Gaussian Regularized Sliced Inverse Regression

**Caroline Bernard-Michel · Laurent Gardes ·
Stéphane Girard**

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. We propose to introduce a
Gaussian prior distribution 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 as well as an application to the
estimation of Mars surface physical properties from hyperspectral images are 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 (Draper and

C. Bernard-Michel · L. Gardes · S. Girard (✉)
Laboratoire Jean-Kuntzmann & INRIA Rhône-Alpes,
Team Mistis, Inovallée, 655, av. de l'Europe, Montbonnot,
38334 Saint-Ismier cedex, France
e-mail: Stephane.Girard@inrialpes.fr

Smith 1998) to support vector regression (Cristianini and
Shawe-Taylor 2000). 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) (Li 1991), sliced average variance estimation (Cook and Weisberg 1991), and graphical regression (Cook 1998) 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 (Gannoun and Saracco
2003; Hsing and Carroll 1992; Saracco 1997, 2005; Zhu and
Ng 1995). We also refer to Aragon and Saracco (1997) for
a study of the SIR finite sample properties, to Ferré (1998),
Schott (1994) for the estimation of $K = \dim(S_{Y|X})$, the dimension of the central subspace, and to Amato et al. (2006),
Ferré and Yao (2005) for extension to functional covariates.

Assuming that K is known, introducing $\Sigma = \text{Cov}(X)$ and
$\Gamma = \text{Cov}(\mathbb{E}(X|Y))$, in the SIR methodology, a basis of the
central subspace is obtained by computing the eigenvectors
associated to the largest K eigenvalues of $\Sigma^{-1}\Gamma$. Unfortunately, the classical n-sample estimate $\hat{\Sigma}$ of Σ can be singular, or at least ill-conditioned, in several situations. Indeed, since $\text{rank}(\hat{\Sigma}) \leq \min(n-1, p)$, 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 Chiaromonte and Martinelli (2002) and Li and Li (2004), 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 of adopting a ridge regression technique (see for instance Draper and Smith 1998, Chap. 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 (Zhong et al. 2005). 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 Scrucca (2006, 2007), regularized discriminant analysis (Friedman 1989) is adapted to the SIR framework. More recently, it is proposed in Li and Yin (2008) 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 Cook (2007) 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 (Chiaromonte and Martinelli 2002; Li and Li 2004; Scrucca 2006, 2007; Zhong et al. 2005) 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 Li and Yin (2008) 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 Sect. 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 on simulations in Sect. 4. An application to the estimation of Mars surface parameters from hyperspectral images is presented in Sect. 5. 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 $K = \dim(S_{Y|X}) = 1$, the case $1 < K < p$ being discussed in Sect. 3. We thus introduce $\beta \in \mathbb{R}^p$ such that $S_{Y|X} = \text{span}(\beta)$. In Sect. 2.1, the considered inverse regression model is presented. The estimation of the unknown parameters is discussed in Sect. 2.2 and the links with the SIR method are established in Sect. 2.3.

2.1 Single-index inverse regression model

The following inverse single-index regression model is considered (see Cook 2007, (2) for the multi-index model):

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

where μ and b are nonrandom $\mathbb{R}^p$-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 random vector located in the direction Vb. From Cook (2007), Proposition 1, b corresponds to the direction β of the central subspace. 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

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

Now, restricting ourselves to the single-index case, the forward model of SIR asserts that there exists a univariate link function g such that $\mathbb{E}(Y|X) = g(b^t X)$ or equivalently, $b^t X = g^{-1}(\mathbb{E}(Y|X))$. Thus, replacing in (2) yields

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

i.e. the coordinate function is, up to an affine transformation, 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 direction 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 $s_j(.)$, $j = 1, \dots, h$:

$$c(.) = \sum_{j=1}^{h} c_j s_j(.), \tag{3}$$

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

$$X = \mu + s^t(Y)cVb + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, V), \quad (4)$$

where $\mathcal{N}(0, V)$ is the multivariate centered Gaussian distribution with covariance matrix V. In the sequel we denote by

$$\rho = \frac{\text{var}(b^t(X - \varepsilon))}{\text{var}(b^t \varepsilon)} = \frac{b^t \Sigma b - b^t V b}{b^t V b},$$

the signal to noise ratio in the direction b. 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 the Gaussian model (4) consists in minimizing

$$G(\mu, V, b, c)$$
$$= \log \det V + \frac{1}{n} \sum_{i=1}^{n} (\mu + s^t(Y_i)cVb - X_i)^t V^{-1}$$
$$\times (\mu + s^t(Y_i)cVb - X_i), \quad (5)$$

with respect to (μ, V, b, c). Note that $G(\mu, V, b, c)$ can also be interpreted as a discrepancy functional, see (5) in Cook and Ni (2005). Up to our knowledge, the introduction of such functional in the inverse regression framework is due to Cook (2004). 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

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

Lemma 1 *Using the above notation, $G(\mu, V, b, c)$ can be rewritten as*

$$G(\mu, V, b, c) = \log \det V + \text{tr}(\hat{\Sigma} V^{-1})$$

$$+ (\mu - \bar{X} + \bar{s}^t cVb)^t V^{-1} (\mu - \bar{X} + \bar{s}^t cVb)$$
$$+ (c^t Wc)(b^t Vb) - 2c^t Mb.$$

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

Proposition 1 *Under (4), 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 $\hat{\Sigma}^{-1} M^t W^{-1} M$,*
- $\hat{c} = \frac{1}{b^t \hat{V} b} W^{-1} M \hat{b}$,
- $\hat{\mu} = \bar{X} - \bar{s}^t \hat{c} \hat{V} \hat{b}$,
- $\hat{V} = \hat{\Sigma} - \frac{\hat{\lambda}}{\hat{b}^t \hat{\Sigma} \hat{b}} \hat{\Sigma} \hat{b} \hat{b}^t \hat{\Sigma}$.

As a consequence of the above equation, one also has $\hat{\lambda} = 1 - \hat{b}^t \hat{V} \hat{b} / \hat{b}^t \hat{\Sigma} \hat{b}$, which provides an estimation of the signal to noise ratio in the direction b through

$$\hat{\rho} = \hat{\lambda}/(1 - \hat{\lambda}). \quad (6)$$

Let us now 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

$$s_j(.) = \mathbb{I}\{. \in S_j\}, \quad j = 1, \ldots, h \quad (7)$$

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

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

The covariance matrix of the regression curve $\Gamma = \text{Cov}(\mathbb{E}(X|Y))$ is then estimated by the $p \times p$ empirical "between slices" covariance matrix

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

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

Corollary 1 *Under (4) and (7), if $\hat{\Sigma}$ is regular, then the maximum likelihood estimator $\hat{b}$ of b is the eigenvector associated to the largest eigenvalue of $\hat{\Sigma}^{-1} \hat{\Gamma}$.*

To summarize, the maximum likelihood estimator of b is (up to a scale factor) the SIR estimator of the direction 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 Sect. 3.1 how the introduction of a prior information on the unknown direction 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 Sect. 3.2. Finally, three new regularizations of the SIR method based on our framework are introduced in Sect. 3.3.

3.1 Introducing a Gaussian prior

We propose to introduce a prior information on $s^t(Y)cb$ appearing in the projection of X on b under the inverse regression model. More precisely, we focus on

$$\frac{\mathbb{E}(b^t(X - \mu)|Y)b}{b^t V b} = c(Y)b = s^t(Y)cb,$$

from (2) and (3), assuming that

$$(1 + \rho)^{-1/2} (s(Y) - \bar{s})^t cb \sim \mathcal{N}(0, \Omega). \qquad (8)$$

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. The scalar $(1 + \rho)^{-1/2}$ is introduced for normalization purposes, permitting to preserve the interpretation (6) of the eigenvalue in terms of signal to noise ratio. As a comparison, in Antoniadis et al. (2004), a Bayesian estimation method is proposed using a B-splines approximation of the link function g in the forward model and a Fisher-von Mises prior on the direction b. In our approach, working on the inverse regression model allows to obtain explicit solutions, see Lemma 2 and Proposition 2 below.

Lemma 2 *Maximum likelihood estimators are obtained by minimizing*

$$G_\Omega(\mu, V, b, c) = G(\mu, V, b, c)$$
$$+ \frac{(b^t \Omega^{-1} b)(b^t V b)(c^t W c)}{b^t \Sigma b} \qquad (9)$$

with respect to (μ, V, b, c).

Comparing to (5), the additional term due to the prior information can be read as a regularization term in Tikhonov theory, see for instance Vogel (2002), Chap. 1, penalizing large projections. The following result can be stated:

Proposition 2 *Under* (4) *and* (8), *if* W *and* $\Omega\hat{\Sigma} + I_p$ *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* $(\Omega\hat{\Sigma} + I_p)^{-1}\Omega M^t W^{-1} M$,
- $\hat{c} = \frac{1}{(1+\eta(\hat{b}))\hat{b}^t \hat{V} \hat{b}} W^{-1} M \hat{b}$, *where* $\eta(\hat{b}) = \frac{\hat{b}^t \Omega^{-1} \hat{b}}{\hat{b}^t \hat{\Sigma} \hat{b}}$,
- $\hat{\mu} = \bar{X} - \bar{s}^t \hat{c} \hat{V} \hat{b}$,
- $\hat{V} = \hat{\Sigma} - \frac{\hat{\lambda}}{\hat{b}^t \hat{\Sigma} \hat{b}} \hat{\Sigma} \hat{b} \hat{b}^t \hat{\Sigma}$.

The inversion of $\hat{\Sigma}$ in Proposition 1 is replaced by the inversion of $\Omega\hat{\Sigma} + I_p$ in Proposition 2. Thus, for a properly chosen prior matrix Ω, the numerical instabilities in the estimation of b disappear. Note that, since the estimation of V is formally the same as in Proposition 1, the interpretation (6) of $\hat{\lambda}$ still holds. As previously, this result can be applied to the particular case of the SIR method.

Corollary 2 *Under* (4), (7) *and* (8), *if* $\Omega\hat{\Sigma} + I_p$ *is regular, then the maximum likelihood estimator of* b *is the eigenvector associated to the largest eigenvalue of* $(\Omega\hat{\Sigma} + I_p)^{-1}\Omega\hat{\Gamma}$.

In the following, the above estimator of the direction b will be referred to as the Gaussian Regularized Sliced Inverse Regression (GRSIR) estimator. Let us emphasize that the GRSIR estimator can be extended to the multi-index situation by considering the eigenvectors $\hat{b}_1, \ldots, \hat{b}_K$ associated to the K largest eigenvalues $\hat{\lambda}_1 > \cdots > \hat{\lambda}_K$ of $(\Omega\hat{\Sigma} + I_p)^{-1}\Omega\hat{\Gamma}$. For instance, one can show that $\hat{b}_2$ maximizes G_Ω under the orthogonality constraint $\hat{b}_1^t(\hat{\Sigma} + \Omega^{-1})\hat{b}_2 = 0$.

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 (9). 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 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 $(\hat{\Sigma} + \tau I_p)^{-1}\hat{\Gamma}$ associated to its largest eigenvalue, which is the ridge estimator introduced independently in Zhong et al. (2005) and Scrucca (2006, 2007).

PCA + SIR approach   As already seen, a popular technique to overcome the singularity problems of $\hat{\Sigma}$ is to use PCA as a preprocessing step (Chiaromonte and Martinelli 2002; Li and Li 2004). The principle is the following. Let $d \in \{1, \ldots, p\}$ be fixed and denote by $\hat{\delta}_1 \geq \cdots \geq \hat{\delta}_d$ the d largest eigenvalues of $\hat{\Sigma}$ (supposed to be positive), $\hat{q}_1, \ldots, \hat{q}_d$ the associated eigenvectors and $S_d = \text{span}(\hat{q}_1, \ldots, \hat{q}_d)$ the linear subspace spanned by $\hat{q}_1, \ldots, \hat{q}_d$. The first step consists of 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*

$$\Omega_2 = \frac{1}{\tau} \sum_{j=1}^{d} \frac{1}{\hat{\delta}_j} \hat{q}_j \hat{q}_j^t,$$

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.

L_2-regularization   The regularization proposed in Li and Yin (2008) consists of 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 (Bernard-Michel et al. 2008, Proposition 1) that the only possible solution $\hat{b}$ of the minimization problem is $\hat{b} = 0$. In view of this result, the proposed alternating least squares algorithm (Li and Yin 2008, Sect. 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 numbers $t \neq 0$. 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 direction b is the eigenvector of $(\hat{\Sigma}^2 + \tau I_p)^{-1} \hat{\Sigma} \hat{\Gamma}$ associated to its largest eigenvalue. In the following, this estimator will be referred to as the Tikhonov estimator. Indeed, let us recall that SIR estimator is obtained by a spectral decomposition of $\hat{\Sigma}^{-1} \hat{\Gamma}$. For all $k = 1, \ldots, p$, denote by x_k the k-th column of this matrix. Computing x_k is equivalent to solve with respect to x the linear system $\hat{\Sigma} x = \hat{\Gamma}_k$ where $\hat{\Gamma}_k$ is the k-th column of $\hat{\Gamma}$. The associated Tikhonov minimization problem (see (1.34) in Vogel 2002) can be written as

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

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

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

Proposition 4 *For all real function φ let*

$$\Omega(\varphi) = \sum_{j=1}^{d} \varphi(\hat{\delta}_j) \hat{q}_j \hat{q}_j^t.$$

Then, the associated GRSIR estimator can be obtained by first projecting the predictors on $S_d = \text{span}(\hat{q}_1, \ldots, \hat{q}_d)$ and second performing GRSIR on the projected predictors with prior covariance matrix

$$\tilde{\Omega}(\varphi) = \sum_{j=1}^{p} \varphi(\hat{\delta}_j) \hat{q}_j \hat{q}_j^t.$$

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

- Choosing $\varphi(t) = 1/(\tau t)$, we obtain $\Omega(1/(\tau \, \text{Id})) = \Omega_2$ and $\tilde{\Omega}(1/(\tau \, \text{Id})) = (\tau \hat{\Sigma})^{-1} = \Omega_0$, where Id is the identity function. It appears that Proposition 3 is a particular case of Proposition 4. 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

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

and $\tilde{\Omega}(1/\tau) = I_p/\tau = \Omega_1$. Consequently, this new method consists of applying the ridge approach (Zhong et al. 2005) 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

$$\Omega_5 \stackrel{def}{=} \Omega(\text{Id}/\tau) = \frac{1}{\tau} \sum_{j=1}^{d} \hat{\delta}_j \hat{q}_j \hat{q}_j^t,$$

and $\tilde{\Omega}(\text{Id}/\tau) = \hat{\Sigma}/\tau = \Omega_3$. This new method consists of 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.

4 Simulation study

4.1 Experimental setup

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 data simulated from a random pair (X, Y) where $X \in \mathbb{R}^p$ with $p = 50$ and $Y \in \mathbb{R}$. The random vector X is Gaussian, centered, with covariance matrix $\Sigma = Q\Delta Q^t$ where Δ is the diagonal matrix containing the eigenvalues of Σ defined by $\delta_j = (p + 1 - j)^\theta$, $j = 1, \ldots, p$ and Q is a matrix drawn from the uniform distribution on the set of orthogonal matrices. Following (Heiberger 1978), the orthogonal matrix Q is obtained by performing a QR-decomposition on a $p \times p$ matrix whose coefficients are independent standard Gaussian random variables. 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 defined by the forward single-index model $Y = g(\beta^t X/\sigma_X) + \xi$ where σ_X is the standard deviation of the projection of X on β i.e. $\sigma_X = (\beta^t \Sigma \beta)^{1/2}$ and ξ is a centered Gaussian random variable independent of X with standard deviation $\sigma_\xi = 0.03$. Two link functions are considered: $g_1(t) = \sin(\pi t/2)$ and $g_2(t) = |t - 1/2|$. The true index $\beta = 5^{-1/2} Q(1, 1, 1, 1, 1, 0, \ldots, 0)^t$ is located on the subspace spanned by $\{q_1, \ldots, q_5\}$ and representing about 27% of the total variance when $\theta = 2$. In our experiments, $N = 100$ samples of size $n = 100$ are considered. More precisely, for all $i = 1, \ldots, n$ and $r = 1, \ldots, N$, the pairs $(X_i, Y_i^{(r)})$ are simulated following the model $Y_i^{(r)} = g(\beta^t X_i/\sigma_X) + \xi_r$, with $g \in \{g_1, g_2\}$. This yields N samples, with common predictors but different noises, as well as N estimators $\hat{b}^{(r)}$, $r = 1, \ldots, N$ of β. Two criteria are used to assess the estimates accuracy: The Proximity Criterion (PC) and the Stability Criterion (SC) defined as

$$\text{PC} = \frac{1}{N} \sum_{r=1}^{N} (\beta^t \hat{b}^{(r)})^2 \quad \text{and}$$

$$\text{SC} = \frac{1}{N(N-1)} \sum_{r=1}^{N} \sum_{s \neq r} ((\hat{b}^{(s)})^t \hat{b}^{(r)})^2.$$

Both criteria take their values in [0, 1] and rely on the computation of the squared cosine between two axes. We refer to Ferré (1998) for an adaptation to multi-index models. PC is a proximity measure between the estimator $\hat{b}$ and the true direction β: A value close to 0 implies a low proximity ($\hat{b}$ is nearly orthogonal to β) whereas a value close to 1 implies a high proximity ($\hat{b}$ is approximatively collinear with β). SC is a measure of the estimator stability: A value close to 0 implies a low stability (two realizations of $\hat{b}$ are approximatively orthogonal) whereas a value close to 1 implies a high stability (two realizations of $\hat{b}$ are approximatively collinear).

4.2 Results

In the sequel, the number of slices is fixed to $h + 1 = 10$.

Influence of the regularization parameter The two previously defined criteria are computed as functions of the regularization parameter τ. A logarithmic scale is adopted, 150 values of $\log(\tau)$ regularly distributed in $[-5, 25]$ were considered. Here, we limit ourselves to $\theta = 2$. Moreover, we choose $d = 20$ in the PCA + SIR, PCA + ridge and PCA + Tikhonov methods. Results obtained with link functions g_1 and g_2 are respectively displayed on Figs. 1, 2 and Figs. 3, 4. It appears that, for all the considered methods, the shape of the link function does not influence much the results. For both link functions, SIR gives poor results, especially in terms of proximity. Ridge and Tikhonov regularizations can bring a significant improvement provided τ is large enough. PCA + SIR obtains reasonable results compared to SIR, with the advantage of not requiring the selection of τ. The selection of d is addressed in our third experiment. Note that PCA + ridge and PCA + Tikhonov methods are less sensitive to the choice of τ than ridge and Tikhonov methods. Besides, they both outperform PCA + SIR for sufficiently large values of the regularization parameter. Unsurprisingly, the stability of GRSIR methods (ridge, Tikhonov, PCA + ridge and PCA + Tikhonov) increases with the regularization parameter.

For one of the $N = 100$ simulated datasets, the pairs $(\beta^t X_i, \hat{b}^t X_i)$, $i = 1, \ldots, n$ are plotted on Fig. 5 for the link function g_1 and on Fig. 6 for the link function g_2. Two estimators are considered for $\hat{b}$: SIR and GRSIR with PCA + Tikhonov prior. GRSIR estimator is computed with the optimal regularization parameter, *i.e.* the regularization parameter maximizing PC. The horizontal shape of the SIR cloud of points indicates that the estimated projection $\hat{b}^t X$ is almost independent of the true one $\beta^t X$. Conversely, the GRSIR points are concentrated along the first diagonal, indicating a high correlation between $\hat{b}^t X$ and $\beta^t X$.

Fig. 1 Influence of the regularization parameter. The link function is g_1 and the condition number is fixed ($\theta = 2$). *Horizontally*: $\log(\tau)$, *vertically*: (**a**) PC and (**b**) SC. *Continuous line*: Ω_0 (SIR), $-\circ-$: Ω_1 (ridge), *dashed line*: Ω_3 (Tikhonov)
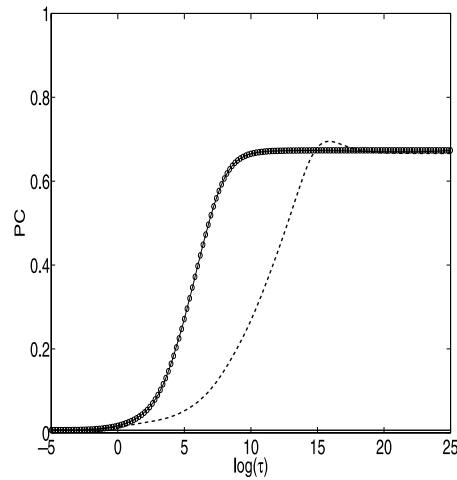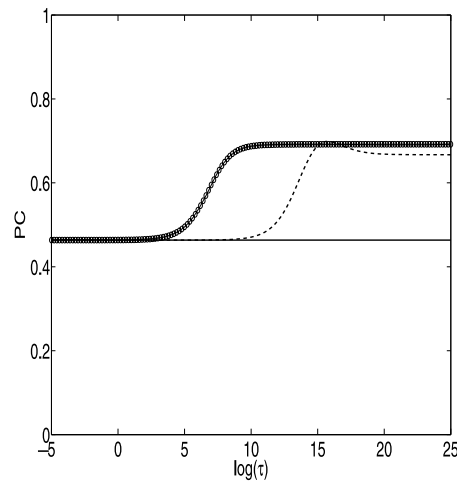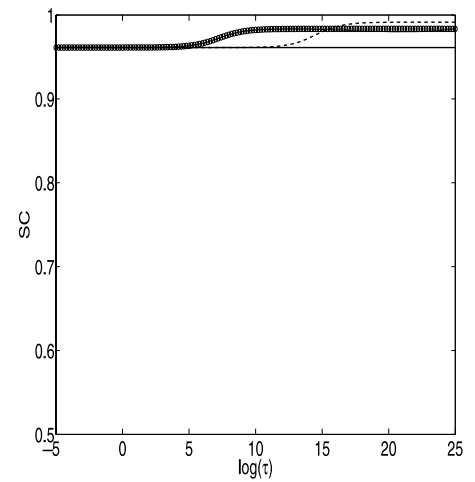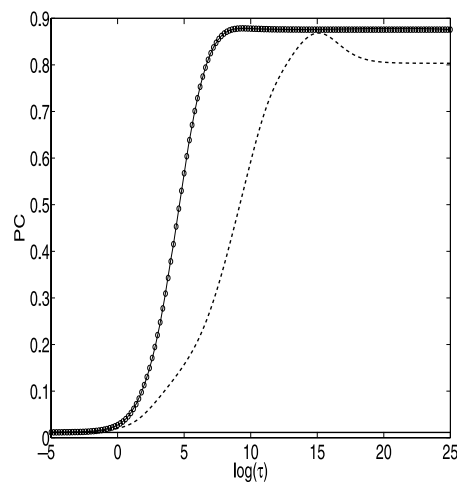

(a) PC as a function of τ
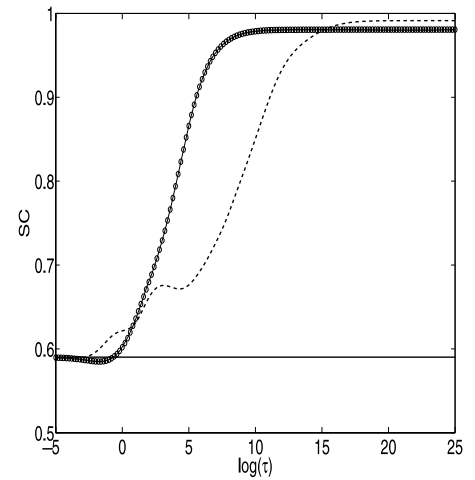
(b) SC as a function of τ

Fig. 2 Influence of the regularization parameter. The link function is g_1, the "cut-off" dimension and the condition number are fixed ($d = 20$ and $\theta = 2$). *Horizontally*: $\log(\tau)$, *vertically*: (**a**) PC and (**b**) SC. *Continuous line*: Ω_2 (PCA + SIR), $-\circ-$: Ω_4 (PCA + ridge), *dashed line*: Ω_5 (PCA+Tikhonov)
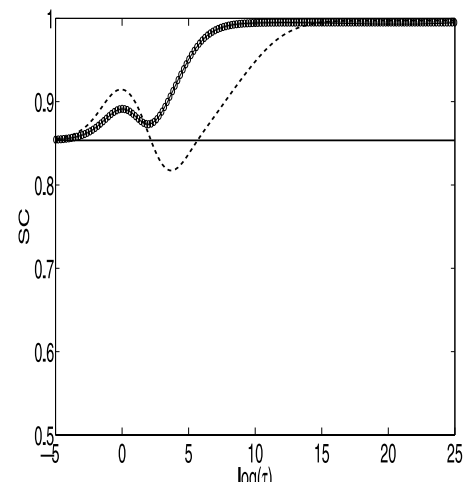

(a) PC as a function of τ

(b) SC as a function of τ

Fig. 3 Influence of the regularization parameter. The link function is g_2 and the condition number is fixed ($\theta = 2$). *Horizontally*: $\log(\tau)$, *vertically*: (**a**) PC and (**b**) SC. *Continuous line*: Ω_0 (SIR), $-\circ-$: Ω_1 (ridge), *dashed line*: Ω_3 (Tikhonov)
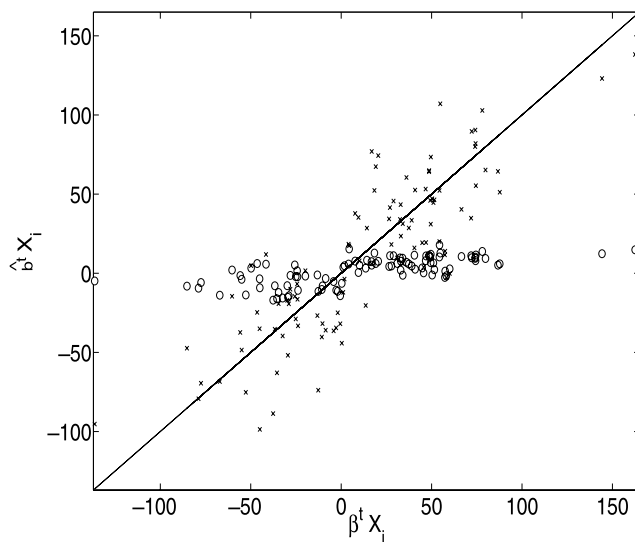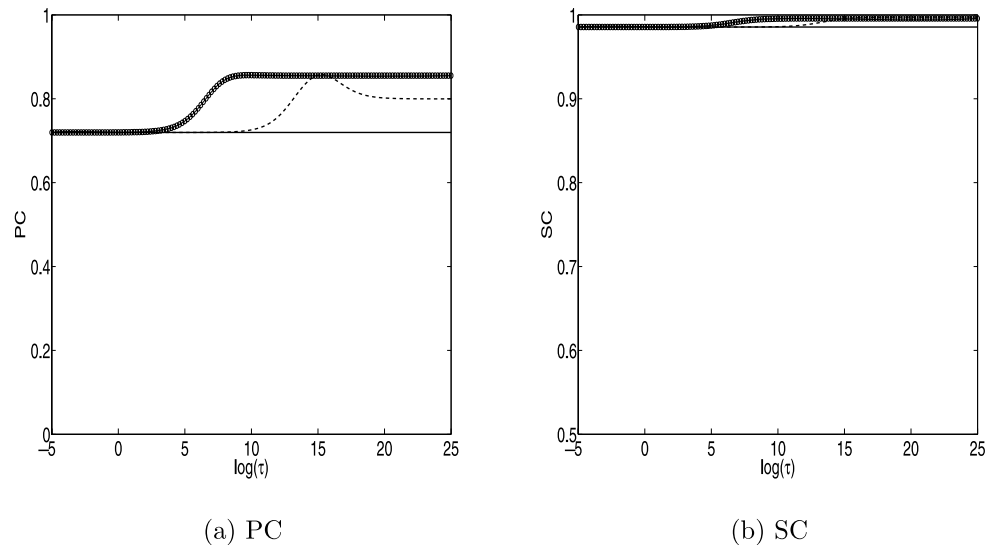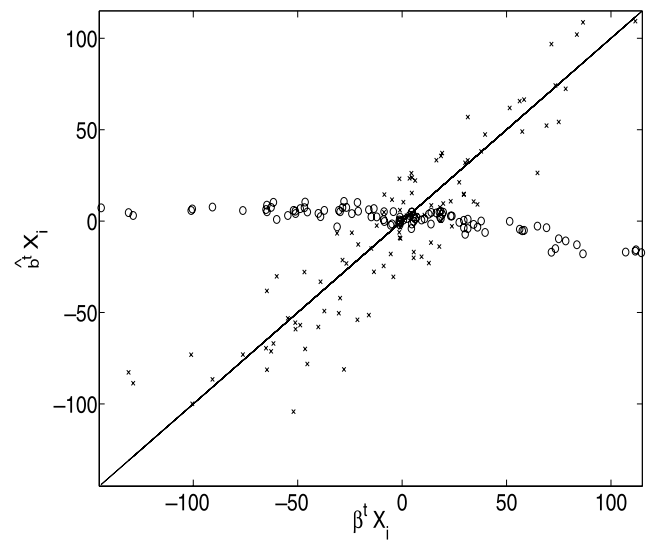

(a) PC

(b) SC

Fig. 4 Influence of the regularization parameter. The link function is g_2, the "cut-off" dimension and the condition number are fixed ($d = 20$ and $\theta = 2$). *Horizontally*: $\log(\tau)$, *vertically*: (**a**) PC and (**b**) SC. *Continuous line*: Ω_2 (PCA + SIR), $- \circ -$: Ω_4 (PCA + ridge), *dashed line*: Ω_5 (PCA + Tikhonov)

(a) PC

(b) SC

Fig. 5 Comparison of pairs $(\beta^t X_i, \hat{b}^t X_i)$, $i = 1, \ldots, n$ obtained with SIR ($\circ$) and GRSIR, PCA + Tikhonov prior ($\times$). The link function is g_1 and the optimal regularization parameter is used. The "cut-off" dimension and the condition number are fixed ($d = 20$ and $\theta = 2$)

Fig. 6 Comparison of pairs $(\beta^t X_i, \hat{b}^t X_i)$, $i = 1, \ldots, n$ obtained with SIR ($\circ$) and GRSIR, PCA + Tikhonov prior ($\times$). The link function is g_2 and the optimal regularization parameter is used. The "cut-off" dimension and the condition number are fixed ($d = 20$ and $\theta = 2$)

Influence of the condition number 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 is selected for each method and the corresponding PC is displayed on Figs. 7, 8 in case of link function g_1. Similar results have been observed with g_2. Clearly, SIR 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.
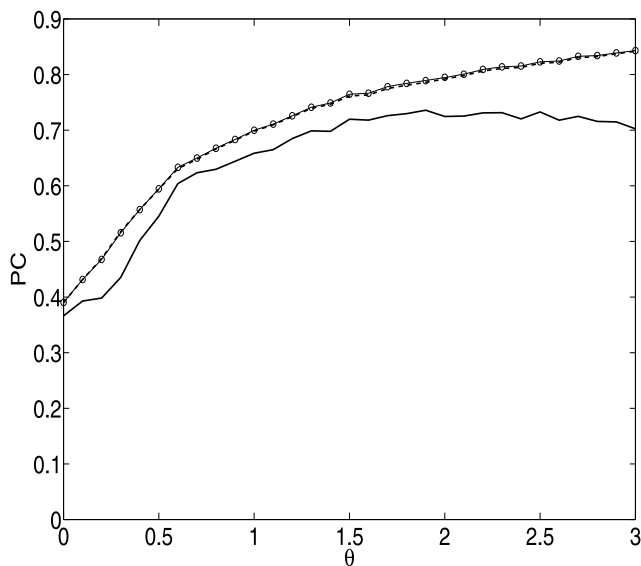
Influence of the "cut-off" dimension This experiment is dedicated to the illustration of the role of d in PCA + SIR,

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 is selected for each method and the corresponding PC is displayed on Fig. 9 in case of link function g_1. Similar results have been observed with g_2. 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.
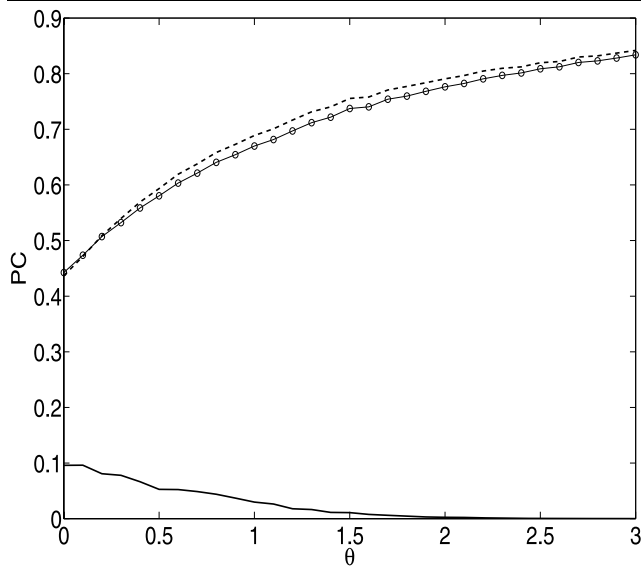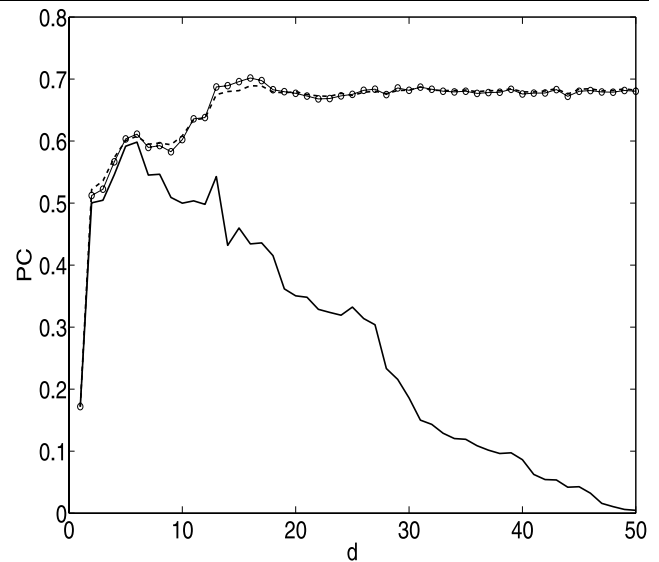
Fig. 7 Sensitivity of GRSIR with respect to the condition number of the covariance matrix. The link function is g_1 and the optimal regularization parameter is used for each value of θ. *Horizontally*: θ, *vertically*: PC. *Continuous line*: Ω_0 (SIR), $- \circ -$: Ω_1 (ridge), *dashed line*: Ω_3 (Tikhonov)

Fig. 9 Sensitivity of GRSIR with respect to the "cut-off" dimension. The link function is g_1, the condition number is fixed ($\theta = 2$) and the optimal regularization parameter is used for each value of d. *Horizontally*: d, *vertically*: PC. *Continuous line*: Ω_2 (PCA + SIR), $- \circ -$: Ω_4 (PCA + ridge), *dashed line*: Ω_5 (PCA + Tikhonov)
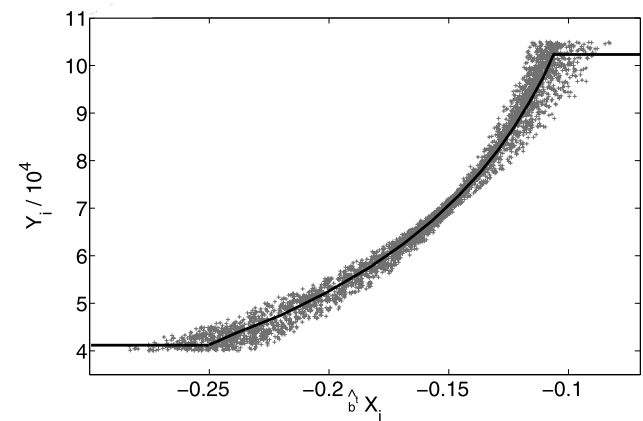

Fig. 8 Sensitivity of GRSIR with respect to the condition number of the covariance matrix. The link function is g_1, the "cut-off" dimension is fixed to $d = 20$ and the optimal regularization parameter is used for each value of θ. *Horizontally*: θ, *vertically*: PC. *Continuous line*: Ω_2 (PCA + SIR), $- \circ -$: Ω_4 (PCA + ridge), *dashed line*: Ω_5 (PCA + Tikhonov)

Fig. 10 Estimated functional relationship using piecewise linear regression between reduced spectra $\hat{\beta}^t X$ on the first GRSIR (PCA + ridge prior) direction and Y, the grain size of CO_2. *Horizontally*: reduced spectra from the learning database on the first GRSIR direction. *Vertically*: Grain size of CO_2

5 Retrieval of Mars surface physical properties from hyperspectral images

We propose to compare SIR and GRSIR on a nonlinear inverse problem in remote sensing. Hyperspectral remote sensing is a promising space technology regularly selected by agencies with regard to the observation of planets. It allows to collect, for each pixel of a scene, the intensity of light energy reflected from materials as it varies across different wavelengths. Hundreds of spectels in the visible and near infra-red are recorded for each image cell. The analysis of these spectral signatures allows for the identification of the physical, chemical or mineralogical properties of the surface that may help to understand the geological history of planets. Our goal is to evaluate the physical properties of surface materials on the planet Mars from hyperspectral images collected by the OMEGA instrument onboard the Mars

express spacecraft. Our approach is based on the estimation of the functional relationship G between some physical parameters Y and observed spectra X. For this purpose, a database of synthetic spectra is generated by a physical radiative transfer model and used to estimate G. The high dimension of spectra ($p = 184$ wavelengths) is reduced with SIR and GRSIR.

5.1 Data

We focus on an observation of the south pole of Mars at the end of summer, collected during orbit 61 by the French imaging spectrometer OMEGA onboard the Mars Express Mission. A detailed analysis of this image (Douté et al. 2007) revealed that this portion of Mars mainly contains water ice, carbon dioxide and dust. This has led to the physical modeling of individual spectra with a surface reflectance model. This model allows, to generate $n = 12,000$ synthetic spectra with the corresponding parameters that constitute a learning database. Here, we study the terrain unit of strong CO_2 concentration determined by a classification method based on wavelets (Schmidt et al. 2007). It contains approximately 9,000 spectra to analyze. The most important parameters characterizing the morphology of these spectra are the proportions of CO_2 ice, water ice and dust as well as the grain sizes of water ice and CO_2 ice. In the sequel only one parameter, the grain size of CO_2 ice is presented. A detailed analysis for the four other parameters can be found in Bernard-Michel et al. (2007).

5.2 Methodology

Two methods are compared in order to reverse the hyperspectral image: SIR and GRSIR with the PCA + ridge prior. Both methods aim at finding a lower dimensional predictor space, sufficient to describe the relationship of interest. In practice, it appears that a unidimensional predictor space gives satisfactory results (Bernard-Michel et al. 2007) leading to the single-index model $G(X) = g(\beta^t X)$. As a consequence, the estimation of the relationship G reduces to compute one direction β, the univariate function g being estimated by piecewise linear regression, see (Bernard-Michel et al. 2007) for further details. The "cut-off" dimension is fixed to $d = 20$ in GRSIR. The regularization parameter, fixed to $\tau = 10^{-4.2}$, is chosen to minimize the mean squared error when estimating the grain size of CO_2 ice on the learning database itself. The estimated relationship between reduced spectra $\beta^t X$ and Y, the grain size of CO_2, is presented on Fig. 10. It appears that the grain size of CO_2 can be accurately modeled by a nonlinear function applied to a linear combination of the 184 wavelengths.

Fig. 11 Grain size of CO_2 ice estimated by piecewise linear regression from the hyperspectral image observed on Mars during orbit 61 when the dimension of the predictor space is reduced by SIR
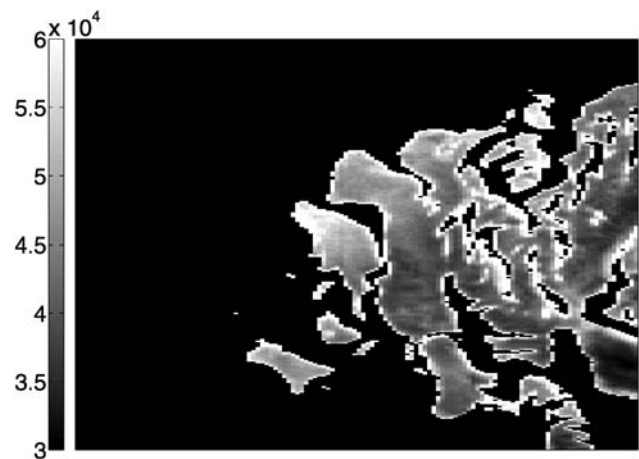
Fig. 12 Grain size of CO_2 ice estimated by piecewise linear regression from the hyperspectral image observed on Mars during orbit 61 when the dimension of the predictor space is reduced by GRSIR (PCA + ridge prior)

5.3 Results

The inversion of the image from orbit 61 using GRSIR shows a smooth mapping of the grain size of CO_2 (see Fig. 12) making it possible to distinguish some areas with large grain sizes of CO_2 ice on the boundaries and some areas with small values inside the cap. On the opposite, SIR estimates assume a small number of values that seem to be randomly distributed (see Fig. 11) and correspond to the extremal values of the parameter in the learning database. These poor results can be explained by the high condition number, about 10^{14}, of the empirical covariance matrix. Since no ground data is available, it is quite difficult to quantify the accuracy of GRSIR estimates. However, comparisons with other approaches or with estimates from other hyperspectral images of the same portion of Mars (Bernard-

Michel et al. 2007) give consistent results. GRSIR then appears promising for model inversion in remote sensing.

6 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 and ridge methods. The use of a "cutoff" dimension d serves to limit the sensitivity 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 $d \simeq p/2$ appears as a good heuristic 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 to obtain L_1-penalizations and thus sparse estimates of β.

Appendix: Proofs

Proof of Lemma 1 Let us remark that

$$R \overset{def}{=} G(\mu, V, b, c) - \log \det V = \frac{1}{n}\sum_{i=1}^{n} Z_i^t V^{-1} Z_i, \quad (10)$$

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

$$
\begin{aligned}
Z_i &= \mu + s^t(Y_i)cVb - X_i \\
&= (\mu - \bar{X} + \bar{s}^t cVb) + (s(Y_i) - \bar{s})^t cVb - (X_i - \bar{X}) \\
&\overset{def}{=} Z_1 + Z_{2,i} - Z_{3,i}.
\end{aligned}
$$

Since $Z_{2,\cdot}$ and $Z_{3,\cdot}$ are centered, replacing the previous expansion in (10) yields

$$
\begin{aligned}
R &= Z_1^t V^{-1} Z_1 + \frac{1}{n}\sum_{i=1}^{n} Z_{2,i}^t V^{-1} Z_{2,i} + \frac{1}{n}\sum_{i=1}^{n} Z_{3,i}^t V^{-1} Z_{3,i} \\
&\quad - \frac{2}{n}\sum_{i=1}^{n} Z_{2,i}^t V^{-1} Z_{3,i},
\end{aligned}
$$

where

$$Z_1^t V^{-1} Z_1 = (\mu - \bar{X} + \bar{s}^t cVb)^t V^{-1}(\mu - \bar{X} + \bar{s}^t cVb)$$

$$\frac{1}{n}\sum_{i=1}^{n} Z_{2,i}^t V^{-1} Z_{2,i} = (c^t W c)(b^t V b)$$

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} Z_{3,i}^t V^{-1} Z_{3,i} &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}((X_i - \bar{X})^t V^{-1}(X_i - \bar{X})) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}(V^{-1}(X_i - \bar{X})(X_i - \bar{X})^t) \\
&= \mathrm{tr}(V^{-1}\hat{\Sigma})
\end{aligned}
$$

$$\frac{1}{n}\sum_{i=1}^{n} Z_{2,i}^t V^{-1} Z_{3,i} = c^t M b,$$

and the conclusion follows.   □

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

$$\frac{1}{2}\nabla_\mu G = \hat{V}^{-1}(\hat{\mu} - \bar{X} + \bar{s}^t \hat{c}\hat{V}\hat{b}) = 0, \quad (11)$$

$$\frac{1}{2}\nabla_b G = \left((\hat{c}^t \bar{s})^2 + \hat{c}^t W \hat{c}\right)\hat{V}\hat{b} - M^t \hat{c} + (\hat{c}^t \bar{s})(\hat{\mu} - \bar{X}) = 0, \quad (12)$$

$$\frac{1}{2}\nabla_c G = (\hat{b}^t \hat{V}\hat{b})(\bar{s}\bar{s}^t + W)\hat{c} - M\hat{b} + (\hat{\mu} - \bar{X})^t \hat{b}\bar{s} = 0, \quad (13)$$

$$
\begin{aligned}
\nabla_V G = \hat{V}^{-1} &+ \hat{b}\hat{b}^t \left((\hat{c}^t \bar{s})^2 + \hat{c}^t W \hat{c}\right) \\
&- \hat{V}^{-1}\left((\hat{\mu} - \bar{X})(\hat{\mu} - \bar{X})^t + \hat{\Sigma}\right)\hat{V}^{-1} = 0. \quad (14)
\end{aligned}
$$

From (11), we have

$$\hat{\mu} = \bar{X} - \bar{s}^t \hat{c}\hat{V}\hat{b}. \quad (15)$$

Replacing in (12) and (13) yields the simplified system of equations

$$(\hat{c}^t W \hat{c})\hat{V}\hat{b} = M^t \hat{c}, \quad (16)$$

$$(\hat{b}^t \hat{V}\hat{b})W\hat{c} = M\hat{b}. \quad (17)$$

Assuming W is regular, (17) entails $\hat{c} = W^{-1}M\hat{b}/\hat{b}^t \hat{V}\hat{b}$ and replacing in (16) yields

$$(\hat{c}^t W \hat{c})(\hat{b}^t \hat{V}\hat{b})\hat{V}\hat{b} = M^t W^{-1} M\hat{b}. \quad (18)$$

Now, multiplying (14) on the left and on the right by $\hat{V}$ and taking account of (15) entail

$$\hat{\Sigma} = \hat{V} + (\hat{c}^t W \hat{c})\hat{V}\hat{b}\hat{b}^t \hat{V}. \quad (19)$$

As a consequence of (19), we have

$$\hat{\Sigma}\hat{b} = \left(1 + (\hat{c}^t W \hat{c})(\hat{b}^t \hat{V} \hat{b})\right)\hat{V}\hat{b}, \tag{20}$$

which means that $\hat{\Sigma}\hat{b}$ is proportional to $\hat{V}\hat{b}$. Consequently,

$$\hat{V}\hat{b} = \theta(\hat{b})\hat{\Sigma}\hat{b}, \tag{21}$$

where we have defined $\theta(\hat{b}) = \hat{b}^t \hat{V}\hat{b}/\hat{b}^t \hat{\Sigma}\hat{b}$. Substituting (21) in (18) yields

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

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

$$\hat{\lambda} = (\hat{c}^t W \hat{c})(\hat{b}^t \hat{V}\hat{b})\theta(\hat{b}). \tag{22}$$

Putting together (20) and (21) yields $1 + (\hat{c}^t W \hat{c})(\hat{b}^t \hat{V}\hat{b}) = 1/\theta(\hat{b})$ and thus the eigenvalue can be rewritten as $\hat{\lambda} = 1 - \theta(\hat{b})$. Moreover, we have, from (16),

$$\hat{c}^t M \hat{b} = \hat{\lambda}/\theta(\hat{b}), \tag{23}$$

$$\text{tr}(\hat{\Sigma}\hat{V}^{-1}) = p + \hat{\lambda}/\theta(\hat{b}), \tag{24}$$

$$\log \det \hat{V} = \log \det \hat{\Sigma} - \log \det\left(I_p + (\hat{c}^t W \hat{c})\hat{V}\hat{b}\hat{b}^t\right)$$

$$= \log \det \hat{\Sigma} - \log\left(1 + \hat{\lambda}/\theta(\hat{b})\right), \tag{25}$$

entailing

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

$$= p + \log \det \hat{\Sigma} + \log(1 - \hat{\lambda}).$$

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

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

we get

$$V_0 + (\hat{c}^t W \hat{c}) V_0 \hat{b}\hat{b}^t V_0$$

$$= \hat{\Sigma} + \left((\hat{c}^t W \hat{c})\theta^2(\hat{b}) - \frac{\hat{\lambda}}{\hat{b}^t \hat{\Sigma}\hat{b}}\right)\hat{\Sigma}\hat{b}\hat{b}^t \hat{\Sigma}$$

$$= \hat{\Sigma} + \frac{\hat{\lambda}\theta(\hat{b})}{\hat{b}^t V_0 \hat{b}}\left((\hat{c}^t W \hat{c})(\hat{b}^t V_0 \hat{b})\theta(\hat{b}) - \hat{\lambda}\right)\hat{\Sigma}\hat{b}\hat{b}^t \hat{\Sigma} = \hat{\Sigma},$$

in view of (22) and thus V_0 verifies (19). $\qquad\square$

Proof of Corollary 1 Let us remark that, under (7), the coefficients W_{ij} of W are given by $W_{ij} = f_i \mathbb{I}\{i = j\} - f_i f_j$ for all $(i, j) \in \{1, \ldots, h\}^2$. The inverse matrix of W is

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

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

$$M^t W^{-1} M = \sum_{j=1}^{h} f_j(\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^t + \frac{1}{f_{h+1}} M^t U M$$

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

$$+ \frac{1}{h f_{h+1}}(M^t U)(M^t U)^t. \tag{26}$$

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

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

$$= -f_{h+1}(\bar{X}_{h+1} - \bar{X}),$$

it 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 (26) yields

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

The result is then a consequence of Proposition 1. $\qquad\square$

Proof of Lemma 2 The joint distribution of (X, b) given Y denoted by $p(X, b|Y)$ is calculated as the product $p(X|Y, b) p(b|Y)$ where $p(X|Y, b)$ is given by (4) and $p(b|Y)$ is given by (8). The estimators are obtained by minimizing

$$J_\Omega(\mu, V, b, c)$$

$$= -\frac{2}{n} \sum_{i=1}^{n} \log p(X_i, b|Y_i)$$

$$= -\frac{2}{n} \sum_{i=1}^{n} \log p(X_i|Y_i, b) - \frac{2}{n} \sum_{i=1}^{n} \log p(b|Y_i)$$

$$= G(\mu, V, b, c) + \frac{b^t \Omega^{-1} b}{1 + \rho} \frac{1}{n} \sum_{i=1}^{n} ((s(Y_i) - \bar{s})^t c)^2 + C^{ste}$$

$$= G(\mu, V, b, c) + \frac{b^t V b}{b^t \Sigma b}(b^t \Omega^{-1} b)(c^t W c) + C^{ste},$$

which is $G_\Omega(\mu, V, b, c)$ up to the constant C^{ste}. $\qquad\square$

Proof of Proposition 2 The proof is similar to the one of Proposition 1. First, remark that (11) still holds and thus $\hat{\mu} =$

$\bar{X} - \bar{s}^t \hat{c} \hat{V} \hat{b}$. Let us recall the following definitions

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

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

$$(\hat{c}^t W \hat{c})\big(\Omega \hat{V} \hat{b} + \theta(\hat{b})\hat{b} + \eta(\hat{b})\Omega(\hat{V}\hat{b} - \theta(\hat{b})\hat{\Sigma}\hat{b})\big)$$
$$= \Omega M^t \hat{c}, \qquad (27)$$

$$(\hat{b}^t \hat{V} \hat{b})(1 + \eta(\hat{b}))W\hat{c} = M\hat{b}, \qquad (28)$$

$$\hat{V}^{-1} + \hat{b}\hat{b}^t(\hat{c}^t W \hat{c})(1 + \eta(\hat{b})) = \hat{V}^{-1}\hat{\Sigma}\hat{V}^{-1}. \qquad (29)$$

Assuming W is regular, (28) entails $\hat{c} = W^{-1}M\hat{b}/((1 + \eta(\hat{b}))(\hat{b}^t \hat{V}\hat{b}))$ and replacing in (27) yields

$$(\hat{c}^t W \hat{c})(\hat{b}^t \hat{V}\hat{b})(1 + \eta(\hat{b}))$$
$$\times \big(\Omega \hat{V}\hat{b} + \theta(\hat{b})\hat{b} + \eta(\hat{b})\Omega(\hat{V}\hat{b} - \theta(\hat{b})\hat{\Sigma}\hat{b})\big)$$
$$= \Omega M^t W^{-1}M\hat{b}. \qquad (30)$$

Now, multiplying (29) on the left and on the right by $\hat{V}$, it follows that

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

leading to

$$\hat{\Sigma}\hat{b} = \big(1 + (\hat{c}^t W \hat{c})(\hat{b}^t \hat{V}\hat{b})(1 + \eta(\hat{b}))\big)\hat{V}\hat{b}, \qquad (31)$$

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

$$\hat{V}\hat{b} = \theta(\hat{b})\hat{\Sigma}\hat{b}, \qquad (32)$$

and replacing in (30) yields

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

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

$$\hat{\lambda} = (\hat{c}^t W \hat{c})(\hat{b}^t \hat{V}\hat{b})(1 + \eta(\hat{b}))\theta(\hat{b}). \qquad (33)$$

Collecting (31), (32) and (33) yields $\hat{\lambda} = 1 - \theta(\hat{b})$. Now, let us remark that (23), (24) and (25) still hold in this context entailing

$$G_\Omega(\hat{\mu}, \hat{V}, \hat{b}, \hat{c}) = p + \log\det\hat{\Sigma} + \log(1 - \hat{\lambda}).$$

As a consequence, to minimize G_Ω, $\hat{\lambda}$ should be the largest eigenvalue. The end of the proof follows the same lines as the one of Proposition 1. □

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

$$P = \sum_{j=1}^d \hat{q}_j \hat{q}_j^t,$$

and, for all $i = 1, \ldots, n$ consider the projected predictor defined by $\tilde{X}_i = P\hat{\Gamma}X_i$. Introducing $\tilde{\Gamma} = P\hat{\Gamma}P$ 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, $P\hat{\Gamma}P\tilde{b} = \tilde{\lambda}P\hat{\Sigma}P\tilde{b}$. Remarking that $P\hat{\Sigma}P = \hat{\Sigma}P$, we have, for all $\tau > 0$,

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

and defining $\hat{b} = P\tilde{b}$ and $\hat{\lambda} = \tilde{\lambda}/(1 + \tau)$, it follows that $P\hat{\Gamma}\hat{b} = \hat{\lambda}(P\hat{\Sigma} + \tau\hat{\Sigma})\hat{b}$. Since $P = \tau\hat{\Sigma}\Omega_2$, we thus have $\hat{\Sigma}\Omega_2\hat{\Gamma}\hat{b} = \hat{\lambda}(\hat{\Sigma}\Omega_2\hat{\Sigma} + \hat{\Sigma})\hat{b}$, which means that $\hat{b}$ is an eigenvector of $(\Omega_2\hat{\Sigma} + I_p)^{-1}\Omega_2\hat{\Gamma}$. Corollary 2 concludes the proof, i.e. $\hat{b}$ is the GRSIR estimator with prior covariance matrix Ω_2. □

Proof of Proposition 4 We adopt the notation 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 $\tilde{\Omega}(\varphi)P\hat{\Gamma}P\tilde{b} = \tilde{\lambda}(\tilde{\Omega}(\varphi)P\hat{\Sigma}P + I_p)\tilde{b}$. Multiplying this equation by P on the left, we obtain $P\tilde{\Omega}(\varphi)P\hat{\Gamma}P\tilde{b} = \tilde{\lambda}(P\tilde{\Omega}(\varphi)P\hat{\Sigma}P + P)\tilde{b}$. Since $P\tilde{\Omega}(\varphi)P = \Omega(\varphi)$, and introducing $\hat{b} = P\tilde{b}$, it follows that $\Omega(\varphi)\hat{\Gamma}\hat{b} = \tilde{\lambda}(\Omega(\varphi)\hat{\Sigma} + I_p)\hat{b}$, which means that $\hat{b}$ is an eigenvector of $(\Omega(\varphi)\hat{\Sigma} + I_p)^{-1}\Omega(\varphi)\hat{\Gamma}$. Corollary 2 concludes the proof, i.e. $\hat{b}$ is the GRSIR estimator with prior covariance matrix $\Omega(\varphi)$. □

References

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

Antoniadis, A., Grégoire, G., McKeague, I.W.: Bayesian estimation in single-index models. Stat. Sin. **14**, 1147–1164 (2004)

Aragon, Y., Saracco, J.: Sliced inverse regression: an appraisal of small sample alternatives to slicing. Comput. Stat. **12**, 109–130 (1997)

Bernard-Michel, C., Gardes, L., Girard, S.: Estimation of Mars surface physical properties from hyperspectral images using Sliced Inverse Regression. Technical Report, INRIA. http://hal.inria.fr/inria-00187444 (2007)

Bernard-Michel, C., Gardes, L., Girard, S.: A note on Sliced Inverse Regression with regularizations. Biometrics (2008, to appear)

Chiaromonte, F., Martinelli, J.: Dimension reduction strategies for analyzing global gene expression data with a response. Math. Biosci. **176**, 123–144 (2002)

Cook, R.D., Weisberg, S.: Discussion of "Sliced Inverse Regression for dimension reduction" by K.C. Li. J. Am. Stat. Assoc. **86**, 328–332 (1991)

Cook, R.D.: Regression Graphics. Ideas for Studying Regressions Through Graphics. Wiley Series in Probability and Statistics. Wiley, New York (1998)

Cook, R.D.: Testing predictor contributions in sufficient dimension reduction. Ann. Stat. **32**, 1062–1092 (2004)

Cook, R.D.: Fisher lecture: Dimension reduction in regression. Stat. Sci. **22**(1), 1–26 (2007)

Cook, R.D., Ni, L.: Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. J. Am. Stat. Assoc. **100**, 410–428 (2005)

Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge (2000)

Douté, S., Schmitt, B., Langevin, Y., Bibring, J.-P., Altieri, F., Bellucci, G., Gondet, B., Poulet, F.: South pole of Mars: Nature and composition of the icy terrains from Mars Express OMEGA observations. Planet. Space Sci. **55**(1–2), 113–133 (2007)

Draper, N.R., Smith, R.: Applied Regression Analysis, 3rd edn. Wiley, New York (1998)

Ferré, L.: Determining the dimension in Sliced Inverse Regression and related methods. J. Am. Stat. Assoc. **93**, 132–140 (1998)

Ferré, L., Yao, A.F.: Smoothed functional inverse regression. Stat. Sin. **15**, 665–683 (2005)

Friedman, J.H.: Regularized discriminant analysis. J. Am. Stat. Assoc. **84**, 165–175 (1989)

Gannoun, A., Saracco, J.: An asymptotic theory for SIR_α method. Stat. Sin. **13**, 297–310 (2003)

Heiberger, R.: Generation of random orthogonal matrices. J. R. Stat. Soc. Ser. C **27**, 199–206 (1978)

Hsing, T., Carroll, R.J.: An asymptotic theory for Sliced Inverse Regression. Ann. Stat. **20**(2), 1040–1061 (1992)

Li, K.C.: Sliced Inverse Regression for dimension reduction. J. Am. Stat. Assoc. **86**, 316–327 (1991)

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

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

Saracco, J.: An asymptotic theory for Sliced Inverse Regression. Commun. Stat. Theory Methods **26**, 2141–2171 (1997)

Saracco, J.: Asymptotics for pooled marginal slicing estimator based on SIR_α approach. J. Multivar. Anal. **96**(1), 117–135 (2005)

Schmidt, F., Douté, S., Schmitt, B.: Wavanglet: an efficient supervised classifier for hyperspectral images. Geosci. Remote Sens. IEEE Trans. **45**(5), 1374–1385 (2007)

Schott, J.R.: Determining the dimensionality in Sliced Inverse Regression. J. Am. Stat. Assoc. **89**, 141–148 (1994)

Scrucca, L.: Regularized Sliced Inverse Regression with applications in classification. In: Zani, S., Cerioli, A., Riani, M., Vichi, M. (eds.) Data Analysis, Classification and the Forward Search, pp. 59–66. Springer, Berlin (2006)

Scrucca, L.: Class prediction and gene selection for DNA microarrays using regularized Sliced Inverse Regression. Comput. Stat. Data Anal. **52**, 438–451 (2007)

Vogel, C.R.: Computational Methods for Inverse Problems. SIAM, Philadelphia (2002)

Zhong, W., Zeng, P., Ma, P., Liu, J.S., Zhu, Y.: RSIR: Regularized Sliced Inverse Regression for motif discovery. Bioinformatics **21**(22), 4169–4175 (2005)

Zhu, L.X., Ng, K.W.: Asymptotics of Sliced Inverse Regression. Stat. Sin. **5**, 727–736 (1995)