Journal de la Société Française de Statistique
Vol. 155 No. 2 (2014)

Comparison of sliced inverse regression approaches for underdetermined cases

Titre: Comparaison d'approches de type SIR (régression inverse par tranches) pour les cas sous-déterminés ($n < p$)

Raphaël Coudret [1], Benoit Liquet [2] and Jérôme Saracco [1]

Abstract: Among methods to analyze high-dimensional data, the sliced inverse regression (SIR) is of particular interest for non-linear relations between the dependent variable and some indices of the covariate. When the dimension of the covariate is greater than the number of observations, classical versions of SIR cannot be applied. Various upgrades were then proposed to tackle this issue such as regularized SIR (RSIR) and sparse ridge SIR (SR-SIR), to estimate the parameters of the underlying model and to select variables of interest. In this paper, we introduce two new estimation methods respectively based on the QZ algorithm and on the Moore-Penrose pseudo-inverse. We also describe a new selection procedure of the most relevant components of the covariate that relies on a proximity criterion between submodels and the initial one. These approaches are compared with RSIR and SR-SIR in a simulation study. Finally we applied SIR-QZ and the associated selection procedure to a genetic dataset in order to find markers that are linked to the expression of a gene. These markers are called expression quantitative trait loci (eQTL).

Résumé : Parmi les méthodes pour analyser des données de grande dimension, la régression inverse par tranches (*sliced inverse regression* ou SIR en anglais) est particulièrement intéressante si des relations non-linéaires existent entre la variable à expliquer et des combinaisons linéaires des prédicteurs (appelées indices). Lorsque la dimension de ces prédicteurs est plus grande que le nombre d'observations, les versions classiques de SIR ne peuvent plus être utilisées. Des améliorations diverses comme RSIR et SR-SIR (pour *regularized SIR* et *sparse ridge SIR*) ont été proposées dans la litérature pour résoudre ce problème, estimer les paramètres du modèle sous-jacent et enfin réaliser une sélection des prédicteurs les plus pertinents (en un certain sens). Dans cet article, nous introduisons deux nouvelles procédures d'estimation basées respectivement sur l'algorithme QZ et sur l'inverse généralisé de Moore-Penrose. Nous décrivons également une méthode qui repose sur un critère de proximité entre des sous-modèles et le modèle initial pour sélectionner les prédicteurs les plus pertinents. Ces approches sont ensuite comparées avec RSIR et SR-SIR par le biais de simulations. Enfin, nous illustrons, sur un jeu de données génétiques, l'intérêt de l'approche SIR-QZ proposée et de l'algorithme de sélection de prédicteurs associé pour trouver des marqueurs liés á l'expression d'un gène. De tels marqueurs sont appelés *expression quantitative trait loci* ou eQTL.

Keywords: dimension reduction, high-dimensional data, semiparametric regression, sparsity

Mots-clés : grande dimension, regression semi-paramétrique, réduction de dimension, sparsité

AMS 2000 subject classifications: 62H12, 62F07

[1] INRIA Bordeaux Sud Ouest, CQFD team & Institut de Mathématiques de Bordeaux, UMR CNRS 5251
Université de Bordeaux 33405 Talence Cedex
E-mail: raphael.coudret@math.u-bordeaux1.fr and E-mail: jerome.saracco@math.u-bordeaux1.fr

[2] ISPED, INSERM U 897
Université de Bordeaux 33076 Bordeaux cedex
E-mail: benoit.liquet@isped.u-bordeaux2.fr

 ISSN: 2102-6238

1. Introduction

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

$$y = f(x'\beta_1, \ldots, x'\beta_K, \varepsilon) \tag{1}$$

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

When the dimension p of x is high and when we have little knowledge about the structure of the relationship between the response and the covariates, this semiparametric regression model is a nice alternative to parametric modeling (since it is really difficult to have knowledge about the structure of the relationship between the response and the covariates) and non-parametric modeling (which suffers from the well-known curse of dimensionality due to the data sparseness in the domain of x). The idea of dimension reduction in model (1) is intuitive because it aims at constructing a low dimensional projection of the covariate without losing information to predict the response y. If the dimension K of the EDR subspace is sufficiently small, it facilitates data visualization and explanation and it alleviates the curse of the dimensionality to non-parametrically estimate f with usual approaches such as kernel or splines smoothing (when the error term is additive).

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

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

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})' = \frac{1}{n}(\tilde{x}_1, \ldots, \tilde{x}_n)(\tilde{x}_1, \ldots, \tilde{x}_n)'. \tag{2}$$

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

 ISSN: 2102-6238

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

In the following, we only focus on the SIR approach. We describe it in Section 2.1 when $n > p$. The goal of this paper is then twofold. On one hand, we present methods to tackle the issue $n < p$. On the other hand, we also provide procedures in order to select which components of x have an effect on y.

In Section 2.2, we consider two different regularizations added to the SIR method, proposed by Zhong et al. (2005) and Li and Yin (2008), to find EDR estimates when $n < p$. Moreover, the SIR method can be seen as a generalized eigenvalue problem and linear algebra algorithms exist to solve this kind of problem without requiring any matrix inversion. The QZ algorithm (see Moler and Stewart (1973) for instance) is one of them and will be used in the SIR context in Section 2.3. In Section 2.4, we also adapt an approach introduced in functional sliced inverse regression (i.e., when x is an explanatory functional variable), based on the Moore-Penrose pseudo-inverse.

Concerning the selection of useful predictors in the indices, Zhong et al. (2005) use a chi-square test to find which components of x affect y, while the approach of Li and Yin (2008) relies on a Lasso penalization (Section 3.1). In Section 3.2, we propose another procedure. We choose randomly some submodels (i.e., using a number $p^0 < p$ of components of x) and we measure how close they are from the initial one with all the p components of x. The latter model is thus taken as a benchmark. Components of x that appear the most in submodels that are the closest to the benchmark are kept. We naturally consider that the other components of x do not affect y.

In Section 4, we compare in a simulation study the numerical behavior of the described methods to estimate EDR directions. We also evaluate the different procedures of selection of the useful components of x. In Section 5, we apply the most efficient one on real data from a genetic framework. Finally, some concluding remarks are given in Section 6.

2. SIR in determined and underdetermined cases

2.1. *Brief review of usual SIR*

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

Inverse regression step. The basic principle of the SIR method is to reverse the role of y and x, that is, instead of regressing the univariate variable y on the multivariate variable x, the covariable x is regressed on the response variable y. The price we have to pay to succeed in inverting the role of x and y is an additional assumption on the distribution of x, named the linearity condition (described hereafter).

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

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

$$(LC): \quad \forall b \in \mathbb{R}^p, \mathbb{E}(x'b|x'\beta_1, \ldots, x'\beta_K) \text{ is linear in } x'\beta_1, \ldots, x'\beta_K. \tag{3}$$

 ISSN: 2102-6238

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

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

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

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

$$\widehat{M}=\sum_{h=1}^{H}\hat{p}_h(\hat{m}_h-\hat{\mu})(\hat{m}_h-\hat{\mu})', \tag{4}$$

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

The choice of the slicing T is discussed in Li (1991); Kötter (2000); Saracco (2001) but, theoretically, there is no optimal one. In practice, we fix the number of observations per slice to $\lfloor n/H\rfloor$ where $\lfloor a\rfloor$ stands for the integer part of a. If the sample size n is not proportional to the number H of slices, some slices will then contain $\lfloor n/H\rfloor+1$ observations. Note that, in order to avoid the choice of a slicing, alternative SIR methods have been investigated. For instance, one can mention kernel-based methods of SIR proposed by Zhu and Fang (1996) or Aragon and Saracco (1997). However, these methods are hard to implement and are computationally slow. Moreover, Bura (1997) and Bura and Cook (2001) proposed a parametric version of SIR.

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

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

 ISSN: 2102-6238

2.2. *Two existing SIR methods when* $n < p$

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

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

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

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

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

where $\hat{\beta}(s) = (\hat{\beta}_1(s), \ldots, \hat{\beta}_K(s))$ is the matrix of the K first generalized eigenvector of $\widehat{M}$ and $\widetilde{\Sigma}(s)$, which is built such that the constraint $\hat{\beta}_k(s)'\widetilde{\Sigma}(s)\hat{\beta}_{\tilde{k}}(s) = \mathbb{I}[k = \tilde{k}]$ is verified for all $(k, \tilde{k}) \in \{1, \ldots, K\}^2$. More details about generalized eigenvectors can be found in Section 2.3. Because β_k is unknown, the authors replaced it with $\mathbb{E}(\hat{\beta}_k(s_0))$ in the expression of $L(s)$, to obtain an approximation $\widetilde{L}(s)$. Note that the parameter s_0 has to be sufficiently small in order for $\mathbb{E}(\hat{\beta}_k(s_0))$ to be close to β_k. In practice, s_0 is chosen equal to 0. Variances and expectations in $\widetilde{L}(s)$ are then estimated with bootstrap samples, which leads to an estimate $\widehat{\widetilde{L}}(s)$ of $\widetilde{L}(s)$. Remark that estimating $\mathbb{E}(\hat{\beta}_k(s_0))$ for $s_0 = 0$, implies using SIR with $\widehat{\Sigma}$. To do so, Zhong et al. (2005) apply the QZ algorithm (see Section 2.3 for details). The optimal regularization parameter is then given by

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

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

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

$$\left(\hat{\beta}, \widehat{V}\right) = \arg\min_{u,v} \sum_{h=1}^{H} \hat{p}_h \left\| (\hat{m}_h - \hat{\mu}) - \widehat{\Sigma} u v_h \right\|^2, \tag{5}$$

with $v = (v_1, \ldots, v_h)$ and where the minimum is taken over the respective sets of $p \times K$ matrices and $K \times H$ matrices. Note that this equation is also defined when $\widehat{\Sigma}$ is not invertible. From (5), the authors proposed thus a ridge version of the estimator $\hat{\beta}$ for a given regularization parameter s:

$$\left(\hat{\beta}(s), \widehat{V}(s)\right) = \arg\min_{u,v} G_s(u, v). \tag{6}$$

 ISSN: 2102-6238

where

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

In practice, $\hat{\beta}(s)$ can be then obtained from (6) with an alternating least-squares algorithm even when $n < p$. The SR-SIR method rely on a generalized crossvalidation criterion to find the optimal regularization parameter s_{opt} (see Li and Yin (2008) for details). Finally the matrix of estimated EDR directions is defined by $\hat{\beta}_{\text{SR-SIR}} = \hat{\beta}(s_{opt})$.

Remark 1. The exisence of a solution for (6) is not proved as explained by Bernard-Michel et al. (2008). Indeed, assume that $\left(\hat{\beta}(s), \widehat{V}(s)\right)$ is such a solution and that $\hat{\beta}(s)$ is not the null vector, we then have

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

which contradicts the fact that $\left(\hat{\beta}(s), \widehat{V}(s)\right)$ verifies (6). This encourages Bernard-Michel et al. (2008) to replace (6) with the following optimization problem:

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

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

2.3. *SIR-QZ: Solving the generalized eigenvalues problem in SIR*

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

$$\widehat{M} v = \lambda \widehat{\Sigma} v. \tag{7}$$

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

Notice that if we have two invertible matrices Q and Z, then finding λ and v in (7) is equivalent to find λ and w in

$$Q\widehat{M}Zw = \lambda Q\widehat{\Sigma}Zw, \tag{8}$$

and to set $v = Zw$. Similarly to the QR algorithm that is designed to find the Schur decomposition of a matrix in order to compute its eigenvalues, the QZ algorithm aims at finding unitary matrices

 ISSN: 2102-6238

Q and Z such that $Q\widehat{M}Z$ and $Q\widehat{\Sigma}Z$ are upper triangular, for square matrices $\widehat{M}$ and $\widehat{\Sigma}$. Such a transformation is called a generalized Schur decomposition. When working with complex matrices, Q and Z always exist (see Golub and Van Loan (1983), Theorem 7.7-1). Possible values of λ that verify (7) are such that $\det(\widehat{M}-\lambda\widehat{\Sigma})=0$, and such that $\det(Q(\widehat{M}-\lambda\widehat{\Sigma})Z)=0$. The latter determinant is the product of the diagonal elements of $Q(\widehat{M}-\lambda\widehat{\Sigma})Z$ since it is an upper triangular matrix. Hence, the generalized eigenvalues of (7) are the ratios of the diagonal elements of $Q\widehat{M}Z$ to the ones of $Q\widehat{\Sigma}Z$, provided that the diagonal elements of $Q\widehat{\Sigma}Z$ are not equal to zero. More specifically, this can be seen with the following formula (Golub and Van Loan (1983), Theorem 7.7-1)

$$\det(\widehat{M}-\lambda\widehat{\Sigma})=\det(Q'Z')\prod_{j=1}^{p}(t_j-\lambda u_j), \tag{9}$$

where $t_1,\dots,t_p$ and $u_1,\dots,u_p$ are the respective diagonal elements of $Q\widehat{M}Z$ and $Q\widehat{\Sigma}Z$. Notice that the generalized Schur decomposition only produces complex upper triangular matrices $Q\widehat{M}Z$ and $Q\widehat{\Sigma}Z$. However, there is a similar available decomposition for real matrices $\widehat{M}$ and $\widehat{\Sigma}$ (see Appendix B.1 or Golub and Van Loan (1983) for details).

Estimating the indices $X'\beta$ using the QZ algorithm. Equation (9) implies that if it exists $j\in\{1,\dots,p\}$ such that $t_j=u_j=0$, then $\det(\widehat{M}-\lambda\widehat{\Sigma})=0$ for all $\lambda\in\mathbb{C}$, and trying to choose eigenvectors corresponding to the greatest eigenvalues to estimate the EDR directions does not make sense. Numerically, for any $j\in\{1,\dots,p\}$, due to rounding errors, t_j and u_j are almost always different from 0, but if both their absolute value are too small, $\det(\widehat{M}-\lambda\widehat{\Sigma})$ is sufficiently unstable to call its value in question. As a consequence, every computed λ_j can be wrong. For similar reasons, if $|u_j|$ is too small for a given $j\in\{1,\dots,p\}$, t_j/u_j should not be considered as an eigenvalue. These remarks and the regularization procedure in Zhong et al. (2005) lead to the algorithm we provide in Appendix B.2 to find an estimate $\hat{\beta}_{\mathrm{QZ}}$ of the EDR directions.

Let $X=(x_1,\dots,x_n)$. When n is sufficiently smaller than p and $H>K$, a generalized eigenvector v of (7) is such that the n indices $X'v$ only takes H distinct values, as explained in Appendix B.3. In practice, the regularization parameter s of the algorithm of Appendix B.2 is small and we can distinguish easily H clusters in the values of $X'\hat{\beta}_{\mathrm{QZ}}$ in Figure 8 in Appendix B.3. This is a drawback of our approach when $n<p$ since the values of $X'\beta$ are a priori distinct. To circumvent this shortcoming, we compute several $\hat{\beta}_{\mathrm{QZ}}$ with different number of slices $H_1,\dots,H_{N_H}$. Let denote the corresponding estimates $\hat{\beta}_{\mathrm{QZ},1},\dots,\hat{\beta}_{\mathrm{QZ},N_H}$. We would like to find a K-dimensional subspace of $\mathbb{R}^n$ which is as close to the KN_H points of the matrix $X'(\hat{\beta}_{\mathrm{QZ},1},\dots,\hat{\beta}_{\mathrm{QZ},N_H})$ as possible. Thus, we would choose a basis $\hat{\gamma}$ of this subspace as an estimate of $X'\beta$. This leads us to consider the following equation:

$$\left(\hat{\gamma},\hat{\delta}\right)=\arg\min_{\gamma,\delta}\left\|X'(\hat{\beta}_{\mathrm{QZ},1},\dots,\hat{\beta}_{\mathrm{QZ},N_H})-\gamma\delta\right\|^2 \tag{10}$$

where the minimum is taken over the respective sets of $n\times K$ matrices and $K\times KN_H$ matrices and each column of $\gamma\delta$ is the approximation of the corresponding column of $X'(\hat{\beta}_{\mathrm{QZ},1},\dots,\hat{\beta}_{\mathrm{QZ},N_H})$ in the K-dimensional subspace spanned by γ. A solution of (10) is given by a principal component analysis (see Besse (2012), p80-81). Note that their exists an infinite number of bases $\hat{\gamma}$ which span the optimal K-dimensional subspace. Thus, the solution provided by the principal component analysis is just one of them. We call the whole approach SIR-QZ.

 ISSN: 2102-6238

Remark 2. When n is smaller enough than p, finding a satisfying estimate of β may not be possible. For example, if $K = 1$, $n < p$ and if the columns of X are not linearly dependent, there are infinitely many solutions u of the system $X'u = X'\hat{\beta}$ for a given estimate $\hat{\beta}$ of β. Recalling the underlying model (1), there is no reason why $\hat{\beta}$ should be a better estimate of β than any of these solutions. That is why, when $n < p$, we focus on estimates of $X'\beta$ rather than on β itself, in (10).

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

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

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

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

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

$$b_k = M^+\Sigma^{1/2}\eta_k \quad \text{for } k = 1, \ldots, K. \tag{11}$$

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

Adaptation for multivariate real covariates. In the context of our paper (that is, $n < p < \infty$), we will evaluate how the functional SIR procedure behaves in the multivariate framework. To do this, we simply substitute the operators of covariance M and Σ by the estimates $\widehat{M}$ and $\widehat{\Sigma}$ previously defined in (2) and (4). The resulting estimated directions are:

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

where the $\widehat{\eta}_k$'s are the eigenvectors of $\widehat{\Sigma}^{1/2}\widehat{M}^+\widehat{\Sigma}^{1/2}$ associated with the smallest eigenvalues (among those not structurally equal to zero, see Remark 3 for details). This adaptation is called SIR-MP.

 ISSN: 2102-6238

Remark 3. For a p-dimensional covariate, the $p \times p$ matrix $\widehat{\Sigma}^{1/2}\widehat{M}^{+}\widehat{\Sigma}^{1/2}$ is symmetric positive semidefinite and its rank r is at most equal to $H-1$ when $H < n < p$. Therefore, the eigenvalues of $\widehat{\Sigma}^{1/2}\widehat{M}^{+}\widehat{\Sigma}^{1/2}$ are such that $\hat{\alpha}_1 \geq \cdots \geq \hat{\alpha}_r > 0$, and the geometric multiplicity of the eigenvalue zero is equal to $p-r$ by construction. Thus we are interested in the eigenvectors $\widehat{\eta}_k$ associated with the K eigenvalues $\hat{\alpha}_r, \ldots, \hat{\alpha}_{r-K+1}$.

3. Selecting relevant components of x which are linked with y

Let $\beta_{j,k}$ denote the jth element of the EDR direction β_k, for $k = 1, \ldots, K$ and $j = 1, \ldots, p$. If $\beta_{j,.} = (\beta_{j,1}, \ldots, \beta_{j,K})$ is the null vector then the jth component of x does not have any effect on y. Finding such components is an important concern when $n < p$ because it allows $x \in \mathbb{R}^p$ to be reduced to $x^\star \in \mathbb{R}^{p^\star}$, where $p^\star < p$, without any loss of information. If in addition $p^\star$ is less enough than n, the EDR directions can then be accurately estimated with a classical SIR procedure applied on y and $x^\star$. In Section 3.1, we describe the methods from Zhong et al. (2005) and Li and Yin (2008) to determine which $\beta_{j,.}$ are null. In Section 3.2, we introduce another method to solve this problem based on proximity measures between models with only a few components of x and the initial model (in which every component of x is taken into account).

3.1. *Review of existing selection procedures*

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

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

SR-SIR: A Lasso method. From $\hat{\beta}(s_{opt})$ and $(\tilde{v}_1, \ldots, \tilde{v}_H) = \widehat{V}(s_{opt})$, given in equation (6), Li and Yin (2008) propose to minimize the following expression under a constraint on the L_1-norm of the vector ϕ:

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

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

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

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

 ISSN: 2102-6238

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

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

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

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

3.2. *CSS based on SIR: Closest submodel selection for SIR methods*

The idea of the procedure described here is to select submodels of (1) with only a given number p^0 of components of x which are the closest to the initial one. The components of x that appear the most in these submodels are asserted to have an effect on y.

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

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

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

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

Step 3. Compute the SIR-QZ indices $\hat{\gamma}^{(a)} \in \mathbb{R}^n$ based on Y and $X^{(a)}$.

Step 4. Calculate the linear correlation between the indices $\hat{\gamma}$ and $\hat{\gamma}^{(a)}$. Let us denote by $\hat{c}^{(a)}$ the square of this correlation.
Let $a = a + 1$.
Repeat N_0 times steps 2-4.

Step 5. Consider the submodels corresponding to the N_1 largest correlations $\hat{c}^{(a)}$.
Either the user set $\zeta \in]0,1[$ and then gets $N_1 = \zeta N_0$, or the user chose a value for ρ and then N_1 is the number of submodels such that $\hat{c}^{(a)} > \rho$.

Step 6. Count the number of occurrences of each component of x in these N_1 submodels. The components that affect y are the ones that have the greater number of occurrences.

For example, in our simulation study, we set $N_0 = 10^4$ and $\zeta = 10\%$ to determine the closest $N_1 = \zeta N_0$ submodels. In our real data application, we use $N_0 = 9 \times 10^5$ and $\rho = 0.9$ to select the top N_1 submodels.

Note that choosing $p^0 < n$ allows us to use classical SIR instead of SIR-QZ in Step 3 which significantly improves the computational time. In addition, any SIR approach that provides

 ISSN: 2102-6238

estimates of the indices (when $n \leq p$ and $n > p$) could be used in the whole algorithm instead of SIR-QZ.

4. A simulation study

In Sections 2-3, we presented 4 methods to estimate EDR directions (or indices) and 3 procedures to select which components of x have effects on y. In Section 4.1, we illustrate them on a single simulated data set. To compare their numerical performances, we then study them on several replications in Section 4.2.

4.1. Analysis of a single data set

4.1.1. Simulated model

We consider the following single index model

$$y = (x'\beta)^3 + \varepsilon \tag{12}$$

where x and β are p-dimensional vectors defined hereafter and $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ with $\sigma = 10^{-3}$. Let $p = 200$ and $p^\star = 20$. We choose $\beta = \frac{1}{10}(\mathbb{I}(1 \leq p^\star),\ldots,\mathbb{I}(p \leq p^\star))'$, so that $p^\star$ is the number of non-null components of β, that is the number of components of x that affect y. We construct $x = (x_1,\ldots,x_p)'$ as follows: for $j = 1,\ldots,p^\star$, generate σ_j^2 from the law $\mathscr{U}([0.05,0.1])$ and x_j from the law $\mathcal{N}(0,\sigma_j^2)$. For $j = p^\star+1,\ldots,p$, set

$$\sigma_j^2 = \left(\frac{12 - \lfloor (j-1)/p^\star \rfloor}{\lfloor (j-1)/p^\star \rfloor}\right)^2 \sigma^2_{(j-1) \bmod p^\star+1},$$

when mod denotes the modulo operation. Generate then $\check{x}_j$ from the law $\mathcal{N}(0,\sigma_j^2)$ and set $x_j = x_{(j-1) \bmod p^\star+1} + \check{x}_j$. This ensures that $\mathrm{cor}(x_j, x_{(j-1) \bmod p^\star+1}) = \lfloor (j-1)/p^\star \rfloor / 12$.

4.1.2. Estimation of EDR indices

We simulate an independent and identically distributed sample (X,Y) of size $n = 100$ from model (12). We plot Y versus the true indices $X'\beta$ in Figure 1. We analyze X and Y with the various methods presented in Section 2.2-2.4.

- For the RSIR method we evaluate $\widehat{\widetilde{L}}(s)$ for $s \in \{0, 10^{-10}, 10^{-9},\ldots,10^5\}$ with 50 bootstrap samples and $H = 10$. In Figure 2(a) we plot the values of Y against the indices provided by the RSIR method. The structure of Figure 1 can not be discerned in Figure 2(a). The regularization parameter that RSIR provides is equal to 10^5 and thus the RSIR procedure is equivalent to an eigendecomposition of $\widehat{M}$.
- Concerning the SR-SIR method, the regularization parameter s_{opt} is chosen among values in $\{10^{-10}, 10^{-9},\ldots,10^5\}$. The number of iterations of the alternating least square algorithm of SR-SIR is set to 50 and we take $H = 10$. For this example, we find $s_{opt} = 10^3$. In Figure 2(b), we draw the values of Y against the estimated indices $X'\hat{\beta}_{\text{SR-SIR}}$. The points of this graphic do not form the same shape as the points in Figure 1.

 ISSN: 2102-6238

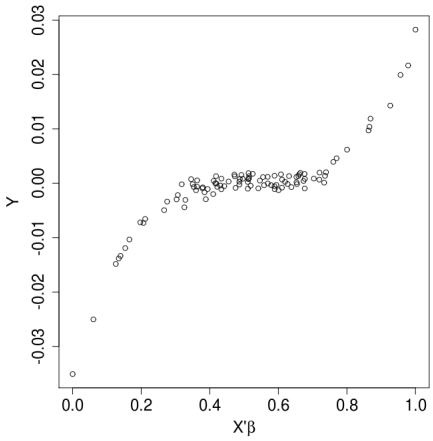

Figure 1. *Plot of Y versus $X'\beta$ generated from model (12), with $n = 100$ and $p = 200$. The horizontal scale was standardized.*

– We run SIR-QZ for $\{H_1, \ldots, H_{N_H}\} = \{5, \ldots, 15\}$. In Figure 2(c), we plot y against the corresponding estimated indices. This graphic exhibits a structure which is similar to the one in Figure 1.
– We finally apply SIR-MP with $H = 10$. We observe in Figure 2(d), which shows how Y and the indices produced by the SIR-MP method are related, that this method also fails to recover the shape of Figure 1.

Thus, Figure 2 shows that for this data set, SIR-QZ provides better estimations of the indices than RSIR, SR-SIR and SIR-MP.

To quantify such conclusions, we can use a criterion that measures how $X'\beta$ and $X'\hat{\beta}$ are close form each other, for a given estimate $\hat{\beta}$. Let $\widetilde{X} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ and let P be the projector on the subspace of $\mathbb{R}^n$ spanned by $\widetilde{X}'\beta$. More precisely, we have

$$P = \widetilde{X}'\beta(\beta'\widetilde{X}\widetilde{X}'\beta)^{-1}\beta'\widetilde{X}. \tag{13}$$

Similarly, define P_{RSIR}, $P_{\text{SR-SIR}}$, and $P_{\text{SIR-MP}}$ by respectively replacing β by $\hat{\beta}_{\text{RSIR}}$, $\hat{\beta}_{\text{SR-SIR}}$, and $\hat{\beta}_{\text{SIR-MP}}$ in (13). Let us also define $P_{\text{SIR-QZ}}$ by

$$P_{\text{SIR-QZ}} = \bar{I}_n\hat{\gamma}(\hat{\gamma}'\bar{I}_n\hat{\gamma})^{-1}\hat{\gamma}'\bar{I}_n,$$

where $\bar{I}_n$ is a matrix that centers $\hat{\gamma}$ (see Appendix B.3). Note that for any $a \neq 0$, we also have $P_{\text{SIR-QZ}} = \bar{I}_n(a\hat{\gamma})((\hat{\gamma}'a)\bar{I}_n(a\hat{\gamma}))^{-1}(\hat{\gamma}'a)\bar{I}_n$, which is coherent since if $(\hat{\gamma}, \hat{\delta})$ is a solution of (10), then $(a\hat{\gamma}, \frac{1}{a}\hat{\delta})$ is another solution of this very equation. For a given method m, we use the squared trace correlation between the subspaces spanned by $\widetilde{X}'\beta$ and by $\widetilde{X}'\hat{\beta}_m$ as a measure of the closeness between these subspaces. It is defined by

$$R(m) = \frac{1}{K}\text{Trace}(PP_m). \tag{14}$$

Notice that if $K = 1$, $R(m)$ is the squared cosine of the angle between the vectors $X'\beta$ and $X'\hat{\beta}_m$. This quality measure belongs to $[0, 1]$ and the higher its value is, the better the indices are estimated.

 ISSN: 2102-6238

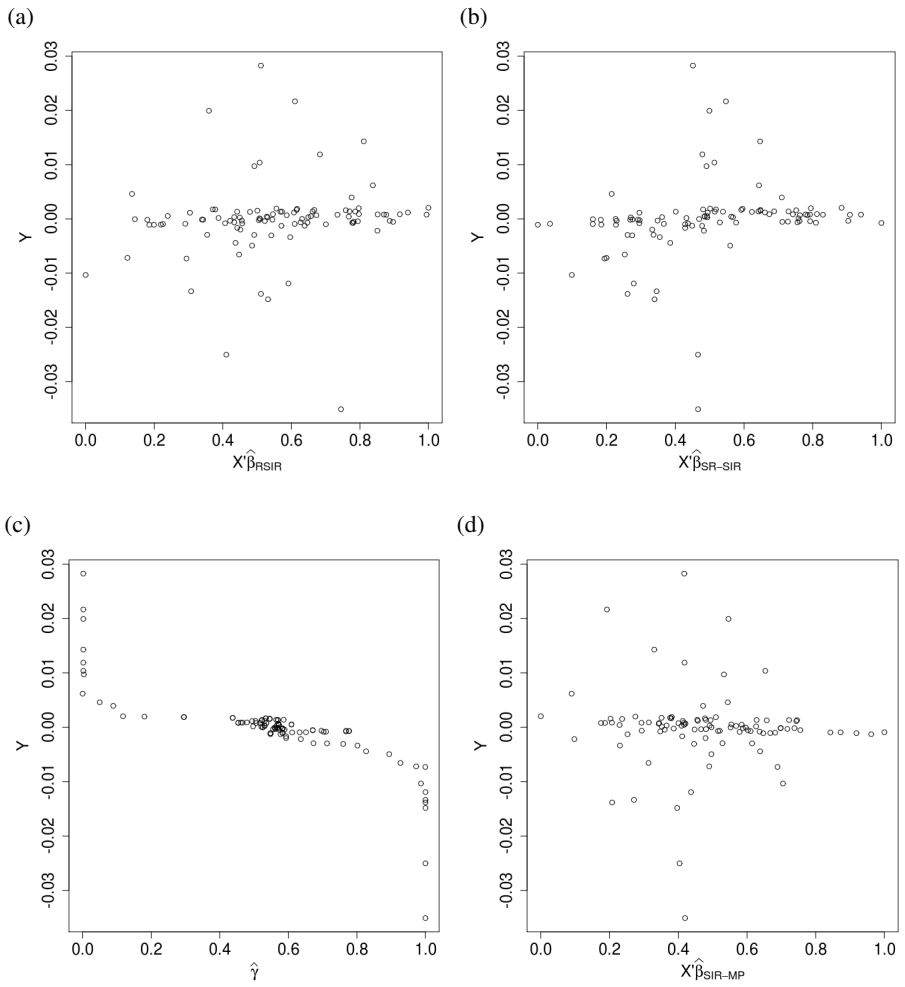

Figure 2. *Plot of Y versus estimated indices obtained with (a) the RSIR method, (b) the SR-SIR method, (c) the SIR-QZ method and (d) the SIR-MP method. The horizontal scale was standardized.*

In Table 1, we present values of $R(m)$ for the four considered methods. For SIR-QZ, $R(m)$ is clearly higher than for RSIR, SR-SIR and SIR-MP (0.74 versus less than 0.1). Notice also that the computational time is a lot greater for RSIR and SR-SIR than for SIR-MP and SIR-QZ.

4.1.3. Selection of components of x

We rely on the true positive rate (TPR) and on the false positive rate (FPR) to evaluate procedures that find which elements of β are equal to 0. The TPR is the number of selected components of x that actually affect y divided by the total number of components of x that affect y. The FPR is the number of selected components of x that do not affect y divided by the total number of components of x that do not affect y.

For the RSIR selection method the returned p-values are ordered and the components that

 ISSN: 2102-6238

Table 1. *Quality measure and computational time of various estimates of the indices $X'\beta$ for the simulated sample of size $n = 100$, with $p = 200$.*

m	RSIR	SR-SIR	SIR-QZ	SIR-MP
$R(m)$	0.051	0.088	0.741	0.000
Computational time (s)	101.88	1,234.70	7.85	0.25

Table 2. *True positive rate (TPR), false positive rate (FPR) and computational time of various methods run on the simulated sample of size $n = 100$, with $p = 200$, to determine which components of β of model (12) are not null.*

m		RSIR				CSS			
Estimates		$\hat{\beta}_{\text{RSIR}}$		$\hat{\beta}_{\text{QZ}}$		$\hat{\gamma}$		$x'\beta$	
Quality criteria		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Number of selected components	10	0.05	0.05	0.10	0.04	0.10	0.04	0.20	0.03
	20	0.15	0.09	0.25	0.08	0.35	0.07	0.40	0.07
	50	0.30	0.24	0.45	0.23	0.60	0.21	0.70	0.20
	100	0.40	0.51	0.70	0.48	0.75	0.47	0.90	0.46
	150	0.85	0.74	0.85	0.74	0.80	0.74	0.95	0.73
Computational time (s)		120.61		120.64		328.73		321.04	

m	SR-SIR					
Estimates	$\hat{\beta}_{\text{SR-SIR}}$		$\hat{\beta}_{\text{QZ}}$		β	
Quality criterion	TPR	FPR	TPR	FPR	TPR	FPR
Selection with AIC	0.00	0.13	0.00	0.16	0.90	0.00
Selection with BIC	0.00	0.03	0.00	0.07	0.65	0.00
Selection with RIC	0.00	0.03	0.00	0.04	0.65	0.00
Computational time (s)	35.53		40.88		33.29	

correspond to the first p-values are selected. For the CSS procedure, the components related to the greatest number of occurrence are selected. We evaluate the results of both methods by selecting the best 10 (resp. 20, 50, 100, 150) components. For the SR-SIR method, we use the different criteria (AIC, BIC and RIC) proposed in Li and Yin (2008) to determine the appropriate Lasso parameter.

The number of bootstrap samples generated for the RSIR method is set to 10^3. It is great enough so that increasing it does not improve significantly the quality criterion. This method relies on a regularization parameter which can also be provided by the Algorithm 1 to find $\hat{\beta}_{\text{QZ}}$ (see Appendix B.2). The estimate $\hat{\beta}_{\text{QZ}}$ can thus be plugged in the RSIR selection method. For the CSS method, we choose $N_0 = 10^4$, $\zeta = 10\%$ and $p^0 = 50$. While increasing N_0 could lead to better quality criteria, the computational time is sufficiently large not to choose it greater. The tested values of the Lasso parameter for the SR-SIR method are in $\{1, 2, \ldots, 100\}$. This algorithm needs an estimate of β in input. We use $\hat{\beta}_{\text{SR-SIR}}$ but, because of the poor results of Table 1 for this estimate, we also take $\hat{\beta}_{\text{QZ}}$ with $H = 10$, and the true EDR direction β.

Results of the corresponding TPR and FPR are displayed in Table 2. The SR-SIR method performs very well if an accurate estimate of β is provided, while the results are really bad otherwise because no selected component of x has any effect of y. Concerning RSIR and the CSS method, their FPR are similar, but the TPR are greater for the latter. The results for RSIR with $\hat{\beta}_{\text{QZ}}$ are slightly better than with the full RSIR procedure. The CSS method also seems to need good estimates of the indices since working with the true ones produces better TPR than using $\hat{\gamma}$.

To get more insights about the numerical performances of the various procedures tested in this

 ISSN: 2102-6238

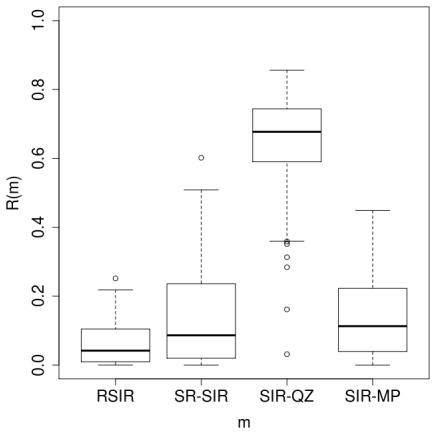

Figure 3. *Boxplots of* 100 *values of* $R(m)$ *for various estimation procedures* m *and samples of size* $n = 100$ *generated from model (12) with* $p = 200$.

example, we run them in several replications in the following section.

4.2. General behaviors of the estimates over several replications

We generate 100 samples of size $n = 100$ from the model (12). For each of them, we launch the RSIR, SR-SIR, SIR-QZ and SIR-MP procedures with the same parameters as in Section 4.1.2, and compute the quality criterion given in (14). We display the values of the criterion in Figure 3. The trend that is exhibited in Table 1 is confirmed in this graphic since the values of $R(\text{SIR-QZ})$ are clearly greater than the others.

Various methods to select important components of x are then run in the 100 samples drawn from model (12):

- RSIR with $\hat{\beta}_{\text{RSIR}}$, 1000 bootstrap samples and the following significance levels of the corresponding test: $0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$,
- SR-SIR with $\hat{\beta}_{\text{SR-SIR}}$ and $\tau \in \{1, 2, \ldots, 100\}$,
- The CSS procedure with $\hat{\gamma}$ from SIR-QZ and parameters $N_0 = 10^4$, $\zeta = 10\%$ and $p^0 = 50$ for various number of selected components: $10, 20, 30, 40, 60, 80, 100, 120, 140, 160$.

The mean ROC curves over the 100 replications are displayed in Figure 4. The CSS method outperforms RSIR while SR-SIR provides poor results. Notice that for RSIR, the values of the FPR are close to the chosen levels of test, in spite of the fact that this test is asymptotic.

5. Real data application

5.1. Description of the Dataset

We illustrate our developed approach on a genetic dataset which contains transcriptomic data and genomic data. In this study, we aim at finding genetic causes of variation in the expression of genes, that is eQTL (expression Quantitative Trait Loci). In this context, the gene expression data

 ISSN: 2102-6238

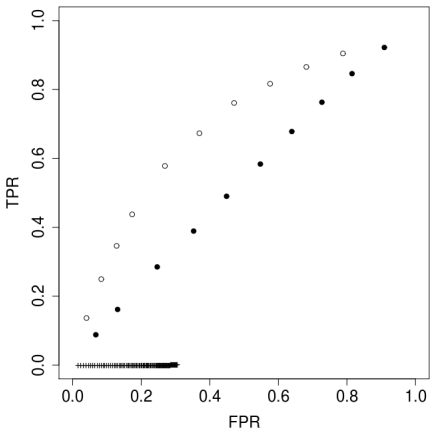

Figure 4. *Mean ROC curves for various procedures to select components of x that affect y in the model of Section 4.1, over* 100 *replications, with* $n = 100$ *and* $p = 200$*. Solid circles: RSIR, crosses: SR-SIR, empty circles: SIR-CSS.*

are treated as a quantitative phenotype and the genotype data (SNPs) are considered as predictors. In this illustration, we study the Hopx gene analyzed in Petretto et al. (2010). We investigate the ability of SIR-QZ combined with the CSS selection procedure to find a parsimonious index that explains the variability of Hopx gene expression in the heart tissue using $p = 770$ SNPs from $n = 29$ inbred line rats.

5.2. SIR-QZ and CSS results

We first run SIR-QZ for $\{H_1, \ldots, H_{N_H}\} = \{2, \ldots, 6\}$. In Figure 5(a), we plot the dependent variable Hopx versus the index based on the whole set of SNPs ($p = 770$). This graphic clearly exhibits a link between the phenotype and the index estimated by a smooth kernel method. In this illustration, this link is almost linear. In contrast, the plot (not given here) of Hopx gene versus the second EDR index does not show any structure. From this graphical diagnostic, it appears that only one EDR direction provides relevant information to explain the variability of the gene expression.

Then, in order to find a parsimonious index, we run our CSS selection procedure with $N_0 = 900.000$. The examined values of p^0 are in $\{10, 20\}$ while ρ takes value in $\{0.75, 0.80, 0.85, 0.90\}$. In Table 3, we present the number of selected SNPs for each combination of this two parameters. The threshold used for the selection will be detailed below. We can observe that, not surprisingly, the numbers of selected SNPs increases with p^0 and decreases with ρ. Moreover for a given value of p^0, we specify, in this table, the number of selected SNPs in common with those selected with $\rho = 0.9$ (corresponding to the parsimonious model). We also indicate, for a given ρ, the number of SNPs in common with those selected when $p^0 = 10$ and when $p^0 = 20$. This table highlights an overlap of 10 SNPs among all the sets of the selected SNPs for the various couple (p^0,ρ). Note that, the smallest set contains 11 SNPs when $p^0 = 10$ and $\rho = 0.9$ which comforts us about the stability of the CSS procedure.

In eQTL study, it is known that only a few number of SNPs can explain the variation of the gene expression. Thus, from the expertise of the biologists, we decide to select the sparsest model, that

 ISSN: 2102-6238

Table 3. *Results on selected SNPs for various values of p_0 and ρ*

	ρ	0.75	0.8	0.85	0.9
$p^0 = 10$	Number of selected SNPs	53	43	29	11
	Number of SNPs in common with those selected when $\rho = 0.9$	9	11	11	11
$p^0 = 20$	Number of selected SNPs	136	125	106	69
	Number of SNPs in common with those selected when $\rho = 0.9$	64	67	68	69
Number of SNPs in common with those selected when $p^0 = 10$ and when $p^0 = 20$		50	36	19	10

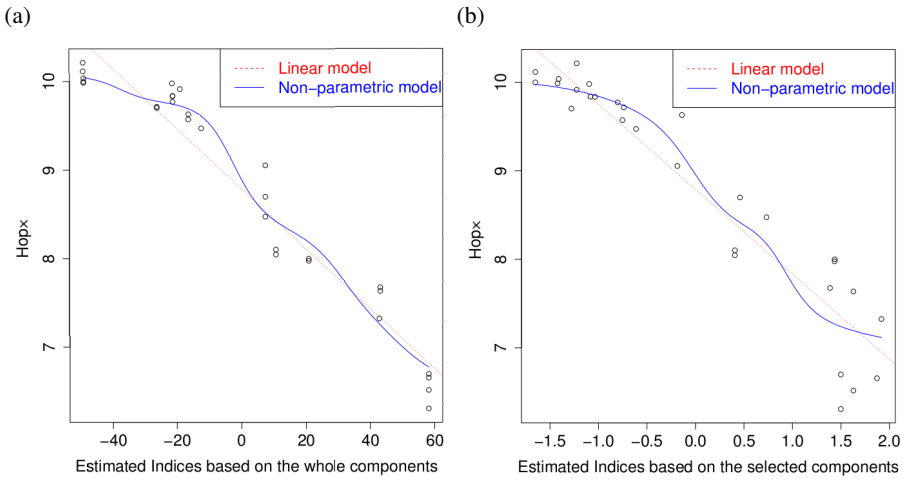

Figure 5. *Plots of the dependent variable* `Hopx` *versus the index based (a) on the whole SNPs, (b) on the 10 selected SNPs. The linear correlation between these two indices (evaluated on $n = 29$ rats) is 0.956. The dotted (red) line is the estimated linear model, the solid (blue) line is the kernel estimate of the link function with a bandwidth chosen by cross-validation.*

is with $(p^0, \rho) = (10, 0.9)$. Figure 6 exhibits the selected 11 SNPs for this choice of p^0 and ρ. The threshold (horizontal red line in the figure) is defined as follows: $N_1 \frac{p^0}{p} + u_{1-\frac{\alpha/2}{p}} \sqrt{N_1 \frac{p^0}{p}(1 - \frac{p^0}{p})}$ where $u_{1-\frac{\alpha/2}{p}}$ is the quantile of order $(1 - \frac{\alpha/2}{p})$ of the standard normal distribution. It corresponds to the upper bound of the prediction interval of the occurrence of a SNP in the selected model under the hypothesis that none of the SNPs are associated with the gene expression. The level of this interval is fixed at $1 - \alpha = 0.95$ and is corrected by a Bonferroni approach.

On Figure 5(b), we plot the dependent variable `Hopx` versus the index based on the 10 SNPs selected according to our previous comments on Table 3. The linear correlation between this index and the one estimated on all the SNPs which is equal to 0.956, highlights the good behaviour of our CCS strategy to select the relevant SNPs. Thus, not surprisingly, we observe the same relation between `Hopx` gene expression and the estimated indices.

 ISSN: 2102-6238

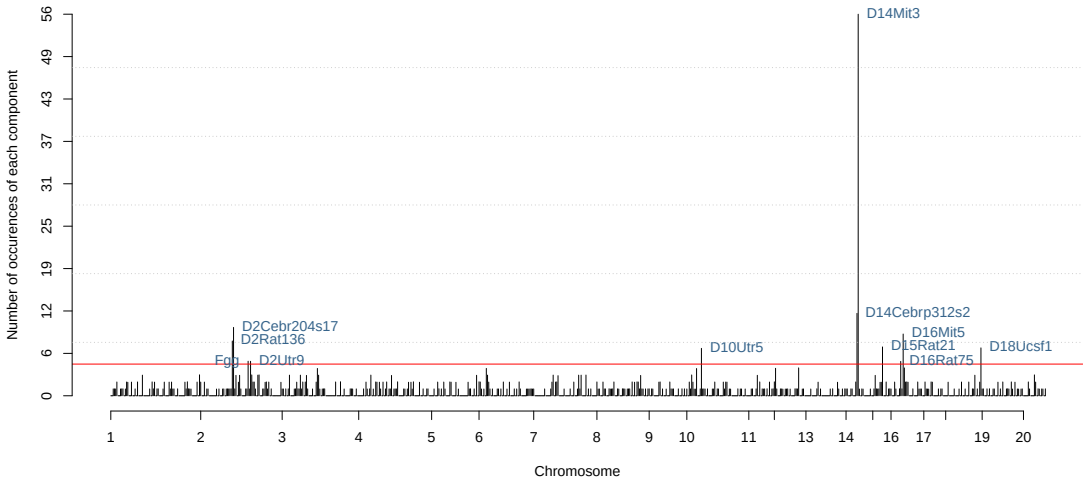

Figure 6. *Plot of the occurrence of the SNPs by the CSS procedure ($p^0 = 10$, $\rho = 0.9$). The horizontal solid red line represent the threshold to select the most relevant SNPs.*

5.3. Comparison methods

We compare our approach with three popular multivariate methods for analyzing high-dimensionnal datasets: A Lasso approach, the sparse Partial Least Squares (sPLS) and a Bayesian variable selection regression (ESS++). The Lasso method (Tibshirani, 1996) often performs poorly in prediction and interpretation especially when n is small and p is large. This technique tends to shrink the regression coefficients towards zero in order to select a sparse subset of covariates and provide a better prediction performance. sPLS (Lê Cao et al., 2009) seeks for the best linear combination of SNPs to predict the outcome. To ensure sparsity, sPLS includes a penalty function on some loading coefficients which is equivalent to a restriction on the number of loading vectors and on the number of SNPs, in each vector, that have a non-null coefficient. Both Lasso and sPLS approaches require a preliminary calibration of the tuning parameters which directly affects the number of selected variables, the estimate of the model parameters and therefore the statistical performances of the models. Calibration procedures usually involve the minimization of the mean square error of prediction through V-fold cross validation. In this illustration, we used the leave-one-out crossvalidation method to choose the tuning parameter for both methods. We finally compare our results with ESS++ a Bayesian variable selection approach for linear regression that can analyze single and multiple responses (Bottolo and Richardson, 2010; Bottolo et al., 2011). ESS++ exploits recent developments in MCMC search algorithms to explore the 2^p-dimensional model space. The performances of this method have been, among others, illustrated on eQTL studies (Petretto et al., 2010).

Figure 7 presents the Venn diagram of the sets of SNPs selected by the different approaches. Two SNPs (`D14Mit3` and `D2Cebr204s17`) are selected by the four methods. `D14Mit3` (chromosome 14) is clearly the first SNP in the list of the SNPs selected by the CSS procedure (see Figure 6) and

 ISSN: 2102-6238

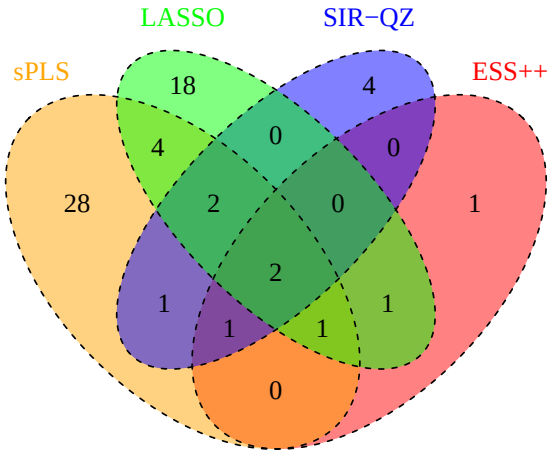

Figure 7. *Venn diagram of the sets of SNPs selected by Lasso, sPLS, ESS++ and SIR-QZ (combined with CSS) approaches*

`D2Cebr204s17` (chromosome 2) is at the third position. Moreover, our proposed approach reveals 4 SNPs (`D18Ucsf1`, `Fgg`, `D2Utr9`, `D16Rat75`) not selected by the other methods. Although three of them are very close to our proposed threshold (red line in Figure 6), while the SNP `D18Ucsf1` (chromosome 18) is clearly selected by our procedure.

The main advantage of our approach is the opportunity to reveal a non-linear link between the gene expression and a parsimonious index while the other compared approaches are based on a linear model. However, theses methods could treat multiple correlated phenotypes (multiple continuous responses). For example, ESS++ has been used to study the joint variability of gene expression in seven tissues (adrenal, gland, aorta, fat, heart, kidney, liver, skeletal) from inbred line rats (Bottolo et al., 2011). An extension of SIR-QZ for multivariate response is under investigation.

6. Concluding remarks

Although regularizing the estimated covariance matrix or constraining the optimization problem are natural ways to extend SIR to underdetermined cases ($n < p$), it may not be clear which one should be chosen and how to set the related parameter for each procedure. For the RSIR method, we illustrated that in such a context, the corresponding parameter should rather be determined with respect to the stability of the linear algebra algorithm (as in SIR-QZ) than with a statistical criterion. Moreover, the SIR-QZ approach introduced in this paper produces better results in simulation than the SR-SIR method that constraint the underlying optimization problem. In addition, the poor performances of SIR-MP suggest that adapting properties of the pseudo-inverse from the functional SIR to our high dimensional context is not well-adapted. We assumed that the dimension K of the EDR subspace was known in our simulation study. While, in the application, an empirical argument was given to determine this dimension, its estimation remains to be done with care.

We also proposed the CSS method that searches which submodels are the most informative

 ISSN: 2102-6238

to select relevant components of x. This procedure relies on a quality measure of the estimated indices, and it outperforms, in simulation, RSIR and SR-SIR selection procedures which are both based on estimates of β. We thus explain these results by pointing out that, when $n < p$, only the indices can be properly estimated, in general cases. Note that the space of the submodels may not be browsed optimally by our proposed CSS algorithm. Improvements should be made using optimization techniques such as genetic algorithms.

All the developed methods have been implemented in R language and are available upon request from the corresponding author. An R package is currently under development.

The illustration of genetic data highlights the opportunity of SIR-QZ combined by the CSS procedure to reveal a few number of SNPs which can explain the variability of the expression of the `Hopx` gene. Then, a non linear link between the gene expression and the parsimonious index could be estimated. The choice of the number of SNPs to keep is still a topic of concern since alternatives to our threshold could be considered.

In genetic datasets, the response variable is often multivariate. For instance, it could represent several phenotypes as in eQTL studies. Some approaches already handle such datasets. Since univariate results for SIR-QZ and the CSS selection procedure are consistent with the other methods presented in our application, it thus appears interesting to extend them to the multivariate case.

Acknowledgements

The authors wish to gratefully thank Marie Chavent for the first discussions that led to this work. They also acknowledge the editor and both anonymous referees for their helpful comments which lead to significant enhancements of this article.

Appendix A: Proof of (11)

We have, for $k = 1, \ldots, K$, $\Sigma^{1/2} M^{+} \Sigma^{1/2} \eta_k = \alpha_k \eta_k$. By pre-multiplying by the matrix $(M^{+}\Sigma^{1/2})$, we obtain:

$$M^{+}\Sigma^{1/2}\Sigma^{1/2}M^{+}\Sigma^{1/2}\eta_k = \alpha_k M^{+}\Sigma^{1/2}\eta_k, \quad \text{thus} \quad M^{+}\Sigma M^{+}\Sigma^{1/2}\eta_k = \alpha_k M^{+}\Sigma^{1/2}\eta_k.$$

Using the definition of b_k given in (11), we get: $M^{+}\Sigma b_k = \alpha_k b_k$. From the comments on the functional SIR context, provided in Section 2.4, the proof is complete.

Appendix B: Details about SIR-QZ

B.1. Generalized real Schur decomposition

We work here with real matrices $\widehat{M}$ and $\widehat{\Sigma}$. Similarly to the generalized Schur decomposition introduced in Section 2.3, that produces complex matrices Q and Z, the generalized real Schur decomposition (see Theorem 7.7-2 of Golub and Van Loan (1983)) ensures that we can find Q and Z such that $Q\widehat{M}Z$ is an upper quasi-triangular real matrix and $Q\widehat{\Sigma}Z$ is an upper triangular real one. An upper quasi-triangular matrix can be defined as the sum of an upper triangular matrix and

 ISSN: 2102-6238

of a block diagonal matrix where the sizes of the block are 1×1 or 2×2. For a 1×1 diagonal block of the matrix $Q\widehat{M}Z$, its unique element is called $\tilde{t}_j$ if it is located at the jth row and at the jth column of $Q\widehat{M}Z$. We write $\tilde{t}_{j_1,j_2}$ the element of a 2×2 diagonal block, in the j_1th row and in the j_2th column of $Q\widehat{M}Z$. An example of such an upper quasi-triangular matrix is given below:

$$Q\widehat{M}Z=\begin{pmatrix}\tilde{t}_1 & * & * & * & *\\ 0 & \tilde{t}_{2,2} & \tilde{t}_{2,3} & * & *\\ 0 & \tilde{t}_{3,2} & \tilde{t}_{3,3} & * & *\\ 0 & 0 & 0 & \tilde{t}_4 & *\\ 0 & 0 & 0 & 0 & \tilde{t}_5\end{pmatrix},$$

where $*$ denotes some real values. Let J be made of the elements $j\in\{1,\dots,p\}$ such that $\tilde{t}_j$ exists and J^c be the set made of $j\in\{1,\dots,p-1\}$ such that $\tilde{t}_{j,j}$ and $\tilde{t}_{j+1,j+1}$ exist. For each $j\in J$, let $\tilde{u}_j$ be the element of $Q\widehat{\Sigma}Z$ at the same location than $\tilde{t}_j$ in $Q\widehat{M}Z$ and define similarly $\tilde{u}_{j_1,j_2}$ for each $\tilde{t}_{j_1,j_2}$. Thus, we have

$$\det(\widehat{M}-\lambda\widehat{\Sigma})=\det(Q'Z')\prod_{j\in J}(\tilde{t}_j-\lambda\tilde{u}_j)\prod_{j\in J^c}\det\begin{pmatrix}\tilde{t}_{j,j}-\lambda\tilde{u}_{j,j} & \tilde{t}_{j,j+1}-\lambda\tilde{u}_{j,j+1}\\ \tilde{t}_{j+1,j} & \tilde{t}_{j+1,j+1}-\lambda\tilde{u}_{j+1,j+1}\end{pmatrix}.$$

Hence, for $j\in J$, if $\tilde{u}_j\neq 0$, then $\lambda_j=\tilde{t}_j/\tilde{u}_j$ is a real generalized eigenvalue. In addition, for $j\in J^c$, Moler and Stewart (1973) succeeded in finding $(\tilde{t}_j,\tilde{t}_{j+1})\in\mathbb{C}^2$, and $(\tilde{u}_j,\tilde{u}_{j+1})\in\mathbb{R}\backslash\{0\}$ such that $\lambda_j=\tilde{t}_j/\tilde{u}_j$ and $\lambda_{j+1}=\tilde{t}_{j+1}/\tilde{u}_{j+1}$ are generalized eigenvalues of $\widehat{M}$ and $\widehat{\Sigma}$. This leads to vectors $\tilde{t}=(\tilde{t}_1,\dots,\tilde{t}_p)'$ and $\tilde{u}=(\tilde{u}_1,\dots,\tilde{u}_p)'$ that are sent back by the QZ algorithm in order to provide generalized eigenvalues.

B.2. Algorithm

As explained in Section 2.3, the QZ algorithm has to be controlled when dealing with singular pencils. The following pseudocode in Scilab language allows the user to do so in the context of sliced inverse regression, for underdetermined cases. Because it is based on generalized real Schur decompositions, the notations involved are related to Appendix B.1 rather than Section 2.3.

```
// Initialize s_min, c and ε.
s = s_min;
keepGoing = %T;
while keepGoing
        //Use the QZ algorithm on M̂ and Σ̃(s) to find
        //vectors ũ and t̃.
        if (sum(abs(ũ) < ε & abs(t̃) < ε) == 0) &
           (length(ũ) - sum(abs(ũ) < ε) >= K) then
                keepGoing = %F;
        else
                s = s * c;
        end
end
// The estimated EDR directions are the eigenvectors sent by the
// last run of the QZ algorithm that corresponds to the K
// greatest values of t̃/ũ.
```

Algorithm 1. *A procedure to estimate EDR directions with the QZ algorithm*

 ISSN: 2102-6238

Typical values for $s_{\min}$, c and ε chosen in the simulation study of Section 4 are respectively 10^{-16}, 10 and 10^{-10}. The QZ algorithm is implemented in Matlab through the **eig** function. With Scilab, one should use **spec**, which is based on the LAPACK library. The R software is able to call functions from this library. Thus the QZ algorithm can be easily tested with this software. Notice that Algorithm 1 is designed to handle real values of $\tilde{u}_j$ and $\tilde{t}_j$ but, as mentioned in Appendix B.1, they can be complex. In that case, knowing if the blocks made of $\tilde{t}_{j,j}$, $\tilde{t}_{j+1,j}$, $\tilde{t}_{j,j+1}$ and $\tilde{t}_{j+1,j+1}$ and of $\tilde{u}_{j,j}$, $\tilde{u}_{j+1,j}$, $\tilde{u}_{j,j+1}$ and $\tilde{u}_{j+1,j+1}$ produce unstable eigenvalues is more difficult. As explained in Section 5 of Moler and Stewart (1973), the QZ algorithm aims at finding stable λ_j and λ_{j+1} corresponding to these 2×2 blocks. Because we do not control this procedure, we simply report if the QZ algorithm send back complex values in $\tilde{t}$. We never encounter this case in the simulation study of Section 4.

B.3. The sliced indices issue

Hereafter, we describe why the Algorithm 1 produces clustered indices as in Figure 8. Recall that $\widehat{\Sigma} = \frac{1}{n}\widetilde{X}\widetilde{X}'$. Define $\bar{I}_n = I_n - \frac{1}{n}\mathbf{1}_{n,n}$, where every element of the $n \times n$ matrix $\mathbf{1}_{n,n}$ is equal to 1. Notice that $\widetilde{X} = X\bar{I}_n$ and then $\widehat{\Sigma} = \frac{1}{n}X\bar{I}_nX'$. Let $\widehat{S}$ be a $n \times H$ matrix made of elements $\widehat{S}_{i,h}$ defined as

$$\widehat{S}_{i,h} = \frac{1}{n}\left(\frac{\mathbb{I}[y_i \in s_h]}{\hat{p}_h}\right).$$

We can also write $\widehat{M} = X\bar{I}_n\widehat{S}\widehat{W}\widehat{S}'\bar{I}_nX'$, where $\widehat{W}$ is defined in Section 2.2. Because of the structure of $\widehat{S}$, for any $H \times \alpha$ matrix A, $\widehat{S}A$ has at most H distinct rows, so has $\bar{I}_n\widehat{S}A$. Let w be the first generalized eigenvector of $\bar{I}_n\widehat{S}\widehat{W}\widehat{S}'\bar{I}_n$ and $\bar{I}_n$ associated with the eigenvalue λ. This means that $\bar{I}_n\widehat{S}\widehat{W}\widehat{S}'\bar{I}_nw = \lambda\bar{I}_nw$ which implies that $\bar{w} = \bar{I}_nw$ has at most H distinct values and then that w has also at most H distinct values.

Assume that X has full column rank, which is likely to happen when $p > n$. Then, $X^{+\prime}w$ is a generalized eigenvector of $\widehat{M}$ and $\widehat{\Sigma}$ and $X'X^{+\prime}w = w$ has at most H distinct values. The eigenvalue that is related to $X^{+\prime}w$ is equal to $n\lambda$.

If it exists $\hat{\beta}_1 \neq X^{+\prime}w$, a generalized eigenvector of $\widehat{M}$ and $\widehat{\Sigma}$ such that the generalized eigenvalue which is related to $\hat{\beta}_1$ is greater than the one corresponding to $X^{+\prime}w$, then we should have

$$\frac{\hat{\beta}_1'X\bar{I}_n\widehat{S}\widehat{W}\widehat{S}'\bar{I}_nX'\hat{\beta}_1}{\hat{\beta}_1'X\bar{I}_nX'\hat{\beta}_1} > \frac{w'\bar{I}_n\widehat{S}\widehat{W}\widehat{S}'\bar{I}_nw}{w'\bar{I}_nw}.$$

But, because w is the first generalized eigenvector of $\bar{I}_n\widehat{S}\widehat{W}\widehat{S}'\bar{I}_n$ and $\bar{I}_n$, it maximizes $\frac{u'\bar{I}_n\widehat{S}\widehat{W}\widehat{S}'\bar{I}_nu}{u'\bar{I}_nu}$ over the vectors u of length n, which contradicts the latter equation. Hence, such a $\hat{\beta}_1$ does not exist, and the first generalized eigenvector of $\widehat{M}$ and $\widehat{\Sigma}$ is $X^{+\prime}w$.

In this paragraph, we show that it exists $H-1$ orthogonal vectors $w_1, \ldots, w_{H-1}$ such that, for $k = 1, \ldots, H-1$,

$$\frac{w_k'\bar{I}_n\widehat{S}\widehat{W}\widehat{S}'\bar{I}_nw_k}{w_k'\bar{I}_nw_k} = \frac{w'\bar{I}_n\widehat{S}\widehat{W}\widehat{S}'\bar{I}_nw}{w'\bar{I}_nw},$$

which means that the first K generalized eigenvectors of $\widehat{M}$ and $\widehat{\Sigma}$ are the vectors $X^{+\prime}w_k$ for $k = 1, \ldots, K \leq H-1$. We assume $n > H$. Let sort Y increasingly and reorder the columns of

 ISSN: 2102-6238

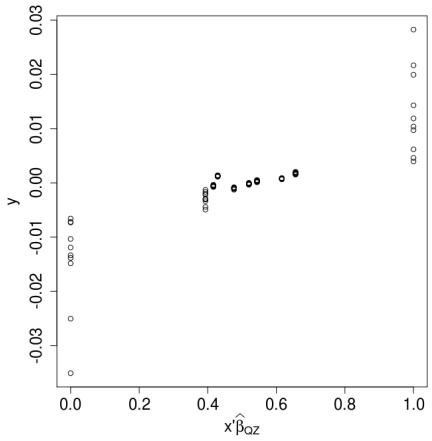

Figure 8. *Plot of Y versus $X'\hat{\beta}_{\mathrm{QZ}}$ with $H = 10$. The horizontal scale was standardized.*

X such that each column corresponds to the appropriate element of Y. This transformation implies that $\widehat{S}\widehat{W}\widehat{S}'$ is block diagonal with H blocks. For $h = 1,\ldots,H$, the size of the block h is equal to $n\hat{p}_h$ and each element it contains is equal to $\frac{1}{n^2\hat{p}_h}$. Hence, the rank of each block is equal to 1 and each block provides a positive eigenvalue for $\widehat{S}\widehat{W}\widehat{S}'$, which is equal to $\frac{n\hat{p}_h}{n^2\hat{p}_h} = \frac{1}{n}$. The corresponding eigenvector is made of elements $\frac{\mathbb{I}[y_i \in s_h]}{\sqrt{n\hat{p}_h}}$ for $i = 1,\ldots,n$. We have now H orthonormal eigenvectors of $\widehat{S}\widehat{W}\widehat{S}'$ for the eigenvalue $\frac{1}{n}$, and then we can find $H-1$ orthonormal centered eigenvectors $\bar{w}_1,\ldots,\bar{w}_{H-1}$ of $\widehat{S}\widehat{W}\widehat{S}'$ for this eigenvalue. In addition, for $k = 1,\ldots,H-1$, $\bar{w}_k$ is also a generalized eigenvector of $\bar{I}_n\widehat{S}\widehat{W}\widehat{S}'\bar{I}_n$ and $\bar{I}_n$ because $\bar{I}_n\bar{w}_k = \bar{w}_k$, and $\bar{w}_k$ maximizes $\frac{u'\bar{I}_n\widehat{S}\widehat{W}\widehat{S}'\bar{I}_n u}{u'\bar{I}_n u}$, over the vectors u of length n. Finally, we have that the first K generalized eigenvectors of $\widehat{M}$ and $\widehat{\Sigma}$ are $X^{+\prime}(\bar{w}_1,\ldots,\bar{w}_K)$, which means that the indices $X'X^{+\prime}(\bar{w}_1,\ldots,\bar{w}_K)$ only have H distinct rows.

This feature is illustrated on the simulated sample of size $n = 100$, with $p = 200$ from Section 4.1. In Figure 8, we plot Y versus the estimated indices obtained with $\hat{\beta}_{\mathrm{QZ}}$ for $H = 10$.

References

Amato, U., Antoniadis, A., and de Feis, I. (2006). Dimension reduction in functional regression with applications. *Comput. Stat. Data Anal.*, 50 (9):2422–2446.

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

Azais, R., Gegout-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.

Bercu, B., Nguyen, T., and Saracco, J. (2011). A new approach of recursive and non recursive sir methods. *Journal of the Korean Statistical Society*, 41:17–36.

Bernard-Michel, C., Gardes, L., and Girard, S. (2008). A note on sliced inverse regression with regularizations. *Biometrics*, 64:982–986.

 ISSN: 2102-6238

Besse, P. (2012). Exploration statistique multidimensionnelle. http://www.math.univ-toulouse.fr/~besse/pub/Explo_stat.pdf.

Bottolo, L., Chadeau-Hyam, M., Hastie, D. I., Langley, S. R., Petretto, E., Tiret, L., Tregouet, D., and Richardson, S. (2011). ESS^{++}: a C^{++} objected-oriented algorithm for bayesian stochastic search model exploration. *Bioinformatics*, 27(4):587–588.

Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for bayesian model exploration. *Bayesian Analysis*, 5(3):583–618.

Bura, E. (1997). *Dodge, Y. (ed), L_1-Statistical Procedures and Related Topics*, chapter Dimension reduction via parametric inverse regression, pages 215–228. Institute of Mathematical Statistics, Hayward.

Bura, E. and Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 63:393–410.

Carroll, R. J. and Li, K.-C. (1992). Measurement error regression with unknown link: dimension reduction and data visualization. *Journal of the American Statistical Association*, 87(420):1040–1050.

Chen, C.-H. and Li, K.-C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica*, 8(2):289–316.

Cook, R. D. (1998). Principal hessian directions revisited (with discussion). *Journal of the American Statistical Association*, 93:84–100.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Annals of Statistics*, 32(3):1062–1092.

Duan, N. and Li, K.-C. (1991). Slicing regression: a link-free regression method. *Annals of Statistics*, 19:505–530.

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

Ferré, L. and Yao, A.-F. (2007). Reply to the paper : "a note on smoothed functional inverse regression". *Statistica Sinica*, 17:1683–1687.

Golub, G. H. and Van Loan, C. F. (1983). *Matrix computations*, volume 3 of *Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins University Press, Baltimore, MD.

Kötter, T. T. (2000). *Schimek, M. G. (ed), Smoothing and regression. Approaches, computation and application.*, chapter Sliced inverse regression, pages 497–512. Wiley, Chichester.

Lê Cao, K.-A., Martin, P. G. P., Robert-Granié, C., and Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10(1):34.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction, with discussion. *Journal of the American Statistical Association*, 86:316–342.

Li, L. and Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics*, 64(1):124–131.

Li, Y. and Zhu, L. (2007). Asymptotics for sliced average variance estimation. *Annals of Statistics*, 35:41–69.

Liquet, B. and Saracco, J. (2008). Application of the bootstrap approach to the choice of dimension and the α parameter in the SIR$_\alpha$ method. *Communications in statistics - Simulation and Computation*, 37(6):1198–1218.

Moler, C. B. and Stewart, G. W. (1973). An algorithm for generalized matrix eigenvalue problems. *SIAM Journal on Numerical Analysis*, 10(2):241–256.

Petretto, E., Bottolo, L., Langley, S. R., Heinig, M., McDermott-Roe, C., Sarwar, R., Pravenec, M., Hübner, N., Aitman, T. J., Cook, S. A., and Richardson, S. (2010). New insights into the genetic control of gene expression using a bayesian multi-tissue approach. *PLoS Comput Biol*, 6(4):e1000737.

Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Communications in statistics - Theory and methods*, 26:2141–2171.

Saracco, J. (2001). Pooled slicing methods versus slicing methods. *Communications in statistics - Simulation and Computation*, 30:489–511.

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, 89:141–148.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Yin, X. and Seymour, L. (2007). Asymptotic distributions for dimension reduction in the sir-ii method. *Statistica Sinica*, 15:1069–1079.

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

Zhu, L. and Zhu, L. (2007). On kernel method for sliced average variance estimation. *Journal of Multivariate Analysis*, 98:970–991.

 ISSN: 2102-6238

Zhu, L. X. and Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Annals of Statistics*, 24:1053–1068.

Zhu, L. X., Ohtaki, M., and Li, Y. (2007). On hybrid methods of inverse regression-based algorithms. *Computational Statistics*, 51:2621–2635.

 ISSN: 2102-6238