Stochastic modelling using large data sets : applications in ecology and genetics

Raphaël Coudret

▶ **To cite this version:**

Raphaël Coudret. Stochastic modelling using large data sets : applications in ecology and genetics. General Mathematics [math.GM]. Université Sciences et Technologies - Bordeaux I, 2013. English. NNT : 2013BOR14838 . tel-00865867

HAL Id: tel-00865867

https://theses.hal.science/tel-00865867

Submitted on 25 Sep 2013

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

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

N° ordre : 4838

Thèse

présentée à

l'Université Bordeaux 1

Ecole doctorale de Mathématiques et Informatique

par

Raphaël Coudret

pour obtenir le grade de

Docteur

Spécialité : Mathématiques appliquées

Stochastic modelling using large data sets: applications in ecology and genetics

---

Soutenue le 16 septembre 2013 devant la commission d'examen composée de

M. Gérard Biau	Prof.	Université Paris VI	Rapporteur
M. Gilles Durrieu	Prof.	Université de Bretagne-Sud	Dir. de thèse
M. Pierrick Legrand	MCF.	Université Bordeaux Segalen	Examinateur
M. Jean-Charles Massabuau	DR.	CNRS & Université Bordeaux 1	Examinateur
M. Stéphane Robin	DR.	INRA & AgroParisTech	Rapporteur
M. Jérôme Saracco	Prof.	Institut Polytechnique de Bordeaux	Dir. de thèse

Institut de Mathématiques de Bordeaux, UMR 5251,
Université Bordeaux 1
352 cours de la Libération
33405 Talence Cedex

Environnements et Paléoenvironnements Océaniques et Continentaux, UMR CNRS 5805,
Université Bordeaux 1
Place du Dr Peyneau
33120 Arcachon

INRIA Bordeaux - Sud-Ouest
200 avenue de la Vieille Tour
33405 Talence Cedex

Remerciements

Je tiens à remercier dans un premier temps mes directeurs de thèse Gilles Durrieu et Jérôme Saracco. Ce travail doit beaucoup à leur pédagogie, à leurs encouragements et à leur capacité à jalonner les projets de recherche auxquels j'ai participé. J'ai ainsi pu, grâce à eux, acquérir de nombreuses connaissances tout en conservant la capacité d'exprimer mes propres idées.

Je suis reconnaissant envers Gérard Biau et Stéphane Robin pour avoir accepté de rapporter ma thèse et pour le temps qu'ils ont passé à considérer ce document. Merci aussi à Pierrick Legrand et à Jean-Charles Massabuau pour avoir bien voulu prendre part à ce jury de thèse.

Je remercie l'ensemble des membres des équipes CQFD et ALEA d'INRIA pour l'environnement stimulant, épanouissant et productif qu'ils ont su créer. Merci aux scientifiques du laboratoire EPOC de m'avoir fourni un contexte biologique qui donne un sens pratique aux résultats de recherche présentés ici. Merci également à Stéphane Girard et à Benoît Liquet pour nos fécondes collaborations.

Enfin, plus généralement, simplement remercier ma famille, mes amis et mes collègues qui m'ont aidé durant ces trois dernières années me semblerait annoncer la fin de ce qui nous unis. Ce n'est pas mon intention.

Modélisation stochastique de grands jeux de données : applications en écologie et en génétique

Résumé

Deux parties principales composent cette thèse. La première d'entre elles est consacrée à la valvométrie, c'est-à-dire ici l'étude de la distance entre les deux parties de la coquille d'une huître au cours du temps. La valvométrie est utilisée afin de déterminer si de tels animaux sont en bonne santé, pour éventuellement tirer des conclusions sur la qualité de leur environnement. Nous considérons qu'un processus de renouvellement à quatre états sous-tend le comportement des huîtres étudiées. Afin de retrouver ce processus caché dans le signal valvométrique, nous supposons qu'une densité de probabilité reliée à ce signal est bimodale. Nous comparons donc plusieurs estimateurs qui prennent en compte ce type d'hypothèse, dont des estimateurs à noyau.

Dans un second temps, nous comparons plusieurs méthodes de régression, dans le but d'analyser des données transcriptomiques. Pour comprendre quelles variables explicatives influent sur l'expression de gènes, nous avons réalisé des tests multiples grâce au modèle linéaire FAMT. La méthode SIR peut être envisagée pour trouver des relations non-linéaires. Toutefois, elle est principalement employée lorsque la variable à expliquer est univariée. Une version multivariée de cette approche a donc été développée. Le coût d'acquisition des données transcriptomiques pouvant être élevé, la taille n des échantillons correspondants est souvent faible. C'est pourquoi, nous avons également étudié la méthode SIR lorsque n est inférieur au nombre de variables explicatives p.

Mots-clés : données transcriptomiques, estimateur à noyau, processus de renouvellement, régression inverse par tranches, tests multiples, valvométrie.

Stochastic modelling using large data sets: applications in ecology and genetics

Abstract

There are two main parts in this thesis. The first one concerns valvometry, which is here the study of the distance between both parts of the shell of an oyster, over time. The health status of oysters can be characterized using valvometry in order to obtain insights about the quality of their environment. We consider that a renewal process with four states underlies the behaviour of the studied oysters. Such a hidden process can be retrieved from a valvometric signal by assuming that some probability density function linked with this signal, is bimodal. We then compare several estimators which take this assumption into account, including kernel density estimators.

In another chapter, we compare several regression approaches, aiming at analysing transcriptomic data. To understand which explanatory variables have an effect on gene expressions, we apply a multiple testing procedure on these data, through the linear model FAMT. The SIR method may find nonlinear relations in such a context. It is however more commonly used when the response variable is univariate. A multivariate version of SIR was then developed. Procedures to measure gene expressions can be expensive. The sample size n of the corresponding datasets is then often small. That is why we also studied SIR when n is less than the number of explanatory variables p.

Keywords: kernel density estimator, multiple testing, renewal process, sliced inverse regression, transcriptomics, valvometry.

Contents

Chapter 1

Introduction générale

Cette thèse contient des exemples de modélisation en biologie basés sur des outils statistiques existants ou originaux et sur des jeux de données de grande taille, principalement fournis par le laboratoire EPOC (Environnements et Paléoenvironnements Océaniques et Continentaux) de l'Université de Bordeaux. Après deux chapitres d'introduction respectivement en français et en anglais, le Chapitre 3 présente les travaux réalisés sur des mesures de l'activité d'huîtres. Le Chapitre 4 est dédié au développement et à l'utilisation de méthodes statistiques adaptées aux données transcriptomiques. Le Chapitre 5 conclut cet ouvrage. Nous y décrivons également quelques perspectives de recherche.

1.1 Un modèle stochastique pour des bioindicateurs

Dans le Chapitre 3, nous cherchons à déterminer comment résumer l'information contenue dans de grands échantillons. En effet, le signal considéré est formé de distances entre les deux parties de la coquille d'une huître, mesurées à haute fréquence. On dit alors qu'il s'agit d'un signal valvométrique. L'un des objectifs de ces enregistrements est de réaliser un diagnostic de l'état de santé de l'animal. Lorsque plusieurs animaux placés à un même endroit sont en mauvaise santé, un changement dans leur environnement est une cause privilégiée. L'étude de données valvométriques est donc une étape importante dans l'utilisation d'huîtres comme bioindicateur de la qualité de l'eau. Nous présentons plus en détail le contexte biologique relatif aux enregistrements valvométriques dans l'introduction du Chapitre 3 et dans les Sections 3.1.1-3.1.2.

Comme le montre la Figure 1.1, il est possible de distinguer deux états de l'animal dans un signal valvométrique. Nous appelons ces états "ouvert" et "fermé". Lorsque l'huître est en bonne santé, les biologistes ont observé que les transitions d'un état à l'autre sont reliées à la marée, ce qui n'est pas forcément visible si l'animal présente des difficultés à s'adapter à son environnement. En utilisant un processus de renouvellement dont "ouvert" et "fermé" sont deux états, et un estimateur de sa fonction de survie, nous montrons dans A hidden renewal model for monitoring aquatic systems biosensors (voir List of works) comment il est possible de réaliser une classification des animaux à partir de leur signal valvométrique. Le résultat de cette classification s'avère être en adéquation avec l'information dont nous disposons concernant leur état de santé. Cet article constitue la Section 3.1.

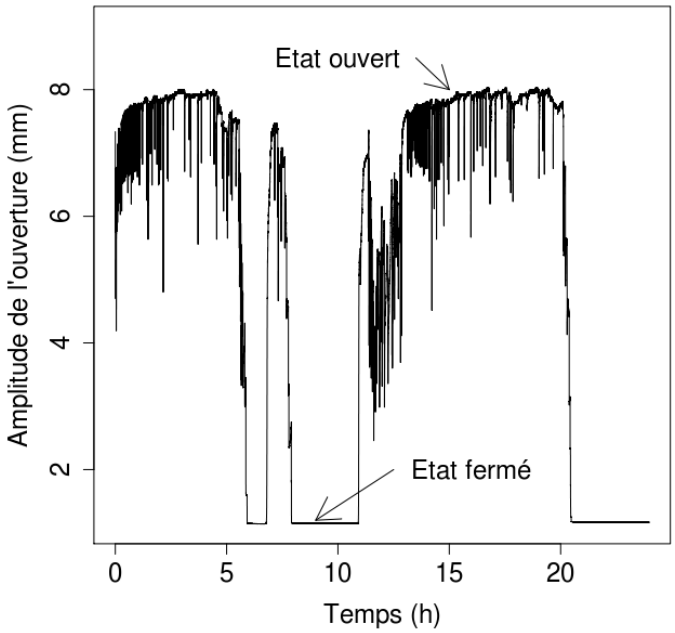


Figure 1.1: Un exemple de données valvométriques, pour une huître de la jetée d'Eyrac, le 27 février 2007

Le processus de renouvellement est constant par morceaux alors que le signal observé (Figure 1.1) est continu. De plus, ce signal présente des zones où il varie très rapidement vers le bas pour revenir ensuite très rapidement à une valeur élevée. Ces pics sont appelés microfermetures et leur minimum peut parfois être tellement bas qu'il est difficile de distinguer microfermetures et passages de l'état "ouvert" à l'état "fermé". Nous présentons donc dans la Section 3.1 une méthode pour estimer le processus de renouvellement caché dans le signal valvométrique observé. Cette procédure permet de déterminer un seuil. Lorsque le signal valvométrique est en dessous de ce seuil, nous décidons que le processus sous-jacent est dans l'état fermé. Sinon, il est dans l'état ouvert. Pour trouver le seuil en question, nous estimons une densité de probabilité reliée au signal valvométrique. Nous supposons que cette densité a deux modes : l'un correspond à l'état ouvert de l'huître et l'autre à l'état fermé. Nous choisissons un estimateur qui a également deux modes. La position de l'unique minimum local de cet estimateur est alors choisi comme seuil.

Plusieurs méthodes ont été conçues afin d'estimer une densité de probabilité f lorsque nous avons à notre disposition des observations indépendantes en provenance d'une même loi de densité f, ainsi que le nombre de modes de f. Nous pouvons par exemple utiliser un estimateur à noyau ou un estimateur basé sur les ensembles de niveaux de f. Dans Comparison of kernel density estimators with assumption on number of modes, nous avons étudié en simulation ces estimateurs ainsi qu'un estimateur à noyau qui n'utilise pas d'hypothèse sur le nombre de modes de f. Cet article constitue la Section 3.2.

Le nombre de microfermetures et leur amplitude peut également être informatif sur l'état de santé de l'animal. Toutefois leur détection est difficile car le signal valvométrique

se révèle être bruité. Dans la Section 3.3, nous présentons une méthode existante de débruitage en utilisant les ondelettes de Haar. Celle-ci a été choisie parmi de nombreuses autres grâce aux résultats de simulation présentés dans Antoniadis *et al.* (2001). Nous étudions également plus amplement la détection de microfermetures directement à partir du signal bruité, dans cette section.

Pour faciliter l'utilisation de ces approches, une implémentation en R a été réalisée et une interface graphique a été crée grâce à la bibliothèque GTK+. De plus, ces programmes permettent d'importer des données valvométriques depuis un serveur SQL distant où elles sont stockées comme nous l'expliquons dans la Section 3.4. Des pistes y sont également décrites pour optimiser le temps de calcul lors d'un travail avec R.

1.2 Des méthodes statistiques pour analyser des données génétiques

Les problèmes biologiques abordés dans le Chapitre 4 concernent les relations entre des données transcriptomiques d'une part et des variables diverses que l'on nomme variables explicatives, d'autre part. Nous appelons données transcriptomiques des mesures de la quantité d'expression de marqueurs génétiques. Nous cherchons à expliquer les variations de ces expressions en utilisant les variables explicatives. Dans ce contexte de régression, la grande taille des jeux de données se caractérise soit par un grand nombre de variables à expliquer, c'est-à-dire un grand nombre de gènes, soit par un grand nombre de variables explicatives. Deux exemples de problème génétique sont fournis dans la Section 4.1.

Pour expliquer les variations dans l'expression d'un seul gène grâce aux variables explicatives, il est possible d'utiliser une régression linéaire. En testant quels paramètres sont significativement non nuls, nous pouvons en déduire quelles variables explicatives influent sur l'expression du gène. Lorsque nous avons affaire aux expressions de nombreux gènes, nous sommes alors confrontés au problème des tests multiples : même en présence d'une seule variable explicative, il faut réaliser un test de nullité d'un paramètre pour chaque gène.

Au lieu de contrôler la probabilité d'avoir un faux positif, nous pouvons alors être tentés de contrôler l'espérance du rapport entre le nombre de faux positifs et le nombre de tests positifs, aussi appelé *False Discovery Rate* ou FDR. Il est facile d'obtenir une série de tests de nullité de certains paramètres basée sur le FDR, si chaque test réalisé pour un gène donné est indépendant des tests correspondant à chacun des autres gènes.

Dans notre cas, cela n'est pas vrai. Nous sommes donc amenés à introduire un modèle plus complexe, afin de retirer une structure des données transcriptomiques et ainsi produire des tests indépendants. Cette approche, développée par Friguet *et al.* (2009), est appelée FAMT pour *Factor Analysis for Multiple Testing*. Elle est présentée dans la Section 4.2 dans laquelle l'implémentation d'outils informatiques pour faciliter l'utilisation de cette méthode est également détaillée.

FAMT se restreint à des relations linéaires entre expressions de gènes et variables explicatives. Pour modéliser des relations non-linéaires, il est possible d'utiliser la méthode SIR (pour *Sliced Inverse Regression*), introduite par Li (1991), qui permet d'estimer B

dans le modèle univarié suivant :

$$y = f(x'B, \varepsilon),$$

où f est une fonction quelconque de $\mathbb{R}^{K+1}$ dans $\mathbb{R}$, x est une variable aléatoire de dimension p, y et ε sont des variables aléatoires univariées, x et ε sont indépendants et B est une matrice $p \times K$ de rang K, avec $K \leq p$. Ainsi présenté, ce modèle ne permet pas de gérer des expressions de plusieurs gènes. Toutefois, si nous avons à notre disposition les expressions de q gènes, que l'on note $y^{(1)}, \ldots, y^{(q)}$, nous pouvons le remplacer par le modèle suivant :

$$\begin{cases} y^{(1)} = f_1(x'B, \varepsilon^{(1)}), \\ \qquad \vdots \\ y^{(q)} = f_q(x'B, \varepsilon^{(q)}), \end{cases} \tag{1.1}$$

où $f_1, \ldots, f_q$ sont des fonctions quelconques de $\mathbb{R}^{K+1}$ dans $\mathbb{R}$ et $\varepsilon^{(1)}, \ldots, \varepsilon^{(q)}$ sont des variables aléatoires univariées indépendantes de x. On appelle espace EDR, l'espace engendré par B. Pour tout $j = 1, \ldots, q$, en utilisant la méthode SIR, nous sommes en mesure d'obtenir un estimateur $\widehat{B}_j$ de B grâce à la $j^{\text{ème}}$ équation du modèle (1.1). Soit $\widehat{\mathbb{B}} = \left[\widehat{B}_1, \ldots, \widehat{B}_q\right]$, et soit $\widehat{\Sigma}$, l'estimateur usuel de la matrice de variance-covariance de x. Dans A new sliced inverse regression method for multivariate response regression, nous montrons que la matrice formée par les K vecteurs propres de la matrice $\widehat{\mathbb{B}}\widehat{\mathbb{B}}'\widehat{\Sigma}$ correspondant aux K plus grandes valeurs propres converge en probabilité vers une matrice qui engendre l'espace EDR. Cet article constitue la Section 4.3 et illustre également le bon comportement de cette approche en simulation ainsi qu'avec deux jeux de données réelles.

Même lorsque l'on étudie les expressions d'un seul gène, la méthode SIR classique peut ne pas convenir à des jeux de données de trop grande taille. En effet, elle nécessite que le nombre p de variables explicatives à prendre en compte soit inférieur à la taille n de l'échantillon observé. Si cette condition n'est pas vérifiée, nous sommes alors dans un cas sous-déterminé. Le problème est que les estimateurs de B fournis par SIR sont généralement obtenus par une décomposition en éléments propres de la matrice $\widehat{\Sigma}^{-1}\widehat{M}$, où $\widehat{M}$ est une matrice $p \times p$ calculée à partir des données et définie dans la Section 4.4. Or lorsque $n < p$, $\widehat{\Sigma}$ n'est pas inversible.

Plusieurs approches ont été développées pour modifier la méthode SIR afin qu'elle puisse fonctionner dans ce cas. Par exemple, la méthode RSIR (pour *Regularized SIR*) de Zhong *et al.* (2005) est basée sur le remplacement de $\widehat{\Sigma}$ par $\widetilde{\Sigma}(s) = \widehat{\Sigma} + sI_p$, où I_p est la matrice identité de taille $p \times p$ et s est un réel positif. Le paramètre s optimal est déterminé par une technique de bootstrap. Nous proposons une méthode alternative appelée SIR-QZ qui choisit s comme étant le plus petit réel positif possible tel que l'algorithme QZ, qui calcule les éléments propres généralisés de $\widehat{M}$ et $\widetilde{\Sigma}(s)$, renvoie un résultat dépourvu d'erreur numérique. Notez que la méthode SIR classique dépend d'un paramètre H : le nombre de tranches. Une autre fonctionnalité de SIR-QZ est de se servir de plusieurs estimations de $x'B$ pour des valeurs de H différentes, afin d'obtenir une estimation qui dépend moins du choix de ce paramètre. Dans Comparison of sliced inverse regression approaches for underdetermined cases, RSIR et SIR-QZ sont comparés en simulation entre elles ainsi qu'avec deux autres approches. SIR-QZ y est également appliqué sur des données transcriptomiques. Cet article constitue la Section 4.4.

Chapter 2

General introduction

This thesis is made of modelling examples in biology, based on existing or original statistical tools and on datasets coming mainly from the EPOC laboratory (Environnements et Paléoenvironnements Océaniques et Continentaux) of the University of Bordeaux. After two introductory chapters, respectively in French and in English, Chapter 3 presents works about measurements of the activity of oysters. In Chapter 4, we describe statistical methods well-suited for transcriptomic data. We conclude this work in Chapter 5 where prospective research projects are also developed.

2.1 A stochastic model for bioindicators

In Chapter 3, we try to sum up the information which is contained in large samples. The signals we considered are indeed composed of high-frequency measures of the distance between both parts of the shell of an oyster. We call them valvometric signals. These data are for instance studied in order to give a diagnosis about the health of the animal. When several oysters from a single location are in poor health, changes in their environment are a plausible explanation. Finding a model to explain valvometric data is thus an important step toward the creation of efficient biosensors based on oysters to monitor the quality of water. We present at length the biological context related to these valvometric measurements in the introduction of Chapter 3 and in Sections 3.1.1-3.1.2.

As shown in Figure 2.1, two states can be observed in a valvometric signal. We name these states "open" and "closed". When an oyster is in good health, biologists observe that the transitions from a state to another are linked with the tide. This is not generally the case when the animal does not adapt well to its environment. Using a renewal process which can visit the states "open" and "closed" and an estimate of its survival function, we show in A hidden renewal model for monitoring aquatic systems biosensors (see the list of works) how we can perform a classification on some animals which only relies on their valvometric signal. This classification appears to be consistent with some information we have about their health. This article forms Section 3.1.

The renewal process is piecewise constant and the observed signal (Figure 2.1) is continuous. This signal also exhibit areas where it varies fast toward low values and quickly returns to a high level. These spikes are called microclosings and their minimum can sometimes be so low that it is difficult to separate microclosings from transitions from the open state to the closed state. In Section 3.1, we describe a method to estimate the

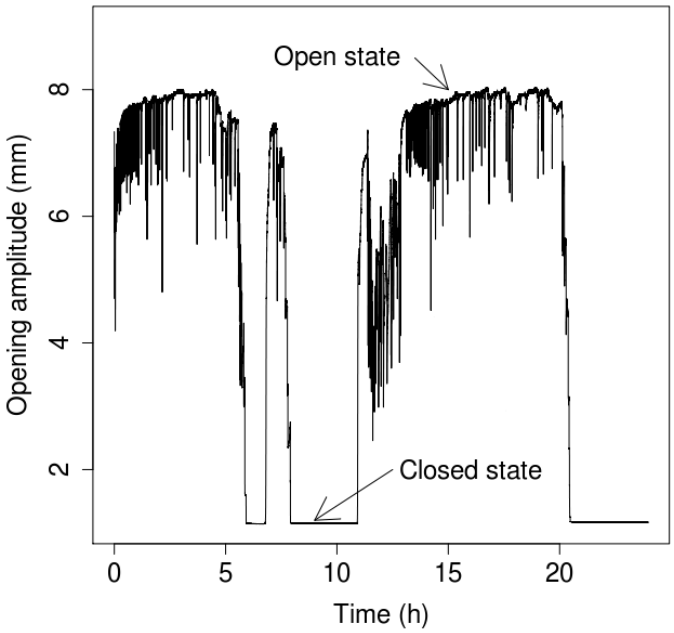


Figure 2.1: An example of valvometric data, for one oyster at Eyrac pier on the 27th of February 2007

renewal process which is hidden in the valvometric signal. This procedure allows us to determine a threshold. When the valvometric signal is below this threshold, we decide that the underlying process is in the closed state, otherwise it is in the open state. To find the threshold we need, we estimate a probability density function related to the valvometric signal. We assume that this density has two modes: one of them corresponds to the open state and the other one is linked to the closed state. We choose a density estimator which produces an estimate with two modes too. The location of the unique antimode of the estimate is then chosen to be the threshold.

Several methods were created to estimate a probability density function f when we have independent observations from the probability distribution defined by f at our disposal, and when we know how many modes f has. For example, we can apply a kernel density estimator or an estimator based on the level sets of f on these data. In Comparison of kernel density estimators with assumption on number of modes, we studied these estimators and another kernel density estimator which does not need any hypothesis about the number of modes of f, in a simulation study. This article forms Section 3.2.

The number of microclosings and their magnitude can also provide information about the health of an animal. It is however difficult to detect them because the valvometric signal can contain noise. In Section 3.3, we present an existing denoising method based on Haar wavelets. It was chosen among many others thanks to the simulation study from Antoniadis *et al.* (2001). We also studied in this section how to locate microclosings directly from the signal with noise.

To make these approaches easier to use, an implementation was realized with R and a user interface was created using the GTK+ library. These programs include additional

features such as the ability to download valvometric data from a remote SQL server where they are stored as it is explained in Section 3.4. Some ideas are also presented in this section to optimize computational time when working with R.

2.2 Statistical methods to analyse genetic data

The biological problems we tackle in Chapter 4 concern the links between transcriptomic data and some other variables. The latter ones are named explanatory variables. Transcriptomic data are measures of how much genetic markers are expressed. We would like to explain how the explanatory variables make these expressions change. This is a regression context and the large size of the datasets is characterized either by a great number of variables to explain, which means a lot of genes, or by many explanatory variables. We provide two examples of such problems in Section 4.1.

We can link the changes in the expression of a single gene with the explanatory variables using a linear regression. By testing which parameters are significantly non null, we can deduce which explanatory variables have an effect on the gene expression. When we deal with expressions of several genes, we are facing multiple tests: Even if we only have one explanatory variable, we have to test one parameter for nullity, for each gene.

Instead of controlling the probability to obtain a false positive, we could prefer to control the expected value of the ratio between the number of false positives and the number of positive tests, also known as False Discovery Rate or FDR. If each test which is realized for a given gene is independent from the tests related to every other genes, we are able to produce a series of tests for nullity based on the FDR.

For our transcriptomic data, this assumption is not true. We then have to introduce a more complex model, in order to withdraw the structure of these data and then produce independent tests. This approach was developed by Friguet *et al.* (2009) and is called Factor Analysis for Multiple Testing or FAMT. We describe it in Section 4.2 where we also give details of the user-friendly tool we implement for this method.

FAMT can only detect linear links between gene expressions and explanatory variables. The Sliced Inverse Regression (SIR) method can however find non-linear relations. It was introduced by Li (1991), and aims at estimating B in the following univariate model:

$$y = f(x'B, \varepsilon),$$

where f can be any function from $\mathbb{R}^{K+1}$ to $\mathbb{R}$, x is a p-dimensional explanatory variable, y and ε are univariate random variables, x and ε are independent and B is a $p \times K$ matrix with full rank and which satisfies $K \leq p$. As it is, this model can not handle expressions of several genes. Let $y^{(1)}, \ldots, y^{(q)}$ be the expressions of q genes. We can replace the previous univariate model by the following one

$$\begin{cases} y^{(1)} = f_1(x'B, \varepsilon^{(1)}), \\ \qquad \vdots \\ y^{(q)} = f_q(x'B, \varepsilon^{(q)}), \end{cases} \tag{2.1}$$

where $f_1, \ldots, f_q$ can be any function from $\mathbb{R}^{K+1}$ to $\mathbb{R}$ and $\varepsilon^{(1)}, \ldots, \varepsilon^{(q)}$ are univariate random variables which are independent from x. The vectorial space spanned by B is

called EDR space. For all $j = 1, \ldots, q$, we are then able to obtain an estimate $\widehat{B}_j$ of B, using the SIR method and the j^{th} equation of model (2.1). Let $\widehat{\mathbb{B}} = \left[\widehat{B}_1, \ldots, \widehat{B}_q\right]$, and let $\widehat{\Sigma}$ be the usual estimator of the variance-covariance matrix of x. In A new sliced inverse regression method for multivariate response regression, we show that the matrix made of the K eigenvectors of the matrix $\widehat{\mathbb{B}}\widehat{\mathbb{B}}'\widehat{\Sigma}$ which correspond to the greatest K eigenvalues converges in probability toward a matrix which spans the EDR space. This article forms Section 4.3 and also illustrate the good behaviour of this approach with analyses on simulated data and on two real datasets.

Even when we study expressions from a single gene, the usual SIR method may not suit too large datasets. More precisely, it needs that the number p of explanatory variables to take into account is less than the sample size n. If this requirement is not met, we have to deal with an underdetermined case. The issue is that estimates of B are generally provided by an eigendecomposition of the matrix $\widehat{\Sigma}^{-1}\widehat{M}$, where $\widehat{M}$ is a $p \times p$ matrix which is computed from the data and defined in Section 4.4. When $n < p$, $\widehat{\Sigma}$ is however not invertible.

Several improvements of the SIR method where then developed to make it work in this case. For example, the RSIR method (for Regularized SIR) from Zhong *et al.* (2005) replaces $\widehat{\Sigma}$ by $\widetilde{\Sigma}(s) = \widehat{\Sigma} + sI_p$, where I_p is the $p \times p$ identity matrix and s is a positive real. The optimal value of s is determined with a bootstrap technique. We also propose an alternative approach, called SIR-QZ. It is based on the QZ algorithm which computes generalized eigenelements of $\widehat{M}$ and $\widetilde{\Sigma}(s)$. It chooses s as small as possible such that the QZ algorithm returns a result without any numerical error. Note that the usual SIR method relies on a parameter H: the number of slices. Another feature of SIR-QZ is to take advantage of several estimates of $x'B$ for various values of H in order to produce a final estimate which is less sensitive to the choice of H. In Comparison of sliced inverse regression approaches for underdetermined cases, RSIR and SIR-QZ are compared with each other and with two other methods in a simulation study. SIR-QZ is also applied to some transcriptomic data. This article forms Section 4.4.

Chapter 3

A stochastic model for bioindicators

We propose in this chapter statistical methods dedicated to the transformation of a crude signal from an animal into a conclusion about its environment. Here, the environment is an aquatic one, the animals are bivalves and the measurement technique is the valvometry. A bivalve is a mollusc with a shell made of two articulated parts named valves. Given a location on each of these valves, a valvometric signal is then a sequence of measurements of the distance between them. The MolluSCAN Eye team currently composed of the engineers and researchers Pierre Ciret, Gilles Durrieu, Jean-Charles Massabuau, Mohamedou Sow and Damien Tran, studied such signals since 1997. This allows them to obtain insights about bivalve's physiology and aquatic toxicology. This team is supported by the EPOC laboratory (UMR CNRS 5805).

While in the following sections we will only focus on the measures they took from oysters, they have installed several valvometers around the world with various species. The signals we will consider come from Locmariaquer, Arcachon and Santander. The first two places are in France. The last one is in Spain. Another device was disposed in New Caledonia and worked from 2007 to 2009. It contained giant clams. Scallops were placed in the Svalbard archipelago and in Russia, where a mussel-based valvometer is also present. Valvometers rely on electrodes positioned on each valve of the animal as in Figure 3.1. More technical details about these devices can be found in Section 3.1.2 or at `http://molluscan-eye.epoc.u-bordeaux1.fr/`.

In Section 3.1, we present a suitable model for the valvometric data in order to diagnose the health status of the animals. This model relies on a kernel density estimator and on an assumption about the number of modes of a probability density function. We compare this kind of estimator with other ones in Section 3.2 in order to find a precise estimation of the location of the lowest local minimum of the probability density function.

Sections 3.1-3.2 aim at finding an overall diagnosis about the health status of oysters. Other interesting works concern more precise indicators. For example, Schwartzmann *et al.* (2011) observe the growth of bivalves using valvometers and another method called sclerochronology. Galtsoff (1938) studied spawning movements with an early type of valvometer. Scientists from EPOC witnessed the same behaviour with their installations. Spikes with a great amplitude are interesting features because bivalves only exhibit a lot of them in particular situations such as spawning or illness. They are for instance considered in Schmitt *et al.* (2011) and in Section 3.3. In this section, we also tackle the issue of a possible noise in the signal. The analyses that lead to the model of Sections 3.1-3.2 needed a programming effort which is explained in Section 3.4.

Figure 3.1: Two electrodes stuck on an oyster. Picture taken by the MolluSCAN Eye team.

3.1 A hidden renewal model for monitoring aquatic systems biosensors

This section is an article under revision and was written with Romain Azaïs and Gilles Durrieu. It was submitted in *Environmetrics*.

3.1.1 Introduction

Protection of the aquatic environment is a top priority for marine managers, policy makers, and the general public. Due to an increasing interest in the health of aquatic systems, there is a compelling need for the use of remote online sensors to instantly and widely distribute information on a daily basis. Among these sensors, bio-indicators are increasingly used and are highly effective to reveal the presence of low concentrations of contaminants through accumulation in tissues (see Tran *et al.*, 2003, 2004, 2007). The ability of mollusk bivalves to adapt with the environment is one of the possible ways to assess water quality. Monitoring their opening and closing activities over time is yet another method to evaluate the behavior of the bivalves in reaction to their environmental exposure.

The interest in investigating the bivalve's activities by recording their valve movements (valvometry) has been explored in ecotoxicology for more than 20 years. The basic idea of valvometry is to use the bivalve's ability to close its shell when exposed to a contaminant as an alarm signal (Sow *et al.*, 2011; Nagai *et al.*, 2006 among others). Nowadays, valvometers are available on the market and use the principle of electromagnetic induction (Sloff *et al.*, 1983; Jenner *et al.*, 1989) such as the Mossel Monitor (Kramer *et al.*, 1989) or the Dreissena Monitor (Borcherding and Volpers, 1994). There has been a clear research interest in the recent years to measuring the bivalve's behaviors directly in real conditions (Robson *et al.*, 2007; Tran *et al.*, 2003; Sow *et al.*, 2011).

These noninvasive valvometric techniques produce high-frequency data and different statistical models were built to analyze them (Sow *et al.*, 2011; Schmitt *et al.*, 2011 and Jou and Liao, 2006). Here we propose a statistical procedure based on a four-state stochastic process to give inferences about oysters' health and provide some arguments about the healthiness of their environment. This method also exhibits a link between the tide and oysters' behavior, as shown in Sow *et al.* (2011).

The paper is organized as follows. In Section 3.1.2, we give a detailed presentation of the real data problem. Briefly, we consider valvometric data samples collected by the laboratory Environnements et Paléoenvironnements Océaniques et Continentaux (EPOC, `http://www.epoc.u-bordeaux.fr/)`. In Section 3.1.3, we describe the state either open, or close, or in transition from open to close, or close to open, of an oyster by a four-state renewal model. For this class of continuous-time stochastic processes, the path is piecewise-constant over time and changes at random times. In our particular framework, the motion is characterized by the conditional jump rate λ or by its integrate version Λ, called the cumulative jump rate. We provide a method to estimate Λ from a single trajectory of the renewal process. However, this process is not directly observed, but instead we have a signal that relies on it. Section 3.1.4 is devoted to extract from this signal the period of time when the oyster is in each state of the process. For this purpose, we estimate a density related to the distance between both parts of its shell, each day. This density is reasonably assumed to have two modes, the first one corresponds to a

closed state of the shell and the second one to an open state. It leads to a threshold that enables us to approximate the time when oysters are in closed or open state. Thus, Section 3.1.3, together with Section 3.1.4, provides estimates of the survival functions related to both states. In Section 3.1.5, we apply the results of previous sections to biosensors for environmental monitoring. We especially exhibit links between groups of oysters and features related to the survey of the environment in different experimental sites, such as temperature variations and knowledge about their imminent death. Concluding remarks are given in Section 3.1.6.

3.1.2 Biological data

We first describe the experimental site and the animal species. Then we give some details on evaluation of valve activity. Afterwards, we provide information on data collection and transmission. Finally, exploring typical features of these environmental data, we explain which inferences are valuable from the biological point of view.

3.1.2.1 Data acquisition

The three monitoring sites we considered are respectively located in France, either in the Bay of Arcachon at the Eyrac pier (Latitude: 44°40 N, Longitude: 1°10 W), or at Locmariaquer (Latitude: 47°57 N, Longitude: 2°94 W) or in Spain, at Santander (Latitude: 43°44 N, Longitude: 3°79 W). A group of sixteen Pacific oysters, *Crassostrea gigas*, measuring from 8 to 10 cm length, are installed on each site. Every oyster has almost the same age (1.5 years old). They were placed in a traditional oyster farmer bag.

The electronic principle of monitoring was described by Tran *et al.* (2003) and further modified by Chambon *et al.* (2007) to handle difficult environmental conditions such as strong water current conditions and even storms. Some information about these specific aspects can be found on `http://molluscan-eye.epoc.u-bordeaux1.fr`. The main challenge was to ensure the complete autonomy of the system without in-situ human intervention for at least one full year. In brief, each animal is equipped with two light coils (sensors), of approximately 53 mg each (unembedded), fixed on the edge of each valve. These coils measure 2.5×2.5×2 mm and were coated with a resin sealing before fixation on the valves. One of the coils emits a high-frequency, sinusoidal signal which is received by the other coil. In each group of sixteen animals, one measurement is received every 0.1s (10 Hz). This means that each animal's behavior is measured every 1.6s. Every day, a data set with 864,000 pairs of values (1 distance value, 1 stamped time value) is generated. A first electronic card manages the electrodes and is in a waterproof case next to the animals. A second electronic card handling the data acquisition and the programmed emission is also in the field but outside the water (on a pier, a buoy, or in a light house). This unit is equipped with a GSM/GPRS modem and uses Linux operating system for driving the first control unit, managing the data storage, accessing the Internet, and transferring the data.

A self-developed software module runs on mobile phone technology. After each 24h period (or any other programmed period of time), the data collected are transmitted to a remote central workstation server where they are stored in text files. Every day, files from each site are inserted in a SQL database. This database is accessible with the software R

Figure 3.2: Example of trajectory of the four-state renewal process for modeling oyster's behavior

(R Development Core Team, 2012) or a text terminal, via Internet or directly from the storage server.

3.1.2.2 The biological issue

These measurements produce some characteristic features that can be examined in Figure 2.1. Oysters generally follow a two-stated pattern of behavior. The first state occurs when the distance between the parts of their shell is minimal and roughly constant, and the other one when this distance is bigger, varies more and is scattered with fast and short changes called microclosings. To date, link between tide and state of the oysters was shown (Sow *et al.*, 2011; Tran *et al.*, 2011) in the sense that they are in open state when the tide is high and in closed state when it is low. However, oysters can close its shell for other reasons such as defense against an inhospitable environment. For example, pollution can affect oysters (see for instance Tran *et al.*, 2003, 2010 and Sow *et al.*, 2011). Thus, detecting these changes of state can provide insights about the health of oysters. Because bivalves that we focus on are penned in where tide is semidiurnal, it makes sense to obtain at least one closed state and one open state every day like in Figure 2.1.

3.1.3 Four-state renewal model

In this section, we propose to model the behavior of oysters over time by a marked continuous-time process with four states. Indeed as shown in Figure 3.2, at each time one may consider that an oyster is either open, or close, or in a transition state from open to close or close to open. Among these four states, the open one and the closed one are of particular interest, given the remarks of Section 3.1.2.2. We thus consider the state space given by $E = \{O, T_{OC}, T_{CO}, C\}$, where O stands for open, while C stands for close and T_{OC} (respectively T_{CO}) models the transition state from open to close (respectively from close to open). Finally, at each time t, $X_t \in E$ models the state of an oyster.

The process (X_t) is assumed to be a marked renewal process on E. This is a class of piecewise-constant stochastic models which change their location at random times, called jump times. Here, (T_k) denotes the sequence of the jump times of (X_t), where $T_0 < 0$ is

unknown. For renewal processes, one often considers the inter-jumping times S_k's given, for any $k \geq 1$, by $S_k = T_k - T_{k-1}$. As a consequence, the first inter-jumping time S_1 is unknown. One may also take into consideration the sequence (Z_k) of the locations of (X_t). For any integer k, $Z_k = X_{T_k}$.

The relation between the continuous-time process (X_t) and the discrete-time one (Z_k, T_k) is given by

$$X_t = Z_k \quad \text{for} \quad T_k \leq t < T_{k+1}.$$

As a consequence, the discrete-time process (Z_k, T_k), and equivalently the Markov chain (Z_k, S_k), contains all the information of (X_t). In our particular four-state case, the motion of the process (X_t) depends only on its conditional jump rate, which determines the distribution of the S_k's. The dynamic of (Z_k, S_k) is defined, for all $k \geq 0$, by

- if $Z_k = O$, $Z_{k+1} = T_{OC}$,
- if $Z_k = T_{OC}$, $Z_{k+1} = C$,
- if $Z_k = C$, $Z_{k+1} = T_{CO}$,
- and if $Z_k = T_{CO}$, $Z_{k+1} = O$,

and for all $k \geq 1$ and $t \geq 0$,

$$\mathbf{P}(S_{k+1} > t | Z_k, \ldots, Z_0, S_k, \ldots, S_1) = \mathbf{P}(S_{k+1} > t | Z_k) = \exp\left(-\int_0^t \lambda(Z_k, s)ds\right).$$

An example of a path of (X_t) is given in Figure 3.2. The function $\lambda : E \times \mathbf{R}_+ \to \mathbf{R}_+$ is called the (conditional) jump rate of the process (X_t). One may consider also the cumulative jump rate $\Lambda : E \times \mathbf{R}_+ \to \mathbf{R}_+$ which is the integrate version of λ,

$$\forall (x,t) \in E \times \mathbf{R}_+, \ \Lambda(x,t) = \int_0^t \lambda(x,s)ds.$$

Each of both these functions fully characterizes the conditional distribution of the inter-jumping times.

Here, we investigate the nonparametric estimation of the cumulative jump rate for an oyster from the observation of its behavior. We consider a nonparametric framework because no prior information about the time spent by an oyster in a given state allows us to choose a parametric family of distributions. We shall obtain for each considered oyster a function which completely characterizes the distribution of the period spent in the state open or in the state close. In the sequel, m denotes the number of observed jumps. The observation of the continuous-time process within one week corresponds with the observation of about eighty jumps.

For any $t \geq 0$, we propose to estimate the cumulative jump rate $\Lambda(x,t)$, $x \in E$, by the Nelson-Aalen estimator $\widehat{\Lambda}_m(x,t)$ (see Andersen *et al.*, 1993 and the references therein) given by

$$\widehat{\Lambda}_m(x,t) = \sum_{k=1}^{m-1} R_m(x, S_{k+1}) \mathbf{1}_{\{Z_k = x\}} \mathbf{1}_{\{S_{k+1} \leq t\}}, \tag{3.1}$$

where $\mathbf{1}_A$ takes the values 1 and 0 according to whether the condition A is satisfied or not and

$$R_m(x,t) = \begin{cases} \frac{1}{L_m(x,t)} & \text{if } L_m(x,t) > 0, \\ 0 & \text{otherwise,} \end{cases}$$

with

$$L_m(x,t) = \sum_{k=1}^{m-1} \mathbf{1}_{\{Z_k = x\}} \mathbf{1}_{\{S_{k+1} \geq t\}}.$$

We exclude the first inter-jumping time S_1 from the estimation procedure since it is unknown.

The Nelson-Aalen estimator is well defined in the framework of the multiplicative intensity model, introduced by Aalen in the seventies. Fortunately, Azaïs *et al.* have shown that this model is well suited for the particular case of marked renewal processes with discrete state space. In this context, we establish in Proposition 1 the consistency of the estimator under some additional conditions imposed on the asymptotic behavior of R_m and L_m when m goes to infinity. We also prove in Lemma 2 that these conditions are verified under a reasonable assumption on λ.

Proposition 1. *Let $(x,t) \in E \times \mathbf{R}_+$. Assume that the conditions*

$$\forall \varepsilon > 0, \ \lim_{m \to \infty} \mathbf{P}\left(\int_0^t R_m(x,s)\lambda(x,s)ds > \varepsilon \right) = 0, \tag{3.2}$$

and

$$\forall \varepsilon > 0, \ \lim_{m \to \infty} \mathbf{P}\left(\int_0^t \mathbf{1}_{\{L_m(x,s)=0\}}\lambda(x,s)ds > \varepsilon \right) = 0, \tag{3.3}$$

are verified. Then, we have

$$\forall \varepsilon > 0, \ \lim_{m \to \infty} \mathbf{P}\left(\sup_{0 \leq s \leq t} \left| \widehat{\Lambda}_m(x,s) - \Lambda(x,s) \right| > \varepsilon \right) = 0. \tag{3.4}$$

Proof. Azaïs *et al.* have shown that the multiplicative intensity model makes sense for the observation of the sequence (Z_k, S_k). The proof of (3.4) under the conditions (3.2) and (3.3) can be found in Andersen *et al.* (1993), Theorem IV.1.1, when the multiplicative intensity assumption is verified. □

Lemma 2. *Let $x \in E$. Assume that $\lambda(x,\cdot)$ is a locally integrable function. Then, for any $t \geq 0$, the conditions (3.2) and (3.3) are verified.*

Proof. Given in Appendix A.3.1. □

When the process (X_t) is hidden, one may only obtain some approximations $\widehat{S}_k$'s of the inter-jumping times S_k's. For every $k \in \{2, \ldots, m\}$, let $\widehat{S}_k \in [S_k, S_k + \zeta]$, for a small constant $\zeta > 0$. We do not compute the Nelson-Aalen estimator $\widehat{\Lambda}_m(x,t)$ but an approximation of this estimator $\widetilde{\Lambda}_m(x,t)$ from the $\widehat{S}_k$'s. Since, for all the integers k,

$$R_m(x, S_{k+1}) = R_m\left(x, \widehat{S}_{k+1}\right),$$

we have

$$\forall x \in \{O, C\}, \ \forall t \in \mathbf{R}_+ \setminus \bigcup_k \left[\widehat{S}_k - \zeta, \widehat{S}_k \right[, \ \widehat{\Lambda}_m(x,t) = \widetilde{\Lambda}_m(x,t).$$

Thus, the error between the Nelson-Aalen estimator and its approximation is equal to 0 for most $t \geq 0$.

Besides the cumulative jump rate, one can also estimate the conditional survival functions H associated with Λ. Indeed, these functions are easier to classify than the Nelson-Aalen estimators because they take values between 0 and 1, whereas the range of values taken by $\widetilde{\Lambda}_m$ depends on m. The Fleming-Harrington estimator (Fleming and Harrington, 1984) of H can easily be computed from our modified Nelson-Aalen estimator. For any $x \in E$, for any $t \geq 0$, it is given by

$$\widetilde{H}_m(x,t) = \exp\left(-\widetilde{\Lambda}_m(x,t)\right).$$

By considering the process $t \mapsto \mathbf{1}_{\{X_t = T_{OC}\}}$, one may also estimate the survival function of the time of the first return in the state T_{OC}. Let $\mathbf{J}_m$ be the subset of $\{2, \ldots, m-4\}$ such that for any $k \in \mathbf{J}_m$, $Z_k = C$. Using, for each $k \in \mathbf{J}_m$, an approximation $\widetilde{S}_{k+1}$ of $S_{k+1} + S_{k+2} + S_{k+3}$, one may simply compute an approximation of the Fleming-Harrington estimator of the survival function of the time of the first return in the state T_{OC}. This estimator is denoted $\widetilde{H}_m^{T_{OC}}$.

3.1.4 Estimation procedure

In the application, since we do not observe the process (X_t), the estimator $\widehat{\Lambda}_m(x,t)$ of $\Lambda(x,t)$ can not be computed. We assume that we observe instead a stochastic process (G_t) that looks like the signal in Figure 2.1. This process should behave like (X_t) by taking low values when $X_t = C$, high values when $X_t = O$ and intermediate values otherwise. The aim of this section is to use (G_t) to construct estimators $\widehat{S}_k$ of the observations of S_k for some appropriate values of k. We would then be able to compute $\widetilde{\Lambda}_m$, $\widetilde{H}_m$ and $\widetilde{H}_m^{T_{OC}}$.

In Section 3.1.4.1, we link (G_t) with a probability density function f that is designed to summarize the information contained in (G_t). We especially assume that f has a given number of modes. In Section 3.1.4.2, we propose a modification of kernel-based estimators of f for equally spaced sample.

3.1.4.1 A probability density for oysters' openings

For a fixed time interval and a fixed oyster, consider that (G_t) is the process that returns the oyster's opening amplitude for a given instant. The process (G_t) is studied for $t \in [0, T_{\max}]$ where $T_{\max}$ is a positive real. It takes values in an interval $[a,d] \subset \mathbf{R}_+$. For all ω in the probability space Ω, let G be a function from $[0, T_{\max}]$ to $[a,d]$ defined by $G : t \mapsto G(t) = G_t(\omega)$. The assumptions about G made in this article are thus made for all $\omega \in \Omega$. We assume that G is continuously differentiable except for a countable set of points on which G is only continous. For all $i \in \{1, \ldots, n\}$, let V_i be a random variable that follows the distribution $\mathcal{U}_{[0,T_{\max}]}$ and Y_i the random variable defined by $Y_i = G(V_i)$. Let f denote the continuous probability density of each Y_i given G. Such a density can be also defined as the occupation density of G on $[0, T_{\max}]$. See Geman and Horowitz (1980, § 2) and Bosq and Blanke (2008, Section 9.2) for details.

The following assumptions allow us to link f and $G(t)$ for t such that $X_t \in \{T_{OC}, T_{CO}\}$.

Assumptions 3.

(i) It exists a couple (b, c)*, with* $a < b < c < d$*, such that,*

$$\forall t \in [0, T_{\max}[,\ X_t = C \Rightarrow G(t) < b \text{ and } X_t = O \Rightarrow G(t) > c.$$

(ii) $\inf\{|G'(t)| :\ G(t) \in [b, c]\} > 0$ *and* G' *is continuous on* $\{t\ :\ G(t) \in [b, c]\}$.

Assumptions 3 are reasonable with respect to the signal observed in Figure 2.1. The first one expresses the fact that we are looking for a threshold that separates out values of $G(t)$ when the oyster is open or closed. Note that for $x \in [a, d]$, $f(x)$ can be linked to the points $t \in [0, T_{\max}]$ such that $G(t) = x$. To compute easily $f(x)$ when $x \in [b, c]$ we assumed that $G(t) = x$ for only one point for each interval $[T_k, T_{k+1}]$, for values of k such that $Z_k \in \{T_{OC}, T_{CO}\}$. This leads to the second assumption. Note that because b and c can be arbitrarily close from each other, Assumptions 3 are mild. Provided theses hypotheses, the mean value theorem enables us to show that,

$$\forall p \in [b, c], f(p) = \frac{1}{T_{\max}} \sum_{k \in \mathbf{I}_m(p)} \frac{1}{|G'(u_k)|}, \tag{3.5}$$

where $m = \max\{k\ :\ T_k < T_{\max}\}$ and $\mathbf{I}_m(p)$ is the subset of $\{1, \ldots, m\}$ made of integers k such that there exists a unique $u_k \in [T_k, T_{k+1}[$ which satisfies $G(u_k) = p$.

For every $x \in \mathbf{R}$, if in points $t \in [0, T_{\max}]$ where $G(t) = x$, the values of $|G'|$ are overall great, $f(t)$ is low and conversely. Let $N(f)$ be the number of modes of the density f. We postulate that the values of $|G'|$ are lower when the oyster is open or closed than when it is in a transition state. We thus assume that $N(f) = 2$ which is reasonable for a period of time $[0, T_{\max}]$ throughout which the oyster visits both open and closed state. We also assume that during transition states $|G'(t)|$ is greater for t such that $G(t) \in [b, c]$ than for t such that $G(t)$ is close to b or c but not in $[b, c]$. This remarks leads to Assumptions 4.

Assumptions 4.

(i) $m \geq 3$.

(ii) $N(f) = 2$*,* f *has no flat part and has a unique antimode located in* θ.

(iii) $\forall t \in \{b, c\}$, $\forall k \in \mathbf{I}_m(t)$, $\exists \varepsilon_k > 0$, $\forall \varepsilon \in]0, \varepsilon_k[$, $\exists v_k \in \{u_k - \varepsilon, u_k + \varepsilon\}$,

$$G(v_k) \notin [b, c] \text{ and } |G'(v_k)| \in]0, |G'(u_k)|[\,,$$

where u_k *is the unique point in* $[T_k, T_{k+1}[$ *such that* $G(u_k) = t$.

Note that it is easy to take $T_{\max}$ great enough so that the first point of Assumptions 4 is satisfied. Given Assumptions 3 and 4 and equation (3.5), we then have

$$\mathbf{P}\left(\theta \in [b, c]\right) = 1.$$

In this section, we showed that we can link the stochastic function G that describes an oyster's opening during a given time interval and a probability density f. In Section 3.1.4.2, we propose some estimators for f and θ. The estimate of θ will then be taken as a threshold and the instants when G crosses it will lead to approximations of the S_k's.

3.1.4.2 Density estimation procedure

For $t \in \mathbf{R}$, we use the following kernel density estimator $\widetilde{f}_n(t)$ of $f(t)$

$$\widetilde{f}_n(t) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{t - G\left(i\frac{T_{\max}}{n}\right)}{h_n} \right), \tag{3.6}$$

where K is the Gaussian kernel defined, for every $t \in \mathbf{R}$, by $K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$ and h_n is a positive real parameter called the bandwidth. Some pointwise properties about kernel estimators for occupation densities are available in Wang and Phillips (2009).

The estimator $\widetilde{f}_n$ allows us to estimate θ under the following assumptions on the bandwidth h_n and G.

Assumptions 5.

(i) *f is uniformly continuous.*

(ii) *$h_n \to 0$ and $nh_n^2 \to \infty$ as n goes to infinity.*

(iii) *G' is bounded.*

(iv) *$\widetilde{f}_n$ has a unique antimode located in $\widetilde{\theta}_n$.*

Under the second point of Assumptions 4 and Assumptions 5, we have

$$\mathbf{P}\left(\lim_{n\to\infty} \widetilde{\theta}_n = \theta \right) = 1.$$

Note that because G is measured from a living being, the third point of Assumptions 5 is reasonably satisfied.

To ensure the convergence of $\widetilde{\theta}_n$ toward θ, the bandwidth h_n remains to be chosen. To satisfy the last point of Assumptions 5, mimicking Silverman (1981), we choose $h_n = \widetilde{h}_{crit}$, which is defined as

$$\widetilde{h}_{crit} = \min\{h \ : N(\widetilde{f}_n) \leq 2\}.$$

When K is the Gaussian kernel, Silverman (1981) showed that $\widetilde{h}_{crit}$ can be easily computed.

For any $t \in \mathbf{R}$, define $\widehat{f}_n$ as

$$\widehat{f}_n(t) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{t - Y_i}{h_n} \right). \tag{3.7}$$

The second point of Assumptions 5 for $\widetilde{h}_{crit}$ is convincing because if K is the Gaussian kernel, we have the following points.

- Let $\widehat{h}_{crit}$ be defined as $\widetilde{h}_{crit}$ using $\widehat{f}_n$ instead of $\widetilde{f}_n$. We have for any $\varepsilon > 0$ that $\mathbf{P}(\widehat{h}_{crit} > \varepsilon) \to 0$ and that $\mathbf{P}(n\widehat{h}_{crit}^2 > \varepsilon) \to 1$ when n goes to infinity (see Mammen *et al.*, 1991 and Section 3.2).

- for such a bandwidth, $\widehat{f}_n$ converges uniformly in probability toward f (see Devroye and Wagner, 1980 and Section 3.2).

- For a given random bandwidth h_{rand} used in both $\widehat{f}_n$ and $\widetilde{f}_n$ and such that for any $\varepsilon > 0$, $\mathbf{P}(h_{rand} > \varepsilon) \to 0$ and $\mathbf{P}(nh_{rand}^2 > \varepsilon) \to 1$ when n goes to infinity, we can prove that $\widetilde{f}_n$ converges uniformly in probability toward $\mathbf{E}\left[\widehat{f}_n\right]$.

When n is large enough, we are now able to build a threshold $\widetilde{\theta}_n$ from G which is only crossed by G at points $t \in [T_k, T_{k+1}[$ such that $Z_k \in \{T_{OC}, T_{CO}\}$. In addition, asymptotically, G crosses only once $\widetilde{\theta}_n$ in each of such intervals $[T_k, T_{k+1}[$, because of the second point of Assumptions 3. For every $k \in \mathbf{I}_m(\widetilde{\theta}_n)$ let t_k be the unique point in $[T_k, T_{k+1}[$ such that $G(t_k) = \widetilde{\theta}_n$. Thus, $k \in \mathbf{I}_m(\widetilde{\theta}_n) \backslash \{\max\{\mathbf{I}_m(\widetilde{\theta}_n)\}\}$, implies that $k+2 \in \mathbf{I}_m(\widetilde{\theta}_n)$ and we have

$$t_{k+2} - t_k \in]S_{k+2}, S_{k+2} + \zeta[,$$

where $\zeta = 2\max_{k \in \mathbf{I}_m(\widetilde{\theta}_n)}\{S_{k+1}\}$. The value of ζ is small because when $k \in \mathbf{I}_m(\widetilde{\theta}_n)$, $Z_k \in \{T_{OC}, T_{CO}\}$. We then set $\widehat{S}_{k+2} = t_{k+2} - t_k$ for all $k \in \mathbf{I}_m(\widetilde{\theta}_n) \backslash \{\max\{\mathbf{I}_m(\widetilde{\theta}_n)\}\}$. Thus, for $x \in \{O, C\}$, and $t \geq 0$, we are able to build the estimators $\widetilde{\Lambda}_m(x,t)$, $\widetilde{H}_m(x,t)$ introduced in Section 3.1.3. Note that $\forall k \in \mathbf{J}_m$, $\{k-1, k+3\} \subset \mathbf{I}_m(\widetilde{\theta}_n)$ and we can define

$$\widetilde{S}_{k+1} = t_{k+3} - t_{k-1} \in]S_{k+1} + S_{k+2} + S_{k+3}, S_{k+1} + S_{k+2} + S_{k+3} + \zeta[,$$

and then compute $\widetilde{H}_m^{T_{OC}}(x,t)$.

3.1.5 Real data application

In this section, we apply results from the previous sections to oysters from different experimental sites. We give $\widetilde{H}_m(x,t)$ for $x \in \{O, C\}$ in Figure 3.3 for four oysters studied during 8 days. We computed a value of $\widetilde{\theta}_n$ for each day and each oyster and pooled the resulting values of $\widehat{S}_i$ in a single estimator $\widetilde{\Lambda}_m(x,t)$. The first two animals, corresponding to solid and dashed lines do not have reason to be in poor health. The oyster from Santander (mixed line), was in a harbor in Spain which is an environment that might not be clean enough for this kind of animal, and the last one (dotted line) died the 20th of May 2009. In Figure 3.3(b), one can observe that estimated survival functions related to the oyster from Santander and to the dying one are close from each other. In addition they can be separated from the two others. The latter functions decrease less than the previous ones for small values of t. This means that the oyster of Santander and the dying one exhibit more short opening periods than the two others. A big amount of such periods is characteristic of oysters that live in an inhospitable environment. Groups of curves of Figure 3.3(b) can also be noticed in Figure 3.3(a), but this is less obvious. Since the oyster from Santander sometimes spent more than 40 hours closed, the mixed line is far from 0 for $t = 12$.

Another feature that can arise for oysters with health problems is that they do not open or do not close themselves during a long time. In this case, we can have $m = 1$ and $N(f) = 1$ when $T_{\max} = 24$, which provides incorrect estimates $\widetilde{H}_m(x,t)$. Hopefully, because of spikes in open state (see Figure 2.1), and because of electronic noise in closed state, the inaccurate $\widetilde{\theta}_n$ leads in both cases to small values of $\widehat{S}_i$. Thus, the corresponding function $\widetilde{H}_m(x,t)$ is close to estimates for oysters that exhibit a lot of microclosings. Because we aim at classifying oysters according to their health status, this does not bother us and we use $\widetilde{H}_m(x,t)$ even if $m = 1$ and $N(f) = 1$.

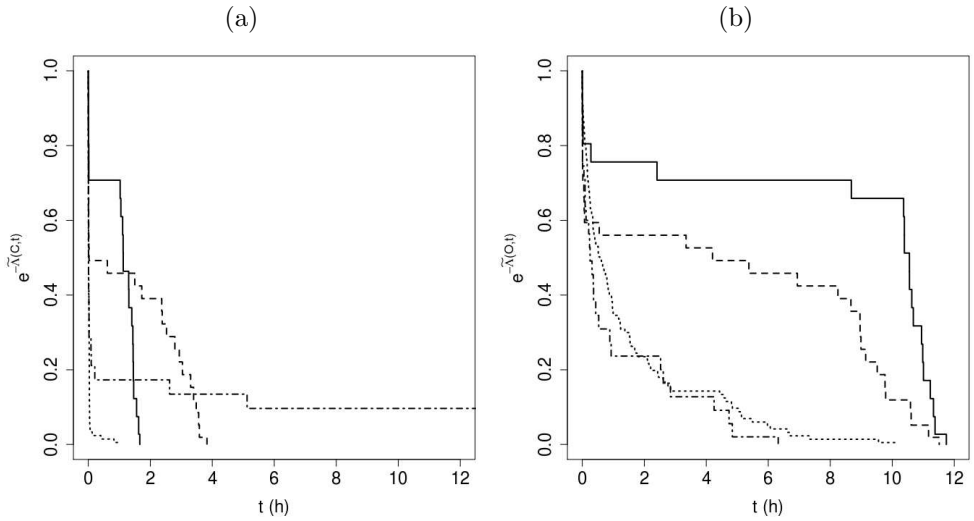


Figure 3.3: Fleming-Harrington estimates of the survival function with respect to t for four oysters: (solid line) oyster of the Eyrac pier between 1st and 8th of May 2007; (dashed line) oyster of Locmariaquer between 1st and 8th of October 2011; (dotted line) oyster of Santander between 1st and 8th of December 2011; (mixed line) oyster of Eyrac pier between 12th and 19th of May 2009. 3.3(a) Survival function estimation for $x = C$; 3.3(b) Survival function estimation for $x = O$.

Remark 1. An alternative way to handle a hidden process with a finite number of states and an observed one taking real values is the Hidden Semi-Markov Models (HSMM). O'Connell and Højsgaard (2011) implement in the R package `mhsmm` a procedure to estimate X_t for $t \in [0, T_{\max}]$ in this context. They assume that the distribution G_t depends only on X_t, which means that G_t given X_t is independent of G_s for any $s < t$. In addition, while the distributions of the S_k's given Z_k can be non-parametrically estimated, the distribution of G_t given X_t has to belong to a parametric family. For oysters' opening measured with high frequency, these assumptions are not reasonable and the method to approximate X_t is time consuming. The corresponding algorithm takes more than 4 hours to handle the signal of an oyster for a given day while finding a threshold with our method requires a few seconds.

For each animal of Figure 3.3, the estimated survival function $\widetilde{H}_m^{T_{OC}}$ of the time of the first return in T_{OC} is plotted in Figure 3.4. Curves in solid line and in dashed line decrease fast when t is around 12h and we know that in the locations where the oysters are placed, the tide has a period of approximately 12.41 hours. Hence, oysters' openings that correspond to these curves seem to be tide-driven, as they should be when the animals are not perturbed (see Sow *et al.*, 2011). This let us think that these oysters are in good health.

It can be unclear why we take $T_{\max} = 24$ to compute $\widetilde{\theta}_n$ while we use data from 8 days for each oyster. Notice that the oysters grow and that the distance between the parts of their shell can be greater when they are closed than when they were open before. That is why $N(f)$ can not be assumed to be equal to 2 for a $T_{\max}$ too large.

We studied 9 other oysters than the 4 previously considered, during 8 days. A value of $\widetilde{\theta}_n$ was computed for each oyster and for each day, which led to estimates $\widetilde{H}_m^{T_{OC}}$. Among

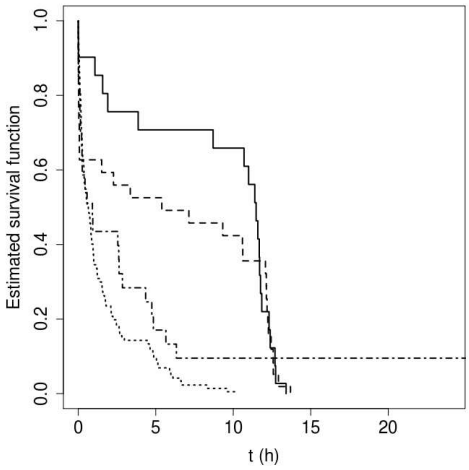


Figure 3.4: Fleming-Harrington estimates of the probability that the first return time to state T_{OC} is greater than t for four oysters: (solid line) oyster of the Eyrac pier between 1st and 8th of May 2007; (dashed line) oyster of Locmariaquer between 1st and 8th of October 2011; (dotted line) oyster of Santander between 1st and 8th of December 2011; (mixed line) oyster of Eyrac pier between 12th and 19th of May 2009.

these 9 animals, 2 of them died at most 2 weeks after the period when the data were collected, 5 of them had no known reason to be in poor health and 3 of them lived in the harbor of Santander when the measurements were taken. The last one, installed at Locmariaquer went through temperature variations that may have perturbed it. We set $N = 100$ and we measured $\widetilde{H}_m^{T_{OC}}(t)$ for $t \in \left\{0, \frac{1}{N-1}T_{\max}, \frac{2}{N-1}T_{\max}, \ldots, T_{\max}\right\}$, for each oyster. We performed a principal components analysis (PCA) on the resulting $N \times 13$ matrix. This first principal component explains a large part of the total variability of the data set (88%) and the second one explains only 9% of it. We then clustered the oysters into two groups by using a k-means algorithm on the scalar products between the values of $\widetilde{H}_m^{T_{OC}}(t)$ of every oyster and the first principal component of the PCA.

The scalar products corresponding to the first two components of the PCA are plotted in Figure 3.5. This graphic also presents the result of the k-means algorithm. In the second group, the dying oysters, those from Santander and the one that suffered temperature variations are gathered. These animals have reasons to be in poor health.

3.1.6 Concluding remarks

This paper presented a framework to analyze a renewal process with a finite number of states when this process is not directly observed. Instead, we have a signal that takes values in different intervals when the process is in different special states of interest. Theses intervals are assumed to be disjoint and the states that are not related to any interval are called transition states. The absolute value of the derivative of the signal should be large in these transition states but it is allowed to take great values in the other states too.

Focusing in a four-state model with two transition states and two states of interest, we described how such a renewal process can underly signals made of measures of the distance between the parts of the shell of oysters. We gave convergences of an estimator

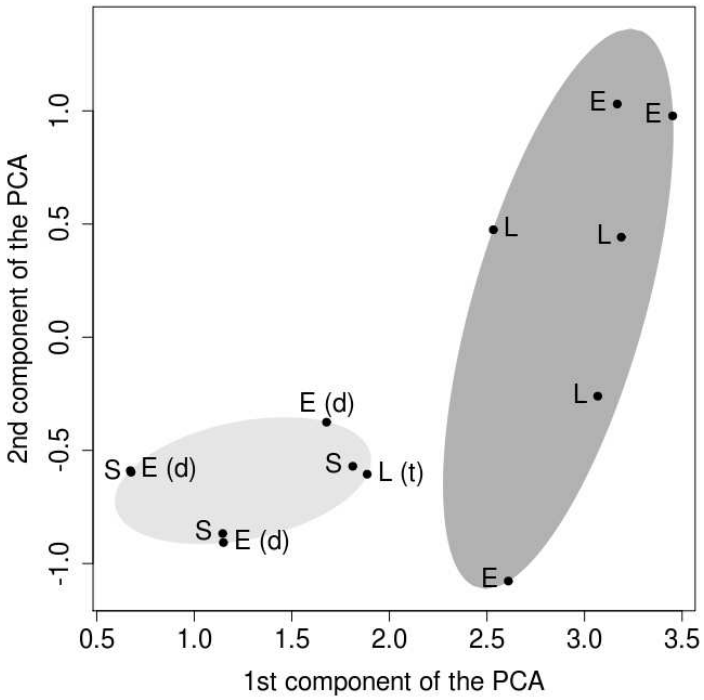


Figure 3.5: Projection of classified oysters on the first two principal component axes which display 97% of the total variability. The label of each oyster corresponds to its location: E: Eyrac, L: Locmariaquer, S: Santander, (d): Dying oysters, (t): Oyster affected by quick temperature variations.

of the cumulative jump rate of an observed process. Then, we summarized the information contained in the signals with a bimodal probability density function that is estimated by a bimodal function. The location of the antimode of this estimate provide a threshold that separate out the instants when the oyster is in each states of interest. We thus are able to estimate the cumulative jump rate for the states of interest, from a signal of oysters' openings.

Finally, for a transition state, we studied an estimator of the survival function of the first return time that appears to be different for oysters that have a tide-driven behavior than for animals that have health problems or live in an inhospitable environment. A classification procedure on these functions succeeded in separating out both kinds of oysters. Because the quality of their environment and their health status are closely related, this approach seems to us to be a new step toward the use of valvometric devices as biosensors for water quality.

Acknowledgements

This work was supported by the University of Bordeaux, the ASPEET project from the University of South Brittany, the Portonovo project and the Centre National de la Recherche Scientifique (UMR EPOC CNRS 5805, UMR CNRS 6205 and UMR CNRS 5251). We are grateful to Jean-Charles Massabuau, Pierre Ciret and Damien Tran from the EPOC laboratory, for the valvometric data and helpful discussions.

3.2 Comparison of kernel density estimators with assumption on number of modes

This section is an article which is accepted for publication in *Communications in Statistics - Simulation and Computation*. It was written with Gilles Durrieu and Jérôme Saracco.

3.2.1 Introduction

Since the seminal papers of Parzen (1962) and Rosenblatt (1956), the use of kernels to find an estimate $\hat{f}_{K,h}$ of a density function f of a random variable X is widely studied because of the advantages of the nonparametric point of view. Let $(X_1, \dots X_n)$ be a vector of independent and identically distributed random variables generated from f. For $t \in \mathbb{R}$, the kernel density estimator $\hat{f}_{K,h}(t)$ of $f(t)$ can be defined as

$$\hat{f}_{K,h}(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right) \tag{3.8}$$

where K is the kernel and h is the bandwidth on which the amount of smoothness of $\hat{f}_{K,h}$ relies. Most of the time, K is a probability density function and h is a positive real. The larger the bandwidth, the smoother the estimate is. The choice of the bandwidth h is an important area in kernel estimators research field. Even if there exists a sufficiently large interval around the optimal bandwidth where $\hat{f}_{K,h}$ stays roughly the same (Scott (1992), p. 161), h needs to be carefully determined. To perform this choice, one can use biased or unbiased cross-validation (Rudemo (1982); Scott and Terrell (1987)) as well as plug-in methods (Sheather and Jones (1991)) among other approaches.

In practical situations, the scientist that brings the data to analyze is able to determine if the estimated density function is smooth enough or not. In this paper, we are interested in using this information on the necessary amount of smoothing in order to set the corresponding bandwidth h for the estimator $\hat{f}_{K,h}$. More precisely, we will assume a fixed number $N(f)$ of modes of f. We will introduce and study the bandwidth $h_{crit,k}$ which is the smallest one such that the estimator $\hat{f}_{K,h}$ has at most k modes. Thus, the definition of $h_{crit,k}$ is:

$$h_{crit,k} = \min_{N(\hat{f}_{K,h}) \leq k} h, \tag{3.9}$$

for a given $k \in \mathbb{N}^*$. The minimum is well-defined for the different kernels we will consider. The link between h and $N(\hat{f}_{K,h})$ has been studied by several authors. With a Gaussian kernel (i.e. K is the density function of the standardized normal distribution), according to Silverman (1981), the function $h \mapsto N(\hat{f}_{K,h})$ is decreasing, which allows him and Mammen *et al.* (1991) to test the number of modes of f. For many other kernels among those with bounded support, we do not have these kind of properties, but we have at our disposal a visualization tool called the "mode tree" (see for details Minnotte and Scott (1993) or Minnotte *et al.* (1998)). Other theoretical results are also available in the literature, see for instance Hall *et al.* (2004). A method to find an estimate of f based on this kind of assumptions already exists (Polonik (1995a)).

The paper is organized as follows. In Section 3.2.2, we give asymptotic results for the density kernel estimator $\hat{f}_{K,h_{crit,k}}$ where K is the Gaussian kernel. We also present theoretical results for the uniform kernel. Define $h_{crit} = h_{crit,N(f)}$. In Section 3.2.3,

we present a simulation study which compares $\hat{f}_{K,h_{crit}}$ to another kernel estimator that is taken as a benchmark and relies on a widely-used bandwidth. The procedure from Polonik (1995a) was also considered in simulation to try to determine the better approach to include information about the number of modes of f in its estimation. Then, in Section 3.2.4, we describe constraints under which one can use $h_{crit,k}$ in the context of mixture models. We apply it to environment monitoring data in Section 3.2.5. Lastly, concluding remarks are given in Section 3.2.6.

3.2.2 Estimating a density with $N(f)$ modes

In this section we study the kernel density estimator $\hat{f}_{K,h}$ given in (3.8) with the bandwidth $h_{crit,k}$ defined in (3.9). For our purpose, we only consider two kernels:

- the uniform kernel defined for $t \in \mathbb{R}$ as $K(t) = \mathbf{1}_{[-\frac{1}{2},\frac{1}{2}]}(t)$,
- and the Gaussian kernel defined for $t \in \mathbb{R}$ as $K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$.

The first kernel has a bounded support, this is not the case for the Gaussian kernel. We also describe two alternatives of $\hat{f}_{K,h_{crit,k}}$ which are respectively $\hat{f}_{K,h_{SJ}}$, where h_{SJ} is the bandwidth given by Sheather and Jones (1991) plug-in method (see Section 3.2.2.4) and Polonik (1995a) estimator based on density contour clusters (see Section 3.2.2.5).

3.2.2.1 Assumptions on the density f of X

We need the following assumptions on the density f of X in order to have a density with $N(f)$ modes which can be properly estimated.

$(\mathcal{H}1)$ f is uniformly continuous on $\mathbb{R}$.

$(\mathcal{H}2)$ $\exists (r,s),\ -\infty < r < s < +\infty,\ f(x) \neq 0 \Rightarrow x \in [r,s]$ and $x \in]r,s[\Rightarrow f(x) \neq 0$.

$(\mathcal{H}3)$ $\exists!\ (z_1, z_2, \ldots, z_{2N(f)-1}) \in]r,s[^{2N(f)-1}, \quad \forall i \in \{1,2,\ldots,2N(f)-1\}$,

$$f^{(1)}(z_i) = 0 \text{ and } sign(f^{(2)}(z_i)) = (-1)^i,$$

where $f^{(q)}$ is the q^{th} derivative of f.

$(\mathcal{H}4)$ $f \in \mathcal{C}^2(]r,s[)$.

$(\mathcal{H}5)$ $\lim_{t \downarrow r} f^{(1)}(t) > 0$ and $\lim_{t \uparrow s} f^{(1)}(t) < 0$.

$(\mathcal{H}6)$ $\forall x \in]r,s[$ that satisfies $f^{(1)}(x) = 0,\ f^{(2)}(x) \neq 0$.

Remark 2. $(\mathcal{H}1)$ follows Devroye and Wagner (1980). The authors give an asymptotic result with the L_∞ norm, that we discuss in Section 3.2.2.3. $(\mathcal{H}2)$ - $(\mathcal{H}6)$ are taken from Mammen *et al.* (1991).

3.2.2.2 A computable bandwidth

For the Gaussian kernel, some interesting results on $h_{crit,k}$ already exist. They underline that the bandwidth $h_{crit,k}$ is easily computable. Indeed, Silverman (1981) shows that the function $h \mapsto N(\hat{f}_{K,h})$ is decreasing and right continuous. This ensures computability of $h_{crit,k}$ with the desired accuracy by a dichotomous search. With the assumption that $h_{crit,k} \in [h_1, h_2]$, and if we want to obtain it with an error less than $\frac{h_2-h_1}{2^{\check{n}}}$, we have to compute $N(\hat{f}_{K,h})$, $\check{n}$ times, for various h. If for each h, to determine $N(\hat{f}_{K,h})$, $\hat{f}_{K,h}$ is computed in $\tilde{n}$ points, then the computational complexity of the whole algorithm to find $h_{crit,k}$ is equal to $O(n\tilde{n}\check{n})$, where n is the sample size. In our simulations, we often take $\tilde{n} = 10000$ and $\check{n} \leq 30$.

For the uniform kernel, we provide a similar result by explaining that $h \mapsto N(\hat{f}_{K,h})$ is piecewise constant and has at most $\frac{n(n-1)}{2}$ jumps. Besides, if $\left\{X_{(i)}\right\}_{i\in\{1,\dots,n\}}$ is a set of ordered random variables from which we want to compute $\hat{f}_{K,h_{crit,k}}$, the locations of the jumps are equal to $X_{(j)} - X_{(i)}$ for some $i \in \{1,\dots,n-1\}$ and some $j \in \{i+1,\dots,n\}$. This means that we are able to find $h_{crit,k}$ by analyzing values of $\hat{f}_{K,h}$ between jumps.

Remark 3. The number of jumps in $h \mapsto N(\hat{f}_{K,h})$ is not bounded by $\frac{n(n-1)}{2}$ for every kernel. Indeed Hall *et al.* (2004) studied the set of points $X(\omega) = (-1, 0, 1)$ and drew $N(\hat{f}_{K_\theta,h})$, with $K_\theta(x) = C_\theta(1-x^2)^\theta \mathbf{1}_{[-1,1]}(x)$ as a function of h and θ, where C_θ ensures that $\|K_\theta\|_{L_1} = 1$. For example for $\theta = 1.5$, one can find 4 different values in $h \mapsto N(\hat{f}_{K_\theta,h})$, which is greater than $\frac{n(n-1)}{2} = 3$.

3.2.2.3 Asymptotic results on $\hat{f}_{K,h_{crit,k}}$

Proof of consistency for this estimator toward f is not trivial since $h_{crit,k}$ is data-driven. However, for the Gaussian kernel, the pointwise convergence in probability was proved by Futschik and Isogai (2006), when $k \geq N(f)$. We have furthermore the uniform convergence of the estimate toward the true density in probability and the convergence of the integrated absolute error to 0 in probability. To explain these results, we first find conditions for a given data-driven bandwidth h_n under which some asymptotic properties can be shown for $\hat{f}_{K,h_n}$. This is realized in Theorem A by combining Theorem 2 of Devroye and Wagner (1980) about the L_∞ distance between $\hat{f}_{K,h_n}$ and f and Theorem 3.3 from Devroye (1987, p. 38), concerning the L_1 distance.

Theorem A (Devroye and Wagner (1980), Devroye (1987)**).** *Let f be a probability density satisfying $(\mathcal{H}1)$, h_n a random bandwidth depending on X. If we assume the following hypotheses:*

$(\mathcal{F}1)$ *K is a Riemann integrable probability density,*

$(\mathcal{F}2)$ $\sup_{x\in\mathbb{R}} K(x) < \infty$,

$(\mathcal{F}3)$ $\int_0^\infty \sup_{|x|\geq z} K(x)dz < \infty$,

$(\mathcal{F}4)$ $\forall \varepsilon > 0,\ \mathbb{P}(h_n > \varepsilon) \to 0,$ *when* $n \to \infty$,

$(\mathcal{F}5)$ $\forall A > 0,\ \mathbb{P}(nh_n > A) \to 1,$ *when* $n \to \infty$,

then, we have, for $n \to \infty$,

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}}\left|\hat{f}_{K,h_n}(t) - f(t)\right| > \varepsilon\right) \to 0,$$

and

$$\mathbb{P}\left(\int_{\mathbb{R}}\left|\hat{f}_{K,h_n}(t) - f(t)\right| dt > \varepsilon\right) \to 0.$$

The following theorem states that Theorem A can be applied with $h_{crit,k}$, when K is the Gaussian kernel and $k \geq N(f)$.

Theorem 6. *Let* f *be a density satisfying* $(\mathcal{H}1) - (\mathcal{H}6)$ *and let* $\hat{f}_{K,h_{crit,k}}$ *be the estimator of* f *with the Gaussian kernel* K *and the bandwidth* $h_{crit,k}$ *given in (3.9), with* $k \geq N(f)$. *Then we have, for* $n \to \infty$,

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}}\left|\hat{f}_{K,h_{crit,k}}(t) - f(t)\right| > \varepsilon\right) \to 0,$$

and

$$\mathbb{P}\left(\int_{\mathbb{R}}\left|\hat{f}_{K,h_{crit,k}}(t) - f(t)\right| dt > \varepsilon\right) \to 0.$$

Proof. When K is the Gaussian kernel, $(\mathcal{F}1)$ and $(\mathcal{F}2)$ are satisfied as well as $(\mathcal{F}3)$ because K is defined and decreasing on $[0, \infty[$ and then $\int_0^\infty \sup_{|x|\geq z} K(x)dz = \int_0^\infty K(x)dx < \infty$. For this kernel, Corollary 2.1 from Mammen *et al.* (1991) implies that if $k \geq N(f)$,

$$\mathbb{P}\left(n^{-1/4} \leq h_{crit,k} \leq n^{-1/6}\right) \to 1,$$

and we thus have $(\mathcal{F}4)$ and $(\mathcal{F}5)$ (see also Lemma 1 and Lemma 2 from Futschik and Isogai (2006)). Theorem A can then be applied. □

Assuming some regularity conditions on the kernel, Hall *et al.* (2004) proved similar results in their Theorems 3.1 and 3.2 (pp. 2130–2131). These conditions on the kernel are stronger than continuity on $\mathbb{R}$ and thus the uniform kernel does not satisfy them. In that case, we prove that we cannot have $(\mathcal{F}4)$ in the following theorem.

Theorem 7. *For any probability density function* f *of* X, *let* $\hat{f}_{K,h_{crit,k}}$ *be the estimator of* f *when* K *is the uniform kernel with* $h_{crit,k}$ *given in (3.9). Then we have* $h_{crit,k}$ *increasing with* n, *for all* $k \in \mathbb{N}$.

The proof is given in A.3.9. See also Appendix A.1 for details.

3.2.2.4 Sheather and Jones' plug-in method to choose a bandwidth

In the more general context of estimating a density without assumption on the number of its modes, algorithms that provide a suitable bandwidth h are of particular interest. Among a large selection of procedures, we focus on the plug-in method developed by Sheather and Jones (1991) (see also Jones and Sheather (1991)), which leads to the bandwidth h_{SJ}. We chose it because h_{SJ} has good asymptotic properties and is easy

to compute. This bandwidth is designed to minimize the asymptotic mean integrated squared error ($AMISE$) between $\hat{f}_{K,h}$ and f, defined as:

$$AMISE(h) = \frac{R(K)}{nh} + \frac{1}{4}\sigma_K^4 h^4 R(f^{(2)}), \tag{3.10}$$

where for any function ψ, $R(\psi) = \int_{-\infty}^{\infty} \psi(x)^2 dx$ and σ_K^2 is the variance of a random variable of density K. To minimize $AMISE(h)$, $R(f^{(2)})$ must be estimated. For this purpose, modifying an estimator studied by Hall and Marron (1987), Sheather and Jones (1991) used the following one:

$$\hat{R}(f^{(2)}) = \frac{1}{n(n-1)\tilde{h}^5} \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{K}^{(4)} \left(\frac{X_i - X_j}{\tilde{h}} \right), \tag{3.11}$$

where $\tilde{K}$ is allowed to be different from the kernel K used in the estimate of f, $\tilde{K}^{(4)}$ is the fourth derivative of $\tilde{K}$ and $\tilde{h} = \left[\frac{2K^{(4)}(0)}{n\sigma_K^2 \hat{R}(f^{(3)})} \right]^{1/7}$. Estimator $\hat{R}(f^{(3)})$ of $R(f^{(3)})$ is similar to the one in (3.11). It requires a new bandwidth $\check{h}$ chosen to be equal to $0.912\hat{\lambda}n^{-1/9}$, where $\hat{\lambda}$ is the sample interquartile range. Finally,

$$h_{SJ} = \arg\min_h \left(\frac{R(K)}{nh} + \frac{1}{4}\sigma_K^4 h^4 \hat{R}(f^{(2)}) \right).$$

Let $h_{opt} = \arg\min_h(MISE(h))$, Sheather and Jones (1991) showed that

$$\frac{h_{SJ}}{h_{opt}} = 1 + O_P(n^{-5/14}).$$

This means that h_{SJ} is close to the bandwidth that minimizes the expected L_2 distance between a kernel density estimator and the true density. Note that, for the Gaussian kernel, this result is valid for a density with three derivatives and which satisfies for all x and y:

$$\exists Z > 0, \ \left| f^{(3)}(x) - f^{(3)}(y) \right| \leq Z \left| x - y \right|^{\frac{1}{4}}.$$

Remark 4. Among other methods to estimate $R(f)$ in (3.10), one can replace f by the density of a normal distribution as described in Silverman (1986, Section 3.4.2). This is called the normal reference rule. Another alternative is to use the density of a mixture of Gaussian distributions instead of f, in $R(f)$. This technique was introduced by Einbeck and Taylor (2013). It relies on several approximations but the resulting bandwidth is fast to compute and provides small values of $MISE(h)$ in simulation. The authors explain that the number m of components of the underlying Gaussian mixture should be set with respect to $N(f)$. Notice that the number of modes of the estimate which is produced by this procedure is not necessarily equal to m.

3.2.2.5 Polonik's estimator based on excess mass location

The kernel density estimator $\hat{f}_{K,h}$ aims to associate a fixed point t with a value $\hat{f}_{K,h}(t)$ as close as possible to $f(t)$. Another approach, described in this subsection, tries to determine for every given $\lambda \in [0, \infty[$, the set $\hat{\Gamma}_{n,\mathbb{C}}(\lambda)$ which is the most similar to $\Gamma(\lambda) =$

$\{t \ : \ f(t) \geq \lambda\}$, where $\mathbb{C}$ is a set of unions of disjoint intervals of $\mathbb{R}$ chosen such that for every $\lambda \in [0, \infty[$, $\Gamma(\lambda)$ lies in $\mathbb{C}$. In our case, $\mathbb{C}$ is made of every unions of at most $N(f)$ disjoint intervals. This procedure was developed by Polonik (1995b,a) and is related to the test for multimodality introduced by Müller and Sawitzki (1991). It leads to an estimator of $f(t) = \int_0^\infty \mathbf{1}_{\Gamma(\lambda)}(t)d\lambda$ defined by:

$$\hat{f}_P(t) = \int_0^\infty \mathbf{1}_{\hat{\Gamma}_{n,\mathbb{C}}(\lambda)}(t)d\lambda,$$

where $\hat{\Gamma}_{n,\mathbb{C}}(\lambda)$, the so-called empirical generalized λ-cluster in $\mathbb{C}$, satisfies

$$\hat{\Gamma}_{n,\mathbb{C}}(\lambda) = \arg \max_{\{C(i)\}_{i\in\{0,\dots n\}}} \left(\frac{i}{n} - \lambda\mu(C(i))\right).$$

For every $C \in \mathbb{C}$, $\mu(C)$ is the sum of the length of all disjoint closed intervals in C. $C(i)$ is one of the narrowest elements of $\mathbb{C}(i)$, which means that $\forall C \in \mathbb{C}(i)$, $\mu(C(i)) \leq \mu(C)$, where $\mathbb{C}(i)$ is the subset of $\mathbb{C}$ such that $\forall C \in \mathbb{C}(i)$, $\sum_{j=1}^n \mathbf{1}_C(X_j) = i$. Then, if $\int_{\mathbb{R}} \hat{f}_P(t)dt = 1$, Polonik (1995a, Theorem 3.1) gives that:

$$\mathbb{P}\left(\lim_{n\to\infty} \int_{\mathbb{R}} \left|\hat{f}_P(t) - f(t)\right| dt = 0\right) = 1.$$

We have $\int_{\mathbb{R}} \hat{f}_P(t)dt = \frac{n-1}{n}$, but, the proof of Theorem 3.1 from Polonik (1995a) still works for $\int_{\mathbb{R}} \hat{f}_P(t)dt \leq 1$. Thus, to obtain a value for $\hat{f}_P(t)$, $C(i)$ should be determined for every $i \in \{0, \dots, n\}$. Then, because $\hat{\Gamma}_{n,\mathbb{C}}(\lambda)$ is the $C(i)$ that maximizes $n+1$ linear functions, it only changes for a finite number k_n of λ, with $k_n \leq n$. Let $\left\{\lambda_{(i)}\right\}_{i\in\{1,\dots,k_n\}}$ be the set of this ordered change points and take $\lambda_{(0)} = 0$. When $\lambda \geq \lambda_{(k_n)}$, $\hat{\Gamma}_{n,\mathbb{C}}(\lambda) = C(0)$ and $\mu(C(0)) = 0$, and we can set $\hat{\Gamma}_{n,\mathbb{C}}(\lambda) = C(0) = \emptyset$. Hence, another expression for $\hat{f}_P$ is given by

$$\forall t \in \mathbb{R}, \ \hat{f}_P(t) = \sum_{i=0}^{k_n-1} \left[\left(\lambda_{(i+1)} - \lambda_{(i)}\right) \mathbf{1}_{\hat{\Gamma}_{n,\mathbb{C}}(\lambda_{(i)})}(t)\right],$$

which can be used to computed $\hat{f}_P(t)$.

3.2.3 Simulation study

In this simulation study, we compare four density estimators: $\hat{f}_{K,h_{crit}}$ based on the bandwidth h_{crit} with both Gaussian and uniform kernels, $\hat{f}_{K,h_{SJ}}$ based on the Sheather and Jones' bandwidth with the Gaussian kernel, and Polonik's estimator $\hat{f}_P$. We specifically show numerical convergences of $\hat{f}_{K,h_{crit}}$ with the Gaussian kernel and illustrate consequences of the unsatisfied requirement of the uniform kernel.

3.2.3.1 Simulated data and quality assessment of the estimates

To generate simulated datasets, we use a beta mixture, a Gaussian mixture and a mixture of t-distributions. The beta mixture model is defined by:

$$X \sim \begin{cases} \mathcal{B}(\alpha_1, \beta_1) & \text{with probability } p_1, \\ \mathcal{B}(\alpha_2, \beta_2) & \text{with probability } p_2 = 1 - p_1. \end{cases} \tag{3.12}$$

Note that the corresponding density of X is:

$$f_1(t) = p_1 \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} t^{\alpha_1-1}(1-t)^{\beta_1-1} + p_2 \frac{\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)} t^{\alpha_2-1}(1-t)^{\beta_2-1}.$$

In this section, we used the parameters $\alpha_1 = 2$, $\beta_1 = 5$, $\alpha_2 = 10$, $\beta_2 = 2$, $p_1 = \frac{2}{3}$. Graphically, we observe that $N(f_1) = 2$, which is theoretically confirmed in Section 3.2.4.

The Gaussian mixture we chose is the asymmetric claw density introduced by Marron and Wand (1992) :

$$X \sim \begin{cases} \mathcal{N}(0,1) & \text{with probability } \frac{1}{2}, \\ \mathcal{N}\left(l+\frac{1}{2}, \left(\frac{2^{-l}}{10}\right)^2\right) & \text{with probability } \frac{2^{1-l}}{31}, \text{ for } l \in \{-2,-1,0,1,2\}. \end{cases} \tag{3.13}$$

Despite the fact that this mixture has 6 components, the underlying density has 5 modes (see Minnotte *et al.* (1998)). Its expression is:

$$f_2(t) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2}t^2} + \sum_{l=-2}^{2} \frac{20}{31\sqrt{2\pi}} e^{-50\left(\frac{t-l-\frac{1}{2}}{2^{-l}}\right)^2}.$$

The last mixture studied in this section is made of two Student's t-distributions and satisfies:

$$X \sim \begin{cases} \mathcal{T}(2,-3) & \text{with probability } \frac{1}{2}, \\ \mathcal{T}(2,3) & \text{with probability } \frac{1}{2}, \end{cases} \tag{3.14}$$

where $X \sim \mathcal{T}(\nu,\rho)$ means that X is equal to ρ plus a random variable generated from a t-distribution with ν degrees of freedom. It can be proven that the density f_3 which is associated with the model (3.14) has two modes. In addition, we have

$$f_3(t) = \frac{1}{4\sqrt{2}}\left(1+\frac{(t+3)^2}{2}\right)^{-3/2} + \frac{1}{4\sqrt{2}}\left(1+\frac{(t-3)^2}{2}\right)^{-3/2}.$$

Because most of theoretical results we present in this paper concern the L_1 distance between an estimator $\hat{f}$ and the true density f, it makes sense to use the following criterion, often called integrated absolute error (IAE), defined as:

$$IAE = \left\|\hat{f}-f\right\|_{L_1} = \int_{\mathbb{R}} \left|\hat{f}(t)-f(t)\right| dt.$$

Given different datasets, it can be difficult to compare their respective estimated densities and to draw conclusions from these functions. Locations of modes and antimodes can thus be studied to build real indicators which are easiest to handle as explained in Section 3.2.5. Estimating such positions is straightforward when f and its estimate have the same number of modes. When it is not the case, we would like to determine if these positions can still be estimated. To do so, we estimate the location of the lowest local minimum

$$z = \arg \min_{z_i \in \{z_{2j}\}_{j\in\{1,\dots,N(f)-1\}}} f(z_i),$$

with z_{2j} defined in ($\mathcal{H}3$). The estimator $\hat{z}$ of z is chosen to satisfy

$$\hat{z} = \arg\min_{z_i \in \hat{Z}} \hat{f}(z_i),$$

where $\hat{Z}$ is made of the points $\tilde{z}$ such that,

$$\exists \varepsilon > 0,\ \forall \tilde{\varepsilon} < \varepsilon,\ \hat{f}(\tilde{z} - \tilde{\varepsilon}) > \hat{f}(\tilde{z}) \text{ and } \hat{f}(\tilde{z} + \tilde{\varepsilon}) > \hat{f}(\tilde{z}).$$

In Sections 3.2.3.2-3.2.3.4, we will study values of IAE as well as estimates $\hat{z}$ for the three models which are presented in this section and for several density estimators.

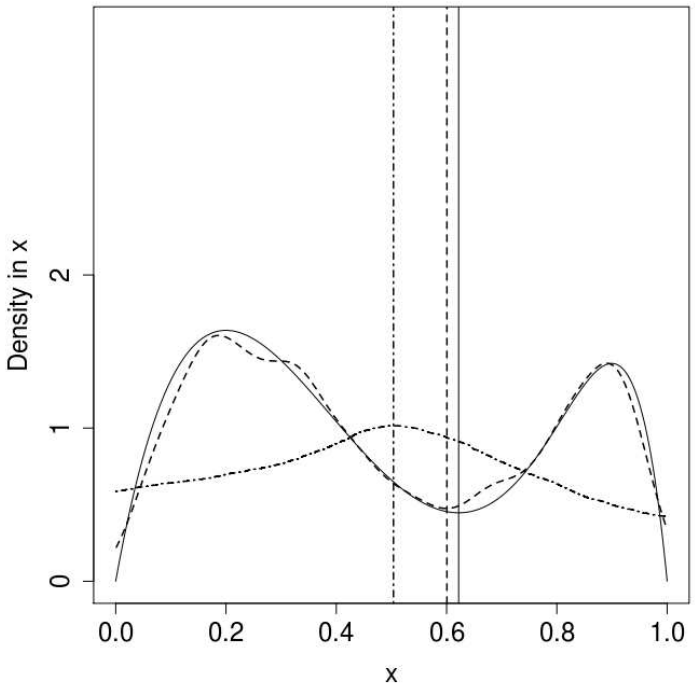
3.2.3.2 Simulation results for the beta mixture

In Figure 3.6, we first draw an example of density estimation with the four considered estimators, using a sample of size $n = 1600$ generated from model (3.12). Apart from $\hat{f}_{K,h_{crit}}$ with the uniform kernel, estimates of the density seem to be close to f, considering the shape. We observe a lot of peaks for $\hat{f}_P$. This feature is directly related to the estimation method as it has been already noticed by Müller and Sawitzki (1991). Polonik (1995a) wrote that $N(\hat{f}_P)$ can be different from $N(\hat{f})$. In Figure 3.6(b), the estimation of z related to $\hat{f}_P$ is close to 0 and far from $z \approx 0.6219$. For a given estimation $\hat{f}$ of f, finding $\hat{z}$ in $]-\infty, z_1[\cup]z_{2N(f)-1}, \infty[$ only happens if there is a mode of $\hat{f}$ in this interval that is sufficiently far from z_1 and from $z_{2N(f)-1}$. This event is especially likely to arise when $N(\hat{f}) > N(f)$ but it can also happen when $N(\hat{f}) = N(f)$ if the modes of f are not clearly separated. In addition, when K is the Gaussian kernel, estimations of z for $\hat{f}_{K,h_{crit}}$ (Figure 3.6(a)) and for $\hat{f}_{K,h_{SJ}}$ (Figure 3.6(b)) are close to z.

Then, we generate 100 replicates from model (3.12) for various sample sizes $n \in \{100 \times 2^i\}_{i\in\{0,\dots,9\}}$. For each sample and each density estimation procedure, IAE is computed. The corresponding values are represented in Figure 3.7 with boxplots. Not surprisingly, for $\hat{f}_{K,h_{crit}}$ with the uniform kernel, we observe in Figure 3.7(a) that IAE increases with n, which is compatible with Theorem 7. Performances of $\hat{f}_{K,h_{crit}}$ for the Gaussian kernel, shown in Figure 3.7(b), are better. In this case, boxplots exhibit the L_1 convergence of Theorem 6. They reach a similar precision to those obtained for $\hat{f}_{K,h_{SJ}}$ which are drawn in Figure 3.7(c). Polonik's method (Figure 3.7(d)) needs extensive computational time. That is why we do not draw boxplots for the greatest sample sizes, but we still observe convergence of this procedure despite of the many peaks of the estimates. Values of IAE appear to be slightly greater for this estimator than those for $\hat{f}_{K,h_{SJ}}$ and $\hat{f}_{K,h_{crit}}$ with the Gaussian kernel.

For each replicate we made, we also compute estimations of z. In Figure 3.8 we draw various values of $\hat{z}$ for the four previously considered estimators. For $\hat{f}_{K,h_{crit}}$ with the uniform kernel, in Figure 3.8(a), $\hat{z}$ values move away from the position of the local minimum of f_1 (which is equal to 0.6219) when sample size increases. This is not surprising because of the poor quality of this estimator which, according to Figure 3.6, provides two modes close to each other, but far from the true ones. In Figure 3.8(b), we notice a convergence of the boxplots toward the aimed location for $\hat{f}_{K,h_{crit}}$ with the Gaussian kernel. For $\hat{f}_{K,h_{SJ}}$ with the Gaussian kernel, in Figure 3.8(c), a similar convergence can be observed. For $\hat{f}_P$, in Figure 3.8(d), values of $\hat{z}$ seem to tend toward 0 or 1 when sample size increases. A possible explanation of this phenomenon is that a spurious mode close to 0 or 1 is sometimes created by the procedure. Because this mode is located in an interval of t where $f_1(t)$ is small, the local minimum near the mode is the minimum over all local minima. This occurs in Figure 3.6(b), for example.

(a) Estimations $\hat{f}_{K,h_{crit}}$ with Gaussian kernel (dashed line) and $\hat{f}_{K,h_{crit}}$ with uniform kernel (mixed line).



(b) Estimations $\hat{f}_P$ (long dashed line) and $\hat{f}_{K,h_{SJ}}$ (dotted line).

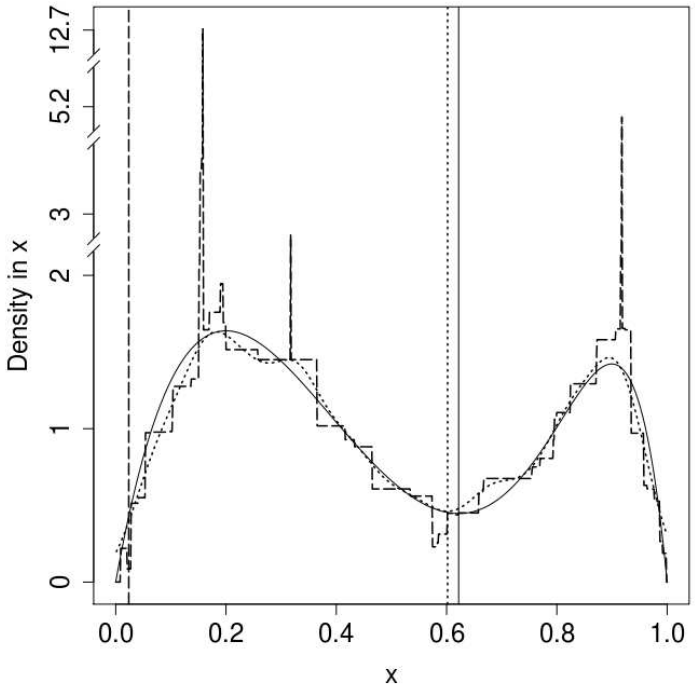
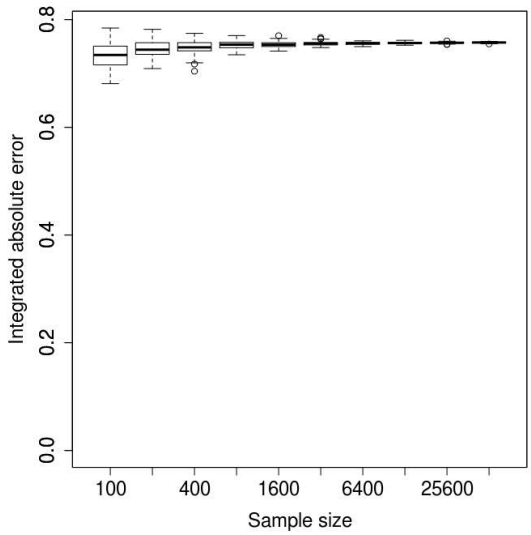
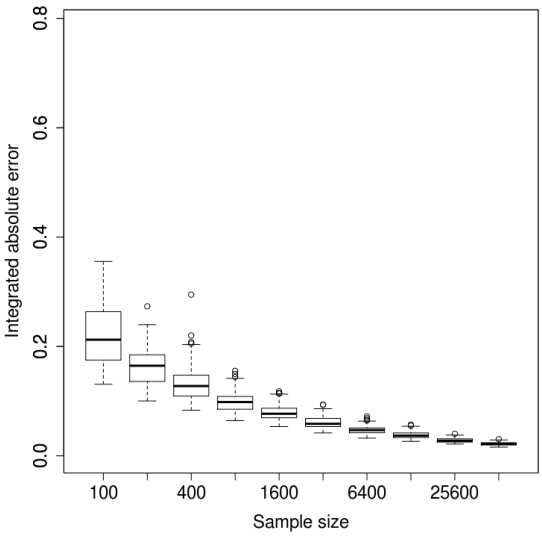


Figure 3.6: Beta mixture density f_1 (solid line), and various estimations; for $n = 1600$. Vertical lines: positions of the minimum of local minima of each plotted density. In Figure 3.6(b), vertical axis is broken between 2.2 and 2.8, between 3.5 and 5 and between 5.5 and 12.5 to be able to clearly see the shape of all curves.

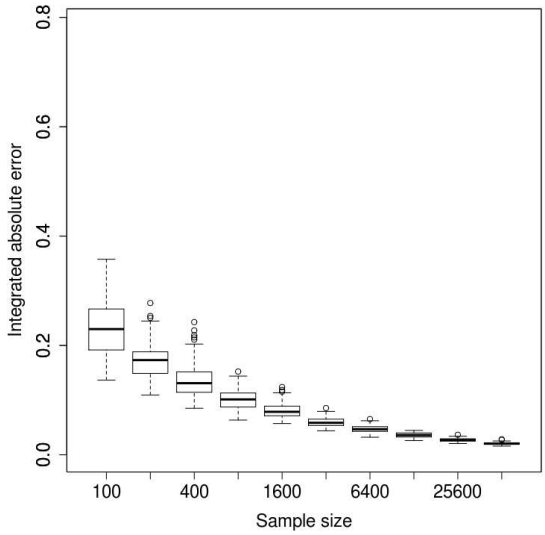
(a) Kernel density estimator with h_{crit} and the uniform kernel.



(b) Kernel density estimator with h_{crit} and the Gaussian kernel.



(c) Kernel density estimator with h_{SJ} and the Gaussian kernel.



(d) Polonik's estimator.

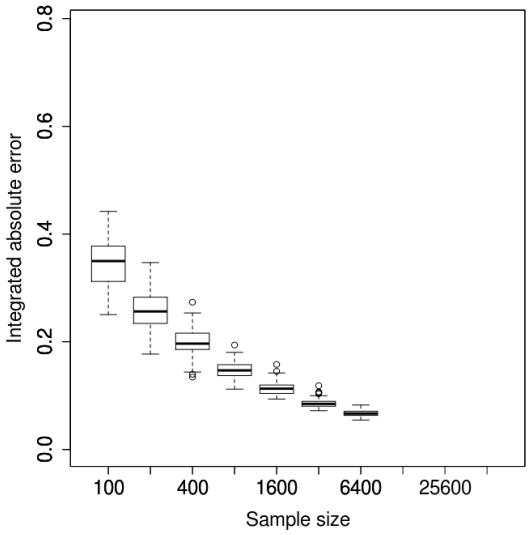
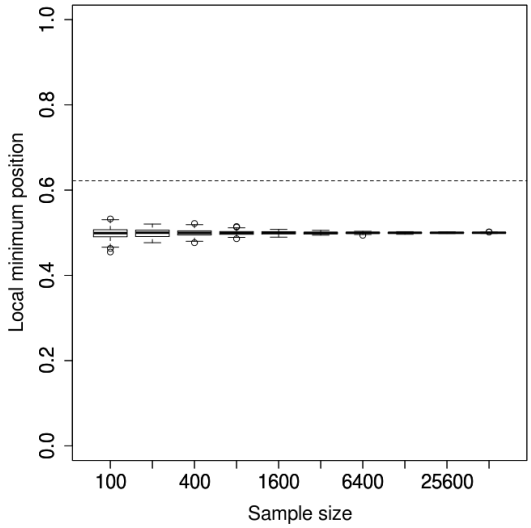
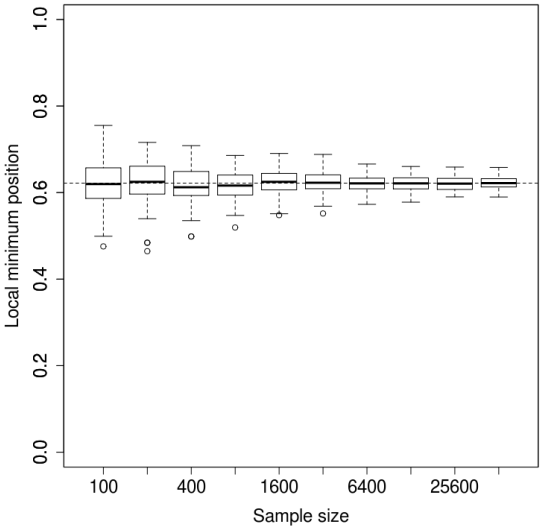


Figure 3.7: IAE for the beta mixture model (3.12) and various density estimators.

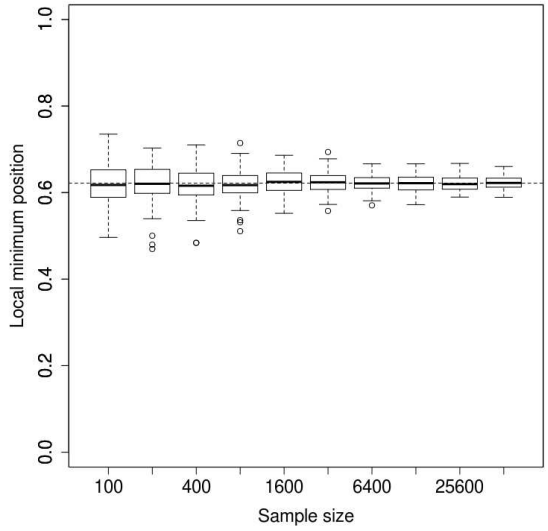
(a) Estimator related to $\hat{f}_{K,h_{crit}}$ with the uniform kernel.



(b) Estimator related to $\hat{f}_{K,h_{crit}}$ with the Gaussian kernel.



(c) Estimator related to $\hat{f}_{K,h_{SJ}}$ with the Gaussian kernel.



(d) Estimator related to $\hat{f}_P$.

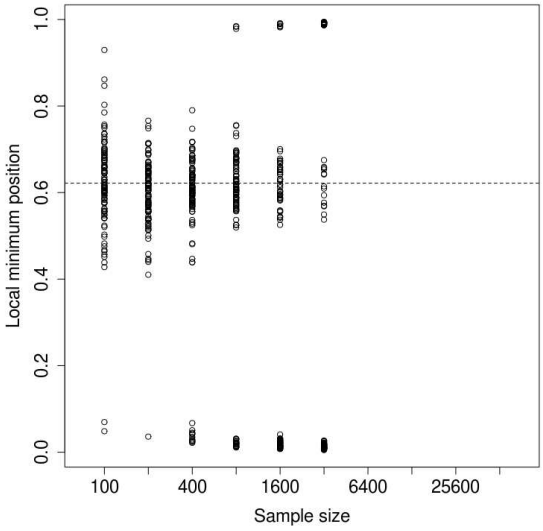


Figure 3.8: Estimations of the position of the local minimum of the density of the beta mixture model.

3.2.3.3 Simulation results for the asymmetric claw density

The simulation results of Section 3.2.3.2 lead us to focus only on the estimators $\hat{f}_{K,h_{SJ}}$ and $\hat{f}_{K,h_{crit}}$ with the Gaussian kernel because both exhibit IAE convergence toward 0 and convergence of $\hat{z}$ toward z. To study them in more depth, we use the model (3.13). Moreover, this selection is also made because of the costs in computational time of the different methods. For example, for a given sample of size $n = 1600$, we measured computational time of the methods we consider in this study, with our Intel Core 2 Quad Q9505 processor. To obtain h_{crit} with a Gaussian kernel we need about 1.6 seconds using the `density()` R function. Finding h_{SJ} requires 0.004 seconds, with the `KernSmooth` R package while our R implementation of $\hat{f}_P$ needs 425 seconds to be computed. Our R algorithm finds h_{crit} in 4 seconds for the uniform kernel.

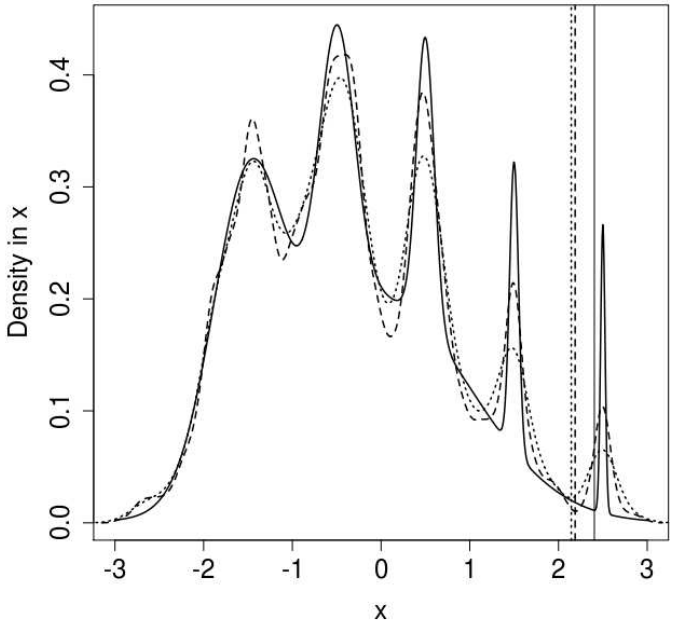
In Figure 3.9(a), we draw another example of density estimation using the estimators $\hat{f}_{K,h_{SJ}}$ and $\hat{f}_{K,h_{crit}}$, and a sample of size $n = 1600$ generated from model (3.13). We observe very similar results for both procedures even if $\hat{f}_{K,h_{crit}}$ seems to produce an estimate slightly more precise than $\hat{f}_{K,h_{SJ}}$ for the estimation of z. We also provide estimates $\hat{f}_{K,h_{crit,2}}$ and $\hat{f}_{K,h_{crit,7}}$ for this sample, in Figure 3.9(b). On one hand, choosing $k = 7 > N(f)$ does not change the quality of the estimation, comparing to $\hat{f}_{K,h_{crit}}$. On the other hand, setting $k = 2 < N(f)$ leads to an estimator which smooths the modes of f_2 and thus may produce poor values of IAE. Because the only antimode of $\hat{f}_{K,h_{crit,2}}$ does not have any reason to be the lowest one, estimating z from this density can also provide inaccurate results, as in Figure 3.9(b). This explains the constraint $k \geq N(f)$ in Theorem 6.

Boxplots presenting IAE values for these methods are drawn in Figure 3.10. The dispersion of the IAE values does not seem to decrease with n in Figure 3.10(a). This result could come from the fact that for the asymmetric claw density, $(\mathcal{H}2)$ does not hold and Theorem 6 cannot be applied, but IAE values of Figure 3.10(a) globally decrease. Thus, another practical explanation is that the bandwidth h_{crit} cannot adapt to the various sharpness of the modes of f_2, while $\hat{f}_{K,h_{SJ}}$ may compensate this phenomenon by creating a new mode such that $N(\hat{f}_{K,h_{SJ}}) > N(f_2)$. Indeed, in Figure 3.10(b), IAE values converge toward 0 like those in Figure 3.7(c).

We notice that for the asymmetric claw density, estimates of z can be located at the extrema of the sample, for both $\hat{f}_{K,h_{crit}}$ and $\hat{f}_{K,h_{SJ}}$, in Figure 3.11. This previously occurs for $\hat{f}_P$ and model (3.12) in Figure 3.8(d). For $\hat{f}_{K,h_{crit}}$, in Figure 3.11(a), samples that produce this kind of estimate are sufficiently rare for the estimates to be considered as outliers when building boxplots. In Figure 3.11(b), for $\hat{f}_{K,h_{SJ}}$ and when $n \geq 6400$, every $\hat{z}$ is in the tails of the estimated distribution. Thus, for the estimation of z, h_{crit} seems to perform better than h_{SJ} despite of the fact that $(\mathcal{H}2)$ does not hold for f_2.

Remark 5. Sheather and Jones (1991) proposed two versions of their bandwidth h_{SJ} respectively called "direct plug-in" and "solve the equation". In this article, only the "direct plug-in" approach is considered. The "solve the equation" version leads to smaller values of IAE in simulation than the "direct plug-in" but needs more computational time. Results still are of the same order of magnitude. Conclusions about the estimation of z are valid for both of them.

(a) Estimations $\hat{f}_{K,h_{crit}}$ (dashed line) and $\hat{f}_{K,h_{SJ}}$ (dotted line).



(b) Estimations $\hat{f}_{K,h_{crit,2}}$ (mixed line) and $\hat{f}_{K,h_{crit,7}}$ (long dashed line).

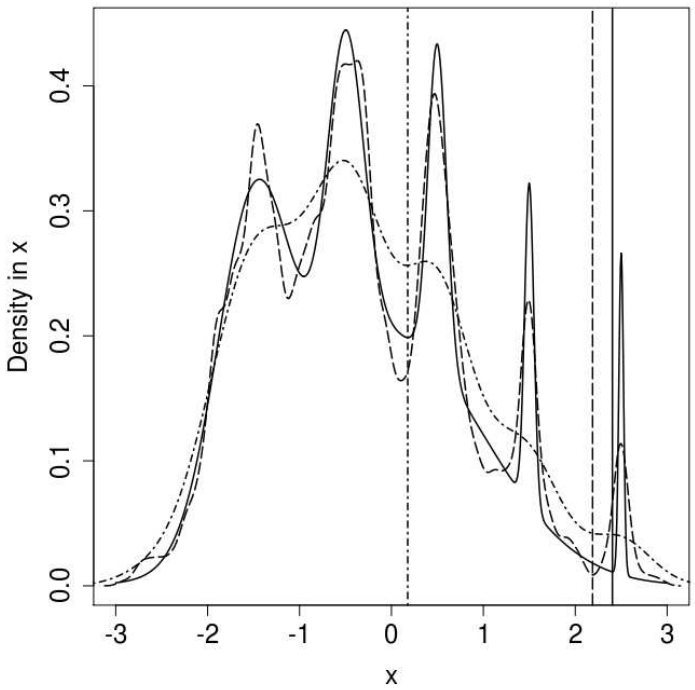


Figure 3.9: Asymmetric claw density f_2 (solid line) and various estimations; for $n = 1600$ and the Gaussian kernel K. Vertical lines: positions of the minimum of local minima of each plotted density.

(a) Kernel density estimator with h_{crit} and the Gaussian kernel.

(b) Kernel density estimator with h_{SJ} and the Gaussian kernel.

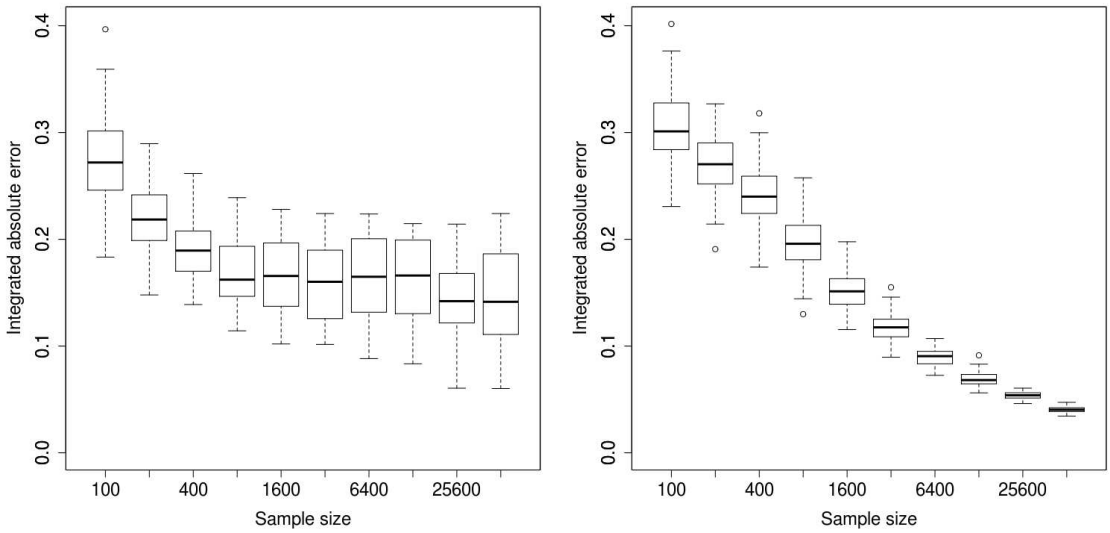


Figure 3.10: IAE for the asymmetric claw model (3.13) and various density estimators.

(a) Estimator related to $\hat{f}_{K,h_{crit}}$ with the Gaussian kernel.

(b) Estimator related to $\hat{f}_{K,h_{SJ}}$ with the Gaussian kernel.

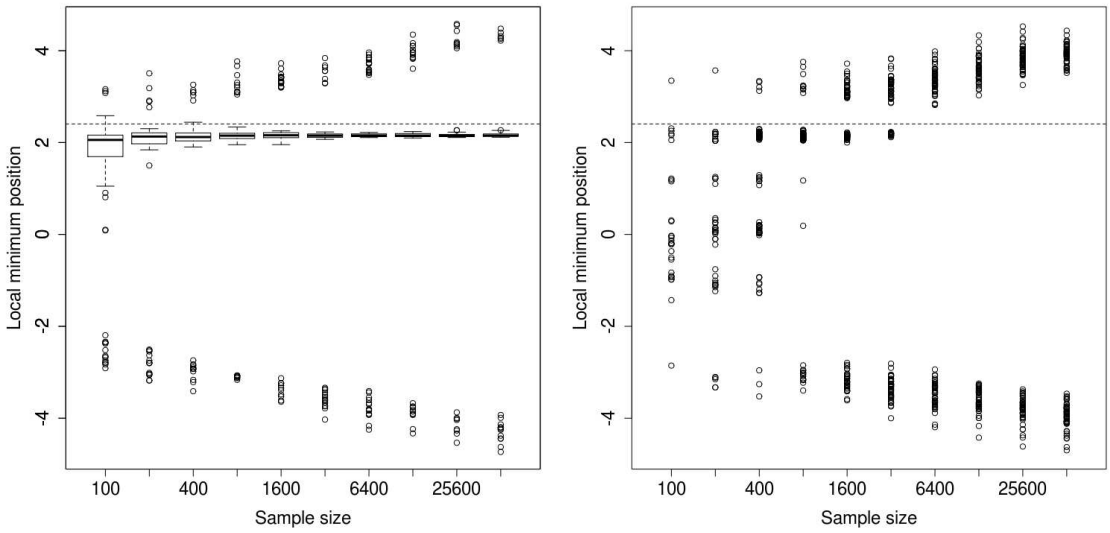


Figure 3.11: Estimations of the position of the local minimum of the density of the asymmetric claw model.

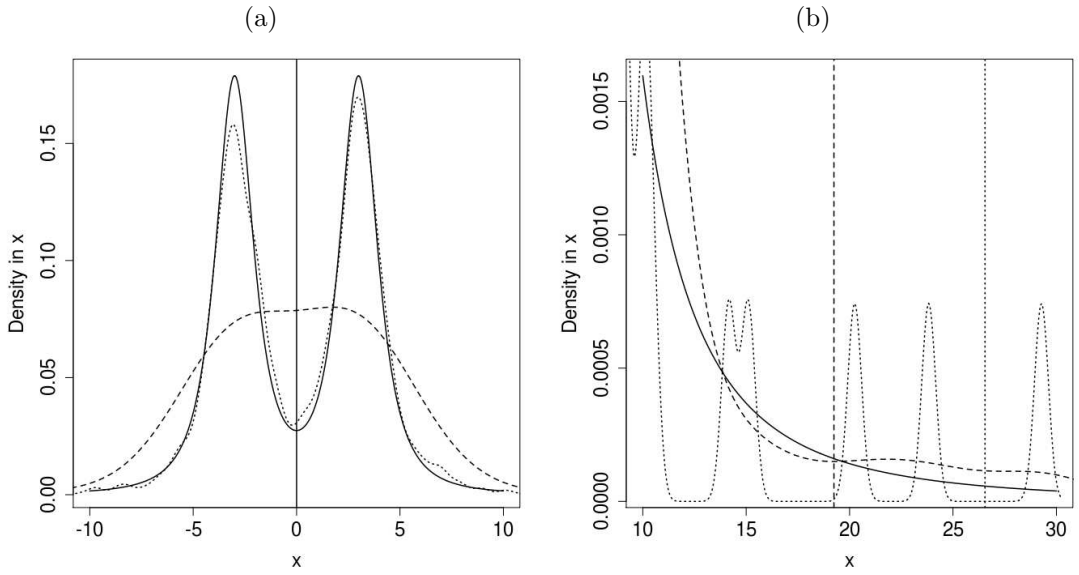


Figure 3.12: Density f_3 of a mixture of t-distributions (solid line). Estimations $\hat{f}_{K,h_{crit}}$ with Gaussian kernel (dashed line) and $\hat{f}_{K,h_{SJ}}$ (dotted line), for $n = 1600$. Vertical lines: positions of the minimum of local minima of each plotted density.

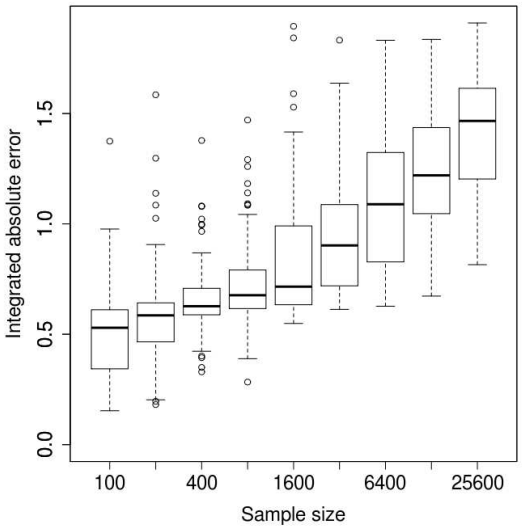
3.2.3.4 Simulation results for the mixture of t-distributions

Studying the model (3.14), we estimate f_3 with $\hat{f}_{K,h_{SJ}}$ and $\hat{f}_{K,h_{crit}}$ from a sample of size $n = 1600$, and with the Gaussian kernel K. In Figure 3.12, we draw the resulting estimates. The bandwidth h_{crit} seems too large since $\hat{f}_{K,h_{crit}}$ is smoother than f_3. Because f_3 has two modes, we have $N(\hat{f}_{K,h_{crit}}) = 2$ but one mode of $\hat{f}_{K,h_{crit}}$ is close to the modes of f_3 (Figure 3.12(a)) and the other one is in its tails (Figure 3.12(b)). The estimate $\hat{f}_{K,h_{SJ}}$ appears to perform better than $\hat{f}_{K,h_{crit}}$ in this example but it also produces a lot of antimodes in the tails. The lowest local minimum of $\hat{f}_{K,h_{SJ}}$ is thus far from 0 as presented in Figure 3.12(b).

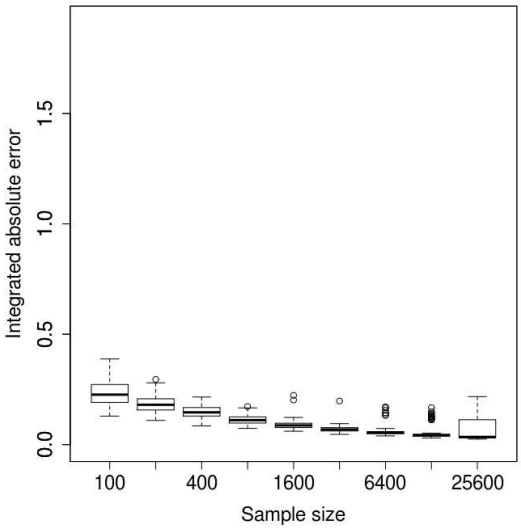
In this paragraph, K is the Gaussian kernel. The tendency in Figure 3.12 is confirmed in Figure 3.13 where values of the integrated absolute error between f_3 and $\hat{f}_{K,h_{crit}}$ (resp. $\hat{f}_{K,h_{SJ}}$), are displayed. When the sample size increases, $\hat{f}_{K,h_{crit}}$ exhibited poor results. Note that f_3 has heavy tails and then when n is sufficiently large, there is a great probability that a few points of the studied sample are really far from 0. If this happens, the estimator $\hat{f}_{K,h_{crit}}$ produces a mode near these points and oversmooths the density estimate. Observing Figure 3.13(b), $\hat{f}_{K,h_{SJ}}$ seems to perform pretty well even when the density to estimate has heavy tails. When n is large, the bandwidth h_{SJ} becomes however really little, implying many spikes and almost constant parts in $\hat{f}_{K,h_{SJ}}$. These features lead us to rely on a Monte-Carlo algorithm to integrate since it performs better in such cases than the QUADPACK library used until now. The poor results of $\hat{f}_{K,h_{crit}}$ encourage us to consider $\hat{f}_P$ again because it does not need that the support of f is bounded. In Figure 3.13(c), we observe that this estimator produces small values of the IAE, even when the tails of f are heavy.

With the Gaussian kernel K, $\hat{f}_{K,h_{SJ}}(t)$ is often numerically equal to 0 where t is one of

(a) Kernel density estimator with h_{crit} and the Gaussian kernel.



(b) Kernel density estimator with h_{SJ} and the Gaussian kernel.



(c) Polonik's estimator.

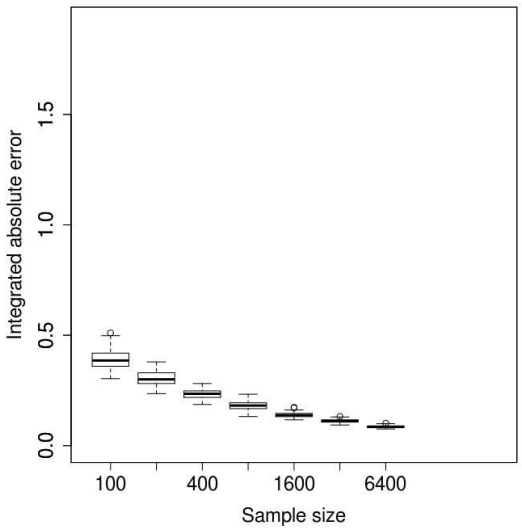


Figure 3.13: IAE for the mixture (3.14) of t-distributions and various density estimators.

(a) Kernel density estimator with h_{crit} and the Gaussian kernel.

(b) Polonik's estimator.

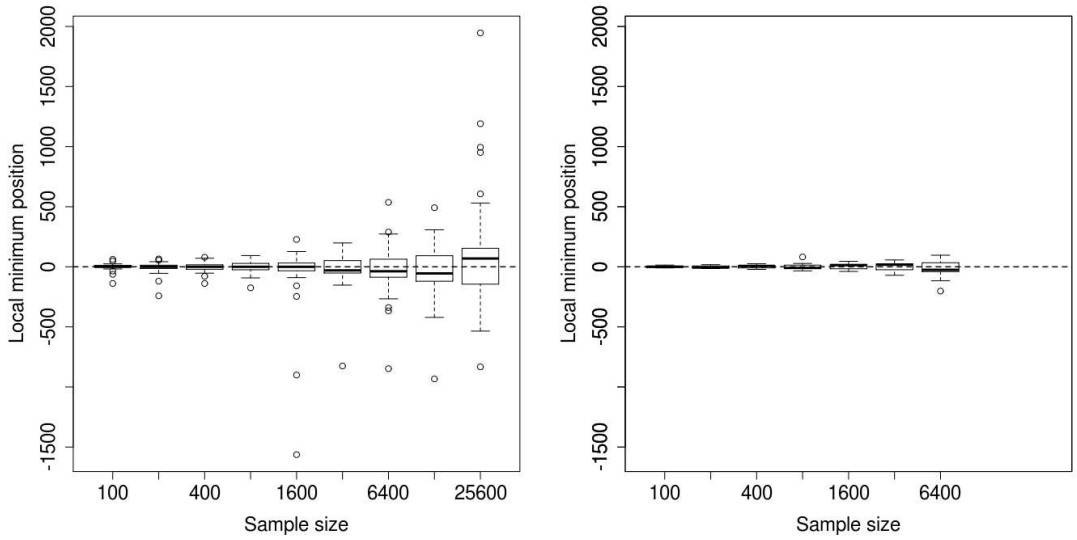


Figure 3.14: Estimations of the position of the local minimum of the density of the mixture of t-distributions with various estimators.

the numerous local minimum of $\hat{f}_{K,h_{SJ}}$. For example, for 100 samples of size $n = 128000$, we found that the mean number of antimodes where $\hat{f}_{K,h_{SJ}}(t) = 0$ is equal to 813.56. It is thus impossible to find a relevant estimate of z using h_{SJ}. It is also difficult to estimate z from $\hat{f}_{K,h_{crit}}$ where K is the Gaussian kernel as one can see in Figure 3.14(a). This graphic shows the increasing dispersion of the position of the unique local minimum of $\hat{f}_{K,h_{crit}}$ when n increases, which results from the presence of modes far from 0. The bandwidth h_{crit} does not lead then to a satisfying estimator of z in this case. Polonik's estimator also exhibits an increasing dispersion in the corresponding values of $\hat{z}$ as displayed in Figure 3.14(b). Although this increase is less important than the one related to $\hat{f}_{K,h_{crit}}$, $\hat{f}_P$ does not provide accurate estimates of z and should not be used for this purpose.

While for model (3.13), the assumption ($\mathcal{H}2$) did not appear to be important, we thus see why it is for this mixture of t-distributions. This is consistent with Hall and York (2001, Section 2.3) in which it is stated that $h_{crit,1}$ does not satisfies ($\mathcal{F}4$) when f is a t-distribution. The authors also indicate that, when f is a normal distribution, if $h_{crit,1}$ satisfies ($\mathcal{F}4$) then its rate of convergence toward 0 is slow. This could explain Figure 3.10(a). To avoid difficulties when studying densities with unbounded support, Hall and York (2001) propose to replace the function N in (3.9) with a function $N_{\mathcal{I}}$ that returns the number of modes of a density which are observed in a given bounded interval $\mathcal{I}$. This interval must be chosen by the user.

3.2.4 Assuming the number of modes of a mixture density

Estimation of a density of a mixture model with a known maximum number of components is the main task that $\hat{f}_{K,h_{crit,k}}$ can realize. Indeed, if each component of the mixture model is an unimodal density, there are various cases where the number of modes of the density of the mixture is at most equal to the number of components. However, it is not true

for every mixture model. For instance, densities made from components $f_{\mu,\theta}$ that satisfy $f_{\mu,\theta}(x) = C_\theta(1-(x-\mu)^2)^\theta \mathbf{1}_{[-\mu,\mu]}(x)$, where μ and θ should be chosen for each component, can have a number of modes greater than its number of components (see Remark 3 and Hall *et al.* (2004)).

Notice that a function which is convex on an open interval does not have any mode on this interval and that the density of a mixture of densities which are convex on an interval is convex on this interval too. Thus, let f be a density of a mixture model with m components. For $i \in \{1,\dots,m\}$, let g_i be the density of the component i. If there exists $\mu_i \in \mathbb{R}$ such that g_i is convex on $]-\infty,\mu_i[$ and on $]\mu_i,\infty[$, then $N(f) \leq m$ and $\forall j \in \{1,\dots,N(f)\}$, $z_{2j-1} \in \cup_{i=1}^m \{\mu_i\}$. For example one can take for g_i the standard two-sided power distribution of van Dorp and Kotz (2002):

$$g_i(t) = \begin{cases} \gamma_i \left(\frac{t}{\theta_i}\right)^{\gamma_i - 1} & \text{for } t \in [0,\theta_i[, \\ \gamma_i \left(\frac{1-t}{1-\theta_i}\right)^{\gamma_i - 1} & \text{for } t \in [\theta_i, 1], \end{cases}$$

with $\theta_i \in [0,1]$ and $\gamma_i > 2$. If we allow ourselves not to respect $(\mathcal{H}2)$, the Laplace distribution is another valid choice for g_i.

Mixture model densities with $N(f) \leq m$ are not restricted to those that satisfy the previous condition of convexity. Considering the beta mixture model of Section 3.2.3 with $\alpha_1 = 2$ and $\beta_2 = 2$, the density f_4 of (3.12) can be written

$$f_4(t) = p_1 g_1(t) + (1-p_1) g_2(t),$$

with $g_1(t) = \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} t(1-t)^{\beta_1 - 1}$ and $g_2(t) = \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} t^{\alpha_2 - 1}(1-t)$. When $\beta_1 > 2$, for $q \in \{1,2,3\}$, the q^{th} derivative of g_1 satisfies:

$$g_1^{(q)} \geq 0 \Leftrightarrow (-1)^{q-1} x \leq \frac{(-1)^{q-1} q}{\beta_1},$$

and we also have for $q \in \{1,2,3\}$:

$$g_2^{(q)} \geq 0 \Leftrightarrow x \leq 1 - \frac{q}{\alpha_2}.$$

These relations hold for strict inequalities and imply that f_4 has no mode in $[0, \frac{1}{\beta_1}[$ and in $]1-\frac{1}{\alpha_2}, 1]$ because $g_1^{(1)}$ and $g_2^{(1)}$ have the same sign on these intervals on which this sign is constant. Assume now that $\frac{3}{\beta_1} < 1 - \frac{2}{\alpha_2}$ and that $\frac{2}{\beta_1} < 1 - \frac{3}{\alpha_2}$, which is true for the values of β_1 and α_2 chosen in Section 3.2.3 for f_1. Then, f_4 has at most one mode on $\left[\frac{1}{\beta_1}, \min\left\{\frac{3}{\beta_1}, 1-\frac{3}{\alpha_2}\right\}\right[$. This can be proven by assuming that f_4 has two modes located in z_1 and z_3 and an antimode located in z_2 with $\frac{1}{\beta_1} \leq z_1 < z_2 < z_3 < \min\left\{\frac{3}{\beta_1}, 1-\frac{3}{\alpha_2}\right\}$. This implies that $f_4^{(2)}(z_2) = p_1 g_1^{(2)}(z_2) + (1-p_1) g_2^{(2)}(z_2) \geq 0$, and because $g_1^{(3)}(t) > 0$ and $g_2^{(3)}(t) > 0$ for $t \in \left]z_2, \min\left\{\frac{3}{\beta_1}, 1-\frac{3}{\alpha_2}\right\}\right[$, $p_1 g_1^{(2)}(t) + (1-p_1) g_2^{(2)}(t) > 0$ which negates the fact that $f_4^{(2)}(z_3) = p_1 g_1^{(2)}(z_3) + (1-p_1) g_2^{(2)}(z_3) \leq 0$. A demonstration of the same type leads to the property that f_4 has at most one mode on $\left]\max\left\{\frac{3}{\beta_1}, 1-\frac{3}{\alpha_2}\right\}, 1-\frac{1}{\alpha_2}\right]$.

Thus, in order to demonstrate that $N(f_4) \leq 2$, we have to show that f_4 has no mode on Δ, with:

$$\Delta = \left[\max\left\{ \frac{3}{\beta_1}, 1 - \frac{3}{\alpha_2} \right\}, \min\left\{ \frac{3}{\beta_1}, 1 - \frac{3}{\alpha_2} \right\} \right].$$

Because of assumptions we made on β_1 and α_2, Δ is included in $\left]\frac{2}{\beta_1}, 1 - \frac{2}{\alpha_2}\right[$ on which both $g_1^{(2)}$ and $g_2^{(2)}$ are positive. This implies that f_4 is convex on this interval and then has no mode on it. Thus f_4 has no mode on Δ and $N(f_4) \leq 2$. Although we anticipate that generalizing this result to a wider set of mixtures is feasible, the demonstration would probably be tedious.

3.2.5 Oyster opening amplitudes modeled with a bimodal density

In this section we describe a real data application. We apply the estimator $\hat{f}_{K,h_{crit,k}}$ to opening amplitudes of oysters. These animals are studied by a laboratory called Environnements et Paléoenvironnements Océaniques et Continentaux (http://molluscan-eye.epoc.u-bordeaux1.fr, EPOC) in order to derive water quality indicators. Their approach is based on the assumption that a water of poor quality leads to perturbations in the oysters' behavior. It consists in measurements of the distance between both parts of the shell of the oysters with a frequency of 0.625 Hz. The procedure carried out to obtain the dataset is non invasive. It relies on electrodes stuck on the shell of the oysters and on the GSM/GPRS service to transfer the data. The animals studied in this article are immersed in the Bay of Arcachon, in France.

We aim at estimating the density f of the distances of the parts of the shell of these animals. During a day, following the tide, an oyster is either open or closed (see for instance Sow *et al.* (2011)). For each of this state, the density of the opening amplitudes is assumed to be unimodal. If we also assume that these densities behave similarly to the two-sided power distribution or to the beta mixture detailed in Section 3.2.4, we have $N(f) \leq 2$. The density f has obviously a bounded support since the measured distance is at least 0 and since the opening of an oyster can not be infinitely large.

Because of these assumptions, we estimate f with $\hat{f}_{K,h_{crit,2}}$ with the Gaussian kernel and two samples of respective size $n = 53932$ and $n = 53728$, which respectively leads to Figure 3.15(a) and Figure 3.15(b). In Figure 3.15(a), the data come from an oyster that does not exhibit any feature of sickness. In Figure 3.15(b), the oyster analyzed died one week after these measures were recorded. Generally, when they are in the open state, dying oysters produce opening amplitudes with a wide variability which implies a larger mode for the corresponding density. Thus, the local minimum z_2 of f located between its two modes comes close to the location z_1 of the mode related to the close state, when oysters' health becomes poor. This feature can also be observed for $\hat{f}_{K,h_{crit,2}}$ in Figure 3.15. Indeed, let $\hat{z}_1$ and $\hat{z}_3$ be the local maxima of $\hat{f}_{K,h_{crit,2}}$ and let $\hat{z}_2$ be its local minimum such that $\hat{z}_1 < \hat{z}_2 < \hat{z}_3$. Then, $\frac{\hat{z}_2 - \hat{z}_1}{\hat{z}_3 - \hat{z}_1}$ is greater in Figure 3.15(a) than in Figure 3.15(b) with respective values 0.3073 and 0.1566. This observation could lead to a detection of oysters in poor health.

Such an indicator can be built with other estimators of f. The point z_2 could be thus estimated by $\hat{z}$. Both local maxima for $t \leq \hat{z}$ and for $t > \hat{z}$ could be chosen as estimators

(a) Oyster assumed to be healthy.

(b) Oyster that dies the week that follows the day when measurements were taken.

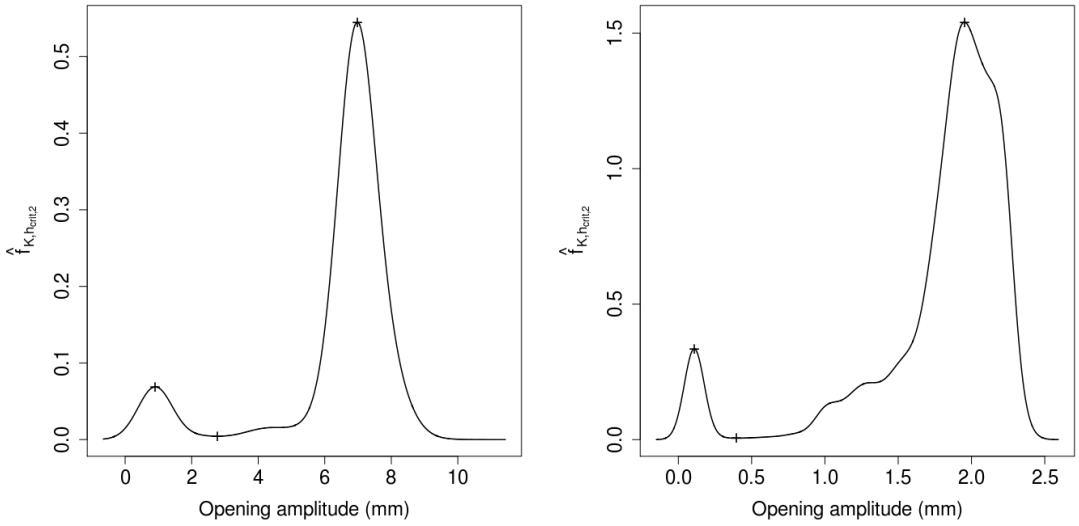


Figure 3.15: Density estimations with $\hat{f}_{K,h_{crit,2}}$ of opening amplitudes of two oysters. Crosses indicate local extrema of the estimated densities.

of the modes. The simulation study of Section 3.2.3 shows however that $\hat{z}$ provides good results when it comes from $\hat{f}_{K,h_{crit}}$, for densities f with bounded support. This explains the choice of the estimator of f in this section.

Maximum amplitude of the openings of the animal is also linked to its health and we can see that it is approximately 4 times greater in Figure 3.15(a) than in Figure 3.15(b). However, this quantity may also vary with the size of the animal, and with the position of the electrodes on it. That is why we prefer not to rely on it.

3.2.6 Concluding remarks

The estimator $\hat{f}_{K,h_{crit,k}}$, when K is a Gaussian kernel, is able to estimate a density f that has a known number of modes because when sample size is large enough, $\hat{f}_{K,h_{crit,k}}$ is close to f from both L_1 and L_∞ points of view. This is not the case when K is the uniform kernel in our simulation study and theoretical reasons were explored (Theorem 7).

When K is the Gaussian kernel and f has a bounded support, the fact that we require that $N(\hat{f}_{K,h_{crit,k}}) = k$ does not imply a great loss of convergence rate of the IAE for $\hat{f}_{K,h_{crit}}$ compared with results of $\hat{f}_{K,h_{SJ}}$, as shown in our simulation study. In addition, $\hat{f}_{K,h_{crit}}$ exhibits slightly better results than $\hat{f}_P$, in this case. When f has heavy tails, $\hat{f}_P$ becomes however more accurate than $\hat{f}_{K,h_{crit}}$ from the point of view of the IAE. To handle this case, the definition of $\hat{f}_{K,h_{crit}}$ can be modified such that spurious modes in its tails are ignored (see for instance Hall and York (2001)).

An interesting feature of $\hat{f}_{K,h_{crit,k}}$ is that it has a deterministic number of modes and can have as many modes as f. In real data analysis, this allows to seek the positions of the various modes and antimodes of the estimate of a density with bounded support, and to derive indicators from them in order to compare densities (see Section 3.2.5). This goal

could also be achieved when $N(f) = 2$ by $\hat{f}_P$ and $\hat{f}_{SJ}$ by choosing $\hat{z}$ as an estimate of the location of the antimode but results for this estimator are better when it is computed from $\hat{f}_{K,h_{crit,k}}$, with the Gaussian kernel K. From a practical point of view, note that the bandwidth h_{crit} required more time to compute than h_{SJ} but less than $\hat{f}_P$.

An important assumption in order for $\hat{f}_{K,h_{crit,k}}$ to converge toward f is that $k \geq N(f)$. In practice, for a mixture model with m unimodal components, this assumption can be replaced by $k \geq m$, provided that $m \geq N(f)$. Note however that for some mixture models, we have $m < N(f)$. The estimator $\hat{f}_{K,h_{crit,k}}$ should thus be used with care in this context.

To conclude, asymptotic properties of $\hat{f}_{K,h_{crit}}$ with the Gaussian kernel, together with its behavior in simulation, and its deterministic number of modes allow this estimator to be applied to real datasets that are assumed to come from mixture model densities. For that matter, an implementation of this work is included in numerical procedures daily performed on the environmental data of EPOC laboratory.

Acknowledgements

The authors would like to thank Jean-Charles Massabuau and Pierre Ciret from the EPOC laboratory (University of Bordeaux, UMR CNRS 5805) for the valvometric data they provide us, and for helpful discussions. They are also grateful to Kristin Clements for improvements on the language of this manuscript. Computer time for this study was partially provided by the computing facilities MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour. The authors also acknowledge the anonymous referee for useful comments that lead to several ameliorations of this article.

3.3 Microclosings, noise and wavelets

Wavelets are common tools to analyse a signal with various resolutions. They share their theoretical background with Fourier series but they may be more appropriate than these series to describe local behaviours. They are involved in the JPEG-2000 standard for image compressing. Such problems are described in Mallat (2008, Chapter 10). We are here interested in wavelets in order to remove a possible noise from a valvometric signal and to automatically detect spikes in this signal. We focus on Haar wavelets which are piecewise constant because we believe that valvometric signals, without noise, vary slowly in the intervals where there is no spike. We introduce these wavelets in Section 3.3.1. More general kinds of wavelets can be found in Daubechies (1992). In Section 3.3.2, we explain how to obtain a noise-free signal thanks to wavelets and using especially the works of Antoniadis *et al.* (2001) and Coifman and Donoho (1995). Among important articles about this subject there are Donoho and Johnstone (1994, 1995, 1998, 1999) and Hall and Patil (1996a,b). The locations where a valvometric signal exhibits high variations are likely to correspond to microclosings or transitions between the open and the closed state. In Section 3.3.3, we propose a method which is linked with wavelets in order to find these locations.

3.3.1 The Haar wavelets

Walnut (2004, Definition 5.11), describes the Haar wavelets as functions $h_{j,k}$ defined for $(j,k) \in \mathbb{Z}^2$ and $x \in \mathbb{R}$ by

$$h_{j,k}(x) := 2^{j/2}\left(\mathbb{I}_{\left[\frac{k}{2^j},\frac{k+1/2}{2^j}\right[}(x) - \mathbb{I}_{\left[\frac{k+1/2}{2^j},\frac{k+1}{2^j}\right[}(x)\right),$$

where for a given subset $A \subset \mathbb{R}$, the indicator function $\mathbb{I}_A$ is equal to 1 when $x \in A$ and 0 when $x \notin A$. Let also define for all $(j,k) \in \mathbb{Z}^2$ the Haar scaling functions $p_{j,k}$ which satisfies

$$p_{j,k}(x) := 2^{j/2}\mathbb{I}_{\left[\frac{k}{2^j},\frac{k+1}{2^j}\right[}(x),$$

for every $x \in \mathbb{R}$. In this subsection, we only consider functions in $L^2([0,1])$, which leads us to study only $h_{j,k}$ and $p_{j,k}$ for $j \in \mathbb{N}$ and $k \in \mathbb{D}_j := \{0,\dots,2^j-1\}$. These functions are interesting because for a given $j_0 \in \mathbb{N}$, the indexed family $\mathbb{U}_{j_0} := \{p_{j_0,k_1}, h_{j,k_2}\}_{j\geq j_0, k_1\in\mathbb{D}_{j_0}, k_2\in\mathbb{D}_j}$ is a complete orthonormal system on $[0,1]$ (Walnut, 2004, Theorem 5.24). This means that $\mathbb{U}_{j_0}$ is an orthonormal basis of the functional space $L^2([0,1])$. We thus have $\forall j_0 \in \mathbb{N}$:

- $\forall G \in L^2([0,1]),\ \forall x \in [0,1]$,

$$G(x) = \sum_{h\in\mathbb{U}_{j_0}} \langle G,h\rangle\, h(x), \tag{3.15}$$

  where for a given $(G,h) \in L^2([0,1]) \times L^2([0,1])$, $\langle G,h\rangle := \int_{\mathbb{R}} G(x)h(x)dx$,

- $\forall (h_1,h_2) \in \mathbb{U}_{j_0}^2$, such that $h_1 \neq h_2$,

$$\langle h_1,h_2\rangle = 0,$$

- $\forall h \in \mathbb{U}_{j_0}$,

$$\langle h,h\rangle = \|h\|_2^2 = 1.$$

Equation (3.15) implies that G is fully characterized by $\mathbb{U}_{j_0}$ and by the set $\{\langle G, h\rangle\}_{h\in\mathbb{U}_{j_0}}$.

Assume now that we only have $y_k = G\left(\frac{k}{2^J}\right)$ for all $k = \mathbb{D}_J$ and a given $J \in \mathbb{N}^*$, at our disposal. Let $\widetilde{G}_J$ be the piecewise constant approximation of G defined for all $x \in [0,1]$ by:

$$\widetilde{G}_J(x) := \sum_{k\in\mathbb{D}_J} y_k \mathbb{I}_{\left[\frac{k}{2^J}, \frac{k+1}{2^J}\right[}(x).$$

If G is continuous, we have $\|G - \widetilde{G}_J\|_\infty \to 0$ and then $\|G - \widetilde{G}_J\|_2 \to 0$, when $n \to \infty$. Following Walnut (2004, Chapter 6), for all $h \in \{h_{j,k}, p_{j,k}\}_{j\in\mathbb{N},k\in\mathbb{D}_j}$,

$$\langle G, h\rangle - \left\langle \widetilde{G}_J, h\right\rangle = \left\langle G - \widetilde{G}_J, h\right\rangle \leq \|G - \widetilde{G}_J\|_2 \|h\|_2 = \|G - \widetilde{G}_J\|_2,$$

and then $\left\langle \widetilde{G}_J, h\right\rangle \to \langle G, h\rangle$ when $n \to \infty$. It thus makes sense to approximate $\langle G, h\rangle$ by $\left\langle \widetilde{G}_J, h\right\rangle$ for all $h \in \{h_{j,k}, p_{j,k}\}_{j\in\mathbb{N},k\in\mathbb{D}_j}$.

Computing these values of $\left\langle \widetilde{G}_J, h\right\rangle$ is easy. We indeed have for all $k \in \mathbb{D}_J$, $\left\langle \widetilde{G}_J, p_{J,k}\right\rangle = 2^{-J/2} y_k$. We also have for all $j > J$, for all $k_1 \in \mathbb{D}_J$ and for all $k_2 \in \{2^{j-J}k_1, \dots 2^{j-J}(k_1+1) - 1\}$, $\left\langle \widetilde{G}_J, p_{j,k_2}\right\rangle = 2^{-j/2} y_{k_1}$. In addition, since for all $x \in [0,1]$,

$$\widetilde{G}_J(x) = \sum_{k\in\mathbb{D}_J} \langle \widetilde{G}_J, p_{J,k}\rangle p_{J,k}(x),$$

and because $\mathbb{U}_J$ is a complete orthonormal system on $[0,1]$, we have $\left\langle \widetilde{G}_J, h_{j,k}\right\rangle = 0$, for all $j \geq J, k \in \mathbb{D}_j$. Only $\left\langle \widetilde{G}_J, p_{j,k}\right\rangle$ and $\left\langle \widetilde{G}_J, h_{j,k}\right\rangle$ for $j \in \{0, \dots, J-1\}$ and $k \in \mathbb{D}_j$ remain to be determined. Note that for all $j \in \mathbb{N}^*$ and for all $k \in \mathbb{D}_j$, we have:

$$\left\langle \widetilde{G}_J, p_{j-1,k}\right\rangle = \frac{1}{\sqrt{2}}\left\langle \widetilde{G}_J, p_{j,2k}\right\rangle + \frac{1}{\sqrt{2}}\left\langle \widetilde{G}_J, p_{j,2k+1}\right\rangle, \tag{3.16}$$

and

$$\left\langle \widetilde{G}_J, h_{j-1,k}\right\rangle = \frac{1}{\sqrt{2}}\left\langle \widetilde{G}_J, p_{j,2k}\right\rangle - \frac{1}{\sqrt{2}}\left\langle \widetilde{G}_J, p_{j,2k+1}\right\rangle, \tag{3.17}$$

as mentioned in Walnut (2004, p142). This allows us to compute recursively $\left\langle \widetilde{G}_J, p_{j,k}\right\rangle$ and $\left\langle \widetilde{G}_J, h_{j,k}\right\rangle$ for $j \in \{0, \dots J-1\}$ and $k \in \mathbb{D}_j$, using $\{\left\langle \widetilde{G}_J, p_{J,k}\right\rangle\}_{k\in\mathbb{D}_J}$.

Thanks to $\left\langle \widetilde{G}_J, h\right\rangle$ for $h \in \{h_{j,k}, p_{j,k}\}_{j\in\mathbb{N},k\in\mathbb{D}_j}$, we are now able to approximate the decomposition of G into the basis $\mathbb{U}_{j_0}$, for any $j_0 \in \mathbb{N}$. In addition, the quality of this approximation is linked to the closeness between G and $\widetilde{G}$ by the Plancherel formula (see for instance Mallat, 2008, Appendix A.3):

$$\|G - \widetilde{G}_J\|_2^2 = \sum_{h\in\mathbb{U}_{j_0}} \left(\langle G, h\rangle - \left\langle \widetilde{G}_J, h\right\rangle\right)^2. \tag{3.18}$$

Note however that it is not really informative to use the bases $\mathbb{U}_{j_0}$ for $j_0 \geq J$.

3.3.2 Denoising with wavelets

Let us now consider the following model for all $k \in \mathbb{D}_J$:

$$\tilde{y}_k = y_k + z_k, \tag{3.19}$$

where $\tilde{y}_0, \dots, \tilde{y}_{2^J-1}$ are available measures, $z_0, \dots, z_{2^J-1}$ are independently and identically distributed from the probability distribution $\mathcal{N}(0, \sigma^2)$ and σ^2 is an unknown positive real. For a given $j_0 \in \mathbb{N}$, we are interested in estimating $\langle G, h \rangle$ for all $h \in \mathbb{U}_{j_0}$. Although wavelet coefficients related to $\{z_k\}_{k \in \mathbb{D}_J}$ are often considered through a Gaussian white noise we study here the following functions $\widetilde{Z}_J$, for the sake of simplicity. They are defined for all $J \in \mathbb{N}^*$ and for all $x \in [0,1]$ by

$$\widetilde{Z}_J(x) := \sum_{k \in \mathbb{D}_J} z_k \mathbb{I}_{\left[\frac{k}{2^J}, \frac{k+1}{2^J}\right[}(x).$$

This definition and equations (3.16) and (3.17) imply that for all $j \in \mathbb{N}$, for all $k \in \mathbb{D}_j$,

$$\langle \widetilde{Z}_J, p_{j,k} \rangle \sim \mathcal{N}(0, 2^{-J}\sigma^2),$$

and that for all $j \in \{0, \dots, J-1\}$, for all $k \in \mathbb{D}_j$,

$$\langle \widetilde{Z}_J, h_{j,k} \rangle \sim \mathcal{N}(0, 2^{-J}\sigma^2).$$

Using wavelets on $\widetilde{G}_J + \widetilde{Z}_J$, it is thus possible to find an estimate $\widehat{G}_J$ of the function G from model (3.19). One may also control $\|\widehat{G}_J - G\|_2^2$ thanks to equation (3.18).

As explained in Antoniadis *et al.* (2001), it is expected that for a given $j_0 \in \{0 \dots, J\}$, only a few coefficients among $\{\langle \widetilde{G}_J, h \rangle\}_{h \in \mathbb{U}_{j_0}}$ are not null. For any $h_1 \in \mathbb{U}_{j_0}$ such that $\langle \widetilde{G}_J, h_1 \rangle = 0$ and any $h_2 \in \mathbb{U}_{j_0}$ such that $\langle \widetilde{G}_J, h_2 \rangle \neq 0$, if σ^2 is small enough, we have with a great probability that $|\langle \widetilde{G}_J + \widetilde{Z}_J, h_1 \rangle|$ is smaller than $|\langle \widetilde{G}_J + \widetilde{Z}_J, h_2 \rangle|$. It then makes sense for all $h \in \mathbb{U}_{j_0}$ to find a threshold $\delta_h > 0$ and to choose

$$\langle \widehat{G}_J, h \rangle = \mathrm{HT}(\langle \widetilde{G}_J + \widetilde{Z}_J, h \rangle) := \begin{cases} 0 & \text{if } |\langle \widetilde{G}_J + \widetilde{Z}_J, h \rangle| < \delta_h, \\ \langle \widetilde{G}_J + \widetilde{Z}_J, h \rangle & \text{otherwise.} \end{cases}$$

This approach is called hard thresholding. Soft thresholding is another popular method to estimate $\langle G, h \rangle$, for all $h \in \mathbb{U}_{j_0}$, it is defined by:

$$\langle \widehat{G}_J, h \rangle = \mathrm{ST}(\langle \widetilde{G}_J + \widetilde{Z}_J, h \rangle) := \begin{cases} 0 & \text{if } |\langle \widetilde{G}_J + \widetilde{Z}_J, h \rangle| < \delta_h, \\ \langle \widetilde{G}_J + \widetilde{Z}_J, h \rangle - \delta_h & \text{if } \langle \widetilde{G}_J + \widetilde{Z}_J, h \rangle > \delta_h, \\ \langle \widetilde{G}_J + \widetilde{Z}_J, h \rangle + \delta_h & \text{if } \langle \widetilde{G}_J + \widetilde{Z}_J, h \rangle < -\delta_h. \end{cases}$$

Both hard and soft thresholding provide then an estimate $\widehat{G}_J$ of G, which satisfies for all $x \in [0,1]$

$$\widehat{G}_J(x) = \sum_{h \in \mathbb{U}_{j_0}} \langle \widehat{G}_J, h \rangle h(x). \tag{3.20}$$

To use model (3.19) for valvometric data we have to decide which thresholding method we want to apply, choose $j_0 \in \mathbb{N}$ to define a basis $\mathbb{U}_{j_0}$, and set the values of δ_h for each $h \in \mathbb{U}_{j_0}$. Let $n := 2^J$. Various choices are available and have good asymptotic properties such as the SureShrink estimator (Donoho and Johnstone, 1995), the optimal thresholds from a MISE point of view (Hall and Patil, 1996a) or the simple so-called universal threshold defined for all $h \in \mathbb{U}_{j_0}$ by $\delta_h = \hat{\sigma}\sqrt{2\log(n)}/\sqrt{n}$, for a given estimator $\hat{\sigma}$ of σ (Donoho and Johnstone, 1994). Antoniadis *et al.* (2001) studied various wavelet thresholding estimators

in simulation to obtain insights about how they behave when n is finite. They especially consider the case where G is the *Bumps* function from Donoho and Johnstone (1994), which is a sum of functions $t \mapsto (1-|t|)^{-4}$ with various translation and scale parameters. This *Bumps* function seems to us to be the most interesting one among the functions Antoniadis *et al.* (2001) tested, because it looks like the valvometric signals we want to denoise. This is mostly due to the spikes this function exhibits. For this function, the best estimator among those considered by Antoniadis *et al.* (2001) appears to be the translation invariant hard thresholding estimator from Coifman and Donoho (1995). We will detail their approach in Section 3.3.3. Note that we set $j_0 = \lfloor \log(\log(n))/\log(2) \rfloor + 1$ and $\delta_h = \hat{\sigma}\sqrt{2\log(n\log(n)/\log(2))}/\sqrt{n}$, where $\lfloor t \rfloor$ is the integer part of t, for all $t \in \mathbb{R}$, like Antoniadis *et al.* (2001) did. We discuss a way to find $\hat{\sigma}$ in Remark 6.

3.3.3 Finding microclosings with wavelets

Assume that we measure a valvometric signal at each time $\frac{k}{2^J}$ for $k \in \mathbb{D}_J$. This can be done easily by choosing suitable starting time and unit of time. We can then use the model (3.19) for these data. Microclosings arise for values of t where $|G'(t)|$ is great. We also consider that changes between states open and close happen at such instants (see Section 3.1). Provided a way to detect these state changes, the only instants left when $|G'|$ is large correspond to the microclosings. We are then interested in studying G'.

Sow *et al.* (2011) estimated G' with a kernel-based estimator and did not have to assume that z_k comes from a Gaussian distribution, for $k \in \mathbb{D}_J$. Here, we would like to approximate values of $G'\left(\frac{k}{2^J}\right)$ by $2^J(y_{k+1} - y_k)$ for $k \in \mathbb{D}_J \backslash \{2^J - 1\}$ or by $2^J(y_k - y_{k-1})$ for $k \in \mathbb{D}_J \backslash \{0\}$. For all even k in $\mathbb{D}_J$, we have

$$2^J(y_{k+1} - y_k) = 2^{3J/2}(\langle \widetilde{G}_J, p_{J,k+1} \rangle - \langle \widetilde{G}_J, p_{J,k} \rangle) = -2^{(3J+1)/2}\langle \widetilde{G}_J, h_{J-1,k/2} \rangle,$$

because of equation (3.17). The same equation implies that for all odd k in $\mathbb{D}_J$

$$2^J(y_k - y_{k-1}) = 2^{3J/2}(\langle \widetilde{G}_J, p_{J,k} \rangle - \langle \widetilde{G}_J, p_{J,k-1} \rangle) = -2^{(3J+1)/2}\langle \widetilde{G}_J, h_{J-1,(k-1)/2} \rangle.$$

Using hard thresholding, and $\delta_h = \hat{\sigma}\sqrt{2\log(n\log(n)/\log(2))}/\sqrt{n}$, this leads us, for all $k \in \mathbb{D}_{J-1}$, to use $-2^{(3J+1)/2}\langle \widehat{G}_J, h_{J-1,k} \rangle$ as an estimator of both $G'\left(\frac{2k}{2^J}\right)$ and $G'\left(\frac{2k+1}{2^J}\right)$.

Let now assume that we observe $\{\mathring{y}_k := \tilde{y}_{k+1}\}_{k \in \mathbb{D}_J}$ instead of $\{\tilde{y}_k\}_{k \in \mathbb{D}_J}$. This allows us to study G' on $\left[2^{-J}, 1 + 2^{-J}\right]$ instead of on $[0,1]$ by translating the studied functions. We then define $\mathring{G}_J$ and $\mathring{Z}_J$ for all $x \in [0,1]$ by

$$\mathring{G}_J(x) := \sum_{k \in \mathbb{D}_J} y_{k+1} \mathbb{I}_{\left[\frac{k}{2^J}, \frac{k+1}{2^J}\right]} \quad \text{and} \quad \mathring{Z}_J(x) := \sum_{k \in \mathbb{D}_J} z_{k+1} \mathbb{I}_{\left[\frac{k}{2^J}, \frac{k+1}{2^J}\right]}.$$

For all $k \in \mathbb{D}_{J-1}$, let

$$\tilde{d}_k := \langle \widetilde{G}_J + \widetilde{Z}_J, h_{J-1,k} \rangle = -2^{-(J+1)/2}(\tilde{y}_{2k+1} - \tilde{y}_{2k}),$$

and

$$\mathring{d}_k := \langle \mathring{G}_J + \mathring{Z}_J, h_{J-1,k} \rangle = -2^{-(J+1)/2}(\mathring{y}_{2k+1} - \mathring{y}_{2k}).$$

We then have that $-2^{(3J+1)/2}\mathrm{HT}(\tilde{d}_k)$ is an estimate of $G'\left(\frac{2k}{2^J}\right)$ and $G'\left(\frac{2k+1}{2^J}\right)$, while we also have that $-2^{(3J+1)/2}\mathrm{HT}(\mathring{d}_k)$ is an estimate of $G'\left(\frac{2k+1}{2^J}\right)$ and $G'\left(\frac{2k+2}{2^J}\right)$. Both estimates can however take significantly different values.

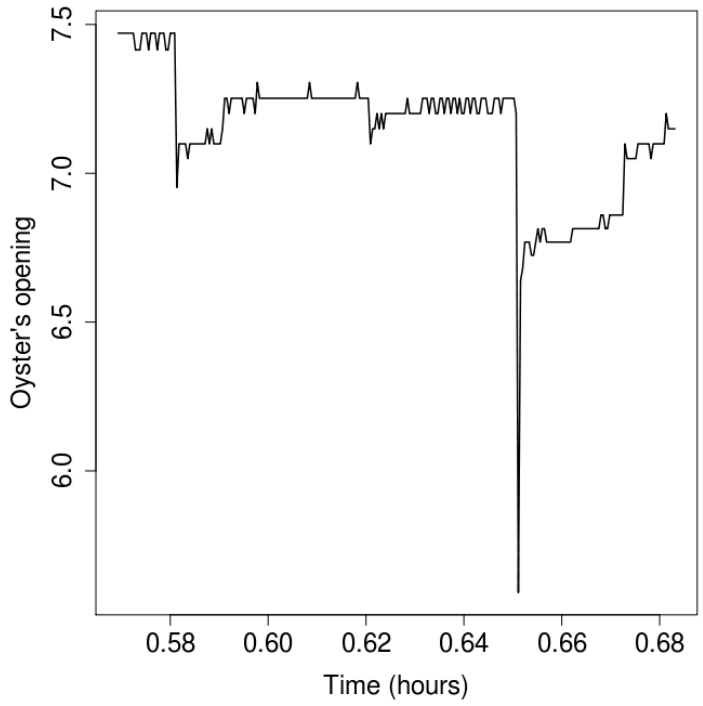


Figure 3.16: Valvometric measures from an oyster living at Eyrac on May 8, 2007 from 00:34:10 to 00:40:59.

To illustrate this idea, we analyse a valvometric signal of $2^J + 1 = 257$ points from an oyster from Eyrac, studied on May 8, 2007 from 00:34:10 to 00:40:59. We display it in Figure 3.16. We can observe a big spike near 0.58 h and another one close to 0.65 h. Smaller ones are located near 0.59 h, 0.62 h and 0.67 h. The corresponding values of $\{|\tilde{d}_k|\}_{k\in\mathbb{D}_{J-1}}$ and $\{|\mathring{d}_k|\}_{k\in\mathbb{D}_{J-1}}$ are plotted in Figure 3.17. They were computed using the *wavethresh* R package. The hard thresholding sets each value of $\{|\tilde{d}_k|\}_{k\in\mathbb{D}_{J-1}}$ and $\{|\mathring{d}_k|\}_{k\in\mathbb{D}_{J-1}}$ below the dashed line to 0. Both $\{\mathrm{HT}(\tilde{d}_k)\}_{k\in\mathbb{D}_{J-1}}$ and $\{\mathrm{HT}(\mathring{d}_k)\}_{k\in\mathbb{D}_{J-1}}$ lead to the detection of both main spikes of Figure 3.16. Only $\{\mathrm{HT}(\mathring{d}_k)\}_{k\in\mathbb{D}_{J-1}}$ manages to detect the second small spike and only $\{\mathrm{HT}(\tilde{d}_k)\}_{k\in\mathbb{D}_{J-1}}$ succeeds in finding the third one. It is then unclear whether or not we should consider that these spikes exist. For denoising purposes this is not an important issue since the inverse wavelet transforms (3.20) respectively computed from $\{\tilde{y}_k\}_{k\in\mathbb{D}_J}$ and $\{\mathring{y}_k\}_{k\in\mathbb{D}_J}$ are very similar. We plot them in Figure 3.18.

Notice that for $k \in \mathbb{D}_{J-1}$, $2^{J-1}(y_{2k+2} - y_{2k})$ is an approximation of $G'\left(\frac{2k+1}{2^J}\right)$ based on central difference. This encourages us to use $-2^{(3J-1)/2}(\mathrm{HT}(\tilde{d}_k) + \mathrm{HT}(\mathring{d}_k))$ as an estimator of $G'\left(\frac{2k+1}{2^J}\right)$. For the signal from Figure 3.16, we obtain the following estimate of $\left(G'\left(\frac{1}{2^J}\right), G'\left(\frac{3}{2^J}\right) \ldots, G'\left(\frac{2^J-1}{2^J}\right)\right)$

$$(\overbrace{0,\ldots,0}^{13}, -47.58, \overbrace{0,\ldots,0}^{43}, -19.60, \overbrace{0,\ldots,0}^{33}, -205.99, 134.03, \overbrace{0,\ldots,0}^{23}, 30.63, \overbrace{0,\ldots,0}^{11}).$$

This vector should be multiplied by $\frac{3600}{1.6n}$ to be compatible with the scale of Figure 3.16. For each detected spike, the corresponding value $-2^{(3J-1)/2}(\mathrm{HT}(\tilde{d}_k) + \mathrm{HT}(\mathring{d}_k))$ is then

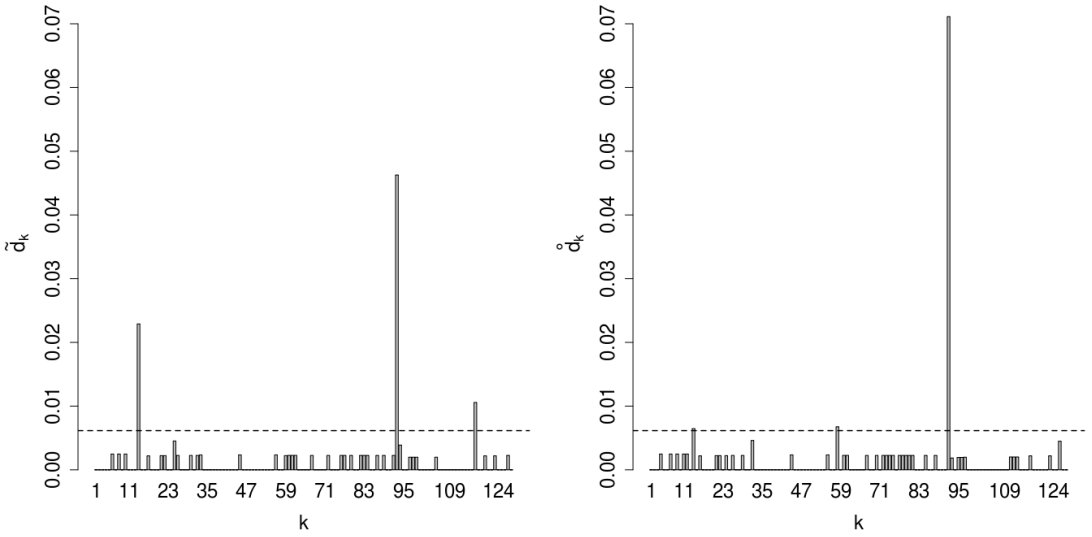


Figure 3.17: Wavelet coefficients computed from $\{\tilde{y}_k\}_{k\in\mathbb{D}_J}$ (left) and $\{\mathring{y}_k\}_{k\in\mathbb{D}_J}$ (right). Dashed line: threshold $\delta_h = \hat{\sigma}\sqrt{2\log(n\log(n)/\log(2))}/\sqrt{n}$.

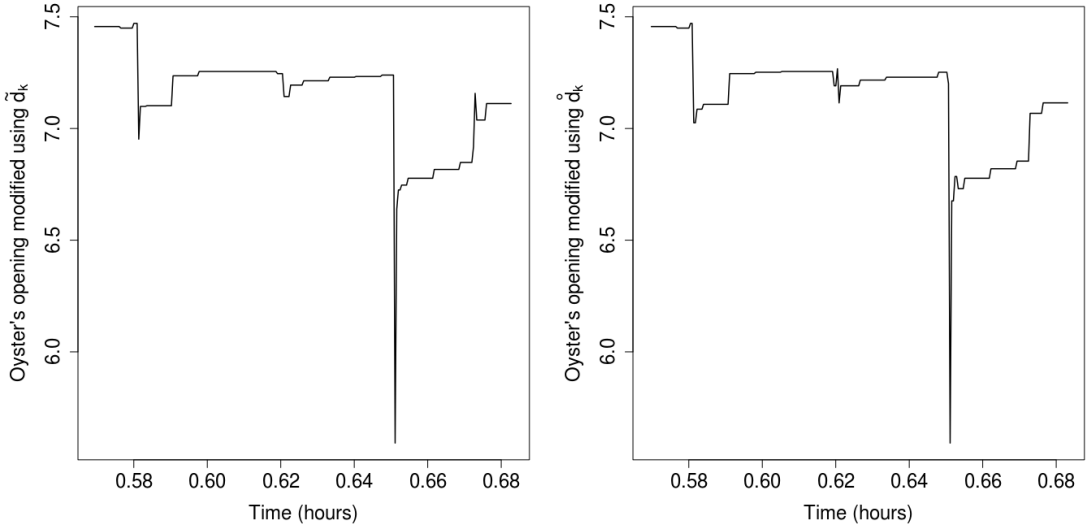


Figure 3.18: Valvometric signals reconstructed after hard thresholding of the wavelet coefficients of $\mathbb{U}_{j_0}$ computed from $\{\tilde{y}_k\}_{k\in\mathbb{D}_J}$ (left) and $\{\mathring{y}_k\}_{k\in\mathbb{D}_J}$ (right), using $j_0 = \lfloor\log(\log(n))/\log(2)\rfloor + 1$ and $\delta_h = \hat{\sigma}\sqrt{2\log(n\log(n)/\log(2))}/\sqrt{n}$.

coherent with a visual inspection of Figure 3.16.

We present now the approach from Coifman and Donoho (1995). For all $(k,l) \in \mathbb{D}_J^2$, let $\tilde{y}_{k,l}$ be defined by:

$$\tilde{y}_{k,l} := \tilde{y}_{(k+l)} \bmod 2^J,$$

where mod denotes the modulo operator. This means that $\{\tilde{y}_{k,l}\}_{k\in\mathbb{D}_J}$ and $\{\tilde{y}_k\}_{k\in\mathbb{D}_J}$ are linked through a circulant shift. Let also each element of $\{z_{k,l}\}_{k\in\mathbb{D}_J}$ satisfy

$$z_{k,l} := z_{(k+l)} \bmod 2^J,$$

and define $\widetilde{G}_{J,l}$ and $\widetilde{Z}_{J,l}$ from $\{\tilde{y}_{k,l}\}_{k\in\mathbb{D}_J}$ and $\{z_{k,l}\}_{k\in\mathbb{D}_J}$ like $\widetilde{G}_J$ and $\widetilde{Z}_J$ were defined from $\{\tilde{y}_k\}_{k\in\mathbb{D}_J}$ and $\{z_k\}_{k\in\mathbb{D}_J}$. we will also consider the elements of

$$\{d_{j,k,l}\}_{j\in\{j_0,\dots,J-1\},k\in\mathbb{D}_j,l\in\mathbb{D}_J} \quad \text{and} \quad \{h_{j,k}\}_{j\in\{j_0,\dots,J-1\},k\in\mathbb{D}_j}$$

where

$$d_{j,k,l} := \langle \widetilde{G}_{J,l} + \widetilde{Z}_{J,l}, h_{j,k} \rangle,$$

and

$$c_{k,l} := \langle \widetilde{G}_{J,l} + \widetilde{Z}_{J,l}, p_{j_0,k} \rangle.$$

One can then use a hard thresholding approach to compute the functions $\widehat{G}_{J,l}$ defined for all $x \in [0,1]$ by

$$\widehat{G}_{J,l}(x) := \sum_{k\in\mathbb{D}_{j_0}} \mathrm{HT}(c_{k,l}) p_{j_0,k}(x) + \sum_{j=j_0}^{J-1} \sum_{k\in\mathbb{D}_j} \mathrm{HT}(d_{j,k,l}) h_{j,k}(x).$$

Note that because of the function HT, for each $l \in \mathbb{D}_j$ $\widehat{G}_{J,l}$ changes with j_0. Coifman and Donoho (1995) finally propose the translation-invariant estimate $\widehat{G}_{\mathrm{TI},J}$ of G, which satisfies for all $x \in [0,1]$

$$\widehat{G}_{\mathrm{TI},J}(x) := \frac{1}{2^J} \sum_{l\in\mathbb{D}_J} \widehat{G}_{J,l}\left(\left(x - \frac{l}{2^J}\right) \bmod 1\right).$$

When $j_0 = J-1$ and $J \geq 1$, we have for all $x \in [0,1[$

$$\widehat{G}_{\mathrm{TI},J}(x) = \frac{1}{2}\left[\widehat{G}_{J,0}(x) + \widehat{G}_{J,1}\left(\left(x - \frac{1}{2^J}\right) \bmod 1\right)\right].$$

This is a straightforward consequence of the following property

Proposition 8. *Let $j_0 = J-1$ and $J \geq 1$. For all $(l_1,l_2) \in \mathbb{D}_J^2$ such that there exists $u \in \mathbb{D}_{J-1}$ which satisfies $l_1 - l_2 = 2u$, and for every $x \in [0,1[$, we have*

$$\widehat{G}_{J,l_1}\left(\left(x - \frac{l_1}{2^J}\right) \bmod 1\right) = \widehat{G}_{J,l_2}\left(\left(x - \frac{l_2}{2^J}\right) \bmod 1\right).$$

Proof. This proposition results mainly from the following points:

- for all $k \in \mathbb{D}_{J-1} \backslash \{2^{J-1}-1\}$, for all $x \in \left[\frac{k}{2^{J-1}}, \frac{k+1}{2^{J-1}}\right[$,

$$h_{J-1,k}(x) = h_{J-1,k+1}\left(x + \frac{1}{2^{J-1}}\right) \quad \text{and} \quad p_{J-1,k}(x) = p_{J-1,k+1}\left(x + \frac{1}{2^{J-1}}\right)$$

- for all $k \in \mathbb{D}_{J-1}$, for all $x \in \left[\frac{k}{2^{J-1}}, \frac{k+1}{2^{J-1}}\right[$, for all $l \in \mathbb{D}_J$,

$$\widehat{G}_{J,l}(x) = \mathrm{HT}(c_{k,l}) p_{J-1,k}(x) + \mathrm{HT}(d_{J-1,k,l}) h_{J-1,k}(x)$$

□

In addition, we have for all $k \in \{0, \ldots, 2^{J-1}-2\}$ and for all $x \in \left[\frac{k}{2^{J-1}} + \frac{1}{2^J}, \frac{k+1}{2^{J-1}}\right[$

$$\widehat{G}_{J,0}(x) = 2^{(J-1)/2}\left[\mathrm{HT}(c_{k,0}) - \mathrm{HT}(d_{J-1,k,0})\right]$$

and

$$\widehat{G}_{J,1}\left(\left(x - \frac{1}{2^J}\right) \bmod 1\right) = 2^{(J-1)/2}\left[\mathrm{HT}(c_{k,1}) + \mathrm{HT}(d_{J-1,k,1})\right].$$

By recalling the definition of $\mathring{d}_k$ and $\tilde{d}_k$ for all $k \in \{0, \ldots, 2^{J-1}-2\}$, we obtain for all $x \in \left[\frac{k}{2^{J-1}} + \frac{1}{2^J}, \frac{k+1}{2^{J-1}}\right[$

$$\widehat{G}_{\mathrm{TI},J}(x) = 2^{(J-3)/2}\left[\mathrm{HT}(c_{k,0}) + \mathrm{HT}(c_{k,1}) + \mathrm{HT}(\mathring{d}_k) - \mathrm{HT}(\tilde{d}_k)\right].$$

This is coherent with Coifman and Donoho's will to remove spurious spikes from the estimate of G while we use $-2^{(3J-1)/2}(\mathrm{HT}(\tilde{d}_k) + \mathrm{HT}(\mathring{d}_k))$ to detect where each spike is located, whether it is spurious or not.

Remark 6. The translation-invariant thresholding from Coifman and Donoho (1995) and our method to find spikes both require an estimate $\hat{\sigma}$ of σ because they are based on a threshold δ_h which depends on $\hat{\sigma}$. We assume that the valvometric signal is constant when the oyster is closed for a sufficiently large time. It is then possible to estimate σ with the empirical standard deviation $\tilde{\sigma}$ measured during these periods, but two other problems have to be tackled: for all $k \in \mathbb{D}_J$, $\tilde{y}_k$ only takes a finite number of distinct values and besides, the possible electronic noise is more amplified for high values of the signal than for low values. We then consider that for all $k \in \mathbb{D}_J$, z_k follows a binomial distribution with an appropriate scale. For a portion of a valvometric signal recorded when the oyster is closed, let $a := \min_{k_1 \in \mathbb{D}_J, k_2 \in \mathbb{D}_J}\{|\tilde{y}_{k_1} - \tilde{y}_{k_2}| \ : \ \tilde{y}_{k_1} \neq \tilde{y}_{k_2}\}$. For another part of this signal during which the oyster is open, let $b := \max_{k \in \mathbb{D}_J}\{\tilde{y}_k\}$ and $c := b - \max_{k \in \mathbb{D}_J}\{\tilde{y}_k \ : \ \tilde{y}_k \neq b\}$. We then set $\hat{\sigma} = \tilde{\sigma}\frac{c}{a}$.

We apply our procedure to find spikes in the signal of Figure 2.1. Because we need a sample size n such that it exists a positive integer J satisfying $n = 2^J + 1$, we focus on the first 32,769 points of this signal. To estimate σ, we consider the valvometric data between 8.1 and 10.5 hours and the procedure in Remark 6. For 5,000 points within the studied sample, the results of our detection method is shown in Figure 3.19. If the absolute value of a wavelet coefficient is greater than the threshold δ_h, we display it with a cross. We observe that every cross is positioned in the beginning of an important variation of the signal. If a spike is made of a sequence of several great variations, several crosses are then placed on it. Further work should thus be done in order to determine whether or not two consecutive crosses represent the same spike. We could then automatically find the location and the amplitude of each microclosing.

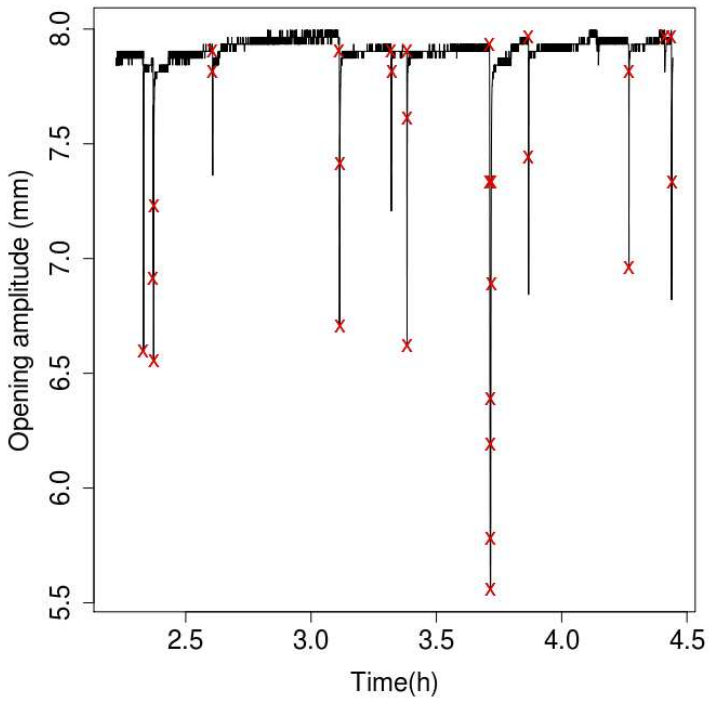


Figure 3.19: Microclosing detection using the first 32,769 points of Figure 2.1. Crosses are locations where they are detected.

3.4 Implementation

In this section, we describe how the various analyses presented in this chapter were made functional in practice. The main software we use is R. We present it in Section 3.4.1. To handle the valvometric data we studied, we stored them in an SQL database. We explain how to import them into R in Section 3.4.2. Because everyone is not comfortable with the R command-line interface, we also provide a procedure to launch R codes with a graphical user interface in Section 3.4.3.

3.4.1 The R software

Overview. R is both a mathematical software and a programming language. It is a major tool for data analysis, data visualisation and stochastic modelling. It is a free software under General Public License (GPL) and is available for Linux, Mac OS X and Windows. It is based on packages built by the community. A lot of packages and their documentation are stored in the Comprehensive R Archive Network (CRAN, `http://cran.r-project.org/`). Other package repositories and bundles exist such as Bioconductor (`http://www.bioconductor.org/`) which contains statistical procedures for biological data analysis. Among commercial alternatives, SAS, Stata and SPSS are commonly chosen by firms.

While the CRAN released informative manuals for both beginners and advanced users, they may find interesting information in books such as Teetor (2011) for quick answers to simple questions and details about tricky pitfalls, or Adler (2010) for a more extensive exploration of R. French speakers may also be interested in Lafaye De Micheaux *et al.* (2010).

A few notable features. R has many assets we did not need to program our methods. Some of them still deserve to be highlighted.

The `data.frame` objects are tables and can contain both strings and real numbers, among other kind of variables. The `read.table` function creates such an object from some data saved in a text file. Let `my.d.f` be a `data.frame` with two lines and three columns. The command `dimnames(my.d.f)[[1]] <- c("Fst.line", "Snd.line")` changes the names of the lines while

```
dimnames(my.d.f)[[2]] <- c("Fst.col.", "Snd.col.", "Trd.col.")
```

does the same for the columns. Then, `my.d.f[1,]` and `my.d.f["Fst.line",]` both return the first line of `my.d.f`, while the commands in Algorithm 3.1 display its second column.

The R language includes a framework to program in an object-oriented way. This framework may be a little difficult to apprehend because, as Adler (2010, Chapter 10) point it out, there exist two different syntaxes to design objects. While new programs should be written with the most recent one (S4), it might be more appropriate to present here the slightly more simple and older one (S3). Using the latter syntax, Algorithm 3.2 creates a function `my.function` which behaves differently whether its parameter is a real or a character string. Usual functions like `print`, `plot` and `summary` were implemented with S3. If `my.function` is an S4 generic function, `showMethods("my.function")` displays

```
> my.d.f[,2]

> my.d.f[,"Snd.col."]

> my.d.f$Snd.col.

> attach(my.d.f)
> Snd.col.
```

Algorithm 3.1: Different methods to obtain the second column of a `data.frame`

```
> my.function.double <- function(x) { x^2 }
> my.function.character <- function(a) { rep(a,2) }
> my.function <- function(object){UseMethod("my.function")}
> my.function(3)
[1] 9
> my.function("test")
[1] "test" "test"
```

Algorithm 3.2: An example of object-oriented programming using S3

the classes of objects for which the method `my.function` exists. It returns an explicit message otherwise.

R provides tools to visualise multivariate data with static graphics. This is not an easy task when the dimension of the data is greater than 2. The `lattice` package gathers some of these tools. Algorithm 3.3 underlies the usefulness of this package. This code creates three samples of $n = 1000$ pairs (X, Y) which satisfies

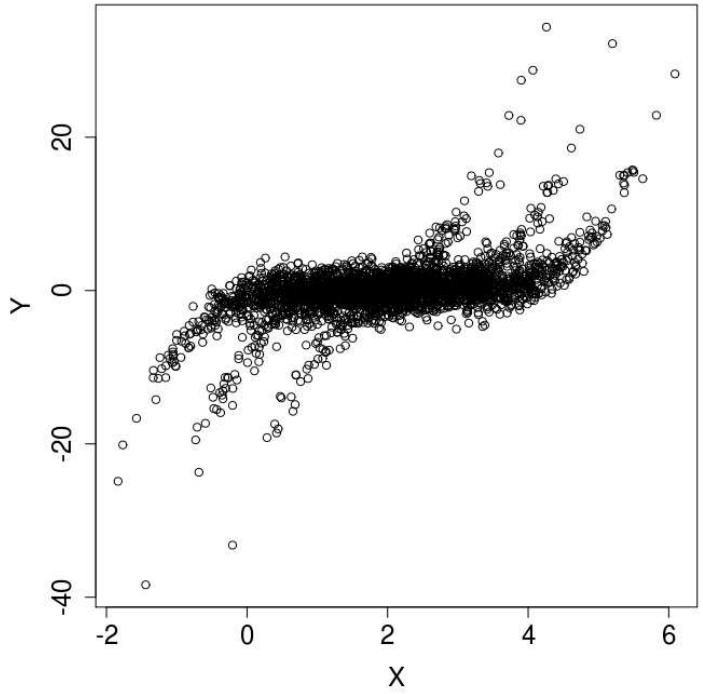
$$\begin{cases} Y = (X - \mu)^3 + \varepsilon, \\ X \sim \mathcal{N}(\mu, 1), \\ \varepsilon \sim \mathcal{N}(0, 2.25). \end{cases} \tag{3.21}$$

Each sample corresponds to a value of μ in $\{1, 2, 3\}$. The random variables X and ε are independent. When the whole samples are pooled, the plot of Y versus X is cluttered and discerning how these variables are linked is difficult (Figure 3.20(a)). The function `xyplot` can plot separately each sample using the pooled one, if the underlying values of μ are known (Figure 3.20(b)). The function that links Y and X appears then clearly. For each subfigure in Figure 3.20(b), the selected value of μ is given by the grey mark in the strip above it.

3.4.2 How to access MySQL databases with R

Databases are efficient structures to gather various data related to a project. They allow their users to easily retrieve a part of their dataset, using a particular language. SQL (for Structured Query Language) is one of them. The MySQL software is a bundle of tools built for importing data into a database and manipulating them with the SQL language.

(a) Plot of Y versus X for the whole samples



(b) Plots of Y versus X for each sample corresponding to a single value of μ

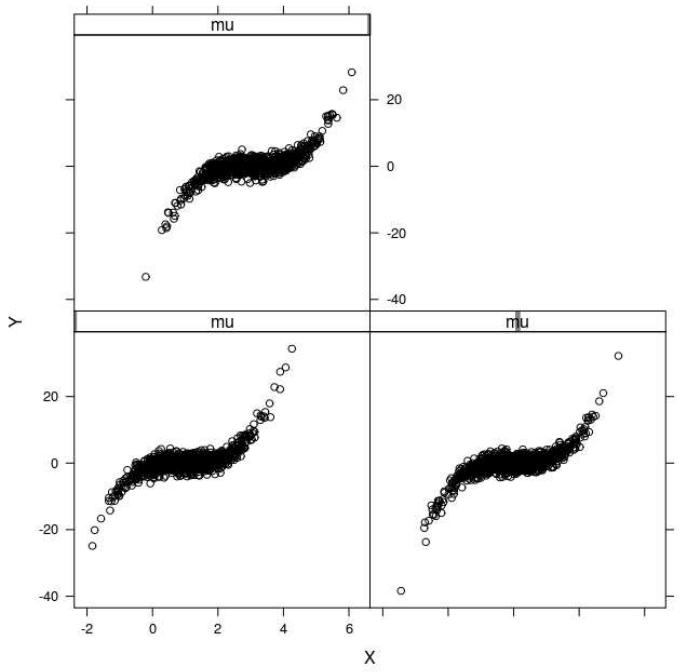


Figure 3.20: Samples of pairs (X, Y) from model (3.21)

Such actions can be remotely done using the Internet. The datasets are then stored in a MySQL server. A MySQL client can be used to access it. If MySQL is included in a commercial product, it is associated with a proprietary license. Otherwise, it is under GPL (General Public License). PostgreSQL and SQLite are fully free alternatives of MySQL. MySQL can be downloaded from http://dev.mysql.com/downloads/mysql/.

From text files to an SQL database. Many tutorials about MySQL are available on-line, we then only focus here on a few important commands. Algorithm 3.4 allows Linux users to connect to a MySQL server from a command-line interface. In this line of code, `0.0.0.0` has to be replaced with the IP address of the MySQL server and `user.name` should be the login the network administrator in charge of the server gave to the user. A password was probably assigned to the latter and he should enter it after typing Algorithm 3.4. Once someone is connected, he can create a new database called `myDb` using Algorithm 3.5. SQL databases are basically groups of tables. Algorithm 3.6 creates a table `myTable` with three columns (`intCol`, `stringCol` and `realCol`) that can respectively contain integers, character strings and real numbers. Thanks to Algorithm 3.7, we can thus transfer data to the database from a text file `myTable.txt` which contains three columns with the appropriate types of variable. Note that the option `--local` can be added if MySQL does not have sufficient permissions to handle the file `myTable.txt`.

The `mysqlimport` program allows us to import valvometric data from text files into a MySQL server every day. This is automatically done with the `cron` program which can be managed with `crontab`. We named our database `dbValvo` and the table that contains the valvometric measures is called `tableValvo`. The column names of this table can be found using Algorithm 3.8. The French labels `lieu`, `animal`, `jour`, `heure` and `mesure` respectively mean "location", "animal", "day", "hour" and "measure".

The `RMySQL` package Accessing the MySQL command-line interface is not the only way to execute SQL commands. This can also be done with the `mysql` program and the option `-e`. In Algorithm 3.9, `file.txt` receives the result of the SQL command `SQL.command`. For example, to obtain the valvometric signal of the first oyster of the Eyrac pier, on May 1, 2007, we use Algorithm 3.10.

Analysing data from an SQL database with R and the previous method may be a burden since it requires a text file. Hopefully, the `RMySQL` package can directly import data from a MySQL server into a `data.frame`, in a R session. Algorithm 3.11 is a framework to do so. In this algorithm, `0.0.0.0` should be the IP address of the MySQL server, `user.name` and `user.pwd` should be valid login password and `SQL.command` should be an SQL statement.

Installing `RMySQL` can be quite difficult for Windows users. We provide here a method that worked for a PC with Windows 7 and `RMySQL` 0.9-3.

- Install R, MySQL, RTools and the `DBI` package.
- Download `RMySQL` source files.
- Extract the `.tar` archive from the `.tar.gz` archive and then extract the files from the `.tar` archive into a `RMySQL` folder.
- Open the Command Prompt.

```
> n <- 1000
> nMu <- 3
> mu <- rep(1:nMu, n)
> X <- rnorm(n * nMu, mu)
> Y <- (X-mu)^3 + rnorm(n * nMu, 0, 1.5)
> par(mar=c(5,5,1,1))
> plot(X,Y,cex.axis=1.5, cex.lab=1.5)
<Figure 3.20(a)>
> trellis.par.set("strip.shingle",
+ list(col=rgb(0.5,0.5,0.5)))
> trellis.par.set("strip.background",list(col=0))
> xyplot(Y~X|mu, col=1)
<Figure 3.20(b)>
```

Algorithm 3.3: Scatter plots with and without the `lattice` package

```
$ mysql -h 0.0.0.0 -u user.name -p
```

Algorithm 3.4: Reaching a MySQL server with `mysql`

```
mysql> CREATE DATABASE myDb;
```

Algorithm 3.5: How to create an SQL database

```
mysql> CREATE TABLE myDb.myTable(intCol INT,
    -> stringCol VARCHAR(30), realCol FLOAT);
```

Algorithm 3.6: How to create an SQL table

```
$ mysqlimport -u user.name -p -h 0.0.0.0 myDb myTable.txt
```

Algorithm 3.7: Data from `myTable.txt` are stored in the database

```
mysql> SHOW COLUMNS FROM dbValvo.tableValvo;
+--------+-------------+------+-----+---------+-------+
| Field  | Type        | Null | Key | Default | Extra |
+--------+-------------+------+-----+---------+-------+
| lieu   | varchar(50) | YES  | MUL | Eyrac   |       |
| animal | tinyint(4)  | YES  |     | NULL    |       |
| jour   | date        | YES  |     | NULL    |       |
| heure  | double      | YES  |     | NULL    |       |
| mesure | double      | YES  |     | NULL    |       |
+--------+-------------+------+-----+---------+-------+
5 rows in set (0.00 sec)
```

Algorithm 3.8: Displaying the column names of a table

- Enter: `SET CYGWIN=nodosfilewarning`.
- Go to the folder where the `RMySQL` folder is, using the command `cd` which is available since RTools is installed.
- If `R_folder` is the path to the folder where R is installed, enter `"R_folder\bin\R.exe" CMD build RMySQL` to build the `RMySQL` package.
- Click on "Start Menu", "Control Panel", "System", "Advanced System Settings" and "Environment Variables".
- Edit the system variable `PATH` by adding `c:\Rtools\gcc-x.y.z\bin;` at the beginning. Use you file explorer to find the appropriate values of `x`, `y` and `z`.
- In a new system variable `MYSQL_HOME`, write the path to the folder where MySQL is installed, like it should be in Linux (for example `C:/Program Files/MySQL/MySQL Server 5.5`).
- In the folder where MySQL is installed, find the `lib` folder and create an `opt` folder in it. Move the file `libmysql.lib` from `lib` to `opt`.
- Back in the Command Prompt, go where the `RMySQL` folder is.
- Enter: `"R_folder\bin\R.exe" CMD INSTALL RMySQL_x-y-z.tar.gz`. Use you file explorer to find the appropriate values of `x`, `y` and `z`.
- Launch R and type `library(RMySQL)`.

Remark 7. We created an index on the table `dbValvo.tableValvo` in order for the server to return fast answers when an appropriate `WHERE` clause is provided. The index sort the rows of `dbValvo.tableValvo` following the column `lieu`, and then the column `animal` and finally `jour`. This means that requesting every measure from a given location and a given animal may be faster than asking for every measure from a given location and a given day. More details about the indexes can be found in the documentation of MySQL.

3.4.3 How to link R with a graphical user interface

Using Algorithm 3.11 implies writing the password in the R console but graphical objects that hide the entered content exist. Some other programs might need a dynamic adjustment to perform well. This can also be realized with a graphical user interface (GUI) as in Mangiarotti *et al.* (2012). Such interfaces can finally be designed to make user-friendly functions.

GTK+ is a set of tools for creating GUIs. It can be linked with R with the `RGtk2` package. GUIs can be designed with the Glade software. We recommend to first install Glade, which will also install GTK+ and then to download `RGtk2`. Glade allows developers to save their GUI in a `.xml` format. An example is given in Algorithm 3.12. This file describes a GUI with a window called `myWindow` and a button named `myButton`.

Algorithm 3.13 shows how to load this GUI with R. The functions `gtkWindowGetType()` and `gtkButtonGetType()` initialize the corresponding types of object. The `.xml` is imported into the variable `builder`. The function `getObject()` allows us to reach

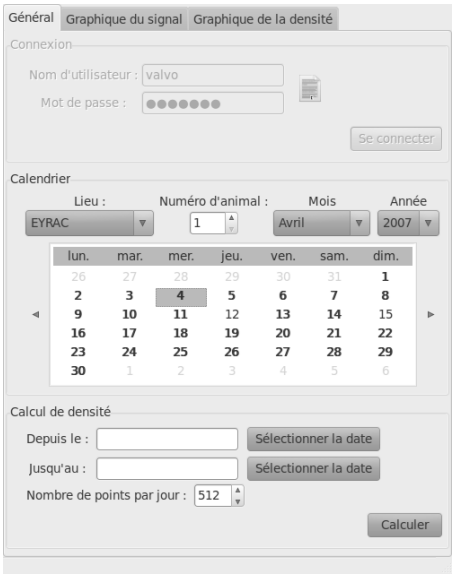


Figure 3.21: We designed a graphical user interface to analyse valvometric data

each graphical object. For a given kind of object, we can then call the functions defined by GTK+, like `show()`. See the following documentation for more information: https://developer.gnome.org/gtk2/stable/. The function `gSignalConnect` should be considered to execute R code when a graphical event arise. In Algorithm 3.13, when `myButton` is clicked, a new window appears.

For the valvometric data of this chapter, we created the GUI shown in Figure 3.21. A user can enter its login and password to connect to the MySQL database. A calendar then becomes visible. After selecting a location, a month and an animal, he can display a signal by clicking on the corresponding day in the calendar. In another tab, he can browse this signal and remove its noise (see Section 3.3). He can also draw estimates $\widetilde{f}_n$ of the density f introduced in Section 3.1.4.1 for a given oyster and several days. He is then able to produce a 3-dimensional figure (day vs oyster's opening vs value of the estimates), using the `rgl` package. For an oyster of Eyrac, from May 2007 to August 2007, an example of such a plot is given in Figure 3.22. In this figure, the location of the mode of $\widetilde{f}_n$ which corresponds to the lowest part of the valvometric signal increases with time. This illustrates the growth of the animal.

3.4.4 Fast computation with R

Because R is an interpreted language, it is computational time consuming. Using matrix calculus instead of loops produces faster programs. For large project development, calling libraries written in other languages like C or Fortran is recommended. The `dyn.load` R function loads such libraries. We can access functions in a library thanks to `.Call`. For example, the `density` function computes kernel density estimators using C code. The QZ algorithm presented in Section 4.4.2.3 can be reached by `.Call` from R.

In a command-line interpreter, `R CMD BATCH` can be executed to obtain the results of some R code without launching R. It is then easy to call R from a C program. We can take advantage of computer clusters using the C language. With this method, we then can do it with R too. An alternative is given by the `doMPI` package.

```
$ mysql -u user.name -p -h 0.0.0.0 -e "SQL.command" > file.txt
```

Algorithm 3.9: How to execute an SQL command with `mysql`

```
mysql> SELECT heure, mesure FROM dbValvo.tableValvo
    -> WHERE lieu='EYRAC' AND animal=1
    -> AND jour='2007-05-01';
```

Algorithm 3.10: An SQL command to extract valvometric data from `dbValvo.tableValvo`

```
> library(RMySQL)
> con <- dbConnect(MySQL(), host="0.0.0.0",
+ user="user.name", password="user.pwd", dbname="dbEels")
> output <- dbSendQuery(con, "SQL.command")
> result <- fetch(output, n=-1)
> print(result)
...
> dbDisconnect(con)
```

Algorithm 3.11: Sending an SQL command from R

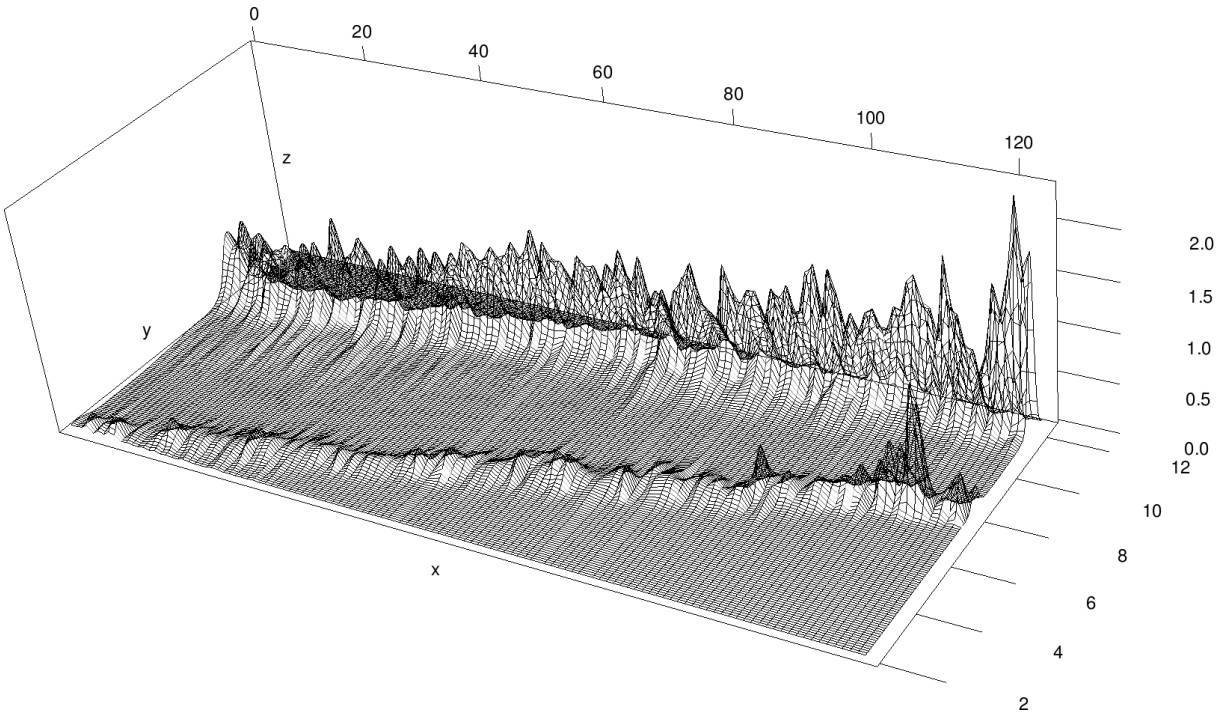


Figure 3.22: Estimates of the probability density f for an oyster at the Eyrac pier from May 1, 2007 to August 31, 2007; `x`: days, `y`: distance between shell parts, `z`: estimated density

```
<?xml version="1.0"?>
<interface>
 <requires lib="gtk+" version="2.16"/>
 <object class="GtkWindow" id="myWindow">
  <child>
   <object class="GtkHButtonBox" id="hbuttonbox1">
    <property name="visible">True</property>
    <child>
     <object class="GtkButton" id="myButton">
      <property name="label" translatable="yes">
      Create a new window</property>
      <property name="visible">True</property>
      <property name="can_focus">True</property>
      <property name="receives_default">True</property>
     </object>
     <packing>
      <property name="expand">False</property>
      <property name="fill">False</property>
      <property name="position">0</property>
     </packing>
    </child>
   </object>
  </child>
 </object>
</interface>
```

Algorithm 3.12: File `ex.xml` which contains characteristics of a graphical interface

```
> library(RGtk2)
> gtkWindowGetType()
> gtkButtonGetType()
> builder <- gtkBuilder()
> builder$addFromFile("ex.xml")
> builder$connectSignals(NULL)
> myButton <- builder$getObject("myButton")
> myWindow <- builder$getObject("myWindow")
> myWindow$show()
>
> gSignalConnect(myButton, "clicked", function(btn){
>   myWindow2 <- gtkWindow()
> })
```

Algorithm 3.13: A window which can create new windows

Chapter 4

Statistical methods to analyse genetic data

Among datasets that help to understand how genetics affect biological observations we will mostly consider gene expressions in this chapter. More specifically, we will study which ribonucleic acid (RNA) sequences are present in a given cell. Such a science is called transcriptomics, because deoxyribonucleic acid (DNA) is said to be transcribed into RNA. While DNA contains information to build proteins, only a part of it is considered at any given moment. A gene is a portion of the DNA. For two distinct genes, we obtain two distinct RNA sequences. An expressed gene is then a gene which corresponds to a RNA sequence we observe.

In Section 4.1, we detail two examples in which gene expressions are involved. Focusing on the second one, we describe in Section 4.2 a method to link gene expressions to an explanatory variable, using linear functions and multiple tests. We propose a Multivariate SIR (MSIR) method based on Sliced Inverse Regression (SIR), which can detect non-linear relationships between two sets of variables, in Section 4.3. We apply it to pupil rating and to a remote sensing dataset, although the MSIR method seems to be able to handle transcriptomic datasets. In Section 4.4 we finally tackle the case when the sample size is smaller than the number of explanatory variables by comparing four improvements of SIR built to solve this problem.

4.1 Challenges

4.1.1 Finding expression Quantitative Trait Loci (eQTL)

Concerning diseases which directly result from precise alleles of genes, an important goal is to locate the parts of the genome which are implicated. These illnesses are traits and they are called quantitative traits if they can be quantitatively measured. The portions of DNA which affect the values of these traits are then Quantitative Trait Loci (QTL). Other diseases may occur when a set of genes is too much or too little expressed. See Schadt *et al.* (2003) and references therein for examples. While knowing which are these genes is informative, finding those which control how the first ones are expressed is also of particular interest. Such genomic regions are named expression Quantitative Trait Loci (eQTL).

In Section 4.4.5, we try to find eQTL for the `Hopx` gene in rats. According to WikiGenes (`http://www.wikigenes.org`) and to Yamaguchi *et al.* (2009), `Hopx` is often called `HOP`, `LAGY` or `NECC1`. This gene allows an organism to create the HOP protein which may be needed for its heart to develop (see Shin *et al.*, 2002). Variations in the expression of this gene can be linked with various types of cancer as in Chen *et al.* (2003). That is why, finding eQTL for the `Hopx` gene seems to be a significant step toward the creation of a therapy for these diseases.

4.1.2 Selecting genes to build an RNA microarray

While in the previous subsection, we focused on finding relevant explanatory variables according to the expression of a given gene, we would like now to keep only genes which exhibit variations in their expression because of some explanatory variable. For example, this need can arise when we have at our disposal gene expressions which come from a DNA sequencing procedure. This technique gives information about every expressed gene of an individual. Because DNA sequencing is an expensive technology, we could prefer microarrays, which are cheaper. Such devices can only measure the expression of a limited number of genes. When the DNA sequencing produces data for a lot of different genes, a selection should then be made.

In the IMMORTEEL project which is supported by both the French ANR and the Canadian CRSNG and is coordinated by Magalie Baudrimont and Patrice Couture, the biologists are facing this problem. They would like to characterise how various pollutants affect eels which spend time either in Canadian rivers or in French ones. For this purpose, gene expressions were measured from these animals, using DNA sequencing. Biologists found 17,866 expressed genes. In her work to obtain a PhD, Lucie Baillon, supervised by Fabien Pierron, would like to build a microarray with only 1,000 genes. This selection should be realized by taking some explanatory variables into account.

These variables aim either at determining how grown the animal is, characterising normal variation of its environment or measuring how much the eel is subject to pollutants. We are especially interested in the last group of variables. For each pollutant, we would like to select genes that are differently expressed whether we detect it in the animal or not. Such genes should correspond to a single pollutant. We thus be able to detect its presence using only gene expressions. The variables which are not related to any pollutant are included in the study in order to select some gene expressions that do not

vary according to the standard evolution of the eel or of its environment, but only with a single pollutant. Concerning the growth of the animal we have

- the proportion Lip of lipid in muscles,
- the length L of the animal,
- the relative condition factor (Kn) which indicates how much a fish is heavy with respect to its length,
- the hepatosomatic index (HSI) which is defined as the ratio between the liver weight and the total weight,
- the splenosomatic index (SSI) which is the spleen weight divided by the total weight.

The following variables were measured in the fishing sites:

- the salinity S of the water in permille,
- the temperature T of the water in ℃.

The pollutants taken into account are enumerated in the next list. Metals and arsenic are studied through their weight by gram of dry kidney. The other pollutants are measured in weight unit by gram of muscle.

- Cadmium, copper, zinc, silver, arsenic, lead, chromium, nickel and mercury respectively correspond to the variables Cd, Cu, Zn, Ag, As, Pb, Cr, Ni and Hg,
- several polychlorinated biphenyls (PCBs) are merged in the variable PCB,
- a dichlorodiphenyldichloroethane (DDD) produces the variable DDD,
- a dichlorodiphenyldichloroethylene (DDE) and dieldrin form the variable DDE_Diel,
- another DDE leads to the variable DDE,
- a dichlorodiphenyltrichloroethane (DDT) provides the variable DDT,
- the variable Ld is built from lindane,
- the variable $PBDE$ stands for a tetrabromodiphenyl ether called PBDE-47,
- the variable HCB represents hexachlorobenzene (HCB),
- the variable Tnc symbolizes trans-nonachlor.

The considered metals and arsenic can be gathered together in the heavy metal group. This family of elements does not have any consensual definition. It generally contains metals and metalloids that are toxic for some living beings. Cadmium, for example, is responsible for renal problems as explained in Bernard (2008) or in Järup *et al.* (1998). Although too much copper, zinc or chromium can be harmful for humans, they are also essential for their organism to work properly. Ingestion of lead can cause saturnism. Mercury provokes hydrargyria. The toxicity of a compound can be measured using the LD_{50} indicator which is the quantity such that half a group of animals will die if they

have to handle it. It is measured in weight of the compound per unit of mass of the animal. Hughes (2002, Table 1) gives the LD_{50} for several molecules that contain arsenic. For example, arsenic trioxide (As_2O_3) has an LD_{50} of 26 mg/kg for rats exposed by oral route. Silver is not very dangerous for human but the possible effect it could have on aquatic animals is still examined (Ratte, 1999; Call *et al.*, 1999). Nickel is studied for its possible carcinogen behaviour and for the inflammations of the skin it may induce. Some toxic metals are not easily excreted by living beings and can stay a long time in their organism. If an animal lives in an environment which contains a low concentration of such a metal, this element can thus affect the animal after a sufficient long time. This is called bioaccumulation. This phenomenon is for example studied in Casas (2005) for mercury, cadmium, lead, copper and zinc and in Durrieu *et al.* (2005) for mercury. PCBs are toxic molecules which could formerly be found in electric devices such as transformers. DDT, dieldrin and lindane are very dangerous insecticides. Living beings can change the first one into DDE and DDD. These molecules are then called metabolites. They are toxic too. HCB is a fungicide. PBDE-47 was used to build fireproof objects. Like PCBs, DDT, dieldrin and lindane, tetrabromodiphenyl ethers and HCB are listed as persistent organic pollutants by a United Nations treaty. Chlordane is another insecticide which belongs to this list. It is a mixture of several components and trans-nonachlor is one of them.

4.2 A regression model with factors

4.2.1 Testing hypotheses with the False Discovery Rate

Assume that we want to study the expression of q genes and that we have n measures for each of them. For $j \in \{1, \dots, q\}$, let $\mathbf{y}^{(j)}$ be the random vector of length n which produces the expressions of the j^{th} gene. Let $\mathbf{X} = (X_1, \dots, X_p)'$ be the $p \times n$ matrix made of n values of each of the p explanatory variables. We suppose that $n > p$. For all $j \in \{1, \dots, q\}$, the following linear regression model may be chosen to link $\mathbf{y}^{(j)}$ to $\mathbf{X}$,

$$\mathbf{y}^{(j)} = \mathbf{X}'\beta^{(j)} + \boldsymbol{\varepsilon}^{(j)}, \tag{4.1}$$

where $\beta^{(j)}$ is a real vector of length p and $\boldsymbol{\varepsilon}^{(j)}$ is a random vector of length n. For the sake of simplicity, we suppose that $\boldsymbol{\varepsilon}^{(j)} \sim \mathcal{N}_n(0, \sigma_j^2 \mathbf{I}_n)$, where $\mathbf{I}_n$ is the identity matrix and σ_j is an unknown positive real.

Let H_0 be a boolean, and let U be a dataset which comes from a model $M(H_0)$. Properties about $M(1)$ and $M(0)$ allows researchers to build statistical tests. A statistical test is a procedure that takes U and a parameter $\alpha \in]0, 1]$ as inputs and produces

- an estimate $\widehat{H}_0$ of H_0,
- a property which relies on α.

This property is an insight about how much $\widehat{H}_0$ and H_0 are likely to be the same. Usual statistical tests provide the following property:

$$\mathbb{P}(\widehat{H}_0 = 0 | H_0 = 1) \leq \alpha. \tag{4.2}$$

Note that these tests also satisfies for all $\alpha \in]0, 1]$,

$$\mathbb{P}(\widehat{H}_0 = 0 | H_0 = 0) \to 1, \tag{4.3}$$

when the sample size n of U goes to ∞. That is why, when $\widehat{H}_0 = 0$ and α is small enough, we can be confident about the fact that $H_0 = 0$.

For $j \in 1, \dots, q$, if $\beta_i^{(j)}$ is the i^{th} element of the vector $\beta^{(j)}$ in (4.1), there exist usual statistical tests for the following hypothesis

$$H_0 = H_{0,i}^{(j)} := \begin{cases} 0 & \text{if } \beta_i^{(j)} \neq 0, \\ 1 & \text{otherwise,} \end{cases}$$

with $U = U^{(j)} := (\mathbf{y}^{(j)}, \mathbf{X})$. Among them, we choose the F-Test which relies on the property that for any $j \in \{1, \dots, q\}$ and for any $i \in \{1, \dots, p\}$, if $H_{0,i}^{(j)} = 1$:

$$(n - p) \frac{\|\mathbf{H}\mathbf{y}^{(j)} - \mathbf{H}_{-i}\mathbf{y}^{(j)}\|^2}{\|\mathbf{y}^{(j)} - \mathbf{H}\mathbf{y}^{(j)}\|^2} \sim \mathcal{F}(1, n - p),$$

where

$$\mathbf{H} := \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X},$$
$$\mathbf{H}_{-i} := \mathbf{X}'_{-i}(\mathbf{X}_{-i}\mathbf{X}'_{-i})^{-1}\mathbf{X}_{-i},$$

$$\mathbf{X}_{-i} := (X_1, \ldots, X_{i-1}, X_{i+1}, X_p)',$$

$\mathcal{F}$ is the Fisher-Snedecor distribution, and for any vector a, $\|a\|^2 = a'a$. We call this test W_i. More details about it can be found in Bercu and Chafaï (2007, Chapter 7). If $\|X_i\|^2 \to \infty$ when $n \to \infty$, W_i satisfies equation (4.3). For all $\alpha \in]0,1]$, the estimate of $H_{0,i}^{(j)}$ provided by $W_i(U^{(j)}, \alpha)$ is named $\widehat{H}_{0,i}^{(j)}$.

Let $U^\star := \left\{U^{(1)}, \ldots, U^{(q)}\right\}$, and let $W_i^\star$ be a sequence of tests defined for a given $\gamma \in]0,1]$ by

$$W_i^\star(U^\star, \gamma) := \{W_i(U^{(1)}, \gamma), \ldots, \ W_i(U^{(q)}, \gamma)\}.$$

The False Discovery Rate (FDR) is formalised in Benjamini and Hochberg (1995). It is the expected proportion of false positive tests among positive tests and then satisfies

$$FDR(W_i^\star(U^\star, \gamma)) := \mathbb{E}\left[\frac{\#\{\widehat{H}_{0,i}^{(j)} = 0 \text{ and } H_{0,i}^{(j)} = 1\}}{\#\{\widehat{H}_{0,i}^{(j)} = 0\}}\right].$$

We consider here that $0/0 = 0$. For all $j \in \{1, \ldots, q\}$, let $\tilde{\alpha}_i^{(j)}$ be the p-value of $W_i(U^{(j)}, .)$, that is

$$\tilde{\alpha}_i^{(j)} := \sup\{\alpha \in]0,1] \ : \ \widehat{H}_{0,i}^{(j)} = 1\}.$$

Without any loss of generality, assume that $\tilde{\alpha}_i^{(1)} \leq \ldots, \leq \tilde{\alpha}_i^{(q)}$. From a usual statistical test W_i, and a dataset $\widetilde{U}_i := (W_i, U^\star)$, provided that for all $(j_1, j_2) \in \{1, \ldots, q-1\} \times \{j_1+1, q\}$, $\tilde{\alpha}_i^{(j_1)}$ and $\tilde{\alpha}_i^{(j_2)}$ are independent, Benjamini and Hochberg (1995) managed to create a sequence of tests $\widetilde{W}(\widetilde{U}_i, \gamma)$ defined by

- estimates $\widetilde{H}_{0,i}^{(j)}$ of $H_{0,i}$ for all $j \in \{1, \ldots, q\}$ satisfying

$$\widetilde{H}_{0,i}^{(j)} := \begin{cases} 0 & \text{if } \exists j^\star \geq j, \ \tilde{\alpha}_i^{(j^\star)} \leq \frac{j^\star}{q}\gamma, \\ 1 & \text{otherwise,} \end{cases}$$

- the property $FDR(\widetilde{W}(\widetilde{U}_i, \gamma)) \leq \gamma$.

Note that for the estimate $\widehat{H}_{0,i}^{(j)}$ produced by $W_i(U^{(j)}, \alpha)$, we have

$$\widehat{H}_{0,i}^{(j)} = \begin{cases} 0 & \text{if } \tilde{\alpha}_i^{(j)} \leq \alpha, \\ 1 & \text{otherwise.} \end{cases}$$

For all $j \in \{1, \ldots, q\}$, we can also define a p-value $\tilde{\gamma}_i^{(j)}$ of $\widetilde{W}_i(\widetilde{U}_i, .)$ by

$$\tilde{\gamma}_i^{(j)} := \sup\{\gamma \in]0,1] \ : \ \widetilde{H}_{0,i}^{(j)} = 1\}.$$

If for a given $i \in \{1, \ldots, p\}$, W_i satisfies (4.3), we also have for all $j \in \{1, \ldots, q\}$, and all $\gamma \in]0,1]$

$$\mathbb{P}(\widetilde{H}_{0,i}^{(j)} = 0 | H_{0,i}^{(j)} = 0) \to 1. \tag{4.4}$$

This can be shown by noticing that if $\alpha = \frac{1}{q}\gamma$, $(\widehat{H}_{0,i}^{(j)} = 0) \Rightarrow (\widetilde{H}_{0,i}^{(j)} = 0)$ and if $\alpha = \frac{j^\star}{q}\gamma$, $(\widetilde{H}_{0,i}^{(j)} = 0) \Rightarrow (\widehat{H}_{0,i}^{(j)} = 0)$.

4.2.2 FAMT: a method to take the correlations between the error terms into account

For a given $i \in \{1, \ldots, p\}$ and a given $j \in \{1, \ldots, q\}$, $\tilde{\alpha}_i^{(j)}$ relies on $\boldsymbol{\varepsilon}^{(j)}$ so that for all $(j_1, j_2) \in \{1, \ldots, q-1\} \times \{j_1 + 1, q\}$, $\tilde{\alpha}_i^{(j_1)}$ and $\tilde{\alpha}_i^{(j_2)}$ are independent if and only if the covariance matrix of $(\boldsymbol{\varepsilon}^{(j_1)\prime}, \boldsymbol{\varepsilon}^{(j_2)\prime})'$ is diagonal, which means that the link that may exist between $\mathbf{y}^{(j_1)}$ and $\mathbf{y}^{(j_2)}$ is fully characterized by $\mathbf{X}'\beta^{(j_1)}$ and $\mathbf{X}'\beta^{(j_2)}$. We would like to weaken this assumption. To do so, like Friguet *et al.* (2009) did, we introduce the following model, for $j \in \{1, \ldots, q\}$

$$\mathbf{y}^{(j)} = \mathbf{X}'\beta^{(j)} + \mathbf{Z}'b^{(j)} + \boldsymbol{\varepsilon}^{(j)}, \tag{4.5}$$

where $b^{(j)}$ is a vector of length r and $\mathbf{Z}$ is a $r \times n$ random matrix. In addition, if $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_n]$, we assume that

$$\begin{pmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_n \end{pmatrix} \sim \mathcal{N}_{nr}(0, \mathbf{I}_{nr}).$$

We also suppose that the covariance matrix of $(\boldsymbol{\varepsilon}^{(1)\prime}, \ldots, \boldsymbol{\varepsilon}^{(q)\prime})'$ is diagonal, but this does not imply this time that for all $(j_1, j_2) \in \{1, \ldots, q-1\} \times \{j_1 + 1, q\}$, $\mathbf{y}^{(j_1)}$ and $\mathbf{y}^{(j_2)}$ are only linked through $\mathbf{X}$.

Friguet *et al.* (2009) use an EM algorithm to find estimates $\widehat{\mathbf{Z}}$ and $\hat{b}^{(1)}, \ldots, \hat{b}^{(q)}$ of $\mathbf{Z}$ and $b^{(1)}, \ldots, b^{(q)}$. Their procedure includes a criterion to choose a suitable value of r. Note that for all $j \in \{1, \ldots, q\}$, if $\mathbf{y}^{(j)}$ satisfies model (4.5), then $\tilde{\mathbf{y}}^{(j)} = \mathbf{y}^{(j)} - \mathbf{Z}'b^{(j)}$ satisfies model (4.1). Let $\hat{\tilde{\mathbf{y}}}^{(j)} = \mathbf{y}^{(j)} - \widehat{\mathbf{Z}}'\hat{b}^{(j)}$. For a given $i \in \{1, \ldots, p\}$ and any $\gamma \in]0, 1]$, we can then perform the sequence of tests $\widetilde{W}(\widetilde{U}_i, \gamma)$ using $U^{(j)} = (\hat{\tilde{\mathbf{y}}}^{(j)}, \mathbf{X})$ for all $j \in \{1, \ldots, q\}$.

4.2.3 Gene expressions of eels studied with FAMT

We analyse the dataset presented in Section 4.1.2. Let X_1 be a vector made of n values of 1. When $p \geq 2$, using model (4.5), we would like to have

$$\beta^{(j)} = (\beta_1^{(j)}, \overbrace{0, \ldots, 0}^{i-2}, \beta_i^{(j)}, \overbrace{0, \ldots, 0}^{p-i})', \tag{4.6}$$

where $\beta_i^{(j)} \neq 0$, in order to link a single explanatory variable $i \in \{2, \ldots, p\}$ to a gene $j \in \{1, \ldots, q\}$. FAMT allows us to test if $\beta_i^{(j)}$ is not null but we may not be able to confidently assure that for a given i and a given j, $\beta_i^{(j)} = 0$. We nevertheless set $\hat{\beta}_i^{(j)} = 0$ if $\tilde{\gamma}_i^{(j)}$ is greater than a chosen value $\gamma^\star$, because of equation (4.4) for the sequence of tests $\widetilde{W}(\widetilde{U}_i, \gamma^\star)$.

For all $j \in \{1, \ldots, q\}$, and a pair $(i_1, i_2) \in \{2, \ldots, p-1\} \times \{i_1+1, \ldots, p\}$, if $\mathrm{cor}(X_{i_1}, X_{i_2})^2$ is close to one, it is very unlikely to have $\tilde{\gamma}_{i_1}^{(j)} \leq \gamma$ and $\tilde{\gamma}_{i_2}^{(j)} \geq \gamma^\star$ for values γ and $\gamma^\star$ such that $\gamma^\star - \gamma$ is large. We would like then to withdraw the components of $\mathbf{X}$ which are too much correlated with other components. To do so, we display in Figure 4.1 the projection of the explanatory variables onto the first two principal components computed from $\mathbf{X}$. The respective proportions of the total empirical variance of $\mathbf{X}$ along these axes are equal

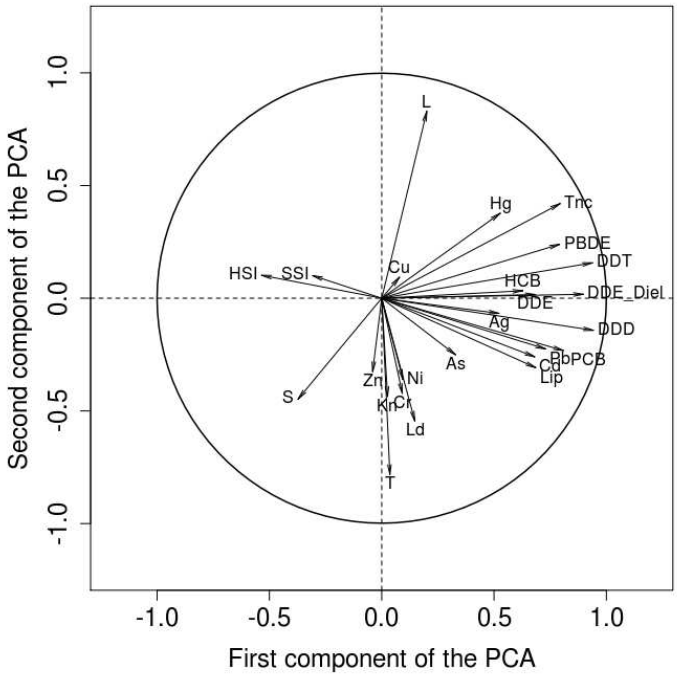


Figure 4.1: Explanatory variables projected onto the first two components of the PCA of $\mathbf{X}$

to $32,3\%$ and $12,8\%$. The variable *Cu* is not well-displayed in this plane and thus could be correlated with other variables through other axes. It is also not the most dangerous pollutant among those considered and we then choose to withdraw it. For the same reasons and because *Zn* is close to *Kn*, we withdraw it too. The variables *Pb* and *PCB* are close to each other. We then only keep *PCB*. The variable *Lip* is withdrawn because it is near *Cd*. The metals *Ni* and *Cr* are not far from *Ld* and *Kn*. That is why, we withdraw the first ones. Because DDE and DDD are metabolites of DDT, and since *HCB* is close to *DDE*, we replace *DDT*, *DDE*, *DDE_Diel* and *DDD* by $DD = DDT + DDE + DDE\_Diel + DDD$.

We have $n = 44$ eels. Gene expression data and observations of the explanatory variables were stored in a MySQL server. We access it from R like we did in Section 3.4 for the valvometric signals. We then study it with the FAMT model thanks to the `FAMT` package (Causeur *et al.*, 2011). We created a GUI to perform easily such analyses. There are three tabs in this interface. The first one (see Figure 4.2) includes a frame to connect to the SQL server and another one to select the explanatory variables to consider. We can launch the `modelFAMT` function from this tab and obtain $\tilde{\gamma}_i^{(j)}$ for $(i,j) \in \{2,\ldots,p\} \times \{1,\ldots,q\}$. These p-values can be saved from the second tab of our interface (see Figure 4.3). Finally, Figure 4.4 shows the last tab of the GUI which lists genes j such that for a variable X_{i_1}, $\tilde{\gamma}_{i_1}^{(j)} \leq \gamma$ and $\tilde{\gamma}_{i_2}^{(j)} \geq \gamma^\star$ for all $i_2 \in \{2,\ldots,p\}\backslash\{i_1\}$. For example, setting $\gamma = 10^{-3}$ and $\gamma^\star = 0$, when X_{i_1} corresponds to *Cd* leads to select 215 genes. If we take $\gamma^\star = 2 \times 10^{-3}$, we only obtain 92 genes which is a more suitable number since we have space for $1,000$ genes in the microarray and we have 10 pollutants left in the set of explanatory variables. Because γ is already small, we prefer to introduce a non-null $\gamma^\star$ rather than choosing a smaller value for γ.

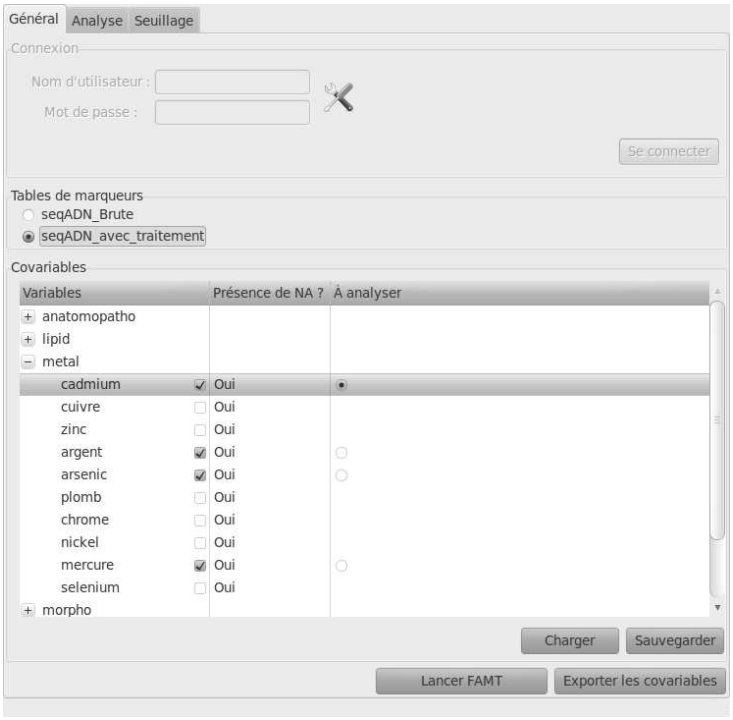


Figure 4.2: A graphical user interface to choose the explanatory variables to take into account

Figure 4.3: Some information is displayed in the second tab while FAMT is running

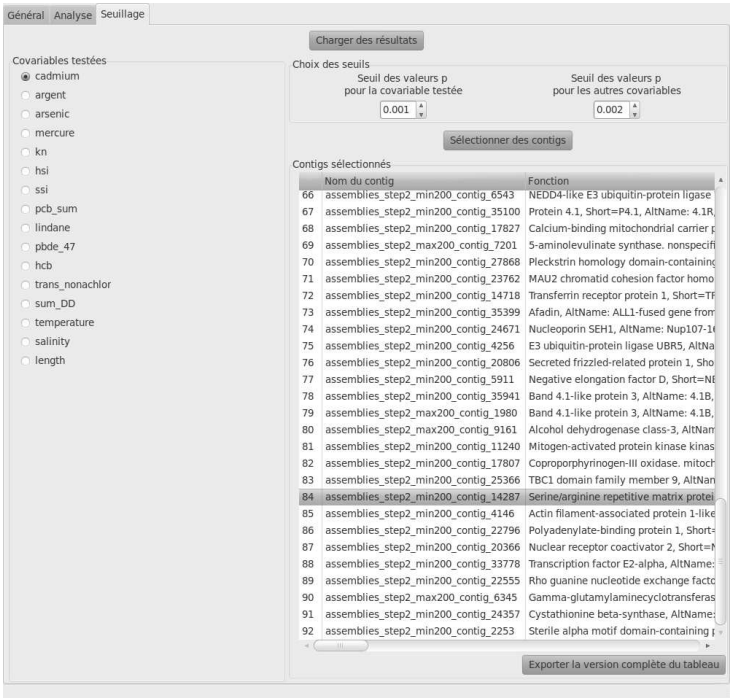


Figure 4.4: A selection is performed on genes using the thresholds γ and $\gamma^\star$

4.3 A new sliced inverse regression method for multivariate response regression

This section is an article under revision and was written with Stéphane Girard and Jérôme Saracco. It was submitted in *Computational Statistics and Data Analysis*.

4.3.1 Introduction

In analyzing large datasets, multivariate response regression analysis with a p-dimensional vector of regressors has been extensively studied in the literature. The reduction of the dimension of the regressors' space is a major concern in this framework. When the response variable is univariate, the issue has been addressed by Li (1991) via the notion of EDR (effective dimension reduction) space. The EDR directions (which form a basis of this subspace) are used to project the p-dimensional covariate $\mathbf{x}$ on a K-dimensional linear subspace (with $K < p$) first for displaying and then for studying its relationship with the response variable y. When the dimension of y is one, it is easy to view the link between the projected predictors and the response variable. The notion of EDR space was also clarified by Cook and his collaborators in their numerous papers introducing the notions of central subspace and central mean subspace, see for details Cook (1998) or Cook and Li (2002). Li (1991) introduced sliced inverse regression (SIR) which is a well-known method to estimate the EDR space. The link function can be estimated with a smoothing method such as kernel or smoothing splines approaches for instance.

In this paper, a q-dimensional response variable $\mathbf{y}$ is considered. Hence, we deal with a high dimensional regression framework which is not a linear one or a prespecified parametric one. The underlying idea of the dimension reduction of the explanatory variable $\mathbf{x}$ without loss of information is to identify the linear combinations $\beta_1'\mathbf{x}, \ldots, \beta_K'\mathbf{x}$ such that

$$\mathbf{y} \perp \mathbf{x} | (\beta_1'\mathbf{x}, \ldots, \beta_K'\mathbf{x}), \tag{4.7}$$

where $\perp$ denotes independence, $K (\leq p)$ is as small as possible and the p-dimensional vectors β_k are linearly independent. Let $\mathbf{B} = [\beta_1, \ldots, \beta_K]$ denote the $p \times K$ matrix of the β_k's. Statement (4.7) means that $\mathbf{y}|\mathbf{x}$ and $\mathbf{y}|\mathbf{B}'\mathbf{x}$ share the same distribution for all values of $\mathbf{x}$. A straightforward consequence is that the p-dimensional covariate $\mathbf{x}$ can be replaced by the K-dimensional predictor $\mathbf{B}'\mathbf{x}$ without loss of regression information. The goal of dimension reduction is achieved for $K < p$. As mentioned in Li (1991) or Cook (1994), statement (4.7) is equivalent to $\mathbf{y} \perp \mathbf{x} | P_{\mathbf{B}}\mathbf{x}$, where $P_{\mathbf{B}}$ denotes the projection operator on Span($\mathbf{B}$) which is the linear subspace of $\mathbb{R}^p$ spanned by the columns of $\mathbf{B}$. In addition, Span($\mathbf{B}$) can be viewed as the EDR space. From a regression model point of view, one can mention that the corresponding underlying model is the following semiparametric one

$$\mathbf{y} = f(\mathbf{B}'\mathbf{x}, \boldsymbol{\varepsilon}), \tag{4.8}$$

where $f : \mathbb{R}^{K+r} \to \mathbb{R}^q$ is an arbitrary and unknown link function, $\boldsymbol{\varepsilon}$ is a r-dimensional random error variable independent of $\mathbf{x}$ (with $r \geq 1$). Li *et al.* (2003) consider a regression model with an additive error term: $\mathbf{y} = g(\mathbf{B}'\mathbf{x}) + \boldsymbol{\varepsilon}$, where g is an unknown link function taking its values in $\mathbb{R}^q$.

In the following, we consider a slightly more restrictive regression model defined as follows:

$$\begin{cases} y^{(1)} = f_1(\mathbf{B}'\mathbf{x}, \varepsilon^{(1)}) \\ \vdots \\ y^{(q)} = f_q(\mathbf{B}'\mathbf{x}, \varepsilon^{(q)}) \end{cases} \tag{4.9}$$

where for $j = 1, \ldots, q$, $y^{(j)}$ (resp. $\varepsilon^{(j)}$) stands for the jth component of $\mathbf{y}$ (resp. of $\boldsymbol{\varepsilon}$) and the link function f_j is an unknown real-valued function. In this paper, we propose an approach to estimate the corresponding EDR space. It is based on combining information from the marginal regression of each component $y^{(j)}$ of $\mathbf{y}$.

Remark 8. The information from the marginal regression is sufficient to recover the whole EDR space in model (4.9). However, this is not always the case when working with model (4.8). Let us illustrate this point with the following regression model proposed by Zhu *et al.* (2010b):

$$\begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sin(\mathbf{B}'\mathbf{x}) \\ \sin(\mathbf{B}'\mathbf{x}) & 1 \end{pmatrix} \right).$$

In this example, since the information of interest is only in the correlations between the components of the response variables, none of the available marginal regression provides useful information about the EDR space while considering the entire $\mathbf{y}$ allows to recover it.

The vectors β_k are not individually identifiable neither in model (4.8), nor in the more restrictive one (4.9). Thus, the main objective is to estimate a basis of the K-dimensional EDR space. Many numerical methods have been introduced to achieve this goal. Let us mention three of them which are relatively simple and easy to implement: SIR introduced by Duan and Li (1991) for the single index model ($K = 1$) and Li (1991) for the multiple indices model ($K > 1$), principal Hessian directions (see for instance Li (1992) or Cook (1998)), and sliced average variance estimation (see for details Cook (2000), Prendergast (2007) or Shao *et al.* (2009)). In this paper, we shall only focus on the SIR approach which is based on a property of the first moment of the inverse distribution of $\mathbf{x}$ given $\mathbf{y}$.

From a theoretical point of view, the following linearity condition is required:

(LC) the conditional expectation $\mathbb{E}[b'\mathbf{x}|\mathbf{B}'\mathbf{x}]$ is linear in $\mathbf{B}'\mathbf{x}$ for any $b \in \mathbb{R}^p$.

One can observe that the linearity condition does not involve the response variable $\mathbf{y}$ and only concerns the distribution of the covariate $\mathbf{x}$. Let us mention that, when the distribution of $\mathbf{x}$ is an elliptically symmetric distribution (such as a p-dimensional normal distribution), this condition is satisfied. Cook and Nachtsheim (1994) proposed a method based on the minimum volume ellipsoid to transform and weight the predictors in order to approximate ellipticity. Kuentz and Saracco (2010) recommended to cluster the predictor space so that the linearity condition approximately holds in the different partitions. To conclude this brief discussion on the linearity condition, using a Bayesian argument of Hall and Li (1993), we can infer that (LC) approximately holds for many high-dimensional datasets (that is when p is large).

As previously mentioned, the proposed approach to estimate the EDR space relies on combining estimates from the marginal regressions of model (4.9). Moreover, we can

naturally take the information from these regressions into account to detect if a common EDR space really exists for all the components of $\mathbf{y}$ as in model (4.9). Otherwise, one shall consider the following more general regression model for multivariate response regression:

$$\begin{cases} y^{(1)} &= f_1(\mathbf{B}_1'\mathbf{x}, \varepsilon^{(1)}), \\ \vdots &= \vdots \\ y^{(q_1)} &= f_{q_1}(\mathbf{B}_1'\mathbf{x}, \varepsilon^{(q_1)}), \\ y^{(q_1+1)} &= f_{q_1+1}(\mathbf{B}_2'\mathbf{x}, \varepsilon^{(q_1+1)}), \\ \vdots &= \vdots \\ y^{(q_1+q_2)} &= f_{q_1+q_2}(\mathbf{B}_2'\mathbf{x}, \varepsilon^{(q_1+q_2)}), \\ \vdots &= \vdots \\ y^{(q)} &= f_q(\mathbf{B}_L'\mathbf{x}, \varepsilon^{(q)}), \end{cases} \tag{4.10}$$

where, for every $l = 1, \ldots, L$, $\mathbf{B}_l$ is a $p \times K$ matrix, K is assumed to be known, $\text{Span}(\mathbf{B}_1) \neq \text{Span}(\mathbf{B}_2) \neq \ldots \neq \text{Span}(\mathbf{B}_L)$ and $\sum_{l=1}^{L} q_l = q$. This means that writing this model as in (4.9) requires the number of columns of $\mathbf{B}$ to be greater than K. When trying to reduce as much as possible the dimension of a model, estimating $\mathbf{B}_1, \ldots, \mathbf{B}_L$ seems thus more appropriate than seeking $\mathbf{B}$. One then needs to cluster the components of $\mathbf{y}$ associated with the same EDR space. Therefore, for each identified cluster of components, one can use only these components to estimate the corresponding (common) EDR space. In a more general case, one can also assume that the dimension K is specific for each $\mathbf{B}_l$.

The goal of this paper is twofold. First we propose a new multivariate SIR approach for estimating the K-dimensional EDR space which is common to the q components of the multivariate response variable in model (4.9). Then we propose a way to cluster the components of $\mathbf{y}$ associated with the same EDR space in model (4.10) in order to apply properly our multivariate SIR on each cluster instead of blindly applying it on all the components of $\mathbf{y}$.

The paper is organized as follows. In Section 4.3.2, we give a brief overview on usual univariate SIR and existing multivariate SIR methods. The population version of the new SIR approach for a multivariate response, named MSIR hereafter, is described in Section 4.3.3.1. The corresponding sample version is introduced in Section 4.3.3.2 and asymptotic results are provided in Section 4.3.3.3. A weighted version of MSIR, named wMSIR hereafter, is proposed in Section 4.3.3.4. Both these methods rely on a tuning parameter H which is the number of slices. The choice of H and of the dimension K is discussed in Section 4.3.3.5. Practical methods to investigate the possible existence of a common EDR space for $\mathbf{y}$ and to detect and identify clusters of components of $\mathbf{y}$ are proposed in Section 4.3.4. Numerical results based on simulations are exhibited in Section 4.3.5 in order to show the good behavior of MSIR and wMSIR approaches and the usefulness of the diagnostic and clustering procedures on the components of $\mathbf{y}$. In Section 4.3.6, two real datasets are considered: the first one concerns hyperspectral remote sensing while the second one is the widely studied Minneapolis elementary schools dataset. Finally, concluding remarks are given in Section 4.3.7.

4.3.2 Brief review of univariate and multivariate SIR approaches

In this section, we consider the regression model (4.9). We first provide an overview of the SIR method when the response y is univariate. Then, some existing SIR methods for a multivariate response are briefly described. The aim of all these approaches is to estimate the EDR space.

4.3.2.1 Univariate SIR

We focus here on a univariate response (i.e. $q = 1$).

Inverse regression step. The basic principle of the SIR method is to reverse the roles of y and $\mathbf{x}$, that is, instead of regressing the univariate variable y on the multivariate variable $\mathbf{x}$, the covariable $\mathbf{x}$ is regressed on the response variable y.

Let T denote a monotone (but not necessarily strictly monotone) transformation of y. Let $\mu = \mathbb{E}(\mathbf{x})$ and $\mathbf{\Sigma} = \mathbb{V}(\mathbf{x})$. We assumed that $\mathbb{E}((\mathbf{x}'\mathbf{x})^2) < \infty$ and that $\mathbf{\Sigma}$ is invertible. Under model (4.9) and (LC), Li (1991) established the following geometric property: the centered inverse regression curve, $\mathbb{E}(\mathbf{x}|T(y)) - \mu$ as y varies, is contained in the linear subspace of $\mathbb{R}^p$ spanned by $\mathbf{\Sigma B}$. A straightforward consequence is that the covariance matrix

$$\mathbf{\Gamma} := \mathbb{V}(\mathbb{E}(\mathbf{x}|T(y)))$$

is degenerated in any direction $\mathbf{\Sigma}$-orthogonal to Span($\mathbf{B}$). Therefore, the eigenvectors associated with the K non-null eigenvalues of $\mathbf{\Sigma}^{-1}\mathbf{\Gamma}$ are some EDR directions.

Slicing step. Li (1991) proposed a transformation T, called "slicing", which categorizes the response y into a new response with $H > K$ levels. The support of y is partitioned into H non-overlapping slices $s_1, \ldots, s_h, \ldots, s_H$. With such a transformation T, the subspace recovered through a slicing (based on the inverse $\mathbf{x}|T(y)$ function) may fall short of the space recovered through y in its entirety (based on the $\mathbf{x}|y$ function). However the main advantage of the slicing is that the matrix of interest is now written as

$$\mathbf{\Gamma} = \sum_{h=1}^{H} p_h(\mathbf{m}_h - \mu)(\mathbf{m}_h - \mu)'$$

where $p_h = \mathbb{P}(y \in s_h)$ and $\mathbf{m}_h = \mathbb{E}(\mathbf{x}|y \in s_h)$.

Estimation process. In the usual statistical framework, when a sample $\{(\mathbf{x}_i, y_i),\ i = 1, \ldots, n\}$ is available, it is straightforward to estimate the matrices $\mathbf{\Sigma}$ and $\mathbf{\Gamma}$, by substituting empirical versions of the moments for their theoretical counterparts, and therefore to obtain the estimation of the EDR directions. Li (1991) showed that each of these estimated EDR directions converges to an EDR direction at rate $\sqrt{n}$. Asymptotic normality of these estimated EDR directions has been obtained by Saracco (1997).

From a practical point of view, the choice of the slicing is discussed in Li (1991), Chen and Li (1998) or Saracco (2001). Since SIR theory makes no assumption about the slicing strategy, the user must choose the number H of slices and how to construct them. In practice, there are naturally two possibilities: to fix the width of the slices or to fix the

number of observations per slice. This second option is often preferred, and from the sample point of view, the slices are often chosen such that the number of observations in each slice is as close to each other as possible. Note that H must be greater than K in order to avoid an artificial reduction of dimension and must be lower than $\lfloor n/2 \rfloor$ in order to have at least two observations in each slice (where $\lfloor a \rfloor$ denotes the integer part of a). The choice of H is less sensitive than the choice of a smoothing parameter in nonparametric regression. This point is clearly illustrated in Liquet and Saracco (2012) with a graphical tool that allows the user to find simultaneously realistic values of the two parameters H and K, see Section 4.3.3.5 for some details on this method.

The SIR estimates based on the first inverse moment have been studied extensively, see for instance Hsing and Carroll (1992), Zhu and Ng (1995), Saracco (1999), Prendergast (2005), Szretter and Yohai (2009) among others for some asymptotic results. Chen and Li (1998) exhibited many features to popularize SIR. Chavent *et al.* (2011) considered the case of a stratified population. In order to avoid the choice of a slicing in SIR, pooled slicing, kernel or spline versions of SIR have been investigated, see for example Zhu and Fang (1996), Aragon and Saracco (1997), Zhu and Yu (2007), Wu (2008), Kuentz *et al.* (2010) or Azaïs *et al.* (2012). However, these methods are hard to implement comparing to the basic SIR approach and are often computationally slow. Regularized versions for SIR have been proposed for high-dimensional covariates, see for instance Zhu *et al.* (2006), Scrucca (2007), Li and Yin (2008), Bernard-Michel *et al.* (2009b). Sparse SIR has been proposed by Li and Nachtsheim (2006). Hybrid methods of inverse regression-based algorithms have been also studied, see for example Gannoun and Saracco (2003) or Zhu *et al.* (2007).

4.3.2.2 Multivariate SIR

In the multivariate framework (that is when $\mathbf{y} \in \mathbb{R}^q$ with $q > 1$), Aragon (1997), Li *et al.* (2003) considered several estimation methods of the EDR space based on SIR. Note that Barreda *et al.* (2007) proposed extensions of the following multivariate SIR methods based on SIR_α approach instead of SIR, where SIR_α is a generalization of SIR which combines information from the first two conditional moments of $\mathbf{x}$ given $T(\mathbf{y})$.

Complete slicing and marginal slicing approaches. In the complete slicing method, the SIR procedure is directly applied on $\mathbf{y}$. To build slices of nearly equal sizes, the following recursive approach is used. The first component of $\mathbf{y}$ is sliced. Then, each slice is separately sliced again according to the next component of $\mathbf{y}$, and so on. This extension of univariate SIR to multivariate $\mathbf{y}$ appears straightforward and the theoretical development can be formally carried over. Computation of such estimators suffers from the so-called curse of dimensionality when the dimension q of $\mathbf{y}$ is large ($q \geq 4$). Note that Hsing (1999) proposed a version of SIR in which the slices are determined by the nearest neighbors approach and showed that the EDR directions can be estimated with rate $\sqrt{n}$ under general conditions. Moreover, Setodji and Cook (2004) extended that univariate SIR to multivariate framework by introducing a new way of performing the slicing of $\mathbf{y}$ based on k-means method. The corresponding method is called k-means inverse regression (KIR).

A natural way to circumvent the curse of dimensionality of the complete slicing approach is proposed in the marginal slicing procedure which consists in applying the SIR method on a transformation of $\mathbf{y}$ depending on one's interest. For instance, it can be the

mean or the median of the $y^{(j)}$'s. One can also take the first few significant components of a principal component analysis of the $y^{(j)}$'s to construct the slices. However, slicing a lower dimensional projection of $\mathbf{y}$ may not lead to recover as many EDR directions as slicing the entire $\mathbf{y}$.

For these reasons, these two multivariate approaches (complete slicing and marginal slicing) are not completely satisfactory.

Pooled marginal slicing approach. The idea of the pooled marginal slicing (PMS) method is to consider the q univariate marginal SIR of each component $y^{(j)}$ of $\mathbf{y}$ on $\mathbf{x}$ and to combine the corresponding matrices of interest $\mathbf{\Gamma}^{(j)} := \mathbb{V}(\mathbb{E}(\mathbf{x}|T_j(y^{(j)})))$ in the following pooling:

$$\mathbf{\Gamma}_P = \sum_{j=1}^{q} w_j \mathbf{\Gamma}^{(j)}, \tag{4.11}$$

for positive weights w_j. It has been shown that the eigenvectors associated with the non-null K eigenvalues of $\mathbf{\Sigma}^{-1}\mathbf{\Gamma}_P$ are EDR directions. Aragon (1997) proposes to use two kinds of weighting for the w_j's: equal weights or weights proportional to the major eigenvalues found by a preliminary univariate SIR analysis of each component of $\mathbf{y}$. Saracco (2005) obtained the asymptotic normality of the pooled marginal slicing estimator based on SIR_α. Lue (2009) derived the asymptotic weighted chi-squared test for dimension. For $j = 1, \ldots, q$, rather than constructing $\mathbf{\Gamma}^{(j)}$ from $y^{(j)}$, one can also build it from a linear combination $\tau'\mathbf{y}$ of $\mathbf{y}$. This method which is called projective resampling was introduced by Li *et al.* (2008). To ensure good performances, the number of linear combinations to handle should be greater than the sample size n.

Some other multivariate SIR approaches. Bura and Cook (2001) introduced the parametric inverse regression that may easily adapt to multivariate response framework. Yin and Bura (2006) proposed a moment-based dimension reduction approach in this context. Moreover, in order to solve the dimensionality problem when p is large and to rationalize the slicing step, Li *et al.* (2003) presented an algorithm based on a duality between SIR variates and MP (most predictable) variates. The term "variate" denotes any linear combination of either the regressor $\mathbf{x}$ or the response variable $\mathbf{y}$. The SIR variates are the variables $b'\mathbf{x}$ formed by an EDR direction b obtained with SIR. The MP variates $\theta'\mathbf{y}$ are defined as those minimizing the ratio $\mathbb{E}[\mathbb{V}(\theta'\mathbf{y}|\mathbf{x})]/\mathbb{V}(\theta'\mathbf{y})$, where $\mathbb{V}(\theta'\mathbf{y}|\mathbf{x})$ is the associated prediction mean squared error of the best nonlinear prediction $\mathbb{E}[\theta'\mathbf{y}|\mathbf{x}]$ for the squared error loss. Equivalently, due to ANOVA identity, the MP variates can be found by maximizing the ratio $\mathbb{V}(\mathbb{E}[\theta'\mathbf{y}|\mathbf{x}])/\mathbb{V}(\theta'\mathbf{y})$, which conducts to the same eigenvalue decomposition as the SIR approach except for the exchanged roles of $\mathbf{x}$ and $\mathbf{y}$. This twin relationship between SIR variates and MP variates underlies the development of the alternating SIR algorithm. The idea of the algorithm is to alternate computations of either $\widehat{\theta}$ or $\widehat{b}$ respectively obtained by the slicing of SIR variates or MP variates constructed at the previous step. Li *et al.* (2003) proposed an iterative procedure for the alternating SIR and showed that choosing the canonical directions as an initial projection of the $\mathbf{y}$'s guarantees the convergence of the corresponding algorithm in a finite number of steps (equal to K, the number of EDR directions).

4.3.3 A new multivariate SIR approach

The population version of the proposed MSIR approach is first described in Section 4.3.3.1. Then, the corresponding sample version is given in Section 4.3.3.2 and some asymptotic results are derived in Section 4.3.3.3. We then modify MSIR to handle a weighting in the components of $\mathbf{y}$ in Section 4.3.3.4. Methods to choose K and H are discussed in Section 4.3.3.5. Finally, procedures to withdraw or cluster components of $\mathbf{y}$ are detailled in Section 4.3.4.

4.3.3.1 Population version

Let us assume that the dimension K of the EDR space is known. Let $P_{\mathbf{M},\mathbf{\Sigma}}$ be the $\mathbf{\Sigma}$-orthogonal projector on the linear subspace spanned by the columns of a $p \times K$ matrix $\mathbf{M}$. A proximity measure between two projectors $P_{\mathbf{M}_1,\mathbf{\Sigma}_1}$ and $P_{\mathbf{M}_2,\mathbf{\Sigma}_2}$ is given by the squared trace correlation:

$$r(\mathbf{M}_1, \mathbf{\Sigma}_1, \mathbf{M}_2, \mathbf{\Sigma}_2) := \frac{1}{K} \text{Trace}\left(P_{\mathbf{M}_1,\mathbf{\Sigma}_1} P_{\mathbf{M}_2,\mathbf{\Sigma}_2}\right),$$

for full column rank matrices $\mathbf{M}_1$ and $\mathbf{M}_2$.

For $j = 1, \ldots, q$, let $\mathbf{B}^{(j)}$ be a $p \times K$ matrix containing a $\mathbf{\Sigma}$-orthonormal basis of the EDR space from the marginal regression of $y^{(j)}$ given $\mathbf{x}$. Let $\mathbf{D}$ be a $p \times K$ matrix such that $\mathbf{D}'\mathbf{\Sigma}\mathbf{D} = \mathbf{I}_K$, where $\mathbf{I}_K$ is the identity matrix of order K. Let $Q(\mathbf{D}, \mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(q)})$ denote the following proximity measure between Span($\mathbf{D}$) and the q marginal EDR spaces Span($\mathbf{B}^{(1)}$), ..., Span($\mathbf{B}^{(q)}$):

$$Q(\mathbf{D}, \mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(q)}) := \frac{1}{q} \sum_{j=1}^{q} r\left(\mathbf{D}, \mathbf{\Sigma}, \mathbf{B}^{(j)}, \mathbf{\Sigma}\right). \tag{4.12}$$

This measure takes its values in [0,1]. Note that Span($\mathbf{D}$) = Span($\mathbf{B}^{(1)}$) = $\cdots$ = Span($\mathbf{B}^{(q)}$) implies $Q(\mathbf{D}, \mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(q)}) = 1$. The closer to one is this measure, the closer to the q marginal EDR spaces is the linear subspace Span($\mathbf{D}$).

Let us now consider the following optimization problem:

$$\mathbf{V} := \arg \max_{\mathbf{D} \in \mathcal{M}_{\mathbf{\Sigma}}} Q(\mathbf{D}, \mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(q)}), \tag{4.13}$$

where $\mathcal{M}_{\mathbf{\Sigma}}$ is the set of $\mathbf{\Sigma}$-orthogonal $p \times K$ matrices. A solution of (4.13) is given by the following theorem.

Theorem 9. *Under model (4.9) and assumption (LC), the $p \times K$ matrix $\mathbf{V}$ is formed by the eigenvectors $v_1, \ldots, v_K$ associated with the K non-null eigenvalues of $\mathbb{B}\mathbb{B}'\mathbf{\Sigma}$ where $\mathbb{B}$ is the $p \times (Kq)$ matrix defined as $\mathbb{B} := [\mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(q)}]$. Moreover we have Span($\mathbf{V}$) = Span($\mathbf{B}$) where Span($\mathbf{B}$) is the EDR space.*

The proof is given in Appendix A.3.2. From Theorem 9, we can estimate a basis of the EDR space based on estimators of matrices $\mathbb{B}$ and $\mathbf{\Sigma}$. This is the goal of the next subsection.

4.3.3.2 Sample version

Let $\{(\mathbf{x}_i, \mathbf{y}_i),\ i = 1, \ldots, n\}$ be a sample of independent observations from model (4.9). Each $\mathbf{y}_i$ is a q-dimensional random variable. We assume that the sample size n is larger than the dimension p of each covariate $\mathbf{x}_i$.

Let $\bar{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ and $\widehat{\mathbf{\Sigma}} := \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ be the empirical mean and the covariance matrix of the $\mathbf{x}_i$'s.

In order to estimate the matrix $\mathbb{B}$, we have to estimate each $p \times K$ matrix $\mathbf{B}^{(j)}$ with usual univariate SIR from the subsample $\{(\mathbf{x}_i, y_i^{(j)}),\ i = 1, \ldots, n\}$ where $y_i^{(j)}$ stands for the jth component of $\mathbf{y}_i$. To this end, let us assume that the support of $y^{(j)}$ is partitioned into a fixed number of slices denoted by $s_1^{(j)}, \ldots, s_h^{(j)}, \ldots, s_{H^{(j)}}^{(j)}$. Let $p_h^{(j)} := \mathbb{P}\left(y^{(j)} \in s_h^{(j)}\right)$ and $\mathbf{m}_h^{(j)} := \mathbb{E}\left(\mathbf{x} | y^{(j)} \in s_h^{(j)}\right)$. Thus, the matrix $\mathbf{\Gamma}^{(j)} := \sum_{h=1}^{H^{(j)}} p_h^{(j)}(\mathbf{m}_h^{(j)} - \mu)(\mathbf{m}_h^{(j)} - \mu)'$ is estimated by $\widehat{\mathbf{\Gamma}}^{(j)} := \sum_{h=1}^{H^{(j)}} \hat{p}_h^{(j)}(\widehat{\mathbf{m}}_h^{(j)} - \bar{\mathbf{x}})(\widehat{\mathbf{m}}_h^{(j)} - \bar{\mathbf{x}})'$ with $\hat{p}_h^{(j)} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[y_i^{(j)} \in s_h^{(j)}]$ and $\widehat{\mathbf{m}}_h^{(j)} := \frac{1}{n\hat{p}_h^{(j)}}\sum_{i=1}^{n} \mathbf{x}_i \mathbb{I}[y_i^{(j)} \in s_h^{(j)}]$ where $\mathbb{I}[.]$ is the indicator function. Then, the matrix $\mathbf{B}^{(j)}$ is estimated by

$$\widehat{\mathbf{B}}^{(j)} := \left[\hat{\mathbf{b}}_1^{(j)}, \ldots, \hat{\mathbf{b}}_K^{(j)}\right],$$

where the vectors $\hat{\mathbf{b}}_k^{(j)},\ k = 1, \ldots, K$ are the $\widehat{\mathbf{\Sigma}}$-orthonormal eigenvectors associated with the K largest eigenvalues of the matrix $\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Gamma}}^{(j)}$. It follows that the matrix $\mathbb{B}$ is directly estimated by

$$\widehat{\mathbb{B}} := \left[\widehat{\mathbf{B}}^{(1)}, \ldots, \widehat{\mathbf{B}}^{(q)}\right].$$

Finally, a $\widehat{\mathbf{\Sigma}}$-orthonormal estimated basis of the EDR space is given by the vectors $\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_K$ defined as the eigenvectors associated with the K largest eigenvalues of the matrix $\widehat{\mathbb{B}}\widehat{\mathbb{B}}'\widehat{\mathbf{\Sigma}}$ and we write $\widehat{\mathbf{V}} := [\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_K]$.

Remark 9. Similarly to the pooled marginal slicing (PMS) presented in Section 4.3.2.2, MSIR relies on the univariate version of SIR, applied to each component of $\mathbf{y}$. While both methods need estimates of $\mathbf{\Gamma}^{(1)}, \ldots, \mathbf{\Gamma}^{(q)}$, estimates of $\mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(q)}$ are only required by MSIR. Computing such estimates is useful to explore the relations between components of $\mathbf{y}$ as explained in Section 4.3.4.

4.3.3.3 An asymptotic result

The following assumptions are necessary to state our asymptotic result. Let $n_h^{(j)} := n\hat{p}_h^{(j)}$ be the number of observations in the slice $s_h^{(j)}$.

- (A1) Observations $\{(\mathbf{x}_i, \mathbf{y}_i),\ i = 1, \ldots, n\}$ are independently drawn from a given regression model.
- (A2) For each component $y^{(j)}$ of $\mathbf{y}$, the support is partitioned into a fixed number $H^{(j)}$ of slices such that $p_h^{(j)} > 0$ for $h = 1, \ldots, H^{(j)}$.
- (A3) For $j = 1, \ldots, q$ and $h = 1, \ldots, H^{(j)}$, $n_h^{(j)} \to \infty$ as $n \to \infty$.

Theorem 10. *Under model (4.9) and assumptions (LC) and (A1)-(A3), we have, for* $k = 1, \ldots, K$,

$$\hat{\mathbf{v}}_k = v_k + O_p(n^{-1/2}),$$

which means that for all $\epsilon > 0$ *there exist* $\zeta_\epsilon > 0$ *and* $\tilde{n}_\epsilon \in \mathbb{N}^*$ *such that for all* $n \geq \tilde{n}_\epsilon$

$$\mathbb{P}\left(\sqrt{n}\,|\hat{\mathbf{v}}_k - v_k| > \zeta_\epsilon\right) < \epsilon,$$

and then the estimated EDR space Span$(\widehat{\mathbf{V}})$ *converges in probability to the EDR space.*

The poof is given in Appendix A.3.3.

Remark 10. Using Delta-method and asymptotic results of Tyler (1981) and Saracco (1997), it is possible to obtain the asymptotic normality of

$$\sqrt{n}\left(\text{vec}(\widehat{\mathbb{B}}\widehat{\mathbb{B}}'\widehat{\mathbf{\Sigma}}) - \text{vec}(\mathbb{B}\mathbb{B}'\mathbf{\Sigma})\right),$$

where vec($\mathbf{M}$) stands for the "vec" operator applied to matrix $\mathbf{M}$. More precisely, this operator rearranges the p^2 elements of $\mathbf{M}$ in the form of a p^2-dimensional column vector by stacking the p columns of $\mathbf{M}$ one under the other. Then, the asymptotic normality of the eigenprojector onto the estimated EDR space can be derived, as well as the asymptotic distribution of the estimated EDR directions $\hat{\mathbf{v}}_k$, associated with eigenvalues assumed to be different (that is $\lambda_1 > \cdots > \lambda_K > 0$).

4.3.3.4 A weighted version of MSIR

Following the idea used in pooled marginal slicing approach in which the matrix of interest $\mathbf{\Gamma}_P$ is a weighted average of the marginal matrices $\mathbf{\Gamma}^{(j)}$, we can consider a weighted version of the multivariate SIR method introduced in this paper, named wMSIR hereafter. As it has already been proposed by Aragon (1997) or Lue (2009), we shall use weights based on the proportion of eigenvalues corresponding to significant eigenvectors (which are EDR directions) in each marginal SIR (i.e. univariate SIR on each marginal component of $\mathbf{y}$).

More precisely, for $j = 1, \ldots, q$, let $\lambda_k^{(j)}, k = 1, \ldots, p$ be the eigenvalues of the eigendecomposition $\mathbf{\Sigma}^{-1}\mathbf{\Gamma}^{(j)}v_k^{(j)} = \lambda_k^{(j)}v_k^{(j)}$ where $\lambda_1^{(j)} \geq \lambda_2^{(j)} \geq \cdots \geq \lambda_p^{(j)}$. Let us define, for each component $y^{(j)}$ of $\mathbf{y}$, the proportion of eigenvalues corresponding to significant eigenvectors: $\pi^{(j)} = \frac{\sum_{k=1}^{K}\lambda_k^{(j)}}{\sum_{k=1}^{p}\lambda_k^{(j)}}$. Let us also define $\pi_\star = \sum_{j=1}^{q}\pi^{(j)}$. Then, we can introduce the following $qK \times qK$ matrix of weights

$$\mathbb{W} = \text{diag}(\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(j)}, \ldots, \mathbf{W}^{(q)})$$

with $\mathbf{W}^{(j)} = \frac{\pi^{(j)}}{\pi_\star}\mathbf{I}_K$ for $j = 1, \ldots, q$. Note that, from a theoretical point of view, under model (4.9) and (LC), the matrix of weights is such that $\mathbb{W} = \frac{1}{q}\mathbf{I}_{qK}$ since $\lambda_k^{(j)} = 0$ for $j = 1, \ldots, q$ and $k = K+1, \ldots, p$.

The population version of wMSIR consists in noticing that the eigenvectors $\tilde{v}_1, \ldots, \tilde{v}_K$ associated with the K largest eigenvalues of the $\mathbf{\Sigma}$-symmetric matrix $\mathbb{B}\mathbb{W}\mathbb{B}'\mathbf{\Sigma}$ span the EDR space. We write $\widetilde{\mathbf{V}} := [\tilde{v}_1, \ldots, \tilde{v}_K]$. To show this result, one can proceed analogously to the proof of Theorem 9.

Let $\hat{\lambda}_k^{(j)}$ be the kth eigenvalue of $\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Gamma}}^{(j)}$. The sample version of wMSIR is obtained by substituting the empirical matrices $\widehat{\mathbb{B}}$, $\widehat{\mathbb{W}}$ and $\widehat{\mathbf{\Sigma}}$ for their theoretical counterparts

$\mathbb{B}$, $\mathbb{W}$ and $\mathbf{\Sigma}$, where $\widehat{\mathbb{W}} = \text{diag}(\widehat{\mathbf{W}}^{(1)}, \ldots, \widehat{\mathbf{W}}^{(q)})$ with, for $j = 1, \ldots, q$, $\widehat{\mathbf{W}}^{(j)} = \frac{\hat{\pi}^{(j)}}{\hat{\pi}_\star}\mathbf{I}_K$, $\hat{\pi}^{(j)} = \frac{\sum_{k=1}^{K}\hat{\lambda}_k^{(j)}}{\sum_{k=1}^{p}\hat{\lambda}_k^{(j)}}$ and $\hat{\pi}_\star = \sum_{j=1}^{q}\hat{\pi}^{(j)}$. Therefore, one can get the corresponding estimated EDR directions $\widehat{\widetilde{\mathbf{V}}} := \left[\widehat{\tilde{\mathbf{v}}}_1, \ldots, \widehat{\tilde{\mathbf{v}}}_K\right]$.

Theorem 11. *Under model (4.9) and assumptions (LC) and (A1)-(A3), we have, for* $k = 1, \ldots, K$,

$$\widehat{\tilde{\mathbf{v}}}_k = \tilde{v}_k + O_p(n^{-1/2}),$$

that is, the estimated EDR space Span$(\widehat{\widetilde{\mathbf{V}}})$ *converges in probability to the EDR space.*

The proof is given in Appendix A.3.4.

4.3.3.5 Discussion on the choice of K and H

Up to now, the dimension K of the EDR space was assumed to be known. However, in most applications based on real datasets, the number K of indices $\beta_k'\mathbf{x}$ in model (4.7) is a priori unknown and hence must be determined from the data. In addition, the number of slices H in MSIR has to be chosen.

Several approaches to determine K have been proposed in the literature for univariate SIR. Some of them are based on hypothesis tests on the nullity of the last $(p - K)$ eigenvalues, see for instance Li (1991), Schott (1994), Bai and He (2004), Barrios and Velilla (2007) or Nkiet (2008). In the multivariate response framework, Lue (2009) derived an asymptotic weighted chi-squared test for dimension adapted to the pooled marginal slicing estimator. In our case, a crude choice of the dimension can be also made by a visual inspection of the eigenvalues scree plot of the matrix $\widehat{\mathbb{B}}\widehat{\mathbb{B}}'\widehat{\mathbf{\Sigma}}$: the idea is to determine the number of the significantly non-null eigenvalues.

In the univariate response model, Liquet and Saracco (2012) proposed to consider a risk function which can be replaced in this multivariate context by:

$$R_{h,k} := \mathbb{E}\left(r(\mathbf{V}_k, \mathbf{\Sigma}, \widehat{\mathbf{V}}_k, \widehat{\mathbf{\Sigma}})\right), \tag{4.14}$$

where $\mathbf{V}_k := [v_1, \ldots, v_k]$, $\widehat{\mathbf{V}}_k := [\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_k]$ and h is the number of slices used to obtain $\mathbf{V}_k$ and $\widehat{\mathbf{V}}_k$. This risk function only makes sense for any dimension k lower than or equal to the true dimension K of the EDR space. For the true dimension K, $R_{h,K}$ converges to one as n tends to infinity. For a fixed n, a reasonable way to assess whether an EDR direction is available is to graphically evaluate how much $R_{h,k}$ departs from one. From a computational point of view, consistent estimates of $R_{h,k}$ are required. Liquet and Saracco (2012) use a bootstrap estimator $\widehat{R}_{h,k}$ of this criterion in order to determine the pair (H, K) of parameters. The proposed graphical method consists in evaluating the $\widehat{R}_{h,k}$ values for all $k = 1, \ldots, p$ and some reasonable values of h, and in observing how much the criterion departs from one. The best choice will be the pair $(\widehat{H}, \widehat{K})$ which gives a value of $\widehat{R}_{h,k}$ close to one, such that $\widehat{K} << p$ in order to get an effective dimension reduction. In practice, there is no objective criterion to find a trade-off between a large value of the criterion $\widehat{R}_{h,k}$ and a small value of the dimension K. Then, a visual expertise of the 3D-plot of the $\widehat{R}_{h,k}$ versus (h, k) allows the selection of the best value. It is also useful to provide, for each (h, k), the boxplots of the bootstrap replication of the squared trace

correlation to have a look on how much the corresponding k-dimensional linear subspace is stable. Although boxplots of $\widehat{R}_{h,k}$ are also useful to determine the optimal number of slices $\widehat{H}$, wMSIR is not really sensitive to this parameter as shown in Section 4.3.5.

4.3.4 Analyzing components of y through MSIR

Recall that the estimate $\widehat{\mathbf{V}}$ (resp. $\widehat{\widetilde{\mathbf{V}}}$) of MSIR (resp. wMSIR) is computed from the estimated EDR directions $\hat{\mathbf{B}}^{(j)}$ associated with each component of $\mathbf{y}$. From these estimates, it is straightforward to calculate the proximity measure $\hat{r}_j := r(\hat{\mathbf{B}}^{(j)}, \hat{\mathbf{\Sigma}}, \widehat{\mathbf{V}}, \hat{\mathbf{\Sigma}})$ (resp. $\hat{\tilde{r}}_j := r(\hat{\mathbf{B}}^{(j)}, \hat{\mathbf{\Sigma}}, \widehat{\widetilde{\mathbf{V}}}, \hat{\mathbf{\Sigma}})$) between each estimated marginal EDR space and the estimated common one, for $j = 1, \ldots, q$. Then it is easy to sort these measures in descending order and to draw the associated scree plot. For $j = 1, \ldots, q$, assuming model (4.9) and observing a low value of $\hat{r}_j$ or $\hat{\tilde{r}}_j$ could indicate an unprecise estimate of $\mathbf{B}^{(j)}$ since $\hat{r}_j$ and $\hat{\tilde{r}}_j$ tends to 1 in probability as n goes to ∞. One can then withdraw the component $y^{(j)}$ of $\mathbf{y}$ to improve the accuracy of $\widehat{\mathbf{V}}$ or $\widehat{\widetilde{\mathbf{V}}}$.

In addition, assuming model (4.9) with a low dimensional common EDR space for the whole components of $\mathbf{y}$ does not always seem realistic in real data analysis. Therefore, applying any multivariate SIR method on $\mathbf{y}$ should not provide a suitable dimension reduction. However, it makes sense to assume that only groups of components of $\mathbf{y}$ rely on model (4.9) with small values of K, as in model (4.10). For this model, we propose a methodology in order to identify the variables $y^{(j)}$ which share the same EDR space. Thus, we obtain clusters of components on which applying a multivariate SIR approach is sensible. Note that performing marginal univariate SIR on each component $y^{(j)}$ of $\mathbf{y}$ leads to consistent estimates of each K-dimensional EDR space, since for $l = 1, \ldots, L$, it is assumed that the rank of $\mathbf{B}_l$ is equal to K. Recalling notations of Section 4.3.3.2, we obtain q_1 estimates $\mathrm{Span}(\hat{\mathbf{B}}^{(1)}), \ldots, \mathrm{Span}(\hat{\mathbf{B}}^{(q_1)})$ of $\mathrm{Span}(\mathbf{B}_1)$, q_2 estimates $\mathrm{Span}(\hat{\mathbf{B}}^{(q_1+1)}), \ldots, \mathrm{Span}(\hat{\mathbf{B}}^{(q_1+q_2)})$ of $\mathrm{Span}(\mathbf{B}_2)$, and so on, but values of $q_1, \ldots, q_L$ are unknown, as well as the number L of clusters.

For $(j, j^\star) \in \{1, \ldots, q\}^2$, we define $\hat{r}_{j,j^\star} := r(\hat{\mathbf{B}}^{(j)}, \hat{\mathbf{\Sigma}}, \hat{\mathbf{B}}^{(j^\star)}, \hat{\mathbf{\Sigma}})$. Without any loss of generality, assume that $L = 2$. Let us define

$$(j_1, j_2, j_3, j_4) \in \{1, \ldots, q_1\}^2 \times \{q_1 + 1, \ldots, q_1 + q_2\}^2 .$$

We thus have the following Lemma.

Lemma 12. *Under model (4.10) and assumptions (LC) and (A1)-(A3), $\hat{r}_{j_1,j_2}$ and $\hat{r}_{j_3,j_4}$ tend to 1 in probability.*

The proof of this Lemma is given in Appendix A.3.5. Let us remark that, however, $\hat{r}_{j_1,j_3}$ does not converge to 1 in probability since $r(\mathbf{B}_1, \mathbf{\Sigma}, \mathbf{B}_2, \mathbf{\Sigma}) < 1$. This leads to the following criterion to cluster components of $\mathbf{y}$: for $(j, j^\star) \in \{2, \ldots, q\} \times \{1, \ldots, j-1\}$, components $y^{(j)}$ and $y^{(j^\star)}$ are put in the same cluster if $\hat{r}_{j,j^\star}$ is close to 1. To do so, we can perform, for instance, a hierarchical ascending classification on the $q \times q$ (symmetrtic) matrix of the proximity measures $\hat{r}_{j,j^\star}$ (with $\hat{r}_{j,j} = 1$ for $j = 1, \ldots, q$). Other clustering procedures can be applied on this matrix, such as the multidimensional scaling together with the k-means method. In Sections 4.3.5-4.3.6, we give illustrations of a clustering step for simulated and real datasets, which clearly improves the estimation of the corresponding EDR spaces.

Remark 11. Computing the common estimated EDR space for each obtained cluster is not time consuming since $\widehat{\mathbf{B}}^{(1)}, \ldots, \widehat{\mathbf{B}}^{(q)}$ have already been computed. This is not the case for the k-means inverse regression (KIR) which requires new computations. In addition, applying PMS instead of MSIR or wMSIR on a cluster of $\mathbf{y}$ requires to store the $p \times p$ matrices $\widehat{\mathbf{\Gamma}}^{(1)}, \ldots, \widehat{\mathbf{\Gamma}}^{(q)}$ and summing some of them, which represent more computational time and more memory space than required by MSIR or wMSIR method, especially when p is large.

4.3.5 A simulation study

This section illustrates the ability of the proposed MSIR and wMSIR approaches, together with the diagnostic procedures on the components of $\mathbf{y}$, to properly estimate EDR spaces. The two following subsections respectively correspond to models (4.9) and (4.10).

4.3.5.1 Single EDR space model

Two simulation models are considered here. For a given sample size n and a dimension p, 100 replications of the covariate $\mathbf{x}$ are generated from the p-dimensional normal distribution $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ with the same pair $(\boldsymbol{\mu}, \mathbf{\Sigma})$. Here, $\boldsymbol{\mu}$ is randomly generated from the $\mathcal{N}_p(0, \mathbf{I}_p)$ distribution and $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}' + 0.1\mathbf{I}_p$ where $\mathbf{L}$ is a $p \times p$ matrix made of entries independently generated from the standard normal distribution $\mathcal{N}_1(0, 1)$. For every model, for all $j \geq 1$, it is assumed that $\varepsilon^{(j)} \sim \mathcal{N}_1(0, 1)$. The dimension K is also assumed to be known in this section. We first study a special case of model (4.9) with $K = 1$. Secondly, MSIR and wMSIR are evaluated with a multiple indices model ($K = 2$).

Single index model. Consider the following single index model:

$$\begin{cases} y^{(1)} &= \mathbf{x}'\beta_1 + \varepsilon^{(1)}, \\ y^{(2)} &= (\mathbf{x}'\beta_1)^3 + 3\varepsilon^{(2)}, \\ y^{(3)} &= \mathbf{x}'\beta_1(1 + \varepsilon^{(3)}), \end{cases} \tag{4.15}$$

with $\beta_1 = [\beta_{1,1}, \ldots, \beta_{1,p}]'$. We choose for all $i = 1, \ldots, p$, $\beta_{1,i} = i\ \mathbb{I}(i \leq 5) + \mathbb{I}(i > 5)$.

We generate samples of size $n = 100$ from (4.15), with $p = 20$. Then, the EDR direction is estimated using the following methods, with $H = 10$ slices:

- univariate SIR for each component of $\mathbf{y}$ which produces estimates $\widehat{\mathbf{B}}^{(1)}$, $\widehat{\mathbf{B}}^{(2)}$ and $\widehat{\mathbf{B}}^{(3)}$,
- MSIR which gives the estimate $\widehat{\mathbf{V}}$,
- wMSIR that leads to the estimate $\widehat{\widehat{\mathbf{V}}}$,
- k-means inverse regression which provides the estimate $\widehat{\mathbf{V}}_{\text{KIR}}$,
- pooled marginal slicing, leading to the estimate $\widehat{\mathbf{V}}_{\text{PMS}}$.

For each estimator $\widehat{\mathbf{B}} \in \left\{\widehat{\mathbf{B}}^{(1)}, \widehat{\mathbf{B}}^{(2)}, \widehat{\mathbf{B}}^{(3)}, \widehat{\mathbf{V}}, \widehat{\widehat{\mathbf{V}}}, \widehat{\mathbf{V}}_{\text{KIR}}, \widehat{\mathbf{V}}_{\text{PMS}}\right\}$, we compute the squared trace correlation $r(\widehat{\mathbf{B}}) := r(\widehat{\mathbf{B}}, \mathbf{\Sigma}, \mathbf{B}, \mathbf{\Sigma})$ between the estimated EDR space and the true

EDR space. The closer to one is $r(\widehat{\mathbf{B}})$, the better is the estimate. Note that for $K = 1$, the criterion $r(\widehat{\mathbf{B}})$ corresponds to the squared cosine of the angle between $\widehat{\mathbf{B}}$ and β_1.

Boxplots of this criterion are drawn on Figure 4.5(a). It appears that $\widehat{\mathbf{B}}^{(3)}$ exhibits low squared trace correlation. This phenomenon can be explained by the heteroscedasticity in the third marginal model of (4.15). Even if $\widehat{\mathbf{B}}^{(3)}$ is necessary to compute $\widehat{\mathbf{V}}$ and $\widehat{\widetilde{\mathbf{V}}}$, the poor estimates of $\mathbf{B}^{(3)}$ do not imply a significant loss in the squared trace correlations related to $\widehat{\mathbf{V}}$ and to $\widehat{\widetilde{\mathbf{V}}}$. One can also observe that the weighting in wMSIR seems to improve the estimation of the EDR space since values of $r(\widehat{\widetilde{\mathbf{V}}})$ are globally greater than those of $r(\widehat{\mathbf{V}})$. This trend is confirmed in Figure 4.5(b) where the values of $r(\widehat{\mathbf{V}})$ are plotted versus those of $r(\widehat{\widetilde{\mathbf{V}}})$. It appears that wMSIR seems to be uniformly better than MSIR in this simulation. In addition, Figure 4.5(a) shows that the pooled marginal slicing produces slightly better estimates than wMSIR in this example and that wMSIR outperforms the k-means inverse regression.

(a) Boxplots of quality measures $r(\widehat{\mathbf{B}})$ for various estimators $\widehat{\mathbf{B}}$

(b) Plot of $r(\widehat{\mathbf{V}})$ versus $r(\widehat{\widetilde{\mathbf{V}}})$

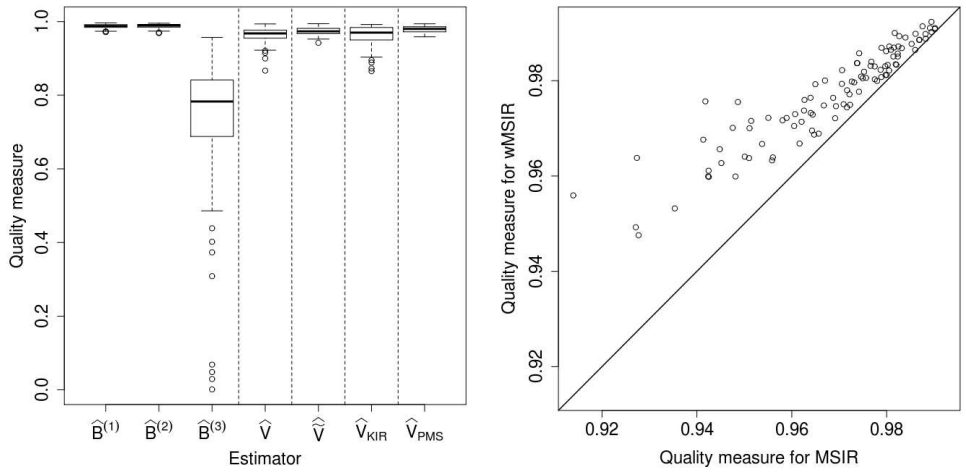


Figure 4.5: Comparison of estimators of $\mathbf{B}$ on 100 samples from model (4.15) with $n = 100$ and $p = 20$. The line in Figure 4.5(b) corresponds to the first bisecting line.

The poor quality of the estimate $\widehat{\mathbf{B}}^{(3)}$ can be observed directly from the simulated data. In Figure 4.6(a), we draw boxplots of values of $\hat{\tilde{r}}_j$ for $j = 1, 2, 3$. Considering that the first quartile of the third boxplot is equal to 0.80 and that the minimum value of the others is equal to 0.96, it makes sense to withdraw the third component of $\mathbf{y}$ from the analysis for at least a forth of the datasets. Let $\widehat{\widetilde{\mathbf{V}}}^{\star}$ be the wMSIR estimate built only from $y^{(1)}$ and $y^{(2)}$. In Figure 4.6(b), we compare quality measures for $\widehat{\widetilde{\mathbf{V}}}^{\star}$ and for the PMS estimate. We observe that the selection based on the $\hat{\tilde{r}}_j$'s improves performances of the wMSIR method, so that it produces better quality measures than the PMS without this selection step.

To study the sensitivity of wMSIR with respect to n and p, several samples are generated for various values of these parameters. In Figure 4.7(a), boxplots of 100 values

of $r(\widehat{\widetilde{\mathbf{V}}})$ are drawn for each pair (n, p). Not surprisingly, estimated EDR spaces become closer to the true one when n increases. Moreover, one can also observe that estimates of the EDR space are more precise when p is small. This phenomenon can be explained by the fact that $p \times p$ matrices have to be estimated. Notice that for every value of (n, p) in Figure 4.7(a), estimates are very precise since most of the values of $r(\widehat{\widetilde{\mathbf{V}}})$ are greater than 0.85.

In previous analyses, the number of slices H was set to 10 to have enough slices to study the functions that link $\mathbf{x}'\beta_1$ to each component of $\mathbf{y}$ and enough points in each slice. In Figure 4.8(a), we observe that for wMSIR, one can arbitrarily choose a number of slices between 8 and 38 and obtain an estimate of the EDR space which is as reliable as the one computed with $H = 10$.

Multiple indices model. Consider now a more complex model than model (4.15). It is defined by:

$$\begin{cases} y^{(1)} &= \exp(\mathbf{x}'\beta_1) \times (\mathbf{x}'\beta_2) + \varepsilon^{(1)}, \\ y^{(2)} &= (\mathbf{x}'\beta_1) \times \exp(\mathbf{x}'\beta_2) + \varepsilon^{(2)}, \end{cases} \tag{4.16}$$

where $\beta_1 = [\beta_{1,1}, \ldots, \beta_{1,p}]'$, $\beta_2 = [\beta_{2,1}, \ldots, \beta_{2,p}]'$, with $\beta_{1,i} = i\ \mathbb{I}(i \leq 5) + \mathbb{I}(i > 5)$ and $\beta_{2,i} = (-1)^{i-1}(1 + \mathbb{I}(i \in \{3, 4\}))$. Note that, in model (4.16), we clearly have $K = 2$.

Samples are generated from model (4.16) for various n and p. Then, these samples are used to estimate the corresponding EDR space with wMSIR with $H = 10$ slices. This leads to boxplots of $r(\widehat{\widetilde{\mathbf{V}}})$ displayed in Figure 4.7(b). We observe lower and more scattered values of $r(\widehat{\widetilde{\mathbf{V}}})$ than in Figure 4.7(a). The complexity of the link function between $\mathbf{y}$ and $\mathbf{x}$ and the greater dimension K are believable reasons for this phenomenon. Apart from this feature, Figure 4.7(b) provides identical evolutions of $r(\widehat{\widetilde{\mathbf{V}}})$ with n and p to those observed for the single index model in Figure 4.7(a). Note that for both model (4.15) and model (4.16), the behavior of MSIR and wMSIR when n and p vary are similar. That is why only results concerning wMSIR are drawn in Figure 4.7.

Examining Figure 4.8(b), it seems that the quality of wMSIR estimates is less connected to the chosen number of slices for model (4.15) than for model (4.16). For the latter, performances of wMSIR are nevertheless similar for values of H from 6 to 12.

4.3.5.2 Multiple EDR spaces model

Consider model (4.10) with $q = 12$, $p = 20$ and $K = 1$. Define for $i = 1, \ldots, p$, the ith component of respectively β_1, β_2 and β_3 as $\beta_{1,i} := i\ \mathbb{I}(i \leq 5) + \mathbb{I}(i > 5)$, $\beta_{2,i} := 6 - i + 5\left\lfloor \frac{i}{5} \right\rfloor$ and $\beta_{3,i} := (-1)^{i-1}(1 + \mathbb{I}(i \in \{3, 4\}))$. The random variable $\mathbf{y}$ is drawn from the following model:

(a) Values of $\hat{\tilde{r}}_j$ for each component $y^{(j)}$ of $\mathbf{y}$

(b) Plot of $r(\widehat{\widetilde{\mathbf{V}}}^{\star})$ versus $r(\widehat{\mathbf{V}}_{\mathrm{PMS}})$

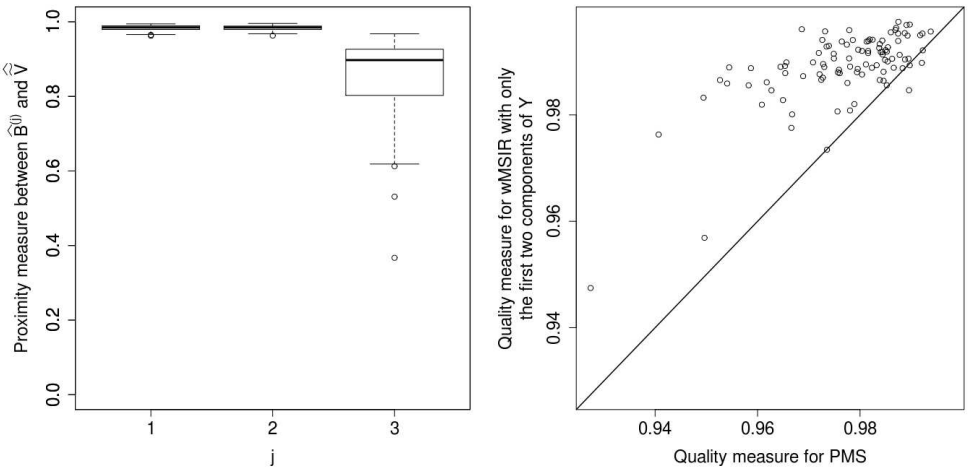


Figure 4.6: Study of the contribution of $y^{(3)}$ to $\widehat{\widetilde{\mathbf{V}}}$ for 100 samples generated from model (4.15) with $n = 100$ and $p = 20$.

(a) Model (4.15)

(b) Model (4.16)

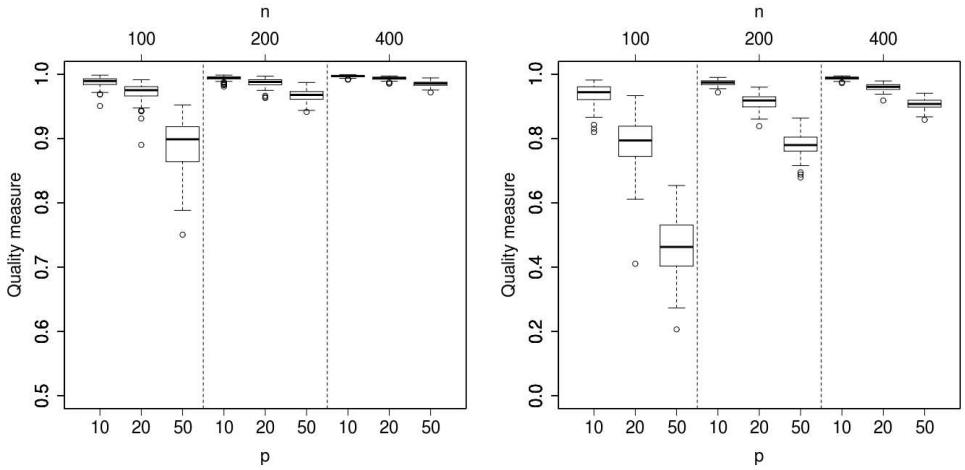


Figure 4.7: Boxplots of quality measure for the wMSIR method, for 100 samples generated from model (4.15) (Figure 4.7(a)) or from model (4.16) (Figure 4.7(b)) with various values of n and p.

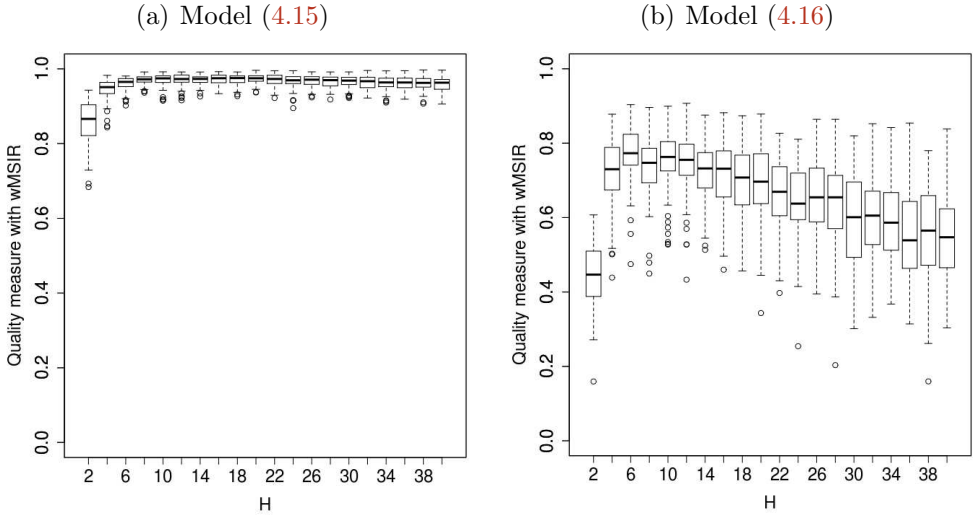


Figure 4.8: Boxplots of quality measure for the wMSIR method for 100 samples generated from the model (4.15) (Figure 4.8(a)) and from the model (4.16) (Figure 4.8(b)), with $n = 100$, $p = 20$ and various values of H.

$$\left\{\begin{array}{rcl} y^{(1)} &=& \mathbf{x}'\beta_1 + \varepsilon^{(1)}, \\ y^{(2)} &=& (\mathbf{x}'\beta_1)^3 + 3\varepsilon^{(2)}, \\ y^{(3)} &=& \mathbf{x}'\beta_1(1+\varepsilon^{(3)}), \\ y^{(4)} &=& \mathbf{x}'((1-\theta_1)\beta_1 + \theta_1\beta_2) + \varepsilon^{(4)}, \\ y^{(5)} &=& (\mathbf{x}'((1-\theta_1)\beta_1 + \theta_1\beta_2))^3 + 3\varepsilon^{(5)}, \\ y^{(6)} &=& \mathbf{x}'((1-\theta_1)\beta_1 + \theta_1\beta_2)(1+\varepsilon^{(6)}), \\ y^{(7)} &=& \mathbf{x}'((1-\theta_2)\beta_1 + \theta_2\beta_3) + \varepsilon^{(7)}, \\ y^{(8)} &=& (\mathbf{x}'((1-\theta_2)\beta_1 + \theta_2\beta_3))^3 + 3\varepsilon^{(8)}, \\ y^{(9)} &=& \mathbf{x}'((1-\theta_2)\beta_1 + \theta_2\beta_3)(1+\varepsilon^{(9)}), \\ y^{(10)} &=& \varepsilon^{(10)}, \\ y^{(11)} &=& \varepsilon^{(11)}, \\ y^{(12)} &=& \varepsilon^{(12)}, \end{array}\right. \tag{4.17}$$

based on model (4.15), where $(\theta_1, \theta_2) \in [0,1]$ and for $j = 1, \ldots, 12$, $\varepsilon^{(j)} \sim \mathcal{N}_1(0,1)$.

A model with 6 clusters. We first choose $\theta_1 = \theta_2 = 1$ which produces the $l = 6$ following clusters:

$$\left\{y^{(1)}, y^{(2)}, y^{(3)}\right\}, \left\{y^{(4)}, y^{(5)}, y^{(6)}\right\}, \left\{y^{(7)}, y^{(8)}, y^{(9)}\right\}, \left\{y^{(10)}\right\}, \left\{y^{(11)}\right\}, \left\{y^{(12)}\right\}.$$

We plot, in Figure 4.9, values of $\hat{r}_{j,j^\star}$ for $(j, j^\star) \in \{2, \ldots, q\} \times \{1, \ldots, j-1\}$, computed from a sample of size $n = 1000$. Darker squares correspond to values of $\hat{r}_{j,j^\star}$ close to 1. Thus, Figure 4.9 leads to cluster components $y^{(1)}$ and $y^{(2)}$ together as well as components $y^{(4)}$, $y^{(5)}$ and $y^{(6)}$. Components $y^{(7)}$ and $y^{(8)}$ can also be grouped together. It is tempting to put $y^{(3)}$ with $y^{(1)}$ and $y^{(2)}$ and to group $y^{(9)}$ with $y^{(8)}$ and $y^{(7)}$ because the related squared trace correlations are high but $\hat{r}_{3,9}$ is approximately as high as them. Note also that for $j \in \{1,2,3\}$, we have $\hat{r}_{3j,3j-1} < \hat{r}_{3j-2,3j-1}$ and $\hat{r}_{3j,3j-2} < \hat{r}_{3j-2,3j-1}$. Recalling that

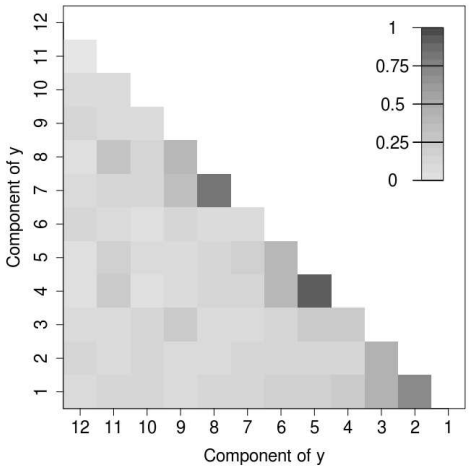


Figure 4.9: Values of $\hat{r}_{j,j^\star}$ in various shades of grays for $(j, j^\star) \in \{2, \ldots, q\} \times \{1, \ldots, j-1\}$ and for a sample of size $n = 1000$ generated from model (4.17) with $\theta_1 = \theta_2 = 1$ and $p = 20$.

for the considered values of j, $y^{(3j)}$ is the third component of each model (4.15) embedded in model (4.17), these inequalities can be explained by the heteroscedasticity in this component which leads to imprecise estimates of the relevant EDR direction as pointed out on Figure 4.5(a). In addition, on Figure 4.9, component $y^{(11)}$ could be unexpectedly grouped with $y^{(4)}$ and $y^{(5)}$ or with $y^{(8)}$.

In this example, the interpretation of Figure 4.9 can easily be done because components of $\mathbf{y}$ are already clustered in the definition of the model. In other words, every component between two others that are related to the same EDR space belongs to a marginal model based on this EDR space. In practical cases, the components of $\mathbf{y}$ may not be ordered that way which means that the corresponding representation of the squared trace correlations may be cluttered. To tackle this problem, we use an agglomerative hierarchical clustering algorithm based on the dissimilarity $1 - \hat{r}_{j,j^\star}$ between $\hat{\mathbf{B}}^{(j)}$ and $\hat{\mathbf{B}}^{(j^\star)}$.

A model with 5 clusters. We generate a new sample of size $n = 1000$ from model (4.17) with $\theta_1 = 1$ and $\theta_2 = 0$ which leads to $l = 5$ clusters:

$$\left\{y^{(1)}, y^{(2)}, y^{(3)}, y^{(7)}, y^{(8)}, y^{(9)}\right\}, \left\{y^{(4)}, y^{(5)}, y^{(6)}\right\}, \left\{y^{(10)}\right\}, \left\{y^{(11)}\right\}, \left\{y^{(12)}\right\}.$$

The hierarchical clustering algorithm applied to the estimates $\widehat{\mathbf{B}}^{(j)}$ for $j = 1, \ldots, q$ and produces the dendrogram of Figure 4.10(a). A classification directly based on this procedure would not be really accurate. For instance, in model (4.17), $y^{(11)}$ and $y^{(10)}$ belong to two different clusters. On Figure 4.10(a), to divide $y^{(11)}$ and $y^{(10)}$ into two groups, the tree has to be cut at a level around 0.75. This implies grouping $y^{(1)}$, $y^{(7)}$, $y^{(2)}$ and $y^{(8)}$ together and putting $y^{(3)}$ and $y^{(9)}$ in another group while components of both groups are actually related with the same EDR space. However, the dendrogram of Figure 4.10(a) allows to order components of $\mathbf{y}$ in such a way that those which are linked by high squared trace correlation are close from each other. Thus, in Figure 4.10(b), we displayed values of $\hat{r}_{j,j^\star}$ for $j^\star \prec j$ where $\prec$ denotes the ordering in Figure 4.10(a). It becomes clear, then, that components $y^{(1)}$, $y^{(7)}$, $y^{(2)}$, $y^{(8)}$, $y^{(3)}$ and $y^{(9)}$ should be clustered in the same group.

(a) Dendrogram of components of $\mathbf{y}$

(b) Values of $\hat{r}_{j,j^\star}$ in various shades of grays for $j^\star \prec j$

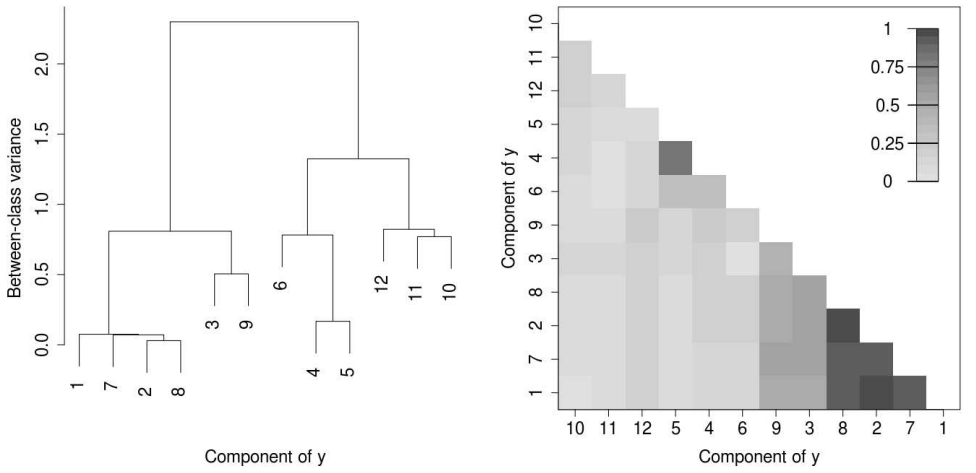


Figure 4.10: Clustering analysis for a sample of size $n = 1000$ generated from model (4.17) with $\theta_1 = 1$, $\theta_2 = 0$ and $p = 20$.

Note that another cluster containing $y^{(6)}$, $y^{(4)}$ and $y^{(5)}$ can be made from Figure 4.10(b). Not surprisingly, we observe again that the squared trace correlation between components $y^{(3)}$, $y^{(6)}$, $y^{(9)}$ and the components that correspond to the group they respectively belong to is quite low. Finally, components $y^{(12)}$, $y^{(11)}$ and $y^{(10)}$ appears to form three distinct clusters.

4.3.6 Real data illustrations

4.3.6.1 Remote sensing data

As an illustration, we consider a nonlinear inverse problem in remote sensing. The goal is to estimate the physical properties of surface materials on the planet Mars from hyperspectral data. The method is based on the estimation of the functional relationship between some physical parameters $\mathbf{x}$ and observed spectra $\mathbf{y}$. We refer to Bernard-Michel *et al.* (2009a) for further details. We focus on an observation of the south pole of Mars at the end of summer 2003, collected by the French imaging spectrometer OMEGA on board the Mars Express Mission. A detailed analysis of this image (Douté, Schmitt, Langevin, Bibring, Altieri, Bellucci, Gondet, and Poulet (2007)) revealed that this portion of Mars mainly contains water ice, carbon dioxide and dust. This led to the physical modeling of individual spectra with a surface reflectance model $\mathbf{y} = g(\mathbf{x})$. The $p = 3$ parameters $x^{(1)}$, $x^{(2)}$ and $x^{(3)}$ are respectively the proportion of carbon dioxide, the proportion of dust, and the grain size of water ice. Let us note that the proportion of water is equal to $1 - x^{(1)} - x^{(2)}$. Each spectra $\mathbf{y}$ is made of $q = 352$ wavelengths. The link function g has no close-form expression, but it can be computed thanks to a dedicated software. This yields the simulation of a sample $\{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, n\}$ of size $n = 6400$.

We ran the clustering procedure described in the above section associated to wMSIR with $H = 10$ slices. The clustering results are depicted on Figure 4.11. Two clusters of

(a) Values for $(j, j^\star) \in \{2, \ldots, q\} \times \{1, \ldots, j-1\}$

(b) Values for $j^\star \prec j$ where $\prec$ denotes an ordering produced by a hierarchical clustering algorithm

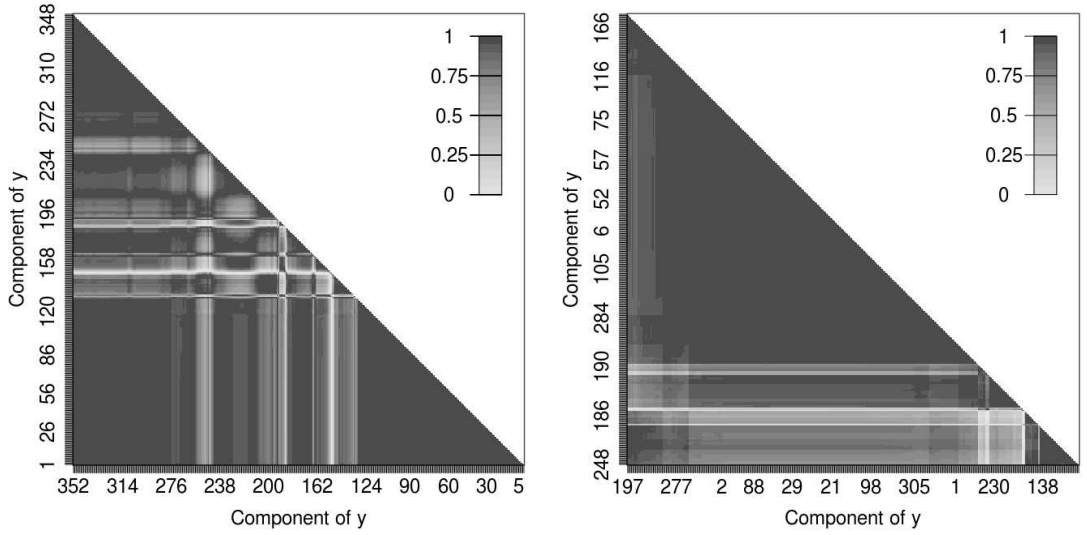


Figure 4.11: Values of $\hat{r}_{j,j^\star}$ in various shades of grays for hyperspectral data.

wavelengths have been identified, corresponding to two different orientations of $\widehat{\widetilde{\mathbf{v}}}_1$. It appears on Figure 4.12(a) that, in the first cluster, only the proportion of dust $x^{(2)}$ has an important contribution to the EDR direction. In the second cluster, the estimated EDR direction is close to $(1, 1, 0)$. This corresponds to the index $x^{(1)} + x^{(2)}$ which is equal to one minus the proportion of water. Let us note that the grain size of water ice $x^{(3)}$ does not appear in the EDR directions. Figure 4.12(b) permits to visualize the clustering of the wavelengths: It reveals which wavelengths are more sensible to the presence of water ice or dust.

4.3.6.2 Minneapolis elementary schools data

Another dataset which is widely studied in the dimension reduction context with a multivariate response is related to test results of students in Minneapolis elementary schools. These data are for example presented in Cook (2009); Cook and Setodji (2003); Yin and Bura (2006). The response variable $\mathbf{y}$ is made of $q = 4$ components. The first (resp. third) component is the proportion of pupils scoring below the average on a fourth (resp. sixth) grade test. The second (resp. fourth) one is the proportion of marks above the average. Following Yin and Bura (2006), we would like to explain $\mathbf{y}$ with a 8-dimensional variable $\mathbf{x}$. The seven first components $\mathbf{x}$ are called $x^{(1)}, \ldots, x^{(7)}$ and are respectively the squared root of the percentages of children receiving an aid called AFDC, children who do not live with both parents, people in the area of a school who completed high school, people who suffer for poverty, minority, mobility and pupils who attend school regularly. The last component of $\mathbf{x}$, named $x^{(8)}$, is the mean number of pupils for each teacher. Note that Yoo (2009) analyzed the variables $x^{(1)}$, $x^{(2)}$ and $x^{(3)}$ while Cook (2009) focused on some increasing functions of these variables and Cook and Setodji (2003) considered $x^{(1)}$, $x^{(2)}$, $x^{(3)}$, $x^{(4)}$ and $x^{(8)}$. We first compute $\widehat{\widetilde{\mathbf{V}}}$ with $p = 8$ components of $\mathbf{x}$. We

(a) Values of $\widehat{\tilde{v}}_{1,1}$ (circles), $\widehat{\tilde{v}}_{1,2}$ (triangles) and $\widehat{\tilde{v}}_{1,3}$ (crosses) for each cluster

(b) Clustering of the wavelengths

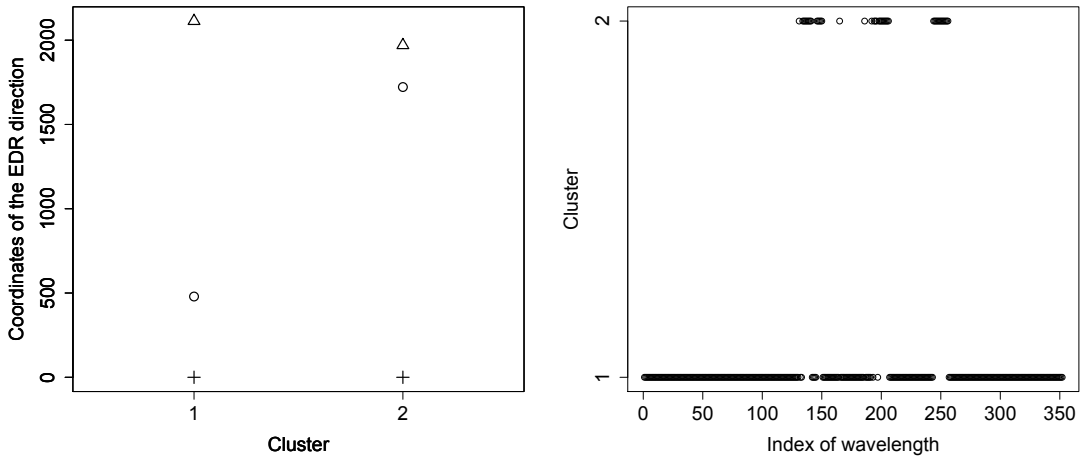


Figure 4.12: Clustering of the components of $\mathbf{y}$ and estimates $\widehat{\tilde{\mathbf{v}}}_1 = [\widehat{\tilde{v}}_{1,1}, \widehat{\tilde{v}}_{1,2}, \widehat{\tilde{v}}_{1,3}]'$ of the EDR direction for each cluster.

then use a stepwise algorithm with the AIC to perform a linear regression of $\mathbf{x}'\widehat{\widetilde{\mathbf{V}}}$ on $\mathbf{x}$ and sort the components of $\mathbf{x}$ with respect to how much they explain $\mathbf{x}'\widehat{\widetilde{\mathbf{V}}}$. The first three sorted variables are $x^{(1)}$, $x^{(2)}$, and $x^{(3)}$. Besides, regressing $\mathbf{x}'\widehat{\widetilde{\mathbf{V}}}$ on $\{x^{(1)}, x^{(2)}, x^{(3)}\}$ produces an adjusted coefficient of determination of 95.42%. This encourages us to take $\mathbf{x} = (x^{(1)}, x^{(2)}, x^{(3)})'$.

Working with this 3-dimensional covariate and with $H = 8$, we compute $\widehat{\mathbf{B}}^{(j)}$ for $j = 1, \ldots, 4$, with $K = 1$, which is the size of the dimension chosen in Cook (2009), Cook and Setodji (2003) and Yin and Bura (2006). We then construct the matrix $\widehat{\mathbb{B}}\widehat{\mathbb{B}}'\widehat{\mathbf{\Sigma}}$. Its eigenvalues are 0.96, 0.04 and 0.01 which confirms the choice of $K = 1$. We can not build several groups of components of $\mathbf{y}$ since the least value of $\hat{r}(j, j^\star)$ for $(j, j^\star) \in \{1, \ldots, 4\}$ is equal to 0.75, neither we can withdraw a component from $\mathbf{y}$ since the least value of $\hat{\tilde{r}}_j$ for $j \in \{1, \ldots, 4\}$ is equal to 0.92. In addition, we find $\widehat{\widetilde{\mathbf{V}}} = (0.673, -0.406, -0.528)'$. Let $\widehat{\mathbf{V}}_Y$ be the EDR direction found by Yoo (2009). We have that $r(\widehat{\widetilde{\mathbf{V}}}, \widehat{\mathbf{\Sigma}}, \widehat{\mathbf{V}}_Y, \widehat{\mathbf{\Sigma}}) = 0.97$. The signs of the elements of $\widehat{\widetilde{\mathbf{V}}}$ make also sense compared to results from Cook (2009) and Cook and Setodji (2003).

4.3.7 Concluding remarks

In this paper, we proposed the new multivariate SIR approaches MSIR and wMSIR for estimating the EDR space. The idea consists in performing first several marginal SIR analyzes. A common EDR space is then deduced from the marginal ones by maximizing the proximity criterion defined in (4.12). This optimization problem benefits from a closed-form solution.

Let us highlight that this two-step approach can be applied to any variant of the SIR method. For instance, it permits to build multivariate regularized SIR approaches

from Bernard-Michel *et al.* (2009b), multivariate SIR_α approaches from Gannoun and Saracco (2003) or multivariate kernel SIR methods from Wu (2008). To avoid the choice of the number H of slices, one may also consider building a multivariate version of the CUME procedure from Zhu *et al.* (2010a).

MSIR and wMSIR can also be run with a graphical procedure that cluster components of the response variable depending on the EDR space they are related with. Therefore, the approach we described allows to deal with datasets that come from a model that includes several different EDR spaces. It is then possible to estimate each of them from clusters of components of the response variable rather than blindly apply a multivariate SIR procedure on the whole variable. In addition, we noticed that estimating the EDR space does not cost a significant amount of computational time when it is done in the context of this clustering procedure. R codes are available on request from the authors.

Acknowledgments

The authors are grateful to Efstathia Bura for the Minneapolis schools data she provided us with. They also thank the associated editor and both anonymous referees for their useful comments that lead to several ameliorations of this article.

4.4 Comparison of sliced inverse regression approaches for underdetermined cases

This section is an article which is accepted for publication in *Journal de la Société Française de Statistique*. It was written with Benoît Liquet and Jérôme Saracco. Additional contents were nonetheless added in the simulation study of Section 4.4.4.

4.4.1 Introduction

For a univariate response variable y and a multivariate covariate $x \in \mathbb{R}^p$, the semiparametric regression model

$$y = f(x'\beta_1, \ldots, x'\beta_K, \varepsilon) \tag{4.18}$$

is an attractive dimension-reduction approach to model the effect of the p-dimensional covariates x on y. Let $\mu = \mathbb{E}(x)$ and $\Sigma = \mathbb{V}(x)$. The error term ε is assumed to be independent of x. Since the link function $f(\cdot)$ is an unknown smooth function, the parameters $\beta_k \in \mathbb{R}^p$ are not entirely identifiable, only the linear subspace spanned by the β_k's can be identified without additional assumptions. Duan and Li (1991) and Li (1991) called this subspace the effective dimension reduction (EDR) subspace. Moreover any direction belonging to this subspace is called an EDR direction. If the β_k's are assumed linearly independent, the EDR subspace is then a K-dimensional linear subspace of $\mathbb{R}^p$. Other authors refer to this subspace as the dimension reduction subspace (DRS) or the central subspace (which is defined as the smallest DRS), see Cook (1998) for more details.

When the dimension p of x is high and when we have little knowledge about the structure of the relationship between the response and the covariates, this semiparametric regression model is a nice alternative to parametric modeling (since it is really difficult to have knowledge about the structure of the relationship between the response and the covariates) and non-parametric modeling (which suffers from the well-known curse of dimensionality due to the data sparseness in the domain of x). The idea of dimension reduction in model (4.18) is intuitive because it aims at constructing a low dimensional projection of the covariate without losing information to predict the response y. If the dimension K of the EDR subspace is sufficiently small, it facilitates data visualization and explanation and it alleviates the curse of the dimensionality to non-parametrically estimate f with usual approaches such as kernel or splines smoothing (when the error term is additive).

In this semiparametric regression model (4.18), an important purpose is to estimate the EDR subspace from a sample $\{(x_i, y_i),\ i = 1, \ldots, n\}$. Most of the existing approaches are usually based on the eigendecomposition of a specific matrix of interest. The most popular one is the sliced inverse regression (SIR) introduced by Duan and Li (1991) and Li (1991) , respectively for single index models ($K = 1$) and multiple indices models ($K \geq 1$). Among alternative methods there are SIR-II, see Li (1991); Yin and Seymour (2007) for instance, and sliced average variance estimation (SAVE), see Zhu and Zhu (2007); Li and Zhu (2007) for example. These approaches require the inverse of Σ. Then, from a practical point of view, it is necessary to inverse an estimate $\widehat{\Sigma}$ of Σ.

Define $\tilde{x}_i = (x_i - \hat{\mu}) \in \mathbb{R}^p$ for $i = 1, \ldots, n$, with $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$. A usual (biased)

estimate of Σ is

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})' = \frac{1}{n}(\tilde{x}_1, \ldots, \tilde{x}_n)(\tilde{x}_1, \ldots, \tilde{x}_n)'. \tag{4.19}$$

Clearly, the rank of the $p \times p$ matrix $\widehat{\Sigma}$ is at most equal to $n-1$ since $\sum_{i=1}^{n} \tilde{x}_i = 0_p$ where 0_p stands for the null vectors of $\mathbb{R}^p$. From this remark on the rank of $\widehat{\Sigma}$, this matrix is singular when $n < p$. Moreover, it is also often ill-conditioned when $n \approx p$.

Therefore, SIR, SIR-II or SAVE methods only work well when the sample size n is greater than the dimension p of the covariate x, but naturally fail when $n < p$. In this underdetermined case, the standard estimate of Σ is not invertible even if the components of x are independent.

In the following, we only focus on the SIR approach. We describe it in Section 4.4.2.1 when $n > p$. The goal of this paper is then twofold. On one hand, we present methods to tackle the issue $n < p$. On the other hand, we also provide procedures in order to select which components of x have an effect on y.

In Section 4.4.2.2, we consider two different regularizations added to the SIR method, proposed by Zhong *et al.* (2005) and Li and Yin (2008), to find EDR estimates when $n < p$. Moreover, the SIR method can be seen as a generalized eigenvalue problem and linear algebra algorithms exist to solve this kind of problem without requiring any matrix inversion. The QZ algorithm (see Moler and Stewart (1973) for instance) is one of them and will be used in the SIR context in Section 4.4.2.3. In Section 4.4.2.4, we also adapt an approach introduced in functional sliced inverse regression (i.e., when x is an explanatory functional variable), based on the Moore-Penrose pseudo-inverse.

Concerning the selection of useful predictors in the indices, Zhong *et al.* (2005) use a chi-square test to find which components of x affect y, while the approach of Li and Yin (2008) relies on a Lasso penalization (Section 4.4.3.1). In Section 4.4.3.2, we propose another procedure. We choose randomly some submodels (i.e., using a number $p^0 < p$ of components of x) and we measure how close they are from the initial one with all the p components of x. The latter model is thus taken as a benchmark. Components of x that appear the most in submodels that are the closest to the benchmark are kept. We naturally consider that the other components of x do not affect y.

In Section 4.4.4, we compare in a simulation study the numerical behavior of the described methods to estimate EDR directions. We also evaluate the different procedures of selection of the useful components of x. In Section 4.4.5, we apply the most efficient one on real data from a genetic framework. Finally, some concluding remarks are given in Section 4.4.6.

4.4.2 SIR in determined and underdetermined cases

4.4.2.1 Brief review of usual SIR

Let β be a $p \times K$ matrix defined by $\beta = (\beta_1, \ldots, \beta_K)$. The EDR subspace is thus spanned by β.

Inverse regression step. The basic principle of the SIR method is to reverse the role of y and x, that is, instead of regressing the univariate variable y on the multivariate

variable x, the covariable x is regressed on the response variable y. The price we have to pay to succeed in inverting the role of x and y is an additional assumption on the distribution of x, named the linearity condition (described hereafter).

Usual SIR estimate is based on the first moment $\mathbb{E}(x|y)$. It has been initially introduced by Duan and Li (1991) for single index model and by Li (1991) for multiple indices model. SIR approaches have been extensively studied, see for instance Carroll and Li (1992); Chen and Li (1998); Zhu *et al.* (2007); Bercu *et al.* (2011); Azaïs *et al.* (2012) among others.

Let us now recall the geometric property on which SIR is based. Let us introduce the linearity condition:

$$(LC): \quad \forall b \in \mathbb{R}^p, \mathbb{E}(x'b|x'\beta_1, \ldots, x'\beta_K) \text{ is linear in } x'\beta_1, \ldots, x'\beta_K. \tag{4.20}$$

Note that this condition is satisfied when x is elliptically distributed (for instance normally distributed). The reader can find an interesting discussion on this linearity condition in Chen and Li (1998).

Assuming model (4.18) and (LC), Li (1991) showed that the centered inverse regression curve is contained in the linear subspace spanned by the K vectors $\Sigma\beta_1, \ldots, \Sigma\beta_K$. Let T denote a monotonic transformation of y. He considered the eigendecomposition of the Σ-symmetric matrix $\Sigma^{-1}M$ where $M = \mathbb{V}(\mathbb{E}(x|T(y)))$. Straightforwardly the eigenvectors associated with the largest K eigenvalues of $\Sigma^{-1}M$ are some EDR directions.

Slicing step. To easily estimate the matrix M, Li (1991) proposed a transformation T, called a slicing, which categorizes the response y into a new response with $H > K$ levels (in order to avoid an artificial reduction of dimension). The support of y is partitioned into H non-overlapping slices $s_1, \ldots, s_h, \ldots, s_H$. With such transformation T, the matrix of interest M can be now written as $M = \sum_{h=1}^{H} p_h(m_h - \mu)(m_h - \mu)'$ where $p_h = \mathbb{P}(y \in s_h)$ and $m_h = \mathbb{E}(x|y \in s_h)$.

Estimation process. When a sample $\{(x_i, y_i), i = 1, \ldots, n\}$ is available, matrices Σ and M are estimated by substituting empirical versions of the moments for their theoretical counterparts. Let

$$\widehat{M} = \sum_{h=1}^{H} \hat{p}_h(\hat{m}_h - \hat{\mu})(\hat{m}_h - \hat{\mu})', \tag{4.21}$$

where $\hat{p}_h = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[y_i \in s_h]$ and $\hat{m}_h = \frac{1}{n\hat{p}_h}\sum_{i=1}^{n} x_i\mathbb{I}[y_i \in s_h]$. Therefore the estimated EDR directions are the eigenvectors associated with the K largest eigenvalues of $\widehat{\Sigma}^{-1}\widehat{M}$. They span the K-dimensional estimated EDR subspace The convergence at rate $\sqrt{n}$ and the asymptotic normality of estimated EDR directions have been obtained, see Li (1991); Saracco (1997) for instance.

The choice of the slicing T is discussed in Li (1991); Kötter (2000); Saracco (2001) but, theoretically, there is no optimal one. In practice, we fix the number of observations per slice to $\lfloor n/H \rfloor$ where $\lfloor a \rfloor$ stands for the integer part of a. If the sample size n is not proportional to the number H of slices, some slices will then contain $\lfloor n/H \rfloor + 1$ observations. Note that, in order to avoid the choice of a slicing, alternative SIR methods

have been investigated. For instance, one can mention kernel-based methods of SIR proposed by Zhu and Fang (1996) or Aragon and Saracco (1997). However, these methods are hard to implement and are computationally slow. Moreover, Bura (1997) and Bura and Cook (2001) proposed a parametric version of SIR.

Concerning the determination of the dimension K of the EDR subspace (which is unknown in practice), several works are available in the literature, see for example Li (1991); Schott (1994); Ferré (1998); Bai and He (2004); Liquet and Saracco (2008) among others.

Standardized version for SIR. Another way to obtain a basis of the EDR subspace is to consider the eigendecomposition of $\Sigma^{-1/2}M\Sigma^{-1/2}$, that is the eigendecomposition of $M^* = \mathbb{V}(\mathbb{E}(z|T(y))$ where $z = \Sigma^{-1/2}(x-\mu)$ is the standardized version of the covariate x. For the multiple indices model (4.18), we then focus on the first K eigenvectors $\eta_1, \ldots, \eta_K$ associated with the largest K eigenvalues of the I_p-symmetric matrix M^*. Transforming back to the original scale, the vectors $\Sigma^{-1/2}\eta_k$, $k = 1, \ldots, K$ are in the EDR subspace. Their estimation procedure is a straightforward replication of the previous estimation process using $\widehat{M^*} = \widehat{\Sigma}^{-1/2}\widehat{M}\widehat{\Sigma}^{-1/2}$.

4.4.2.2 Two existing SIR methods when $n < p$

As previously mentioned, the rank of $\widehat{\Sigma}$ implies that this matrix is singular when $n < p$ and ill-conditioned when $n \approx p$. In this section we present two methods, respectively from Zhong *et al.* (2005) and Li and Yin (2008), to tackle these cases.

RSIR: A modified estimated variance matrix. Zhong *et al.* (2005) introduce an upgrade of the SIR method, called RSIR, that relies on a modification of $\widehat{\Sigma}$ such that the result can be inverted. This leads to the following estimate of Σ:

$$\widetilde{\Sigma}(s) = \widehat{\Sigma} + sI_p,$$

where s is a positive real parameter and I_p is the $p \times p$ identity matrix. For a given matrix A, let $\|A\|^2 = \text{Trace}(A'A)$. To find a suitable s, Zhong *et al.* (2005) propose to minimize the mean squared error

$$L(s) = \sum_{k=1}^{K} \text{Trace}(\mathbb{V}(\hat{\beta}_k(s))) + \sum_{k=1}^{K} \|\mathbb{E}(\hat{\beta}_k(s)) - \beta_k\|^2,$$

where $\hat{\beta}(s) = (\hat{\beta}_1(s), \ldots, \hat{\beta}_K(s))$ is the matrix of the K first generalized eigenvector of $\widehat{M}$ and $\widetilde{\Sigma}(s)$, which is built such that the constraint $\hat{\beta}_k(s)'\widetilde{\Sigma}(s)\hat{\beta}_{\tilde{k}}(s) = \mathbb{I}[k = \tilde{k}]$ is verified for all $(k, \tilde{k}) \in \{1, \ldots, K\}^2$. More details about generalized eigenvectors can be found in Section 4.4.2.3. Because β_k is unknown, the authors replaced it with $\mathbb{E}(\hat{\beta}_k(s_0))$ in the expression of $L(s)$, to obtain an approximation $\widetilde{L}(s)$. Note that the parameter s_0 has to be sufficiently small in order for $\mathbb{E}(\hat{\beta}_k(s_0))$ to be close to β_k. In practice, s_0 is chosen equal to 0. Variances and expectations in $\widetilde{L}(s)$ are then estimated with bootstrap samples, which leads to an estimate $\widehat{\widetilde{L}}(s)$ of $\widetilde{L}(s)$. Remark that estimating $\mathbb{E}(\hat{\beta}_k(s_0))$ for $s_0 = 0$,

implies using SIR with $\widehat{\Sigma}$. To do so, Zhong *et al.* (2005) apply the QZ algorithm (see Section 4.4.2.3 for details). The optimal regularization parameter is then given by

$$s_{opt} = \arg\min_{s} \widehat{\widetilde{L}}(s).$$

The corresponding matrix of estimated EDR directions is finally defined by $\hat{\beta}_{\text{RSIR}} = \hat{\beta}(s_{opt})$.

SR-SIR: A ridge sliced inverse regression. We describe here the SR-SIR method from Li and Yin (2008). When $\widehat{\Sigma}$ is invertible, let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_K)$ be the $p \times K$ matrix made of the eigenvectors of $\widehat{\Sigma}^{-1}\widehat{M}$. According to Li and Yin (2008) (see also Cook (2004)), $\hat{\beta}$ also satisfies

$$\left(\hat{\beta}, \widehat{V}\right) = \arg\min_{u,v} \sum_{h=1}^{H} \hat{p}_h \left\| (\hat{m}_h - \hat{\mu}) - \widehat{\Sigma} u v_h \right\|^2, \tag{4.22}$$

with $v = (v_1, \ldots, v_h)$ and where the minimum is taken over the respective sets of $p \times K$ matrices and $K \times H$ matrices. Note that this equation is also defined when $\widehat{\Sigma}$ is not invertible. From (4.22), the authors proposed thus a ridge version of the estimator $\hat{\beta}$ for a given regularization parameter s:

$$\left(\hat{\beta}(s), \widehat{V}(s)\right) = \arg\min_{u,v} G_s(u, v). \tag{4.23}$$

where

$$G_s(u, v) = \sum_{h=1}^{H} \hat{p}_h \left\| (\hat{m}_h - \hat{\mu}) - \widehat{\Sigma} u v_h \right\|^2 + s\|u\|^2$$

In practice, $\hat{\beta}(s)$ can be then obtained from (4.23) with an alternating least-squares algorithm even when $n < p$. The SR-SIR method rely on a generalized crossvalidation criterion to find the optimal regularization parameter s_{opt} (see Li and Yin (2008) for details). Finally the matrix of estimated EDR directions is defined by $\hat{\beta}_{\text{SR-SIR}} = \hat{\beta}(s_{opt})$.

Remark 12. The exisence of a solution for (4.23) is not proved as explained by Bernard-Michel *et al.* (2008). Indeed, assume that $\left(\hat{\beta}(s), \widehat{V}(s)\right)$ is such a solution and that $\hat{\beta}(s)$ is not the null vector, we then have

$$G_s\left(\frac{1}{2}\hat{\beta}(s), 2\widehat{V}(s)\right) < G_s\left(\hat{\beta}(s), \widehat{V}(s)\right),$$

which contradicts the fact that $\left(\hat{\beta}(s), \widehat{V}(s)\right)$ verifies (4.23). This encourages Bernard-Michel *et al.* (2008) to replace (4.23) with the following optimization problem:

$$\left(\hat{\beta}(s), \widehat{V}(s)\right) = \arg\min_{u,v} \left\{ \sum_{h=1}^{H} \hat{p}_h \left\| (\hat{m}_h - \hat{\mu}) - \widehat{\Sigma} u v_h \right\|^2 + s \left\| u v \widehat{W}^{1/2} \right\|^2 \right\},$$

where $\widehat{W} = \text{diag}(\hat{p}_1, \ldots, \hat{p}_H)$. The value of $\hat{\beta}(s)$ in this problem is actually the estimate of the RSIR method, for a regularization parameter s.

4.4.2.3 SIR-QZ: Solving the generalized eigenvalues problem in SIR

When $\widehat{\Sigma}$ is regular, usual SIR estimates of the EDR directions are eigenvectors of $\widehat{\Sigma}^{-1}\widehat{M}$. This eigendecomposition is actually a special case of a generalized eigenvalues problem which consists in finding real numbers λ and non-null vectors v such that:

$$\widehat{M}v = \lambda\widehat{\Sigma}v. \tag{4.24}$$

The generalized Schur decomposition. When $\widehat{\Sigma}$ is singular the generalized eigenvalue problem can still be solved if the function $\lambda \mapsto \widehat{M} - \lambda\widehat{\Sigma}$ behave properly. We call this function a matrix pencil. In this section, we present the QZ algorithm which allows us to find couples (λ, v) that verify (4.24) for a wide range of matrix pencils including some with singular matrices $\widehat{\Sigma}$. The QZ algorithm can be viewed as an extension of the QR algorithm and was proposed by Moler and Stewart (1973). The reader can refer to chapter 7 of Golub and Van Loan (1983) for details. A brief description of this algorithm is provided in the following.

Notice that if we have two invertible matrices Q and Z, then finding λ and v in (4.24) is equivalent to find λ and w in

$$Q\widehat{M}Zw = \lambda Q\widehat{\Sigma}Zw, \tag{4.25}$$

and to set $v = Zw$. Similarly to the QR algorithm that is designed to find the Schur decomposition of a matrix in order to compute its eigenvalues, the QZ algorithm aims at finding unitary matrices Q and Z such that $Q\widehat{M}Z$ and $Q\widehat{\Sigma}Z$ are upper triangular, for square matrices $\widehat{M}$ and $\widehat{\Sigma}$. Such a transformation is called a generalized Schur decomposition. When working with complex matrices, Q and Z always exist (see Golub and Van Loan (1983), Theorem 7.7-1). Possible values of λ that verify (4.24) are such that $\det(\widehat{M} - \lambda\widehat{\Sigma}) = 0$, and such that $\det(Q(\widehat{M} - \lambda\widehat{\Sigma})Z) = 0$. The latter determinant is the product of the diagonal elements of $Q(\widehat{M} - \lambda\widehat{\Sigma})Z$ since it is an upper triangular matrix. Hence, the generalized eigenvalues of (4.24) are the ratios of the diagonal elements of $Q\widehat{M}Z$ to the ones of $Q\widehat{\Sigma}Z$, provided that the diagonal elements of $Q\widehat{\Sigma}Z$ are not equal to zero. More specifically, this can be seen with the following formula (Golub and Van Loan (1983), Theorem 7.7-1)

$$\det(\widehat{M} - \lambda\widehat{\Sigma}) = \det(Q'Z')\prod_{j=1}^{p}(t_j - \lambda u_j), \tag{4.26}$$

where $t_1, \ldots, t_p$ and $u_1, \ldots, u_p$ are the respective diagonal elements of $Q\widehat{M}Z$ and $Q\widehat{\Sigma}Z$. Notice that the generalized Schur decomposition only produces complex upper triangular matrices $Q\widehat{M}Z$ and $Q\widehat{\Sigma}Z$. However, there is a similar available decomposition for real matrices $\widehat{M}$ and $\widehat{\Sigma}$ (see Appendix A.2.1 or Golub and Van Loan (1983) for details).

Estimating the indices $X'\beta$ using the QZ algorithm. Equation (4.26) implies that if it exists $j \in \{1, \ldots, p\}$ such that $t_j = u_j = 0$, then $\det(\widehat{M} - \lambda\widehat{\Sigma}) = 0$ for all $\lambda \in \mathbb{C}$, and trying to choose eigenvectors corresponding to the greatest eigenvalues to estimate

the EDR directions does not make sense. Numerically, for any $j \in \{1, \dots, p\}$, due to rounding errors, t_j and u_j are almost always different from 0, but if both their absolute value are too small, $\det(\widehat{M} - \lambda\widehat{\Sigma})$ is sufficiently unstable to call its value in question. As a consequence, every computed λ_j can be wrong. For similar reasons, if $|u_j|$ is too small for a given $j \in \{1, \dots, p\}$, t_j/u_j should not be considered as an eigenvalue. These remarks and the regularization procedure in Zhong *et al.* (2005) lead to the algorithm we provide in Appendix A.2.2 to find an estimate $\hat{\beta}_{\text{QZ}}$ of the EDR directions.

Let $X = (x_1, \dots, x_n)$. When n is sufficiently smaller than p and $H > K$, a generalized eigenvector v of (4.24) is such that the n indices $X'v$ only takes H distinct values, as explained in Appendix A.2.3. In practice, the regularization parameter s of the algorithm of Appendix A.2.2 is small and we can distinguish easily H clusters in the values of $X'\hat{\beta}_{\text{QZ}}$ in Figure A.3 in Appendix A.2.3. This is a drawback of our approach when $n < p$ since the values of $X'\beta$ are a priori distinct. To circumvent this shortcoming, we compute several $\hat{\beta}_{\text{QZ}}$ with different number of slices $H_1, \dots, H_{N_H}$. Let denote the corresponding estimates $\hat{\beta}_{\text{QZ},1}, \dots, \hat{\beta}_{\text{QZ},N_H}$. We would like to find a K-dimensional subspace of $\mathbb{R}^n$ which is as close to the KN_H points of the matrix $X'(\hat{\beta}_{\text{QZ},1}, \dots, \hat{\beta}_{\text{QZ},N_H})$ as possible. Thus, we would choose a basis $\hat{\gamma}$ of this subspace as an estimate of $X'\beta$. This leads us to consider the following equation:

$$\left(\hat{\gamma}, \hat{\delta}\right) = \arg\min_{\gamma,\delta} \left\| X'(\hat{\beta}_{\text{QZ},1}, \dots, \hat{\beta}_{\text{QZ},N_H}) - \gamma\delta \right\|^2 \tag{4.27}$$

where the minimum is taken over the respective sets of $n \times K$ matrices and $K \times KN_H$ matrices and each column of $\gamma\delta$ is the approximation of the corresponding column of $X'(\hat{\beta}_{\text{QZ},1}, \dots, \hat{\beta}_{\text{QZ},N_H})$ in the K-dimensional subspace spanned by γ. A solution of (4.27) is given by a principal component analysis (see Besse (2012), p80-81). Note that their exists an infinite number of bases $\hat{\gamma}$ which span the optimal K-dimensional subspace. Thus, the solution provided by the principal component analysis is just one of them. We call the whole approach SIR-QZ.

Remark 13. When n is smaller enough than p, finding a satisfying estimate of β may not be possible. For example, if $K = 1$, $n < p$ and if the columns of X are not linearly dependent, there are infinitely many solutions u of the system $X'u = X'\hat{\beta}$ for a given estimate $\hat{\beta}$ of β. Recalling the underlying model (4.18), there is no reason why $\hat{\beta}$ should be a better estimate of β than any of these solutions. That is why, when $n < p$, we focus on estimates of $X'\beta$ rather than on β itself, in (4.27).

4.4.2.4 SIR-MP: a generalization of the inverse for singular matrices

We describe in this section a method which mimics the SIR approach developed for a functional covariate.

Dimension reduction in functional regression. In the functional SIR context, x is an explanatory functional variable (assumed square integrable in order to have its covariance operator well-defined) while y is still a real response variable. In this context, while the covariance operator of x is invertible, it has unbounded inverse so that its estimator is ill-conditioned. Then several methods have been proposed when the covariance operator does not need to be inverted.

One of them consists in using the eigendecomposition of $M^+\Sigma$ instead of $\Sigma^{-1}M$, where M^+ is the Moore-Penrose generalized inverse of M, also called Moore-Penrose pseudoinverse of M. In the particular context of functional sliced inverse regression, the reader can find a discussion on the fact that the eigenvectors of $\Sigma^{-1}M$ are eigenvectors of $M^+\Sigma$, in Ferré and Yao (2007) and references cited therein.

Let us now focus on an alternative approach introduced by Amato *et al.* (2006). They used the fact that $\Sigma^{-1/2}M\Sigma^{-1/2}$ is a finite rank operator, where Σ (resp. M) stands here for the covariance operator of x (resp. $\mathbb{E}(x|T(y))$) in this functional context. The eigenvectors of this operator are eigenvectors of $\Sigma^{1/2}M^+\Sigma^{1/2}$. The authors claimed that the reason of their approach is that a smooth estimate of M produces more stable estimates of the eigenvalue decomposition of M than that of the empirical estimate of Σ. Thus the eigenfunctions $\eta_1, \ldots, \eta_K$ associated with the smallest K eigenvalues $\alpha_1, \ldots, \alpha_K$ of $\Sigma^{1/2}M^+\Sigma^{1/2}$ are also the eigenfunctions associated with the largest eigenvalues of $\Sigma^{-1/2}M\Sigma^{-1/2}$ equal to $1/\alpha_k$ for $k = 1, \ldots, K$. In order to transform back to the original scale, we can not use the transformation $\Sigma^{-1/2}\eta_k$. A basis of the (functional) EDR space is instead given by

$$b_k = M^+\Sigma^{1/2}\eta_k \quad \text{for } k = 1, \ldots, K. \tag{4.28}$$

We provide in Appendix A.3.6 a brief proof of this result of Amato *et al.* (2006).

Adaptation for multivariate real covariates. In the context of our paper (that is, $n < p < \infty$), we will evaluate how the functional SIR procedure behaves in the multivariate framework. To do this, we simply substitute the operators of covariance M and Σ by the estimates $\widehat{M}$ and $\widehat{\Sigma}$ previously defined in (4.19) and (4.21). The resulting estimated directions are:

$$\widehat{b}_k = \widehat{M}^+\widehat{\Sigma}^{1/2}\widehat{\eta}_k \quad \text{for } k = 1, \ldots, K,$$

where the $\widehat{\eta}_k$'s are the eigenvectors of $\widehat{\Sigma}^{1/2}\widehat{M}^+\widehat{\Sigma}^{1/2}$ associated with the smallest eigenvalues (among those not structurally equal to zero, see Remark 14 for details). This adaptation is called SIR-MP.

Remark 14. For a p-dimensional covariate, the $p \times p$ matrix $\widehat{\Sigma}^{1/2}\widehat{M}^+\widehat{\Sigma}^{1/2}$ is symmetric positive semidefinite and its rank r is at most equal to $H-1$ when $H < n < p$. Therefore, the eigenvalues of $\widehat{\Sigma}^{1/2}\widehat{M}^+\widehat{\Sigma}^{1/2}$ are such that $\hat{\alpha}_1 \geq \cdots \geq \hat{\alpha}_r > 0$, and the geometric multiplicity of the eigenvalue zero is equal to $p-r$ by construction. Thus we are interested in the eigenvectors $\widehat{\eta}_k$ associated with the K eigenvalues $\hat{\alpha}_r, \ldots, \hat{\alpha}_{r-K+1}$.

4.4.3 Selecting relevant components of x which are linked with y

Let $\beta_{j,k}$ denote the jth element of the EDR direction β_k, for $k = 1, \ldots, K$ and $j = 1, \ldots, p$. If $\beta_{j,\cdot} = (\beta_{j,1}, \ldots, \beta_{j,K})$ is the null vector then the jth component of x does not have any effect on y. Finding such components is an important concern when $n < p$ because it allows $x \in \mathbb{R}^p$ to be reduced to $x^\star \in \mathbb{R}^{p^\star}$, where $p^\star < p$, without any loss of information. If in addition $p^\star$ is less enough than n, the EDR directions can then be accurately estimated with a classical SIR procedure applied on y and $x^\star$. In Section 4.4.3.1, we describe the

methods from Zhong *et al.* (2005) and Li and Yin (2008) to determine which $\beta_{j,.}$ are null. In Section 4.4.3.2, we introduce another method to solve this problem based on proximity measures between models with only a few components of x and the initial model (in which every component of x is taken into account).

4.4.3.1 Review of existing selection procedures

RSIR: Bootstrap estimates and a chi-squared test. Let $\hat{\beta}_{\text{RSIR},j,k}$ be the elements of the matrix $\hat{\beta}_{\text{RSIR}}$. Zhong *et al.* (2005) claim that for $j = 1, \ldots, p$, the vector $\hat{\beta}_{\text{RSIR},j,.} = (\hat{\beta}_{\text{RSIR},j,1}, \ldots, \hat{\beta}_{\text{RSIR},j,K})$ follows asymptotically a multivariate normal distribution with mean $\beta_{j,.}(s_{opt})$ and covariance matrix Γ_j. Provided that Γ_j can be inverted, if $\beta_{j,.}(s_{opt})$ is the null vector then $\hat{\beta}'_{\text{RSIR},j,.}\Gamma_j^{-1}\hat{\beta}_{\text{RSIR},j,.}$ follows asymptotically a chi-squared distribution with K degrees of freedom. This encouraged Zhong *et al.* (2005) to use a chi-squared test on $\hat{\beta}'_{\text{RSIR},j,.}\widehat{\Gamma}_j^{-1}\hat{\beta}_{\text{RSIR},j,.}$ to select which components of x have effect on y, where $\widehat{\Gamma}_j$ is an estimate of Γ_j computed from bootstrap estimates of $\beta_{j,.}(s_{opt})$.

Difficulties using this procedure could arise when $n < p$ because the distribution under the null hypothesis of this test is asymptotic. In addition, it leads to inferences about the vectors $\beta_{j,.}(s_{opt})$ for $j = 1, \ldots, p$ which are not necessary the same than for $\beta_{j,.}$.

SR-SIR: A Lasso method. From $\hat{\beta}(s_{opt})$ and $(\tilde{v}_1, \ldots, \tilde{v}_H) = \widehat{V}(s_{opt})$, given in equation (4.23), Li and Yin (2008) propose to minimize the following expression under a constraint on the L_1-norm of the vector φ:

$$G(\varphi) = \sum_{h=1}^{H} \left(\hat{p}_h \left\| (\hat{m}_h - \hat{\mu}) - \widehat{\Sigma}\text{diag}(\varphi)\hat{\beta}(s_{opt})\tilde{v}_h \right\|^2 \right).$$

This leads to the following optimization problem for a parameter $\tau > 0$:

$$\hat{\varphi}_\tau = \arg\min_{\varphi} \{G(\varphi)\}, \quad \text{s.t. } |\varphi| \leq \tau$$

where the minimum is taken over the set of vectors φ of length p. The Lasso procedure (Tibshirani (1996)) can be used to find $\hat{\varphi}_\tau$.

Li and Yin (2008) consider that a component of x that corresponds to a zero in $\hat{\varphi}_\tau$ does not have any effect on y. Let p_τ be the number of non-null elements of $\hat{\varphi}_\tau$. In practice, choosing τ implies choosing the amount of selection provided by $\hat{\varphi}_\tau$. To do so, Li and Yin (2008) propose to use classical model selection criteria. More specifically, this involves minimizing one of the following expression over a set of tested values of τ:

$$AIC = pH \log\left(\frac{G(\hat{\varphi}_\tau)}{pH}\right) + 2p_\tau,$$

$$BIC = pH \log\left(\frac{G(\hat{\varphi}_\tau)}{pH}\right) + \log(pH)p_\tau,$$

$$RIC = (pH - p_\tau)\log\left(\frac{G(\hat{\varphi}_\tau)}{pH - p_\tau}\right) + p_\tau(\log(pH) - 1) + \frac{4}{pH - p_\tau - 2}.$$

4.4.3.2 CSS based on SIR: Closest submodel selection for SIR methods

The idea of the procedure described here is to select submodels of (4.18) with only a given number p^0 of components of x which are the closest to the initial one. The components of x that appear the most in these submodels are asserted to have an effect on y.

Let $Y = (y_1, \ldots, y_n)'$. To do this, we propose the following algorithm.

Initialize $p^0 \in]1, p[$, $N_0 \in \mathbb{N}^*$ and $\zeta \in]0, 1[$ or $\rho \in]0, 1[$.

Step 1. Compute the estimated indices $\hat{\gamma} \in \mathbb{R}^n$ on Y and the whole covariate matrix X using SIR-QZ.

Let $a = 1$.

Step 2. Select randomly p^0 components of x and build the corresponding matrix $X^{(a)}$.

Step 3. Compute the SIR-QZ indices $\hat{\gamma}^{(a)} \in \mathbb{R}^n$ based on Y and $X^{(a)}$.

Step 4. Calculate the linear correlation between the indices $\hat{\gamma}$ and $\hat{\gamma}^{(a)}$. Let us denote by $\hat{c}^{(a)}$ the square of this correlation.

Let $a = a + 1$.

Repeat N_0 times steps 2-4.

Step 5. Consider the submodels corresponding to the N_1 largest correlations $\hat{c}^{(a)}$.

Either the user set $\zeta \in]0, 1[$ and then gets $N_1 = \zeta N_0$, or the user chose a value for ρ and then N_1 is the number of submodels such that $\hat{c}^{(a)} > \rho$.

Step 6. Count the number of occurrences of each component of x in these N_1 submodels. The components that affect y are the ones that have the greater number of occurrences.

For example, in our simulation study, we set $N_0 = 10^4$ and $\zeta = 10\%$ to determine the closest $N_1 = \zeta N_0$ submodels. In our real data application, we use $N_0 = 9 \times 10^5$ and $\rho = 0.9$ to select the top N_1 submodels.

Note that choosing $p^0 < n$ allows us to use classical SIR instead of SIR-QZ in Step 3 which significantly improves the computational time. In addition, any SIR approach that provides estimates of the indices (when $n \leq p$ and $n > p$) could be used in the whole algorithm instead of SIR-QZ.

4.4.4 A simulation study

In Sections 4.4.2-4.4.3, we presented 4 methods to estimate EDR directions (or indices) and 3 procedures to select which components of x have effects on y. In Section 4.4.4.1, we illustrate them on a single simulated data set. To compare their numerical performances, we then study them on several replications in Section 4.4.4.2.

4.4.4.1 Analysis of a single data set

Simulated model. We consider the following single index model

$$y = (x'\beta)^3 + \varepsilon \tag{4.29}$$

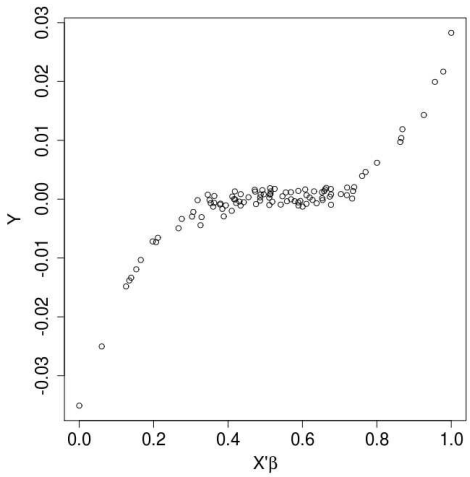


Figure 4.13: Plot of Y versus $X'\beta$ generated from model (4.29), with $n = 100$ and $p = 200$. The horizontal scale was standardized.

where x and β are p-dimensional vectors defined hereafter and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 10^{-3}$. Let $p = 200$ and $p^\star = 20$. We choose $\beta = \frac{1}{10}(\mathbb{I}(1 \leq p^\star), \ldots, \mathbb{I}(p \leq p^\star))'$, so that $p^\star$ is the number of non-null components of β, that is the number of components of x that affect y. We construct $x = (x_1, \ldots, x_p)'$ as follows: for $j = 1, \ldots, p^\star$, generate σ_j^2 from the law $\mathcal{U}([0.05, 0.1])$ and x_j from the law $\mathcal{N}(0, \sigma_j^2)$. For $j = p^\star + 1, \ldots, p$, set

$$\sigma_j^2 = \left(\frac{12 - \lfloor (j-1)/p^\star \rfloor}{\lfloor (j-1)/p^\star \rfloor} \right)^2 \sigma^2_{(j-1) \bmod p^\star + 1},$$

when mod denotes the modulo operation. Generate then $\check{x}_j$ from the law $\mathcal{N}(0, \sigma_j^2)$ and set $x_j = x_{(j-1) \bmod p^\star + 1} + \check{x}_j$. This ensures that $\mathrm{cor}(x_j, x_{(j-1) \bmod p^\star + 1}) = \lfloor (j-1)/p^\star \rfloor / 12$.

Estimation of EDR indices. We simulate an independent and identically distributed sample (X,Y) of size $n = 100$ from model (4.29). We plot Y versus the true indices $X'\beta$ in Figure 4.13. We analyze X and Y with the various methods presented in Section 4.4.2.2-4.4.2.4.

- For the RSIR method we evaluate $\widehat{\widetilde{L}}(s)$ for $s \in \{0, 10^{-10}, 10^{-9}, \ldots, 10^5\}$ with 50 bootstrap samples and $H = 10$. In Figure 4.14(a) we plot the values of Y against the indices provided by the RSIR method. The structure of Figure 4.13 can not be discerned in Figure 4.14(a). The regularization parameter that RSIR provides is equal to 10^5 and thus the RSIR procedure is equivalent to an eigendecomposition of $\widehat{M}$.

- Concerning the SR-SIR method, the regularization parameter s_{opt} is chosen among values in $\{10^{-10}, 10^{-9}, \ldots, 10^5\}$. The number of iterations of the alternating least square algorithm of SR-SIR is set to 50 and we take $H = 10$. For this example, we find $s_{opt} = 10^3$. In Figure 4.14(b), we draw the values of Y against the estimated indices $X'\hat{\beta}_{\text{SR-SIR}}$. The points of this graphic do not form the same shape as the points in Figure 4.13.

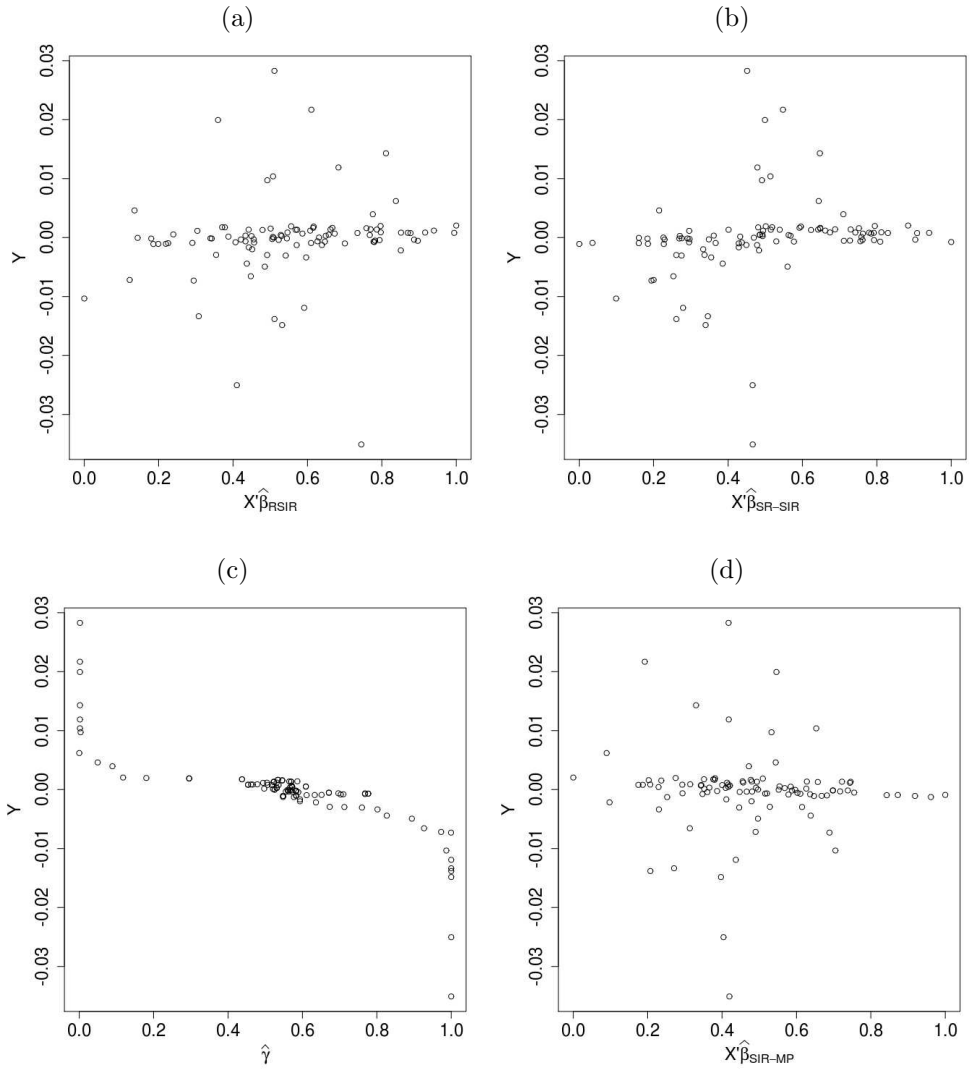


Figure 4.14: Plot of Y versus estimated indices obtained with the RSIR method (Figure 4.14(a)), the SR-SIR method (Figure 4.14(b)), the SIR-QZ method (Figure 4.14(c)) and the SIR-MP method (Figure 4.14(d)). The horizontal scale was standardized.

- We run SIR-QZ for $\{H_1, \ldots, H_{N_H}\} = \{5, \ldots, 15\}$. In Figure 4.14(c), we plot y against the corresponding estimated indices. This graphic exhibits a structure which is similar to the one in Figure 4.13.

- We finally apply SIR-MP with $H = 10$. We observe in Figure 4.14(d), which shows how Y and the indices produced by the SIR-MP method are related, that this method also fails to recover the shape of Figure 4.13.

Thus, Figure 4.14 shows that for this data set, SIR-QZ provides better estimations of the indices than RSIR, SR-SIR and SIR-MP.

To quantify such conclusions, we can use a criterion that measures how $X'\beta$ and $X'\hat{\beta}$ are close form each other, for a given estimate $\hat{\beta}$. Let $\widetilde{X} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ and let P be the

m	RSIR	SR-SIR	SIR-QZ	SIR-MP
$R(m)$	0.051	0.088	0.741	0.000
Computational time (s)	101.88	1,234.70	7.85	0.25

Table 4.1: Quality measure and computational time of various estimates of the indices $X'\beta$ for the simulated sample of size $n = 100$, with $p = 200$.

projector on the subspace of $\mathbb{R}^n$ spanned by $\widetilde{X}'\beta$. More precisely, we have

$$P = \widetilde{X}'\beta(\beta'\widetilde{X}\widetilde{X}'\beta)^{-1}\beta'\widetilde{X}. \tag{4.30}$$

Similarly, define P_{RSIR}, $P_{\text{SR-SIR}}$, and $P_{\text{SIR-MP}}$ by respectively replacing β by $\hat{\beta}_{\text{RSIR}}$, $\hat{\beta}_{\text{SR-SIR}}$, and $\hat{\beta}_{\text{SIR-MP}}$ in (4.30). Let us also define $P_{\text{SIR-QZ}}$ by

$$P_{\text{SIR-QZ}} = \bar{I}_n\hat{\gamma}(\hat{\gamma}'\bar{I}_n\hat{\gamma})^{-1}\hat{\gamma}'\bar{I}_n,$$

where $\bar{I}_n$ is a matrix that centers $\hat{\gamma}$ (see Appendix A.2.3). Note that for any $a \neq 0$, we also have $P_{\text{SIR-QZ}} = \bar{I}_n(a\hat{\gamma})((\hat{\gamma}'a)\bar{I}_n(a\hat{\gamma}))^{-1}(\hat{\gamma}'a)\bar{I}_n$, which is coherent since if $(\hat{\gamma}, \hat{\delta})$ is a solution of (4.27), then $(a\hat{\gamma}, \frac{1}{a}\hat{\delta})$ is another solution of this very equation. For a given method m, we use the squared trace correlation between the subspaces spanned by $\widetilde{X}'\beta$ and by $\widetilde{X}'\hat{\beta}_m$ as a measure of the closeness between these subspaces. It is defined by

$$R(m) = \frac{1}{K}\text{Trace}(PP_m). \tag{4.31}$$

Notice that if $K = 1$, $R(m)$ is the squared cosine of the angle between the vectors $X'\beta$ and $X'\hat{\beta}_m$. This quality measure belongs to $[0, 1]$ and the higher its value is, the better the indices are estimated.

In Table 4.1, we present values of $R(m)$ for the four considered methods. We also present, in this table, the computational time needed for our Intel Core 2 Quad Q9505 processor to execute these methods. For SIR-QZ, $R(m)$ is clearly higher than for RSIR, SR-SIR and SIR-MP (0.74 versus less than 0.1). Notice also that the computational time is a lot greater for RSIR and SR-SIR than for SIR-MP and SIR-QZ.

Selection of components of x. We rely on the true positive rate (TPR) and on the false positive rate (FPR) to evaluate procedures that find which elements of β are equal to 0. The TPR is the number of selected components of x that actually affect y divided by the total number of components of x that affect y. The FPR is the number of selected components of x that do not affect y divided by the total number of components of x that do not affect y.

For the RSIR selection method the returned p-values are ordered and the components that correspond to the first p-values are selected. For the CSS procedure, the components related to the greatest number of occurrence are selected. We evaluate the results of both methods by selecting the best 10 (resp. 20, 50, 100, 150) components. For the SR-SIR method, we use the different criteria (AIC, BIC and RIC) proposed in Li and Yin (2008) to determine the appropriate Lasso parameter.

The number of bootstrap samples generated for the RSIR method is set to 10^3. It is great enough so that increasing it does not improve significantly the quality criterion.

m		RSIR				CSS			
Estimates		$\hat{\beta}_{\text{RSIR}}$		$\hat{\beta}_{\text{QZ}}$		$\hat{\gamma}$		$x'\beta$	
Quality criteria		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Number of selected components	10	0.05	0.05	0.10	0.04	0.10	0.04	0.20	0.03
	20	0.15	0.09	0.25	0.08	0.35	0.07	0.40	0.07
	50	0.30	0.24	0.45	0.23	0.60	0.21	0.70	0.20
	100	0.40	0.51	0.70	0.48	0.75	0.47	0.90	0.46
	150	0.85	0.74	0.85	0.74	0.80	0.74	0.95	0.73
Computational time (s)		120.61		120.64		328.73		321.04	

m	SR-SIR					
Estimates	$\hat{\beta}_{\text{SR-SIR}}$		$\hat{\beta}_{\text{QZ}}$		β	
Quality criterion	TPR	FPR	TPR	FPR	TPR	FPR
Selection with AIC	0.00	0.13	0.00	0.16	0.90	0.00
Selection with BIC	0.00	0.03	0.00	0.07	0.65	0.00
Selection with RIC	0.00	0.03	0.00	0.04	0.65	0.00
Computational time (s)	35.53		40.88		33.29	

Table 4.2: True positive rate (TPR), false positive rate (FPR) and computational time of various methods run on the simulated sample of size $n = 100$, with $p = 200$, to determine which components of β of model (4.29) are not null.

This method relies on a regularization parameter which can also be provided by the Algorithm A.1 to find $\hat{\beta}_{\text{QZ}}$ (see Appendix A.2.2). The estimate $\hat{\beta}_{\text{QZ}}$ can thus be plugged in the RSIR selection method. For the CSS method, we choose $N_0 = 10^4$, $\zeta = 10\%$ and $p^0 = 50$. While increasing N_0 could lead to better quality criteria, the computational time is sufficiently large not to choose it greater. The tested values of the Lasso parameter for the SR-SIR method are in $\{1, 2, \ldots, 100\}$. This algorithm needs an estimate of β in input. We use $\hat{\beta}_{\text{SR-SIR}}$ but, because of the poor results of Table 4.1 for this estimate, we also take $\hat{\beta}_{\text{QZ}}$ with $H = 10$, and the true EDR direction β.

Results of the corresponding TPR and FPR are displayed in Table 4.2. The SR-SIR method performs very well if an accurate estimate of β is provided, while the results are really bad otherwise because no selected component of x has any effect of y. Concerning RSIR and the CSS method, their FPR are similar, but the TPR are greater for the latter. The results for RSIR with β_{QZ} are slightly better than with the full RSIR procedure. The CSS method also seems to need good estimates of the indices since working with the true ones produces better TPR than using $\hat{\gamma}$.

To get more insights about the numerical performances of the various procedures tested in this example, we run them in several replications in the following section.

4.4.4.2 General behaviors of the estimates over several replications.

We generate 100 samples of size $n = 100$ from the model (4.29). For each of them, we launch the RSIR, SR-SIR, SIR-QZ and SIR-MP procedures with the previously described parameters, and compute the quality criterion given in (4.31). We display the values of the criterion in Figure 4.15. The trend that is exhibited in Table 4.1 is confirmed in this graphic since the values of R(SIR-QZ) are clearly greater than the others.

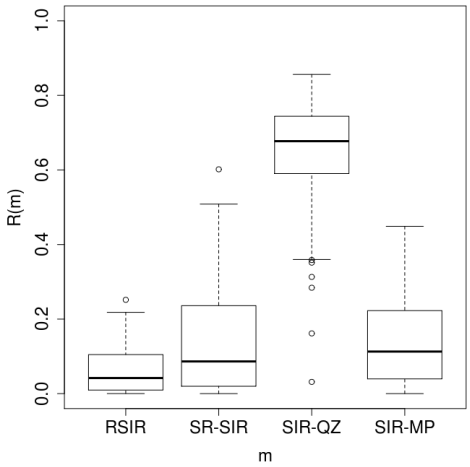


Figure 4.15: Boxplots of 100 values of $R(m)$ for various estimation procedures m and samples of size $n = 100$ generated from model (4.29) with $p = 200$.

Various methods to select important components of x are then run in the 100 samples drawn from model (4.29):

- RSIR with $\hat{\beta}_{\text{RSIR}}$, 1000 bootstrap samples and the following significance levels of the corresponding test: $0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$,
- SR-SIR with $\hat{\beta}_{\text{SR-SIR}}$ and $\tau \in \{1, 2, \ldots, 100\}$,
- The CSS procedure with $\hat{\gamma}$ from SIR-QZ and parameters $N_0 = 10^4$, $\zeta = 10\%$ and $p^0 = 50$ for various number of selected components: 10, 20, 30, 40, 60, 80, 100, 120, 140, 160.

The mean ROC curves over the 100 replications are displayed in Figure 4.16. The CSS method outperforms RSIR while SR-SIR provides poor results. Notice that for RSIR, the values of the FPR are close to the chosen levels of test, in spite of the fact that this test is asymptotic.

Remark 15. The simulated model (4.29) may appear complicated. We explain here why we chose it by considering the following one:

$$y = \text{sign}(x'\beta_1)\log(|x'\beta_2 + 5|) + \varepsilon, \tag{4.32}$$

where $p = 200$ and $\varepsilon \sim \mathcal{N}(0, 0.09)$. In addition, $\beta = (\beta_1, \beta_2)$ satisfies

$$\beta_1 = (1, 1, 1, 1, \overbrace{0, \ldots, 0}^{p-4})',$$

and

$$\beta_2 = (\overbrace{0, \ldots, 0}^{16}, 1, 1, 1, 1, \overbrace{0, \ldots, 0}^{p-20})'.$$

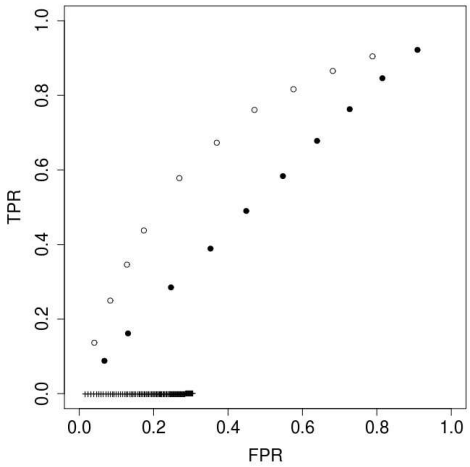


Figure 4.16: Mean ROC curves for various procedures to select components of x that affect y in the model of Section 4.4.4.1, over 100 replications, with $n = 100$ and $p = 200$. Solid circles: RSIR, crosses: SR-SIR, empty circles: SIR-CSS.

The p-dimensional random vector x is generated from the law $\mathcal{N}(0, \Sigma)$ where Σ is a $p \times p$ matrix created as follows. Let $D = \text{diag}(d_1, \ldots, d_p)$ and for $j = 1, \ldots, p$ define d_j by

$$d_j = \begin{cases} \frac{(j-1) \bmod 20+1}{10} & \text{if } (j-1) \bmod 20 \leq 9, \\ (j-1) \bmod 20 - 8 & \text{if } 9 < (j-1) \bmod 20 \leq 16, \\ 25 \times 2^{(j-1) \bmod 20-17} & \text{otherwise.} \end{cases}$$

Let also U_1 be a $p \times p$ matrix satisfying $\text{vec}(U_1) \sim \mathcal{N}(0, I_{pp})$. See Remark 10 for more information about the vec operator. Let U_2 be the $p \times p$ matrix corresponding to an orthogonal basis made of eigenvectors of $U_1 U_1'$. We then choose $\Sigma = U_2 D U_2'$.

Model (4.32) and the model from the simulation study in Zhong *et al.* (2005) are the same except that $p = 20$ in the latter one. For this value of p and $n = 500$, RSIR outperforms the other methods presented in Zhong *et al.* (2005).

Let SIR-I be the method that produces the estimate $\hat{\beta}_{\text{SIR-I}}$. We define it as the matrix that contains the K eigenvectors of $\widehat{M}$ which are related with the first K eigenvalues. This is equivalent to a classical SIR estimation, using $\widehat{\Sigma} = I_p$. We evaluate the performances of RSIR, SR-SIR, SIR-QZ, SIR-MP and SIR-I when trying to accurately estimate $X'\beta$ from samples (X, Y) of size $n = 50$ generated from model (4.32). To do so we compute $R(m)$ for each method m and 100 samples. The results are presented in Figure 4.17. There is no approach m such that $R(m)$ appears significantly higher than $R(\text{SIR-I})$. The choice of the parameter s in RSIR and SIR-QZ is thus not sensitive in this case and then comparing these approaches is difficult. That is why, in model (4.29) we introduced a covariance matrix Σ which differs a lot from I_p.

4.4.5 Real data application

4.4.5.1 Description of the Dataset

We illustrate our developed approach on a genetic dataset which contains transcriptomic data and genomic data. In this study, we aim at finding genetic causes of variation in the

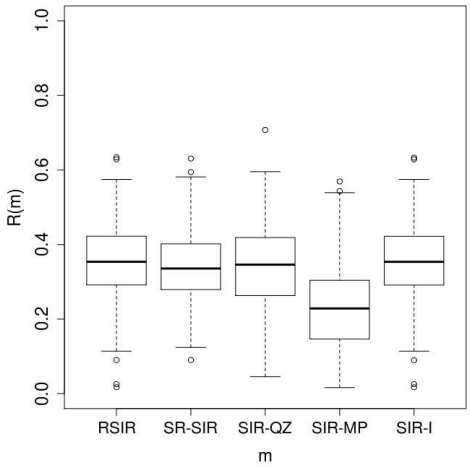


Figure 4.17: Boxplots of 100 values of $R(m)$ for various estimation procedures m and samples of size $n = 50$ generated from model (4.32) with $p = 200$..

expression of genes, that is eQTL (expression Quantitative Trait Loci). In this context, the gene expression data are treated as a quantitative phenotype and the genotype data (SNPs) are considered as predictors. In this illustration, we study the `Hopx` gene analyzed in Petretto *et al.* (2010). We investigate the ability of SIR-QZ combined with the CSS selection procedure to find a parsimonious index that explains the variability of `Hopx` gene expression in the heart tissue using $p = 770$ SNPs from $n = 29$ inbred line rats.

4.4.5.2 SIR-QZ and CSS results

We first run SIR-QZ for $\{H_1, \ldots, H_{N_H}\} = \{2, \ldots, 6\}$. In Figure 4.18(a), we plot the dependent variable `Hopx` versus the index based on the whole set of SNPs ($p = 770$). This graphic clearly exhibits a link between the phenotype and the index estimated by a smooth kernel method. In this illustration, this link is almost linear. In contrast, the plot (not given here) of `Hopx` gene versus the second EDR index does not show any structure. From this graphical diagnostic, it appears that only one EDR direction provides relevant information to explain the variability of the gene expression.

Then, in order to find a parsimonious index, we run our CSS selection procedure with $N_0 = 900.000$. The examined values of p^0 are in $\{10, 20\}$ while ρ takes value in $\{0.75, 0.80, 0.85, 0.90\}$. In Table 4.3, we present the number of selected SNPs for each combination of this two parameters. The threshold used for the selection will be detailed below. We can observe that, not surprisingly, the numbers of selected SNPs increases with p^0 and decreases with ρ. Moreover for a given value of p^0, we specify, in this table, the number of selected SNPs in common with those selected with $\rho = 0.9$ (corresponding to the parsimonious model). We also indicate, for a given ρ, the number of SNPs in common with those selected when $p^0 = 10$ and when $p^0 = 20$. This table highlights an overlap of 10 SNPs among all the sets of the selected SNPs for the various couple (p^0,ρ). Note that, the smallest set contains 11 SNPs when $p^0 = 10$ and $\rho = 0.9$ which comforts us about the stability of the CSS procedure.

In eQTL study, it is known that only a few number of SNPs can explain the variation of the gene expression. Thus, from the expertise of the biologists, we decide to select the

sparsest model, that is with $(p^0, \rho) = (10, 0.9)$. Figure 4.19 exhibits the selected 11 SNPs for this choice of p^0 and ρ. The threshold (horizontal red line in the figure) is defined as follows: $N_1 \frac{p^0}{p} + u_{1-\frac{\alpha/2}{p}} \sqrt{N_1 \frac{p^0}{p}(1 - \frac{p^0}{p})}$ where $u_{1-\frac{\alpha/2}{p}}$ is the quantile of order $(1 - \frac{\alpha/2}{p})$ of the standard normal distribution. It corresponds to the upper bound of the prediction interval of the occurrence of a SNP in the selected model under the hypothesis that none of the SNPs are associated with the gene expression. The level of this interval is fixed at $1 - \alpha = 0.95$ and is corrected by a Bonferroni approach.

On Figure 4.18(b), we plot the dependent variable `Hopx` versus the index based on the 10 SNPs selected according to our previous comments on Table 4.3. The linear correlation between this index and the one estimated on all the SNPs which is equal to 0.956, highlights the good behaviour of our CCS strategy to select the relevant SNPs. Thus, not surprisingly, we observe the same relation between `Hopx` gene expression and the estimated indices.

ρ		0.75	0.8	0.85	0.9
$p^0 = 10$	Number of selected SNPs	53	43	29	11
	Number of SNPs in common with those selected when $\rho = 0.9$	9	11	11	11
$p^0 = 20$	Number of selected SNPs	136	125	106	69
	Number of SNPs in common with those selected when $\rho = 0.9$	64	67	68	69
Number of SNPs in common with those selected when $p^0 = 10$ and when $p^0 = 20$		50	36	19	10

Table 4.3: Results on selected SNPs for various values of p_0 and ρ

4.4.5.3 Comparison methods

We compare our approach with three popular multivariate methods for analyzing high-dimensionnal datasets: A Lasso approach, the sparse Partial Least Squares (sPLS) and a Bayesian variable selection regression (ESS++). The Lasso method (Tibshirani (1996)) often performs poorly in prediction and interpretation especially when n is small and p is large. This technique tends to shrink the regression coefficients towards zero in order to select a sparse subset of covariates and provide a better prediction performance. sPLS (Lê Cao *et al.* (2009)) seeks for the best linear combination of SNPs to predict the outcome. To ensure sparsity, sPLS includes a penalty function on some loading coefficients which is equivalent to a restriction on the number of loading vectors and on the number of SNPs, in each vector, that have a non-null coefficient. Both Lasso and sPLS approaches require a preliminary calibration of the tuning parameters which directly affects the number of selected variables, the estimate of the model parameters and therefore the statistical performances of the models. Calibration procedures usually involve the minimization of the mean square error of prediction through V-fold cross validation. In this illustration, we used the leave-one-out crossvalidation method to choose the tuning parameter for both methods. We finally compare our results with ESS++ a Bayesian variable selection approach for linear regression that can analyze single and multiple responses (Bottolo

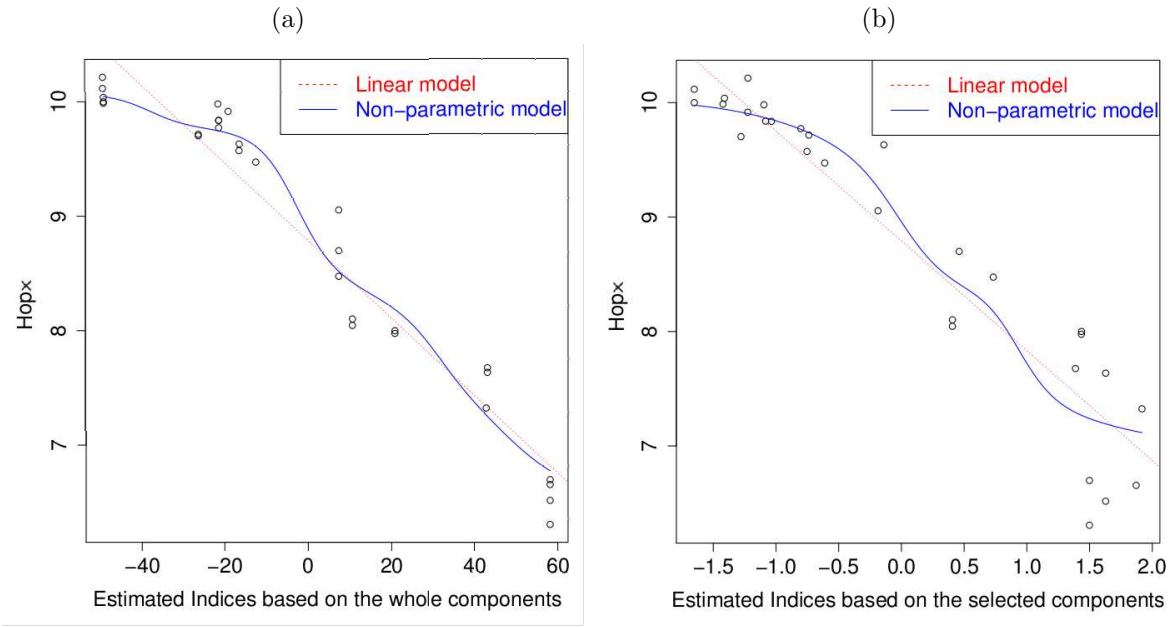


Figure 4.18: Plots of the dependent variable `Hopx` versus the index based on the whole SNPs (Figure 4.18(a)) and on the 10 selected SNPs (Figure 4.18(b)). The linear correlation between these two indices (evaluated on $n = 29$ rats) is 0.956. The dotted (red) line is the estimated linear model, the solid (blue) line is the kernel estimate of the link function with a bandwidth chosen by cross-validation.

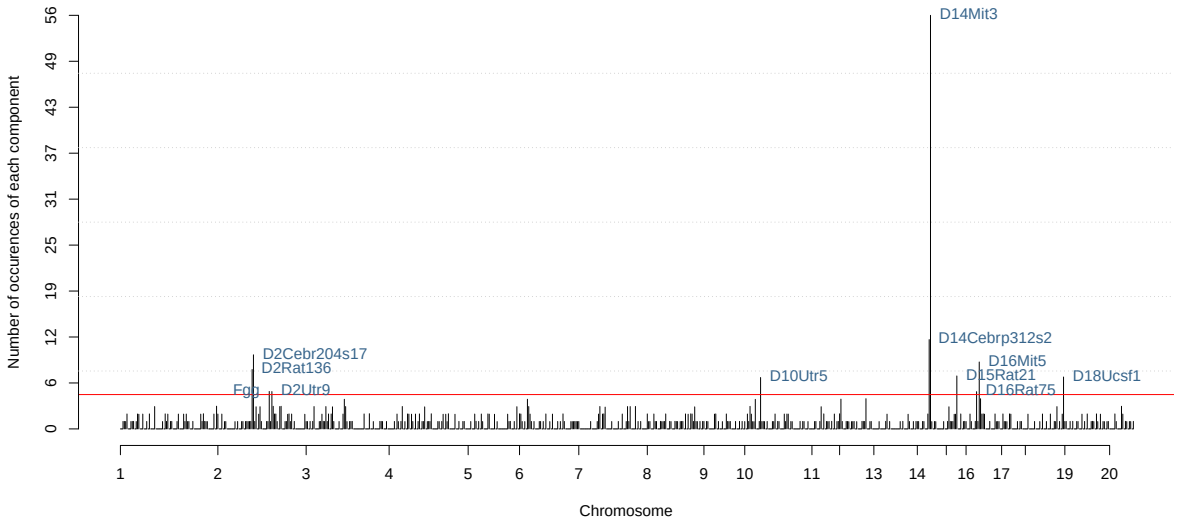


Figure 4.19: Plot of the occurrence of the SNPs by the CSS procedure ($p^0 = 10$, $\rho = 0.9$). The horizontal solid red line represent the threshold to select the most relevant SNPs.

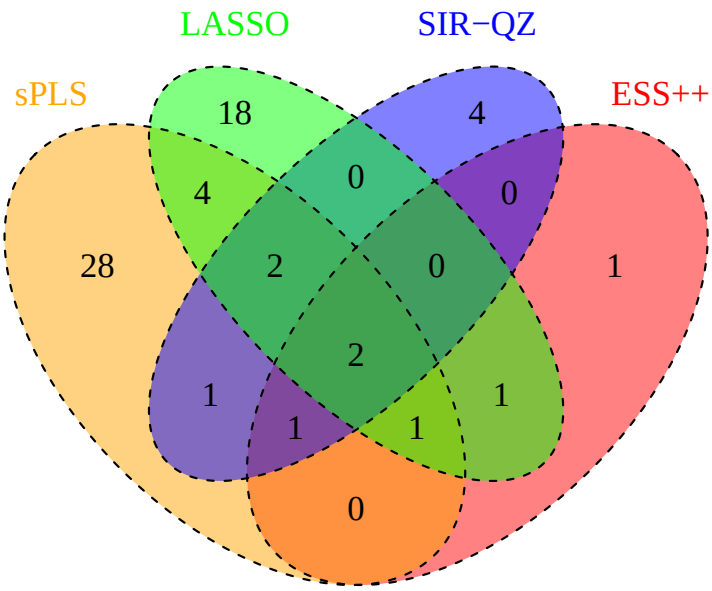


Figure 4.20: Venn diagram of the sets of SNPs selected by Lasso, sPLS, ESS++ and SIR-QZ (combined with CSS) approaches

and Richardson (2010); Bottolo *et al.* (2011)). ESS++ exploits recent developments in MCMC search algorithms to explore the 2^p-dimensional model space. The performances of this method have been, among others, illustrated on eQTL studies (Petretto *et al.* (2010)).

Figure 4.20 presents the Venn diagram of the sets of SNPs selected by the different approaches. Two SNPs (`D14Mit3` and `D2Cebr204s17`) are selected by the four methods. `D14Mit3` (chromosome 14) is clearly the first SNP in the list of the SNPs selected by the CSS procedure (see Figure 4.19) and `D2Cebr204s17` (chromosome 2) is at the third position. Moreover, our proposed approach reveals 4 SNPs (`D18Ucsf1`, `Fgg`, `D2Utr9`, `D16Rat75`) not selected by the other methods. Although three of them are very close to our proposed threshold (red line in Figure 4.19), while the SNP `D18Ucsf1` (chromosome 18) is clearly selected by our procedure.

The main advantage of our approach is the opportunity to reveal a non-linear link between the gene expression and a parsimonious index while the other compared approaches are based on a linear model. However, theses methods could treat multiple correlated phenotypes (multiple continuous responses). For example, ESS++ has been used to study the joint variability of gene expression in seven tissues (adrenal, gland, aorta, fat, heart, kidney, liver, skeletal) from inbred line rats (Bottolo *et al.* (2011)). An extension of SIR-QZ for multivariate response is under investigation.

4.4.6 Concluding remarks

Although regularizing the estimated covariance matrix or constraining the optimization problem are natural ways to extend SIR to underdetermined cases ($n < p$), it may not be

clear which one should be chosen and how to set the related parameter for each procedure. For the RSIR method, we illustrated that in such a context, the corresponding parameter should rather be determined with respect to the stability of the linear algebra algorithm (as in SIR-QZ) than with a statistical criterion. Moreover, the SIR-QZ approach introduced in this paper produces better results in simulation than the SR-SIR method that constraint the underlying optimization problem. In addition, the poor performances of SIR-MP suggest that adapting properties of the pseudo-inverse from the functional SIR to our high dimensional context is not well-adapted. We assumed that the dimension K of the EDR subspace was known in our simulation study. While, in the application, an empirical argument was given to determine this dimension, its estimation remains to be done with care.

We also proposed the CSS method that searches which submodels are the most informative to select relevant components of x. This procedure relies on a quality measure of the estimated indices, and it outperforms, in simulation, RSIR and SR-SIR selection procedures which are both based on estimates of β. We thus explain these results by pointing out that, when $n < p$, only the indices can be properly estimated, in general cases. Note that the space of the submodels may not be browsed optimally by our proposed CSS algorithm. Improvements should be made using optimization techniques such as genetic algorithms.

All the developed methods have been implemented in R language and are available upon request from the corresponding author. An R package is currently under development.

The illustration of genetic data highlights the opportunity of SIR-QZ combined by the CSS procedure to reveal a few number of SNPs which can explain the variability of the expression of the `Hopx` gene. Then, a non linear link between the gene expression and the parsimonious index could be estimated. The choice of the number of SNPs to keep is still a topic of concern since alternatives to our threshold could be considered.

In genetic datasets, the response variable is often multivariate. For instance, it could represent several phenotypes as in eQTL studies. Some approaches already handle such datasets. Since univariate results for SIR-QZ and the CSS selection procedure are consistent with the other methods presented in our application, it thus appears interesting to extend them to the multivariate case.

Acknowledgements

The authors wish to gratefully thank Marie Chavent for the first discussions that led to this work. They also acknowledge the editor and both anonymous referees for their helpful comments which lead to significant enhancements of this article.

Chapter 5

Conclusion

In this document, we exhibit various stochastic models adjusted for some biological issues. A particular attention was given to include knowledge and needs from the tackled problems into the statistical analyses. When several methods have been developed for a similar goal, another concern was to compare them using theoretical arguments or simulation studies. The procedures we consider were made accessible by implementing them in a user-friendly way.

The model of Section 3.1 was designed to handle reasonable assumptions about the valvometric signal. Various models exist to detect states in a stochastic process defined as a piecewise function on which a noise is added. For valvometric signals, we would rather assume that the underlying function is continuous because the distance between both parts of a shell of an animal does not vary with an infinite speed. In addition, two consecutive points in the signal do not seem to be independent even with respect to the current state of the hidden process. For example, if a point belongs to a microclosing, the following one may not follows the same distribution as if it does not. This can be observed in Figure 3.16. The procedure we present to estimate the hidden process does not thus need independent observations but only equally spaced points.

Statistical needs related to genetic datasets, such as multiple testing, are well-known and widely studied. That is why we worked in an existing framework called FAMT which can only handle linear links. Because we anticipate that non-linear ones may also be of interest, we gave in Section 4.3 a method to apply SIR in such a multivariate context. Due to the cost of DNA sequencing, the sample size of a transcriptomic dataset may be quite small. That is why we also provided some modifications of SIR adapted to cases when $n < p$, in Section 4.4.

Finding a threshold to separate the open state from the closed one in the valvometric data requires to estimate a probability density function f with a single anti-mode located in θ. Various alternatives are available for this purpose. We showed in the simulation study of Section 3.2.3 that computing Polonik's estimator may not produce good enough estimates of θ. When f have bounded support, the kernel-based estimators with Gaussian kernel seems to succeed in this task whether h_{SJ} or h_{crit} is chosen. Appendix A.1 gives a possible reason why using h_{crit} with the uniform kernel may provide inconsistent estimates of θ.

Multivariate sliced inverse regressions can be performed with various procedures like pooled marginal slicing or k-means inverse regression. The results of the simulation study

of Section 4.3.5 let us think that these methods manage to properly estimate EDR spaces, as so do MSIR and wMSIR. The latter ones can however straightforwardly be paired with a procedure to detect clusters of components in the multivariate response variable.

When the sample size is smaller than the number of explanatory variables and in the particular context of model (4.29), SIR-QZ outperforms RSIR, SR-SIR and SIR-MP when trying to estimate a basis of the indices $X'\beta$. The CSS procedure, together with SIR-QZ, also achieves a better selection of the explanatory variables than the full RSIR and SR-SIR methods, for this model.

A lot of theoretical results in statistics are given when $n \to \infty$. Statistics is also a privileged science when considering high-dimensional variables. It thus makes sense to have to handle large data sets to apply statistical methods. Computer programs which help to do so are then welcome. SQL databases are appropriate structures for storing and easily accessing such data sets. Free programs exist to manage these databases. We then imported most real data we can into databases in the analyses presented in this thesis. We also explained how to retrieve data from the SQL databases with the R software in Section 3.4.

Executing R functions built by statisticians and sending SQL requests to another server may take a lot of time for occasional R users. Deploying a graphical user interface (GUI) with an R package may circumvent these difficulties. We created a GUI to estimate the probability density f from valvometric data (see Section 3.4.3) and another one in order to apply functions from the `FAMT` package on the IMMORTEEL transcriptomic data (Section 4.2.3).

Concerning the estimator $\widetilde{f}_n$ described in equation (3.6) with $h_n = h_{crit}$, further work would be interesting. The asymptotic behaviour of the bandwidth h_{crit} should be studied in order to know whether or not it goes toward 0 and if $nh_{crit}^2 \to \infty$, when $n \to \infty$. This is important in order to have Assumptions 5 and then convergence of the position of the unique local minimum of $\widetilde{f}_n$ toward θ.

For a kernel density estimator $\widehat{f}_n$ of f, with a random bandwidth, Devroye and Wagner (1980) and Devroye (1987) gave the convergences of $\|\widehat{f}_n - f\|_\infty$ and of $\|\widehat{f}_n - f\|_1$. For a kernel density estimator $\widehat{f}_n$ with a real bandwidth, there exist results about $\|\widehat{f}_n - f\|_2$. To our knowledge this is not the case for random bandwidths. For these bandwiths, proving that $\|\widehat{f}_n - f\|_2$ converges toward 0 or explaining why it is not possible may be an exciting challenge. It could for example bring new arguments in the comparison between Polonik's estimator and a kernel estimator relying on the bandwidth h_{crit}.

In the example of Section 3.2.3, we actually found that for a kernel density estimator with the Gaussian kernel, choosing the bandwidth $h_{crit,k}$ with a large integer k may produce smaller IAE values than taking the bandwidth $h_{crit} = h_{crit,N(f)}$. It also creates spurious modes in $\hat{f}_{K,h_{crit,k}}$. In his test of the existence of modes, Minnotte (1997) proposed to replace some modes with constant parts. Using this smoothing technique, we could have for all $k \geq N(f)$ an estimator based on $\hat{f}_{K,h_{crit,k}}$ with only $N(f)$ modes. Theoretical properties of this estimator remain to be found.

In Section 3.1 we classified some oysters into two groups. This procedure only relies on their transitions between their open and their closed state. Their microclosings may however furnish additional information about their health status. Considering their amplitude and the time between two consecutive microclosings, we could for example study

the bivariate distribution of these variables like in Schmitt *et al.* (2011). This requires a microclosing detection technique, which could be built from Section 3.3.3.

The R package we created to access valvometric data sometimes need a significant amount of computational time. To improve it, we could create another index in the SQL database which would consider the location of the studied animal, the date and the bivalve's identification number, in this order. We also could implement a cache for the valvometric signal. For example, when a user studies data from a given day, the program could silently download signals related to the next or to the previous ones in order to display them faster if the user ask for them.

The FAMT model (4.5) is a particular case of model (4.9), with $K = 1$. We can then apply the MSIR method on the IMMORTEEL dataset in order to find clusters in the set of selected genes. If for $j_1 \in \{1, \dots, q\}$ and $i_1 \in \{2, \dots, p\}$, we have equation (4.6) with $j = j_1$ and $i = i_1$ and if we have it too for $j = j_2$ and $i = i_2$ where $j_2 \in \{1, \dots, q\}$ and $i_2 \in \{2, \dots, p\} \backslash \{i_1\}$, then $y^{(j_1)}$ and $y^{(j_2)}$ should not appear in the same cluster.

In model (4.5), let $\mathbf{Y} := \left[\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(q)}\right]$, and $\boldsymbol{\beta} := \left[\beta^{(1)}, \dots, \beta^{(q)}\right]$. Let also $\boldsymbol{\varepsilon} := \left[\varepsilon^{(1)}, \dots, \varepsilon^{(q)}\right]$ and $\mathbf{B} := \left[b^{(1)}, \dots, b^{(q)}\right]$. If the rank of $\boldsymbol{\beta}$ is equal to p, because $p < q$ for the IMMORTEEL dataset, we have

$$\mathbf{X}' = \mathbf{Y}\boldsymbol{\beta}^+ - \mathbf{Z}'\mathbf{B}\boldsymbol{\beta}^+ - \boldsymbol{\varepsilon}\boldsymbol{\beta}^+,$$

where $\boldsymbol{\beta}^+$ is the Moore-Penrose pseudoinverse of $\boldsymbol{\beta}$. Let $\boldsymbol{\beta}^+ =: \left[\beta_1^+, \dots, \beta_p^+\right]$. We can write for all $i \in \{1, \dots, p\}$

$$X_i = \mathbf{Y}\beta_i^+ - \mathbf{Z}'\mathbf{B}\beta_i^+ - \boldsymbol{\varepsilon}\beta_i^+.$$

We have $n = 44$ and $q = 17,866$ which are appropriate values to execute SIR-QZ and the CSS procedure on the response variable X_i and on the matrix of explanatory variables $\mathbf{Y}$.

While SIR was introduced in 1991, new extensions can still be proposed nowadays. The underlying model (4.18) could for example be modified so that for all $k \in \{1, \dots, K\}$, $\beta_k \in \mathbb{N}^p$. This might make the null elements of β_k easier to detect, when trying to select relevant explanatory variables. This approach was studied in Sabatier and Reynès (2008) and Vines (2000) for the principal component analysis.

Finally, it seems to us that the statistical models studied in this document may also be useful for other biological problems and in other fields than biology. Continuous variations in a signal in which a process with a finite number of states is hidden may not only happen for valvometric measures. Graphics in Watwood *et al.* (2006) suggest that cetaceans' dive could also be studied with the framework of Section 3.1. SIR and FAMT are suitable methods when aiming at finding a few useful explanatory variables among a lot of considered ones. In marketing, we may find the reasons why people buy a given product with these statistical tools.

Appendix A

Details and proofs

A.1 The critical bandwidth for the uniform kernel

A.1.1 Introduction

Let consider a sample $\mathbf{X} = (X_1, \dots X_n)$ made of independent and identically distributed random variables generated from the density f. To estimate f from this sample of size n, Parzen (1962) and Rosenblatt (1956) introduced the kernel density estimator $\hat{f}_{K,h}$, defined for every real t by

$$\hat{f}_{K,h}(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right),$$

where K is called the kernel and is most of time a density function while h is a positive real parameter that controls the smoothness of $\hat{f}_{K,h}$. We also refer the interested reader to Scott (1992) and Silverman (1986).

Let $N(\hat{f}_{K,h})$ be the number of modes of $\hat{f}_{K,h}$. To decide how smooth $\hat{f}_{K,h}$ should be, an approach is to set $N(\hat{f}_{K,h})$. To do so, one can use the critical bandwidth h_{crit} introduced by Silverman (1981) for h. It can be defined by

$$h_{crit,k} = \min_{N(\hat{f}_{K,h}) \leq k} h, \tag{A.1}$$

for any $k \in \mathbb{N}^*$.

When K is the Gaussian kernel, that is $\forall t \in \mathbb{R}, K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$, the bandwidth $h_{crit,k}$ has interesting properties. It can easily be computed and, provided that $k \geq N(f)$, allows $\hat{f}_{K,h}$ to exhibit, in probability, a pointwise convergence, an L_1-convergence and a uniform convergence toward f. Details can be found in Futschik and Isogai (2006), Mammen *et al.* (1991), Devroye and Wagner (1980), Devroye (1987) and Section 3.2.

These properties are proven using the key point that $h_{crit,k}$ converges in probability toward 0 when K is the Gaussian kernel. To extend them to other kernels, the asymptotic behavior of $h_{crit,k}$ should thus be studied. In this Note, we focus on the case when K is the uniform kernel defined as $\forall t \in \mathbb{R}, K(t) = \mathbf{1}_{[-\frac{1}{2}, \frac{1}{2}]}(t)$, where $\mathbf{1}$ is the indicator function. We will prove in the following section that for this kernel, $h_{crit,k}$ does not converge toward 0, after giving some properties about $\hat{f}_{K,h}$ and $N(\hat{f}_{K,h})$.

A.1.2 Properties for the uniform kernel

Let us first introduce some additional notations. Let

$$\mathbf{A}_h = \cup_{i=1}^n \{X_i - \frac{h}{2}\} = \left\{a_{h,(i)}\right\}_{i\in\{1,\dots,\text{card}(\mathbf{A}_h)\}}$$

and

$$\mathbf{B}_h = \cup_{i=1}^n \{X_i + \frac{h}{2}\} = \left\{b_{h,(i)}\right\}_{i\in\{1,\dots,\text{card}(\mathbf{B}_h)\}}.$$

In order to deduce the value of $N(\hat{f}_{K,h})$, we only need to investigate how the points in $\mathbf{A}_h \cup \mathbf{B}_h$ are ordered because of Proposition 13 below. We write $w = \text{card}(\mathbf{A}_h \cup \mathbf{B}_h) \leq 2n$. We set $\left\{c_{h,(i)}\right\}_{i\in\{1,\dots,w\}}$ as the ordered points in $\mathbf{A}_h \cup \mathbf{B}_h$. Let us also write $b_{h,(0)} = c_{h,(0)} = -\infty$ and $a_{h,(\text{card}(\mathbf{A}_h)+1)} = c_{h,(w+1)} = +\infty$.

Proposition 13. *Let $(X_1, \dots X_n)$ be a vector of independent random variables generated from f. Let $\hat{f}_{K,h}$ be the kernel estimator of f for the uniform kernel K. Then, $\forall h > 0$, $\forall i \in \{0, 1, \dots, w\}$, the function $\hat{f}_{K,h}$ is constant on $]c_{h,(i)}, c_{h,(i+1)}[$.*

The proof is given in Appendix A.3.7.

Remark 16. The reasoning in the proof of Proposition 13 can also be used to obtain the following results:

- $\forall i \in \{1, \dots, w\}$, $c_{h,(i)} \in \mathbf{A}_h \Leftrightarrow \exists \zeta \in \mathbb{N}^*,\ \forall u \in]c_{h,(i-1)}, c_{h,(i)}[,\ f(u) = f(c_{h,(i)}) - \frac{\zeta}{nh}$,
- $\forall i \in \{1, \dots, w\}$, $c_{h,(i)} \in \mathbf{B}_h \Leftrightarrow \exists \zeta \in \mathbb{N}^*,\ \forall u \in]c_{h,(i)}, c_{h,(i+1)}[,\ f(u) = f(c_{h,(i)}) - \frac{\zeta}{nh}$,
- $\forall i \in \{1, \dots, w\}$, $c_{h,(i)} \notin \mathbf{A}_h \Leftrightarrow \forall u \in]c_{h,(i-1)}, c_{h,(i)}[,\ f(u) = f(c_{h,(i)})$,
- $\forall i \in \{1, \dots, w\}$, $c_{h,(i)} \notin \mathbf{B}_h \Leftrightarrow \forall u \in]c_{h,(i)}, c_{h,(i+1)}[,\ f(u) = f(c_{h,(i)})$.

Proposition 13 and Remark 16 are illustrated in Figure A.1 where the sample (3, 6.5, 6, 1.5, 5) of size $n = 5$ is used to compute the function $\hat{f}_{K,2}$. We have here $w = 9$. Note that $N(\hat{f}_{K,2}) = 3$ and that a mode of $\hat{f}_{K,2}$ is actually a single point. This mode vanishes if we use a bandwidth slightly less than 2 so that there is a jump in the estimator $h \mapsto N(\hat{f}_{K,h})$ located at $h = 2$. Figure 16 gives a lead to find $N(\hat{f}_{K,h})$ from $\mathbf{A}_h$ and $\mathbf{B}_h$, as explained in the following property.

Proposition 14. *Let $(X_1, \dots X_n)$ be a vector of independent random variables generated from f. Let $\hat{f}_{K,h}$ be the kernel estimator of f for the uniform kernel K. The number of modes $N(\hat{f}_{K,h})$ of $\hat{f}_{K,h}$ is such that*

$$N(\hat{f}_{K,h}) = \text{card}\left(\left\{(i,j)\ :\ a_{h,(i)} \in]b_{h,(j-1)}, b_{h,(j)}] \text{ and } b_{h,(j)} \in [a_{h,(i)}, a_{h,(i+1)}[\right\}\right),$$

where $i \in \{1, \dots, \text{card}(\mathbf{A}_h)\}$ and $j \in \{1, \dots, \text{card}(\mathbf{B}_h)\}$.

The proof is given in Appendix A.3.8.

This characterization of $N(\hat{f}_{K,h})$ based on the sets $\mathbf{A}_h$ and $\mathbf{B}_h$, together with an argument about totally positive matrices, allows us to show the following theorem:

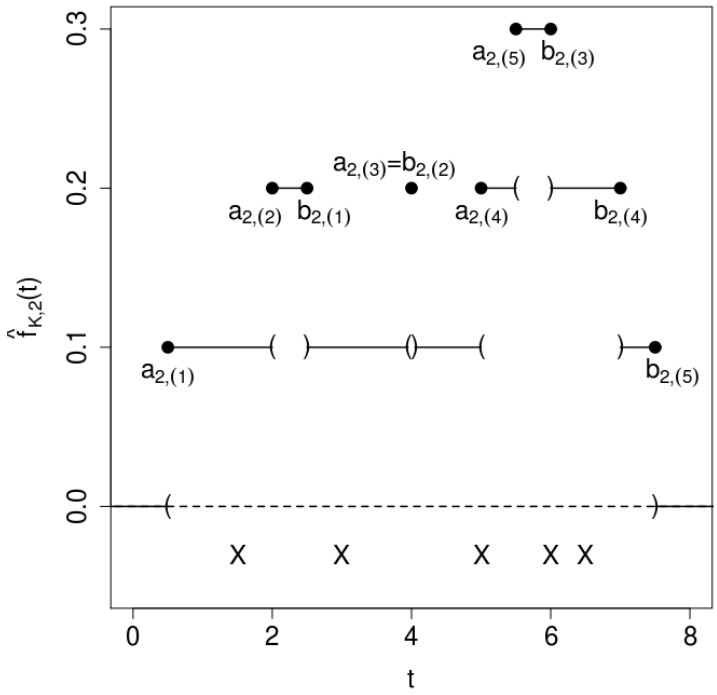


Figure A.1: Estimation of a density with the estimator $\hat{f}_{K,2}$ and the uniform kernel K (solid line). Solid circles are special points of this estimated density while brackets indicate a difference between the limit and the value of $\hat{f}_{K,2}$. Crosses represent the underlying sample.

Theorem 15. *For any probability density function f of $\mathbf{X}$, let $\hat{f}_{K,h_{crit,k}}$ be the estimator of f when K is the uniform kernel with $h_{crit,k}$ given in (A.1). Then we have $h_{crit,k}$ increasing with n, for all $k \in \mathbb{N}$.*

The proof is given in Appendix A.3.9.

To illustrate this theorem, we consider again the sample $(3, 6.5, 6, 1.5, 5)$ and every subsamples made of the first n elements of $(3, 6.5, 6, 1.5, 5)$ for $n \in \{1, \ldots, 5\}$. For each subsample, we compute the function $h \mapsto N(\hat{f}_{K,h})$, where K is the uniform kernel. In Figure A.2, we display ranges of values of h and n for which $N(\hat{f}_{K,h})$ is equal to a given number. The increase of $N(\hat{f}_{K,h})$ with n can be observed for small values of h. In the general case, this feature implies that for any $k \in \mathbb{N}$, $h_{crit,k}$ also increases with n.

Theorem 15 means that we can not have that $h_{crit,k}$ goes toward 0 in probability for the uniform kernel. Thus, the proof of the pointwise convergence of $\hat{f}_{K,h_{crit,k}}$ toward f, as given in Futschik and Isogai (2006), does not hold for this kernel. It is also impossible to use the work of Devroye and Wagner (1980) and Devroye (1987) to show the corresponding L_1-convergence and uniform convergence. In addition, the simulation study from Section 3.2 shows that $\hat{f}_{K,h_{crit,k}}$ is not an accurate estimator of f, when K is the uniform kernel. For these reasons, we recommend not to use $\hat{f}_{K,h_{crit,k}}$ with this kernel in practice.

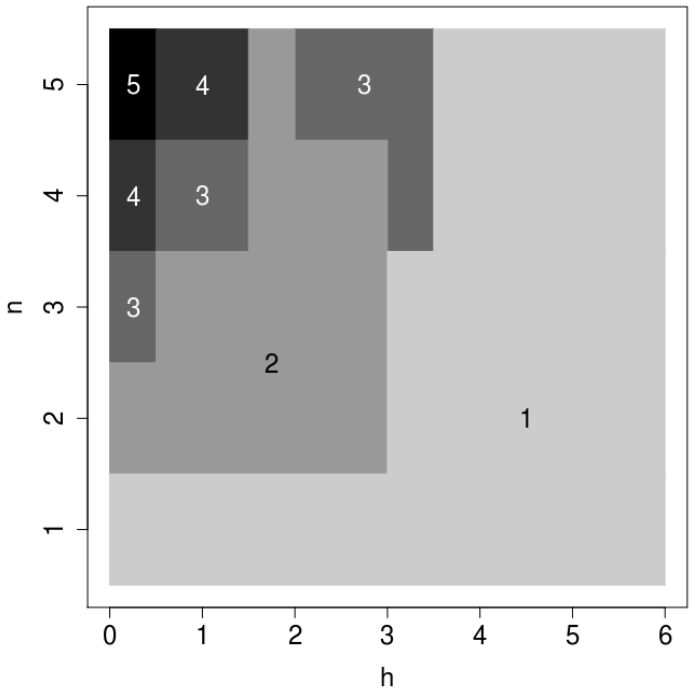


Figure A.2: Evolution of $N(\hat{f}_{K,h})$ with respect to the bandwidth h and to the sample size n, for the uniform kernel K. The considered samples are the first n values of $(3, 6.5, 6, 1.5, 5)$.

A.2 Details about SIR-QZ

A.2.1 Generalized real Schur decomposition

We work here with real matrices $\widehat{M}$ and $\widehat{\Sigma}$. Similarly to the generalized Schur decomposition introduced in Section 4.4.2.3, that produces complex matrices Q and Z, the generalized real Schur decomposition (see Theorem 7.7-2 of Golub and Van Loan (1983)) ensures that we can find Q and Z such that $Q\widehat{M}Z$ is an upper quasi-triangular real matrix and $Q\widehat{\Sigma}Z$ is an upper triangular real one. An upper quasi-triangular matrix can be defined as the sum of an upper triangular matrix and of a block diagonal matrix where the sizes of the block are 1×1 or 2×2. For a 1×1 diagonal block of the matrix $Q\widehat{M}Z$, its unique element is called $\tilde{t}_j$ if it is located at the jth row and at the jth column of $Q\widehat{M}Z$. We write $\tilde{t}_{j_1,j_2}$ the element of a 2×2 diagonal block, in the j_1th row and in the j_2th column of $Q\widehat{M}Z$. An example of such an upper quasi-triangular matrix is given below:

$$Q\widehat{M}Z = \begin{pmatrix} \tilde{t}_1 & * & * & * & * \\ 0 & \tilde{t}_{2,2} & \tilde{t}_{2,3} & * & * \\ 0 & \tilde{t}_{3,2} & \tilde{t}_{3,3} & * & * \\ 0 & 0 & 0 & \tilde{t}_4 & * \\ 0 & 0 & 0 & 0 & \tilde{t}_5 \end{pmatrix},$$

where $*$ denotes some real values. Let J be made of the elements $j \in \{1,\dots,p\}$ such that $\tilde{t}_j$ exists and J^c be the set made of $j \in \{1,\dots,p-1\}$ such that $\tilde{t}_{j,j}$ and $\tilde{t}_{j+1,j+1}$ exist. For each $j \in J$, let $\tilde{u}_j$ be the element of $Q\widehat{\Sigma}Z$ at the same location than $\tilde{t}_j$ in $Q\widehat{M}Z$ and define similarly $\tilde{u}_{j_1,j_2}$ for each $\tilde{t}_{j_1,j_2}$. Thus, we have

$$\det(\widehat{M}-\lambda\widehat{\Sigma}) = \det(Q'Z')\prod_{j\in J}(\tilde{t}_j-\lambda\tilde{u}_j)\prod_{j\in J^c}\det\begin{pmatrix} \tilde{t}_{j,j}-\lambda\tilde{u}_{j,j} & \tilde{t}_{j,j+1}-\lambda\tilde{u}_{j,j+1} \\ \tilde{t}_{j+1,j} & \tilde{t}_{j+1,j+1}-\lambda\tilde{u}_{j+1,j+1}\end{pmatrix}.$$

Hence, for $j \in J$, if $\tilde{u}_j \neq 0$, then $\lambda_j = \tilde{t}_j/\tilde{u}_j$ is a real generalized eigenvalue. In addition, for $j \in J^c$, Moler and Stewart (1973) succeeded in finding $(\tilde{t}_j,\tilde{t}_{j+1}) \in \mathbb{C}^2$, and $(\tilde{u}_j,\tilde{u}_{j+1}) \in \mathbb{R}\backslash\{0\}$ such that $\lambda_j = \tilde{t}_j/\tilde{u}_j$ and $\lambda_{j+1} = \tilde{t}_{j+1}/\tilde{u}_{j+1}$ are generalized eigenvalues of $\widehat{M}$ and $\widehat{\Sigma}$. This leads to vectors $\tilde{t} = (\tilde{t}_1,\dots,\tilde{t}_p)'$ and $\tilde{u} = (\tilde{u}_1,\dots,\tilde{u}_p)'$ that are sent back by the QZ algorithm in order to provide generalized eigenvalues.

A.2.2 Algorithm

As explained in Section 4.4.2.3, the QZ algorithm has to be controlled when dealing with singular pencils. The following pseudocode in Scilab language allows the user to do so in the context of sliced inverse regression, for underdetermined cases. Because it is based on generalized real Schur decompositions, the notations involved are related to Appendix A.2.1 rather than Section 4.4.2.3.

```
// Initialize s_min, c and ε.
s = s_min;
keepGoing = %T;
while keepGoing
        //Use the QZ algorithm on M̂ and Σ̃(s) to find
```

```
        //vectors ũ and t̃.
        if (sum(abs(ũ) < ε & abs(t̃) < ε) == 0) &
           (length(ũ) - sum(abs(ũ) < ε) >= K) then
                keepGoing = %F;
        else
                s = s * c;
        end
end
// The estimated EDR directions are the eigenvectors sent
// by the last run of the QZ algorithm that corresponds to
// the K greatest values of t̃/ũ.
```

Algorithm A.1: A procedure to estimate EDR directions with the QZ algorithm

Typical values for $s_{\min}$, c and ε chosen in the simulation study of Section 4.4.4 are respectively 10^{-16}, 10 and 10^{-10}. The QZ algorithm is implemented in Matlab through the **eig** function. With Scilab, one should use **spec**, which is based on the LAPACK library. The R software is able to call functions from this library. Thus the QZ algorithm can be easily tested with this software. Notice that Algorithm A.1 is designed to handle real values of $\tilde{u}_j$ and $\tilde{t}_j$ but, as mentioned in Appendix A.2.1, they can be complex. In that case, knowing if the blocks made of $\tilde{t}_{j,j}$, $\tilde{t}_{j+1,j}$, $\tilde{t}_{j,j+1}$ and $\tilde{t}_{j+1,j+1}$ and of $\tilde{u}_{j,j}$, $\tilde{u}_{j+1,j}$, $\tilde{u}_{j,j+1}$ and $\tilde{u}_{j+1,j+1}$ produce unstable eigenvalues is more difficult. As explained in Section 5 of Moler and Stewart (1973), the QZ algorithm aims at finding stable λ_j and λ_{j+1} corresponding to these 2×2 blocks. Because we do not control this procedure, we simply report if the QZ algorithm send back complex values in $\tilde{t}$. We never encounter this case in the simulation study of Section 4.4.4.

A.2.3 The sliced indices issue

Hereafter, we describe why the Algorithm A.1 produces clustered indices as in Figure A.3. Recall that $\widehat{\Sigma} = \frac{1}{n}\widetilde{X}\widetilde{X}'$. Define $\bar{I}_n = I_n - \frac{1}{n}\mathbf{1}_{n,n}$, where every element of the $n \times n$ matrix $\mathbf{1}_{n,n}$ is equal to 1. Notice that $\widetilde{X} = X\bar{I}_n$ and then $\widehat{\Sigma} = \frac{1}{n}X\bar{I}_nX'$. Let $\widehat{S}$ be a $n \times H$ matrix made of elements $\widehat{S}_{i,h}$ defined as

$$\widehat{S}_{i,h} = \frac{1}{n}\left(\frac{\mathbb{I}[y_i \in s_h]}{\hat{p}_h}\right).$$

We can also write $\widehat{M} = X\bar{I}_n\widehat{SW}\widehat{S}'\bar{I}_nX'$, where $\widehat{W}$ is defined in Section 4.4.2.2. Because of the structure of $\widehat{S}$, for any $H \times \alpha$ matrix A, $\widehat{S}A$ has at most H distinct rows, so has $\bar{I}_n\widehat{S}A$. Let w be the first generalized eigenvector of $\bar{I}_n\widehat{SW}\widehat{S}'\bar{I}_n$ and $\bar{I}_n$ associated with the eigenvalue λ. This means that $\bar{I}_n\widehat{SW}\widehat{S}'\bar{I}_nw = \lambda\bar{I}_nw$ which implies that $\bar{w} = \bar{I}_nw$ has at most H distinct values and then that w has also at most H distinct values.

Assume that X has full column rank, which is likely to happen when $p > n$. Then, $X^{+\prime}w$ is a generalized eigenvector of $\widehat{M}$ and $\widehat{\Sigma}$ and $X'X^{+\prime}w = w$ has at most H distinct values. The eigenvalue that is related to $X^{+\prime}w$ is equal to $n\lambda$.

If it exists $\hat{\beta}_1 \neq X^{+\prime}w$, a generalized eigenvector of $\widehat{M}$ and $\widehat{\Sigma}$ such that the generalized eigenvalue which is related to $\hat{\beta}_1$ is greater than the one corresponding to $X^{+\prime}w$, then we

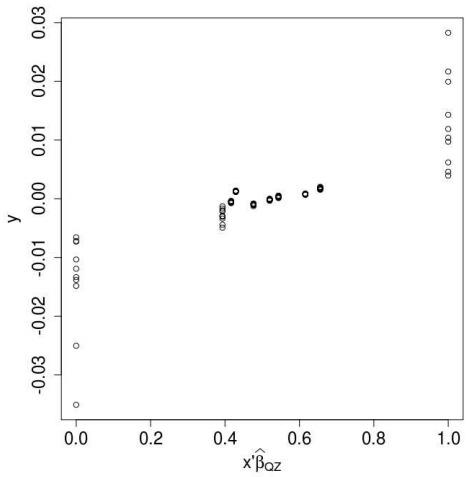


Figure A.3: Plot of Y versus $X'\hat{\beta}_{\mathrm{QZ}}$ with $H = 10$. The horizontal scale was standardized.

should have

$$\frac{\hat{\beta}_1' X \bar{I}_n \widehat{SWS}' \bar{I}_n X' \hat{\beta}_1}{\hat{\beta}_1' X \bar{I}_n X' \hat{\beta}_1} > \frac{w' \bar{I}_n \widehat{SWS}' \bar{I}_n w}{w' \bar{I}_n w}.$$

But, because w is the first generalized eigenvector of $\bar{I}_n \widehat{SWS}' \bar{I}_n$ and $\bar{I}_n$, it maximizes $\frac{u' \bar{I}_n \widehat{SWS}' \bar{I}_n u}{u' \bar{I}_n u}$ over the vectors u of length n, which contradicts the latter equation. Hence, such a $\hat{\beta}_1$ does not exist, and the first generalized eigenvector of $\widehat{M}$ and $\widehat{\Sigma}$ is $X^{+\prime} w$.

In this paragraph, we show that it exists $H-1$ orthogonal vectors $w_1, \ldots, w_{H-1}$ such that, for $k = 1, \ldots, H-1$,

$$\frac{w_k' \bar{I}_n \widehat{SWS}' \bar{I}_n w_k}{w_k' \bar{I}_n w_k} = \frac{w' \bar{I}_n \widehat{SWS}' \bar{I}_n w}{w' \bar{I}_n w},$$

which means that the first K generalized eigenvectors of $\widehat{M}$ and $\widehat{\Sigma}$ are the vectors $X^{+\prime} w_k$ for $k = 1, \ldots, K \leq H-1$. We assume $n > H$. Let sort Y increasingly and reorder the columns of X such that each column corresponds to the appropriate element of Y. This transformation implies that $\widehat{SWS}'$ is block diagonal with H blocks. For $h = 1, \ldots, H$, the size of the block h is equal to $n\hat{p}_h$ and each element it contains is equal to $\frac{1}{n^2 \hat{p}_h}$. Hence, the rank of each block is equal to 1 and each block provides a positive eigenvalue for $\widehat{SWS}'$, which is equal to $\frac{n\hat{p}_h}{n^2 \hat{p}_h} = \frac{1}{n}$. The corresponding eigenvector is made of elements $\frac{\mathbb{I}[y_i \in s_h]}{\sqrt{n\hat{p}_h}}$ for $i = 1, \ldots, n$. We have now H orthonormal eigenvectors of $\widehat{SWS}'$ for the eigenvalue $\frac{1}{n}$, and then we can find $H-1$ orthonormal centered eigenvectors $\bar{w}_1, \ldots, \bar{w}_{H-1}$ of $\widehat{SWS}'$ for this eigenvalue. In addition, for $k = 1, \ldots, H-1$, $\bar{w}_k$ is also a generalized eigenvector of $\bar{I}_n \widehat{SWS}' \bar{I}_n$ and $\bar{I}_n$ because $\bar{I}_n \bar{w}_k = \bar{w}_k$, and $\bar{w}_k$ maximizes $\frac{u' \bar{I}_n \widehat{SWS}' \bar{I}_n u}{u' I_n u}$, over the vectors u of length n. Finally, we have that the first K generalized eigenvectors of $\widehat{M}$ and $\widehat{\Sigma}$ are $X^{+\prime}(\bar{w}_1, \ldots, \bar{w}_K)$, which means that the indices $X' X^{+\prime}(\bar{w}_1, \ldots, \bar{w}_K)$ only have H distinct rows.

This feature is illustrated on the simulated sample of size $n = 100$, with $p = 200$ from Section 4.4.4.1. In Figure A.3, we plot Y versus the estimated indices obtained with $\hat{\beta}_{\mathrm{QZ}}$ for $H = 10$.

A.3 Proofs

A.3.1 Proof of Lemma 2

The Markov chain (Z_k, S_{k+1}) is ergodic and admits a unique invariant measure given by

$$\mathcal{U}_E(dx)\lambda(x,t)\exp\left(-\int_0^t \lambda(x,s)ds\right)dt,$$

where $\mathcal{U}_E$ denotes the uniform distribution on the set E. One may refer the reader to Azaïs *et al.* for more details on the properties of (Z_k, S_{k+1}) in the general case. Let $0 \leq s \leq t$. As a consequence, we have

$$\mathbf{P}\left(\lim_{m\to\infty}\frac{L_m(x,s)}{m} = \frac{\exp\left(-\int_0^s \lambda(x,u)du\right)}{4}\right) = 1, \tag{A.2}$$

according to the almost sure ergodic theorem. In particular, the limit for $s = t$ is strictly positive. Besides, since $L_m(x,\cdot)$ is a decreasing function, we have $R_m(x,s) \leq R_m(x,t)$ at least for m large enough. In this context, we have

$$\int_0^t R_m(x,s)\lambda(x,s)ds \;\leq\; R_m(x,t)\int_0^t \lambda(x,s)ds.$$

Thus,

$$\mathbf{P}\left(\lim_{m\to\infty}\int_0^t R_m(x,s)\lambda(x,s)ds = 0\right) = 1,$$

using (A.2) and the definition of R_m. This states (3.2). For the second limit, we have

$$\mathbf{P}\left(\lim_{m\to\infty}\mathbf{1}_{\{L_m(x,s)=0\}} = 0\right) = 1,$$

in light of foregoing. By the bounded convergence theorem, this shows (3.3).

A.3.2 Proof of Theorem 9

Recall that, since the bases $\mathbf{D}$, $\mathbf{B}^{(1)}$, …, $\mathbf{B}^{(q)}$ are assumed to be $\mathbf{\Sigma}$-orthonormal, we have $P_{\mathbf{D},\mathbf{\Sigma}} = \mathbf{D}\mathbf{D}'\mathbf{\Sigma}$ and $P_{\mathbf{B}^{(j)},\mathbf{\Sigma}} = \mathbf{B}^{(j)}\mathbf{B}^{(j)\prime}\mathbf{\Sigma}$. It follows that

$$\begin{aligned}
Kq \times Q(\mathbf{D},\mathbf{B}^{(1)},\ldots,\mathbf{B}^{(q)}) &= \sum_{j=1}^{q}\mathrm{Trace}(\mathbf{D}\mathbf{D}'\mathbf{\Sigma}\mathbf{B}^{(j)}\mathbf{B}^{(j)\prime}\mathbf{\Sigma}) \\
&= \sum_{j=1}^{q}\mathrm{Trace}(\mathbf{D}'\mathbf{\Sigma}\mathbf{B}^{(j)}\mathbf{B}^{(j)\prime}\mathbf{\Sigma}\mathbf{D}) \\
&= \mathrm{Trace}\left(\mathbf{D}'\mathbf{\Sigma}\left\{\sum_{j=1}^{q}\mathbf{B}^{(j)}\mathbf{B}^{(j)\prime}\right\}\mathbf{\Sigma}\mathbf{D}\right) \\
&= \mathrm{Trace}(\mathbf{D}'\mathbf{\Sigma}\mathbb{B}\mathbb{B}'\mathbf{\Sigma}\mathbf{D}).
\end{aligned}$$

Hence, it is well known that the matrix $\mathbf{V}$ which maximizes $\mathrm{Trace}(\mathbf{D}'\mathbf{\Sigma}\mathbb{B}\mathbb{B}'\mathbf{\Sigma}\mathbf{D})$ over the set of matrices $\mathbf{D}$ such that $\mathbf{D}'\mathbf{\Sigma}\mathbf{D} = \mathbf{I}_K$ is made of the K generalized eigenvectors

of $\boldsymbol{\Sigma}\mathbb{B}\mathbb{B}'\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}$ associated with K non-null eigenvalues. Thus, $\mathbf{V}$ contains the K eigenvectors of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}\mathbb{B}\mathbb{B}'\boldsymbol{\Sigma}) = \mathbb{B}\mathbb{B}'\boldsymbol{\Sigma}$ which are associated with K non-null eigenvalues.

In addition, we have that

$$\text{Span}(\boldsymbol{\Sigma}\mathbb{B}\mathbb{B}'\boldsymbol{\Sigma}) = \text{Span}(\boldsymbol{\Sigma}\mathbb{B})$$

because $\text{Span}(\boldsymbol{\Sigma}\mathbb{B}\mathbb{B}'\boldsymbol{\Sigma}) \subset \text{Span}(\boldsymbol{\Sigma}\mathbb{B})$ and $\dim(\text{Span}(\boldsymbol{\Sigma}\mathbb{B}\mathbb{B}'\boldsymbol{\Sigma})) = K$. Since $\boldsymbol{\Sigma}$ is invertible, this implies that

$$\text{Span}(\mathbb{B}\mathbb{B}'\boldsymbol{\Sigma}) = \text{Span}(\mathbb{B}).$$

Finally, we have under model (4.9) and assumption (LC) that for all $j \in \{1, \ldots, q\}$, $\text{Span}(\mathbf{B}) = \text{Span}(\mathbf{B}^{(j)})$ using univariate SIR theory. We then have $\text{Span}(\mathbb{B}) = \text{Span}(\mathbf{B})$ and then $\text{Span}(\mathbf{V}) = \text{Span}(\mathbf{B})$.

A.3.3 Proof of Theorem 10

For each component $y^{(j)}$ of $\mathbf{y}$ and under the assumptions (LC), (A1)-(A3), from univariate SIR theory of Li (1991), each estimated EDR space $\widehat{\mathbf{B}}^{(j)}$ converges to $\mathbf{B}^{(j)}$ at root n rate: that is, for $j = 1, \ldots, q$, $\widehat{\mathbf{B}}^{(j)} = \mathbf{B}^{(j)} + O_p(n^{-1/2})$. It follows that $\widehat{\mathbb{B}} = \mathbb{B} + O_p(n^{-1/2})$. Since $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + O_p(n^{-1/2})$, we get $\widehat{\mathbb{B}}\widehat{\mathbb{B}}'\widehat{\boldsymbol{\Sigma}} = \mathbb{B}\mathbb{B}'\boldsymbol{\Sigma} + O_p(n^{-1/2})$. Therefore, the eigenvectors associated with the largest K eigenvalues of $\widehat{\mathbb{B}}\widehat{\mathbb{B}}'\widehat{\boldsymbol{\Sigma}}$ converge to the corresponding ones of $\mathbb{B}\mathbb{B}'\boldsymbol{\Sigma}$ at the same rate: $\hat{\mathbf{v}}_k = v_k + O_p(n^{-1/2})$ for $k = 1, \ldots, K$. Consequently, the estimated EDR space $\text{Span}(\widehat{\mathbf{V}})$ converges to $\text{Span}(\mathbf{V})$ at rate $\sqrt{n}$. Since $\text{Span}(\mathbf{V}) = \text{Span}(\mathbf{B})$, the estimated EDR space converges to the true one in probability.

A.3.4 Proof of Theorem 11

We can proceed analogously to the proof of Theorem 10. It is sufficient to show that $\widehat{\mathbb{W}} = \mathbb{W} + O_p(n^{-1/2})$. Since, for each component $y^{(j)}$ of $\mathbf{y}$ and under assumptions (LC) and (A1)-(A3), from univariate SIR theory of Li (1991), each estimated eigenvalue $\hat{\lambda}_k^{(j)}$ converges to $\lambda_k^{(j)}$ at rate $\sqrt{n}$, we have $\hat{\pi}^{(j)} = \pi^{(j)} + O_p(n^{-1/2})$ and $\hat{\pi}_\star = \pi_\star + O_p(n^{-1/2})$. We thus get $\widehat{\mathbf{W}}^{(j)} = \mathbf{W}^{(j)} + O_p(n^{-1/2})$ and $\widehat{\mathbb{W}} = \mathbb{W} + O_p(n^{-1/2})$.

Consequently, $\widehat{\mathbb{B}}\widehat{\mathbb{W}}\widehat{\mathbb{B}}'\widehat{\boldsymbol{\Sigma}} = \mathbb{B}\mathbb{W}\mathbb{B}'\boldsymbol{\Sigma} + O_p(n^{-1/2})$ and the eigenvectors associated with the largest K eigenvalues of $\widehat{\mathbb{B}}\widehat{\mathbb{W}}\widehat{\mathbb{B}}'\widehat{\boldsymbol{\Sigma}}$ converge to the corresponding ones of $\mathbb{B}\mathbb{W}\mathbb{B}'\boldsymbol{\Sigma}$ at the same rate: $\widehat{\tilde{\mathbf{v}}}_k = \tilde{v}_k + O_p(n^{-1/2})$ for $k = 1, \ldots, K$. Therefore, the estimated EDR space $\text{Span}(\widehat{\widetilde{\mathbf{V}}})$ converges to $\text{Span}(\widetilde{\mathbf{V}})$ at root n rate. Since $\text{Span}(\widetilde{\mathbf{V}}) = \text{Span}(\mathbf{B})$, the estimated EDR space converges to the true one in probability.

A.3.5 Proof of Lemma 12

For $j = 1, \ldots, q$, recall that $P_{\hat{\mathbf{B}}^{(j)}, \hat{\boldsymbol{\Sigma}}}$ is the $\hat{\boldsymbol{\Sigma}}$-orthogonal projector onto $\text{Span}(\hat{\mathbf{B}}^{(j)})$. According to Theorem 4 of Saracco (1997), we have that $P_{\hat{\mathbf{B}}^{(j)}, \hat{\boldsymbol{\Sigma}}} = P_{\mathbf{B}_1, \boldsymbol{\Sigma}} + O_P(n^{-1/2})$, for

$j \in \{j_1, j_2\}$ and $P_{\hat{\mathbf{B}}^{(j)},\hat{\mathbf{\Sigma}}} = P_{\mathbf{B}_2,\mathbf{\Sigma}} + O_P(n^{-1/2})$, for $j \in \{j_3, j_4\}$. It follows that

$$\begin{aligned} r(\hat{\mathbf{B}}^{(j_1)}, \hat{\mathbf{\Sigma}}, \hat{\mathbf{B}}^{(j_2)}, \hat{\mathbf{\Sigma}}) &= \frac{1}{K}\text{Trace}(P_{\hat{\mathbf{B}}^{(j_1)},\hat{\mathbf{\Sigma}}} P_{\hat{\mathbf{B}}^{(j_2)},\hat{\mathbf{\Sigma}}}) \\ &= \frac{1}{K}\text{Trace}((P_{\mathbf{B}_1,\mathbf{\Sigma}} + O_P(n^{-1/2}))(P_{\mathbf{B}_1,\mathbf{\Sigma}} + O_P(n^{-1/2}))) \\ &= \frac{1}{K}\text{Trace}(P_{\mathbf{B}_1,\mathbf{\Sigma}}) + O_P(n^{-1/2}) \\ &= 1 + O_P(n^{-1/2}). \end{aligned}$$

Similarly, $r(\hat{\mathbf{B}}^{(j_3)}, \hat{\mathbf{\Sigma}}, \hat{\mathbf{B}}^{(j_4)}, \hat{\mathbf{\Sigma}})$ tends to 1 in probability when $n \to \infty$.

A.3.6 Proof that (4.28) provides a basis of the EDR space

We have, for $k = 1, \ldots, K$, $\Sigma^{1/2} M^+ \Sigma^{1/2} \eta_k = \alpha_k \eta_k$. By multiplying by the matrix $(M^+\Sigma^{1/2})$, we obtain:

$$M^+\Sigma^{1/2}\Sigma^{1/2}M^+\Sigma^{1/2}\eta_k = \alpha_k M^+\Sigma^{1/2}\eta_k, \quad \text{thus} \quad M^+\Sigma M^+\Sigma^{1/2}\eta_k = \alpha_k M^+\Sigma^{1/2}\eta_k.$$

Using the definition of b_k given in (4.28), we have: $M^+\Sigma b_k = \alpha_k b_k$. From the comments on the functional SIR context, provided in Section 4.4.2.4, the proof is complete.

A.3.7 Proof of Proposition 13

Let $(u, v) \in]c_{h,(i)}, c_{h,(i+1)}[\times]u, c_{h,(i+1)}[$ and $\left\{X_{(i)}\right\}_{i\in\{1,\ldots n\}}$ be the ordered sequence of the elements of $\mathbf{X}$. We will show that $\hat{f}_{K,h}(u)$ is neither greater nor lesser than $\hat{f}_{K,h}(v)$ with a proof by contradiction. Note that for the uniform kernel we have $\hat{f}_{K,h}(u) = \frac{1}{nh}\text{card}\left(\left\{X_k \in [u - \frac{h}{2}, u + \frac{h}{2}]\right\}\right)$.

If $\hat{f}_{K,h}(u) > \hat{f}_{K,h}(v)$, this implies that there exists at least one $k \in \{1, \ldots, n\}$, for which we have $X_{(k)} \in [u - \frac{h}{2}, v - \frac{h}{2}[$, which means that there exists $k' \in \{1, \ldots, w\}$ which satisfies $c_{h,(k')} = b_{h,(k)} \in [u, v[$. Because $[u, v[\subset]c_{h,(i)}, c_{h,(i+1)}[$, $c_{h,(k')} \in]c_{h,(i)}, c_{h,(i+1)}[$, which is impossible.

Conversely, $\hat{f}_{K,h}(v) > \hat{f}_{K,h}(u)$ implies that there exists $X_{(k)} \in]u + \frac{h}{2}, v + \frac{h}{2}]$. Then there exists $k' \in \{1, \ldots, w\}$ such that $c_{h,(k')} = a_{h,(k)} \in]u, v] \subset]c_{h,(i)}, c_{h,(i+1)}[$ and it is impossible.

A.3.8 Proof of Proposition 14

We will show the equivalence between the presence of a mode between $a_{h,(i)}$ and $b_{h,(j)}$ and the inequality $b_{h,(j-1)} < a_{h,(i)} \leq b_{h,(j)} < a_{h,(i+1)}$.

At first, we notice that ordered like this, there is no element of $\mathbf{A}_h$ or $\mathbf{B}_h$ that can be between $a_{h,(i)}$ and $b_{h,(j)}$. This is why the last inequality is equivalent to $\exists k \in \{1, \ldots, w - 1\}$, $a_{h,(i)} = c_{h,(k)}$ and $b_{h,(j)} = c_{h,(k+1)}$, provided that $a_{h,(i)} \neq b_{h,(j)}$.

From Proposition 13, $\hat{f}_{K,h}$ is constant on $]a_{h,(i)}, b_{h,(j)}[$, and thanks to Remark 16, it is equivalent to: $\hat{f}_{K,h}$ is constant on $[a_{h,(i)}, b_{h,(j)}] = [c_{h,(k)}, c_{h,(k+1)}]$. In order for this interval to be a mode, we must prove that there exists $\varepsilon > 0$ for which $\hat{f}_{K,h}$ is increasing on $[c_{h,(k)} - \varepsilon, c_{h,(k)}[$ and decreasing on $]c_{h,(k+1)}, c_{h,(k+1)} + \varepsilon]$, which is also made in Remark 16.

When $a_{h,(i)} = b_{h,(j)} = c_{h,(k)}$, $\hat{f}_{K,h}$ is increasing on $[c_{h,(k)} - \varepsilon, c_{h,(k)}[$ too and decreasing on $]c_{h,(k)}, c_{h,(k)} + \varepsilon]$. The mode is reduced to a single point.

A.3.9 Proof of Theorem 15

First, note that when K is the uniform kernel, for some $h > 0$, we can find $N(\hat{f}_{K,h})$ by counting the number of variations of sign of the following function

$$g'_{h,\varepsilon}(x) = \begin{cases} 1 & \text{for } x \in [a_{h,(i)} - \varepsilon, a_{h,(i)}[, \ \forall i \in \{1, \dots, \text{card}(\mathbf{A}_h)\}, \\ -1 & \text{for } x \in [b_{h,(i)}, b_{h,(i)} + \varepsilon[, \ \forall i \in \{1, \dots, \text{card}(\mathbf{B}_h)\}, \\ 0 & \text{elsewhere,} \end{cases}$$

where ε is chosen in a way that ensures that

$$\forall (i,j) \in \{1, \dots, \text{card}(\mathbf{A}_h)\} \times \{1, \dots, \text{card}(\mathbf{B}_h)\}, (a_{h,(i)} - b_{h,(j)}) \in]-\infty, 0] \cup]\varepsilon, \infty[,$$

in order to obtain a unique value of $g'_{h,\varepsilon}(x)$ for each x. The aim of $g'_{h,\varepsilon}$ is to mimic the derivative of $\hat{f}_{K,h}$. It seems easier to use than dirac functions involved in $\hat{f}'_{K,h}$. Besides, one can see that $N(g_{h,\varepsilon}) = N(\hat{f}_{K,h})$, using the fact that Proposition 14 is valid for $g_{h,\varepsilon}$. That is why the number of variations of sign of $g'_{h,\varepsilon}$ is equal to $2N(\hat{f}_{K,h}) - 1$.

Let $\mathbf{C}_{\varepsilon,n} = \left\{c_{h,\varepsilon,(i)}\right\}_{i \in \{1,\dots,w\}}$ be the ordered sequence made of the sets

$$\left\{a_{h,(i)} - \frac{\varepsilon}{2}\right\}_{i \in \{1,\dots,\text{card}(\mathbf{A}_h)\}} \text{ and } \left\{b_{h,(i)} + \frac{\varepsilon}{2}\right\}_{i \in \{1,\dots,\text{card}(\mathbf{B}_h)\}}.$$

Let

$$\begin{aligned} d_{h,\varepsilon,(i)} &= \mathbf{1}\left(c_{h,\varepsilon,(i)} \in \left\{a_{h,(i)} - \tfrac{\varepsilon}{2}\right\}_{i \in \{1,\dots,\text{card}(\mathbf{A}_h)\}}\right) \\ &- \mathbf{1}\left(c_{h,\varepsilon,(i)} \in \left\{b_{h,(i)} + \tfrac{\varepsilon}{2}\right\}_{i \in \{1,\dots,\text{card}(\mathbf{B}_h)\}}\right), \end{aligned}$$

and $\mathbf{D}_{\varepsilon,n} = \left\{d_{h,\varepsilon,(i)}\right\}_{i \in \{1,\dots,w\}}$. Every interval where $g'_{h,\varepsilon}(x) \neq 0$ is represented by a $c_{h,\varepsilon,(i)}$, then the number of variations of sign is the same for $g'_{h,\varepsilon}$ and for $\mathbf{D}_{\varepsilon,n}$. We write $v(\mathbf{D}_{\varepsilon,n})$ for the number of variations of sign of $\mathbf{D}_{\varepsilon,n}$ like Schoenberg (1950) did in his article.

Now, we prove that $v(\mathbf{D}_{\varepsilon,n}) \geq v(\mathbf{D}_{\varepsilon,n-1})$, for $n > 1$. This property is satisfied if $\mathbf{D}_{\varepsilon,n-1} = \mathbf{J}\mathbf{D}_{\varepsilon,n}$ where $\mathbf{J}$ is a totally positive matrix, following Schoenberg (1950). To define $\mathbf{J}$, we first focus on the case where the last point in the sample is different from the others. This means that if $\mathbf{\Omega}$ is our sample space, we define $\mathbf{\Omega}_1$ as:

$$\mathbf{\Omega}_1 = \{\omega \ : \ \forall i \in \{1, \dots, n-1\}, \ X_i(\omega) \neq X_n(\omega)\}.$$

We remark that, when our sample comes from $\omega \in \mathbf{\Omega}_1$, $\mathbf{D}_{\varepsilon,n-1}$ is constructed by removing two points in $\mathbf{D}_{\varepsilon,n}$. These points correspond to $c_{h,\varepsilon,(\gamma_1)} = X_n - h - \frac{\varepsilon}{2}$ and $c_{h,\varepsilon,(\gamma_2)} = X_n + h + \frac{\varepsilon}{2}$. This is why we have :

$$\mathbf{J} = \begin{pmatrix} & & 0 & \cdots & 0 & \cdots & 0 \\ & \mathbf{I}_{\gamma_1 - 1} & \vdots & & \vdots & & \vdots \\ & & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & & 0 & \cdots & 0 \\ \vdots & & \vdots & \mathbf{I}_{\gamma_2 - \gamma_1 - 1} & \vdots & & \vdots \\ 0 & \cdots & 0 & & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 0 & & \\ \vdots & & \vdots & & \vdots & \mathbf{I}_{w - \gamma_2} & \\ 0 & \cdots & 0 & \cdots & 0 & & \end{pmatrix}$$

where $\mathbf{I}_\gamma$ is the $\gamma \times \gamma$ identity matrix. It is straightforward to show that $\mathbf{J}$ is a totally positive matrix since all its minors are not negative.

If $\omega \notin \mathbf{\Omega}_1$, then $\mathbf{D}_{\varepsilon,n} = \mathbf{D}_{\varepsilon,n-1}$, because $\mathbf{A}_h$ and $\mathbf{B}_h$ stay the same if we build them with $(X_1, \ldots, X_n)$ or with $(X_1, \ldots, X_{n-1})$. Then $\mathbf{J} = \mathbf{I}_w$ and is totally positive.

To conclude, we write $\tilde{N}_{K,h} : n \mapsto \tilde{N}_{K,h}(n) = N(\hat{f}_{K,k})$. Recall that $\tilde{N}_{K,h}(n) = \frac{v\left(g'_{h,\varepsilon}\right)+1}{2} = \frac{v(\mathbf{D}_{\varepsilon,n})+1}{2}$. Because $n \mapsto v(\mathbf{D}_{\varepsilon,n})$ is increasing, $\tilde{N}_{K,h}$ is also an increasing function. Let $h_{crit,k,n}$ be the critical bandwidth defined in (A.1) for a sample of size n, then we have:

$$\forall h < h_{crit,k,n}, \quad \tilde{N}_{K,h}(n) > k.$$

Because $\tilde{N}_{K,h}$ increases with n, it follows that,

$$\forall \eta \in \mathbb{N},\ \forall h < h_{crit,k,n}, \quad \tilde{N}_{K,h}(n+\eta) > k.$$

Thus, with the definition of $h_{crit,k}$,

$$\forall \eta \in \mathbb{N}, \quad h_{crit,k,n+\eta} \geq h_{crit,k,n},$$

and the proof is complete.

List of works

- ## Articles

Articles accepted for publication:

Coudret R., Durrieu G. and Saracco J. Comparison of kernel density estimators with assumption on number of modes. *Communications in Statistics - Simulation and Computation.* See also Section 3.2.

Coudret R., Liquet B. and Saracco J. Comparison of sliced inverse regression approaches for underdetermined cases. *Journal de la Société Française de Statistique.* See also Section 4.4.

Articles in revision:

Azaïs R., Coudret R. and Durrieu G. A hidden renewal model for monitoring aquatic systems biosensors. Submitted in: *Environmetrics.* See also Section 3.1.

Coudret R., Girard S. and Saracco J. A new sliced inverse regression method for multivariate response regression. Submitted in: *Computational Statistics and Data Analysis.* See also Section 4.3.

- ## Technical report

Coudret R., Durrieu G. and Saracco J. (2012) A note about the critical bandwidth for a kernel density estimator with the uniform kernel. See also http://hal.archives-ouvertes.fr/hal-00765843 and Appendix A.1.

- **Conferences**

  Coudret R., Durrieu G. and Saracco J. (2012). Comparison of kernel density estimators with assumption on number of modes. *20^{th} International Conference on Computational Statistics*, Limassol, Cyprus.

  Coudret R., Durrieu G. and Saracco J. (2012). Estimateurs à noyau bimodaux d'une densité bimodale et comparaison avec d'autres estimateurs non paramétriques. *$44^{èmes}$ Journées de Statistique de la SFdS*, Brussels, Belgium.

  Coudret R., Durrieu G. and Saracco J. (2012). Une interface graphique pour analyser des données distantes sous R. *$1^{ères}$ Rencontres R*, Bordeaux.

  Coudret R. (2011). Estimateurs non paramétriques appliqués à des données valvométriques. *$4^{èmes}$ Rencontres des Jeunes Statisticiens*, Aussois.

- **Seminars**

  Azaïs, R., Coudret R. and Durrieu G. (2013). Un processus de renouvellement pour des bioindicateurs de surveillance des milieux aquatiques. *Séminaire de Probabilités et Statistique*, Bordeaux.

  Azaïs, R. and Coudret R. (2013). Huître ouvre-toi ! (Des huîtres comme bioindicateurs de la qualité de l'eau). *Unithé ou café*, Inria, Bordeaux.

  Azaïs, R., Coudret R. and Durrieu G. (2013). Un processus de renouvellement pour des bioindicateurs de surveillance des milieux aquatiques. *Journée systèmes dynamiques, probabilités et statistique*, Quimper.

  Coudret R. and Fourestier S. (2013). Utiliser R : Spécificités du logiciel et exemples d'application. *Mardis du développement technologique*, Inria, Bordeaux.

  Coudret R., Durrieu G. and Saracco J. (2012). Comparaison d'estimateurs d'une densité avec hypothèse sur son nombre de modes. *Séminaire de mathématiques du LMBA*, Vannes.

Bibliography

Adler, J. (2010). *R in a Nutshell*. O'Reilly Media.

Amato, U., Antoniadis, A., and de Feis, I. (2006). Dimension reduction in functional regression with applications. *Computational Statistics & Data Analysis*, 50 (9):2422–2446.

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

Antoniadis, A., Bigot, J., and Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: a comparative study. *Journal of Statistical Software*, 6(6):1–83.

Aragon, Y. (1997). A gauss implementation of multivariate sliced inverse regression. *Computational Statistics*, 12:355–372.

Aragon, Y. and Saracco, J. (1997). Sliced Inverse Regression (SIR): an appraisal of small sample alternatives to slicing. *Computational Statistics*, 12:109–130.

Azaïs, R., Dufour, F., and Gégout-Petit, A. (In press). Nonparametric estimation of the jump rate for non-homogeneous marked renewal processes. *Annales de l'Institut Henri Poincaré*. Preprint arXiv:1202.2211v2.

Azaïs, R., Gégout-Petit, A., and Saracco, J. (2012). Optimal quantization applied to sliced inverse regression. *Journal of Statistical Planning and Inference*, 142(2):481–492.

Bai, Z. D. and He, X. (2004). A chi-square test for dimensionality for non-gaussian data. *Journal of Multivariate Analysis*, 88:109–117.

Barreda, L., Gannoun, A., and Saracco, J. (2007). Some extensions of multivariate SIR. *Journal of Statistical Computation and Simulation*, 77(1-2):1–17.

Barrios, M. P. and Velilla, S. (2007). A bootstrap method for assessing the dimension of a general regression problem. *Statistics & Probability Letters*, 77(3):247–255.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 57(1):289–300.

Bercu, B. and Chafaï, D. (2007). *Modélisation stochastique et simulation: Cours et applications*. Sciences sup. Dunod.

Bercu, B., Nguyen, T., and Saracco, J. (2011). A new approach of recursive and non recursive sir methods. *Journal of the Korean Statistical Society*, 41:17–36.

Bernard, A. (2008). Cadmium & its adverse effects on human health. *Indian Journal of Medical Research*, 128(4):557.

Bernard-Michel, C., Douté, S., Fauvel, M., Gardes, L., and Girard, S. (2009a). Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression. *Journal of Geophysical Research - Planets*, 114(E06005).

Bernard-Michel, C., Gardes, L., and Girard, S. (2008). A note on sliced inverse regression with regularizations. *Biometrics*, 64:982–986.

— (2009b). Gaussian Regularized Sliced inverse Regression. *Statistics and Computing*, 19:85–98.

Besse, P. (2012). Exploration statistique multidimensionnelle. http://www.math.univ-toulouse.fr/~besse/pub/Explo_stat.pdf.

Borcherding, J. and Volpers, M. (1994). The dreissena monitor - first results on the application of this biological early warning system in the continuous monitoring of water quality. *Water Science Technology*, 29:199–201.

Bosq, D. and Blanke, D. (2008). *Inference and prediction in large dimensions*. John Wiley & Sons.

Bottolo, L., Chadeau-Hyam, M., Hastie, D. I., Langley, S. R., Petretto, E., Tiret, L., Tregouet, D., and Richardson, S. (2011). ESS^{++}: a C^{++} objected-oriented algorithm for bayesian stochastic search model exploration. *Bioinformatics*, 27(4):587–588.

Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for bayesian model exploration. *Bayesian Analysis*, 5(3):583–618.

Bura, E. (1997). *L_1-Statistical Procedures and Related Topics*, chapter Dimension reduction via parametric inverse regression. Institute of Mathematical Statistics, Hayward, pages 215–228.

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*, 63:393–410.

Call, D. J., Polkinghorne, C. N., Markee, T. P., Brooke, L. T., Geiger, D. L., Gorsuch, J. W., and Robillard, K. A. (1999). Silver toxicity to chironomus tentans in two freshwater sediments. *Environmental Toxicology and Chemistry*, 18(1):30–39.

Carroll, R. J. and Li, K.-C. (1992). Measurement error regression with unknown link: dimension reduction and data visualization. *Journal of the American Statistical Association*, 87(420):1040–1050.

Casas, S. (2005). *Modélisation de la bioaccumulation de métaux traces (Hg, Cd, Pb, Cu et Zn) chez la moule, Mytilus Galloprovincialis, en milieu méditerranéen*. Ph.D. thesis, Université du Sud Toulon-Var.

Causeur, D., Friguet, C., Houee-Bigot, M., and Kloareg, M. (2011). Factor analysis for multiple testing (FAMT): an R package for large-scale significance testing under dependence. *Journal of Statistical Software.*, 40(14).

Chambon, C., Legeay, A., Durrieu, G., Gonzalez, P., Ciret, P., and Massabuau, J.-C. (2007). Influence of the parasite worm Polydora sp. on the behaviour of the oyster crassostrea gigas: a study of the respiratory impact and associated oxidative stress. *Marine Biology*, 152(2):329–338.

Chavent, M., Kuentz, V., Liquet, B., and Saracco, J. (2011). A sliced inverse regression approach for a stratified population. *Communications in statistics - Theory and Methods*, 40:1–22.

Chen, C.-H. and Li, K.-C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica*, 8(2):289–316.

Chen, Y., Petersen, S., Pacyna-Gengelbach, M., Pietas, A., and Petersen, I. (2003). Identification of a novel homeobox-containing gene, lagy, which is downregulated in lung cancer. *Oncology*, 64(4):450–458.

Coifman, R. R. and Donoho, D. L. (1995). *Wavelets and Statistics*, volume 103 of *Lecture Notes in Statistics*, chapter Translation-invariant de-noising. Springer-Verlag, pages 125–150.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89:177–189.

— (1998). Principal hessian directions revisited (with discussion). *Journal of the American Statistical Association*, 93:84–100.

— (2000). SAVE: a method for dimension reduction and graphics in regression. *Communications in statistics - Theory and Methods*, 29:2109–2121.

— (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, 32(3):1062–1092.

— (2009). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley Series in Probability and Statistics. Wiley.

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30:450–474.

Cook, R. D. and Nachtsheim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association*, 89:592–599.

Cook, R. D. and Setodji, C. M. (2003). A model-free test for reduced rank in multivariate regression. *Journal of the American Statistical Association*, 98(462):340–351.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics.

Devroye, L. P. (1987). *A Course in Density Estimation*. Birkäuser, Boston.

Devroye, L. P. and Wagner, P. J. (1980). The strong uniform consistency of kernel density estimates. In P. R. Krishnaiah (Ed.), *Multivariate Analysis V: Proceedings of the fifth International Symposium on Multivariate Analysis, Volume 5*. North-Holland Pub. Co., pages 59–77.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

— (1995). Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224.

— (1998). Minimax Estimation via Wavelet Shrinkage. *The Annals of Statistics*, 26(3):879–921.

— (1999). Asymptotic Minimaxity of Wavelet Estimators with Sampled Data. *Statistica Sinica*, 9(1):1–32.

Douté, S., Schmitt, B., Langevin, Y., Bibring, J.-P., Altieri, F., Bellucci, G., Gondet, B., and Poulet, F. (2007). South pole of Mars: Nature and composition of the icy terrains from Mars Express OMEGA observations. *Planetary and Space Science*, 55(1-2):113–133.

Duan, N. and Li, K.-C. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics*, 19:505–530.

Durrieu, G., Maury-Brachet, R., and Boudou, A. (2005). Goldmining and mercury contamination of the piscivorous fish Hoplias aimara in French Guiana (Amazon basin). *Ecotoxicology and Environmental Safety*, 60(3):315–323.

Einbeck, J. and Taylor, J. (2013). A number-of-modes reference rule for density estimation under multimodality. *Statistica Neerlandica. Journal of the Netherlands Society for Statistics and Operations Research*, 67:54–66.

Ferré, L. (1998). Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association*, 93:132–140.

Ferré, L. and Yao, A.-F. (2007). Reply to the paper : "A note on smoothed functional inverse regression". *Statistica Sinica*, 17:1683–1687.

Fleming, T. R. and Harrington, D. P. (1984). Nonparametric estimation of the survival distribution in censored data. *Communications in Statistics - Theory and Methods*, 13(20):2469–2486.

Friguet, C., Kloareg, M., and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104(488):1406–1414.

Futschik, A. and Isogai, E. (2006). On the consistency of kernel density estimates under modality constraints. *Statistics & Probability Letters*, 76(16):431–437.

Galtsoff, P. S. (1938). Physiology of reproduction of ostrea virginica I. spawning reactions of the female and male. *The Biological Bulletin*, 74(3):461–486.

Gannoun, A. and Saracco, J. (2003). An asymptotic theory for SIR_α method. *Statistica Sinica*, 13:297–310.

Geman, D. and Horowitz, J. (1980). Occupation densities. *The Annals of Probability*, 8(1):1–67.

Golub, G. H. and Van Loan, C. F. (1983). *Matrix computations*, volume 3 of *Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins University Press, Baltimore, MD.

Hall, P. and Li, K.-C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, 21:867–889.

Hall, P. and Marron, J. S. (1987). Estimation of integrated squared density derivatives. *Statistics & Probability Letters*, 6(2):109–115.

Hall, P., Minnotte, M. C., and Zhang, C. (2004). Bump hunting with non-gaussian kernels. *The Annals of Statistics*, 32(5):2124–2141.

Hall, P. and Patil, P. (1996a). Effect of threshold rules on performances of wavelet-based curve estimators. *Statistica Sinica*, 6(2):331–345.

— (1996b). On the Choice of Smoothing Parameter, Threshold and Truncation in Nonparametric Regression by Non-linear Wavelet Methods. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 58(2):361–377.

Hall, P. and York, M. (2001). On the calibration of Silverman's test for multimodality. *Statistica Sinica*, 11:515–536.

Hsing, T. (1999). Nearest neighbor inverse regression. *The Annals of Statistics*, 27:697–731.

Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, 20:1040–1061.

Hughes, M. F. (2002). Arsenic toxicity and potential mechanisms of action. *Toxicology letters*, 133(1):1–16.

Järup, L., Berglund, M., Elinder, C. G., Nordberg, G., and Vanter, M. (1998). Health effects of cadmium exposure–a review of the literature and a risk estimate. *Scandinavian Journal of Work, Environment & Health*, 24:1–51.

Jenner, H., Noppert, F., and Sikking, T. (1989). A new system for the detection of valve movement response of bivalves. *Kema Scientific and Technical Reports*, 7(2):91–98.

Jones, M. C. and Sheather, S. J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistics & Probability Letters*, 11(6):511–514.

Jou, L. J. and Liao, C. M. (2006). A dynamic artificial clam (corbicula fluminea) allows parcimony on-line measurement of waterborne metals. *Environmental Pollution*, 144:172–183.

Kötter, T. T. (2000). *Smoothing and regression. Approaches, computation and application.*, chapter Sliced inverse regression. Wiley, Chichester, pages 497–512.

Kramer, K., Jenner, H., and De Zwart, D. (1989). The valve movement response of mussels: a tool in biological monitoring. *Hydrobiologia*, 188/189:433–443.

Kuentz, V., Liquet, B., and Saracco, J. (2010). Bagging versions of sliced inverse regression. *Communications in statistics - Theory and Methods*, 39(11):1985–1996.

Kuentz, V. and Saracco, J. (2010). Cluster-based sliced inverse regression. *Journal of the Korean Statistical Society*, 39(2):251–267.

Lafaye De Micheaux, P., Drouilhet, R., and Liquet, B. (2010). *Le logiciel R - Maîtriser le langage - Effectuer des analyses statistiques.* Statistiques et probabilités appliquées. Springer.

Li, B., Wen, S., and Zhu, L. (2008). On a projective resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association*, 103(483):1177–1186.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

— (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association*, 87:1025–1039.

Li, K.-C., Aragon, Y., Shedden, K., and Agnan, C. T. (2003). Dimension reduction for multivariate response data. *Journal of the American Statistical Association*, 98(461):99–109.

Li, L. and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. *Technometrics*, 48(4):503–510.

Li, L. and Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics*, 64(1):124–131.

Li, Y. and Zhu, L.-X. (2007). Asymptotics for sliced average variance estimation. *The Annals of Statistics*, 35:41–69.

Liquet, B. and Saracco, J. (2008). Application of the bootstrap approach to the choice of dimension and the α parameter in the SIR_α method. *Communications in statistics - Simulation and Computation*, 37(6):1198–1218.

— (2012). A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches. *Computational Statistics*, 27:103–125.

Lue, H.-H. (2009). Sliced inverse regression for multivariate response regression. *Journal of Statistical Planning and Inference*, 139(8):2656–2664.

Lê Cao, K.-A., Martin, P., Robert-Granie, C., and Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10(1):34.

Mallat, S. (2008). *A Wavelet Tour of Signal Processing: The Sparse Way*. Elsevier Science.

Mammen, E., Marron, J. S., and Fisher, N. J. (1991). Some asymptotics for multimodality tests based on kernel density estimates. *Probability Theory and Related Fields*, 91:115–132.

Mangiarotti, S., Coudret, R., Drapeau, L., and Jarlan, L. (2012). Polynomial search and global modeling: Two algorithms for modeling chaos. *Physical Review E. Statistical, Nonlinear, and Soft Matter Physics*, 86.

Marron, J. and Wand, M. (1992). Exact mean integrated squared error. *The Annals of Statistics*, 20(2):712–736.

Minnotte, M. C. (1997). Nonparametric testing of the existence of modes. *The Annals of Statistics*, 25(4):1646–1660.

Minnotte, M. C., Marchette, D. J., and Wegman, E. J. (1998). The Bumpy Road to the Mode Forest. *Journal of Computational and Graphical Statistics*, 7(2):239–251.

Minnotte, M. C. and Scott, D. W. (1993). The mode tree: A tool for visualization of nonparametric density features. *Journal of Computational and Graphical Statistics*, 2(1):51–68.

Moler, C. B. and Stewart, G. W. (1973). An algorithm for generalized matrix eigenvalue problems. *SIAM Journal on Numerical Analysis*, 10(2):241–256.

Müller, D. W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86(415):738–746.

Nagai, K., Honjo, T., Go, J., Yamashita, H., and Oh, S. (2006). Detecting the shellfish killer heterocapsa circularisquama (dinophyceae) by measuring the bivalve valve activity with hall element sensor. *Aquaculture*, 255:395–401.

Nkiet, G.-M. (2008). Consistent estimation of the dimensionality in sliced inverse regression. *Annals of the Institute of Statistical Mathematics*, 60(2):257–271.

O'Connell, J. and Højsgaard, S. (2011). Hidden semi markov models for multiple observation sequences: The mhsmm package for R. *Journal of Statistical Software*, 39(4).

Parzen, E. (1962). On estimation of probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Petretto, E., Bottolo, L., Langley, S. R., Heinig, M., McDermott-Roe, C., Sarwar, R., Pravenec, M., Hübner, N., Aitman, T. J., Cook, S. A., and Richardson, S. (2010). New insights into the genetic control of gene expression using a bayesian multi-tissue approach. *PLoS Computational Biology*, 6(4):e1000737.

Polonik, W. (1995a). Density estimation under qualitative assumptions in higher dimensions. *Journal of Multivariate Analysis*, 55(2):61–81.

— (1995b). Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics*, 23(3):855–881.

Prendergast, L. A. (2005). Influence functions for sliced inverse regression. *Scandinavian Journal of Statistics*, 32(3):385–404.

— (2007). Implications of influence function analysis for sliced inverse regression and sliced average variance estimation. *Biometrika*, 94(3):585–601.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ratte, H. T. (1999). Bioaccumulation and toxicity of silver compounds: A review. *Environmental Toxicology and Chemistry*, 18(1):89–108.

Robson, A., Wilson, R., and Garcia de Leaniz, C. (2007). Mussels flexing their muscles: a new method for quantifying bivalve behavior. *Marine Biology*, 151:1195–1204.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 42(1):43–47.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(2):65–78.

Sabatier, R. and Reynès, C. (2008). Extensions of simple component analysis and simple linear discriminant analysis using genetic algorithms. *Computational Statistics & Data Analysis*, 52(10):4779–4789.

Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Communications in statistics - Theory and Methods*, 26:2141–2171.

— (1999). Sliced inverse regression under linear constraints. *Communications in statistics - Theory and Methods*, 28(10):2367–2393.

— (2001). Pooled slicing methods versus slicing methods. *Communications in statistics - Simulation and Computation*, 30:489–511.

— (2005). Asymptotics for pooled marginal slicing estimator based on SIR_α approach. *Journal of Multivariate Analysis*, 96:117–135.

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., *et al.* (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302.

Schmitt, F., De Rosa, M., Durrieu, G., Sow, M., Ciret, P., Tran, D., and Massabuau, J.-C. (2011). Statistical study of bivalve high frequency microclosing behavior: Scaling properties and shot noise analysis. *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, 21(12):1–12.

Schoenberg, I. J. (1950). On pólya frequency functions II: Variation diminishing integral operators of the convolution type. *Acta Universitatis Szegediensis. Acta Scientiarum Mathematicarum*, 12:97–106.

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, 89:141–148.

Schwartzmann, C., Durrieu, G., Sow, M., Ciret, P., Lazareth, C. E., and Massabuau, J.-C. (2011). In situ giant clam growth rate behavior in relation to temperature: A one-year coupled study of high-frequency noninvasive valvometry and sclerochronology. *Limnology and Oceanography*, 56(5):1940–1951.

Scott, D. W. (1992). *Multivariate Density Estimation : Theory, Practice and Visualization*. John Wiley & Sons.

Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 62(400):1131–1146.

Scrucca, L. (2007). Class prediction and gene selection for DNA microarrays using regularized sliced inverse regression. *Computational Statistics & Data Analysis*, 52(1):438–451.

Setodji, C. M. and Cook, R. D. (2004). K-means inverse regression. *Technometrics*, 46:421–429.

Shao, Y., Cook, R. D., and Weisberg, S. (2009). Partial central subspace and sliced average variance estimation. *Journal of Statistical Planning and Inference*, 139(3):952–961.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 53(3):683–690.

Shin, C. H., Liu, Z.-P., Passier, R., Zhang, C.-L., Wang, D.-Z., Harris, T. M., Yamagishi, H., Richardson, J. A., Childs, G., and Olson, E. N. (2002). Modulation of cardiac growth and development by HOP, an unusual homeodomain protein. *Cell*, 110(6):725–735.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 43(1):97–99.

— (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Sloff, W., De Zwart, D., and Marquenie, J. (1983). Detection limits of a biological monitoring system for chemical water pollution based on mussel activity. *Environmental Contamination and Toxicology*, 30(4):400–405.

Sow, M., Durrieu, G., Briollais, L., Ciret, P., and Massabuau, J.-C. (2011). Water quality assessment by means of HFNI valvometry and high-frequency data modeling. *Environmental Monitoring and Assessment*, 182(1-4):155–170.

Szretter, M. E. and Yohai, V. J. (2009). The sliced inverse regression algorithm as a maximum likelihood procedure. *Journal of Statistical Planning and Inference*, 139(10):3570–3578.

Teetor, P. (2011). *R Cookbook*. O'Reilly Media.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tran, D., Ciret, P., Ciutat, A., Durrieu, G., and Massabuau, J.-C. (2003). Estimation of potential and limits of bivalve closure response to detect contaminants: application to cadmium. *Environmental Toxicology and Chemistry*, 22(4):914–920.

Tran, D., Fournier, E., Durrieu, G., and Massabuau, J.-C. (2004). Copper detection in the Asiatic clam Corbicula fluminea: Optimum valve closure response. *Aquatic Toxicology*, 66:333–343.

— (2007). Inorganic mercury detection by valve closure response in the freshwater clam corbicula fluminea: integration of time and water metal concentration changes. *Environmental Toxicology and Chemistry*, 26:1545–1551.

Tran, D., Haberkorn, H., Soudant, P., Ciret, P., and Massabuau, J.-C. (2010). Behavioral responses of Crassostrea gigas exposed to the harmful algae Alexandrium minutum. *Aquaculture*, 298:338–345.

Tran, D., Nadau, A., Durrieu, G., Ciret, P., Parisot, J.-P., and Massabuau, J.-C. (2011). Field chronobiology of a molluscan bivalve: How the moon and the sun cycles interact to drive oyster activity rhythms. *Chronobiology International*, 28(4):307–317.

Tyler, D. E. (1981). Asymptotic inference for eigenvectors. *The Annals of Statistics*, 9(4):725–736.

van Dorp, J. R. and Kotz, S. (2002). A Novel Extension of the Triangular Distribution and Its Parameter Estimation. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 51(1):63–79.

Vines, S. (2000). Simple principal components. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(4):441–451.

Walnut, D. F. (2004). *An Introduction to Wavelet Analysis*. Birkhäuser.

Wang, Q. and Phillips, P. C. B. (2009). Asymptotic theory for local time density estimation and nonparametric cointegrating regression. *Econometric Theory*, 25(3):710–738.

Watwood, S. L., Miller, P. J. O., Johnson, M., Madsen, P. T., and Tyack, P. L. (2006). Deep-diving foraging behaviour of sperm whales (physeter macrocephalus). *Journal of Animal Ecology*, 75(3):814–825.

Wu, H.-M. (2008). Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17(3):590–610.

Yamaguchi, S., Asanoma, K., Takao, T., Kato, K., and Wake, N. (2009). Homeobox gene HOPX is epigenetically silenced in human uterine endometrial cancer and suppresses estrogen-stimulated proliferation of cancer cells by inhibiting serum response factor. *International journal of cancer*, 124(11):2577–2588.

Yin, X. and Bura, E. (2006). Moment-based dimension reduction for multivariate response regression. *Journal of Statistical Planning and Inference*, 136:3675–3688.

Yin, X. and Seymour, L. (2007). Asymptotic distributions for dimension reduction in the sir-ii method. *Statistica Sinica*, 15:1069–1079.

Yoo, J. K. (2009). Iterative optimal sufficient dimension reduction for conditional mean in multivariate regression. *Journal of Data Science*, 7:267–276.

Zhong, W., Zeng, P., Ma, P., Liu, J. S., and Zhu, Y. (2005). RSIR: regularized sliced inverse regression for motif discovery. *Bioinformatics*, 21(22):4169–4175.

Zhu, L., Miao, B., and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474):630–643.

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.-P. and Zhu, L.-X. (2007). On kernel method for sliced average variance estimation. *Journal of Multivariate Analysis*, 98:970–991.

Zhu, L.-P., Zhu, L.-X., and Feng, Z.-H. (2010a). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association*, 105(492):1455–1466.

Zhu, L.-P., Zhu, L.-X., and Wen, S.-Q. (2010b). On dimension reduction in regressions with multivariate responses. *Statistica Sinica*, 20(3):1291–1307.

Zhu, L.-X. and Fang, K.-T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, 24:1053–1068.

Zhu, L.-X. and Ng, K.-W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica*, 5:727–736.

Zhu, L.-X., Ohtaki, M., and Li, Y. (2007). On hybrid methods of inverse regression-based algorithms. *Computational Statistics*, 51:2621–2635.