Advanced topics in Sliced Inverse Regression

Stéphane Girard, Hadrien Lorenzo, Jérôme Saracco

Advanced topics in Sliced Inverse Regression

Stéphane Girard*

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

Hadrien Lorenzo, Jérôme Saracco

Univ. Bordeaux, CNRS, Bordeaux INP, IMB, UMR 5251, F-33400, Talence, France
Inria, IMB, UMR 5251, F-33400, Talence, France

Abstract

Since its introduction in the early 90's, the Sliced Inverse Regression (SIR) methodology has evolved adapting to increasingly complex data sets in contexts combining linear dimension reduction with non linear regression. The assumption of dependence of the response variable with respect to only a few linear combinations of the covariates makes it appealing for many computational and real data application aspects. This work proposes an overview of the most active research directions in SIR modeling from multivariate regression models to regularization and variable selection.

Keywords: Curse of dimensionality, Multivariate response, Regularization, Semi-parametric regression model, Sufficient dimension reduction, Variable selection.

Classification codes: 62H12, 62J07, 62J99.

1. Introduction

Let us consider a regression setting where the goal is to estimate the link between a univariate response variable Y and a covariate $\mathbf{X}$. When the dimension p of the covariate is one or two, a simple two-dimensional or three-dimensional plot can reveal the relationships between $\mathbf{X}$ and Y, and thus be useful in determining the regression strategy to be used. Such an approach is not feasible when p is large. A possibility to overcome dimensionality problems arising in the regression context is to make the assumption that the response variable does not depend on the whole predictor space but only 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 the notion of dimension-reduction subspace (DRS) [22]. A subspace S is a DRS if Y is independent of $\mathbf{X}$ given the orthogonal projection of $\mathbf{X}$ onto S. In other words, all the information carried by the covariate $\mathbf{X}$ on Y can be compressed in its projection onto S. It is then of particular interest to estimate a DRS since, once it is identified, the initial regression problem can be solved equivalently using the low-dimensional projection of $\mathbf{X}$ onto the subspace.

Among methods providing an estimation of the dimension-reduction subspace, Sliced Inverse Regression (SIR), introduced 30 years ago in [51], is one of the most popular with 33500 entries on Google Scholar. The basic principles of SIR are recalled in Section 2. Since it is not possible to discuss all extensions and applications of SIR, we focus on three main hot topics. In Section 3, some extensions of SIR to a multidimensional response variable $\mathbf{Y}$ are presented. Section 4 discusses the adaptation of SIR to a (very) high-dimensional covariate $\mathbf{X}$ through the use of regularization methods. Finally, Section 5 is dedicated to variable selection techniques adapted to the SIR context. The paper is concluded with a short discussion.

*Corresponding author

Email addresses: stephane.girard@inria.fr (Stéphane Girard), hadrien.lorenzo@inria.fr (Hadrien Lorenzo), jerome.saracco@inria.fr (Jérôme Saracco)

2. Sliced Inverse Regression

The semi-parametric regression model associated with SIR is described first, with an emphasis on the underlying assumptions. The inference principle of SIR is then presented. Some popular extensions to SIR conclude this section.

2.1. Semi-parametric regression model

Consider a univariate response variable Y and a square-integrable multidimensional covariate $\mathbf{X} \in \mathbb{R}^p$ with $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X})$ and $\boldsymbol{\Sigma} = \mathbb{V}(\mathbf{X})$. Let $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K]$ be a $p \times K$ matrix (with $K \leq p$) where the $\boldsymbol{\beta}_k$'s are unknown p-dimensional vectors assumed to be linearly independent. The semi-parametric regression model

$$Y = g(\boldsymbol{\beta}^\top \mathbf{X}, \varepsilon) \tag{1}$$

is an attractive dimension reduction approach to model the effect of $\mathbf{X}$ on Y. The function g is an unknown arbitrary link function (with no shape assumptions) and the error term ε is assumed to be independent of $\mathbf{X}$ (with no distribution assumption). This kind of model is called link-free and distribution-free. Another way to understand the underlying dimension reduction framework is to write

$$Y \perp \mathbf{X} \,|\, \boldsymbol{\beta}^\top \mathbf{X}, \tag{2}$$

which means that Y is independent of $\mathbf{X}$ given $\boldsymbol{\beta}^\top \mathbf{X}$. Therefore, one can replace $\mathbf{X} \in \mathbb{R}^p$ by the index $\boldsymbol{\beta}^\top \mathbf{X} \in \mathbb{R}^K$ without loss of information on the regression of Y on $\mathbf{X}$. Note that $\boldsymbol{\beta}$ always exists since one may consider $\boldsymbol{\beta} = \mathbf{I}_p$ where $\mathbf{I}_p$ is the $p \times p$ identity matrix, but there is no dimension reduction in this case. In contrast, when $K < p$, there is an effective dimension reduction.

Let $S(\boldsymbol{\beta})$ be the K-dimensional linear subspace of $\mathbb{R}^p$ spanned by the columns of $\boldsymbol{\beta}$. Without additional assumptions on g and $\boldsymbol{\beta}$, the parameter $\boldsymbol{\beta}$ is not entirely identifiable, only the subspace $S(\boldsymbol{\beta})$ is. It is referred to as a DRS or an effective dimension reduction (EDR) subspace [32, 51]. Moreover, any direction belonging to this subspace is called an EDR direction. The smallest DRS is referred to as the central subspace, see [22] for further details.

When the dimension p of $\mathbf{X}$ is high, the relationship between the response and the covariate may be difficult to handle. Hence, the semi-parametric regression model (1) appears to be a nice alternative to parametric and non-parametric modeling, both of which seem to suffer from the well-known curse of dimensionality. Note that the idea of dimension reduction in model (1) is natural since it aims at constructing a low dimensional projection of the covariate without loosing information to predict the response Y. When the dimension K of the EDR subspace is small ($K \ll p$), it first facilitates data visualization and exploration, and it alleviates, in a second step, the curse of dimensionality in the estimation of the link function g.

The goal of SIR is to estimate the EDR subspace in model (1) from a sample $\{(\mathbf{X}_i, Y_i),\ i \in \{1, \ldots, n\}\}$. This yields estimated K-dimensional indices $\{\widehat{\mathbf{B}}^\top \mathbf{X}_i,\ i \in \{1, \ldots, n\}\}$ where $\widehat{\mathbf{B}} = [\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_K]$ denotes an estimated basis of the EDR subspace. Then, in a second step, the link function g can be estimated from the pseudo-sample $\{(Y_i, \widehat{\mathbf{B}}^\top \mathbf{X}_i),\ i \in \{1, \ldots, n\}\}$ with parametric or non-parametric estimators.

2.2. Estimation of the EDR subspace with SIR

Most of estimation approaches are based on the eigen-decomposition of a certain matrix of interest. The most popular one is SIR, introduced by [32] and [51], respectively for single index models ($K = 1$) and multiple index models ($K > 1$). Since then, SIR has been extensively studied, see for instance [5, 16, 19, 89], among others.

In this section, we focus on the SIR approach when the sample size n is larger than the dimension p of the covariate, see Section 4 and Section 5 for solutions to tackle the case where $n \leq p$. We first provide a characterization of the EDR subspace from a population point of view. Then, the corresponding estimation process of the EDR subspace is derived. In the name of the SIR method, Inverse corresponds to the use of a geometrical property of the expectation of $\mathbf{X}$ given Y, that is of $\mathbb{E}[\mathbf{X}|Y]$, while Sliced refers the discretization of Y in order to simplify the estimation of the moments appearing in the above-mentioned geometrical property.

The basic principle of the SIR method is to reverse the role of Y and $\mathbf{X}$, that is, instead of regressing the univariate variable Y on the p-dimensional covariate $\mathbf{X}$, the covariate $\mathbf{X}$ is regressed on the response variable Y. This is often denoted as the inverse regression step. The price to pay for inverting the roles of $\mathbf{X}$ and Y and retrieve the EDR subspace is an additional assumption on the distribution of $\mathbf{X}$, named the linearity condition:

$$\text{For all } \mathbf{b} \in \mathbb{R}^p, \mathbb{E}[\mathbf{b}^\top \mathbf{X} \,|\, \boldsymbol{\beta}^\top \mathbf{x}] \text{ is linear w.r.t. } \boldsymbol{\beta}^\top \mathbf{x}. \tag{3}$$

This linearity condition is discussed in details by [19]. Note that (3) is satisfied when $\mathbf{X}$ is elliptically distributed (for instance, normally distributed). Moreover, simulation studies showed that, in practice, SIR is robust to minor violations of (3). Besides, [43] mentioned that, for large dimension p, condition (3) is approximately fulfilled.

Assuming condition (3) together with model (1), or equivalently model (2), [51] showed that the centered inverse regression curve is included in the linear subspace spanned by the columns of the $p \times K$ matrix $\mathbf{\Sigma}\boldsymbol{\beta}$. More specifically, let $\mathbf{M} = \mathbb{V}[\mathbb{E}\{\mathbf{X} \,|\, T(Y)\}]$, where T denotes a monotonic transformation of Y. Then, the eigenvectors associated with the largest K eigenvalues of the $\mathbf{\Sigma}$-symmetric matrix $\mathbf{\Sigma}^{-1}\mathbf{M}$ are EDR directions.

To estimate the matrix $\mathbf{M}$, [51] proposed a transformation T, called a slicing, which categorizes the response Y into a new discrete response with H levels. This is often denoted as the slicing step. The condition $H > K$ is required to avoid an artificial reduction of dimension. To this end, the support of Y is partitioned into H non-overlapping slices $s_1, \ldots, s_H$. With such a transformation, the matrix of interest $\mathbf{M}$ can be rewritten as

$$\mathbf{M} = \sum_{h=1}^{H} p_h (\mathbf{m}^{(h)} - \boldsymbol{\mu})(\mathbf{m}^{(h)} - \boldsymbol{\mu})^\top,$$

where $p_h = \mathbb{P}[Y \in s_h]$ and $\mathbf{m}^{(h)} = \mathbb{E}[\mathbf{X} \,|\, Y \in s_h]$. In the following, the eigenvectors associated with the largest K eigenvalues of $\mathbf{\Sigma}^{-1}\mathbf{M}$ which are EDR directions are denoted by $\mathbf{b}_k$, $k \in \{1, \ldots, K\}$. It appears that $\mathbf{M}$ is a "between-class" covariance matrix and therefore SIR can be interpreted as a Fisher discriminant analysis (FDA) computed on sliced data. We refer to Figure 1 for an illustration and to [39, Chapter 4] for a general account on FDA.

For the estimation step, starting from a sample $\{(\mathbf{X}_i, Y_i), i \in \{1, \ldots, n\}\}$, matrices $\mathbf{\Sigma}$ and $\mathbf{M}$ are estimated by their empirical counterparts:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i, \quad \widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{X}_i - \hat{\boldsymbol{\mu}})(\mathbf{X}_i - \hat{\boldsymbol{\mu}})^\top \text{ and } \widehat{\mathbf{M}} = \sum_{h=1}^{H} \hat{p}^{(h)} (\widehat{\mathbf{m}}^{(h)} - \hat{\boldsymbol{\mu}})(\widehat{\mathbf{m}}^{(h)} - \hat{\boldsymbol{\mu}})^\top, \tag{4}$$

where

$$\hat{p}^{(h)} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_i \in s_h\} \text{ and } \widehat{\mathbf{m}}^{(h)} = \frac{1}{n\hat{p}^{(h)}}\sum_{i=1}^{n} \mathbf{X}_i \mathbb{I}\{Y_i \in s_h\}.$$

Here, $\mathbb{I}\{\cdot\}$ denotes the indicator function. As a consequence, the estimated EDR directions are the eigenvectors $\widehat{\mathbf{b}}_k$, $k \in \{1, \ldots, K\}$ associated with the K largest eigenvalues of $\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{M}}$ and spanning the estimated K-dimensional EDR subspace. From the theoretical point of view, there is no optimal slicing T. Practical choices are discussed in [51, 66]. In practice, the number of observations per slice is fixed to $\lfloor n/H \rfloor$ where $\lfloor \cdot \rfloor$ stands for the integer part. Note that, when the sample size n is not proportional to the number H of slices, some slices may contain $\lfloor n/H \rfloor + 1$ observations. Finally, let us also highlight that the main advantage of SIR method is its low computational cost.

Some asymptotic properties of the SIR estimator have been obtained. The $n^{1/2}$ consistency of the estimated EDR directions has been proved: $\widehat{\mathbf{b}}_k = \mathbf{b}_k + \mathbf{0}_p(n^{-1/2})$ for $k \in \{1, \ldots, K\}$. Moreover, the asymptotic normality of the estimated EDR directions is established, for instance in [47, 51, 65, 94].

The important issue of selecting the number of dimensions K of the EDR subspace has been addressed through two main lines of works, available in the literature. The selection of K can first be based on hypothesis testing procedures designed from the previously mentioned asymptotic results, see for example [6, 36, 51, 68]. Second, a number of graphical tools have been proposed [57, 58] for selecting simultaneously the number of slices H and the dimension K.

2.3. Some extensions to SIR

Among the numerous extensions to SIR proposed in the statistical literature, two research directions are briefly described hereafter: the use of higher conditional moments of $\mathbf{X}$ given Y for estimating the EDR directions and the proposal of alternatives to the slicing step. Recent contributions to SIR are then described in more details: the extension to a multidimensional response (multivariate SIR, Section 3), regularization of SIR and variable selection in SIR (Section 4 and Section 5 respectively) both to deal with non-regular cases.

The use of higher conditional moments of $\mathbf{X}$ given Y has been considered. Indeed, the original SIR approach is based on the conditional expectation of $\mathbf{X}$ given Y. However, it is also possible to retrieve information on the EDR subspace from higher (inverse) conditional moments of $\mathbf{X}$ given Y, such as the conditional variance $\mathbb{V}[\mathbf{X} \,|\, T(Y)]$.

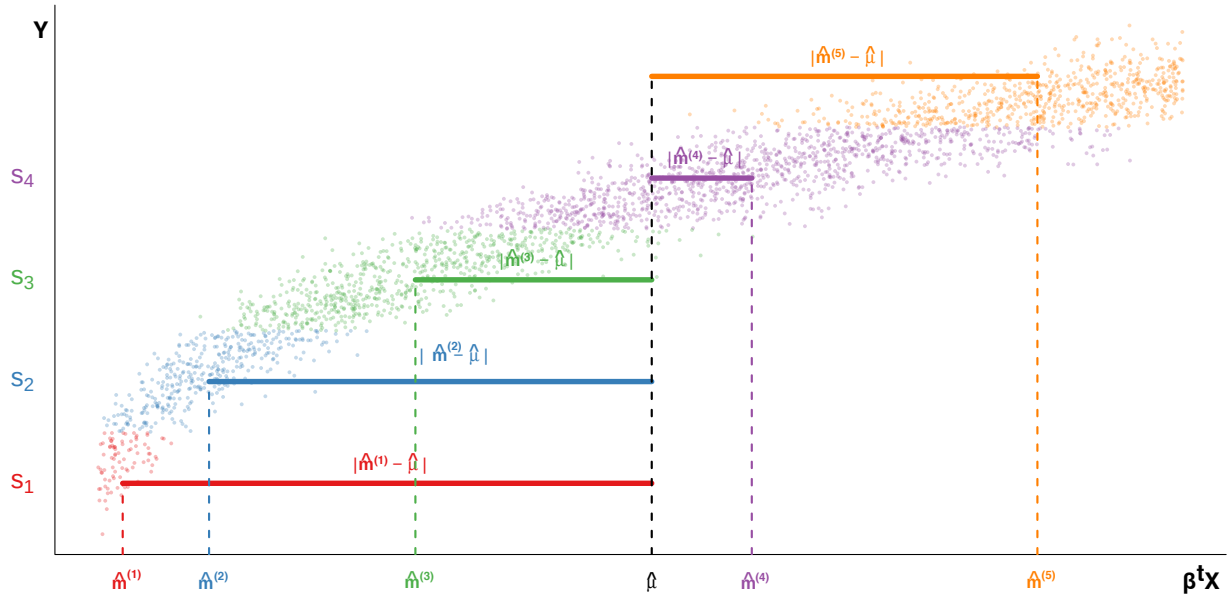

Fig. 1: Principle of the sample version of SIR on single index model ($K = 1$), an example with $H = 5$ slices.

This gives rise to the SIR-II method, see [51, 84] for further details, or sliced average variance estimation (SAVE), see [23, 56, 64, 90] among others.

Alternative SIR methods have been investigated in order to circumvent the slicing step. For instance, one can mention kernel-based methods [80, 82, 93] which may be hard to implement in practice and are computationally heavy. A parametric version of SIR is introduced in [12], and [46] proposed a nearest neighbor inverse regression method. Note that [4] introduced a pooled slicing approach to combine information from several slicings, while [49] recommend bagging versions of SIR.

3. Multivariate SIR

Originally, sliced inverse regression and, more generally, dimension reduction approaches were introduced for a scalar response variable Y and a p-dimensional covariate $\mathbf{X}$ to capture the relationship between $\mathbf{X}$ and Y with only a few linear combinations of $\mathbf{X}$ components and without imposing the form of the distribution of Y given $\mathbf{X}$. An important recent extension of the SIR methodology considers a multivariate response. Suppose that $\mathbf{Y} = (Y_1, \ldots, Y_q)^\top$ is a q-dimensional random vector (with $q > 1$). The underlying model assumption is:

$$\mathbf{Y} \perp \mathbf{X} \,|\, \boldsymbol{\beta}^\top \mathbf{X}, \tag{5}$$

which is a direct multivariate extension of (2). From a semi-parametric point of view, the corresponding regression model can be written similarly to (1) as

$$\mathbf{Y} = \mathbf{g}(\boldsymbol{\beta}^\top \mathbf{X}, \boldsymbol{\varepsilon}), \tag{6}$$

where $\mathbf{g} : \mathbb{R}^{K+r} \to \mathbb{R}^q$ is an arbitrary and unknown link function, and $\boldsymbol{\varepsilon}$ is a r-dimensional random error variable independent of $\mathbf{X}$. Slightly restrictive regression models can be also considered. For instance, a model with a q-dimensional additive error term is studied by [52]: $\mathbf{Y} = \mathbf{g}(\boldsymbol{\beta}^\top \mathbf{X}) + \boldsymbol{\varepsilon}$, while [28] focused on the following model

$$\left\{ \begin{array}{rcl} Y^{(1)} & = & g_1(\boldsymbol{\beta}^\top \mathbf{X}, \varepsilon^{(1)}), \\ & \vdots & \\ Y^{(q)} & = & g_q(\boldsymbol{\beta}^\top \mathbf{X}, \varepsilon^{(q)}), \end{array} \right. \tag{7}$$

where $Y^{(j)}$ (resp. $\varepsilon^{(j)}$) stands for the jth component of $\mathbf{Y}$ (resp. of ε) and the link functions $g_j,\ j \in \{1,\dots,q\}$ are unknown real-valued functions.

It is possible to generalize to the multivariate response case the various existing SIR methods by straightforwardly replacing the conditional distribution of Y given $\mathbf{X}$ by that of $\mathbf{Y}$ given $\mathbf{X}$ to retrieve the EDR subspace. However, in this high dimensional context, dedicated SIR methodologies have been developed, and new issues also emerged. Let us now present some of these multivariate SIR approaches.

3.1. Complete slicing approach and associated methods

The first basic idea extends directly the slicing step to the multivariate response $\mathbf{Y}$. The complete slicing procedure consists in building recursively the slices (of nearly equal weights) from the slicing of one $\mathbf{Y}$ component by slicing each class according to the next component of $\mathbf{Y}$. From a practical point of view, several computational problems arise as the dimension q of $\mathbf{Y}$ increases, since the number H of slices grows exponentially fast with q.

To circumvent this difficulty, [46] proposed to use a slicing of the observations $\mathbf{Y}_i$'s based on the nearest neighbors approach. In the same vein, [72] introduced the k-means inverse regression (KIR) where simple slices are replaced by clusters obtained from the k-means algorithm, here $T(\mathbf{Y}) = \tilde{\mathbf{Y}} \in \mathbb{R}$ the vector containing the assignments to the clusters. Note that both the number of neighbors and clusters in KIR have to be chosen by the user (instead of the number H of slices in SIR). Similarly, k-medoids inverse regression was proposed by [11].

Remark. Multivariate SIR or KIR methods only use the information from the inverse conditional expectation $\mathbb{E}[\mathbf{X} \,|\, T(\mathbf{Y})]$ where $T(\mathbf{Y})$ is univariate and contains the assignments to the slices (for SIR) or clusters (for KIR). All these methods do not take into account intra-cluster (or intra-slice) information, which could be valuable and substantial for a continuous response variable. Based on the method introduced by [27] which allows to recover the intra-slice information for a univariate Y, [79] provides an extension to a multivariate continuous response $\mathbf{Y}$. The corresponding method is named generalized k-means inverse regression estimation (GM.KIRE).

3.2. Marginal slicing approaches

Another way to overcome the above mentioned issue is implemented in marginal slicing approaches which focus on a specific function f of $\mathbf{Y}$ (which is not a slicing) so that the dimension of $f(\mathbf{Y})$ is smaller. The choice of f depends on the problem of interest. For instance, $f(\mathbf{Y})$ can be the average or the median of the q components of $\mathbf{Y}$, and the usual slicing step of SIR is then applied to the univariate dependent part $f(\mathbf{Y})$. The user can also choose for $f(\mathbf{Y})$, the first relevant components of the principal component analysis of the $\mathbf{Y}_i$'s, and then the slices are constructed via a complete slicing method based on this lower dimensional projection of $\mathbf{Y}$.

3.3. Pooled marginal slicing approaches

Based on the regression model (7), the underlying idea of the pooled marginal slicing method is to combine information from the q univariate SIR methods applied to each component $Y^{(j)}$ of $\mathbf{Y}$ and $\mathbf{X}$. The way to aggregate the q marginal SIR results is as follows: for some positive weights $w_1,\dots,w_q$, let

$$\boldsymbol{M}_P = \sum_{j=1}^{q} w_j \mathbb{V}[\mathbb{E}[\mathbf{X} \,|\, T_j(Y^{(j)})]],$$

where $T_j(Y^{(j)})$ stands for the usual slicing applied to the jth component of $\mathbf{Y}$. The eigenvectors associated with the largest K eigenvalues of $\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_P$ are EDR directions. Two ways to choose the w_j's have been proposed in [3]: weights proportional to the largest eigenvalues of each marginal SIR, or simply equal weights. A very similar multivariate SIR approach was also introduced by [60].

An alternative way to combine information from the q marginal SIR methods is proposed in [28]. Denote by $\mathbf{B}_j$ the $p \times K$ matrix spanning the EDR subspace from the marginal regression of $Y^{(j)}$ given $\mathbf{X}$, for $j \in \{1,\dots,q\}$. Let us introduce the following proximity measure between the q marginal EDR subspaces, $\mathrm{Span}(\mathbf{B}_j),\ j \in \{1,\dots,q\}$, and the linear subspace $\mathrm{Span}(\mathbf{D})$, where $\mathbf{D}$ is a $p \times K$ matrix such that $\mathbf{D}^\top \boldsymbol{\Sigma} \mathbf{D} = \mathbf{I}_k$:

$$Q(\mathbf{D}, \mathbf{B}_1, \dots, \mathbf{B}_q) = \frac{1}{q} \sum_{j=1}^{q} \left(\frac{1}{K} \mathrm{trace}(P_{\mathbf{D}} P_{\mathbf{B}_j}) \right) \in [0,1], \tag{8}$$

where $P_{\mathbf{M}}$ stands for the $\mathbf{\Sigma}$-orthogonal projector on the linear subspace spanned by the columns of the $p \times K$ matrix $\mathbf{M}$. Note that $\text{Span}(\mathbf{D}) = \text{Span}(\mathbf{B}_1) = \cdots = \text{Span}(\mathbf{B}_q)$ implies $Q(\mathbf{D}, \mathbf{B}_1, \ldots, \mathbf{B}_q) = 1$. The closer to the q marginal EDR subspaces is the linear subspace $\text{Span}(\mathbf{D})$, the closer to one is this measure. Under model (7), the following estimator was introduced by [28]:

$$\mathbf{V} = \arg\max_{\mathbf{D}} Q(\mathbf{D}, \mathbf{B}_1, \ldots, \mathbf{B}_q).$$

It has been shown that $\text{Span}(\mathbf{V}) = \text{Span}(\mathbf{B})$, i.e. $\text{Span}(\mathbf{V})$ is an EDR subspace. Moreover, to compute such an estimator, the $p \times K$ matrix $\mathbf{V}$ is made with the eigenvectors associated with the K non-null eigenvalues of $\mathcal{B}\mathcal{B}^\top\mathbf{\Sigma}$ where $\mathcal{B}$ is the $p \times (qK)$ matrix defined as $\mathcal{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_q]$. A weighted version of this multivariate SIR method has been also proposed, following the idea of the first pooled marginal slicing approach in which the matrix of interest $\boldsymbol{M}_P$ is a weighted average of the marginal matrices $\mathbb{V}[\mathbb{E}[\mathbf{X} \,|\, T_j(Y^{(j)})]]$. Marginal weights' $K \times K$ matrices $\mathbf{W}_j$ can be used and then be aggregated in the global weighting matrix $\mathcal{W} = \text{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_q)$. The estimated EDR subspace is thus spanned by the eigenvectors associated with the largest K eigenvalues of $\mathcal{B}\mathcal{W}\mathcal{B}^\top\Sigma$.

3.4. Alternating SIR approach

The philosophy of this approach uses the duality existing between the "SIR variates" and the "most predictable (MP) variates" as introduced by [52]. To simplify the notation, let us consider the case where $K = 1$. The SIR variate is defined as $\mathbf{b}^\top \mathbf{X}$ where $\mathbf{b}$ is an EDR direction estimated by SIR. The MP variate is defined as $\boldsymbol{\theta}^\top \mathbf{Y}$ obtained from the minimization of the ratio $\mathbb{E}[\mathbb{V}[\boldsymbol{\theta}^\top \mathbf{Y} \,|\, \mathbf{X}]]/\mathbb{V}[\boldsymbol{\theta}^\top \mathbf{Y}]$ where $\mathbb{V}[\boldsymbol{\theta}^\top \mathbf{Y} \,|\, \mathbf{X}]$ can be interpreted as the associated prediction mean square error of the best nonlinear prediction $\mathbb{E}[\boldsymbol{\theta}^\top \mathbf{Y} \,|\, \mathbf{X}]$ under the square loss error. The MP variate can be equivalently obtained by maximizing the ratio $\mathbb{V}[\mathbb{E}[\boldsymbol{\theta}^\top \mathbf{Y} \,|\, \mathbf{X}]]/\mathbb{V}[\boldsymbol{\theta}^\top \mathbf{Y}]$ which conducts to the same eigen-decomposition problem as SIR by reversing the roles played by $\mathbf{Y}$ and $\mathbf{X}$. Thus, the alternating SIR procedure proposed by [52] works as follows: alternate MP steps and SIR steps where the slicing of each step is made either on the current SIR variate $\mathbf{b}^\top \mathbf{X}$ or on the current MP variate $\boldsymbol{\theta}^\top \mathbf{Y}$. The initialization is done using the canonical direction $\boldsymbol{\theta}_c^\top \mathbf{Y}$ to ensure the convergence of the algorithm. An application to reference curves estimation based on a methodology combing alternating SIR and kernel estimation of conditional quantiles is described in [40] in the context of a study of biophysical properties of the skin of healthy French women.

3.5. Clustering of the components of $\mathbf{Y}$ associated with the same EDR subspace

The marginal pooled slicing approach relies on model (7) which assumes an unique common EDR subspace, $\text{Span}(\boldsymbol{\beta})$, for all q components of $\mathbf{Y}$. This assumption may appear too restrictive in many real data applications. Therefore, applying a multivariate SIR method on $\mathbf{Y}$ is unlikely to provide a suitable dimension reduction. In contrast, it makes sense to assume that only groups of components of $\mathbf{Y}$ rely on model (7) with small values of K. In [28], a more general semi-parametric regression model is considered for a multivariate response, with potentially $L > 1$ underlying EDR subspaces $\text{Span}(\boldsymbol{\beta}_1) \neq \text{Span}(\boldsymbol{\beta}_2) \neq \cdots \neq \text{Span}(\boldsymbol{\beta}_L)$:

$$\left\{\begin{array}{lcl} Y^{(1)} & = & g_1(\boldsymbol{\beta}_1^\top \mathbf{X}, \varepsilon^{(1)}), \\ & \vdots & \\ Y^{(q_1)} & = & g_{q_1}(\boldsymbol{\beta}_1^\top \mathbf{X}, \varepsilon^{(q_1)}), \\ Y^{(q_1+1)} & = & g_{q_1+1}(\boldsymbol{\beta}_2^\top \mathbf{X}, \varepsilon^{(q_1+1)}), \\ & \vdots & \\ Y^{(q_1+q_2)} & = & g_{q_1+q_2}(\boldsymbol{\beta}_2^\top \mathbf{X}, \varepsilon^{(q_1+q_2)}), \\ & \vdots & \\ Y^{(q)} & = & g_q(\boldsymbol{\beta}_L^\top \mathbf{X}, \varepsilon^{(q)}), \end{array}\right. \tag{9}$$

with $\sum_{\ell=1}^{L} q_\ell = q$. Under this model, estimating the underlying L EDR subspaces is clearly more appropriate than seeking a common EDR subspace when trying to reduce as much as possible the dimension. To this end, it is necessary to cluster the components of $\mathbf{Y}$ associated with the same EDR subspace. Then, for each identified cluster, it is possible to estimate the associated common EDR subspace. Note that the dimension of each EDR subspace may vary. In this complex framework, [28] proposed a way to properly cluster the components of $\mathbf{Y}$ and to detect the possible existence of a common EDR subspace.

3.6. Some extensions of multivariate SIR

Extensions of multivariate SIR methods (complete slicing, marginal slicing, pooled marginal slicing, alternating SIR) to the case where the underlying SIR method is replaced by the SIR_α method have been proposed in [7], where SIR_α is a hybrid approach combining information from SIR and SIR-II via a tuning parameter $\alpha \in [0, 1]$. Note that if $\alpha = 0$ (resp. 1, 0.5), SIR_α is equivalent to SIR (resp. SIR-II, SAVE). For instance, [67] provides asymptotic results for the pooled marginal slicing estimator based on SIR_α approach, while moment-based dimension reduction methods for a multivariate response $\mathbf{Y}$ were developed by[83]. The case of missing values in $\mathbf{Y}$ has been explored by [34] since multivariate SIR approaches cannot be used directly. The specific case of a bivariate response was studied by [78], especially when $\mathbf{Y}$ is a mix of continuous and categorical responses. Considering semi-parametric multivariate sample selection models, [18] used multivariate SIR (and more specifically pooled marginal slicing) to estimate the slope vectors in the selection equation and in the outcome equation. Prediction regions through multivariate inverse regression were studied by [30]. In [50], an interesting projective resampling procedure for estimating the dimension reduction space in model (5) was proposed. Other efficient dimension reduction approaches are available in the multivariate response context, see for example [85, 87, 92].

4. Regularization for SIR

Assuming that K is known, and recalling that $\boldsymbol{\Sigma} = \mathbb{V}[\mathbf{X}]$ and $\mathbf{M} = \mathbb{V}[\mathbb{E}[\mathbf{X} \mid Y]]$, EDR directions are obtained by computing the eigenvectors associated to the largest K eigenvalues of $\boldsymbol{\Sigma}^{-1}\mathbf{M}$. Unfortunately, the classical $n-$sample estimate $\widehat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$ can be singular, or at least ill-conditioned, in several situations. Indeed, since $\text{rank}(\widehat{\boldsymbol{\Sigma}}) \leq \min(n-1, p)$, if $n \leq p$ then $\widehat{\boldsymbol{\Sigma}}$ is necessarily singular. Even when n and p are of the same order, $\widehat{\boldsymbol{\Sigma}}$ may be ill-conditioned, and its inversion introduces numerical instabilities in the estimation of the EDR directions. Similar phenomena occur when the coordinates of $\mathbf{X}$ are highly correlated. As an illustration, let us consider $n = 100$ simulated data from model (1) with link function $g(t, \varepsilon) = \sin(\pi t/2) + \varepsilon$ and where ε is a Gaussian random variable: $\varepsilon \sim \mathcal{N}_1(0, 9.10^{-4})$. The $p-$dimensional covariate $\mathbf{X}$ is Gaussian distributed: $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \mathbf{Q}\boldsymbol{\Delta}\mathbf{Q}^\top$ with $\boldsymbol{\Delta} = \text{diag}(p^2, \ldots, 1^2)$ and $\mathbf{Q}$ is an orientation matrix drawn from the uniform distribution on the set of orthogonal matrices. The condition number of $\boldsymbol{\Sigma}$ is thus given by $\kappa(\boldsymbol{\Sigma}) = p^2$. Finally, we focus on a single-index model, $K = 1$ and $\boldsymbol{\beta} = 5^{-1/2}\mathbf{Q}(1, 1, 1, 1, 1, 0, \ldots, 0)^\top$. The results obtained by the SIR method are depicted on Figure 2 in two situations: dimension $p = 10$ (left panel) and $p = 50$ (right panel). It appears that SIR performs well in dimension $p = 10$ (with associated condition number $\kappa(\boldsymbol{\Sigma}) = 100$): the estimated projections are highly correlated with the true ones and the projected sample $\{(\hat{\mathbf{b}}^\top \mathbf{X}_i, Y_i), i \in \{1, \ldots, n\}\}$ allows to recover the shape of the link function. At the opposite, in dimension $p = 50$ (with associated condition number $\kappa(\boldsymbol{\Sigma}) = 2500$), SIR no longer works properly: the estimated projections are not correlated anymore with the true ones and the projected sample does not allow to recover the shape of the link function.

Some regularizations of the SIR method have been proposed to overcome this limitation due to the ill-conditioning of $\boldsymbol{\Sigma}$. They can be classified in three main families.

4.1. Dealing directly with inversion issues

SIR-QZ method [29] uses the QZ algorithm [62] to solve the generalized eigenvalue problem

$$\widehat{\mathbf{M}}\mathbf{b} = \lambda\widehat{\boldsymbol{\Sigma}}\mathbf{b}, \tag{10}$$

rather than the original one $\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{M}}\mathbf{b} = \lambda\mathbf{b}$. The SIR-MP method [29] adapts an approach introduced in functional sliced inverse regression (i.e., when $\mathbf{X}$ is an explanatory functional variable), described in [2, 37]. This method consists in using the eigen-decomposition of $\mathbf{M}^+\boldsymbol{\Sigma}$ instead of $\boldsymbol{\Sigma}^{-1}\mathbf{M}$, where $\mathbf{M}^+$ is the Moore-Penrose generalized inverse of $\mathbf{M}$.

Similarly, in [21, 54], a principal component analysis (PCA) is used as a preprocessing step in order to eliminate the directions in which the random vector $\mathbf{X}$ is degenerated. Thus, for a properly chosen dimension d of the projection subspace, the covariance matrix of the projected observations is regular, since null eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ have been discarded. Another method consists in adopting a ridge regression technique (see for instance [31, Chapter 17]) which replaces the sample estimate $\widehat{\boldsymbol{\Sigma}}$ by a perturbed version $\widehat{\boldsymbol{\Sigma}} + \tau\mathbf{I}_p$ where τ is a positive real number [88]. Here, the idea is that, for τ large enough, $\widehat{\boldsymbol{\Sigma}} + \tau\mathbf{I}_p$ is regular and its condition number $\kappa(\widehat{\boldsymbol{\Sigma}} + \tau\mathbf{I}_p)$ decreases with τ since all eigenvalues have been increased by τ. Similarly, in [70, 71], regularized discriminant analysis [38] is adapted to the

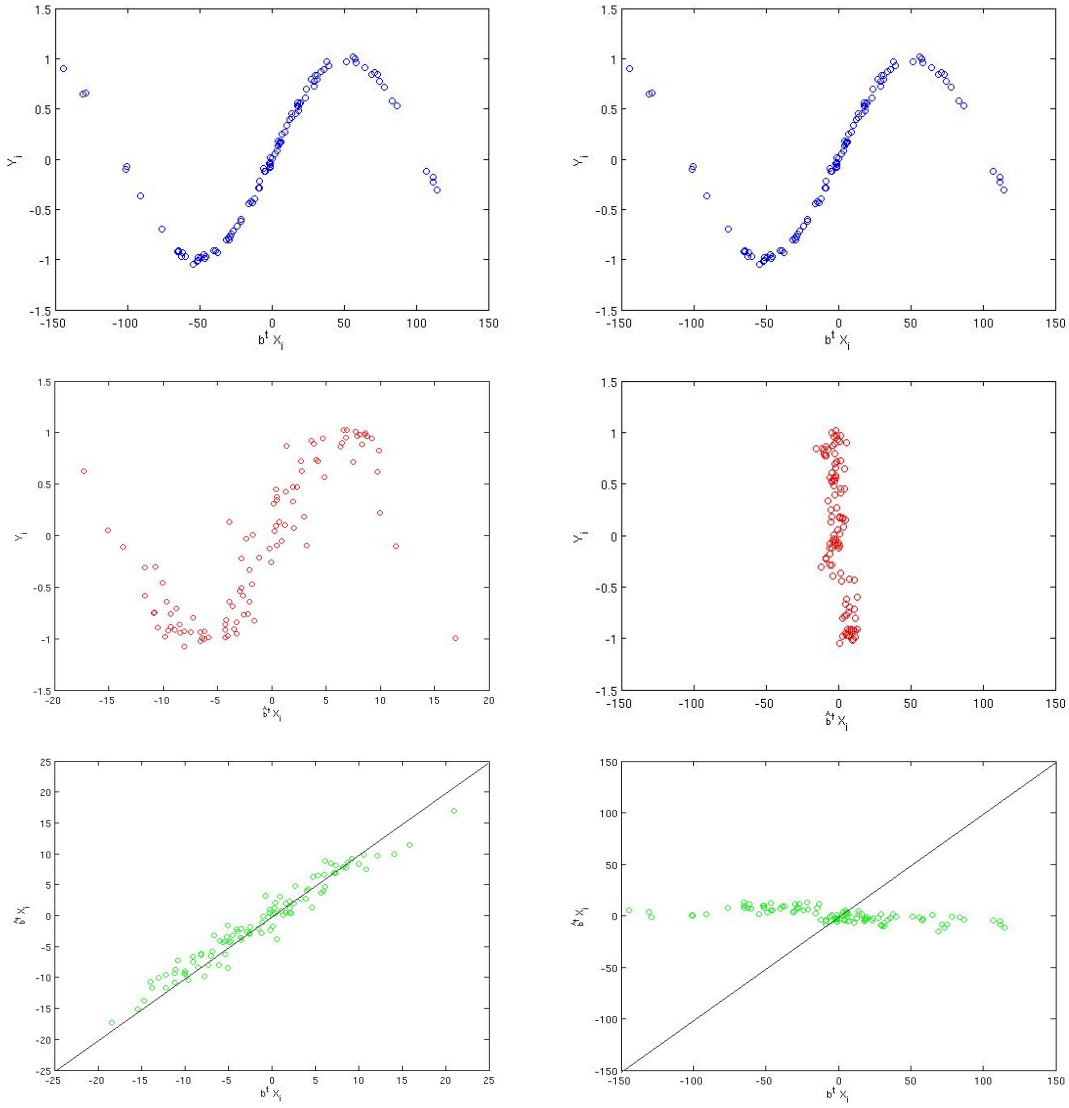

Fig. 2: Some SIR results on simulate data in dimension $p = 10$ (left panel) and $p = 50$ (right panel). Top (blue): Y_i versus the projection $\boldsymbol{\beta}^\top \mathbf{X}_i$ on the true direction $\boldsymbol{\beta}$, center (red): Y_i versus the projection $\hat{\mathbf{b}}^\top \mathbf{X}_i$ on the estimated direction $\hat{\mathbf{b}}$, bottom (green): estimated projection $\hat{\mathbf{b}}^\top \mathbf{X}_i$ versus true projection $\boldsymbol{\beta}^\top \mathbf{X}_i$.

SIR framework thanks to the interpretation of SIR as a particular FDA method. The sample estimate $\widehat{\boldsymbol{\Sigma}}$ is replaced by a shrunk version $\widehat{\boldsymbol{\Sigma}}(\tau) = (1-\tau)\widehat{\boldsymbol{\Sigma}} + (\tau/p)\text{trace}(\widehat{\boldsymbol{\Sigma}})\mathbf{I}_p$. As τ tends to 1, $\widehat{\boldsymbol{\Sigma}}(\tau)$ approaches a diagonal matrix proportional to $\mathbf{I}_p$ whose condition number tends to 1. The practical choice of τ may be challenging. A cross-validation criterion adapted to the regression framework may be used, provided that the estimation of $\boldsymbol{\beta}$ is coupled with the estimation of the link function g, see (22) below.

4.2. Optimization-based approaches

The SIR method can be interpreted as a least-square optimization problem [24] introducing the function

$$F(\boldsymbol{B}, \boldsymbol{C}) = \sum_{h=1}^{H} \hat{p}^{(h)} \|\widehat{\boldsymbol{Z}}^{(h)} - \boldsymbol{B}\boldsymbol{C}_h\|^2, \tag{11}$$

where $\boldsymbol{C} = (\boldsymbol{C}_1, \ldots, \boldsymbol{C}_H)$ is an auxiliary variable, and $\widehat{\boldsymbol{Z}}^{(h)} = \widehat{\boldsymbol{\Sigma}}^{-1/2}(\widehat{\mathbf{m}}^{(h)} - \hat{\boldsymbol{\mu}})$ is the associated empirical estimate of the standardized mean in slice $\bar{\boldsymbol{Z}}^{(h)} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{m}^{(h)} - \boldsymbol{\mu})$. Then the optimization problem becomes

$$\left(\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{C}}\right) = \arg\min_{\boldsymbol{B},\boldsymbol{C}} F(\boldsymbol{B}, \boldsymbol{C}) \tag{12}$$

and $\widehat{\boldsymbol{B}}$ spans an estimate of the EDR subspace. One may then introduce L_1- and L_2- penalty terms in (12) to regularize the estimate of $\boldsymbol{B}$. This approach is detailed in [55] and discussed in [8]. More precisely, following [63], they focus on a new formulation of (11) in the original $\mathbb{R}^p$-space introducing the matrix $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{B}$ and involving the new function

$$G(\mathbf{A}, \boldsymbol{C}) = \sum_{h=1}^{H} \hat{p}^{(h)} \|\widehat{\mathbf{m}}^{(h)} - \hat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\Sigma}}\mathbf{A}\boldsymbol{C}_h\|^2. \tag{13}$$

The estimate of the central subspace becomes $\text{Span}(\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\mathbf{A}})$ where $\widehat{\mathbf{A}}$ minimizes $G(\mathbf{A}, \boldsymbol{C})$. This allows the introduction of a ridge-like optimization problem w.r.t. the first variable $\mathbf{A}$ for all $\tau > 0$:

$$\left(\widehat{\mathbf{A}}, \widehat{\boldsymbol{C}}\right) = \arg\min_{\mathbf{A},\boldsymbol{C}} \; G(\mathbf{A}, \boldsymbol{C}) + \tau\|\mathbf{A}\|_F^2, \tag{14}$$

where $\|.\|_F$ denotes the Frobenius norm. The regularized optimization problem (14) is solved thanks an alternating least-square algorithm and showed good performances both on simulated and real cases, see [55]. However, it is noted in [8] that the optimization problem (14) suffers from a non-invariance with respect to linear transformations. This point is detrimental since the objective of the SIR method is to estimate a linear subspace, which should therefore be invariant w.r.t. linear transformations. Starting from this observation, it is proved in [8] that the only solution for $\mathbf{A}$ is the degenerate null solution.

To fix this vexing effect, one may replace (14) by

$$\min_{\mathbf{A},\boldsymbol{C}} G(\mathbf{A}, \boldsymbol{C}) + \tau \sum_{h=1}^{H} \|\mathbf{A}\boldsymbol{C}_h\|^2. \tag{15}$$

It is easily shown that this problem is invariant w.r.t. linear bijective transformations [8]. Starting from a solution $(\widehat{\mathbf{A}}, \widehat{\boldsymbol{C}})$, any pair $(\widehat{\mathbf{A}}\mathbf{K}, \mathbf{K}^{-1}\widehat{\boldsymbol{C}})$ is still a solution for all regular matrix $\mathbf{K}$. It is also shown that $\widehat{\mathbf{A}}$, the solution of (15), coincides with the Ridge estimator, introduced by [55], since its columns are the first K eigenvectors of the matrix $(\widehat{\boldsymbol{\Sigma}} + \tau\mathbf{I})^{-1}\widehat{\mathbf{M}}$. In [55], it is also proposed to introduce a L_1-penalty term in (14) to obtain parsimonious estimates of the EDR directions, see Section 5 for variable selection techniques.

Finally, it is noted in [10] that the generalized eigenvalue problem (10) can be rewritten as

$$\arg\max_{\mathbf{B}\in\mathbb{R}^{p\times K}} \frac{\boldsymbol{B}^\top\mathbf{M}\boldsymbol{B}}{\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B}}. \tag{16}$$

Graph Laplacian-based regularization is then introduced ensuring that the solution of the associated optimization problem is invariant w.r.t. scalar and orthogonal transformations. In practice, the solution is computed using a conjugate gradient method on the Grassmann manifold [33].

4.3. Interpretation in a Bayesian framework

The approach of [9] is based on the interpretation of the axes spanning the central subspace as solutions of an inverse regression problem [25]. Focusing on the single-index situation ($K = 1$), the model is written as

$$\mathbf{X} = \boldsymbol{\mu} + c(Y)\boldsymbol{\Sigma}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{17}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\varepsilon}$ is a random vector and $c(\cdot)$ is a univariate real function. The latter is expanded as a linear combination of H basis functions $S_h(\cdot)$, $h \in \{1, \ldots, H\}$:

$$c(\cdot) = \sum_{h=1}^{H} \gamma^{(h)} S_h(\cdot),$$

where the coefficients $\gamma^{(h)}$, $h \in \{1, \ldots, H\}$ are unknown. Introducing $\boldsymbol{\gamma} = \left(\gamma^{(1)}, \ldots, \gamma^{(H)}\right)^\top$ and $\mathbf{S}(\cdot) = (S_1(\cdot), \ldots, S_H(\cdot))^\top$, model (17) can be rewritten as

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{S}^\top(Y)\boldsymbol{\gamma}\boldsymbol{\Sigma}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{18}$$

Assuming that $\boldsymbol{\varepsilon} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ can be estimated by maximum likelihood (ML), or equivalently by considering the optimization problem

$$\arg\min_{\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta},\boldsymbol{\gamma}} \log\det\boldsymbol{\Sigma} + \operatorname{trace}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) + (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}} + \bar{\mathbf{S}}^\top\boldsymbol{\gamma}\boldsymbol{\Sigma}\boldsymbol{\beta})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}} + \bar{\mathbf{S}}^\top\boldsymbol{\gamma}\boldsymbol{\Sigma}\boldsymbol{\beta}) + (\boldsymbol{\gamma}^\top W\boldsymbol{\gamma})(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}) - 2\boldsymbol{\gamma}^\top\Gamma\boldsymbol{\beta}, \tag{19}$$

where

$$\Gamma = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{S}(Y_i) - \bar{\mathbf{S}})(\mathbf{X}_i - \hat{\boldsymbol{\mu}})^\top, \quad W = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{S}(Y_i) - \bar{\mathbf{S}})(\mathbf{S}(Y_i) - \bar{\mathbf{S}})^\top, \quad \bar{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{S}(Y_i).$$

It is shown in [9] that, when $S_h(\cdot) = \mathbb{I}\{\cdot \in s_h\}$, $h \in \{1, \ldots, H\}$, the ML estimators of $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ are explicit and the ML estimator of $\boldsymbol{\beta}$ coincides with the SIR estimator. A Gaussian prior is then introduced on the unknown parameters of (18) in order to regularize their estimation, which amounts to introducing quadratic penalty terms in (19). The maximum a posteriori estimator is still explicit and it is shown that the previously mentioned techniques [21, 54, 70, 71, 88] can all enter this framework, providing a new interpretation in terms of Gaussian priors. New priors are also proposed leading to new Tikhonov regularizations [76, Chapter 1] of the SIR method.

5. Variable selection in SIR

Variable selection is a general framework in regression methods with several possible benefits. First, it dampens the curse of dimensionality by reducing the dimension of the covariate. Variable selection can therefore be interpreted as a regularization technique. Second, it simplifies the interpretation of the regression model by reducing its potential complexity. Finally, variable selection may also speed up computations since most of inference procedures are strongly affected by the dimension p.

The approaches detailed below are divided in two main families. The first one gathers optimization-based approaches inspired from regularization techniques, while the second one gathers computational-based approaches involving the assessment of numerous possible configurations. In the following, "$\underline{\bullet}$" denotes the matrix/vector of n realizations (collected in rows) of the associated random vector/variable "$\bullet$".

5.1. Optimization-based approaches

As discussed in Section 4, re-interpreting the SIR method as an optimization problem (see (11) and (12)) opens the door to the introduction of L_1-regularization techniques to obtain sparse estimates. Shrinkage-SIR [63] is a two-step approach based first on a preliminary non-sparse estimate of the SIR directions and second, on the use of a shrinkage index $\boldsymbol{\alpha}$ to get a sparse estimate of the EDR subspace. Assuming that the pair $(\widehat{\mathbf{A}}, \widehat{\boldsymbol{C}})$ minimizing the function F, defined in (11), has been computed, a new optimization problem is introduced, considering the function

$$H(\boldsymbol{\alpha}) = \sum_{h=1}^{H} \hat{p}^{(h)} \left\| \widehat{\mathbf{m}}^{(h)} - \hat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\Sigma}}\,\widehat{\mathbf{A}}\,\widehat{\boldsymbol{C}}_h \operatorname{diag}(\boldsymbol{\alpha}) \right\|^2. \tag{20}$$

This cost function can be interpreted as an application of the function G, see (13), to the pair $(\widehat{\mathbf{A}}, \widehat{\boldsymbol{C}})$ where each coordinate $j \in \{1, \ldots, p\}$ is weighted by α_j. This weighting procedure, associated with the following optimization problem

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} H(\boldsymbol{\alpha}) + \lambda \|\boldsymbol{\alpha}\|_1, \tag{21}$$

where $\lambda > 0$ allows for variable selection in the SIR context thanks to the L_1 penalty. The parameter λ can be fixed by the user minimizing, for example, the Generalized Cross-Validation (GCV), Akaike's Information Criterion (AIC [1]), the Bayesian Information Criterion (BIC [69]) or the Residual Information Criterion (RIC [73]) defined as,

$$\begin{gathered} \text{GCV} = \text{RSS}/\left(n - p(\lambda)\right), \quad \text{AIC} = n \log\left(\text{RSS}/n\right) + 2p(\lambda), \quad \text{BIC} = n \log\left(\text{RSS}/n\right) + \log(n) p(\lambda), \\ \text{RIC} = (n - p(\lambda)) \log\left(\text{RSS}/(n - p(\lambda))\right) + (\log(n) - 1) p(\lambda) + 4/(n - p(\lambda) - 2), \end{gathered} \tag{22}$$

where $p(\lambda)$ is the effective number of parameters, defined in [75], and $\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Note that the computation of the prediction $\hat{y}_i$ of y_i requires the use of an adapted regression technique. Finally, $\text{Span}\left(\text{diag}(\widehat{\boldsymbol{\alpha}})\widehat{\mathbf{A}}\right)$ is a sparse estimator of the EDR subspace.

Sparse Ridge-SIR method [55] is very similar to the previous Shrinkage-SIR solution. The only difference lies on the preliminary estimation of $\left(\widehat{\mathbf{A}}, \widehat{\boldsymbol{C}}\right)$ which is achieved by considering the optimization problem (14).

In [86], the Dantzig selector, introduced in the context of the linear regression estimation [15] to deal with the $p \gg n$ case, is adapted to the SIR framework. In the original linear context, the Dantzig selector is defined through the following optimization problem

$$\arg\min_{\mathbf{b} \in \mathbb{R}^p} \|\mathbf{b}\|_1 \quad \text{s.t.} \quad \left\| \underline{\boldsymbol{X}}^\top (\underline{Y} - \underline{\boldsymbol{X}}\, \mathbf{b}) \right\|_\infty \leq (1 + 1/t)\, \sigma \sqrt{2 \log p}, \tag{23}$$

where $t > 0$ and under the additive Gaussian noise assumption $Y = \boldsymbol{\beta}^\top \mathbf{X} + E$, with $E \sim \mathcal{N}(0, \sigma^2)$. The role of the constraints $|(\underline{\boldsymbol{X}}^\top (\underline{Y} - \underline{\boldsymbol{X}}\mathbf{b}))_j| \leq (1 + 1/t)\sigma \sqrt{2 \log p}$ for all $j \in \{1, \ldots, p\}$, is to impose the residuals to be smaller than the noise level. Following this idea, the generalized eigenvalue problem (10) is rewritten as

$$\arg\min_{\mathbf{b} \in \mathbb{R}^p} \|\mathbf{W}_k\, \mathbf{b}\|_1 \quad \text{s.t.} \quad \left\| \mathbf{W}_k^{-1} (\widehat{\mathbf{M}}\, \widehat{\mathbf{b}}_k^{[0]} - \widehat{\lambda}_k^{[0]}\, \widehat{\boldsymbol{\Sigma}}\, \mathbf{b}) \right\|_\infty \leq \gamma_k, \tag{24}$$

for all $k \in \{1, \ldots, K\}$, where $(\widehat{\mathbf{b}}_k^{[0]}, \widehat{\lambda}_k^{[0]})$ is kth coordinate of the solution of (10) and where $\mathbf{W}_k$ is a diagonal matrix of weights "that should vary inversely with the magnitude of [the components of] $\widehat{\mathbf{b}}_k^{[0]}$" as defined by the authors. Let us stress that, similarly to the above shrinkage methods, this version of the Dantzig selector SIR requires a preliminary estimate of the SIR directions $\widehat{\mathbf{b}}_k^{[0]}$, $k \in \{1, \ldots, K\}$, which may difficult to obtain, see Section 4. Here, the Partial inverse regression solution of [53] is adopted and the regularization parameters γ_k, $k \in \{1, \ldots, K\}$ are fixed thanks to a cross-validation technique using one of the criteria defined in (22).

The next approach is denoted as refined sparse SIR. In [74], the sparse canonical correlation analysis [41] is adapted to the SIR framework, to define the so-called natural sparse SIR estimator as

$$\widehat{\mathbf{B}} = \arg\max_{\mathbf{B} \in \mathbb{R}^{p \times K}} \text{trace}\left(\mathbf{B}^\top \widehat{\mathbf{M}} \mathbf{B}\right) \quad \text{s.t.} \quad \mathbf{B}^\top \widehat{\boldsymbol{\Sigma}} \mathbf{B} = \mathbf{I}_K, \quad \left|\text{supp}\,(\mathbf{B})\right| \leq s, \tag{25}$$

where $\text{supp}\,(\mathbf{B})$ denotes the support of $\mathbf{B}$ and then $\left|\text{supp}\,(\mathbf{B})\right|$ represents the number of non null coefficients in $\mathbf{B}$. Note that this optimization problem can be interpreted as a constrained version of (16) where s is a sparse parameter to be chosen by the user. The latter optimization problem can be relaxed to build a convex optimization problem (for some $\rho_1 > 0$):

$$\widehat{\mathbf{B}}\widehat{\mathbf{B}}^\top = \arg\max_{\mathbf{F} \in \mathbb{R}^{p \times p}} \text{trace}\left(\widehat{\mathbf{M}}^\top \mathbf{F}\right) - \rho_1 \|\mathbf{F}\|_1 \quad \text{s.t.} \quad \left\| \widehat{\boldsymbol{\Sigma}}^{1/2} \mathbf{F} \widehat{\boldsymbol{\Sigma}}^{1/2} \right\|_\star \leq K \quad \text{and} \quad \left\| \widehat{\boldsymbol{\Sigma}}^{1/2} \mathbf{F} \widehat{\boldsymbol{\Sigma}}^{1/2} \right\|_{\text{op}} \leq 1, \tag{26}$$

where $\|\cdot\|_\star$ is the nuclear norm and $\|\cdot\|_{\text{op}}$ is the operator norm. $\widehat{\mathbf{B}}$ is not identifiable in this context but the estimation of $\widehat{\mathbf{B}}\widehat{\mathbf{B}}^\top$ gives clues on $\widehat{\mathbf{B}}$ itself. To estimate $\widehat{\mathbf{B}}$ if K is known, one can simply perform an eigen-decomposition of $\widehat{\mathbf{B}}\widehat{\mathbf{B}}^\top$ such as $\widehat{\mathbf{B}}\widehat{\mathbf{B}}^\top = \sum_{k=1}^K \phi_k \mathbf{v}_k \mathbf{v}_k^\top$ where $\phi_1, \ldots, \phi_K$ are the eigenvalues and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_K)$ is the matrix of associated eigenvectors. The so-called natural sparse SIR estimator is then defined as

$$\widehat{\mathbf{B}}^\star = \widehat{\mathbf{V}} \left(\widehat{\mathbf{V}}^\top \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{V}}\right)^{-1/2}. \tag{27}$$

Remarking that this estimator is not rate optimal, the authors propose a refined sparse SIR estimator, which is a three-step estimator. It is adapted from the methodology of [42] where the overall sample is divided into three independent subsamples S_1, S_2 and S_3 with nearly equal sizes. Let us denote by $\widehat{\mathbf{\Sigma}}^{[i]}$ and $\widehat{\mathbf{M}}^{[i]}$ the sample estimators of $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{M}}$ on the subsample S_i respectively, for $i \in \{1, 2, 3\}$. The method goes as follows.

- Step 1: The solution of (25) is computed using the sample estimators $\left(\widehat{\mathbf{\Sigma}}^{[1]}, \widehat{\mathbf{M}}^{[1]}\right)$ and is denoted by $\widehat{\mathbf{B}}^{[1]}$.
- Step 2: Based on S_2, let

$$\widehat{\boldsymbol{\eta}} = \arg\min_{\mathbf{B}\in\mathbb{R}^{p\times K}} \operatorname{trace}\left(\mathbf{B}^\top \widehat{\mathbf{\Sigma}}^{[2]}\mathbf{B}\right) - 2\operatorname{trace}\left(\mathbf{B}^\top \widehat{\mathbf{M}}^{[2]}\widehat{\mathbf{B}}^{[1]}\right) \text{ s.t. } \left|\operatorname{supp}(\mathbf{B})\right| \leq s,$$

  which is due to [19].
- Step 3: Finally, the refined sparse SIR estimator is the normalized version of $\widehat{\boldsymbol{\eta}}$ defined similarly to (27) as $\widetilde{\boldsymbol{B}} = \widehat{\boldsymbol{\eta}}\left(\widehat{\boldsymbol{\eta}}^\top \widehat{\mathbf{\Sigma}}^{[3]} \widehat{\boldsymbol{\eta}}\right)^{-1/2}$.

In practice, the implementation of this estimator requires exhaustive search and convex relaxation methods.

5.2. Computational variable selection methods

Basing on a measure of variable importance, [48] proposes an approach allowing to perform simultaneously SIR estimation and variable selection within the ultra-high dimensional context ($p \gg n$) which is denoted as variable importance assessment. Two ingredients are needed: first, a perturbation method of the initial sample, and second a proximity measure between the estimated EDR subspace on the perturbed sample and a gold standard estimator. In this paper, the selected measure of proximity is the trace correlation coefficient between projectors on the EDR subspaces (see for instance (8)). Besides, the perturbations are achieved by randomly permuting the n observed values of the jth coordinate $X^{(j)}$, $j \in \{1, \ldots, p\}$. The EDR subspace estimation is performed thanks to the SIR-QZ algorithm (see Section 5.1) both on the original data set (gold standard) and on the perturbed ones. The quality measure finally considered is the trace correlation coefficient averaged on a given number of permutations. The selected variables are the ones for which permutations induce the largest modifications of the EDR estimations, i.e. the smallest quality measures. To this end, the resulting quality measures are sorted in increasing order, and a threshold is chosen based whether on visual inspection or on changepoint detection.

A similar idea is implemented by [29] in the single-index situation and is denoted as closest submodel selection (CSS). The perturbation method consists in randomly selecting some covariates among the p initial ones such as the SIR problem is no more singular (i.e. selecting less variables than n to inverse $\widehat{\mathbf{\Sigma}}$). The proximity measure is the correlation between the observations projected on the index computed on the original sample on the one hand, and projected on the index computed on the perturbed sample on the other hand. The idea of the method is to keep a fixed number m of sub-models maximizing the proximity measure. Finally the variables $X^{(j)}$, the most frequently encountered in the m sub-models, are selected.

The sure independent screening (SIS) methodology has also been developed by [35] to deal with $p \gg n$ data sets in the linear framework $Y = \boldsymbol{\beta}^\top \mathbf{X} + E$. The approach is based on the ranking of the square covariance coefficients $(\operatorname{cov}(X^{(j)}, Y))^2$, $j \in \{1, \ldots, p\}$. Each of these coefficients reflects the linear association of the corresponding variable with the response. The idea is to screen the s variables of $\mathbf{X}$ with largest estimated coefficient (on the standardized variables). There is no general law to fix s, but it can be chosen "from p to a relatively large scale, say, below sample size n". In practice, the authors adopt $s = \lfloor n/\log(n) \rfloor$ or $s = n - 1$ in their simulation studies and further reduce the number of variables using a Lasso or Dantzig selector as post-processing steps.

SIRS [91] is an adaptation of the SIS methodology to the model-free setting using a marginal utility measure defined as $\omega_j = \mathbb{E}\left\{\Omega_j^2(Y)\right\}$ where $\Omega_j(y) = \operatorname{cov}\left(X^{(j)}, \mathbb{I}(Y < y)\right)$, $j \in \{1, \ldots, p\}$. The measure is estimated by

$$\hat{\omega}_j = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{X}_i^{(j)} \mathbb{I}_{Y_i<Y_k}\right)^2$$

on the standardized data. The marginal utility is used to rank the variables $X^{(j)}$, $j \in \{1, \ldots, p\}$. The consistency in probability of the ranking is established under classical conditions, such as the elliptical distribution condition

associated with SIR estimator (see Section 2.2). Such a consistency insures that an active variable (i.e., associated with Y) is ranked before an inactive one (i.e., not associated with Y) with high probability. Finally, a threshold has to be fixed in the rankings to classify variables between active/inactive classes. Two solutions are investigated.

- Generate $\underline{\boldsymbol{Z}}$ a matrix of n realizations of p auxiliary variables, inactive by construction. Perform the ranking process on the stacked matrix $[\underline{\boldsymbol{X}}, \underline{\boldsymbol{Z}}]$ and define the required threshold as $s = \arg\max_{\ell \in \{p+1,\ldots,2p\}} \hat{\omega}_\ell$. This solution was introduced in [61] and [81] in order to control the false discovery rate.
- The other solution [35] consists in selecting the $s = \lfloor n/\log(n) \rfloor$ firstly ranked variables as the active ones.

It appears that the first solution is the most efficient as the number of active predictors is large. Conversely, the second solution is the best as the number of active variables is small compared to p (sparse case). The final choice of the authors is to consider the reunion of both selected subsets.

6. Conclusion

This work presented a general overview of the SIR approach, and three topics of particular relevance in modern multivariate data contexts have been highlighted. Indeed, initially dedicated to the prediction of univariate components, SIR for multivariate $\mathbf{Y}$ is today accompanied by a large literature. Besides, data analyses on unidentifiable high-dimensional structures are recurrent nowadays and regularized versions of SIR allow to manage this aspect. The same remains true for variable selection, where many research directions have been proposed by the statistical community.

Many other extensions of SIR have been proposed, such as kernel sliced inverse regression allowing the estimation of a nonlinear subspace [80, 82, 93], Student sliced inverse regression dealing with heavy-tailed data [20], SIR versions dedicated to elliptically contoured distributions [13], as well as on-line versions [14, 17] to deal with data streams, among others. Besides, due to conditional expectation and covariance matrix estimations, SIR models are sensible to outliers. As a consequence, several solutions have been investigated to detect such observations or introduce robust SIR versions, see [59] and the references therein. In this sense, this review work does not aim at exhaustiveness.

It is also interesting to note that the SIR approach can be associated with many other dimension reduction methods, in particular through the envelopes' theory [26]. This theory indeed includes SIR as an example of sufficient dimension reduction method, PCA, principal component regression and PLS (Partial Least Squares) approaches as well as their derivatives.

Finally, let us emphasize that numerous alternatives to SIR can be found in the dimension reduction literature dedicated to regression. For instance, similarly, single-index models overcome the curse of dimensionality by modeling the non-linear relationship between Y and $\mathbf{X}$ through an unknown link function and a single linear combination of the covariates referred to as the index, see [45, Chapter 2]. Among the numerous works dedicated to the simultaneous estimation of the index and the link function, the most popular ones include the average derivative estimation method in the context of kernel smoothing [44], and the M-estimation technique based on spline regression [77].

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, Information Theory and an Extension of the Maximum Likelihood Principle, Springer New York, New York, NY, 1998, pp. 199–213.

[2] U. Amato, A. Antoniadis, I. De Feis, Dimension reduction in functional regression with applications, Computational Statistics & Data Analysis 50 (2006) 2422–2446.

[3] Y. Aragon, A Gauss implementation of multivariate sliced inverse regression, Computational Statistics 12 (1997) 355–372.

[4] Y. Aragon, J. Saracco, Sliced Inverse Regression (SIR): an appraisal of small sample alternatives to slicing, Computational Statistics 12 (1997) 109–130.

[5] R. Azaïs, A. Gégout-Petit, J. Saracco, Optimal quantization applied to sliced inverse regression, Journal of Statistical Planning and Inference 142 (2012) 481–492.

[6] Z. D. Bai, X. He, A chi-square test for dimensionality for non-Gaussian data, Journal of Multivariate Analysis 88 (2004) 109–117.

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

[8] C. Bernard-Michel, L. Gardes, S. Girard, A note on sliced inverse regression with regularizations, Biometrics 64 (2008) 982–984.

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

[10] W. Bian, D. Tao, Manifold regularization for SIR with rate root-n convergence, Advances in Neural Information Processing Systems 22 (2009) 117–125.
[11] M. J. Brusco, D. Steinley, J. Stevens, K-medoids inverse regression, Communications in Statistics - Theory and Methods 48 (2019) 4999–5011.
[12] E. Bura, R. Cook, Estimating the structural dimension of regressions via parametric inverse regression, Journal of the Royal Statistical Society, Series B 63 (2001) 393–410.
[13] E. Bura, L. Forzani, Sufficient reductions in regressions with elliptically contoured inverse predictors, Journal of the American Statistical Association 110 (2015) 420–434.
[14] Z. Cai, R. Li, L. Zhu, Online sufficient dimension reduction through sliced inverse regression., J. Mach. Learn. Res. 21 (2020) 1–25.
[15] E. Candes, T. Tao, The Dantzig selector: Statistical estimation when p is much larger than n, The Annals of Statistics 35 (2007) 2313–2351.
[16] R. Carroll, K.-C. Li, Measurement error regression with unknown link: dimension reduction and data visualization, Journal of the American Statistical Association 87 (1992) 1040–1050.
[17] 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 (2014) 1129–1152.
[18] M. Chavent, B. Liquet, J. Saracco, A semiparametric approach for a multivariate sample selection model, Statistica Sinica 20 (2010) 513–536.
[19] C.-H. Chen, K.-C. Li, Can SIR be as popular as multiple linear regression?, Statistica Sinica 8 (1998) 289–316.
[20] A. Chiancone, F. Forbes, S. Girard, Student sliced inverse regression, Computational Statistics and Data Analysis 113 (2017) 441–456.
[21] F. Chiaromonte, J. Martinelli, Dimension reduction strategies for analyzing global gene expression data with a response, Mathematical Biosciences 176 (2002) 123–144.
[22] R. Cook, Principal Hessian directions revisited (with discussion), Journal of the American Statistical Association 93 (1998) 84–100.
[23] R. Cook, SAVE: a method for dimension reduction and graphics in regression, Communications in statistics - Theory and Methods 29 (2000) 2109–2121.
[24] R. Cook, Testing predictor contributions in sufficient dimension reduction, The Annals of Statistics 32 (2004) 1062 – 1092.
[25] R. Cook, Fisher lecture: Dimension reduction in regression, Statistical Science 22 (2007) 1–26.
[26] R. Cook, An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics, Wiley Series in Probability and Statistics, Wiley, Hoboken, NJ, 2018.
[27] R. Cook, L. Ni, Using intra slice covariances for improved estimation of the central subspace in regression, Biometrika 93 (2006) 65–74.
[28] R. Coudret, S. Girard, J. Saracco, A new sliced inverse regression method for multivariate response, Computational Statistics & Data Analysis 77 (2014) 285–299.
[29] R. Coudret, B. Liquet, J. Saracco, Comparison of sliced inverse regression approaches for underdetermined cases, Journal de la Société Française de Statistique 155 (2014) 72–96.
[30] E. Devijver, E. Perthame, Prediction regions through inverse regression, Journal of Machine Learning Research 21 (2020) 1–24.
[31] N. Draper, R. Smith, Applied regression analysis (3rd edition), Wiley, New-York, 1998.
[32] N. Duan, K.-C. Li, Slicing regression: a link-free regression method, The Annals of Statistics 19 (1991) 505–530.
[33] A. Edelman, T. Arias, S. Smith, The geometry of algorithms with orthogonality constraints, SIAM journal on Matrix Analysis and Applications 20 (1998) 303–353.
[34] G.-L. Fan, H.-X. Xu, H.-Y. Liang, Dimension reduction estimation for central mean subspace with missing multivariate response, Journal of Multivariate Analysis 174 (2019) 104542.
[35] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, Journal of the Royal Statistical Society: Series B 70 (2008) 849–911.
[36] L. Ferré, Determining the dimension in sliced inverse regression and related methods, Journal of the American Statistical Association 93 (1998) 132–140.
[37] L. Ferré, A. Yao, Reply to the paper 'A note on smoothed functional inverse regression' by Liliana Forzani and R. Dennis Cook, Statistica Sinica 17 (2007) 1534–1544.
[38] J. Friedman, Regularized discriminant analysis, Journal of the American Statistical Association 84 (1989) 165–175.
[39] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, Springer series in statistics, New York, 2001.
[40] A. Gannoun, C. Guinot, J. Saracco, Reference curve estimation via alternating sliced inverse regression, Environmetrics 15 (2004) 81–99.
[41] C. Gao, Z. Ma, Z. Ren, H. H. Zhou, Minimax estimation in sparse canonical correlation analysis, The Annals of Statistics 43 (2015) 2168 – 2197.
[42] C. Gao, Z. Ma, H. H. Zhou, Sparse CCA: Adaptive estimation and computational barriers, The Annals of Statistics 45 (2017) 2074 – 2101.
[43] P. Hall, K.-C. Li, On almost linearity of low dimensional projections from high dimensional data, The Annals of Statistics 21 (1993) 867–889.
[44] W. Hardle, T. M. Stoker, Investigating smooth multiple regression by the method of average derivatives, Journal of the American Statistical Association 84 (1989) 986–995.
[45] J. L. Horowitz, Single-Index Models, Semiparametric and Nonparametric Methods in Econometrics, New-York, 2009.
[46] T. Hsing, Nearest neighbor inverse regression, The Annals of Statistics 27 (1999) 697 – 731.
[47] T. Hsing, R. J. Carroll, An asymptotic theory for sliced inverse regression, The Annals of Statistics 20 (1992) 1040–1061.
[48] I. Jlassi, J. Saracco, Variable importance assessment in sliced inverse regression for variable selection, Communications in Statistics - Simulation and Computation 48 (2019) 169–199.
[49] V. Kuentz, B. Liquet, J. Saracco, Bagging versions of sliced inverse regression, Communications in Statistics - Theory and Methods 39 (2010) 1985–1996.
[50] B. Li, S. Wen, L. Zhu, On a projective resampling method for dimension reduction with multivariate responses, Journal of the American Statistical Association 103 (2008) 1177–1186.
[51] K.-C. Li, Sliced inverse regression for dimension reduction, with discussion, Journal of the American Statistical Association 86 (1991) 316–342.
[52] K.-C. Li, Y. Aragon, K. Shedden, C. T. Agnan, Dimension reduction for multivariate response data, Journal of the American Statistical

Association 98 (2003) 99–109.
[53] L. Li, R. Cook, C.-L. Tsai, Partial inverse regression, Biometrika 94 (2007) 615–625.
[54] L. Li, H. Li, Dimension reduction methods for micro-arrays with application to censored survival data, Bioinformatics 20 (2004) 3406–3412.
[55] L. Li, X. Yin, Sliced inverse regression with regularizations, Biometrics 64 (2008) 124–131.
[56] Y. Li, L. Zhu, Asymptotics for sliced average variance estimation, The Annals of Statistics 35 (2007) 41–69.
[57] B. Liquet, J. Saracco, Application of the bootstrap approach to the choice of dimension and the α parameter in the SIR_α method, Communications in statistics - Simulation and Computation 37 (2008) 1198–1218.
[58] 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 (2012) 103–125.
[59] H. Lorenzo, J. Saracco, Computational outlier detection methods in sliced inverse regression, in: Advances in Contemporary Statistics and Econometrics, Springer, Cham, Switzerland, 2021, pp. 101–122.
[60] H.-H. Lue, Sliced inverse regression for multivariate response regression, Journal of Statistical Planning and Inference 139 (2009) 2656–2664.
[61] X. Luo, L. A. Stefanski, D. D. Boos, Tuning variable selection procedures by adding noise, Technometrics 48 (2006) 165–175.
[62] C. Moler, G. Stewart, An algorithm for generalized matrix eigenvalue problems, SIAM Journal on Numerical Analysis 10 (1973) 241–256.
[63] L. Ni, R. Cook, C.-L. Tsai, A note on shrinkage sliced inverse regression, Biometrika 92 (2005) 242–247.
[64] L. A. Prendergast, Implications of influence function analysis for sliced inverse regression and sliced average variance estimation, Biometrika 94 (2007) 585–601.
[65] J. Saracco, An asymptotic theory for sliced inverse regression, Communications in Statistics - Theory and Methods 26 (1997) 2141–2171.
[66] J. Saracco, Pooled slicing methods versus slicing methods, Communications in statistics - Simulation and Computation 30 (2001) 489–511.
[67] J. Saracco, Asymptotics for pooled marginal slicing estimator based on SIR_α approach, Journal of Multivariate Analysis 96 (2005) 117–135.
[68] J. R. Schott, Determining the dimensionality in sliced inverse regression, Journal of the American Statistical Association 89 (1994) 141–148.
[69] G. Schwarz, Estimating the dimension of a model, The Annals of Statistics 6 (1978) 461–464.
[70] L. Scrucca, Regularized sliced inverse regression with applications in classification, in: Data Analysis, Classification and the Forward Search, Springer-Verlag, Berlin, 2006, pp. 59–66.
[71] L. Scrucca, Class prediction and gene selection for DNA microarrays using regularized sliced inverse regression, Computational Statistics & Data Analysis 52 (2007) 438–451.
[72] C. M. Setodji, R. Cook, K-means inverse regression, Technometrics 46 (2004) 421–429.
[73] P. Shi, C.-L. Tsai, Regression model selection—a residual likelihood approach, Journal of the Royal Statistical Society: Series B 64 (2002) 237–252.
[74] K. Tan, L. Shi, Z. Yu, Sparse SIR: Optimal rates and adaptive estimation, The Annals of Statistics 48 (2020) 64 – 85.
[75] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B 58 (1996) 267–288.
[76] C. Vogel, Computational methods for inverse problems, Society for Industrial and Applied Mathematics, Philadelphia, 2002.
[77] L. Wang, L. Yang, Spline estimation of single-index models, Statistica Sinica 19 (2009) 765–783.
[78] X. Wen, R. Cook, New approaches to model-free dimension reduction for bivariate regression, Journal of Statistical Planning and Inference 139 (2009) 734–748.
[79] X. Wen, C. Setodji, A. Adekpedjou, A minimum discrepancy approach to multivariate dimension reduction via k-means inverse regression, Statistics and Its Interface 2 (2009) 503–511.
[80] H.-M. Wu, Kernel sliced inverse regression with applications to classification, Journal of Computational and Graphical Statistics 17 (2008) 590–610.
[81] Y. Wu, D. D. Boos, L. A. Stefanski, Controlling variable selection by the addition of pseudovariables, Journal of the American Statistical Association 102 (2007) 235–243.
[82] Y.-R. Yeh, S.-Y. Huang, Y.-J. Lee, Nonlinear dimension reduction with kernel sliced inverse regression, IEEE transactions on Knowledge and Data Engineering 21 (2008) 1590–1603.
[83] X. Yin, E. Bura, Moment-based dimension reduction for multivariate response regression, Journal of Statistical Planning and Inference 136 (2006) 3675–3688.
[84] X. Yin, L. Seymour, Asymptotic distributions for dimension reduction in the sir-ii method, Statistica Sinica 15 (2007) 1069–1079.
[85] J. K. Yoo, Iterative optimal sufficient dimension reduction for conditional mean in multivariate regression, Journal of Data Science (2009).
[86] Z. Yu, L. Zhu, H. Peng, L. Zhu, Dimension reduction and predictor selection in semiparametric models, Biometrika 100 (2013) 641–654.
[87] Y. Zhang, L. Zhu, Y. Ma, Efficient dimension reduction for multivariate response data, Journal of Multivariate Analysis 155 (2017) 187–199.
[88] W. Zhong, P. Zeng, P. Ma, J. Liu, Y. Zhu, RSIR: regularized sliced inverse regression for motif discovery, Bioinformatics 21 (2005) 4169–4175.
[89] L. Zhu, B. Miao, H. Peng, On sliced inverse regression with high-dimensional covariates, Journal of the American Statistical Association 101 (2006) 630–643.
[90] L. Zhu, L. Zhu, On kernel method for sliced average variance estimation, Journal of Multivariate Analysis 98 (2007) 970–991.
[91] L.-P. Zhu, L. Li, R. Li, L.-X. Zhu, Model-free feature screening for ultrahigh-dimensional data, Journal of the American Statistical Association 106 (2011) 1464–1475.
[92] L.-P. Zhu, L.-X. Zhu, S.-Q. Wen, On dimension reduction in regressions with multivariate responses, Statistica Sinica 20 (2010) 1291–1307.
[93] L. X. Zhu, K. T. Fang, Asymptotics for kernel estimate of sliced inverse regression, The Annals of Statistics 24 (1996) 1053–1068.
[94] L. X. Zhu, K. W. Ng, Asymptotics of sliced inverse regression, Statistica Sinica 5 (1995) 727–736.