



HAL
open science

A new sliced inverse regression method for multivariate response regression

Raphaël Coudret, Stéphane Girard, Jerome Saracco

► **To cite this version:**

Raphaël Coudret, Stéphane Girard, Jerome Saracco. A new sliced inverse regression method for multivariate response regression. 2013. hal-00714981v2

HAL Id: hal-00714981

<https://inria.hal.science/hal-00714981v2>

Preprint submitted on 6 Mar 2013 (v2), last revised 5 Oct 2013 (v3)

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.

A new sliced inverse regression method for multivariate response regression

Raphaël Coudret¹, Stéphane Girard³ and Jérôme Saracco^{1,2}

¹*INRIA Bordeaux Sud-Ouest and Institut de Mathématiques de Bordeaux*, ²*Institut Polytechnique de Bordeaux*,

³*INRIA Grenoble Rhône-Alpes and Laboratoire Jean Kuntzmann*.

Abstract

We consider a semiparametric regression model of a q -dimensional multivariate response \mathbf{y} on a p -dimensional covariate \mathbf{x} . In this paper, a new approach is proposed based on sliced inverse regression (SIR) for estimating the effective dimension reduction (EDR) space without requiring a prespecified parametric model. The convergence at rate \sqrt{n} of the estimated EDR space is shown. We discuss the choice of the dimension of the EDR space. Moreover, we provide a way to cluster components of \mathbf{y} related to the same EDR space. One can thus use properly the proposed multivariate SIR method on each cluster instead of blindly applying it on all components of \mathbf{y} . The numerical performances of multivariate SIR are illustrated on a simulation study. Applications to a remote sensing dataset and to the Minneapolis elementary schools data are also provided.

Keywords: dimension reduction, semiparametric regression model, multivariate response, sliced inverse regression.

1 Introduction

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

In this paper, a q -dimensional response variable \mathbf{y} is considered. Hence, we deal with a high dimensional regression framework which is not a linear one or a prespecified parametric one. The

underlying idea of the dimension reduction of the explanatory variable \mathbf{x} without loss of information is to identify the linear combinations $\beta'_1\mathbf{x}, \dots, \beta'_K\mathbf{x}$ such that

$$\mathbf{y} \perp \mathbf{x} | (\beta'_1\mathbf{x}, \dots, \beta'_K\mathbf{x}), \quad (1)$$

where \perp denotes independence, $K(\leq p)$ is as small as possible and the p -dimensional vectors β_k are linearly independent. Let $\mathbf{B} = [\beta_1, \dots, \beta_K]$ denote the $p \times K$ matrix of the β_k 's. Statement (1) means that $\mathbf{y}|\mathbf{x}$ and $\mathbf{y}|\mathbf{B}'\mathbf{x}$ share the same distribution for all values of \mathbf{x} . A straightforward consequence is that the p -dimensional covariate \mathbf{x} can be replaced by the K -dimensional predictor $\mathbf{B}'\mathbf{x}$ without loss of regression information. The goal of dimension reduction is achieved for $K < p$. As mentioned in Li (1991) or Cook (1994), statement (1) is equivalent to $\mathbf{y} \perp \mathbf{x} | P_{\mathbf{B}}\mathbf{x}$, where $P_{\mathbf{B}}$ denotes the projection operator on $\text{Span}(\mathbf{B})$ which is the linear subspace of \mathbb{R}^p spanned by the columns of \mathbf{B} . In addition, $\text{Span}(\mathbf{B})$ can be viewed as the EDR space. From a regression model point of view, one can mention that the corresponding underlying model is the following semiparametric one

$$\mathbf{y} = f(\mathbf{B}'\mathbf{x}, \boldsymbol{\varepsilon}), \quad (2)$$

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

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

$$\begin{cases} y^{(1)} = f_1(\mathbf{B}'\mathbf{x}, \varepsilon^{(1)}) \\ \vdots \\ y^{(q)} = f_q(\mathbf{B}'\mathbf{x}, \varepsilon^{(q)}) \end{cases} \quad (3)$$

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

Remark 1. The information from the marginal regression is sufficient to recover the whole EDR space in model (3). However, this is not always the case when working with model (2). Let us illustrate this point with the following regression model proposed by Zhu, Zhu, and Wen (2010):

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

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

The vectors β_k are not individually identifiable neither in model (2), nor in the more restrictive one (3). Thus, the main objective is to estimate a basis of the K -dimensional EDR space. Many numerical methods have been introduced to achieve this goal. Let us mention three of them which are relatively simple and easy to implement: SIR introduced by [Duan and Li \(1991\)](#) for the single index model ($K = 1$) and [Li \(1991\)](#) for the multiple indices model ($K > 1$), principal Hessian directions (see for instance [Li \(1992\)](#) or [Cook \(1998\)](#)), and sliced average variance estimation (see for details [Cook \(2000\)](#), [Prendergast \(2007\)](#) or [Shao, Cook, and Weisberg \(2009\)](#)). In this paper, we shall only focus on the SIR approach which is based on a property of the first moment of the inverse distribution of \mathbf{x} given \mathbf{y} .

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

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

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

As previously mentioned, the proposed approach to estimate the EDR space relies on combining estimates from the marginal regressions of model (3). Moreover, we can naturally take the information from these regressions into account to detect if a common EDR space really exists for all the components of \mathbf{y} as in model (3). Otherwise, one shall consider the following more general

regression model for multivariate response regression:

$$\left\{ \begin{array}{l} y^{(1)} = f_1(\mathbf{B}'_1 \mathbf{x}, \varepsilon^{(1)}), \\ \vdots = \vdots \\ y^{(q_1)} = f_{q_1}(\mathbf{B}'_1 \mathbf{x}, \varepsilon^{(q_1)}), \\ y^{(q_1+1)} = f_{q_1+1}(\mathbf{B}'_2 \mathbf{x}, \varepsilon^{(q_1+1)}), \\ \vdots = \vdots \\ y^{(q_1+q_2)} = f_{q_1+q_2}(\mathbf{B}'_2 \mathbf{x}, \varepsilon^{(q_1+q_2)}), \\ \vdots = \vdots \\ y^{(q)} = f_q(\mathbf{B}'_L \mathbf{x}, \varepsilon^{(q)}), \end{array} \right. \quad (4)$$

where, for every $l = 1, \dots, L$, \mathbf{B}_l is a $p \times K$ matrix, K is assumed to be known, $\text{Span}(\mathbf{B}_1) \neq \text{Span}(\mathbf{B}_2) \neq \dots \neq \text{Span}(\mathbf{B}_L)$ and $\sum_{l=1}^L q_l = q$. This means that writing this model as in (3) requires the number of columns of \mathbf{B} to be greater than K . When trying to reduce as much as possible the dimension of a model, estimating $\mathbf{B}_1, \dots, \mathbf{B}_L$ seems thus more appropriate than seeking \mathbf{B} . One then needs to cluster the components of \mathbf{y} associated with the same EDR space. Therefore, for each identified cluster of components, one can use only these components to estimate the corresponding (common) EDR space. In a more general case, one can also assume that the dimension K is specific for each \mathbf{B}_l .

The goal of this paper is twofold. First we propose a new multivariate SIR approach for estimating the K -dimensional EDR space which is common to the q components of the multivariate response variable in model (3). Then we propose a way to cluster the components of \mathbf{y} associated with the same EDR space in model (4) in order to apply properly our multivariate SIR on each cluster instead of blindly applying it on all the components of \mathbf{y} .

The paper is organized as follows. In Section 2, we give a brief overview on usual univariate SIR and existing multivariate SIR methods. The population version of the new SIR approach for a multivariate response, named MSIR hereafter, is described in Section 3.1. The corresponding sample version is introduced in Section 3.2 and asymptotic results are provided in Section 3.3. A weighted version of MSIR, named wMSIR hereafter, is proposed in Section 3.4. Both these methods rely on a tuning parameter H which is the number of slices. The choice of H and of the dimension K is discussed in Section 3.5. Practical methods to investigate the possible existence of a common EDR space for \mathbf{y} and to detect and identify clusters of components of \mathbf{y} are proposed in Section 3.6. Numerical results based on simulations are exhibited in Section 4 in order to show the good behavior of MSIR and wMSIR approaches and the usefulness of the diagnostic and clustering procedures on the components of \mathbf{y} . In Section 5, two real datasets are considered: the first one

concerns hyperspectral measures while the second one is the widely studied Minneapolis elementary schools dataset. Finally, concluding remarks are given in Section 6.

2 Brief review of univariate and multivariate SIR approaches

In this section, we consider the regression model (3). We first provide an overview of the SIR method when the response y is univariate. Then, some existing SIR methods for a multivariate response are briefly described. The aim of all these approaches is to estimate the EDR space.

2.1 Univariate SIR

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

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

Let T denote a monotone (but not necessarily strictly monotone) transformation of y . Let $\mu = \mathbb{E}(\mathbf{x})$ and $\Sigma = \mathbb{V}(\mathbf{x})$. We assumed that $\mathbb{E}((\mathbf{x}'\mathbf{x})^2) < \infty$ and that Σ is invertible. Under model (3) and (LC), Li (1991) established the following geometric property: the centered inverse regression curve, $\mathbb{E}(\mathbf{x}|T(y)) - \mu$ as y varies, is contained in the linear subspace of \mathbb{R}^p spanned by $\Sigma\mathbf{B}$. A straightforward consequence is that the covariance matrix

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

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

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

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

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

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

From a practical point of view, the choice of the slicing is discussed in Li (1991), Chen and Li (1998) or Saracco (2001). Since SIR theory makes no assumption about the slicing strategy, the user must choose the number H of slices and how to construct them. In practice, there are naturally two possibilities: to fix the width of the slices or to fix the number of observations per slice. This second option is often preferred, and from the sample point of view, the slices are often chosen such that the number of observations in each slice is as close to each other as possible. Note that H must be greater than K in order to avoid an artificial reduction of dimension and must be lower than $\lfloor n/2 \rfloor$ in order to have at least two observations in each slice (where $\lfloor a \rfloor$ denotes the integer part of a). The choice of H is less sensitive than the choice of a smoothing parameter in nonparametric regression. This point is clearly illustrated in Lique and Saracco (2012) with a graphical tool that allows the user to find simultaneously realistic values of the two parameters H and K , see Section 3.5 for some details on this method.

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

2.2 Multivariate SIR

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

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

A natural way to circumvent the curse of dimensionality of the complete slicing approach is proposed in the marginal slicing procedure which consists in applying the SIR method on a transformation of \mathbf{y} depending on one's interest. For instance, it can be the mean or the median of the $y^{(j)}$'s. One can also take the first few significant components of a principal component analysis of the $y^{(j)}$'s to construct the slices. However, slicing a lower dimensional projection of \mathbf{y} may not lead to recover as many EDR directions as slicing the entire \mathbf{y} .

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

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

$$\mathbf{\Gamma}_P = \sum_{j=1}^q w_j \mathbf{\Gamma}^{(j)}, \quad (5)$$

for positive weights w_j . It has been shown that the eigenvectors associated with the non-null K eigenvalues of $\mathbf{\Sigma}^{-1} \mathbf{\Gamma}_P$ are EDR directions. [Aragon \(1997\)](#) proposes to use two kinds of weighting

for the w_j 's: equal weights or weights proportional to the major eigenvalues found by a preliminary univariate SIR analysis of each component of \mathbf{y} . Saracco (2005) obtained the asymptotic normality of the pooled marginal slicing estimator based on SIR_α . Lue (2009) derived the asymptotic weighted chi-squared test for dimension. For $j = 1, \dots, q$, rather than constructing $\mathbf{\Gamma}^{(j)}$ from $y^{(j)}$, one can also build it from a linear combination $\tau'\mathbf{y}$ of \mathbf{y} . This method which is called projective resampling was introduced by Li, Wen, and Zhu (2008). To ensure good performances, the number of linear combinations to handle should be greater than the sample size n .

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

3 A new multivariate SIR approach

The population version of the proposed MSIR approach is first described in Section 3.1. Then, the corresponding sample version is given in Section 3.2 and some asymptotic results are derived in Section 3.3. We then modify MSIR to handle a weighting in the components of \mathbf{y} in Section 3.4. Methods to choose K and H are discussed in Section 3.5. Finally, procedures to withdraw or cluster

components of \mathbf{y} are detailed in Section 3.6.

3.1 Population version

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

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

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

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

$$Q(\mathbf{D}, \mathbf{B}^{(1)}, \dots, \mathbf{B}^{(q)}) := \frac{1}{q} \sum_{j=1}^q r(\mathbf{D}, \boldsymbol{\Sigma}, \mathbf{B}^{(j)}, \boldsymbol{\Sigma}). \quad (6)$$

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

Let us now consider the following optimization problem:

$$\mathbf{V} := \arg \max_{\mathbf{D} \in \mathcal{M}_{\boldsymbol{\Sigma}}} Q(\mathbf{D}, \mathbf{B}^{(1)}, \dots, \mathbf{B}^{(q)}), \quad (7)$$

where $\mathcal{M}_{\boldsymbol{\Sigma}}$ is the set of $\boldsymbol{\Sigma}$ -orthogonal $p \times K$ matrices. A solution of (7) is given by the following theorem.

Theorem 1. *Under model (3) and assumption (LC), the $p \times K$ matrix \mathbf{V} is formed by the eigenvectors v_1, \dots, v_K associated with the K non-null eigenvalues of $\mathbb{B}\mathbb{B}'\boldsymbol{\Sigma}$ where \mathbb{B} is the $p \times (Kq)$ matrix defined as $\mathbb{B} := [\mathbf{B}^{(1)}, \dots, \mathbf{B}^{(q)}]$. Moreover we have $\text{Span}(\mathbf{V}) = \text{Span}(\mathbf{B})$ where $\text{Span}(\mathbf{B})$ is the EDR space.*

The proof is given in Appendix A.2. From Theorem 1, we can estimate a basis of the EDR space based on estimators of matrices \mathbb{B} and $\boldsymbol{\Sigma}$. This is the goal of the next subsection.

3.2 Sample version

Let $\{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \dots, n\}$ be a sample of independent observations from model (3). Each \mathbf{y}_i is a q -dimensional random variable. We assume that the sample size n is larger than the dimension p of each covariate \mathbf{x}_i .

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

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

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

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

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

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

Remark 2. Similarly to the pooled marginal slicing (PMS) presented in Section 2.2, MSIR relies on the univariate version of SIR, applied to each component of \mathbf{y} . While both methods need estimates of $\boldsymbol{\Gamma}^{(1)}, \dots, \boldsymbol{\Gamma}^{(q)}$, estimates of $\mathbf{B}^{(1)}, \dots, \mathbf{B}^{(q)}$ are only required by MSIR. Computing such estimates is useful to explore the relations between components of \mathbf{y} as explained in Section 3.6.

3.3 An asymptotic result

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

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

Theorem 2. Under model (3) and assumptions (LC) and (A1)-(A3), we have, for $k = 1, \dots, K$,

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

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

The proof is given in Appendix A.3.

Remark 3. Using Delta-method and asymptotic results of Tyler (1981) and Saracco (1997), it is possible to obtain the asymptotic normality of

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

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

3.4 A weighted version of MSIR

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

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

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

with $\mathbf{W}^{(j)} = \frac{\pi^{(j)}}{\pi_\star} \mathbf{I}_K$ for $j = 1, \dots, q$. Note that, from a theoretical point of view, under model (3) and (LC), the matrix of weights is such that $\mathbb{W} = \frac{1}{q} \mathbf{I}_{qK}$ since $\lambda_k^{(j)} = 0$ for $j = 1, \dots, q$ and $k = K + 1, \dots, p$.

The population version of wMSIR consists in noticing that the eigenvectors $\tilde{v}_1, \dots, \tilde{v}_K$ associated with the K largest eigenvalues of the Σ -symmetric matrix $\mathbb{B}\mathbb{W}\mathbb{B}'\Sigma$ span the EDR space. We write $\tilde{\mathbf{V}} := [\tilde{v}_1, \dots, \tilde{v}_K]$. To show this result, one can proceed analogously to the proof of Theorem 1.

Let $\hat{\lambda}_k^{(j)}$ be the k th eigenvalue of $\widehat{\Sigma}^{-1}\widehat{\Gamma}^{(j)}$. The sample version of wMSIR is obtained by substituting the empirical matrices $\widehat{\mathbb{B}}$, $\widehat{\mathbb{W}}$ and $\widehat{\Sigma}$ for their theoretical counterparts \mathbb{B} , \mathbb{W} and Σ , where $\widehat{\mathbb{W}} = \text{diag}(\widehat{\mathbb{W}}^{(1)}, \dots, \widehat{\mathbb{W}}^{(q)})$ with, for $j = 1, \dots, q$, $\widehat{\mathbb{W}}^{(j)} = \frac{\hat{\pi}^{(j)}}{\hat{\pi}_*} \mathbf{I}_K$, $\hat{\pi}^{(j)} = \frac{\sum_{k=1}^K \hat{\lambda}_k^{(j)}}{\sum_{k=1}^p \hat{\lambda}_k^{(j)}}$ and $\hat{\pi}_* = \sum_{j=1}^q \hat{\pi}^{(j)}$. Therefore, one can get the corresponding estimated EDR directions $\widehat{\mathbf{V}} := [\widehat{\mathbf{v}}_1, \dots, \widehat{\mathbf{v}}_K]$.

Theorem 3. *Under model (3) and assumptions (LC) and (A1)-(A3), we have, for $k = 1, \dots, K$,*

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

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

The proof is given in Appendix A.4.

3.5 Discussion on the choice of K and H

Up to now, the dimension K of the EDR space was assumed to be known. However, in most applications based on real datasets, the number K of indices $\beta'_k \mathbf{x}$ in model (1) is a priori unknown and hence must be determined from the data. In addition, the number of slices H in MSIR has to be chosen.

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

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

$$R_{h,k} := \mathbb{E} \left(r(\mathbf{V}_k, \Sigma, \widehat{\mathbf{V}}_k, \widehat{\Sigma}) \right), \quad (8)$$

where $\mathbf{V}_k := [v_1, \dots, v_k]$, $\widehat{\mathbf{V}}_k := [\widehat{v}_1, \dots, \widehat{v}_k]$ and h is the number of slices used to obtain \mathbf{V}_k and $\widehat{\mathbf{V}}_k$. This risk function only makes sense for any dimension k lower than or equal to the true dimension K of the EDR space. For the true dimension K , $R_{h,K}$ converges to one as n tends to infinity. For a fixed n , a reasonable way to assess whether an EDR direction is available is to graphically evaluate how much $R_{h,k}$ departs from one. From a computational point of view, consistent estimates of $R_{h,k}$ are required. Liquet and Saracco (2012) use a bootstrap estimator $\widehat{R}_{h,k}$ of this criterion in order to determine the pair (H, K) of parameters. The proposed graphical method consists in evaluating

the $\widehat{R}_{h,k}$ values for all $k = 1, \dots, p$ and some reasonable values of h , and in observing how much the criterion departs from one. The best choice will be the pair $(\widehat{H}, \widehat{K})$ which gives a value of $\widehat{R}_{h,k}$ close to one, such that $\widehat{K} \ll p$ in order to get an effective dimension reduction. In practice, there is no objective criterion to find a trade-off between a large value of the criterion $\widehat{R}_{h,k}$ and a small value of the dimension K . Then, a visual expertise of the 3D-plot of the $\widehat{R}_{h,k}$ versus (h, k) allows the selection of the best value. It is also useful to provide, for each (h, k) , the boxplots of the bootstrap replication of the squared trace correlation to have a look on how much the corresponding k -dimensional linear subspace is stable. Although boxplots of $\widehat{R}_{h,k}$ are also useful to determine the optimal number of slices \widehat{H} , wMSIR is not really sensitive to this parameter as shown in Section 4.

3.6 Analyzing components of \mathbf{y} through MSIR

Recall that the estimate $\widehat{\mathbf{V}}$ (resp. $\widehat{\widehat{\mathbf{V}}}$) of MSIR (resp. wMSIR) is computed from the estimated EDR directions $\widehat{\mathbf{B}}^{(j)}$ associated with each component of \mathbf{y} . From these estimates, it is straightforward to calculate the proximity measure $\hat{r}_j := r(\widehat{\mathbf{B}}^{(j)}, \widehat{\boldsymbol{\Sigma}}, \widehat{\mathbf{V}}, \widehat{\boldsymbol{\Sigma}})$ (resp. $\hat{\hat{r}}_j := r(\widehat{\mathbf{B}}^{(j)}, \widehat{\boldsymbol{\Sigma}}, \widehat{\widehat{\mathbf{V}}}, \widehat{\boldsymbol{\Sigma}})$) between each estimated marginal EDR space and the estimated common one, for $j = 1, \dots, q$. Then it is easy to sort these measures in descending order and to draw the associated scree plot. For $j = 1, \dots, q$, assuming model (3) and observing a low value of \hat{r}_j or $\hat{\hat{r}}_j$ could indicate an unprecise estimate of $\mathbf{B}^{(j)}$ since \hat{r}_j and $\hat{\hat{r}}_j$ tends to 1 in probability as n goes to ∞ . One can then withdraw the component $y^{(j)}$ of \mathbf{y} to improve the accuracy of $\widehat{\mathbf{V}}$ or $\widehat{\widehat{\mathbf{V}}}$.

In addition, assuming model (3) with a low dimensional common EDR space for the whole components of \mathbf{y} does not always seem realistic in real data analysis. Therefore, applying any multivariate SIR method on \mathbf{y} should not provide a suitable dimension reduction. However, it makes sense to assume that only groups of components of \mathbf{y} rely on model (3) with small values of K , as in model (4). For this model, we propose a methodology in order to identify the variables $y^{(j)}$ which share the same EDR space. Thus, we obtain clusters of components on which applying a multivariate SIR approach is sensible. Note that performing marginal univariate SIR on each component $y^{(j)}$ of \mathbf{y} leads to consistent estimates of each K -dimensional EDR space, since for $l = 1, \dots, L$, it is assumed that the rank of \mathbf{B}_l is equal to K . Recalling notations of Section 3.2, we obtain q_1 estimates $\text{Span}(\widehat{\mathbf{B}}^{(1)}), \dots, \text{Span}(\widehat{\mathbf{B}}^{(q_1)})$ of $\text{Span}(\mathbf{B}_1)$, q_2 estimates $\text{Span}(\widehat{\mathbf{B}}^{(q_1+1)}), \dots, \text{Span}(\widehat{\mathbf{B}}^{(q_1+q_2)})$ of $\text{Span}(\mathbf{B}_2)$, and so on, but values of q_1, \dots, q_L are unknown, as well as the number L of clusters.

For $(j, j^*) \in \{1, \dots, q\}^2$, we define $\hat{r}_{j,j^*} := r(\widehat{\mathbf{B}}^{(j)}, \widehat{\boldsymbol{\Sigma}}, \widehat{\mathbf{B}}^{(j^*)}, \widehat{\boldsymbol{\Sigma}})$. Without any loss of generality, assume that $L = 2$. Let us define $(j_1, j_2, j_3, j_4) \in \{1, \dots, q_1\}^2 \times \{q_1 + 1, \dots, q_1 + q_2\}^2$. We thus have the following Lemma.

Lemma 1. Under model (4) and assumptions (LC) and (A1)-(A3), \hat{r}_{j_1, j_2} and \hat{r}_{j_3, j_4} tend to 1 in probability.

The proof of this Lemma is given in Appendix A.1. Let us remark that, however, \hat{r}_{j_1, j_3} does not converge to 1 in probability since $r(\mathbf{B}_1, \boldsymbol{\Sigma}, \mathbf{B}_2, \boldsymbol{\Sigma}) < 1$. This leads to the following criterion to cluster components of \mathbf{y} : for $(j, j^*) \in \{2, \dots, q\} \times \{1, \dots, j-1\}$, components $y^{(j)}$ and $y^{(j^*)}$ are put in the same cluster if \hat{r}_{j, j^*} is close to 1. To do so, we can perform, for instance, a hierarchical ascending classification on the $q \times q$ (symmetric) matrix of the proximity measures \hat{r}_{j, j^*} (with $\hat{r}_{j, j} = 1$ for $j = 1, \dots, q$). Other clustering procedures can be applied on this matrix, such as the multidimensional scaling together with the k-means method. In Sections 4-5, we give illustrations of a clustering step for simulated and real datasets, which clearly improves the estimation of the corresponding EDR spaces.

Remark 4. Computing the common estimated EDR space for each obtained cluster is not time consuming since $\hat{\mathbf{B}}^{(1)}, \dots, \hat{\mathbf{B}}^{(q)}$ have already been computed. This is not the case for the k-means inverse regression (KIR) which requires new computations. In addition, applying PMS instead of MSIR or wMSIR on a cluster of \mathbf{y} requires to store the $p \times p$ matrices $\hat{\boldsymbol{\Gamma}}^{(1)}, \dots, \hat{\boldsymbol{\Gamma}}^{(q)}$ and summing some of them, which represent more computational time and more memory space than required by MSIR or wMSIR method, especially when p is large.

4 A simulation study

This section illustrates the ability of the proposed MSIR and wMSIR approaches, together with the diagnostic procedures on the components of \mathbf{y} , to properly estimate EDR spaces. The two following subsections respectively correspond to models (3) and (4).

4.1 Single EDR space model

Two simulation models are considered here. For a given sample size n and a dimension p , 100 replications of the covariate \mathbf{x} are generated from the p -dimensional normal distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the same pair $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here, $\boldsymbol{\mu}$ is randomly generated from the $\mathcal{N}_p(0, \mathbf{I}_p)$ distribution and $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + 0.1\mathbf{I}_p$ where \mathbf{L} is a $p \times p$ matrix made of entries independently generated from the standard normal distribution $\mathcal{N}_1(0, 1)$. For every model, for all $j \geq 1$, it is assumed that $\varepsilon^{(j)} \sim \mathcal{N}_1(0, 1)$. The dimension K is also assumed to be known in this section. We first study a special case of model (3) with $K = 1$. Secondly, MSIR and wMSIR are evaluated with a multiple indices model ($K = 2$).

Single index model. Consider the following single index model:

$$\begin{cases} y^{(1)} &= \mathbf{x}'\beta_1 + \varepsilon^{(1)}, \\ y^{(2)} &= (\mathbf{x}'\beta_1)^3 + 3\varepsilon^{(2)}, \\ y^{(3)} &= \mathbf{x}'\beta_1(1 + \varepsilon^{(3)}), \end{cases} \quad (9)$$

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

We generate samples of size $n = 100$ from (9), with $p = 20$. Then, the EDR direction is estimated using the following methods, with $H = 10$ slices:

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

For each estimator $\widehat{\mathbf{B}} \in \left\{ \widehat{\mathbf{B}}^{(1)}, \widehat{\mathbf{B}}^{(2)}, \widehat{\mathbf{B}}^{(3)}, \widehat{\mathbf{V}}, \widehat{\widehat{\mathbf{V}}}, \widehat{\mathbf{V}}_{\text{KIR}}, \widehat{\mathbf{V}}_{\text{PMS}} \right\}$, we compute the squared trace correlation $r(\widehat{\mathbf{B}}) := r(\widehat{\mathbf{B}}, \boldsymbol{\Sigma}, \mathbf{B}, \boldsymbol{\Sigma})$ between the estimated EDR space and the true EDR space. The closer to one is $r(\widehat{\mathbf{B}})$, the better is the estimate. Note that for $K = 1$, the criterion $r(\widehat{\mathbf{B}})$ corresponds to the squared cosine of the angle between $\widehat{\mathbf{B}}$ and β_1 .

Boxplots of this criterion are drawn on Figure 1(a). It appears that $\widehat{\mathbf{B}}^{(3)}$ exhibits low squared trace correlation. This phenomenon can be explained by the heteroscedasticity in the third marginal model of (9). Even if $\widehat{\mathbf{B}}^{(3)}$ is necessary to compute $\widehat{\mathbf{V}}$ and $\widehat{\widehat{\mathbf{V}}}$, the poor estimates of $\mathbf{B}^{(3)}$ do not imply a significant loss in the squared trace correlations related to $\widehat{\mathbf{V}}$ and to $\widehat{\widehat{\mathbf{V}}}$. One can also observe that the weighting in wMSIR seems to improve the estimation of the EDR space since values of $r(\widehat{\widehat{\mathbf{V}}})$ are globally greater than those of $r(\widehat{\mathbf{V}})$. This trend is confirmed in Figure 1(b) where the values of $r(\widehat{\mathbf{V}})$ are plotted versus those of $r(\widehat{\widehat{\mathbf{V}}})$. It appears that wMSIR seems to be uniformly better than MSIR in this simulation. In addition, Figure 1(a) shows that the pooled marginal slicing produces slightly better estimates than wMSIR in this example and that wMSIR outperforms the k-means inverse regression.

The poor quality of the estimate $\widehat{\mathbf{B}}^{(3)}$ can be observed directly from the simulated data. In Figure 2(a), we draw boxplots of values of \hat{r}_j for $j = 1, 2, 3$. Considering that the first quartile of the third boxplot is equal to 0.80 and that the minimum value of the others is equal to 0.96, it makes sense to withdraw the third component of \mathbf{y} from the analysis for at least a forth of the datasets. Let $\widehat{\widehat{\mathbf{V}}}^*$ be the wMSIR estimate built only from $y^{(1)}$ and $y^{(2)}$. In Figure 2(b), we compare

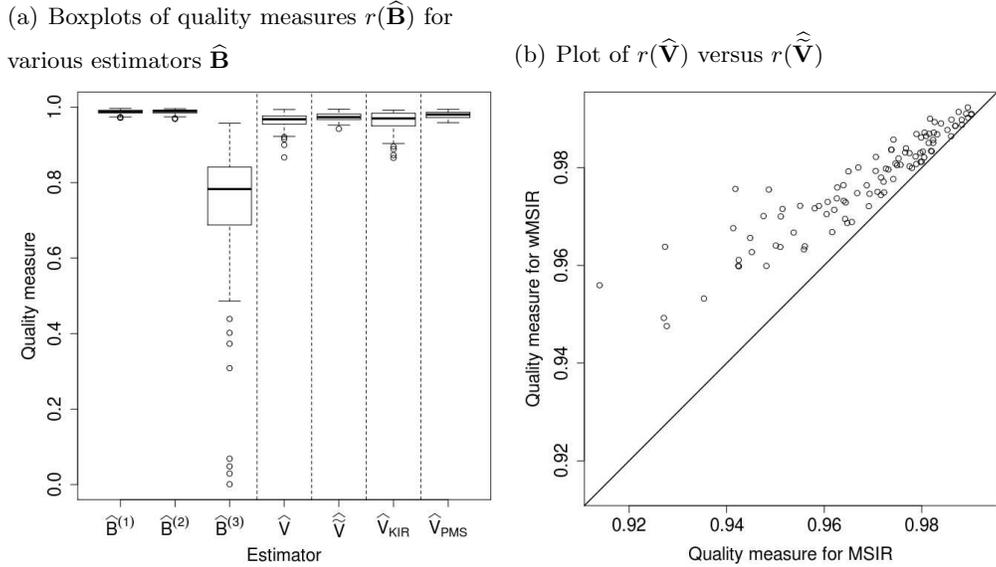


Figure 1: Comparison of estimators of \mathbf{B} on 100 samples from model (9) with $n = 100$ and $p = 20$. The line in (b) corresponds to the first bisecting line.

quality measures for $\hat{\mathbf{V}}^*$ and for the PMS estimate. We observe that the selection based on the \hat{r}_j 's improves performances of the wMSIR method, so that it produces better quality measures than the PMS without this selection step.

To study the sensitivity of wMSIR with respect to n and p , several samples are generated for various values of these parameters. In Figure 3(a), boxplots of 100 values of $r(\hat{\mathbf{V}})$ are drawn for each pair (n, p) . Not surprisingly, estimated EDR spaces become closer to the true one when n increases. Moreover, one can also observe that estimates of the EDR space are more precise when p is small. This phenomenon can be explained by the fact that $p \times p$ matrices have to be estimated. Notice that for every value of (n, p) in Figure 3(a), estimates are very precise since most of the values of $r(\hat{\mathbf{V}})$ are greater than 0.85.

In previous analyses, the number of slices H was set to 10 to have enough slices to study the functions that link $\mathbf{x}'\beta_1$ to each component of \mathbf{y} and enough points in each slice. In Figure 4(a), we observe that for wMSIR, one can arbitrarily choose a number of slices between 8 and 38 and obtain an estimate of the EDR space which is as reliable as the one computed with $H = 10$.

Multiple indices model. Consider now a more complex model than model (9). It is defined by:

$$\begin{cases} y^{(1)} &= \exp(\mathbf{x}'\beta_1) \times (\mathbf{x}'\beta_2) + \varepsilon^{(1)}, \\ y^{(2)} &= (\mathbf{x}'\beta_1) \times \exp(\mathbf{x}'\beta_2) + \varepsilon^{(2)}, \end{cases} \quad (10)$$

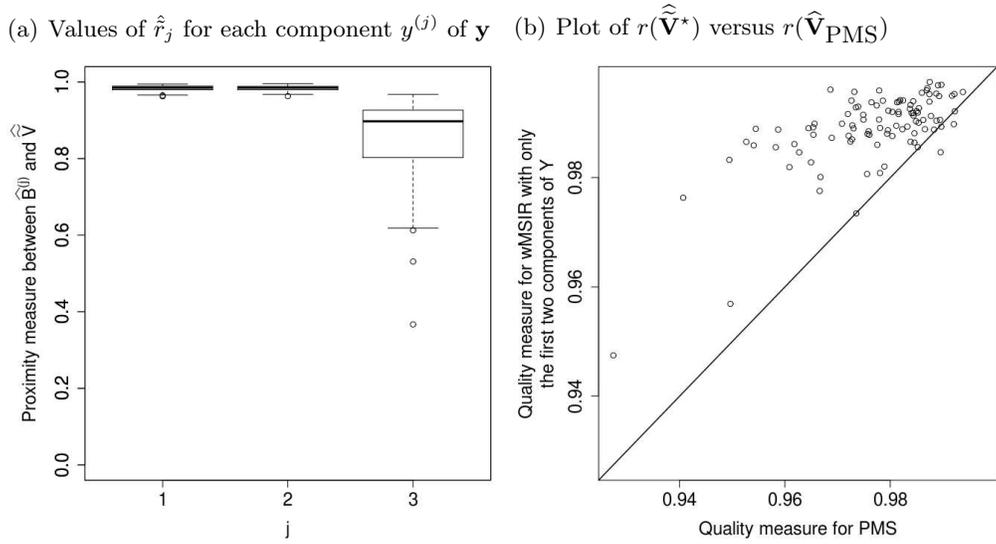


Figure 2: Study of the contribution of $y^{(3)}$ to $\hat{\mathbf{V}}$ for 100 samples generated from model (9) with $n = 100$ and $p = 20$.

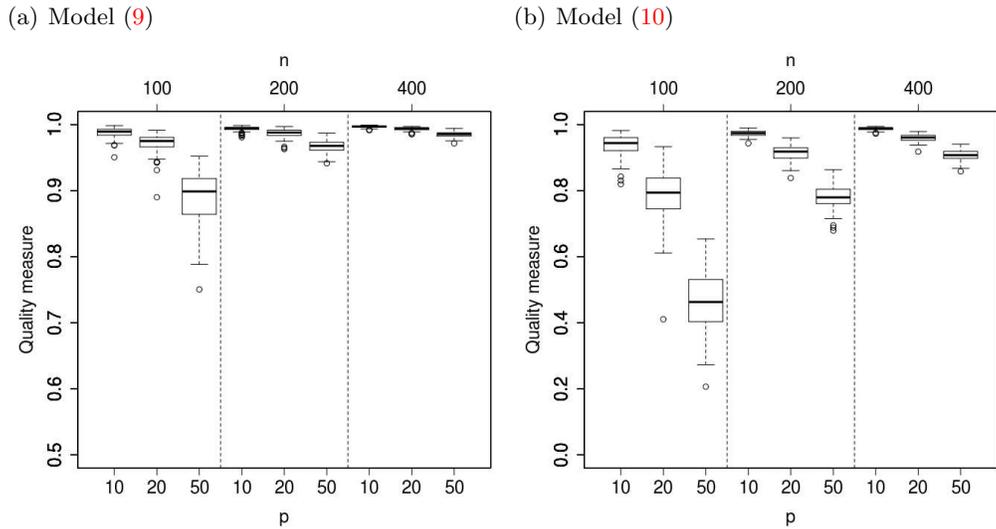


Figure 3: Boxplots of quality measure for the wMSIR method, for 100 samples generated from (a) model (9) or from (b) model (10) with various values of n and p .

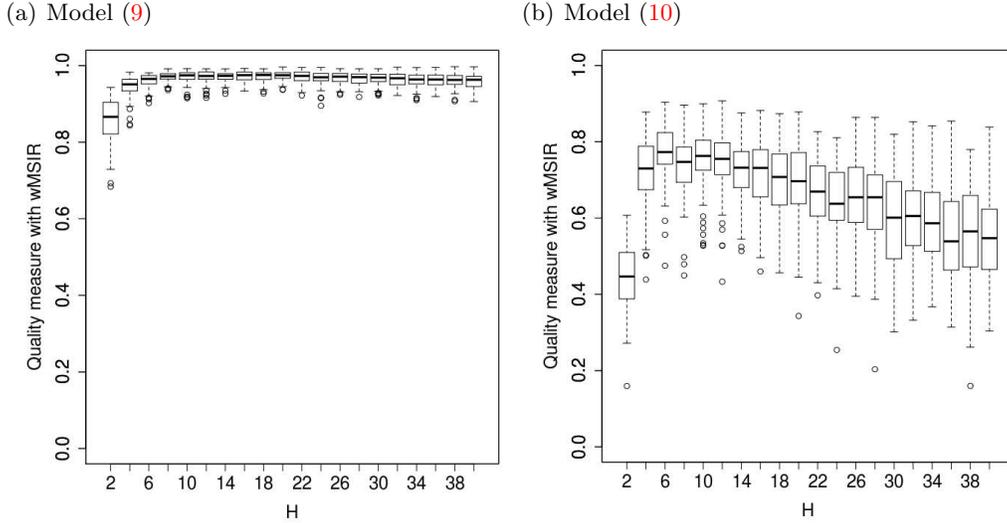


Figure 4: Boxplots of quality measure for the wMSIR method for 100 samples generated (a) from the model (9) and (b) from the model (10), with $n = 100$, $p = 20$ and various values of H .

where $\beta_1 = [\beta_{1,1}, \dots, \beta_{1,p}]'$, $\beta_2 = [\beta_{2,1}, \dots, \beta_{2,p}]'$, with $\beta_{1,i} = i \mathbb{I}(i \leq 5) + \mathbb{I}(i > 5)$ and $\beta_{2,i} = (-1)^{i-1}(1 + \mathbb{I}(i \in \{3, 4\}))$. Note that, in model (10), we clearly have $K = 2$.

Samples are generated from model (10) for various n and p . Then, these samples are used to estimate the corresponding EDR space with wMSIR with $H = 10$ slices. This leads to boxplots of $r(\widehat{\mathbf{V}})$ displayed in Figure 3(b). We observe lower and more scattered values of $r(\widehat{\mathbf{V}})$ than in Figure 3(a). The complexity of the link function between \mathbf{y} and \mathbf{x} and the greater dimension K are believable reasons for this phenomenon. Apart from this feature, Figure 3(b) provides identical evolutions of $r(\widehat{\mathbf{V}})$ with n and p to those observed for the single index model in Figure 3(a). Note that for both model (9) and model (10), the behavior of MSIR and wMSIR when n and p vary are similar. That is why only results concerning wMSIR are drawn in Figure 3.

Examining Figure 4(b), it seems that the quality of wMSIR estimates is less connected to the chosen number of slices for model (9) than for model (10). For the latter, performances of wMSIR are nevertheless similar for values of H from 6 to 12.

4.2 Multiple EDR spaces model

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

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

based on model (9), where $(\theta_1, \theta_2) \in [0, 1]$ and for $j = 1, \dots, 12$, $\varepsilon^{(j)} \sim \mathcal{N}_1(0, 1)$.

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

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

We plot, in Figure 5, values of \hat{r}_{j,j^*} for $(j, j^*) \in \{2, \dots, q\} \times \{1, \dots, j-1\}$, computed from a sample of size $n = 1000$. Darker squares correspond to values of \hat{r}_{j,j^*} close to 1. Thus, Figure 5 leads to cluster components $y^{(1)}$ and $y^{(2)}$ together as well as components $y^{(4)}$, $y^{(5)}$ and $y^{(6)}$. Components $y^{(7)}$ and $y^{(8)}$ can also be grouped together. It is tempting to put $y^{(3)}$ with $y^{(1)}$ and $y^{(2)}$ and to group $y^{(9)}$ with $y^{(8)}$ and $y^{(7)}$ because the related squared trace correlations are high but $\hat{r}_{3,9}$ is approximately as high as them. Note also that for $j \in \{1, 2, 3\}$, we have $\hat{r}_{3j,3j-1} < \hat{r}_{3j-2,3j-1}$ and $\hat{r}_{3j,3j-2} < \hat{r}_{3j-2,3j-1}$. Recalling that for the considered values of j , $y^{(3j)}$ is the third component of each model (9) embedded in model (11), these inequalities can be explained by the heteroscedasticity in this component which leads to imprecise estimates of the relevant EDR direction as pointed out on Figure 1(a). In addition, on Figure 5, component $y^{(11)}$ could be unexpectedly grouped with $y^{(4)}$ and $y^{(5)}$ or with $y^{(8)}$.

In this example, the interpretation of Figure 5 can easily be done because components of \mathbf{y} are already clustered in the definition of the model. In other words, every component between two others that are related to the same EDR space belongs to a marginal model based on this EDR space. In practical cases, the components of \mathbf{y} may not be ordered that way which means that

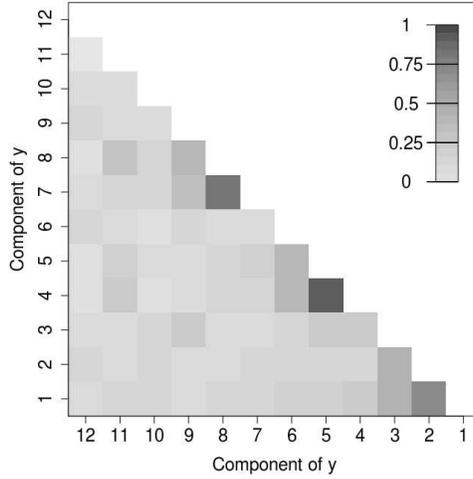


Figure 5: Values of \hat{r}_{j,j^*} in various shades of grays for $(j, j^*) \in \{2, \dots, q\} \times \{1, \dots, j-1\}$ and for a sample of size $n = 1000$ generated from model (11) with $\theta_1 = \theta_2 = 1$.

the corresponding representation of the squared trace correlations may be cluttered. To tackle this problem, we use an agglomerative hierarchical clustering algorithm based on the dissimilarity $1 - \hat{r}_{j,j^*}$ between $\hat{\mathbf{B}}^{(j)}$ and $\hat{\mathbf{B}}^{(j^*)}$.

A model with 5 clusters. We generate a new sample of size $n = 1000$ from model (11) with $\theta_1 = 1$ and $\theta_2 = 0$ which leads to $l = 5$ clusters:

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

The hierarchical clustering algorithm applied to the estimates $\hat{\mathbf{B}}^{(j)}$ for $j = 1, \dots, q$ and produces the dendrogram of Figure 6(a). A classification directly based on this procedure would not be really accurate. For instance, in model (11), $y^{(11)}$ and $y^{(10)}$ belong to two different clusters. On Figure 6(a), to divide $y^{(11)}$ and $y^{(10)}$ into two groups, the tree has to be cut at a level around 0.75. This implies grouping $y^{(1)}, y^{(7)}, y^{(2)}$ and $y^{(8)}$ together and putting $y^{(3)}$ and $y^{(9)}$ in another group while components of both groups are actually related with the same EDR space. However, the dendrogram of Figure 6(a) allows to order components of \mathbf{y} in such a way that those which are linked by high squared trace correlation are close from each other. Thus, in Figure 6(b), we displayed values of \hat{r}_{j,j^*} for $j^* \prec j$ where \prec denotes the ordering in Figure 6(a). It becomes clear, then, that components $y^{(1)}, y^{(7)}, y^{(2)}, y^{(8)}, y^{(3)}$ and $y^{(9)}$ should be clustered in the same group. Note that another cluster containing $y^{(6)}, y^{(4)}$ and $y^{(5)}$ can be made from Figure 6(b). Not surprisingly, we observe again that the squared trace correlation between components $y^{(3)}, y^{(6)}, y^{(9)}$ and the components that correspond to the group they respectively belong to is quite low. Finally,

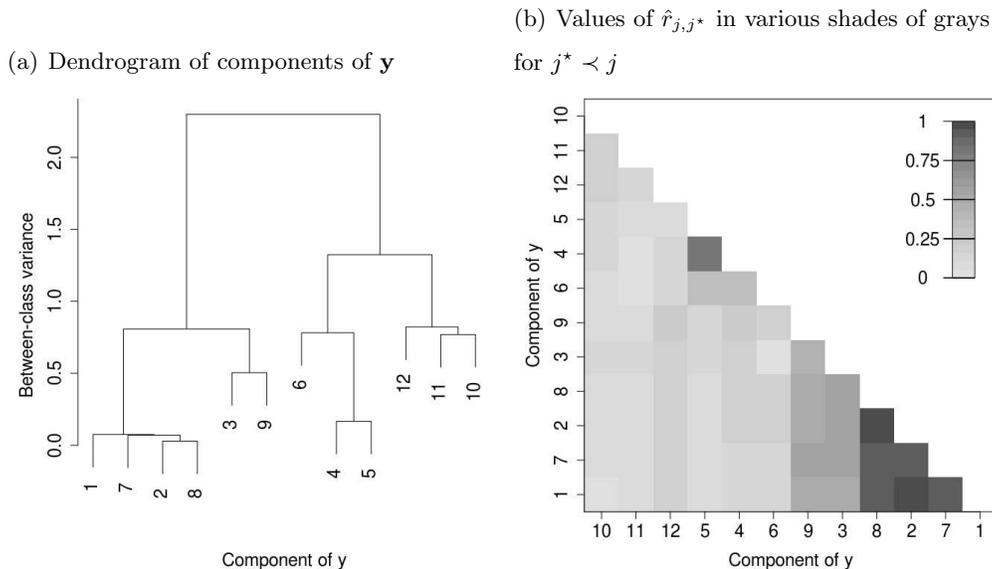


Figure 6: Clustering analysis for a sample of size $n = 1000$ generated from model (11) with $\theta_1 = 1$ and $\theta_2 = 0$.

components $y^{(12)}$, $y^{(11)}$ and $y^{(10)}$ appears to form three distinct clusters.

5 Real data illustrations

5.1 Remote sensing data

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

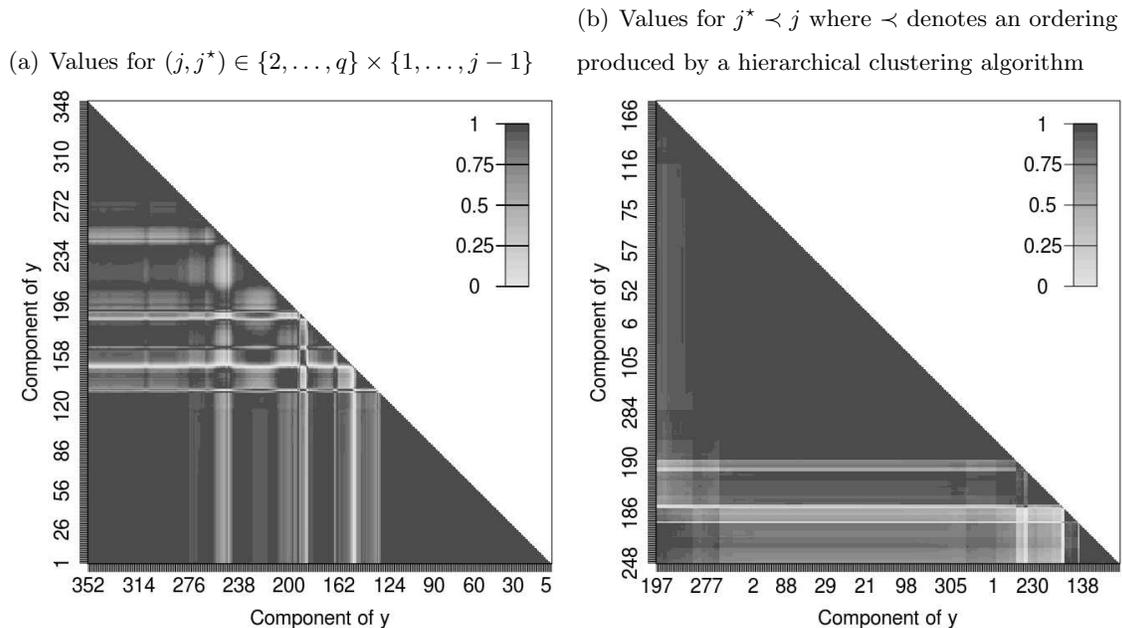


Figure 7: Values of \hat{r}_{j,j^*} in various shades of grays for hyperspectral data.

We ran the clustering procedure described in the above section associated to wMSIR with $H = 10$ slices. The clustering results are depicted on Figure 7. Two clusters of wavelengths have been identified, corresponding to two different orientations of $\hat{\mathbf{v}}_1$. It appears on Figure 8(a) that, in the first cluster, only the proportion of dust $x^{(2)}$ has an important contribution to the EDR direction. In the second cluster, the estimated EDR direction is close to $(1, 1, 0)$. This corresponds to the index $x^{(1)} + x^{(2)}$ which is equal to one minus the proportion of water. Let us note that the grain size of water ice $x^{(3)}$ does not appear in the EDR directions. Figure 8(b) permits to visualize the clustering of the wavelengths: It reveals which wavelengths are more sensible to the presence of water ice or dust.

5.2 Minneapolis elementary schools data

Another dataset which is widely studied in the dimension reduction context with a multivariate response is related to test results of students in Minneapolis elementary schools. These data are for example presented in Cook (2009); Cook and Setodji (2003); Yin and Bura (2006). The response variable \mathbf{y} is made of $q = 4$ components. The first (resp. third) component is the proportion of pupils scoring below the average on a fourth (resp. sixth) grade test. The second (resp. fourth) one is the proportion of marks above the average. Following Yin and Bura (2006), we would like to explain \mathbf{y} with a 8-dimensional variable \mathbf{x} . The seven first components \mathbf{x} are called $x^{(1)}, \dots, x^{(7)}$ and are respectively the squared root of the percentages of children receiving an aid called AFDC,

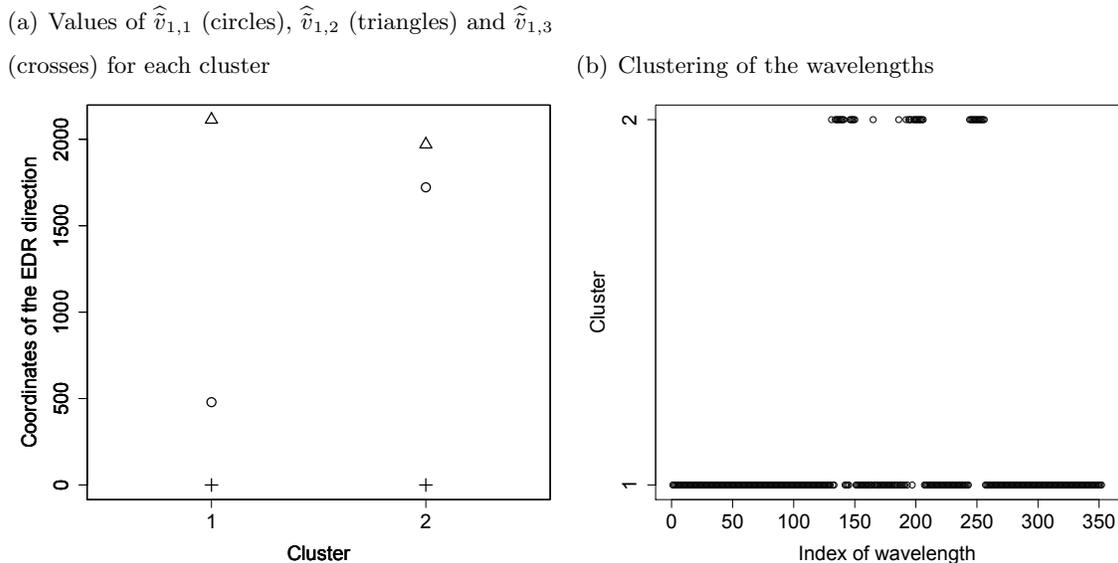


Figure 8: Clustering of the components of \mathbf{y} and estimates $\widehat{\mathbf{v}}_1 = [\widehat{v}_{1,1}, \widehat{v}_{1,2}, \widehat{v}_{1,3}]'$ of the EDR direction for each cluster.

children who do not live with both parents, people in the area of a school who completed high school, people who suffer for poverty, minority, mobility and pupils who attend school regularly. The last component of \mathbf{x} , named $x^{(8)}$, is the mean number of pupils for each teacher. Note that [Yoo \(2009\)](#) analyzed the variables $x^{(1)}$, $x^{(2)}$ and $x^{(3)}$ while [Cook \(2009\)](#) focused on some increasing functions of these variables and [Cook and Setodji \(2003\)](#) considered $x^{(1)}$, $x^{(2)}$, $x^{(3)}$, $x^{(4)}$ and $x^{(8)}$. We first compute $\widehat{\mathbf{V}}$ with $p = 8$ components of \mathbf{x} . We then use a stepwise algorithm with the AIC to perform a linear regression of $\mathbf{x}'\widehat{\mathbf{V}}$ on \mathbf{x} and sort the components of \mathbf{x} with respect to how much they explain $\mathbf{x}'\widehat{\mathbf{V}}$. The first three sorted variables are $x^{(1)}$, $x^{(2)}$, and $x^{(3)}$. Besides, regressing $\mathbf{x}'\widehat{\mathbf{V}}$ on $\{x^{(1)}, x^{(2)}, x^{(3)}\}$ produces an adjusted coefficient of determination of 95.42%. This encourages us to take $\mathbf{x} = (x^{(1)}, x^{(2)}, x^{(3)})'$.

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

6 Concluding remarks

In this paper, we proposed the new multivariate SIR approaches MSIR and wMSIR for estimating the EDR space. The idea consists in performing first several marginal SIR analyzes. A common EDR space is then deduced from the marginal ones by maximizing the proximity criterion defined in (6). This optimization problem benefits from a closed-form solution.

Let us highlight that this two-step approach can be applied to any variant of the SIR method. For instance, it permits to build multivariate regularized SIR approaches from [Bernard-Michel, Gardes, and Girard \(2009\)](#), multivariate SIR_α approaches from [Gannoun and Saracco \(2003\)](#) or multivariate kernel SIR methods from [Wu \(2008\)](#). To avoid the choice of the number H of slices, one may also consider building a multivariate version of the CUME procedure from [Zhu, Zhu, and Feng \(2010\)](#).

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

Acknowledgments

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

A Proofs

A.1 Proof of Lemma 1

For $j = 1, \dots, q$, recall that $P_{\hat{\mathbf{B}}^{(j)}, \hat{\Sigma}}$ is the $\hat{\Sigma}$ -orthogonal projector onto $\text{Span}(\hat{\mathbf{B}}^{(j)})$. According to Theorem 4 of [Saracco \(1997\)](#), we have that $P_{\hat{\mathbf{B}}^{(j)}, \hat{\Sigma}} = P_{\mathbf{B}_1, \Sigma} + O_P(n^{-1/2})$, for $j \in \{j_1, j_2\}$ and

$P_{\hat{\mathbf{B}}^{(j)}, \hat{\Sigma}} = P_{\mathbf{B}_2, \Sigma} + O_P(n^{-1/2})$, for $j \in \{j_3, j_4\}$. It follows that

$$\begin{aligned} r(\hat{\mathbf{B}}^{(j_1)}, \hat{\Sigma}, \hat{\mathbf{B}}^{(j_2)}, \hat{\Sigma}) &= \frac{1}{K} \text{Trace}(P_{\hat{\mathbf{B}}^{(j_1)}, \hat{\Sigma}} P_{\hat{\mathbf{B}}^{(j_2)}, \hat{\Sigma}}) \\ &= \frac{1}{K} \text{Trace}((P_{\mathbf{B}_1, \Sigma} + O_P(n^{-1/2}))(P_{\mathbf{B}_1, \Sigma} + O_P(n^{-1/2}))) \\ &= \frac{1}{K} \text{Trace}(P_{\mathbf{B}_1, \Sigma}) + O_P(n^{-1/2}) \\ &= 1 + O_P(n^{-1/2}). \end{aligned}$$

Similarly, $r(\hat{\mathbf{B}}^{(j_3)}, \hat{\Sigma}, \hat{\mathbf{B}}^{(j_4)}, \hat{\Sigma})$ tends to 1 in probability. \square

A.2 Proof of Theorem 1

Recall that, since the bases $\mathbf{D}, \mathbf{B}^{(1)}, \dots, \mathbf{B}^{(q)}$ are assumed to be Σ -orthonormal, we have $P_{\mathbf{D}, \Sigma} = \mathbf{D}\mathbf{D}'\Sigma$ and $P_{\mathbf{B}^{(j)}, \Sigma} = \mathbf{B}^{(j)}\mathbf{B}^{(j)'}\Sigma$. It follows that

$$\begin{aligned} Kq \times Q(\mathbf{D}, \mathbf{B}^{(1)}, \dots, \mathbf{B}^{(q)}) &= \sum_{j=1}^q \text{Trace}(\mathbf{D}\mathbf{D}'\Sigma\mathbf{B}^{(j)}\mathbf{B}^{(j)'}\Sigma) \\ &= \sum_{j=1}^q \text{Trace}(\mathbf{D}'\Sigma\mathbf{B}^{(j)}\mathbf{B}^{(j)'}\Sigma\mathbf{D}) \\ &= \text{Trace}\left(\mathbf{D}'\Sigma\left\{\sum_{j=1}^q \mathbf{B}^{(j)}\mathbf{B}^{(j)'}\right\}\Sigma\mathbf{D}\right) \\ &= \text{Trace}(\mathbf{D}'\Sigma\mathbb{B}\mathbb{B}'\Sigma\mathbf{D}). \end{aligned}$$

Hence, it is well known that the matrix \mathbf{V} which maximizes $\text{Trace}(\mathbf{D}'\Sigma\mathbb{B}\mathbb{B}'\Sigma\mathbf{D})$ over the set of matrices \mathbf{D} such that $\mathbf{D}'\Sigma\mathbf{D} = \mathbf{I}_K$ is made of the K generalized eigenvectors of $\Sigma\mathbb{B}\mathbb{B}'\Sigma$ and Σ associated with K non-null eigenvalues. Thus, \mathbf{V} contains the K eigenvectors of $\Sigma^{-1}(\Sigma\mathbb{B}\mathbb{B}'\Sigma) = \mathbb{B}\mathbb{B}'\Sigma$ which are associated with K non-null eigenvalues.

In addition, we have that $\text{Span}(\Sigma\mathbb{B}\mathbb{B}'\Sigma) = \text{Span}(\Sigma\mathbb{B})$ because $\text{Span}(\Sigma\mathbb{B}\mathbb{B}'\Sigma) \subset \text{Span}(\Sigma\mathbb{B})$ and $\dim(\text{Span}(\Sigma\mathbb{B}\mathbb{B}'\Sigma)) = K$. Since Σ is invertible, this implies that $\text{Span}(\mathbb{B}\mathbb{B}'\Sigma) = \text{Span}(\mathbb{B})$. Finally, we have under model (3) and assumption (LC) that for all $j \in \{1, \dots, q\}$, $\text{Span}(\mathbf{B}) = \text{Span}(\mathbf{B}^{(j)})$ using univariate SIR theory. We then have $\text{Span}(\mathbb{B}) = \text{Span}(\mathbf{B})$ and then $\text{Span}(\mathbf{V}) = \text{Span}(\mathbf{B})$.

A.3 Proof of Theorem 2

For each component $y^{(j)}$ of \mathbf{y} and under the assumptions (LC), (A1)-(A3), from univariate SIR theory of Li (1991), each estimated EDR space $\hat{\mathbf{B}}^{(j)}$ converges to $\mathbf{B}^{(j)}$ at root n rate: that is, for $j = 1, \dots, q$, $\hat{\mathbf{B}}^{(j)} = \mathbf{B}^{(j)} + O_p(n^{-1/2})$. It follows that $\hat{\mathbb{B}} = \mathbb{B} + O_p(n^{-1/2})$. Since $\hat{\Sigma} = \Sigma + O_p(n^{-1/2})$, we get $\hat{\mathbb{B}}\hat{\mathbb{B}}'\hat{\Sigma} = \mathbb{B}\mathbb{B}'\Sigma + O_p(n^{-1/2})$. Therefore, the eigenvectors associated with

the largest K eigenvalues of $\widehat{\mathbb{B}}\widehat{\mathbb{B}}'\widehat{\Sigma}$ converge to the corresponding ones of $\mathbb{B}\mathbb{B}'\Sigma$ at the same rate: $\widehat{\mathbf{v}}_k = \mathbf{v}_k + O_p(n^{-1/2})$ for $k = 1, \dots, K$. Consequently, the estimated EDR space $\text{Span}(\widehat{\mathbf{V}})$ converges to $\text{Span}(\mathbf{V})$ at root n rate. Since $\text{Span}(\mathbf{V}) = \text{Span}(\mathbf{B})$, the estimated EDR space converges to the true one in probability.

A.4 Proof of Theorem 3

We can proceed analogously to the proof of Theorem 2. It is sufficient to show that $\widehat{\mathbb{W}} = \mathbb{W} + O_p(n^{-1/2})$. Since, for each component $y^{(j)}$ of \mathbf{y} and under assumptions (LC) and (A1)-(A3), from univariate SIR theory of Li (1991), each estimated eigenvalue $\widehat{\lambda}_k^{(j)}$ converges to $\lambda_k^{(j)}$ at root n rate, we have $\widehat{\pi}^{(j)} = \pi^{(j)} + O_p(n^{-1/2})$ and $\widehat{\pi}_* = \pi_* + O_p(n^{-1/2})$. We thus get $\widehat{\mathbf{W}}^{(j)} = \mathbf{W}^{(j)} + O_p(n^{-1/2})$ and $\widehat{\mathbb{W}} = \mathbb{W} + O_p(n^{-1/2})$.

Consequently, $\widehat{\mathbb{W}}\widehat{\mathbb{W}}'\widehat{\Sigma} = \mathbb{W}\mathbb{W}'\Sigma + O_p(n^{-1/2})$ and the eigenvectors associated with the largest K eigenvalues of $\widehat{\mathbb{W}}\widehat{\mathbb{W}}'\widehat{\Sigma}$ converge to the corresponding ones of $\mathbb{W}\mathbb{W}'\Sigma$ at the same rate: $\widehat{\mathbf{v}}_k = \mathbf{v}_k + O_p(n^{-1/2})$ for $k = 1, \dots, K$. Therefore, the estimated EDR space $\text{Span}(\widehat{\mathbf{V}})$ converges to $\text{Span}(\mathbf{V})$ at root n rate. Since $\text{Span}(\mathbf{V}) = \text{Span}(\mathbf{B})$, the estimated EDR space converges to the true one in probability.

References

- Aragon, Y. (1997). A gauss implementation of multivariate sliced inverse regression. *Computational Statistics* **12** 355–372.
- Aragon, Y. and Saracco, J. (1997). Sliced Inverse Regression (SIR): an appraisal of small sample alternatives to slicing. *Computational Statistics* **12** 109–130.
- Azaïs, R., Gégout-Petit, A., and Saracco, J. (2012). Optimal quantization applied to sliced inverse regression. *Journal of Statistical Planning and Inference* **142** 481–492.
- Bai, Z. D. and He, X. (2004). A chi-square test for dimensionality for non-gaussian data. *Journal of Multivariate Analysis* **88** 109–117.
- Barreda, L., Gannoun, A., and Saracco, J. (2007). Some extensions of multivariate SIR. *Journal of Statistical Computation and Simulation* **77** 1–17.
- Barrios, M. P. and Velilla, S. (2007). A bootstrap method for assessing the dimension of a general regression problem. *Statistics & Probability Letters* **77** 247–255.
- Bernard-Michel, C., Douté, S., Fauvel, M., Gardes, L., and Girard, S. (2009). Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression. *Journal of Geophysical Research - Planets* **114**.

- Bernard-Michel, C., Gardes, L., and Girard, S. (2009). Gaussian Regularized Sliced inverse Regression. *Statistics and Computing* **19** 85–98.
- Bura, E. and Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society, Series B* **63** 393–410.
- Chavent, M., Kuentz, V., Lique, B., and Saracco, J. (2011). A sliced inverse regression approach for a stratified population. *Communications in statistics - Theory and methods* **40** 1–22.
- Chen, C.-H. and Li, K.-C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica* **8** 289–316.
- Cook, R. (2009). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley Series in Probability and Statistics, Wiley.
- Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association* **89** 177–189.
- (1998). Principal hessian directions revisited (with discussion). *Journal of the American Statistical Association* **93** 84–100.
- (2000). SAVE: a method for dimension reduction and graphics in regression. *Communications in statistics - Theory and methods* **29** 2109–2121.
- Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics* **30** 450–474.
- Cook, R. D. and Nachtsheim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association* **89** 592–599.
- Cook, R. D. and Setodji, C. M. (2003). A model-free test for reduced rank in multivariate regression. *J. Amer. Statist. Assoc.* **98** 340–351.
- Douté, S., Schmitt, B., Langevin, Y., Bibring, J.-P., Altieri, F., Bellucci, G., Gondet, B., and Poulet, F. (2007). South pole of Mars: Nature and composition of the icy terrains from Mars Express OMEGA observations. *Planetary and Space Science* **55** 113–133.
- Duan, N. and Li, K.-C. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics* **19** 505–530.
- Gannoun, A. and Saracco, J. (2003). An asymptotic theory for SIR_α method. *Statistica Sinica* **13** 297–310.
- Hall, P. and Li, K.-C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics* **21** 867–889.
- Hsing, T. (1999). Nearest neighbor inverse regression. *The Annals of Statistics* **27** 697–731.
- Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics* **20** 1040–1061.
- Kuentz, V., Lique, B., and Saracco, J. (2010). Bagging versions of sliced inverse regression. *Communications in statistics - Theory and methods* **39** 1985–1996.

- Kuentz, V. and Saracco, J. (2010). Cluster-based sliced inverse regression. *Journal of the Korean Statistical Society* **39** 251–267.
- Li, B., Wen, S., and Zhu, L. (2008). On a projective resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association* **103** 1177–1186.
- Li, K.-C. (1991). Sliced inverse regression for dimension reduction, with discussion. *Journal of the American Statistical Association* **86** 316–342.
- (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Steins lemma. *Journal of the American Statistical Association* **87** 1025–1039.
- Li, K.-C., Aragon, Y., Shedden, K., and Agnan, C. T. (2003). Dimension reduction for multivariate response data. *Journal of the American Statistical Association* **98** 99–109.
- Li, L. and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. *Technometrics* **48** 503–510.
- Li, L. and Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics* **64** 124–131.
- Liquet, B. and 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** 103–125.
- Lue, H.-H. (2009). Sliced inverse regression for multivariate response regression. *Journal of Statistical Planning and Inference* **139** 2656–2664.
- Nkiet, G.-M. (2008). Consistent estimation of the dimensionality in sliced inverse regression. *Annals of the Institute of Statistical Mathematics* **60** 257–271.
- Prendergast, L. A. (2005). Influence functions for sliced inverse regression. *Scandinavian Journal of Statistics* **32** 385–404.
- (2007). Implications of influence function analysis for sliced inverse regression and sliced average variance estimation. *Biometrika* **94** 585–601.
- Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Communications in statistics - Theory and methods* **26** 2141–2171.
- (1999). Sliced inverse regression under linear constraints. *Communications in statistics - Theory and methods* **28** 2367–2393.
- (2001). Pooled slicing methods versus slicing methods. *Communications in statistics - Simulation and Computation* **30** 489–511.
- (2005). Asymptotics for pooled marginal slicing estimator based on SIR_α approach. *Journal of Multivariate Analysis* **96** 117–135.
- Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association* **89** 141–148.

- Scrucca, L. (2007). Class prediction and gene selection for DNA microarrays using regularized sliced inverse regression. *Computational Statistics & Data Analysis* **52** 438–451.
- Setodji, C. M. and Cook, R. D. (2004). K-means inverse regression. *Technometrics* **46** 421–429.
- Shao, Y., Cook, R. D., and Weisberg, S. (2009). Partial central subspace and sliced average variance estimation. *Journal of Statistical Planning and Inference* **139** 952–961.
- Szretter, M. E. and Yohai, V. J. (2009). The sliced inverse regression algorithm as a maximum likelihood procedure. *Journal of Statistical Planning and Inference* **139** 3570–3578.
- Tyler, D. E. (1981). Asymptotic inference for eigenvectors. *The Annals of Statistics* **9** 725–736.
- Wu, H.-M. (2008). Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics* **17** 590–610.
- Yin, X. and Bura, E. (2006). Moment-based dimension reduction for multivariate response regression. *Journal of Statistical Planning and Inference* **136** 3675–3688.
- Yoo, J. K. (2009). Iterative optimal sufficient dimension reduction for conditional mean in multivariate regression. *Journal of Data Science* **7** 267–276.
- Zhu, L., Miao, B., and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association* **101** 630–643.
- Zhu, L.-P. and Yu, Z. (2007). On spline approximation of sliced inverse regression. *Science in China Series A: Mathematics* **50** 1289–1302.
- Zhu, L.-P., Zhu, L.-X., and Feng, Z.-H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association* **105** 1455–1466.
- Zhu, L.-P., Zhu, L.-X., and Wen, S.-Q. (2010). On dimension reduction in regressions with multivariate responses. *Statistica Sinica* **20** 1291–1307.
- Zhu, L. X. and Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics* **24** 1053–1068.
- Zhu, L. X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica* **5** 727–736.
- Zhu, L. X., Ohtaki, M., and Li, Y. (2007). On hybrid methods of inverse regression-based algorithms. *Computational Statistics* **51** 2621–2635.

CQFD team, INRIA Bordeaux Sud-Ouest and Institut de Mathématiques de Bordeaux, UMR CNRS 5251, 33405 Talence, France

E-mail: raphael.coudret@math.u-bordeaux1.fr

Mistis team, INRIA Grenoble Rhône-Alpes, 38334 Saint-Ismier, France

E-mail: stephane.girard@inria.fr

CQFD team, INRIA Bordeaux Sud-Ouest, Institut de Mathématiques de Bordeaux, UMR CNRS 5251 and Institut Polytechnique de Bordeaux, 33405 Talence, France
E-mail: jerome.saracco@math.u-bordeaux1.fr