THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par:
Université Toulouse III Paul Sabatier (UT3 Paul Sabatier)

Discipline ou spécialité:
Mathématiques appliquées

Présentée et soutenue par
Maikol SOLÍS
le: 23 juin 2014

Titre:
Conditional covariance estimation for dimension reduction and sensitivity analysis

---

École doctorale :
Mathématiques Informatique Télécommunications (MITT)

Unité de recherche:
UMR 5219

Directeurs de thèse:
Jean-Michel LOUBES - Université Paul Sabatier
Clément MARTEAU - Institut National de Sciences Apliquées

Rapporteurs:
Benoît CADRE - École Normale Superior de Rennes
Anne-Françoise YAO - Université Blaise Pascal

Membres du jury:
Fabrice GAMBOA - Université Paul Sabatier
Theofanis SAPATINAS - University of Cyprus

"Gracias a la vida, que me ha dado tanto.
Me dio el corazón que agita su marco.
Cuando miro el fruto del cerebro humano.
Cuando miro al bueno tan lejos del malo.
Cuando miro al fondo de tus ojos claros."

Gracias a la vida
Violeta Parra (1966)

Acknowledgments

First, I would like to thank to my advisors Jean-Michel Loubes and Clément Marteau. I am grateful for their support when I did not know what to do, their time that they spent trying to improve my work and their scolding when I did the things wrong. I learned a lot with them. If I could summarize my experience with my advisors, I would say: "They taught me how to find my own path as scientific researcher". Thank you.

Also, I would like to thank to the MELISSA [1] and PREFALC [2] projects They were the access key to meet top french professor as Jean-Michel Loubes, François Dupuy and Thierry Klein in Costa Rica. I am grateful to the professors of the Universidad de Costa Rica Alexander Ramirez, Pedro Méndez, Javier Trejos and Jorge Poltronieri, that participated actively into those programs. Thanks to all them, in 2008 Jean-Michel offered me to make the *Master 2 Recherche* in the Institute de Mathématiques de Toulouse. I started the M2R in 2009 and then my PhD project in 2010.

I want to mention with special gratitude to Béatrice Laurent for the useful and always appropriate advises for my research.

Thank you very much to Benoît Cadre and Anne-Françoise Yao for their willingness to evaluate this thesis, their useful comments and for being members of the jury. Thanks also to Fabrice Gamboa and Theofanis Sapatinas for agreeing to be part of the jury. All they do me a great honor with their valuable comments and constructive suggestions.

[1] Mathematical E-Learning in Statistics for Latin and South America. `www.math.univ-toulouse.fr/MELISSA_UPS/`

[2] Programme Régional France-Amérique Latine-Caraïbes. `www.prefalc.msh-paris.fr/`

The Universidad de Costa Rica together with the Escuela de Matemática provided mainly my scholarship. I am deeply grateful for the confidence placed in me. Moreover, the French government gave me an intense support for my project through the Institut Français de Amérique Central in Costa Rica and Campus France (formerly called Égide) in Toulouse. I want to express my gratitude particularly to Vivian Madrigal, Norma Peña and Natalie Roubert.

The administrative staff of the Institut de Mathématiques de Toulouse were very kind to me, specially when deadlines and administrative problems emerged. I want to recognize all the patient and the kindness of Marie-Laure Ausset and Agnès Requis.

Every researcher needs a place to search for information and the Bibliothèque de Mathématiques et Mécanique was this place for me. Particularly, Dominique Barrère helped me to find all the books and articles that I needed for my research. Dominique: Thanks for all the documents that I could not find and you could.

I learned from my father that I have to fight for everything that I want to achieve. In life there is not shortcuts or easy ways, just work and hard work. My mother supported every decision that I took in my career (even if she did not know nothing about it). They formed me in the very same way that I am now. They will always have my gratitude and respect.

This journey could not be possible without the support of one single person: my wife Laura. She was my cornerstone when I started this project in 2009. She always pushed me and encouraged me when I needed the most. In this long and hard process we have sacrificed many things. Only she knows what I am talking about. Thank you my love. Of course, I could not forget to Pascal, our guinea pig that give us plenty of joy when we felt sad or tired.

Last but not least, I want to thank to all the friends and colleagues that I made these years: Tibo, Adil, Salomón, Santiago (les membres du bureau latin), Chloé, Ricardo, Lilian, Daniel, Willy and so many that I am overwhelmed. To all them, thank you.

Toulouse, France
23 Juin 2014
Maikol Solís

Résumé

Cette thèse se concentre autour du problème de l'estimation de matrices de covariance conditionnelles et ses applications, en particulier sur la réduction de dimension et l'analyse de sensibilités. En supposant qu'on observe $\boldsymbol{X} \in \mathbb{R}^p$ et $Y \in \mathbb{R}$ des variables aléatoires qui possèdent certaine distribution que, par rapport à la mesure de Lebesgue, ont une densité jointe $f(\boldsymbol{x}, y)$. La matrice conditionnelle de covariance $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y]) = (\sigma_{ij})_{p\times p}$ pour $i, j = 1, \ldots, p$, est constituée d'éléments σ_{ij} à estimer qui néanmoins, dépendent de la fonction f.

Ainsi, nous nous intéressons aux problématiques suivantes :

1. Construire des estimateurs des paramètres σ_{ij} les plus optimaux possibles. Le cadre de l'estimation est la statistique semi-paramétrique puisque nous estimons des paramètres dépendant d'une distribution non paramétrique.
2. Nous nous interrogeons sur l'efficacité de notre méthode d'estimation.
3. Estimer la matrice de covariance conditionnelle en entier en tant qu'operateur matricial.
4. Nous proposerons des applications de notre nouvelle méthode d'estimation.

La thèse est structurée de la manière suivante :

Chapitre 1

Le Chapitre 1 appliquera les enjeux développes dans cette thèse. Il présentera les différentes définitions et les méthodes mathématiques abordées dans ce

travail de recherche.

L'estimation de variance conditionnelle occupe une place importante dans deux problématiques : La réduction de la dimension en régression non linéaire et l'étude de l'analyse de sensibilité d'un modèle. Ces deux problèmes ont de nombreuses applications modernes dans la chimie, la biologie, l'économie ou le marketing, pour en nombrer que quelqu'un-es.

Réduction de la dimension Le travail de Li (1991a) présente la méthode de *régression inverse par tranches*. Cette méthode réduit le nombre de variables d'une régression non linéaire en grande dimension. Il suppose un modèle avec $\boldsymbol{X} \in \mathbb{R}^p$ et $Y \in \mathbb{R}$ qui sont les variables indépendantes et dépendante respectivement. La technique approxime l'espace central pour la réduction de la dimension, basée sur l'estimation des valeurs propres de la matrice conditionnelle $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$. Une fois que nous avons cet espace, nous sélectionnons les vecteurs propres associés aux plus grandes valeurs propres. Ces vecteurs sont la projection du modèle d'origine vers un nouvel ensemble réduit de variable. La question principal ici est l'estimation de $\mathbb{E}[\mathbb{E}[X_i|Y]\mathbb{E}[X_j|Y]]$ pour $i, j = 1, \ldots, p$.

Analyses de sensibilité Supposons un modèle avec entrées $(X_1, \ldots, X_p)$ et une sortie Y. Sobol' (1993) a montré que la valeur $\mathrm{Var}(\mathbb{E}[Y|X_i])/\mathrm{Var}(Y)$, pour $i = 1, \ldots, p$, mesure la sensibilité de X_i par rapport à Y. Cela signifie que ces indices quantifient combien d'entrée X_i apporte à la variabilité de la sortie Y. La littérature dans ce sujet se concentre dans la quantification de $\mathrm{Var}(\mathbb{E}[Y|X_i])$, en particulier dans l'espérance conditionnelle $\mathbb{E}[(\mathbb{E}[Y|X_i])^2]$.

Chapitre 2

Dans ce chapitre nous plaçons dans un modèle d'observation de type régression en grande dimension pour lequel nous souhaitons utiliser une méthodologie de type *régression inverse par tranches*. Pour cela nous proposons un nouvel estimateur de $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ qui repose sur un generalisation du travail de Da Veiga and Gamboa (2013). Ils ont proposé un estimateur basé sur une décomposition de Taylor pour $\mathbb{E}[(\mathbb{E}[Y|X_i])^2]$ pour $i = 1, \ldots, p$, qui nécessite l'estimation d'intégrales quadratiques en deux dimensions, étudiés avant par Laurent (1996). Notre travail cherche à generaliser leur travail au cas multidimensionel pour estimer $\mathbb{E}[\mathbb{E}[X_i|Y]\mathbb{E}[X_j|Y]]$ pour $i, j = 1, \ldots, p$.

Notons $f(x_i, x_j, y)$ la fonction de densité de (X_i, X_j, Y) et $f_Y(y)$ la densité marginale de Y. Nous allons commencer par la réécriture du terme non linéaire conditionnel

$$\mathbb{E}[\mathbb{E}[X_i|Y]E[X_j|Y]] = \int \left(\frac{\int x_i f(x_i, x_j, y)\, dx_i\, dx_j}{f_Y(y)} \right) \left(\frac{\int x_j f(x_i, x_j, y)\, dx_i\, dx_j}{f_Y(y)} \right) f(x_i, x_j, y)\, dx_i\, dx_j\, dy.$$

L'utilisation d'un opérateur fonctionnel en f, nous permettra d'appliquer la décomposition de Taylor autour d'un estimateur préliminaire de f appelé $\hat{f}$. Cet opérateur sera divisé en une partie linéaire, une quadratique et un terme d'erreur. Cette décomposition nous servira de base pour développer notre estimateur. En outre, la convergence asymptotique de la partie quadratique et du terme d'erreur est négligeable par rapport à la partie linéaire. Cette propriété nous permet de prouver deux choses : notre estimateur est asymptotiquement normal avec une variance que dépend de la partie linéaire, et cette variance est efficace selon le point de vue de Cramér-Rao. Nous allons également démontrer la normalité asymptotique pour la matrice complète à l'aide du *"half-vectorization"* opérateur. Encore une fois, la variance asymptotique de l'estimateur de la matrice complète sera uniquement dépendant de la partie linéaire de la matrice.

Chapitre 3

Dans ce chapitre, nous étudions l'estimation de matrices de covariance conditionnelles dans un cadre général. Il s'agit d'estimer dans un premier temps les matrices coordonnée par coordonnée, soit le paramètres $\mathbb{E}[\mathbb{E}[X_i|Y]\mathbb{E}[X_j|Y]]$ pour $i, j = 1, \ldots, p$. Ces espérances dépendent de la densité jointe inconnue que nous remplacerons par un estimateur à noyaux inspiré avec les idées de Härdle and Tsybakov (1991) et Zhu and Fang (1996).

Le principal résultat de ce chapitre se présente comme suit : si la distribution jointe de $(\boldsymbol{X}, Y)$ appartient à une classe de fonctions *lisses*, nous pouvons prouver que l'erreur quadratique moyenne de l'estimateur converge à une vitesse paramétrique. Sinon, nous aurons une vitesse plus lente en fonction de la régularité de la densité jointe.

Pour éviter des incohérences dans notre estimateur de la matrice, en raison de la grande dimension des données, nous allons appliquer une transformation de *"banding"* étudiée par Bickel and Levina (2008b). Nous allons montrer que sous une hypothèse de régularité sur la structure des matrices de covariance conditionnelles, nous obtiendrons de nouveau une vitesse paramétrique de convergence pour le risque quadratique sous la norme de Frobenius.

Chapitre 4

Nous allons dans ce chapitre utiliser nos résultats pour estimer des indices de Sobol utilisés en analyses de sensibilité, lorsqu'on observe une sortie d'un code numérique Y dépendant de variables d'entrée X_i, $i = 1, \dots, p$. Ces indices mesurent l'influence des variables et sont définis par

$$S_i = \frac{\mathrm{Var}(\mathbb{E}[Y|X_i])}{\mathrm{Var}(Y)} \quad \text{for} \quad i = 1, \dots, p.$$

Nous allons utiliser la méthodologie appliquée dans le Chapitre 3 pour estimer la valeur de $\mathbb{E}[(\mathbb{E}[Y|X_i])^2]$. En supposant, au moins, que la fonction de densité conjointe de (X_i, Y) est deux fois différentiable, nous pouvons prouver un comportement paramétrique de notre estimateur semi-paramétrique. L'avantage de notre implémentation est d'estimer les indices de Sobol sans l'utilisation de coûteuses méthodes de type Monte-Carlo. Certaines illustrations sont présentées dans le chapitre pour montrer les capacités de notre estimateur.

Abstract

This thesis will be focused in the estimation of conditional covariance matrices and their applications, in particular, in dimension reduction and sensitivity analyses. Suppose that we observe $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ two random variables with certain distribution. Denote as $f(\boldsymbol{x}, y)$ the joint density of $(\boldsymbol{X}, Y)$ with respect to the Lebesgue measure. The conditional covariance $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y]) = (\sigma_{ij})_{p\times p}$ for $i, j = 1, \ldots, p$ is formed by the elements $\boldsymbol{\sigma_{ij}}$ which depend on the function f.

Thus, we will be interested in the following problems:

1. Construct the estimator for the parameters σ_{ij} the most optimal possible. We will be in a semiparametric framework since we will estimate those parameters depending on one nonparametric distribution.
2. We will study the efficiency of our estimators.
3. Estimate the conditional covariance matrix as an operator.
4. We will propose some applications of our new estimation methods.

We structure this thesis as follows:

Chapter 1

The Chapter 1 introduces the challenges discovered in this thesis. We address all the different definitions and the mathematical methods used in all the text.

The estimation of the conditional covariance is linked to two problems: The dimension reduction in nonlinear regression and the study of sensitivity analysis in a model. Both problems have many modern applications in chemistry, biology, economics or marketing, just to name a few.

Dimension reduction The work of Li (1991a) presents the *sliced inverse regression* method. This method reduces the number of variables on a high-dimensional nonlinear regression. He assumes a model with $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ being the independent and dependent variables respectively. The core of its technique is the estimation of the spectral space of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$. Once we have this space, we select the eigenvectors associated to the largest eigenvalues. Those vectors are the projection of the original model to a new reduced set of variables. The main issue here is the estimation of $\mathbb{E}[\mathbb{E}[X_i|Y]\mathbb{E}[X_j|Y]]$ for $i, j = 1, \ldots, p$.

Sensitivity analysis Assume a model with inputs $(X_1, \ldots, X_p)$ and one output Y. Sobol' (1993) showed that the value $\mathrm{Var}(\mathbb{E}[Y|X_i])/\mathrm{Var}(Y)$, for $i = 1, \ldots, p$, measures the sensitivity of X_i with respect to Y. It means, these indices quantify how much the input X_i affects the variability of the output Y. The literature in this topic focuses in the quantification of $\mathrm{Var}(\mathbb{E}[Y|X_i])$, specifically in the conditional expectation $\mathbb{E}[(\mathbb{E}[Y|X_i])^2]$.

Chapter 2

In this chapter, we are in a context of high-dimensional nonlinear regression. The main objective is to use the *sliced inverse regression* methodology. For this, we propose a new estimator of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ generalizing the work of Da Veiga and Gamboa (2013). They have proposed one estimator based on a Taylor decomposition of $\mathbb{E}[(\mathbb{E}[Y|X_i])^2]$ for $i = 1, \ldots, p$. It requires the estimation of quadratic integrals in two dimensions, studied priorly by Laurent (1996). We search to generalize their work to the multidimensional case to estimate $\mathbb{E}[\mathbb{E}[X_i|Y]\mathbb{E}[X_j|Y]]$ for $i, j = 1, \ldots, p$.

Denote as $f(x_i, x_j, y)$ the density function of (X_i, X_j, Y) and $f_Y(y)$ the marginal density of Y. We start rewriting the conditional nonlinear term as

$$\mathbb{E}[\mathbb{E}[X_i|Y]E[X_j|Y]] = \int \left(\frac{\int x_i f(x_i, x_j, y)\, dx_i\, dx_j}{f_Y(y)} \right)$$

$$\left(\frac{\int x_j f(x_i, x_j, y)\, dx_i\, dx_j}{f_Y(y)} \right) f(x_i, x_j, y)\, dx_i\, dx_j\, dy.$$

Using a functional depending on f, we apply Taylor decomposition around a preliminary estimator of f called $\hat{f}$. This operator is split into a linear part, a quadratic one and an error term. The Taylor decomposition serve us as a base to develop our estimator. Moreover, the asymptotic convergence of the quadratic part and the error term are negligible with respect to the linear part. This property enable us to prove two things: our estimator is asymptotical normal with variance depending only on the linear part, and this variance is efficient from the Cramér-Rao point of view. We also prove the asymptotic normality for the whole matrix using a *"half-vectorization"* operator. Again, the asymptotic variance for the whole matrix estimator depend only in the linear part of the matrix.

Chapter 3

In this chapter, we study the estimation of conditional covariance matrices in a general framework. First, we estimate the matrix coordinate-wise given by the parameters $\mathbb{E}[\mathbb{E}[X_i|Y]\mathbb{E}[X_j|Y]]$ for $i, j = 1, \ldots, p$. These expectations depend on the unknown joint density. We will replace this density by a kernel estimator inspired by the ideas of Härdle and Tsybakov (1991) and Zhu and Fang (1996).

The main result of this chapter stands as follows: if the joint distribution of $(\boldsymbol{X}, Y)$ belongs to some class of *smooth* functions; we can prove that the mean squared error for the Nadaraya-Watson estimator of $\mathbb{E}[\mathbb{E}[X_i|Y]\mathbb{E}[X_j|Y]]$ has a parametric rate of convergence. Otherwise, we get a slower rate depending on the regularity of the model.

Therefore, we can expand our estimator to the whole matrix $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$. To avoid inconsistencies in our matrix estimator due to the high dimensionality of the data, we apply a banding transformation studied by Bickel and Levina (2008b). We prove that assuming some regularity structure in the covariance matrices, we get again a parametric rate for mean squared error under the Frobenius norm.

Chapter 4

In this chapter, we apply our results to estimate the Sobol or sensitivity indices. Assume that we observe one output Y from a numeric code depending on several inputs variables X_i, $i = 1, \ldots, p$. These indices measure the influence of the inputs with respect to the output and are defined by,

$$S_i = \frac{\mathrm{Var}(\mathbb{E}[Y|X_i])}{\mathrm{Var}(Y)} \quad \text{for} \quad i = 1, \ldots, p.$$

We will use the methodology applied in the Chapter 3 to estimate the value $\mathbb{E}[(\mathbb{E}[Y|X_i])^2]$. Assuming, at least, that the joint density function of (X_i, Y) is twice differentiable, we can prove a parametric behavior of our semiparametric estimator. The advantage of our implementation is that we can estimate the Sobol indices without use computing expensive Monte-Carlo methods. Some illustrations are presented in the chapter showing the capabilities of our estimator.

Contents

List of Figures

List of Tables

Notation

$\mathbb{N}$	Set of natural numbers
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^p$	p-dimensional Euclidean space
$(X_1, \ldots, X_p)$	p-tuple of random variables
$\mathcal{F}$	Space of functions
$\mathcal{G}$	Class of matrices
$\mathbb{L}^2$	Space of square integrable functions
$\mathcal{H}(\beta, L)$	Holder functional space with parameters β and L
$\lvert \cdot \rvert$	Absolute value
$\lVert A \rVert_F$	Frobenius norm of the matrix A
$(a_{ij})_{p\times p}$	$p \times p$ matrix with elements a_{ij} for $i, j = 1, \ldots, p$
$A^\top$	Transpose of the matrix A
A^{-1}	Inverse of the matrix A
I_p	$p \times p$ identity matrix
$\text{diag}(\lambda_1, \ldots, \lambda_p)$	Diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_p$
$\text{span}\{v_1, \ldots, v_K\}$	Vectorial space spanned by the vectors $\{v_1, \ldots, v_K\}$
$\mathbb{P}$	Probability
$f(x_1, \ldots, x_p)$	Joint density function of the random variable $(X_1, \ldots, X_p)$
$\mathbb{E}[X]$	Expectation of X
$\mathbb{E}[X\|Y]$	Conditional expectation of X given Y
$\text{Var}(X)$	Variance of X
$\text{Cov}(X)$	Covariance of X
$\mathcal{N}\left(m, s^2\right)$	Gaussian distribution with mean m and variance s^2
$\mathcal{N}\left(\boldsymbol{m}, \boldsymbol{S}\right)$	Multivariate Gaussian distribution with mean $\boldsymbol{m}$ and covariance

	matrix $\boldsymbol{S}$
K	Kernel function
ϕ', ϕ'', ϕ'''	First, second and third derivative of ϕ
$\phi^{(k)}$,	kth derivative of ϕ
$\ll$	Much less than
$\approx$	Same order of magnitude
vech	Half-vectorization operator
sup, inf	Supremum, Infimum
max, min	Maximum, Minimum
supp f	Support of the function f
argmin	Argument of the minimum
$\xrightarrow{\mathcal{D}}$	Converge in distribution
$\xrightarrow{\mathbb{P}}$	Converge in probability
$\xrightarrow{\mathbb{L}^2}$	Converge in $\mathbb{L}^2$-norm

Chapter 1

General presentation

1.1 Introduction

Real-world data usually lives in a high-dimensional space, such as in biology (Heeger and Ress (2002); Stears et al. (2003)), economics (Fan et al. (2011)), marketing (Dyer and Owen (2011)), among others. We illustrate some typical cases when the data has high-dimensionality:

- The samples of DNA microarrays are of the order of thousands of genes, but only a reduced set are the significant ones. The correct identification of these variables, allows detecting effectively malign diseases such as cancer.
- In finance and risk management, million of transactions run every second for every stock. Those transactions have multiple characteristics as the price, time to maturity, historic of interest rate, among others. The analysts have to compress, transform and interpret this information into a relevant subset to take immediate decisions.
- Machine learning and data mining aim to classify, predict and estimate a variety of process automatically. The size of the data in these sets can be astronomical. For instance, grocery sales, biomedical images, financial market trading, natural resources surveys or web services.

For any statistician, giving sense to some phenomenon with hundreds or thousands features is an overwhelming task. Given the complexity of those

problems, he needs to control the quantity of variables in the data set. He could select some relevant variables, and make his analyses with partials data sets or change the model.

The dimension reduction scheme is another technique that intends to handle the complexities of a model. It aims to reduce the number of variables by transforming the original data into a small data set. The new representation should have the minimum number of variables needed to observe the main properties of the original data (see Fukunaga (1972)). An effective reduction method allows, among others facilities: classification, visualization, and compression of the high-dimensional data. See for example, Donoho (2000) and Fan and Li (2006) for overviews of statistical challenges in high-dimensional applications.

In this thesis, we will motivate our contributions to the dimension reduction techniques starting with the following example. Let $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ be random variables. Then, the general nonparametric regression model in this framework is,

$$Y = m(\boldsymbol{X}) + \varepsilon. \tag{1.1}$$

Here ε is a noisy random variable independent of $\boldsymbol{X}$. The function $m : \mathbb{R}^p \mapsto \mathbb{R}$ is unknown and represents the conditional expectation of Y given $\boldsymbol{X}$.

In the model (1.1), we aim to estimate explicitly the function m by some estimator $\hat{m}$. It is possible to estimate $\hat{m}$ restricting to m to some parametric (linearity, quadratic, exponential) structure. In this case, it is only necessary to 3adjust the parameters that fit better in the model. Our objective is the estimation of m without imposing any predefined structure. However, nonparametric problems impose regularity conditions on m such belonging to some *smooth* functional class. The semiparametric models mix the two techniques to use the best of both of them. For a complete overview in nonparametric and semiparametric models we refer, to Hardle (1990), Green and Silverman (1994), Wand and Jones (1995), Fan and Gijbles (1996), Eubank (1999), Härdle (2004), Tsybakov (2009) and references therein.

In nonparametric regression, the dimensionality of $\boldsymbol{X}$ penalizes the rate of convergence of m to $\hat{m}$ (see Hastie et al. (2009)). In other words, if the dimensionality of $\boldsymbol{X}$ is large compared to the number of observations available, then the estimation of the m by a nonparametric method will be inaccurate. Some popular methods in nonparametric regression are kernel regression, local polynomial regression, smoothing splines, Fourier, or wavelet regression. The

literature split all those methods in two general schemes: The linear smoothers and the orthogonal series.

Linear smoothers: The linear smoothers are a popular approach to tackle the nonparametric regression. Recall that we have available an independent and identically distributed sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. In general, the estimator has the form,

$$\hat{m}(\cdot) = \sum_{k=1}^{n} Y_k \ell_k(\cdot) \tag{1.2}$$

where the ℓ_k's are functional weights depending on the sample. For any $x \in \mathbb{R}$, $\hat{m}(x)$ is the weighted average of Y_k's.

Nadaraya (1964) and Watson (1964) proposed, independently, the first nonparametric univariate regression estimator. They used a linear smoother with weights given by

$$\ell_k(\cdot) = \frac{K((\cdot - X_k)/h)}{\sum_{k=1}^{n} K((\cdot - X_k)/h)}$$

where K is a kernel function. The function K is usually a univariate density function, symmetric and supported on $[-1, 1]$. The first moment of K is equal to zero and h is a bandwidth depending on n. The function K express the way the weights decrease with the distance and h quantifies the closeness between points. In general, we could generalize the function $\ell(x)$ to be a polynomial of any degree. This extension is called *local polynomial regression*.

The local polynomial regression is easily adapted to that multivariate setting. For the Nadaraya-Watson estimator, Stone (1982) showed if m is s times differentiable, the optimal achievable rate of uniform convergence in norm $q < \infty$ of $\hat{m}$ to m is $n^{s/(2s+p)}$. This rate has sense only if $p \ll n$, otherwise the dispersion of the data in such high-dimensional space hampers the performance of the estimator.

Orthogonal series: Another type of methods are based in the called projection estimators or an orthogonal series estimators. The idea is to approximate $m(x)$ to its projection $\sum_{j=1}^{M} \theta_k b_k(x)$, where $b_1(x), \ldots, b_M(x)$ is a functional basis. The number M acts as a smoothing parameter in the same way as h. We can estimate the coefficients θ's by its empirical version. Even more, we can rewrite a projection or orthogonal regression estimator into the form (1.2) taking

$$\ell_k(x) = \frac{1}{n} \sum_{j=1}^{M} b_j(X_k) b_j(x)$$

The bases $\{b_j\}$ most used in projection are the trigonometric (Fourier), the polynomials (splines) and the wavelets. For classical references about orthogonal projections, we refer to Friedman and Silverman (1989), Friedman (1991), Moussa and Cheema (1992) and Stone et al. (1997).

Some important early contributions in nonparametric regression are Shibata (1981) and Rice (1984). The discussion continued with the references of Eubank (1999), Efromovich (1999), Wasserman (2007) and Massart (2007). During the 1990s, the wavelets techniques dominated the literature in nonparametric regression. From the invention of the wavelets by Meyer (1990), other authors started the use in nonparametric regression like Donoho and Johnstone (1994) and Donoho et al. (1996). For an extensive overview and references we refer to Härdle et al. (1998). Also other models in the same spirit are the project pursuit regression (PPR) by Friedman and Stuetzle (1981) and the alternating conditional expectations (ACE) by Breiman and Friedman (1985). In general all these models fit into the area of generalized additive models. The classic reference is given by Hastie and Tibshirani (1990).

The nonparametric models that we have reviewed in this section suffer from the *"curse of dimensionality"*, term coined by Bellman (2003); Bellman et al. (1961). This means that the sample needed to estimate some process, to a given degree of accuracy, grows exponentially with the number of variables. In other words, if $p \approx n$ or $p \gg n$ the model complexity blurs the relation between $\boldsymbol{X}$ and Y, hindering its properties. Recall the rate of convergence of $n^{s/(2s+p)}$ for the nonparametric regression estimator found by Stone (1982). Given certain regularity s on the density, we see that as p goes infinity faster than n, then the rate of convergence loses its efficiency. Scott and Thompson (1983) remarked that the responsible of the curse of dimensionality is the *empty space phenomenon*: high-dimensional spaces are inherently *sparse*. The following example illustrate the problem: One-dimensional normal standard normal distribution $\mathcal{N}(0,1)$ have 70% of its mass contained in the interval $[-1,1]$. For a 10-dimensional $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_{10})$, the hyper-sphere with radius 1 contains only 0.02% of its mass. We have to extend the radius to more than 3 to get the 70%.

Summarizing, when the dimension is high, any model-fitting method will be unsuccessful. For this reason, it is necessary reduce the dimensional space before examining the data further.

1.2 Dimension Reduction

Nowadays, we find high-dimensional problems in almost every modern statistical applications. Such problems have the number of features (p) bigger than the number of observations (n) available. One of the properties with high-dimensional datasets is that, frequently, only a reduced set of variables are "important" to the understanding the underlying process. The variables are redundant for two main reasons: their variances are lower than the model noise; or they are correlated with each other (through some functional dependence). In any case, it is necessary to extract only the independent and relevant sources of information in data. Fodor (2002) defines the dimension reduction scheme as:

> *Given the p-dimensional random variable* $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ *find a lower dimensional representation of it,* $\boldsymbol{S} = (S_1, \ldots, S_K)^\top$ *with* $K \leq p$*, that captures the content in the original data, according to some criterion.*

Therefore, analyze directly a high-dimensional problem with the raw data is not only naive but impractical. Generally, we need to reduce the dimensionality of the data into a manageable size, preserving as much of the original information as possible. Thus, we apply—to the reduced-dimensional data—some technique to explain our model such as classification, visualization, hypothesis tests, parametric or nonparametric regression and so on.

An effective dimension reduction technique finds the minimum number of variables that explains with high fidelity the original process. Bennett (1969) called this minimum the *intrinsic dimensionality* while studying collection of signals. Determinate the intrinsic dimensionality is a core problem because it avoids the possibility of over- or under-fit the model. In this thesis we will not study the estimation of intrinsic dimensionality. However, a diversity of methods exist to estimate it, for instance: the correlation dimension method, local PCA or the reconstruction error. We refer to Lee and Verleysen (2007) for general references on intrinsic dimension estimation.

In general we can find three general types of dimension reduction techniques in the literature: the variable selection, the manifold learning and the projection pursuit. We will mention briefly the two first approaches, and we will explain with some detail the projection pursuit. The projection pursuit will introduce a dimension reduction technique called *sufficient dimension reduction*, which will be one of the basis for this thesis.

1.2.1 Variable selection

The variable or feature selection, discards the irrelevant features of the model preserving only the most "interesting" variables. Before presenting a review on variable selection, we have to establish what *relevant* or *irrelevant* means first. Diverse classifications exist in the literature, but the most popular is:

- **Relevant features:** They explain, by themselves or in a subset with other features, information about the model.
- **Redundant features:** We can remove these features because, another feature or subset of feature, already have the same information about the model.
- **Noisy features:** Those features only have noisy information and do not contribute to explain any information of the model.

For a more mathematical definition on feature relevance, Gennari et al. (1989), John et al. (1994) and Kohavi and John (1997) characterize the selection problem in detail.

The issue in feature or variable selection is how to pick the relevant feature and discard the others. For instance, Blum and Langley (1997), Guyon and Elisseeff (2003) and Perkins and Theiler (2003) classify the type of selection into two classes: wrapper/embedded methods and filter methods.

Wrapper/embedded methods: The wrapper methods score different subsets of features according to their predictive power via a black-box learning machine. They were popularized by Kohavi and John (1997) given the powerful way to address the variable selection problem. Those methods work under a "brute-force" approach requiring a massive amount of computational time. In fact, Amaldi and Kann (1998) proved that the variable selection is NP-hard. However, efficient techniques could alleviate the performance issue. Two popular methods in this spirit are the *forward selection* and *backward elimination*. In forward selection, we start with an empty set of features and then add variables progressively. Backward elimination starts with the full set of features and removes the least promising ones. Both cases search to maximize some score function in each step.

In a similar context, the embedded methods perform variable selection in the training process step. In other words, they embed the feature selection in the

induction algorithm for classification. Some examples on embedding methods are the decision tree algorithms ID3 (Quinlan (2007)) or the weighted models (Payne and Edwards (1998) and Perkins and Theiler (2003)).

Filter methods: They are preprocessing methods that attempt to identify the best features from the data, without take into account the properties of the predictor. The simplest filtering method uses the mutual correlation between each variable and the target function. It computes all the correlations and takes the K features with the highest values. The filter methods are faster than the wrapper methods, since they avoid to search over all the variable space. However, given the independence with the predictor, some investigations argue that filter methods are not tuned for a given learning machine. For instance Almuallim and Dietterich (1994) developed the FOCUS algorithm which first searches individual features, then pairs, then triplets and so on, until finding the minimal feature set. Kira and Rendell (1992) presents the RELIEF algorithm which evaluates individually the features and keeps only the best K features. Gilad-Bachrach et al. (2004) describes a margin based feature selection algorithm.

1.2.2 Manifold learning

Other recent techniques to study this dimension reduction use the underlying nonlinear complex structure of the data and treat it as an abstract object—or manifold.

For example, Tenenbaum et al. (2000) developed the Isometric featuring mapping—Isomap—algorithm which can be viewed as a generalization of the multidimensional scaling method. It estimates the nonlinear proximity between the variables with the geodesic distance instead of the euclidean one. The Isomap learns the geodesic distances by linearly approximating the nonlinear manifold. Thus, it constructs an undirected neighborhood graph where each point is a node. Finally, it computes a geodesic square distance matrix where we project the original manifold to another in low-dimension.

Another popular method is the Local Linear Embedding—LLE—(Roweis and Saul (2000)) which profits the intrinsic geometry of the manifold. The LLE project a near-group of points on the manifold to an euclidean space via a convex linear transformation. We estimated the new representation solving a series of least squares problems based on a k-neighborhood graph.

Other authors have developed alternative algorithms over the last decade. Some of these are: Laplacian Eigenmap (Belkin and Niyogi (2003)), Hessian Eigenmap (Donoho and Grimes (2003)), Diffusion maps (Coifman and Lafon (2006)) and Local Tangent Space Alignment—LTSA—(Zhang and Zha (2004)). For surveys on manifold algorithms see Cayton (2005), Lee and Verleysen (2007), Izenman (2008) and Engel et al. (2012).

1.2.3 Projection pursuit

Pearson (1901) proposed one of the oldest technique in dimension reduction, the principal component analysis. This technique constructs a linear subspace in low-dimension minimizing the distance to the original data. Numerous authors have rediscovered or extended the principal component analysis in diverse areas. We can cite for example, the Hotelling transform (Hotelling (1933)), the Karhunen-Loève transform (Karhunen (1946) and Loève (1955)), the empirical orthogonal functions (Lorenz (1956)) and the proper orthogonal decomposition (Lumley (1967)).

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent and identical distributed observations taken from random variable $\boldsymbol{X} \in \mathbb{R}^p$. Assume for simplicity that each $\boldsymbol{X}$ is centered. The method of moments defines the classic sample covariance estimator for our sample,

$$\widehat{\Sigma}_{\boldsymbol{X}} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{X}_k \boldsymbol{X}_k^\top \tag{1.3}$$

The matrix $\widehat{\Sigma}_{\boldsymbol{X}}$ is symmetric semidefinite positive and admits a spectral decomposition

$$\widehat{\Sigma}_{\boldsymbol{X}} = U \Lambda U^\top.$$

Here $U = (u_1, \ldots, u_p)_{p \times p}$ is an orthogonal matrix with columns vectors u_i, the matrix Λ is equal to $\operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ and $\lambda_i \geq 0$ for $i = 1, \ldots, p$. Each vector u_i represents the normalized eigenvector of $\widehat{\Sigma}_{\boldsymbol{X}}$ associated with the eigenvalue λ_i. We create a new set of coordinates transforming the original variables to $\boldsymbol{Z} = U^\top \boldsymbol{X}$. In this new reference system, the variables Z has mean 0 and diagonal covariance matrix Λ. Therefore, we can discard the variables with small variance, reducing the space to only the most $K \leq p$ relevant variables. The principal component method finds the best linear subspace (in the least square sense). It projects the original data $\boldsymbol{X}$ to the subspace spanned by $U_K = (u_1, \ldots, u_K)$.

The principal component analysis belongs to a general set of methods called projection pursuit. This is an unsupervised technique that picks relevant low-dimensional linear orthogonal projection of a high-dimensional space, maximizing an objective function called the *projection index*. To obtain uncorrelated direction, the projection pursuit constraints the search to normalized and mutually orthogonal spaces. A projection index J is a real function on the space of square integrable distributions, i.e., $J\colon f \in \mathbb{L}^2 \mapsto I(f) \in \mathbb{R}$ where f is a probability density function. Abusing of notation, we will write $I(\boldsymbol{X})$ instead of $I(f)$ where $\boldsymbol{X}$ is random variable having the distribution of f. The optimization problem is then

$$\max_{\boldsymbol{A}^\top \boldsymbol{A}=I} J(\boldsymbol{A}\boldsymbol{X}).$$

with $\boldsymbol{A}$ an orthonormal matrix. Two classic examples of projection pursuit techniques are:

- The principal component analysis (Hotelling (1933)) maximizes the variances of the data with the variable $\boldsymbol{Z}$ restricted to a linear space of dimension K.

$$\max_{\boldsymbol{A}^\top \boldsymbol{A}=I} \left\{ \operatorname{tr}\left(\frac{1}{n} \sum_{k=1}^{n} \boldsymbol{Z}_k \boldsymbol{Z}_k^\top \right) \right\} = \sum_{k=1}^{K} \lambda_k$$

  with $\boldsymbol{Z} = U^\top \boldsymbol{X}$ restricted to $A = U_K$.

- The Fisher matrix information (Huber (1985)) measures the information of the variable $\boldsymbol{X}$ assuming a distribution parameterized by θ.

$$J(\boldsymbol{X}) \equiv J'(\theta) \equiv \mathbb{E}\left\{ \left(\frac{\partial f(\boldsymbol{X};\theta)}{\partial \theta} \right) \left(\frac{\partial f(\boldsymbol{X};\theta)}{\partial \theta} \right)^\top \right\}^{-1}$$

  where $f(\boldsymbol{X};\theta)$ depends on some parameter θ.

The vector-based approach is useful to facilitate our intuitive understanding. But, in practice, we use a matrix-based theory which is usually preferred for their brevity and rich mathematical support. Recall that the covariance matrix represents the variability or *"interestingness"* of each variable with respect to another. For any covariance matrix Σ, the eigenvalues λ and eigenvectors v of Σ are constructed by the relation $\Sigma v = \lambda v$. The eigenvalues represent a score of variability for each feature in the matrix Σ and the eigenvectors are the directions that maximize this score. Thus, we seek those directions v that amplify this variability for the covariance matrix.

The main aim in projected pursuit methods is the estimation of the spectral space of the sample covariance matrix $\widehat{\Sigma}_X$ (or some equivalent form). Some examples illustrate how the projection pursuit methods use the covariance matrix to find the directions:

Kernel PCA: Assumes that a nonlinear function or Kernel approximates the inner product of data point in the feature space. Thereby, the Kernel PCA computes a reduced set of direction through the eigenvectors of the covariance matrix of the data in the feature space.

Canonical correlation: This method takes two pairs of variables and seeks the directions that create the maximum correlation between the variables. The eigenvectors of the largest eigenvalues of the cross-correlation matrix produce these directions.

Multidimensional Scaling: Assume a matrix of dissimilarities (in some norm) between a set of features. The multidimensional scaling searches a low dimensional space—embedded into the original one—that preserve the distance between the variables. The eigenvectors associated to the largest eigenvalues of the matrix of dissimilarities generate this new space.

Other projection pursuit method are factor analysis, linear discriminant analysis, correspondence analysis, among others. The reader can found general reference on projection pursuit models in Friedman and Tukey (1974), Jones and Sibson (1987), Burges (2009), Engel et al. (2012) and Huber (1981).

The covariance matrix is the essential point to execute any projection pursuit technique. In the next section we will review some approaches to estimate it in a high-dimensional context. Then, in Section 1.3 we will explore a supervised paradigm that improves the projection pursuit.

1.3 The sliced inverse regression

We have discussed about the *"curse of dimensionality"* on the nonlinear regression and some techniques to overcome it. The classical projection pursuit methods belong to the unsupervised techniques that aim to reduce the space of some variable in high-dimension. Nevertheless, we look for a supervised scheme that shrinks the dimension in a model driven by equation (1.1).

In particular, projection pursuit methods have a long history in the literature and recently a branch called *sufficient dimension reduction* has gained much popularity. Li (1991a) wrote one of the historical articles in sufficient dimension reduction which introduced the method called sliced inverse regression. He started changing the model (1.1) by transforming the original variables $\boldsymbol{X} \in \mathbb{R}^p$ on a reduced subset only. In particular, this reduction will keep the structural information of the model, using the independent variable $\boldsymbol{X}$ and the dependent variable Y. For example, recall that the principal component analysis only reduces the dimensionality of $\boldsymbol{X}$ to later apply a regression technique in model (1.1). The sliced inverse regression finds a reduced subspace, similar to principal component analysis, but incorporating the information of the output Y directly.

The works of Cook and Nachtsheim (1994), Cook (1994), Cook and Li (2002) and Cook (2003) characterize this novel paradigm. Assume that we have $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. The goal of sufficient dimension reduction is to find a map $\mathcal{M} : \mathbb{R}^p \mapsto \mathbb{R}^K$, with $K \ll p$, such as the distribution of $Y|\boldsymbol{X}$ is equal to the distribution of $Y|\mathcal{M}(\boldsymbol{X})$. Cook showed that the last assertion is equivalent to

$$Y \perp\!\!\!\perp \boldsymbol{X}|\mathcal{M}(\boldsymbol{X}).$$

where $\perp\!\!\!\perp$ stands for probabilistic independence.

Thus, we can model Y only with a subset generated by $\mathcal{M}(\boldsymbol{X})$. In particular, Li (1991a) presented an equivalent form of the latter equation, using a linear transformation $\mathcal{M}(\boldsymbol{X}) = (v_1^\top \boldsymbol{X}, \ldots, v_K^\top \boldsymbol{X})$ where the v's are unknown vectors to be estimated from data. This mapping transforms the model (1.1) into the following semiparametric model,

$$Y = \phi(v_1^\top \boldsymbol{X}, \ldots, v_K^\top \boldsymbol{X}, \varepsilon). \tag{1.4}$$

Here ϵ is independent of $\boldsymbol{X}$ and ϕ is an arbitrary function in $\mathbb{R}^{K+1}$.

The variable Y explains a high-dimensional event via the covariate $\boldsymbol{X}$. Model (1.4) represents the weakest form that the information of Y could be retrieved by the low-dimensional space $v_1^\top \boldsymbol{X}, \ldots, v_K^\top \boldsymbol{X}$ when K is small. This model gathers all the relevant information about the variable Y, with only a projection onto the $K \ll p$ dimensional subspace $(v_1^\top X, \ldots, v_K^\top X)$.

The Figure 1.1 presents a simplified scheme of the change of model suggested by Li. If K is small, the method reduces the dimension by estimating the v's efficiently. We call the v's the effective dimension reduction directions and the

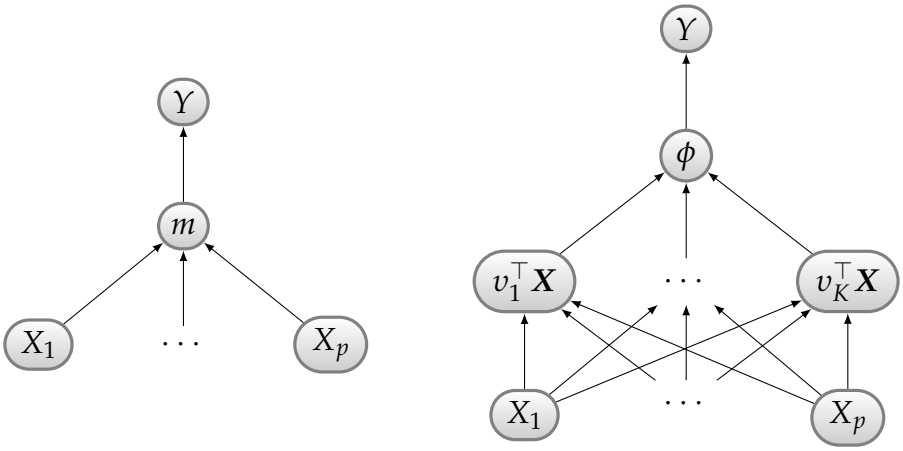


Figure 1.1: **Left diagram:** Interactions in the model $Y = m(\boldsymbol{X}) + \varepsilon$. The variables $X_1, \ldots, X_p$ affect directly the output Y through the function m. **Right diagram:** Interactions in the model $Y = \phi(v_1^\top \boldsymbol{X}, \ldots, v_K^\top \boldsymbol{X}, \varepsilon)$. The original variables $X_1, \ldots, X_p$ are transformed to a reduced subset $v_1^\top \boldsymbol{X}, \ldots, v_K^\top \boldsymbol{X}$ with $K \ll p$. Then we use this reduced subset to model Y through ϕ.

$\text{span}\{v_1, \ldots, v_K\}$ the effective dimension reduction space. This method is used to search nonlinear structures in data and to estimate the projection directions.

The estimation of ϕ is unnecessary to find the effective dimension-reduction directions, contrary to model-fitting algorithms reviewed in Section 1.1. That is, the sliced inverse regression method only requires the dataset information to find the effective dimension reduction directions. Nevertheless, we could fit a nonparametric model to ϕ (or any statistical scheme, e.g., classification) in the new coordinate system to understand better our model.

The sliced inverse regression method requires technical conditions about the law of $\boldsymbol{X}$,

Condition 1.1. For any direction $b \in \mathbb{R}^p$, $\mathbb{E}[b\boldsymbol{X}|v_1^\top \boldsymbol{X}, \ldots, v_K^\top \boldsymbol{X}]$ is a linear combination of $v_1^\top \boldsymbol{X}, \ldots, v_K^\top \boldsymbol{X}$.

Condition 1.1 seems to impose a restrictive requirement on the distribution of $\boldsymbol{X}$. In fact, Li (1991a) mentioned that the condition occurs on the elliptical distributions (e.g., the normal distribution). Later, Cook and Weisberg (1991) showed explicitly that is a weaker version that characterizes elliptical distributions. In a further work, Li (1991b) and then Hall and Li (1993) proved that for many high-dimension random vectors the condition holds approximately. We

can ensure Condition 1.1 doing a normal resampling suggested by Brillinger (1991); or averaging the data (by a discrete probability measure) such as they get closer to an elliptical random vector studied by Cook and Nachtsheim (1994).

We state the following two results in sliced inverse regression.

Theorem 1.1 (Li (1991a)). *Under the model* (1.4) *and Condition 1.1, the centered inverse regression curve* $\mathbb{E}[\boldsymbol{X}|Y] - \mathbb{E}[\boldsymbol{X}]$ *is contained in the linear subspace spanned by* $v_k\Sigma_X$ *(*$k = 1, \ldots, K$*), where* Σ_X *denotes de covariance matrix of* $\boldsymbol{X}$*.*

Corollary 1.1 (Li (1991a)). *Assume that* $\boldsymbol{X}$ *has been standardized to* $\boldsymbol{Z}$ *. Under the model*

$$Y = \phi(\eta_1^\top \boldsymbol{Z}, \ldots, \eta_K^\top \boldsymbol{Z}, \varepsilon)$$

and Condition 1.1, the standardized inverse regression curve $\mathbb{E}[\boldsymbol{Z}|Y]$ *is contained in the linear subspace spanned by* $\eta_k = v_k\Sigma_X^{1/2}$ *(*$k = 1, \ldots, K$*).*

Under Condition 1.1, Li showed that the effective dimension reduction space belongs to the spectral subspace spanned by $\mathrm{Cov}(\mathbb{E}[\boldsymbol{Z}|Y])$. In other words, the covariance matrix $\mathrm{Cov}(\mathbb{E}[\boldsymbol{Z}|Y])$ is degenerated in any direction orthogonal to the η_k's. Therefore, the eigenvectors η_k associated with the largest K eigenvalues of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{Z}|Y])$ are the standardized effective dimension reduction directions.

We present the classic sliced inverse regression procedure in the Algorithm 1.1. Notice that steps 2 and 3 create an approximation of $\mathbb{E}[\boldsymbol{Z}|Y]$ slicing the support of Y dividing it in H parts. Using this roughly approximation, the step 4 calculates empirically $\mathrm{Cov}(\mathbb{E}[\boldsymbol{Z}|Y])$ through a weighted covariance matrix $\hat{V}$. Then, it finds the eigenvalues and eigenvectors of $\hat{V}$ and returns the eigenvectors associated to the largest K eigenvalues of the estimated matrix.

We have to emphasize the following result, product from Corollary 1.1 and Algorithm 1.1:

> **To find the effective dimension reduction space of model** (1.4)**, it is enough to estimate the matrix** $\mathrm{Cov}(\mathbb{E}[\boldsymbol{Z}|Y])$**. Once this matrix is estimated, the eigenvectors associated to the** $K \ll p$ **largest eigenvalues of** $\mathrm{Cov}(\mathbb{E}[\boldsymbol{Z}|Y])$ **are the effective dimension reduction directions.**

Summarizing, we will address the high-dimension problem from the sufficient dimension reduction point of view. However, the estimation of the

Algorithm 1.1 Classic sliced inverse regression method

Require: A sample $(\boldsymbol{X}_i, Y_i)$ $i = 1, \ldots, n$.
Ensure: $\hat{v}_k$ for $k = 1, \ldots, K$.

1: Standardize $\boldsymbol{X}$ by an affine transformation to get $\boldsymbol{Z}_i = \Sigma_{\boldsymbol{X}}^{-1/2}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})$ $(i = 1, ..., n)$, where $\widehat{\Sigma}_{\boldsymbol{X}}$ and $\bar{\boldsymbol{X}}$ are the sample covariance matrix and sample mean of $\boldsymbol{X}$ respectively.
2: Divide range of Y into H slices, $I_1, \ldots, I_H$; let the proportion of the Y_i that falls in the slice h be $\hat{p}_h$; that is $\hat{p}_h = (1/n)\sum_{i=1}^{n} \delta(Y_i)$, where $\delta(Y_i)$ takes the values 0 or 1 depending on whether Y falls into the h-th slice I_h or not.
3: Within each slice, compute the sample mean of the $\boldsymbol{Z}_i$'s denoted by $\hat{\mu}_h$, i.e., for each $h = 1, \ldots, H$ estimate $\hat{\mu}_h = (1/n\hat{p}_h)\sum_{y \in I_h} \boldsymbol{Z}_i$.
4: Conduct a (weighted) principal component analysis for the data $\hat{\mu}_h$ in the following way: Form the weighted covariance matrix $\hat{V} = \sum_{h=1}^{H} \hat{p}_h \hat{\mu}_h \hat{\mu}_h^\top$ then find the eigenvalues and the eigenvectors for $\hat{V}$.
5: Let the K largest eigenvectors (row vectors) be $\hat{\eta}_k$ $(k = 1, \ldots, K)$. Estimate $\hat{v}_k = \hat{\eta}_k \widehat{\Sigma}_{\boldsymbol{X}}^{-1/2}$

empirical covariance matrix is insufficient for this purpose. Therefore, we have to use another richer object. In particular, we shall use the conditional covariance matrix $\mathrm{Cov}(\mathbb{E}[\boldsymbol{Z}|Y])$ to solve a connected problem called sliced inverse regression. In particular, we will use the conditional covariance matrix to reduce the dimensionality in problems like nonparametric regression and estimation of sensitive indices.

1.4 Covariance estimation

The purpose in this thesis will be the estimation of high-dimensional conditional covariances. This estimator will be used to reduce the dimensionality on the nonlinear regression problem (1.1). In Section 1.4.1, we will present a bibliographic review of techniques to overcome the dimensionality for the sample covariance matrix $\widehat{\Sigma}_X$ introduced in (1.3). Afterwards, in Section 1.4.2, we will examine some classic approaches for the estimation of conditional covariances in the sliced inverse regression context.

1.4.1 High-dimensional covariance estimation

As we mentioned in Section 1.2.3, the estimation of the covariance matrix—or some related form—is an essential tool for the analysis. The estimator $\widehat{\Sigma}_{\boldsymbol{X}}$

defined in (1.3) has well known properties as: unbiasedness, consistency, parametric convergence, among others. However, these properties become useless when the dimension of the model turns high and the empirical covariance matrix has unexpected features.

For example, take a sample from a multivariate Gaussian distribution $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma_p)$ where $\Sigma_p = I_p$ denotes the identity matrix of size p. If p/n converges to some constant c, then the empirical distribution of the eigenvalues of the sample covariance matrix $\widehat{\Sigma}_p$ follows the Marĉenko-Pastur law (Marčenko and Pastur (1967)), which is supported on $((1-\sqrt{c})^2, (1+\sqrt{c})^2)$. Therefore, if the dimension grows faster than the number of observations, the eigenvalues will be more spread out. Some general references about similar issues in numerous contexts are presented in Muirhead (1987), Johnstone (2001), Fan et al. (2008), and references therein.

The authors have proposed solutions to mitigate the irregularities in covariance estimation. The general procedure on these situations occurs in two stages: first we need to restrict our attention to *"well-behaved"* subsets of covariance matrices and then incorporate some *regularization* into the estimation. The term *"well-behaved"* is vague because it depends on the problem. Sparsity is one common hypothesis on these cases and it seems realistic for real-world applications. Another popular matrix class is to assume that the entries of the covariance decrease at some rate far away of the diagonal.

Roughly speaking, the literature classifies regularized estimation of high-dimensional covariance matrices into two major categories:

Shrinkage type: These estimators shrink the covariance matrix to some well-conditioned one under different performance measures. For instance, Ledoit and Wolf (2004) and Chen et al. (2010) proposed a convex combination between $\widehat{\Sigma}_X$ and $p^{-1}\operatorname{Tr}(\widehat{\Sigma}_X)I_p$, with the optimal combination weight that minimizes the mean-squared errors. Recently, Fisher and Sun (2011) proposed using $\operatorname{diag}(\widehat{\Sigma}_X)$ as the shrinkage target with possibly unequal variances. Other estimators in this line are: Lam and Fan (2009) minimizes a penalized log-likelihood with a L_1-penalty; Meinshausen and Bühlmann (2006) uses Lasso in graph models and Levina et al. (2008) studies a nested Lasso together with a banding technique both to find the structural zeros in sparse covariance matrices.

Matrix transformation type: Estimators in this category operate directly in the covariance matrix through operators. For instance, thresholding (Bickel

and Levina (2008a)), banding (Bickel and Levina (2008b)) and tapering (Cai et al. (2010)). When the natural ordering exists in the variables (for example in time-series) the banding and tapering techniques regularize better the empirical covariance. The banding sets to zero the elements of the matrix away of some chosen subdiagonal and keep the other entries unchanged. Tapering, improves the banding by applying some smoother (originally a linear one) to shrink gradually the values outside some subdiagonal. We can view banding as a hard-thresholding rule while tapering is a soft-thresholding rule, up to a certain unknown permutation (Bickel and Lindner (2012)). The thresholding regularization deals with general permutation-invariant covariance matrices cutting-off the entries below some level. Assuming certain covariance class of matrices, the banding, tapering and thresholding regularizations are statistical consistent. In fact, the tapering and thresholding techniques produce optimal rates of convergence (see Cai et al. (2010), Cai et al. (2012) and Cai and Zhou (2012)), contrary to the banding technique (Bickel and Levina (2008b)).

1.4.2 Conditional covariance estimation

The original sliced inverse regression has suffered alternative improvements through the last decades. Recall that the key point in the method is the estimation of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ for $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$.

Cook and Weisberg (1991) discussed about the validity of Condition 1.1 and suggested a variance checking condition. He proposed a new method called sliced average variance estimation (SAVE) using the first two moments of the inverse conditional expectation. In particular, they proved that the sliced inverse regression might fail under symmetric models because $\mathbb{E}[\boldsymbol{X}|Y]$ could be 0. Thus, they suggested the use of $\mathrm{Var}[\boldsymbol{X}|Y]$ to recover the effective dimension reduction directions. Using the same notation as Algorithm 1.1, they proposed the following estimator

$$\mathrm{SAVE} = \sum_{h=1}^{H} (I - \mathrm{Var}(\boldsymbol{Z}|Y \in I_h)),$$

where I is identity matrix, $\boldsymbol{Z}$ is the standardized version of $\boldsymbol{X}$ and I_h is the indicator for a particular slice h. To estimate $\mathrm{Var}(\boldsymbol{Z}|Y \in I_h)$ we can use an empirical procedure similar to Algorithm 1.1. Other alternatives to estimate the conditional variance can be found in Ruppert et al. (1997), Fan (1998) or Pérez-González et al. (2010).

In a further discussion, Li (1991b) generalized the SAVE estimator using that $\text{Cov}(\boldsymbol{Z}) = \text{Cov}(\mathbb{E}[\boldsymbol{Z}|Y]) + \mathbb{E}[\text{Cov}(\boldsymbol{Z}|Y)]$. In particular, he defined the SIRII method which is represented by the matrix

$$\text{SIRII} = \mathbb{E}[(\text{Cov}(\boldsymbol{Z}|Y) - \mathbb{E}[\text{Cov}(\boldsymbol{Z}|Y)])^2]$$

Härdle and Tsybakov (1991) criticized the estimator of the weighted covariance matrix in step 4 of the Algorithm 1.1. Instead, they considered the estimation of each conditional expectation $R_i(Y) = \mathbb{E}[X_i|Y]$ by nonparametric methods. In other words, the coefficient (i, j) of $\text{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ can be estimated by

$$\frac{1}{n}\sum_{k=1}^{n} \hat{R}_i(Y_k)\hat{R}_j(Y_k). \tag{1.5}$$

The functions $\hat{R}_i$, $\hat{R}_j$ may be kernel, orthogonal series (e.g., splines, wavelets), or any other estimates. If $\hat{R}$ is a regressogram, then we get an estimator similar to the sliced inverse regression. Zhu and Fang (1996) showed the asymptotic normality for $\text{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ and its eigenvalues when $\hat{R}$ is estimated by the Nadaraya-Watson estimator. Later, Zhu and Yu (2007) proved the exact normality behavior as Zhu and Fang, but estimating $\hat{R}$ by a B-splines series. Yin et al. (2010) proposed a recent study on nonparametric conditional covariance estimation.

Li (1992) presented an application of the Stein lemma (Stein (1981)) to find the effective dimension reduction space named principal Hessian directions (pHd). He used the average Hessian of $\mathbb{E}[Y|\boldsymbol{X}]$ and its eigenvectors to create a new reduced coordinate system. To construct this new space, he applied the Stein's Lemma to find a root-n consistency.

Hsing (1999) estimated $\text{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ by

$$\frac{1}{2n}\sum_{k=1}^{n} \boldsymbol{Z}_k\boldsymbol{Z}_{k*}^\top + \boldsymbol{Z}_{k*}\boldsymbol{Z}_k^\top$$

where $\boldsymbol{Z}_{k*}$ is the nearest neighbor of $\boldsymbol{Z}_k$. Hsing proved the root-n rate of convergence of the new estimator. Cadre and Dong (2010) present a modern reference on the estimation of the central subspaces for dimension reduction using nearest neighbors.

Bura and Cook (2001) estimated the sliced inverse regression through an additive parametric form with q elements. Formally, they fixed $\mathbb{E}[X|Y]$ to have

the form $\boldsymbol{F_nB} + \boldsymbol{\varepsilon_n}$, where $\boldsymbol{F}_n$ is an $n \times q$ matrix of parametric fixed functions, $\boldsymbol{B}$ is a $q \times p$ coefficient matrix and $\boldsymbol{\varepsilon_n}$ is an error matrix.

Ferré and Yao (2003); Ferré et al. (2005) worked with the functional version of the sliced inverse regression. They assumed a set of curves $X(t)$ as input with some real value Y as output. They rewrote model (1.4) in the following way

$$Y = \phi\left(\int v_1(t)X(t)\,dt, \ldots, \int v_K(t)X(t)\,dt, \varepsilon\right).$$

where $v_1(t), \ldots, v_K(t)$ are squared integrable functions in some interval. Using a kernel estimate, they obtained similar results as Zhu and Fang (1996).

Another useful application of the sliced inverse regression is in graphics displays. Cook (2003, 1998) studied the use of the sliced inverse regression to create representative plots of the data giving novel ideas in the subject. Other techniques in sliced inverse regression include: Cook and Ni (2005), Zhu et al. (2007), Yoo and Cook (2007) Yoo (2008a,b) use ordinary least squares to detect the central mean subspace in different settings or Loubes and Yao (2013) studied the sliced inverse regression under a framework of strong mixing conditions. Also, other investigations have extended the sliced inverse regression to accommodate multiple outcomes. For instance Cook (2003) slices the bivariate outcome into hypercubes or Setodji and Cook (2004) use k-means clustering. For further disscussion on the classical sliced inverse regression method, we refer to Brillinger (1991), Cook and Weisberg (1991), Härdle and Tsybakov (1991), Kent (1991), Li (1991b), and references therein.

We have investigated the implications of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ to solve high-dimensional problems. However the study of the conditional covariance has other applications. For example, Duffée (2005) modeled the relations between aggregate stocks returns and aggregate consumption growth via conditional covariance matrices. Another modern application is the sensitivity analysis where we have a model with numerous inputs and one output. The researcher wants to know how much the variation of one input affects the rest of the model. In particular, we can model this relation by

$$\mathrm{Var}(\text{Input} \mid \text{Output}).$$

Therefore, in the next section we will explore estimation of conditional variances for sensitivity analysis.

1.5 Sensitivity analysis

Models are built to approximate or mimic process and systems of different nature, e.g., physics, economics, chemistry, etc. The mathematical formulation of them, receives a series of equations, parameters, input factors and variables to generate an output similar to the real process. Some sources of uncertainty affect the input variables, such as measurement errors, lack of information or poor or partial understanding of the system mechanism.

To handle complex models, uncertainty analysis and sensitivity analysis are two essential tools. In one hand, uncertainty analysis refers to the determination of the uncertainty in the outputs that derives from uncertainty in the inputs. In other hand, sensitivity analysis studies the effects of the variations in the inputs on the calculated output. Terms such as influence, importance, ranking and dominance are all related to sensitivity analysis. Ideally, both analysis should be run in tandem.

The model complexity is independent of the model size. A small model might have complex interactions hardening its analysis. Instead, Simon (1969) describes the complexity by the numbers of hierarchies, the "span" of each level in the hierarchy and the number of levels.

Indeed, we can see the sensitivity analysis as a kind of reduction dimension method. While the reduction dimension methods aims to *"summarize"* the complete data space into a lower one, the sensitivity analysis tries to identify the most relevant inputs in the model according to some score function. The larger is the score function for some variable, the larger will be the influence of that variable into the model. Below we will review some scores used to identify relevant variable in sensitivity analysis. In any case, once we have the new set of relevant variables, we can proceed to do further studies like classification, regression, estimation and so on.

Saltelli et al. (2004) and later Pappenberger et al. (2010), emphasize the importance of specifying the objectives before to apply any algorithm or technique. In this way, the scientist knows a priori what it is searching and what kind of result he wants. In this way, the most important objectives are

- identify and select the most influent inputs, and identify the nonrelevant ones to set them constant,
- map the output behavior with respect to the inputs and focus the analysis

in some interests zones,

- adjust the model variables taking in account the information of the most important inputs.

Commonly, models in sensitivity analysis ignore the stochastic error. In this framework, we will assume a nonparametric model similar to equation (1.1), but with ε equal to 0. Therefore, for a set of variables $(X_1, \ldots, X_p) \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ we have,

$$Y = m(X_1, \ldots, X_p). \tag{1.6}$$

Also, assume that each X has a valid range of variation and Y is non-null. The explicit form of m is unavailable or simply is hard to obtain. However, the analyst can create a computer code to describe some phenomenon, including two-phase flow, multi-mode heat transfer, clad oxidation chemistry, stress and strain, and reactor kinetics. For example, the work of Janon (2012) studies the behavior of fluids in oceanographic and hydrologic models. Parameters like viscosity, friction, wind speed, initial height among others, make impossible have an explicit equation for the function m. For other illustrations on the use of computer codes, we refer to Oakley and O'Hagan (2004), Langewisch (2010) and references therein.

To solve model (1.6) there exist methods such as: one-at-a-time algorithms, differential analysis, response surface methodology, Monte Carlo procedures and variance decomposition procedures. Overview of these approaches are available in general literature, for instance Saltelli et al. (2000), Christopher Frey and Patil (2002), Helton et al. (2006), Saltelli et al. (2008).

Screening method: These methods are a low computational way to discard factors in the analysis. For example, we can evaluate the model output by varying only one variable across its entire range, while fixing the others in their nominal values. Thus, we measure the difference between the nominal and changed outputs to extract the most relevant factors. Even if the technique is easy to implement, it provides only a limited quantity of information. For example, we cannot recover the crossed interactions with this method. For a general review in screening we refer to Cullen and Frey (1999), Campolongo et al. (2011) and Saltelli et al. (2009)

Automatic differentiation: If we derivate the output variable with respect to each factor, we will get a local measure of the model for every variable. In

general, we want to estimate the following normalized partial derivative,

$$\frac{\sigma_{X_i}\partial Y}{\sigma_Y \partial X_i} \tag{1.7}$$

where σ_{X_i} and σ_Y are variances of X_i and Y respectively. When the model is complex, the estimation of the quantity (1.7) turns hard. Some methods to estimate numerically those partial derivates from a model are available in Rall (1980), Kedem (1980) and Carmichael et al. (1997).

Regression analysis: Recall the classical linear regression,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

We can interpret the regression coefficients β_i as the change in the output when the in input X_i increase or decreases by one unit (Devore and Peck (1996)). This interpretation is valid if the other factors remain constant. Therefore, we can use the regression coefficient as nominal measures of the sensitivity in the model. Methods like stepwise regression serve to exclude statistical insignificant inputs. Moreover, we can measure the distance of a linear model and the true using the coefficient of determination or R^2. This coefficient determines the percentage of variance in the Y explained by the linear model (see Draper and Smith (1981)). For example, if R^2 is equal to 0.9, then the model is 90% linear and one could use the β's for sensitivity analysis at risk of lose 10% of information.

Response surface method: The response surface uses least squared regression to fit a standardized first- or second-order equation to the data obtained from the original model. The amount of time and effort required is as big as the number of inputs. For this reason it is recommended to use it only in the latest steps of the sensitivity investigation. Some general references in the area are Myers et al. (2009) and Goos (2002).

Variance decomposition methods: Fix the variable X_i for some $i = 1, \ldots, p$. Recall that $\mathbb{E}[Y|X_i]$ represents the best approximation (in $\mathbb{L}^2$) of Y given all the knowledge of X_i. The variance of $\mathbb{E}[Y|X_i]$, with respect X_i, quantifies the dispersion of the best approximation of Y. Therefore, when the value $\mathrm{Var}(\mathbb{E}[Y|X_i])$ is large, the variable X_i *"influences"* more respect to the output Y. In other words, the variation of Y due to the variation of X_i is large. Normalizing this conditional variance by the total variance of Y, we obtain the first order

Sobol index associated to X_i,

$$S_i = \frac{\mathrm{Var}(\mathbb{E}[Y|X_i])}{\mathrm{Var}(Y)}.$$

Sobol' (1993)–inspired by the prior work of Cukier et al. (1978)–initially studied the estimation of S_i while studying functional decomposition via Monte-Carlo procedures.

We are mainly interested in the variance decomposition methodology applied to sensitivity analysis. In fact, in the next section we will present some techniques used to estimate the value S_i. Additionally, our knowledge in conditional covariances, presented in Section 1.4.2, will serve us as a tool for the estimation of S_i.

1.5.1 Estimation of Sobol indices

Recall, from the previous Section, the first order Sobol indices of Y associated to the variables X_i ($i = 1, \ldots, p$) are

$$S_i = \frac{\mathrm{Var}(\mathbb{E}[Y|X_i])}{\mathrm{Var}(Y)} = \frac{\mathbb{E}[\mathbb{E}[Y|X_i]^2] - \mathbb{E}[Y]}{\mathrm{Var}(Y)}. \tag{1.8}$$

Given the normalization by $\mathrm{Var}(Y)$, each S_i belongs to the interval $[0, 1]$. The Sobol indices measure the relevance of each factor and with respect the output. They represent the percent of the variability of Y when we alter each variable. Thus, S_i reveals the sensitivity of X_i with respect to Y. We can define indices taking account the interactions between the variables. Theoretically, the sum of all Sobol indices—including interactions—adds up to 1.

Sobol' (1993) showed that we can decompose any function into terms of increasing dimension. We call this kind of decomposition as High-Dimensional Model Representation (HDMR). Li et al. (2001) presents, for example, a complete review about HDMRs. Sobol proved that if each term of the representation has mean zero and all the term are orthogonal in pairs, then we have the following equation

$$Y = \sum_i V_i + \sum_{ij} V_{ij} + \sum_{ijk} V_{ijk} + \cdots + V_{1,\ldots,p}.$$

where

$$V_i = \mathbb{E}[Y|X_i], \quad V_{ij} = \mathbb{E}[Y|X_iX_j] - V_i - V_j, \quad \ldots$$

In our context, if we take variances in both sides of the HDMR equation, we define the higher sensitivity Sobol indices by

$$S_i = \frac{\text{Var}(\mathbb{E}(Y|X_i))}{\text{Var}(Y)} = \frac{V_i}{\text{Var}(Y)}, \quad S_{ij} = \frac{V_{ij}}{\text{Var}(Y)}, \quad S_{ijk} = \frac{V_{ijk}}{\text{Var}(Y)}, \quad \cdots$$

Finally, we can also define another type of index called *Total effects*. This index accounts the total contribution to the output variation due to factor X_i, i.e., its first-order effect plus all higher-order effects due to interactions. Define the variable $\boldsymbol{X}_{\sim i}$ as $\boldsymbol{X}$ with the i^{th} factor removed. The total effect index for the variable X_i is defined by

$$S_{T_i} = 1 - \frac{\text{Var}(\mathbb{E}[Y|\boldsymbol{X}_{\sim i}])}{\text{Var}(Y)}.$$

The aim of this thesis will be the estimation of first-order Sobol indices. To achieve a good estimate of S_i, it is necessary to estimate $\text{Var}(\mathbb{E}[Y|X_i])$. The most difficult part in equation (1.8) is the term $\mathbb{E}[\mathbb{E}[Y|X_i]^2]$. Indeed, the output of the model m could be complex and the integrals for the conditional variance could be intractable analytically. The Monte-Carlo estimation for integrals (Hammersley (1960)) seems to be the simplest but most expensive method to estimate S_i. We start by selecting a set of random points to estimate the inner conditional expectation for a fixed value of X_i. Then, for each of these values make another sample of values to estimate the outer variance. We see that as the number of points taken for the evaluation, and the number of samples increase, the computational complexity of the algorithm increases as well. Hence, numerous studies have developed numerical approximations suitable to the practical needs in the applications.

The challenge in this context is given a finite sample, construct a computational effective estimator of S_i. Ishigami and Homma (1990) studied one of the first solutions in this path. They reduced the computational complexity to only one Monte-Carlo loop. The idea was to rewrite the Sobol indices S_i's by resampling $\boldsymbol{X}$ and by creating a two-fold function instead of the original. This technique costs only $2p - 1$ calculations. Later, Saltelli (2002), found that with $n(2p + 2)$ calculations it could be possible to estimate $p - 2$ Sobol indices.

The Fourier Amplitude Sensitivity Test (FAST) is another classic method. Cukier et al. (1973) and Cukier et al. (1978) created the FAST which was originally used to analyze sensitivity in nonlinear rate equations. The method was

developed further by Koda et al. (1979), McRae et al. (1982) and Saltelli et al. (1999). The aim of FAST is to transform a multidimensional integral over all the factors to a one-dimensional integral. This approach is done using a Fourier expansion, i.e., we redefine each coefficient as following,

$$X_i = G_i(\sin(w_i t))$$

where G_i are suitable transformation functions, $\{w_i\}$ is a set of integer angular frequencies and $t \in (-\pi, \pi)$. Some recent investigations with this method can be found for example in Tarantola et al. (2006) and Tissot and Prieur (2012).

The Sobol pick-freeze (SPF) scheme is widely used and was proposed by Sobol' (1993, 2001). Rewrite the equation (1.8) in the following way,

$$S_i = \frac{\mathrm{Cov}(Y, Y')}{\mathrm{Var}(Y)}$$

where

$$Y' = m(X'_1, X'_1, \ldots, X'_{i-1}, X_i, X'_{i+1}, \ldots, X'_p)$$

and X'_j for $j \neq i$ correspond to a random variable independent with the same distribution as X_j. This scheme freezes the variable of interest while resampling the other variables to measure the sensitivity in the former. In SPF we can interpret the Sobol index as the regression coefficient between the output and its pick-freezed replication. As a Monte-Carlo-based method, SPF requires thousands of machine computations to get one index. And it gets harder if the model is created by a complex computer code (e.g., numerical partial differential equations). To workaround the computational complexity some authors use *meta-models* which are—low-fidelity—approximations to the true model. Other implementations of the SPF algorithm are proposed in Janon et al. (2013).

Alternatively, Da Veiga and Gamboa (2013) explored another estimator for Sobol indices via a Taylor decomposition. As before, the quantity $\mathbb{E}[\mathbb{E}[Y|X_i]^2]$ represents the hardest part in equation (1.8). They see the conditional expectation as functional depending on the joint distribution f of $(\boldsymbol{X}, Y)$. Specifically, we have the representation

$$\mathbb{E}[\mathbb{E}[Y|X_i]^2] = \int \left(\frac{\int x_i f(x_i, y)\, dy}{\int f(x_i, y)\, dx_i\, dy} \right)^2 dy.$$

Using a third order Taylor series development around a preliminary estimator of f called $\hat{f}$, the latter functional turns into,

$$\int H(\hat{f}, x, y) f(x, y)\, dx\, dy + \int K(\hat{f}, x, y, z) f(x, y) f(x, z)\, dx\, dy\, dz + \Gamma_n.$$

where H and K are functions depending on $\hat{f}$ and Γ_n is an error term.

The Taylor-based estimator has two elements: An empirical estimator for the first addend; and a projection to some functional orthonormal basis for the second one. This type of projections were studied for instance by Laurent (1996) in the estimation of integral functionals. They proved that its estimator is asymptotical normal with variance depending only in the linear part of the functional and that it is efficient from the Cramér-Rao point of view.

Further details in sensitivity analysis are in Saltelli et al. (2000), Saltelli et al. (2004) and Saltelli et al. (2008).

1.6 Thesis scope and outline

In Section 1.3, we discussed how the matrix $\Sigma = \mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ works into the sliced inverse regression method. Once we have an estimator for Σ, the eigenvectors associated to the largest eigenvalues of Σ provides an efficient dimension reduction space. In an apparently unrelated issue, we explore, in Section 1.5.1, techniques to estimate sensitivity indices through the value $S_i = \mathrm{Var}(Y|X_i)/\mathrm{Var}(Y)$. Both issues identify or extract the real influences of the variables into the model and both are based in the estimation of some conditional variance-covariance.

Our work falls in the following chapters:

Chapter 2. Chapter 2 present an Taylor-based estimator inspired by the work of Da Veiga and Gamboa (2013). Motivated by application to sliced inverse regression, we will estimate

$$\sigma_{ij} = \mathbb{E}[\mathbb{E}[X_i|Y]\mathbb{E}[X_j|Y]] \quad \text{for} \quad i,j = 1,\ldots,p$$

Notice that this estimator generalizes the Da Veiga and Gamboas's procedure to three dimensions.

The strategy adopted in this chapter is the following: Denote as $f(x_i, x_j, y)$ the density function of (X_i, X_j, Y) and $f_Y(y)$ the marginal density of Y. We can rewrite σ_{ij} as

$$\sigma_{ij} = \int \left(\frac{\int x_i f(x_i, x_j, y)\, dx_i\, dx_j}{f_Y(y)} \right)$$

$$\left(\frac{\int x_j f(x_i, x_j, y)\, dx_i\, dx_j}{f_Y(y)} \right) f(x_i, x_j, y)\, dx_i\, dx_j\, dy.$$

We will develop the estimator $\widehat{\Sigma}_T = (\hat{\sigma}_{ij,T})_{p\times p}$ where $\hat{\sigma}_{ij,T}$ is based in the Taylor approximation explained in Section 1.5.1. The asymptotic convergence of the quadratic part and the error term are negligible with respect to the linear part. For this reason the variance in the normal distribution is built only by the linear term.

Summing up, we will prove that the random variable $\sqrt{n}\ (\hat{\sigma}_{ij,T} - \sigma_{ij})$ converges to a normal distribution, with mean 0 and variance depending on the linear part of $T_{ij}(f)$, when $n \to \infty$.

Moreover we will show that our estimator is asymptotic efficient using a Cramér Rao criterion. Finally, we can extend these results to the whole matrix Σ proving that

$$\sqrt{n}\,\mathrm{vech}(\widehat{\Sigma}_T - \Sigma) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{C}(f))$$

where $\boldsymbol{C} = (C_{ij})_{p\times p}$ depends on the linear part of $\boldsymbol{T}(f) = (T_{ij}(f))_{p\times p}$.

Chapter 3. The main issue with the estimator Taylor-based estimator $\widehat{\Sigma}_T$ in Chapter 2 is the lack of flexibility to estimate rates of convergence in some general cases. Thus, in Chapter 3 we propose a nonparametric estimator to handle this issue. If the joint-density function of $(\boldsymbol{X}, Y)$ is smooth enough, we can achieve parametric rates of convergence. We will use the ideas presented by Zhu and Fang (1996).

We will define an estimator of the form (1.5) with $\hat{R}_i$, $\hat{R}_j$ being the Nadaraya-Watson of $\mathbb{E}[\boldsymbol{X}|Y]$. We will call it $\widehat{\Sigma}_K$. In this work, we will find rates of convergence for $\widehat{\Sigma}_K$ depending on the smoothness of the joint distribution f. We will set that f belong to some Hölder functional class with parameters $\beta > 0$ and radius $L > 0$.

The main result in Chapter 3 establishes that under some mild conditions and $\beta \geq 2$, then $\mathbb{E}[(\sigma_{ij,K} - \sigma_{ij})^2]$ decreases at rate $1/n$. In other words, with enough regularity in the model, we can achieve a parametric rate of convergence with a nonparametric estimator. Otherwise, if $\beta < 2$, we will get a nonparametric rate.

If $p \gg n$, then the estimator $\widehat{\Sigma}_K = (\hat{\sigma}_{ij,K})_{ij}$ loses its performance. As we explain in Section 1.4.1, as the dimension grows, we lose the estimator convergence. To recover it, we need to regularize the matrix $\widehat{\Sigma}_K$ assuming

certain structure in the matrix Σ. We will apply the solution giving by Bickel and Levina (2008b) with a banding method. We shall use the Frobenuis norm under some regular covariance class. Moreover, the rates will depend on the dimensionality and the smoothness of the joint density function $f(x, y)$ and they will be coherent with the work of Cai et al. (2010). If $p \ll n$, we will use the original covariance matrix $\widehat{\Sigma}_K$ given the low-dimension setting. Otherwise, it is necessary to regularize the estimator to conserve the convergence.

Chapter 4. Finally, we will adjust the nonparametric estimator for conditional variances, developed in Chapter 3, to study the quantity $\mathbb{E}[\mathbb{E}[Y|X_i]]$ for $i = 1, \ldots, p$. We have seen how the latter expectation is related to the estimation of the Sobol index S_i presented in equation (1.8). Assuming at least that the joint density of (X_i, Y) is twice times differentiable, we show that the estimator of S_i converges at a parametric rate to S_i. Otherwise, for densities with less than two derivatives, we obtain a nonparametric rate of convergence depending on the regularity. Our method avoids the run computational expensive Monte-Carlo simulations because it estimates each Sobol index directly with the inherent structure of the data. Furthermore, we present some numerical results with popular sensitivity analysis test models, like the Ishigami function. Our numerical simulations identify correctly the relevant and irrelevant factors of the models.

Chapter 2

Efficient estimation of conditional covariance matrices for dimension reduction

joint work with S. Da Veiga[1] and J-M. Loubes[2].

Abstract: Let $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. In this chapter we estimate the conditional covariance matrix $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ in an inverse regression setting. We develop a functional Taylor expansion of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ under some mild conditions. This methodology provides a new efficient estimator from the Cramer-Rao point of view. Also, we study its asymptotic properties.

2.1 Introduction

Consider the nonparametric regression

$$Y_k = \varphi(\boldsymbol{X}_k) + \epsilon_k \quad k = 1, \ldots, n,$$

where $\boldsymbol{X}_k = (X_{1k}, \ldots, X_{pk}) \in \mathbb{R}^p$, $Y_k \in \mathbb{R}$ and ϵ_k are random noises with $\mathbb{E}[\epsilon_k] = 0$. If we face a model with more variables than observed data (i.e., $p \gg n$); the high-dimensional setting blurs the relation between X and Y, unless

[1]Institut Français du Pétrole, Paris, France.
[2]Institut de Mathématiques de Toulouse, Université Paul Sabatier, Toulouse, France.

we have at hand a large sample. The literature calls this phenomenon the *curse of dimensionality*.

Many authors have studied the dimensionality problem. For instance, the generalized linear model in Brillinger (1983), the additive models in Hastie and Tibshirani (1990), sparsity constraint models in Li (2007) and references therein.

Alternatively, Li (1991a) propose the sliced inverse regression method. He considers the semiparametric model,

$$Y_k = \phi(v_1^\top X_k, \ldots, v_K^\top X_k, \epsilon_k) \tag{2.1}$$

where the v's are unknown vectors in $\mathbb{R}^p$, the ϵ_k's are independent of X_k and ϕ is an arbitrary function in $\mathbb{R}^{K+1}$. This model gathers all the relevant information about the variable Y, with only a projection onto the $K \ll p$ dimensional subspace $(v_1^\top X, \ldots, v_K^\top X)$. If K is small, the method reduces the dimension by estimating the v's efficiently. We call the v's effective dimension reduction directions. This method is also used to search nonlinear structures in data and to estimate the projection directions v's.

For a review on the sliced inverse regression methods, we refer to Li (1991a), Li (1991b), Duan and Li (1991) Hardle and Tsybakov (1991) and references therein. In short, the eigenvectors associated with the largest eigenvalues of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ are the model (2.1) effective dimension reduction directions.

In this context, it is enough to estimate $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ to find the effective dimension reduction directions. Some previous works include: Zhu and Fang (1996) and Ferré and Yao (2003) and Ferré et al. (2005) use kernel estimators; Hsing (1999) combines nearest neighbor and the sliced inverse regression; Bura and Cook (2001) assume that $\mathbb{E}[X|Y]$ has some parametric form; Setodji and Cook (2004) use k-means and Cook and Ni (2005) write the sliced inverse regression to least square form.

We propose an alternate estimation of the matrix

$$\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y]) = \mathbb{E}[\mathbb{E}[\boldsymbol{X}|Y]\mathbb{E}[\boldsymbol{X}|Y]^\top] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\top,$$

using the ideas of Da Veiga and Gamboa (2013), inspired by the prior work of Laurent (1996). Since we can compute $\mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\top$ easily, we will focus on the $\mathbb{E}[\mathbb{E}[\boldsymbol{X}|Y]\mathbb{E}[\boldsymbol{X}|Y]^\top]$ estimation.

We will present a quadratic functional estimator of $\mathbb{E}[\mathbb{E}[\boldsymbol{X}|Y]\mathbb{E}[\boldsymbol{X}|Y]^\top]$ via a coordinate-wise Taylor development. Given a preliminary approximation of

the (X_i, X_j, Y)'s density, we will make a Taylor's expansion up to thrice order of $\mathbb{E}[\mathbb{E}[\boldsymbol{X}|Y]\mathbb{E}[\boldsymbol{X}|Y]^\top]$. The first order term drives the asymptotic convergence meanwhile the other ones rest negligible. We will also prove this kind of convergence for the whole matrix. Besides, we shall prove also its efficiency in a semiparametric framework using a Cramer-Rao criterion.

Instead of compete with the methods mentioned before, we intent to make it complementary. Indeed, we offer an alternative method on plug-in methods for conditional covariance matrices with minimum variance properties.

We organize this chapter as follows. Section 2.2 motivates our investigation of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ using a Taylor approximation. In Section 2.3.1 we set up all the notations and hypotheses. We demonstrate, in Section 2.3.2, the efficient convergence of each coordinate for our estimator. Also, we state the asymptotic normality for the whole matrix. For the quadratic term of the Taylor's expansion of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$, we find an asymptotic bound for the variance in Section 2.4. We postpone all the technical Lemmas and related proofs to Sections 2.6.2 and 2.6.1 respectively.

2.2 Methodology

Let $\boldsymbol{X} \in \mathbb{R}^p$ be a squared integrable random vector with $p \geq 1$ and $Y \in \mathbb{R}$ be a random variable. Denote as X_i, X_j the i-th and j-th coordinates of $\boldsymbol{X}$, respectively for i and j different. We denote by $f_{ij}(x_i, x_j, y)$ the joint density of the vector (X_i, X_j, Y) for $i, j = 1 \dots p$. Remark that the density function f_{ij} depends on the indices i and j, namely for each triplet (X_i, X_j, Y) there exist a joint density function called $f_{ij}(x_i, x_j, y)$. For the sake of simplicity, we will denote f_{ij} only by f to avoid cumbersome notations. When i is equal to j, we will call to the joint density of (X_i, Y) by $f_i(x_i, y)$, for $i = 1, \dots, p$. When the context is clear, we can name f_i simply by f. Finally, let $f_Y(\cdot) = \int_{\mathbb{R}} f(x_i, x_j, \cdot)\, dx_i\, dx_j$ be the marginal density function with respect to Y

The studies of Laurent (1996) and Da Veiga and Gamboa (2013) have already considered the estimation of the diagonal of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$. According to the paragraph precedent, they assumed a context where the indices i and j are equal. The former studied the estimation of density integrals in the unidimensional case and the latter applied this estimation into sensibility analysis for computing Sobol indices. In this work, we shall extend their methodologies to the case i different of j to find an alternative estimator for the sliced inverse regression directions.

Recall that

$$\Sigma = \mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y]) = \mathbb{E}[\mathbb{E}[\boldsymbol{X}|Y]\mathbb{E}[\boldsymbol{X}|Y]^\top] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\top,$$

where $A^\top$ means the transpose of A. We can easily estimate $\mathbb{E}[\boldsymbol{X}]$ with the empirical sample mean. Also, without loss of generality, we always can assume the variables $\boldsymbol{X}$ centered, i.e., $\mathbb{E}[\boldsymbol{X}] = 0$.

Define each entry of the conditional covariance matrix Σ as

$$\sigma_{ij} = \mathbb{E}[\mathbb{E}[X_i|Y]\mathbb{E}[X_j|Y]] \quad i, j = 1, \ldots, p.$$

Given a sample of $(\boldsymbol{X}, Y)$, we aim to study the asymptotic and efficiency properties of σ_{ij}. Also, we will find similar results for the whole matrix Σ.

Notice that we can write each σ_{ij} for $i \neq j$ as

$$\sigma_{ij} = \int \left(\frac{\int x_i f(x_i, x_j, y)\, dx_i\, dx_j}{f_Y(y)} \right) \left(\frac{\int x_j f(x_i, x_j, y)\, dx_i\, dx_j}{f_Y(y)} \right) f(x_i, x_j, y)\, dx_i\, dx_j\, dy. \quad (2.2)$$

We will need the following functional operator.

Definition 2.1. Let ψ be a square-integrable function in $\mathbb{L}^2(dx_i\, dx_j, dy)$. Define the functional mapping $\psi \mapsto T_{ij}(\psi)$ where

$$T_{ij}(\psi) = \int \left(\frac{\int x_i \psi(x_i, x_j, y)\, dx_i\, dx_j}{\int \psi(x_i, x_j, y)\, dx_i\, dx_j} \right) \left(\frac{\int x_j \psi(x_i, x_j, y)\, dx_i\, dx_j}{\int \psi(x_i, x_j, y)\, dx_i\, dx_j} \right) \psi(x_i, x_j, y)\, dx_i\, dx_j\, dy. \quad (2.3)$$

A word to clarify the notations it is necessary at this point. The functional $T_{ij}(\psi)$ takes any squared-integrable ψ an estimates the value described in equation (2.3). If we take specifically $\psi = f$, then the functional $T_{ij}(f)$ is equal to the parameter σ_{ij} defined in 2.2. We will apply a Taylor development to $T_{ij}(f)$ to obtain an estimator for σ_{ij}. Thus, we should be able of transfer all the properties of $T_{ij}(f)$ to σ_{ij}.

Suppose that (X_{ik}, X_{jk}, Y_k), $k = 1, \ldots, n$ is an independent and identically distributed sample of (X_i, X_j, Y). Assume that $\hat{f}$ is a preliminary estimator of f estimated with a subsample of size $n_1 < n$. The main idea is to expand $T_{ij}(f)$ in a Taylor's series around a neighborhood of $\hat{f}$.

More precisely, define an auxiliar function $F : [0,1] \to \mathbb{R}$;

$$F(u) = T_{ij}(uf + (1-u)\hat{f})$$

with $u \in [0,1]$. The Taylor's expansion of F between 0 and 1 up to the third order is

$$F(1) = F(0) + F'(0) + \frac{1}{2}F''(0) + \frac{1}{6}F'''(\xi)(1-\xi)^3 \tag{2.4}$$

for some $\xi \in [0,1]$. Moreover, we have

$$F(1) = T_{ij}(f)$$
$$F(0) = T_{ij}(\hat{f})$$

To simplify the notations set

$$m_i(f_u, y) = \frac{\int x_i f_u(x_i, x_j, y)\, dx_i\, dx_j}{\int f_u(x_i, x_j, y)\, dx_i\, dx_j},$$

where $f_u = uf + (1-u)\hat{f}$, for all u belonging to $[0,1]$. Notice that if $u = 0$ then $m_i(f_0, y) = m_i(\hat{f}, y)$.

We can rewrite $F(u)$ as

$$F(u) = \int m_i(f_u, y) m_j(f_u, y) f_u(x_i, x_j, y)\, dx_i\, dx_j\, dy.$$

The next Proposition gives the $T_{ij}(f)$ Taylor's expansion.

Proposition 1 (Linearization of the operator T)**.** *For the functional $T_{ij}(f)$ defined in* (2.3)*, the following decomposition holds*

$$\begin{aligned} T_{ij}(f) = &\int H_1(\hat{f}, x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy \\ &+ \int H_2(\hat{f}, x_{i1}, x_{j2}, y) f(x_{i1}, x_{j1}, y) f(x_{i2}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy + \Gamma_n \end{aligned} \tag{2.5}$$

where

$$H_1(\hat{f}, x_i, x_j, y) = x_i m_j(\hat{f}, y) + x_j m_i(\hat{f}, y) - m_i(\hat{f}, y) m_j(\hat{f}, y)$$

$$H_2(\hat{f}, x_{i1}, x_{j2}, y) = \frac{1}{\int \hat{f}(x_i, x_j, y)\, dx_i\, dx_j}(x_{i1} - m_i(\hat{f}, y))(x_{j2} - m_j(\hat{f}, y))$$
$$\Gamma_n = \frac{1}{6}F'''(\xi)(1-\xi)^3, \tag{2.6}$$

for some $\xi \in]0,1[$.

This decomposition separates $T_{ij}(f)$ in three parts. A linear functional of f—which is easily estimable—, a quadratic one and an error term Γ_n. In Section 2.3.2, we prove that the term H_1 drives the asymptotic convergence of $T_{ij}(f)$. Besides, Theorem 2.2 gives the semiparametric efficiency. We control the quadratic functional variance in Section 2.4.

2.3 Main Results

In this section we estimate σ_{ij} efficiently employing decomposition (2.5). Since we used $n_1 < n$ to build a preliminary approximation $\hat{f}$, we will use a sample of size $n_2 = n - n_1$ to estimate σ_{ij}. Since the first term in equation (2.5) is a linear functional in f, then its empirical estimator is

$$\frac{1}{n_2}\sum_{k=1}^{n_2} H_1(\hat{f}, X_{ik}, X_{jk}, Y_k). \tag{2.7}$$

Conversely, the second addend complicates the estimate because it is a nonlinear functional of f. However, in Section 2.4 we will study deeply the general functional

$$\theta(f) = \int \eta(x_{i1}, x_{j2}, y) f(x_{i1}, x_{j1}, y) f(x_{i2}, x_{j2}, y) dx_{i1} dx_{j1} dx_{i2} dx_{j2} dy$$

where $\eta : \mathbb{R}^3 \to \mathbb{R}$ is a bounded function. The value $\hat{\theta}_n$ of $\theta(f)$ gives an approximation of the second term in (2.5). This technique extends the method developed by Da Veiga and Gamboa (2013).

2.3.1 Hypothesis and Assumptions

Throughout the chapter, we will use the following notations. Let a_k and b_k for $k = 1, 2, 3$ be real numbers where $a_k < b_k$. Let, for i and j fixed, $\mathbb{L}^2(\, dx_i\, dx_j\, dy)$ be the squared integrable functions in the cube $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$. Moreover,

let $(p_l(x_i, x_j, y))_{l\in D}$ be an orthonormal basis of $\mathbb{L}^2(\,dx_i\,dx_j\,dy)$, where D is a subset of $\{1,\dots,p\}$. Let $a_l = \int p_l f$ denote the scalar product of f with p_l.

Furthermore, denote by $\mathbb{L}^2(\,dx_i\,dx_j)$ (resp. $\mathbb{L}^2(\,dy)$) the set of squared integrable functions in $[a_1,b_1]\times[a_2,b_2]$ (resp. $[a_3,b_3]$). If $\left(\alpha_{l_\alpha}(x_i,x_j)_{l_\alpha\in D_1}\right)$ (resp. $\left(\beta_{l_\beta}(y)_{l_\beta\in D_2}\right)$) is an orthonormal basis of $\mathbb{L}^2(\,dx_i\,dx_j)$ (resp. $\mathbb{L}^2(\,dy)$) then $p_l(x_i,x_j,y) = \alpha_{l_\alpha}(x_i,x_j)\beta_{l_\beta}(y)$ with $l=(l_\alpha,l_\beta)\in D_1\times D_2$.

We also use the following subset of $\mathbb{L}^2(\,dx_i\,dx_j\,dy)$

$$\mathcal{E} = \left\{ \sum_{l\in D} e_l p_l :\ (e_l)_{l\in D} \text{ is such that } \sum_{l\in D}\left|\frac{e_l}{c_l}\right|^2 < 1 \right\}$$

where $(c_l)_{l\in D}$ is a decreasing fixed sequence.

Moreover assume that (X_i, X_j, Y) have a bounded joint density f_{ij} on $[a_1,b_1]\times[a_2,b_2]\times[a_3,b_3]$ which lies in the ellipsoid $\mathcal{E}$.

Besides, $X_n \xrightarrow{\mathcal{D}} X$ (resp. $X_n \xrightarrow{\mathcal{P}} X$) denotes the convergence *in distribution* or *weak* convergence (resp. convergence *in probability*) of X_n to X. Additionally, we denote by $\operatorname{supp} f$ the support of f.

Let $(M_n)_{n\geq 1}$ denote a sequence of subsets of D. For each n there exists M_n such that $M_n \subset D$. Write by $|M_n|$ the cardinal of M_n. We shall make three main assumptions:

Assumption 2.1. *For all $n\geq 1$ there is a subset $M_n\subset D$ such that*

$$\sup_{l\notin M_n} |c_l|^2 \approx \sqrt{|M_n|}/n$$

($A_n\approx B$ means $\lambda_1 \leq A_n/B_n \leq \lambda_2$ for some positives constants λ_1 and λ_2). Moreover, for all $f\in\mathbb{L}^2(\,dx\,dy\,dz)$, $\int (S_{M_n}f - f)^2\,dx\,dy\,dz \to 0$ when $n\to\infty$, where $S_{M_n}f = \sum_{l\in M_n} a_l p_l$.

Assumption 2.2. *We assume that* $\operatorname{supp} f \subset [a_1,b_1]\times[a_2,b_2]\times[a_3,b_3]$ *and for all $(x,y,z)\in\operatorname{supp} f$, $0<\alpha\leq f(x,y,z)\leq\beta$ with $\alpha,\beta\in\mathbb{R}$.*

Assumption 2.3. *It is possible to find an estimator $\hat{f}$ of f built with $n_1\approx n/\log(n)$ observations, such that for $\epsilon>0$,*

$$\forall (x,y,z)\in\operatorname{supp} f,\ 0<\alpha-\epsilon\leq\hat{f}(x,y,z)\leq\beta+\epsilon$$

and,

$$\forall\, 2 \leq q \leq +\infty,\ \forall l \in \mathbb{N}^*,\ \mathbb{E}_f \|\hat{f} - f\|_q^l \leq C(q,l) n_1^{-l\lambda}$$

for some $\lambda > 1/6$ *and some constant* $C(q,l)$ *not depending on* f *belonging to the ellipsoid* $\mathcal{E}$*.*

Assumption 2.1 is necessary to bound the bias and variance of $\hat{\theta}_n$. It depends on n to obtain a good quadratic approximation of the density projection using only M_n coefficients. This condition relates the growing size of the set M_n with the decay rate of the sequence c_l.This behavior relates strongly the smoothness of the density function f with the size of the coefficients c_l.

For instance, we will use the Example 2 in Laurent (1996) and its notation. Assume that f belongs to some Hölder space with index greater than s. If the wavelet $\tilde{\psi}$ has regularity $r > s$, then $f \in \mathcal{E}$ where

$$\mathcal{E} = \left\{ \sum_{j\geq 0} \sum_{\lambda \in \Lambda_j} \alpha(\lambda) \tilde{\psi}_\lambda, \sum_{j\geq 0} \sum_{\lambda \in \Lambda_j} 2^{2js} |\alpha(\lambda)|^2 \leq 1 \right\}.$$

See Meyer and Salinger (1993) for further details.

Moreover, they show that if $s > p/4$ and

$$M_n = \left\{ \lambda \in \Lambda_j, j \leq j_0, 2^{j_0} = n^{2/(p+4s)} \right\}$$

then $\sup_{l \notin M_n} |c_l|^2 \approx \sqrt{|M_n|}/n$. Also, $|M_n|/n \to 0$ with

$$|M_n| \approx n^{2p/(d+4s)}, \quad \sup_{l \notin M_n} |c_l|^2 \approx 2^{-2j_0^2} = n^{-4s/(p+4s)}.$$

Assumption 2.2 and 2.3 establish that $\Gamma_n = O(1/n)$, i.e., the error term in (2.5) is negligible. In Assumption 2.3, the function $\hat{f}$ converges faster than some given rate. Of course, the existence of such an estimator deeply relies on the regularity of the true function.

On the one hand, a large literature exists to determine which estimator we have to choose to achieve a rate of convergence for some kind of functions. See for instance Tsybakov (2009). On the other hand, suppose we set a class of estimators and a rate of convergence. Another type of theory specify the kind of functions that we might approximate with these estimators and achieving

this convergence rate. Particularly, the maxiset theory and we refer to Autin et al. (2009) or Kerkyacharian and Picard (2002) for general articles.

The Nikol'skii functional class provides an example for our framework. For $x \in \mathbb{R}^p$, $s > 0$ and $L > 0$, we consider the class $\mathcal{H}_q(s, L)$ of Nikol'skii of functions $f \in \mathbb{L}^q(dx)$ with partials derivatives up to order $r = \lfloor s \rfloor$ inclusive. For each of these derivatives $f^{(r)}$, we assume that

$$\|f^{(r)}(\cdot + h) - f^{(r)}(\cdot)\|_q \leq L|h^{s-r}| \quad \forall h \in \mathbb{R}.$$

If $f \in \mathcal{H}_q(s, L)$ with $s > p/4$, then Assumption 2.3 is satisfied (see Ibragimov and Khas' minskii (1983, 1984)).

2.3.2 Efficient Estimation of σ_{ij}

We consider the following estimator of σ_{ij} adopting the decomposition of $T_{ij}(f)$ in equation (2.5) along equations (2.7) and (2.12),

$$\begin{aligned}\hat{\sigma}_{ij,T} = \frac{1}{n_2}\sum_{k=1}^{n_2} H_1\big(\hat{f}, X_{ik}, X_{jk}, Y_k\big) + \frac{1}{n_2(n_2-1)}\sum_{l\in M}\sum_{k\neq k'=1}^{n_2} p_l\big(X_{ik}, X_{jk}, Y_k\big) \\ \int p_l\big(x_i, x_j, Y_{k'}\big) H_3\big(\hat{f}, x_i, x_j, X_{ik'}, X_{jk'}, Y_{k'}\big)\, dx_i\, dx_j \\ - \frac{1}{n_2(n_2-1)}\sum_{l,l'\in M}\sum_{k\neq k'=1}^{n_2} p_l\big(X_{ik}, X_{jk}, Y_k\big) p_{l'}\big(X_{ik'}, X_{jk'}, Y_{k'}\big) \\ \int p_l\big(x_{i1}, x_{j1}, y\big) p_{l'}\big(x_{i2}, x_{j2}, y\big) H_2\big(\hat{f}, x_{i1}, x_{j2}, y\big)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2} dy. \quad (2.8)\end{aligned}$$

where $H_3(f, x_{i1}, x_{j1}, x_{i2}, x_{j2}, y) = H_2(f, x_{i1}, x_{j2}, y) + H_2(f, x_{i2}, x_{j1}, y)$ and $n_2 = n - n_1$. We remove the term Γ_n of (2.5) in equation (2.8) because we will prove that it is negligible compared to the others terms.

Notice that if we take for example $n_1 = n/\log n$, then $n_1/n \to 0$ and $n_2/n \to 1$ as $n \to \infty$. Thus, we will use n instead of n_2 (asymptotically speaking) from this point to simplify the notation.

The next theorem gives the asymptotic behavior of $\hat{\sigma}_{ij,T}$ for i and j.

Theorem 2.1. *Let Assumptions 2.1-2.3 hold and $|M_n|/n \to 0$ when $n \to \infty$. Then,*

$$\sqrt{n}\big(\hat{\sigma}_{ij,T} - \sigma_{ij}\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, C_{ij}(f)), \quad (2.9)$$

and

$$\lim_{n\to\infty} n\,\mathbb{E}[\hat{\sigma}_{ij,T} - \sigma_{ij}]^2 = C_{ij}(f), \tag{2.10}$$

where

$$C_{ij}(f) = \mathrm{Var}(H_1(f, X_i, X_j, Y))$$

The asymptotic variance of σ_{ij} depends only on $H_1(f, X_i, X_j, Y)$. In other words, the linear part of (2.5) controls the asymptotic normality of σ_{ij}. This property entails the natural efficiency of $\hat{\sigma}_{ij,T}$.

The next theorem produces the σ_{ij}'s semiparametric Cramér-Rao bound.

Theorem 2.2 (Semiparametric Cramér Rao bound)**.** *Consider the estimation of*

$$\sigma_{ij} = \mathbb{E}[\mathbb{E}[X_i|Y]\mathbb{E}[X_j|Y]^\top]$$

for a random vector (X_i, X_j, Y) with joint density $f \in \mathcal{E}$.

Let $f_0 \in \mathcal{E}$ be a density verifying the assumptions of Theorem 2.1. Then, for any estimator $\hat{\sigma}_{ij,T}$ of σ_{ij} and every family $\{\mathcal{V}_r(f_0)\}_{r>0}$ of neighborhoods of f_0 we have

$$\inf_{\{\mathcal{V}_r(f_0)\}_{r>0}} \liminf_{n\to\infty} \sup_{f\in\mathcal{V}_r(f_0)} n\,\mathbb{E}_f\left[\hat{\sigma}_{ij,T} - \sigma_{ij}\right]^2 \geq C_{ij}(f_0)$$

where $\mathcal{V}_r(f_0) = \{f : \|f - f_0\|_2 < r\}$ for $r > 0$.

Theorems 2.1 and 2.2 establish the asymptotic efficiency of the estimator $\hat{\sigma}_{ij,T}$ defined in (2.8).

We have proved asymptotic normality entry by entry of the matrix $\Sigma = (\sigma_{ij})_{p\times p}$ using the estimator $\widehat{\Sigma}_T = (\hat{\sigma}_{ij,T})_{p\times p}$ defined in equation (2.8). To extend the result to the whole matrix, we introduce the half-vectorization operator vech. This operator stacks only the columns from the principal diagonal of a square matrix downwards in a column vector. Formally, for a $p \times p$ matrix $A = (a_{ij})$,

$$\mathrm{vech}(A) = \left[a_{11}, \cdots, a_{p1}, a_{22}, \cdots, a_{p2}, \cdots, a_{33}, \cdots, a_{pp}\right]^\top.$$

Name $\boldsymbol{H_1(f)}$ the matrix with entries defined by $\left(H_1(f_{ij}, x_i, x_j, y)\right)_{i,j}$ if i different of j and $(H_1(f_i, x_i, x_i, y))_{i,i}$ when i is equal to j and $i, j = 1, \ldots, p$.

Corollary 2.1 generalizes our previous results.

Corollary 2.1. *Let Assumptions 2.1-2.3 hold and $|M_n|/n \to 0$ when $n \to \infty$. Then $\widehat{\boldsymbol{T}}_n$ has the following properties:*

$$\sqrt{n}\,\mathrm{vech}(\widehat{\Sigma}_T - \Sigma) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{C}(f)),$$
$$\lim_{n\to\infty} n\,\mathbb{E}\left[\mathrm{vech}\left(\widehat{\Sigma}_T - \Sigma\right)\mathrm{vech}\left(\widehat{\Sigma}_T - \Sigma\right)^\top\right] = \boldsymbol{C}(f)$$

where the limit is taking element-wise and

$$\boldsymbol{C}(f) = \mathrm{Cov}(\mathrm{vech}(\boldsymbol{H_1}(f))).$$

The estimator $\hat{\sigma}_{ij,T}$ is asymptotically normal with a variance depending on the linear term of the Taylor development. Given this particular nature, it was possible to show the asymptotic efficiency of $\hat{\sigma}_{ij,T}$. It means, among all the estimators of σ_{ij}, the estimator defined in equation (2.8) has the lowest variance. The conclusions in Theorems 2.1 and 2.2 depend on to estimate accurately the quadratic term of $\hat{\sigma}_{ij,T}$. We will handle this issue in the next section.

2.4 Estimation of quadratic functionals

We have proved, in Section 2.3.2, the asymptotic normality and found an efficient semiparametric Cramér-Rao bound of the estimator $\hat{\sigma}_{ij,T}$ defined in equation (2.8). We used the Taylor decomposition (2.5) to construct the estimator $\hat{\sigma}_{ij,T}$. In the present section, we will build an estimator for the quadratic term

$$\int H_2(\hat{f}, x_{i1}, x_{j2}, y) f(x_{i1}, x_{j1}, y) f(x_{i2}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy.$$

To this end, we build a general estimator of the parameter with the form:

$$\theta = \int \eta(x_{i1}, x_{j2}, y) f(x_{i1}, x_{j1}, y) f(x_{i2}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy,$$

for $f \in \mathcal{E}$ and $\eta : \mathbb{R}^3 \to \mathbb{R}$ a bounded function.

Given $M = M_n$ a subset of D, consider the estimator

$$\hat{\theta}_n = \frac{1}{n(n-1)} \sum_{l\in M} \sum_{k\neq k'=1}^{n} p_l(X_{ik}, X_{jk}, Y_k)$$
$$\int p_l(x_i, x_j, Y_{k'}) \left(\eta(x_i, X_{jk'}, Y_{k'}) + \eta(X_{ik'}, x_j, Y_{k'})\right) dx_i\, dx_j$$

$$- \frac{1}{n(n-1)} \sum_{l,l' \in M} \sum_{k \neq k'=1}^{n} p_l(X_{ik}, X_{jk}, Y_k) p_{l'}(X_{ik'}, X_{jk'}, Y_{k'})$$
$$\int p_l(x_{i1}, x_{j1}, y) p_{l'}(x_{i2}, x_{j2}, y) \eta(x_{i1}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy. \quad (2.11)$$

To simplify the presentation of Theorem 2.1, write $\psi(x_{i1}, x_{j1}, x_{i1}, x_{j2}, y) = \eta(x_{i1}, x_{j2}, y) + \eta(x_{i2}, x_{j1}, y)$ verifying

$$\int \psi(x_{i1}, x_{j1}, x_{i2}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2} dy$$
$$= \int \psi(x_{i2}, x_{j2}, x_{i1}, x_{j1}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy.$$

With this notation we can simplify (2.11) into

$$\hat{\theta}_n = \frac{1}{n(n-1)} \sum_{l \in M} \sum_{k \neq k'=1}^{n} p_l(X_{ik}, X_{jk}, Y_k)$$
$$\int p_l(x_i, x_j, Y_{k'}) \psi(x_i, x_j, X_{ik'}, X_{jk'}, Y_{k'})\, dx_i\, dx_j$$
$$- \frac{1}{n(n-1)} \sum_{l,l' \in M} \sum_{k \neq k'=1}^{n} p_l(X_{ik}, X_{jk}, Y_k) p_{l'}(X_{ik'}, X_{jk'}, Y_{k'})$$
$$\int p_l(x_{i1}, x_{j1}, y) p_{l'}(x_{i2}, x_{j2}, y) \eta(x_{i1}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy. \quad (2.12)$$

The bias of $\hat{\theta}$ is equal to

$$- \int (S_M f(x_{i1}, x_{j1}, y) - f(x_{i1}, x_{j1}, y))(S_M f(x_{i2}, x_{j2}, y) - f(x_{i2}, x_{j2}, y))$$
$$\eta(x_{i1}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy.$$

The following Theorem gives an explicit bound for the $\hat{\theta}_n$ variance.

Theorem 2.3. *Let Assumption 2.1 hold. Then if $|M_n|/n \to 0$ when $n \to \infty$, then $\hat{\theta}_n$ has the following property*

$$\left| n\, \mathbb{E}\left[\left(\hat{\theta}_n - \theta \right)^2 \right] - \Lambda(f, \eta) \right| \leq \gamma \left[\frac{|M_n|}{n} + \| S_{M_n} f - f \|_2 + \| S_{M_n} g - g \|_2 \right],$$

where $g(x_i, x_j, y) = \int f(x_{i2}, x_{j2}, y)\psi(x_i, x_j, x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2}$ *and*

$$\Lambda(f, \eta) = \int g(x_i, x_j, y)^2 f(x_i, x_j, y)\, dx_i\, dx_j\, dy - \left(\int g(x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy \right)^2,$$

where γ *is constant depending only on* $\|f\|_\infty$, $\|\eta\|_\infty$, *and* $\Delta = (b_1 - a_1) \times (b_2 - a_2)$. *Moreover, this constant is an increasing function of these quantities.*

Note that equation (2.3) implies that

$$\lim_{n\to\infty} n\, \mathbb{E}[(\hat{\theta}_n - \theta)^2] = \Lambda(f, \eta).$$

We can control the quadratic term of $\hat{\sigma}_{ij,T}$, which is a particular case of θ choosing $\eta(x_{i1}, x_{j2}, y) = H_2(\hat{f}, x_{i1}, x_{j2}, y)$.

We will show, in proof 2.6.1, that $\Lambda(f, \eta) \to 0$ when $n \to \infty$. Consequently, the linear part of $\hat{\sigma}_{ij,T}$ governs its asymptotic variance, yielding also asymptotic efficiency.

2.5 Conclusion

In this chapter, we propose a new way to estimate $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$, different from the usual plugin type estimators. We use a general functional $T_{ij}(f)$ depending on the joint density function f of (X_i, X_j, Y). In particular, we grab a suitable approximation $\hat{f}$ of f and construct a coordinate-wise Taylor's expansion up to order three around it. We call this estimator $\hat{\sigma}_{ij,T}$. This expansion serves to estimate $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ using an orthonormal base of $\mathbb{L}^2(\, dx_i\, dx_j\, dy)$.

We highlight that $\hat{\sigma}_{ij,T}$ is asymptotic normal with variance leaded by the first order term. This behavior also causes an efficiency from the Cramér-Rao's point of view. Again, the Cramér-Rao bound depends only on the linear part of the Taylor's series.

With the help of the vech operator, we expanded our results to the matrix-estimator $\widehat{\Sigma}$ formed with the entries $\hat{\sigma}_{ij,T}$. We showed that the $\boldsymbol{T}(f)$'s linear term guides the variance for the $\widehat{\Sigma}_T$'s asymptotic normality.

Even if we aim principally to study a new class of estimators for $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$, we refer to Da Veiga and Gamboa (2013) for some simulations in a context similar to ours. In general, their numerical result behaves reasonably well despite its

implementation's complexity. These results could also work in our framework and we will consider them in a future article.

The estimator $\widehat{\Sigma}_T$ could have negative eigenvalues, violating the semipositive definiteness of the covariance. From a practical point of view, we could project $\widehat{\Sigma}_T$ into the space of positive-semidefinite matrices. It means, we first diagonalize $\widehat{\Sigma}_T$ and then replace negative eigenvalues by 0. The resulting estimator is then semipositive-definite. The works of Bickel and Levina (2008a,b), and Cai et al. (2010), present an extended discussion about techniques on matrix regularization.

This research constitutes a first step in the study of estimators based in Taylor's series with minimum variance. To simplify the implementation's complexity of this estimator, we will explore another kind of techniques like nonparametric methods for example.

2.6 Appendix

2.6.1 Proofs

Proof of Proposition 1.

We need to calculate the three first derivatives of $F(u)$. to ease the calculation, notice first that

$$\frac{d}{du} m_i(f_u, y)u = \frac{\int (x_i - m_i(f_u, y))(f(x_i, x_j, y) - \hat{f}(x_i, x_j, y))\, dx_i\, dx_j}{\int f_u(x_i, x_j, y)\, dx_i\, dx_j}. \tag{2.13}$$

It is possible to interchange the derivate with the integral sign because f and $\hat{f}$ are bounded. Now, using (2.13) and taking $u = 0$ we have

$$F'(0) = \int \left[x_i m_j(\hat{f}, y) + x_j m_i(\hat{f}, y) - m_i(\hat{f}, y) m_j(\hat{f}, y) \right] (f(x_i, x_j, y) - \hat{f}(x_i, x_j, y))\, dx_i\, dx_j\, dy. \tag{2.14}$$

Deriving $m_i(f_u, y) m_j(f_u, y)$ using the same arguments as in (2.13) and again taking $u = 0$ we get,

$$F''(0) = \int \frac{2}{\int \hat{f}(x_i, x_j, y)\, dx_i\, dx_j} (x_{i1} - m_i(\hat{f}, y))(x_{j2} - m_j(\hat{f}, y))(f(x_{i1}, x_{j1}, y) - \hat{f}(x_{i1}, x_{j1}, y)) (f(x_{i2}, x_{j2}, y) - \hat{f}(x_{i2}, x_{j2}, y))\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy. \quad (2.15)$$

Using the previous arguments we find also that

$$F'''(u) = \int \frac{-6}{\int f_u(x_i, x_j, y) dx_i dx_j} \big(x_{i1} - m_j(f_u, y)\big)\big(x_{j2} - m_j(f_u, y)\big) \big(f(x_{i1}, x_{j1}, y) - \hat{f}(x_{i1}, x_{j1}, y)\big)\big(f(x_{i2}, x_{j2}, y) - \hat{f}(x_{i2}, x_{j2}, y)\big) \big(f(x_{i3}, x_{j3}, y) - \hat{f}(x_{i3}, x_{j3}, y)\big)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dx_{i3}\, dx_{j3}\, dy \quad (2.16)$$

Replacing (2.14), (2.15) and (2.16) into (2.4) we get the desired decomposition. □

Proof of Theorem 2.1. We will first control the remaining term (2.6),

$$\Gamma_n = \frac{1}{6} F'''(\xi)(1 - \xi)^3.$$

Remember that

$$F'''(\xi) = -6 \int \frac{\big(x_{i1} - m_i(f_\xi, y)\big)\, \big(x_{j2} - m_j(f_\xi, y)\big)}{\big(\int f_\xi(x_i, x_j, y)\, dx_i\, dx_j\big)^2} \Big(f(x_{i1}, x_{j1}, y) - \hat{f}(x_{i1}, x_{j1}, y)\Big)\, \Big(f(x_{i2}, x_{j2}, y) - \hat{f}(x_{i2}, x_{j2}, y)\Big) \Big(f(x_{i3}, x_{j3}, y) - \hat{f}(x_{i3}, x_{j3}, y)\Big)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dx_{i3}\, dx_{j3}\, dy,$$

Assumptions 2.1 and 2.2 ensure that the first part of the integrand is bounded by a constant μ. Furthermore,

$$\begin{aligned} |\Gamma_n| &\leq \mu \int \Big|f(x_{i1}, x_{j1}, y) - \hat{f}(x_{i1}, x_{j1}, y)\Big| \Big|f(x_{i2}, x_{j2}, y) - \hat{f}(x_{i2}, x_{j2}, y)\Big| \\ &\quad \Big|f(x_{i3}, x_{j3}, y) - \hat{f}(x_{i3}, x_{j3}, y)\Big|\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dx_{i3}\, dx_{j3}\, dy \\ &= \mu \int \left(\int \Big|f(x_i, x_j, y) - \hat{f}(x_i, x_j, y)\Big|\, dx_i\, dx_j\right)^3 dy \\ &\leq \mu \Delta^3 \int \Big|f(x_i, x_j, y) - \hat{f}(x_i, x_j, y)\Big|^3\, dx_i\, dx_j\, dy \end{aligned}$$

by the Hölder inequality. Then $\mathbb{E}[\Gamma_n^2]$ is equal to $O(\mathbb{E}\|f - \hat{f}\|_3^6)$. Since $\hat{f}$ verifies Assumption 2.3, this quantity is of order $O(n_1^{-6\lambda})$. Since we also assume $n_1 \approx n/\log(n)$ and $\lambda > 1/6$, then $n_1^{-6\lambda} = o(1/n)$. Therefore, we get $\mathbb{E}[\Gamma_n^2] = o(1/n)$ which implies that the remaining term Γ_n is negligible.

To prove the asymptotic normality of $\hat{\sigma}_{ij,T}$, we shall show that $\sqrt{n}\left(\hat{\sigma}_{ij,T} - T_{ij}(f)\right)$ and

$$Z_{ij}^{(n)} = \frac{1}{n_2}\sum_{k=1}^{n_2} H_1(f, X_{ik}, X_{jk}, Y_k) - \int H_1(f, x_i, x_j, y) f(x_i, x_j, y))\, dx_i\, dx_j\, dy \quad (2.17)$$

have the same asymptotic behavior. We can get for $Z_{ij}^{(n)}$ a classic central limit theorem with variance

$$\begin{aligned} C_{ij}(f) &= \mathrm{Var}(H_1(f, x_i, x_j, y)) \\ &= \int H_1(f, x_i, x_j, y)^2 f(x_i, x_j, y))\, dx_i\, dx_j\, dy \\ &\quad - \left(\int H_1(f, x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy\right)^2 \end{aligned}$$

which implies (2.9) and (2.10). In order to establish our claim, we will show that

$$R_{ij}^{(n)} = \sqrt{n}\left[\hat{\sigma}_{ij,T} - T_{ij}(f) - Z_{ij}^{(n)}\right] \quad (2.18)$$

has second-order moment converging to 0.

Define $\widehat{Z}_{ij}^{(n)}$ as $Z_{ij}^{(n)}$ with f replaced by $\hat{f}$. Let us note that $R_{ij}^{(n)} = R_1 + R_2$ where

$$\begin{aligned} R_1 &= \sqrt{n}\left[\hat{\sigma}_{ij,T} - T_{ij}(f) - \widehat{Z}_{ij}^{(n)}\right] \\ R_2 &= \sqrt{n}\left[\widehat{Z}_{ij}^{(n)} - Z_{ij}^{(n)}\right]. \end{aligned}$$

It only remains to state that $\mathbb{E}[R_1^2]$ and $\mathbb{E}[R_2^2]$ converges to 0. We can rewrite R_1 as

$$R_1 = -\sqrt{n}\left[\widehat{Q} - Q + \Gamma_n\right]$$

with

$$Q = \int H_2(\hat{f}, x_{i1}, x_{j2}, y) f(x_{i1}, x_{j1}, y) f(x_{i2}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy$$
$$H_2(\hat{f}, x_{i1}, x_{j2}, y) = \frac{1}{\int \hat{f}(x_i, x_j, y)\, dx_i\, dx_j} \Big(x_{i1} - m_i(\hat{f}, y)\Big) \Big(x_{j2} - m_j(\hat{f}, y)\Big).$$

We can estimate $H_2(\hat{f}, x_{i1}, x_{j2}, y) = \eta(x_{i1}, x_{j2}, y)$ as done in Section 2.4. Let $\widehat{Q}$ the estimator of Q. Since $\mathbb{E}[\Gamma_n^2] = o(1/n)$, we only have to control the term $\sqrt{n}(\widehat{Q} - Q)$ which is such that $\lim_{n\to\infty} n\, \mathbb{E}[\widehat{Q} - Q]^2 = 0$ by Lemma 2.7 in Section 2.6.2 below. This Lemma implies that $\mathbb{E}[R_1^2] \to 0$ as $n \to \infty$. For R_2 we have

$$\mathbb{E}[R_2^2] = \frac{n}{n_2} \left[\int \Big(H_1(f, x_i, x_j, y) - H_1(\hat{f}, x_i, x_j, y)\Big)^2 f(x_i, x_j, y))\, dx_i\, dx_j\, dy \right]$$
$$- \frac{n}{n_2} \left[\int H_1(f, x_i, x_j, y) f(x_i, x_j, y))\, dx_i\, dx_j\, dy \right.$$
$$\left. - \int H_1(\hat{f}, x_i, x_j, y)^2 f(x_i, x_j, y))\, dx_i\, dx_j\, dy \right]^2.$$

The same arguments as the ones of Lemma 2.7 (mean value and Assumptions 2.2 and 2.3) show that $\mathbb{E}[R_2^2] \to 0$. □

Proof of Theorem 2.2 . To prove the inequality we will use the usual framework described in Ibragimov and Khas'Minskii (1991). The first step is to calculate the Fréchet derivative of $T_{ij}(f)$ at some point $f_0 \in \mathcal{E}$. Assumptions 2.2 and 2.3 and equation (2.5), imply that

$$T_{ij}(f) - T_{ij}(f_0) = \int \Big(x_i m_j(f_0, y) + x_j m_i(f_0, y) - m_i(f_0, y) m_j(f_0, y)\Big)$$
$$\Big(f(x_i, x_j, y) - f_0(x_i, x_j, y)\Big)\, dx_i\, dx_j\, dy + O\left(\int (f - f_0)^2\right)$$

where $m_i(f_0, y) = \int x_i f_0(x_i, x_j, y)\, dx_i\, dx_j\, dy / \int f_0(x_i, x_j, y)\, dx_i\, dx_j\, dy$. Therefore, the Fréchet derivative of $T_{ij}(f)$ at f_0 is $T'_{ij}(f_0) \cdot h = \langle H_1(f_0, \cdot), h \rangle$ with

$$H_1(f_0, x_i, x_j, y) = x_i m_j(f_0, y) + x_j m_i(f_0, y) - m_i(f_0, y) m_j(f_0, y).$$

Using the results of Ibragimov and Khas'Minskii (1991), denote $H(f_0) = \left\{ u \in \mathbb{L}^2(\, dx_i\, dx_j\, dy), \int u(x_i, x_j, y) \sqrt{f_0(x_i, x_j, y)}\, dx_i\, dx_j\, dy = 0 \right\}$ the set

of functions in $\mathbb{L}^2(\,dx_i\,dx_j\,dy)$ orthogonal to $\sqrt{f_0}$, $\mathbb{P}_{H(f_0)}$ the projection onto $H(f_0)$, $A_n(t) = \left(\sqrt{f_0}\right) t/\sqrt{n}$ and $P_{f_0}^{(n)}$ the joint distribution of $\left(X_{ik}, X_{jk}\right)$ $k = 1, \ldots, n$ under f_0. Since $\left(X_{ik}, X_{jk}\right)$ $k = 1, \ldots, n$ are i.i.d., the family $\left\{P_{f_0}^{(n)}, f \in \mathcal{E}\right\}$ is differentiable in quadratic mean at f_0 and therefore locally asymptotically normal at all points $f_0 \in \mathcal{E}$ in the direction $H(f_0)$ with normalizing factor $A_n(f_0)$ (see the details in Van der Vaart (2000)). Then, by the results of Ibragimov and Khas'Minskii (1991) say that under these conditions, denoting $K_n = B_n\theta'(f_0)A_n\mathbb{P}_{H(f_0)}$ with $B_n = \sqrt{n}u$, if $K_n \xrightarrow{\mathcal{D}} K$ and if $K(u) = \langle t, u\rangle$, then for every estimator $\hat{\sigma}_{ij,T}$ of $T_{ij}(f)$ and every family $\mathcal{V}(f_0)$ of vicinities of f_0, we have

$$\inf_{\{\mathcal{V}(f_0)\}} \liminf_{n\to\infty} \sup_{f\in\mathcal{V}(f_0)} n\,\mathbb{E}[\hat{\sigma}_{ij,T} - T_{ij}(f_0)]^2 \geq \|t_{\mathbb{L}^2(\,dx_i\,dx_j\,dy)}\|^2.$$

Here,

$$K_n(u) = \sqrt{n}T'(f_0)\cdot\frac{\sqrt{f_0}}{\sqrt{n}}\mathbb{P}_{H(f_0)}(u) = T'(f_0)\left(\sqrt{f_0}\left(u - \sqrt{f_0}\int u\sqrt{f_0}\right)\right),$$

since for any $u \in \mathbb{L}^2(\,dx_i\,dx_j\,dy)$ we can write it as $u = \sqrt{f_0}\left\langle\sqrt{f_0}, u\right\rangle + \mathbb{P}_{H(f_0)}(u)$. In this case $K_n(u)$ does not depend on n and

$$\begin{aligned} K(h) &= T'(f_0)\cdot\left(\sqrt{f_0}\left(u - \sqrt{f_0}\int h\sqrt{f_0}\right)\right) \\ &= \int H_1(f_0,\cdot)\sqrt{f_0}u - \int H_1(f_0,\cdot)\sqrt{f_0}\int u\sqrt{f_0} \\ &= \langle t, u\rangle \end{aligned}$$

with

$$t(x_i, x_j, y) = H_1(f_0, x_i, x_j, y)\sqrt{f_0} - \left(\int H_1(f_0, x_i, x_j, y)f_0\right)\sqrt{f_0}.$$

The semi-parametric Cramér-Rao bound for this problem is thus

$$\begin{aligned} \|t_{\mathbb{L}^2(\,dx_i,\,dx_j,\,dy)}\| = &\int H_1(f_0, x_i, x_j, y)^2 f_0\,dx_i\,dx_j\,dy \\ &- \left(\int H_1(f_0, x_i, x_j, y)f_0\;dx_i\,dx_j\,dy\right)^2 \end{aligned}$$

and we recognize the expression $C_{ij}(f_0)$ found in Theorem 2.1. □

Proof of Corollary 2.1. The proof is based in the following observation. Employing equation (2.18) we have

$$\widehat{\boldsymbol{T}}_n - \boldsymbol{T}(f) = \boldsymbol{Z}_n(f) + \frac{\boldsymbol{R}_n}{\sqrt{n}}$$

where $\boldsymbol{Z}_n(f)$ and $\boldsymbol{R}_n$ are matrices with elements $Z_{ij}^{(n)}$ and $R_{ij}^{(n)}$, defined in (2.17) and (2.18), respectively.

Hence we have,

$$n\,\mathbb{E}\left[\left\|\operatorname{vech}\left(\widehat{\boldsymbol{T}}_n - \boldsymbol{T}(f) - \boldsymbol{Z}_n(f)\right)\right\|^2\right] = \mathbb{E}\left[\|\operatorname{vech}(\boldsymbol{R}_n)\|^2\right] = \sum_{i\le j}\mathbb{E}\left[\left(R_{ij}^{(n)}\right)^2\right].$$

We see by Lemma 2.7 that $\mathbb{E}[R_{ij}^2] \to 0$ as $n \to 0$. It follows that

$$n\,\mathbb{E}[\|\operatorname{vech}\left(\widehat{\boldsymbol{T}}_n - \boldsymbol{T}(f) - \boldsymbol{Z}_n(f)\right)\|^2] \to 0 \text{ as } n \to 0.$$

We know that if X_n, X and Y_n are random variables, then if $X_n \xrightarrow{\mathcal{D}} X$ and $(X_n - Y_n) \xrightarrow{\mathcal{P}} 0$, follows that $Y_n \xrightarrow{\mathcal{D}} X$.

Remember also that convergence in $\mathbb{L}^2$ implies convergence in probability, therefore

$$\sqrt{n}\operatorname{vech}\left(\widehat{\boldsymbol{T}}_n - \boldsymbol{T}(f) - \boldsymbol{Z}_n(f)\right) \xrightarrow{\mathcal{P}} 0.$$

By the multivariate central limit theorem we have that $\sqrt{n}\operatorname{vech}(\boldsymbol{Z}_n(f)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{C}(f))$. Therefore, $\sqrt{n}\operatorname{vech}\left(\hat{\boldsymbol{T}}_n - \boldsymbol{T}(f)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{C}(f))$. □

Proof of Theorem 2.3. For abbreviation, we write M instead of M_n and set $m = |M_n|$. We first compute the mean squared error of $\hat{\theta}_n$ as

$$\mathbb{E}[\hat{\theta}_n - \theta]^2 = \operatorname{Bias}^2(\hat{\theta}_n) + \operatorname{Var}(\hat{\theta}_n)$$

where $\operatorname{Bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta$.

We begin the proof by bounding $\operatorname{Var}(\hat{\theta}_n)$. Let A and B be $m \times 1$ vectors with components

$$a_l = \int p_l(x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy \quad l = 1, \ldots, m,$$

$$
\begin{aligned}
b_l &= \int p_l(x_{i1}, x_{j1}, y) f(x_{i2}, x_{j2}, y) \psi(x_{i1}, x_{j1}, x_{i2}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy \\
&= \int p_l(x_i, x_j, y) g(x_i, x_j, y)\, dx_i\, dx_j\, dy \quad l = 1, \ldots, m
\end{aligned}
$$

where $g(x_i, x_j, y) = \int f(x_{i2}, x_{j2}, y)\psi(x_i, x_j, x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2}$. Let Q and R be $m \times 1$ vectors of centered functions

$$
\begin{aligned}
q_l(x_i, x_j, y) &= p_l(x_i, x_j, y) - a_l \\
r_l(x_i, x_j, y) &= \int p_l(x_{i2}, x_{j2}, y)\psi(x_i, x_j, x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2} - b_l
\end{aligned}
$$

for $l = 1, \ldots, m$. Let C a $m \times m$ matrix of constants with indices $l, l' = 1, \ldots, m$ defined by

$$
c_{ll'} = \int p_l(x_{i1}, x_{j1}, y) p_{l'}(x_{i2}, x_{j2}, y) \eta(x_{i1}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy.
$$

Let us denote by U_n the process

$$
U_n h = \frac{1}{n(n-1)} \sum_{k \neq k'=1}^{n} h\big(X_{ik}, X_{jk}, Y_k, X_{ik'}, X_{jk'}, Y_{k'}\big)
$$

and P_n the empirical measure

$$
P_n h = \frac{1}{n} \sum_{k=1}^{n} h\big(X_{ik}, X_{jk}, Y_k\big)
$$

for some h in $\mathbb{L}^2(\, dx_i, \, dx_j, \, dy)$. With these notations, $\hat{\theta}_n$ has the Hoeffding's decomposition

$$
\begin{aligned}
\hat{\theta}_n &= \frac{1}{n(n-1)} \sum_{l \in M} \sum_{k \neq k'=1}^{n} \big(q_l(X_{ik}, X_{jk}, Y_k) + a_l\big)\big(r_l(X_{ik'}, X_{jk'}, Y_{k'}) + b_l\big) \\
&\quad - \frac{1}{n(n-1)} \sum_{l,l' \in M} \sum_{k \neq k'=1}^{n} \big(q_l(X_{ik}, X_{jk}, Y_k) + a_l\big)\big(q_{l'}(X_{ik'}, X_{jk'}, Y_{k'}) + a_{l'}\big) c_{ll'} \\
&= U_n K + P_n L + A^\top B - A^\top C A
\end{aligned}
$$

where

$$
\begin{aligned}
K\left(x_{i1}, x_{j1}, y_1, x_{i2}, x_{j2}, y_2\right) =& Q^\top(x_{i1}, x_{j1}, y_1) R(x_{i2}, x_{j2}, y_2) \\
& - Q^\top(x_{i1}, x_{j1}, y_1) C Q(x_{i2}, x_{j2}, y_2)
\end{aligned}
$$

$$L(x_i, x_j, y) = A^\top R(x_i, x_j, y) + BQ(x_i, x_j, y) - 2A^\top CQ(x_i, x_j, y).$$

Therefore $\mathrm{Var}(\hat{\theta}_n) = \mathrm{Var}(U_n K) + \mathrm{Var}(P_n L) - 2\,\mathrm{Cov}(U_n K, P_n L)$. These three terms are bounded in Lemmas 2.2 - 2.4, which gives

$$\mathrm{Var}(\hat{\theta}_n) \leq \frac{20}{n(n-1)} \|\eta\|_\infty^2 \|f\|_\infty^2 \Delta^2 (m+1) + \frac{12}{n} \|\eta\|_\infty^2 \|f\|_\infty^2 \Delta^2.$$

For n enough large and a constant $\gamma \in \mathbb{R}$,

$$\mathrm{Var}(\hat{\theta}_n) \leq \gamma \|\eta\|_\infty^2 \|f\|_\infty^2 \Delta^2 \left(\frac{m}{n^2} + \frac{1}{n} \right).$$

The term $\mathrm{Bias}(\hat{\theta}_n)$ is easily computed, as proven in Lemma 2.5, is equal to

$$-\int \left(S_M f(x_{i1}, x_{j1}, y) - f(x_{i1}, x_{j1}, y)\right) \left(S_M f(x_{i2}, x_{j2}, y) - f(x_{i2}, x_{j2}, y)\right) \eta(x_{i1}, x_{j1}, x_{i2}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy.$$

From Lemma 2.5, we bound the bias of $\hat{\theta}_n$ by

$$|\mathrm{Bias}(\hat{\theta}_n)| \leq \Delta \|\eta\|_\infty \sup_{l \notin M} |c_l|^2.$$

The assumption of $\left(\sup_{l \notin M} |c_i|^2\right)^2 \approx m/n^2$ and since $m/n \to 0$, we deduce that $\mathbb{E}[\hat{\theta}_n - \theta]^2$ has a parametric rate of convergence $O(1/n)$.

Finally to prove (2.3), note that

$$\begin{aligned} n\,\mathbb{E}[\hat{\theta}_n - \theta]^2 &= n\,\mathrm{Bias}^2(\hat{\theta}_n) + n\,\mathrm{Var}(\hat{\theta}_n) \\ &= n\,\mathrm{Bias}^2(\hat{\theta}_n) + n\,\mathrm{Var}(U_n K) + n\,\mathrm{Var}(P_n L). \end{aligned}$$

We previously proved that for some $\lambda_1, \lambda_2 \in \mathbb{R}$

$$\begin{aligned} n\,\mathrm{Bias}^2(\hat{\theta}_n) &\leq \lambda_1 \Delta^2 \|\eta\|_\infty^2 \frac{m}{n} \\ n\,\mathrm{Var}(U_n K) &\leq \lambda_2 \Delta^2 \|f\|_\infty^2 \|\eta\|_\infty^2 \frac{m}{n}. \end{aligned}$$

Thus, Lemma 2.6 implies

$$|n\,\mathrm{Var}(P_n L) - \Lambda(f, \eta)| \leq \lambda \left[\|S_M f - f\|_2 + \|S_M g - g\|_2 \right],$$

where λ is an increasing function of $\|f_\infty\|^2$, $\|\eta\|_\infty^2$ and Δ. From all this we deduce (2.3) which ends the proof of Theorem 2.3. □

2.6.2 Technical Results

Lemma 2.1 (Bias of $\hat{\theta}_n$)**.** *The estimator $\hat{\theta}_n$ defined in* (2.12) *estimates θ with bias equal to*

$$-\int \left(S_M f(x_{i1}, x_{j1}, y) - f(x_{i1}, x_{j1}, y)\right) \left(S_M f(x_{i2}, x_{j2}, y) - f(x_{i2}, x_{j2}, y)\right) \eta(x_{i1}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy.$$

Proof. Let $\hat{\theta}_n = \hat{\theta}_n^1 - \hat{\theta}_n^2$ where

$$\hat{\theta}_n^1 = \frac{1}{n(n-1)} \sum_{l \in M} \sum_{k \neq k'=1} p_l(X_{ik}, X_{jk}, Y_k) \int p_l(x_i, x_j, Y_{k'}) \psi(x_i, x_j, X_{ik'}, X_{jk'}, Y_{k'}) dx_i dx_j$$

$$\hat{\theta}_n^2 = -\frac{1}{n(n-1)} \sum_{l,l' \in M} \sum_{k \neq k'=1}^{n} p_l(X_{ik}, X_{jk}, Y_k) p_{l'}(X_{ik'}, X_{jk'}, Y_{k'}) \int p_l(x_{i1}, x_{j1}, y) p_{l'}(x_{i2}, x_{j2}, y) \eta(x_{i1}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy.$$

Let us first compute $\mathbb{E}[\hat{\theta}_n^1]$.

$$\begin{aligned}
&\mathbb{E}[\hat{\theta}_n^1] \\
&= \sum_{l \in M} \int p_l(x_{i1}, x_{j1}, y) f(x_{i1}, x_{j1}, y)\, dx_{i1}\, dx_{j1}\, dy \\
&\quad \int p_l(x_{i1}, x_{j1}, y) \psi(x_{i1}, x_{j1}, x_{i2}, x_{j2}, y) f(x_{i2}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy \\
&= \sum_{l \in M} a_l \int p_l(x_{i1}, x_{j1}, y) \psi(x_{i1}, x_{j1}, x_{i2}, x_{j2}, y) f(x_{i2}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy \\
&= \int \left(\sum_{l \in M} a_l p_l(x_{i2}, x_{j2}, y) \right) \psi(x_{i1}, x_{j1}, x_{i2}, x_{j2}, y) f(x_{i2}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy \\
&= \int S_M f(x_{i1}, x_{j1}, y) f(x_{i2}, x_{j2}, y) \eta(x_{i1}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy \\
&+ \int S_M f(x_{i2}, x_{j2}, y) f(x_{i1}, x_{j1}, y) \eta(x_{i1}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy
\end{aligned}$$

Now for $\hat{\theta}_n^2$, we get

$$\mathbb{E}[\hat{\theta}_n^2]$$

$$= \sum_{l,l' \in M} \int p_l(x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy \int p_{l'}(x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy$$
$$\int p_l(x_{i1}, x_{j1}, y) p_{l'}(x_{i2}, x_{j2}, y) \eta(x_{i1}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy$$
$$= \sum_{l,l' \in M} a_l a_{l'} \int p_l(x_{i1}, x_{j1}, y) p_{l'}(x_{i2}, x_{j2}, y) \eta(x_{i1}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy$$
$$= \int \left(\sum_{l \in M} a_l p_l(x_{i1}, x_{j1}, y) \right) \left(\sum_{l' \in M} a_{l'} p_{l'}(x_{i2}, x_{j2}, y) \right) \eta(x_{i1}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy$$
$$= \int S_M f(x_{i1}, x_{j1}, y) S_M f(x_{i2}, x_{j2}, y) \eta(x_{i1}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy.$$

Arranging these terms and using

$$\text{Bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta = \mathbb{E}[\hat{\theta}_n^1] - \mathbb{E}[\hat{\theta}_n^2] - \theta$$

we obtain the desire bias. □

Lemma 2.2 (Bound of $\text{Var}(U_n K)$)**.** *Under the assumptions of Theorem 2.3, we have*

$$\text{Var}(U_n K) \leq \frac{20}{n(n-1)} \|\eta\|_\infty^2 \|f\|_\infty^2 \Delta^2 (m+1)$$

Proof. Note that $U_n K$ is centered because Q and R are centered and (X_{ik}, X_{jk}, Y_k), $k = 1, \ldots, n$ is an independent sample. So $\text{Var}(U_n K)$ is equal to

$$\begin{aligned}
\mathbb{E}[U_n K]^2 =& \mathbb{E}\Bigg(\frac{1}{(n(n-1))^2} \sum_{k_1 \neq k_1'=1}^{n} \sum_{k_2 \neq k_2'=1}^{n} K\big(X_{ik_1}, X_{jk_1}, Y_{k_1}, X_{ik_1'}, X_{jk_1'}, Y_{k_1'}\big) \\
& K\big(X_{ik_2}, X_{jk_2}, Y_{k_2}, X_{ik_2'}, X_{jk_2'}, Y_{k_2'}\big) \Bigg) \\
=& \frac{1}{n(n-1)} \mathbb{E}\Bigg(K^2\big(X_{i1}, X_{j1}, Y_1, X_{i2}, X_{j2}, Y_2\big) \\
& + K\big(X_{i1}, X_{j1}, Y_1, X_{i2}, X_{j2}, Y_2\big) K\big(X_{i2}, X_{j2}, Y_2, X_{i1}, X_{j1}, Y_1\big) \Bigg)
\end{aligned}$$

By the Cauchy-Schwarz inequality, we get

$$\text{Var}(U_n K) \leq \frac{2}{n(n-1)} \mathbb{E}[K^2\left(X_{i1}, X_{j1}, Y_1, X_{i2}, X_{j2}, Y_2\right)].$$

Moreover, using the fact that $2|\mathbb{E}[XY]| \leq \mathbb{E}[X^2] + \mathbb{E}[Y^2]$, we obtain

$$\begin{aligned}
&\mathbb{E}[K^2\left(X_{i1}, X_{j1}, Y_1, X_{i2}, X_{j2}, Y_2\right)] \\
&\qquad \leq 2\Bigg[\mathbb{E}[\left(Q^\top(X_{i1}, X_{j1}, Y_1)R(X_{i2}, X_{j2}, Y_2)\right)^2] \\
&\qquad + \mathbb{E}[\left(Q^\top(X_{i1}, X_{j1}, Y_1)CQ(X_{i2}, X_{j2}, Y_2)\right)^2]\Bigg].
\end{aligned}$$

We will bound these two terms. The first one is

$$\begin{aligned}
&\mathbb{E}[\left(Q^\top(X_{i1}, X_{j1}, Y_1)R(X_{i2}, X_{j2}, Y_2)\right)^2] \\
=&\sum_{l,l' \in M} \left(\int p_l(x_i, x_j, y) p_{l'}(x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy - a_l a_{l'} \right) \\
&\left(\int p_l(x_{i2}, x_{j2}, y) p_{l'}(x_{i3}, x_{j3}, y) \psi(x_{i1}, x_{j1}, x_{i2}, x_{j2}, y) \right. \\
&\left. \psi(x_{i1}, x_{j1}, x_{i3}, x_{j3}, y) f(x_{i1}, x_{j1}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dx_{i3}\, dx_{j3}\, dy - b_l b_{l'} \right) \\
=&W_1 - W_2 - W_3 + W_4
\end{aligned}$$

where

$$\begin{aligned}
W_1 = &\int \sum_{l,l' \in M} p_l(x_{i1}, x_{j1}, y) p_{l'}(x_{i1}, x_{j1}, y) \\
&p_l(x_{i2}, x_{j2}, y') p_{l'}(x_{i3}, x_{j3}, y') \psi(x_{i4}, x_{j4}, x_{i2}, x_{j2}, y') \\
&\psi(x_{i4}, x_{j4}, x_{i3}, x_{j3}, y') f(x_{i1}, x_{j1}, y) \\
&f(x_{i4}, x_{j4}, y')\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dx_{i3}\, dx_{j3}\, dx_{i4}\, dx_{j4}\, dy\, dy' \\
W_2 = &\int \sum_{l,l' \in M} b_l b_{l'} p_l(x_{i1}, x_{j1}, y) p_{l'}(x_{i1}, x_{j1}, y) f(x_{i1}, x_{j1}, y)\, dx_{i1}\, dx_{j1} dy \\
W_3 = &\int \sum_{l,l' \in M} a_l a_{l'} p_l(x_{i2}, x_{j2}, y') p_{l'}(x_{i3}, x_{j3}, y') \\
&\psi(x_{i4}, x_{j4}, x_{i2}, x_{j2}, y') \psi(x_{i4}, x_{j4}, x_{i3}, x_{j3}, y') \\
&f(x_{i4}, x_{j4}, y')\, dx_{i2}\, dx_{j2}\, dx_{i3}\, dx_{j3}\, dx_{i4}\, dx_{j4}\, dy' \\
W_4 = &\sum_{l,l' \in M} a_l a_{l'} b_l b_{l'}.
\end{aligned}$$

W_2 and W_3 are positive, hence

$$\mathbb{E}[\Big(2Q^\top(X_{i1}, X_{j1}, Y_1)R(X_{i2}, X_{j2}, Y_2)\Big)^2] \leq W_1 + W_4.$$

$$
\begin{aligned}
W_1 = &\int \sum_{l,l' \in M} p_l(x_{i1}, x_{j1}, y) p_{l'}(x_{i1}, x_{j1}, y) \\
&\left(\int p_l(x_{i2}, x_{j2}, y') \psi(x_{i4}, x_{j4}, x_{i2}, x_{j2}, y') \, dx_{i2} \, dx_{j2} \right) \\
&\left(\int p_{l'}(x_{i3}, x_{j3}, y') \psi(x_{i4}, x_{j4}, x_{i3}, x_{j3}, y') \, dx_{i3} \, dx_{j3} \right) \\
&f(x_{i1}, x_{j1}, y) f(x_{i4}, x_{j4}, y') \, dx_{i1} \, dx_{j1} \, dx_{i4} \, dx_{j4} \, dy \, dy' \\
\leq \|f\|_\infty^2 &\sum_{l,l' \in M} \int p_l(x_{i1}, x_{j1}, y) p_{l'}(x_{i1}, x_{j1}, y) \, dx_{i1} \, dx_{j1} dy \\
&\int \left(\int p_l(x_{i2}, x_{j2}, y') \psi(x_{i4}, x_{j4}, x_{i2}, x_{j2}, y') \, dx_{i2} \, dx_{j2} \right) \\
&\left(\int p_{l'}(x_{i3}, x_{j3}, y') \psi(x_{i4}, x_{j4}, x_{i3}, x_{j3}, y') \, dx_{i3} \, dx_{j3} \right) \, dx_{i2} \, dx_{j2} \, dx_{i4} \, dx_{j4} \, dy'
\end{aligned}
$$

Since p_l's are orhonormal we have

$$
W_1 \leq \|f\|_\infty^2 \sum_{l \in M} \int \left(\int p_l(x_{i2}, x_{j2}, y') \psi(x_{i4}, x_{j4}, x_{i2}, x_{j2}, y') \, dx_{i2} \, dx_{j2} \right)^2 dx_{i4} \, dx_{j4} \, dy'.
$$

Moreover by the Cauchy-Schwarz inequality and $\|\psi\|_\infty \leq 2\|\eta\|_\infty$

$$
\begin{aligned}
&\left(\int p_l(x_{i2}, x_{j2}, y') \psi(x_{i4}, x_{j4}, x_{i2}, x_{j2}, y') \, dx_{i2} \, dx_{j2} \right)^2 \\
&\qquad \leq \int p_l(x_{i2}, x_{j2}, y')^2 \, dx_{i2} \, dx_{j2} \int \psi(x_{i4}, x_{j4}, x_{i2}, x_{j2}, y')^2 \, dx_{i2} \, dx_{j2} \\
&\qquad \leq \|\psi\|_\infty^2 \Delta \int p_l(x_{i2}, x_{j2}, y')^2 \, dx_{i2} \, dx_{j2} \\
&\qquad \leq 4\|\eta\|_\infty^2 \Delta \int p_l(x_{i2}, x_{j2}, y')^2 \, dx_{i2} \, dx_{j2},
\end{aligned}
$$

and then

$$
\begin{aligned}
&\int \left(\int p_l(x_{i2}, x_{j2}, y') \psi(x_{i4}, x_{j4}, x_{i2}, x_{j2}, y') \, dx_{i2} \, dx_{j2} \right)^2 dx_{i4} \, dx_{j4} \, dy' \\
&\leq 4\|\eta\|_\infty^2 \Delta^2 \int p_l(x_{i2}, x_{j2}, y')^2 \, dx_{i2} \, dx_{j2} \, dy' \\
&= 4\|\eta\|_\infty^2 \Delta^2.
\end{aligned}
$$

Finally,

$$W_1 \leq 4\|\eta\|_\infty^2 \|f\|_\infty^2 \Delta^2 m.$$

For the term W_4 using the facts that $S_M f$ and $S_M g$ are projection and that $\int f = 1$, we have

$$W_4 = \left(\sum_{l \in M} a_l b_l \right)^2 \leq \sum_{l \in M} a_l^2 \sum_{l \in M} b_l^2 \leq \|f\|_2^2 \|g\|_2^2 \leq \|f\|_\infty \|g\|_2^2.$$

By the Cauchy-Schwartz inequality we have $\|g\|_2^2 \leq 4\|\eta\|_\infty^2 \|f\|_\infty \Delta^2$ and then

$$W_4 \leq 4\|\eta\|_\infty^2 \|f\|_\infty^2 \Delta^2$$

which leads to

$$\mathbb{E}[\left(Q^\top(X_{i1}, X_{j1}, Y_1) R(X_{i2}, X_{j2}, Y_2)\right)^2] \leq 4\|\eta\|_\infty^2 \|f\|_\infty^2 \Delta^2 (m+1). \tag{2.19}$$

The second term $\mathbb{E}[\left(Q^\top(X_{i1}, X_{j1}, Y_1 C Q(X_{i2}, X_{j2}, Y_2)\right)] = W_5 - 2W_6 + W_7$ where

$$\begin{aligned}
W_5 = & \int \sum_{l_1, l_1'} \sum_{l_2, l_2'} c_{l_1 l_1'} c_{l_2 l_2'} p_{l_1}(x_{i1}, x_{j1}, y) p_{l_2}(x_{i1}, x_{j1}, y) p_{l_1'}(x_{i2}, x_{j2}, y') p_{l_2'}(x_{i2}, x_{j2}, y') \\
& f(x_{i1}, x_{j1}, y) f(x_{i2}, x_{j2}, y') \, dx_{i1} \, dx_{j1} \, dx_{i2} \, dx_{j2} \, dy' \, dy \\
W_6 = & \int \sum_{l_1, l_1'} \sum_{l_2, l_2'} c_{l_1 l_1'} c_{l_2 l_2'} a_{l_1} a_{l_2} p_{l_1'}(x_i, x_j, y) p_{l_2'}(x_i, x_j, y) \, dx_i \, dx_j \, dy \\
W_7 = & \sum_{l_1, l_1'} \sum_{l_2, l_2'} c_{l_1 l_1'} c_{l_2 l_2'} a_{l_1} a_{l_1'} a_{l_2} a_{l_2'}.
\end{aligned}$$

Using the previous manipulation, we show that $W_6 \geq 0$. Thus

$$\mathbb{E}[\left(Q^\top(X_{i1}, X_{j1}, Y_1) C Q(X_{i2}, X_{j2}, Y_2)\right)] \leq W_5 + W_7.$$

First, observe that

$$\begin{aligned}
W_5 = & \sum_{l_1, l_1'} \sum_{l_2, l_2'} c_{l_1 l_1'} c_{l_2 l_2'} \left(\int p_{l_1}(x_{i1}, x_{j1}, y) p_{l_2}(x_{i1}, x_{j1}, y) f(x_{i1}, x_{j1}, y) \, dx_{i1} \, dx_{j1} \, dy \right) \\
& \left(\int p_{l_1'}(x_{i2}, x_{j2}, y') p_{l_2'}(x_{i2}, x_{j2}, y') f(x_{i2}, x_{j2}, y') \, dx_{i2} \, dx_{j2} \, dy' \right) \\
\leq & \|f\|_\infty^2 \sum_{l_1, l_1'} \sum_{l_2, l_2'} c_{l_1 l_1'} c_{l_2 l_2'} \left(\int p_{l_1}(x_{i1}, x_{j1}, y) p_{l_2}(x_{i1}, x_{j1}, y) \, dx_{i1} \, dx_{j1} \, dy \right)
\end{aligned}$$

$$\left(\int p_{l'_1}(x_{i2}, x_{j2}, y')p_{l'_2}(x_{i2}, x_{j2}, y')\,dx_{i2}\,dx_{j2}\,dy'\right)$$
$$= \|f\|_\infty^2 \sum_{l,l'} c_{ll'}^2$$

again using the orthonormality of the p_l's. Besides given the decomposition $p_l(x_i, x_j, y) = \alpha_{l_\alpha}(x_i, x_j)\beta_{l_\beta}(y)$,

$$\sum_{l,l'} c_{ll'}^2 = \int \sum_{l_\beta, l'_\beta} \beta_{l_\beta}(y)\beta_{l'_\beta}(y)\beta_{l_\beta}(y')\beta_{l'_\beta}(y') \sum_{l_\alpha, l'_\alpha} \left(\int \alpha_{l_\alpha}(x_{i1}, x_{j1})\alpha_{l'_\alpha}(x_{i2}, x_{j2})\eta(x_{i1}, x_{j2}, y)\,dx_{i1}\,dx_{j1}\,dx_{i2}\,dx_{j2}\right) \left(\int \alpha_{l_\alpha}(x_{i3}, x_{j3})\alpha_{l'_\alpha}(x_{i4}, x_{j4})\eta(x_{i3}, x_{j4}, y')\,dx_{i3}\,dx_{j3}\,dx_{i4}\,dx_{j4}\right) dy dy'$$

But

$$\sum_{l_\alpha, l'_\alpha} \left(\int \alpha_{l_\alpha}(x_{i1}, x_{j1})\alpha_{l'_\alpha}(x_{i2}, x_{j2})\eta(x_{i1}, x_{j2}, y)\,dx_{i1}\,dx_{j1}\,dx_{i2}\,dx_{j2}\right) \left(\int \alpha_{l_\alpha}(x_{i3}, x_{j3})\alpha_{l'_\alpha}(x_{i4}, x_{j4})\eta(x_{i3}, x_{j4}, y')\,dx_{i3}\,dx_{j3}\,dx_{i4}\,dx_{j4}\right)$$
$$= \sum_{l_\alpha, l'_\alpha} \int \alpha_{l_\alpha}(x_{i1}, x_{j1})\alpha_{l'_\alpha}(x_{i2}, x_{j2})\eta(x_{i1}, x_{j2}, y)\alpha_{l_\alpha}(x_{i3}, x_{j3})$$
$$\alpha_{l'_\alpha}(x_{i4}, x_{j4})\eta(x_{i3}, x_{j4}, y')\,dx_{i1}\,dx_{j1}\,dx_{i2}\,dx_{j2}\,dx_{i3}\,dx_{j3}\,dx_{i4}\,dx_{j4}$$
$$= \int \sum_{l_\alpha} \left(\int \alpha_{l_\alpha}(x_{i1}, x_{j1})\eta(x_{i1}, x_{j2}, y)\,dx_{i1}\,dx_{j1}\right)\alpha_{l_\alpha}(x_{i3}, x_{j3})$$
$$\sum_{l'_\alpha} \left(\int \alpha_{l'_\alpha}(x_{i4}, x_{j4})\eta(x_{i3}, x_{j4}, y')\,dx_{i4}\,dx_{j4}\right)\alpha_{l'_\alpha}(x_{i2}, x_{j2})\,dx_{i2}\,dx_{j2}\,dx_{i3}\,dx_{j3}$$
$$\leq \int \eta(x_{i3}, x_{j3}, x_{i2}, x_{j2}, y)\eta(x_{i3}, x_{j2}, y')\,dx_{i2}\,dx_{j2}\,dx_{i3}\,dx_{j3}$$
$$\leq \Delta^2\|\eta\|_\infty^2$$

using the orthonormality of the basis α_{l_α}. Then we get

$$\sum_{l,l'} c_{ll'}^2 \leq \Delta^2\|\eta\|_\infty^2 \left(\int \sum_{l_\beta, l'_\beta} \beta_{l_\beta}(y)\beta_{l'_\beta}(y)\beta_{l_\beta}(y')\beta_{l'_\beta}(y')\,dy\,dy'\right)$$

$$
\begin{aligned}
&= \Delta^2 \|\eta\|_\infty^2 \sum_{l_\beta, l'_\beta} \left(\int \beta_{l_\beta}(y) \beta_{l'_\beta}(y)\, dy \right)^2 \\
&\leq \Delta^2 \|\eta\|_\infty^2 \sum_{l_\beta} \left(\int \beta_{l_\beta}^2(y)\, dy \right)^2 \\
&\leq \Delta^2 \|\eta\|_\infty^2 m
\end{aligned}
$$

since the β_{l_β} are orthonormal. Finally

$$W_5 \leq \|f\|_\infty^2 \|\eta\|_\infty^2 \Delta^2 m.$$

Now for W_7 we first will bound,

$$
\begin{aligned}
|\sum_{l,l'} c_{ll'} a_l a_{l'}| &= |\int \sum_{l,l' \in M} a_l a_{l'} p_{l_2}(x_{i1}, x_{j1}, y) p_{l'_1}(x_{i2}, x_{j2}, y) \eta(x_{i1}, x_{j2}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy| \\
&\leq \int |S_M(x_{i1}, x_{j1}, y) S_M(x_{i2}, x_{j2}, y) \eta(x_{i1}, x_{j2}, y)|\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy \\
&\leq \|\eta\|_\infty \int \left(\int |S_M(x_{i1}, x_{j1}, y) S_M(x_{i2}, x_{j2}, y)|\, dy \right) dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}.
\end{aligned}
$$

Taking squares in both sides and using the Cauchy-Schwartz inequality twice, we get

$$
\begin{aligned}
&\left(\sum_{l,l'} c_{ll'} a_l a_{l'} \right)^2 \\
&= \|\eta\|_\infty^2 \left(\int \left(\int |S_M(x_{i1}, x_{j1}, y) S_M(x_{i2}, x_{j2}, y)|\, dy \right) dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2} \right)^2 \\
&\leq \|\eta\|_\infty^2 \Delta^2 \int \left(\int |S_M(x_{i1}, x_{j1}, y) S_M(x_{i2}, x_{j2}, y)|\, dy \right)^2 dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2} \\
&\leq \|\eta\|_\infty^2 \Delta^2 \int \left(\int S_M(x_{i1}, x_{j1}, y)^2\, dy \right) \left(\int S_M(x_{i2}, x_{j2}, y')^2\, dy' \right) dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2} \\
&= \|\eta\|_\infty^2 \Delta^2 \int S_M(x_{i1}, x_{j1}, y)^2 S_M(x_{i1}, x_{j1}, y')^2\, dx_{i1}\, dx_{j1}\, dx_{i2} dx_{j2}\, dy\, dy' \\
&= \|\eta\|_\infty^2 \Delta^2 \left(\int S_M(x_i, x_j, y)^2\, dx_i\, dx_j\, dy \right) \\
&\leq \|\eta\|_\infty^2 \Delta^2 \|f\|_\infty^2.
\end{aligned}
$$

Finally,

$$\mathbb{E}[\left(Q^\top(X_{i1}, X_{j1}, Y_1) C Q(X_{i2}, X_{j2}, Y_2)\right)^2] \leq \|\eta\|_\infty^2 \|f\|_\infty^2 \Delta^2 (m+1). \tag{2.20}$$

Collecting (2.19) and (2.20), we obtain

$$\mathrm{Var}(U_n K) \leq \frac{20}{n(n-1)} \|\eta\|_\infty^2 \|f\|_\infty^2 \Delta^2 (m+1)$$

which concludes the proof of Lemma 2.2. □

Lemma 2.3 (Bound for $\mathrm{Var}(P_n L)$)**.** *Under the assumptions of Theorem 2.3, we have*

$$\mathrm{Var}(P_n L) \leq \frac{12}{n} \|\eta\|_\infty^2 \|f\|_\infty^2 \Delta^2.$$

Proof. First note that given the independence of $\big(X_{ik}, X_{jk}, Y_k\big)$ $k = 1, \ldots, n$ we have

$$\mathrm{Var}(P_n L) = \frac{1}{n} \mathrm{Var}(L\big(X_{i1}, X_{j1}, Y_1\big))$$

we can write $L\big(X_{i1}, X_{j1}, Y_1\big)$ as

$$\begin{aligned}
&A^\top R\left(X_{i1}, X_{j1}, Y_1\right) + B^\top Q\left(X_{i1}, X_{j1}, Y_1\right) - 2A^\top C Q\left(X_{i1}, X_{j1}, Y_1\right) \\
&= \sum_{l \in M} a_l \left(\int p_l(x_i, x_j, Y_1) \psi(x_i, x_j, X_{i1}, X_{j1}, Y_1)\, dx_i\, dx_j - b_l \right) \\
&+ \sum_{l \in M} b_l \left(p_l(X_{i1}, X_{j1}, Y_1) - a_l \right) - 2 \sum_{l,l' \in M} c_{ll'} a_{l'} \left(p_l(X_{i1}, X_{j1}, Y_1) - a_l \right) \\
&= \int \sum_{l \in M} a_l p_l(x_i, x_j, Y_1) \psi(x_i, x_j, X_{i1}, X_{j1}, Y_1)\, dx_i\, dx_j \\
&+ \sum_{l \in M} b_l p_l(X_{i1}, X_{j1}, Y_1) - 2 \sum_{l,l' \in M} c_{ll'} a_{l'} p_l(X_{i1}, X_{j1}, Y_1) - 2A^t B - 2A^t C A. \\
&= \int S_M f(x_i, x_j, Y_1) \psi(x_i, x_j, X_{i1}, X_{j1}, Y_1)\, dx_i\, dx_j + S_M g(X_{i1}, X_{j1}, Y_1) \\
&- 2 \sum_{l,l' \in M} c_{ll'} a_{l'} p_l(X_{i1}, X_{j1}, Y_1) - 2A^\top B - 2A^\top C A.
\end{aligned}$$

Let $h(x_i, x_j, y) = \int S_M f(x_{i2}, x_{j2}, y) \psi(x_i, x_j, x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2}$, we have

$$\begin{aligned}
S_M h(x_i, x_j, y) &= \sum_{l \in M} \left(\int h(x_{i2}, x_{j2}, y) p_l(x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2}\, dy \right) p_l(x_i, x_j, y) \\
&= \sum_{l \in M} \Big(\int S_M f(x_{i3}, x_{j3}, y) \psi(x_{i2}, x_{j2}, x_{i3}, x_{j3}, y) \\
&\qquad p_l(x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2}\, dx_{i3}\, dx_{j3}\, dy \Big) p_l(x_i, x_j, y)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{l,l' \in M} \Big(\int a_{l'} p_{l'}(x_{i3}, x_{j3}, y) \psi(x_{i2}, x_{j2}, x_{i3}, x_{j3}, y) \\
&\quad p_l(x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2}\, dx_{i3}\, dx_{j3}\, dy \Big) p_l(x_i, x_j, y) \\
&= 2 \sum_{l,l' \in M} \Big(\int a_{l'} p_{l'}(x_{i3}, x_{j3}, y) \eta(x_{i2}, x_{j3}, y) \\
&\quad p_l(x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2}\, dx_{i3}\, dx_{j3}\, dy \Big) p_l(x_i, x_j, y) \\
&= 2 \sum_{l,l' \in M} a_{l'} c_{ll'} p_l(x_i, x_j, y)
\end{aligned}
$$

and we can write

$$
\begin{aligned}
L(X_{i1}, X_{j1}, Y_1) = h(X_{i1}, X_{j1}, Y_1) &+ S_M g(X_{i1}, X_{j1}, Y_1) \\
&- S_M h(X_{i1}, X_{j1}, Y_1) - 2A^\top B - 2A^\top C A.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\mathrm{Var}(L(X_{i1}, X_{j1}, Y_1)) \\
&= \mathrm{Var}(h(X_{i1}, X_{j1}, Y_1) + S_M g(X_{i1}, X_{j1}, Y_1) + S_M h(X_{i1}, X_{j1}, Y_1)) \\
&\leq \mathbb{E}[(h(X_{i1}, X_{j1}, Y_1) + S_M g(X_{i1}, X_{j1}, Y_1) + S_M h(X_{i1}, X_{j1}, Y_1))^2] \\
&\leq \mathbb{E}[(h(X_{i1}, X_{j1}, Y_1))^2 + (S_M g(X_{i1}, X_{j1}, Y_1))^2 + (S_M h(X_{i1}, X_{j1}, Y_1))^2].
\end{aligned}
$$

Each of these terms can be bounded

$$
\begin{aligned}
&\mathbb{E}[(h(X_{i1}, X_{j1}, Y_1))^2] \\
&= \int \left(\int S_M f(x_{i2}, x_{j2}, y) \psi(x_{i1} x_{j2}, x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2} \right)^2 f(x_{i1}, x_{j1}, y)\, dx_{i1}\, dx_{j1}\, dy \\
&\leq \Delta \int S_M f(x_{i2}, x_{j2}, y)^2 \psi(x_{i1} x_{j2}, x_{i2}, x_{j2}, y)^2 f(x_{i1}, x_{j1}, y)\, dx_{i1}\, dx_{j1}\, dx_{i2}\, dx_{j2}\, dy \\
&\leq 4\Delta^2 \|f\|_\infty \|\eta\|_\infty^2 \int S_M f(x_i, x_j, y)^2\, dx_i\, dx_j\, dy \\
&= 4\Delta^2 \|f\|_\infty \|\eta\|_\infty^2 \|S_M f\|_2^2 \\
&\leq 4\Delta^2 \|f\|_\infty \|\eta\|_\infty^2 \|f\|_2^2 \\
&\leq 4\Delta^2 \|f\|_\infty^2 \|\eta\|_\infty^2
\end{aligned}
$$

and similar calculations are valid for the others two terms,

$$
\mathbb{E}[(S_M g(X_{i1}, X_{j1}, Y_1))^2] \leq \|f\|_\infty \|S_M g\|_2^2 \leq \|f\|_\infty \|g\|_2^2 \leq 4\Delta^2 \|f\|_\infty^2 \|\eta\|_\infty^2
$$

$$\mathbb{E}[\big(S_M h\big(X_{i1}, X_{j1}, Y_1\big)\big)^2] \leq \|f\|_\infty \|S_M h\|_2^2 \leq \|f\|_\infty \|h\|_2^2 \leq 4\Delta^2 \|f\|_\infty^2 \|\eta\|_\infty^2.$$

Finally we get,

$$\mathrm{Var}(P_n L) \leq \frac{12}{n} \|\eta\|_\infty^2 \|f\|_\infty^2 \Delta^2.$$

□

Lemma 2.4 (Computation of $\mathrm{Cov}(U_n K, P_n L)$)**.** *Under the assumptions of Theorem 2.3, we have*

$$\mathrm{Cov}(U_n K, P_n L) = 0.$$

Proof of Lemma 2.4. Since $U_n K$ and $P_n L$ are centered, we have

$$\begin{aligned}
&\mathrm{Cov}(U_n K, P_n L) \\
&= \mathbb{E}[U_n K P_n L] \\
&= \mathbb{E}\left[\frac{1}{n^2(n-1)} \sum_{k \neq k'=1}^{n} K\big(X_{ik}, X_{jk}, Y_k, X_{ik'}, X_{jk'}, Y_{k'}\big) \sum_{k=1}^{n} L\big(X_{ik}, X_{jk}, Y_k\big)\right] \\
&= \frac{1}{n} \mathbb{E}[K\big(X_{i1}, X_{j1}, Y_1, X_{i2}, X_{j2}, Y_2\big)\big(L\big(X_{i1}, X_{j1}, Y_1\big) + L\big(X_{i2}, X_{j2}, Y_2\big)\big)] \\
&= \frac{1}{n} \mathbb{E}\big[\big(Q^\top(X_{i1}, X_{j1}, Y_1) R(X_{i2}, X_{j2}, Y_2) - Q^\top(X_{i1}, X_{j1}, Y_1) C Q(X_{i2}, X_{j2}, Y_2)\big) \\
&\big(A^\top R(X_{i1}, X_{j1}, Y_1) + B^\top Q(X_{i1}, X_{j1}, Y_1) - 2A^\top C Q(X_{i1}, X_{j1}, Y_1) \\
&+ A^\top R(X_{i2}, X_{j2}, Y_2) + B^\top Q X_{i2}, X_{j2}, Y_2) - 2A^\top C Q(X_{i2}, X_{j2}, Y_2)\big)\big] \\
&= 0.
\end{aligned}$$

Since K, L, Q and R are centered. □

Lemma 2.5 (Bound of $\mathrm{Bias}(\hat{\theta}_n)$)**.** *Under the assumptions of Theorem 2.3, we have*

$$|\mathrm{Bias}(\hat{\theta}_n)| \leq \Delta \|\eta\|_\infty \sup_{l \notin M} |c_l|^2.$$

Proof.

$$\begin{aligned}
|\mathrm{Bias}\, \hat{\theta}_n| \leq \|\eta\|_\infty \int \left(\int |S_M f(x_{i1}, x_{j1}, y) - f(x_{i1}, x_{j1}, y)|\, dx_{i1}\, dx_{j1}\right) \\
\left(\int |S_M f(x_{i2}, x_{j2}, y) - f(x_{i2}, x_{j2}, y)|\, dx_{i2}\, dx_{j2}\right) dy
\end{aligned}$$

$$
\begin{aligned}
&= \|\eta\|_\infty \int \left(\int |S_M f(x_i, x_j, y) - f(x_i, x_j, y)| \, dx_i \, dx_j \right)^2 dy \\
&\leq \Delta \|\eta\|_\infty \int \left(S_M f(x_i, x_j, y) - f(x_i, x_j, y) \right)^2 dx_i \, dx_j \, dy \\
&= \Delta \|\eta\|_\infty \sum_{l, l' \notin M} a_l a_{l'} \int p_l(x_i, x_j, y) p_{l'}(x_i, x_j, y) \, dx_i \, dx_j \, dy \\
&= \Delta \|\eta\|_\infty \sum_{l \notin M} |a_l|^2 \leq \Delta \|\eta\|_\infty \sup_{l \notin M} |c_l|^2.
\end{aligned}
$$

We use the Hölder's inequality and the fact that $f \in \mathcal{E}$ then $\sum_{l \notin M} |a_l|^2 \leq \sup_{l \notin M} |c_l|^2$. □

Lemma 2.6 (Asymptotic variance of $\sqrt{n}(P_n L)$.)**.** *Under the assumptions of Theorem 2.3, we have*

$$n \ \mathrm{Var}(P_n L) \to \Lambda(f, \eta)$$

where

$$
\begin{aligned}
\Lambda(f, \eta) = \int g(x_i, x_j, y)^2 f(x_i, x_j, y) \, dx_i \, dx_j \, dy \\
- \left(\int g(x_i, x_j, y) f(x_i, x_j, y) \, dx_i \, dx_j \, dy \right)^2.
\end{aligned}
$$

Proof. We proved in Lemma 2.3 that

$$
\begin{aligned}
&\mathrm{Var}(L(X_{i1}, X_{j1}, Y_1)) \\
&= \mathrm{Var}(h(X_{i1}, X_{j1}, Y_1) + S_M g(X_{i1}, X_{j1}, Y_1) + S_M h(X_{i1}, X_{j1}, Y_1)) \\
&= \mathrm{Var}(A_1 + A_2 + A_3) \\
&= \sum_{k,l=1}^{3} \mathrm{Cov}(A_k, A_l).
\end{aligned}
$$

We claim that $\forall k, l \in \{1, 2, 3\}^2$, we have

$$
\begin{aligned}
&\Big| \mathrm{Cov}(A_k, A_l) \\
&- \epsilon_{kl} \left[\int g(x_i, x_j, y)^2 f(x_i, x_j, y) \, dx_i \, dx_j \, dy - \left(\int g(x_i, x_j, y) f(x_i, x_j, y) \, dx_i \, dx_j \, dy \right)^2 \right] \Big| \\
&\leq \lambda \left[\|S_M f - f\|_2 + \|S_M g - g\|_2 \right]
\end{aligned}
\tag{2.21}
$$

where

$$\epsilon_{kl} = \begin{cases} -1 & \text{if } k = 3 \text{ or } l = 3 \text{ and } k \neq l \\ 1 & \text{otherwise} \end{cases},$$

and λ depends only on $\|f\|_\infty$, $\|\eta\|_\infty$ and Δ. We will do the details only for the case $k = l = 3$ since the calculations are similar for others configurations.

$$\mathrm{Var}(A_3) = \int S_M^2 h\left(x_i, x_j, y\right) f(x_i, x_j, y)\, dx_i\, dx_j\, dy - \left(\int S_M h\left(x_i, x_j, y\right) f(x_i, x_j, y)\, dx_i\, dx_j\, dy\right)^2.$$

The computation will be done in two steps. We first bound the quantity by the Cauchy-Schwartz inequality

$$\begin{aligned}
&\Big|\int S_M^2 h(x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy - \int g(x_i, x_j, y)^2 f(x_i, x_j, y)\, dx_i\, dx_j\, dy\Big| \\
&\leq \int |S_M^2 h(x_i, x_j, y) f(x_i, x_j, y) - S_M^2 g(x_i, x_j, y) f(x_i, x_j, y)|\, dx_i\, dx_j\, dy \\
&+ \int |S_M^2 g(x_i, x_j, y) f(x_i, x_j, y) - g(x_i, x_j, y)^2 f(x_i, x_j, y)|\, dx_i\, dx_j\, dy \\
&\leq \|f\|_\infty \|S_M h + S_M g\|_2 \|S_M h - S_M g\|_2 + \|f\|_\infty \|S_M g + g\|_2 \|S_M g - g\|_2.
\end{aligned}$$

Using several times the fact that since S_M is a projection, $\|S_M g\|_2 \leq \|g\|_2$, the sum is bounded by

$$\begin{aligned}
&\|f\|_\infty \|h + g\|_2 \|h - g\|_2 + 2\|f\|_\infty \|g\|_2 \|S_M g - g\|_2 \\
&\leq \|f\|_\infty \left(\|h\|_2 + \|g\|_2\right) \|h - g\|_2 + 2\|f\|_\infty \|g\|_2 \|S_M g - g\|_2.
\end{aligned}$$

We saw previously that $\|g\|_2 \leq 2\Delta \|f\|_\infty^{1/2} \|\eta\|_\infty$ and $\|h\|_2 \leq 2\Delta \|f\|_\infty^{1/2} \|\eta\|_\infty$. The sum is then bound by

$$4\Delta \|f\|_\infty^{3/2} \|\eta\|_\infty \|h - g\|_2 + 4\Delta \|f\|_\infty^{3/2} \|\eta\|_\infty \|S_M g - g\|_2.$$

We now have to deal with $\|h - g\|_2$:

$$\begin{aligned}
&\|h - g\|_2^2 \\
&= \int \left(\int \left(S_M f(x_{i2}, x_{j2}, y) - f(x_{i2}, x_{j2}, y)\right) \psi(x_{i1}, x_{j1}, x_{i2}, x_{j2}, y) dx_{i2} dx_{j2}\right)^2 dx_{i1}\, dx_{j1}\, dy \\
&\leq \int \left(\int \left(S_M f(x_{i2}, x_{j2}, y) - f(x_{i2}, x_{j2}, y)\right)^2 dx_{i2}\, dx_{j2}\right)
\end{aligned}$$

$$\left(\int \psi^2(x_{i1}, x_{j1}, x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2}\right) dx_{i1}\, dx_{j1}\, dy$$
$$\leq 4\Delta^2 \|\eta\|_\infty^2 \|S_M f - f\|_2^2.$$

Finally this first part is bounded by

$$\left|\int S_M^2 h\,(x_i, x_j, y)\, f(x_i, x_j, y)\, dx_i\, dx_j\, dy - \int g(x_i, x_j, y)^2 f(x_i, x_j, y)\, dx_i\, dx_j\, dy\right|$$
$$\leq 4\Delta \|f\|_\infty^{3/2} \|\eta\|_\infty \left(2\Delta \|\eta\|_\infty \|S_M f - f\|_2 + \|S_M g - g\|_2\right).$$

Following with the second quantity

$$\left|\left(\int S_M h(x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy\right)^2 - \left(\int g(x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy\right)^2\right|$$
$$= \left|\left(\int \left(S_M h(x_i, x_j, y) - g(x_i, x_j, y)\right) f(x_i, x_j, y)\, dx_i\, dx_j\, dy\right)\right.$$
$$\left.\left(\int \left(S_M h(x_i, x_j, y) + g(x_i, x_j, y)\right) f(x_i, x_j, y)\, dx_i\, dx_j\, dy\right)\right|.$$

By using the Cauchy-Schwartz inequality, it is bounded by

$$\begin{aligned}
&\|f\|_2 \|S_M h - g\|_2 \|f\|_2 \|S_M h + g\|_2 \\
&\leq \|f\|_2^2 \left(\|h\|_2 + \|g\|_2\right) \left(\|S_M h - S_M g\|_2 + \|S_M g - g\|_2\right) \\
&\leq 4\Delta \|f\|_\infty^{3/2} \|\eta\|_\infty \left(\|h - g\|_2 + \|S_M g - g\|_2\right) \\
&\leq 4\Delta \|f\|_\infty^{3/2} \|\eta\|_\infty \left(2\Delta \|\eta\|_\infty \|S_M f - f\|_2 + \|S_M g - g\|_2\right)
\end{aligned}$$

using the previous calculations. Collecting the two inequalities gives (2.21) for $k = l = 3$. Finally, since by assumption $\forall t \in \mathbb{L}^2(d\mu)$, $\|S_M t - t\|_2 \to 0$ when $n \to \infty$ a direct consequence of (2.21) is

$$\begin{aligned}
&\lim_{n\to\infty} \mathrm{Var}(L(X_{i1}, X_{j1}, Y_1)) \\
&= \int g^2(x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy - \left(\int g(x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy\right)^2 \\
&= \Lambda(f, \eta).
\end{aligned}$$

We conclude by noticing that $\mathrm{Var}(\sqrt{n}(P_n L)) = \mathrm{Var}(L(X_{i1}, X_{j1}, Y_1))$. □

Lemma 2.7 (Asymptotics for $\sqrt{n}(\hat{Q} - Q)$)**.** *Under the assumptions of Theorem 2.1, we have*

$$\lim_{n\to\infty} n\,\mathbb{E}[\hat{Q} - Q]^2 = 0.$$

Proof. The bound given in (2.3) states that if $|M_n|/n \to 0$ we have

$$\begin{aligned}
&\left| n\,\mathbb{E}[(\widehat{Q} - Q)^2 | \hat{f}] - \left[\int \hat{g}(x_i, x_j, y)^2 f(x_i, x_j, y)\, dx_i\, dx_j\, dy \right.\right. \\
&\left.\left. - \left(\int \hat{g}(x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy \right)^2 \right] \right| \\
&\leq \gamma(\|f\|_\infty, \|\eta\|_\infty, \Delta) \left[\frac{|M_n|}{n} + \|S_M f - f\|_2 + \|S_M \hat{g} - \hat{g}\|_2 \right]
\end{aligned}$$

where

$$\hat{g}(x_i, x_j, y) = \int H_3(\hat{f}, x_i, x_j, x_{i2}, x_{j2}, y) f(x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2},$$

where we recall that $H_3(f, x_{i1}, x_{j1}, x_{i2}, x_{j2}, y) = H_2(f, x_{i1}, x_{j2}, y) + H_2(f, x_{i2}, x_{j1}, y)$ with

$$H_2(\hat{f}, x_{i1}, x_{j2}, y) = \frac{\big(x_{i1} - m_i(\hat{f}, y)\big)\big(x_{j2} - m_j(\hat{f}, y)\big)}{\int \hat{f}(x_i, x_j, y)\, dx_i\, dx_j}.$$

By deconditioning we get

$$\begin{aligned}
&\left| n\,\mathbb{E}[(\widehat{Q} - Q)^2] - \mathbb{E}\left[\int \hat{g}(x_i, x_j, y)^2 f(x_i, x_j, y)\, dx_i\, dx_j\, dy \right.\right. \\
&\left.\left. - \left(\int \hat{g}(x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy \right)^2 \right] \right| \\
&\leq \gamma(\|f\|_\infty, \|\eta\|_\infty, \Delta) \left[\frac{|M_n|}{n} + \|S_M f - f\|_2 + \mathbb{E}[\|S_M \hat{g} - \hat{g}\|_2] \right]
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}[\|S_{M_n}\hat{g} - \hat{g}\|_2] \leq& \mathbb{E}[\|S_M\hat{g} - S_M g\|_2] + \mathbb{E}[\|\hat{g} - g\|_2] + \mathbb{E}[\|S_{M_n} g - g\|_2] \\
&\leq 2\mathbb{E}[\|\hat{g} - g\|_2] + \mathbb{E}[\|S_{M_n} g - g\|_2]
\end{aligned}$$

where $g(x_i, x_j, y) = \int H_3(f, x_i, x_j, x_{i2}, x_{j2}, y) f(x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2}$. The second term converges to 0 since $g \in \mathbb{L}^2(dx\, dy\, dz)$ and $\forall t \in \mathbb{L}^2(dx\, dy\, dz)$, $\int (S_M t - t)^2\, dx\, dy\, dz \to 0$. Moreover

$$
\begin{aligned}
&\|\hat{g} - g\|_2^2 \\
&= \int \left[\hat{g}(x_i, x_j, y) - g(x_i, x_j, y)\right]^2 dx_i\, dx_j\, dy \\
&= \int \left[\int \left(H_3(\hat{f}, x_i, x_j, x_{i2}, x_{j2}, y) - H_3(f, x_i, x_j, x_{i2}, x_{j2}, y) \right) \right. \\
&\quad \left. f(x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2} \right]^2 dx_i\, dx_j\, dy \\
&\leq \int \left[\int \left(H_3(\hat{f}, x_i, x_j, x_{i2}, x_{j2}, y) - H_3(f, x_i, x_j, x_{i2}, x_{j2}, y) \right)^2 dx_{i2}\, dx_{j2} \right] \\
&\quad \left[\int f(x_{i2}, x_{j2}, y)^2\, dx_{i2}\, dx_{j2} \right] dx_i\, dx_j\, dy \\
&\leq \Delta \|f\|_\infty^2 \int \left(H_2(\hat{f}, x_i, x_j, x_{i2}, x_{j2}, y) - H_2(f, x_i, x_j, x_{i2}, x_{j2}, y) \right)^2 dx_i\, dx_j\, dx_{i2}\, dx_{j2}\, dy \\
&\leq \delta \Delta^2 \|f\|_\infty^2 \int \left(\hat{f}(x_i, x_j, y) - f(x_i, x_j, y) \right)^2 dx_i\, dx_j\, dy
\end{aligned}
$$

for some constant δ that comes out of applying the mean value theorem to $H_3(\hat{f}, x_i, x_j, x_{i2}, x_{j2}, y) - H_3(f, x_i, x_j, x_{i2}, x_{j2}, y)$. The constant δ was taken under Assumptions 2.1-2.3. Since $\mathbb{E}[\|f - \hat{f}\|_2] \to 0$ then $\mathbb{E}[\|g - \hat{g}\|_2] \to 0$. Now show that the expectation of

$$
\int \hat{g}(x_i, x_j, y)^2 f(x_i, x_j, y)\, dx_i\, dx_j\, dy - \left(\int \hat{g}(x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j\, dy \right)^2
$$

converges to 0. We develop the proof for only the first term. We get

$$
\begin{aligned}
&\left| \int \hat{g}(x_i, x_j, y)^2 f(x_i, x_j, y)\, dx_i\, dx_j\, dy - \int g(x_i, x_j, y)^2 f(x_i, x_j, y)\, dx_i\, dx_j\, dy \right| \\
&\leq \int |\hat{g}(x_i, x_j, y)^2 - g(x_i, x_j, y)^2| f(x_i, x_j, y)\, dx_i\, dx_j\, dy \\
&\leq \lambda \int \left(\hat{g}(x_i, x_j, y) - g(x_i, x_j, y)\right)^2 dx_i\, dx_j\, dy \\
&= \lambda \|\hat{g} - g\|_2^2
\end{aligned}
$$

for some constant λ. By taking the both sides expectation, we see it is enough to show that $\mathbb{E}[\|\hat{g} - g\|_2^2] \to 0$. Besides, we can verify

$$\begin{aligned} g(x_i, x_j, y) &= \int H_3(f, x_i, x_j, x_{i2}, x_{j2}, y) f(x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2} \\ &= \frac{2}{\int f(x_i, x_j, y)\, dx_i\, dx_j} (x_i - \hat{m}_i(y)) \\ &\left(\int x_{j2} f(x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2} - \hat{m}_j(y) \int f(x_{i2}, x_{j2}, y)\, dx_{i2}\, dx_{j2} \right) \\ &= 0 \end{aligned}$$

which proves that the expectation of $\int \hat{g}(x_i, x_j, y)^2 f(x_i, x_j, y)\, dx_i\, dx_j$ converges to 0. Similar computations shows that the expectation of $(\int \hat{g}(x_i, x_j, y) f(x_i, x_j, y)\, dx_i\, dx_j)^2$ also converges to 0.

Finally we have

$$\lim_{n \to \infty} n\, \mathbb{E}[\widehat{Q} - Q]^2 = 0.$$

□

Chapter 3

Rates of convergence in conditional covariance matrix estimation

joint work with J-M. Loubes[1] and C. Marteau[2].

Abstract: Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ two random variables. In this chapter we are interested in the estimation of the conditional covariance matrix $\mathrm{Cov}\left(\mathbb{E}\left[\boldsymbol{X}|Y\right]\right)$. To this end, we will use a plug-in kernel based algorithm. Then, we investigate the related performance under smoothness assumptions on the density function of $(\boldsymbol{X}, Y)$. Moreover, in high-dimensional context, we shall improve our estimator to avoid inconsistency issues. In the matrix case, the convergence depends on some structural conditions over the $\mathrm{Cov}\left(\mathbb{E}\left[\boldsymbol{X}|Y\right]\right)$.

Keywords: Conditional covariance, Frobenius norm, Hölder functional class, nonparametric estimator, parametric rate.

3.1 Introduction

[1]Institut de Mathématiques de Toulouse, Université Paul Sabatier, Toulouse, France.
[2]Institut National des Sciences Appliquées de Toulouse, Toulouse, France.

Given a couple of random variables $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$, the conditional covariance matrix

$$\Sigma = \mathrm{Cov}\left(\mathbb{E}\left[\boldsymbol{X}|Y\right]\right) = \left(\sigma_{ij}\right)_{p\times p}, \tag{3.1}$$

plays a key role in reduction dimension techniques. For example, the sliced inverse regression method is a popular one based on the accurate estimation of the matrix Σ. See for instance Li (1991a). On Section 3.5 we will give more details about this application.

According to the statistical literature, we can use parametric and nonparametric techniques to build an estimator of the matrix Σ. For instance, Hsing (1999) tackles the nearest neighbor and the sliced inverse regression, Setodji and Cook (2004) use a version of k-means algorithm and Cook and Ni (2005) transform the sliced inverse regression into a least square minimization program. Usual non parametric methods involving kernel estimators are fully used to model Σ as explained in, for instance, Ferré and Yao (2003), Ferré et al. (2005), Zhu and Fang (1996) among others. From a parametric point of view Bura and Cook (2001) assume that $\mathbb{E}\left[\boldsymbol{X}|Y\right]$ has some parametric form while Da Veiga et al. (2011) use a functional Taylor approximation on $\mathrm{Cov}\left(\mathbb{E}\left[\boldsymbol{X}|Y\right]\right)$ to build an efficient estimate.

The aim of this chapter is to build an estimator of Σ when the joint density of $(\boldsymbol{X}, Y)$ is unknown. For this, we will plug an estimate of the marginal density of the observations into a parametric estimator of the conditional covariance and study its asymptotic behavior. Under some smoothness assumptions, we will prove that this density can be considered as a nuisance parameter which does not hamper the rate of the covariance. We face issues for the estimation of the covariance matrix, which arise in the high-dimensional context. Specifically, if p is larger than n the empirical covariance matrix has unexpected features like the lack of consistency or the significant spreading of the eigenvalues. In a slightly different context, we refer to Marčenko and Pastur (1967), Johnstone (2001) and references therein.

Hence regularization methods are necessary to get a consistent estimator for the sample covariance matrix. Such estimation techniques include: banding methods in Wu and Pourahmadi (2009) and Bickel and Levina (2008b), tapering in Furrer and Bengtsson (2007) and Cai et al. (2010), thresholding in Bickel and Levina (2008a) and El Karoui (2008), penalized estimation in Huang (2006), Lam and Fan (2009) and Rothman et al. (2008), regularizing principal components in Zou et al. (2006).

In this chapter, we will use the nonparametric estimator used by Zhu and Fang (1996) to compute entrywise the elements of $\Sigma = (\sigma_{ij})$. More precisely, the marginal density of Y and the vector of conditionals densities $\mathbb{E}[\boldsymbol{X}|Y]$ are unknown and we will estimate them by a kernel estimator. Then, we will propose a new estimator of the conditional covariance matrix Σ based on a plug-in version of the banding estimator. We will employ the normalized Frobenius norm to measure the squared risk over a given class of matrices.

We will prove, provided that the model is regular enough, that it is possible to obtain a pointwise parametric rate of convergence for the estimator of σ_{ij}. A parametric behavior is also found for the estimator of Σ with respect to the Frobenius norm. In these cases, the estimation of the conditional covariance matrix Σ turns into an efficient semiparametric issue.

This chapter falls into the following parts. Section 3.2 describes the nonparametric algorithm introduced by Zhu and Fang (1996) to estimate entry-wise the matrix defined in (3.1). In Section 3.3.1, we present all the required assumptions to assure the consistency and convergence of our estimator. The core of this article is in Section 3.3.2 where we provide the convergence rate for the element-wise estimator of Σ. We extend, in Section 3.4, our study for whole matrix assuming some structure on Σ. Section 3.5 is devoted to the relation between the matrix Σ and the sliced inverse regression method. In Section 3.5.1 we run some simulations comparing them with the classic the sliced inverse regression algorithm. Finally, the conclusions of this work are drawn in Section 3.6. All the technical proofs are gathered in Section 3.7.

3.2 Methodology

Let $\boldsymbol{X} \in \mathbb{R}^p$ be a random vector and $Y \in \mathbb{R}$ be a random variable. We denote by $f(x, y)$ the joint density of the couple $(\boldsymbol{X}, Y)$. Let $f_Y(\cdot) = \int_{\mathbb{R}^p} f(x, \cdot)dx$ be the marginal density function with respect to Y.

Suppose that $\boldsymbol{X}_k^\top = (X_{1k}, \ldots, X_{pk})$ and Y_k, $k = 1, \ldots, n$ are i.i.d. observations from the random vector $\boldsymbol{X}^\top = (X_1, \ldots, X_p)$ and Y respectively. Without loss of generality, we suppose $\mathbb{E}[X_i] = 0,\ k = 1, \ldots, p$.

Our aim is to estimate, based on the sample $(\boldsymbol{X}_k, Y_k)$'s, the covariance matrix

$$\Sigma = \big(\sigma_{ij}\big)_{i,j=1,\ldots,p} = \big(\mathrm{Cov}(\mathbb{E}(X_i|Y), \mathbb{E}(X_j|Y))\big)_{i,j=1,\ldots,p}.$$

For the sake of convenience, we introduce the following notation,

$$\begin{aligned} R_i(Y) &= \mathbb{E}(X_i|Y) \\ g_i(Y) &= R_i(Y) f_Y(Y) \\ &= \int x_i f(x_i, Y) dx_i \quad \forall i \in \{1, \ldots, p\}. \end{aligned} \tag{3.2}$$

For any $i, j \in \{1, \ldots, p\}$, the (i, j)-entry of the matrix Σ can then be written as

$$\sigma_{ij} = \mathbb{E}[\mathbb{E}[X_i|Y]\mathbb{E}[X_i|Y]] = \mathbb{E}\left[\frac{g_i(Y)g_j(Y)}{f_Y^2(Y)}\right]. \tag{3.3}$$

Equation (3.3) has two functions to estimate, $g_i(\cdot)$ and $f_Y(\cdot)$. We will estimate them using a nonparametric method (based on the work of Zhu and Fang (1996)) in an inverse regression framework.

Firstly, assume that we know $f_Y(\cdot)$. Denote $\hat{\sigma}_{ij}^{f_Y}$ the following estimator of σ_{ij},

$$\hat{\sigma}_{ij}^{f_Y} = \frac{1}{n}\sum_{k=1}^{n} \frac{\hat{g}_i(Y_k)\hat{g}_j(Y_k)}{f_Y^2(Y_k)}, \tag{3.4}$$

where

$$\hat{g}_i(Y) = \frac{1}{nh}\sum_{l=1}^{n} X_{il} K\left(\frac{Y - Y_l}{h}\right).$$

Here, $K(u)$ is a kernel function that satisfies some assumptions that we will make precise later and h is a bandwidth depending on n.

The next step is to replace the function $f_Y(\cdot)$, in equation (3.4), by the nonparametric estimator

$$\hat{f}_Y(Y) = \frac{1}{nh}\sum_{l=1}^{n} K\left(\frac{Y - Y_l}{h}\right).$$

The drawback with this approach is the observation superposition. In other words, we need the whole sample twice to estimate first the $\hat{g}_i$'s and then $\hat{f}_Y$. We can have dependency issues in the proper estimation of these functions.

To prevent any further difficulties, we will use a subsample of size $n_1 < n$ to estimate $\hat{g}_j$, and the remaining data $n_2 = n - n_1$ to estimate $\hat{f}_Y$. For the sake of simplicity, we choose $n_1 = n_2 = n/2$. Moreover, we will remove all the repeated observations in the $\hat{g}_i$'s estimation. This assumption entails the mutually independence between the $\hat{g}_i$'s and $\hat{f}_Y$.

In this context, given the bandwidths h_1 and h_2 depending on n_1 and n_2 respectively, we estimate Σ by $\hat{\Sigma} = (\hat{\sigma}_{ij})_{p\times p}$ where for all $i, j \in \{1, \dots, p\}$

$$\tilde{\sigma}_{ij} = \frac{1}{n_1} \sum_{k=1}^{n_1} \frac{\hat{g}_i(Y_k)\hat{g}_j(Y_k)}{\hat{f}_Y^2(Y_k)} \tag{3.5}$$

and

$$\hat{g}_i(Y) = \frac{1}{(n_1 - 1)h_1} \sum_{\substack{l=1 \\ l\neq k}}^{n_1} X_{il} K\left(\frac{Y - Y_l}{h_1}\right),$$
$$\hat{f}_Y(Y) = \frac{1}{n_2 h_2} \sum_{l=1}^{n_2} K\left(\frac{Y - Y_l}{h_2}\right).$$

However, to avoid any trouble due to small values in the denominator, let $b > 0$ a sequence of values satisfying

$$\hat{f}_{Y,b}(y) = \max\{\hat{f}_Y(y), b\}.$$

Besides, we assume that $0 < \eta < f_Y(y)$. Then, we propose the following estimator for σ_{ij},

$$\hat{\sigma}_{ij,K} = \frac{1}{n_1} \sum_{k=1}^{n_1} \frac{\hat{g}_i(Y_k)\hat{g}_j(Y_k)}{\hat{f}_{Y,b}^2(Y_k)}. \tag{3.6}$$

The estimator defined (3.6) relies on an unknown nonparametric density distribution function $f_Y(y)$. Hence, we are dealing with a semiparametric framework. Our aim is to study under which conditions the density of Y is a mere blurring parameter, that does not play any role in the estimation procedure. In this case, the plug-in method does not hamper the estimation rate of the conditional covariance Σ, i.e. leading to an efficient estimation rate. In the next section we shall establish the rate of convergence for the mean squared risk of $\hat{\sigma}_{ij,K}$.

3.3 Pointwise performance for $\hat{\sigma}_{ij,K}$

3.3.1 Assumptions

We denote C, C_1, C_2 and so on, constants (independent of n) that may take different values throughout all the chapter.

Assume that $f(x,y)$ has compact support. Let β be a positive real value and define $\lfloor\beta\rfloor$ as the largest integer such that $\lfloor\beta\rfloor \leq \beta$. We define the continuous kernel function $K(\cdot)$ of order $\lfloor\beta\rfloor$ to every function satisfying the following three conditions:

(a) the support of $K(\cdot)$ is the interval $[-1,1]$;

(b) $K(\cdot)$ is symmetric around 0;

(c) $\int_{-1}^{1} K(u)du = 1$ and $\int_{-1}^{1} u^k K(u)du = 0$ for $k = 1,\ldots,\lfloor\beta\rfloor$.

To guarantee a parametric consistency in our model, we need to impose some regularity conditions. In our case, define the Hölder class of smooth functions as follows.

Definition 3.1. Denote as $\mathcal{H}(\beta,L)$ the *Hölder class* of density functions with smoothness $\beta > 0$ and radius $L > 0$, defined as the set of $\lfloor\beta\rfloor$ times differentiable functions $\phi : T \to \mathbb{R}$ where T is an interval in $\mathbb{R}$, whose derivative $\phi^{(\lfloor\beta\rfloor)}$ satisfies

$$\left|\phi^{(\lfloor\beta\rfloor)}(x) - \phi^{(\lfloor\beta\rfloor)}(x')\right| \leq L\left|x - x'\right|^{\beta-\lfloor\beta\rfloor},\ \forall x,x' \in T.$$

The following assumption will be used recurrently in the proofs:

Assumption 3.1. *For x fixed, the function $f(x,y)$ belongs to a Hölder class of regularity β and constant L, i.e. $f(x,\cdot) \in \mathcal{H}(\beta,L)$.*

Moreover, $f_Y(y)$ belongs to a Hölder class of smoothness $\beta' > \beta$ and radius L', i.e. $f_Y \in \mathcal{H}(\beta',L')$.

Remark 3.1. Notice that function g_i defined in (3.2) also belongs to $\mathcal{H}(\beta,L)$ for $i = 1,\ldots,p$. Recall that

$$g_i(y) = \int x_i f(x_i,y)dx_i.$$

Let $x_i \in \mathbb{R}$ fixed. If $f(x_i,\cdot) \in \mathcal{H}(\beta,L)$ then by a direct calculation we can prove our assertion.

Denote as $\mathcal{F} = \mathcal{F}_\beta(L)$ the class of functions that fulfill Assumption 3.1. In the next section, we shall found the rates of convergence for $\hat{\sigma}_{ij,K}$ depending on the parameter β.

Remark 3.2. To control the term $\hat{f}_{Y,b}$ in the denominator on equation (3.6), we need to fix h_2 and b—both sequences converging to zero—to ensure the convergence of $\hat{\sigma}_{ij,K}$. As $n \to \infty$, set $h_2 \sim n^{-c_1}$ and $b \sim n^{-c_2}$ with the positives number c_1 and c_2 satisfying that $c_1/\beta < c_2 < 1/2 - c_1$, and the notation "$\sim$" means that two quantities have the same convergence order.

3.3.2 Rate of convergence for the matrix entries estimates

We derive in this section the risk upper bound for the element-wise estimator of (3.1) defined in (3.5).

Theorem 3.1. *Assume that* $\mathbb{E}\,|X_i|^4 < \infty$, $i = 1, \ldots, p$. *The upper bound risk of the estimator* $\hat{\sigma}_{ij,K}$ *defined in* (3.5) *over the functional class* $\mathcal{F}$ *satisfies:*

$$\sup_{\mathcal{F}} \mathbb{E}[(\hat{\sigma}_{ij,K} - \sigma_{ij})^2] \leq C_1 h_1^{2\beta} + \frac{C_2 \log^4 n_1}{n_1^2 h_1^4}.$$

In particular,

- *if* $\beta \geq 2$ *and choosing* $n_1^{-1/4} \leq h_1 \leq n_1^{-1/2\beta}$ *then*

$$\sup_{\mathcal{F}} \mathbb{E}[(\hat{\sigma}_{ij,K} - \sigma_{ij})^2] \leq \frac{C}{n_1}. \tag{3.7}$$

- *if* $\beta < 2$ *and choosing* $h_1 = n_1^{-1/(\beta+2)}$ *then,*

$$\sup_{\mathcal{F}} \mathbb{E}[(\hat{\sigma}_{ij,K} - \sigma_{ij})^2] \leq \left(\frac{\log^2(n_1)}{n_1}\right)^{2\beta/(\beta+2)}. \tag{3.8}$$

We provide the guideline for the proof of Theorem 3.1. The proofs of the auxiliary lemmas are postponed to the Appendix.

Proof. First consider the usual bias-variance decomposition.

$$\mathbb{E}\left[(\hat{\sigma}_{ij,K} - \sigma_{ij})^2\right] = \text{Bias}^2(\hat{\sigma}_{ij,K}) + \text{Var}(\hat{\sigma}_{ij,K})$$

where $\text{Bias}(\hat{\sigma}_{ij,K}) = \mathbb{E}[\hat{\sigma}_{ij,K}] - \sigma_{ij}$ and $\text{Var}(\hat{\sigma}_{ij,K}) = \mathbb{E}[\hat{\sigma}_{ij,K}^2] - \mathbb{E}[\hat{\sigma}_{ij,K}]^2$.

The following lemma provides a control of the bias of the estimator.

Lemma 3.1. *Under the same assumptions as Theorem 3.1 and supposing that $n_1 h_1 \to 0$ as $n_1 \to \infty$ and $n_2 h_2 \to 0$ as $n_1 \to \infty$, we have*

$$\text{Bias}^2(\hat{\sigma}_{ij,K}) \leq C_1 h_1^{2\beta} + \frac{C_2}{n_1^2 h_1^2}$$

for C_1 and C_2 positives constants depending only on L, s, β and on the kernel K.

Then, the next lemma gives an upper bound for the variance term

Lemma 3.2. *Under the same assumptions as Theorem 3.1 supposing that $n_1 h_1 \to 0$ as $n_1 \to \infty$ and $n_2 h_2 \to 0$ as $n_1 \to \infty$, we have*

$$\text{Var}\left(\hat{\sigma}_{ij,K}\right) \leq C_1 h_1^{2\beta} + \frac{C_2 \log^4 n_1}{n_1^2 h_1^4}$$

for C_1, C_2 and C_3 positives constants depending only on L, s, β and the kernel K.

Therefore, we obtain the following upper bound for the estimation error

$$\mathbb{E}[(\hat{\sigma}_{ij,K} - \sigma_{ij})^2] \leq C_1 h_1^{2\beta} + \frac{C_2 \log^4 n_1}{n_1^2 h_1^4}.$$

Depending on the regularity of the model, we consider two cases

- if $\beta \geq 2$ then we can choose h_1 such that

  $$\frac{1}{n_1^{1/4}} \leq h \leq \frac{1}{n_1^{1/2\beta}},$$

  then

  $$h^{2\beta} \leq \frac{1}{n_1}, \quad \frac{1}{n_1^2 h^4} \leq \frac{1}{n_1},$$

  concluding the result.

- Otherwise, if $\beta < 2$, we need to find h_1 such that

  $$h_1 = \underset{h}{\text{argmin}} \left(h^{2\beta} + \frac{\log^4(n_1)}{n_1^2 h^4} \right).$$

  We get $h_1 = (\log^2(n_1)/n_1)^{1/(\beta+2)}$ and the risk is bounded by

  $$\sup_{\mathcal{F}} \mathbb{E}[(\hat{\sigma}_{ij,K} - \sigma_{ij})^2] \leq \left(\frac{\log^2(n_1)}{n_1} \right)^{2\beta/(\beta+2)}.$$

□

From Theorem 3.1 we find an "elbow" effect in the rates of convergence. It means, we recover a parametric rate for regular enough functions $f_Y(y)$ and $g_i(y)$ $i = 1, \dots, p$. Otherwise, the mean squared error has a slower—indeed logarithmic—rate depending on the regularity of the functional class $\mathcal{F}$. This type of behaviour is common on functional analysis, for instance Donoho et al. (1996).

In other words, the problem can be solved in a semiparametric context as soon as the unknown functions $f_Y(y)$ and $g_i(y)$ for $i = 1, \dots, p$ are regular enough.

We have provided rates of convergence for any $\beta > 0$, particularly the n-consistency for $\beta \geq 2$. Additionally, the results obtained here are coherent with Zhu and Fang (1996). In their case, they have a n-consistency of the mean squared error assuming a regularity of $\beta = 4$ supporting our method.

Now, under some mild conditions, it seems natural to investigate the rate of convergence of the whole matrix estimator $\widehat{\Sigma}_K = (\hat{\sigma}_{ij,K})$. The next section will be dedicated to extend the result from Theorem 3.1 to find the rate of convergence of $\widehat{\Sigma}_K$ under the Frobenius norm.

3.4 Rate of convergence for the nonparametric covariance estimator

We have obtained in the previous section upper bounds of the quadratic risk related to the estimation of each coefficient of the matrix $\Sigma = \mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y]) = (\sigma_{ij})$. We have estimated them with $\widehat{\Sigma}_K = (\hat{\sigma}_{ij,K})$ where $\hat{\sigma}_{ij,K}$ is defined in equation (3.6). In this section, we shall extend this study to the whole matrix Σ.

There are several ways to measure the matrix mean squared error. In this work, we have chosen the Frobenius norm defined as follows,

Definition 3.2. The Frobenius norm of a matrix $A = (a_{ij})_{p \times p}$ is defined as the ℓ^2 vector norm of all entries in the matrix

$$\|A\|_F^2 = \sum_{i,j} a_{ij}^2.$$

In other words, this is equivalent to consider the matrix A as a vector of length p^2.

The mean squared error over the normalize Frobenius norm between $\widehat{\Sigma}_K$ and Σ is

$$\sup_{f\in\mathcal{F}} \frac{1}{p}\mathbb{E}\|\widehat{\Sigma}_K - \Sigma\|_F^2 \leq \frac{p}{n_1}.$$

This approach is impractical when $p \gg n_1$ since it causes the lost of convergence. Some references about this issue in several contexts are: Muirhead (1987), Johnstone (2001), Bickel and Levina (2008a), Bickel and Levina (2008b) and Fan et al. (2008).

To avoid consistency problems, we consider a modified version of $\widehat{\Sigma}_K$. We build this modification setting on zero the coefficients of the matrix from some point. This technique is also called "banding" and was studied by Bickel and Levina (2008b) for instance.

Formally, for a given integer m with $1 \leq m \leq p$, we define the banding estimator of $\widehat{\Sigma}_K$ as

$$\widehat{\Sigma}_{K,m} = (w_{ij}\hat{\sigma}_{ij,K})_{p\times p}, \tag{3.9}$$

where the function w_{ij} is defined as

$$w_{ij} = \begin{cases} 1, & \text{when } |i-j| \leq m, \\ 0, & \text{otherwise.} \end{cases}$$

If $p \gg n$, we require that Σ belongs to some space where it is sufficiently regular. Otherwise, it is not possible ensure any kind of convergence for $\widehat{\Sigma}_K$. The next assumption fixes Σ in a subset of the definite positive matrices.

Assumption 3.2. *The positive-definite covariance matrix Σ belongs to the following parameter space:*

$$\begin{aligned}\mathcal{G}_\alpha &= \mathcal{G}_\alpha(M_0, M1) \\ &= \{\Sigma : |\sigma_{ij}| \leq M_1|i-j|^{-(\alpha+1)} \text{ for } i \neq j \text{ and } \lambda_{max}(\Sigma) \leq M_0\}\end{aligned}$$

where $\lambda_{max}(\Sigma)$ is the maximum eigenvalue of the matrix Σ, and $M_0 > 0$ and $M_1 > 0$.

Notation 3.1. Set $\mathcal{G}' = \mathcal{G}'_{\alpha,\beta}(L)$ as the functional class formed by the intersection between $\mathcal{F}_\beta(L)$ and $\mathcal{G}_\alpha$.

In our case, Assumption 3.2 defines a matrix space indexed by a regularity parameter α. This parameter α states a rate of decay for the conditional covariance as they move away from the diagonal. A detailed discussion over this

subject can be found in the articles of Bickel and Levina (2008b) and Cai et al. (2010).

The following theorem gives an upper bound for the rate of convergence of the estimate defined in (3.9) under the normalized Frobenius norm based on the sample $\{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$.

Theorem 3.2. *Assume that* $\mathbb{E}\left|X_i\right|^4 < \infty$, $k = 1, \ldots, p$. *The upper bound risk, under the Frobenius norm, of the estimator* $\widehat{\Sigma}_K = (\hat{\sigma}_{ij,K})$ *defined in* (3.5) *over the functional class* $\mathcal{G}'$ *satisfies:*

- *if* $\beta \geq 2$,

$$\sup_{\mathcal{G}'} \frac{1}{p} \mathbb{E} \|\widehat{\Sigma}_{K,m} - \Sigma\|_F^2 \leq \min \left\{ n_1^{-\frac{2\alpha+1}{2(\alpha+1)}}, \frac{p}{n_1} \right\} \tag{3.10}$$

  where $m = p$ *if* $n_1^{1/(2(\alpha+1))} > p$ *or* $m = n_1^{1/(2(\alpha+1))}$ *otherwise.*

- *if* $\beta < 2$,

$$\sup_{\mathcal{G}'} \frac{1}{p} \mathbb{E} \|\widehat{\Sigma}_{K,m} - \Sigma\|_F \leq \min \left\{ \left(\frac{\log^2 n_1}{n_1} \right)^{\frac{2\beta(2\alpha+1)}{(\beta+2)(2(\alpha+1))}}, p \left(\frac{\log^2 n_1}{n_1} \right)^{\frac{2\beta}{\beta+2}} \right\} \tag{3.11}$$

  where $m = p$ *if* $(\log^2 n_1 / n_1)^{-2\beta/(2(\alpha+1)(\beta+2))} > p$ *or* $m = (\log^2 n_1 / n_1)^{-2\beta/(2(\alpha+1)(\beta+2))}$ *otherwise.*

The minimum in equations (3.10) and (3.11) depend if p is smaller or greater than n_1. If $p \ll n_1$, we use the original covariance matrix $\widehat{\Sigma}_K$. Otherwise, it is necessary to regularize the estimator to conserve the consistency. For example, in the case $\beta \geq 2$, if $n_1^{1/(2(\alpha+1))} > p$, we are in a relative low-dimensional framework and we use the full matrix $\widehat{\Sigma}_K$. In other case, when $n_1^{1/(2(\alpha+1))} \leq p$, regularize the matrix is mandatory and we choose $m = n_1^{1/(2(\alpha+1))}$ to generate the matrix $\widehat{\Sigma}_{K,m}$ defined in (3.9). A similar analysis can be done if $\beta < 2$.

Proof of Theorem 3.2. For the estimator (3.9), we have

$$\mathbb{E} \|\widehat{\Sigma}_{K,m} - \Sigma\|_F^2 = \sum_{i,j=1}^{p} \mathbb{E}(w_{ij} \hat{\sigma}_{ij,K} - \sigma_{ij})^2.$$

Let $i, j \in \{1, \ldots, p\}$ be fixed. Then

$$\begin{aligned}\mathbb{E}(w_{ij}\hat{\sigma}_{ij,K} - \sigma_{ij})^2 &= w_{ij}^2\mathbb{E}[(\hat{\sigma}_{ij,K} - \sigma)^2] + (1 - w_{ij})^2\sigma_{ij}^2 \\ &\leq w_{ij}^2\gamma_{n_1} + (1 - w_{ij})^2\sigma_{ij}^2\end{aligned}$$

where $\gamma_{n_1}$is the rate (3.7) or (3.8) depending on the value of β. Furthermore,

$$\begin{aligned}\frac{1}{p}\mathbb{E}\|\widehat{\Sigma}_{K,m} - \Sigma\|_F &\leq \frac{1}{p}\sum_{\{(i,j)\colon |i-j|>m\}}\sigma_{ij}^2 + \frac{1}{p}\sum_{\{(i,j)\colon |i-j|\leq m\}}\gamma_{n_1} \\ &\equiv R_1 + R_2.\end{aligned}$$

Since the cardinality of $\{(i,j)\colon |i-j| \leq m\}$ is bounded by mp we have directly that $R_2 \leq Cm\gamma_{n_1}$.

Thus, using Assumption 3.2 we show that

$$\sup_{\mathcal{G}'}\frac{1}{p}\sum_{\{(i,j)\colon |i-j|>m\}}\sigma_{ij}^2 \leq Cm^{-2\alpha-1},$$

where $|\sigma_{ij}| \leq C_1|i-j|^{-(\alpha+1)}$ for all$j \neq i$. Thus,

$$\sup_{\mathcal{G}'}\frac{1}{p}\mathbb{E}\|\widehat{\Sigma}_{K,m} - \Sigma\|_F^2 \leq Cm^{-2\alpha-1} + Cm\gamma_{n_1} \leq C_2\gamma_{n_1}^{(2\alpha+1)/(2(\alpha+1))} \tag{3.12}$$

by choosing

$$m = \gamma_{n_1}^{-1/(2(\alpha+1))}$$

if $\gamma_{n_1}^{-1/(2(\alpha+1))} \leq p$. In the case of $\gamma_{n_1}^{-1/(2(\alpha+1))} > p$ we will choose $m = p$, then the bias part is 0 and consequently

$$\frac{1}{p}\mathbb{E}\|\widehat{\Sigma}_{K,m} - \Sigma\|_F \leq Cm\gamma_{n_1}. \tag{3.13}$$

Using the result of Theorem 3.1, we distinguish two cases depending on the regularity of the model. If $\beta \geq 2$ then we take $\gamma_{n_1} = 1/n_1$ and if $\beta < 2$ then $\gamma_{n_1} = (\log^2 n_1/n_1)^{2\beta/(\beta+2)}$. The result is obtained combining the latter with (3.12) and (3.13).

□

3.5 Application to dimension reduction

We introduced the sliced inverse regression method in the context of reduction dimension. In general, we want to predict the behavior of a quantity of interest. We have a multidimensional explanatory variables $\boldsymbol{X} \in \mathbb{R}^p$ and noisy measurements of this quantity denoted by $Y \in \mathbb{R}$. In a nonparametric context, we deal with the model

$$Y = \psi(X) + \varepsilon,$$

where ψ is a unknown function. It is known in the literature that as p increases, the quality of estimation of the function ψ deteriorates as well. This is called the curse of dimensionality.

To cope this issue, Li (1991a) proposed a methodology to reduce the dimensionality called sliced inverse reduction . He considered the following regression model

$$Y = \varphi\left(\nu_1^\top \boldsymbol{X}, \ldots, \nu_k^\top \boldsymbol{X}, \varepsilon\right)$$

where k is much less than p denoted as $k \ll p$, $\boldsymbol{X}$ is a $p-$dimensional random vector, the ν's are unknown vectors but fixed, ε is independent of $\boldsymbol{X}$ and φ is a $\mathbb{R}^{k+1}$ arbitrary real valued function. This model implies, via projection, the extraction of all the Y's relevant information by only a k-dimensional subspace generated by the ν's. These directions are called effective dimension reduction (EDR) directions.

The main idea of the sliced inverse regression method is to estimate the unknown matrix

$$\Sigma = \mathrm{Cov}\left(\mathbb{E}\left[\boldsymbol{X}|Y\right]\right) = \left(\sigma_{ij}\right)_{p\times p},$$

where we denote σ_{ij} the (i,j) matrix element. This matrix is degenerate in any direction orthogonal to the ν's. Therefore, the eigenvectors, ν_j $(j = 1, \ldots, k)$, associated with the largest k eigenvalues of Σ are the EDR directions. This lead the classical the sliced inverse regression method consisting in slicing the inverse regression curve, calculate the empirical covariance of this curve and then estimate the largest eigenvalues with its corresponding eigenvectors. The first k eigenvectors span the space.

Theorem 3.2 claims that if $\beta \geq 2$ and other mild conditions, the nonparametric estimator (3.9) behaves asymptotically as $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$. As consequence, the eigenvectors of $\widehat{\Sigma}_{K,m}$ estimates correctly the EDR directions in this context. In other words, we have an accurate characterization of the EDR space span by the eigenvectors of $\hat{\Sigma}_{K,m}$.

3.5.1 Simulation study

We consider numerical simulations to asses the performance of our estimator. We will use two different models for the simulations.

Linear case: Define $\alpha > 0$. Set the linear model,

$$\begin{aligned} X_i &= a_i Y + (1 - a_i)\epsilon, \\ a_i &= i^{-(\alpha+1)}, \quad i = 1, \ldots, p. \end{aligned}$$

The random variables Y and ϵ are independent having the standard normal distribution. This model provides the conditional covariance matrix $\Sigma = (\sigma_{ij})$ with,

$$\sigma_{ij} = \mathrm{Cov}(\mathbb{E}[X_i|Y]\mathbb{E}[X_j|Y]) = \mathrm{Var}(Y^2)(ij)^{-(\alpha+1)} = 2(ij)^{-(\alpha+1)}$$

Moreover, σ_{ij} satisfies Condition 3.2.

Polar case: Define the two dimensional model with independent random variables $r \in \mathbb{R}$ and $0 \leq \theta \leq \pi/4$,

$$\begin{aligned} X_1 &= 1.5\, r\cos(\theta) \\ X_2 &= 1.5\, r\sin(\theta) \\ Y &= X_1^2 + X_2^2 + \epsilon. \end{aligned}$$

We use r and ϵ with standard normal distributions and θ with a uniform distribution from 0 to $\pi/4$. In addition to these variables, we also generate $X_3, \ldots, X_p$, all variables being independent and following the standard normal distribution. The true conditional covariance matrix is,

$$\begin{aligned} \Sigma &= \mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y]) \\ &= 1.5^2\,\mathrm{Var}(r^2) \begin{pmatrix} \mathbb{E}[\cos\theta]^2 & \mathbb{E}[\cos\theta]\mathbb{E}[\sin\theta] & 0 & \cdots & 0 \\ \mathbb{E}[\cos\theta]\mathbb{E}[\sin\theta] & \mathbb{E}[\sin\theta]^2 & 0 & \cdots & 0 \\ 0 & 0 & & & \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & & & 0 \end{pmatrix}_{p\times p}. \end{aligned}$$

In both cases, we do not assume any structure for the conditional covariance structure in the estimation of $\hat{\Sigma}_{K,m}$. We choose the value $m = \lfloor n_1^{1/(2(\alpha+1))} \rfloor$ to

regularize the estimated matrix $\widehat{\Sigma}_K$. We set the threshold b in the first quantile of each estimation of $\hat{f}$.

We consider a range of parameter values for n, p and α. The simulation uses values of α equals to 0.05, 0.1 and 0.5. Both p and n varying between 100 to 600. To estimate the kernel functions $\hat{g}_i$ and $\hat{f}$ we use the software `R` along the library `np` ().

Table 3.1 reports the average error under the Frobenius norm over 100 replications for the linear and polar case respectively. The mean error decreases with a rate equivalent to p/n as n increases due to the regularity of the model. In other words, in both cases f_Y and g_i belong to a Hölder class with $\beta \geq 2$. These results highlight the convergence of $\hat{\Sigma}_{K,m}$ without any assumption in the data structure.

3.5.2 Graphical representations

We next examine a simple application to the projection of data in a reduced space. The models remains exactly the sames as before (linear and polar cases). In this section, we have generated 500 data points for each model over 200 variables.

We compare the projected points calculated by our method against the classical the sliced inverse regression method proposed by Li (1991a) implemented in the package `dr`.

The results in general are satisfactory. Figure 3.1 shows the data representation in the linear model. In this case, both methods capture well most of the model information. The polar case is presented in Figure 3.2. We observe how our method performs similar to the classic sliced inverse regression capturing the data curvature for the first direction.

In Figure 3.3 and Figure 3.4, we observe that our method behaves better with respect to the explained variance of the model. The nonparametric estimator explains around 60% of the variance with around 50 variables. In the meantime, the classic sliced inverse regression describes at most 40% of the variance with the same number of variables. We have set in 0 all the negative eigenvalues, therefore the curves in the nonparametric cases remains constant at the end.

3.6 Conclusion

In this chapter we have investigated the rate of convergence for a nonparametrical estimator for the conditional covariance $\mathrm{Cov}(\mathbb{E}[X|Y])$. We have started

		α		
p	n	**0.05**	**0.1**	**0.5**
100	100	0.0090	0.0073	0.0049
	200	0.0054	0.0050	0.0031
	300	0.0042	0.0039	0.0024
	400	0.0040	0.0035	0.0021
	500	0.0037	0.0033	0.0019
	600	0.0034	0.0031	0.0019
200	100	0.0070	0.0063	0.0041
	200	0.0037	0.0036	0.0022
	300	0.0028	0.0026	0.0016
	400	0.0024	0.0022	0.0013
	500	0.0022	0.0020	0.0011
	600	0.0021	0.0018	0.0010
300	100	0.0066	0.0065	0.0040
	200	0.0034	0.0034	0.0019
	300	0.0024	0.0021	0.0014
	400	0.0020	0.0018	0.0010
	500	0.0017	0.0015	0.0009
	600	0.0016	0.0014	0.0008
400	100	0.0059	0.0055	0.0041
	200	0.0029	0.0028	0.0018
	300	0.0021	0.0019	0.0012
	400	0.0017	0.0016	0.0009
	500	0.0015	0.0013	0.0008
	600	0.0013	0.0012	0.0007
500	100	0.0071	0.0052	0.0060
	200	0.0028	0.0028	0.0016
	300	0.0020	0.0018	0.0012
	400	0.0016	0.0014	0.0008
	500	0.0013	0.0012	0.0007
	600	0.0012	0.0011	0.0006
600	100	0.0062	0.0057	0.0036
	200	0.0027	0.0026	0.0016
	300	0.0019	0.0017	0.0011
	400	0.0015	0.0014	0.0008
	500	0.0012	0.0011	0.0007
	600	0.0011	0.0010	0.0006

(a) Linear case.

		α		
p	n	**0.05**	**0.1**	**0.5**
100	100	0.0513	0.0403	0.0336
	200	0.0178	0.0187	0.0155
	300	0.0139	0.0142	0.0116
	400	0.0118	0.0114	0.0104
	500	0.0110	0.0108	0.0098
	600	0.0106	0.0104	0.0096
200	100	0.0536	0.0508	0.0346
	200	0.0154	0.0161	0.0117
	300	0.0093	0.0097	0.0077
	400	0.0081	0.0077	0.0067
	500	0.0071	0.0070	0.0059
	600	0.0064	0.0063	0.0054
300	100	0.0594	0.0505	0.0251
	200	0.0145	0.0138	0.0103
	300	0.0094	0.0081	0.0075
	400	0.0068	0.0067	0.0051
	500	0.0057	0.0054	0.0045
	600	0.0050	0.0048	0.0042
400	100	0.0664	0.0411	0.0260
	200	0.0156	0.0118	0.0095
	300	0.0085	0.0080	0.0064
	400	0.0060	0.0058	0.0043
	500	0.0050	0.0048	0.0038
	600	0.0043	0.0042	0.0033
500	100	0.0377	0.0395	0.0350
	200	0.0129	0.0140	0.0092
	300	0.0076	0.0074	0.0054
	400	0.0060	0.0054	0.0041
	500	0.0048	0.0044	0.0034
	600	0.0040	0.0038	0.0029
600	100	0.0570	0.0430	0.0350
	200	0.0134	0.0132	0.0091
	300	0.0078	0.0071	0.0049
	400	0.0057	0.0052	0.0037
	500	0.0045	0.0042	0.0032
	600	0.0037	0.0034	0.0027

(b) Polar case.

Table 3.1: Average errors under the Frobenius norm of the nonparametric conditional covariance estimator over 100 replications.

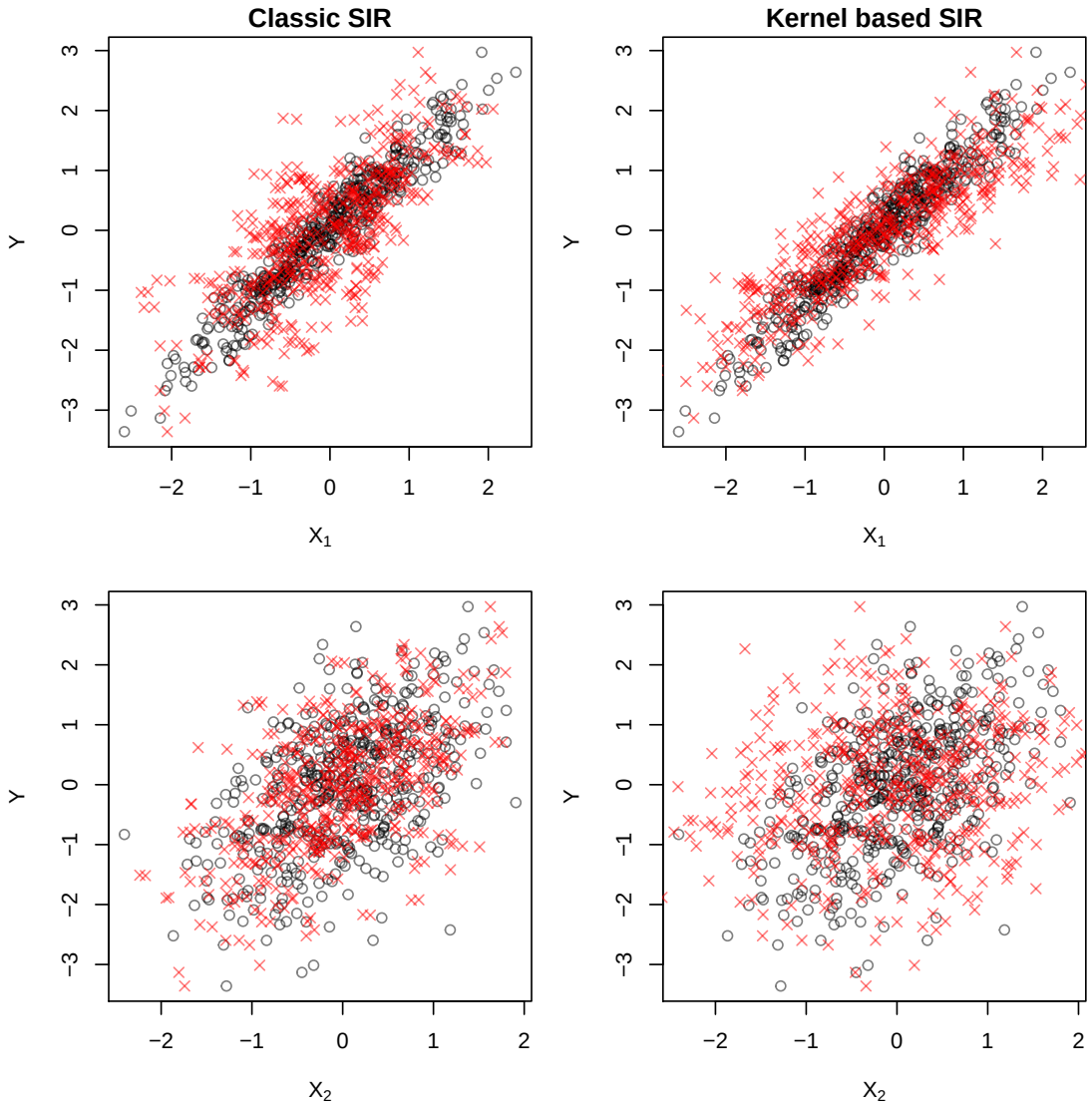


Figure 3.1: Linear case: Real data against projected data. The black circles represents the original data. The red crosses represent the projected points for the linear model.

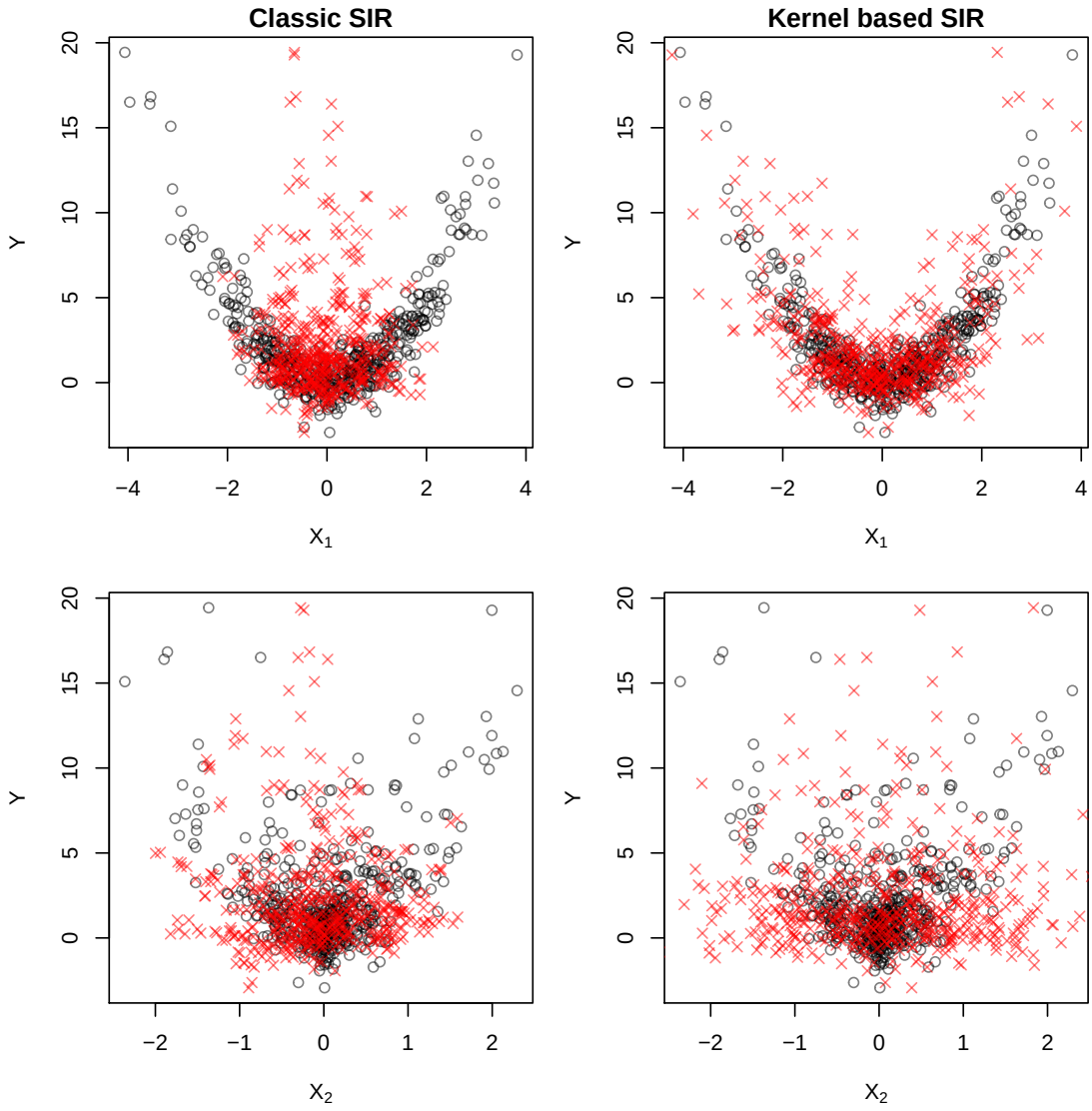


Figure 3.2: Polar case: Real data against projected data. The black circles represents the original data. The red crosses represent the projected points for the polar model.

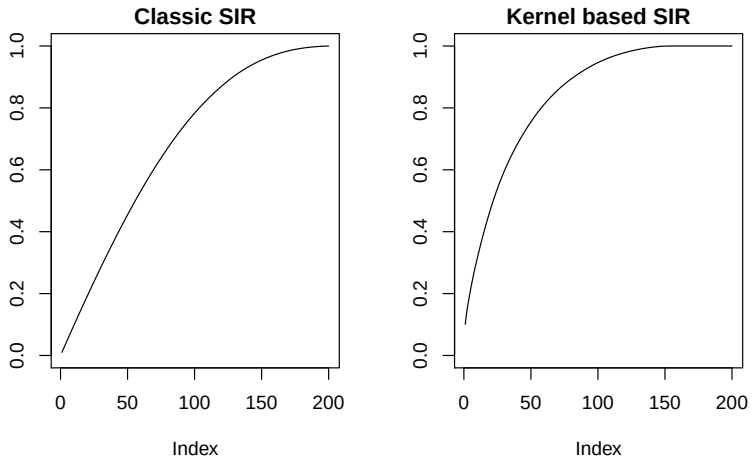


Figure 3.3: Linear case: Comparison of the cumulative variance explained between the nonparametric and classic sliced inverse regression methods

.

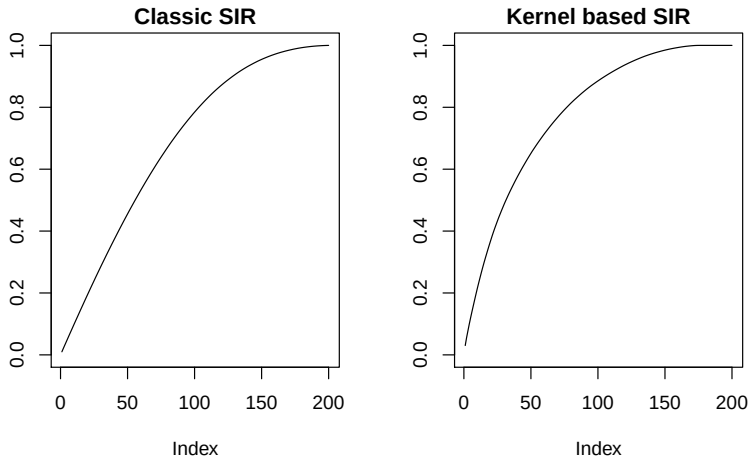


Figure 3.4: Polar case: Comparison of the cumulative variance explained between the nonparametric and classic sliced inverse regression methods

.

studying the nonparametric behavior of each element of the matrix based in the work of Zhu and Fang (1996). This approach, allow us to find rate of convergence depending on the value of β for the classical mean squared error. We have shown that if the model is regular enough, it is possible to achieve the parametric rate $1/n$. Otherwise, we get a slower rate $(\log^2(n)/n)^{2\beta/(\beta+2)}$, depending on the regularity β of certain Hölder class.

As a natural extension, we studied the performance of mean squared risk of $\widehat{\Sigma}_K$ under the Frobenius norm. In order to keep the consistency and avoid issues due to the high-dimension data, it is necessary regularize the matrix $\widehat{\Sigma}_K$. We used a regularized version of $\widehat{\Sigma}_K$ called $\hat{\Sigma}_{K,m}$, obtained by doing a Schur multiplication between $\widehat{\Sigma}_K$ and a definite-positive matrix of weights. Those weights are 1 until some point away from the diagonal when they turn in 0.

This method could not ensure the positive definiteness of the estimate—but since we proved that under some mild conditions, our estimator is consistent given either $p/n \to 0$ or $p(\log^2(n)/n)^{2\beta/(\beta+2)} \to 0$, the estimator will be positive definite with probability tending to 1. Other practical solution for this issue is suggested by Cai et al. (2010). He proposes project $\widehat{\Sigma}_K$ to the space of positive-semidefinite matrices under the operator norm. In other words, first diagonalize $\widehat{\Sigma}_K$ and replace the negative eigenvalues by 0. The matrix obtained is then semidefinite positive.

There are several alternatives norms to compute and matrix-based error, among them the operator norm. However, this approach would require the estimation of concentration inequalities for some specific matrices blocks of $\widehat{\Sigma}_K$ which is not suitable in our context. Nevertheless, the use of the operator norm in the nonparametric estimation of Σ is worth to investigate.

3.7 Appendix

3.7.1 Technical lemmas

Lemma 3.3 (cf. Rao and Prakasa Rao (1983), Theorem 2.1.8)**.** *Suppose that K is a kernel of order $s = \lfloor \beta \rfloor$ and Assumption 3.1 is fulfilled. Then*

$$\sup_y |\hat{f}(y) - f(y)| = O\left(h^\beta + \frac{\log n}{n^{1/2}h}\right).$$

The following lemma is a modified version of Theorem 2.37 of Pollard (1984).

Lemma 3.4. *Suppose that K is a kernel of order $s = \lfloor \beta \rfloor$, $\mathbb{E}[X_i^4] < \infty$ and Assumption 3.1 is fulfilled. Then for any $\varepsilon > 0$,*

$$\begin{aligned}\mathbb{P}\left(\sup_y |\hat{g}_i(y) - \mathbb{E}\left[\hat{g}_i(y)\right]| > 8n^{-1/2}h^{-1}\varepsilon\right) \\ \leq 2c\left(\frac{\varepsilon}{\sqrt{n}d}\right)^{-4}\exp\left\{-\frac{1}{2}\varepsilon^2/\left(32d(\log n)^{1/2}\right)\right\} \\ + 8cd^{-8}\exp\left(-nd^2\right) + \mathbb{E}\left[X_i^4\right] I\left(|X_i| > cd^{-1/2}(\log n)^{1/4}\right),\end{aligned}$$

where

$$d \geq \sup_y \left\{\mathrm{Var}\left(K\left(\frac{y-Y}{h}\right)\right)\right\}^{1/2}.$$

We refer to Zhu and Fang (1996) for the proof of Lemma 3.4. Using the last result the uniform convergence rate of $\hat{g}_i(y)$ can be obtained.

Lemma 3.5. *Suppose that K is a kernel of order $s = \lfloor \beta \rfloor$, $\mathbb{E}[X_i^4] < \infty$ and Assumption 3.1 is fulfilled. Then*

$$\sup_y |\hat{g}_i(y) - g_i(y)| = O_p\left(h^\beta + \frac{\log n}{n^{1/2}h}\right).$$

Proof. Since the kernel function K is uniformly continuous on $[-1, 1]$, writing $c_1 = \sup_{|u|\leq 1} |K(u)|$, we have

$$\sup_y \left(\mathrm{Var}\left(K\left(\frac{y-Y}{h}\right)\right)\right)^{1/2} \leq \sup_y \left(\int K^2\left(\frac{y-Y}{h}\right) f(Y)dY\right)^{1/2} \leq c_1.$$

Choose $\varepsilon = \log n$, then as $n \to \infty$, we have

$$\sup_y |\hat{g}_i(y) - \mathbb{E}\left(\hat{g}_i(y)\right)| = O_p\left(n^{-1/2}h^{-1}\log n\right).$$

On the contrary, we will expand $g_i(y)$ in a Taylor series with the Lagrange form of the remainder term (see Rao and Prakasa Rao (1983), page 47). Using Assumption 3.1 and Remark 3.1, for any $0 < \tau < 1$ and $s = \lfloor \beta \rfloor$ we have,

$$\sup_y |\mathbb{E}\left(\hat{g}_i(y)\right) - g_i(y)|$$

$$
\begin{aligned}
&= \sup_y \left| \int K_h(y-Y)\left\{g_i(Y)-g_i(y)\right\} dY \right| \\
&= \sup_y \left| \int K(u)\left\{g_i(y+uh)-g_i(y)\right\} du \right| \\
&= \sup_y \left| \int K(u)\left\{ g_i(y) + uhg_i'(y) + \cdots + (uh)^s \frac{g_i^{(s)}(y+\tau uh)}{s!} - g_i(y) \right\} du \right| \\
&= \sup_y \left| \int K(u) \frac{(uh)^s}{s!} \left(g_i^{(s)}(y+\tau uh) - g_i^{(s)}(y) \right) du \right|,
\end{aligned}
$$

as $g_i \in \mathcal{H}(\beta, L)$ we conclude that,

$$
\sup_y \left| \mathbb{E}\left(\hat{g}_i(y)\right) - g_i(y) \right| \leq c \int \left| u^\beta K(u) \right| du \cdot h^\beta. \qquad \square
$$

3.7.2 Proof of Lemmas

Proof of Lemma 3.1. The proof follows three steps.

Step 1: Prove that

$$
\mathbb{E}\left[\frac{\hat{g}_i(Y_1)\hat{g}_j(Y_1)}{\hat{f}^2_{Y,b}(Y_1)} \right] \leq \mathbb{E}\left[\frac{\hat{g}_i(Y_1)\hat{g}_j(Y_1)}{f^2_Y(Y_1)} \right] \left(1 + C_1 h_2^\beta + C_2 \frac{\log^2 n_2}{n_2 h_2^2} \right).
$$

Notice that $\hat{g}_i(Y_k)\hat{g}_j(Y_k)/\hat{f}^2_{Y,b}(Y_k)$ are not independent for $k = 1, \ldots, n_1$ but have the same distribution. Thus,

$$
\mathbb{E}[\hat{\sigma}_{ij,K}] = \mathbb{E}\left[\frac{1}{n_1} \sum_{k=1}^{n_1} \frac{\hat{g}_i(Y_k)\hat{g}_j(Y_k)}{\hat{f}^2_{Y,b}(Y_k)} \right] = \mathbb{E}\left[\frac{\hat{g}_i(Y_1)\hat{g}_j(Y_1)}{\hat{f}^2_{Y,b}(Y_1)} \right]
$$

Start by multiplying and dividing by $f^2_Y(Y)$ inside the expectation. Then, conditioning by respect Y_1 and using the independence between $\hat{g}_i$ and $\hat{f}_Y$ we can see that,

$$
\mathbb{E}\left[\frac{\hat{g}_i(Y_1)\hat{g}_j(Y_1)}{\hat{f}^2_{Y,b}(Y_1)} \right] = \mathbb{E}\left[\frac{\hat{g}_i(Y_1)\hat{g}_j(Y_1)}{f^2_Y(Y_1)} \mathbb{E}\left[\frac{f^2_Y(Y_1)}{\hat{f}^2_{Y,b}(Y_1)} \middle| Y_1 \right] \right].
$$

We observe that

$$
\frac{f^2_Y(y)}{\hat{f}^2_{Y,b}(y)} = 1 + 2\frac{(f_Y(y) - \hat{f}_{Y,b}(y))}{\hat{f}_{Y,b}(y)} + \frac{(f_Y(y) - \hat{f}_{Y,b}(y))^2}{\hat{f}_{Y,b}(y)},
$$

By Lemma 3.3 and Remark 3.2, equation (3.7.2) turns into

$$\begin{aligned}
&\mathbb{E}\left[\frac{\hat{g}_i(Y_1)\hat{g}_j(Y_1)}{\hat{f}^2_{Y,b}(Y_1)}\right] \\
&\leq \mathbb{E}\left[\frac{\hat{g}_i(Y_1)\hat{g}_j(Y_1)}{f^2_Y(Y_1)}\right]\left(1 + C_1 n^{c_1}\left(h_2^\beta + \frac{\log n_2}{n_2^{1/2}h_2}\right) + C_2 n^{2c_1}\left(h_2^\beta + \frac{\log n_2}{n_2^{1/2}h_2}\right)^2\right) \\
&\leq \mathbb{E}\left[\frac{\hat{g}_i(Y_1)\hat{g}_j(Y_1)}{f^2_Y(Y_1)}\right](1 + C_1(n_2^{c_1-\beta c_2} + n_2^{1/2-c_1-c_2}\log n_2)).
\end{aligned}$$

To simplify the notations, we will write simply $n = n_1$ and $h = h_1$ through Steps 2 to 4

Step 2: Prove that,

$$\begin{aligned}
\mathbb{E}\left[\frac{\hat{g}_i(Y_1)\hat{g}_j(Y_1)}{f^2_Y(Y_1)}\right] \leq \frac{C_1}{n-1}\mathbb{E}\left(\frac{X_{i2}X_{j2}K_h^2(Y_1-Y_2)}{f^2_Y(Y_1)}\right) \\
+ C_2\left(\frac{n-2}{n-1}\right)\mathbb{E}\left(\frac{X_{i2}X_{j3}K_h(Y_1-Y_2)K_h(Y_1-Y_3)}{f^2_Y(Y_1)}\right).
\end{aligned}$$

Denote

$$B = \mathbb{E}\left[\frac{\hat{g}_i(Y_1)\hat{g}_j(Y_1)}{f^2_Y(Y_1)}\right].$$

Again, conditioning with respect to Y_1 and developing all the terms, we obtain,

$$\begin{aligned}
B =& \frac{1}{(n-1)^2}\mathbb{E}\left[\frac{1}{f^2_Y(Y_1)}\mathbb{E}\left[\left(\sum_{k=2}^n X_{ik}K_h(Y_1-Y_k)\right)\left(\sum_{k=2}^n X_{jk}K_h(Y_1-Y_l)\right)\middle| Y_1\right]\right] \\
=& \frac{1}{(n-1)^2}\mathbb{E}\left[\frac{1}{f^2_Y(Y_1)}\mathbb{E}\left[\sum_{k=2}^n X_{ik}X_{jk}K_h^2(Y_1-Y_k)\,\middle|\, Y_1\right]\right] \\
&+ \frac{1}{(n-1)^2}\mathbb{E}\left[\frac{1}{f^2_Y(Y_1)}\mathbb{E}\left[\sum_{\substack{k,r=2\\k\neq r}}^n X_{ik}X_{jr}K_h(Y_1-Y_k)K_h(Y_1-Y_r)\,\middle|\, Y_1\right]\right].
\end{aligned}$$

Therefore,

$$\mathbb{E}[\hat{\sigma}_{ij,K}] = \frac{C_1}{n-1}\mathbb{E}\left[\frac{X_{i2}X_{j2}K_h^2(Y_1-Y_2)}{f_Y^2(Y_1)}\right] + C_2\left(\frac{n-2}{n-1}\right)\mathbb{E}\left[\frac{X_{i2}X_{j3}K_h(Y_1-Y_2)K_h(Y_1-Y_3)}{f_Y^2(Y_1)}\right] \equiv B_1 + B_2. \quad (3.14)$$

Step 3: Prove that

$$B_1 \leq \frac{C}{nh}.$$

We start with

$$\begin{aligned} B_1 &= \frac{1}{n-1}\mathbb{E}\left(\frac{X_{i2}X_{j2}K_h^2(Y_1-Y_2)}{f_Y^2(Y_1)}\right) \\ &= \frac{1}{n-1}\int \frac{1}{f_Y^2(y_1)} x_i x_j K_h^2(y_1-y_2) f(x_i,y_2) f(x_j,y_2) f_Y(y_1) dx_i dx_j dy_2 dy_1 \\ &= \frac{1}{n-1}\int \frac{1}{f_Y(y_1)} x_i x_j K_h^2(y_1-y_2) f(x_i,y_2) f(x_j,y_2) dx_i dx_j dy_2 dy_1 \\ &= \frac{1}{(n-1)\,h}\int \frac{1}{f_Y(y_1)} x_i x_j K^2(u) f(x_i,y_1+uh) f(x_j,y_1+uh) du dx_i dx_j dy_1. \end{aligned}$$

Remark that given a_m and b_m, $m = 1,2$ being real numbers such as $a_m < b_m$, the integrals containing the coordinate (x,y) will be evaluated in the cube $[a_1,b_1]\times[a_2,b_2]$.

Define the supremum norm of f as $\|f\|_\infty = \sup\{f(x,y) \in [a_1,b_1]\times[a_2,b_2]\}$. Therefore,

$$B_1 \leq \frac{\|f\|_\infty^2}{(n-1)\,h}\int \frac{1}{f_Y(y_1)} x_i x_j K^2(u) du dx_i dx_j dy_1,$$

which leads to

$$B_1 \leq \frac{C}{nh}.$$

Step 4: Show that

$$|B_2 - \sigma_{ij}| \leq Ch^{2\beta}.$$

The second term of (3.14) can be bounded as follows

$$B_2 = \left(\frac{n-2}{n-1}\right) \mathbb{E}\left(\frac{X_{i2}X_{j3}K_h(Y_1-Y_2)K_h(Y_1-Y_3)}{f_Y^2(Y_1)}\right)$$
$$\leq \int \left(\int x_i K_h(y_1-y) f(x,y) dx_i dy\right)$$
$$\left(\int x_j K_h(y_1-y) f(x,y) dx_j dy\right) \frac{1}{f_Y(y_1)} dy_1. \quad (3.15)$$

By Assumption 3.1 with $s = \lfloor \beta \rfloor$ and for $0 < \tau < 1$ we have,

$$\begin{aligned}
&\int K_h(y_1-y) f(x,y) dy - f(x,y) \\
&\quad = \int K(u) f(x, uh+y) du - f(x,y) \\
&\quad = \int K(u) \left\{ f(x,y) + uhf'(x,y) + \cdots + \frac{f^{(s)}(x, y+\tau uh)}{s!} (uh)^j \right\} du \\
&\quad = \frac{1}{s!} \int K(u) (uh)^s f^{(s)}(x, y+\tau uh) du.
\end{aligned}$$

Adding $K(u)(uh)^s f^{(s)}(x,y)$ to the last integral and the fact that $f(x,\cdot) \in \mathcal{H}(\beta, L)$ we can see that,

$$\begin{aligned}
&\int K_h(y_1-y) f(x,y) dy - f(x,y) \\
&\quad = \frac{1}{s!} \int K(u) (uh)^s \left\{ f^{(s)}(x, y+\tau uh) - f^{(s)}(x,y) \right\} du \\
&\quad \leq \frac{1}{s!} \int K(u) (uh)^s (\tau uh)^{\beta - s} du \\
&\quad \leq \left(\frac{1}{s!} \int |u^\beta K(u)| du \right) \tau h^\beta = C h^\beta.
\end{aligned}$$

Plugging this into (3.15) leads to,

$$\begin{aligned}
|B_2 - \sigma_{ij}| \leq &\left| \int \left\{ \left(\int x_i \left(Ch^\beta + f(x_i, y) \right) dx_i \right) \left(\int x_j \left(Ch^\beta + f(x_j, y) \right) dx_j \right) \right. \right. \\
&\left. \left. - \left(\int x_i f(x_i, y) dx_i \right) \left(\int x_j f(x_j, y) dx_j \right) \right\} \frac{1}{f_Y(y)} dy \right| \\
&\leq C^2 h^{2\beta} \left| \int \frac{x_i x_j}{f_Y(y)} \, dx_i dx_j dy \right|
\end{aligned}$$

$$
\begin{aligned}
&\quad + Ch^{\beta} \left| \left(\int x_i R_i(y) dx_i dy + \int x_j R_j(y) dx_j dy \right) \right| \\
&\leq C_1 h^{\beta} + C_2 h^{2\beta} \\
&\leq C_1 h^{\beta}.
\end{aligned}
$$

Step 5: Gathering the results from Steps 1 to 4, we get

$$
\text{Bias}(\hat{\sigma}_{ij,K}) \leq \left(C_1 h_1^{\beta} + \frac{C_2}{n_1 h_1} \right) (1 + C_3 (n_2^{c_1 - \beta c_2} + n_2^{1/2 - c_1 - c_2} \log n_2)).
$$

Since $h_1 \to 0$ and $n_1 h_1 \to \infty$ as $n_1 \to \infty$, we obtain

$$
\text{Bias}^2(\hat{\sigma}_{ij,K}) \leq C_1 h_1^{2\beta} + \frac{C_2}{n_1^2 h_1^2}.
$$

□

Proof of Lemma 3.2. The proof will be done in several steps. Define

$$
\begin{aligned}
R_{i,b}(Y) &= \frac{g_i(Y)}{f_Y(Y)}, \\
V_1(Y) &= \frac{g_i(Y) g_j(Y)}{f_{Y,b}^2(Y)} = R_{i,b} R_{j,b}, \\
V_2(Y) &= \frac{g_i(Y)}{f_{Y,b}^2(Y)} \left(\hat{g}_j(Y) - g_j(Y) \right) = \frac{R_{i,b}}{f_Y(Y)} \left(\hat{g}_j(Y) - g_j(Y) \right), \\
V_3(Y) &= \frac{g_j(Y)}{f_{Y,b}^2(Y)} \left(\hat{g}_i(Y) - g_i(Y) \right) = \frac{R_{j,b}}{f_Y(Y)} \left(\hat{g}_i(Y) - g_i(Y) \right), \\
V_4(Y) &= \frac{1}{f_{Y,b}^2(Y)} \left(\hat{g}_i(Y) - g_i(Y) \right) \left(\hat{g}_j(Y) - g_j(Y) \right), \\
J_n(Y) &= (V_1(Y) + V_2(Y) + V_3(Y) + V_4(Y)) \\
&\qquad \left(2 \frac{(f_Y(Y) - \hat{f}_{Y,b}(Y))}{\hat{f}_{Y,b}(Y)} + \frac{(f_Y(Y) - \hat{f}_{Y,b}(Y))^2}{\hat{f}_{Y,b}^2(Y)} \right).
\end{aligned}
$$

It is clear that $\hat{\sigma}_{ij,K} = n_1^{-1} \sum_{k=1}^{n_1} V_1(Y_k) + V_2(Y_k) + V_3(Y_k) + V_4(Y_k) + J_n(Y_k)$. If $C > 0$, then the variance $\text{Var}(\hat{\sigma}_{ij,K})$ is bounded by

$$
C \left\{ \text{Var} \left(\frac{1}{n_1} \sum_{k=1}^{n_1} V_1(Y_k) \right) + \text{Var} \left(\frac{1}{n_1} \sum_{k=1}^{n_1} V_2(Y_k) \right) + \text{Var} \left(\frac{1}{n_1} \sum_{k=1}^{n_1} V_3(Y_k) \right) \right.
$$

$$+ \operatorname{Var}\left(\frac{1}{n_1}\sum_{k=1}^{n_1} V_4(Y_k)\right) + \operatorname{Var}\left(\frac{1}{n_1}\sum_{k=1}^{n_1} J_n(Y_k)\right)\Bigg\}.$$

We are going to bound every term separately.

Step 1: Prove that

$$\operatorname{Var}\left(\frac{1}{n_1}\sum_{k=1}^{n_1} J_n(Y_k)\right) \leq C(n^{2c_1-4\beta c_2} + n^{2c_1+2c_2-1}\log n).$$

We bound first the term

$$J_{1n} = \frac{1}{n_1}\sum_{k=1}^{n_1} V_1(Y_k)\frac{(f_Y(Y_k) - \hat{f}_{Y,b}(Y_k))}{\hat{f}_{Y,b}(Y_k)}.$$

Using Cauchy-Schwartz's inequality, it is straightforward that

$$\begin{aligned}\operatorname{Var}(J_{1n}) &\leq \frac{1}{n_1^2}\mathbb{E}\left[\left(\sum_{k=1}^{n_1}\frac{g_i(Y_k)g_j(Y_k)}{f_{Y,b}^2}\left(\frac{f_Y(y) - \hat{f}_{Y,b}(y)}{\hat{f}_{Y,b}(y)}\right)\right)^2\right] \\ &\leq \frac{1}{n_1}\mathbb{E}\left[\sum_{k=1}^{n_1}\left(\frac{g_i(Y_k)g_j(Y_k)}{f_{Y,b}^2(Y_k)}\left(\frac{f_Y(Y_k) - \hat{f}_{Y,b}(Y_k)}{\hat{f}_{Y,b}(Y_k)}\right)\right)^2\right].\end{aligned}$$

By Lemma 3.3 and Remark 3.2 we have

$$\begin{aligned}\operatorname{Var}(J_{1n}) &\leq C\mathbb{E}\left[\frac{1}{n_1}\sum_{k=1}^{n_1}\frac{g_i^2(Y_k)g_j^2(Y_k)}{f_{Y,b}^4(Y_k)}\right] b^{-2}(h^{2\beta} + n^{-1/2}h^{-1}\log n)^2 \\ &\leq C(n^{2c_1-4\beta c_2} + n^{2c_1-1+2c_2}\log n),\end{aligned}$$

where the second inequality is due to the law of large numbers for

$$\frac{1}{n_1}\sum_{k=1}^{n_1} R_{i,b}(Y_k)R_{j,b}(Y_k).$$

To simplify the notation, we will write simply $n = n_1$ and $h = h_1$ through Steps 2 to 4. Moreover we will denote $Z_k = (\boldsymbol{X}_k, Y_k)$ $k = 1, \ldots, n$.

Step 2. Prove that

$$\mathrm{Var}(V_1) \leq \frac{C}{n}.$$

By independence of the Z_k's and given that g_j, g_l and $\hat{f}_Y$ are functions built with the second sample, it is clear that

$$\mathrm{Var}\left(\frac{1}{n}\sum_{k=1}^{n} V_1(Y_k)\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{k=1}^{n} R_{i,b}(Y_k)R_{j,b}(Y_k)\right)$$
$$= \frac{1}{n}\mathrm{Var}\left(R_{i,b}(Y)R_{j,b}(Y)\right) \leq \frac{C}{n}.$$

Step 3. Show that

$$\mathrm{Var}\left(V_2\right) + \mathrm{Var}\left(V_3\right) \leq \frac{C_1}{n} + \frac{C_2}{n^2 h}.$$

First we get a bound of $\mathrm{Var}(V_2)$. Note that,

$$\begin{aligned}
V_2 =& \frac{1}{n}\sum_{k=1}^{n} \frac{R_{i,b}(Y_k)}{f_Y(Y_k)}\left(\hat{g}_j(Y_k) - g_j(Y_k)\right) \\
=& \frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{\substack{l=1\\ l\neq k}}^{n} \frac{R_{i,b}(Y_k)}{f_Y(Y_k)}\left(X_{jl}K_h(Y_k - Y_l) - g_j(Y_k)\right) \\
=& \frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{\substack{l=1\\ l\neq k}}^{n} \frac{R_{i,b}(Y_k)}{f_Y(Y_k)} X_{jl}K_h(Y_k - Y_l) \\
& - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{l=1\\ l\neq k}}^{n} R_{i,b}(Y_k)R_{j,b}(Y_k) \\
=& V_{21} - V_{22}.
\end{aligned}$$

Notice that

$$\mathrm{Var}(V_{22}) = \frac{1}{n}\mathrm{Var}\left(R_{i,b}(Y)R_{j,b}(Y)\right) = \frac{C}{n}.$$

The term V_{21} is indeed a one sample U-statistic of order two. Hence, if we define $Z_{li} = (X_{li}, Y_k)$ and rewrite the expression, we get

$$V_{21} = \frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{\substack{l=1\\ l\neq\ k}}^{n} \frac{R_{i,b}(Y_k)}{f_Y(Y_k)} X_{jl}K_h(Y_k - Y_l)$$

$$= \frac{1}{2}\binom{n}{2}^{-1} \sum_{(k,l)\in C_2^n} h_i(Z_{ik}, Z_{il}),$$

where $C_2^n = \{(k,l); 1 \leq k < l \leq n\}$ and

$$h_i(Z_{jk}, Z_{jl}) = \frac{R_{i,b}(Y_k)}{f_Y(Y_k)} X_{jk} K_h(Y_k - Y_l).$$

Employing the symmetric version of h_i

$$\widetilde{h}_i(Z_{jk}, Z_{jl}) = \frac{1}{2}\left(\frac{R_{i,b}(Y_k)}{f_Y(Y_k)} X_{jl} + \frac{R_{i,b}(Y_l)}{f_Y(Y_l)} X_{jk}\right) K_h(Y_k - Y_l),$$

it is possible (see Kowalski and Tu (2007) or Van der Vaart (2000)) to decompose $\mathrm{Var}(V_{21})$ as

$$\mathrm{Var}(V_{21}) = \binom{n}{2}^{-1} \left\{ \binom{2}{1}\binom{n-2}{1} \mathrm{Var}(\mathbb{E}(\widetilde{h}_i(Z_{j1}, Z_{j2})|Z_{j1})) + \binom{2}{2}\binom{n-2}{0} \mathrm{Var}(\widetilde{h}_i(Z_{j1}, Z_{j2})) \right\}.$$

Since $f_y(y) \leq f_Y(y)$ we have,

$$\begin{aligned}
&\mathbb{E}(\widetilde{h}_i(Z_{j1}, Z_{j2})|Z_{j1}) \\
&\quad = \frac{1}{2}\int K_h(Y_1 - y)\left(\frac{x_j R_{i,b}(Y_1)}{f_Y(Y_1)} + \frac{X_{j1} R_i(y)}{f_Y(y)}\right) f(x_j, y) dx_j dy \\
&\quad = \frac{R_{i,b}(Y_1)}{2 f_Y(Y_1)} \int K_h(Y_1 - y) R_j(y) f_Y(y) dy \\
&\qquad + \frac{1}{2} X_{j1} \int K_h(Y_1 - y) \frac{R_{i,b} f_Y(y)}{f_Y(y)} dy \\
&\quad \leq \frac{1}{2} R_i(Y_1) R_j(Y_1) + \frac{1}{2} X_{j1} R_i(Y_1) \\
&\qquad + \frac{R_i(Y_1)}{2 f_Y(Y_1)} \int K_h(Y_1 - y)\{R_j(y) f_Y(y) - R_j(Y_1) f_Y(Y_1)\} dy \\
&\qquad + \frac{1}{2} X_{j1} \int K_h(Y_1 - y)\left\{R_i(y) - R_i(Y_1)\right\} dy \\
&\quad \leq \frac{1}{2} R_i(Y_1) R_j(Y_1) + \frac{1}{2} X_{j1} R_i(Y_1) + J_1(Z_{j1}) + J_2(Z_{j1}).
\end{aligned}$$

Using Assumption 3.1 and applying the same arguments as in the proof of Lemma 3.4, we can conclude that

$$\begin{aligned}\mathrm{Var}(J_1(Z_{j1})) &\leq \mathbb{E}[(J_1(Z_{j1}))^2] \\ &\leq Ch^{2\beta}\int\left(\frac{R_i(y)}{f_Y(y)}\right)^2\left(\int|u^s K(u)du|\right)^2 f_Y(y)dy \\ &\leq Ch^{2\beta}\mathbb{E}[R_i^2(Y_1)] \leq Ch^{2\beta}.\end{aligned}$$

Moreover, as $f \in \mathcal{H}(\beta, L)$ and $0 < \eta < f_Y(y)$, we have

$$\begin{aligned}\mathrm{Var}(J_2(Z_{j1})) &\leq \mathbb{E}[(J_2(Z_{j1}))^2] \\ &\leq \frac{1}{4}\mathbb{E}\left[X_{j1}^2\left(\int K(u)\left(R_i(Y_1+uh) - R_i(Y_1)du\right)du\right)^2\right] \\ &\leq Ch^{2\beta}.\end{aligned}$$

Therefore,

$$\mathrm{Var}(\mathbb{E}(\widetilde{h}_i(Z_{j1}, Z_{j2})|Z_{j1})) \leq \mathrm{Var}\left(\frac{1}{2}R^2(Y_1) + \frac{1}{2}X_{l1}R(Y_1)\right) + C_1h^{2\beta}.$$

By similar calculations we bound,

$$\mathrm{Var}(\widetilde{h}_i(Z_{j1}, Z_{j2})) \leq \frac{C_2}{h}.$$

Using the same procedure we can bound $\mathrm{Var}(V_3)$. We conclude that,

$$\begin{aligned}\mathrm{Var}(V_2) + \mathrm{Var}(V_3) &\leq \frac{2}{n(n-1)}\left\{(n-2)\left(C_1 + C_2h^{2\beta}+\right) + \frac{C_4}{h}\right\} \\ &\leq C_1h^{2\beta} + \frac{C_2}{n^2h}.\end{aligned}$$

Step 4. Show that

$$\mathrm{Var}(V_4) \leq C(h^{4\beta} + n^{-2}h^{-4}\log^4 n).$$

Using Lemma 3.4 we obtain

$$\mathrm{Var}\,(V_4) \leq \mathbb{E}[V_4^2]$$

$$
\begin{aligned}
&= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{k=1}^{n}\frac{\left(\hat{g}_i(Y_k)-g_i(Y_k)\right)\left(\hat{g}_j(Y_k)-g_j(Y_k)\right)}{f_Y^2(Y_k)}\right)^2\right]\\
&\leq \frac{Cn^2}{n^2}(h^\beta + n^{-1/2}h^{-1}\log n)^4\\
&\leq C(h^\beta + n^{-1/2}h^{-1}\log n)^4\\
&\leq C(h^{4\beta} + n^{-2}h^{-4}\log^4 n).
\end{aligned}
$$

Final bound Gathering all previous results, we have

$$
\mathrm{Var}\left(\hat{\sigma}_{ij,K}\right) \leq C_1 h^{2\beta} + \frac{C_2 \log^4 n}{n^2 h^4}.
$$

□

Chapter 4

Nonparametric estimator of Sobol indices

4.1 Introduction

In the areas of chemistry, biology, psychology or finance, the process of implementing some policy or taking some decision is often supported by complex models. The researcher has to process, analyze and interpret those systems with numerous variables, high interactions and complexity. Two examples of complex models are the following:

- One bank could buy some financial contract (e.g., European option) indexed to some rate of interest. This rate is unknown and depends on complex factors: market dynamics, inflation or political policies (see Hull (2011)). If the bank wants to sell the option, they face the risk associated with the interest rate movements. The goal, in general, is to avoid the risk exposure by holding this product and make the most profit with it.

- The Michaelis-Menten kinetics (Cornish-Bowden (2013)) models a saturable enzymatic degradation of a substance. The model is highly complex because it needs several experiments to estimate their parameters. In general, the scientists search to approximate stably the time of the reaction equilibrium. This process models the oxidation of glucose, which produces water, carbon dioxide and energy for instance.

To handle those complexities, it is necessary identify the most important variables in the model. In the first example, the risk manager has to take a decision based in some financial model. Given the inherent complexity, he needs to select the most relevant features on the model of rates. With the new set of *relevant* variables, he will gain insight on the model and it is possible take a better decision. Also, as the Michaelis-Menten kinetics, the model has to run stably to find the correct equilibrium. In other words, if the model suffer of small alterations in the input, those should not produce large variations on the output. In any case, the analyst has to validate, check and correct the model if it is necessary.

We assume a set of inputs variables $\boldsymbol{X} = (X_1, \ldots, X_p) \in \mathbb{R}^p$ producing an output $Y \in \mathbb{R}$ related by the model

$$Y = m(X_1, \ldots, X_p). \tag{4.1}$$

The function m is, generally, an unknown and complex function. However, a computer code can gauge it in some cases (e.g., Oakley and O'Hagan (2004)). When the running time for the evaluation of such complex function are important, one could replace the original model by a *meta-model* (see Box and Draper (1987)).

The literature present techniques to rank the influences of the inputs $(X_1, \ldots, X_p)$ in the model (4.1). Some of them are, for example, the screening method (Cullen and Frey (1999) Campolongo et al. (2011)), the automatic differentiation (Rall (1980), Carmichael et al. (1997)), the regression analysis (Draper and Smith (1981), Devore and Peck (1996)) or the response surface method (Myers et al. (2009), Goos (2002)).

Alternatively, Sobol' (1993), inspired by a ANOVA (or Hoeffding) decomposition, split down the variance of the model in partial variances generated by the conditionals expectations of Y giving each input X_i for $i = 1, \ldots, p$. These partial variances represent the uncertainty created by each input or its interactions. Dividing each partial variance by the model total variance, we obtain a normalized index of importance. Specifically, we call the first-order Sobol indices to the quantities,

$$S_i = \frac{\mathrm{Var}(\mathbb{E}[Y|X_i])}{\mathrm{Var}(Y)} \quad \text{for} \quad i = 1, \ldots, p.$$

Notice that $\mathbb{E}[Y|X_i]$ is the best approximation of Y given the information of X_i. Thus, if the variance of $\mathbb{E}[Y|X_i]$ is large, it means a large influence of X_i into

Y. These indices are a widely used in theoretical and applied techniques and identify the most relevant and sensible inputs on the model. We can construct indices that measure the interactions between variables or the total effect of certain input in the whole model. We refer the reader to Saltelli et al. (2000) for the exact computation of higher-order Sobol indices.

The principle endeavor with the Sobol indices relays in its computation. Some methods propose the use of multiple samples (of the order of hundreds or thousands) for the evaluation of the model outputs. From applied problems in engineering, biology, oceanography and others; the scientists have developed Monte-Carlo or quasi Monte-Carlo methods. For instance, the Fourier amplitude sensitivity test (FAST) or the Sobol pick-freeze (SPF). Cukier et al. (1973, 1978) created the FAST method which transforms the partial variances in Fourier expansions. This method allows the aggregated and simple estimation of Sobol indices in an escalated way. The SPF scheme regresses the model output against a pick-frozen replication. The principle is to create a replication holding the interest variable (frozen variable) and resampling the other variables (picked variables). We refer to the reader to Sobol' (1993, 2001) and Janon et al. (2013). Other methods include to Ishigami and Homma (1990) which improved the classic Monte-Carlo procedure by resampling the inputs and reducing the whole process to only one Monte-Carlo loop. Moreover, Saltelli (2002) proposed an algorithm to estimate higher-order indices with the minimum computation effort.

The Monte-Carlo methods suffer of the high-computational stress in its implementation. For example, the FAST method requires estimate a set of suitable transformation functions and integer angular frequencies for each variable. The SPF scheme creates a new copy of the variable in each iteration. For complex and high-dimensional models, those techniques could be expensive in computational time.

The aim of this chapter propose an alternative way to compute the Sobol indices. In particular, we will take the ideas of Zhu and Fang (1996) and we shall apply a nonparametric Nadaraya-Watson to estimate the value S_i for $i = 1, \ldots, p$. With this estimator, we avoid the stochastic techniques and we use the structure of the data to fit the nonparametric model. We consider only the indices with simple interaction between one variable with respect the output. We leave out the sensitivity indices with interactions for a further study. Furthermore, we will show that if the joint distribution of (X_i, Y) is twice differentiable, the nonparametric estimator of S_i, has a parametric rate of convergence. Otherwise,

we will get a nonparametric rate of convergence depending on the regularity of the density.

In Section 4.2 we will propose the nonparametric estimator for the first-order Sobol indices. All the hypotheses and assumptions are gathered in Section 4.3. We display the main result of this chapter in Section 4.4. We illustrate our methodology with some numerical examples in Section 4.5. Finally, in Section 4.6, we discuss the results and expose some conclusions.

4.2 Methodology

In our context we suppose that $\boldsymbol{X}_k^\top = (X_{1k}, \ldots, X_{pk})$ and Y_k, $k = 1, \ldots, n$ are independent and identically distributed observations from the random vector $\boldsymbol{X}^\top = (X_1, \ldots, X_p)$ and Y respectively. We denote by $f(x_i, y)$ the joint density of the couple (X_i, Y). Let $f_i(x_i) = \int_{\mathbb{R}^p} f(x_i, y)dy$ be the marginal density function with respect to X_i for $i = 1, \ldots, p$.

Recall the definition of Sobol indices presented in the introduction,

$$S_i = \frac{\mathrm{Var}(\mathbb{E}[Y|X_i])}{\mathrm{Var}(Y)} = \frac{\mathbb{E}[\mathbb{E}[Y|X_i]]^2 - \mathbb{E}[Y]^2}{\mathrm{Var}(Y)} \quad \text{for} \quad i = 1, \ldots, p. \tag{4.2}$$

We have expanded the variance of the numerator to simplify the presentation. Notice that we can estimate the terms $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$ in equation (4.2) by their empirical counterparts

$$\overline{Y} = \frac{1}{n}\sum_{k=1}^{n} Y_k \tag{4.3}$$

and

$$\hat{\sigma}_Y = \frac{1}{n}\sum_{k=1}^{n} (Y_k - \overline{Y})^2 \tag{4.4}$$

respectively.

Conversely, the estimation of the term $\mathbb{E}[\mathbb{E}[Y|X_i]^2]$ requires more effort. For any $i = 1, \ldots, p$ we introduce the following notation,

$$V_i = \mathbb{E}[\mathbb{E}[Y|X_i]^2] = \int \left(\frac{\int y\, f(x_i, y)\, dy}{f_i(x_i)} \right)^2 f_i(x_i)\, dx_i$$

$$= \int \left(\frac{g_i(x_i)}{f_i(x_i)} \right)^2 f_i(x_i)\, dx_i,$$

where

$$g_i(x_i) = \int y\, f(x_i, y)\, dy.$$

We will use a slightly modified version of the nonparametric estimator developed in Loubes et al. (2013). This paper estimates the conditional expectation covariance for a reduction dimension problem. Taking the same methodology applied to our case, define $n_1 < n$ and $n_2 = n - n_2$. For the sake of simplicity we use $n_1 = n_2 = n/2$. Let h_1 and h_2 be two bandwidths depending on n_1 and n_2 respectively.

We will estimate the functions $g_i(x)$ and $f_i(x)$ respectively by their nonparametric estimators,

$$\hat{g}_i(x) = \frac{1}{(n_1 - 1)h_1} \sum_{\substack{l=1 \\ l \neq k}}^{n_1} Y_l K\left(\frac{x - X_{il}}{h_1} \right), \tag{4.5}$$

$$\hat{f}_i(x) = \frac{1}{n_2 h_2} \sum_{l=1}^{n_2} K\left(\frac{x - X_{il}}{h_2} \right). \tag{4.6}$$

Also, to avoid small values in the V's estimation due to small values in the denominator, let $b > 0$ a sequence of values satisfying

$$\hat{f}_{i,b}(x) = \max\{\hat{f}_i(x), b\}.$$

The nonparametric estimator for V_i is,

$$\widehat{V}_i = \frac{1}{n_1} \sum_{k=1}^{n_1} \left(\frac{\hat{g}_i(X_{ik})}{\hat{f}_{i,b}(X_{ik})} \right)^2. \tag{4.7}$$

Therefore, gathering the estimators (4.3) and (4.7), we define the nonparametric estimator for S_i as

$$\widehat{S}_i = \frac{\widehat{V}_i - \overline{Y}^2}{\hat{\sigma}_Y}. \tag{4.8}$$

The estimator (4.8) provides a direct way to estimate the first-order Sobol index S_i. We want to control the square risk over some regular class of functions

for the joint density $f(x_i, y)$. Therefore, we can choose a bandwidth h_1 according to the regularity to get a convergent estimator. Our objective is to find sufficient conditions to have a parametric rate of convergence for $\widehat{S}_i$.

4.3 Hypothesis and Assumptions

We denote C_1, C_2 and so on, constants (independent of n) that may take different values throughout all the chapter.

Let β be a positive real value and define $\lfloor \beta \rfloor$ as the largest integer such that $\lfloor \beta \rfloor \leq \beta$. We define the continuous kernel function $K(\cdot)$ of order $\lfloor \beta \rfloor$ to every function satisfying the following three conditions:

(a) the support of $K(\cdot)$ is the interval $[-1, 1]$;

(b) $K(\cdot)$ is symmetric around 0;

(c) $\int_{-1}^{1} K(u)du = 1$ and $\int_{-1}^{1} u^k K(u)du = 0$ for $k = 1, \ldots, \lfloor \beta \rfloor$.

To guarantee a parametric consistency in our model, we need to impose some regularity conditions. In our case, define the Hölder class of smooth functions as follows.

Definition 4.1. Denote as $\mathcal{H}(\beta, L)$ the *Hölder class* of density functions with smoothness $\beta > 0$ and radius $L > 0$, defined as the set of $\lfloor \beta \rfloor$ times differentiable functions $\phi : T \to \mathbb{R}$ where T is an interval in $\mathbb{R}$, whose derivative $\phi^{(\lfloor \beta \rfloor)}$ satisfies

$$\left| \phi^{(\lfloor \beta \rfloor)}(x) - \phi^{(\lfloor \beta \rfloor)}(x') \right| \leq L \left| x - x' \right|^{\beta - \lfloor \beta \rfloor}, \ \forall x, x' \in T.$$

The following technical assumption is important to establish the class of function where we will find the upper bound of the risk.

Assumption 4.1. *For y fixed, the function $f(x, y)$ belongs to a Hölder class of regularity β and constant L, i.e. $f(\cdot, y) \in \mathcal{H}(\beta, L)$.*

Moreover, $f_i(x)$ for $i = 1, \ldots, p$ belongs to a Hölder class of smoothness $\beta' > \beta$ and radius L', i.e. $f_i \in \mathcal{H}(\beta', L')$.

Remark 4.1. Notice that function g_i defined in (4.5) also belongs to $\mathcal{H}(\beta, L)$ for $i = 1, \ldots, p$. Recall that

$$g_i(x_i) = \int x_i f(x_i, y)dy.$$

Let $y \in \mathbb{R}$ fixed. If $f(\cdot, y) \in \mathcal{H}(\beta, L)$ then by a direct calculation we can prove our assertion.

Denote as $\mathcal{F} = \mathcal{F}_\beta(L)$ the class of functions that fulfill Assumption 4.1. In the next section, we shall found the rates of convergence for $\hat{\sigma}_{ij}$ depending on the parameter β.

Remark 4.2. To control the term $\hat{f}_{X,b}$ in the denominator on equation (4.7), we need to fix h_2 and b—both sequences converging to zero—to ensure the convergence of $\widehat{V}_i$. As $n \to \infty$, set $h_2 \sim n^{-c_1}$ and $b \sim n^{-c_2}$ with the positives number c_1 and c_2 satisfying that $c_1/\beta < c_2 < 1/2 - c_1$, and the notation "$\sim$" means that two quantities have the same convergence order.

4.4 Main result

The following theorem emphasizes the performance of our estimator.

Theorem 4.1. *Assume that* $\mathbb{E}\,|X_i|^4 < \infty$, $i = 1, \ldots, p$. *The upper bound risk of the estimator* $\widehat{S}_i$ *defined in* (4.8) *over the functional class* $\mathcal{F}$ *satisfies:*

- *if* $\beta \geq 2$ *and choosing* $h_1 \approx n_1^{-1/4}$ *then*

$$\sup_{\mathcal{F}} \mathbb{E}[(\widehat{S}_i - S_i)^2] \leq \frac{C}{n_1}. \tag{4.9}$$

- *if* $\beta < 2$ *and choosing* $h_1 \approx n_1^{-1/(\beta+2)}$ *then,*

$$\sup_{\mathcal{F}} \mathbb{E}[(\widehat{S}_i - S_i)^2] \leq C \left(\frac{\log^2(n_1)}{n_1} \right)^{2\beta/(\beta+2)}. \tag{4.10}$$

The proof of the Theorem 4.1 will be postponed to the Appendix 4.7.

Theorem 4.1 presents an *elbow effect* on the rates of convergences. This is a typical behavior in linked to studies on functional estimation, for instance Baraud et al. (2003) or Laurent (2005). The regularity of the joint density function f defines the rate of convergence for the mean squared risk of $\widehat{S}_i$. It means, we can get a parametric rate n_1^{-1} paying a price on the regularity of $\beta \geq 2$. When $\beta < 2$, the rate turn into a nonparametric one which depends on the parameter β.

In the regular case, when $\beta \geq 2$, we avoid any adaptability problem with respect to the choose of the bandwidth h_1. We can ensure a parametric rate of convergence for $\widehat{S}_i$ taking a bandwidth of $h_i = n^{-1/4}$. In other case, it is necessary to be aware of the regularity of our model to choose the bandwidth.

4.5 Numerical illustrations

In this section, we illustrate the asymptotic results with some tests function for sensitivity analysis. The reader can find an abundant list of well-known test functions in the Section 2.9 of Saltelli et al. (2000).

In all the simulations we will took n equal to 1000, 2000, 5000 and 10000 for each case. We repeated the experiment 100 times selecting a different sample in each iteration. The inputs are gaussian random variables with mean 0 and variance specified in each configuration. The horizontal full lines in the graphics represent the theoretical Sobol index for each variable. We used the package `moments` (Komsta and Novomestky (2011)) using a sample of 10^6 observations to estimate those theoretical values. Also, we used the package `np` (Hayfield and Racine (2008)) for the kernel estimators.

4.5.1 Ishigami model

The Ishigami function (cf. Ishigami and Homma (1990)) will generate our first simulation study. We use the following configuration,

$$Y = \sin X_1 + 7\ \sin^2 X_2 + 0.1\ X_3^4 \sin X_1 \tag{4.11}$$

where $X_i \sim \mathcal{N}(0, 1.25^2)$ for $i = 1, 2, 3$.

The Ishigami function is a popular model in sensitivity analysis because presents a strong nonlinearity and nonmonotonicity with interactions in X_3. For a further explanation of this function, we refer the reader to Sobol' and Levitan (1999). We will present only the first-order Sobol indices of the Ishigami function in this work. We choose a bandwidth of $h = n^{-1/4}$ for each sample.

Figure 4.1 summarizes the simulation results. Our method has estimated correctly the Sobol indices for each input. The bias is inherent in the model due to the choose of our bandwidth. We appreciate strongly this effect for the inputs X_2 and X_3. The bandwidth $h = n^{-1/4}$ is sub-optimal. We expected it due to we have established only an upper bound for the kernel estimator. Even so, the variance is well controlled and decrease as we raise the number

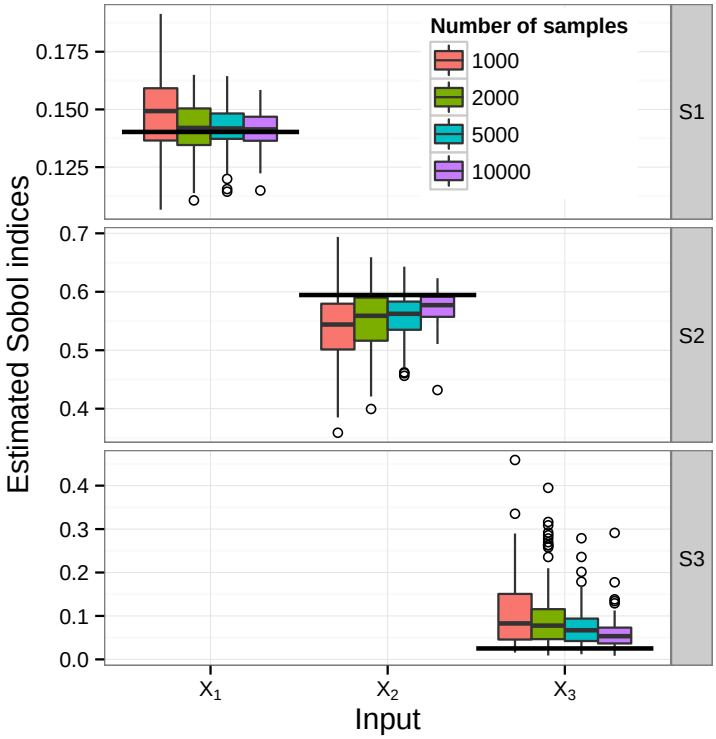


Figure 4.1: Box plot of the Sobol indices for the Ishigami model over 100 replications. Each color represent the size of the sample (1000, 2000, 5000 and 10000). The bandwidth h used in the simulation was $n^{-1/4}$. The horizontal full lines represent the theoretical Sobol indices $S_1 = 0.14024$, $S_2 = 0.59448$ and $S_3 = 0.02515$.

of samples. In general, our method finds the most influential factor X_2 and discards the less influential ones X_1 and X_3 which their theoretical Sobol indices are $S_1 = 0.14024037$, $S_2 = 0.59448009$ and $S_3 = 0.02514975$.

4.5.2 Quartic model

The next model that we investigate is

$$Y = X_1 + X_2^4 \tag{4.12}$$

where we consider three configurations for X_1 and X_2:

(Q1) $X_i \sim \mathcal{N}(0, 0.25^2)$, $i = 1, 2$.
The most influent variable is X_1 with $S_1 = 0.97750$ while X_2 has $S_2 =$

0.022510.

(Q2) $X_i \sim \mathcal{N}(0, 0.5^2)$, $i = 1, 2$.
Both variables have comparable influence with $S_1 = 0.39710$ and $S_2 = 0.60150$.

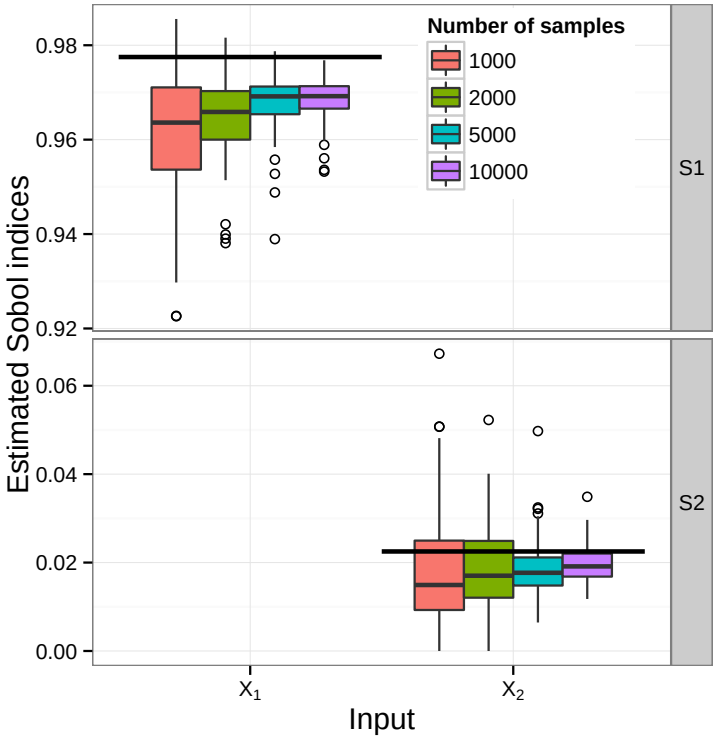
(Q3) $X_i \sim \mathcal{N}(0, 1)$, $i = 1, 2$.
The variable X_2 has the greater Sobol index $S_2 = 0.98977$ against $S_1 = 0.01031$ for X_1.

We have simulated the quartic model with two settings for the bandwidth values: (a) selected by the cross-validation method implemented in the package `np`, and (b) fixed to $h = 1/16\, n^{-1/4}$ according to Theorem 4.1. For the theoretical bandwidths we used the values 0.01111, 0.00935, 0.00743 and 0.00625 for the samples 1000, 2000, 5000 and 10000 respectively. The Table 4.1 presents the average bandwidths estimated by cross-validation for each Sobol index and each configuration.

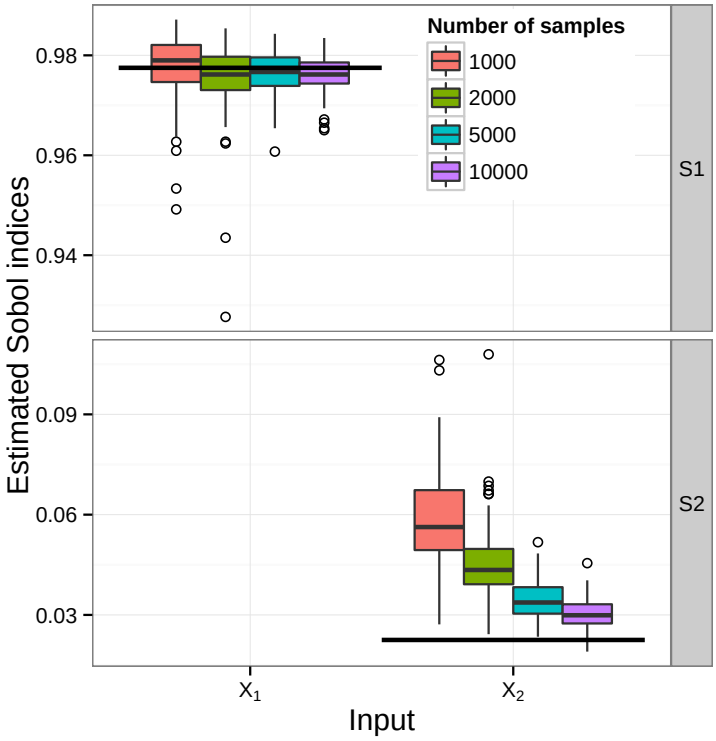
	Q1		Q2		Q3	
	S_1	S_2	S_1	S_2	S_1	S_2
1000	0.02260	0.09704	0.11040	0.06954	0.63539	0.06877
2000	0.02074	0.07039	0.09777	0.05354	0.47266	0.13996
5000	0.01741	0.05959	0.08991	0.03869	0.41863	0.03519
10000	0.01524	0.05323	0.08064	0.04176	0.38029	0.04588

Table 4.1: Average bandwidths estimated by the built-in cross-validation method of the package `np` for the Sobol indices S_1 and S_2 with the configurations Q1, Q2 and Q3.

Figures 4.2, 4.3 and 4.4 represent the box plots of the Sobol indices for the configurations Q1, Q2 and Q3 respectively. First, notice that the simulations taking the bandwidth chosen by the `np` cross-validation estimator performed poorly. In particular, they present a bias consistently in all the configurations. This is due that the package `np` optimizes its procedure for nonparametric functional or regression estimation. It ignores any particular structure of the conditional variance $\mathrm{Var}(\mathbb{E}[Y|X_i])$. Still, using these values the method can approximate the theoretical values and control the variance as the sample increases.

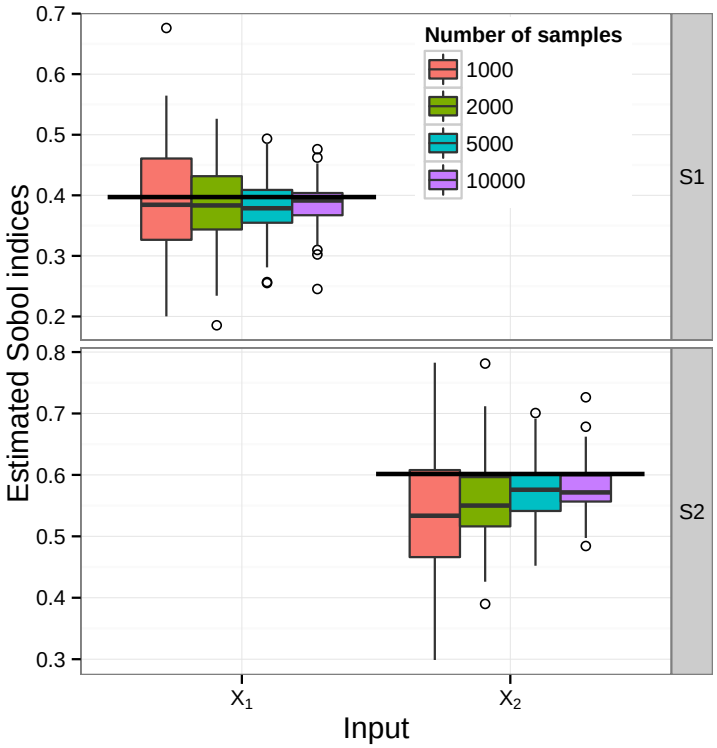


(a) Bandwidth chosen by cross-validation using the package `np`.

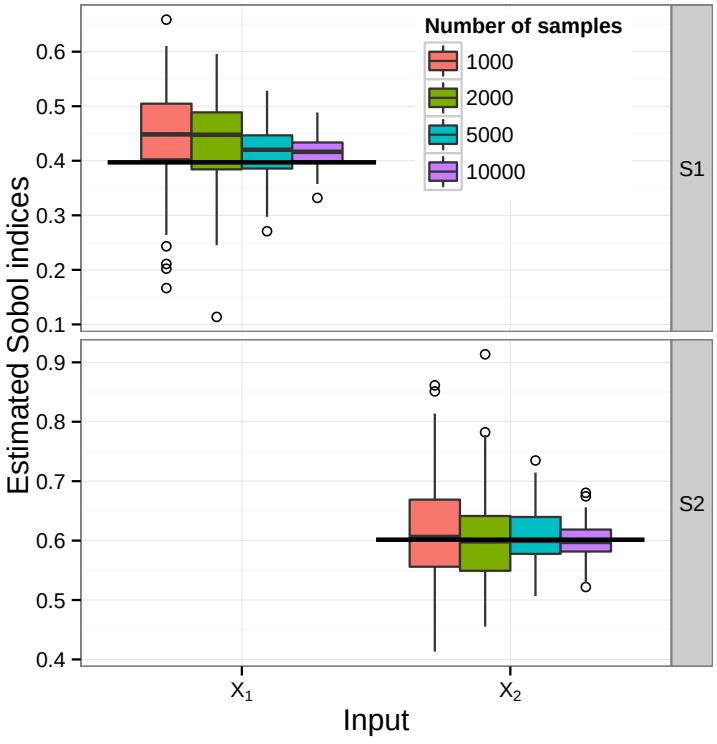


(b) Bandwidth set to $1/16\ n^{-1/4}$.

Figure 4.2: Box plot of the Sobol indices for the analytical model $Y = X_1 + X_2^4$ over 100 replications for the configuration (Q1). Each color represent the size of the sample (1000, 2000, 5000 and 10000). The horizontal full lines represent the theoretical Sobol indices $S_1 = 0.97750$ and $S_2 = 0.02251$.

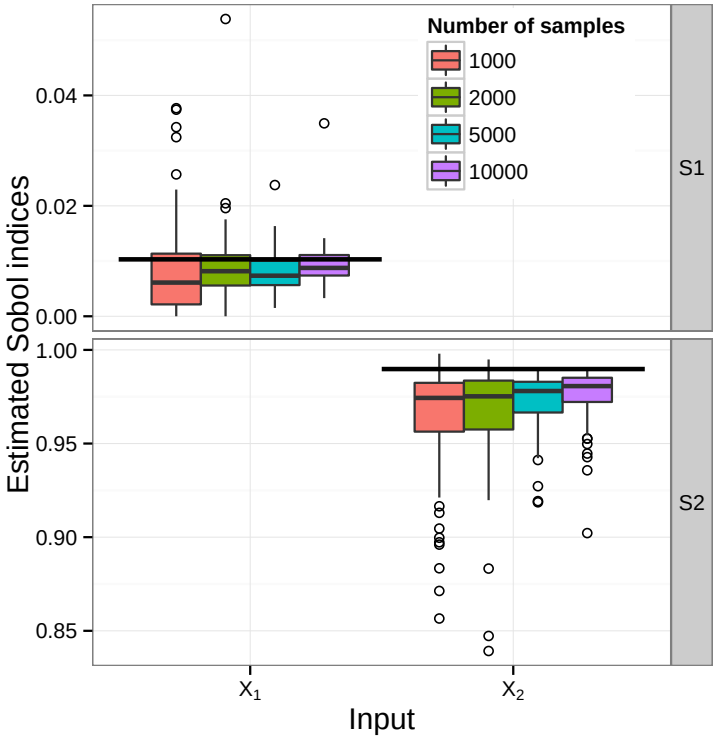


(a) Bandwidth chosen by cross-validation using the package `np`.

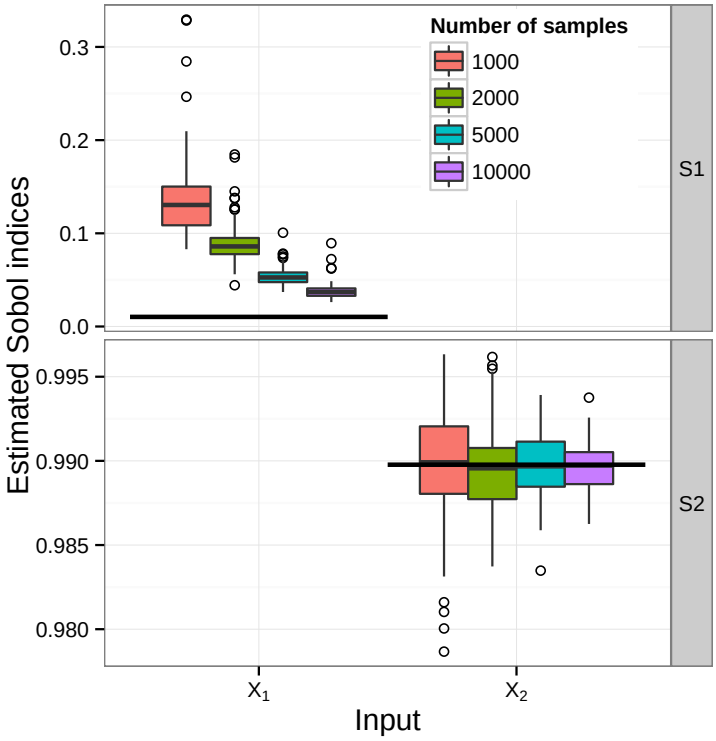


(b) Bandwidth set to $1/16\ n^{-1/4}$.

Figure 4.3: Box plot of the Sobol indices for the analytical model $Y = X_1 + X_2^4$ over 100 replications for the configuration (Q2). Each color represent the size of the sample (1000, 2000, 5000 and 10000). The horizontal full lines represent the theoretical Sobol indices $S_1 = 0.39710$ and $S_2 = 0.60150$.



(a) Bandwidth chosen by cross-validation using the package `np`.



(b) Bandwidth set to $1/16\ n^{-1/4}$.

Figure 4.4: Box plot of the Sobol indices for the analytical model $Y = X_1 + X_2^4$ over 100 replications for the configuration (Q3). Each color represent the size of the sample (1000, 2000, 5000 and 10000). The horizontal full lines represent the theoretical Sobol indices $S_1 = 0.01031$ and $S_2 = 0.98977$.

Using the theoretical bandwidth $1/16\ n^{-1/4}$, we improve our results. The method reduces both the bias and the variance in every case as we raise the number of observations. Compared with the cross-validation choice, we see how the averages of our method stay close of the theoretical Sobol indices. Figures 4.2b and 4.4b present strong biases for the sample of 1000, but they shrink as we increment to 10000 observations. Independently of the choice of the bandwidth, our methodology has identified the most relevant input for each case.

4.6 Conclusion

In this work, we focused our attention in the estimation of first-order Sobol indices. These indices measure the impact of the inputs to the output in general models. We presented a nonparametric estimator based in the work proposed previously by Loubes et al. (2013).

This procedure provides a direct way to estimate the Sobol indices without run any expensive Monte-Carlo simulation. In real sensitivity analysis applications, this advantage could give fast answers and create better models.

We have shown that the mean square risk of our estimator attains a parametric rate of convergence, if the regularity of the model is smooth enough. In other case, we can only ensure a slower rate depending on the regularity of the model. This is interesting from the theoretical point of view because sets parametric properties to a nonparametric object. Moreover, it is possible to extend this technique to higher-order indices. The idea is use multivariate kernels in equations (4.5) and (4.6). We will explore higher-order indices in a further work. Moreover, we have to improve the rates of convergence finding a lower bound to proof the optimality.

The numerical simulations shown that our model identify correctly the Sobol indices in each case presented. The configurations used were set with gaussian inputs. We compared the bandwidth proposed in Theorem 4.1 against the estimated by the `np` package. The results for the choice of our bandwidth performed better than the cross-validation one. This is because the `np` is not optimized for conditional variance regression. In general, it is hard understand the regularity of the joint densities. However, in a future work, we could develop a cross-validation scheme suited for this framework.

We have to test further our estimator, but the first results are promising. Also, it would be interesting apply this algorithm to real data and test its capabilities.

4.7 Appendix

4.7.1 Proof of Theorem 4.1

Proof. Notice that

$$\begin{aligned}\mathbb{E}[(\widehat{S}_i - S_i)^2] &= \mathbb{E}\left[\left(\frac{\widehat{V}_i - \overline{Y}^2}{\hat{\sigma}_Y} - \frac{V_i - \mathbb{E}[Y]}{\mathrm{Var}(Y)}\right)^2\right] \\ &\leq \mathbb{E}\left[\left(\frac{\overline{Y}^2}{\hat{\sigma}_Y} - \frac{\mathbb{E}[Y]}{\mathrm{Var}(Y)}\right)^2\right] + \mathbb{E}\left[\left(\frac{\widehat{V}_i}{\hat{\sigma}_Y} - \frac{V_i}{\mathrm{Var}(Y)}\right)^2\right]. \qquad (4.13)\end{aligned}$$

The first term in equation (4.13) the it is easy to bound by n^{-1} applying the delta-method. Now adding and subtracting $\widehat{V}_i/\sigma_Y$ in the right term, we can decompose it in the following way,

$$\mathbb{E}\left[\left(\frac{\widehat{V}_i}{\hat{\sigma}_Y} - \frac{V_i}{\mathrm{Var}(Y)}\right)^2\right] = \mathbb{E}\left[\left(\frac{\widehat{V}_i}{\sigma_Y}\frac{(\hat{\sigma}_Y - \sigma_Y)}{\hat{\sigma}_Y}\right)^2\right] + \mathbb{E}\left[\left(\frac{\widehat{V}_i - V_i}{\sigma_Y}\right)^2\right] \qquad (4.14)$$

Given that the kernel function K and the value $\mathbb{E}[X_i^4]$ are bounded, we can apply the Cauchy-Schwarz inequality to left term in equation (4.14). We can establish that there exist a real number $M > 0$ such as $\mathbb{E}[(\widehat{V}_i/\sigma_Y)^4] \leq M$. Moreover, applying once again the delta-method, we can control the value

$$\mathbb{E}\left[\left(\frac{\hat{\sigma}_Y - \sigma_Y}{\hat{\sigma}_Y}\right)^2\right] \leq \frac{1}{n}.$$

Finally, for the term

$$\mathbb{E}[(\widehat{V}_i - V_i)^2]$$

in equation (4.14) we will apply directly the Theorem 1 of Loubes et al. (2013). We use the same reasoning in that proof: If $\beta \geq 2$ and choosing $h_1 = n_1^{-1/4}$ we obtain

$$\sup_{\mathcal{F}} \mathbb{E}[(\widehat{S}_i - S_i)^2] \leq \frac{C}{n_1}.$$

Otherwise, taking $h_1 = n_1^{-1/(\beta+2)}$ the upper bound turns into

$$\sup_{\mathcal{F}} \mathbb{E}[(\widehat{S}_i - S_i)^2] \leq C \left(\frac{\log^2(n_1)}{n_1} \right)^{2\beta/(\beta+2)}.$$

This proves the theorem.

□

Conclusions and perspectives

In this thesis we studied the estimation of the conditional covariance matrix for two main problems: reduction dimension and sensibility analysis. We assumed, generally, a nonlinear regression model with $\boldsymbol{X} \in \mathbb{R}^p$ the independent variable and $Y \in \mathbb{R}$ the dependent variable. Also we assume that p is bigger than the number of observations available in the experiment.

In the first part we concentrate our efforts finding an estimator for the conditional covariance matrix $\Sigma = \mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$. We introduced this matrix as a collateral result for the sliced inverse regression method. In short, once we have the spectral space of Σ, we can identify a reduced and more informative set of coordinates. This new coordinates will serve us to explore further the data, applying other technique like classification, regression, etc.

Then we moved our interest to study the value $\mathrm{Var}(E[Y|X])$ for $X \in \mathbb{R}$ and $Y \in \mathbb{R}$. This quantity is directly linked with the estimation of Sobol indices for sensitivity analysis. The Sobol indices measure the impact of the inputs into the output inside a model.

We proposed different nonparametric methods to tackle those issues. We will summarize briefly the results that we obtained and give some perspectives about the ongoing work.

The Taylor estimator

In Chapter 2 we presented an estimator based on the Taylor decomposition of a general functional T associated with the coordinate-wise conditional covariance. We used a general basis of orthonormal functions to project the functional T

onto its span. We found that the variance of the Taylor estimator depends only on the linear part of its development. This characteristic, causes that the Taylor estimator is efficient from the Cramér-Rao point of view. In other words, it attains the minimum variance among the other estimators. Finally we found the asymptotic variance for the whole matrix estimator and proved its asymptotic normality.

There are still several topics about this subject to explore. It would be interesting study the behavior of our estimator using particular sets of functional basis like wavelets, splines or polynomial functions (Legendre, Laguerre and so on). We explained briefly this topic with a general wavelet basis, but the idea is to expand it to a wider set of basis and find some simplifications in the estimator. This steps will allow us to implement it easier in future applications.

Another point worth to investigate the Cramér-Rao efficiency of the whole matrix estimator. We can provide the result with a half-vectorized version of the conditional covariance matrix.

The kernel-based estimator

We proposed a nonparametric estimator for the conditional covariance in Chapter 3 and used the same idea for to find an estimator for the first-order Sobol indices in Chapter 4. We proved that under some mild conditions, the element-wise estimator's mean squared error achieves a parametric rate of convergence. In the case of Chapter 3, we observed a similar behavior for the full matrix estimator under the Frobenius norm. We performed simulations with some test cases.

One of the key point in the methodology was split the sample in two equal parts to get the independence between the numerator and denominator. One improvement to our methodology is to find some way to split the sample adaptatively preserving the result. Also, we limited our work to the Frobenius norm, but it would be interesting find rates of convergence under other norms like the operator-norm.

One of the most encouraging ideas ongoing in both estimators are the minimax and adaptive rates of convergence. There are plenty of literature about minimax lower bound for the sample covariance, for instance Cai et al. (2010). Find a rate of this kind, would prove that our estimators are optimal under some regularity conditions. Moreover, given that our rate depends on the regularity level β, search adaptive rates of convergence will produce attractive results.

Real case simulations

We made some simulations in Chapters 3 and 4 with testing models for the conditional covariance and variance respectively. We used the nonparametric estimators in examples of reduction dimension and sensitivity analysis. For the Taylor estimator, proposed in Chapter 2, we have already an incomplete implementation which needs to be improved for real simulations. Therefore, we have two main objectives in the near future: First, polish the Taylor estimator to get a functional code base and use those implementations in real-case data and test its capabilities.

Appendix

These programs were made using the R project version 3.0.2.

Nonparametric estimator for conditional covariance

```
library(np)
library(Matrix)

estimate_conditional_expectation ← function(i, dataX, dataY,
    fhat){
  n ← nrow(dataY)
  # Estimating the \hat{g} for each i
  ghat ← npksum(
    txdat = dataY[[1]],
    tydat = dataX[[i]],
    bws = n^(-1/4),
    leave.one.out = TRUE,
    bandwidth.divide = TRUE)$ksum/n

  g_div_f ← ghat / fhat

  return(g_div_f)
}

compute_sigma←function(data){
  # Function to estimate the regularized estimator of
  # the conditional covariance C(E(X | Y))
```

```
# Setting the values of n and p
n ← nrow(data$df)
p ← ncol(data$df)−1
alpha ← data$alpha

# Split the sample in two equal parts
n1←floor(n/2)
rows1 ← 1:n1
rows2 ← (n1+1):n
colsY ← 1
colsX ← 2:(p+1)

# use np to estimate \hat{f}_Y
fhat ← npudens(
  tdat=data$df[rows2,colsY,drop=FALSE],
  edat=data$df[rows1,colsY,drop=FALSE])

# Cut the small value to avoid incosistencies
threshold ← fivenum(fhat$dens)[2]
fhat$dens[fhat$dens<threshold] ← threshold

# Estimate hat{g} function (vector)
ghat←sapply(seq(1,p), estimate_conditional_expectation,
            dataX=data$df[rows1,colsX,drop=FALSE],
            dataY=data$df[rows1,colsY,drop=FALSE],
            fhat=fhat$dens)

# With the vector of \hat{g}, estimate
# its covariance
matSigma ← cov(ghat)

# Parameter for the banding matrix
k ← n1^(1/(2*(alpha+1)))

if (k ≤ p && k > 0){
  # Construct a banding regularizator
  Regularizator ← toeplitz(
    c(rep(1,k),
      numeric(p−(k−1))))
  # Apply the banding matrix to Sigma
  matSigma ← matSigma * Regularizator
```

```
  }else{}
  matSigma ← Matrix(matSigma, sparse=TRUE)

  return(list(SigmaHat=matSigma))
}
```

Nonparametric estimator for Sobol indices

```
Ksobol ← function(X,Y){
  # Function to estimate the Sobol index
  # S = Var(E[Y | X])\V(Y).
  # We assume some model with
  # X is the input and,
  # Y is the output

  # Setting the parametres
  n ← dim(X)[1]
  p ← dim(X)[2]
  g ← matrix(nrow=n,ncol=p)

  # Estimating the conditional expectations
  # with the package np
  for(k in 1:p){
    g[,k] ← npreg(txdat=X[,k],tydat=Y)$mean
  }

  # Estimating the Variance of Y
  sigma.Y ← var(Y)
  # Creating a vector of V(E[Y | X])
  var.g ← apply(g,2,var)

  # Computing the Sobol index
  sobol.idx ← var.g/sigma.Y
}
```

References

Almuallim, H. and Dietterich, T. G. (1994). Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence*, 69(1-2):279–305.

Amaldi, E. and Kann, V. (1998). On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1-2):237–260.

Autin, F., Pennec, E., Loubes, J. M., and Rivoirard, V. (2009). Maxisets for Model Selection. *Constructive Approximation*, 31(2):195–229.

Baraud, Y., Huet, S., and Laurent, B. (2003). Adaptive tests of qualitative hypotheses. *ESAIM: Probability and Statistics*, 7(January):147–159.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396.

Bellman, R. (2003). *Dynamic Programming*. Dover Books on Computer Science Series. Dover Publications.

Bellman, R., Bellman, R. E., Bellman, R. E., and Bellman, R. E. (1961). *Adaptive control processes: a guided tour*, volume 4. Princeton university press Princeton.

Bennett, R. (1969). The intrinsic dimensionality of signal collections. *IEEE Transactions on Information Theory*, 15(5):517–525.

Bickel, P. and Lindner, M. (2012). Approximating the inverse of banded matrices by banded matrices with applications to probability and statistics. *Theory of Probability & Its Applications*, 56(1):1–20.

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271.

Box, G. and Draper, N. (1987). *Empirical model-building and response surfaces*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Breiman, L. and Friedman, J. H. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association*, 80(391):580–598.

Brillinger, D. R. (1983). A generalized linear model with "gaussian" regressor variables. In *A Festschrift for Erich L. Lehmann in Honor of His Sixty-Fifth Birthday*, Wadsworth Statist./Probab. Ser., page 97–114. Wadsworth, Belmont, CA.

Brillinger, D. R. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):333–335.

Bura, E. and Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):393–410.

Burges, C. J. C. (2009). Dimension reduction: A guided tour. *Foundations and Trends® in Machine Learning*, 2(4):275–364.

Cadre, B. and Dong, Q. (2010). Dimension reduction for regression estimation with nearest neighbor method. *Electronic Journal of Statistics*, 4:436–460.

Cai, T. T., Ren, Z., and Zhou, H. H. (2012). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probability Theory and Related Fields*, 156(1-2):101–143.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144.

Cai, T. T. and Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420.

Campolongo, F., Saltelli, A., and Cariboni, J. (2011). From screening to quantitative sensitivity analysis. A unified approach. *Computer Physics Communications*, 182(4):978–988.

Carmichael, G. R., Sandu, A., and Potra, f. A. (1997). Sensitivity analysis for atmospheric chemistry models via automatic differentiation. *Atmospheric Environment*, 31(3):475–489.

Cayton, L. (2005). Algorithms for manifold learning. *University of California, San Diego, Tech. Rep. CS2008-0923*.

Chen, Y., Wiesel, A., Eldar, Y. C., and Hero, A. O. (2010). Shrinkage Algorithms for MMSE Covariance Estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029.

Christopher Frey, H. and Patil, S. R. (2002). Identification and Review of Sensitivity Analysis Methods. *Risk Analysis*, 22(3):553–578.

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30.

Cook, R. (1998). *Regression Graphics*, volume 482 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Cook, R. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89(425):177–189.

Cook, R. D. (2003). Dimension reduction and graphical exploration in regression including survival analysis. *Statistics in medicine*, 22(9):1399–413.

Cook, R. D. and Nachtsheim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association*, 89(426):592–599.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association*, 100(470):410–428.

Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332.

Cornish-Bowden, A. (2013). *Fundamentals of enzyme kinetics*. Portland Press, London, UK, revised ed edition.

Cukier, R. I., Levine, H. B., and Shuler, K. E. (1978). Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics*, 26(1):1–42.

Cukier, R. I., Schaibly, J. H., and Shuler, K. E. (1973). Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. i theory. *The Journal of Chemical Physics*, 59(8):3873.

Cullen, A. and Frey, H. (1999). *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs*. Springer.

Da Veiga, S. and Gamboa, F. (2013). Efficient estimation of sensitivity indices. *Journal of Nonparametric Statistics*, 25(3):573–595.

Da Veiga, S., Loubes, J.-M., and Solís, M. (2011). Efficient estimation of conditional covariance matrices for dimension reduction. *Arxiv preprint arXiv:*.

Devore, J. and Peck, R. (1996). Statistics: the exploration and analysis of data. brooks. *Cole, Pacific Grove, CA*.

Donoho, D. L. and Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1996). Density estimation by wavelet thresholding. *The Annals of Statistics*, 24(2):508–539.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Donoho, D. L. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, page 1–32.

Draper, N. and Smith, H. (1981). *Applied regression analysis*. Wiley, 2, illustr edition.

Duan, N. and Li, K.-C. (1991). Slicing regression: A link-free regression method. *The Annals of Statistics*, 19(2):505–530.

Duffée, G. R. (2005). Time Variation in the Covariance between Stock Returns and Consumption Growth. *The Journal of Finance*, 60(4):1673–1712.

Dyer, J. S. and Owen, A. B. (2011). Visualizing bivariate long-tailed data. *Electronic Journal of Statistics*, 5:642–668.

Efromovich, S. (1999). *Nonparametric curve estimation: methods, theory and applications*. Springer.

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, 36(6):2717–2756.

Engel, D., Hüttenberger, L., and Hamann, B. (2012). A Survey of Dimension Reduction Methods for High-dimensional Data Analysis and Visualization. In Garth, C., Middel, A., and Hagen, H., editors, *Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering - Proceedings of IRTG 1131 Workshop 2011*, volume 27 of *OpenAccess Series in Informatics (OASIcs)*, pages 135–149, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Eubank, R. (1999). *Nonparametric regression and spline smoothing*. CRC press.

Fan, J. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3):645–660.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197.

Fan, J. and Gijbles, I. (1996). Local polynomial smoothing.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *Proceedings of the International Congress of Mathematicians,*, page 595–622, Madrid, Spain. European Mathematical Society Publishing House.

Fan, J., Lv, J., and Qi, L. (2011). Sparse high dimensional models in economics. *Annual review of economics*, 3(2006):291–317.

Ferré, L., Yao, A., et al. (2005). Smoothed functional inverse regression. *Statistica Sinica*, 15(3):665.

Ferré, L. and Yao, A.-F. (2003). Functional sliced inverse regression analysis. *Statistics*, 37(6):475–488.

Fisher, T. J. and Sun, X. (2011). Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Computational Statistics & Data Analysis*, 55(5):1909–1918.

Fodor, I. K. (2002). A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Laboratory (LLNL), Livermore, CA.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31(1):3–21.

Friedman, J. H. and Stuetzle, W. (1981). Projection Pursuit Regression. *Journal of the American Statistical Association*, 76(376):817–823.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C-23(9):881–890.

Fukunaga, K. (1972). *Introduction to statistical pattern recognition*. Electrical science series. Academic Press.

Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255.

Gennari, J. H., Langley, P., and Fisher, D. (1989). Models of incremental concept formation. *Artificial intelligence*, 40(1):11–61.

Gilad-Bachrach, R., Navot, A., and Tishby, N. (2004). Margin based feature selection-theory and algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 43. ACM.

Goos, P. (2002). *The Optimal Design of Blocked and Split-Plot Experiments*, volume 164 of *Lecture Notes in Statistics*. Springer New York, New York, NY.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Springer US, Boston, MA.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.

Hall, P. and Li, K.-C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The annals of Statistics*, pages 867–889.

Hammersley, J. (1960). Monte Carlo methods for solving multivariable problems. *Annals of the New York Academy of . . .* , 86(3):844–874.

Hardle, W. (1990). *Applied nonparametric regression*, volume 5. Cambridge Univ Press.

Härdle, W. (2004). *Nonparametric and semiparametric models*. Springer.

Hardle, W. and Tsybakov, A. B. (1991). Sliced Inverse Regression for Dimension Reduction: Comment. *Journal of the American Statistical Association*, 86(414):pp. 333——335.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Monographs on Statistics and Applied Probability ;; 43. Chapman & Hall/CRC, 1st ed. edition.

Hayfield, T. and Racine, J. (2008). Nonparametric econometrics: The np package. *Journal of statistical software*, 27(5).

Heeger, D. J. and Ress, D. (2002). What does fmri tell us about neuronal activity? *Nature reviews. Neuroscience*, 3(2):142–51.

Helton, J. C., Johnson, J. D., Sallaberry, C. J., and Storlie, C. B. (2006). Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering & System Safety*, 91(10-11):1175–1209.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Hsing, T. (1999). Nearest neighbor inverse regression. *Annals of statistics*, 27(2):697–731.

Huang, J. Z. J. Z. Z. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

Huber, P. J. (1981). *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, 13(2):435–475.

Hull, J. (2011). *Options, futures, and other derivatives*. Prentice Hall, Toronto, 8 edition.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications*, volume 129 of *Lecture Notes in Statistics*. Springer New York, New York, NY.

Härdle, W. and Tsybakov, A. B. (1991). Comment. *J. Am. Stat. Assoc.*, 86(414):333.

Ibragimov, I. and Khas' minskii, R. (1983). Estimation of distribution density. *Journal of Soviet Mathematics*, 21(1):40–57.

Ibragimov, I. and Khas' minskii, R. (1984). More on the estimation of distribution densities. *Journal of Soviet Mathematics*, 25(3):1155–1165.

Ibragimov, I. A. and Khas'Minskii, R. Z. (1991). Asymptotically normal families of distributions and efficient estimation. *The Annals of Statistics*, 19(4):1681–1724.

Ishigami, T. and Homma, T. (1990). An importance quantification technique in uncertainty analysis for computer models. In *Uncertainty Modeling and Analysis, 1990. Proceedings., First International Symposium on*, page 398–403. IEEE.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. Springer Texts in Statistics. Springer New York, New York, NY, 1 edition.

Janon, A. (2012). *Analyse de sensibilité et réduction de dimension. Application à l'océanographie.* PhD thesis, Université de Grenoble.

Janon, A., Klein, T., Lagnoux, A., Nodet, M., and Prieur, C. (2013). Asymptotic normality and efficiency of two sobol index estimators. *ESAIM: Probability and Statistics*, eFirst.

John, G. H., Kohavi, R., Pfleger, K., et al. (1994). Irrelevant features and the subset selection problem. *ICML*, 94:121–129.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327.

Jones, M. C. and Sibson, R. (1987). What is projection pursuit? *Journal of the Royal Statistical Society. Series A (General)*, page 1–37.

Karhunen, K. (1946). *Zur spektraltheorie stochastischer prozesse*. Suomalainen tiedeakatemia.

Kedem, G. (1980). Automatic differentiation of computer programs. *ACM Transactions on Mathematical Software (TOMS)*, 6(2):150–165.

Kent, J. T. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):336–337.

Kerkyacharian, G. and Picard, D. (2002). Minimax or maxisets? *Bernoulli*, 8(2):219–253.

Kira, K. and Rendell, L. A. (1992). A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, page 249–256. Morgan Kaufmann Publishers Inc.

Koda, M., Mcrae, G. J., and Seinfeld, J. H. (1979). Automatic sensitivity analysis of kinetic mechanisms. *International Journal of Chemical Kinetics*, 11(4):427–444.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324.

Komsta, L. and Novomestky, F. (2011). *moments: Moments, cumulants, skewness, kurtosis and related tests*.

Kowalski, J. and Tu, X. M. (2007). *Modern Applied U-Statistics*, volume 714 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Lam, C. and Fan, J. (2009). Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation. *Annals of statistics*, 37(6B):4254–4278.

Langewisch, D. (2010). *Uncertainty and sensitivity analysis for long-running computer codes: a critical review*. PhD thesis, Massachusetts Institute of Technology.

Laurent, B. (1996). Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2):659–681.

Laurent, B. (2005). Adaptive estimation of a quadratic functional of a density by model selection. *ESAIM: Probability and Statistics*, 9(February):1–18.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Lee, J. and Verleysen, M. (2007). *Nonlinear Dimensionality Reduction*. Information Science and Statistics. Springer New York, New York, NY.

Levina, E., Rothman, A., and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, 2(1):245–263.

Li, G., Rosenthal, C., and Rabitz, H. (2001). High dimensional model representations. *The Journal of Physical Chemistry A*, 105(33):7765–7777.

Li, K.-C. (1991a). Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.*, 86(414):316–327.

Li, K.-C. (1991b). Sliced inverse regression for dimension reduction: Rejoinder. *Journal of the American Statistical Association*, 86(414):337–342.

Li, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039.

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613.

Lorenz, E. N. (1956). Empirical orthogonal functions and statistical weather prediction. *Technical report Statistical Forecast Project Report 1 Department of Meteorology MIT 49*, 1:52.

Loubes, J. and Yao, A. (2013). Kernel inverse regression for random fields. *International Journal of Applied Mathematics and Statistics*, 32(2):1–26.

Loubes, J.-M., Marteau, C., and Solís, M. (2013). Rates of convergence in conditional covariance matrix estimation. *arXiv preprint arXiv:1310.8244*, page 1–29.

Loève, M. (1955). *Probability Theory. Foundations. Random Sequences*. Van Nostrand, New York.

Lumley, J. L. (1967). The structure of inhomogeneous turbulent flows. *Atmospheric turbulence and radio wave propagation*, page 166–178.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg. Editor:Picard, Jean.

McRae, G. J., Tilden, J. W., and Seinfeld, J. H. (1982). Global sensitivity analysis—a computational implementation of the Fourier amplitude sensitivity test (fast). *Computers & Chemical Engineering*, 6(1):15–25.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Meyer, Y. (1990). Ondelettes et op{é}rateurs, vol. i. *Hermann, Paris*.

Meyer, Y. and Salinger, D. H. (1993). *Wavelets and Operators*. Cambridge Studies in Advanced Mathematics ; 37. Cambridge University Press, Cambridge.

Moussa, M. A. A. and Cheema, M. Y. (1992). Non-parametric regression in curve fitting. *The Statistician*, 41(2):209.

Muirhead, R. J. (1987). Developments in eigenvalue estimation. In *Advances in multivariate statistical analysis*, Theory Decis. Lib. Ser. B: Math. Statist. Methods, page 277–288. Springer, Dordrecht.

Myers, R. H., Montgomery, D. C., and Anderson-Cook, C. M. (2009). *Response surface methodology: process and product optimization using designed experiments*. Wiley, New York.

Nadaraya, E. (1964). On Estimating Regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Oakley, J. E. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):751–769.

Pappenberger, F., Ratto, M., and Vandenberghe, V. (2010). Review of sensitivity analysis methods. In Vanrolleghem, P. A., editor, *Modelling aspects of water framework directive implementation*, page 191–265. IWA Publishing.

Payne, T. R. and Edwards, P. (1998). Implicit feature selection with the value difference metric. In *European Conference on Artificial Intelligence*, page 450–454.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572.

Perkins, S. and Theiler, J. (2003). Online feature selection using grafting. In *In International Conference on Machine Learning*, page 592–599. ACM Press.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag.

Pérez-González, A., Vilar-Fernández, J. M., and González-Manteiga, W. (2010). Nonparametric variance function estimation with missing data. *Journal of Multivariate Analysis*, 101(5):1123–1142.

Quinlan, J. R. (2007). Induction of decision trees. *Readings in Machine Learning*, page 81–106.

Rall, L. B. (1980). *Applications of software for automatic differentiation in numerical computation*. Springer.

Rao, B. and Prakasa Rao, B. L. S. (1983). *Nonparametric Functional Estimation*. Probability and mathematical statistics. Academic Press.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 12(4):1215–1230.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2(January):494–515.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Ruppert, D., Wand, M. P., Holst, U., and HöSJER, O. (1997). Local polynomial variance-function estimation. *Technometrics*, 39(3):262–273.

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145(2):280–297.

Saltelli, A., Campolongo, F., and Cariboni, J. (2009). Screening important inputs in models with strong interaction properties. *Reliability Engineering & System Safety*, 94(7):1149–1155.

Saltelli, A., Chan, K., and Scott, E. (2000). *Sensitivity analysis*, volume 134. Wiley New York.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global sensitivity analysis: the primer*. Wiley. com.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004). *Sensitivity analysis in practice: a guide to assessing scientific models*. John Wiley & Sons.

Saltelli, A., Tarantola, S., and Chan, K.-S. (1999). A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, 41(1):39–56.

Scott, D. W. D. and Thompson, J. R. J. (1983). Probability density estimation in higher dimensions. In *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, volume 528, page 173–179. North-Holland, Amsterdam.

Setodji, C. M. and Cook, R. D. (2004). K -means inverse regression. *Technometrics*, 46(4):421–429.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, 68(1):45–54.

Simon, H. (1969). *The sciences of the artificial*. MIT press, 3rd ed. edition.

Sobol', I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical modeling and computational experiment*, 1(4):407–414.

Sobol', I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1-3):271–280.

Sobol', I. M. and Levitan, Y. (1999). On the use of variance reducing multipliers in Monte Carlo computations of a global sensitivity index. *Computer Physics Communications*, 117(1-2):52–61.

Stears, R. L., Martinsky, T., and Schena, M. (2003). Trends in microarray analysis. *Nature medicine*, 9(1):140–145.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053.

Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling: 1994 Wald memorial lecture. *The Annals of Statistics*, 25(4):1371–1470.

Tarantola, S., Gatelli, D., and Mara, T. A. (2006). Random balance designs for the estimation of first order global sensitivity indices. *Reliability Engineering & System Safety*, 91(6):717–727.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Tissot, J.-Y. and Prieur, C. (2012). Bias correction for the estimation of sensitivity indices based on random balance designs. *Reliability Engineering & System Safety*, 107(0):205–213.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer Verlag.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3 of *Cambridge Series on Statistical and Probabilistic Mathematics*. Cambridge University Press.

Wand, M. M. P. and Jones, M. M. C. (1995). *Kernel smoothing*, volume 60. Crc Press.

Wasserman, L. (2007). *All of nonparametric statistics*. Springer Texts in Statistics. Springer, New York.

Watson, G. (1964). Smooth regression analysis. *Sankhy\={a}: The Indian Journal of Statistics, Series A*, 26:359–372.

Wu, W. and Pourahmadi, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, 19:1755–1768.

Yin, J., Geng, Z., Li, R., and Wang, H. (2010). Nonparametric covariance model. *Statistica Sinica*, 20:469–479.

Yoo, J. K. (2008a). A novel moment-based sufficient dimension reduction approach in multivariate regression. *Computational Statistics & Data Analysis*, 52(7):3843–3851.

Yoo, J. K. (2008b). Sufficient dimension reduction for the conditional mean with a categorical predictor in multivariate regression. *Journal of Multivariate Analysis*, 99(8):1825–1839.

Yoo, J. K. and Cook, R. D. (2007). Optimal sufficient dimension reduction for the conditional mean in multivariate regression. *Biometrika*, 94(1):231–242.

Zhang, Z.-y. and Zha, H.-y. (2004). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *Journal of Shanghai University (English Edition)*, 8(4):406–424.

Zhu, L., Zhu, L., and Li, X. (2007). Transformed partial least squares for multivariate data. *Statistica Sinica*, 17:1657–1675.

Zhu, L.-p. and Yu, Z. (2007). On spline approximation of sliced inverse regression. *Science in China Series A: Mathematics*, 50(9):1289–1302.

Zhu, L.-X. and Fang, K.-T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, 24(3):1053–1068.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

Résumé

Cette thèse se concentre autour du problème de l'estimation de matrices de covariance conditionnelles et ses applications, en particulier sur la réduction de dimension et l'analyse de sensibilités.

Dans le Chapitre 2 nous plaçons dans un modèle d'observation de type régression en grande dimension pour lequel nous souhaitons utiliser une méthodologie de type *régression inverse par tranches*. L'utilisation d'un opérateur fonctionnel, nous permettra d'appliquer une décomposition de Taylor autour d'un estimateur préliminaire de la densité jointe. Nous prouverons deux choses : notre estimateur est asymptoticalement normale avec une variance que dépend de la partie linéaire, et cette variance est efficace selon le point de vue de Cramér-Rao.

Dans le Chapitre 3, nous étudions l'estimation de matrices de covariance conditionnelle dans un premier temps coordonnée par coordonnée, lesquelles dépendent de la densité jointe inconnue que nous remplacerons par un estimateur à noyaux. Nous trouverons que l'erreur quadratique moyenne de l'estimateur converge à une vitesse paramétrique si la distribution jointe appartient à une classe de fonctions *lisses*. Sinon, nous aurons une vitesse plus lent en fonction de la régularité de la densité de la densité jointe. Pour l'estimateur de la matrice complète, nous allons appliquer une transformation de régularisation de type *"banding"*.

Finalement, dans le Chapitre 4, nous allons utiliser nos résultats pour estimer des indices de Sobol utilisés en analyses de sensibilité. Ces indices mesurent l'influence des entrées par rapport a la sortie dans modèles complexes. L'avantage de notre implémentation est d'estimer les indices de Sobol sans l'utilisation de coûteuses méthodes de type Monte-Carlo. Certaines illustrations sont présentées dans le chapitre pour montrer les capacités de notre estimateur.

Abstract

This thesis will be focused in the estimation of conditional covariance matrices and their applications, in particular, in dimension reduction and sensitivity analyses.

In Chapter 2, we are in a context of high-dimensional nonlinear regression. The main objective is to use the *sliced inverse regression* methodology. Using a functional operator depending on the joint density, we apply a Taylor decomposition around a preliminary estimator. We will prove two things: our estimator is asymptotical normal with variance depending only the linear part, and this variance is efficient from the Cramér-Rao point of view.

In the Chapter 3, we study the estimation of conditional covariance matrices, first coordinate-wise where those parameters depend on the unknown joint density which we will replace it by a kernel estimator. We prove that the mean squared error of the nonparametric estimator has a parametric rate of convergence if the joint distribution belongs to some class of *smooth* functions. Otherwise, we get a slower rate depending on the regularity of the model. For the estimator of the whole matrix estimator, we will apply a regularization of type *"banding"*.

Finally, in Chapter 4, we apply our results to estimate the Sobol or sensitivity indices. These indices measure the influence of the inputs with respect to the output in complex models. The advantage of our implementation is that we can estimate the Sobol indices without use computing expensive Monte-Carlo methods. Some illustrations are presented in the chapter showing the capabilities of our estimator.