if (3.1) holds we have*

$$\widehat{\Lambda}_1 \xrightarrow{d} \sum \nu_k W_k^2$$

where the ν_k's are the eigenvalues of the matrix $(Q_2 \otimes Q_1)\Gamma(Q_2 \otimes Q_1)$ and the W_k's are i.i.d. standard Gaussian variables. If moreover (3.2) holds, we have

$$(\widehat{\nu}_1, ..., \widehat{\nu}_{pH}) \xrightarrow{\mathbb{P}} (\nu_1, ..., \nu_{pH}),$$

where the $\widehat{\nu}_k$'s are the eigenvalues of the matrix $(\widehat{Q}_2 \otimes \widehat{Q}_1)\widehat{\Gamma}(\widehat{Q}_2 \otimes \widehat{Q}_1)$.

5. A similar proposition can be stated applying Proposition 3.13. Following this way, the asymptotic depends on g which is difficult to estimate for rank constraints (see Remark 3.2).

6. We no longer need the results of [35] about the asymptotic behaviour of singular values.

Remark 3.10. Unlike Theorem 1 in [15] or Theorem 1 in [13], we prefer to state this theorem with the quantities Q_1 and Q_2 rather than with U_0 and V_0. Because we do not assume that the kernel of M has dimension 1, the vectors that form U_0 or V_0 are not unique because vector spaces with dimension larger than 2 have an infinite number of basis. As a consequence it does not make sense to estimate either U_0 or V_0. To characterize convergence of spaces, a suitable object is their associated orthogonal projectors.

In general, we do not know the asymptotic distribution of $\widehat{\Lambda}_1$ because it depends on $(Q_2 \otimes Q_1)\Gamma(Q_2 \otimes Q_1)$. On the first hand, one can estimate consistently this matrix to get an approximation of the law of $\widehat{\Lambda}_1$ under H_0. Some conditions providing the consistency of the estimation are stated in Proposition 3.9. On the other hand, one can apply the CS bootstrap to estimate the quantile of $\widehat{\Lambda}_1$ in order to test. The main advantage of such an approach is that we no longer need to have a consistent estimator of Γ so that (3.2) is not needed anymore. Following section 3.2.1 and by using Lemma 3.8, we define

$$M_0^* = \widehat{P}_1 \widehat{M} \widehat{P}_2 + n^{-1/2} W^* \qquad \text{with} \quad W^* | \widehat{P} \xrightarrow{\mathrm{d}} W \quad \text{a. s.}, \tag{3.20}$$

with W defined in (3.1). Accordingly, we introduce the CS bootstrap statistic

$$\Lambda_1^* = n \sum_{k=m+1}^{p} \lambda_k^{*2},$$

with $\lambda_{m+1}^*, ..., \lambda_p^*$ the smallest singular values of M^*. The following proposition is a straightforward application of Theorem 3.6 with the submanifold $\{\mathrm{rank}(M) = m\}$.

Proposition 3.11. *If (3.1), (3.20) and $\widehat{M} \xrightarrow{\text{a.s.}} M_0$ hold, then the test described in (3.4) with the statistic $\widehat{\Lambda}_1$ and calculation of quantile with $\widehat{\Lambda}_1^*$ is consistent.*

3.3.2 Wald-type statistic

The Wald-type statistic $\widehat{\Lambda}_2 = \mathrm{vec}(\widehat{Q}_1 \widehat{M} \widehat{Q}_2)^T [(\widehat{Q}_2 \otimes \widehat{Q}_1)\widehat{\Gamma}(\widehat{Q}_2 \otimes \widehat{Q}_1)]^+ \mathrm{vec}(\widehat{Q}_1 \widehat{M} \widehat{Q}_2)$ has been introduced in [13] to get a pivotal statistic [7]. They obtained the following theorem for which we provide a different proof in the appendix.

Proposition 3.12. *If (3.1) and (3.2) hold, we have*

$$\widehat{\Lambda}_2 \xrightarrow{\mathrm{d}} \chi_s^2,$$

with $s = \min(\mathrm{rank}(\Gamma), (p-d)(H-d))$.

7. We write the expression of $\widehat{\Lambda}_2$ another way for the reasons explained in Remark 3.10 but one can recover the original expression by noting that for any symmetric matrix A, $A^+ H = (AH)^+$ if H is an orthonormal basis of a vector subspace of $\mathrm{Im}(A)$.

Following (3.19), we define the associated bootstrap statistic by

$$\widehat{\Lambda}_2^* = \operatorname{vec}(Q_1^* M_0^* Q_2^*)^T [(Q_2^* \otimes Q_1^*)\Gamma^*(Q_2^* \otimes Q_1^*)]^+ \operatorname{vec}(Q_1^* M_0^* Q_2^*),$$

where M_0^* is defined in (3.20), $\Gamma^* \in \mathbb{R}^{pH \times pH}$, $\widehat{Q}_1^*$, and $\widehat{Q}_2^*$ are the eigenprojectors associated with the smallest eigenvalues of $M_0^* M_0^{*T}$ and $M_0^{*T} M_0^*$. As Proposition 3.11, the following one is an easy application of Theorem 3.6.

Proposition 3.13. *If (3.1), (3.2), (3.20),* $\widehat{M} \overset{\text{a.s.}}{\to} M_0$ *and* $\Gamma^* \overset{\mathbb{P}}{\to} \Gamma$ *hold, then the test described in (3.4) with the statistic* $\widehat{\Lambda}_2$ *and calculation of quantile with* Λ_2^* *is consistent.*

3.3.3 Minimum Discrepancy approach

Noting that $\{\operatorname{rank}(M) = m\}$ has co-dimension $(H-m)(p-m)$ and applying (3.13) we get the following proposition [8].

Proposition 3.14. *If (3.1), (3.2), and (3.3) hold, we have*

$$\widehat{\Lambda}_3 \xrightarrow{\text{d}} \chi^2_{(H-m)(p-m)}.$$

In general a minimizer

$$\widehat{M_c} = \underset{\operatorname{rank}(M)=m}{\operatorname{argmin}} \operatorname{vec}(\widehat{M} - M)^T \widehat{\Gamma}^{-1} \operatorname{vec}(\widehat{M} - M)$$

does not have an explicit form as it was for the constrained matrix associated with $\widehat{\Lambda}_1$ and $\widehat{\Lambda}_2$. Therefore, we define

$$M_0^* = \widehat{M_c} + n^{-1/2} W^* \qquad \text{with} \quad W^* | \widehat{P} \overset{\text{d}}{\to} W \quad \text{a. s.}, \tag{3.21}$$

where W is defined in (3.1). We also define the associated CS bootstrap statistic

$$\Lambda_3^* = n \min_{\operatorname{rank}(M)=m} \operatorname{vec}(M_0^* - M)^T \Gamma^{*-1} \operatorname{vec}(M_0^* - M),$$

and applying Theorem 3.6 we have the following result.

Proposition 3.15. *If (3.1), (3.2), (3.3), (3.21),* $\Gamma^* \overset{\mathbb{P}}{\to} \Gamma$*, and* $\widehat{M} \overset{\text{a.s.}}{\to} M_0$ *hold, then the test described in (3.4) with the statistic* $\widehat{\Lambda}_3$ *and calculation of quantiles with* Λ_3^* *is consistent.*

Remark 3.16. The set of assumptions needed to obtain Proposition 3.14 is stronger than the ones stated in propositions 3.9 and 3.12 ensuring the convergence of $\widehat{\Lambda}_1$ and $\widehat{\Lambda}_2$. As a consequence this is also true for Proposition 3.15 with respect to propositions 3.11 and 3.13. The main difference is that we add the assumption on Γ to be non deficient. This assumption cannot be alleviated in the statement but is not as restrictive in practice. On the first hand, if Γ is deficient the optimization under constraint has a free coordinate

8. See [26] for the original proof.

which implies the non-convergence of the minimizer. On the other hand, because of the semi-definite character of Γ the projection of $\widehat{M}$ on the null space of Γ is null. Then one can apply the proposition to the restriction of $\widehat{M}$ on the range of Γ. This is the case in the application to SDR in Section 3.4.

Remark 3.17. Unlike the situation of $\widehat{\Lambda}_1$ and $\widehat{\Lambda}_2$, an optimization algorithm is needed to obtain $\widehat{\Lambda}_3$ and Λ_3^*, this points out an important issue of such a procedure. In [22], the authors noticed that

$$\widehat{\Lambda}_3 = n \min_{A\in H_d, B\in\mathbb{R}^{d\times l}} (\mathrm{vec}(\widehat{M}) - \mathrm{vec}(AB))^T \widehat{\Gamma}^{-1} (\mathrm{vec}(\widehat{M}) - \mathrm{vec}(AB))$$

where H_d is the set of orthogonal basis lying in $\mathbb{R}^p$ with dimension d. We follow their algorithm in the computation of $\widehat{\Lambda}_3$ (see [22], Section 3.3 for the details).

3.3.4 The statistics $\widehat{\Lambda}_1$, $\widehat{\Lambda}_2$, $\widehat{\Lambda}_3$ through an example

In the introduction, we already mentioned several drawbacks and advantages of the use of $\widehat{\Lambda}_1$, $\widehat{\Lambda}_2$, or $\widehat{\Lambda}_3$. The remark relied on both pivotality of the statistics and large matrix inversion. Here we develop another point of view related to the algebraic nature of the statistics. Facing the representation provided by Table 3.1, each statistic $\widehat{\Lambda}_1$ and $\widehat{\Lambda}_2$ evaluates a different distance between $\widehat{M}$ and $\widehat{M}_c$. The first one is the distance that is optimized, but the second is another one. This has raised the issue we present here through the following example. For the sake of clarity, we consider

$$\widehat{M} = \begin{pmatrix} \widehat{\lambda}_1 & 0 \\ 0 & \widehat{\lambda}_2 \end{pmatrix} \qquad \text{with } \widehat{\lambda}_k = \frac{1}{n}\sum_{i=1}^n \lambda_{k,i}, \text{ for } k = 1, 2, \text{ and } (\lambda_{k,i})_{k,i} \text{ i.i.d.},$$

and we test $H_0 : d_0 = 1$ against $H_1 : d_0 > 1$. We assume that $\widehat{\lambda}_1 > \widehat{\lambda}_2$, we have $\widehat{\Lambda}_1 = n\widehat{\lambda}_2^2$. Otherwise, one can show that $\widehat{\Lambda}_2 = n\frac{\widehat{\lambda}_2^2}{\widehat{v}_2} + o_{\mathbb{P}}(1)$, with $\widehat{v}_k = \overline{(\lambda_k - \overline{\lambda_k})^2}$. For $\widehat{\Lambda}_3$ it is clear that the minimization can be done over the diagonal matrix $\mathrm{diag}(\lambda_1, \lambda_2)$ and one has

$$\widehat{\Lambda}_3 = n \operatorname*{argmin}_{\lambda_1\lambda_2=0} \left\{ \frac{\widehat{\lambda}_1 - \lambda_1}{\widehat{v}_1} + \frac{\widehat{\lambda}_2 - \lambda_2}{\widehat{v}_2} \right\} + o_{\mathbb{P}}(1) = n \min\left(\frac{\widehat{\lambda}_1^2}{\widehat{v}_1}, \frac{\widehat{\lambda}_2^2}{\widehat{v}_2} \right) + o_{\mathbb{P}}(1).$$

Accordingly, by Proposition 3.11, 3.13 and 3.15, the three tests can be summarized by

$$\begin{aligned} n\widehat{\lambda}_2^2 \qquad & \text{compared to} \qquad v_2\chi_1^2, \\ n\frac{\widehat{\lambda}_2^2}{\widehat{v}_2} \qquad & \text{compared to} \qquad \chi_2^2, \\ n\min\left(\frac{\widehat{\lambda}_1^2}{\widehat{v}_1}, \frac{\widehat{\lambda}_2^2}{\widehat{v}_2} \right) \qquad & \text{compared to} \qquad \chi_2^2, \end{aligned}$$

where $v_k = \text{var}(\lambda_{k,1})$. Assume there is less variance on the estimate of the smallest eigenvalue, i.e. $v_1 > v_2$ such that $\frac{\widehat{\lambda}_1^2}{\widehat{v}_1} < \frac{\widehat{\lambda}_2^2}{\widehat{v}_2}$, this situation may arise when $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ have similar values but different variances. Then to conduct the test, the statistic $\frac{\widehat{\lambda}_1^2}{\widehat{v}_1}$ is a better choice than $\frac{\widehat{\lambda}_2^2}{\widehat{v}_2}$. As a consequence, unlike $\widehat{\Lambda}_1$ and $\widehat{\Lambda}_2$, the statistic $\widehat{\Lambda}_3$ appears as a coherent choice because its associated minimization takes into account the variance of the estimation.

3.4 Application to sufficient dimension reduction

We focus on a particularly famous method in SDR called sliced inverse regression (SIR) which has been introduced in [66] to deal with the regression model

$$Y = f(PX, \varepsilon) \tag{3.22}$$

where $\varepsilon \perp\!\!\!\perp X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, and P is a projector on the vector space E with dimension $d_0 < p$, called the central subspace. The objective is to estimate E. If X is elliptically distributed, then we have that $\Sigma^{-1}(\mathbb{E}[(X - \mathbb{E}[X])\psi(Y)] \in E$ with $\Sigma = \text{var}(X)$, for any measurable function ψ. Accordingly, in order to recover the whole central subspace one needs to consider many functions ψ. For a given family of functions $(\psi_h)_{1\leq h\leq H}$ we define $\Psi = (\psi_1(Y), ..., \psi_H(Y))^T$. Under some additional conditions [70], the image of the matrix $\Sigma^{-1/2}\,\text{cov}(X, \Psi(Y))$ is equal to $\Sigma^{1/2}E$. Then one can make the svd of an estimator of this matrix to obtain d_0 vectors that form an estimated basis of $\Sigma^{1/2}E$. Motivated by the curse of dimensionality, the estimation of d_0 is one of the most crucial points in SDR. To make that possible, a popular way consists in estimating the rank of $\Sigma^{-1/2}\,\text{cov}(X, \Psi)$ using the hypothesis testing framework given by (3.4) (see for example [66], [15] and [22]). Since we are interested in estimating the rank, we prefer to deal directly with $\text{cov}(X, \Psi)$ to avoid the introduction of an additional noise due to the estimation of the matrix Σ. Assume that $((X_1, Y_1), \cdots, (X_n, Y_n))$ is an i.i.d. sequence from model (3.22), denote by $\widehat{P}$ its associated empirical c.d.f. and define the quantity

$$C = \mathbb{E}[K], \qquad \text{with } K = (X - \mathbb{E}[X])(\Psi(Y) - \mathbb{E}[\Psi(Y)])^T,$$

associated with its empirical estimator

$$\widehat{C} = \overline{\widehat{K}}, \qquad \text{with } \widehat{K}_i = (X_i - \overline{X})(\Psi_i - \overline{\Psi})^T, \quad \text{and } \Psi_i = \Psi(Y_i).$$

We apply the CS bootstrap to calculate the quantiles of each statistic. Facing (3.20) and (3.21), we use an independent weighted bootstrap to reproduce the asymptotic law of $\sqrt{n}(\widehat{C} - C)$, that is we define the bootstrap matrix

$$C^* = \widehat{C}_c + \overline{K^*}, \quad \text{with} \quad K_i^* = w_i(\widehat{K}_i - \overline{\widehat{K}}) \tag{3.23}$$

where $\widehat{C}_c$ stands for the solution of an optimization problem depending on the selected statistic Λ_1, Λ_2 or Λ_3 (see Section 3.3 for the details) and (w_i) is a sequence of i.i.d.

random variables. We also define

$$V = \mathrm{var}(\mathrm{vec}(K)) \qquad \text{and} \qquad V^* = \frac{1}{n}\sum_{i=1}^{n} \mathrm{vec}(K_i^* - \overline{K^*})\,\mathrm{vec}(K_i^* - \overline{K^*})^T.$$

To apply propositions 3.11, 3.13, and 3.15, we need the following result which is of particular interest since it provides a new bootstrap procedure for SIR that is different than the one proposed in [3].

Proposition 3.18. *Assume that* $\mathbb{E}[\|X\|^2] < +\infty$, $\mathbb{E}[\|\Psi(Y)\|^2]$ *and* $\mathbb{E}[\|K\|_F^4]$ *are finite, if moreover* (w_i) *is an i.i.d. sequence of real random variables with mean* 0 *and variance* 1*, then we have*

$$\mathcal{L}_\infty(n^{1/2}\ \overline{K^*}|\widehat{P}) = \mathcal{L}_\infty(n^{1/2}(\widehat{C} - C)) \quad \textit{a.s. and} \quad V^* \xrightarrow{\mathbb{P}} V \quad \textit{conditionally a.s..}$$

Remark 3.19. Taking a partition $\{I(h),\ h = 1, \ldots, H\}$ of the range of Y we recover the original SIR method with the family formed by the $p_h^{-1/2}\mathbb{1}_{\{Y\in I(h)\}}$'s with $p_h = \mathbb{P}(Y \in I(h))$. Then $C_{\mathrm{SIR}} = \Sigma^{-1/2}\,\mathrm{cov}(X, \mathbb{1})D^{-1/2}$ with $\mathbb{1} = (\mathbb{1}_{\{Y_i\in I(1)\}}, \ldots, \mathbb{1}_{\{Y_i\in I(H)\}})^T$ and $D = \mathrm{diag}(p_h)$, is estimated by $\widehat{C}_{\mathrm{SIR}} = \widehat{\Sigma}^{-1/2}\overline{(X - \overline{X})\mathbb{1}^T}\widehat{D}^{-1/2}$ with $\widehat{D} = \mathrm{diag}(\widehat{p}_h)$, $\widehat{p}_h = \overline{\mathbb{1}_{\{Y\in I(h)\}}}$, $\widehat{\Sigma} = \overline{(X - \overline{X})(X - \overline{X})^T}$. We have the expansion

$$\begin{aligned} n^{-1/2}(\widehat{C}_{\mathrm{SIR}} - C_{\mathrm{SIR}}) &= n^{-1/2}\Sigma^{-1/2}(\overline{(X - \mathbb{E}[X])\mathbb{1}^T} - \mathrm{cov}(X, \mathbb{1}))D^{-1/2} \\ &\quad - \Sigma^{-1/2}n^{-1/2}(\widehat{\Sigma}^{1/2} - \Sigma^{1/2})C_{\mathrm{SIR}} - C_{\mathrm{SIR}}n^{-1/2}(\widehat{D}_p^{1/2} - D_p^{1/2})D_p^{-1/2} + o_{\mathbb{P}}(1). \end{aligned}$$

As a consequence, the matrix $\Sigma^{-1/2}$ and the weights p_h's are playing an important role on the asymptotic of the matrix SIR. They introduce some other terms in the asymptotic distribution and clearly the simple bootstrap presented before does not work for SIR as it was originally defined. Even if we believe that a more evolved weighted bootstrap works to bootstrap $\sqrt{n}(\widehat{C}_{\mathrm{SIR}} - C_{\mathrm{SIR}})$, we emphasize that it may be less accurate than the one we propose since it complicates the asymptotic without being necessary for testing the rank.

Recall that m is a non-negative integer, for $k \in \{1, 2, 3\}$ and $B \in \mathbb{N}^*$ we calculate independent copies $\Lambda_{k,1}^*, ..., \Lambda_{k,B}^*$ with the CS bootstrap algorithm corresponding to each statistic. Then we estimate the quantile with

$$q_k^*(\alpha) = \inf_{t\in\mathbb{R}}\{F_k^*(t) > \alpha\} = \Lambda_{k,(\lceil B\alpha\rceil)}^*, \qquad \text{where } F_k^*(t) = \frac{1}{B}\sum_{b=1}^{B}\mathbb{1}_{\{\Lambda_{k,b}^*\le t\}},$$

$\lceil\cdot\rceil$ is the integer ceiling function and $\Lambda_{k,(\cdot)}^*$ stands for the rank statistic associated to the sample $\Lambda_{k,1}^* \ldots \Lambda_{k,B}^*$. On the first hand, we conduct the test described by (3.4) using the CS bootstrap, i.e.

$$H_0 \text{ is rejected if} \qquad \widehat{\Lambda}_k > \widehat{q}_k^*(\alpha). \tag{3.24}$$

On the other hand, the traditional test is conducted by comparing the statistic $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$ to the quantile of their asymptotic law respectively given by propositions 3.12 and 3.14. For $\widehat{\Lambda}_1$, in general the limit in law is quite complicated [9] (see Proposition 3.9), so that we use approximations : the Wood's approximation (see [79]) as it is computed in the R software, an adjusted version $\widehat{\Lambda}_{1,\text{adj.}} = \widehat{\Lambda}_1/a \xrightarrow{\text{d}} \chi^2_b$, with $a = \sum_{k=1}^s \omega^2 / \sum_{k=1}^s \omega_k$, $b = (\sum_{k=1}^s \omega_k)^2 / \sum_{k=1}^s \omega_k^2$, and a re-scaled version $\widehat{\Lambda}_{1,\text{sc}} = \widehat{\Lambda}_1/c \xrightarrow{\text{d}} \chi^2_s$, $c = \overline{\omega}$ (see [5] for these two corrections).

In all the simulations we compute the matrix $\widehat{C}$ by taking $\Psi(t) = (\mathbb{1}_{\{y \in I(1),\ldots,y \in I(H)\}})$ where the $I(h)$'s form an equi-partition of the range of the data $Y_1, \ldots, Y_n$. In the whole study we put $(p, H) = (6, 5)$, $B = 1000$ and we consider $n = 50, 100, 200, 500$. Although the parameter H does not really affect the SIR method, we choose it globally good with respect to all the situations.

The first model we study is the following standard model :

$$\text{Model I :} \quad Y = X_1 + .1e \quad \text{with } e \perp\!\!\!\perp X, \quad X \overset{\text{d}}{=} \mathcal{N}(0, I), \quad e \overset{\text{d}}{=} \mathcal{N}(0, 1).$$

In order to highlight guidelines (A) and (B), we produce in figure 3.1 two graphics each representing situation under H_1 and H_0 for the statistic $\widehat{\Lambda}_3$. Similar graphics dealing with $\widehat{\Lambda}_2$ have been drawn but are not presented here. On the first one we see that even if the sample is under H_1 the bootstrap distribution reflects H_0. As a consequence, guideline (A) is satisfied and the power of the bootstrap test is going to 1. The second graph shows that the statistic distribution is closer to the bootstrap distribution than its asymptotic distribution. This has no reason to occur when the statistic is not pivotal (see the introduction and [48] for the details). As a consequence, we believe that this good fitting is due to Guideline B.

In figure 3.2 we analyse the asymptotic distribution of $\widehat{q}(\alpha)$ in model I for each statistic. To measure the error we consider the behaviour of

$$F_n(\widehat{q}(\alpha)),$$

which is optimally equal to $1 - \alpha$. To make that possible, F_n is estimated with a large sample size so that the estimation error is negligible. Then we run over 100 samples the CS bootstrap to provide, for each sample, a bootstrap estimation of the quantile $\widehat{q}(\alpha)$. The associated boxplot for $n = 100$, 200, 500 are provided in Figure 3.2. As a consequence, we may notice that the behaviour of $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$ are quite similar facing the one of $\widehat{\Lambda}_1$. Even if every boxplot argues for convergence to $1 - \alpha$, testing with $\widehat{\Lambda}_1$ seems a better choice when n is small because of a quasi immediate convergence of the bias. When n increase, this is no longer evident because the variance of either $\widehat{\Lambda}_2^*$ or$\widehat{\Lambda}_3^*$ is smaller.

Furthermore, we go into details in Table 3.2 by running Model I over 5000 samples. For each of them and every statistic, we conduct the bootstrap test (3.24) and its traditional

9. When the predictors are normally distributed, it has been shown that $\widehat{\Lambda}_1$ is asymptotically chi-squared distributed (see [15]). The authors also pointed out that it was less robust than the weighted chi-squared asymptotic as soon as the predictors distribution deviates from normality. As a result, we keep in the nonparametric framework by avoiding such asymptotic in this simulation study.

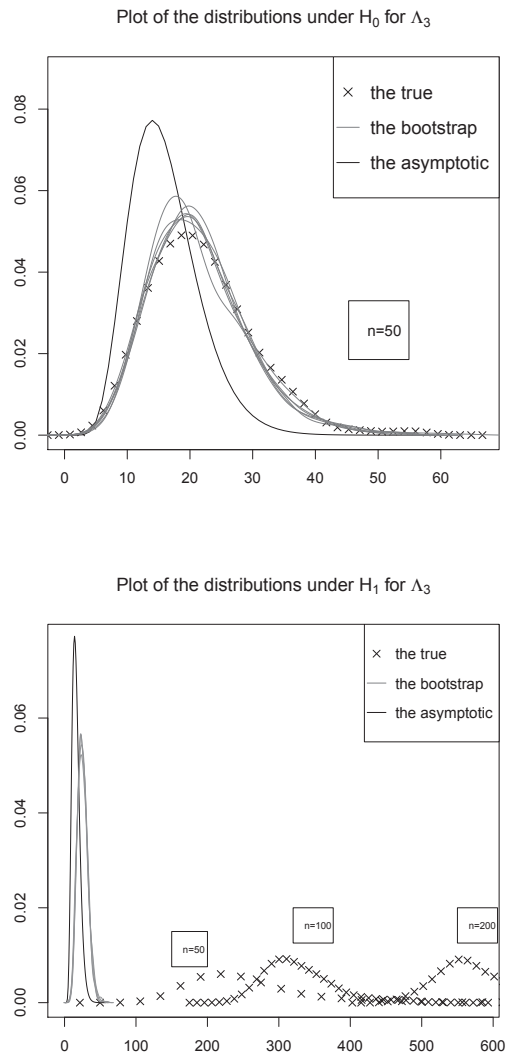

FIGURE 3.1 – Plot of the asymptotic distribution, and the estimated distribution of the statistic and the bootstrap statistic for $\widehat{\Lambda}_3$ in the case of Model I.

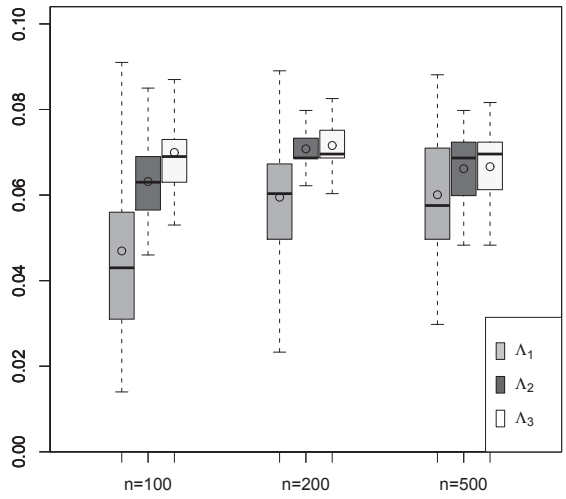

FIGURE 3.2 – Bowplot over 100 samples of $\widehat{q}(\alpha)$ for $\widehat{\Lambda}_1$, $\widehat{\Lambda}_2$, $\widehat{\Lambda}_3$ and $\alpha = 0.95$ in the case of Model I for different values of n .

version. The table presents for each $m = 0, \cdots, d_0$, the proportion of rejected tests. This corresponds to either estimate of the power or estimate of the level of the test.

Although it has not the best power, the clear winner is the tests based on $\widehat{\Lambda}_1$. Inside this group, for any sample number, the bootstrap and the rescaled version are the closest to the nominal level. Concerning $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$ the results are quite impressive when n is small : for $n = 50$, whereas traditional testing makes a type I error 30% of the time, the bootstrap testing goes wrong around 7%. This confirms observation on the second graph of Figure 3.1.

In Table 3.3 and Table 3.4 we consider the same model than Model I excepted that we change the distribution of the predictors : in Model Ia, X has independent coordinates with a student distribution with 5 degrees of freedom, in Model Ib, $X \overset{\mathrm{d}}{=} .1X_1\epsilon + X_2(1-\epsilon)$ with $\epsilon \overset{\mathrm{d}}{=} \mathcal{B}(1/2)$, $X_1 \overset{\mathrm{d}}{=} \mathcal{N}((6, 0, \cdots, 0), I)$, $X_2 \overset{\mathrm{d}}{=} \mathcal{N}(0, I)$. For this two models, we have similar conclusions to model I with two new things. First, the rescaled version is not robust to the distribution of the predictors (Table 3.4). Second, the algorithm employed to optimized $\widehat{\Lambda}_3$ could fail at very small sample size.

We introduce a non linear relationship by considering the model

$$\text{Model II :} \quad Y = \tanh(X_1) + .1e \quad \text{with } e \perp\!\!\!\perp X, \quad X \overset{\mathrm{d}}{=} \mathcal{N}(0, I), \quad e \overset{\mathrm{d}}{=} \mathcal{N}(0, 1).$$

In Table 3.5, we present similar results as before with the difference that the nominal level is $\alpha = 1\%$ in order to highlight differences in the power of each test. Again, the CS

n	m	$\widehat{\Lambda}_1$				$\widehat{\Lambda}_2$		$\widehat{\Lambda}_3$	
		Wood	Resc.	Adj.	CB $\widehat{\Lambda}_1$	$\widehat{\Lambda}_2$	CB $\widehat{\Lambda}_2$	$\widehat{\Lambda}_3$	CB $\widehat{\Lambda}_3$
50	0	0.9988	0.9998	0.9988	0.9988	1.0000	1.0000	1.0000	1.0000
	1	0.0326	0.0590	0.0336	0.0494	0.3466	0.0744	0.3098	0.07
100	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.0386	0.052	0.0388	0.0456	0.1494	0.0676	0.1466	0.0722
200	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.0474	0.055	0.0476	0.0514	0.096	0.0646	0.0954	0.0664
500	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.0492	0.0514	0.0494	0.0516	0.0656	0.0584	0.0654	0.0584

Table 3.2 – Estimated levels and power in Model I for $\alpha = 5\%$.

n	m	$\widehat{\Lambda}_1$				$\widehat{\Lambda}_2$		$\widehat{\Lambda}_3$	
		Wood	Resc.	Adj.	CB $\widehat{\Lambda}_1$	$\widehat{\Lambda}_2$	CB $\widehat{\Lambda}_2$	$\widehat{\Lambda}_3$	CB $\widehat{\Lambda}_3$
50	0	0.9646	0.9928	0.9656	0.9682	1.0000	1.0000	1.0000	1.0000
	1	0.0318	0.0628	0.0324	0.0496	0.3412	0.0588	0.3042	0.0628
100	0	0.9996	1.0000	0.9996	0.9996	1.0000	1.0000	1.0000	1.0000
	1	0.0336	0.0486	0.0344	0.0412	0.1516	0.0696	0.1432	0.0718
200	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.0378	0.0486	0.038	0.0424	0.0844	0.0602	0.0832	0.0604
500	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.0454	0.0502	0.0458	0.0474	0.0638	0.0606	0.0634	0.0608

Table 3.3 – Estimated levels and power in Model Ia for $\alpha = 5\%$.

n	m	$\widehat{\Lambda}_1$				$\widehat{\Lambda}_2$		$\widehat{\Lambda}_3$	
		Wood	Resc.	Adj.	CB $\widehat{\Lambda}_1$	$\widehat{\Lambda}_2$	CB $\widehat{\Lambda}_2$	$\widehat{\Lambda}_3$	CB $\widehat{\Lambda}_3$
50	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.034	0.1072	0.034	0.0378	0.2122	0.0396	0.1394	0.015
100	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.037	0.0904	0.0374	0.0404	0.0986	0.0572	0.0614	0.0284
200	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.0484	0.096	0.0488	0.0518	0.0708	0.066	0.056	0.0506
500	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.0486	0.0912	0.0486	0.0490	0.0598	0.0664	0.0612	0.0674

Table 3.4 – Estimated levels and power in Model Ib for $\alpha = 5\%$.

n	m	$\widehat{\Lambda}_1$				$\widehat{\Lambda}_2$		$\widehat{\Lambda}_3$	
		Wood	Resc.	Adj.	CB $\widehat{\Lambda}_1$	$\widehat{\Lambda}_2$	CB $\widehat{\Lambda}_2$	$\widehat{\Lambda}_3$	CB $\widehat{\Lambda}_3$
50	0	0.9308	0.9884	0.9428	0.9448	1.0000	0.9988	1.0000	0.9988
	1	0.0036	0.0148	0.0050	0.0086	0.1816	0.0148	0.1404	0.0130
100	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.0072	0.0122	0.0082	0.0096	0.0536	0.02	0.0496	0.021
200	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.0076	0.0114	0.0086	0.0102	0.0252	0.0192	0.0248	0.02
500	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.0068	0.0076	0.007	0.0082	0.012	0.011	0.012	0.011

TABLE 3.5 – Estimated levels and power in Model II for $\alpha = 1\%$.

bootstrap induces a large improvement of the accuracy of the test with $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$. At $n = 50$, the test based on $\widehat{\Lambda}_1$ is less powerful than the others but it is more accurate under H_0. The winner remains the CS bootstrap with $\widehat{\Lambda}_1$. A new important things is that at $n = 500$, it seems better to use the CS bootstrap with $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$. Actually this is due to the variance of the formers which is smaller than the variance of Λ_1^* as it was already highlighted in Figure 3.2.

We conclude by increasing difficulty considering the following model, introduced in [66],

$$\text{Model III}: \qquad Y = \frac{X_1}{.5 + (X_2 + 2)^2} + e \qquad\qquad e \perp\!\!\!\perp X, \qquad X \overset{\mathrm{d}}{=} \mathcal{N}(0, I)$$

We still present in Table 3.6 the estimated level and power with the nominal level $\alpha = 2\%$ for each test. For such a model the conclusions are quite mitigated because it induces a trade-off between high power and accurate level. Indeed when n is small, the better powers are provided by the traditional tests with $\widehat{\Lambda}_2$ and $\widehat{\Lambda}_3$. Nevertheless the more accurate levels can be found looking at the CS bootstrap with $\widehat{\Lambda}_2$ (n = 100) or $\widehat{\Lambda}_1$ (n = 200). Moreover the tests associated to $\widehat{\Lambda}_1$ without bootstrap are the worst concerning this model. Accordingly, the simulation study highlighted the good behaviour of the CS bootstrap : in every model it improves the accuracy of the traditional test for each statistic. One may remember that the bias of the CS bootstrap with $\widehat{\Lambda}_1$ has the faster rate of convergence with respect to the CS bootstrap of $\widehat{\Lambda}_2$ or $\widehat{\Lambda}_3$. Otherwise, the variance of $\widehat{\Lambda}_1^*$ may be greater than the variance of $\widehat{\Lambda}_2^*$ or $\widehat{\Lambda}_3^*$. Finally, for the simple models it seems better to use the CS bootstrap with the statistic $\widehat{\Lambda}_1$.

3.5 Concluding remarks

Along this study, we found that the main advantages of the CS bootstrap are the following.

1. The CS bootstrap is a powerful alternative to the asymptotic comparison. This argument is even stronger since the asymptotic law can be unknown (or difficult to

n	m	$\widehat{\Lambda}_1$				$\widehat{\Lambda}_2$		$\widehat{\Lambda}_3$	
		Wood	Resc.	Adj.	CB $\widehat{\Lambda}_1$	$\widehat{\Lambda}_2$	CB $\widehat{\Lambda}_2$	$\widehat{\Lambda}_3$	CB $\widehat{\Lambda}_3$
50	0	0.9950	0.9992	0.9962	0.9960	1.0000	0.9966	1.0000	0.9966
	1	0.3750	0.5342	0.3990	0.4676	0.9074	0.5066	0.8344	0.3270
	2	0.0078	0.0156	0.0086	0.0240	0.0620	0.0164	0.0344	0.0136
100	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.9330	0.9556	0.9368	0.9446	0.9952	0.9842	0.9934	0.9806
	2	0.0134	0.0176	0.0138	0.0210	0.0306	0.0228	0.0266	0.0278
200	0	1.000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	1.000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000	1.0000
	2	0.0154	0.0182	0.0158	0.0198	0.025	0.024	0.0244	0.026
500	0	1.0000	1.000	1.0000	1.0000	1.0000	1.000	1.0000	1.0000
	1	1.0000	1.000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	2	0.0184	0.0194	0.0184	0.02	0.0228	0.0228	0.0228	0.023

Table 3.6 – Estimated levels and power in Model II for $\alpha = 2\%$.

estimate) or the asymptotic law remains too much different from the statistic law (e.g. large matrix inversion).

2. By Theorem 3.6, which provides its consistency, the CS bootstrap works under mild assumptions. Essentially, we ask the submanifold to be locally smooth, and we require to be able to bootstrap the unconstrained estimator $\widehat{M}$.

3. The CS bootstrap is computationally as simple as the considered statistic.

4. In the case of rank testing, the CS bootstrap clearly improves the accuracy of traditional testing (cf. the simulation study).

Besides, there exist some natural extensions of the previous work. First although it is suitable for testing, the form of the objective function we minimize is quiet restrictive. For example, we believe that the CS bootstrap could be extended to M and Z estimation. Secondly, conditions that guarantee

$$\widehat{q}(\alpha) = q_n(\alpha) + o_{\mathbb{P}}(n^{-1/2})$$

have not been provided yet. This would valid theoretically the use of the CS bootstrap with respect to traditional testing.

3.6 Proofs

Proof of Lemma 3.3

The whole proof is made conditionally on the sample. By definition of $\widehat{\theta}_c$, with high probability, A^* is full rank for n large enough, we have

$$\|A^{*1/2}(\theta_c^* - \theta_c)\| \leq \|A^{*1/2}(\theta_c^* - \theta_0^*)\| + \|A^{*1/2}(\theta_0^* - \theta_c)\| \leq 2\|A^{*1/2}(\theta_0^* - \theta_c)\|. \quad (3.25)$$

Then since $\theta_0^* - \widehat{\theta}_c \xrightarrow{\mathbb{P}} 0$, $\widehat{\theta}_c \to \theta_c$ and because $A^* \xrightarrow{\mathbb{P}} A$ is full rank, one gets that $\theta_c^* \xrightarrow{\mathbb{P}} \theta_c$. Therefore, since θ_c is $\mathcal{M}$-nonsingular and reffering to Definition 3.1, we get

$$\operatorname*{argmin}_{\theta \in \mathcal{M}} \|\Gamma^{*1/2}(\theta_0^* - \theta)\| = \operatorname*{argmin}_{g(\theta)=0} \|\Gamma^{*1/2}(\theta_0^* - \theta)\|,$$

with g continuously differentiable on θ_c and $J_g(\theta_c)$ full rank. By assumption on g, θ_c^*, at least for n large enough, satisfies the first order conditions, that are

$$\begin{cases} A^*(\theta_0^* - \theta_c^*) - J_g^T(\theta_c^*)\lambda_n^* = 0 \\ g(\theta_c^*) = 0 \end{cases}$$

where λ_n^* is the Lagrange multiplier. Using a Taylor expansion of g around $\widehat{\theta}_c$, we get $g(\theta_c^*) = g(\widehat{\theta}_c) + J_g^T(\widehat{\theta}_c)(\theta_c^* - \widehat{\theta}_c) + o_{\mathbb{P}}(\|\theta_c^* - \widehat{\theta}_c\|)$, and with the previous equations we have

$$\begin{pmatrix} A^* & J_g^T(\theta_c^*) \\ J_g(\widehat{\theta}_c) & 0 \end{pmatrix} \begin{pmatrix} \theta_c^* - \widehat{\theta}_c \\ \lambda_n^* \end{pmatrix} = \begin{pmatrix} A^*(\theta_0^* - \widehat{\theta}_c) \\ o_{\mathbb{P}}(\|\theta_c^* - \widehat{\theta}_c\|) \end{pmatrix}.$$

Now by Slutsky's lemma, we get

$$\begin{pmatrix} A & J_g^T(\theta_c) \\ J_g(\theta_c) & 0 \end{pmatrix} \begin{pmatrix} n^{1/2}(\theta_c^* - \widehat{\theta}_c) \\ n^{1/2}\lambda_n^* \end{pmatrix} = n^{1/2} \begin{pmatrix} A(\theta_0^* - \widehat{\theta}_c) \\ 0 \end{pmatrix} + o_{\mathbb{P}}(1),$$

and the conclusion follows by multiplying on the left by the matrix

$$\left(A^{-1} - PA^{-1}, \quad A^{-1}J_g^T(\theta_c)(J_g(\theta_c)A^{-1}J_g^T(\theta_c))^{-1}\right)$$

with $P = A^{-1}J_g^T(\theta_c)(J_g(\theta_c)A^{-1}J_g^T(\theta_c))^{-1}J_g(\theta_c)$.

□

Proof of Theorem 3.6

The proof is divided in two parts each corresponding to the level and the power of the test. Assume H_0 and define F_n and F_∞ respectively as the c.d.f. of $\widehat{\Lambda}$ and the weak limit of F_n. Note that we can apply Proposition 3.4 to get

$$n^{1/2} \begin{pmatrix} \widehat{\theta} - \theta_0 \\ \widehat{\theta}_c - \theta_0 \end{pmatrix} = n^{1/2} \begin{pmatrix} I \\ I - P \end{pmatrix} (\widehat{\theta} - \theta_0) + o_{\mathbb{P}}(1),$$

and Theorem 3.5 to get conditionally a.s.

$$n^{1/2} \begin{pmatrix} \theta_0^* - \widehat{\theta}_c \\ \theta_c^* - \widehat{\theta}_c \end{pmatrix} = n^{1/2} \begin{pmatrix} I \\ I - P \end{pmatrix} (\theta_0^* - \widehat{\theta}_c) + o_{\mathbb{P}}(1).$$

with P detailed in the statement of Proposition 3.4. Using (3.12), (3.19) and Slutsky's theorem we have

$$\mathcal{L}_\infty(\Lambda^*|\widehat{P}) = \mathcal{L}_\infty(\widehat{\Lambda}) \quad \text{a. s.}.$$

In other words, with probability 1, $\widehat{F}$ converges pointwise to F_∞. As in [76] chapter 23, Lemma 3, consider Δ the set of discontinuity of F_∞^{-1}. For every $\alpha \in (0,1)\backslash\Delta$, we have $\widehat{q}(\alpha) \longrightarrow q(\alpha)$ a.s. (see for instance [76], chapter 21). Using Slutsky's theorem, we get $\mathcal{L}_\infty(\widehat{\Lambda} - \widehat{q}(\alpha)) = \mathcal{L}_\infty(\widehat{\Lambda} - q(\alpha))$, accordingly

$$\mathbb{P}(\widehat{\Lambda} \leq \widehat{q}(\alpha)) \longrightarrow F_\infty(q(\alpha)) \quad \text{for all } \alpha \in (0,1)\backslash\Delta.$$

Because F_∞ is continuous $F_\infty(q(\alpha)) = \alpha$. Since F_∞ is non-decreasing, Δ is denumerable, since $\alpha \mapsto \mathbb{P}(\widehat{\Lambda} \leq \widehat{q}(\alpha))$ is non-decreasing with continuous limit, the convergence is uniform and so holds for every $\alpha \in (0,1)$. This concludes the proof for the level. It remains to show that the power of the test goes to 1. Assume H_1 and let $\alpha \in (0,1)$, the statistic $\widehat{\Lambda}$ goes to infinity in probability and it suffices to show that with probability 1 the bootstrap quantile $\widehat{q}(\alpha)$ remains bounded. This means exactly that conditionally a.s. the sequence Λ^* is tight. Note that conditionally a.s. we have

$$\Lambda^* \leq n\|A^{*1/2}(\widehat{\theta}_c - \theta_0^*)\|^2 = \widetilde{\Lambda}^*,$$

where $\widetilde{\Lambda}^*$ converges in distribution by (3.18), and is therefore tight. □

Proof of Proposition 3.9

We have

$$\widehat{\Lambda}_1 = \|n^{1/2}\widehat{Q}_1\widehat{M}\widehat{Q}_2\|_F^2 = \|n^{1/2}\,\mathrm{vec}(\widehat{Q}_1\widehat{M}\widehat{Q}_2)\|^2.$$

By the Delta method and because H_0 is realized, we can apply convergence results about eigenprojectors to both matrices $\widehat{M}^T\widehat{M}$ and $\widehat{M}\widehat{M}^T$ to obtain the $\sqrt{n}$-convergence for $\widehat{Q}_1$ and $\widehat{Q}_2$. Then we write

$$\begin{aligned} n^{1/2}\widehat{Q}_1\widehat{M}\widehat{Q}_2 &= n^{1/2}\widehat{Q}_1(\widehat{M} - M)\widehat{Q}_2 + n^{1/2}(\widehat{Q}_1 - Q_1)M(\widehat{Q}_2 - Q_2) \\ &= n^{1/2}Q_1(\widehat{M} - M)Q_2 + O_{\mathbb{P}}(n^{-1/2}), \end{aligned}$$

which suffices to obtained the first statement of the theorem. For the second statement, the symmetric matrix $(Q_2 \otimes Q_1)\Gamma(Q_2 \otimes Q_1)$ is estimated consistently by $(\widehat{Q}_2 \otimes \widehat{Q}_1)\widehat{\Gamma}(\widehat{Q}_2 \otimes \widehat{Q}_1)$ and so are its eigenvalues. □

Proof of Proposition 3.12

We can notice that $\sqrt{n}\widehat{Q}_1\widehat{M}\widehat{Q}_2$ has the same asymptotic law than $\sqrt{n}Q_1(\widehat{M} - M)Q_2$ whose asymptotic variance is consistently estimated by $[(\widehat{Q}_2 \otimes \widehat{Q}_1)\widehat{\Gamma}(\widehat{Q}_2 \otimes \widehat{Q}_1)]^+$ (see the proof of Proposition 3.9). □

Proof of Proposition 3.18

Recall that $\widehat{K}_i = (X_i - \overline{X})(\Psi_i - \overline{\Psi})$, $K_i^* = w_i(\widehat{K}_i - \overline{\widehat{K}})$ and define $K_i = (X_i - \mathbb{E}[X])(\Psi_i - \mathbb{E}[\Psi])$. First note that, by Slutsky's theorem, $\sqrt{n}\ \overline{K^*}$ has the same asymptotic law than $n^{-1/2}\sum_{i=1}^n w_i(\widehat{K}_i - \mathbb{E}[K])$. Then we can develop

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^n w_i(\widehat{K}_i - \mathbb{E}[K]) \\
&= n^{-1/2}\sum_{i=1}^n w_i((X_i - \mathbb{E}[X])(\Psi_i - \overline{\Psi})^T - \mathbb{E}[K]) + (\mathbb{E}[X] - \overline{X})n^{-1/2}\sum_{i=1}^n w_i(\Psi_i - \overline{\Psi})^T \\
&= n^{-1/2}\sum_{i=1}^n w_i(K_i - \mathbb{E}[K]) + n^{-1/2}\sum_{i=1}^n w_i(X_i - \mathbb{E}[X])(\mathbb{E}[\Psi] - \overline{\Psi})^T \\
&\qquad + (\mathbb{E}[X] - \overline{X})n^{-1/2}\sum_{i=1}^n w_i(\Psi_i - \overline{\Psi})^T.
\end{aligned}$$

Checking a Lindeberg condition as bellow to ensure the weak convergence of $n^{-1/2}\sum_{i=1}^n w_i(X_i - \mathbb{E}[X])$ and $n^{-1/2}\sum_{i=1}^n w_i(\Psi_i - \overline{\Psi})^T$, and using the Slutsky's theorem we get conditionally a.s.

$$n^{1/2}\ \overline{K^*} = n^{-1/2}\sum_{i=1}^n w_i(K_i - \mathbb{E}[K]) + O_{\mathbb{P}}(n^{-1/2}).$$

We can apply the multidimensional version of the Lindeberg's central limit theorem (see for instance [7], Corollary 18.2), provided that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\widehat{V}^{-1/2}w_i\xi_i\|^2 \mathbb{1}_{\{\|\widehat{V}^{-1/2}w_i\xi_i\| > \nu n^{1/2}\}} | \widehat{P}] \xrightarrow{\text{a.s.}} 0,$$

where $\xi_i = \mathrm{vec}(K_i - \mathbb{E}[K])$ and $\widehat{V} = \frac{1}{n}\sum_{i=1}^n (\xi_i - \overline{\xi})(\xi_i - \overline{\xi})^T$. The above convergence is a consequence of the Lebesgue domination theorem which ensure that each term of the sum goes to 0, afterwards we can conclude by the Cesaro's Lemma. Thus we have proved that conditionally a.s.

$$n^{-1/2}\widehat{V}^{-1/2}\sum_{i=1}^n w_i\xi_i \xrightarrow{\text{d}} \mathcal{N}(0, I),$$

and it remains to note that $\widehat{V} \xrightarrow{\text{a.s.}} V$ the variance of the limit in law of $\sqrt{n}(\widehat{C} - C)$ provided that K has a finite order 2 moment. For the second convergence, we note that conditionally a.s.

$$V^* - \widehat{V} = \frac{1}{n}\sum_{i=1}^n (w_i^2 - 1)\xi_i\xi_i^T + o_{\mathbb{P}}(1),$$

then by noting v_i a coordinate of $\xi_i \xi_i^T$ we calculate

$$\mathbb{E}\left[\left(n^{-1}\sum_{i=1}^{n}(w_i^2-1)v_i\right)^2\right] = n^{-2}\mathbb{E}[(w_i^2-1)^2]\sum_{i=1}^{n} v_i^2$$

which goes to 0 a.s. provided that K has a finite order 4 moment. We conclude by using the Markov inequality to get that $V^* \xrightarrow{\mathbb{P}} \widehat{V}$ conditionally a.s.. □

Chapter 4

Semiparametric estimation of the central mean subspace

Abstract : In this chapter, we consider the multiple index model

$$Y = g_0(\beta^T X) + \sigma(X)e,$$

where $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$ and $\beta \in \mathbb{R}^{p\times d_0}$. Our main purpose is the estimation of the space span(β). In 2001, the authors Hristache, Juditsky, Polzehl and Spokoiny studied an estimator of the quantity $\mathbb{E}[\nabla g(X)\psi(X)]$, which belongs to the index space, where $g = g_0 \circ \beta^T$, $\psi : \mathbb{R}^p \to \mathbb{R}$. More precisely, they used a local linear estimator to estimate the function ∇g and they showed the $\sqrt{n}$ consistency whenever $d_0 < 4$. Another related work is [53] where the authors Härdle and Stoker (1989) proposed to estimate the same quantity using an integration by part. The $\sqrt{n}$ consistency is achieved provided that the density of the predictors is estimated with a high order kernels. Nevertheless their method estimates β only when the index space has dimension 1. In this work, we follow the second approach, by estimating the vector

$$\mathbb{E}\left[\frac{Y\nabla\psi(X)}{f(X)}\right],$$

where f is the density of X. The latter quantity is equal to $\mathbb{E}[\nabla g(X)\psi(X)]$ and so lies in the index space under mild conditions. The estimation is done using a kernel estimate for the density f. Considering different functions Ψ, we show the $\sqrt{n}$ consistency of the associated method whatever the value of the dimension d_0.

Key words : Dimension reduction ; Multiple index model ; Kernel estimation.

4.1 Introduction

Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be a i.i.d. sequence of random variables generated from the model

$$Y_i = g(X_i) + \sigma(X_i)e_i, \tag{4.1}$$

where $X_i \in \mathbb{R}^p$, $Y_i \in \mathbb{R}$, $e_i \perp\!\!\!\perp X_i$, $\sigma : \mathbb{R}^p \to \mathbb{R}$ and $g : \mathbb{R}^p \to \mathbb{R}$ is called the link function. The functions σ and g are unknown and $p > 1$, $d_0 \geq 1$ (to avoid trivial cases). The *multiple index model* (MIM) specifies the link function as

$$g(x) = g_0(\beta^T x) \qquad \text{for every } x \in \mathbb{R}^p, \tag{4.2}$$

where $\beta \in \mathbb{R}^{p \times d}$. Under the model (4.1), such a β always exists, but is not necessarily unique. One prefers to define the index space, noted E_m as the space with minimal dimension generated by one β that verifies (4.2). Under some conditions, essentially X_1 has a density [70], $E_m = \text{span}(\beta)$ is unique, it is called the *index space* and the term *index* denotes any of its basis. Form now, we assume that E_m is unique.

One of the most popular approach to estimate E_m is called *inverse regression* [66] and is based on the estimation of the function $\mathbb{E}[X_1|Y_1 = \cdot]$. This approach usually assumes the *linearity condition* which is a strong restriction on the law of the explanatory variables [1]. This kind of approach has been studied in the chapters 1 and 2. In this work, we are interested in a different approach that does not rely on the linearity condition.

This work studies a *semiparametric* procedure whose first step is the estimation of the density of the vector X_1. In a second step, our procedure derives an estimator of the index space. More precisely a *kernel smoothing* is employed to estimate the density of X, then this one is used to construct an estimator of E_m. Typically, such an approach is rather different from the one mentioned before because it is free from strong assumptions on X_1 and only requires regularity conditions on f. The main issue raised by this approach relates to the rates of convergence of the estimated index space. Indeed, it is well known that nonparametric curve estimation provides slower rates of convergence than the parametric rates usually in $\sqrt{n}$. As a result, the challenge is to achieve the $\sqrt{n}$-consistency while estimating the index space without making strong assumption on X, as the linearity condition.

Several approaches have been proposed in order to estimate E_m non-parametrically. One of them is called the *semiparametric M-estimation*. Roughly speaking, this approach consists in reproducing classical M-estimation theory with the difference that the link function is estimated nonparametrically. One can see [59] for the original work. We also refer to work of Delecroix, Hristache and Patilea in 2003 [30] for some extensions. Both estimators introduced in these articles achieve the asymptotic normality with the rate root n. This approach suffers from an underlying optimisation problem which is often difficult to solve because of the non-convexity of the estimated link function.

1. Essentially, if the density of X_1 depends only on the radius, i.e. has a spherical distribution, then the linearity condition is satisfied.

Another approach focus on the gradient of the regression curve [2]. Under (4.2), we have $\nabla g(x) \in E_m$. In [56], the authors proposed to estimate the average of ∇g. For instance, in the case $d = 1$, their first estimator is given by

$$n^{-1} \sum_{i=1}^{n} \widehat{\nabla g}(X_i),$$

where $\widehat{\nabla g}$ is derived from an estimation of g by a local linear estimator (see [39] for some details). Under the fixed-design assumption, they show that the later estimator converges slowly, with rates less than $\sqrt{n}$. The original idea of their work is to use model (4.2) to provide an adaptive procedure. More precisely, a second estimator is defined as the first one, but the window of the second procedure, instead of being spherical, is now elliptical with minor axe which is equal to the first estimated direction. As a result, the second estimate uses informations provided by the first one to stretch the window in the interesting direction, i.e. the direction where g varies. This procedure is iterated until the convergence is achieved. They showed that the final estimator recovers the $\sqrt{n}$ consistency when the dimension of E_m does not exceed 3. Based on a similar idea, Dalalyan, Juditsky and Spokoiny in 2008 improved the algorithm by changing the way of extracting the basis of the index [27]. They demonstrated the $\sqrt{n}$ consistency even when $d = 4$.

There exists another way to infer about the derivatives of the regression. This one is based on the following formula which linked the average derivatives of g to a more simple covariance calculation. By an integration by part, we have

$$\beta = \mathbb{E}[\nabla g(X_1)] = \mathbb{E}[Y_1 l(X_1)], \tag{4.3}$$

with $l(X_1) = \nabla f(X_1)/f(X_1)$, provided regularity condition, in particular, f needs to vanish on its boundary. In the context of MIM, this formula was first employed by Stoker in 1986, but in the framework of a density that belongs to a parametric family [74]. The authors Härdle and Stoker in 1989 proposed the method *Average Derivative Estimation* (ADE) which estimates the index space when the density is unknown [53] (see also [71] for a similar approach). Their estimator is the empirical version of the right hand side term of (4.3), and is given by

$$\widehat{\beta} = n^{-1} \sum_{i=1}^{n} Y_i \widehat{l}(X_i) I_i \tag{4.4}$$

with $I_i = \mathbb{1}_{\{\widehat{f}(X_i) > b\}}$, $\widehat{l}(X_i) = \frac{\nabla \widehat{f}(X_i)}{\widehat{f}(X_i)}$ and

$$\widehat{f}(x) = (nh^p)^{-1} \sum_{j=1}^{n} K(h^{-1}(X_j - x)),$$

2. This approach is often refereed as *average derivative based method* but also as *outer product of gradient*.

$K : \mathbb{R}^p \to \mathbb{R}$ is a symmetric kernel. The terms I_i's are called the trimming terms and are required to provide the consistency of the estimator. For some configurations of the parameters b, h and K, under regularity conditions on f, essentially that its $r > p$ first derivatives exist and are bounded, and also moment conditions, the authors showed that $\sqrt{n}(\widehat{\beta} - \beta)$ converges to a Gaussian vector. The main point here is that we keep parametric rates while the function f is pointwise estimated with slower rates. Nevertheless ADE targets only one vector in E_m so that it only works in the case of the *single index model*, i.e. when $\dim(E_m) = 1$. Generalizing the later approach, Zeng and Zhu in 2010 proposed an integral method that estimates E_m with parametric rates whatever the dimension [84]. We will refer to gADE. They introduce the kernel $W : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ by considering the space

$$\operatorname{span}\left(\mathbb{E}[\nabla g(X_1) W(X_1, t)],\ t \in \mathbb{R}^p\right),$$

they show that if $W(x,t) = H(x-t)$, with $H > 0$ on a set with positive Lebesgue measure, then the previous space is equal to E_m. Regarding the estimation, in the same spirit as (4.3), one can note that

$$\beta_t = \mathbb{E}[\nabla g(X_1) W(X_1, t)] = -\mathbb{E}[Y_1 l_t(X_1)],$$

with $l_t(x) = \nabla_x W(x,t) + W(x,t) l(x)$, then following (4.4), the authors defined the estimator

$$\widehat{\beta}_t = n^{-1} \sum_{i=1}^{n} Y_i \widehat{l}_t(X_i) I_i,$$

with $\widehat{l}_t(x) = \nabla_x W(x,t) + W(x,t)\widehat{l}(x)$. As a corollary of Theorem 3.1 in [53], one could show that $\sqrt{n}(\widehat{\beta}_t - \beta_t)$ has a limit in law for each t. Nevertheless this does not provide results about the estimation of the space E_m. To get this, the autors make an eigendecomposition of the matrix

$$\int_{\mathbb{R}^p} \widehat{\beta}_t \widehat{\beta}_t^T \, \mathrm{d}t = n^{-2} \sum_{i,j}^{n} Y_i Y_j L(X_i, X_j) I_i I_j, \tag{4.5}$$

where $L : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}^{p \times p}$ can be derived straightforwardly by some calculus. The expression above is quite important since it permits to avoid proving weak convergence in functional spaces. Indeed, the underlying proof relies on U-statistic theory and more precisely on the asymptotic equivalence between a U-statistic and its projection (see [73]). Generalizing the proof of [84], the authors demonstrated that $M_{\widehat{\beta}}$ is asymptotically normal.

Compared to the adaptive approach describes latter ([56] and [27]), even if ADE or gADE requires more regularity condition on the density f, the root n consistency holds whatever the dimension d_0.

In this work, we study the behavior of the estimator

$$\widehat{\eta}_t = n^{-1} \sum_{i=1}^{n} \frac{Y_i \Psi(X_i, t)}{\widehat{f}(X_i)}, \tag{4.6}$$

where $\Psi : \mathbb{R}^{p+m} \to \mathbb{R}^p$ is a function with support $Q \times T$ such that Q is a bounded convex subset of $\mathbb{R}^p$ and $T = [0,1]^m$. Under some conditions, the probabilistic limit of $\widehat{\eta}_t$ is

$$\eta_t = \int g(x)\Psi(x,t)\mathrm{d}x = \int \nabla g(x)\psi(x,t)\mathrm{d}x,$$

with $\nabla_x \psi(x,t) = \Psi(x,t)$, and clearly, η_t lies in E_m for each t. Typically, the function ψ will be chosen to recover the whole space E_m when t varies, as a result, it is suitable to assume that t is living in a multi-dimensional space. For instance, a natural choice for ψ is the Gaussian kernel defined by $\psi(x,t) = \exp(-(x-t)^T \Sigma_t^{-1}(x-t))$.

We will assume that f is separated from 0 on Q. This assumption is needed to control the denominator $\widehat{f}(X_i)$ in (4.6) and in order to obtain some convergence properties. Theoretically, this is a simpler approach than the method called *trimming* used in [53] and [84]. This is made possible thanks to the flexibility in the choice of Ψ. Moreover, contrary to both previous references, we no longer need to estimate ∇f.

In the following, we study the asymptotic of the random element

$$\widehat{G}(t) = \sqrt{n}(\widehat{\eta}_t - \eta_t).$$

either for each t, or in the space $C(T)$ of continuous functions on T. In both cases, we will consider the decomposition

$$\widehat{G}(t) = \widehat{S}(t) + \widehat{R}(t), \tag{4.7}$$

where

$$\begin{aligned}
\widehat{R}(t) &= n^{-1/2} \left(\sum_{i=1}^{n} \frac{g(X_i)\Psi(X_i,t)}{\widehat{f}(X_i)} - \int g(x)\Psi(x,t)\mathrm{d}x \right) \\
\widehat{S}(t) &= n^{-1/2} \sum_{i=1}^{n} \frac{\sigma(X_i)\Psi(X_i,t)}{\widehat{f}(X_i)} e_i.
\end{aligned}$$

The sketch of the chapter is as follows. As a first step, we study the asymptotic behavior of $\widehat{R}(t)$ for each t. This is of particular interest because it deals with integral approximation. Considering (4.7), the tools we develop in the proof are different from the study in [53] or

[84]. They are related to the work of Vial in 2003 [78] and make full use of the model (4.1). Such considerations permit to describe a new procedure for integral estimation. This is addressed in section 4.2. In section 4.3, we provide the asymptotic normality of $\widehat{\eta}_t$ for each t. In section 4.4, following the gADE method, equation (4.5), we study the asymptotic of the transformation

$$\widehat{M} = \int_T \widehat{\eta}_t \widehat{\eta}_t^T \mathrm{d}t \qquad \text{whose limit is} \qquad M = \int_T \eta_t \eta_t^T \mathrm{d}t. \tag{4.8}$$

In particular we provide conditions such that $\mathrm{span}(M) = E_m$ and we obtain the root n consistency under mild assumptions. The convergence in the space $C(T)$ is studied in section 4.5. In the last two sections, we should use some results presented in two first sections.

4.2 Integral approximation by kernel smoothing

As we have already mentioned in the introduction, the index coefficient β can be estimated at parametric rates without the knowledge of the regression function f. This occurs for ADE type methods as [53], [71], [84], [56], [27] but also in others such as [59], [30]. It is somewhat intriguing since the classical nonparametric estimators of f are not pointwise root n consistent. As a result, the quantity $\widehat{\beta}_t$ converges quickly than each of the terms of its sum. The reason of this speed-up is that by averaging once again, we naturally obtain some U-statistics, which are known to converge at faster rates than $\sqrt{n}$. In this section we show that this can be used to speed up classical rates in integral approximation by monte-carlo procedure.

We are interested in the behavior of the random variable

$$n^{-1/2} \left(\sum_{i=1}^{n} \frac{\varphi(X_i)}{\widehat{f}(X_i)} - \int \varphi(x) \mathrm{d}x \right),$$

where $\varphi : \mathbb{R}^p \to \mathbb{R}$ is Hölder on its bounded convex support Q. Note that if we replace $\widehat{f}$ by the true density, then it converges in law to a Gaussian variable. Our point here is that under some mild assumptions on φ, this random variable goes to 0 in probability.

We introduce the following assumptions. The assumptions denoting A will be used in section 4.3, 4.4, and 4.5 for dimension reduction. The assumptions denoting B are related to integral approximation and are only used in this section.

(B1) The function φ is α-Hölder on $\mathbb{R}^p$ with compact support [3] Q, $0 < \alpha \leq 1$.

(A1) The variable X_1 has a density f on $\mathbb{R}^p$ such that its r-th first order derivatives are bounded.

3. This assumption can be weakened by considering φ as a Hölder function on a bounded convex support (see the remark stated after Lemma 4.B). Nevertheless, this induces slower rates of convergence and for this reason, we do not deal with such function in the following.

(A2) For every $x \in Q$, $f(x) \geq b > 0$.

(A3) The kernel K is bounded and symmetric. Moreover, we have $K(x_1, \cdots, x_p) = \Pi_{k=1}^{p} K_1(x_k)$ with

$$\int_{\mathbb{R}} K_1(x)\mathrm{d}x = 1, \quad \forall k = 1, \cdots, r-1 \int_{\mathbb{R}} |x|^k K_1(x)\mathrm{d}x = 0,$$
$$\int_{\mathbb{R}} |x|^r K_1(x)\mathrm{d}x < +\infty,$$

where r is called the order of the Kernel. Moreover [4] for every $x \in \mathbb{R}^p$, $K(x) \leq C_1 \exp(-C_2\|x\|)$ for some constants C_1 and C_2.

(A4) The window $(h)_{n\in\mathbb{N}}$ is a sequence going to 0 such that $n^{1/2}h^r \to 0$, $n^{1/2}h^p \to +\infty$.

The following Lemma has an important place in our dimension reduction approach. Indeed it will be useful to derive the convergence in law of $\widehat{G}(t)$, either for each t or in the space $C(T)$.

Lemma 4.1. *Assume that (A1-A4) and (B1) hold, we have*

$$\widehat{R} = n^{-1/2}\left(\sum_{i=1}^{n} \frac{\varphi(X_i)}{\widehat{f}(X_i)} - \int \varphi(x)\mathrm{d}x\right) = O_{\mathbb{P}}\left(h^{\alpha} + n^{1/2}h^r + n^{-1/2}h^{-p}\right).$$

Moreover, for $q \geq 2$ we have

$$\left\| \widehat{R} - n^{-1/2}\sum_{i=1}^{n} \frac{\varphi(X_i)(f(X_i) - \widehat{f}(X_i))^2}{f(X_i)^2\widehat{f}(X_i)} \right\|_q \leq C\|\varphi(X_1)\|_q[\varphi]\left(h^{\alpha} + n^{1/2}h^r + n^{-1/2}h^{-p}\right),$$

where $[\varphi]$ stands for the Hölder's constant and C depends only on f and K.

Proof. Let $\phi_1(X_i) = \frac{\varphi(X_i)}{f(X_i)}$, $\phi_2(X_i) = \frac{\varphi(X_i)}{f(X_i)^2}$ and for any function $f : \mathbb{R}^p \to \mathbb{R}$, define $f_h = f * h^{-p}K\left(h^{-1}(\cdot)\right)$. We first use a decomposition provided in [78], Section 7.2. We have

$$\frac{1}{\widehat{f}(X_i)} = \frac{2}{f(X_i)} - \frac{\widehat{f}(X_i)}{f(X_i)^2} + \frac{(f(X_i) - \widehat{f}(X_i))^2}{\widehat{f}(X_i)f(X_i)^2},$$

one can see that

$$\widehat{R} = \widehat{R}_1 - \widehat{R}_2 + \widehat{R}_3, \tag{4.9}$$

4. This additional condition is needed to have a control on the denominator of the estimator, more precisely to obtain that $(\inf_{x\in Q} \widehat{f}(x))^{-1} = O_{\mathbb{P}}(1)$. This condition is satisfied if the kernel has a compact support or if the kernel is Gaussian.

with

$$\widehat{R}_1 = n^{-1/2}\sum_{i=1}^{n}\{\phi_1(X_i) - \phi_{1,h}(X_i)\} + \{\phi_2(X_i)(f(X_i) - f_h(X_i))\} \\ + \mathbb{E}[\phi_2(X_i)(f_h(X_i) - f(X_i))]$$

$$\widehat{R}_2 = n^{-1/2}\sum_{i=1}^{n}\phi_2(X_i)\widehat{f}(X_i) - \phi_{1,h}(X_i) - \phi_2(X_i)f_h(X_i) + \mathbb{E}[\phi_2(X_i)f_h(X_i)]$$

$$\widehat{R}_3 = n^{-1/2}\sum_{i=1}^{n}\phi_2(X_i)\frac{(f(X_i) - \widehat{f}(X_i))^2}{\widehat{f}(X_i)}.$$

The term $\widehat{R}_1 - \widehat{R}_2$ is often called the linearization term (see for instance [53]). Let $q \geq 2$, we have

$$\left\|\widehat{R} - n^{-1/2}\sum_{i=1}^{n}\frac{\varphi(X_i)(f(X_i) - \widehat{f}(X_i))^2}{f(X_i)^2\widehat{f}(X_i)}\right\|_q = \|\widehat{R}_1 - \widehat{R}_2\|_q \leq \|\widehat{R}_1\|_q + \|\widehat{R}_2\|_q. \tag{4.10}$$

The rest of the proof consists in providing upper bounds for each term $\|\widehat{R}_l\|_q$, $l = 1, 2$. This should lead to the second statement of the proposition. Then by (4.9), what remains to be obtained is the rate of convergence to 0 in probability of $\widehat{R}_3$. The case $l = 1$ deals with classical regularization theorems, $l = 2$ needs results on the convergence of degenerate U-statistics and $l = 3$ uses classical tools of kernel estimation.

• Proof for $\|\widehat{R}_1\|_q \leq C\|\varphi(X)\|_q[\varphi](h^\alpha + n^{1/2}h^r)$: For the left-hand side term in $\widehat{R}_1$, we use Lemma 4.B to provide that its behavior is in $h^\alpha + n^{1/2}h^r$ (since φ is α-Hölder on Q, the function $\frac{\varphi}{f}$ is also α-Hölder on Q). For the middle term, defining $\delta_i = \phi_2(X_i)(f(X_i) - f_h(X_i))$, we have

$$\left\|\sum_{i=1}^{n}\delta_i\right\|_q \leq \left\|\sum_{i=1}^{n}\delta_i - \mathbb{E}[\delta_1]\right\|_q + n|\mathbb{E}[\delta_1]|,$$

and by the Rosenthal's inequality [5], we get

$$\left\|n^{-1/2}\sum_{i=1}^{n}\delta_i\right\|_q \leq C'\|\delta_1\|_q + n^{1/2}|\mathbb{E}[\delta_1]|.$$

Clearly, we have

$$|\mathbb{E}[\delta_1]| \leq \int_Q \frac{|\varphi(x)|}{f(x)}|f_h(x) - f(x)|\mathrm{d}x \leq Ch^r\int_Q |\varphi(x)|\mathrm{d}x,$$

5. For a martingale $(S_i, \mathcal{F}_i)_{i\in\mathbb{N}}$ and $2 \leq p < +\infty$, we have $\mathbb{E}[|S_n|^p] \leq C\{\mathbb{E}[(\sum_{i=1}^n \mathbb{E}[X_i^2|\mathcal{F}_{i-1}])^{p/2}] + \sum_{i=1}^n \mathbb{E}|X_i|^p\}$, where $X_i = S_i - S_{i-1}$ (see for instance [47], p. 23-24).

and similarly

$$\|\delta_1\|_q \leq Ch^r \left(\int_Q \frac{|\varphi(x)|^q}{f(x)^{2q-1}} \mathrm{d}x \right)^{1/q}.$$

The last term of $\widehat{R}_1$ is not random and has been treated two equations above.

• Proof for $\|\widehat{R}_2\|_q \leq C\|\varphi(X)\|_q \left(n^{-\frac{3q-4}{2q}} h^{-p} \right)$: We introduce the linear operator H defined by [6]

$$H(\Phi)(x_1, x_2) = \Phi(x_1, x_2) - \mathbb{E}[\Phi(X_1, X_2)|X_2 = x_2] - \mathbb{E}[\Phi(X_1, X_2)|X_1 = x_1] + \mathbb{E}[\Phi(X_1, X_2)],$$

and we define the function

$$\Phi_h(x_1, x_2) = \phi_2(x_1) h^{-p} K\left(\frac{x_1 - x_2}{h} \right).$$

By a quick calculation, we get

$$H(\Phi_h)(x_1, x_2) = \Phi_h(x_1, x_2) - \phi_{1,h}(x_1) - \phi_2(x_2) f_h(x_2) + \mathbb{E}[\phi_2(X_1) f_h(X_1)],$$

and then we have

$$\begin{aligned} \widehat{R}_2 &= n^{-3/2} \sum_{i,j}^{n} H(\Phi_h)(X_i, X_j) \\ &= n^{-3/2} \sum_{i \neq j}^{n} H(\Phi_h)(X_i, X_j) + n^{-3/2} \sum_{i=1}^{n} H(\Phi_h)(X_i, X_i) \\ &= \widehat{U} + \widehat{D}, \end{aligned} \tag{4.11}$$

with $\widehat{U} = n^{-3/2} \sum_{i \neq j} u_{ij}$ and $\widehat{D} = n^{-3/2} \sum_{i=1}^{n} d_i$. It is easy to verify that $\widehat{U}$ is a completely degenerate U-statistic, that is

$$\mathbb{E}[u_{ij}|X_i] = \mathbb{E}[u_{ij}|X_j] = 0 \quad \text{for } i \neq j.$$

Consequently, we can use proposition (2.4") p.11 of [46] (with a decoupling inequality, for instance the one stated in Theorem 3.1.1 of [29]) implies that

$$\mathbb{E}[\widehat{U}^q] \leq n^{-3q/2} Cn(n-1)\mathbb{E}[|u_{12}|^q],$$

where C does not depend on ϕ.Using Minkowski and Jensen inequalities, we get

$$\|u_{12}\|_q = \|H(\Phi_1)(X_1, X_2)\|_q \leq 4\|\Phi_h(X_1, X_2)\|_q. \tag{4.12}$$

6. Note that $H(\Psi)$ is a classical term in the Hoeffding decomposition used for example in [73] page 178.

Defining $v_K = \int_{\mathbb{R}^p} |K(u)|^q \mathrm{d}u$, which exists because of (3) we compute

$$\begin{aligned}\|\Phi_h(X_1, X_2)\|_q^q &= h^{-qp} \int |\phi_2(x)|^q |K(h^{-1}(x-y))|^q f(x) f(y) \mathrm{d}x \, \mathrm{d}y \\ &\leq h^{-p(q-1)} \int |\phi_2(x)|^q f(x) \int |K(u)|^q f(x - hu) \mathrm{d}u \mathrm{d}x \\ &\leq h^{-p(q-1)} v_K \|\phi_2(X_1)\|_q^q.\end{aligned}$$

Using (4.12), we have shown that

$$\|\widehat{U}\|_2 = O(n^{-3/2} n^{2/q} h^{-p(q-1)/q}) = O(n^{(-3q+4)/2q} h^{-p}). \tag{4.13}$$

It remains to obtain the rates for $\|\widehat{D}\|$. By Minkowski and Jensen inequality, we get [7]

$$\begin{aligned}\|\widehat{D}\|_q &\leq n^{-1/2} \|H(\Phi_h)(X_1, X_1)\|_q \\ &\leq n^{-1/2} (\|\phi_2(X_1) h^{-p}\|_q + 3 \|\Phi_h(X_1, X_2)\|_q) \\ &= n^{-1/2} h^{-p} \|\phi_2(X_1)\|_q + O(n^{-1/2} h^{-p(q-1)/q}),\end{aligned} \tag{4.14}$$

Finally, putting together (4.11), (4.13), (4.14), we obtain that

$$\|\widehat{R}_2\|_q \leq C \|\phi_2(X_1)\|_q \left(n^{-1/2} h^{-p} \right).$$

- Proof for $\widehat{R}_3 = O\left(n^{-1/2} h^{-p} + n^{1/2} h^{2r}\right)$: Firstly, we show that for n large enough,

$$\inf_{x \in Q} \widehat{f}(x) > b/2, \tag{4.15}$$

with high probability. We have, since $f > b$ on Q,

$$\mathbb{P}(\inf_{x \in Q} \widehat{f}(x) < b/2) = \mathbb{P}(\sup_{x \in Q} |\widehat{f}(x) - f(x)| > b/2),$$

where the last term is going to 0 by [31], Theorem 1. Now, since

$$|\widehat{R}_3| \leq (b^2 \inf_{x \in Q} \widehat{f}(x))^{-1} \; n^{-1/2} \sum_{i=1}^{n} |\varphi(X_i)| (f(X_i) - \widehat{f}(X_i))^2,$$

we can show the convergence in probability of the right-hand side term. We remark that $(X_i, \widehat{f}(X_i))$ is identically distributed and also that $\widehat{f}(X_1) = (nh^p)^{-1}(1 + \sum_{j \neq 1} K(h^{-1}(X_1 -$

7. This result clearly argues for the use of the leave-one-out estimate for the density, it seems to improve the rate of convergence because the diagonal term of the sum $\widehat{D}$ is no longer here.

$X_j)$), then defining $\widetilde{f}(x) = (nh^p)^{-1}(1 + \sum_{j\neq 1} K(h^{-1}(x - X_j))$, we get

$$\begin{aligned}\mathbb{E}[n^{-1/2}\sum_{i=1}^{n}|\varphi(X_i)|(f(X_i) - \widehat{f}(X_i))^2] &= n^{1/2}\mathbb{E}[|\varphi(X_1)|(f(X_1) - \widehat{f}(X_1))^2] \\ &= n^{1/2}\int_Q |\varphi(x)|\mathbb{E}[(f(x) - \widetilde{f}(x))^2]f(x)\mathrm{d}x \\ &\leq 2n^{1/2}\int_Q |\varphi(x)|(\mathbb{E}[(f(x) - \widehat{f}(x))^2] + \mathbb{E}[(\widetilde{f}(x) - \widehat{f}(x))^2])f(x)\mathrm{d}x \\ &\leq 2n^{1/2}\{C_1\left((nh^p)^{-1} + h^{2r}\right)\int_Q |\varphi(x)|f(x)\mathrm{d}x + C_2(nh^p)^{-2}\}, \end{aligned} \tag{4.16}$$

where the last inequality uses a classical result on the convergence of the MSE in kernel estimation provided that $\int_Q |\varphi(x)|f(x)\mathrm{d}x$ is finite. As a consequence we have shown that

$$\left(\frac{1}{n^{1/2}h^p} + n^{1/2}h^{2r}\right)^{-1}|\widehat{R}_3| \leq T_n Z_n, \tag{4.17}$$

with T_n bounded in L_2 and $\mathbb{P}(Z_n < \frac{2}{b^3}) \to 1$. □

4.3 Pointwise convergence

In the following proposition we obtain the convergence of the vector $\widehat{G}(t)$ for each t. We introduce some assumptions on the regression model.

(A5) The function g is Hölder.

(A6) The integral $\int_Q \sigma(x)^2 f(x)\mathrm{d}x$ is finite.

(A7) The function $x \mapsto \Psi(x,t)$ is Hölder on $\mathbb{R}^p$ with compact support Q.

Proposition 4.2. *Assume that (A1-A7) hold, we have*

$$n^{1/2}(\widehat{\eta}_t - \eta_t) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma_1).$$

where Σ_1 is the variance of the random vector $Z_1 = \frac{Y_1 - g(X_1)}{f(X_1)}\Psi(X_1, t)$.

Proof. In the proof, since t does not intervene, we put $\Psi_0(x) = \Psi(x,t)$ for every $x \in \mathbb{R}^p$. By decomposition (4.7), we are interested in the asymptotic law of the vector

$$n^{-1/2}\sum_{i=1}^{n}\frac{\sigma(X_i)\Psi_0(X_i)}{\widehat{f}(X_i)}e_i + \; n^{-1/2}\left(\sum_{i=1}^{n}\frac{g(X_i)\Psi_0(X_i)}{\widehat{f}(X_i)} - \int g(x)\Psi_0(x)\mathrm{d}x\right),$$

where by Lemma 4.1, the right hand-side term goes to 0 in probability. For the other term, we use the decomposition $\widehat{S}_1 + \widehat{S}_2$, with

$$\widehat{S}_1 = n^{-1/2} \sum_{i=1}^{n} \frac{s(X_i)}{f(X_i)} e_i \quad \text{and} \quad \widehat{S}_2 = n^{-1/2} \sum_{i=1}^{n} \frac{s(X_i)(f(X_i) - \widehat{f}(X_i))}{\widehat{f}(X_i) f(X_i)} e_i. \tag{4.18}$$

where $s(x) = \sigma(x)\Psi_0(x)$. We define $\mathcal{F}$ as the σ-field generated by the set of random variables $\{X_1, X_2, \cdots\}$. We get

$$\mathbb{E}[\widehat{S}_2^2|\mathcal{F}] = n^{-1} \sum_{i=1}^{n} \frac{s(X_i)^2 (f(X_i) - \widehat{f}(X_i))^2}{\widehat{f}(X_i)^2 f(X_i)^2},$$

then, one has

$$\mathbb{E}[\widehat{S}_2^2|\mathcal{F}] \leq (b^2 \inf_{x \in Q} \widehat{f}(x)^2)^{-1} n^{-1} \sum_{i=1}^{n} s(X_i)^2 (f(X_i) - \widehat{f}(X_i))^2.$$

For the term on the left, we have a similar result as (4.15), that is with high probability it is bounded by $b/2$. For the right hand-side term, one can follow (4.16) to obtain that

$$n^{-1} \sum_{i=1}^{n} s(X_i)^2 (f(X_i) - \widehat{f}(X_i))^2 \xrightarrow{\mathbb{P}} 0, \tag{4.19}$$

provided that $\int_Q \sigma(x)^2 f(x) \mathrm{d}x$ is finite. Therefore, we have shown that $\mathbb{E}[\widehat{S}_2^2|\mathcal{F}] \to 0$ in probability. Since for any $\epsilon > 0$, $\mathbb{P}(|\widehat{S}_2| > \epsilon|\mathcal{F}) \leq \epsilon^{-2}\mathbb{E}[\widehat{S}_2^2|\mathcal{F}]$, it remains to note that the sequence $\mathbb{P}(|\widehat{S}_2| > \epsilon|\mathcal{F})$ is uniformly integrable to apply the Lebesgue domination theorem to get

$$\mathbb{P}(\widehat{S}_2 > \epsilon) \longrightarrow 0.$$

To conclude, we apply the CLT to $\widehat{S}_1$ to obtain the statement. □

The previous Lemma provides vectors in the space E_m but this is not enough to estimate this subspace since we have not specified how to obtain an estimated basis of E_m. This is the topic of the following section.

4.4 Estimation of the index space

4.4.1 Asymptotic normality of the estimator $\widehat{M}$

In this section we obtain the asymptotic normality of $\widehat{M}$ defined in (4.8). We need the following assumption.

(A7') The function $\Psi : \mathbb{R}^{p+m} \to \mathbb{R}^p$ is continuous and $x \mapsto \Psi(x,t)$ is Hölder on $\mathbb{R}^p$ with bounded convex support Q_t, uniformly in t and x [8]. Moreover $Q = \cup_t Q_t$ is a compact set.

We also recall some definition

$$\widehat{M} = \int_T \widehat{\eta}_t \widehat{\eta}_t^T \mathrm{d}t, \qquad M = \int_T \eta_t \eta_t^T \mathrm{d}t.$$

Theorem 4.3. *Assume that (A1-A6) and (A7') hold, we have*

$$\sqrt{n}(\widehat{M} - M) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma_2),$$

where Σ_2 is the variance of the random variable $Z_2 + Z_2^T$ with

$$Z_2 = \frac{Y_1 - g(X_1)}{f(X_1)} \int_T \Psi(X_1, t) \eta_t^T \mathrm{d}t.$$

Proof. The proof is based on the following decomposition

$$n^{1/2}(\widehat{M} - M) = \widehat{M}_1 + \widehat{M}_1^T + \widehat{M}_2,$$

with

$$\widehat{M}_1 = \sqrt{n} \int_T (\widehat{\eta}_t - \eta_t) \eta_t^T \mathrm{d}t \qquad \text{and} \qquad \widehat{M}_2 = \sqrt{n} \int_T (\widehat{\eta}_t - \eta_t)(\widehat{\eta}_t - \eta_t)^T \mathrm{d}t.$$

One has

$$\widehat{M}_1 = n^{-1/2} \left(\sum_{i=1}^n \frac{Y_i \widetilde{\Psi}(X_i)}{\widehat{f}(X_i)} - \int g(x) \widetilde{\Psi}(x) \mathrm{d}x \right),$$

with $\widetilde{\Psi}(x) = \int_T \Psi(x,t) \eta_t^T \mathrm{d}t$. Now we can apply Proposition 4.2 with $\int_T \eta_t \eta_t^T \mathrm{d}t$ and $\widetilde{\Psi}$ in place of η_t and $\Psi(\cdot, t)$ to obtain that $\widehat{M}_1 + \widehat{M}_1^T$ converges in law to the distribution that is expressed in the statement. For that, we just have to notice that the support of $\widetilde{\Psi}$ is included in Q which is a compact, and also that for any $(x,y) \in Q^2$, one has

$$\|\widetilde{\Psi}(x) - \widetilde{\Psi}(y)\|_\infty \leq \int_T \|\eta_t\| \|\Psi(x,t) - \Psi(y,t)\| \mathrm{d}t \leq C \|x - y\|^\alpha,$$

where α stands for the Hölder's regularity of the function $x \mapsto \Psi(x,t)$. Now what remains to be shown is that $\widehat{M}_2$ goes to 0 in probability. One has

$$\|\widehat{M}_2\|_\infty \leq n^{-1/2} \int_T \|n^{1/2}(\widehat{\eta}_t - \eta_t)\|^2 \mathrm{d}t.$$

8. This assumption is required to get the second proposition in Lemma 4.1, but especially that the constant denoted by C does not depend on t. The next theorem is also true if $x \mapsto \Psi(x,t)$ is Hölder on its support but this involves some technicalities dealing with the supports Q_t in the bound of Lemma 4.B.

Following (4.7), we write $n^{1/2}(\widehat{\eta}_t - \eta_t) = \widehat{R}(t) + \widehat{S}(t)$ with

$$\widehat{R}(t) = n^{-1/2}\left(\sum_{i=1}^{n} \frac{\varphi_t(X_i)}{\widehat{f}(X_i)} - \int \varphi_t(x)\mathrm{d}x\right) \quad \text{and} \quad \widehat{S}(t) = n^{-1/2}\sum_{i=1}^{n} \frac{s_t(X_i)}{\widehat{f}(X_i)} e_i, \tag{4.20}$$

where $\varphi_t(x) = g(x)\Psi(x,t)$ and $s_t(x) = \sigma(x)\Psi(x,t)$. Then, using the same decomposition as in (4.9) and (4.18), one has

$$\begin{aligned}\|n^{1/2}(\widehat{\eta}_t - \eta_t)\| &\leq \|\widehat{R}(t) - \widehat{R}_3(t)\| + \|\widehat{R}_3(t)\| + \|\widehat{S}(t)\| \\ &\leq \|\widehat{R}(t) - \widehat{R}_3(t)\| + \|\widehat{S}_1(t)\| + \|\widehat{R}_3(t)\| + \|\widehat{S}_2(t)\|,\end{aligned}$$

with

$$\widehat{R}_3(t) = n^{-1/2}\sum_{i=1}^{n} \frac{\varphi_t(X_i)(f(X_i) - \widehat{f}(X_i))^2}{f(X_i)^2\widehat{f}(X_i)}, \qquad \widehat{S}_1(t) = n^{-1/2}\sum_{i=1}^{n} \frac{s_t(X_i)}{f(X_i)} e_i, \tag{4.21}$$

$$\widehat{S}_2(t) = n^{-1/2}\sum_{i=1}^{n} \frac{s_t(X_i)(f(X_i) - \widehat{f}(X_i))}{\widehat{f}(X_i)f(X_i)} e_i, \tag{4.22}$$

and clearly

$$\|\widehat{M}_2\|_\infty \leq n^{-1/2} C \int \|\widehat{R}(t) - \widehat{R}_3(t)\|^2 + \|\widehat{S}_1(t)\|^2 + \|\widehat{S}_2(t)\|^2 + \|\widehat{R}_3(t)\|^2 \mathrm{d}t.$$

In the following, we provide the convergence in probability of each term in the above equation. Firstly, by taking the expectation and applying the second assertion of Lemma 4.1 with $q = 2$, we have that the first one is going to 0 provided that $\int_T \|\varphi_t(X_1)\|_2 \mathrm{d}t$ is finite. For the second term, by a quick calculation, we have

$$\int_T \mathbb{E}[\|\widehat{S}_1(t)\|^2]\mathrm{d}t = \int_{Q\times T} \frac{\|s_t(x)\|^2}{f(x)} \mathrm{d}t\mathrm{d}x.$$

For the last remaining terms with $\widehat{R}_3(t)$ and $\widehat{S}_2(t)$, we follow the same approach as for the terms $\widehat{R}_3$ and $\widehat{S}_2$ in the proof of the Lemma 4.1 and Proposition 4.2. For the term with $\widehat{S}_2(t)$, let $\epsilon > 0$ and write

$$\begin{aligned}\mathbb{P}\Big(\int \|\widehat{S}_2(t)\|^2 \mathrm{d}t > \epsilon | \mathcal{F}\Big) &\leq \epsilon^{-1}\mathbb{E}\Big[\int \|\widehat{S}_2(t)\|^2 \mathrm{d}t | \mathcal{F}\Big] \\ &\leq \epsilon^{-1}(b^2 \inf_{x\in Q} \widehat{f}(x)^2)^{-1} n^{-1} \sum_{i=1}^{n} \int \|s_t(X_i)\|^2 \mathrm{d}t (f(X_i) - \widehat{f}(X_i))^2,\end{aligned}$$

which is going to 0 in probability by (4.19) provided that $\int_{Q\times T} \|\Psi(x,t)\|^2 \, \sigma(x)^2 f(x) \mathrm{d}x\mathrm{d}t$ is finite. Then one can apply the Lebesgue domination theorem to obtain that $\mathbb{P}(\int \|\widehat{S}_2(t)\|^2 \mathrm{d}t >$

$\epsilon) \to 0$. For the term with $\widehat{R}_3(t)$, using the Minkowski inequality, we have

$$\Big(\int_T \|\widehat{R}_3(t)\|^2 \mathrm{d}t \Big)^{1/2} \leq n^{-1/2} \sum_{i=1}^{n} \frac{\left(\int_T \|\varphi_t(X_i)\|^2 \mathrm{d}t\right)^{1/2} (\widehat{f}(X_i) - f(X_i))^2}{f(X_i)^2 \widehat{f}(X_i)}$$
$$\leq (b^2 \inf_{x \in Q} \widehat{f}(x))^{-1} n^{-1/2} \sum_{i=1}^{n} \left(\int_T \|\varphi_t(X_i)\|^2 \mathrm{d}t \right)^{1/2} (\widehat{f}(X_i) - f(X_i))^2.$$

By (4.17) we obtain the convergence in probability of the above term to 0 provided that the quantity $\int_Q \|\Psi(x,t)\|^2 |g(x)| f(x) \mathrm{d}x$ is finite.

□

4.4.2 Exhaustivity of the estimation of the space E_m

The previous section was dedicated to the asymptotics of the random matrix $\widehat{M}$. In particular we showed that it converges to M at the rate $\sqrt{n}$. A key issue is now to know whether the matrix M spanned E_m. This problem is raised in this section by first considering the vector $\eta_t = \mathbb{E}[Y_1 \nabla_x \psi(X_1, t)]$, with $\nabla_x \psi(x,t) = \Psi(x,t)$. By an integration by parts, under some regularity conditions, we have $\eta_t \in E_m$ for each t. This is the topic of the following lemma.

Lemma 4.4. *Assume that Q is a bounded convex set and ψ and g are continuously differentiable on Q with $\psi(x) = 0$ if $x \notin \mathring{Q}$, then*

$$\mathbb{E}\left[\frac{Y_1 \nabla \psi(X_1)}{f(X_1)}\right] = -\mathbb{E}\left[\frac{\nabla g(X_1) \psi(X_1)}{f(X_1)}\right].$$

Proof. We make the proof for the first coordinate since the others can be treated similarly. The difference between both sides of the statement is

$$\int_Q g(x) \partial_{x_1} \psi(x) + \partial_{x_1} g(x) \psi(x) \mathrm{d}x = \int_Q \partial_{x_1}(g\psi)(x) \mathrm{d}x.$$

By noting $x = (x_1, \widetilde{x})$ and using the Fubini's theorem we get

$$\int_Q \partial_{x_1}(g\psi)(x) \mathrm{d}x = \int_{\widetilde{Q}} \int_{Q(\widetilde{x})} \partial_{x_1}(g\psi)(x) \mathrm{d}x_1 \mathrm{d}\widetilde{x},$$

where $Q(\widetilde{x})$ is a bounded interval of the real line. Since $\partial_{x_1}(g\psi)$ is continuous and ψ vanishes on the boundary of Q, we obtain the statement. □

The previous Lemma gives us conditions for η_t to belongs to E_m. Now it is natural to wonder when does the family $\{\eta_t,\ t \in T\}$ generates the whole space E_m. Some conditions on the family $\{\psi(\cdot, t),\ t \in T\}$ are given in the following lemma where the matrix M is considered.

Lemma 4.5. *There exists a bounded convex set Q such that if each ψ_t meets the assumptions of Lemma 4.4 and if $\{\psi(\cdot,t),\ t\in T\}$ is a total family in the space of functions $\{\psi:\mathbb{R}^p\to\mathbb{R}\quad :\quad \int_Q\|\nabla g(x)\|\psi(x)\mathrm{d}x<+\infty\}$, we have*

$$\mathrm{span}(M)=E_m.$$

Proof. Firstly, we show that

$$\mathrm{span}(\nabla g(x), x\in\mathrm{supp}(X_1))=E_m.$$

The sense "$\subset$" is trivial. For the other inclusion, assume that $\beta_1\in E_m$ with $\beta_1^T\nabla g(x)=0$ for every $x\in\mathrm{supp}(X_1)$ and let us show that this is impossible. We have $g(x)=h(\beta^Tx)$ with $\beta=(\beta_1,\widetilde{\beta})$. Then for every $x\in\mathrm{supp}(X_1)$, we have that $\beta_1^T\nabla h(x)=0$ which implies that $\partial_{x_1}h(x)=0$. As a consequence $h(\beta^Tx)=h(0,\widetilde{\beta}^Tx)$ and $\mathrm{span}(\widetilde{\beta})=E_m$ which is impossible because of the definition of the central mean subspace.

As a consequence, there exist some points $(x_1,\cdots,x_H)$ such that

$$\mathrm{span}(\nabla g(x_1),\cdots,\nabla g(x_H))=E_m.$$

Let Q be a bounded convex subset containing the points $(x_1,\cdots,x_H)$. Now we can use the fact that the family ψ is total (see Theorem 1.3 and its proof) to show that

$$\mathrm{span}\left(\int_Q\nabla g(x)\psi_t(x)\mathrm{d}x,\ t\text{ varies in a finite subset}\right)=E_m.$$

This implies the statement.

□

In Chapter 1 Remark 1.4, several examples of families that are total in some L_p spaces are given. They include indicators, polynomials, and complex exponentials.

4.5 Convergence in the space $C(T)$

We have already studied the pointwise convergence of $\widehat{G}(t)=\sqrt{n}(\widehat{\eta}_t-\eta_t)$ and also an integral transform. In this section, we focus on the convergence of the process $\widehat{G}$ in the space of continuous functions. Such a convergence can be used to derive the consistency of the following Kolmogorov test. Given a vector γ, we test

$$H_0:\ \gamma\in E_m^{\perp}\qquad\text{against}\qquad\gamma\notin E_m^{\perp},$$

with the statistic

$$\sup_{t\in T}|\gamma^T\widehat{G}(t)|.$$

We introduce the following norm, if G is a stochastic process on T,

$$|||G|||_q = \left(\int_T \mathbb{E}[|G_t|^q] \mathrm{d}t \right)^{1/q},$$

and we will denote by $(t_1, \cdots, t_m)$ the coordinate of the variable t. We give other assumptions that are needed to obtain the convergence in the space $C(T)$.

(A6') The function σ is continuous.

(A7") The function $\Psi : \mathbb{R}^{p+m} \to \mathbb{R}^p$ is continuously differentiable on its compact support Q_t and $x \mapsto \partial_t \Psi(x,t)$ is Hölder uniformly on t and x. Moreover $Q = \cup_t Q_t$ is a compact set and the quantity $\int_{Q \times T} (\partial_{t_k} \partial_{t_l} \Psi(x,t))^q \mathrm{d}x\mathrm{d}t$ is finite for $k, l = 1, \cdots, m$.

Theorem 4.6. *Assume that (A1-A5), (A6') and (A7") hold, we have*

$$\widehat{G} \Rightarrow G,$$

where G is a Gaussian process with the same covariance function as the process

$$\frac{(Y_1 - g(X_1))^2}{f(X_1)^2} \Psi(X_1, t).$$

Proof. We use the same notation as the ones introduced in (4.20), (4.21) and (4.22). The proof is divided in two parts. Firstly, we use Lemma 4.1 to obtain that $\sup_T |\widehat{R}(t)| = o_{\mathbb{P}}(1)$. Secondly, we study the weak limit of the process $\widehat{S}$.

• Proof for $\sup_T |\widehat{R}(t)| = o_{\mathbb{P}}(1)$. Applying the Sobolev's inequality stated in Lemma 4.D, we have, for $q > m$,

$$\mathbb{E}[\sup_T |\widehat{R}(t) - \widehat{R}_3(t)|] \leq C \left(\left|\left|\left| \widehat{R}_1 - \widehat{R}_2 \right|\right|\right|_q + \sum_{k=1}^m \left|\left|\left| \partial_{t_k} (\widehat{R}_1 - \widehat{R}_2) \right|\right|\right|_q \right).$$

Then the rate of each term is obtained thanks to Lemma 4.1. For $\left|\left|\left| \widehat{R}_1 - \widehat{R}_2 \right|\right|\right|_q$, since Ψ is Lipschitz, we get

$$\int_T \|\widehat{R}_1(t) - \widehat{R}_2(t)\|_q^q \mathrm{d}t \leq C'(h + n^{1/2} h^r + n^{-1/2} h^{-p}).$$

The terms with the derivatives can be treated similarly since $x \mapsto \partial_t \Psi(x,t)$ is Hölder uniformly in t. Therefore, we have showed that $\sup_T |\widehat{R}(t) - \widehat{R}_3(t)| = o_{\mathbb{P}}(1)$. For the remaining term, we have that

$$\sup_T |\widehat{R}_3(t)| \leq (b^2 \inf_{x \in Q} \widehat{f}(x))^{-1} n^{-1/2} \sum_{i=1}^n \sup_{t \in T} \{|\phi_t(X_i)|\} (\widehat{f}(X_i) - f(X_i))^2,$$

and then we use (4.17) to obtain the desired convergence, provided that $\int_Q \sup_{t\in T}\{|\Psi(x,t)|\}|g(x)|f(x)\mathrm{d}x$ is finite.

• Proof for $\widehat{S} \Rightarrow G$. Firstly, we follow Proposition 4.2 to show that the process $\widehat{S}_1$ has the same limit as $\widehat{S} = \widehat{S}_1 + \widehat{S}_2$. Let $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{P}(\sup_T|\widehat{S}_2(t)| > \epsilon|\mathcal{F}) &\leq \epsilon^{-1}\mathbb{E}[\sup_T|\widehat{S}_2(t)|\ |\mathcal{F}] \\
&\leq \epsilon^{-1}\mathbb{E}\left[\left(\int_T|\widehat{S}_2(t)|^q\mathrm{d}t\right)^{1/q} + \sum_{k=1}^m\left(\int_T|\partial_{t_k}\widehat{S}_2(t)|^q\mathrm{d}t\right)^{1/q}\ |\mathcal{F}\right] \\
&\leq \epsilon^{-1}\left(\int_T\mathbb{E}\left[|\widehat{S}_2(t)|^q|\ \mathcal{F}\right]\mathrm{d}t\right)^{1/q} + \sum_{k=1}^m\left(\int_T\mathbb{E}\left[|\partial_{t_k}\widehat{S}_2(t)|^q|\ \mathcal{F}\right]\mathrm{d}t\right)^{1/q}.
\end{aligned}$$

where each lines are consequences of the Markov, Sobolev stated in Lemma 4.D, and Hölder inequalities, respectively. For the term on the left, one can notice that $\widehat{S}_2$ is a martingale with respect to the σ-field $\mathcal{F}_n$ generated by $\{e_1,\cdots,e_n,X_1,X_2,\cdots\}$. Defining $\mu_i = \varphi_t(X_i)(f(X_i)-\widehat{f}(X_i))(f(X_i)\widehat{f}(X_i))^{-1}$, and applying the Rosenthal's inequality conditionally to $\mathcal{F}$ (stated page 134 in a footnote), we obtain

$$\begin{aligned}
\mathbb{E}[|\widehat{S}_2(t)|^q\ |\mathcal{F}] &= n^{-q/2}\mathbb{E}[|\sum_{i=1}^n\mu_ie_i|^q|\mathcal{F}] \\
&\leq C\left(n^{-1}\sum_{i=1}^n\mu_i^2\right)^{q/2} + n^{-q/2}\mathbb{E}[|e_1|^q]\sum_{i=1}^n|\mu_i|^q.
\end{aligned}$$

Applying the Hölder's inequality to the term on the left, we obtain

$$\begin{aligned}
\mathbb{E}[|\widehat{S}_2(t)|^q|\mathcal{F}] &\leq C(1+\mathbb{E}[|e_1|^q])n^{-1}\sum_{i=1}^n|\mu_i|^q \\
&\leq C(1+\mathbb{E}[|e_1|^q])(b^q\inf_{x\in Q}\widehat{f}(x)^q)^{-1}n^{-1}\sum_{i=1}^n\varphi_t(X_i)^q(f(X_i)-\widehat{f}(X_i))^q.
\end{aligned}$$

Then following (4.16), we obtain the convergence in probability to 0 provided that the quantity $\int_{Q\times T}|\sigma(x)|^q|\Psi(x,t)|^qf(x)\mathrm{d}x\mathrm{d}t$. As a consequence $\mathbb{P}(\sup_T|\widehat{S}_2(t)| > \epsilon|\mathcal{F}) \to 0$, and we can apply the Lebesgue domination theorem to obtain the convergence to 0 of $\mathbb{P}(\sup_T|\widehat{S}_2(t)| > \epsilon)$. For the terms with the derivatives, we can follow the same path provided that the quantity $\int_{Q\times T}\sigma(x)^q(\partial_{t_k}\Psi(x,t))^qf(x)\mathrm{d}x\mathrm{d}t$ is finite.

The convergence of the finite dimensional laws of $\widehat{S}_1$ to a center Gaussian law with the covariance function claimed in the statement is a straightforward consequence of the CLT.

We conclude by showing the tightness of the process $\widehat{S}_1$. Since we are working in the space $(C(T),\|\cdot\|_\infty)$, by Theorem 7.3 in [8], the tightness is equivalent to the stochastic equicontinuity. For any $(u,v)\in T$, by the mean value Theorem, there exists $t\in T$ such that

$$|\widehat{S}_1(u)-\widehat{S}_1(v)| = (u-v)^T\nabla\widehat{S}_1(t),$$

then taking the supremum, we get

$$\sup_{|u-v|\leq\delta} |\widehat{S}_1(u) - \widehat{S}_1(v)| \leq \delta \sup_T |\nabla_t \widehat{S}_1(t)|,$$

which implies the equicontinuity provided that each $\sup_T |\partial_{t_k} \widehat{S}_1(t)|$ equals $O_{\mathbb{P}}(1)$ for $k = 1, \cdots, m$. To show that, we use the Sobolev's inequality of Lemma 4.D, to get

$$\mathbb{E}[\sup_T |\partial_{t_k} \widehat{S}_1(t)|] \leq C \left(\left\| \partial_{t_k} \widehat{S}_1 \right\|_q + \sum_{l=1}^{m} \left\| \partial_{t_l} \partial_{t_k} \widehat{S}_1 \right\|_q \right).$$

Then we have

$$\left\| \partial_{t_k} \widehat{S}_1 \right\|_q = \int_{Q\times T} \frac{\sigma(x)^q |(\partial_{t_k} \Psi(x,t))|^q}{f(x)^q} f(x) \mathrm{d}x\mathrm{d}t$$
$$\left\| \partial_{t_l} \partial_{t_k} \widehat{S}_1 \right\|_q = \int_{Q\times T} \frac{\sigma(x)^q |(\partial_{t_k} \partial_{t_l} \Psi(x,t))|^q}{f(x)^q} f(x) \mathrm{d}x\mathrm{d}t.$$

Since $(C(T), \|\cdot\|_\infty)$ is complete and separable, tightness implies the weak convergence (see [8]).

□

4.6 Implementation and simulation results

The asymptotic results we obtained previously deals with the matrix

$$\widehat{M} = \int_T \widehat{\eta}_t \widehat{\eta}_t^T \mathrm{d}t \quad \text{with} \quad \widehat{\eta}_t = n^{-1} \sum_{i=1}^{n} \frac{Y_i \Psi(X_i, t)}{\widehat{f}(X_i)},$$

which is difficult to compute in practice since we do not know the quantity $\int_T \Psi(X_i,t)\Psi(X_i,t)^T \mathrm{d}t$ for any choice of Ψ. A natural way to proceed is to use a Monte-Carlo type procedure, i.e. to approximate $\widehat{M}$ by the matrix $N^{-1} \sum_{i=1}^{N} \widehat{\eta}_{u_i} \widehat{\eta}_{u_i}^T$, where the sequence (u_i) is randomly drawn on the support of X_1. Simulations have shown that the larger is the approximation number N, the more accurate is the estimation. A convenient choice was $(u_i) = (X_i)$, probably because it permits to focus on the area where the points are. As a result, our method is as follows.

The method.

1. Compute $\widehat{M}_{\mathrm{SP}} = n^{-1} \sum_{i=1}^{n} \widehat{\eta}_{X_i} \widehat{\eta}_{X_i}^T$.

2. Eigen decomposition of $\widehat{M}_{\mathrm{SP}}$. The d eigenvectors associated to the d largest eigenvalues form the estimated basis of E_m.

4.6.1 Parameter setting

Choice of the kernel K. Theoretical results provided by Theorem 4.3 require that

$$\lim_{n\to+\infty} n^{1/2}h^r = 0 \quad \text{and} \quad \lim_{n\to+\infty} n^{1/2}h^p = +\infty.$$

If we put $h = Cn^{-a}$, we obtain the condition $\frac{1}{2r} < a < \frac{1}{2p}$, in particular, the order of the kernel r has to be larger than the dimension p. The balance between the variance term, in $n^{-1/2}h^{-p}$, and the bias term, in $n^{1/2}h^r$, gives the choice $Cn^{-\frac{1}{r+p}}$ for the window. In the whole study we put $h = 2n^{-\frac{1}{r+p}}$ with $r = p+1$.

To compute some high order kernels, we use a radial kernel given by

$$K(x) = \widetilde{K}(\|x\|), \tag{4.23}$$

where $\widetilde{K}$ is a polynomial function that satisfies equation (C.6) in Appendix C page 160. We also perform our method with other kernels such as product kernels of order r satisfying equation (C.5) page 159, and the Epanechnikov kernel. It does not seem to have much impact on the estimation. Besides, the use of a radial kernel provides a good heuristic for the adaptive method developed in the next section.

Choice of the function Ψ. As highlighted by the theoretical study, it is better if Ψ is Hölder on $\mathbb{R}^p$. As a result we put

$$\begin{aligned} \psi(x,t) = \widetilde{\psi}(h_0^{-1}\|x-t\|) \qquad &\text{with} \quad \widetilde{\psi}(z) = (1-z)^2(z+1)^2\mathbb{1}_{\{|z|<1\}}, \\ \text{and} \qquad &\Psi = \nabla_x \psi. \end{aligned} \tag{4.24}$$

Note that the parameter h_0 is not a traditional window that goes to 0 with n. It depends on the discrepancy of the sample to ensure that the points x_i such that $f(X_i) \simeq 0$ do not affect badly the estimation. Indeed, because of the denominator in $\widehat{\eta}_t$, such points can produce unusual large values of $\widehat{\eta}_t$. In the whole study, we set h_0 equal to the empirical estimator of $\mathbb{E}[\|X - \mathbb{E}[X]\|]$.

4.6.2 Simulation results

In the whole section, we consider regression models of the form $Y = g(\beta^T X) + e$ where e is normal. For each method that estimates $E_m = \text{span}(\beta)$ by $\widehat{E}_m$, we compute the error with

$$\text{dist}(E_m, \widehat{E}_m) = \|P - \widehat{P}\|_F,$$

where P and $\widehat{P}$ are the orthogonal projectors on the spaces E_m and $\widehat{E}_m$, respectively. In each case, we evaluate the mean of this error over 100 random samples.

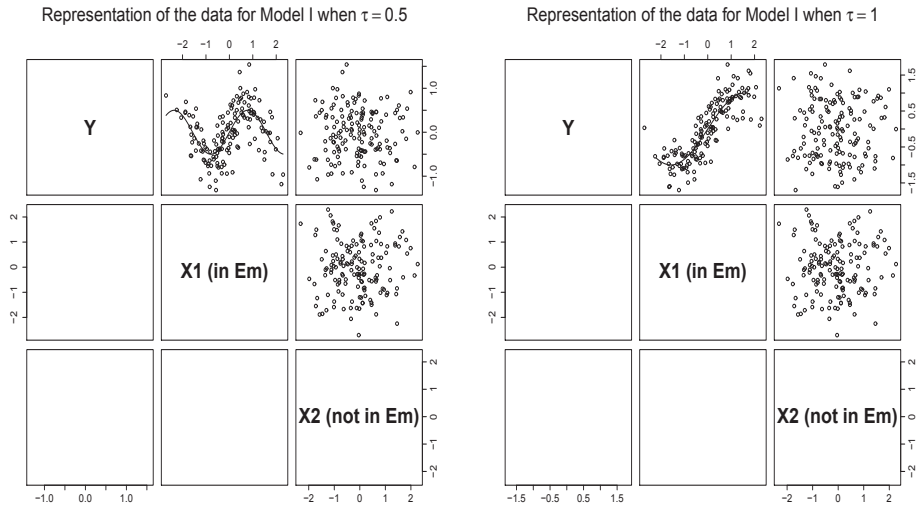

Figure 4.1 – Plot of the data for Model I when $\sigma = 0.2$.

4.6.2.1 Comparison with the inverse regression methods

In this section, we compare our method with the inverse regression methods *Sliced inverse regression* (SIR) [66] and *Sliced average variance estimation* (SAVE) [23]. As a result, the distribution of X will always verify the linearity condition and CCV. To highlight the behavior of our method with respect to SIR, we first focus on a family of regression models that reflects linear and nonlinear situations. We consider the model

$$\text{Model I}: \qquad Y = \tau \sin(X^{(1)}/\tau) + \sigma e,$$

where $X = (X^{(1)}, \ldots, X^{(6)}) \overset{\mathrm{d}}{=} \mathcal{N}(0, I)$, $e \overset{\mathrm{d}}{=} \mathcal{N}(0,1)$ and $\sigma \in \mathbb{R}$. Our first goal is to evaluate the effect of variations of the parameter τ. In Figure 4.1, we provide amongst others the plot of the response versus the good variable $X^{(1)}$. As a result, in Figure 4.2, we provide boxplots of the estimation error for τ varying between 0.5 and 1 and for two different values of σ, 0.2 and 0.4. In figure 4.3, we analyse the asymptotic behavior for $\tau = 0.6$ and $\sigma = 0.4$.

Figure 4.2 shows that the more there is nonlinearity in the model, the more the accuracy of our method exceeds the accuracy of the SIR method. This remains true at any noise level.

We also consider the following model

$$\text{Model II}: \qquad Y = \cos\left(\frac{\pi}{2}(X^{(1)} - \mu)\right) + 0.4e,$$

where $X = (X^{(1)}, \ldots, X^{(6)}) \overset{\mathrm{d}}{=} \mathcal{N}(0, I)$, $e \overset{\mathrm{d}}{=} \mathcal{N}(0,1)$ and $\mu \in \mathbb{R}$. The objective is to study the behavior of the methods facing a link function with some symmetry. When $\mu = 0$ the function is even so that SIR does not work (because of the so called SIR pathology, see in

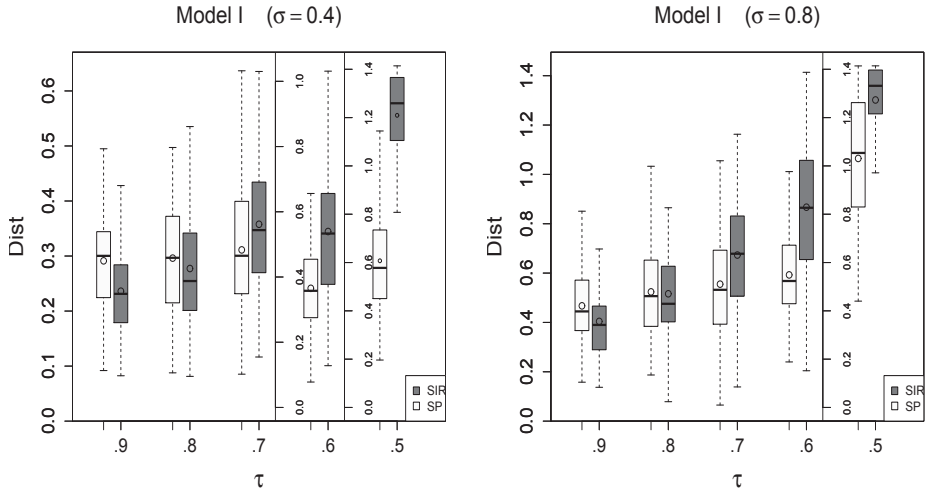

FIGURE 4.2 – Boxplot over 100 samples of the estimation error of SIR and our method in the case of Model I, when $n = 150$ and for different values of τ and σ.

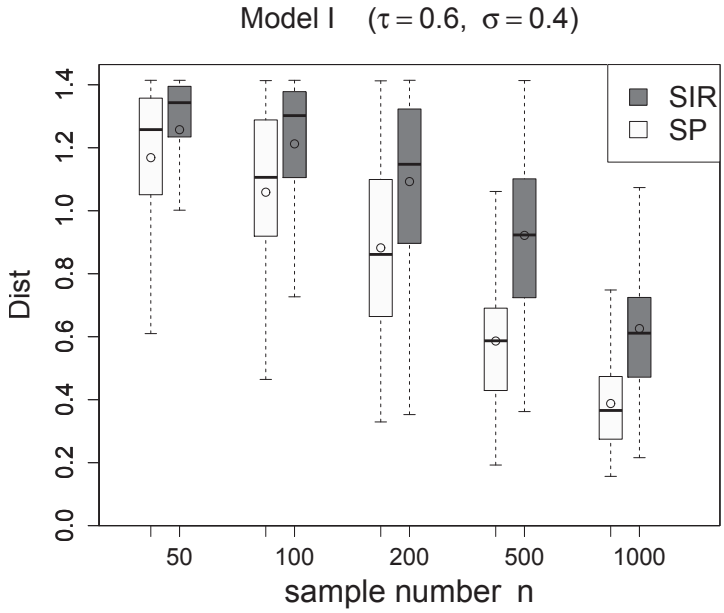

FIGURE 4.3 – Boxplot over 100 samples of the estimation error of SIR and our method in the case of Model I, when $\tau = 0.6$, $\sigma = 0.4$, and for different values of n.

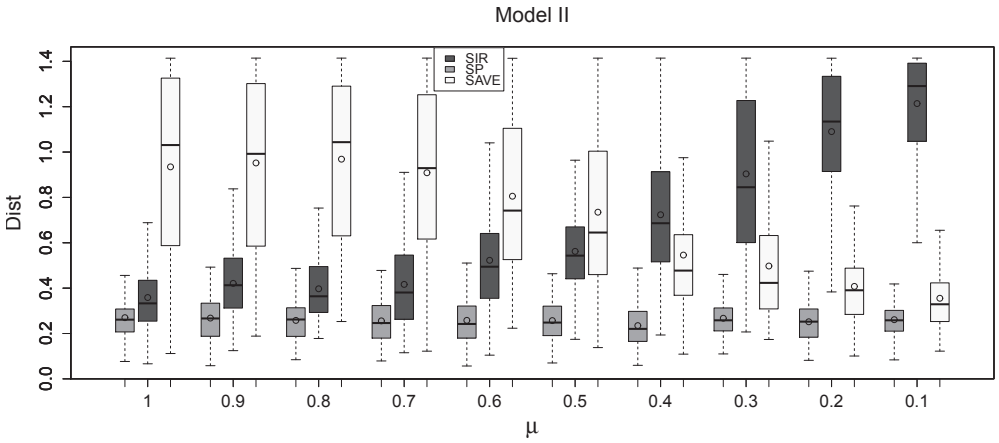

Figure 4.4 – Boxplot over 100 samples of the estimation error of SIR, SAVE and our method in the case of Model II, when $n = 150$ and for different values of μ.

Chapter 1 around equation 1.10, page 56), so that we consider some values of μ that are going to 0. We also include to our simulations the estimation given by the SAVE method, which can handle symmetric cases (among others). The boxplot are provided in Figure 4.4.

For every value of μ considered in Figure 4.4, our method has the best accuracy. Indeed, whereas SAVE and SIR seems to perform symmetrically with respect to the value of μ, our method remains stable. For $\mu = 1$, our method performs (in means) 4 times better than SAVE. For $\mu = .1$, our method performs (in means) 5 times better than SIR.

4.6.2.2 Adaptivity

In Figure 4.2, when the model is close to linear, our method is less accurate than SIR. To overcome this problem, we propose in the following an adaptive version of our method. Essentially we follow an idea proposed by Hristache, Juditsky and Spokoiny in [56].

Our method is based on the formula

$$\mathbb{E}\left[\frac{Y_1 \nabla\psi(X_1)}{f(X_1)}\right] = -\mathbb{E}\left[\frac{\nabla g(X_1)\psi(X_1)}{f(X_1)}\right] \in E_m,$$

demonstrated in Lemma 4.3, where f stands for the density of X_1. This formula can be extended to the following one

$$A\mathbb{E}\left[\frac{Y_1 \nabla\psi(AX_1)}{f_{|AX_1}(AX_1)}\right] = -\mathbb{E}\left[\frac{\nabla g(X_1)\psi(AX_1)}{f_{|AX_1}(AX_1)}\right] \in E_m, \tag{4.25}$$

where $f_{|AX_1}$ is the density of AX_1, provided that $E_m \subset \mathrm{span}(A)$. If the latter condition is not realized, this is not true anymore and the directions of E_m that are not in $\mathrm{span}(A)$ can not be reached by this quantity. The question now is how to choose the matrix A ? An

optimal choice is $A_0 = \beta\beta^T$ where β is a basis of E_m. The new variables become the $\beta^T X_i$'s. Since each of them lies in a subspace of dimension d, this ensures faster convergence rates of our estimators. This choice can not be possible because β is unknown, but following this spirit, we can conduct a plug-in iterative strategy.

The idea is simple : instead of analysing each direction in the same way in the construction of our estimator, one can zoom on directions that belong to or are close to the space E_m. In our implementation, characterized by a radial kernel and a radial function ψ, a natural way to proceed is to shrink the original window in the interesting directions, the one where g varies, and to stretch it in the directions where g does not vary much. Let us denote by $\widehat{\beta}$ the estimated basis of our method as it was originally defined, this is the first-step estimate of the adaptive method. We define

$$\widehat{A}_\epsilon = \widehat{\beta}\widehat{\beta}^T + \epsilon I,$$

By (4.25), we should estimate the vector

$$\eta_t(A_0) = A_0 \mathbb{E}\left[\frac{Y_1 \Psi(A_0 X_1)}{f_{|A_0 X_1}(A_0 X_1)}\right],$$

then, a second step estimate can be defined by

$$\widehat{\eta}_t(\widehat{A}_\epsilon) = n^{-1} \sum_{i=1}^{n} \frac{Y_i \Psi(\widehat{A}_\epsilon(X_i - t)))}{\widehat{f}_{|\widehat{A}_\epsilon X_1}(\widehat{A}_\epsilon X_i)},$$

with

$$\widehat{f}_{|\widehat{A}_\epsilon X_1}(x) = (nh^p)^{-1} \sum_{i=1}^{n} K(h^{-1}(\widehat{A}_\epsilon X_i - x)),$$

where K and Ψ are defined in (4.23) and (4.24). We can iterate this procedure. This leads to the following algorithm.

The adaptive method

1. Initialization. Put $l = 0$ and $\widehat{A}_\epsilon^{(0)} = I$.
2. Compute the estimator $\widehat{\beta}^{(l)}$ as the eigenvectors of the matrix

$$n^{-1} \sum_{i=1}^{n} \widehat{\eta}_{X_i}(\widehat{A}_\epsilon^{(l)}) \widehat{\eta}_{X_i}(\widehat{A}_\epsilon^{(l)})^T.$$

3. Put $\widehat{A}_\epsilon^{(l+1)} = \widehat{\beta}^{(l)}\widehat{\beta}^{(l)T} + \epsilon_l I$ and $l := l + 1$.
4. Repeat the last two steps until the convergence is achieved.

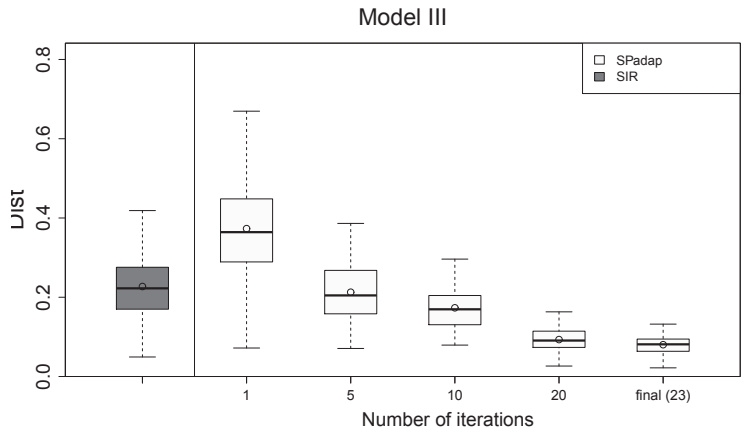

Figure 4.5 – Boxplot over 100 samples of the estimation error of SIR and our adaptive method in the case of Model III, when $n = 150$.

We follow the same implementation than the initial method. The only parameter it remains to implement is the sequence ϵ_l. This one is going to 0 slowly to ensure that we do not lost any direction that belongs to E_m. In the whole study, we compute

$$\epsilon_l = 0.9^l.$$

Following [56], it seems reasonable to stop the algorithm when $\epsilon_l = n^{-1/3}$.

To evaluate the behavior of the adaptive method, we consider a model where the SIR method beats our initial method. Note that the estimator of the initial method is the first-step estimator of the adaptive method. We compute this model

$$\text{Model III :} \qquad Y = X^{(1)} + 0.4e,$$

where $X = (X^{(1)}, \ldots, X^{(6)}) \overset{\mathrm{d}}{=} \mathcal{N}(0, I)$ and $e \overset{\mathrm{d}}{=} \mathcal{N}(0, 1)$. The boxplots of the errors are provided in Figure 4.5.

The plot in Figure 4.5 argues for the convergence of our algorithm. Whereas the first step estimator is a crude one, the last one is much accurate since it has divided by 2 the error of the SIR estimator.

Finally, we also compare our method to the semiparametric method developped in [56]. We consider the same model than their [9], that is

$$\text{Model IV :} \qquad Y = (\beta^T X)^2 \exp(\beta^T X) + 0.2e,$$

$e \overset{\mathrm{d}}{=} \mathcal{N}(0, 1)$, $\beta = (1, 2, 0, \ldots, 0)^T/\sqrt{5}$, $X = (X^{(1)}, \ldots, X^{(p)})$ where all the components are independently distributed with law $2(B(\tau, 1) - 1)$. The quantity τ controls the skewness

9. We also measure the estimation error with the same distance than their, i.e. the $l1$-norm.

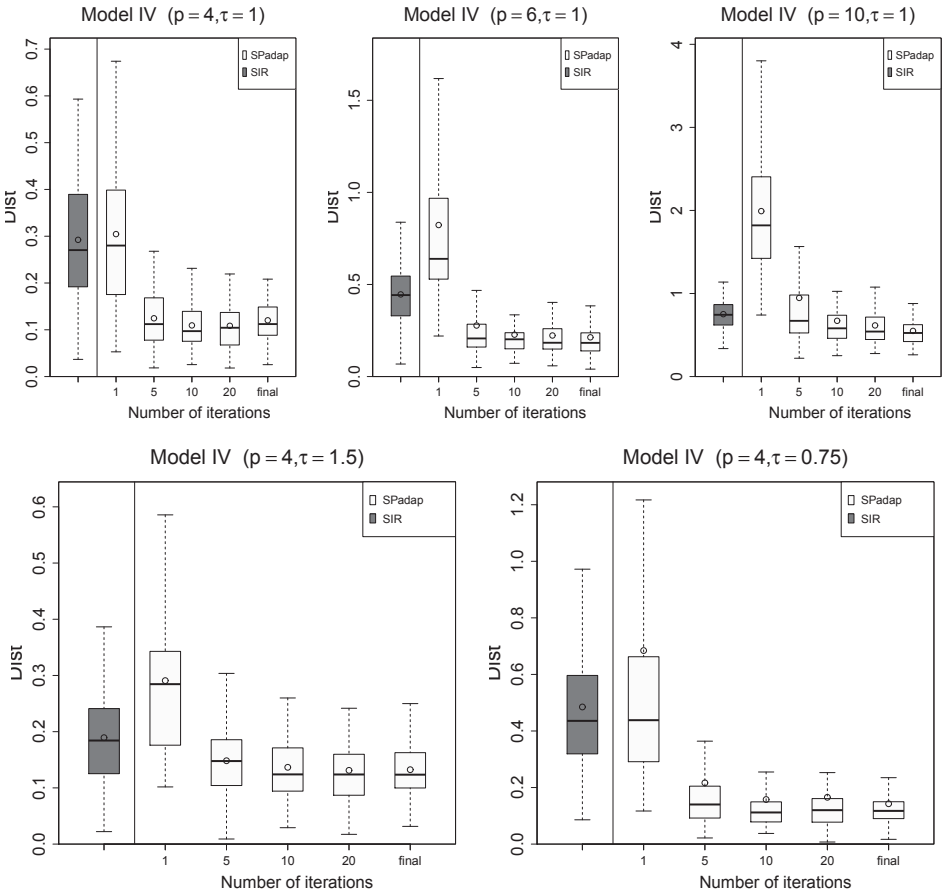

FIGURE 4.6 – Boxplots over 100 samples of the estimation error of SIR and our adaptive method in the case of Model IV, when $n = 200$ and for different values of p and τ.

of the beta distribution $B(\tau, 1)$ (when $\tau = 1$ it corresponds to the uniform distribution). The sample number is fixed to $n = 200$ but we consider different values of the parameters p and τ. We give some boxplot in Figure 4.6.

In the case of Model IV in every situation considered, our first-step estimator is less accurate than the SIR whereas the last-step estimator performs better than SIR. Moreover, the final-step estimator is always at least 2 times more accurate than the first-step estimator. Nevertheless in the case of model IV, the results obtained here are not as good than the one exposed in [56]. Work to understand this underperformance is under progress.

4.7 Further research

The simulation results of the adaptive method are quite impressive since in beats SIR even in some advantageous situation for SIR (characterised in particular by Gaussian

predictors and linear models). Nevertheless, the theoretical results of the adaptive method are still missing. For instance one can ask if the asymptotic normality still holds. If this is true what are the condition on the window ? And what becomes the dominated term in the asymptotic expansion of Lemma 4.1 ?

4.8 Some lemmas

Lemma 4.A. *Under the assumptions (A-1) and (A-2), we have*

$$\forall x \in \mathbb{R}^p \qquad |f_h(x) - f(x)| \leq h^r C,$$

where C does not depend on x.

Lemma 4.B. *Let $q \geq 2$. Under the Assumptions (A-1) and (A-2), if φ has a compact support such that*

$$\int |\varphi(x+u) - \varphi(x)|^q \mathrm{d}x \leq C|u|^{q\alpha}, \tag{4.26}$$

we have

$$\left\| n^{-1/2} \sum_{i=1}^{n} \varphi(X_i) - \varphi_h(X_i) \right\|_q \leq C'(h^\alpha + n^{1/2} h^r).$$

Remark 4.C. *One can notice that the condition (4.26) holds if φ is Hölder on $\mathbb{R}^p$. In this case $C' = C''\|\varphi(X)\|_1[\varphi]$ where C'' depends only on K an f. Moreover, provided that the support of φ is a nonempty bounded convex set and φ is Hölder inside its support, it still holds but with different rates of convergence. We have*

$$\begin{aligned}\int |\varphi(x+u) - \varphi(x)|^q 1_{\{x \notin Q\}} \mathrm{d}x &\leq \|\varphi\|_\infty \int \mathbb{1}_{\{x+u \in Q\}} \mathbb{1}_{\{x \notin Q\}} \mathrm{d}x \\ &\leq \|\varphi\|_\infty \lambda(y \ : \ \mathrm{dist}(y, \partial Q) < |x-u|),\end{aligned}$$

and one can use Steiner's formula stated for instance in [41], Theorem 3.2.35 page 271, to conclude.

Proof. Define the i.i.d. variables

$$\Delta_i = \varphi(X_i) - \varphi_h(X_i),$$

and note that

$$\left\| \sum_{i=1}^{n} \Delta_i \right\|_q \leq \left\| \sum_{i=1}^{n} \Delta_i - \mathbb{E}[\Delta_i] \right\|_q + n|\mathbb{E}[\Delta_1]|.$$

For the first term, because the process $(\sum_{i=1}^n \Delta_i - \mathbb{E}[\Delta_i])_{n\in\mathbb{N}}$ is a martingale with respect to the σ-field generated by $\{\Delta_i,\ i=1,\cdots,n\}$, we use the the Rosenthal's inequality (stated in a footnote page 134) to get

$$\|n^{-1/2}\sum_{i=1}^n \Delta_i - \mathbb{E}[\Delta_i]\|_q^q \leq C\{\|\Delta_1 - \mathbb{E}[\Delta_1]\|_{L^2}^q + \|\Delta_1 - \mathbb{E}[\Delta_1]\|_q^q\} \leq C\|\Delta_1\|_q^q,$$

where the last inequality is a consequence of Hölder's inequality. Then, by the assumption on φ, we get

$$\begin{aligned}\mathbb{E}|\Delta_1|^q &= \int\left(\int(\varphi(x)-\varphi(x+hu))\,K(u)\mathrm{d}u\right)^q f(x)\mathrm{d}x \\ &\leq \iint(\varphi(x)-\varphi(x+hu))^q\,\mathrm{d}xK(u)\mathrm{d}u \leq Ch^{q\alpha}\int|u|^{q\alpha}K(u)\mathrm{d}u.\end{aligned}$$

For the last term, we have

$$\begin{aligned}\mathbb{E}[\Delta_1] &= \int(\varphi(x)-\varphi_h(x))\,f(x)\mathrm{d}x \\ &= \int\varphi(x)f(x) - \varphi(x)f_h(x)\mathrm{d}x \\ &= \int\varphi(x)(f(x)-f_h(x))\mathrm{d}x,\end{aligned}$$

and we obtain, by Lemma 4.A, that $|\mathbb{E}[\Delta_1]| \leq h^r C$. □

Lemma 4.D. *Let $q > m$. Assume that $\varphi : \mathbb{R}^m \to \mathbb{R}$ is differentiable, we have*

$$\|\varphi\|_\infty \leq C\left(\left(\int|\varphi(z)|^q\mathrm{d}z\right)^{1/q} + \sum_{k=1}^m\left(\int|\partial_k\varphi(z)|^q\mathrm{d}z\right)^{1/q}\right),$$

where C only depends on q.

Proof. See for instance [10], p.167, where the previous inequality is the main argument for proving Morrey's Theorem. □

Appendices

A Linearity condition, elliptical distribution, central subspace

This section is dedicated to the link between the linearity condition and the assumption of sphericity. We assume in the following that E_c exists, Σ is non-deficient and $\beta \in \mathbb{R}^{p\times d}$ has full rank and span E_c, $\eta = \Sigma^{1/2}\beta$ and P_c is the orthogonal projection on $\Sigma^{1/2}E_c$. In the whole thesis, for the sake of clarity, we work with the standardized variables. As a result, we express LC stated originally in [66] in terms of standardized variables. The original statement is the weakest which can be ensconced to express the linearity condition. Actually it involves some specification and its original version is not really usable.

Proposition A.1. *The following points are equivalent :*

(i) *For any $u \in \mathbb{R}^p$, $\mathbb{E}[u^T X|\beta^T X]$ is linear.*

(ii) $\mathbb{E}[Z|\eta^T Z] = P_c Z$

Proof. We write (i) the following way : there exists $b \in \mathbb{R}^p$ and $B \in \mathbb{R}^{p\times d}$ such that

$$\mathbb{E}[X|\beta^T X] = b + B\beta^T X.$$

Now it is easy to see that (ii) implies (i). For the other sense, we write (i) as follows

$$\mathbb{E}[X - \mathbb{E}[X]|\beta^T X] = B\beta^T(X - \mathbb{E}[X]),$$

then one can note that the matrix B has full rank by expressing it in the basis (β, γ). We calculate

$$\mathrm{var}(\mathbb{E}[X|\beta^T X]) = \mathbb{E}[\mathbb{E}[X - \mathbb{E}[X]|\beta^T X](X - \mathbb{E}[X])^T] = B\beta^T\Sigma, \tag{A.1}$$

this gives that the random variable $\mathbb{E}[X - \mathbb{E}[X]|\beta^T X]$ is degenerate in the directions $\Sigma^{-1}\gamma$ with $\gamma \in E_c^{\perp}$. Then since $\mathrm{var}(\mathbb{E}[X|\beta^T X])$ is symmetric, we have $B = \Sigma\beta A$ with $A \in \mathbb{R}^{d\times d}$ symmetric and full rank. We also have

$$\mathrm{var}(\mathbb{E}[X|\beta^T X]) = B\beta^T\Sigma\beta B^T,$$

putting together with equation (A.1) gives $\Sigma\beta A\beta^T\Sigma\beta A\beta^T\Sigma = \Sigma\beta A\beta^T\Sigma$, therefore $A = (\beta^T\Sigma\beta)^{-1}$. This leads to another formulation of LC

$$\mathbb{E}[X|\beta^T X] = \mathbb{E}[X] + P_\Sigma(X - \mathbb{E}[X])$$

where $P_\Sigma = \Sigma\beta(\beta^T\Sigma\beta)^{-1}\beta^T$ is the projector on the space span$(\Sigma\beta)$ with null space $E_c^\perp$. Turning the previous relationship in terms of standardized variables gives (ii). □

Spherical and elliptical variables have been introduced to gives a natural extension to Gaussian variable. They can be defined as follows.

Definition A.2. *The random variable $Z \in \mathbb{R}^p$ is said to be spherical if $Z \overset{\mathrm{d}}{=} HZ$ for any orthonormal matrix H. In particular spherical laws are standardized. In addition if $A \in \mathbb{R}^{p\times p}$ has full rank and $\mu \in \mathbb{R}^p$ then $X = AZ + \mu$ is said to be elliptical with with mean μ and variance AA^T.*

The following proposition is useful to generate elliptical variables.

Proposition A.3. (Johnson (1987) [60], Chapter 7) *The radom variable $Z \in \mathbb{R}^p$ is spherical if and only if $\rho = \|Z\|$ et $U = \frac{Z}{\|Z\|}$ are independent, and U is uniformly distributed on the unit sphere of $\mathbb{R}^p$.*

We now provide a family of random variables that checks LC.

Lemma A.4. *If $Z \overset{\mathrm{d}}{=} (2P_c - I)Z$, then LC holds.*

Proof. For every f measurable we have

$$\begin{aligned}\mathbb{E}[Zf(P_cZ)] &= \mathbb{E}[(2P_c - I)Z\ f(P_c(2P_c - I)Z)] \\ &= 2\mathbb{E}[P_cZf(P_cZ)] - \mathbb{E}[Zf(P_cZ)]\end{aligned}$$

equivalently $\mathbb{E}[Zf(P_cZ)] = \mathbb{E}[P_cZf(P_cZ)]$. □

Proposition A.5. *Spherical distributions verify LC.*

Proof. Since $2P_c - I$ is orthonormal, we can apply Lemma A.5. □

Eaton in 1986 has provided the following result : if $\mathbb{E}[Z|PZ] = PZ$ holds for every P orthogonal projection with rank 1, then Z has a spherical distribution [34]. Putting this with the previous proposition gives the following characterization of the spherical distributions.

Proposition A.6. *Z is spherical if and only if $\mathbb{E}[Z|PZ] = PZ$ for every orthogonal projector P.*

This highlight how strong is the assumption of sphericity with respect to LC. Indeed, LC only ask the previous assertion to be true for P_c. Nevertheless since we do not know P_c we are often forced to assumed sphericity.

Now we provide a simple proof of Theorem 7. We recall that by proposition (1), model (3) is equivalent to

$$Y \perp\!\!\!\perp Z | P_c Z. \tag{A.2}$$

Theorem A.7. *Assume that (3) and LC hold, $\mathbb{E}[X] < \infty$, then we have*

$$\mathbb{E}[Z|Y] \in \Sigma^{1/2} E_c.$$

Proof. Using first (A.2) and LC we have

$$\mathbb{E}[Z|Y] = \mathbb{E}[\mathbb{E}[Z|Y, P_c Z]|Y] = \mathbb{E}[\mathbb{E}[Z|P_c Z]|Y] = P_c \mathbb{E}[Z|Y].$$

□

B Asymptotic of the SIR based methods

Let us define

$$\widehat{B} = n^{-1} \sum_{i=1}^{n} (X_i - \mathbb{E}[X])(\mathbb{1}_i - \mathbb{E}[\mathbb{1}]) \qquad \text{and} \qquad B = \mathbb{E}[\widehat{B}]$$

with $\mathbb{1}_i = (\mathbb{1}_{\{Y_i \in I(1)\}}, \cdots, \mathbb{1}_{\{Y_i \in I(H)\}})$ and the $I(h)$'s formed a partition of the range of Y (it has been introduced page 19). We recall that

$$\widehat{D}_p = \operatorname{diag}(\overline{\mathbb{1}})^{-1} \qquad \text{and} \qquad D_p = \operatorname{diag}(\mathbb{E}[\mathbb{1}_i])]^{-1},$$

and

$$\widehat{C} = (\widehat{C}_1, \cdots, \widehat{C}_H), \qquad \text{and} \qquad \widehat{C}_h = n^{-1} \sum_{i=1}^{n} \widehat{Z}_i \mathbb{1}_{\{Y_i \in I(h)\}}.$$

In the following proposition, we provide an asymptotic decomposition of the matrix $\widehat{M}_{\text{SIR}} = \widehat{C}\widehat{D}_p\widehat{C}^T$.

Theorem B.8. *Assume that $\mathbb{E}[\|X\|^4] < +\infty$, we have Assume that $\mathbb{E}[\|X\|^4] < +\infty$, we have*

$$n^{1/2}(\widehat{C}\widehat{D}_p^{1/2} - CD_p^{1/2}) \xrightarrow{\text{d}} W,$$

where $\operatorname{vec}(W)$ *is the Gaussian limit of the vector*

$$(D_p^{1/2} B^T \otimes I, D_p^{1/2} \otimes \Sigma^{-1/2}, I \otimes \Sigma^{-1/2} B) \operatorname{vec}(\widehat{G}_1, \widehat{G}_2, \widehat{G}_3),$$

with $\widehat{G}_1 = n^{1/2}(\widehat{\Sigma}^{-1/2} - \Sigma^{-1/2})$, $\widehat{G}_2 = n^{1/2}(\widehat{B} - B)$, $\widehat{G}_3 = n^{1/2}(\widehat{D}_p^{1/2} - D_p^{1/2})$.

Proof. Using Slutsky's theorem, we have

$$\begin{aligned}
&n^{1/2}(\widehat{C}\widehat{D}_p^{1/2} - CD_p^{1/2}) \\
&= n^{1/2}(\widehat{\Sigma}^{-1/2} n^{-1} \sum_{i=1}^{n} (X_i - \overline{X})(\mathbb{1}_i - \overline{\mathbb{1}})\widehat{D}_p^{1/2} - \Sigma^{-1/2} B D_p^{1/2}) \\
&\underset{n\to\infty}{\overset{\mathrm{d}}{=}} n^{1/2}(\widehat{\Sigma}^{-1/2}\widehat{B}\widehat{D}_p^{1/2} - \Sigma^{-1/2} B D_p^{1/2}) \\
&\underset{n\to\infty}{\overset{\mathrm{d}}{=}} n^{1/2}((\widehat{\Sigma}^{-1/2} - \Sigma^{-1/2}) B D_p^{1/2} \\
&\qquad\qquad + \Sigma^{-1/2}(\widehat{B} - B) D_p^{1/2} + \Sigma^{-1/2} B (\widehat{D}_p^{1/2} - D_p^{1/2})).
\end{aligned}$$

Taking the vec$(\cdot)$ operator, we get that

$$\begin{aligned}
&n^{1/2} \operatorname{vec}(\widehat{C}\widehat{D}_p^{1/2} - CD_p^{1/2}) \\
&\qquad \underset{n\to\infty}{\overset{\mathrm{d}}{=}} (D_p^{1/2} B^T \otimes I, D_p^{1/2} \otimes \Sigma^{-1/2}, I \otimes \Sigma^{-1/2} B) \operatorname{vec}(\widehat{G}_1, \widehat{G}_2, \widehat{G}_3),
\end{aligned}$$

and we can use the Delta method to provide that the asymptotic law is gaussian. □

Corollary B.9. *Assume that* $\mathbb{E}[\|X\|^4] < +\infty$*, we have*

$$n^{1/2}(\widehat{P}_{SIR} - P_{SIR}) \xrightarrow{\mathrm{d}} QWC_{SIR}^T M_{SIR}^+ + M_{SIR}^+ C_{SIR} W^T Q,$$

where vec(W) *is defined in Theorem B.8.*

Proof. We first note that

$$n^{1/2}(\widehat{P}_{SIR} - P_{SIR}) = -n^{1/2}((I - \widehat{P}_{SIR}) - (I - P_{SIR})),$$

and then we apply Theorem B.10 stated bellow, noting that $(I - P_{\mathrm{SIR}})C_{\mathrm{SIR}} = 0$. □

Theorem B.10. (Tyler (1981) [75]) *Assume that* $\widehat{M}$*,* M *are symmetric matrices such that* $n^{1/2}(\widehat{M} - M) \xrightarrow{\mathrm{d}} W$ *and* $\dim(\ker(M)) = d$*, then we have*

$$n^{1/2}(\widehat{Q} - Q) \xrightarrow{\mathrm{d}} -QWM^+ - M^+WQ,$$

where $\widehat{Q}$ *(resp.* Q*) is the orthogonal projector on the sum of the eigenspaces of* $\widehat{M}$ *(resp.* M*) associated to the* d *smallest eigenvalues.*

C High order kernels

We will use the following notation. If $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ and $(m) = (m_1, \ldots, m_p)$ is a sequence of p integers, we define

$$x^{(m)} = x_1^{m_1} \times \cdots \times x_p^{m_p},$$

and

$$|m| = \sum_{k=1}^{p} m_k.$$

A kernel of order r is a function $K : \mathbb{R}^p \to \mathbb{R}$ such that

$$\int K(x)\mathrm{d}x = 1, \tag{C.3}$$

and, for any sequence (m) of p integers such that $1 \leq |m| \leq r-1$, we have [1]

$$\int x^{(m)} K(x)\mathrm{d}x = 0. \tag{C.4}$$

There exists two common ways for building a multi-dimensional kernel using a one dimensional kernel.

Product Kernel. If $\widetilde{K} : \mathbb{R} \to \mathbb{R}$ is of order r, then by Fubini's theorem, the kernel defined by

$$K(x) = \prod_{k=1}^{p} \widetilde{K}(x_k),$$

is a kernel of order r. It is easy to construct one dimensional kernel $\mathbb{R} \to \mathbb{R}$ with given order [2] using for instance polynomial functions with compact support. Since the kernels used in nonparametric estimation often need to be continuous, we add this conditions in the following construction. Assume that

$$\widetilde{K}(x) = \begin{cases} \sum_{l=0}^{r+1} \alpha_l x^l & \text{if } |x| < 1 \\ 0 & \text{elsewhere,} \end{cases}$$

then, one can obtain the weights α_i's by solving the system

$$\begin{cases} \sum_{l=0}^{r+1} \frac{\mathrm{odd}(l+1)\alpha_l}{l+1} = 1 \\ \sum_{l=0}^{r+1} \frac{\mathrm{odd}(k+l+1)\alpha_l}{k+l+1} = 0 \quad \text{for every } 1 \leq k \leq r-1 \\ \sum_{l=0}^{r+1} \alpha_l = 0 \\ \sum_{l=0}^{r+1} (-1)^l \alpha_l = 0, \end{cases} \tag{C.5}$$

where $\mathrm{odd}(k)$ gives 1 if k is odd and 0 elsewhere. The first two lines in the previous system correspond to (C.3) and (C.4), the last two lines ensure the continuity of the kernel [3].

Radial kernel. A radial kernel is such that

$$K(x) = \widetilde{K}(\|x\|),$$

1. As a result, we can develop the bias of the kernel estimator of the density in Taylor series to provide that the bias converges with the rates h^r.
2. For one dimensional kernel, (C.4) becomes $\int x^k K(x)\mathrm{d}x = 0$ for every $1 \leq k \leq r-1$.
3. Obviously, by increasing the degree of the polynomial, one can ask for more smoothness conditions.

with $\widetilde{K} : \mathbb{R}^+ \to \mathbb{R}$. By the usual spherical variable change, we have

$$\int x^{(m)} \widetilde{K}(x) \mathrm{d}x = \int \rho^{|m|+p-1} \widetilde{K}(\rho) \mathrm{d}\rho \int u^{(m)} \mathrm{d}\sigma(u),$$

where σ is the uniform distribution on the unit sphere. Clearly, the latter quantity equals 0 when $|m|$ is odd. Then the condition for a radial kernel to be of order r is that

$$\begin{cases} \int \rho^{p-1} \widetilde{K}(\rho) \mathrm{d}\rho = S_p^{-1} \\ \int \rho^{k+p-1} \widetilde{K}(\rho) \mathrm{d}\rho = 0 \quad \text{for every } 2 \leq k \leq r-1 \text{ and } k \text{ even,} \end{cases} \tag{C.6}$$

where $S_p = \frac{2\pi^{p/2}}{\Gamma(p/2)}$ is the surface of the sphere of $\mathbb{R}^p$. Assuming the form

$$\widetilde{K}(x) = \begin{cases} \sum_{l=0}^{\lfloor \frac{r+1}{2} \rfloor} \alpha_l x^l & \text{if } 0 < x < 1 \\ 0 & \text{elsewhere,} \end{cases}$$

we obtain the system

$$\begin{cases} \sum_{l=0}^{\lfloor \frac{r+1}{2} \rfloor} \frac{\alpha_l}{p+l} = S_p^{-1} \\ \sum_{l=0}^{\lfloor \frac{r+1}{2} \rfloor} \frac{\alpha_l}{k+p+l} = 0 \quad \text{for every } 2 \leq k \leq r-1 \text{ and } k \text{ even} \\ \sum_{l=0}^{\lfloor \frac{r+1}{2} \rfloor} \alpha_l = 0, \end{cases} \tag{C.7}$$

where the first two equations correspond to (C.3) and (C.4) and the last one implies the continuity of the kernel.

The bibliography

[1] Miguel A. Arcones and Evarist Giné. On the bootstrap of M-estimators and other statistical functionals. In *Exploring the limits of bootstrap (East Lansing, MI, 1990)*, pages 13–47. Wiley, New York, 1992.

[2] Philippe Barbe and Patrice Bertail. *The weighted bootstrap*, volume 98. Springer-Verlag, New York, 1995.

[3] M. Pilar Barrios and Santiago Velilla. A bootstrap method for assessing the dimension of a general regression problem. *Statist. Probab. Lett.*, 77(3) :247–255, 2007.

[4] Michel Benaïm and Nicole El Karoui. *Promenade aléatoire : Chaînes de Markov et simulations ; martingales et stratégies.* Editions Ecole Polytechnique, 2005.

[5] M. Peter Bentler and Jun Xie. Corrections to test statistics in principal hessian directions. *Statist. Probab. Lett.*, 47(4) :381–389, 2000.

[6] Philippe Besse. PCA stability and choice of dimensionality. *Statist. Probab. Lett.*, 13(5) :405–410, 1992.

[7] R. N. Bhattacharya and R. Ranga Rao. *Normal approximation and asymptotic expansions.* John Wiley & Sons, New York-London-Sydney, 1976.

[8] Patrick Billingsley. *Convergence of probability measures.* John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[9] Dennis D. Boos. On generalized score tests. *Amer. Statist.*, 46 :327–333, 1990.

[10] Haim Brezis. *Analyse fonctionelle*, volume 5. Masson, 1983.

[11] Wlodzimierz Bryc. *The normal distribution*, volume 100. Springer-Verlag, New York, 1995. Characterizations with applications.

[12] E. Bura and J. Yang. Dimension estimation in sufficient dimension reduction : a unifying approach. *J. Multivariate Anal.*, 102(1) :130–142, 2011.

[13] E. Bura and J. Yang. Dimension estimation in sufficient dimension reduction : a unifying approach. *J. Multivariate Anal.*, 102(1) :130–142, 2011.

[14] Efstathia Bura. Dimension reduction via parametric inverse regression. In L_1-*statistical procedures and related topics (Neuchatel, 1997)*, volume 31, pages 215–228. Inst. Math. Statist., Hayward, CA, 1997.

[15] Efstathia Bura and R. Dennis Cook. Extending sliced inverse regression : the weighted chi-squared test. *J. Amer. Statist. Assoc.*, 96(455) :996–1003, 2001.

[16] Snigdhansu Chatterjee and Arup Bose. Generalized bootstrap for estimating equations. *Ann. Statist.*, 33(1) :414–436, 2005.

[17] DM Chibisov. An investigation of the asymptotic power of the tests of fit. *Theory of Probability & Its Applications*, 10(3) :421–437, 1965.

[18] R. Dennis Cook. On the interpretation of regression plots. *J. Amer. Statist. Assoc.*, 89(425) :177–189, 1994.

[19] R. Dennis Cook. *Regression graphics*. John Wiley & Sons Inc., New York, 1998.

[20] R. Dennis Cook and Liliana Forzani. Likelihood-based sufficient dimension reduction. *J. Amer. Statist. Assoc.*, 104(485) :197–208, 2009.

[21] R. Dennis Cook and Bing Li. Dimension reduction for conditional mean in regression. *Ann. Statist.*, 30(2) :455–474, 2002.

[22] R. Dennis Cook and Liqiang Ni. Sufficient dimension reduction via inverse regression : a minimum discrepancy approach. *J. Amer. Statist. Assoc.*, 100(470) :410–428, 2005.

[23] R. Dennis Cook and Sanford Weisberg. Discussion of "sliced inverse regression for dimension reduction". *J. Amer. Statist. Assoc.*, pages 28–33, 1991.

[24] Y. Coudène. Une version mesurable du théorème de Stone-Weierstrass. *Gaz. Math.*, (91) :10–17, 2002.

[25] John G. Cragg and Stephen G. Donald. On the asymptotic properties of LDU-based tests of the rank of a matrix. *J. Amer. Statist. Assoc.*, 91(435) :1301–1309, 1996.

[26] John G. Cragg and Stephen G. Donald. Inferring the rank of a matrix. *J. Econometrics*, 76(1-2) :223–250, 1997.

[27] Arnak S. Dalalyan, Anatoly Juditsky, and Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *J. Mach. Learn. Res.*, 9 :1648–1678, 2008.

[28] J.-J. Daudin, C. Duby, and P. Trécourt. PCA stability studied by the bootstrap and the infinitesimal jackknife method. *Statistics*, 20, 1989.

[29] Víctor H. de la Peña and Evarist Giné. *Decoupling*. Springer-Verlag, New York, 1999.

[30] Michel Delecroix, Marian Hristache, and Valentin Patilea. On semiparametric M-estimation in single-index regression. *J. Statist. Plann. Inference*, 136(3) :730–769, 2006.

[31] LP Devroye and TJ Wagner. The strong uniform consistency of kernel density estimates. *Multivariate analysis*, 5 :59–77, 1980.

[32] Monroe D. Donsker. Justification and extension of Doob's heuristic approach to the Komogorov-Smirnov theorems. *Ann. Math. Statistics*, 23 :277–281, 1952.

[33] Jan Draisma. Small maximal spaces of non-invertible matrices. *Bull. London Math. Soc.*, 38(5) :764–776, 2006.

[34] Morris L. Eaton. A characterization of spherical distributions. *J. Multivariate Anal.*, 20(2) :272–276, 1986.

[35] Morris L. Eaton and David Tyler. The asymptotic distribution of singular values with applications to canonical correlations and correspondence analysis. *J. Multivariate Anal.*, 50(2) :238–264, 1994.

[36] Morris L. Eaton and David E. Tyler. On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Ann. Statist.*, 19(1) :260–271, 1991.

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

[38] Bradley Efron. *The jackknife, the bootstrap and other resampling plans*, volume 38. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1982.

[39] J. Fan and I. Gijbels. *Local polynomial modelling and its applications*, volume 66. Chapman & Hall, London, 1996.

[40] Jianqing Fan and Irène Gijbels. Variable bandwidth and local linear regression smoothers. *Ann. Statist.*, 20(4) :2008–2036, 1992.

[41] Herbert Federer. *Geometric measure theory*. Springer-Verlag New York Inc., New York, 1969.

[42] Jean-David Fermanian, Dragan Radulović, and Marten Wegkamp. Weak convergence of empirical copula processes. *Bernoulli*, 10(5) :847–860, 2004.

[43] Louis Ferré. Determining the dimension in sliced inverse regression and related methods. *J. Amer. Statist. Assoc.*, 93(441) :132–140, 1998.

[44] Ali Gannoun and Jérôme Saracco. An asymptotic theory for SIR_α method. *Statist. Sinica*, 13(2) :297–310, 2003.

[45] Len Gill and Arthur Lewbel. Testing the rank and definiteness of estimated matrices with applications to factor, state-space and ARMA models. *J. Amer. Statist. Assoc.*, 87(419) :766–776, 1992.

[46] Evarist Gine, Rafal Latala, and Joel Zinn. Exponential and moment inequalities for U-statistics. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47, pages 13–38.

[47] P. Hall and C. C. Heyde. *Martingale limit theory and its application*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1980. Probability and Mathematical Statistics.

[48] Peter Hall. *The bootstrap and Edgeworth expansion*. Springer-Verlag, New York, 1992.

[49] Peter Hall and Ker-Chau Li. On almost linearity of low-dimensional projections from high-dimensional data. *Ann. Statist.*, 21(2) :867–889, 1993.

[50] Peter Hall and Brett Presnell. Intentionally biased bootstrap methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 61(1) :143–158, 1999.

[51] Peter Hall and Susan R. Wilson. Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47(2) :757–762, 1991.

[52] W. Härdle, J. S. Marron, and A. B. Tsybakov. Bandwidth choice for average derivative estimation. *J. Amer. Statist. Assoc.*, 87(417) :218–226, 1992.

[53] Wolfgang Härdle and Thomas M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.*, 84(408) :986–995, 1989.

[54] Jorgen Hoffmann-Jorgensen. *Stochastic processes on Polish spaces*. Number 39. Aarhus Universitet. Matematisk Institut, 1991.

[55] Marian Hristache, Anatoli Juditsky, Jörg Polzehl, and Vladimir Spokoiny. Structure adaptive approach for dimension reduction. *Ann. Statist.*, 29(6) :1537–1566, 2001.

[56] Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, 29(3) :595–623, 2001.

[57] Feifang Hu and John D. Kalbfleisch. The estimating function bootstrap. *Canad. J. Statist.*, 28(3) :449–499, 2000. With discussion and rejoinder by the authors.

[58] I. A. Ibragimov and R. Z. Has′minskiĭ. *Statistical estimation*, volume 16. Springer-Verlag, New York, 1981.

[59] Hidehiko Ichimura. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics*, 58(1-2) :71–120, 1993.

[60] Mark E. johnson. *Multivariate statistical simulation*. New-York : John Wiley and sons, Inc., 1987.

[61] Frank Kleibergen and Richard Paap. Generalized reduced rank tests using the singular value decomposition. *J. Econometrics*, 133(1) :97–126, 2006.

[62] S. Lele. Resampling using estimating equations. In *Estimating functions*, volume 7, pages 295–304. Oxford Univ. Press, New York, 1991.

[63] Bing Li and Yuexiao Dong. Dimension reduction for nonelliptically distributed predictors. *Ann. Statist.*, 37(3) :1272–1298, 2009.

[64] Bing Li and Shaoli Wang. On directional regression for dimension reduction. *J. Amer. Statist. Assoc.*, 102(479) :997–1008, 2007.

[65] Bing Li, Hongyuan Zha, and Francesca Chiaromonte. Contour regression : a general approach to dimension reduction. *Ann. Statist.*, 33(4) :1580–1616, 2005.

[66] Ker-Chau Li. Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, 86(414) :316–342, 1991.

[67] Ker-Chau Li. On principal Hessian directions for data visualization and dimension reduction : another application of Stein's lemma. *J. Amer. Statist. Assoc.*, 87(420) :1025–1039, 1992.

[68] Yingxing Li and Li-Xing Zhu. Asymptotics for sliced average variance estimation. *Ann. Statist.*, 35(1) :41–69, 2007.

[69] D. Pollard. *Convergence of stochastic processes*. Springer-Verlag, New York, 1984.

[70] F. Portier and B. Delyon. Optimal transformation : A new approach for covering the central subspace. *Journal of Multivariate Analysis*, 115(0) :84 – 107, 2013.

[71] James L. Powell, James H. Stock, and Thomas M. Stoker. Semiparametric estimation of index coefficients. *Econometrica*, 57(6) :1403–1430, 1989.

[72] Jean-Marc Robin and Richard J. Smith. Tests of rank. *Econometric Theory*, 16(2) :151–175, 2000.

[73] Robert J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons Inc., New York, 1980.

[74] Thomas M. Stoker. Consistent estimation of scaled coefficients. *Econometrica*, 54(6) :1461–1481, 1986.

[75] David E. Tyler. Asymptotic inference for eigenvectors. *Ann. Statist.*, 9(4) :725–736, 1981.

[76] A. W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, Cambridge, 1998.

[77] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes.* Springer-Verlag, New York, 1996. With applications to statistics.

[78] Céline Vial. *Deux contributions à l'étude semi-paramétrique d'un modèle de régression.* PhD thesis, University of Rennes 1, 2003.

[79] Andrew T. A. Wood. An f approximation to the distribution of a linear combination of chi-squared variables. *Comm. Statist. Simulation and Computation*, 18(4) :1439–1456, 1989.

[80] Yingcun Xia. A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.*, 35(6) :2654–2690, 2007.

[81] Yingcun Xia, Howell Tong, W. K. Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3) :363–410, 2002.

[82] Zhishen Ye and Robert E. Weiss. Using the bootstrap to select one of a new class of dimension reduction methods. *J. Amer. Statist. Assoc.*, 98(464) :968–979, 2003.

[83] Xiangrong Yin and R. Dennis Cook. Dimension reduction for the conditional kth moment in regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(2) :159–175, 2002.

[84] Peng Zeng and Yu Zhu. An integral transform method for estimating the central mean and central subspaces. *J. Multivariate Anal.*, 101(1) :271–290, 2010.

[85] Li-Xing Zhu and Kai-Tai Fang. Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.*, 24(3) :1053–1068, 1996.

[86] Li Xing Zhu and Kai W. Ng. Asymptotics of sliced inverse regression. *Statist. Sinica*, 5(2) :727–736, 1995.