Dimension reductio via Sliced Inverse Regression : ideas and extensions

Alessandro Chiancone

▶ **To cite this version:**

Alessandro Chiancone. Dimension reductio via Sliced Inverse Regression : ideas and extensions. Complex Variables [math.CV]. Université Grenoble Alpes, 2016. English. NNT : 2016GREAM051 . tel-01571824v2

HAL Id: tel-01571824

https://theses.hal.science/tel-01571824v2

Submitted on 10 Jan 2018

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.

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques Appliquées**

Arrêté ministériel : 7 août 2006

Présentée par

Alessandro CHIANCONE

Thèse dirigée par **Stéphane GIRARD** et
codirigée par **Jocelyn CHANUSSOT**

préparée au sein d' **Inria Grenoble Rhône Alpes et du GIPSA-Lab** dans **l'École Doctorale MSTII**

Réduction de dimension via Sliced Inverse Regression: Idées et nouvelles propositions

Thèse soutenue publiquement le **28 Octobre 2016**
devant le jury composé de :

Mme Anne-Françoise YAO
Professeur, Université Balise Pascal, Président
Mme Marie CHABERT
Professeur, INP Toulouse, Rapporteur
M. Jérôme SARACCO
Professeur, Institut Polytechnique de Bordeaux, Rapporteur
Mme Florence FORBES
Directeur de Recherche, Inria Grenoble Rhône-Alpes, Examinateur
M. Stéphane GIRARD
Directeur de Recherche, Inria Grenoble Rhône-Alpes, Directeur de thèse
M. Jocelyn CHANUSSOT
Professeur, Grenoble-INP, co-Directeur de thèse

ABSTRACT

In this thesis, Sliced Inverse Regression (SIR), a method for semi-parametric dimension reduction is discussed, analyzed and extended. Three different contributions namely, Collaborative SIR, Student SIR and Knockoff SIR are presented and dis-discussed. Collaborative SIR aims at finding subgroups in the data that have different characteristics and that are better described dividing the dataset. Student SIR is a robustified version of SIR where the error is described by a multivariate t-Student distribution, a heavy tailed distribution that is flexible to outliers. Finally Knockoff SIR is a method to perform variable selection and to provide sparse solutions at the same time. The basic idea comes from a paper of R. F. Barber and E. J. Candes that controls the false discovery rate in regression procedure such as LASSO.

In the first chapter of the thesis, SIR is presented and discussed, an analysis of the state of the art is detailed. The last part of the chapter is dedicated to give an overview of the three different contributions. The second chapter focuses on Collaborative SIR and includes the paper published in Communications in Statistics - Theory and Methods. Student SIR is treated in chapter 3 where the paper is published in Computational Statistics & Data Analysis is shown. Finally Knockoff SIR is outlined and the main results are presented and discussed providing applications on simulated and real data.
Dulcis in fundo the conclusion is drawn.

At the beginning of each chapter a quote from a famous writer is placed with the aim to help the reader to enter in the theory that is then proposed.

Dedication and acknowledgements

To the ones that are part of my life that I like to call family.

To love. To be loved. To never forget your own insignificance.
To never get used to the unspeakable violence and the vulgar disparity of life around you.
To seek joy in the saddest places. To pursue beauty to its lair.
To never simplify what is complex or complicate what is simple.
To respect strength, never power.
Above
all, to watch. To try and understand. To never look away. And never, never to forget.

Arundhati Roy, The Cost of Living

This thesis would not have been possible without the constant help and support of my supervisors: Stéphane Girard, Jocelyn Chanussot and Florence Forbes the best possible guides I can imagine in the wildlife of Statistics. I deeply thank Jérôme Saracco and Marie Chabert for their valuable comments on the manuscript, for their time and disponibility before and during the defense. A special thanks goes to professor Anne-Françoise Yao for being the president of the jury and for the interesting questions raised during the defense.

TABLE OF CONTENTS

LIST OF TABLES

TABLE **Page**

List of Figures

Figure **Page**

CHAPTER 1

INTRODUCTION

You see things; and you say, 'Why?'
But I dream things that never were; and I say 'Why not?
B. Shaw.

Suppose to observe a group of bombers returning after a mission. The undamaged plane (figure 1.1) on the left and on the right, in black, all parts hit by bullets.

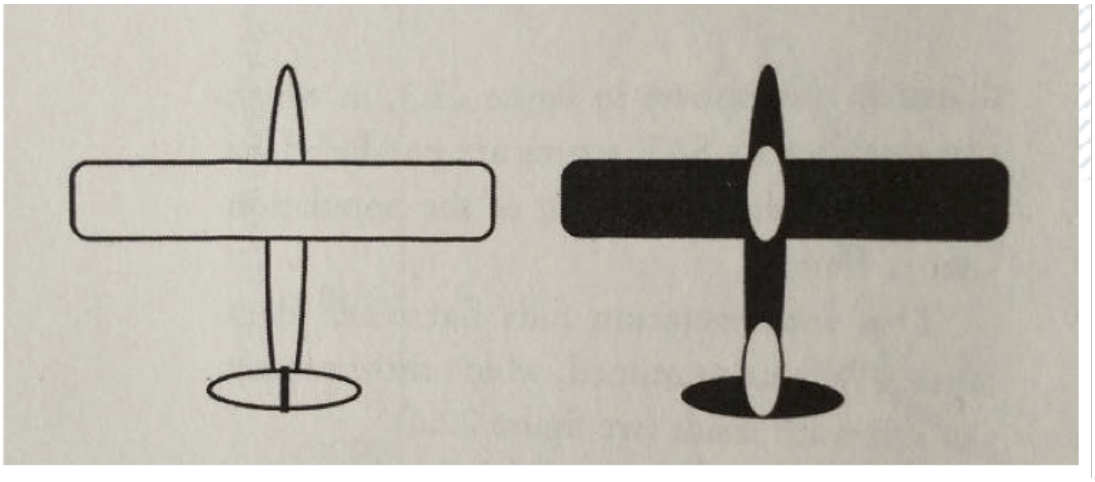


FIGURE 1.1. Undamaged plane on the left, scheme of all damages (in black) reported by bombers in action.

During World War Two, the Army Air Force asked how could they improve the odds of a bomber making it home. Military engineers explained to the statistician that they already knew the allied bombers needed more armor, but not where to place it since it was not possible to cover all the aircraft. The military looked at the bombers that had returned from enemy territory. They recorded where those planes had taken the most damage. According to where the hits tended to group they wanted to place the armor. The Mathematicians and Statistician Abraham Wald changed perspective dramatically and said: the holes show where a bomber can be shot and still survive the flight home. They idea of Wald was to deeply take into account the fact that some very useful information was buried with the planes that never made it to home. Based on this idea he developed a framework to deal with such a situation [57] (Wald was part of the Statistical Research Group (SRG) founded in that period to assist the army during the war).

Taking the scheme of this experiment it is possible to build a simple but explicative model, let us divide the airplane in five different areas, following Wald's example, $A_1, A_2, ..., A_5$ and store the area hit by bullets for different planes in five corresponding continuous variables $(x_1, x_2, ..., x_5)$. The goal of the analysis is to estimate the *full model* function:

$$y = f(x_1, x_2, ..., x_5) \tag{1.1}$$

where $y \in [0,1]$ is the damage of the bomber, 0 is undamaged and 1 is downed. One first assumption could be that the damage is a function of the total area hit by bullets $y = f(x_1 + x_2 + ... + x_5) = f(\beta^T \mathbf{X})$, where $\beta = (1,1,...,1)$ and $\mathbf{X} = (x_1, x_2, ..., x_5)^T$. A different assumption could take into account the vulnerability of different parts giving a weight β proportional to the expected vulnerability of the different parts (e.g. engine, fuselage, fuel system). A hit on the engine should be more critical than one on the fuselage. The model:

$$y = f(\beta^T \mathbf{X}) \tag{1.2}$$

gives some freedom to take into account different settings since there are no assumptions on the link function f but only on the argument which is supposed to be a linear combination of the initial predictors. In general β is unknown, it is of interest to try to look for such a vector since the link function, in our example, under the *full model* (1.1) is defined on $\mathbb{R}^5$ while under model (1.2) is defined on $\mathbb{R}$. In other words the information is packed in just one number (e.g. the total number of hits) instead of living in higher

dimension, a dimension reduction is achieved. Regression is well known to be hard when the dimension of the predictors is high.

The chapter is organized as follows: first a brief introduction about regression in the general setting is sketched, a detailed description of SIR is given in section 1.2 together with comments, analysis, the algorithm and the state of the art. The following three sections are dedicated to an overview of respectively Collaborative SIR, Student SIR and Knockoff SIR.

1.1 Regression

During the first years of 1800 Legendre and Gauss shaped a form of reasoning and approach that has been then named Regression. It was a crucial moment for Statistics and the first and most famous priority dispute over the discovery in this field [65].
Regression analysis is a complex field with the aim of finding relationships among variables, in particular when a dependent variable Y and independent variables $\mathbf{X}$ are taken into account. The assumption is that:

$$Y = f(\mathbf{X}, \epsilon) \tag{1.3}$$

where ϵ is a random noise independent of $\mathbf{X}$. Once the link function f is found it is possible to forecast Y based on the observed value of $\mathbf{X}$. Suppose, for example, that the professor is asked to forecast the grades of his students based on some parameters (e.g. number of lectures attended, grades in other subjects). It is evident from the example that this analysis can be challenging and that the assumption of a link function f between the response variable Y and the predictor space $\mathbf{X}$ is non trivial and highly debated. Recently the problem of correlation vs causation has been of certain interest in economics (in 2003 Clive Granger and Robert Engle were jointly awarded the Nobel Memorial Prize in Economic Sciences). Depending on the assumption on the function f regression analysis is commonly divided into parametric and non parametric. In the first case is assumed that the function f depends on a set of parameters $f(\cdot, \beta)$ and that the function belongs to a pre specified parametric family. In the second case f is not assumed to be part of a specific parametric family and the analysis is carried out based on the data point positions in the space. Generally speaking, the flexibility of the non parametric models has the drawback of a higher number of points needed to correctly guess the shape of the function (in particular when the dimension of the predictor space is

high). Along the spectra of possible methods in between parametric and non-parametric models semi-parametric models try to combine the two approaches. Almost two hundred years after the beginning of regression analysis, in 1991, the advances in technology brought the attention of the statisticians to new problems concerning the amount of variables that one could explore in a regression procedure. Visualization was developing really fast but the capability of gathering data was even faster. Scanning a large pool of variable became challenging and new theories emerged from this need to surf and face the amount of data. Sliced Inverse Regression [47] opened a new way to achieve dimensionality reduction when dealing with a regression problem, in such a way to avoid parametric or non parametric model-fitting.

1.2 Sliced Inverse Regression

SIR solid ground is based on two assumptions discussed in the following paragraphs: a model assumption and an assumption on the predictor space. In section 1.2.3 the algorithm is presented and in paragraph 1.2.4 a simple explanation of why SIR works is detailed. A brief discussion on the selection of the parameter k is given in paragraph 1.2.5 and then an application of SIR algorithm to a simulated dataset is shown in paragraph 1.2.6. In the last two paragraphs asymptotic results and an overview of the state of the art close the part relative to the basic knowledge of SIR.

1.2.1 Assumption on the model

The model assumption of SIR is that f depends only on k linear combinations (or projections) of the predictors:

$$Y = f(\beta_1^T \mathbf{X}, \beta_2^T \mathbf{X}, ..., \beta_k^T \mathbf{X}, \epsilon) \tag{1.4}$$

where ϵ is a random noise independent of X. The parameter k is unknown and the β_i's $\in \mathbb{R}^p$ are the directions (that identify the weights of the linear combinations) that we want to retrieve. Under this model, once we find the β_i's, the regression problem is in dimension $k \leq p$ i.e. the link function is from $\mathbb{R}^k \to \mathbb{R}$ and no longer from $\mathbb{R}^p \to \mathbb{R}$. Referring to the example in the introduction we pass from a regression problem where the predictors are in dimension 5 to a problem where the link function depends only on one linear combination (i.e. $k = 1$). If the assumption holds a dimensionality reduction is achieved not affecting the "quality of the predictors" projecting the predictor space

in lower dimension. It must be noted that since no assumptions are provided for the link function f it is not possible to directly retrieve $\beta = (\beta_1, ..., \beta_k)$ for any symmetric invertible matrix A of order k it follows that:

$$Y = f(\beta^T \mathbf{X}, \epsilon) = f(A^{-1}(\beta A)^T \mathbf{X}, \epsilon) \tag{1.5}$$

A^{-1} can be absorbed by f. Hence β_i's are not directly identifiable but they span a unique space called effective dimension reduction space e.d.r. The goal of SIR is to provide a basis of the e.d.r. space.

1.2.2 Linearity Design Condition

SIR was welcomed by the community with enthusiasm and several papers have been published to comment and think about the new idea. The main and most debated [25, 39, 41] point is the assumption that the predictors X satisfy the following, so called, Linearity Design Condition:

$$\mathbb{E}(b^T \mathbf{X} | \beta_1^T \mathbf{X}, ..., \beta_k^T \mathbf{X}) \text{ is linear in } \beta_1^T \mathbf{X}, ..., \beta_k^T \mathbf{X} \text{ for any } b \in \mathbb{R}^p \ (LDC).$$

It must be noted that this condition depends on the unobserved β_i's and therefore cannot be directly checked. If the condition holds for each $\beta_i \in \mathbb{R}^p$ then X is elliptically symmetric (e.g. Gaussian, t-Student). The (LDC) is the crucial assumption of SIR, an encouraging and very interesting results from [38] shows, under mild assumptions, that if the dimension p tends to infinity the measure of the set of standardized directions β that violate (LDC) tends to zero. This result is closely related to a previous study [28] where the authors show that for most high dimensional datasets almost all low dimensional projections are nearly Gaussian. This is a very important result that should be considered when exploring low dimensional projections since a standard approach, once n vectors in $\mathbb{R}^p$ are given, is to explore low dimensional projections to infer something on the shape of the data. This result tells that low dimensional projections are misleading since despite the distribution of the points the projections will tend to be Gaussian (when $n, p \to \infty$). When the dimension p is intermediate and no clues can be drawn from the asymptotic results a different strategy can be employed: starting from the original predictors $\mathbf{X}$ produce a new set $\tilde{\mathbf{X}}$ which is "close" to $\mathbf{X}$ and is elliptically symmetric. A resampling strategy has been proposed in [11] where a Gaussian distribution of sharing the same mean and variance as the original dataset is generated and then used to select points of the original dataset lying close to the points following the Gaussian distribution. A

slightly more general approach is proposed in [24] where a non zero weight is assigned to points that are lying close to an elliptic distribution, a fraction of points, selected by the user, far from ellipticity is then removed. The (LDC) condition is weaker than elliptic symmetry, there are cases in which the distribution is not elliptic but the condition holds nonetheless since it must be verified only for the unknown k vectors β_i. Under model (1.4) and the (LDC) the following result is stated in [47]:

Theorem 1.1. *The centered inverse regression curve* $\mathbb{E}(\mathbf{X}|Y) - \mathbb{E}(\mathbf{X})$ *is contained in the linear subspace spanned by the* $\Sigma\beta_i$*'s, where* Σ *is the covariance matrix of* $\mathbf{X}$*.*

Proof. We want to show that for each $b \in \mathbb{R}^p$ in the orthogonal complement of $\text{Span}(\Sigma\beta_1, ..., \Sigma\beta_k)$ $b^T\mathbb{E}(\mathbf{X}|Y) = 0$. Remark that

$$b^T\mathbb{E}(\mathbf{X}|Y) = b^T\mathbb{E}(\mathbb{E}(\mathbf{X}|Y, \beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X})|Y) \tag{1.6}$$

$$= \mathbb{E}(\mathbb{E}(b^T\mathbf{X}|\beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X})|Y). \tag{1.7}$$

If $\mathbb{E}(b^T\mathbf{X}|\beta_1^T\mathbf{X}, ..., \beta_k^T) = 0$ the result holds. This is indeed the case, since it is possible to show alternatively that $\mathbb{E}(\mathbb{E}(b^T\mathbf{X}|\beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X})^2) = 0$.

$$\begin{aligned}
\mathbb{E}(\ \ \mathbb{E}(b^T\mathbf{X}|\beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X})^2\ \ ) &= \mathbb{E}(\ \ \mathbb{E}(b^T\mathbf{X}|\beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X})\mathbb{E}(b^T\mathbf{X}|\beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X})\ \ ) \\
&= \mathbb{E}(\ \ \mathbb{E}(\ \ b^T\mathbf{X}\ \ \mathbb{E}(b^T\mathbf{X}|\beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X})|\beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X})\ \ ) \\
&= \mathbb{E}(\ \ b^T\mathbf{X}\ \ \mathbb{E}(b^T\mathbf{X}|\beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X})|\beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X}) \\
&= b^T\mathbb{E}(\mathbf{X}\sum_{i=1}^{k} c_i\beta_i^T\mathbf{X}) \\
&= b^T\mathbb{E}(\mathbf{X}(\sum_{i=1}^{k} c_i\beta_i^T\mathbf{X})^T) \\
&= b^T\sum_{i=1}^{k} c_i\mathbb{E}(\mathbf{X}\mathbf{X}^T)\beta_i \\
&= b^T\sum_{i=1}^{k} c_i\Sigma\beta_i \\
&= 0,
\end{aligned}$$

under the hypothesis $b^T\Sigma\beta_i = 0$ for each $i = 1, ..., k$ this concludes the proof. ■

It is interesting to underline that the first equality of the proof is always true for the, so called, tower property, and the fact that the sigma algebra $\sigma(Y) \subseteq \sigma(Y, \beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X})$. From this result in [47] follows that the covariance matrix $\Sigma^{-1}cov(\mathbb{E}(\mathbf{X}|Y)$ is degenerated

in any direction orthogonal to the $\beta_i' s$. Consequently the eigenvectors associated to the highest eigenvalues of $\Sigma^{-1} cov(\mathbb{E}(\mathbf{X}|Y))$ form a basis of the e.d.r. space. This sets the path to develop an algorithm.

1.2.3 SIR algorithm

In order to provide a basis of the e.d.r. space the estimation of $\Gamma = cov(\mathbb{E}(\mathbf{X}|Y))$ is needed, to this end it is useful to observe that is possible to apply a monotone transformation $T : \mathbb{R} \to \mathbb{R}$ to (1.4) obtaining:

$$T(Y) = T(f(\beta_1^T \mathbf{X}, \beta_2^T \mathbf{X}, ..., \beta_k^T \mathbf{X}, \epsilon)). \tag{1.8}$$

It is straightforward to see that the centered regression curve is nonetheless contained in the space spanned by the $\Sigma\beta_i$'s since the transformation can be absorbed in the function f, since there are no assumption on f in model (1.4). Using this idea Li proposed to slice Y in h-constant slices $s_1, ..., s_h$ to give a crude estimate of the centered inverse regression curve. Consequently in each slice

$$m_i = \mathbb{E}(\mathbf{X}|Y \in s_i) \tag{1.9}$$

and therefore it is possible to define Γ as:

$$\Gamma = \sum_{j=1}^{h} p_j (m_j - \mu)(m_j - \mu)^T \tag{1.10}$$

where $p_j = P(Y \in s_j)$ and $\mu = \mathbb{E}(X)$. A principal component analysis is then applied to Γ to extract the eigenvectors related to the k highest eigenvalues that for (1.1) are spanning the e.d.r space.
Given the response variable $Y = \{y_1, ..., y_n\}$ and the predictors $\mathbf{X} = \{x_1, ..., x_n\}$ the algorithm proceed as follows:

(i) Divide Y in h slices and compute $\hat{p}_j = \dfrac{1}{n} \sum_{i=1}^{n} \mathbb{I}[y_i \in s_j]$, where $\mathbb{I}[\cdot]$ is the indicator function

(ii) Compute $\hat{m}_j = \dfrac{1}{n\hat{p}_j} \sum_{i=1}^{n} x_i \mathbb{I}[y_i \in s_j]$

(iii) Obtain the sample covariance matrix:

$$\hat{\Gamma} = \sum_{j=1}^{h} \hat{p}_j (\hat{m}_j - \hat{\mu})(\hat{m}_j - \hat{\mu})^T \tag{1.11}$$

where $\hat{\mu}$ denotes the sample mean of $\mathbf{X}$. Find the eigenvectors $\hat{\beta}_1, ..., \hat{\beta}_k$ corresponding to the highest eigenvalues.

Since the matrix $\Sigma^{-1}\Gamma$ is degenerated in any direction orthogonal to the β_i's the $p-k$ last eigenvalues are null. In practice, it is rare to find zero values and, similarly to PCA, the highest values are retained. When the covariance matrix is the identity the eigenvalues represent the amount of variance of the inverse regression curve retained. A more detailed description on the selection of k is given in paragraph 1.2.5.

Comments on the number of slices h: The number of slices h must be given by the user, to avoid artificial dimension reduction h must be greater than k. A graphical tool for the selection of h is presented in [54] where is shown that SIR has low sensitivity to the choice of h, indeed for $k < h \leq n/2$ the estimated e.d.r. directions converge, in probability, at $1/\sqrt{n}$ rate to the true directions.

1.2.4 Intuition behind SIR

Let us assume the following model for $\mathbf{X} \in \mathbb{R}^2$:

$$Y = g(\beta^T \mathbf{X}) = g(b_1 x_1 + b_2 x_2) \tag{1.12}$$

where $\beta = (b_1, b_2)$, the link function depends on one linear combination ($k = 1$) of the predictors $\mathbf{X} = (x_1, x_2)$. Given the dependence on $\mathbf{X}$, it follows that when $b_1 x_1 + b_2 x_2$ remains constant the value of Y does not change. In other words the contour lines of function g are perpendicular to the direction β. It must be noted that this fact does not depend on the function g which is unknown, in figure 1.2 it is evident how the direction of the contour lines do not change despite the difference in function g. Setting $\beta = (1,1)$, slicing the range of Y allows to give a crude estimate of the inverse regression curve. In figure 1.3 the blue points are the values of the curve in each of the four considered slices. In this example X follows a standard normal distribution and is straightforward to see that the points tend to be distributed symmetrically with respect to the direction $\beta = (1,1)$. Therefore the mean in each slice approximately lies on the unknown direction β goal of the analysis. Given the blue points a Principal Component Analysis of the matrix $\Gamma = Cov(\mathbb{E}(\mathbf{X}|Y))$ is conducted to find the direction that maximizes the variance, which is an estimation of β, as it is easy to see from figure 1.3. Since X follows a standard normal distribution the covariance matrix is the identity and the spherical symmetry causes the points to be distributed symmetrically with respect to the unknown direction β. In general this is not the case and the distribution should look like an ellipsoid and a correction is needed, that is exactly the role of Σ when computing PCA on $\Sigma^{-1}\Gamma$.

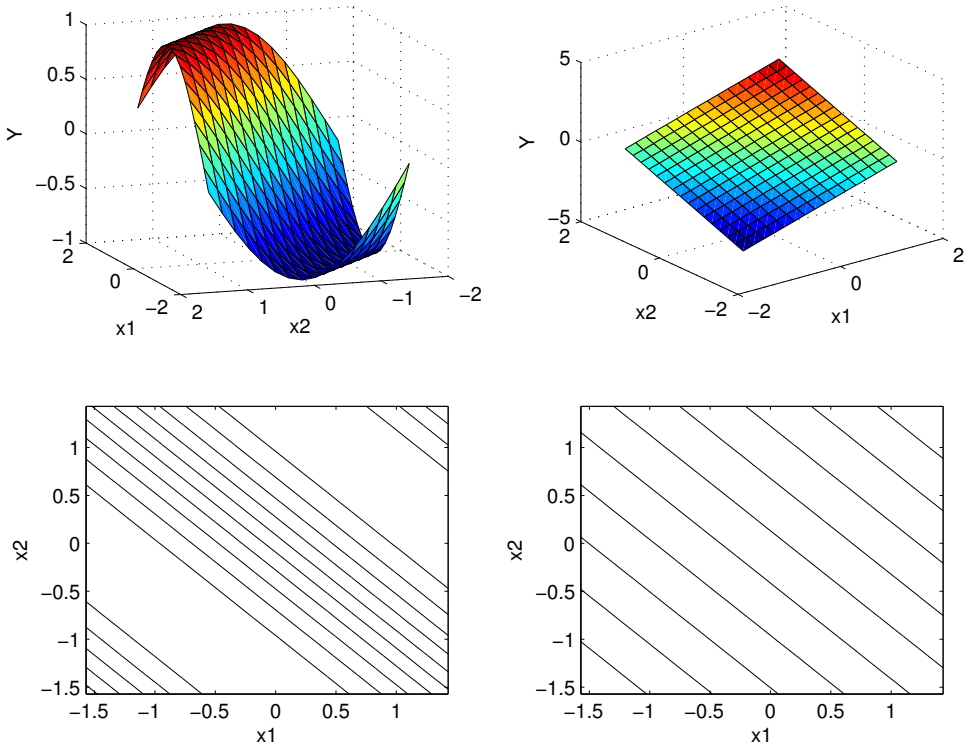


FIGURE 1.2. Upper left: The function g is the sin function. Upper right: The function g is linear. Bottom left: Contour lines of the sin function. Bottom right: Contour lines when g is linear.

1.2.5 Discussion on the unknown parameter k

The dimension k of the e.d.r space is assumed to be known when dealing with SIR, unfortunately this is not the case in real applications. Li proposed, in the pioneering paper [47], an hypothesis test on the nullity of the last $(p-k)$ eigenvalues of the matrix $\Sigma^{-1}\Gamma$. In the special case of normal distribution the $(p-k)$ smallest eigenvalues follow a χ^2 distribution with $(p-k)(h-k-1)$ degrees of freedom. This approach has been extended for elliptic distributions in [59] and [7]. A different approach is to define for each possible value of k a distance between the true e.d.r. space and the estimated one, Ferré in [31] introduced a consistent estimator for this quantity. In this direction, recently, a bootstrap approach has been implemented in [53] and refined as a useful graphical tool in [54]. Right after the publication of SIR many comments pointed out the problem of the estimation of the dimension k, as Li says in his Rejoinder to comments doubting the

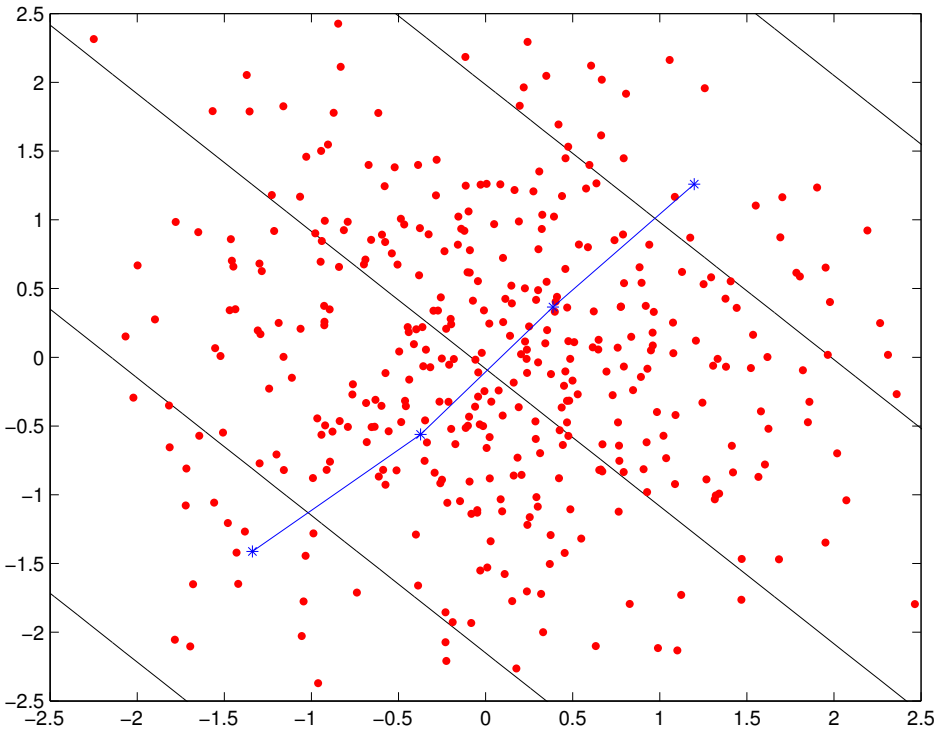


FIGURE 1.3. Slices (in black) and points of X following a standard normal distribution (red), values of $\mathbb{E}(\mathbf{X}|Y \in h_i)$, $i = 1,...,4$ (in blue).

validity of the chi-squared test: *invalidity does not demolish usefulness*. Exactly like in PCA, when it comes to deal with real data, the problem of selecting the right dimension is almost philosophical, it may be well defined once the data is gathered but the situation may change if a new sample is coming. Analyzing all the eigenvalues gives, in practice, a quick view on the problem and useful hints on where to focus the analysis and compare the results with the prior knowledge provided by the experts, which is by far, the most important reference to take into account.

1.2.6 SIR in action

In this section an application of SIR is shown and discussed. Let us assume the following model:

$$Y = (5 + x_1 + x_2 + x_3)^2 + \varepsilon, \tag{1.13}$$

where $\mathbf{X} \in \mathbb{R}^5$ follows a standard multivariate normal distribution and the error $\varepsilon \sim \mathcal{N}(0,1)$. A dataset of $n = 500$ samples, $\{Y_i, \mathbf{X}_i\}_{i=1,..,n}$, is generated following model (1.13). Y depends only on one linear combination of the predictors $\beta^T \mathbf{X} = x_1 + x_2 + x_3$, $\beta = (1,1,1,0,0)$. The application of SIR (number of slices set to $h = 10$) shows, first of all, that the eigenvalues in figure 1.4 are pointing to a one-dimensional e.d.r. space: one linear combination carries all the information needed to regress Y. This is indeed what was expected, the quadratic trend is shown in figure 1.5 where the predictors are projected along the first direction found by SIR. It is interesting to look at the scatterplot of the second direction found by SIR, there is clearly no trend possible to guess from the scatterplot, this strategy is widely used to check if residual information is contained in the directions with small eigenvalues. In this example SIR reduces the dimension from 5 to 1 with no loss of information.

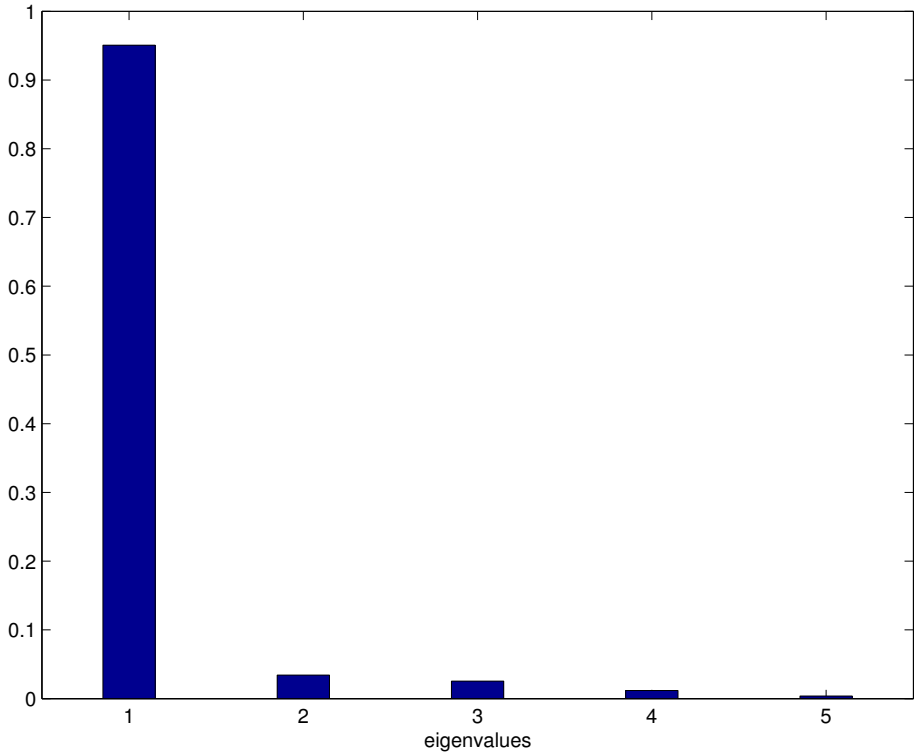
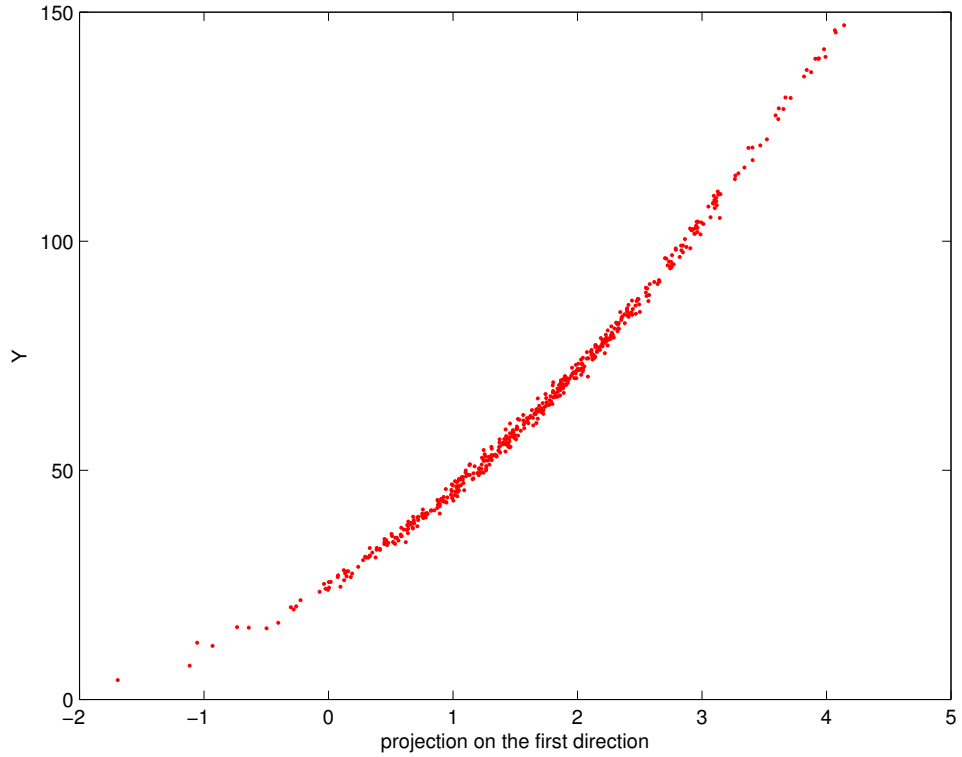


FIGURE 1.4. Bar plot of the five eigenvalues found by SIR. It is evident that all the information is packed in the first eigenvalue and therefore in the linear combination defined by the corresponding eigenvector.

(a)



(b)

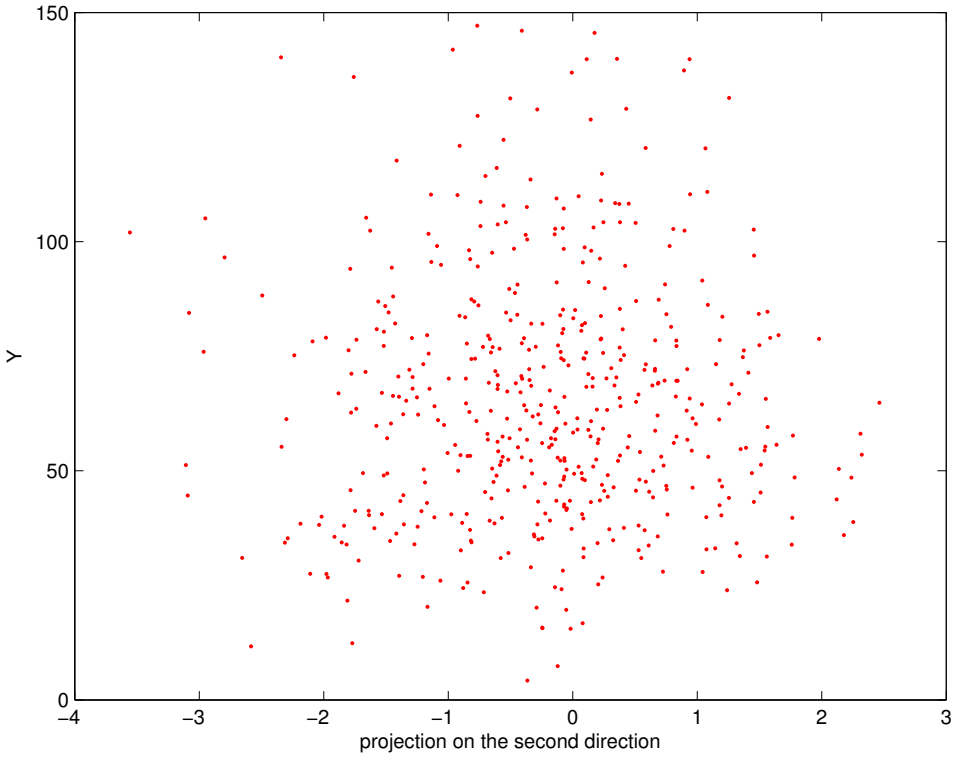


FIGURE 1.5. (a). Scatterplot of Y against the first variate found by SIR. (b). Scatterplot of Y against the second variate found by SIR.

1.2.7 Asymptotic results

The outcome of SIR is root n consistent [47] even when the range of each slice varies to balance the number of observations in each slice. Useful considerations are drawn in [77] where using the law of total covariance, $\mathbb{E}(cov(\mathbf{X}|Y)) = cov(\mathbf{X}) - cov(\mathbb{E}\mathbf{X}|Y)$ the asymptotic behavior is derived when the number of points in each slice is fixed. The asymptotic normality, with zero mean, of the estimator $\hat{\Gamma}$ is achieved. From the expression of the covariance matrix is evident how a large number of samples in each slice helps to control the asymptotic variance. The asymptotic theory of SIR is discussed in [58] where the asymptotic normality is shown for the matrix of interest $\hat{\Sigma}^{-1}\hat{\Gamma}$, for the projector on the e.d.r. space and the e.d.r. directions. The convergence in law is to a normal distribution with zero mean and the expression of the covariance matrices of the three quantity of interest is given in explicit form. Let us assume that:

(A1) $\{(y_i, x_i), i = 1, ..., n\}$ is a sample of independent observations from model (1.4).

(A2) The support of Y is partitioned into h fixed slices $s_1, s_2, ..., s_h$ such that $p_j \neq 0$ for each $j = 1, ..., h$.

(A3) The covariance matrix Σ is positive definite.

(A4) The $k+1$ largest eigenvalues of $\Sigma^{-1}\Gamma$ are non-null and satisfy:

$$\lambda_1 > \lambda_2 > ... > \lambda_{k+1},\ k+1 \leq p. \tag{1.14}$$

The asymptotic behavior of SIR is described by the following theorems, stated in [58]:

Theorem 1.2. *Under assumptions (A1), (A2) and (A3):*

$$\sqrt{n}(\hat{\Sigma}^{-1}\hat{\Gamma} - \Sigma^{-1}\Gamma) \rightarrow_d \Phi, \tag{1.15}$$

where Φ is such that its vectorization, $vec(\Phi)$, is normally distributed with mean zero and covariance matrix given in [58].

Theorem 1.3. *Under assumptions (A1), (A2), (A3) and (A4):*

$$\sqrt{n}(\hat{P} - P) \rightarrow_d \Phi_P, \tag{1.16}$$

where Φ_P is such that $vec(\Phi_P)$ is normally distributed with mean zero and covariance matrix given in [58]. P is the Σ-orthogonal eigen-projector on the e.d.r. space, $P = \sum_{i=1}^{k} P_{\lambda_i}$ and $P_{\lambda_i} = \beta_i \beta_i^T \Sigma$. Similarly $\hat{P}$ is defined using the sample version of Σ and β_i's.

Theorem 1.4. *Under assumptions (A1), (A2), (A3) and (A4):*

$$\sqrt{n}(\hat{\beta}_j - \beta_j) \rightarrow_d \Phi_{\beta_j} \forall j = 1, ..., k, \tag{1.17}$$

where Φ_{β_j} has the normal distribution with mean zero and covariance matrix given in [58].

Theorem 1.5. *Under assumptions (A1), (A2), (A3) and (A4):*

$$\sqrt{n}(\hat{\lambda}_j - \lambda_j) \rightarrow_d \Phi_{\lambda_j} \forall j = 1, ..., k, \tag{1.18}$$

where Φ_{λ_j} has the normal distribution with mean zero and covariance matrix given in [58].

All the theorems are stated in [58] for SIR_α, SIR is obtained setting $\alpha = 0$.

1.2.8 SIR skyline

First reaction of the statistical Community. The original paper of SIR is cited over 1000 times, after his publication has gained increasing attention [25, 39, 41] and many started to think more about the inverse regression curve and its applications. The focus on the paper evidenced the strengths and weaknesses of the method contributing to a better understanding of the original idea. Asymptotic theory has been discussed in [47, 58, 77] where the normality of the estimators has been well established. Despite its solid foundations SIR is not a well known and popular tool: *Can SIR be as popular as multiple linear regression?*, is asked in [14]. The impossibility of SIR to deal with functions symmetric to vertical lines in $(Y, \beta^T \mathbf{X})$ has led to the development of second moment based methods like SAVE [25, 27], SIR-II and SIR_α (Rejoinder [47]). Finite sample properties are investigated in [3] and a bagging version to face small sample size is given in [43]. Particular attention has been given to the (LDC) which is the basic assumption of SIR. Starting from its consequences R.D. Cook proposed the idea of the central subspace [19] and different alternatives to SIR ([21, 22]) based on maximum likelihood approach to dimension reduction. The (LDC) holds if $\mathbf{X}$ is elliptically distributed, given a non elliptical $\mathbf{X}$ is nonetheless possible to move the initial points to get closer to an elliptical distribution: normal resampling [11] or re-weighting [24]. When $\mathbf{X}$ is a mixture of elliptical distributions ellipticity is not global but in each component, locally, holds true. The idea to clusterize the predictor space to force the (LDC) locally [44, 50] is the first step of Collaborative SIR, our first contribution described in Chapter 2. The case

of stratified population encoded in a categorical variable is treated in [13] and recently a different approach is taken in [68]. A solution under the assumption of a Gaussian mixture model using EM is discussed in [62] with application on classification.

Regularizations and robustified version of SIR. When the dimension p increases and $n \leq p$ the covariance matrix Σ becomes singular and its inverse, used in the PCA step (see subsection 1.2.3), brings numerical instabilities, a different page of SIR literature tries to overcome this limitation. Different versions of SIR have been developed to overcome this isssue: starting from the use of PCA on the predictors space before conducting the analysis ([16, 51]) to different approaches using ridge regression ([73]) or regularized discriminant analysis ([60, 61]). Recently in [10] a link has been established among these methods, under the Gaussian assumption of the predictors, and an application to Mars hyperspectral data is detailed in [9]. An interesting application to classification combining SIR and SAVE with a shrinkage is described in [60, 61]. Not extensively studied is the outlier sensitivity of SIR, PCA based methods are well known to be non robust to outliers and this applies with no exception to SIR [37, 64]. To downweight this sensitivity, robust versions of SIR have been proposed, mainly starting from the standard model free estimators and trying to make them more resistant to outliers. Typically, in [36] classical estimators are replaced by high breakdown robust estimators and, recently in [30] two approaches are built: a weighted version of SIR and a solution based on the intra slice multivariate median estimator. The second contribution, Student SIR, wants to enrich this corner of the literature using an approach derived from the formulation of SIR given by Cook in [19]. The idea is to introduce a noise modeled by a multivariate t-student to robustify SIR and overcome, at the same time, the limitation arising by the non elliptical distributed predictors and the (LDC) [34, 35, 45].

Beyond the slices. The slicing step produces a crude estimate of the centered inverse regression curve, different strategies to estimate this curve have been developed in the literature. In case of small sample size different slicing strategies may lead to different results. To overcome the sensitivity to a specific choice of slicing a combination of slicing has been proposed in [3] and its asymptotic properties derived. Furthermore a kernel approach can be used to give an estimation of the centered inverse regression curve, in [76] a family of estimators is presented and its convergence in distribution achieved. The main theorem shows that the asymptotic variance does not depend on the choice of

the kernel function. This observation supports the low sensitivity of SIR to the number of slices. $\sqrt{n}$-consistency of the eigenvalues and corresponding eigenvectors is shown. Based on this results an extension using splines is proposed in [75], $\sqrt{n}$-consistency is shown using perturbation theory as in [76].

Kernel SIR. The e.d.r. space is, by definition, a linear space and a reasonable question to ask is if there is a way to extend this approach to find non-linear e.d.r directions. In [69] the *kernel trick* consists in defining a similarity function $K(\cdot,\cdot)$ that can be represented through an inner product, $<\cdot,\cdot>_{\mathbf{H}}$, in a higher dimension space via an unknown map $\Phi : \mathbf{X} \to \mathbf{H}$, where $\mathbf{H}$ is a reproducing kernel Hilbert space:

$$K(x,y) = <\Phi(x), \Phi(y)>_{\mathbf{H}} . \tag{1.19}$$

Linear functions in this Hilbert space are non-linear in the original predictor space $\mathbf{X}$. Using the new predictors $\Phi(X)$ the SIR strategy can be applied under the model:

$$Y = f(<\beta_1, \Phi(X)>_{\mathbf{H}}, ..., <\beta_k, \Phi(X)>_{\mathbf{H}}). \tag{1.20}$$

Unfortunately the map function is unknown and furthermore the high (or infinite) dimension of the predictor space makes all the analysis not feasible. Using the kernel function K, it is shown in [69], that using a strategy mimicking SIR is possible to retrieve $<\beta_1, \Phi(X)>_{\mathbf{H}}, ..., <\beta_k, \Phi(X)>_{\mathbf{H}}$. Nonetheless, it is important to underline that the e.d.r directions are not available, only the projected predictors are found. In presence of new samples arriving an explicit formula is available for projection. For fast implementation, details are given in [71].

Multivariate SIR. It is natural to try to extend SIR in the multivariate case where the response $Y \in \mathbb{R}^q$, $q > 2$. A common strategy is to analyze each Y-variate separately and then merge the results to obtain a global solution using the information from each individual univariate regression model of type (1.4). A strategy in [2, 48] is to find which Y-variate is most predictable from $\mathbf{X}$ and discard the others. An extension to SIR_α can be found in [6]. On the other hand in [55] all the variates are combined using weights proportional to the eigenvalues of each independent univariate regression. When the slicing strategy is complete (i.e. a "grid" is sequentially computed considering all the Y-variates) the estimation is not feasible when the dimension q increases. K-means is used in [63] as an alternative to slicing to overcome this issue. A more general case when the e.d.r. space varies according to the Y-variate is discussed in [26] where the, much tractable, marginal slicing strategy is adopted.

Variable selection and sparsity. Nowadays with the increasing capability of technology to gather data the number of variables p considered is enormous. SIR components are a linear combination of all the original predictors and since is desirable to have a direct interpretation of the new variables, sparsity constraints can be introduced. Using the generalized eigenvalue formulation penalizations terms are introduced in [49] to obtain spare solutions. A representation of SIR as a regression-type optimization problem combining the shrinkage idea of the lasso with SIR is provided by [52]. An application to classification combining ridge and adaptive lasso can be found in [70]. Our last contribution explores a different approach not enforcing sparsity using an idea from [5] where the false discovery rate is controlled creating copies of the original predictors with some useful characteristics to discover rejectable variables.

Other approaches and new trends. An iterative version of SIR is proposed in [8] and an extension meant to deal with a data-stream providing a strategy to use the information of the previous blocks to help the analysis is given in [12]. Recently, optimal quantization has been introduced to project the predictor space on a grid and then proceed with the analysis, property of the estimators are given in [4]. A different page of SIR is the one concerning its application to functional data, two main papers, [32, 33], extended the use of SIR in this framework. Attention on the assumptions has been raised by [18]. Furthermore a robustified version of functional SIR can be found in [66]. In case of modern biomedical images Tensor-SIR has been proposed in [29], theoretical developments partially overlap with [46].

In this thesis Sliced Inverse Regression (SIR) (1991), is analyzed and extended in three different contributions. The first contribution, namely Collaborative SIR, based on an observation on the design hypothesis of SIR, is presented in chapter 2. A robustified version of SIR, Student SIR, is then developed to take into account the well known sensitivity to outliers of the techniques based on linear projections. The last contribution is based on a paper of R. F. Barber and E. Candes and tackles the problem of quantifying the false discovery rate in SIR. Conclusion and comments are finally outlined. All those contributions are summarized in the next three paragraphs.

1.3 Collaborative SIR: an overview

Collaborative SIR is our first contribution. One of the weak points of SIR is the impossibility to check if the (LDC) holds. It is known that if $\mathbf{X}$ follows an elliptic distribution the

condition holds true, in case of a mixture of elliptic distributions there are no guaranties that the condition is satisfied globally, but locally holds. Starting from this consideration an extension of the model (1.4) is proposed.

1.3.1 The model

Let X be a random vector, $X \in \mathbb{R}^p$, from a mixture model and be Z an unobserved latent random variable $Z \in \{1,...,c\}$, where c is the number of components. Given $Z = i$ we have the following model:

$$Y = f_{F(i)}(\beta_{F(i)}^T X) + \epsilon_i, \tag{1.21}$$

where Y is the random variable to predict, $Y \in \mathbb{R}$, F is an unknown deterministic function $F : \{1,...,c\} \to \{1,...,D\}$, $D \in \mathbb{N}$. The functions $f_j : \mathbb{R} \to \mathbb{R}$, $j = 1,...,D$ are unknown link functions between $\beta_j^T X$ and Y. Finally ϵ_i are random errors $\forall i \ \epsilon_i \in \mathbb{R}$, i.e. each component is allowed to have a different related error. The underlying idea is to allow the e.d.r. direction to change depending on the mixture, different components may lead to different results. If $D = 1$ then $F^{-1}(1) = \{1,...,c\}$, the e.d.r. direction and the link functions do not vary through all the mixture. This specific case is addressed in [44].

1.3.2 Collaborative SIR in practice

Given the predictor variable $\mathbf{X}$, a clustering is performed and the variable Z is estimated. In each cluster, SIR is applied independently. The result from each component collaborates to give an estimation of D. To estimate D a hierarchical merging procedure is introduced based on the proximity measure

$$m(a,b) = \cos^2(a,b) = (a^T b)^2, \tag{1.22}$$

between the estimated e.d.r. directions $\hat{b}_1,...,\hat{b}_c$. A similar procedure has been used in [26] to cluster the components of the multivariate response variable Y related to the same e.d.r. spaces. Let $V = \{v_1, v_2, ..., v_{|V|}\}$ be a set of vectors in dimension p with associated weights w_i. We define the quantity:

$$\lambda(V) = \max_{v \in \mathbb{R}^p} \frac{1}{w_V} \sum_{i=1}^{|V|} w_i m(v_i, v) \text{ s.t. } \|v\| = 1$$
$$= \text{largest eigenvalue of } \frac{1}{w_V} \sum_{i=1}^{|V|} w_i v_i v_i^T$$

where $w_V = \sum_{i=1}^{|A|} w_i$ is the normalization. Vector v maximizing $\lambda(V)$ is the most collinear vector to our set of vectors given the proximity criteria (1.22) and the weights w_i. To build the hierarchy we consider the following iterative algorithm initialized with the set $A = \{\{\hat{b}_1\}, ..., \{\hat{b}_c\}\}$:

while $card(A) \neq 1$

Let $a, b \in A$ **such that** $\lambda(a \cup b) > \lambda(c \cup d) \forall c, d \in A$
$A = (A \setminus \{a, b\}) \bigcup a \cup b$
end

The weights are set equal to the number of samples in each components, i.e. $w_i = n_i$, $i = 1, ..., c$. At each step the cardinality of the set A decreases merging the most collinear sets of directions. The bottom up greedy algorithm proceeds as follows:

- First the two most similar elements of A are merged considering all the $|A| \times (|A| - 1) = c \times (c - 1)$ pairs.
- In the following steps the two most similar sets of vectors are merged, considering all $|A| \times (|A| - 1)$ pairs in A.

An analysis of the cost function allows to give an estimation of D. Once the estimation $\hat{D}$ is available using the information encoded in the hierarchical tree each initial cluster is assigned to its group, i.e. F is estimated, and a final solution is calculated. Each node at level $\hat{D}$ corresponds to a different e.d.r. space.

1.3.3 Asymptotic results

Asymptotic results can be established similarly to [12]. We fix $j \in \{1, ..., D\}$ and consider $\{X_t,\ t \in \bigcup_{i \in F^{-1}(j)} \mathscr{C}_i\}$, where $\mathscr{C}_i = \{t \text{ such that } Z_t = i\}$, and a sample size $n^j = \sum_{i \in F^{-1}(j)} n_i$ which tends to ∞. The following three assumptions are considered:

- (A1) $\{X_t,\ t \in \bigcup_{i \in F^{-1}(j)} \mathscr{C}_i\}$ is a sample of independent observations from the single index model (1.21).
- (A2) For each i, the support of $\{Y_t, t \in \mathscr{C}_i\}$ is partitioned into a fixed number H_t of slices such that $p_i^h > 0, h = 1, \ldots, H_t$.

- (A3) For each i and $h = 1, \ldots, H_t$, $n_{h,i} \to \infty$ (and therefore $n_i \to \infty$) as $n \to \infty$.

Theorem 1.6. *Under model (1.4), linearity condition (LDC) and assumptions (A1)-(A3), we have:*

(*i*) $\widehat{\beta}_j = \beta_j + O_p(\underline{n}^{j\,-1/2})$, *where* $\underline{n}^j = \min_{i \in F^{-1}(j)} n_i$;

(*ii*) *If, in addition* $n_i = \theta_{ij} n^j$, $\theta_{ij} \in (0,1)$ *for each* $i \in F^{-1}(j)$, *then* $\sqrt{n^j}(\widehat{\beta}_j - \beta_j)$ *converges to a centered Gaussian distribution.*

1.3.4 Experimental results

Simulation studies have been established to assess the sensitivity to clustering of Collaborative SIR. Under different configurations it has been shown that Collaborative SIR performs, not surprisingly, better than SIR which is not designed to face multiple e.d.r. spaces as in model (1.21). A significative gain in the accuracy of the results is found through the analysis. Two real datasets are then discussed, in both datasets two distinct e.d.r. spaces have been found supporting the strategy addressed in the paper. In the Galaxy data in figure 1.3.4 two e.d.r. spaces are retrieved by Collaborative SIR. It is interesting to notice that an analysis of the components of the two estimated directions evidence that different variables contribute in different way to predict Y (figure 1.7). There is indeed a difference between the two groups that explains the results.The experts confirm that the subdivision is not unexpected and reflects some property of the galaxies. The same dataset has been analyzed by Student SIR and Knockoff SIR, the solution given by Collaborative SIR is, in general, supported by this comparison. On the other hand Knockoff SIR suggests that variable 6 that is found to be significative in one of the groups should not be considered. Finally Collaborative SIR analyzed only the first e.d.r. direction while our later study via Student SIR the dimension of the e.d.r space has been estimated to $k = 3$. The results of the three methods show the complexity of real data, a comparison of single solutions gives a better understanding of the general analysis.

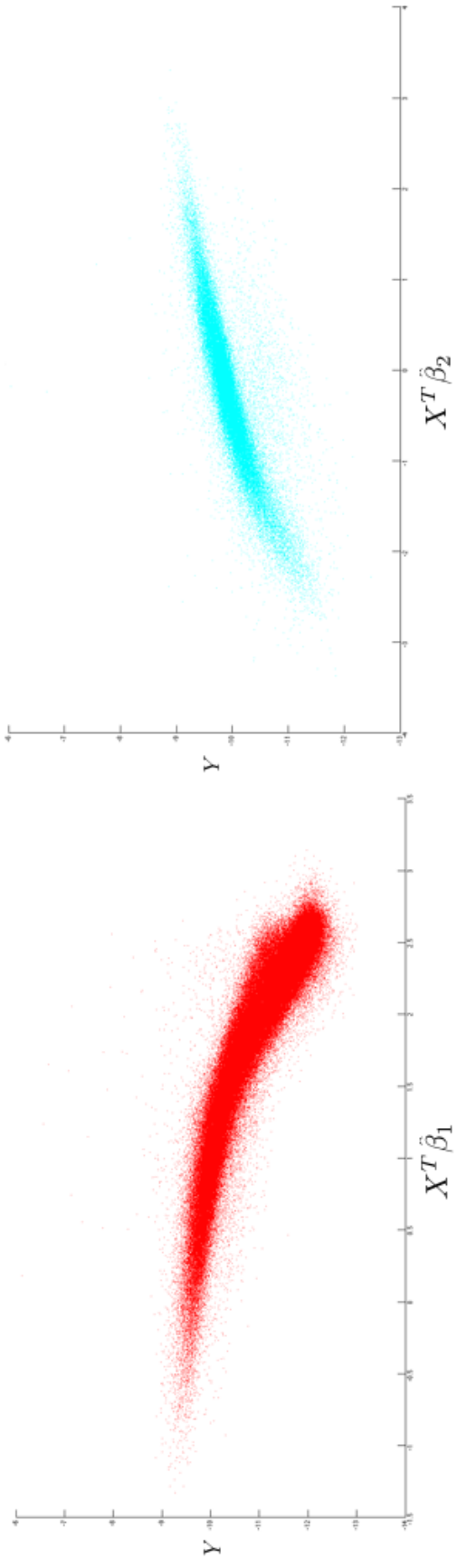


Figure 1.6: (Left) Graph of Y and the projection along the first e.d.r. direction $\hat{\beta}_1$ found by Collaborative SIR. (Right) Graph of Y and the projection along the second e.d.r. direction $\hat{\beta}_2$ found by Collaborative SIR. It is evident the nonlinear behavior of the two link functions.

Figure 1.7: Differences in e.d.r directions $\hat{\beta}_1$ and $\hat{\beta}_2$. Many elements in the vectors are close to zero resulting in a variable selection. Differences in the two lines show how different variables contribute in regressing Y. The squared cosine between the two directions is 0.42.

1.4 Student SIR: an overview

Student SIR comes from the need to robustify SIR. Since SIR is based on the estimation of the covariance, and contains a PCA step, it is indeed sensitive to noise (see [37, 64]). To extend SIR, the approach suggested by Cook in [19] has been used.

1.4.1 The model

A subspace S is a d.r.s. if Y is independent of $\mathbf{X}$ given $\mathbf{P}_S\mathbf{X}$, where $\mathbf{P}_S$ is the orthogonal projection onto S. In other words, all the information carried by the predictors $\mathbf{X}$ on Y can be compressed in $\mathbf{P}_S\mathbf{X}$. It has been shown under weak assumptions that the intersection of all d.r.s., the central subspace, is itself a d.r.s. [72]. The space found by SIR is a d.r.s. Let us assume the following model ([10, 19]):

$$\mathbf{X} = \mu + \mathbf{VBc}(Y) + \varepsilon, \tag{1.23}$$

where $\mu \in \mathbb{R}^p$ is a non random vector, $\mathbf{B}$ is a non random $p \times d$ matrix with $\mathbf{B}^T\mathbf{B} = \mathbf{I}_d$, $\varepsilon \in \mathbb{R}^p$ is assumed to be Gaussian distributed, ε is assumed independent of Y, with scale matrix $\mathbf{V}$, $\mathbf{c} : \mathbb{R} \to \mathbb{R}^d$ is a non random function. It directly follows from (1.23) that

$$\mathbb{E}(\mathbf{X}|Y = y) = \mu + \mathbf{VBc}(y), \tag{1.24}$$

and thus, after translation by μ, the conditional expectation of $\mathbf{X}$ given Y is a random vector located in the space spanned by the columns of $\mathbf{VB}$. When ε is assumed to be Gaussian distributed, Proposition 6 in [19] states that $\mathbf{B}$ is indeed a basis of the central subspace. In [10, 19], it appears then that, under appropriate conditions, the maximum likelihood estimator of $\mathbf{B}$ corresponds to (up to a full rank linear transformation) the SIR estimator of the d.r.s.
The idea is to consider a different error ε modeled by a multivariate Student distribution. Among the elliptically contoured distributions, the multivariate Student is a natural generalization of the multivariate Gaussian but its heavy tails can better accommodate outliers. Considering Student distributed errors it is shown that Proposition 6 in [19] can be generalized and the inverse regression remains tractable via an Expectation-Maximization (EM) algorithm:

Proposition 1.1. *Let $\mathbf{X}_y$ be a random variable distributed as $\mathbf{X}|Y = y$, let us assume that*

$$\mathbf{X}_y = \mu + \mathbf{VBc}(y) + \varepsilon, \tag{1.25}$$

with ε following a generalized Student distribution with certain parameters, $\mathbf{c}(y) \in \mathbb{R}^d$ *is function of y and* $\mathbf{VB}$ *is a $p \times d$ matrix of rank d. Under model (1.25), the distribution of $Y|\mathbf{X} = \mathbf{x}$ is the same as the distribution of $Y|\mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x}$ for all values* $\mathbf{x}$.

1.4.2 Expectation-Maximization algorithm

In order to estimate the model parameters the following generalization to the multivariate Student distribution is considered. Thanks to a useful representation of the t-distribution as a so-called *infinite mixture of scaled Gaussians* or *Gaussian scale mixture* [1] the EM algorithm remains tractable. A Gaussian scale mixture distribution has a probability density function of the form

$$P(\mathbf{x};\mu,\Sigma,\psi) = \int_0^\infty \mathcal{N}_p(\mathbf{x};\mu,\Sigma/u)\, f_U(u;\psi)\, \mathrm{d}u, \tag{1.26}$$

where $\mathcal{N}_p(\,.\,;\mu,\Sigma/u)$ denotes the density function of the p-dimensional Gaussian distribution with mean μ and covariance Σ/u and f_U is the probability distribution of a univariate positive variable U referred as the weight variable. When f_U is a Gamma distribution $\mathcal{G}(\nu/2,\nu/2)$[1] where ν denotes the degrees of freedom, expression (1.26) leads to the standard p-dimensional t-distribution denoted by $t_p(\mathbf{x};\mu,\Sigma,\nu)$ with parameters μ (location vector), Σ ($p \times p$ positive definite scale matrix) and ν (positive degrees of freedom parameter). Its density is given by

$$\begin{aligned} t_p(\mathbf{x};\mu,\Sigma,\nu) &= \int_0^\infty \mathcal{N}_p(\mathbf{x};\mu,\Sigma/u)\, \mathcal{G}(u;\nu/2,\nu/2)\, \mathrm{d}u \\ &= \frac{\Gamma((\nu+p)/2)}{|\Sigma|^{1/2}\,\Gamma(\nu/2)\,(\pi\nu)^{p/2}}\, [1+\delta(\mathbf{x},\mu,\Sigma)/\nu]^{-(\nu+p)/2}, \end{aligned} \tag{1.27}$$

where $\delta(\mathbf{x},\mu,\Sigma) = (\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)$ is the Mahalanobis distance between $\mathbf{x}$ and μ. If $f_U(u;\psi)$ is set equal to a Gamma distribution $\mathcal{G}(\alpha,\gamma)$ without imposing $\alpha = \gamma$, (1.26) results in a multivariate Pearson type VII distribution (see *e.g.* [40] vol.2 chap. 28) also referred to as the Arellano-Valle and Bolfarine's Generalized t distribution in [42]. This generalized version is the multivariate version of the t-distribution considered in this work, its density is given by:

$$\mathcal{S}_p(\mathbf{x};\mu,\Sigma,\alpha,\gamma) = \int_0^\infty \mathcal{N}_p(\mathbf{x};\mu,\Sigma/u)\, \mathcal{G}(u;\alpha,\gamma)\, \mathrm{d}u \tag{1.28}$$

$$= \frac{\Gamma(\alpha+p/2)}{|\Sigma|^{1/2}\,\Gamma(\alpha)\,(2\pi\gamma)^{p/2}}\, [1+\delta(\mathbf{x},\mu,\Sigma)/(2\gamma)]^{-(\alpha+p/2)}\,. \tag{1.29}$$

[1] The Gamma distribution has probability density function $\mathcal{G}(u;\alpha,\gamma) = u^{\alpha-1}\Gamma(\alpha)^{-1}\exp(-\gamma u)\gamma^\alpha$ where Γ denotes the Gamma function.

For a random variable $\mathbf{X}$ following distribution (1.29), an equivalent representation useful for simulation is $\mathbf{X} = \mu + U^{-1/2}\tilde{\mathbf{X}}$ where U follows a $\mathcal{G}(\alpha,\gamma)$ distribution and $\tilde{\mathbf{X}}$ follows a $\mathcal{N}(0,\Sigma)$ distribution.

Remark 1.1 (Identifiability)**.** *The expression* (1.29) *depends on* γ *and* Σ *only through the product* $\gamma\Sigma$ *which means that to make the parameterization unique, an additional constraint is required. One possibility is to impose that* Σ *is of determinant 1. It is easy to see that this is equivalent to have an unconstrained* Σ *with* $\gamma = 1$.

Unconstrained parameters are easier to deal with in inference algorithms. Therefore, it is assumed without loss of generality that $\gamma = 1$. The Arellano-Valle and Bolfarine's Generalized t distribution has the property that marginal and conditional distributions remain in the Generalized Student family. This property is used to estimate the parameters $\{\mu, \mathbf{V}, \mathbf{B}, \alpha, c\}$, the function c is expanded as a linear combination of known basis function and the coefficients matrix $\mathbf{C}$ estimated. The E-step of the algorithm provides an estimation of quantities related to the weight U, then in the M-step $\{\mu, \mathbf{V}, \mathbf{B}, \alpha, \mathbf{C}\}$ are calculated. The initialization of the weights is such that the first iteration of the algorithm corresponds to the standard SIR estimators.

1.4.3 Simulation results

During the simulations different models have been analyzed and Student SIR has been compared to standard SIR and four other approaches. Contour Projection (CP-SIR) [56, 67] applies the SIR procedure on a rescaled version of the predictors. Weighted Canonical Correlation (WCAN) [74] uses a basis of B-splines first estimating the dimension d of the central subspace and then the directions from the nonzero robustified version of the correlation matrices between the predictors and the B-splines basis functions. The idea of Weighted Inverse Regression (WIRE) [30] is to use a different weight function capable of dealing with both outliers and inliers. SIR is a particular case of WIRE with constant weighting function. Slice Inverse Median Estimation (SIME) replaces the intra slice mean estimator with the median which is well known to be more robust. All values referring to CP-SIR, WCAN, WIRE, SIME in the tables are directly extracted from [30].

Three different regression models (**I**, **III**, **III**) and three different distributions of $\mathbf{X}$ (**i**, **ii**, **iii**) are considered. Models **I**, **III** are homoscedastic while model **II** is heteroscedastic. Case (**ii**) is built to test the sensitivity to outliers while the distribution of $\mathbf{X}$ is elliptical. In (**iii**) a non-elliptical distribution of $\mathbf{X}$ is considered. The dimension is set to $p = 10$,

the dimension of the e.d.r. space is $d = 1$ for $\mathbf{I}, \mathbf{II}$ and $d = 2$ for $\mathbf{III}$. The nine different configurations of $\mathbf{X}$ and Y are simulated with a number of samples varying depending on the experiment. In all tables Student SIR is compared Values relative to SIR have been recomputed using [23]. To compare the methods the following proximity criteria has been adopted:

$$r(\mathbf{B}, \hat{\mathbf{B}}) = \frac{trace(\mathbf{B}\mathbf{B}^T\hat{\mathbf{B}}\hat{\mathbf{B}}^T)}{d}. \tag{1.30}$$

The above quantity r ranges from 0 to 1 and evaluates the distance between the subspaces spanned by the columns of $\mathbf{B}$ and $\hat{\mathbf{B}}$. If $d = 1$, r is the squared cosine (1.22) between the two spanning vectors. Values close to one show a good performance of the algorithms. In table 1.1 (a) Student SIR shows its capability to deal with different configurations. The proximity criterion (1.30) in Table 1.1 (a) is very close to one, for the first two regression models independently of the distributions of the predictors. In the Gaussian case, Student SIR and SIR are performing equally well showing that our approach has no undesirable effects when dealing with *simple* cases. For configuration $\mathbf{III} - (\mathbf{iii})$, a slightly different value has been found for SIR compared to [30]. In this configuration however the trend is clear: standard SIR, Student SIR, WIRE and SIME show similar performance. In contrast, configurations $\mathbf{I} - (\mathbf{ii}), \mathbf{II} - (\mathbf{ii}), \mathbf{III} - (\mathbf{ii})$ illustrate that Student SIR can significantly outperform SIR. This is not surprising since the standard multivariate Cauchy has heavy tails and SIR is sensitive to outliers Table 1.1 (b) illustrates on model $\mathbf{I}$ the effect of the sample size n: Student SIR exhibits the best performance among all methods. It is interesting to observe that, in case $(\mathbf{ii})$, the smaller value of r for standard SIR does not depend on the sample size n. In contrast, adding observations results in a better estimation for Student SIR.

A test on real data in high dimension has been performed to compare SIR and Student SIR. The Galaxy dataset corresponds to $n = 362,887$ different galaxies. This dataset has been already used in [15] with a preprocessing based on expert supervision to remove outliers. In this study all the original observations are considered, removing only points with missing values, which requires no expertise. The response variable Y is the stellar formation rate. The predictor $\mathbf{X}$ is made of spectral characteristics of the galaxies and is of dimension $p = 46$. The results show that when small sample sizes are considered Student SIR is more reliable being robust to eventual outliers. The BIC estimated the dimension of the e.d.r. space to $k = 3$, unfortunately in our previous study we analyzed only the first component of Collaborative SIR, an interesting parallel can be done with

our last contribution Knockoff SIR (paragraph 1.5.2) that is supporting the decision of Student SIR pointing out that only three directions are informative.

Model	**X**	Method					
		SIR	CP-SIR	WCAN	WIRE	SIME	st-SIR
I	(i)	**.99(.01)**	.99(.01)	.98(.01)	.98(.01)	**.99(.01)**	**.99(.01)**
	(ii)	.63(.18)	.92(.04)	.88(.06)	.87(.07)	.91(.04)	**.98(.01)**
	(iii)	**.99(.01)**	.86(.12)	.72(.27)	.98(.01)	.97(.01)	**.99(.01)**
II	(i)	**.99(.01)**	.98(.01)	.98(.01)	.98(.01)	98(.01)	**.99(.01)**
	(ii)	.61(.18)	.92(.04)	.89(.06)	.87(.08)	.91(.05)	**.98(.01)**
	(iii)	**.99(.01)**	.67(.25)	.69(.28)	.98(.01)	.97(.02)	**.99(.01)**
III	(i)	.88(.06)	.87(.06)	**.89(.05)**	.86(.06)	.87(.06)	.87(.06)
	(ii)	.40(.13)	.78(.10)	.78(.11)	.76(.11)	.78(.10)	**.85(.06)**
	(iii)	.84(.07)	.63(.12)	.67(.13)	**.85(.07)**	**.85(.07)**	.84(.07)

(a)

Model	**X**	n	Method					
			SIR	CP-SIR	WCAN	WIRE	SIME	st-SIR
I	(i)	50	**.95(.03)**	.91(.09)	.86(.11)	.88(.11)	.90(.08)	**.95(.03)**
		100	**.98(.01)**	.96(.03)	.96(.03)	.95(.03)	.96(.02)	**.98(.01)**
		200	**.99(.01)**	**.99(.01)**	.98(.01)	.98 (.01)	**.99(.01)**	**.99(.01)**
		400	**1(.00)**	.99(.00)	.99(.00)	.99 (.01)	.99(.00)	**1(.00)**
	(ii)	50	.60(.22)	.66(.18)	.57(.23)	.49(.24)	.59(.21)	**.90(.07)**
		100	.62(.21)	.85 (.08)	.78(.11)	.73(.15)	.81(.10)	**.96(.02)**
		200	.62(.20)	.92(.04)	.88(.06)	.87(.07)	.91(.04)	**.98(.01)**
		400	.62(.18)	.96(.02)	.94(.03)	.93(.03)	.96(.02)	**.99(.00)**
	(iii)	50	**.95(.02)**	.45(.29)	.18(.19)	.73(.25)	.86(.09)	**.95(.02)**
		100	**.98(.01)**	.66(.25)	.35(.29)	.94(.04)	.94(.04)	**.98(.01)**
		200	**.99(.01)**	.86(.12)	.72(.27)	.98(.01)	.97(.01)	**.99(.00)**
		400	**.99(.00)**	.96(.04)	.96(.04)	.93(.03)	**.99(.01)**	**.99(.00)**

(b)

Table 1.1: **(a)** Average of the proximity measure r (eq. (1.30)) for sample size $n = 200$; and **(b)** effect of sample size n on the average proximity measure r, both over 200 repetitions with standard deviation in brackets. Six methods are compared. SIR: sliced inverse regression; CP-SIR: contour projection for SIR; WCAN: weighted canonical correlation; WIRE: weighted sliced inverse regression estimation; SIME: sliced inverse multivariate median estimation and st-SIR: Student SIR. In all cases, the number of slices is $h = 5$ and the predictor dimension $p = 10$. Best r values are in bold.

1.5 Knockoff SIR: an overview

Knockoff SIR is an extension of SIR to perform variable selection and give sparse solution that has its foundations in a recently published paper [5] that focuses on the false discovery rate in the regression framework.

1.5.1 The idea

The underlying idea of [5] is to construct copies of the original variables that have some properties. From the comparison between the true and the false variables informations can be used to decide weather the variable is active or not in the specific regression framework. Let us assume to have a $\mathbf{X} \in \mathbb{R}^{p\times n}$ dataset and to construct such copies $\tilde{\mathbf{X}} \in \mathbb{R}^{p\times n}$ the following theorem establishes the behavior of SIR on the concatenation $[\mathbf{X},\tilde{\mathbf{X}}] \in \mathbb{R}^{2p\times n}$:

Theorem 1.7. *Given the predictors $\mathbf{X} = \{x_1,...,x_n\} \in \mathbb{R}^{p\times n}$ and a response variable $Y = \{y_1,...,y_n\} \in \mathbb{R}^{n\times 1}$ let us denote by $\hat{\mathbf{B}}$ the SIR estimator of $\mathbf{B} \in \mathbb{R}^{pxk}$ in the following regression model:*

$$Y = f(\mathbf{XB},\epsilon) \tag{1.31}$$

where f is an unknown link function and ϵ is a random error independent of $\mathbf{X}$. The k-columns of $\mathbf{B}$ span the e.d.r. space [47]. When $n > 2p$ let us consider a knockoff filter $\tilde{\mathbf{X}} \in \mathbb{R}^{p\times n}$ of the form:

$$\tilde{\mathbf{X}} = A^T\mathbf{X} + (\tilde{U}C)^T, \tag{1.32}$$

defined in section 4.2, and the concatenation $[\mathbf{X},\tilde{\mathbf{X}}] \in \mathbb{R}^{2p\times n}$. The SIR estimator $\tilde{\mathbf{B}} \in \mathbb{R}^{2pxk}$ for the concatenation $[\mathbf{X},\tilde{\mathbf{X}}]$ has each column $\tilde{\mathbf{B}}_j$ of the form:

$$\tilde{\mathbf{B}}_j = [\hat{\mathbf{B}}_j, 0] \tag{1.33}$$

where 0 is a p-dimensional vector of all zeros, and $\hat{\mathbf{B}}_j$ is SIR estimation computed on $\mathbf{X}$

This theorem establishes that the true variables are preferred by SIR when the concatenation is analyzed. When a variable has a non zero weight in a linear combination found by SIR its value can be compared to the one of the corresponding copy variable. If they differ it means that the true variable is indeed active, if the result shows that the weight is similar it means that the algorithm cannot distinguish the true from the copy suggesting that the weight found is due to numerical instabilities of the solution.

1.5.2 Simulation results

Extensive analysis have been made both on simulated and real data. On simulated data two regression models are considered with different dimension of the e.d.r. space. Results from the first test case refers to the following regression model:

$$Y = (x_1 + x_2 + x_3 - 10)^2 + \varepsilon, \tag{1.34}$$

where $\mathbf{X} = (x_1, ..., x_{10}) \in \mathbb{R}^{10}$ is a vector of independent standard normal distributions and ε is a standard normal error independent of $\mathbf{X}$. To asses the quality of the estimations two indexes are considered: True Inclusion Rate (TIR), the ratio of the number of correctly identified active predictors to the number of truly active predictors; and the False Inclusion Rate (FIR), the ratio of the number falsely identified active predictors to the total number of inactive predictors. In our test there are 3 active predictors and 7 inactive. A study on the sensitivity to the number of sample n is shown in Table 1.5.2.

n	TIR	FIR	#-slices
25	.81(.25)	.48(.20)	2
50	1(.0)	.16(.16)	5
75	1(.0)	.09(.12)	7
100	1(.0)	.08(.10)	10
150	1(.0)	.08(.11)	15
200	1(.0)	.06(.11)	20
250	1(.0)	.05(.08)	25
300	1(.0)	.04(.08)	30
400	1(.0)	.04(.06)	30

TABLE 1.2. Study on the sensitivity to the number of sample n, averages (and standard deviation in brackets) are obtained over 100 iterations. True Inclusion Rate (TIR) and False Inclusion Rate (FIR) are shown. The number of slices has been selected such that at least 10 samples are contained in each slice.

The quality of the estimation is good and the standard deviation decreases with the increasing number of samples. An application to the Galaxy dataset supports results that have been already obtained by Collaborative SIR and Student SIR. The predictor space is made of spectral characteristics of the galaxies and is of dimension $p = 46$ with $n = 362,887$ points. In the first e.d.r. directions the only active variables found are $\{2,3,23,40,45\}$ this is exactly matching the results of Collaborative SIR, the variable 6 selected by Collaborative SIR is estimated inactive in all e.d.r directions, doubts can

be cast to the selection of variable 6 for the analysis. According to the result of BIC in Student SIR we tested the e.d.r. directions relative to the five highest eigenvalues obtaining that for the first three e.d.r. directions active variables have been found. Grouping the active variables through the first three directions gives only seven variables: $\{2,3,23,40,42,43,45\}$. This means that by default the analysis could be directly performed on the seven predictors avoiding the other 39. The fourth and fifth and further directions with smaller eigenvalues resulted with no active variables supporting the decision of Student SIR.

CHAPTER 2

COLLABORATIVE SIR

How can we live without our lives?
How will we know it's us without our past?
J. Steinbeck.

Collaborative SIR has been accepted for publication in *Communications in Statistics-Theory and Methods.*

2.1 Overall Idea

To give an intuitive idea of what Collaborative SIR is meant for, the example of the bombers in the Introduction will be considered. Suppose that $\mathbf{X} = (x_1, x_2, ..., x_5)$ is the area hit by bullets for different aircrafts in five corresponding continuos variables (as in Wald's paper each aircraft is divided in five areas). Our goal is to predict $Y \in [0,1]$, the damage of the bomber, 0 is undamaged and 1 is downed. After a campaign of different days and actions the data is gathered in X. It is reasonable to assume that different missions required different classes of aircrafts, each class of aircraft with his own characteristics and therefore vulnerability (list of US bombers in WW II in figure 3.1). If the information about the class of each fleet is available one could apply SIR independently in each group, in case this information is not observable, a clustering can be used to obtain a reasonable partition. Collaborative SIR uses the information from each cluster and, taking into account the sample size in each group, merges the groups with *similar* direction β (in the example this corresponds to similar estimated vulnerability characteristics). Solutions

in different clusters collaborate to form the final outcome, the reliability of each solution is function of the sample size.

FIGURE 2.1. U.S. aircrafts during WW2. A story within the history is the one of unofficial plane spotters. Unfortunately, even children were asked to support war and served as unofficial auxiliary of the Army Air Forces Ground Observer Corps (aka GOC). Coca-Cola offered a popular manual called *Know Your Planes* for only ten cents. Even card games served at this scope.

Collaborative Sliced Inverse Regression

Alessandro Chiancone[a,b,c], Stephane Girard[a], Jocelyn Chanussot[b]

[a] *Laboratoire Jean Kuntzmann & INRIA Rhone-Alpes, team Mistis, Inovallee, 655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France*
[b] *GIPSA-Lab, Grenoble INP, Saint Martin d'Heres, France*
[c] *Institute of Statistics Graz University of Technology, Kopernikusgasse 24/III A-8010 Graz, Austria*

Abstract

Sliced Inverse Regression (SIR) is an effective method for dimensionality reduction in high-dimensional regression problems. However the method has requirements on the distribution of the predictors that are hard to check because they depend on unobserved variables. It has been shown that if the distribution of the predictors is elliptical then these requirements are satisfied. In case of mixture models the ellipticity is violated and in addition there is no assurance of a single underlying regression model among the different components. Our approach clusterizes the predictors space to force the condition to hold on each cluster and includes a merging technique to look for different underlying models in the data. SIR, not surprisingly, is not capable of dealing with a mixture of Gaussians with different underlying models whereas our approach is able to correctly investigate the mixture. A study on simulated data as well as two real applications is provided.

Keywords: Mixture models, inverse regression, sufficient dimension reduction

1. Introduction

In multidimensional data analysis, one has to deal with a dataset made of n points in dimension p. When p is large, classical statistical analysis methods and models fail. Supervised and unsupervised dimensionality reduction (d.r.) techniques are widely used to preprocess high dimensional data retaining the information useful to solve the original problem. Recently, more and more investigations aim at developing non-linear unsupervised techniques to better adapt to the complexity of our, often non-linear, World. Van der

Maaten et al. [24] provide an interesting review concluding that even if the variety of non-linear methods is huge, Principal component Analysis (PCA) [19], despite its intrinsic limitations, is still one of the best choices. PCA is not the best in specific cases (i.e. when additional information on the structure of the data are available) but, as expected, is rather general and can be easily controlled and applied. What about the case of supervised d.r.? In unsupervised d.r. one is interested in preserving all the information getting rid of the redundancies in the data. In other words, to catch the intrinsic dimensionality of the data, which is the minimum numbers of parameters needed to describe it [11]. In supervised d.r. a response variable Y is given and the analysis aims at providing a prediction (classification, when Y is categorical, or regression, when Y is continuous). Encoded in Y there is additional information of what we want to select in the data. Estimating the intrinsic dimensionality is no more our goal since we are oriented by the information present in Y.
Regression framework is characterized by the assumption of a link function between X and Y i.e. $Y = f(X, \epsilon)$, where ϵ is a random noise. In this environment it can be assumed that only a portion of X is needed to correctly explain Y. This is a reasonable assumption since data nowadays are rarely tailored on the application and filled by too many details. If Y depends on the multivariate predictor through an unknown number of linear projections $Y = f(X^T\beta_1, ..., X^T\beta_k, \epsilon)$ the effective dimension reduction (e.d.r) space is what we are looking for [15]. It is defined as the smallest linear space containing the information needed to correctly regress the function f. Under the previous assumption the e.d.r space is spanned by $\beta_1, ..., \beta_k$. Sliced Inverse Regression (SIR) [15] has proven to achieve good results retrieving a basis of the e.d.r. space. Recently, many papers focused on the complex structure of real data showing that often the data is organized in subspaces (see [14] or [23] for a detailed discussion and references). Our hypothesis is that the e.d.r. space is not unique all over the data and varies through the components. We introduce a novel technique to identify the number of e.d.r. spaces based on a weighted distance. With this paper we try to give an answer to the question: Can SIR be as popular as multiple linear regression? [5].
In section 2 we rapidly describe SIR and provide a discussion on the limitations of the method. The following section 3 is the core of our paper, where our contribution, Collaborative SIR is introduced. Motivation and main problem are described. Asymptotic results are established under mild conditions. The simulation study, section 4, is where the performances of

Collaborative SIR are shown and analyzed under specific test cases. The stability of the results is detailed and commented. In section 5 two real data applications are reported showing the interest of this technique. A discussion and conclusion are finally drawn encouraging the community to improve our idea.

2. Sliced Inverse Regression (SIR)

2.1. Method

Back in 1991, Li [15] called SIR a *data-analytic tool*: Even if the performance of computers and the capability to explore huge dataset increased tremendously, SIR remains a useful *tool* for d.r. in the framework of regression. The visualization of high dimensional datasets are nowadays of extreme importance because human beings are still, unfortunately, limited by a perception which only allows us to display 3 dimensions at a time while the capability to gather data is amazingly increasing. When p is large a possible approach is to suppose that *interesting features of high-dimensional data are retrievable from low-dimensional projections*, in other words the model Li proposed is:

$$Y = f(X^T\beta_1, ..., X^T\beta_k, \epsilon) \tag{1}$$

where $Y \in \mathbb{R}$ is the response variable, X is a random variable, $X \in \mathbb{R}^p$ ($\Sigma = \mathrm{Cov}(X)$, $\mu = \mathbb{E}(X)$). ϵ is a random error independent of X. If $k \ll p$ the functions depends on k linear combinations of the original predictors and the d.r. is achieved. The goal of SIR is to retrieve a basis of the e.d.r space. Under the Linearity Design Condition:

(LDC) $\mathbb{E}(X^Tb|X^T\beta_1, ..., X^T\beta_k)$ is linear in $X^T\beta_1, ..., X^T\beta_k$ for any $b \in \mathbb{R}^p$

Duan and Li [10] showed that the centered inverse regression curve is contained in the k-dimensional linear subspace of $\mathbb{R}^p$ spanned by $\Sigma\beta_1, ..., \Sigma\beta_k$. If we consider a monotone transformation $T(\cdot)$ of Y, the matrix $\Sigma^{-1}\Gamma$ is degenerated in any direction orthogonal to $\beta_1, ..., \beta_k$, where $\Gamma = \mathrm{Cov}(\mathbb{E}(X|T(Y)))$. Therefore the k eigenvectors corresponding to the k non zero eigenvalues form a basis of the e.d.r. space. To estimate Γ, Li [15] used a slicing procedure as candidate for $T(\cdot)$. Dividing the range of Y in non-overlapping slices, $s^1, ..., s^H (H > 1)$. Γ can then be written as:

$$\Gamma = \sum_{h=1}^{H} p^h (m^h - \mu)(m^h - \mu)^T,$$

where $p^h = P(Y \in s^h)$ and $m^h = \mathbb{E}(X|Y \in s^h)$. The estimator $\hat{\Gamma}$ can then be defined substituting p^h, m^h with the corresponding sample versions. The range of Y can be divided setting the width or the proportion of samples p^h in each slice, through the paper we adopted the second slicing strategy [5]. The k eigenvectors corresponding to the largest eigenvalues of $\hat{\Sigma}^{-1}\hat{\Gamma}$ are the estimation of a basis of the e.d.r. space.

2.2. Limitations

SIR's theory is well established and comes fully equipped by asymptotic results [13, 20]. Two main limitations affect the building:

- The inversion of the estimated covariance matrix $\hat{\Sigma}$;
- The impossibility to check if the (LDC) holds.

When the number of samples is $n \leq p$ the sample covariance matrix is singular, and when the variables are highly correlated (e.g. in hyperspectral images) the covariance matrix is ill conditioned. To compute the e.d.r directions the inversion of $\hat{\Sigma}$ must be achieved, recently many papers faced this problem and provided solutions ([6, 17, 21, 22, 25]). An homogeneous framework to perform regularized SIR has been proposed in [2] where, depending on the choice of the prior covariance matrix, the above mentioned techniques can be obtained and extended.

The (LDC), less studied in literature, is the central assumption of the theory and it depends on the unobserved e.d.r. directions, therefore it cannot be directly checked [26]. It can be proved that if X is elliptical distributed the condition holds. This condition is much stronger than (LDC) but easier to verify in practice since it does not depend on the β_1,...,β_k. Good hope comes from a result of Hall and Li [12] that shows that, when the dimension p tends to infinity, the measure of the set of directions for which the (LDC) does not hold tends to zero. The condition becomes weaker and weaker as soon as the dimension increases. The intuition comes from [9] where the authors show that high dimensional dataset are nearly normal in most of the low dimensional projections. If X follows an elliptical distribution the (LDC) condition holds, it is desirable to work in the direction that allows us to use

this property. Unfortunately when X follows a mixture of elliptical distributions this property is not globally verified. Kuentz and Saracco [14] using an idea from [16] proposed to clusterize the space to look locally for ellipticity rather than globally. Chavent et al. [4] introduced categorical predictors to distinguish different populations. This is our very start, assuming X from a mixture model we focus on decomposing the mixture and we extend the basic model to improve SIR's capability to explore complex datasets.

3. Collaborative SIR

First, we give a motivation and introduce in subsection 3.1 the population version of Collaborative SIR. Second, a sample version in different steps is detailed and an algorithm is outlined (subsections 3.2-3.5). For sake of simplicity we will focus on the case when $k = 1$ i.e. the effective dimension reduction space is of dimension one.

3.1. Population version

In SIR the underlying model through the whole predictors space is $Y = f(\beta^T X, \epsilon)$. When dealing with complex data one could allow the underlying model to change depending on the predictor space. Mixture models provide a good framework to deal with such hypothesis considering the data a realization from a weighted sum of distributions with different parameters. As mentioned before, in such case there is no straightforward way to check if the (LDC) holds. Let X be a random vector, $X \in \mathbb{R}^p$, from a mixture model and be Z an unobserved latent random variable $Z \in \{1, ..., c\}$, where c is the number of components. Given $Z = i$ we have the following model:

$$Y = f_{F(i)}(\beta_{F(i)}^T X) + \epsilon_i, \tag{2}$$

where Y is the random variable to predict, $Y \in \mathbb{R}$, F is an unknown deterministic function $F : \{1, ..., c\} \to \{1, ..., D\}$, $D \in \mathbb{N}$. The functions $f_j : \mathbb{R} \to \mathbb{R}$, $j = 1, ..., D$ are unknown link functions between $\beta_j X$ and Y. Finally ϵ_i are random errors $\forall i \ \epsilon_i \in \mathbb{R}$, i.e. each component is allowed to have a different related error.

Under the model (2), D is the number of different e.d.r spaces. The goal is to find a basis of the D one-dimensional spaces spanned by $\beta_1, ..., \beta_D$. The number D ($D \leq c$) of e.d.r. spaces is unknown and the link function

may change depending on the component. Function F selects the underlying model for the specific component. It is assumed that the (LDC) holds in each component:

(LDC) $\forall i = 1, ..., c\ \mathbb{E}(X^T b | X^T \beta_{F(i)}, Z = i)$ is linear in $X^T \beta_{F(i)}$ for any b.

Given $Z = i$, we define the mean $\mu_i = \mathbb{E}(X|Z = i)$, the covariance matrix $\Sigma_i = Cov(X|Z = i)$ and $\Gamma_i = Cov(\mathbb{E}(X|Y, Z = i))$. Hence the eigenvector b_i corresponding to the highest eigenvalue of $\Sigma_i^{-1}\Gamma_i$, is a basis of the e.d.r. space: $Span\{b_i\} = Span\{\beta_{F(i)}\}$ from SIR theory [15].
If $F : \{1, ..., c\} \to \{1, ..., D\}$ is known, the inverse image of the elements $j \in \{1, ..., D\}$ can be defined:

$$F^{-1}(j) = \{i \in \{1, ..., c\} \text{ s.t. } F(i) = j\},$$

since F is not required to be injective, an e.d.r direction β_i may be associated with several components. Suppose that $\{b_i, i \in F^{-1}(j)\}$ are observed, given the proximity criteria

$$m(a, b) = \cos^2(a, b) = (a^T b)^2, \tag{3}$$

the "most collinear vector" to the set of directions $\{b_i, i \in F^{-1}(j)\}$ is the solution of the following problem:

$$\max_{v \in \mathbb{R}^p, \|v\|=1} \sum_{i \in F^{-1}(j)} m(v, b_i) = \max_{v \in \mathbb{R}^p, \|v\|=1} \sum_{i \in F^{-1}(j)} (v^T b_i)^2 =$$
$$= \max_{v \in \mathbb{R}^p, \|v\|=1} v^T \Big(\sum_{i \in F^{-1}(j)} (b_i b_i^T) \Big) v = \max_{v \in \mathbb{R}^p, \|v\|=1} v^T (B_j^T B_j) v,$$

where $B_j = [b_{i, i \in F^{-1}(j)}]$. Using Lagrange multipliers is easy to show that vector v must be an eigenvector of the matrix $(B_j^T B_j)$ and, since we want to maximize, it will be the one associated with the largest eigenvalue. The following lemma motivates this argument.

Lemma 1. *Assuming the (LCD) and model* (2) *the eigenvector* $\tilde{\beta}_j$ *associated to the only non-zero eigenvalue of the matrix* $[B_j B_j^T]$ *is collinear with* β_j.

Proof. For each $i \in F^{-1}(j)$, b_i is collinear with β_j, $b_i = \alpha_i \beta_{F(i)}$, $\alpha_i \in \mathbb{R} \setminus \{0\}$. Since $B_j = [\alpha_i \beta_i, i \in F^{-1}(j)]$ we have:

$[B_j B_j^T] = \sum_{i \in F^{-1}(j)} \alpha_i^2 \beta_j \beta_j^T = \|\alpha\|^2 \beta_j \beta_j^T$. This concludes the proof. □

This lemma shows that $\tilde{\beta}_j$ is an e.d.r. direction for each j and the precedent argument gives a strategy to estimate the directions β_j based on the proximity criteria (3).

Remark. If $D = 1$ then $F^{-1}(1) = \{1, ..., c\}$, the e.d.r. direction and the link functions do not vary through all the mixture. This specific case is addressed in [14].

3.2. Sample version: Z is observed, F and D known

Let $\{Y_1, ..., Y_n\}$ be a sample from Y, $\{X_1, ..., X_n\}$ a sample from X, $\{Z_1, ..., Z_n\}$ a sample from Z. We suppose Z_i observed at this stage. Let $\mathcal{C}_i = \{t \text{ such that } Z_t = i\}$ and $n_i = card(\mathcal{C}_i)$.
We can now estimate for each $\mathcal{C}_i$ the mean and covariance matrix:

$$\bar{X}_i = \frac{1}{n_i} \sum_{t \in \mathcal{C}_i} X_t, \; \hat{\Sigma}_i = \frac{1}{n_i} \sum_{t \in \mathcal{C}_i} (X_t - \bar{X}_i)(X_t - \bar{X}_i)^T, \text{ for each } i = 1, ..., c.$$

To obtain an estimator for Γ_i, we introduce as in classical SIR a slicing. For each $\mathcal{C}_i$ we can define the slicing T_i of Y_i into $H_i \in \mathbb{N}$ slices ($H_i > 1 \; \forall i = 1, ..., c$). Let $s_i^1, ..., s_i^{H_i}$ be the slicing associated to $\mathcal{C}_i$, $\Gamma_i = Cov(\mathbb{E}(X|Y, Z = i))$ can be written as:

$$\Gamma_i = \sum_{h=1}^{H_i} p_i^h (m_i^h - \mu_i)(m_i^h - \mu_i)^T,$$

where $p_i^h = P(Y \in s_i^h | Z = i)$, $m_i^h = \mathbb{E}(X | Z = i, Y \in s_i^h)$. Let us recall that $\mu_i = \mathbb{E}(X|Z = i)$ and $\Sigma_i = Cov(X|Z = i)$, as defined in section 3.1. Let $n_{h,i} = \sum_{t \in \mathcal{C}_i} \mathbb{I}[Y_t \in s_t^h]$, where $\mathbb{I}$ is the indicator function. Replacing p_i^h, m_i^h with the corresponding sample versions, it is possible to estimate Γ_i:

$$\hat{\Gamma}_i = \sum_{h=1}^{H_j} \hat{p}_i^h (\hat{m}_i^h - \bar{X}_i)(\hat{m}_i^h - \bar{X}_i)^T,$$

where $\hat{p}_i^h = \frac{n_{h,i}}{n_i}$ and $\hat{m}_i^h = \frac{1}{n_{h,i}} \sum_{t \in \mathcal{C}_i} X_t \mathbb{I}[Y_t \in s_t^h]$. The estimated e.d.r. directions are then $\hat{b}_1, ..., \hat{b}_c$ where $\hat{b}_i$ is the major eigenvector of the matrix $\hat{\Sigma}_i^{-1}\hat{\Gamma}_i$. This allows us to estimate B_j and β_j:

- (i) $\hat{B}_j = [\hat{b}_{i,i \in F^{-1}(j)}]$, $i \in \{1, ..., c\}$, $\hat{B}_j$ is a $p \times |F^{-1}(j)|$ matrix;
- (ii) $\hat{\beta}_j$ $\forall j = 1, ..., D$ is the major eigenvalue of $\hat{B}_j^T \hat{B}_j$.

Asymptotic results can be establish similarly to Chavent et al. [3]. We fix $j \in \{1, ..., D\}$ and consider $\{X_t,\ t \in \bigcup_{i \in F^{-1}(j)} \mathcal{C}_i\}$ and a sample size $n^j = \sum_{i \in F^{-1}(j)} n_i$ which tends to ∞. The following three assumption are considered:

- (A1) $\{X_t,\ t \in \bigcup_{i \in F^{-1}(j)} \mathcal{C}_i\}$ is a sample of independent observations from the single index model (2).
- (A2) For each i, the support of $\{Y_t, t \in \mathcal{C}_i\}$ is partitioned into a fixed number H_t of slices such that $p_i^h > 0, h = 1, \ldots, H_t$.
- (A3) For each i and $h = 1, \ldots, H_t$, $n_{h,i} \to \infty$ (and therefore $n_i \to \infty$) as $n \to \infty$.

Theorem 1. *Under model (2), linearity condition (LDC) and assumptions (A1)-(A3), we have:*

(*i*) $\widehat{\beta}_j = \beta_j + O_p(\underline{n}^{j^{-1/2}})$, *where* $\underline{n}^j = \min_{i \in F^{-1}(j)} n_i$;

(*ii*) *If, in addition* $n_i = \theta_{ij} n^j$, $\theta_{ij} \in (0, 1)$ *for each* $i \in F^{-1}(j)$, *then* $\sqrt{n^j}(\widehat{\beta}_j - \beta_j)$ *converges to a centered Gaussian distribution.*

Proof. (*i*) For each $i \in F^{-1}(j)$ and under the assumptions (*LC*), (A1)-(A3), from the SIR theory [15] each estimated EDR direction $\widehat{b}_i$ converges to β_j at root $\underline{n}^j$ rate: that is, for $i \in F^{-1}(j)$, $\widehat{b}_i = \beta_j + O_p(\underline{n}^{j^{-1/2}})$. We then have $\widehat{B}_j^T \widehat{B}_j = B_j^T B_j + O_p(n^{j^{-1/2}})$. Therefore the principal eigenvector of $\widehat{B}_j^T \widehat{B}_j$ converges to that corresponding to $B_j^T B_j$ at the same rate: $\hat{\beta}_j =$

$\beta_j + O_p(\underline{n}^{j^{-1/2}})$. The estimated e.d.r. direction $\hat{\beta}_j$ converges to an e.d.r. direction at root $\underline{n^j}$ rate. □

(*ii*) The proof is similar to the one of Chavent et al. [3], Theorem 2.

In the following sections a merging algorithm is introduced to infer the number D based on the collinearity of the vectors b_i and a procedure is given to estimate the function F.

3.3. Sample version: D unknown, Z is observed and F known

We assumed, so far, D known. To estimate D a hierarchical merging procedure is introduced based on the proximity measure (3) between the estimated e.d.r. directions $\hat{b}_1, ..., \hat{b}_c$. A similar procedure has been used in Coudret et al. [8] to cluster the components of the multivariate response variable Y related to the same e.d.r. spaces.

Definition. Let $V = \{v_1, v_2, ..., v_{|V|}\}$ be a set of vectors in dimension p with associated weights w_i. We define the quantity $\lambda(V)$:

$$\lambda(V) = \max_{v \in \mathbb{R}^p} \frac{1}{w_V} \sum_{i=1}^{|V|} w_i m(v_i, v) \text{ s.t. } \|v\| = 1$$

$$= \text{largest eigenvalue of } \frac{1}{w_V} \sum_{i=1}^{|V|} w_i v_i v_i^T$$

where $w_V = \sum_{i=1}^{|A|} w_i$ is the normalization. Vector v maximizing $\lambda(V)$ is the most collinear vector to our set of vectors given the proximity criteria (3) and the weights w_i. To build the hierarchy we consider the following iterative algorithm initialized with the set $A = \{\{\hat{b}_1\}, ..., \{\hat{b}_c\}\}$:

while $card(A) \neq 1$

Let $a, b \in A$ **such that** $\lambda(a \cup b) > \lambda(c \cup d) \forall c, d \in A$
$A = (A \setminus \{a, b\}) \bigcup a \cup b$
end

the weights are set equal to the number of samples in each components,

i.e. $w_i = n_i$, $i = 1, ..., c$. At each step the cardinality of the set A decreases merging the most collinear sets of directions (Fig. 1). The bottom up greedy algorithm proceeds as follows:

- First the two most similar elements of A are merged considering all the $|A| \times (|A| - 1) = c \times (c - 1)$ pairs ($\hat{b}_1, \hat{b}_2$ are selected to be merged in Fig. 1).
- In the following steps the two most similar sets of vectors are merged, considering all $|A| \times (|A| - 1)$ pairs in A (e.g. in the second step $A = \{\{\hat{b}_1, \hat{b}_2\}, \{\hat{b}_3\}, ..., \{b_{12}\}\}$ in Fig. 1)

Therefore it is possible to infer the number D of underlying e.d.r. spaces analyzing the values of λ in the hierarchy (Fig. 2) looking for a discontinuity that will occur when two sets with different underlying β_j (i.e. non collinear) are merged. We automatically estimate D with the following procedure:

(*i*) Draw a line from the first value of the graph $(1, \lambda_1)$ to the last (c, λ_c).

(*ii*) Compute the distance between points in the graph and the line.

(*iii*) Select the merging point maximizing that distance. $\hat{D} = c-$number of merge selected.

Once achieved an estimation of D, $\hat{D}$, function F can be estimated. Even if we used an automatic procedure, a visual selection of $\hat{D}$ depending on the task and previous knowledge is strongly recommended.

3.4. Sample version: F unknown

For each node of the tree at level $\hat{D}$, the "most collinear direction", using (3), is computed. Solving the related $\hat{D}$ diagonalization problems gives $\hat{\beta}_1, ..., \hat{\beta}_{\hat{D}}$. In the following paragraph a procedure for the estimation $\hat{F}$ of the function F is detailed.

Once the candidates $\hat{\beta}_1, ..., \hat{\beta}_{\hat{D}}$ are estimated, the whole data (X,Y) is considered to estimate F. Starting from $i \in \{1, ..., \hat{c}\}$ the goal is to find $j \in \{1, ..., \hat{D}\}$ such that $F(i) = j$, under certain conditions. The $\hat{D}$ covariance matrices of the distributions $(X_t^T \hat{\beta}_j, Y_t)$, $t \in \mathcal{C}_i$, $j \in \{1, ..., \hat{D}\}$ are considered. The idea is to select the direction that best explains Y_t, $t \in \mathcal{C}_i$ among the

estimated directions $\hat{\beta}_1, ..., \hat{\beta}_{\hat{D}}$.
Let us assume f_j functions "locally" linear $(A4)$: f_j can be approximated with piecewise linear functions so that $Y_t = f_j(X_t^T\beta_j) = k_i X_t^T \beta_j$, $\forall t \in \mathcal{C}_i$, $i \in F^{-1}(j)$.

Lemma 2. *Let $j \in \{1, ..., D\}$. Under assumption $(A4)$ the e.d.r. direction β_j is the vector minimizing the second eigenvalue of the covariance matrix of the pairs $(X^T\beta_s, Y)_{s=1,...,D}$.*

Proof. We have that:

$$cov(X^T\beta_s, Y) = cov(X^T\beta_s, k_i X^T\beta_j) = \begin{pmatrix} \beta_s^T\Sigma\beta_s & k_i\beta_s^T\Sigma\beta_j \\ k_i\beta_s T\Sigma\beta_j & k_i^2\beta_j^T\Sigma\beta_j \end{pmatrix} =$$
$$= \begin{pmatrix} \langle\beta_s, \beta_s\rangle & k_i\langle\beta_s, \beta_j\rangle \\ k_i\langle\beta_s, \beta_j\rangle & k_i^2\langle\beta_j, \beta_j\rangle \end{pmatrix} = \begin{pmatrix} \|\beta_s\|^2 & k_i\langle\beta_s, \beta_j\rangle \\ k_i\langle\beta_s, \beta_j\rangle & k_i^2\|\beta_j\|^2 \end{pmatrix}$$

where the scalar product and the norm are induced by Σ. The characteristic polynomial is $p(\lambda) = \lambda^2 - \lambda(\|\beta_s\|^2 + k_j^2\|\beta_j\|^2) + k_j^2(\|\beta_s\|^2\|\beta_j\|^2 - \langle\beta_s, \beta_j\rangle^2)$. We have $\Delta = (\|\beta_s\|^2 - k_j^2\|\beta_j\|^2) + 4k_j^2\langle\beta_s, \beta_j\rangle^2 > 0$. From Cauchy-Schwarz inequality $\lambda_1, \lambda_2 \geq 0$ and $\lambda_2 = 0$ if and only if the equality holds. Since β_s, $s = 1, ..., D$ are linearly independent it follows that $\lambda_2 = 0$ if and only if $\beta_s = \beta_j \Leftrightarrow s = j$. □

In practice, fixed $i = \{1, ..., \hat{c}\}$, vectors $\hat{\beta}_j$, $j = 1, ..., \hat{D}$ are the candidates for (X_t, Y_t), $t \in \mathcal{C}_i$. Lemma 2 is stating that under the assumption $(A4)$ the vector $\hat{\beta}_j$ minimizing the second eigenvalue of $(X_t^T\hat{\beta}_s, Y_t)$, s=1,...,D, $t \in \mathcal{C}_i$ is such that $j = F(i)$. We require the functions to be locally linear, if the functions are approximately linear the estimation will work. In case of dramatic non linearities the method may lead to unreasonable results. A possibility is to resize the interval where we want to regress the functions and zoom until we find a reasonable local behavior of the functions.
It must be noted that in case D is overestimated $\hat{D} > D$ (e.g. due to instabilities in the estimation of the direction in some components) in the simulation we observed that the estimation of F mitigates this error often avoiding to select the aberrant directions β_j, $j > D$.

3.5. Estimation of Z via clustering

To estimate the latent variable Z the explanatory space X is partitioned using a k-means algorithm. It is worth noticing that we decided to use k-means for simplicity and also to compare our results with [14]. Twenty initial

random centroids are chosen as initialization of k-means, the one minimizing the sum of squares is retained.

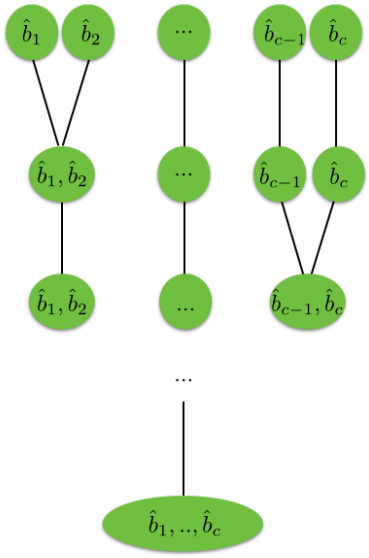


Figure 1: Hierarchy built following the proximity criteria (3).

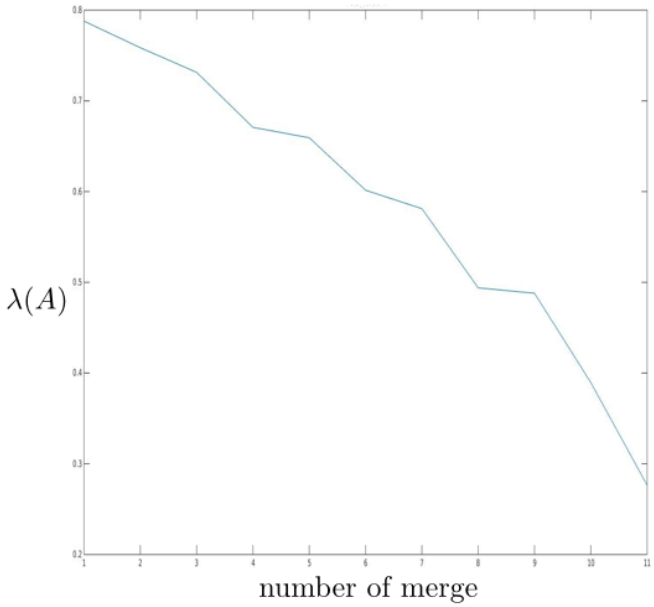


Figure 2: Cost function $\lambda(A)$, the number D of unknown e.d.r directions decreases at each step by one. $\hat{D} = c-$number of merge selected. In the example above $c = 12$. The algorithm selects merge step 9 which corresponds to the correct estimation of the parameter: $\hat{D} = 3$.

4. Simulation study

We performed a study on simulated data, this was the opportunity to test in a controlled setting and evidence the weaknesses and strengths of the method. Two aspects are of interest:

(A) Study the sensitivity to clustering (estimation of Z).

(B) Analyze the quality of the estimation compared to SIR performed independently in each cluster.

The first experiment is performed on the same dataset to study the effect of different initializations of k-means and how the quality of clustering affects the result. In the second experiment different simulated datasets are analyzed to test the method under a variety of different conditions.

4.1. Test case A

To study the sensitivity to clustering $n = 2500$ samples from Gaussian mixture model are drawn with uniform mixing proportions and $c = 10$ components. Each component follows a Gaussian distribution $\mathcal{N}(\mu_i, \Sigma_i)$, $\Sigma_i = Q_i \Delta Q_i^t$ where Q_i is a matrix drawn from the uniform distribution on the set of orthogonal matrices and $\Delta_{ii} = (\frac{(p+1-i)}{p})^{\theta_i}$. The parameter θ_i is randomly drawn from the standard uniform distribution. To prevent too close centroids, each entry of the μ_i is the result of adding two samples from the standard uniform distribution. In figure 3 the projection on the two first principal components of the considered mixture is reported, different colors represent different components. Data in figure 3 appear mixed and clustering non-trivial. Clustering centroids are randomly initialized 100 times, the iterations of k-means are limited to five to prevent the clustering to converge. The number of clusters is supposed to be known. Y is simulated as follows:

- For each $i \in \{1, ..., c\}$, one of the two possible directions $\beta_j \in \{\beta_1, \beta_2\}$ is randomly selected with probability 1/2.
- $Y_t = sinh(X_t^T \beta_j) + \epsilon$, $\forall t \in \mathcal{C}_i$, $i \in F^{-1}(j)$ where $\epsilon \sim \mathcal{N}(0, 0.1^2)$ is an error independent of X_t.

The two e.d.r. spaces are randomly generated and orthogonalized: $\beta_1^t \beta_2 = 0$. We are interested in the case when we insert in the same cluster samples from different components. This is the case when we estimate Z by $\hat{Z}$ such that for some (t_1, t_2) we have $\hat{Z}_{t_1} = \hat{Z}_{t_2}$ but $Z_{t_1} \neq Z_{t_2}$.

For each of the 100 runs of k-means the estimated directions for Collaborative SIR $\{\hat{\beta}_{\hat{F}(1)}, ..., \hat{\beta}_{\hat{F}(c)}\}$ are considered. The number of samples in each slice is set to 250 resulting in $H = 10$ uniform slices. The average of the squared cosines (3) between the estimated and real direction $\{\beta_{F(1)}, ..., \beta_{F(c)}\}$ is computed (see column 2 Table 1). The 100 results are then averaged. In the cases where clustering has zero error (fig. 4) the average of the quality measure is 0.8958. Averaging only on the runs of k-means with more than 10 percent of error (fig. 4) the quality measure decreases to 0.8273. This shows that even if, not surprisingly, an error on the estimation of Z affects the solution, the influence is, empirically proved, not to be severe. It must be noted that we obtain the worst results when we insert in the same clusters samples with different underlying models: $\hat{Z}_{t_1} = \hat{Z}_{t_2}$ but $Z_{t_1} \neq Z_{t_2}$ and there is no j such that $Z_{t_1}, Z_{t_2} \in F^{-1}(j)$. This is indeed the reason why we extended SIR's theory.

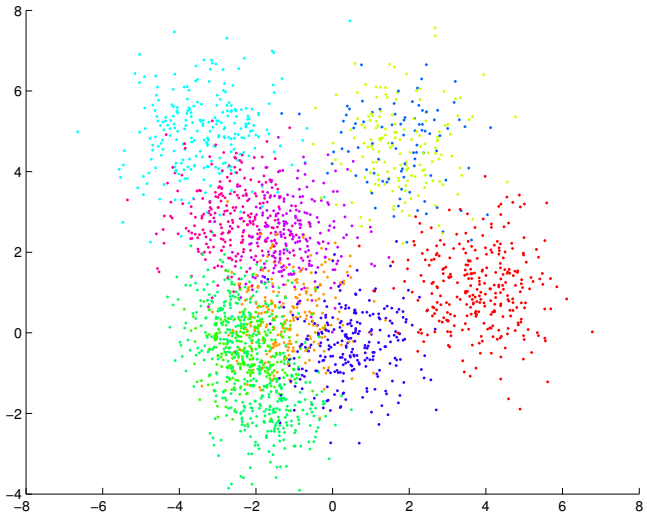


Figure 3: Projection on the two first principal components of the considered mixture, different colors represent different components.

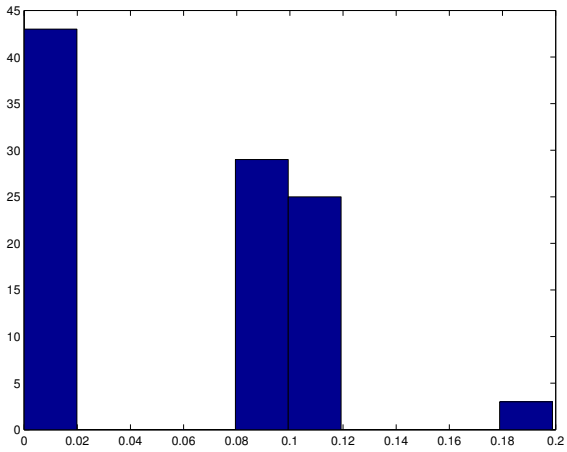


Figure 4: Histograms of the percentage of badly clustered samples over 100 runs of k-means.

4.2. Test case B

To investigate the strengths and limitations of the method 100 different mixture of Gaussian models for different numbers of total samples (10000, 5000, 2500) are generated. Only the case where $n = 2500$, dimension $p = 200$, $D = 2$, $c = 10$ and $\beta_1^T \beta_2 = 0$ is displayed here. The response variable Y is generated as in test case A for each of the 100 datasets. The same slicing strategy with $H = 10$ is applied. We selected such dimension p to mimic the dimensionality of hyperspectral satellite sensors that are of interest in future works. The number of clusters is supposed to be known. Not surprisingly, as soon as the dimension decreases the performance of the algorithm are more and more stable, e.g. at dimension $p = 50$ the performance are still stable and accurate. Analyzing the histograms of the differences of the average of the squared cosines (Table 1) between Collaborative SIR and SIR (figure 5) it is evident that Collaborative SIR is always improving the quality of the estimation leading to a significant difference. The average and standard deviation of the 100 quality measures is 0.50 ± 0.05 for SIR and 0.80 ± 0.07 for Collaborative SIR. Since the quality measure is bounded to 1, a relevant improvement is found using Collaborative SIR. In figure 6 we show the estimation $\hat{D}$ of the number of e.d.r. spaces. The estimation is concentrated around the true value, $D = 2$.

Table 1: Quality measure

SIR	Collaborative SIR
$\frac{1}{c}\sum_{i=1}^{c} \cos^2(\hat{b}_i, \beta_{F(i)})$	$\frac{1}{c}\sum_{i=1}^{c} \cos^2(\hat{\beta}_{\hat{F}(i)}, \beta_{F(i)})$

4.3. Comments on simulation results

In the simulations the sensitivity to clustering and the effective gain in using Collaborative SIR are analyzed. Several tests changing the dimension p, and the collinearity of the β_j were carried out. As soon as the directions get collinear our model is no more identifiable, despite that, the results are not affected. When the vectors are, in the limit, collinear the e.d.r spaces simply reduce to one. Non orthogonal e.d.r. directions and multiple e.d.r. spaces ($D = 3$) have been analyzed reporting good results in case of orthogonality

and non orthogonality of the β_j's. Simulations are interesting but cannot cover the complexity of the real application. In the following, two real dataset where Collaborative SIR shows its capabilities are discussed and analyzed.

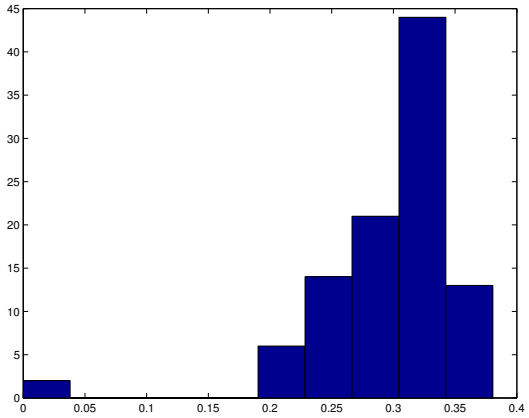


Figure 5: Histograms of the difference between the quality measure (table 1) of Collaborative SIR and SIR obtained over 100 different dataset.

5. Real data application

We show, in the following, two real applications where the number D of different effective dimension spaces differs from one. Nevertheless, it must be underlined that for many different datasets $D = 1$ was found. This is extremely satisfying because it means that in those cases a single underlying model, $Y = f(\beta^T X, \epsilon)$, is the best choice for the considered dataset. First, the Horse-mussel dataset, that can be found in Kuentz and Saracco [14], is considered. Second, a dataset composed of different parameters on galaxies is investigated. Finally a discussion on possible improvements, strengths and limitations is drawn.

5.1. Horse-mussel dataset

The horse-mussel dataset X is composed of $n = 192$ samples of different numerical measures of the shell: length, width, height and weight ($p = 4$, a detailed description can be found in Cook and Weisberg [7]). The response

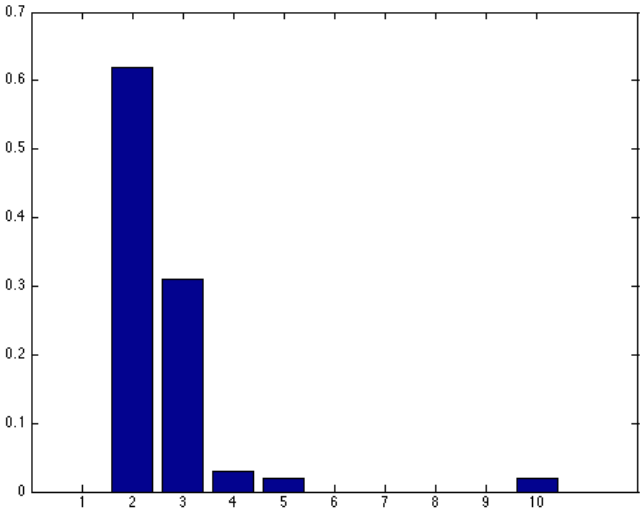


Figure 6: Bar plot of frequencies of the number of estimated e.d.r spaces $\hat{D}$ over 100 different dataset, $D = 2$.

variable Y to predict is the weight of the edible portion of the mussel. To compare to [14] the discrete response variable was transformed into a continuous variable $Y = Y + \epsilon$, $\epsilon \sim N(0, 0.01^2)$. The clustering obtained by [14] was adopted and the number of slices set to four: $H_i = 4$ for all $i \in \{1, ..., 5\}$. The following algorithm is used to analyze and compare SIR, cluster SIR and Collaborative SIR:

(1) Randomly select 80% of X for training T and 20% for validation, V.

(2) Apply SIR, cluster SIR and collaborative SIR on the training.

(3) Project and regress the functions using the training samples (we fitted a polynomial of degree 2)

(4) Compute the Mean Absolute Relative Error (MARE) on the test.
MARE$= \frac{1}{|V|} \sum_{Y \in V} \frac{Y - \hat{Y}}{Y}$, where $\hat{Y}$ is our estimation.

We computed 100 different training and validation set. In figure 7 the box plots of the three different methods are shown. It must be noted that this dataset is low dimensional: $p = 4$. However it is of interest that the number

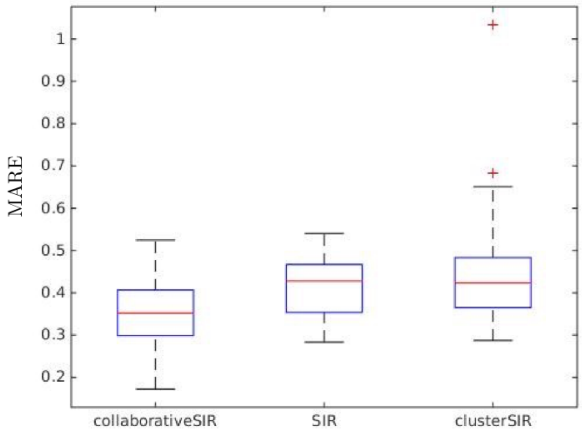


Figure 7: Box plots of MARE for Collaborative SIR, SIR and Cluster SIR using 100 different initializations.

of e.d.r. spaces found is $\hat{D} = 2$. In figure 9 the data is decomposed and the regression of the two link functions appears easier compared to the regression in figure 8 where the cloud of point is thicker and not well shaped. Using different regression techniques (Gaussian kernel and polynomial regression) the results do not change significantly. On this dataset Collaborative SIR performed better than SIR and cluster SIR. In addition, this result suggests that two subgroups are present in the data.

5.2. Galaxy dataset

The Galaxy dataset is composed by $n = 292766$ different galaxies. Aberrant samples have been removed from the dataset after a careful observation of the histograms in each variable supervised by experts. The response variable Y is the specific stellar formation rate. The predictor X is of dimension $p = 46$ and is composed of spectral characteristics of the galaxies. For all the tests the number of samples in the first $H - 1$ slices is the closest integer to n/H, $H = 1000$. We applied Collaborative SIR on the whole dataset to investigate the presence of subgroups and different directions.
After different runs and number $\hat{c} = \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$ of clusters we observed two different subgroups and hence directions $\hat{\beta}_1, \hat{\beta}_2$.

Best results are reported with $\hat{c} = 5$, in figure 10 the two non linear link

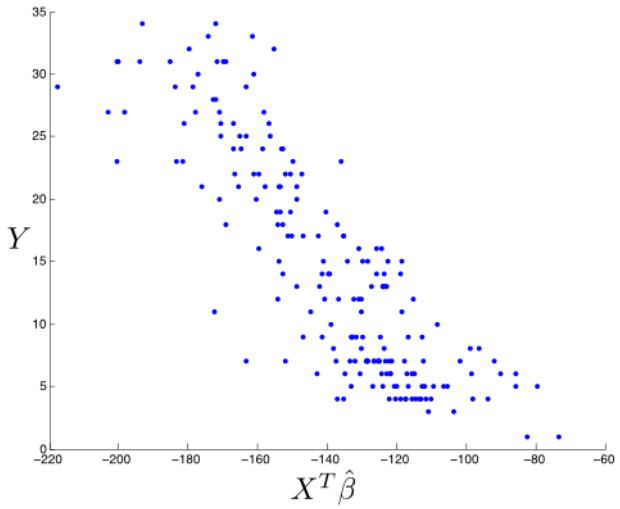


Figure 8: Graph of Y and the projection along the direction $\hat{\beta}$ found by SIR.

functions are shown. Clouds are thick but they show a very clear trend in the distributions. This dataset is a good example of how, in high dimension, two families can be found in a dataset using Collaborative SIR.
In figure 11 the distribution of the coefficients of the two directions are presented. It is interesting to observe how some variables are contributing in both linear combinations but that there is a reasonable difference in four variables (variables 2, 3, 6 and 23). The $d4000_n$ (variable 40), found to be relevant for both directions, is often used to estimate the specific stellar formation rate. Experts are working on a possible physical interpretation of the results. Even if the link functions look similar, we observe a significant difference in the coefficient of the two directions. This could lead to a better understanding and designing of further analysis of this kind of data.

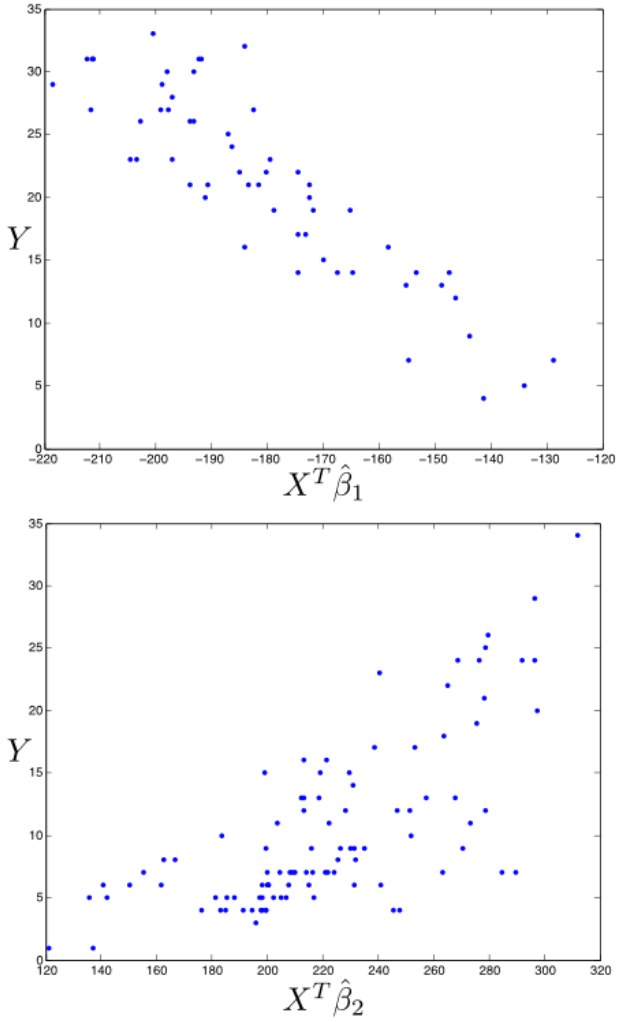


Figure 9: (Top) Graph of Y and the projection along the first direction $\hat{\beta}_1$ found by Collaborative SIR. (Bottom) Graph of Y and the projection along the second direction $\hat{\beta}_2$ found by Collaborative SIR. The directions $\hat{\beta}_1, \hat{\beta}_2$ found are nearly orthogonal $\cos^2(\hat{\beta}_1, \hat{\beta}_2) = 0.01$

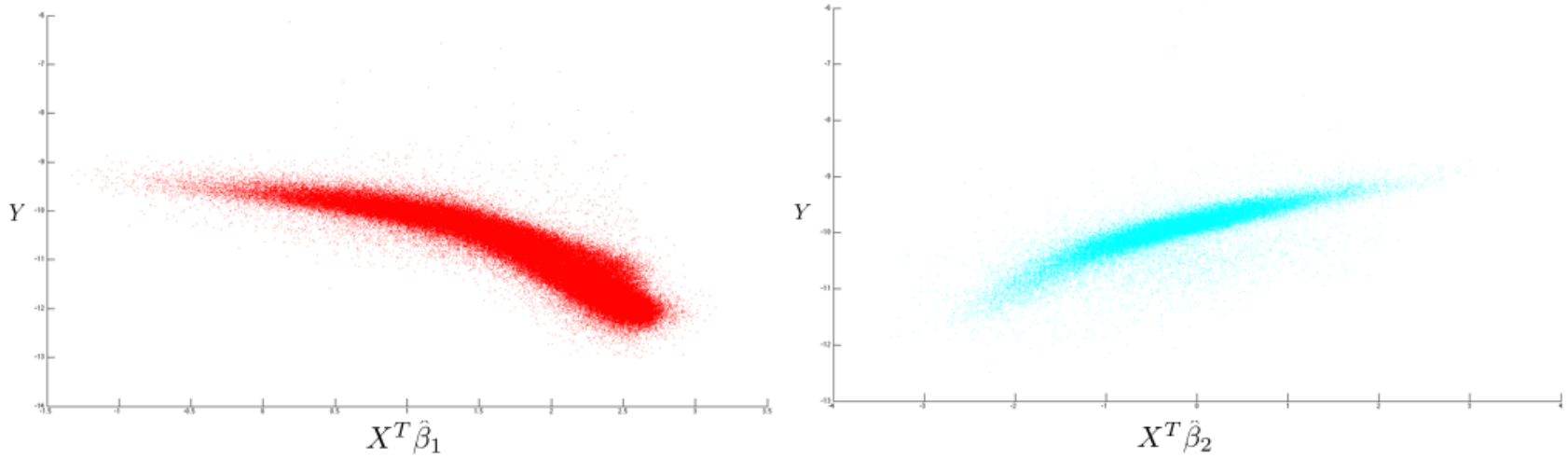


Figure 10: (Left) Graph of Y and the projection along the first e.d.r. direction $\hat{\beta}_1$ found by Collaborative SIR. (Right) Graph of Y and the projection along the second e.d.r. direction $\hat{\beta}_2$ found by Collaborative SIR. It is evident the nonlinear behavior of the two link functions.

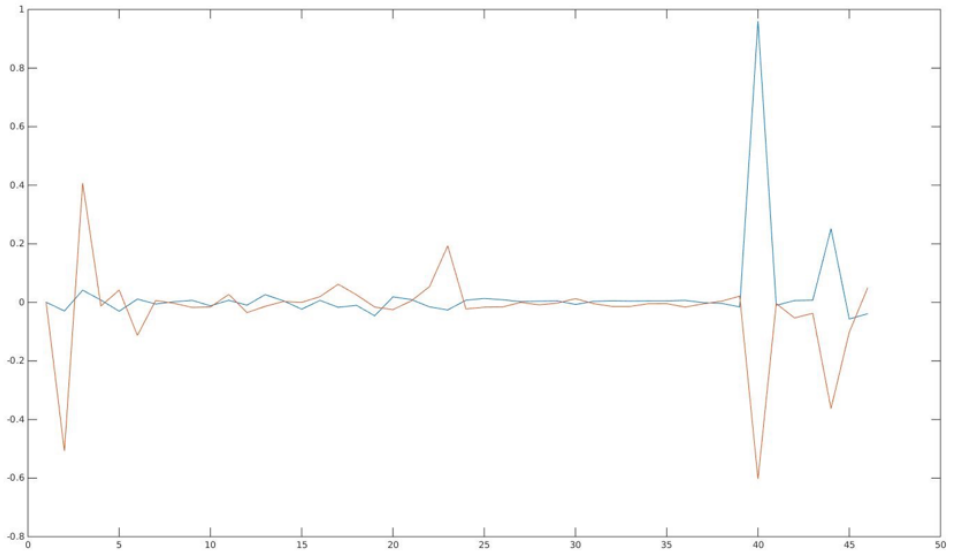


Figure 11: Differences in e.d.r directions $\hat{\beta}_1$ and $\hat{\beta}_2$. Many elements in the vectors are close to zero resulting in a variable selection. Differences in the two lines show how different variables contribute in regressing Y. The squared cosine between the two directions is 0.42

5.3. Discussion on dimension k and the number of clusters c

In the whole paper we presented results for dimension $k = 1$ ($Y = f(X^T\beta_1, ..., X^T\beta_k)$), the assumption is that e.d.r. spaces are one-dimensional. It is worth noticing that the entire approach can be easily extended to a higher k, it is sufficient to give a proximity measure between the linear subspaces (e.g. $Trace$ in [3]). If the dimension k is uniform in all the e.d.r. spaces the same strategy can be applied leading to a hierarchical merging tree. In case the dimension k varies depending on the mixture the proximity between e.d.r. spaces with different k is set to zero. It must be noted that the estimation of the dimension k is a classical problem for SIR ([15, 1]), a solution in real application is described in [18] where a graphical approach is proposed to analyze the information of the single projections $X^T\beta_j$ versus Y. SIR is a method to reduce dimensionality to "better" perform regression. When a regression is performed the visualization of the results is crucial, that is one of the reasons for dimensionality reduction. If the dimension k is greater than 2 visualization is not possible. This explains why SIR and its variants have mainly been applied with $k = 1$. Collaborative SIR is first dividing the predictors space into clusters, it seems natural to assume that dimension k locally would be smaller than globally i.e. that considering $k = 1$ is not a severe restriction if a visualization is needed. Finally another drawback of increasing dimensionality is that the samples become be more and more sparse and not cover enough the surface we want to regress, different regression techniques may lead to dramatically different results. The problem of dimension k could be the reason why SIR is not yet widely used.
We did not give an automatic way of selecting the number of clusters. In SIR literature Kuentz and Saracco [14] translate the selection in an optimization problem. Nowadays, with the increasing capabilities of sensors, data are complex and complicated and is hard to define a general criteria, ignoring previous knowledge, that could work for any kind of data. The number of clusters is deeply connected with how we want to group elements, the same data can show two possible "correct" clustering, depending on the task. Since SIR and collaborative SIR are fast and simple techniques the user, using prior information, should orient the clustering and try different values for the parameters and empirically check which is the most suitable for the purpose. Developing flexible clustering capable of incorporating prior knowledge is one of our interests.

6. Conclusion and future work

Sliced Inverse Regression is an interesting and fast tool to explore data in regression, it is yet not so popular [5] but has well established theory and simple implementation. If the link function turns out to be linear SIR, not surprisingly, is outperformed by linear regression techniques, but in case of evidence of non linearity, linear regression techniques force the model resulting in poor estimations. Collaborative SIR is meant to deal with the increasing complexity of the dataset that statisticians are asked to analyze. Often there is no reasonable criteria of gathering the samples resulting in dataset that are, at least, a mixture of different phenomena and/or full of ambiguous samples. The hypothesis of having different families with different underlying models gives flexibility not affecting tractability. We encourage the community to improve our idea. A robustified version of SIR will be our main field of research for the next period.

Acknowledgement

The authors thank Didier Fraix-Burnet for his contribution to the data. They are grateful to Vanessa Kuentz and Jerome Saracco for providing their results on Horse-mussel dataset. This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01).

References

[1] Barrios, M. P., Velilla, S., 2007. A bootstrap method for assessing the dimension of a general regression problem. Statistics & probability letters 77 (3), 247–255.

[2] Bernard-Michel, C., Gardes, L., Girard, S., 2009. Gaussian regularized sliced inverse regression. Statistics and Computing 19 (1), 85–98.

[3] Chavent, M., Girard, S., Kuentz-Simonet, V., Liquet, B., Nguyen, T. M. N., Saracco, J., 2014. A sliced inverse regression approach for data stream. Computational Statistics 29 (5), 1129–1152.

[4] Chavent, M., Kuentz, V., Liquet, B., Saracco, J., 2011. A sliced inverse regression approach for a stratified population. Communications in Statistics-Theory and Methods 40 (21), 3857–3878.

[5] Chen, C.-H., Li, K.-C., 1998. Can sir be as popular as multiple linear regression? Statistica Sinica 8 (2), 289–316.

[6] Chiaromonte, F., Martinelli, J., 2002. Dimension reduction strategies for analyzing global gene expression data with a response. Mathematical Biosciences 176 (1), 123–144.

[7] Cook, R. D., Weisberg, S., 2009. Applied regression including computing and graphics. Vol. 488. John Wiley & Sons.

[8] Coudret, R., Girard, S., Saracco, J., 2014. A new sliced inverse regression method for multivariate response. Computational Statistics & Data Analysis 77, 285–299.

[9] Diaconis, P., Freedman, D., 1984. Asymptotics of graphical projection pursuit. The Annals of Statistics 12 (3), 793–815.

[10] Duan, N., Li, K.-C., 1991. Slicing regression: a link-free regression method. The Annals of Statistics 19 (2), 505–530.

[11] Fukunaga, K., 2013. Introduction to statistical pattern recognition. Academic press.

[12] Hall, P., Li, K.-C., 1993. On almost linearity of low dimensional projections from high dimensional data. The Annals of Statistics 21 (2), 867–889.

[13] Hsing, T., Carroll, R. J., 1992. An asymptotic theory for sliced inverse regression. The Annals of Statistics 20 (2), 1040–1061.

[14] Kuentz, V., Saracco, J., 2010. Cluster-based sliced inverse regression. Journal of the Korean Statistical Society 39 (2), 251–267.

[15] Li, K.-C., 1991. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association 86 (414), 316–327.

[16] Li, L., Cook, R. D., Nachtsheim, C. J., 2004. Cluster-based estimation for sufficient dimension reduction. Computational Statistics & Data Analysis 47 (1), 175–193.

[17] Li, L., Li, H., 2004. Dimension reduction methods for microarrays with application to censored survival data. Bioinformatics 20 (18), 3406–3412.

[18] Liquet, B., Saracco, J., 2012. A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches. Computational Statistics 27 (1), 103–125.

[19] Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2 (11), 559–572.

[20] Saracco, J., 1997. An asymptotic theory for sliced inverse regression. Communications in Statistics-Theory and Methods 26 (9), 2141–2171.

[21] Scrucca, L., 2006. Regularized sliced inverse regression with applications in classification. In: Data Analysis, Classification and the Forward Search. Springer, pp. 59–66.

[22] Scrucca, L., 2007. Class prediction and gene selection for dna microarrays using regularized sliced inverse regression. Computational Statistics & Data Analysis 52 (1), 438–451.

[23] Soltanolkotabi, M., Elhamifar, E., Candes, E. J., 2014. Robust subspace clustering. The Annals of Statistics 42 (2), 669–699.

[24] Van der Maaten, L. J., Postma, E. O., van den Herik, H. J., 2009. Dimensionality reduction: A comparative review. Journal of Machine Learning Research 10 (1-41), 66–71.

[25] Zhong, W., Zeng, P., Ma, P., Liu, J. S., Zhu, Y., 2005. Rsir: regularized sliced inverse regression for motif discovery. Bioinformatics 21 (22), 4169–4175.

[26] Zhu, L.-P., 2010. Extending the scope of inverse regression methods in sufficient dimension reduction. Communications in Statistics-Theory and Methods 40 (1), 84–95.

CHAPTER 3

STUDENT SIR

Four legs good, two legs better! All Animals Are Equal.
But Some Animals Are More Equal Than Others.
G. Orwell.

Student SIR has been accepter for publication in Computational Statistics and Data Analysis - Special issue on Robust Analysis of Complex Data.

3.1 Overall Idea

To give an intuitive idea of what Student SIR is meant for, the example of the bombers in the Introduction will be considered. Suppose that $\mathbf{X} = (x_1, x_2, ..., x_5)$ is the area hit by bullets for different aircrafts in five corresponding continuos variables (as in Wald's paper each aircraft is divided in five). Our goal is to predict $Y \in [0,1]$, the damage of the aircraft, 0 is undamaged and 1 is downed. The presence of outliers always brings problems in the estimation of statistical parameters (e.g. covariance matrix). Like PCA, SIR makes no exception, the presence of outliers affects the estimation of β in model (1.4). Two approaches are common in this area: identify and remove the outliers before the analysis or downweight their importance. Student SIR takes the second option. The following episode motivates the use of Student SIR to analyze bombers in action showing the possible presence of outliers:

On Dec. 20, 1943, a young American named Charles "Charlie" Brown was on his first World War II mission. Flying in the German skies, Brown's B-17 bomber was shot

and badly damaged and the crew was helpless: one could not walk, one could not use his hands, one with a leg bone off and one dead. In such desperate time a Luftwaffe ace Franz Stigler appeared with his fighter. All was lost. But Franz Stigler could not shoot. He escorted the B-17 on the border in direction of Sweden. When Charles decided to try to make it to England Franz gave a wave salute and left.

"Have you ever seen a bomber so severely damaged?" Has been asked in 1997 in an interview to Franz Stigler: "Not flying". This bomber with respect to our analysis can be, with no doubt, considered an outlier. The government decided to classify this episode because it showed the humanity of the enemy. Charles Brown managed to find Stigler several years later. They became close friends, and remained so, until their deaths within several months of each other in 2008. Student SIR downweights the importance of outliers during the estimation of the parameters, importance that in other fields must be enhanced and brought as an example.

FIGURE 3.1. The crew of "Ye Olde Pub." Kneeling L-R: Charlie Brown, Spencer Luke, Al Sadok, and Robert Andrews. Standing L-R: "Frenchy" Coulombe, Alex Yelesanko, Richard Pechout, Lloyd Jennings, Hugh Eckenrode, and Sam Blackford. PHOTO COURTESY ADAM MAKOS.

Student Sliced Inverse Regression

Alessandro Chiancone[a,b,c,*], Florence Forbes[a], Stéphane Girard[a]

[a] *Inria Grenoble Rhône-Alpes & LJK, team Mistis, 655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France.*
[b] *GIPSA-Lab, Grenoble INP, Saint Martin d'Hères, France.*
[c] *Institute of Statistics, Graz University of Technology, Kopernikusgasse 24/III, A-8010 Graz, Austria.*

Abstract

Sliced Inverse Regression (SIR) has been extensively used to reduce the dimension of the predictor space before performing regression. SIR is originally a model free method but it has been shown to actually correspond to the maximum likelihood of an inverse regression model with Gaussian errors. This intrinsic Gaussianity of standard SIR may explain its high sensitivity to outliers as observed in a number of studies. To improve robustness, the inverse regression formulation of SIR is therefore extended to non-Gaussian errors with heavy-tailed distributions. Considering Student distributed errors it is shown that the inverse regression remains tractable via an Expectation-Maximization (EM) algorithm. The algorithm is outlined and tested in the presence of outliers, both in simulated and real data, showing improved results in comparison to a number of other existing approaches.

Keywords: Dimension reduction, Inverse regression, Outliers, Robust estimation, Generalized Student distribution.

*Corresponding Author
Email address: al.chiancone@gmail.com (Alessandro Chiancone)

1. Introduction

Let us consider a regression setting where the goal is to estimate the relationship between a univariate response variable Y and a predictor $\mathbf{X}$. When the dimension p of the predictor space is 1 or 2, a simple 2D or 3D plot can visually reveal the relationship and can be useful to determine the regression strategy to be used. If p becomes large such an approach is not feasible. A possibility to overcome problems arising in the context of regression is to make the assumption that the response variable does not depend on the whole predictor space but just on a projection of $\mathbf{X}$ onto a subspace of smaller dimension. Such a dimensionality reduction leads to the concept of sufficient dimension reduction and to that of central subspace [1]. The central subspace is the intersection of all dimension-reduction subspaces (d.r.s.). A subspace S is a d.r.s. if Y is independent of $\mathbf{X}$ given $\mathbf{P}_S\mathbf{X}$, where $\mathbf{P}_S$ is the orthogonal projection onto S. In other words, all the information carried by the predictors $\mathbf{X}$ on Y can be compressed in $\mathbf{P}_S\mathbf{X}$. It has been shown under weak assumptions that the intersection of all d.r.s., and therefore the central subspace, is itself a d.r.s. [2]. It is of particular interest to develop methods to estimate the central subspace as once it is identified, the regression problem can be solved equivalently using the lower-dimensional representation $\mathbf{P}_S\mathbf{X}$ of $\mathbf{X}$ in the subspace.

Among methods that lead to an estimation of the central subspace, Sliced Inverse Regression (SIR) [3] is one of the most popular. SIR is a semiparametric method assuming that the link function depends on d linear combinations of the predictors and a random error independent of $\mathbf{X}$: $Y = f(\beta_1^T\mathbf{X}, \ldots, \beta_d^T\mathbf{X}, \epsilon)$. When this model holds, the projection of $\mathbf{X}$ onto the space spanned by the vectors $\{\beta_i, i = 1, \ldots, d\}$ captures all the information about Y. In addition, [3] shows that a basis of this space can be recovered using an inverse regression strategy provided that the so called *linearity condition* holds. It has been shown that the *linearity condition* is satisfied as soon as $\mathbf{X}$ is elliptically distributed. Moreover, this condition approximately holds in high-dimensional datasets, see [4]. However, solutions have been

proposed to deal with non elliptical distributed predictors and to overcome the *linearity condition* limitation [5, 6, 7].

The inverse regression approach to dimensionality reduction gained then rapid attention [8] and was generalized in [9] which shows the link between the axes spanning the central subspace and an inverse regression problem with Gaussian distributed errors. More specifically, in [10, 9], it appears that, for a Gaussian error term and under appropriate conditions, the SIR estimator can be recovered as the maximum likelihood estimator of the parameters of an inverse regression model. In other words, although SIR is originally a model free method, the standard SIR estimates are shown to correspond to maximum likelihood estimators for a Gaussian inverse regression model. It is then not surprising that SIR has been observed, *e.g.* in [11], to be at best under normality and that its performance may degrade otherwise. Indeed, the Gaussian distribution is known to have tails too light to properly accommodate extreme values. In particular, [12] observes that SIR was highly sensitive to outliers, with additional studies, evidence and analysis given in [13]. To downweight this sensitivity, robust versions of SIR have been proposed, mainly starting from the standard *model free* estimators and trying to make them more resistant to outliers. Typically, in [14] classical estimators are replaced by high breakdown robust estimators and, recently in [15] two approaches are built: a weighted version of SIR and a solution based on the intra slice multivariate median estimator.

As an alternative, we propose to rather exploit the inverse regression formulation of SIR [10, 9]. A new error term modeled by a multivariate Student distribution [16] is introduced. Among the elliptically contoured distributions, the multivariate Student is a natural generalization of the multivariate Gaussian but its heavy tails can better accommodate outliers. The result in Proposition 6 of [9] is extended from Gaussian to Student errors showing that the inverse regression approach of SIR is still valid outside the Gaussian case, meaning that the central subspace can still be estimated by maximum likelihood estimation of the inverse regression parameters. It is then shown that

the computation of the maximum likelihood estimators remains tractable in the Student case via an Expectation-Maximization (EM) algorithm which has a simple implementation and desirable properties.

The paper is organized as follows. In Section 2 general properties of the multivariate Student distribution and some of its variants are first recalled. The inverse regression model is introduced in Section 3 followed by the EM strategy to find the maximum likelihood estimator, the link with SIR and the resulting Student SIR algorithm. A simulation study is carried out in Section 4 and a real data application, showing the interest of this technique, is detailed in Section 5. The final section contains concluding remarks and perspectives. Proofs are postponed to the Appendix.

2. Multivariate generalized Student distributions

Multivariate Student, also called t-distributions, are useful when dealing with real-data because of their heavy tails. They are a robust alternative to the Gaussian distribution, which is known to be very sensitive to outliers. In contrast to the Gaussian case though, no closed-form solution exists for the maximum likelihood estimation of the parameters of the t-distribution. Tractability is, however, maintained both in the univariate and multivariate case, via the EM algorithm [17] and thanks to a useful representation of the t-distribution as a so-called *infinite mixture of scaled Gaussians* or *Gaussian scale mixture* [18]. A Gaussian scale mixture distribution has a probability density function of the form

$$P(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\psi}) = \int_0^\infty \mathcal{N}_p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}/u)\; f_U(u;\boldsymbol{\psi})\; \mathrm{d}u, \tag{1}$$

where $\mathcal{N}_p(\,.\,;\boldsymbol{\mu},\boldsymbol{\Sigma}/u)$ denotes the density function of the p-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}/u$ and f_U is the probability distribution of a univariate positive variable U referred to hereafter as the weight variable. When f_U is a Gamma distribution $\mathcal{G}(\nu/2,\nu/2)$ where ν denotes the degrees of freedom, expression (1) leads to the standard p-dimensional t-distribution denoted by $t_p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ with parameters $\boldsymbol{\mu}$ (lo-

cation vector), $\mathbf{\Sigma}$ ($p \times p$ positive definite scale matrix) and ν (positive degrees of freedom parameter). Its density is given by

$$t_p(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{\Sigma}, \nu) = \int_0^\infty \mathcal{N}_p(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{\Sigma}/u)\ \mathcal{G}(u; \nu/2, \nu/2)\ \mathrm{d}u$$
$$= \frac{\Gamma((\nu+p)/2)}{|\mathbf{\Sigma}|^{1/2}\ \Gamma(\nu/2)\ (\pi\nu)^{p/2}}\ [1 + \delta(\boldsymbol{x}, \boldsymbol{\mu}, \mathbf{\Sigma})/\nu]^{-(\nu+p)/2}, \quad (2)$$

where $\delta(\boldsymbol{x}, \boldsymbol{\mu}, \mathbf{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ is the Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$. The Gamma distribution has probability density function $\mathcal{G}(u; \alpha, \gamma) = u^{\alpha-1}\Gamma(\alpha)^{-1} \exp(-\gamma u)\gamma^\alpha$, where Γ denotes the Gamma function.

If $f_U(u; \boldsymbol{\psi})$ is set equal to a Gamma distribution $\mathcal{G}(\alpha, \gamma)$ without imposing $\alpha = \gamma$, (1) results in a multivariate Pearson type VII distribution (see *e.g.* [19] vol.2 chap. 28) also referred to as the Arellano-Valle and Bolfarine's Generalized t distribution in [16]. This generalized version is the multivariate version of the t-distribution considered in this work, its density is given by:

$$\mathcal{S}_p(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{\Sigma}, \alpha, \gamma) = \int_0^\infty \mathcal{N}_p(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{\Sigma}/u)\ \mathcal{G}(u; \alpha, \gamma)\ \mathrm{d}u \quad (3)$$
$$= \frac{\Gamma(\alpha + p/2)}{|\mathbf{\Sigma}|^{1/2}\ \Gamma(\alpha)\ (2\pi\gamma)^{p/2}}\ [1 + \delta(\boldsymbol{x}, \boldsymbol{\mu}, \mathbf{\Sigma})/(2\gamma)]^{-(\alpha+p/2)}\ . \quad (4)$$

For a random variable $\mathbf{X}$ following distribution (4), an equivalent representation useful for simulation is $\mathbf{X} = \boldsymbol{\mu} + U^{-1/2}\tilde{\mathbf{X}}$ where U follows a $\mathcal{G}(\alpha, \gamma)$ distribution and $\tilde{\mathbf{X}}$ follows a $\mathcal{N}(0, \mathbf{\Sigma})$ distribution.

Remark 1 (Identifiability). *The expression* (4) *depends on* γ *and* $\mathbf{\Sigma}$ *only through the product* $\gamma\mathbf{\Sigma}$ *which means that to make the parameterization unique, an additional constraint is required. One possibility is to impose that* $\mathbf{\Sigma}$ *is of determinant 1. It is easy to see that this is equivalent to have an unconstrained* $\mathbf{\Sigma}$ *with* $\gamma = 1$.

Unconstrained parameters are easier to deal with in inference algorithms. Therefore, we will rather assume without loss of generality that $\gamma = 1$ with the notation $\mathcal{S}_p(0, \mathbf{V}, \alpha, 1) \equiv \mathcal{S}_p(0, \mathbf{V}, \alpha)$ adopted in the next Section.

3. Student Sliced Inverse Regression

Let $\mathbf{X} \in \mathbb{R}^p$ be a random vector, $Y \in \mathbb{R}$ the real response variable and $S_{Y|X}$ the central subspace spanned by the columns of the matrix $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$. In the following, it is assumed that $\dim(S_{Y|X}) = d$ where d is known and $d \leq p$. To address the estimation of the central subspace, we consider the inverse regression formulation of [9], which models the link from Y to $\mathbf{X}$. In addition to be a simpler regression problem, the inverse regression approach is of great interest because Proposition 6 in [9] states that in the Gaussian case, an estimation of the central subspace is provided by the estimation of the inverse regression parameters. In Subsection 3.1, the inverse regression model of [9] is extended by considering Student distributed errors. It is then shown in Subsection 3.2 that the estimation of the extended model is tractable via an Expectation-Maximization algorithm (EM). A link with SIR is presented in Subsection 3.3 and the resulting Student SIR algorithm is described in Subsection 3.4.

3.1. Student multi-index inverse regression model

In the spirit of [9, 10] the following regression model is considered

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{VBc}(Y) + \boldsymbol{\varepsilon}, \tag{5}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is a non random vector, $\mathbf{B}$ is a non random $p \times d$ matrix with $\mathbf{B}^T\mathbf{B} = \mathbf{I}_d$, $\boldsymbol{\varepsilon} \in \mathbb{R}^p$ is a centered generalized Student random vector following the distribution given in (4), $\boldsymbol{\varepsilon}$ is assumed independent of Y, with scale matrix $\mathbf{V}$, $\mathbf{c} : \mathbb{R} \to \mathbb{R}^d$ is a non random function. It directly follows from (5) that

$$\mathbb{E}(\mathbf{X}|Y = y) = \boldsymbol{\mu} + \mathbf{VBc}(y), \tag{6}$$

and thus, after translation by $\boldsymbol{\mu}$, the conditional expectation of $\mathbf{X}$ given Y is a random vector located in the space spanned by the columns of $\mathbf{VB}$. When $\boldsymbol{\varepsilon}$ is assumed to be Gaussian distributed, Proposition 6 in [9] states that $\mathbf{B}$ corresponds to the directions of the central subspace $\boldsymbol{\beta}$. In [9, 10], it appears then that, under appropriate conditions, the maximum likelihood

estimator of $\mathbf{B}$ is (up to a full rank linear transformation) the SIR estimator of $\boldsymbol{\beta}$, *i.e.* Span$\{\mathbf{B}\}$ = Span$\{\boldsymbol{\beta}\}$. Proposition 6 in [9] can be generalized to our Student setting, so that $\mathbf{B}$ still corresponds to the central subspace. The generalization of Proposition 6 of [9] is given below.

Proposition 1. *Let $\mathbf{X}_y$ be a random variable distributed as $\mathbf{X}|Y = y$, let us assume that*

$$\mathbf{X}_y = \boldsymbol{\mu} + \mathbf{VBc}(y) + \boldsymbol{\varepsilon}, \tag{7}$$

with $\boldsymbol{\varepsilon}$ following a generalized Student distribution $\mathcal{S}_p(0, \mathbf{V}, \alpha)$, $\mathbf{c}(y) \in \mathbb{R}^d$ is function of y and $\mathbf{VB}$ is a $p \times d$ matrix of rank d. Under model (7), the distribution of $Y|\mathbf{X} = \mathbf{x}$ is the same as the distribution of $Y|\mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x}$ for all values $\mathbf{x}$.

The proof is given in Appendix 7.1. According to this proposition, $\mathbf{X}$ can be replaced by $\mathbf{B}^T\mathbf{X}$ without loss of information on the regression of Y on $\mathbf{X}$. A procedure to estimate $\mathbf{B}$ is then proposed in the next Section

3.2. Maximum likelihood estimation via EM algorithm

Let $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$ be a set of independent random variables distributed according to the distribution of $(\mathbf{X}, Y)$ as defined in (5). The unknown quantities to be estimated in model (5) are $\{\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \alpha\}$ and the function $\mathbf{c}(.)$. Regarding $\mathbf{c}$, we focus on projection estimators for each coordinate of $\mathbf{c}(.) = (c_1(.), \ldots, c_d(.))$. For $k = 1, \ldots, d$, function $c_k(.)$ is expanded as a linear combination of h basis functions $s_j(.)$, $j = 1, \ldots, h$ as

$$c_k(.) = \sum_{j=1}^{h} c_{jk} s_j(.), \tag{8}$$

where the coefficients c_{jk}, $j = 1, \ldots, h$ and $k = 1, \ldots, d$ are unknown and to be estimated while h is supposed to be known. Let $\mathbf{C}$ be a $h \times d$ matrix with the kth column given by $(c_{1k}, \ldots, c_{hk})^T$ and $\mathbf{s}(.) = (s_1(.), \ldots, s_h(.))^T$. Then, model (5) can be rewritten as

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{VBC}^T\mathbf{s}(Y) + \boldsymbol{\varepsilon}, \qquad \text{with } \boldsymbol{\varepsilon} \sim \mathcal{S}_p(0, \mathbf{V}, \alpha), \tag{9}$$

where $\mathcal{S}_p(0, \mathbf{V}, \alpha)$ is the multivariate centered generalized Student distribution with scale matrix $\mathbf{V}$. For each i, it follows that conditionally to Y_i, $\mathbf{X}_i \sim \mathcal{S}_p(\boldsymbol{\mu} + \mathbf{VBC}^T\mathbf{s}_i, \mathbf{V}, \alpha)$ where $\mathbf{s}_i$=$\mathbf{s}(Y_i)$. The density of the generalized Student distribution is available in closed form and given in (4). However to perform the estimation, a more useful representation of this distribution is given by its Gaussian scale mixture representation (3). Introducing an additional set of latent variables $\mathbf{U} = \{U_1, \ldots, U_n\}$ with U_i independent of Y_i, one can equivalently write:

$$\mathbf{X}_i | U_i = u_i, Y_i = y_i \sim \mathcal{N}_p(\boldsymbol{\mu} + \mathbf{VBC}^T\mathbf{s}_i, \mathbf{V}/u_i), \tag{10}$$

$$U_i | Y_i = y_i \sim \mathcal{G}(\alpha, 1). \tag{11}$$

Let us denote by $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C}, \alpha\}$ the parameters to estimate from realizations $\{\mathbf{x}_i, y_i, i = 1, \ldots, n\}$. In contrast to the Gaussian case, the maximum likelihood estimates are not available in closed-form for the t-distributions. However, they are reachable using an Expectation-Maximization (EM) algorithm. More specifically, at iteration (t) of the algorithm, $\boldsymbol{\theta}$ is updated from a current value $\boldsymbol{\theta}^{(t-1)}$ to a new value $\boldsymbol{\theta}^{(t)}$ defined as $\boldsymbol{\theta}^{(t)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)})$. Considering the scale mixture representation above, a natural choice for Q is the following expected value of the complete log-likelihood:

$$\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) &= E_U[\sum_{i=1}^{n} \log P(\mathbf{x}_i, U_i | Y_i = y_i; \boldsymbol{\theta}) | \mathbf{X}_i = \mathbf{x}_i, Y_i = y_i; \boldsymbol{\theta}^{(t-1)}] \qquad (12)\\
&= \sum_{i=1}^{n} E_{U_i}[\log P(\mathbf{x}_i | U_i, y_i; \boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C}) | \mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}] + E_{U_i}[\log P(U_i; \alpha) | \mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}]\\
&= -\frac{1}{2} n \log\det \mathbf{V} + \frac{1}{2} p \sum_{i=1}^{n} E_{U_i}[\log(U_i) | \mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}]\\
&- \frac{1}{2} \sum_{i=1}^{n} E_{U_i}[U_i | \mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}] \; (\boldsymbol{\mu} + \mathbf{VBC}^T\mathbf{s}_i - \mathbf{x}_i)^T \mathbf{V}^{-1} (\boldsymbol{\mu} + \mathbf{VBC}^T\mathbf{s}_i - \mathbf{x}_i)\\
&+ \sum_{i=1}^{n} E_{U_i}[\log P(U_i; \alpha) | \mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}] \; .
\end{aligned}$$

Note that all computations are conditionally to the Y_i's and no assumption

is made on the distribution of the Y_i's. The E-step therefore consists of computing the quantities

$$\bar{u_i}^{(t)} = E_{U_i}[U_i|\mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}] \,, \tag{13}$$

$$\tilde{u_i}^{(t)} = E_{U_i}[\log U_i|\mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}] \,, \tag{14}$$

while the M-step divides into two-independent M-steps involving separately parameters $(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C})$ and α. The second quantity (14) is needed only in the estimation of α. The following notation is introduced for the next sections:

$$\bar{u}^{(t)} = \frac{\sum_{i=1}^n \bar{u_i}^{(t)}}{n} \tag{15}$$

$$\tilde{u}^{(t)} = \frac{\sum_{i=1}^n \tilde{u_i}^{(t)}}{n} \,. \tag{16}$$

E-step. The quantities (13) and (14) above require the posterior distribution of the U_i's. This distribution can be easily determined using the well known conjugacy of the Gamma and Gaussian distributions for the mean. It follows then from standard Bayesian computations that the posterior distribution is still a Gamma distribution with parameters specified below,

$$\begin{aligned}
&p(u_i|\mathbf{X}_i = \mathbf{x}_i, Y_i = y_i; \boldsymbol{\theta}^{(t-1)}) \\
&\quad \propto \mathcal{N}_p(\mathbf{x}_i; \boldsymbol{\mu}^{(t-1)} + \mathbf{V}^{(t-1)}\mathbf{B}^{(t-1)}\mathbf{C}^{(t-1)T}\mathbf{s}_i, \mathbf{V}^{(t-1)}/u_i)\ \mathcal{G}(u_i; \alpha^{(t-1)}, 1) \\
&\quad = \mathcal{G}(u_i; \alpha^{(t-1)} + \frac{p}{2}, 1 + \frac{1}{2}\delta(\mathbf{x}_i, \mu^{(t-1)} + \mathbf{V}^{(t-1)}\mathbf{B}^{(t-1)}\mathbf{C}^{(t-1)T}\mathbf{s}_i, \mathbf{V}^{(t-1)})),
\end{aligned}$$

where $\delta(\mathbf{x}_i, \boldsymbol{\mu}+\mathbf{VBC}^T\mathbf{s}_i, \mathbf{V}) = (\boldsymbol{\mu}+\mathbf{VBC}^T\mathbf{s}_i-\mathbf{x}_i)^T\mathbf{V}^{-1}(\boldsymbol{\mu}+\mathbf{VBC}^T\mathbf{s}_i-\mathbf{x}_i)$ is the Mahalanobis distance between $\mathbf{x}_i$ and $\boldsymbol{\mu} + \mathbf{VBC}^T\mathbf{s}_i$ when the covariance is $\mathbf{V}$.

The required moments (13) and (14) are then well known for a Gamma distribution, so that it comes,

$$\bar{u}_i^{(t)} = \frac{\alpha^{(t-1)} + \frac{p}{2}}{1 + \frac{1}{2}\delta(\mathbf{x}_i, \mu^{(t-1)} + \mathbf{V}^{(t-1)}\mathbf{B}^{(t-1)}\mathbf{C}^{(t-1)T}\mathbf{s}_i, \mathbf{V}^{(t-1)})} \text{ and}$$

$$\tilde{u}_i^{(t)} = \Psi(\alpha^{(t-1)} + \frac{p}{2}) - \log(1 + \frac{1}{2}\delta(\mathbf{x}_i, \boldsymbol{\mu}^{(t-1)} + \mathbf{V}^{(t-1)}\mathbf{B}^{(t-1)}\mathbf{C}^{(t-1)T}\mathbf{s}_i, \mathbf{V}^{(t-1)})) \,,$$

where Ψ is the Digamma function. As it will become clear in the following M-step, $\bar{u}_i^{(t)}$ acts as a weight for $\mathbf{x}_i$. Whenever the Mahalanobis distance of $\mathbf{x}_i$ to $\boldsymbol{\mu}^{(t-1)} + \mathbf{V}^{(t-1)}\mathbf{B}^{(t-1)}\mathbf{C}^{(t-1)T}\mathbf{s}_i$ increases, the weight $\bar{u}_i^{(t)}$ of $\mathbf{x}_i$ decreases and the influence of $\mathbf{x}_i$ in the estimation of the parameters will be downweighted in the next iteration. The idea of using weights to handle outliers is common in the literature, Weighted Inverse Regression (WIRE) [15] gives weights through a deterministic kernel function to ensure the existence of the first moment. Our approach does not require previous knowledge to select an appropriate kernel and refers to the wide range of t-distributions (the Cauchy distribution for which the first moment is not defined lies in this family).

M- step. The M-step divides into the following two independent sub-steps. **M-$(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C})$ substep.** Omitting terms that do not depend on the parameters in (12), estimating $(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C})$ by maximization of Q consists, at iteration (t), of minimizing with respect to $(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C})$ the following G function,

$$G(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C}) = \log\det \mathbf{V} + \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}^{(t)} \ (\boldsymbol{\mu} + \mathbf{V}\mathbf{B}\mathbf{C}^T\mathbf{s}_i - \mathbf{x}_i)^T\mathbf{V}^{-1}(\boldsymbol{\mu} + \mathbf{V}\mathbf{B}\mathbf{C}^T\mathbf{s}_i - \mathbf{x}_i) \ . \quad (17)$$

To this aim, let us introduce (omitting the index iteration (t) in the notation) the $h \times h$ weighted covariance matrix $\mathbf{W}$ of $\mathbf{s}(Y)$ defined by:

$$\mathbf{W} = \frac{1}{n}\sum_{i=1}^{n} \bar{u_i} \ (\mathbf{s}_i - \bar{\mathbf{s}})(\mathbf{s}_i - \bar{\mathbf{s}})^T,$$

the $h \times p$ weighted covariance matrix $\mathbf{M}$ of $(\mathbf{s}, \mathbf{X})$ defined by

$$\mathbf{M} = \frac{1}{n}\sum_{i=1}^{n} \bar{u_i} \ (\mathbf{s}_i - \bar{\mathbf{s}})(\mathbf{x}_i - \bar{\mathbf{x}})^T,$$

and $\boldsymbol{\Sigma}$ the $p \times p$ weighted covariance matrix of $\mathbf{X}$

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \bar{u_i} \ (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T, \quad (18)$$

where

$$\bar{\mathbf{x}} = \frac{1}{\sum_{i=1}^{n} \bar{u_i}} \sum_{i=1}^{n} \bar{u_i} \mathbf{x}_i \text{ and} \tag{19}$$

$$\bar{\mathbf{s}} = \frac{1}{\sum_{i=1}^{n} \bar{u_i}} \sum_{i=1}^{n} \bar{u_i} \mathbf{s}_i. \tag{20}$$

We derive then the following lemma.

Lemma 1. *Using the above notations,* $G(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C})$ *can be rewritten as*

$$G(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C}) = \log\det \mathbf{V} + \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{V}^{-1}) + \mathrm{tr}(\mathbf{C}^T\mathbf{W}\mathbf{C}\mathbf{B}^T\mathbf{V}\mathbf{B}) - 2\mathrm{tr}(\mathbf{C}^T\mathbf{M}\mathbf{B}) + \bar{u}\ (\boldsymbol{\mu} - \bar{\mathbf{x}} + \mathbf{V}\mathbf{B}\mathbf{C}^T\bar{\mathbf{s}})^T\mathbf{V}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}} + \mathbf{V}\mathbf{B}\mathbf{C}^T\bar{\mathbf{s}})\ .$$

The proof is given in Appendix 7.2. Thanks to this representation of $G(.)$ it is possible to derive the following proposition which is a generalization to the multi-index case and Student setting of the result obtained in case of Gaussian error ϵ in [10].

Proposition 2. *Under (9), if* $\mathbf{W}$ *and* $\boldsymbol{\Sigma}$ *are regular, then the M-step for* $(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C})$ *leads to the updated estimations* $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}}, \hat{\mathbf{B}}, \hat{\mathbf{C}})$ *given below*

- $\hat{\mathbf{B}}$ *is made of the eigenvectors associated to the largest eigenvalues of* $\boldsymbol{\Sigma}^{-1}\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}$*,*
- $\hat{\mathbf{V}} = \boldsymbol{\Sigma} - (\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}\mathbf{B})(\mathbf{B}^T\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}\mathbf{B})^{-1}(\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}\mathbf{B})^T$*,*
- $\hat{\mathbf{C}} = \mathbf{W}^{-1}\mathbf{M}\hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}})^{-1}$ *and*
- $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} - \hat{\mathbf{V}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T\bar{\mathbf{s}}$*.*

The proof is detailed in Appendix 7.3. Regarding parameter α it can be updated using an independent part of Q as detailed in the next M-step.

M-α substep.

Parameter α can be estimated by maximizing independently with regards to α,

$$\sum_{i=1}^{n} E_{U_i}[\log P(U_i;\alpha)|\mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}] \ . \tag{21}$$

Then, since

$$E_{U_i}[\log p(U_i;\alpha)|\mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}] = -\bar{u_i}^{(t)} + (\alpha-1)\tilde{u_i}^{(t)} - \log\Gamma(\alpha) \ , \tag{22}$$

setting the derivative with respect to α to zero, we obtain that $\hat{\alpha} = \Psi^{-1}(\tilde{u})$, where $\Psi(.)$ is the Digamma function.

In practice, for the procedure to be complete, the choice of the h basis functions s_j needs to be specified. Many possibilities for basis functions are available in the literature such as classical Fourier series, polynomials, etc. In the next section, we discuss a choice of basis functions which provides the connection with Sliced Inverse Regression (SIR) [3].

3.3. Connection to Sliced Inverse Regression

As in the Gaussian case [9, 10], a clear connection with SIR can be established for a specific choice of the h basis functions. When Y is univariate a natural approach is to first partition the range of Y into $h+1$ bins S_j for $j = 1, \ldots, h+1$ also referred to as slices, and then defining h basis functions by considering the first h slices as follows,

$$s_j(.) = \mathbb{1}\{. \in S_j\}, \quad j = 1, \ldots, h, \tag{23}$$

where $\mathbb{1}$ is the indicator function. Note that it is important to remove one of the slices so that the basis functions remain independent. However, the following related quantities are defined for $j = 1, \ldots, h+1$:

$$\begin{aligned} n_j &= \sum_{i=1}^{n} \bar{u}_i \mathbb{1}\{y_i \in S_j\}, \\ f_j &= \frac{n_j}{n} \ . \end{aligned} \tag{24}$$

They represent respectively the number of y_i in slice j weighted by the $\bar{u}_i$ and the weighted proportion in slice j. The following weighted mean of $\mathbf{X}$ given $Y \in S_j$ is then denoted by

$$\bar{\mathbf{x}}_j = \frac{1}{n_j} \sum_{i=1}^{n} \bar{u}_i \mathbb{1}\{y_i \in S_j\} \mathbf{x}_i, \tag{25}$$

and the $p \times p$ "between slices" covariance matrix by

$$\mathbf{\Gamma} = \sum_{j=1}^{h+1} f_j (\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T.$$

In this context, the following consequence of Proposition 2 can be established.

Corollary 1. *Under (9) and (23), if* $\mathbf{\Sigma}$ *is regular, then the updated estimation* $\hat{\mathbf{B}}$ *of* $\mathbf{B}$ *is given by the eigenvectors associated to the largest eigenvalues of* $\mathbf{\Sigma}^{-1}\mathbf{\Gamma}$. *In addition,* $\mathbf{\Gamma} = \mathbf{M}^T \mathbf{W}^{-1} \mathbf{M}$.

The proof is given in Appendix 7.4. When all $\bar{u}_i = 1$, the iterative EM algorithm reduces to one M-step and the quantities defined in this section correspond to the standard SIR estimators. The EM algorithm resulting from this choice of basis functions is referred to as the Student SIR algorithm. It is outlined in the next section.

3.4. Central subspace estimation via Student SIR algorithm

The EM algorithm can be outlined using Proposition 2 and Corollary 1. It relies on two additional features to be specified, initialization and stopping rule. As the algorithm alternates the E and M steps, it is equivalent to start with one of this step. It is convenient to start with the Maximization step since the initialization of quantities $\bar{u}_i, \tilde{u}_i$ can be better interpreted. If $\bar{u}_i$ is constant and $\tilde{u}_i = 0$, the first M-step of the algorithm results in performing standard SIR. Regarding an appropriate stopping rule of the algorithm, EM's fundamental property is to increase the log-likelihood at each iteration. A standard criteria is then the relative increase in log-likelihood, denoted by

$\Delta(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t-1)})$, between two iterations. At each iteration, for current parameters values, the log-likelihood is easy to compute using (4) and (9). Another natural criterion is to assess when parameter estimation stabilizes. Typically, focusing on the central subspace $\mathbf{B}$, the following proximity measure [20, 21] can be considered:

$$r(\mathbf{B}, \hat{\mathbf{B}}) = \frac{trace(\mathbf{B}\mathbf{B}^T\hat{\mathbf{B}}\hat{\mathbf{B}}^T)}{d}. \tag{26}$$

The above quantity r ranges from 0 to 1 and evaluates the distance between the subspaces spanned by the columns of $\mathbf{B}$ and $\hat{\mathbf{B}}$. If $d = 1$, r is the squared cosine between the two spanning vectors. Although not directly related to the EM algorithm, in practice this criterion gave similar results in terms of parameter estimation. Experiments on simulated and real data are reported in the next two sections.

3.5. Determination of the central subspace dimension

Determining the dimension d of the central subspace is an important issue for which different solutions have been proposed in the literature. Most users rely on graphical considerations, *e.g.* [22]. A more quantitative approach is to use cross validation after the link function is found. Although in that case, d may vary depending on the specific regression approach that the user selected. Other methods that can be easily used on real data, are mainly based on (sequential) tests [3, 23, 24, 20, 25, 11]. An alternative that uses a penalized likelihood criterion has been proposed in [26]. In our setting, formulated as a maximum likelihood problem, the penalized likelihood approach is the most natural. For a given value d of the central subspace dimension, we therefore propose to compute the Bayesian information criterion [27] defined as $BIC(d) = -2L(d) + \eta \log n$, where $\eta = \frac{p(p+3)}{2} + 1 + \frac{d(2p-d-1+2h)}{2}$ is the number of free parameters in the model and $L(d)$ is the maximized log-likelihood computed at the parameters values obtained at convergence of the EM algorithm. Computing $L(d)$ is a straightforward byproduct of the algorithm described above as this quantity is already used in our stopping

Algorithm 1 Student SIR algorithm

Set h and partition the Y range into $h+1$ slices.

Set the e.d.r. space dimension d and the desired tolerance value for convergence δ.

Initialize the $\bar{u}_i^{(0)}, \tilde{u}_i^{(0)}$'s with $\bar{u}_i^{(0)} = 1$ and $\tilde{u}_i^{(0)} = 0$ for all $i = 1, \ldots, n$

(this first iteration of the algorithm gives the SIR estimation of Γ and $\mathbf{B}$).

while $\Delta(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t-1)}) < \delta$ **do**

M-step

Compute:

- $\bar{u}^{(t)}$ and $\tilde{u}^{(t)}$ (eq. (15) and (16)), $f^{(t)} = (f_1^{(t)}, \ldots f_h^{(t)})^T$ and $f_{h+1}^{(t)}$ (eq. (24)),
- $\bar{\mathbf{x}}_j^{(t)}$ and $\bar{\mathbf{x}}^{(t)}$ (eq. (25) and (19)),
- $\boldsymbol{\Sigma}^{(t)}$ (eq. (18)),
- $\mathbf{M}^{(t)}$ where each row is given by $\mathbf{M}_{j,\cdot}^{(t)} = f_j^{(t)}(\bar{\mathbf{x}}_j^{(t)} - \bar{\mathbf{x}}^{(t)})^T$ for $j = 1, \ldots, h$,
- $\mathbf{W}^{(t)-1} = \text{diag}\left(\frac{1}{f_1^{(t)}}, \ldots, \frac{1}{f_h^{(t)}}\right) + \frac{1}{f_{h+1}^{(t)}}\mathbf{O}$, where $\mathbf{O}$ is the $h \times h$ matrix defined by $O_{ij} = 1$,
- $\boldsymbol{\Gamma}^{(t)} = \mathbf{M}^{(t)T}\mathbf{W}^{(t)-1}\mathbf{M}^{(t)}$,
- $\mathbf{B}^{(t)}$ matrix of the d eigenvectors associated to the d largest eigenvalues of $\boldsymbol{\Sigma}^{(t)-1}\boldsymbol{\Gamma}^{(t)}$,
- $\mathbf{V}^{(t)} = \boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Gamma}^{(t)}\mathbf{B}^{(t)}(\mathbf{B}^{(t)^T}\boldsymbol{\Gamma}^{(t)}\mathbf{B}^{(t)})^{-1}(\boldsymbol{\Gamma}^{(t)}\mathbf{B}^{(t)})^T$,
- $\mathbf{C}^{(t)} = \mathbf{W}^{(t)-1}\mathbf{M}^{(t)}\mathbf{B}^{(t)}(\mathbf{B}^{(t)T}\mathbf{V}^{(t)}\mathbf{B}^{(t)})^{-1}$,
- $\boldsymbol{\mu}^{(t)} = \bar{\mathbf{x}}^{(t)} - \mathbf{V}^{(t)}\mathbf{B}^{(t)}\mathbf{C}^{(t)T}\bar{\mathbf{s}}^{(t)}$,
- $\alpha^{(t)} = \Psi^{-1}(\tilde{u}^{(t)})$.

E-step

Update the $\bar{u}_i, \tilde{u}_i$'s using the quantities estimated in the M-step:

$$\bar{u_i}^{(t+1)} = \frac{\alpha^{(t)} + \frac{p}{2}}{1 + \frac{1}{2}\delta(\mathbf{x}_i, \boldsymbol{\mu}^{(t)} + \mathbf{V}^{(t)}\mathbf{B}^{(t)}\mathbf{C}^{(t)T}\mathbf{s}_i, \mathbf{V}^{(t)})} ,$$

$$\tilde{u_i}^{(t+1)} = \Psi(\alpha^{(t)} + \frac{p}{2}) - \log(1 + \frac{1}{2}\delta(\mathbf{x}_i, \boldsymbol{\mu}^{(t)} + \mathbf{V}^{(t)}\mathbf{B}^{(t)}\mathbf{C}^{(t)T}\mathbf{s}_i, \mathbf{V}^{(t)})) .$$

end while

criterion. Following the BIC principle, an estimator of d can then be defined as the minimizer of $BIC(d)$ over $d \in \{1, \ldots, min(p, h)\}$. The performance of this criterion is investigated in the simulation study in Section 4 and used on the real data example of Section 5. The simulation study reveals that BIC can provide correct selections but requires large enough sample sizes. This limitation has been already pointed out in the literature (see *e.g.* [28]).

4. Simulation study

Student SIR is tested on simulated data under a variety of different models and distributions for the p-dimensional random variable $\mathbf{X}$. The behavior of Student SIR is compared to SIR and four other techniques arising from the literature that claim some robustness. For comparison, the simulation setup described in [15, 14] is adopted.

4.1. Simulation setup

Three different regression models are considered:

I : $Y = 1 + 0.6X_1 - 0.4X_2 + 0.8X_3 + 0.2\varepsilon$,

II : $Y = (1 + 0.1\varepsilon)X_1$,

III : $Y = X_1/(0.5 + (X_2 + 1.5)^2)) + 0.2\varepsilon$,

where ε follows a standard normal distribution. The three models are combined with three possible distributions for the predictors $\mathbf{X}$:

(i) $\mathbf{X}$ is multivariate normal distributed with mean vector $\mathbf{0}$ and covariance matrix defined by its entries as $\sigma_{ij} = 0.5^{|i-j|}$;

(ii) $\mathbf{X}$ is standard multivariate Cauchy distributed;

(iii) $\mathbf{X} = (X_1, \ldots, X_p)^T$, where each X_i is generated independently from a mixture of normal and uniform distributions denoted by $0.8\mathcal{N}(0, 1) + 0.2\mathcal{U}(-\nu, \nu)$ where ν is a positive scalar value.

Models **I**, **III** are homoscedastic while model **II** is heteroscedastic. Case (**ii**) is built to test the sensitivity to outliers while the distribution of $\mathbf{X}$ is elliptical. In (**iii**) a non-elliptical distribution of $\mathbf{X}$ is considered. The dimension is set to $p = 10$, the dimension of the e.d.r. space is $d = 1$ for **I**, **II** and $d = 2$ for **III**. The nine different configurations of $\mathbf{X}$ and Y are simulated with a number of samples varying depending on the experiment. In all tables Student SIR is compared with standard SIR and four other approaches. Contour Projection (CP-SIR) [29, 30] applies the SIR procedure on a rescaled version of the predictors. Weighted Canonical Correlation (WCAN) [31] uses a basis of B-splines first estimating the dimension d of the central subspace and then the directions from the nonzero robustified version of the correlation matrices between the predictors and the B-splines basis functions. The idea of Weighted Inverse Regression (WIRE) [15] is to use a different weight function capable of dealing with both outliers and inliers. SIR is a particular case of WIRE with constant weighting function. Slice Inverse Median Estimation (SIME) replaces the intra slice mean estimator with the median which is well known to be more robust. All values referring to CP-SIR, WCAN, WIRE, SIME in the tables are directly extracted from [15]. Values relative to SIR have been recomputed using [32].

4.2. Results

To assess the sensitivity of the compared methods to different setting parameters, four sets of tests are carried out and reported respectively in Tables 1 and 2. First, the 9 configurations of $\mathbf{X}$ and Y models are tested for fixed sample size $n = 200$, number of slices $h = 5$ and $p = 10$ (Table 1 (a)). Then, the effect of the sample size is illustrated for model **I** (Table 1 (b)). The number of slices is varied to evaluate the sensitivity to the h value (Table reftb3 (a)) and at last, different values of ν are tested in the model (**iii**) case (Table 2 (b)). In all cases and tables, the different methods performance is assessed based on their ability to recover the central subspace which is measured via the value of the proximity measure r (26).

Student SIR shows its capability to deal with different configurations.

The proximity criterion (26) in Table 1 (a) is very close to one, for the first two regression models independently of the distributions of the predictors. In the Gaussian case, Student SIR and SIR are performing equally well showing that our approach has no undesirable effects when dealing with *simple* cases. For configuration $\mathbf{III}-(\mathbf{iii})$, a slightly different value has been found for SIR compared to [15]. In this configuration however the trend is clear: standard SIR, Student SIR, WIRE and SIME show similar performance. In contrast, configurations $\mathbf{I}-(\mathbf{ii}), \mathbf{II}-(\mathbf{ii}), \mathbf{III}-(\mathbf{ii})$ illustrate that Student SIR can significantly outperform SIR. This is not surprising since the standard multivariate Cauchy has heavy tails and SIR is sensitive to outliers [14].

Table 1 (b) illustrates on model **I** the effect of the sample size n: Student SIR exhibits the best performance among all methods. It is interesting to observe that, in case (**ii**), the smaller value of r for standard SIR does not depend on the sample size n. In contrast, adding observations results in a better estimation for Student SIR.

It is then known that SIR is not very sensitive to the number of slices h [22]. In Table 2 (a), an analysis is performed with varying h. Student SIR appears to be as well not very sensitive to the number of slices.

Extra inliers as well as outliers can affect the estimation. In case (**iii**), parameter ν is controlling the extra observations magnitude. Under different values of $\nu = 0.5, 0.2, 0.1, 0.05$,Table 2 (b) shows that both SIR and Student SIR are robust to inliers while CP-SIR and WCAN fail when ν is small and extra observations behave as inliers concentrated around the average.

In addition, a study on the behavior of SIR and Student SIR when $\mathbf{X}$ follows a standard multivariate Student distribution, with different degrees of freedom (df), is shown in Table 3 (a). The multivariate Cauchy of model (**ii**) coincides with the multivariate Student with one degree of freedom. This setting is favorable to our model which is designed to handle heavy tails. Not surprisingly, Student SIR provides better results for small degrees of freedom but the difference with SIR is reduced as the degree of freedom increases and the multivariate Student gets closer to a Gaussian. The stan-

dard deviation follows the same trend. In case $\mathbf{III}-(\mathbf{ii})$ the convergence of SIR becomes extremely slow. Regarding computational time, results are reported in Table 3 (b). Student SIR has multiple iterations, which increases computational time compared to SIR. It is interesting that, in the cases in which SIR fails $(\mathbf{I}-(\mathbf{ii}), \mathbf{II}-(\mathbf{ii}), \mathbf{III}-(\mathbf{ii})$ see Table 1), the convergence of Student SIR is fast, requiring less than a second on a standard laptop (Our Matlab code is available at https://hal.inria.fr/hal-01294982). All reported results have been obtained using a threshold of 0.01 for the relative increase of the Log-likelihood.

At last, the use of BIC as a selection criterion for the central subspace dimension d is investigated. As an illustration, last column of Table 3 (b) shows the number of times the criterion succeeded in selecting the correct dimension (*i.e.* $d = 2$ in this example) over 200 repetitions. BIC performs very well provided the sample size is large enough, this phenomenon being more critical as the number of outlying data increases. This is not surprising as this limitation of BIC has often been reported in the literature.

To summarize, through these simulations Student SIR shows good performance, outperforming SIR when the distribution of $\mathbf{X}$ is heavy-tailed (case $(\mathbf{ii})$) and preserving good properties such as insensitivity to the number of slices or robustness to inliers that are peculiar of SIR.

5. Real data application: The galaxy dataset

5.1. Data

The Galaxy dataset corresponds to $n = 362,887$ different galaxies. This dataset has been already used in [33] with a preprocessing based on expert supervision to remove outliers. In this study all the original observations are considered, removing only points with missing values, which requires no expertise. The response variable Y is the stellar formation rate. The predictor $\mathbf{X}$ is made of spectral characteristics of the galaxies and is of dimension $p = 46$.

5.2. Evaluation setting

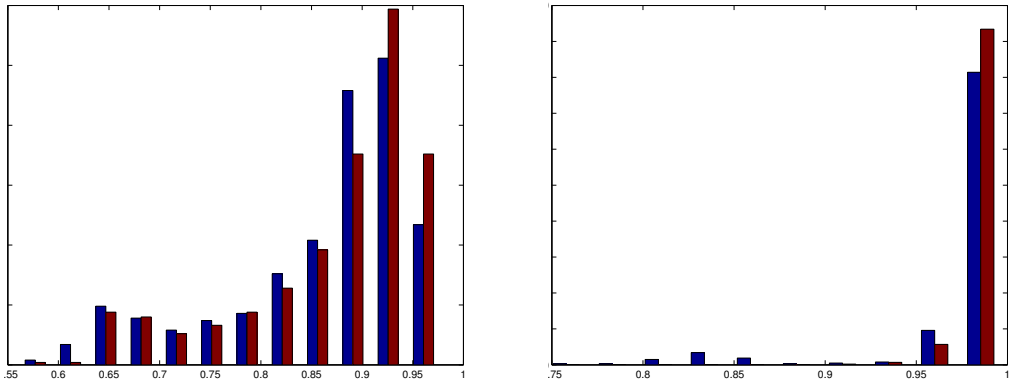
The number of samples n is very large and the proportion of outliers is very small compared to the whole dataset. The following strategy is adopted: 1000 random subsets of $\mathbf{X}$ of size $n_a = 3,000$, $\mathbf{X}_i^a$, $i = 1, \ldots, 1000$ and size $n_b = 30,000$, $\mathbf{X}_i^b$, $i = 1, \ldots, 1000$ are considered to compare the performance of SIR and Student SIR. First a reference result $\hat{\mathbf{B}}^{\text{SIR}}, \hat{\mathbf{B}}^{\text{st-SIR}}$ is obtained using the whole dataset $\mathbf{X}$, using respectively SIR and Student SIR with the dimension of the e.d.r. space set to $d = 3$ and the number of slices to $h = 1000$. The value $d = 3$ was selected via BIC computed for $d = 1, ..., 20$, which is reliable for such a large sample size. The proximity measure r (26) between the two reference spaces is $r(\hat{\mathbf{B}}^{\text{SIR}}, \hat{\mathbf{B}}^{\text{st-SIR}}) = 0.95$. SIR and st-SIR are identifying approximately the same same e.d.r. space.

5.3. Results

Let $\hat{\mathbf{B}}_i^{\text{SIR}}$, $\hat{\mathbf{B}}_i^{\text{st-SIR}}$ be the estimations of the basis of the e.d.r. space for the random subsets $\mathbf{X}_i^a$, $i = 1, \ldots, 1000$ using respectively SIR and Student SIR. The proximity measures $r_i^{\text{SIR}} = r(\hat{\mathbf{B}}^{\text{SIR}}, \hat{\mathbf{B}}_i^{\text{SIR}})$ and $r_i^{\text{st-SIR}} = r(\hat{\mathbf{B}}^{\text{st-SIR}}, \hat{\mathbf{B}}_i^{\text{st-SIR}})$ are considered. All results are obtained setting the number of slices to $h = 10$. The means (and standard deviations) of the resulting proximity measures r are respectively 0.86(0.09) for SIR and 0.87(0.09) for Student SIR. The experiment is better visualized in Figure 1 (a) where histograms show that Student SIR performs better than SIR most of the time. As expected SIR is less robust than Student SIR, obtaining with a higher frequency low values of r. The histograms show a difference between values around $r = 0.96$ (23.8% of random subsets for Student SIR, 17.2% for SIR).

In the second test, the sample size of the subsets is increased to $n_b = 30,000$. Accordingly, the number of slices is increased to $h = 100$. Not surprisingly, the means (and standard deviations) of r_i^{SIR} and $r_i^{\text{st-SIR}}$ are increasing to 0.97(0.04) and 0.99(0.00). Student SIR however still performs better than SIR (Figure 1 (b)) with some low values of the proximity measure for SIR while Student SIR has almost all the values (93.4% of random

subsets) concentrated around $r = 0.98$. The difference between the two approaches is then further emphasized in Figure 2 where the cloud of points in the upper left corner of the plot corresponds to datasets for which SIR was not able to estimate a correct basis of the e.d.r space while Student SIR shows good performance. Even if the true e.d.r space is unknown, this analysis suggests that Student SIR is robust to outliers and can be profitably used in real applications.



(a) (b)

Figure 1: Histograms of the proximity measure (26) $r_i^{\text{SIR}} = r(\hat{\mathbf{B}}^{\text{SIR}}, \hat{\mathbf{B}}_i^{\text{SIR}})$ (blue) and $r_i^{\text{st-SIR}} = r(\hat{\mathbf{B}}^{\text{st-SIR}}, \hat{\mathbf{B}}_i^{\text{st-SIR}})$ (red) for $i = 1, \ldots, 1000$ random subsets of $\mathbf{X}$ of size n_a=3000 **(a)** and n_b=30,000 **(b)**.

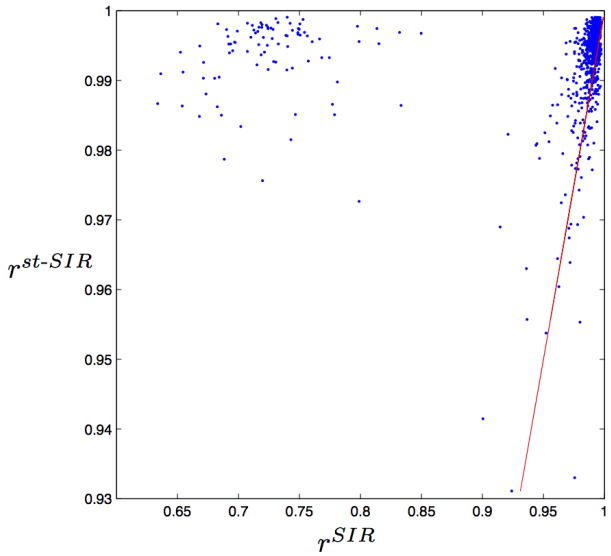


Figure 2: Horizontal axis r_i^{SIR}, vertical axis $r_i^{\text{st-SIR}}$, $i = 1, \ldots, 1000$, proximity measures computed using subsets of $\mathbf{X}$ of size $n_b = 30,000$. Almost all points are lying above the line $y = x$ indicating that Student SIR improves SIR results and significantly so for the subsets in the upper left corner.

Model	**X**	Method					
		SIR	CP-SIR	WCAN	WIRE	SIME	st-SIR
I	(i)	**.99(.01)**	.99(.01)	.98(.01)	.98(.01)	**.99(.01)**	**.99(.01)**
	(ii)	.63(.18)	.92(.04)	.88(.06)	.87(.07)	.91(.04)	**.98(.01)**
	(iii)	**.99(.01)**	.86(.12)	.72(.27)	.98(.01)	.97(.01)	**.99(.01)**
II	(i)	**.99(.01)**	.98(.01)	.98(.01)	.98(.01)	98(.01)	**.99(.01)**
	(ii)	.61(.18)	.92(.04)	.89(.06)	.87(.08)	.91(.05)	**.98(.01)**
	(iii)	**.99(.01)**	.67(.25)	.69(.28)	.98(.01)	.97(.02)	**.99(.01)**
III	(i)	.88(.06)	.87(.06)	**.89(.05)**	.86(.06)	.87(.06)	.87(.06)
	(ii)	.40(.13)	.78(.10)	.78(.11)	.76(.11)	.78(.10)	**.85(.06)**
	(iii)	.84(.07)	.63(.12)	.67(.13)	**.85(.07)**	**.85(.07)**	.84(.07)

(a)

Model	**X**	n	Method					
			SIR	CP-SIR	WCAN	WIRE	SIME	st-SIR
I	(i)	50	**.95(.03)**	.91(.09)	.86(.11)	.88(.11)	.90(.08)	**.95(.03)**
		100	**.98(.01)**	.96(.03)	.96(.03)	.95(.03)	.96(.02)	**.98(.01)**
		200	**.99(.01)**	**.99(.01)**	.98(.01)	.98 (.01)	**.99(.01)**	**.99(.01)**
		400	**1(.00)**	.99(.00)	.99(.00)	.99 (.01)	.99(.00)	**1(.00)**
	(ii)	50	.60(.22)	.66(.18)	.57(.23)	.49(.24)	.59(.21)	**.90(.07)**
		100	.62(.21)	.85 (.08)	.78(.11)	.73(.15)	.81(.10)	**.96(.02)**
		200	.62(.20)	.92(.04)	.88(.06)	.87(.07)	.91(.04)	**.98(.01)**
		400	.62(.18)	.96(.02)	.94(.03)	.93(.03)	.96(.02)	**.99(.00)**
	(iii)	50	**.95(.02)**	.45(.29)	.18(.19)	.73(.25)	.86(.09)	**.95(.02)**
		100	**.98(.01)**	.66(.25)	.35(.29)	.94(.04)	.94(.04)	**.98(.01)**
		200	**.99(.01)**	.86(.12)	.72(.27)	.98(.01)	.97(.01)	**.99(.00)**
		400	**.99(.00)**	.96(.04)	.96(.04)	.93(.03)	**.99(.01)**	**.99(.00)**

(b)

Table 1: **(a)** Average of the proximity measure r (eq. (26)) for sample size $n = 200$; and **(b)** effect of sample size n on the average proximity measure r, both over 200 repetitions with standard deviation in brackets. Six methods are compared. SIR: sliced inverse regression; CP-SIR: contour projection for SIR; WCAN: weighted canonical correlation; WIRE: weighted sliced inverse regression estimation; SIME: sliced inverse multivariate median estimation and st-SIR: Student SIR. In all cases, the number of slices is $h = 5$ and the predictor dimension $p = 10$. Best r values are in bold.

Model	$\mathbf{X}$	h	Method					
			SIR	CP-SIR	WCAN	WIRE	SIME	st-SIR
I	(i)	2	**.96(.02)**	.95(.03)	.98(.01)	.94(.03)	.95(.03)	.95(.02)
		5	**.99(.01)**	.98(.01)	.98(.01)	.98(.02)	.98(.01)	**.99(.00)**
		10	.99(.00)	.99(.01)	.98(.01)	.98 (.01)	.98(.01)	**1(.00)**
		20	**1(.00)**	.99(.01)	.98(.02)	.98 (.02)	.98(.01)	**1(.00)**
	(ii)	2	.60(.18)	.90(.05)	.60(.34)	.87(.06)	.89(.06)	**.95(.02)**
		5	.62(.18)	.92 (.04)	.89(.06)	.88(.07)	.92(.04)	**.98(.01)**
		10	.63(.19)	.92(.04)	.88(.07)	.87(.07)	.86(.08)	**.99(.00)**
		20	.65(.21)	.91(.05)	.85(.08)	.85(.08)	.69(.14)	**1(.00)**
	(iii)	2	**.96(.02)**	.91(.06)	.84(.20)	.95(.02)	.94(.05)	.95(.02)
		5	**.99(.00)**	.64(.26)	.67(.28)	.98(.01)	.98(.01)	**.99(.00)**
		10	**1(.00)**	.63(.26)	.48(.31)	.98(.01)	.98(.01)	**1(.00)**
		20	**1(.00)**	.53(.28)	.43(.30)	.98(.01)	.98(.01)	**1(.00)**

(a)

Model	Y	ν	Method					
			SIR	CP-SIR	WCAN	WIRE	SIME	st-SIR
(iii)	I	.5	**.99(.01)**	.98(.01)	.96(.02)	.96(.02)	.98(.01)	**.99(.01)**
		.2	**.99(.01)**	.96(.02)	.87(.15)	.97(.01)	.97(.01)	**.99(.01)**
		.1	**.99(.01)**	.86(.12)	.72(.27)	.98 (.01)	.97(.01)	**.99(.01)**
		.05	**.99(.01)**	.58(.24)	.65(.30)	.98 (.01)	.97(.01)	**.99(.01)**
	II	.5	**.99(.01)**	.98(.01)	.96(.02)	.96(.02)	.98(.01)	**.99(.01)**
		.2	**.99(.01)**	.96 .03)	.86(.16)	.98(.01)	.98(.01)	**.99(.01)**
		.1	**.99(.01)**	.67(.25)	.69(.28)	98(.01)	.97(.02)	**.99(.01)**
		.05	**.99(.01)**	.28(.24)	.59(.29)	98(.01)	.97(.01)	**.99(.01)**
	III	.5	**.88(.06)**	.85(.07)	.84(.08)	.77(.11)	.87(.06)	**.88(.05)**
		.2	.84(.07)	.76(.12)	.71(.13)	.84(.08)	**.86(.06)**	.84(.07)
		.1	.84(.07)	.63(.12)	.67(.13)	**.85(.07)**	**.85(.07)**	.84(.07)
		.05	.83(.07)	.58(.10)	.65(.13)	**.86(.07)**	**.86(.07)**	.82(.07)

(b)

Table 2: Effect of the number of slices **(a)** and of inlier magnitude ν **(b)** on the average proximity measure r (eq. (26)), over 200 repetitions with related standard deviation in brackets. Six methods are compared. SIR: sliced inverse regression; CP-SIR: contour projection for SIR; WCAN: weighted canonical correlation; WIRE: weighted sliced inverse regression estimation; SIME: sliced inverse multivariate median estimation and st-SIR: Student SIR. In all cases, the sample size is $n = 200$ and the predictor dimension $p = 10$. Best r values are in bold.

Model - **X**	*df*	Method	
		SIR	st-SIR
I - (ii)	3	.94(.05)	**.99(.00)**
	5	.98(.02)	**.99(.00)**
	7	.98(.01)	**.99(.00)**
	10	.99(.01)	**.99(.00)**
II - (ii)	3	.94(.05)	**.99(.00)**
	5	.97(.02)	**.99(.00)**
	7	.98(.01)	**.99(.00)**
	10	.99(.01)	**.99(.00)**
III - (ii)	3	.82(.08)	**.90(.04)**
	5	.88(.05)	**.92(.03)**
	7	.90(.04)	**.92(.03)**
	10	.90(.04)	**.92(.03)**
	30	.91(.03)	**.92(.03)**

(a)

Model - **X**	n	Method		BIC
		SIR	st-SIR	
III-(i)	200	.00(.00)	.13(.05)	25/200
	300	.01(.00)	.09(.03)	109/200
	400	.04(.01)	.33(.16)	156/200
	500	.05(.01)	.43(.17)	189/200
	1000	.10(.02)	.51(.17)	200/200
III-(ii)	200	.00(.00)	.13(.05)	21/200
	300	.01(.00)	.09(.05)	19/200
	400	.04(.01)	.33(.20)	39/200
	500	.05(.01)	.43(.18)	90/200
	1000	.10(.02)	.51(.20)	200/200
III-(iii)	200	.00(.00)	.13(.05)	0/200
	300	.01(.00)	.13(.04)	12/200
	400	.04(.01)	.38(.16)	22/200
	500	.05(.01)	.34(.13)	16/200
	1000	.10(.02)	.51(.17)	198/200

(b)

Table 3: **(a)** Effect of the degree of freedom (df) on the average of the proximity measure r (eq.(26)) for sample size $n = 200$, the number of slices is $h = 5$ and the predictor dimension $p = 10$; and **(b)** Effect of the sample size on the computational time in seconds (standard deviations in brackets) and ratio of correct selections ($d = 2$) for BIC over 200 runs.

6. Conclusion and future work

We proposed a new approach referred to as Student SIR to robustify SIR. In contrast to most existing approaches which aim at replacing the standard SIR estimators by robust versions, we considered the intrinsic characterization of SIR as a Gaussian inverse regression model [9] and modified it into a Student model with heavier tails. While SIR is not robust to outliers, Student SIR has shown to be able to deal with different kind of situations that depart from normality. As expected, when SIR provides good results, Student SIR is performing similarly but at a higher computational cost due to the need for an EM iterative algorithm for estimation.

Limitations of the approach include the difficulty in dealing with the case $p > n$ or when there are strong correlations between variables. Student SIR as well as SIR still suffer from the need to inverse large covariance matrices. A regularization, to overcome this problem, has been proposed in [10] and could be extended to our Student setting. Another practical issue is how to set the dimension d of the central subspace. We have proposed the use of BIC as a natural tool in our maximum likelihood setting. It provided good results but may be not suited when the sample size is too small. A more complete study and comparison with other solutions would be interesting.

To conclude, Student SIR shows good performance in the presence of outliers and is performing equally well in case of Gaussian errors. In our experiments, the algorithm has shown fast convergence being a promising alternative to SIR since nowadays most datasets include outliers. Future work would be to extend this setting to a multivariate response following the lead of [34, 35].

7. Appendix: Proofs

7.1. Proof of Proposition 1

The proof generalizes the proof of Proposition 6 in [9] to the generalized Student case. It comes from (7) that $\mathbf{X}_y$ follows a generalized Student distribution $\mathcal{S}_p(\boldsymbol{\mu}_y, \mathbf{V}, \alpha)$ where $\boldsymbol{\mu}_y = \boldsymbol{\mu} + \mathbf{VBc}(y)$. Generalized Student distributions have similar properties to Gaussian distributions (see for instance section 5.5 in [16]). In particular any affine transformation of a generalized Student distribution remains in this family. It follows that $\mathbf{B}^T\mathbf{X}|Y = y$ is distributed as $\mathcal{S}_d(\mathbf{B}^T\boldsymbol{\mu}_y, \mathbf{B}^T\mathbf{VB}, \alpha)$. Similarly, marginals and conditional distributions are retained in the family. It follows that $\mathbf{X}|\mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x}, Y = y$ is also a generalized Student distribution $\mathcal{S}_p(\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{V}}, \tilde{\alpha}, \tilde{\gamma})$ with

$$
\begin{aligned}
\tilde{\boldsymbol{\mu}} &= \boldsymbol{\mu}_y + \mathbf{VB}(\mathbf{B}^T\mathbf{VB})^{-1}(\mathbf{B}^T\mathbf{x} - \mathbf{B}^T\boldsymbol{\mu}_y) \\
&= \boldsymbol{\mu} + \mathbf{VB}(\mathbf{B}^T\mathbf{VB})^{-1}(\mathbf{B}^T\mathbf{x} - \mathbf{B}^T\boldsymbol{\mu}) \\
\tilde{\mathbf{V}} &= \mathbf{V} - \mathbf{VB}(\mathbf{B}^T\mathbf{VB})^{-1}\mathbf{B}^T\mathbf{V} \\
\tilde{\alpha} &= \alpha + d \\
\tilde{\gamma} &= \frac{1}{2} + (\mathbf{B}^T\mathbf{x} - \mathbf{B}^T\boldsymbol{\mu}_y)^T(\mathbf{B}^T\mathbf{VB})^{-1}(\mathbf{B}^T\mathbf{x} - \mathbf{B}^T\boldsymbol{\mu}_y) \\
&= \frac{1}{2} + \boldsymbol{\varepsilon}^T\mathbf{B}(\mathbf{B}^T\mathbf{VB})^{-1}\mathbf{B}^T\boldsymbol{\varepsilon}\ ,
\end{aligned}
$$

from which it is clear that $\tilde{\mathbf{V}}, \tilde{\alpha}, \tilde{\gamma}$ and $\tilde{\boldsymbol{\mu}}$ do not depend on y. It follows that $\mathbf{X}|\mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x}, Y = y$ has the same distribution as $\mathbf{X}|\mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x}$ for all values $\mathbf{x}$. Consequently Y is independent on $\mathbf{X}$ conditionally to $\mathbf{B}^T\mathbf{X}$ which implies that $Y|\mathbf{X} = \mathbf{x}$ and $Y|\mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x}$ have identical distributions for all values $\mathbf{x}$. ■

Note that for the proof of the proposition, it was necessary to show that the independence on y holds for each parameter of the distribution and not only for the mean. The independence on y of the mean is actually straightforward using [9] where it appears that the proof that $E[\mathbf{X}|\mathbf{B}^T\mathbf{X}, Y = y]$

does not depend on y is independent on the distribution of $\boldsymbol{\varepsilon}$. Indeed the proof uses only the properties of the conditional expectation seen as a projection operator. This means that in our case also, $\mathbf{B}$ corresponds to the *mean* central subspace as defined by $E[\mathbf{X}|\mathbf{B}^T\mathbf{X}, Y = y] = E[\mathbf{X}|\mathbf{B}^T\mathbf{X}]$.

7.2. Proof of Lemma 1

The proof is adapted from the proof of lemma 1 in [10] taking into account the additional quantities $\bar{u_i}$'s. Let us remark that

$$R \stackrel{def}{=} G(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C}) - \log\det \mathbf{V} = \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathbf{Z}_i^T\mathbf{V}^{-1}\mathbf{Z}_i, \tag{27}$$

where we have defined for $i = 1, \ldots, n$,

$$\begin{aligned}
\mathbf{Z}_i &= \boldsymbol{\mu} + \mathbf{V}\mathbf{B}\mathbf{C}^T\mathbf{s}_i - \mathbf{x}_i && (28)\\
&= (\boldsymbol{\mu} - \bar{\mathbf{x}} + \mathbf{V}\mathbf{B}\mathbf{C}^T\bar{\mathbf{s}}) + \mathbf{V}\mathbf{B}\mathbf{C}^T(\mathbf{s}_i - \bar{\mathbf{s}}) - (\mathbf{x}_i - \bar{\mathbf{x}}) && (29)\\
&\stackrel{def}{=} \mathbf{Z}_1 + \mathbf{Z}_{2,i} - \mathbf{Z}_{3,i}. && (30)
\end{aligned}$$

Since $\mathbf{Z}_{2,\cdot}$ and $\mathbf{Z}_{3,\cdot}$ are centered, replacing the previous expansion in (27) yields

$R = \bar{u}\ \mathbf{Z}_1^T\mathbf{V}^{-1}\mathbf{Z}_1 + \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathbf{Z}_{2,i}^T\mathbf{V}^{-1}\mathbf{Z}_{2,i} + \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathbf{Z}_{3,i}^T\mathbf{V}^{-1}\mathbf{Z}_{3,i} - \frac{2}{n}\sum_{i=1}^{n} \bar{u_i}\mathbf{Z}_{2,i}^T\mathbf{V}^{-1}\mathbf{Z}_{3,i}$,

where

$$\begin{aligned}
\mathbf{Z}_1^T\mathbf{V}^{-1}\mathbf{Z}_1 &= (\boldsymbol{\mu} - \bar{\mathbf{x}} + \mathbf{V}\mathbf{B}\mathbf{C}^T\bar{\mathbf{s}})^T\mathbf{V}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}} + \mathbf{V}\mathbf{B}\mathbf{C}^T\bar{\mathbf{s}}),\\
\frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathbf{Z}_{2,i}^T\mathbf{V}^{-1}\mathbf{Z}_{2,i} &= tr(\mathbf{C}^T\mathbf{W}\mathbf{C}\mathbf{B}^T\mathbf{V}\mathbf{B}),\\
\frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathbf{Z}_{3,i}^T\mathbf{V}^{-1}\mathbf{Z}_{3,i} &= \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathrm{tr}((\mathbf{x}_i - \bar{\mathbf{x}})^T\mathbf{V}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}))\\
&= \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathrm{tr}(\mathbf{V}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T)\\
&= \mathrm{tr}(\mathbf{V}^{-1}\boldsymbol{\Sigma}) \text{ and}\\
\frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathbf{Z}_{2,i}^T\mathbf{V}^{-1}\mathbf{Z}_{3,i} &= tr(\mathbf{C}^T\mathbf{M}\mathbf{B}),
\end{aligned}$$

and the conclusion follows. ∎

7.3. Proof of Proposition 2

Cancelling the gradients of $G(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C})$ yields the system of equations

$$\frac{1}{2}\nabla_\mu G = \hat{\mathbf{V}}^{-1}(\hat{\boldsymbol{\mu}} - \bar{\mathbf{x}} + \hat{\mathbf{V}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T\bar{\mathbf{s}}) = 0, \tag{31}$$

$$\frac{1}{2}\nabla_B G = \hat{\mathbf{V}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T(\bar{u}\bar{\mathbf{s}}\bar{\mathbf{s}}^T + \mathbf{W})\hat{\mathbf{C}} - \mathbf{M}^T\hat{\mathbf{C}} + \bar{u}(\hat{\boldsymbol{\mu}} - \bar{\mathbf{x}})\bar{\mathbf{s}}^T\hat{\mathbf{C}} = 0, \tag{32}$$

$$\frac{1}{2}\nabla_C G = \bar{u}(\bar{\mathbf{s}}\bar{\mathbf{s}}^T\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}} + \bar{\mathbf{s}}(\hat{\boldsymbol{\mu}} - \bar{\mathbf{x}})^T\hat{\mathbf{B}}) + \mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}} - \mathbf{M}\hat{\mathbf{B}} = 0, \tag{33}$$

$$\nabla_V G = \hat{\mathbf{V}}^{-1} + \hat{\mathbf{B}}\hat{\mathbf{C}}^T(\bar{u}\bar{\mathbf{s}}\bar{\mathbf{s}}^T + \mathbf{W})\hat{\mathbf{C}}\hat{\mathbf{B}}^T + \tag{34}$$

$$- \hat{\mathbf{V}}^{-1}\left(\bar{u}(\hat{\boldsymbol{\mu}} - \bar{\mathbf{x}})(\hat{\boldsymbol{\mu}} - \bar{\mathbf{x}})^T + \boldsymbol{\Sigma}\right)\hat{\mathbf{V}}^{-1} = 0. \tag{35}$$

From (31), we have

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} - \hat{\mathbf{V}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T\bar{\mathbf{s}}. \tag{36}$$

Replacing in (32) and (33) yields the simplified system of equations

$$\hat{\mathbf{V}}\hat{\mathbf{B}}(\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}) = \mathbf{M}^T\hat{\mathbf{C}}, \tag{37}$$

$$\mathbf{W}\hat{\mathbf{C}}(\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}}) = \mathbf{M}\hat{\mathbf{B}}. \tag{38}$$

It follows from the last equality that

$$\hat{\mathbf{C}} = \mathbf{W}^{-1}\mathbf{M}\hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}})^{-1}\ . \tag{39}$$

Multiplying (37) by $\mathbf{B}^T\mathbf{V}\mathbf{B}$ on the left, we get

$$\hat{\mathbf{V}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}} = \mathbf{M}^T\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}}, \tag{40}$$

and assuming $\mathbf{W}$ is regular, (38) entails $\hat{\mathbf{C}}(\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}}) = \mathbf{W}^{-1}\mathbf{M}\hat{\mathbf{B}}$. Replacing in (40) yields

$$\hat{\mathbf{V}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}} = \mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}\hat{\mathbf{B}}. \tag{41}$$

Now, multiplying (34) on the left and on the right by $\hat{\mathbf{V}}$ and taking account of (36) entails

$$\boldsymbol{\Sigma} = \hat{\mathbf{V}} + \hat{\mathbf{V}}\hat{\mathbf{B}}(\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}})\hat{\mathbf{B}}^T\hat{\mathbf{V}}. \tag{42}$$

As a consequence of (42), it comes

$$\boldsymbol{\Sigma}\hat{\mathbf{B}} = \hat{\mathbf{V}}\hat{\mathbf{B}}(\mathbf{I} + \hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}}), \tag{43}$$

and

$$\hat{\mathbf{V}}\hat{\mathbf{B}} = \hat{\mathbf{\Sigma}}\hat{\mathbf{B}}(\mathbf{I} + \hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}})^{-1}. \tag{44}$$

Using this expression of $\hat{\mathbf{V}}\hat{\mathbf{B}}$ above in (41), it comes

$$\hat{\mathbf{B}}\left(\mathbf{I} + (\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}})^{-1}\right)^{-1} = \mathbf{\Sigma}^{-1}\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}\hat{\mathbf{B}}\ , \tag{45}$$

which means that the columns of $\hat{\mathbf{B}}$ are stable by $\mathbf{\Sigma}^{-1}\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}$ and thus are eigenvectors of $\mathbf{\Sigma}^{-1}\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}$. Let us denote by $\lambda_1, \ldots, \lambda_d$ the associated eigenvalues. Matrix $\mathbf{\Sigma}^{-1}\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}$ is of size $p \times p$ and of rank at most $min(h, p)$ since $\mathbf{W}$ is assumed to be regular. In practice we will assume $h \geq d$ and $p \geq d$. Therefore $d \leq min(h, p)$. It remains to show that $\lambda_1, \ldots, \lambda_d$ are the d largest eigenvalues. To this aim, we observe that using successively (38) and (42),

$$\begin{aligned} G(\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}) &= \log\det\hat{\mathbf{V}} + \mathrm{trace}(\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}}\mathbf{C}^T\mathbf{W}) + \mathrm{trace}(\mathbf{V}^{-1}\mathbf{\Sigma}) - 2\mathrm{trace}(\hat{\mathbf{B}}\hat{\mathbf{C}}^T\mathbf{M}) \\ &= \log\det\hat{\mathbf{V}} + \mathrm{trace}(\hat{\mathbf{M}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T) + p + \mathrm{trace}(\hat{\mathbf{M}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T) - 2\mathrm{trace}(\hat{\mathbf{B}}\hat{\mathbf{C}}^T\mathbf{M}) \\ &= p + \log\det\hat{\mathbf{V}}. \end{aligned}$$

Let us consider the two following matrices, $\mathbf{\Delta}_1 = \mathbf{B}\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}$ and $\mathbf{\Delta}_2 = \hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\mathbf{B}$. $\mathbf{\Delta}_1$ is $p \times p$ of rank at most d and $\mathbf{\Delta}_2$ is $d \times d$ of rank d, invertible with positive eigenvalues denoted by $\delta_1, \ldots, \delta_d$. The eigenvalues of $\mathbf{\Delta}_2$ are that of $\mathbf{\Delta}_1$ too. Indeed consider $\mathbf{y}_k$ an eigenvector for δ_k, then $\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\mathbf{B}\mathbf{y}_k = \delta_k\mathbf{y}_k$. Multiplying on the left by $\hat{\mathbf{B}}$ and considering $\mathbf{z}_k = \hat{\mathbf{B}}\mathbf{y}_k$, it comes that δ_k is also an eigenvalue for $\mathbf{\Delta}_1$. Using (42), it follows then

$$\log\det\hat{\mathbf{V}} = \log\det\mathbf{\Sigma} - \log\det(\mathbf{I} + \mathbf{\Delta}_1) = \log\det\mathbf{\Sigma} - \sum_{k=1}^{d}\log(1 + \delta_k)\ .$$

Multiplying (45) by $\hat{\mathbf{B}}^T$ and using $\hat{\mathbf{B}}^T\hat{\mathbf{B}} = \mathbf{I}$, it comes
$\mathbf{I} + \mathbf{\Delta}_2^{-1} = (\hat{\mathbf{B}}^T\mathbf{\Sigma}^{-1}\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}\hat{\mathbf{B}})^{-1} = diag(1/\lambda_k)$ from which $\delta_k = \frac{1}{1-\lambda_k} - 1$ can be deduced. Finally,

$$G(\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}) = p + \log\det\mathbf{\Sigma} + \sum_{k=1}^{d}\log(1 - \lambda_k)\ .$$

G is then minimized when the λ_k are the largest. As a consequence of (42), it also comes that

$$\hat{\mathbf{V}} = \mathbf{\Sigma} - \hat{\mathbf{V}}\hat{\mathbf{B}}(\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}})\hat{\mathbf{B}}^T\hat{\mathbf{V}}. \tag{46}$$

Replacing $\hat{\mathbf{V}}\hat{\mathbf{B}}$ in (46) by the expression given in (37), it comes

$$\hat{\mathbf{V}} = \mathbf{\Sigma} - \mathbf{M}^T\hat{\mathbf{C}}(\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}})^{-1}\hat{\mathbf{C}}^T\mathbf{M}. \tag{47}$$

Grouping the results in (47), (39), (36) and the considerations after (45) gives the Proposition. ■

7.4. Proof of Corollary 1.

Let us remark that, under (23), the coefficients W_{ij} of $\mathbf{W}$ have an explicit form:

$$\begin{aligned}
\mathbf{W} &= \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\ (\mathbf{s}_i - \bar{\mathbf{s}})(\mathbf{s}_i - \bar{\mathbf{s}})^T \\
&= \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\ \mathbf{s}_i\mathbf{s}_i^T - \frac{2}{n}\sum_{i=1}^{n} \bar{u_i}\ \mathbf{s}_i\bar{\mathbf{s}}^T + \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\ \bar{\mathbf{s}}\bar{\mathbf{s}}^T \\
&= \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\ \mathbf{s}_i\mathbf{s}_i^T - \frac{2ff^t}{\bar{u}} + \frac{ff^t}{\bar{u}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\ \mathbf{s}_i\mathbf{s}_i^T - \frac{ff^t}{\bar{u}},
\end{aligned}$$

where $f = (f_1, ..., f_h)$. Using (23) the first sum corresponds to $\mathrm{diag}(f_1, ..., f_h)$ leading to $\mathbf{W} = \mathrm{diag}(f_1, ..., f_h) - \dfrac{ff^t}{\bar{u}}$. The inverse matrix of $\mathbf{W}$ can be calculated using Sherman-Morrison formula:

$$\mathbf{W}^{-1} = \mathrm{diag}\left(\frac{1}{f_1}, \ldots, \frac{1}{f_h}\right) + \frac{1}{f_{h+1}}\mathbf{O},$$

where $\mathbf{O}$ is the $h \times h$ matrix defined by $O_{ij} = 1$ for all $(i, j) \in \{1, \ldots, h\} \times \{1, \ldots, h\}$. Using (23) the jth row of $\mathbf{M}$ is given by:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\bar{u_i}(\mathbb{1}\{y_i \in S_j\} - \bar{s_j})(\mathbf{x}_i - \bar{\mathbf{x}})^T &= \frac{1}{n}\sum_{i=1}^{n}\bar{u_i}\mathbb{1}\{y_i \in S_j\}\mathbf{x}_i^T - \frac{1}{n}\sum_{i=1}^{n}\bar{u_i}\mathbb{1}\{y_i \in S_j\}\bar{\mathbf{x}}^T \\ &\quad - \frac{1}{n}\sum_{i=1}^{n}\bar{u_i}\bar{s_j}\mathbf{x}_i^T + \frac{1}{n}\sum_{i=1}^{n}\bar{u_i}\bar{s_j}\bar{\mathbf{x}}^T \\ &= f_j\bar{\mathbf{x}_j}^T - f_j\bar{\mathbf{x}}^T - f_j\bar{\mathbf{x}}^T + f_j\bar{\mathbf{x}}^T \\ &= f_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T,\end{aligned}$$

for all $j = 1, \ldots, h$. Now taking into account that $\mathbf{O}^2 = h\mathbf{O}$, we have

$$\begin{aligned}\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M} &= \sum_{j=1}^{h} f_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T + \frac{1}{f_{h+1}}\mathbf{M}^T\mathbf{O}\mathbf{M} \\ &= \sum_{j=1}^{h} f_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T + \frac{1}{hf_{h+1}}(\mathbf{M}^T\mathbf{O})(\mathbf{M}^T\mathbf{O})^T. \qquad (48)\end{aligned}$$

Now, remarking that all the columns of $\mathbf{M}^T\mathbf{O}$ are equal to

$$\sum_{j=1}^{h} f_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}) = \sum_{j=1}^{h+1} f_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}) - f_{h+1}(\bar{\mathbf{x}}_{h+1} - \bar{\mathbf{x}}) = -f_{h+1}(\bar{\mathbf{x}}_{h+1} - \bar{\mathbf{x}}),$$

where $f_{h+1} = \frac{1}{n}\sum_{i=1}^{n}\bar{u_i}\mathbb{1}\{y_i \in S_{h+1}\} = \bar{u} - \sum_{j=1}^{h} f_j$ it follows that

$$(\mathbf{M}^T\mathbf{O})(\mathbf{M}^T\mathbf{O})^T = hf_{h+1}^2(\bar{\mathbf{x}}_{h+1} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{h+1} - \bar{\mathbf{x}})^T$$

and thus replacing in (48) yields

$$\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M} = \sum_{j=1}^{h+1} f_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T = \mathbf{\Gamma}.$$

The result is then a consequence of Proposition 2. ■

Acknowledgments

The authors would like to deeply thank the two reviewers and the AE for their comments and remarks. The authors would like to thank Didier Fraix-Burnet for his contribution to the data. This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01).

References

[1] R. D. Cook, Graphics for regressions with a binary response, Journal of the American Statistical Association 91 (435) (1996) 983–992.

[2] X. Yin, B. Li, R. D. Cook, Successive direction extraction for estimating the central subspace in a multiple-index regression, Journal of Multivariate Analysis 99 (8) (2008) 1733–1757.

[3] K.-C. Li, Sliced inverse regression for dimension reduction, Journal of the American Statistical Association 86 (414) (1991) 316–327.

[4] P. Hall, K.-C. Li, On almost linearity of low dimensional projections from high dimensional data, The Annals of Statistics 21 (2) (1993) 867–889.

[5] K. Fukumizu, F. R. Bach, M. I. Jordan, Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces, The Journal of Machine Learning Research 5 (2004) 73–99.

[6] B. Li, Y. Dong, Dimension reduction for nonelliptically distributed predictors, The Annals of Statistics 37 (3) (2009) 1272–1298.

[7] K. Fukumizu, F. R. Bach, M. I. Jordan, Kernel dimension reduction in regression, The Annals of Statistics 37 (4) (2009) 1871–1905.

[8] R. D. Cook, S. Weisberg, Sliced inverse regression for dimension reduction: Comment, Journal of the American Statistical Association 86 (414) (1991) 328–332.

[9] R. D. Cook, Fisher lecture: Dimension reduction in regression, Statistical Science 22 (1) (2007) 1–26.

[10] C. Bernard-Michel, L. Gardes, S. Girard, Gaussian regularized sliced inverse regression, Statistics and Computing 19 (1) (2009) 85–98.

[11] E. Bura, R. D. Cook, Extending sliced inverse regression: The weighted chi-squared test, Journal of the American Statistical Association 96 (455) (2001) 996–1003.

[12] S. J. Sheather, J. W. McKean, A comparison of procedures based on inverse regression, Lecture Notes-Monograph Series 31 (1997) 271–278.

[13] U. Gather, T. Hilker, C. Becker, A note on outlier sensitivity of sliced inverse regression, Statistics: A Journal of Theoretical and Applied Statistics 36 (4) (2002) 271–281.

[14] U. Gather, T. Hilker, C. Becker, A robustified version of sliced inverse regression, in: Statistics in Genetics and in the Environmental Sciences, Trends in Mathematics, Springer, 2001, Ch. 2, pp. 147–157.

[15] Y. Dong, Z. Yu, L. Zhu, Robust inverse regression for dimension reduction, Journal of Multivariate Analysis 134 (2015) 71–81.

[16] S. Kotz, S. Nadarajah, Multivariate t Distributions and their Applications, Cambridge, 2004.

[17] G. McLachlan, D. Peel, Robust mixture modelling using the t distribution, Statistics and Computing 10 (2000) 339–348.

[18] D. F. Andrews, C. L. Mallows, Scale mixtures of normal distributions, Journal of the Royal Statistical Society. Series B (Methodological) 36 (1) (1974) 99–102.

[19] N. L. Johnson, S. Kotz, N. Balakrishnan, Continuous Univariate Distributions, vol.2, 2nd edition, John Wiley & Sons, New York, 1994.

[20] L. Ferré, Determining the dimension in sliced inverse regression and related methods, Journal of the American Statistical Association 93 (441) (1998) 132–140.

[21] M. Chavent, S. Girard, V. Kuentz-Simonet, B. Liquet, T. M. N. Nguyen, J. Saracco, A sliced inverse regression approach for data stream, Computational Statistics 29 (5) (2014) 1129–1152.

[22] B. Liquet, J. Saracco, A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches, Computational Statistics 27 (1) (2012) 103–125.

[23] J. R. Schott, Determining the Dimensionality in Sliced Inverse Regression, Journal of the American Statistical Association 89 (425) (1994) 141–148.

[24] S. Velilla, Assessing the number of linear components in a general regression problem, Journal of the American Statistical Association 93 (443) (1998) 1088–1098.

[25] M. P. Barrios, S. Velilla, A bootstrap method for assessing the dimension of a general regression problem, Statistics & Probability Letters 77 (3) (2007) 247–255.

[26] L. Zhu, B. Miao, H. Peng, On Sliced Inverse Regression With High-Dimensional Covariates, Journal of the American Statistical Association 101 (2006) 630–643.

[27] R. E. Kass, A. E. Raftery, Bayes factors, Journal of the American Statistical Association 90 (1995) 773–795.

[28] C. Giraud, Introduction to high-dimensional statistics, Chapman and Hall/CRC Monographs on Statistics and Applied Probability, 2014.

[29] H. Wang, L. Ni, C.-L. Tsai, Improving dimension reduction via contour-projection, Statistica Sinica 18 (1) (2008) 299–311.

[30] R. Luo, H. Wang, C.-L. Tsai, Contour projected dimension reduction, The Annals of Statistics 37 (6B) (2009) 3743–3778.

[31] J. Zhou, Robust dimension reduction based on canonical correlation, Journal of Multivariate Analysis 100 (1) (2009) 195–209.

[32] R. D. Cook, L. Forzani, D. R. Tomassi, Ldr: a package for likelihood-based sufficient dimension reduction, Journal of Statistical Software 39 (3) (2011) 1–20.

[33] A. Chiancone, S. Girard, J. Chanussot, Collaborative sliced inverse regression, Communications in Statistics - Theory and Methods. To appear.

[34] L. Barreda, A. Gannoun, J. Saracco, Some extensions of multivariate SIR, Journal of Statistical Computation and Simulation 77 (2007) 1–17.

[35] R. Coudret, S. Girard, J. Saracco, A new sliced inverse regression method for multivariate response, Computational Statistics and Data Analysis 77 (2014) 285–299.

CHAPTER 4

KNOCKOFF SIR

The inferno of the living is not something that will be; if there is one,
it is what is already here, the inferno where we live every day,
that we form by being together. There are two ways to escape suffering it.
The first is easy for many: accept the inferno and become such a part of it
that you can no longer see it. The second is risky and demands constant
vigilance and apprehension: seek and learn to recognize who and what,
in the midst of inferno, are not inferno, then make them endure,
give them space.
I.Calvino.

This last chapter is dedicated to the development of an extension of SIR providing sparse solutions and able to perform variable selection. The strategy to achieve sparsity differs from what can be found in [49] or [52] where the shrinkage idea if lasso is adapted to SIR. The solution given in the following makes use of knockoff filters ([5]). A knockoff filter is a copy of the original dataset $\mathbf{X}$ with certain properties that will be discussed in the next section. In section 4.2 the adaptation to SIR is described and motivated by a theorem. Finally the two remaining sections are dedicated respectively to the analysis of simulation results and a real data application.

4.1 Knockoff filter

Let $\mathbf{X} = \{x_1, ..., x_n\} \in \mathbb{R}^{p \times n}$ be the set of observed predictors and denote by $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}$. It is further assumed without loss of generality $\mathbb{E}(\mathbf{X}) = 0$ and $diag(\hat{\Sigma}) = 1$. A knockoff filter $\tilde{\mathbf{X}} = \{\tilde{x}_1, ..., \tilde{x}_n\} \in \mathbb{R}^{p \times n}$ is a set of points such that:

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \hat{\Sigma} \quad \text{and} \quad \mathbf{X}^T\tilde{\mathbf{X}} = \hat{\Sigma} - diag\{s\} \tag{4.1}$$

where s is a p-dimensional nonnegative vector. The knockoff has the same structure of the covariance matrix. Couples variables of $\mathbf{X}$, $(\mathbf{X}_j, \mathbf{X}_k)$ (columns of $\mathbf{X}$) and $(\mathbf{X}_j, \tilde{\mathbf{X}}_k)$ have the same correlation for $k \neq j$. On the diagonal, results that:

$$\mathbf{X}_j^T\tilde{\mathbf{X}}_j = \hat{\Sigma}_{jj} - s_j = 1 - s_j. \tag{4.2}$$

In other words each variable $\mathbf{X}_j$ interacts with the other variables $\mathbf{X}_k$ in the same way as the knockoffs $\tilde{\mathbf{X}}_k$ for $k \neq j$. The comparison of a variable $\mathbf{X}_j$ and its knockoff $\tilde{\mathbf{X}}_j$ gives a correlation of $1 - s_j$; the choice of s is crucial to allow our procedure to distinct a knockoff copy from the *true* variable. Let us concatenate $\mathbf{X}$ and the knockoff $\tilde{\mathbf{X}}$, $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{2p \times n}$ and look at:

$$\tilde{\Sigma} = [X, \tilde{X}]^T[X, \tilde{X}] = \begin{bmatrix} \hat{\Sigma} & \hat{\Sigma} - \text{diag}\{s\} \\ \hat{\Sigma} - \text{diag}\{s\} & \hat{\Sigma} \end{bmatrix}, \tag{4.3}$$

for our purpose $\tilde{\Sigma}$ must be a covariance matrix and therefore positive semidefinite. This is true, as stated in [5], when $diag\{s\}$ and $2\hat{\Sigma} - diag\{s\}$ are positive semidefinite. A knockoff filter can be obtained using the following formula:

$$\tilde{\mathbf{X}} = A^T\mathbf{X} + (\tilde{U}C)^T, \tag{4.4}$$

where $A = (I - \hat{\Sigma}^{-1}diag\{s\})$, $\tilde{U}$ is a $n \times p$ matrix orthogonal to the span of the columns of $\mathbf{X}$, Span$\{X\}$, and $CC^T = 2\text{diag\{s\}} - \text{diag}\{s\}\hat{\Sigma}^{-1}\text{diag}\{s\}$. Such $\tilde{U}$ exists only if $n \geq 2p$. This assumption is not restrictive and lies in the comfort zone for SIR, it is in fact well known that when $n \leq p$ instabilities arise in the inversion of the covariance matrix [9]. Depending on the choice of s knockoffs with different properties are considered in [5]:

Equi-correlated knockoffs. In this case all couples of variables are required to have the same correlation, for all j:

$$\mathbf{X}_j^T\tilde{\mathbf{X}}_j = 1 - min\{2\lambda_{min}(\hat{\Sigma}), 1\}. \tag{4.5}$$

where $\lambda_{min}(\hat{\Sigma})$ is the smallest eigenvalue of the matrix $\hat{\Sigma}$. In case of equi-correlated knockoffs this choice of s minimizes the absolute value of the correlation $|\mathbf{X}_j^T\tilde{\mathbf{X}}_j|$.

SDP knockoffs. A different possibility is to drop the equi-correlated assumption and provide the minimal average correlation for each pair of variables:

(4.6)

$$\text{minimize} \quad \sum_j (1-s_j) \quad \text{such that} \quad 0 \le s_j \le 1 \quad \text{and} \quad 2\hat{\Sigma} - diag\{s\} \quad \text{is positive semidefinite.}$$

This optimization problem can be efficiently solved via semidefinite programming (SDP).

A fast implementation in Matlab allows to generate both (see Matlab package), through the analysis the first typology of knockoffs variables has been used.

4.2 Main result: Knockoff SIR

In this section the main result is presented: a theorem on the behavior of SIR when the knockoffs are added to the analysis. Let us show first the following Lemma:

Lemma 4.1. *The SIR covariance matrix* $\tilde{\Gamma}$ *(see subsection 1.2.3) for the concatenation* $[\mathbf{X}, \tilde{\mathbf{X}}]$ *has the form:*

$$\tilde{\Gamma} = \begin{bmatrix} \tilde{\Gamma}_1 & \tilde{\Gamma}_2 \\ \tilde{\Gamma}_3 & \tilde{\Gamma}_4 \end{bmatrix} \tag{4.7}$$

where the four $p \times p$ *matrices are*

(i) $\tilde{\Gamma}_1 = \hat{\Gamma}$ *the covariance matrix for SIR calculated on* $\mathbf{X}$

(ii) $\tilde{\Gamma}_2 = \hat{\Gamma} A$

(iii) $\tilde{\Gamma}_3 = A^T \hat{\Gamma}$

(iv) $\tilde{\Gamma}_4 = A^T \hat{\Gamma} A + \sum_{j=1}^{h} \frac{1}{n_j n} \sum_{i=1}^{n} \mathbb{I}[y_i \in s_j] \tilde{f}_i \tilde{f}_i^T$, *where* $\tilde{f}_i = (\tilde{U}C)^T{}_i$ *is a column vector,*

$n_j = \sum_{i=1}^{n} \mathbb{I}[y_i \in s_j]$ *and* A, C *and* $\tilde{U}$ *are from equation* (4.12).

Proof. After dividing Y in h slices $\{s_1, ..., s_h\}$ (subsection 1.2.3) the SIR covariance matrix has the form:

$$\hat{\Gamma} = \sum_{j=1}^{h} \frac{n_j}{n} \hat{m}_j \hat{m}_j^T \tag{4.8}$$

where $\hat{m}_j = \frac{1}{n_j}\sum_{i=1}^{n} x_i \mathbb{I}[y_i \in s_j]$. From (4.8) is straightforward to see that $\tilde{\Gamma}_1 = \hat{\Gamma}$ since only the first p variables of $[\mathbf{X},\tilde{\mathbf{X}}]$ are involved in the calculation and thus only $\mathbf{X}$ contributes. For $\tilde{\Gamma}_2$ from the definition follows that:

$$\tilde{\Gamma}_2 = \sum_{j=1}^{h} \frac{n_j}{n} \hat{m}_j \tilde{m}_j^T, \tag{4.9}$$

where $\tilde{m}_j = \frac{1}{n_j}\sum_{i=1}^{n} \tilde{x}_i \mathbb{I}[y_i \in s_j] = \frac{1}{n_j}\sum_{i=1}^{n} (A^T x_i + (\tilde{U}C)^T{}_i)\mathbb{I}[y_i \in s_j]$. Therefore it follows that:

$$\tilde{\Gamma}_2 = \sum_{j=1}^{h} \frac{n_j}{n} \hat{m}_j \hat{m}_j^T A + \sum_{j=1}^{h} m_j \left(\frac{1}{n_j} \sum_{i=1}^{n} (\tilde{U}C)_i \mathbb{I}[y_i \in s_j] \right). \tag{4.10}$$

The first term of this equation is simply $\hat{\Gamma} A$, the second term is a $p \times p$ zero matrix since by construction $\mathbf{X}\tilde{U} = 0$. A similar procedure gives $\tilde{\Gamma}_3$ and $\tilde{\Gamma}_4$.

■

Theorem 4.1. *Given the predictors $\mathbf{X} = \{x_1, ..., x_n\} \in \mathbb{R}^{p \times n}$ and a response variable $Y = \{y_1, ..., y_n\} \in \mathbb{R}^{n \times 1}$ let us denote by $\hat{\mathbf{B}}$ the SIR estimator of $\mathbf{B} \in \mathbb{R}^{p \times k}$ in the following regression model:*

$$Y = f(\mathbf{XB}, \epsilon) \tag{4.11}$$

where f is an unknown link function and ϵ is a random error independent of $\mathbf{X}$. The k-columns of $\mathbf{B}$ span the e.d.r. space [47]. When $n > 2p$ let us consider a knockoff filter $\tilde{\mathbf{X}} \in \mathbb{R}^{p \times n}$ of the form:

$$\tilde{\mathbf{X}} = A^T \mathbf{X} + (\tilde{U}C)^T, \tag{4.12}$$

defined in the previous section, and the concatenation $[\mathbf{X},\tilde{\mathbf{X}}] \in \mathbb{R}^{2p \times n}$. The SIR estimator $\tilde{\mathbf{B}} \in \mathbb{R}^{2pxk}$ for the concatenation $[\mathbf{X},\tilde{\mathbf{X}}]$ has each column $\tilde{\mathbf{B}}_j$ of the form:

$$\tilde{\mathbf{B}}_j = [\hat{\mathbf{B}}_j, 0] \tag{4.13}$$

where 0 is a p-dimensional vector of all zeros.

Proof. Without loss of generality let $\mathbb{E}([\mathbf{X},\tilde{\mathbf{X}}]) = 0$ and let us focus on the case where $k = 1$, $\tilde{\mathbf{B}} \in \mathbb{R}^{2p \times 1}$. For construction it follows:

$$\tilde{\Sigma} = [X, \tilde{X}]^T [X, \tilde{X}] = \begin{bmatrix} \hat{\Sigma} & \hat{\Sigma} - \text{diag}\{s\} \\ \hat{\Sigma} - \text{diag}\{s\} & \hat{\Sigma} \end{bmatrix}, \tag{4.14}$$

for a given vector s, where $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}$. We want to show that

$$\tilde{\Gamma}\tilde{\mathbf{B}} = \lambda\tilde{\Sigma}\tilde{\mathbf{B}}. \tag{4.15}$$

Using Lemma 4.1 and $\tilde{\mathbf{B}}_j = [\hat{\mathbf{B}}_j, 0]$ is easy to see that the problem can be decomposed in two parts:

$$\hat{\Gamma}\hat{\mathbf{B}} = \lambda\hat{\Sigma}\hat{\mathbf{B}} \tag{4.16}$$

$$A^T\hat{\Gamma}\hat{\mathbf{B}} = \lambda(\hat{\Sigma} - \text{diag}(s))\hat{\mathbf{B}}, \tag{4.17}$$

since $\hat{\Gamma}$ is the covariance matrix of SIR the first equation holds true, the second follows immediately using $A = (I - \hat{\Sigma}^{-1}diag\{s\})$. ■

This theorem shows that when the knockoff filter is added we can expect that SIR will privilege the true variables against the copies. Analyzing each component in the estimated directions $\hat{\mathbf{B}}_i$, $i = 1,..,k$ it is possible to compare the values obtained with the corresponding one of the copy. What is expected is that when the true direction $\mathbf{B}_i$ has non null components $\mathbf{B}_{i,1},...,\mathbf{B}_{i,p}$ the estimated values $\hat{\mathbf{B}}_{i,1},...,\hat{\mathbf{B}}_{i,p}$ and the copies will differ significantly. Therefore for each direction found by SIR on the concatenation $[\mathbf{X},\tilde{\mathbf{X}}]$ we claim that it is possible to distinguish the variables involved and the one that are not. In particular when different knockoff filters are applied to the same dataset $\mathbf{X}$ a statistic can be extracted, the components that behave like their copies are to be discarded while the ones that differ are to be considered informative.

Knockoff SIR in practice. Knockoff SIR proceeds first to the calculation of N different knockoff filters starting from $\mathbf{X} \in \mathbb{R}^{p\times n}$. Let us assume that the e.d.r space has dimension $k = 1$ to lighten the notation. The N e.d.r. directions found by SIR applied to the N concatenations produce $\tilde{\mathbf{B}}_1,...,\tilde{\mathbf{B}}_N$. The following p statistics are considered:

$$\{(\tilde{\mathbf{B}}_{1,i},...,\tilde{\mathbf{B}}_{N,i}),(\tilde{\mathbf{B}}_{1,i+p},...,\tilde{\mathbf{B}}_{N,i+p})\} \qquad \text{for} \quad i = 1,...,p \tag{4.18}$$

A Wilcoxon-Mann-Whitney test is used to establish if the two samples $\{\tilde{\mathbf{B}}_{1,i},...,\tilde{\mathbf{B}}_{N,i}\}$ and $\{\tilde{\mathbf{B}}_{1,i+p},...,\tilde{\mathbf{B}}_{N,i+p}\}$ are coming from the same distribution, which means the variable i **should not** be selected, or from different ones which means the variable i is informative and must be selected. This result enforce sparsity of the solution without any constraint on the number of non null entries (that is unknown). The following two sections are dedicated to show the results on simulated and real data.

4.3 Simulation results

In this section two test cases are considered with different e.d.r. space dimension ($k = 1, 2$).

Test case A: $k = 1$**.** Let us consider the following regression problem:

$$Y = (x_1 + x_2 + x_3 - 10)^2 + \varepsilon, \tag{4.19}$$

where $\mathbf{X} = (x_1, ..., x_{10}) \in \mathbb{R}^{10}$ is a vector of independent standard normal distributions and ε is a standard normal error independent of $\mathbf{X}$. In accordance to [52] we consider the True Inclusion Rate (TIR), the ratio of the number of correctly identified active predictors to the number of truly active predictors; and the False Inclusion Rate (FIR), the ratio of the number falsely identified active predictors to the total number of inactive predictors. In our test there are 3 active predictors and 7 inactive. A study on the sensitivity to the number of sample n is shown in Table 4.3. Results are obtained over 100 repetitions, in each repetition 1000 knockoff filters are generated to provide a statistic and the Wilcoxon-Mann-Whitney test at significance level $\alpha = 0.05$ has been applied to each of the p-predictors to establish if they must be considered active or inactive. In Table 4.3 the capability of Knockoff SIR are shown, it is evident how, even when the number of samples is small, good results are obtained with high value of TIR and low value of FIR.

n	TIR	FIR	#-slices
25	.81(.25)	.48(.20)	2
50	1(.0)	.16(.16)	5
75	1(.0)	.09(.12)	7
100	1(.0)	.08(.10)	10
150	1(.0)	.08(.11)	15
200	1(.0)	.06(.11)	20
250	1(.0)	.05(.08)	25
300	1(.0)	.04(.08)	30
400	1(.0)	.04(.06)	30

TABLE 4.1. Study on the sensitivity to the number of sample n, averages (and standard deviation in brackets) are obtained over 100 iterations. True Inclusion Rate (TIR) and False Inclusion Rate (FIR) are shown. The number of slices has been selected such that at least 10 samples are contained in each slice.

Selection of the number of e.d.r. directions k**.** An analysis of the active predictors for the second direction found by Knockoff SIR evidence that this method can be used to

select the dimension k. In the example $k = 1$ only the first component brings information. On the second direction found ($n = 200$) over 100 repetitions on the average 0.04 (0.06) predictors have been found active pointing out that this direction is not reliable. For the third direction the average of active predictors is decreasing (as expected for the property of SIR) to 0.01(.04), the trend is common in all directions.

Test case A: $k = 2$**.** Let us consider the following regression problem:

$$Y = \text{sign}\,(\beta_1^T \mathbf{X}) \log(|\beta_2^T \mathbf{X} + 5|) + 0.2\varepsilon \tag{4.20}$$

where $\mathbf{X} = (x_1, ..., x_{20}) \in \mathbb{R}^{20}$ is a vector of independent standard normal distributions and ε is a standard normal error independent of $\mathbf{X}$. Two configurations are considered for the e.d.r. directions β_1, β_2:

(i) $\beta_1 = (1,1,1,1,0,...,0)$, $\beta_2 = (0,...,0,1,1,1,1)$

(ii) $\beta_1 = (1,1,0.1,0.1,0,...,0)$, $\beta_2 = (0,...,0,0.1,0.1,1,1)$

For each configuration 100 replications are considered to evaluate the average TIR and FIR. The e.d.r. directions found by SIR have the property that Span$\{\hat{\beta}_1, \hat{\beta}_2\}$ =Span$\{\beta_1, \beta_2\}$ this does not imply that the e.d.r. directions correspond directly to the true. A generalization of TIR and FIR to multiple dimension is needed to account this property. TIR and FIR are calculated as follows:

- For the two e.d.r. directions compute two binary p-vectors, h_1 and h_2. When the i-variable is active assign one, otherwise zero.
- TIR is the ratio of the number of identified active components of $\max(h_1, h_2)$ over the number of truly active components of $\beta_1 \bigcup \beta_2$.
- FIR is the ration of the number of inactive components of $\max(h_1, h_2)$ over the number of inactive components of $\beta_1 \bigcup \beta_2$.

The results in Table 4.3 show that Knockoff SIR is capable of dealing with multiple dimension when the relative importance of the variable is equal (case (i)) but, in analogy with [49] has troubles in identifying the variables with a small relative importance. In the second configuration (ii) half of the active variables have 0.1 weight that is not retrieved by Knockoff SIR.

	TIR	FIR
(i)	.95(.09)	.01(.03)
(ii)	.52(.05)	.01(.03)

TABLE 4.2. Study under different configurations, $(i)-(ii)$, of the e.d.r. directions. The average and the standard deviation (in brackets) are calculated over 100 iterations.

Selection of the number of e.d.r. directions k. An extremely interesting behavior is shown when the procedure is applied to the third direction, the analysis of the eigenvalues in Figure 4.1 suggests that the first two directions are to be considered but the third and the following are uncertain. The result of Knockoff SIR shows that, in both configurations, all variables in the third direction are inactive. The same trend is observed in each direction associated with smaller eigenvalues.

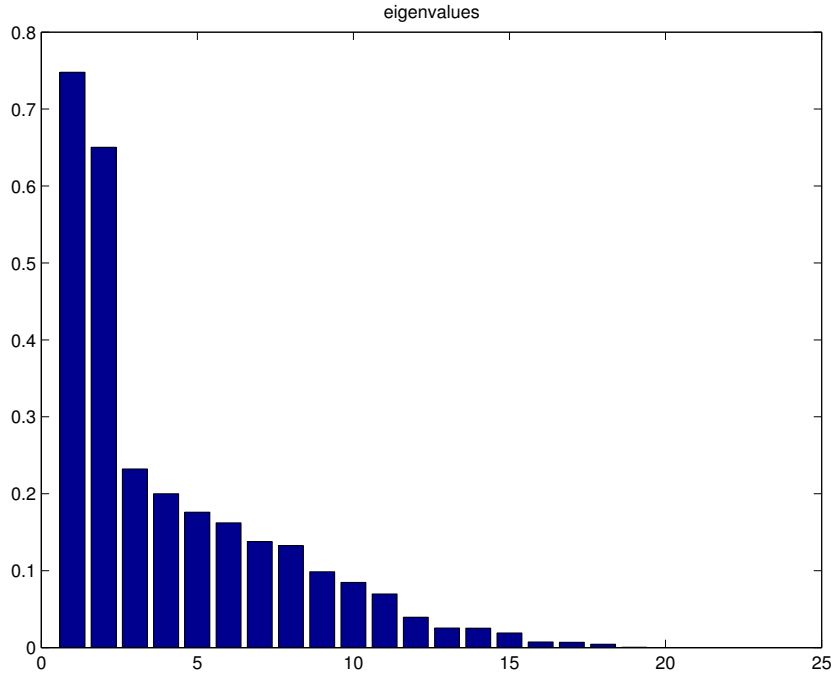


Figure 4.1: Barplot of the eigenvalues relative to the e.d.r. direction found by SIR applied to $\mathbf{X}$ with no knockoffs.

4.4 A real data application

The Galaxy dataset discussed in Collaborative SIR and Student SIR has been used as a real data application and to compare to the other contributions to verify if consistent results are given through the analysis. The Galaxy dataset corresponds to $n = 362,887$ different galaxies. This dataset is used in [15] with a preprocessing based on expert supervision to remove outliers. In this study all the original observations are considered, removing only points with missing values, which requires no expertise. The response variable Y is the stellar formation rate. The predictor space $\mathbf{X}$ is made of spectral characteristics of the galaxies and is of dimension $p = 46$. Knockoff SIR has been used to identify which are the active variables in the e.d.r. directions, from Student SIR we have a hint that the e.d.r. space dimension k should be 3. This is confirmed by the naive analysis of the eigenvalues in Figure 4.2 obtained by SIR on the original Galaxy data $\mathbf{X}$. Extensive tests have been made adding knockoff filters to the analysis. For eigenvectors corresponding to the first five higher eigenvalues 1000 different knockoff filters have been used to provide a statistic to assign to each variable an active or inactive label. In the first e.d.r. directions the only active variables found are $\{2,3,23,40,45\}$ matching the results of Collaborative SIR, the variable 6 selected by Collaborative SIR is not estimated active in any of the first three e.d.r directions, doubts can be cast to the selection of variable 6 for the analysis. According to the result of BIC in Student SIR we tested the e.d.r. directions relative to the five highest eigenvalues obtaining that for the first three e.d.r. directions active variables have been found. Grouping the active variables through the first three directions gives only seven variables: $\{2,3,23,40,42,43,45\}$. This means that by default the analysis could be directly performed on the seven predictors avoiding the other 39. The fourth and fifth and further directions with smaller eigenvalues resulted with no active variables supporting the decision of Student SIR.

Comments. Knockoff SIR is a procedure that rather than providing sparse solutions provides insight to orient the analysis selecting only some variables in the predictor space. If inactive variables through all the considered directions are found they can be removed from the beginning as a preliminary dimension reduction before rerunning SIR. The use of Knockoff SIR in the selection of dimension k, even if no theoretical results have been established is of great interest. Further studies will be dedicated to a better understanding of this phenomena and to find better statistics to discard active and inactive variables. The computational complexity is for each run the one of SIR on the concatenation of $\mathbf{X}$ and its knockoff filter.

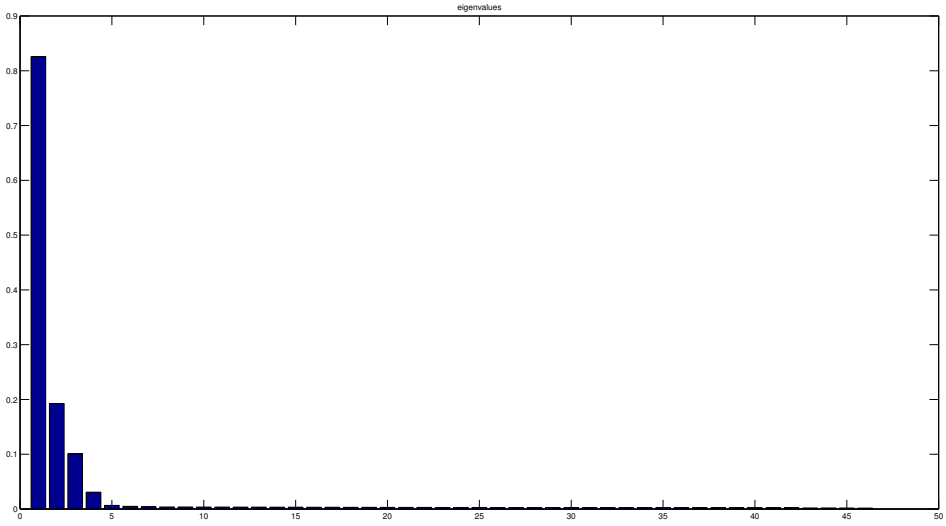


Figure 4.2: Barplot of the eigenvalues relative to the e.d.r. direction found by SIR applied to the Galaxy dataset $\mathbf{X}$ with no knockoffs.

CHAPTER 5

CONCLUSION

I realized then that a man who had lived
only one day could easily live for a hundred years in prison.
He would have enough memories to keep him from being bored
A.Camus.

Dimension reduction is a broad field that can be seen through the eyes of a Mathematician as well as from a Poet. Both approaches select or enhance different characteristics of a phenomena. In this thesis three different extensions of the well known method SIR have been proposed. Each contribution is focusing on a different aspect of the original method trying to improve or at least to better explore under which conditions this method can be applied, when could not work and what can bring to the analysis. I believe that Statistics is a powerful tool to analyze and look at the data, but I also tend to agree completely with the following sentence from R.D. Cook extracted by [20]:

Findings that are not accompanied by an understanding of how the data and model interacted to produce them should ordinarily be accompanied by a good dose of skepticism.

The increasing capability of the technology to gather data is bringing new exciting problems to Statistics and dimension reduction seems to be of greater importance now. A result of an algorithm, though, should be carefully analyzed before drawing conclusions on the phenomena itself. Tim Harford delivered the 2014 Significance lecture at the

Royal Statistical Society International Conference where he raised a question: *Big data: are we making a big mistake?*. The idea of Harford is that when the dimension increases the traps lying in the data become more and more tricky to discover. In my personal view I see an analogy with the recycling process. If all products are thrown in the same bucket is really hard to find a fast and effective solution to differentiate categories, indeed many countries started the process of recycling from each house, separating in categories before the "analysis". As is easy to see in the Galaxy dataset from the three different analysis it appears that few variables are contributing to forecast the stellar formation rate, experts should be more careful and try to avoid to cluster variables having faith in statistical approaches to discard what is not informative. The understanding of the original problem should be good enough to judge the results of the algorithm that can orient further analysis. The choice to analyze the Galaxy data in all our contributions gave us the possibility to compare and check if the results obtained under different methodologies were consistent. It is indeed the case and is encouraging for further analysis of this dataset in collaboration with experts.

In the short term our project is to organize the material regarding Knockoff SIR in a paper to have a feedback from the community. Extensions in the framework of Student SIR are possible introducing a generalization of the multivariate t-Student distribution with variable marginal amounts of tailweight. The different aspects of dimensionality reduction explored brought me to develop a genuine interest for Computational Topology which is under investigation thanks to the links established at TU Graz.

Not part of this thesis but part of my PhD is a paper [17] on classification published in Pattern Recognition Letters and an unpublished paper with philosophical reasoning on dimensionality reduction and classification that can be found here (by far my best work).

BIBLIOGRAPHY

[1] D. F. ANDREWS AND C. L. MALLOWS, *Scale mixtures of normal distributions*, Journal of the Royal Statistical Society. Series B (Methodological), 36 (1974), pp. 99–102.

[2] Y. ARAGON, *A Gauss implementation of multivariate sliced inverse regression*, Computational Statistics, 12 (1997), pp. 355–372.

[3] Y. ARAGON AND J. SARACCO, *Sliced inverse regression (SIR): an appraisal of small sample alternatives to slicing*, Compitational Statistics, 12 (1997), pp. 109–130.

[4] R. AZAÏS, A. GÉGOUT-PETIT, AND J. SARACCO, *Optimal quantization applied to sliced inverse regression*, Journal of Statistical Planning and Inference, 142 (2012), pp. 481–492.

[5] R. F. BARBER AND E. J. CANDÈS, *Controlling the false discovery rate via knockoffs*, The Annals of Statistics, 43 (2015), pp. 2055–2085.

[6] L. BARREDA, A. GANNOUN, AND J. SARACCO, *Some extensions of multivariate sliced inverse regression*, Journal of Statistical Computation and Simulation, 77 (2007), pp. 1–17.

[7] M. P. BARRIOS AND S. VELILLA, *A bootstrap method for assessing the dimension of a general regression problem*, Statistics & probability letters, 77 (2007), pp. 247–255.

[8] B. BERCU, T. M. N. NGUYEN, AND J. SARACCO, *A new approach on recursive and non-recursive sir methods*, Journal of the Korean Statistical Society, 41 (2012), pp. 17–36.

[9] C. BERNARD-MICHEL, S. DOUTÉ, M. FAUVEL, L. GARDES, AND S. GIRARD, *Retrieval of Mars surface physical properties from OMEGA hyperspectral images*

using regularized sliced inverse regression, Journal of Geophysical Research: Planets, 114 (2009).

[10] C. BERNARD-MICHEL, L. GARDES, AND S. GIRARD, *Gaussian regularized sliced inverse regression*, Statistics and Computing, 19 (2009), pp. 85–98.

[11] D. R. BRILLINGER AND P. R. KRISHNAIAH, *Time series in the frequency domain*, Handbook of Statistics, Amsterdam: North-Holland, 1983, edited by Brillinger, DR; Krishnaiah, PR, 1 (1983).

[12] M. CHAVENT, S. GIRARD, V. KUENTZ-SIMONET, B. LIQUET, T. M. N. NGUYEN, AND J. SARACCO, *A sliced inverse regression approach for data stream*, Computational Statistics, 29 (2014), pp. 1129–1152.

[13] M. CHAVENT, V. KUENTZ, B. LIQUET, AND J. SARACCO, *A sliced inverse regression approach for a stratified population*, Communications in Statistics-Theory and Methods, 40 (2011), pp. 3857–3878.

[14] C.-H. CHEN AND K.-C. LI, *Can SIR be as popular as multiple linear regression?*, Statistica Sinica, 8 (1998), pp. 289–316.

[15] A. CHIANCONE, S. GIRARD, AND J. CHANUSSOT, *Collaborative sliced inverse regression*, Communications in Statistics - Theory and Methods, (Accepted for publication).

[16] F. CHIAROMONTE AND J. MARTINELLI, *Dimension reduction strategies for analyzing global gene expression data with a response*, Mathematical Biosciences, 176 (2002), pp. 123–144.

[17] M. CHINI, A. CHIANCONE, AND S. STRAMONDO, *Scale object selection (sos) through a hierarchical segmentation by a multi-spectral per-pixel classification*, Pattern Recognition Letters, 49 (2014), pp. 214–223.

[18] R. COOK, L. FORZANI, AND A. YAO, *Necessary and sufficient conditions for consistency of a method for smoothed functional inverse regression*, Statistica Sinica, (2010), pp. 235–238.

[19] R. D. COOK, *Fisher lecture: Dimension reduction in regression*, Statistical Science, 22 (2007), pp. 1–26.

[20] R. D. COOK, *Reflections on a statistical career and their implications*, Past, Present, and Future of Statistical Science, (2014), pp. 97–108.

[21] R. D. COOK AND L. FORZANI, *Likelihood-based sufficient dimension reduction*, Journal of the American Statistical Association, 104 (2009), pp. 197–208.

[22] R. D. COOK, L. FORZANI, ET AL., *Principal fitted components for dimension reduction in regression*, Statistical Science, 23 (2008), pp. 485–501.

[23] R. D. COOK, L. FORZANI, AND D. R. TOMASSI, *Ldr: a package for likelihood-based sufficient dimension reduction*, Journal of Statistical Software, 39 (2011), pp. 1–20.

[24] R. D. COOK AND C. J. NACHTSHEIM, *Reweighting to achieve elliptically contoured covariates in regression*, Journal of the American Statistical Association, 89 (1994), pp. 592–599.

[25] R. D. COOK AND S. WEISBERG, *Sliced inverse regression for dimension reduction: Comment*, Journal of the American Statistical Association, 86 (1991), pp. 328–332.

[26] R. COUDRET, S. GIRARD, AND J. SARACCO, *A new sliced inverse regression method for multivariate response*, Computational Statistics & Data Analysis, 77 (2014), pp. 285–299.

[27] R. DENNIS COOK, *Save: a method for dimension reduction and graphics in regression*, Communications in statistics-Theory and methods, 29 (2000), pp. 2109–2121.

[28] P. DIACONIS AND D. FREEDMAN, *Asymptotics of graphical projection pursuit*, The Annals of Statistics, 12 (1984), pp. 793–815.

[29] S. DING AND R. D. COOK, *Tensor sliced inverse regression*, Journal of Multivariate Analysis, 133 (2015), pp. 216–231.

[30] Y. DONG, Z. YU, AND L. ZHU, *Robust inverse regression for dimension reduction*, Journal of Multivariate Analysis, 134 (2015), pp. 71–81.

[31] L. FERRÉ, *Determining the dimension in sliced inverse regression and related methods*, Journal of the American Statistical Association, 93 (1998), pp. 132–140.

[32] L. FERRÉ AND A.-F. YAO, *Functional sliced inverse regression analysis*, Statistics, 37 (2003), pp. 475–488.

[33] ——, *Smoothed functional inverse regression*, Statistica Sinica, (2005), pp. 665–683.

[34] K. FUKUMIZU, F. R. BACH, AND M. I. JORDAN, *Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces*, The Journal of Machine Learning Research, 5 (2004), pp. 73–99.

[35] ——, *Kernel dimension reduction in regression*, The Annals of Statistics, 37 (2009), pp. 1871–1905.

[36] U. GATHER, T. HILKER, AND C. BECKER, *A robustified version of sliced inverse regression*, in Statistics in Genetics and in the Environmental Sciences, Trends in Mathematics, Springer, 2001, ch. 2, pp. 147–157.

[37] ——, *A note on outlier sensitivity of sliced inverse regression*, Statistics: A Journal of Theoretical and Applied Statistics, 36 (2002), pp. 271–281.

[38] P. HALL AND K.-C. LI, *On almost linearity of low dimensional projections from high dimensional data*, The Annals of Statistics, 21 (1993), pp. 867–889.

[39] W. HARDLE AND A. TSYBAKOV, *Sliced inverse regression for dimension reduction: Comment*, Journal of the American Statistical Association, 86 (1991), pp. 333–335.

[40] N. L. JOHNSON, S. KOTZ, AND N. BALAKRISHNAN, *Continuous Univariate Distributions, vol.2, 2nd edition*, John Wiley & Sons, New York, 1994.

[41] J. T. KENT, *Comment*, Journal of the American Statistical Association, 86 (1991), pp. 336–337.

[42] S. KOTZ AND S. NADARAJAH, *Multivariate t Distributions and their Applications*, Cambridge, 2004.

[43] V. KUENTZ, B. LIQUET, AND J. SARACCO, *Bagging versions of sliced inverse regression*, Communications in Statistics-Theory and Methods, 39 (2010), pp. 1985–1996.

[44] V. KUENTZ AND J. SARACCO, *Cluster-based sliced inverse regression*, Journal of the Korean Statistical Society, 39 (2010), pp. 251–267.

[45] B. LI AND Y. DONG, *Dimension reduction for nonelliptically distributed predictors*, The Annals of Statistics, 37 (2009), pp. 1272–1298.

[46] B. LI, M. K. KIM, AND N. ALTMAN, *On dimension folding of matrix-or array-valued statistical objects*, The Annals of Statistics, 38 (2010), pp. 1094–1121.

[47] K.-C. LI, *Sliced inverse regression for dimension reduction*, Journal of the American Statistical Association, 86 (1991), pp. 316–327.

[48] K.-C. LI, Y. ARAGON, K. SHEDDEN, AND C. THOMAS AGNAN, *Dimension reduction for multivariate response data*, Journal of the American Statistical Association, 98 (2003), pp. 99–109.

[49] L. LI, *Sparse sufficient dimension reduction*, Biometrika, 94 (2007), pp. 603–613.

[50] L. LI, R. D. COOK, AND C. J. NACHTSHEIM, *Cluster-based estimation for sufficient dimension reduction*, Computational Statistics & Data Analysis, 47 (2004), pp. 175–193.

[51] L. LI AND H. LI, *Dimension reduction methods for microarrays with application to censored survival data*, Bioinformatics, 20 (2004), pp. 3406–3412.

[52] L. LI AND C. J. NACHTSHEIM, *Sparse sliced inverse regression*, Technometrics, 48 (2006), pp. 503–510.

[53] B. LIQUET AND J. SARACCO, *Application of the bootstrap approach to the choice of dimension and the α parameter in the sirα method*, Communications in Statistics—Simulation and Computation, 37 (2008), pp. 1198–1218.

[54] ——, *A graphical tool for selecting the number of slices and the dimension of the model in sir and save approaches*, Computational Statistics, 27 (2012), pp. 103–125.

[55] H.-H. LUE, *Sliced inverse regression for multivariate response regression*, Journal of Statistical Planning and Inference, 139 (2009), pp. 2656–2664.

[56] R. LUO, H. WANG, AND C.-L. TSAI, *Contour projected dimension reduction*, The Annals of Statistics, 37 (2009), pp. 3743–3778.

[57] M. MANGEL AND F. J. SAMANIEGO, *Abraham wald's work on aircraft survivability*, Journal of the American Statistical Association, 79 (1984), pp. 259–267.

[58] J. SARACCO, *An asymptotic theory for sliced inverse regression*, Communications in Statistics-Theory and Methods, 26 (1997), pp. 2141–2171.

[59] J. R. SCHOTT, *Determining the dimensionality in sliced inverse regression*, Journal of the American Statistical Association, 89 (1994), pp. 141–148.

[60] L. SCRUCCA, *Regularized sliced inverse regression with applications in classification*, in Data Analysis, Classification and the Forward Search, Springer, 2006, pp. 59–66.

[61] ——, *Class prediction and gene selection for dna microarrays using regularized sliced inverse regression*, Computational Statistics & Data Analysis, 52 (2007), pp. 438–451.

[62] ——, *Model-based sir for dimension reduction*, Computational Statistics & Data Analysis, 55 (2011), pp. 3010–3026.

[63] C. M. SETODJI AND R. D. COOK, *K-means inverse regression*, Technometrics, 46 (2004), pp. 421–429.

[64] S. J. SHEATHER AND J. W. MCKEAN, *A comparison of procedures based on inverse regression*, Lecture Notes-Monograph Series, 31 (1997), pp. 271–278.

[65] S. M. STIGLER, *Gauss and the invention of least squares*, The Annals of Statistics, 9 (1981), pp. 465–474.

[66] G. WANG, J. ZHOU, W. WU, AND M. CHEN, *Robust functional sliced inverse regression*, Statistical Papers, (2015), pp. 1–19.

[67] H. WANG, L. NI, AND C.-L. TSAI, *Improving dimension reduction via contour-projection*, Statistica Sinica, 18 (2008), pp. 299–311.

[68] T. WANG, X. M. WEN, AND L. ZHU, *Multiple-population shrinkage estimation via sliced inverse regression*, Statistics and Computing, (2015), pp. 1–12.

[69] Q. WU, *Regularized sliced inverse regression for kernel models*, PhD thesis, Duke University, Durham, 2007.

[70] X.-L. XU, C.-X. REN, R.-C. WU, AND H. YAN, *Sliced inverse regression with adaptive spectral sparsity for dimension reduction*, IEEE Transactions on Cybernetics, PP (2016), pp. 1–13.

[71] Y.-R. YEH, S.-Y. HUANG, AND Y.-J. LEE, *Nonlinear dimension reduction with kernel sliced inverse regression*, IEEE Transactions on Knowledge and Data Engineering, 21 (2009), pp. 1590–1603.

[72] X. YIN, B. LI, AND R. D. COOK, *Successive direction extraction for estimating the central subspace in a multiple-index regression*, Journal of Multivariate Analysis, 99 (2008), pp. 1733–1757.

[73] W. ZHONG, P. ZENG, P. MA, J. S. LIU, AND Y. ZHU, *Rsir: regularized sliced inverse regression for motif discovery*, Bioinformatics, 21 (2005), pp. 4169–4175.

[74] J. ZHOU, *Robust dimension reduction based on canonical correlation*, Journal of Multivariate Analysis, 100 (2009), pp. 195–209.

[75] L.-P. ZHU AND Z. YU, *On spline approximation of sliced inverse regression*, Science in China Series A: Mathematics, 50 (2007), pp. 1289–1302.

[76] L.-X. ZHU, K.-T. FANG, ET AL., *Asymptotics for kernel estimate of sliced inverse regression*, The Annals of Statistics, 24 (1996), pp. 1053–1068.

[77] L.-X. ZHU AND K. W. NG, *Asymptotics of sliced inverse regression*, Statistica Sinica, 5 (1995), pp. 727–736.