Gaussian Regularized Sliced Inverse Regression

Caroline Bernard-Michel, Laurent Gardes, Stéphane Girard

Gaussian Regularized Sliced Inverse Regression

Caroline Bernard-Michel, Laurent Gardes, Stéphane Girard*

Laboratoire Jean-Kuntzmann & INRIA Rhône-Alpes, team Mistis,
Inovallée, 655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France,
(* corresponding author, Stephane.Girard@inrialpes.fr)

Abstract

Sliced Inverse Regression (SIR) is an effective method for dimension reduction in high-dimensional regression problems. The original method, however, requires the inversion of the predictors covariance matrix. In case of collinearity between these predictors or small sample sizes compared to the dimension, the inversion is not possible and a regularization technique has to be used. Our approach is based on a Fisher Lecture given by R.D. Cook where it is shown that SIR axes can be interpreted as solutions of an inverse regression problem. In this paper, a Gaussian prior distribution is introduced on the unknown parameters of the inverse regression problem in order to regularize their estimation. We show that some existing SIR regularizations can enter our framework, which permits a global understanding of these methods. Three new priors are proposed leading to new regularizations of the SIR method. A comparison on simulated data is provided.

Keywords: Inverse regression, regularization, sufficient dimension reduction.

1 Introduction

Many methods have been developed for inferring the conditional distribution of an univariate response Y given a predictor X in $\mathbb{R}^p$, ranging from linear regression [11] to support vector regression [10]. When p is large, sufficient dimension reduction aims at replacing the predictor X by its projection onto a subspace of smaller dimension without loss of information on the conditional distribution of Y given X. In this context, the central subspace, denoted by $S_{Y|X}$ plays an important role. It is defined as the smallest subspace such that, conditionally on the projection of X on $S_{Y|X}$, Y and X are independent. In other words, the projection of X on $S_{Y|X}$ contains all the information on Y that is available in the predictor X.

The estimation of the central subspace has received considerable attention since the past twenty years. Without intending to be exhaustive, we refer to Sliced Inverse Regression (SIR) [17], sliced average variance estimation [5], and graphical regression [6] methods. Among them, SIR seems to be the most popular one. The original method

has been adapted to various frameworks and the relative asymptotic properties have been derived, see for instance [16, 26, 20, 15, 21]. We also refer to [3] for a study of the SIR finite sample properties, to [12, 22] for the estimation of $K = \dim(S_{Y|X})$, the dimension of the central subspace, and to [1, 13] for extension to functional covariates.

Assuming that K is known and introducing $\Sigma = \mathrm{Cov}(X)$, in the SIR methodology, a basis of the central subspace is obtained by computing the eigenvectors associated to the largest K eigenvalues of $\Sigma^{-1}\mathrm{Cov}(\mathbb{E}(X|Y))$. Unfortunately, the classical $n-$ sample estimate $\hat{\Sigma}$ of Σ can be singular, or at least ill-conditioned, in several situations. Indeed, since $\mathrm{rank}(\hat{\Sigma}) \leq \min(n-1, p)$, if $n \leq p$ then $\hat{\Sigma}$ is singular. Even when n and p are of the same order, $\hat{\Sigma}$ is ill-conditioned, and its inversion introduces numerical instabilities in the estimation of the central subspace. Similar phenomena occur when the coordinates of X are highly correlated.

Some regularizations of the SIR method have been proposed to overcome this limitation. In [4] and [18], a Principal Component Analysis (PCA) is used as a preprocessing step in order to eliminate the directions in which the random vector X is degenerated. Thus, for a properly chosen dimension d of the projection subspace, the covariance matrix of the projected observations is regular. In the sequel, this technique will be referred to as PCA+SIR. Another method consists in adopting a ridge regression technique (see for instance [11], Chapter 17) *i.e.* replaces the sample estimate $\hat{\Sigma}$ by a perturbed version $\hat{\Sigma}+\tau I_p$ where I_p is the $p \times p$ identity matrix and τ is a positive real number [25]. Here, the idea is that, for τ large enough, $\hat{\Sigma} + \tau I_p$ is regular and its condition number increases with τ. Similarly, in [23], regularized discriminant analysis [14] is adapted to the SIR framework. More recently, it is proposed in [19] to interpret SIR as an optimization problem and to introduce L_1- and L_2- penalty terms in the optimized criterion.

Our approach is based on a Fisher Lecture given by R.D. Cook [8] where it is shown that the axes spanning the central subspace can be interpreted as the solutions of an inverse regression problem. In this paper, a Gaussian prior is introduced on the unknown parameters of the inverse regression problem in order to regularize their estimation. We show that the previously mentioned techniques [4, 18, 25] can enter our framework, which permits a global understanding of these methods. Three new priors are proposed leading to new regularizations of the SIR method. A comparison with the L_2- regularization introduced in [19] is also provided. It is shown that, from the theoretical point of view, the proposed L_2- regularization cannot be justified.

This paper is organized as follows. In Section 2, an adaptation of the inverse regression model to our framework is presented. Section 3 is dedicated to the regularization aspects. Theoretical comparisons with existing approaches as well as new methods are provided. Finite sample properties are illustrated in Section 4. Proofs are postponed to the Appendix.

2 Inverse regression without regularization

Consider X a $\mathbb{R}^p-$ random vector, Y the real response variable and let us denote by $S_{Y|X}$ the central subspace. In the following, for the sake of simplicity, we assume that $K = \dim(S_{Y|X}) = 1$. We thus introduce $\beta \in \mathbb{R}^p$ such that $S_{Y|X} = \text{span}(\beta)$. In Subsection 2.1, the considered inverse regression model [8] is presented. The estimation of the unknown parameters is discussed in Subsection 2.2 and the links with the SIR method are established in Subsection 2.3.

2.1 Inverse regression model

The following inverse regression model is considered:

$$X = \mu + c(Y)Vb + \varepsilon, \tag{1}$$

where μ and b are non-random $\mathbb{R}^p-$ vectors, ε is a centered $\mathbb{R}^p-$ Gaussian random vector, independent of Y, with covariance matrix $\text{Cov}(\varepsilon) = V$ and $c : \mathbb{R} \to \mathbb{R}$ is a nonrandom function. Under this model,

$$\mathbb{E}(X|Y = y) = \mu + c(y)Vb,$$

and thus, after translation by μ, the conditional expectation of X given Y is a degenerate random vector located on the axis Vb. In the sequel, it will appear that, under appropriate conditions, the maximum likelihood estimator of b is (up to a scale parameter) the SIR estimator of β. Moreover, note that, under (1), one has

$$c(y) = \frac{\mathbb{E}(b^t(X - \mu)|Y = y)}{b^tVb}. \tag{2}$$

Now, neglecting the noise term and restricting ourselves to the single-index case, the forward model of SIR asserts that there exists a univariate link function g such that $Y = g(b^tX)$ or equivalently, $b^tX = g^{-1}(Y)$. Thus, replacing in (2) yields

$$c(y) = \frac{g^{-1}(y) - b^t\mu}{b^tVb},$$

i.e. the coordinate function is, up to an affine function, the inverse of the link function in the single-index forward model of SIR.

2.2 Maximum likelihood estimation

We now address the estimation of the coordinate function $c(.)$, the axis b, the covariance matrix V and the location parameter μ in model (1). To this end, we focus on a projection estimator of the unknown function $c(.)$. More precisely, it is expanded as a linear combination of h basis functions $s_j(.)$, $j = 1, \ldots, h$:

$$c(.) = \sum_{j=1}^{h} c_j s_j(.),$$

where the coefficients c_j, $j = 1, \ldots, h$ are unknown whereas h is supposed to be known. Introducing $c = (c_1, \ldots, c_h)^t$ and $s(.) = (s_1(.), \ldots, s_h(.))^t$, model (1) can be rewritten as

$$X = \mu + s^t(Y)cVb + \varepsilon. \tag{3}$$

Let (X_i, Y_i), $i = 1, \ldots, n$ be a sample of independent random variables distributed as (X, Y). Clearly, estimating (μ, V, b, c) by maximization of the likelihood in model (3) consists in minimizing

$$G(\mu, V, b, c) = \log\det V + \frac{1}{n}\sum_{i=1}^{n}(\mu + s^t(Y_i)cVb - X_i)^t V^{-1}(\mu + s^t(Y_i)cVb - X_i), \tag{4}$$

with respect to (μ, V, b, c). Note that $G(\mu, V, b, c)$ can also be interpreted as a discrepancy functional, see equation (5) in [9]. Up to our knowledge, the introduction of such functional in the inverse regression framework is due to [7]. The functional $G(\mu, V, b, c)$ generalizes equation (3) in [19] since, here, no assumptions are made on the basis functions $s_j(.)$, $j = 1, \ldots, h$. Let us introduce the $h \times h$ empirical covariance matrix W of $s(Y)$ defined by

$$W = \frac{1}{n}\sum_{i=1}^{n}(s(Y_i) - \bar{s})(s(Y_i) - \bar{s})^t,$$

the $h \times p$ matrix M defined by

$$M = \frac{1}{n}\sum_{i=1}^{n}(s(Y_i) - \bar{s})(X_i - \bar{X})^t,$$

and $\hat{\Sigma}$ the empirical $p \times p$ covariance matrix of X

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^t,$$

where

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \bar{s} = \frac{1}{n}\sum_{i=1}^{n} s(Y_i).$$

Using these notations, standard calculations show that $G(\mu, V, b, c)$ can be rewritten as

$$\begin{aligned} G(\mu, V, b, c) &= \log\det V + \mathrm{tr}(\hat{\Sigma}V^{-1}) + (\mu - \bar{X} + \bar{s}^t cVb)^t V^{-1}(\mu - \bar{X} + \bar{s}^t cVb) \\ &+ (c^t Wc)(b^t Vb) - 2c^t Mb, \end{aligned}$$

and thus the maximum likelihood estimators of μ, V, b and c are closed-form.

Proposition 1 *Under (3), if W and $\hat{\Sigma}$ are regular, then the maximum likelihood estimator of (μ, V, b, c) is defined by:*

- *$\hat{b}$ is the eigenvector associated to the largest eigenvalue $\hat{\lambda}$ of $\hat{\Sigma}^{-1}M^tW^{-1}M$,*
- $\hat{c} = \dfrac{1}{\hat{b}^t\hat{V}\hat{b}}W^{-1}M\hat{b}$,

- $\hat{\mu} = \bar{X} - \bar{s}^t \hat{c} \hat{V} \hat{b}$,
- $\hat{V} = \hat{\Sigma} - \dfrac{\hat{\lambda}}{\hat{b}^t \hat{\Sigma} \hat{b}} \hat{\Sigma} \hat{b} \hat{b}^t \hat{\Sigma}$.

In the next paragraph, we show that the SIR method corresponds to the particular case of piecewise constant basis functions $s_j(.)$, $j = 1, \ldots, h$.

2.3 Sliced Inverse Regression (SIR)

Suppose the range of Y is partitioned into $h+1$ non-overlapping slices S_j, $j = 1, \ldots, h+1$ and consider the h basis functions defined by

$$s_j(.) = \mathbb{I}\{. \in S_j\}, \quad j = 1, \ldots, h \tag{5}$$

where $\mathbb{I}$ is the indicator function. Let us denote by n_j the number of Y_i on the slice $j = 1, \ldots, h+1$, define the corresponding proportion by $f_j = n_j/n$, the empirical mean of X given $Y \in S_j$ by

$$\bar{X}_j = \frac{1}{n_j} \sum_{Y_i \in S_j} X_i$$

and let Γ the $p \times p$ empirical "between slices" covariance matrix defined by

$$\Gamma = \sum_{j=1}^{h+1} f_j (\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^t.$$

In this context, the following consequence of Proposition 1 can be established.

Corollary 1 *Under (3) and (5), if $\hat{\Sigma}$ is regular, then the maximum likelihood estimator $\hat{b}$ of b is the eigenvector associated to the largest eigenvalue of $\hat{\Sigma}^{-1}\Gamma$.*

To summarize, the maximum likelihood estimator of b is (up to a scale factor) the classical SIR estimator of the axis β spanning the central subspace. The next section is dedicated to the introduction of a regularization in the inverse regression problem in order to avoid the inversion of $\hat{\Sigma}$.

3 Regularized inverse regression

First, we present in Subsection 3.1 how the introduction of a prior information on the unknown axis b can overcome the SIR limitations due to the ill-conditioning or singularity of $\hat{\Sigma}$. Second, some links with the existing SIR regularizations are highlighted in Subsection 3.2. Finally, basing on our framework, three new regularizations of the SIR method are introduced in Subsection 3.3.

3.1 Our approach

We propose to introduce a prior information on the projection of X on b appearing in model (3). More precisely, we assume that

$$\left(\frac{b^tVb}{b^t\Sigma b}\right)^{1/2}(s(Y)-\bar{s})^tcb \sim \mathcal{N}(0,\Omega), \tag{6}$$

where $\mathcal{N}(0,\Omega)$ is the multivariate centered Gaussian distribution with covariance matrix Ω. The role of the matrix Ω is to describe which directions in $\mathbb{R}^p$ are the most likely to contain b. Some examples are provided in the next two paragraphs. As a comparison, in [2], a Bayesian estimation method is proposed basing on B-splines approximation of the link function g in the forward model and a Fisher-von Mises prior on the axis b. In our approach, working on the inverse regression model allows to obtain explicit solutions, see Proposition 2 below. Maximum A Posteriori (MAP) estimators are obtained by minimizing

$$G_\Omega(\mu,V,b,c) = G(\mu,V,b,c) + \frac{(b^t\Omega^{-1}b)(b^tVb)(c^tWc)}{b^t\Sigma b} \tag{7}$$

with respect to (μ,V,b,c). Comparing to (4), the additional term due to the prior information can be read as a regularization term in Tikhonov theory, see for instance [24], Chapter 1, penalizing large projections on the axis. The scalar factor $(b^tVb/b^t\Sigma b)^{1/2}$ is introduced in (6) to normalize the amplitude of the penalization. The following result can be stated:

Proposition 2 *Under (3) and (6), if W and $\Omega\hat{\Sigma}+I_p$ are regular, then the MAP estimator of (μ,V,b,c) is defined by:*

- *$\hat{b}$ is the eigenvector associated to the largest eigenvalue $\hat{\lambda}$ of $(\Omega\hat{\Sigma}+I_p)^{-1}\Omega M^tW^{-1}M$,*
- *$\hat{c} = \dfrac{1}{(1+\eta(\hat{b}))\hat{b}^t\hat{V}\hat{b}}W^{-1}M\hat{b}$, where $\eta(\hat{b}) = \dfrac{\hat{b}^t\Omega^{-1}\hat{b}}{\hat{b}^t\hat{\Sigma}\hat{b}}$,*
- *$\hat{\mu} = \bar{X} - \bar{s}^t\hat{c}\hat{V}\hat{b}$,*
- *$\hat{V} = \hat{\Sigma} - \dfrac{\hat{\lambda}}{\hat{b}^t\hat{\Sigma}\hat{b}}\hat{\Sigma}\hat{b}\hat{b}^t\hat{\Sigma}$.*

Compared to Proposition 1, the inversion of $\hat{\Sigma}$ is replaced by the inversion of $\Omega\hat{\Sigma}+I_p$. Thus, for a properly chosen prior matrix Ω, the numerical instabilities in the estimation of b disappear. As previously, this result can be applied to the particular case of the SIR method.

Corollary 2 *Under (3), (5) and (6), if $\Omega\hat{\Sigma}+I_p$ is regular, then the MAP estimator of b is the eigenvector associated to the largest eigenvalue of $(\Omega\hat{\Sigma}+I_p)^{-1}\Omega\Gamma$.*

In the following, the above estimator of the axis b will be referred to as the Gaussian Regularized Sliced Inverse Regression (GRSIR) estimator. Some examples of possible prior covariance matrices Ω are now presented.

3.2 Links with existing methods

In all the next examples, a non-negative parameter τ is introduced in the prior covariance matrix in order to tune the importance of the penalty term in (7). Consequently, in the sequel, τ is called a regularization parameter.

Classical SIR approach. It is easily seen from Corollary 2 that choosing the prior covariance matrix $\Omega_0 = (\tau\hat{\Sigma})^{-1}$ in GRSIR gives back SIR, and this for all $\tau > 0$. This prior matrix indicates that directions corresponding to small variances are most likely, *i.e* the SIR method favors directions in which $\hat{\Sigma}$ is close to singularity. In practice, this choice yields instabilities in the estimation.

Ridge approach [25]. The simplest choice for the prior covariance matrix is $\Omega_1 = \tau^{-1} I_p$. In this case, the identity matrix indicates that no privileged direction for b is available. Following Corollary 2, the corresponding GRSIR estimator of b is the eigenvector of $(\hat{\Sigma}+\tau I_p)^{-1}\Gamma$ associated to its largest eigenvalue, which is the ridge estimator introduced in [25].

PCA+SIR approach [4, 18]. As already seen, a popular technique to overcome the singularity problems of $\hat{\Sigma}$ is to use PCA as a preprocessing step [4, 18]. The principle is the following. Let $d \in \mathbb{N}$ be fixed and denote by $\lambda_1 \geq \cdots \geq \lambda_d$ the d largest eigenvalues of $\hat{\Sigma}$ (supposed to be positive), $q_1, \ldots, q_d$ the associated eigenvectors and $S_d = \text{span}(q_1, \ldots, q_d)$ the linear subspace spanned by $q_1, \ldots, q_d$. The first step consists in projecting the predictors on S_d. The second step is to perform SIR in this subspace. The next result shows that this method corresponds to a particular prior covariance matrix.

Proposition 3 *PCA+SIR corresponds to GRSIR with prior covariance matrix*

$$\Omega_2 = \frac{1}{\tau}\sum_{j=1}^{d} \frac{1}{\lambda_j} q_j q_j^t,$$

where $\tau > 0$ is arbitrary.

Let us note that although Ω_2 depends on τ, the GRSIR estimator does not, since there is no regularization parameter in the PCA+SIR methodology.

Li and Yin's approach [19]. In this paper, the proposed L_2- regularization consists in estimating (b, c) by minimization of

$$H_\tau(b,c) = \sum_{j=1}^{h} f_j(\mu + \hat{\Sigma}c_j b - \bar{X}_j)^t N(\mu + \hat{\Sigma}c_j b - \bar{X}_j) + \tau b^t b,$$

the matrix N being either $N = I_p$ or $N = \hat{\Sigma}^{-1}$. In our opinion, this approach suffers from a lack of invariance since the functional H_τ does not penalize the same way two different

axes (b and $2b$ for instance) defining the same direction. As a consequence, one can show that the only possible solution $\hat{b}$ of the minimization problem is $\hat{b} = 0$:

Proposition 4 *Let $\tau > 0$ and let N be a regular and symmetric matrix.*
If $(\hat{b}, \hat{c}) \in \arg\min_{b,c} H_\tau(b, c)$ then $\hat{b} = 0$.

In view of this result, the proposed alternating least squares algorithm ([19], Section 2) cannot be justified theoretically. As a comparison, our method does not yield this kind of problem thanks to the invariance property: $G_\Omega(\mu, V, tb, c/t) = G_\Omega(\mu, V, b, c)$ for all real number $t \neq 0$. We now propose some alternative choices of the covariance matrix Ω yielding new regularizations of the SIR method.

3.3 Three new SIR regularizations

Tikhonov regularization. An alternative choice of the prior covariance matrix is $\Omega_3 = \tau^{-1}\hat{\Sigma}$. Comparing Ω_3 to the matrix Ω_0 associated to the SIR method, it appears that the underlying ideas are opposite. Here, directions corresponding to large variances are most likely. The associated GRSIR estimator of the axis b is the eigenvector of $(\hat{\Sigma}^2 + \tau I_p)^{-1}\hat{\Sigma}\Gamma$ associated to its largest eigenvalue. In the following, this estimator will be referred to as the Tikhonov estimator. Indeed, let us recall that the classical SIR estimator is obtained by a spectral decomposition of $\hat{\Sigma}^{-1}\Gamma$. For all $k = 1, \ldots, p$, denote by x_k the $k-$th column of this matrix. Computing x_k is equivalent to solving with respect to x the linear system $\hat{\Sigma}x = \Gamma_k$ where Γ_k is the $k-$th column of Γ. The associated Tikhonov minimization problem (see (1.34) in [24]) can be written as

$$x_k = \arg\min_x \|\hat{\Sigma}x - \Gamma_k\|^2 + \tau\|x\|^2 = (\hat{\Sigma}^2 + \tau I_p)^{-1}\hat{\Sigma}\Gamma_k.$$

Thus, in this framework, $(\hat{\Sigma}^{-1}\Gamma)_k$ is estimated by $(\hat{\Sigma}^2 + \tau I_p)^{-1}\hat{\Sigma}\Gamma_k$ and consequently $\hat{\Sigma}^{-1}\Gamma$ is estimated by $(\hat{\Sigma}^2 + \tau I_p)^{-1}\hat{\Sigma}\Gamma$.

Dimension reduction approaches. It has been seen in Proposition 3, that the PCA+SIR approach is equivalent to using the prior covariance matrix Ω_2 in GRSIR. The following result is an extension to more general covariance matrices.

Proposition 5 *For all real function φ let*

$$\Omega(\varphi) = \sum_{j=1}^{d} \varphi(\lambda_j) q_j q_j^t.$$

Then, the associated GRSIR estimator can be obtained by first projecting the predictors on $S_d = span(q_1, \ldots, q_d)$ and second performing GRSIR on the projected predictors with prior covariance matrix

$$\tilde{\Omega}(\varphi) = \sum_{j=1}^{p} \varphi(\lambda_j) q_j q_j^t.$$

The dimension d plays the role of a "cut-off" parameter, since when computing $\hat{b}$, all directions $q_{d+1}, \ldots, q_p$ are discarded. Three illustrations of this result can be given:

- Choosing $\varphi(t) = 1/(\tau t)$, we obtain $\Omega(1/(\tau \mathrm{Id})) = \Omega_2$ and $\tilde{\Omega}(1/(\tau \mathrm{Id})) = (\tau\hat{\Sigma})^{-1} = \Omega_0$, where Id is the identity function. It appears that Proposition 3 is a particular case of Proposition 5. As already discussed, since the choice of Ω_0 as a prior covariance matrix seems not very natural, we thus propose two new choices.

- First, $\varphi(t) = 1/\tau$ yields

$$\Omega_4 \stackrel{def}{=} \Omega(1/\tau) = \frac{1}{\tau}\sum_{j=1}^{d} q_j q_j^t,$$

  and $\tilde{\Omega}(1/\tau) = I_p/\tau = \Omega_1$. Consequently, this new method consists in applying the ridge approach [25] on the projected predictors, the interpretation being that, in the subspace S_d, all directions share the same prior probability. This method will be referred to as PCA+ridge.

- Second, $\varphi(t) = t/\tau$ yields

$$\Omega_5 \stackrel{def}{=} \Omega(\mathrm{Id}/\tau) = \frac{1}{\tau}\sum_{j=1}^{d} \lambda_j q_j q_j^t,$$

  and $\tilde{\Omega}(\mathrm{Id}/\tau) = \hat{\Sigma}/\tau = \Omega_3$. This new method consists in applying Tikhonov approach on the projected predictors. In this context, directions of S_d carrying a large fraction of the total variance of X are more likely. This method will be referred to as PCA+Tikhonov.

In the next section, GRSIR methods associated to the prior covariance matrices Ω_0 (SIR), Ω_1 (ridge), Ω_2 (PCA+SIR), Ω_3 (Tikhonov), Ω_4 (PCA+ridge) and Ω_5 (PCA+Tikhonov) are compared on simulated data.

4 Numerical experiments

In these experiments, a sample of size $n = 100$ of the random pair (X, Y) is considered, where $X \in \mathbb{R}^p$ with $p = 50$ and $Y \in \mathbb{R}$. The random vector X is centered with covariance matrix $\Sigma = Q\Delta Q^t$ where Q is an orthogonal matrix randomly chosen and Δ is the diagonal matrix containing the eigenvalues of Σ defined by $\Delta = \mathrm{diag}\{1^\theta, 2^\theta, \ldots, p^\theta\}$. Several values of θ will be considered. Note that the condition number of Σ is given by p^θ and is thus an increasing function of θ. The random variable Y is given by

$$Y = \sin\left(\frac{\pi}{2\sigma}\beta^t X\right),$$

where the true index is $\beta = p^{-1/2}(1, \ldots, 1)^t$ and σ is the standard deviation of the projection of X on β *i.e.* $\sigma = (\beta^t\Sigma\beta)^{1/2}$. In all the experiments, the quality of the estimate

$\hat{b}$ (supposed to be normed) is evaluated by computing the mean squared cosine $(\beta^t\hat{b})^2$ on $N = 100$ replications of the simulated model.

The first part of the study is dedicated to the illustration of the dependence of the above methods with respect to the regularization parameter τ. A logarithmic scale was adopted, 150 values of $\log_{10}(\tau)$ regularly distributed in $[-5, 25]$ were considered. Here, we limit ourselves to $\theta = 2$ and $d = 20$ in the PCA+SIR, PCA+ridge and PCA+Tikhonov methods. Results are displayed on Figure 1 and Figure 2. It appears that the classical SIR approach gives very poor results in such a situation where n and p are of the same order. Ridge and Tikhonov regularizations can bring a significant improvement provided τ is large enough. One can see that Tikhonov regularization can outperform the ridge regression but the choice of the regularization parameter is more crucial. PCA+SIR obtains reasonable results compared to SIR, with the advantage of do not requiring the selection of τ. The selection of d is addressed in our third experiment. Note that PCA+ridge and PCA+Tikhonov methods both outperform PCA+SIR for all values of the regularization parameter. Moreover, PCA+ridge and PCA+Tikhonov methods are less sensitive to the choice of τ than ridge and Tikhonov methods. The pairs $(\beta^t X_i, Y_i)$, $i = 1, \ldots, n$ are represented on Figure 3 for one of the $N = 100$ simulated datasets. As a comparison, for the same dataset, the pairs $(\hat{b}^t X_i, Y_i)$, $i = 1, \ldots, n$ are plotted on Figure 4, where $\hat{b}$ is the axis estimated by the PCA+ridge method, with the optimal regularization parameter, *i.e.* maximizing the squared cosine. It appears that, even in this difficult situation, the shape of the link function still appears along the estimated axis.

In the second part of the experiment, the robustness with respect to the condition number is investigated by varying θ in $\{0, 0.1, 0.2, \ldots, 3\}$. For each value within this set, the optimal regularization parameter has been selected for each method and the corresponding squared cosine is displayed on Figure 5 and Figure 6. Clearly, the classical SIR method is very sensitive to the ill-conditioning of the covariance matrix. For all the other considered methods, results are getting better while the condition number increases. Note that ridge and Tikhonov methods as well as PCA+ridge and PCA+Tikhonov yield very similar results.

The third experiment in dedicated to the illustration of the role of d in PCA+SIR, PCA+ridge and PCA+Tikhonov methods. Here, the condition number is fixed by choosing $\theta = 2$. For each value of d in $\{0, 1, \ldots, p\}$, the optimal regularization parameter has been selected for each method and the corresponding squared cosine is displayed on Figure 7. One can see that the PCA+SIR method is very sensitive to d. Indeed, if d is large, then this approach reduces to SIR, whose accuracy is low for large dimensions. At the opposite, PCA+ridge and PCA+Tikhonov results remain stable as d increases, since these methods get close to ridge and Tikhonov methods respectively.

5 Concluding remarks

A new framework has been presented to regularize the SIR method. It provides new interpretations of the existing methods as well as the construction of new regularization techniques. Among them, it appears that the PCA+ridge and PCA+Tikhonov methods are interesting alternatives to the existing PCA+SIR [4, 18] and ridge [25] methods. The use of a "cut-off" dimension d permits to limit the sensibility to the choice of the regularization parameter τ. The choice of this dimension itself seems not crucial since for large value the above methods are close to the ridge and Tikhonov techniques. In our experiments, the choice $d = p/2$ appears as a good heuristics in most situations. Of course, an automatic selection of (τ, d) would be of interest. To this end, the construction of a generalized cross-validation criterion is under investigation. We also plan to study the introduction on non-Gaussian priors in order to obtain L_1- penalizations and thus sparse estimates of β.

References

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

[2] Antoniadis, A., Grégoire, G. and McKeague, I.W. (2004) Bayesian estimation in single-index models. *Statistica Sinica*, **14**, 1147–1164.

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

[4] Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Bio- sciences*, **176**, 123–144.

[5] Cook, R.D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction" by K.C. Li, *Journal of the American Statistical Association*, **86**, 328–332.

[6] Cook, R.D. (1998). *Regression graphics. Ideas for studying regressions through graphics.* Wiley Series in Probability and Statistics, New York.

[7] Cook, R.D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, **32**, 1062–1092.

[8] Cook, R.D. (2005). Fisher Lecture: Dimension reduction in regression. *Joint Statistical Meetings*, Minneapolis.

[9] Cook, R.D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, **100**, 410–428.

[10] Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

[11] Draper, N.R. and Smith, R. (1998). *Applied regression analysis (3rd edition)*, Wiley.

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

[13] Ferré, L. and Yao, A.F. (2005). Smoothed functional inverse regression. *Statistica Sinica*, **15**, 665–683.

[14] Friedman, J.H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, **84**, 165–175.

[15] Gannoun, A. and Saracco, J. (2003). An asymptotic theory for SIR_α method. *Statistica Sinica*, **13**, 297–310.

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

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

[18] Li, L. and Li, H. (2004). Dimension reduction methods for micro-arrays with application to censored survival data. *Bioinformatics*, **20** (18), 3406–3412.

[19] Li, L. and Yin, X. (2007). Sliced inverse regression with regularizations. *Biometrics*, to appear.

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

[21] Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on SIR_α approach. *Journal of Multivariate Analysis*, **96**(1), 117–135.

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

[23] Scrucca, L. (2006). Regularized sliced inverse regression with applications in classification. In Zani S., Cerioli A., Riani M. and Vichi M., editors, *Data Analysis, Classification and the Forward Search*, pp. 59-66, Berlin, Springer-Verlag,

[24] Vogel, C.R. (2002). *Computational methods for inverse problems*, Society for Industrial and Applied Mathematics, Philadelphia.

[25] 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.

[26] Zhu, L.X. and Ng, K.W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica*, **5**, 727–736.

Acknowledgments

This research is partially supported by IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy).

Appendix: Proofs

Proof of Proposition 1. Annulling the gradients of $G(\mu, V, b, c)$ yields the system of equations

$$\begin{aligned}
\frac{1}{2}\nabla_\mu G &= \hat{V}^{-1}(\hat{\mu} - \bar{X} + \bar{s}^t\hat{c}\hat{V}\hat{b}) = 0, && (8)\\
\frac{1}{2}\nabla_b G &= \left((\hat{c}^t\bar{s})^2 + \hat{c}^t W\hat{c}\right)\hat{V}\hat{b} - M^t\hat{c} + (\hat{c}^t\bar{s})(\hat{\mu} - \bar{X}) = 0, && (9)\\
\frac{1}{2}\nabla_c G &= (\hat{b}^t\hat{V}\hat{b})(\bar{s}\bar{s}^t + W)\hat{c} - M\hat{b} + (\hat{\mu} - \bar{X})^t\hat{b}\bar{s} = 0, && (10)\\
\nabla_V G &= \hat{V}^{-1} + \hat{b}\hat{b}^t\left((\hat{c}^t\bar{s})^2 + \hat{c}^t W\hat{c}\right) - \hat{V}^{-1}\left((\hat{\mu} - \bar{X})(\hat{\mu} - \bar{X})^t + \hat{\Sigma}\right)\hat{V}^{-1} = 0. && (11)
\end{aligned}$$

From (8), we have

$$\hat{\mu} = \bar{X} - \bar{s}^t\hat{c}\hat{V}\hat{b}. \tag{12}$$

Replacing in (9) and (10) yields the simplified system of equations

$$\begin{aligned}
(\hat{c}^t W\hat{c})\hat{V}\hat{b} &= M^t\hat{c}, && (13)\\
(\hat{b}^t\hat{V}\hat{b})W\hat{c} &= M\hat{b}. && (14)
\end{aligned}$$

Assuming W is regular, equation (14) entails

$$\hat{c} = W^{-1}M\hat{b}/\hat{b}^t\hat{V}\hat{b}$$

and replacing in (13) yields

$$(\hat{c}^t W\hat{c})(\hat{b}^t\hat{V}\hat{b})\hat{V}b = M^t W^{-1}M\hat{b}. \tag{15}$$

Now, multiplying (11) on the left and on the right by $\hat{V}$ and taking account of (12), the following equation follows

$$\hat{\Sigma} = \hat{V} + (\hat{c}^t W\hat{c})\hat{V}\hat{b}\hat{b}^t\hat{V}. \tag{16}$$

As a consequence of (16), we have

$$\hat{\Sigma}\hat{b} = \left(1 + (\hat{c}^t W\hat{c})(\hat{b}^t\hat{V}\hat{b})\right)\hat{V}\hat{b}, \tag{17}$$

which means that $\hat{\Sigma}\hat{b}$ is proportional to $\hat{V}\hat{b}$. As a consequence, one also has

$$\hat{V}\hat{b} = \theta(\hat{b})\hat{\Sigma}\hat{b}, \tag{18}$$

where we have defined

$$\theta(\hat{b}) = \frac{\hat{b}^t\hat{V}\hat{b}}{\hat{b}^t\hat{\Sigma}\hat{b}}.$$

Substituting (18) in (15) yields

$$(\hat{c}^tW\hat{c})(\hat{b}^t\hat{V}\hat{b})\theta(\hat{b})\hat{\Sigma}b = M^tW^{-1}M\hat{b}$$

and thus b is an eigenvector of $\hat{\Sigma}^{-1}M^tW^{-1}M$. Let us denote by $\hat{\lambda}$ the associated eigenvalue

$$\hat{\lambda} = (\hat{c}^tW\hat{c})(\hat{b}^t\hat{V}\hat{b})\theta(\hat{b}). \tag{19}$$

Collecting (17) and (18) yields

$$\frac{1}{\theta(\hat{b})} = 1 + (\hat{c}^tW\hat{c})(\hat{b}^t\hat{V}\hat{b}),$$

and thus the eigenvalue can be rewritten as

$$\hat{\lambda} = 1 - \theta(\hat{b}).$$

Moreover, we have, from (13),

$$\begin{aligned}
\hat{c}^tM\hat{b} &= \hat{\lambda}/\theta(\hat{b}),\\
\mathrm{tr}(\hat{\Sigma}\hat{V}^{-1}) &= p + \hat{\lambda}/\theta(\hat{b}),\\
\log\det\hat{V} &= \log\det\hat{\Sigma} - \log\det\left(I_p + (\hat{c}^tW\hat{c})\hat{V}\hat{b}\hat{b}^t\right)\\
&= \log\det\hat{\Sigma} - \log\left(1 + \hat{\lambda}/\theta(\hat{b})\right),
\end{aligned}$$

entailing

$$\begin{aligned}
G(\hat{\mu},\hat{V},\hat{b},\hat{c}) &= p + \log\det\hat{\Sigma} - \log\left(1 + \hat{\lambda}/\theta(\hat{b})\right)\\
&= p + \log\det\hat{\Sigma} + \log(1-\hat{\lambda}).
\end{aligned}$$

As a consequence, to minimize G, $\hat{\lambda}$ should be the largest eigenvalue. Finally, let us consider

$$V_0 = \hat{\Sigma} - \frac{\hat{\lambda}}{\hat{b}^t\Sigma\hat{b}}\hat{\Sigma}\hat{b}\hat{b}^t\hat{\Sigma},$$

leading to

$$\begin{aligned}
V_0 + (\hat{c}^tW\hat{c})V_0\hat{b}\hat{b}^tV_0 &= \hat{\Sigma} + \left((\hat{c}^tW\hat{c})\theta^2(\hat{b}) - \frac{\hat{\lambda}}{\hat{b}^t\Sigma\hat{b}}\right)\hat{\Sigma}\hat{b}\hat{b}^t\hat{\Sigma}\\
&= \hat{\Sigma} + \frac{\hat{\lambda}\theta(\hat{b})}{\hat{b}^tV_0\hat{b}}\left((\hat{c}^tW\hat{c})(\hat{b}^tV_0\hat{b})\theta(\hat{b}) - \hat{\lambda}\right)\hat{\Sigma}\hat{b}\hat{b}^t\hat{\Sigma}\\
&= \hat{\Sigma},
\end{aligned}$$

in view of (19) and thus V_0 verifies equation (16). ■

Proof of Corollary 1. Let us remark that, under (5), the coefficients of W_{ij} of W are given by $W_{ij} = f_i \mathbb{I}\{i=j\} - f_i f_j$ for all $(i,j) \in \{1,\dots,h\}^2$. The inverse matrix of W is

$$W^{-1} = \text{diag}\left(\frac{1}{f_1},\dots,\frac{1}{f_j}\right) + \frac{1}{f_{h+1}}U,$$

where U is the $h \times h$ matrix defined by $U_{ij} = 1$ for all $(i,j) \in \{1,\dots,h\} \times \{1,\dots,h\}$. Since the $j-$ line of M is given by $f_j(\bar{X}_j - \bar{X})^t$ for all $j = 1,\dots,h$ and taking account of $U^2 = f_{h+1}U$, we have

$$\begin{aligned} M^t W^{-1} M &= \sum_{j=1}^{h} f_j(\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^t + \frac{1}{f_{h+1}} M^t U M \\ &= \sum_{j=1}^{h} f_j(\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^t + \frac{1}{h f_{h+1}} (M^t U)(M^t U)^t. \end{aligned} \tag{20}$$

Now, remarking that all the columns of $M^t U$ are equal to

$$\begin{aligned} \sum_{j=1}^{h} f_j(\bar{X}_j - \bar{X}) &= \sum_{j=1}^{h} f_j(\bar{X}_j - \bar{X}) - f_{h+1}(\bar{X}_{h+1} - \bar{X}) \\ &= -f_{h+1}(\bar{X}_{h+1} - \bar{X}), \end{aligned}$$

if follows that

$$(M^t U)(M^t U)^t = h f_{h+1}^2 (\bar{X}_{h+1} - \bar{X})(\bar{X}_{h+1} - \bar{X})^t$$

and thus replacing in (20) yields

$$\begin{aligned} M^t W^{-1} M &= \sum_{j=1}^{h+1} f_j(\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^t \\ &= \Gamma. \end{aligned}$$

The result is then a consequence of Proposition 1. ∎

Proof of Proposition 2. The proof is similar to the one of Proposition 1. First, remark that equation (8) still holds and thus $\hat{\mu} = \bar{X} - \bar{s}^t \hat{c} \hat{V} \hat{b}$. Let us recall the following definitions

$$\theta(\hat{b}) = \frac{\hat{b}^t \hat{V} \hat{b}}{\hat{b}^t \hat{\Sigma} \hat{b}} \text{ and } \eta(\hat{b}) = \frac{\hat{b}^t \Omega^{-1} \hat{b}}{\hat{b}^t \hat{\Sigma} \hat{b}}.$$

Annulling the gradients of $G_\Omega(\mu, V, b, c)$ yields the system of equations

$$(\hat{c}^t W \hat{c}) \left(\Omega \hat{V} \hat{b} + \theta(\hat{b}) \hat{b} + \eta(\hat{b}) \Omega (\hat{V} \hat{b} - \theta(\hat{b}) \hat{\Sigma} \hat{b}) \right) = \Omega M^t \hat{c}, \tag{21}$$

$$(\hat{b}^t \hat{V} \hat{b})(1 + \eta(\hat{b})) W \hat{c} = M \hat{b}, \tag{22}$$

$$\hat{V}^{-1} + \hat{b} \hat{b}^t (\hat{c}^t W \hat{c})(1 + \eta(\hat{b})) = \hat{V}^{-1} \hat{\Sigma} \hat{V}^{-1}. \tag{23}$$

Assuming W is regular, equation (22) entails

$$\hat{c} = W^{-1}M\hat{b}/((1+\eta(\hat{b}))(\hat{b}^t\hat{V}\hat{b}))$$

and replacing in (21) yields

$$(\hat{c}^tW\hat{c})(\hat{b}^t\hat{V}\hat{b})(1+\eta(\hat{b}))\left(\Omega\hat{V}\hat{b}+\theta(\hat{b})\hat{b}+\eta(\hat{b})\Omega(\hat{V}\hat{b}-\theta(\hat{b})\hat{\Sigma}\hat{b})\right) = \Omega M^tW^{-1}M\hat{b}. \quad (24)$$

Now, multiplying (23) on the left and on the right by $\hat{V}$, it follows that

$$\hat{\Sigma} = \hat{V} + (\hat{c}^tW\hat{c})(1+\eta(\hat{b}))\hat{V}\hat{b}\hat{b}^t\hat{V}, \quad (25)$$

leading to

$$\hat{\Sigma}\hat{b} = \left(1+(\hat{c}^tW\hat{c})(\hat{b}^t\hat{V}\hat{b})(1+\eta(\hat{b}))\right)\hat{V}\hat{b}, \quad (26)$$

which means that $\hat{\Sigma}\hat{b}$ is proportional to $\hat{V}\hat{b}$. As a consequence, one also has

$$\hat{V}\hat{b} = \theta(\hat{b})\hat{\Sigma}\hat{b}, \quad (27)$$

and replacing in (24) yields

$$(\hat{c}^tW\hat{c})(\hat{b}^t\hat{V}\hat{b})(1+\eta(\hat{b}))\theta(\hat{b})\left(\Omega\hat{\Sigma}+I_p\right)\hat{b} = \Omega M^tW^{-1}M\hat{b},$$

and thus b is an eigenvector of $(\Omega\hat{\Sigma}+I_p)^{-1}\Omega M^tW^{-1}M$. Let us denote by $\hat{\lambda}$ the associated eigenvalue

$$\hat{\lambda} = (\hat{c}^tW\hat{c})(\hat{b}^t\hat{V}\hat{b})(1+\eta(\hat{b}))\theta(\hat{b}). \quad (28)$$

Collecting (26) and (27) yields

$$\frac{1}{\theta(\hat{b})} = 1+(\hat{c}^tW\hat{c})(\hat{b}^t\hat{V}\hat{b})(1+\eta(\hat{b}))$$

and consequently the eigenvalue can be rewritten as

$$\hat{\lambda} = 1-\theta(\hat{b}).$$

Now, let us remark that

$$\begin{aligned}
\hat{c}^tM\hat{b} &= \hat{\lambda}/\theta(\hat{b}),\\
\mathrm{tr}(\hat{\Sigma}\hat{V}^{-1}) &= p+\hat{\lambda}/\theta(\hat{b}),\\
\log\det\hat{V} &= \log\det\hat{\Sigma}-\log\left(1+\hat{\lambda}/\theta(\hat{b})\right),
\end{aligned}$$

entailing

$$\begin{aligned}
G_\Omega(\hat{\mu},\hat{V},\hat{b},\hat{c}) &= p+\log\det\hat{\Sigma}-\log(1+\hat{\lambda}/\theta(\hat{b}))\\
&= p+\log\det\hat{\Sigma}+\log(1-\hat{\lambda})
\end{aligned}$$

As a consequence, to minimize G_Ω, $\hat{\lambda}$ should be the largest eigenvalue. Finally, let us consider

$$V_0 = \hat{\Sigma} - \frac{\hat{\lambda}}{\hat{b}^t\Sigma\hat{b}}\hat{\Sigma}\hat{b}\hat{b}^t\hat{\Sigma},$$

leading to

$$\begin{aligned} V_0 + (\hat{c}^tW\hat{c})(1+\eta(\hat{b}))V_0\hat{b}\hat{b}^tV_0 &= \hat{\Sigma} + \left((\hat{c}^tW\hat{c})(1+\eta(\hat{b}))\theta^2(\hat{b}) - \frac{\hat{\lambda}}{\hat{b}^t\Sigma\hat{b}}\right)\hat{\Sigma}\hat{b}\hat{b}^t\hat{\Sigma} \\ &= \hat{\Sigma} + \frac{\hat{\lambda}\theta(\hat{b})}{\hat{b}^tV_0\hat{b}}\left((\hat{c}^tW\hat{c})(\hat{b}^tV_0\hat{b})(1+\eta(\hat{b}))\theta(\hat{b}) - \hat{\lambda}\right)\hat{\Sigma}\hat{b}\hat{b}^t\hat{\Sigma} \\ &= \hat{\Sigma}, \end{aligned}$$

in view of (28) and thus V_0 verifies equation (25). ∎

Proof of Proposition 3. Let P be the projection matrix on S_d:

$$P = \sum_{j=1}^{d} q_j q_j^t,$$

and, for all $i = 1, \ldots, n$ consider the projected predictor defined by $\tilde{X}_i = PX_i$. Introducing $\tilde{\Gamma} = P\Gamma P$ the empirical "between slices" matrix associated to $\tilde{X}_1, \ldots, \tilde{X}_n$ and $\tilde{\Sigma} = P\hat{\Sigma}P$ the corresponding covariance matrix, the PCA+SIR method finds $\tilde{b}$ such that

$$\tilde{\Gamma}\tilde{b} = \tilde{\lambda}\tilde{\Sigma}\tilde{b},$$

where $\tilde{\lambda} \in \mathbb{R}$, or equivalently,

$$P\Gamma P\tilde{b} = \tilde{\lambda}P\hat{\Sigma}P\tilde{b}.$$

Remarking that $P\hat{\Sigma}P = \hat{\Sigma}P$, we have, for all $\tau > 0$,

$$P\Gamma P\tilde{b} = \frac{\tilde{\lambda}}{1+\tau}(P\hat{\Sigma} + \tau\hat{\Sigma})P\tilde{b},$$

and defining $b = P\tilde{b}$ and $\lambda = \tilde{\lambda}/(1+\tau)$, it follows that

$$P\Gamma b = \lambda(P\hat{\Sigma} + \tau\hat{\Sigma})b.$$

Since $P = \tau\hat{\Sigma}\Omega_2$, we thus have

$$\hat{\Sigma}\Omega_2\Gamma b = \lambda(\hat{\Sigma}\Omega_2\hat{\Sigma} + \hat{\Sigma})b,$$

which means that b is an eigenvector of $(\Omega_2\hat{\Sigma} + I_p)^{-1}\Omega_2\Gamma$. Corollary 2 concludes the proof, *i.e.* b is the GRSIR estimator with prior covariance matrix Ω_2. ∎

Proof of Proposition 4. Let $(\hat{b}, \hat{c}) \in \arg\min H_\tau$. We thus have $\nabla_b H_\tau(\hat{b}, \hat{c}) = 0$ and $\frac{\partial H_\tau(\hat{b},\hat{c})}{\partial c_j} = 0$ for all $j = 1, \ldots, h$. These equations can be rewritten as

$$\left(\sum_{j=1}^{h} f_j \hat{c}_j^2\right) \hat{\Sigma} N \hat{\Sigma} \hat{b} - \hat{\Sigma} N \sum_{j=1}^{h} f_j \hat{c}_j (\bar{X}_j - \mu) = -\tau \hat{b}, \tag{29}$$

$$\hat{c}_j \hat{b}^t \hat{\Sigma} N \hat{\Sigma} \hat{b} - \hat{b}^t \hat{\Sigma} N (\bar{X}_j - \mu) = 0. \tag{30}$$

Multiplying (29) on the left by $\hat{b}^t$, it yields,

$$\left(\sum_{j=1}^{h} f_j \hat{c}_j^2\right) \hat{b}^t \hat{\Sigma} N \hat{\Sigma} \hat{b} - \hat{b}^t \hat{\Sigma} N \sum_{j=1}^{h} f_j \hat{c}_j (\bar{X}_j - \mu) = -\tau \|\hat{b}\|^2,$$

and, from (30), we get $\tau \|\hat{b}\|^2 = 0$. Consequently, $\hat{b} = 0$. ∎

Proof of Proposition 5. Here, we adopt the notations introduced in the proof of Proposition 3. The GRSIR estimator $\tilde{b}$ computed on the projected predictors $\tilde{X}_1, \ldots, \tilde{X}_n$ verifies

$$\tilde{\Omega}(\varphi) \tilde{\Gamma} \tilde{b} = \tilde{\lambda} (\tilde{\Omega}(\varphi) \tilde{\Sigma} + I_p) \tilde{b},$$

or equivalently

$$\tilde{\Omega}(\varphi) P \Gamma P \tilde{b} = \tilde{\lambda} (\tilde{\Omega}(\varphi) P \hat{\Sigma} P + I_p) \tilde{b}.$$

Multiplying this equation by P on the left, we obtain

$$P \tilde{\Omega}(\varphi) P \Gamma P \tilde{b} = \tilde{\lambda} (P \tilde{\Omega}(\varphi) P \hat{\Sigma} P + P) \tilde{b}.$$

Since $P \tilde{\Omega}(\varphi) P = \Omega(\varphi)$, and introducing $b = P\tilde{b}$, it follows that

$$\Omega(\varphi) \Gamma b = \tilde{\lambda} (\Omega(\varphi) \hat{\Sigma} + I_p) b,$$

which means that b is an eigenvector of $(\Omega(\varphi)\hat{\Sigma} + I_p)^{-1} \Omega(\varphi) \Gamma$. Corollary 2 concludes the proof, *i.e.* b is the GRSIR estimator with prior covariance matrix $\Omega(\varphi)$. ∎

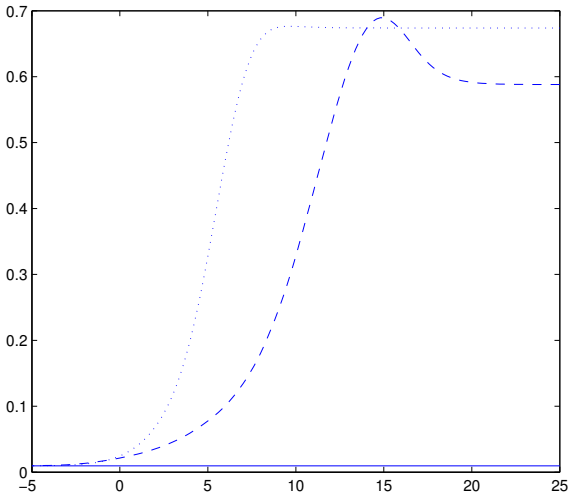

Figure 1: Sensibility of GRSIR with respect to the regularization parameter. Horizontally: $\log_{10}(\tau)$, vertically: $(\beta^t \hat{b})^2$. Continuous line: Ω_0 (SIR), dotted line: Ω_1 (ridge), dashed line: Ω_3 (Tikhonov).

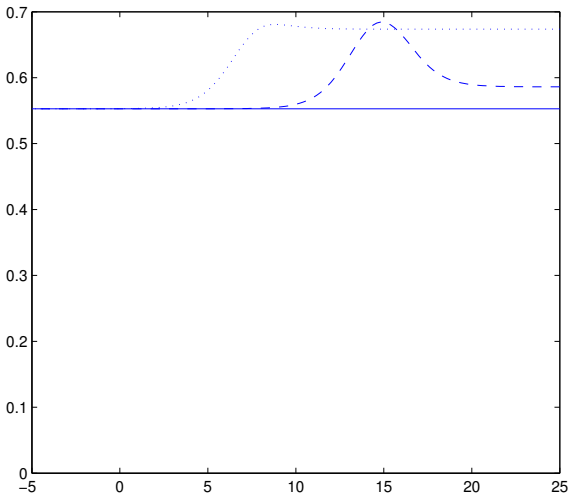

Figure 2: Sensibility of GRSIR with respect to the regularization parameter. The cut-off dimension is chosen to $d = 20$. Horizontally: $\log_{10}(\tau)$, vertically: $(\beta^t \hat{b})^2$. Continuous line: Ω_2 (PCA+SIR), dotted line: Ω_4 (PCA+ridge), dashed line: Ω_5 (PCA+Tikhonov).

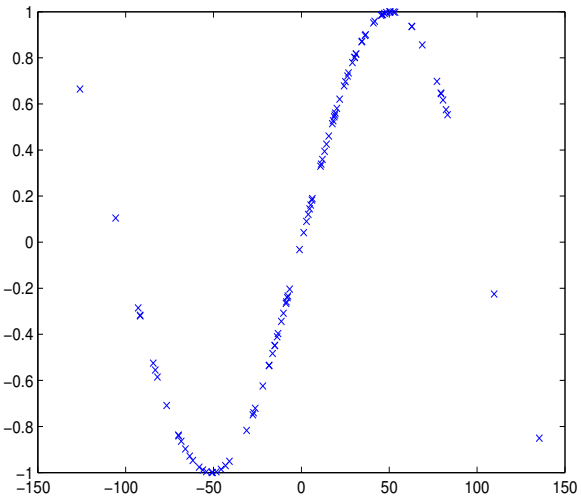

Figure 3: An example of pairs $(\beta^t X_i, Y_i)$, $i = 1, \ldots, n$ obtained after projecting the covariates on the true axis β.

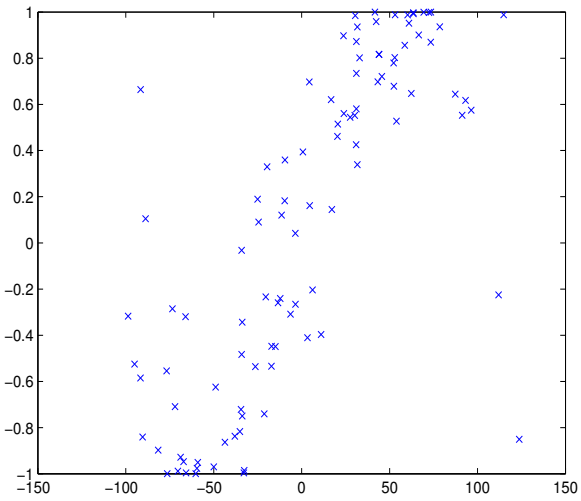

Figure 4: Corresponding pairs $(\hat{b}^t X_i, Y_i)$, $i = 1, \ldots, n$ obtained after projecting the covariates on the axis $\hat{b}$ computed by GRSIR with prior covariance matrix Ω_4 (PCA+ridge).

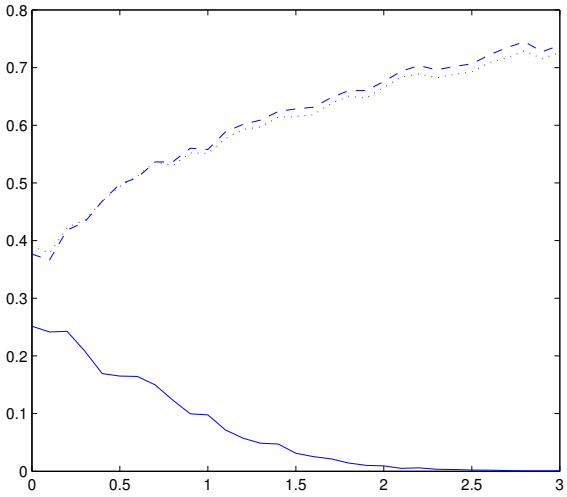

Figure 5: Sensibility of GRSIR with respect to the condition number of the covariance matrix. Horizontally: θ, vertically: $(\beta^t \hat{b})^2$. Continuous line: Ω_0 (SIR), dotted line: Ω_1 (ridge), dashed line: Ω_3 (Tikhonov).

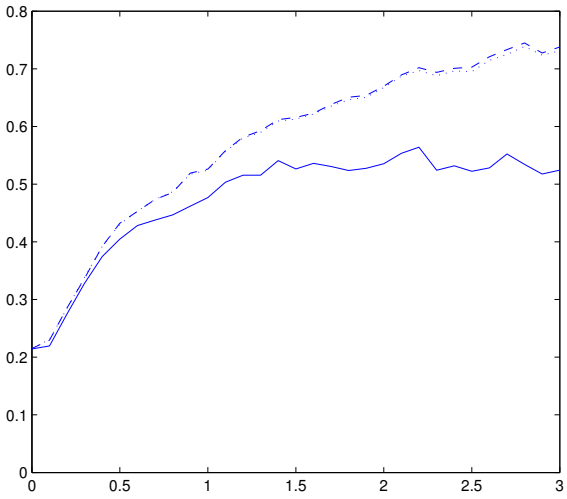

Figure 6: Sensibility of GRSIR with respect to the condition number of the covariance matrix. The cut-off dimension is chosen to $d = 20$. Horizontally: θ, vertically: $(\beta^t \hat{b})^2$. Continuous line: Ω_2 (PCA+SIR), dotted line: Ω_4 (PCA+ridge), dashed line: Ω_5 (PCA+Tikhonov).

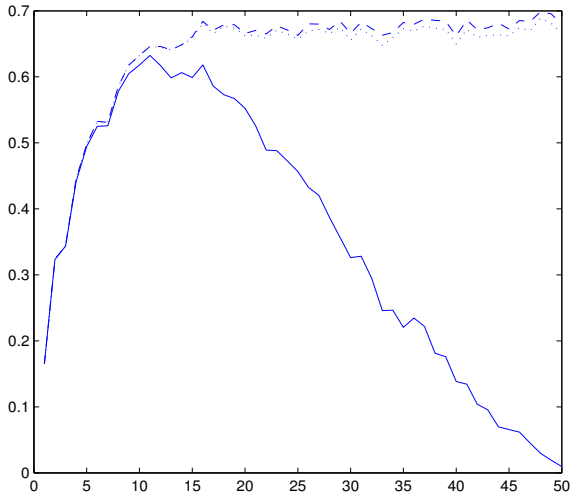

Figure 7: Sensibility of GRSIR with respect to the cut-off dimension. Horizontally: d, vertically: $(\beta^t \hat{b})^2$. Continuous line: Ω_2 (PCA+SIR), dotted line: Ω_4 (PCA+ridge), dashed line: Ω_5 (PCA+Tikhonov).