A sliced inverse regression approach for data stream

Marie Chavent, Stéphane Girard, Vanessa Kuentz, Benoit Liquet, Thi Mong Ngoc Nguyen, Jérôme Saracco

A sliced inverse regression approach for data stream

Marie Chavent[1,2], Stéphane Girard[3], Vanessa Kuentz-Simonet[4], Benoit Liquet[5], Thi Mong Ngoc Nguyen[6] and Jérôme Saracco[1,2]

[1] Institut de Mathématiques de Bordeaux, UMR CNRS 5251
Université de Bordeaux / Institut Polytechnique de Bordeaux,
351 cours de la libération, 33405 Talence Cedex, France
e-mail: {marie.chavent,Thi.Mong.Ngoc.Nguyen,jerome.saracco}@math.u-bordeaux1.fr

[2] INRIA Bordeaux Sud-Ouest, CQFD team, France

[3] INRIA Rhône-Alpes, MISTIS team, France
Inovallée, 655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France
e-mail: Stephane.Girard@inria.fr

[4] IRSTEA, Unité ADBX "Aménités et Dynamiques des Espaces Ruraux"
50 avenue de Verdun - Gazinet, 33612 Cestas Cedex, France
e-mail: vanessa.kuentz-simonet@irstea.fr

[5] INSERM U897, ISPED
Université Victor Segalen Bordeaux 2,
146 rue Leo Saignat, 33076 Bordeaux cedex, France
e-mail: Benoit.Liquet@isped.u-bordeaux2.fr

[6] Université de Strasbourg, IRMA, UMR 7501
7 rue René Descartes, 67084 Strasbourg cedex, France
e-mail: tmnnguyen@math.unistra.fr

Abstract. In this article, we focus on data arriving sequentially by block in a stream. A semiparametric regression model involving a common EDR (Effective Dimension Reduction) direction β is assumed in each block. Our goal is to estimate this direction at each arrival of a new block. A simple direct approach consists in pooling all the observed blocks and estimate the EDR direction by the SIR (Sliced Inverse Regression) method. But some disadvantages appear in practice such as the storage of the blocks and the running time for high dimensional data. To overcome these drawbacks, we propose an adaptive SIR estimator of β based on the SIR approach for a stratified population developed by Chavent *et al.* (2011). The proposed approach is faster both from computational complexity and running time points of view, and provides data storage benefits. We show the consistency of our estimator at the root-n rate and give its asymptotic distribution. We propose an extension to multiple indices model. We also provide a graphical tool in order to detect if a drift occurs in the EDR direction or if some aberrant blocks appear in the data stream. In a simulation study, we illustrate the good numerical behavior of our estimator. One important advantage of this approach is its adaptability to changes in the underlying model. Finally we apply it on real data concerning the estimation

of Mars surface physical properties.

Keywords: Effective Dimension Reduction (EDR), Sliced Inverse Regression (SIR), data stream

1 Introduction

Regression analysis is used to highlight the relationship between one response variable Y and a p-dimensional explanatory variable X. In parametric regression, the link function is specified as a simple algebraic function of the covariable, and the search for the best global fit is reached for instance by least squares or maximum likelihood methods. In nonparametric regression, the class of fitted functions is enlarged to get greater flexibility via sophisticated smoothing procedures (such kernel or smoothing splines methods). However as the dimension p of the covariable X becomes large, increased difficulties in modeling are often encountered in practice. This is the well-known curse of dimensionality.

In this framework of high dimensional regression, Duan and Li (1991) proposed the following semiparametric dimension reduction single index model:

$$Y = f(X'\beta, \varepsilon), \tag{1}$$

where the univariate response variable Y is linked with the p-dimensional regressor X (with expectation $\mathbb{E}(X) = \mu$ and covariance matrix $\mathbb{V}(X) = \Sigma$) only through the single index $X'\beta$. The error term ε is independent of X. The link function f and the vector β are unknown. Since β is not totally identifiable in this model, we search for the linear subspace spanned by β, called the Effective Dimension Reduction (EDR) space.

Li (1991) introduced the SIR (Sliced Inverse Regression) method which is a computationally simple and fast method to estimate the EDR space without assuming neither the functional form of f nor the distribution of ε. This method is based on some properties of the conditional distribution of X given Y and exploits a property of the first inverse moment $\mathbb{E}(X|Y)$; see for instance Duan and Li (1991), Chen and Li (1998), Zhu *et al.* (2007) among others. The basic principle of SIR methods is to reverse the role of Y and X and to study the property of the conditional moments of X given Y. In this paper, we will only focus on the SIR-I method (denoted by SIR hereafter) which is based on the first conditional moment. The term "Sliced" refers to the fact that a slicing is realized on the response variable Y to facilitate the estimation of the inverse conditional mean.

In this paper we focus on data stream, that is data arriving sequentially by block in a stream. Since a large number of data sets are currently no more fixed but do evolve in

time, the study of dimension reduction model in this case appears to be very useful. We assume that each data block t is composed of an independent and identically distributed (i.i.d.) sample $\{(X_i, Y_i), i = 1, \ldots, n_t\}$ available from model (1). A first simple approach to estimate the EDR direction consists in waiting for all the blocks to be observed, pooling them and then estimating the EDR direction by SIR. While SIR is a computationally simple and fast method, the drawback of pooling the data is the storage of the blocks since the size of the dataset considerably increases with the number of blocks. To avoid this, we propose an adaptive SIR method based on the SIR approach for a stratified population developed by Chavent *et al.* (2011).

The idea of the approach is to combine in a matrix the EDR directions of each observed block weighted by one constant term which evaluates the squared cosine between the actual direction of the block and the previous one. This enables to have an adaptive procedure which can detect if some aberrant blocks appear and which can recover some possible changes of EDR direction in the data. Another main advantage of this method is in terms of storage since we do not have to keep all blocks of observations but only their EDR directions.

In Section 2, after a brief recall on SIR, we introduce our SIR approach for data stream (named SIRds) and we describe its population and sample versions. We show the convergence in probability and the asymptotic distribution of the corresponding estimator of the EDR direction. We extend this approach to multiple indices models in Section 3. A simulation study is carried out in Section 4 in order to show the good behavior of our estimator and to compare it to classical SIR applied to all the blocks. Our approach is profitable in terms of computational complexity and running time. Another main advantage is its adaptability since it well recovers the potential changes of direction in some blocks. In Section 5 the proposed adaptive SIR method is used to evaluate the physical properties of surface materials on the planet Mars from hyperspectral images. Concluding remarks are then given in Section 6.

2 An adaptive SIR estimator for data stream: SIRds

We first recall in Section 2.1 the population and sample versions of SIR only based on one block. Then in Section 2.2 we present the population and sample versions of our adaptive method for a data stream (of blocks), SIRds. We give asymptotic results for SIRds estimator in Section 2.3. Finally in Section 2.4 we provide some comments on computational complexity

and data storage for SIRds and usual SIR applied to the union of blocks.

2.1 Recall on SIR in block t

In this section, we focus on only one block, the block t. We first present the population version of SIR. Then we derive its sample version.

The population version SIR relies on the following linear condition:

$$(LC): \quad \forall b \in \mathbb{R}^p, \mathbb{E}(X'b|X'\beta) \text{ is linear in } X'\beta,$$

which is fulfilled when X is elliptically distributed. Moreover, in the presence of high-dimensional data, this condition is often approximately fulfilled, see Hall and Li (1993) for details. Let us consider a monotone transformation $T(\cdot)$ of Y. Under condition (LC) and model (1), Li (1991) showed that the centered inverse regression curve is contained in the one-dimensional linear subspace of $\mathbb{R}^p$ spanned by $\Sigma\beta$. As a consequence, the eigenvector u_t of $\Sigma^{-1}\Gamma_t$ associated with the non-null eigenvalue is an EDR direction (i.e. is collinear with β) where $\Gamma_t = \mathbb{V}(\mathbb{E}(X|T(Y)))$. The vector u_t is Σ-normalized. Let us define b_t the I_p-normalized version of u_t as $b_t = u_t/||u_t||$ with $||u_t||^2 = u_t'u_t$.

To obtain an estimator of Γ_t which can be easily used in practice, Li (1991) proposed for $T(.)$ a slicing into $H_t \geq 2$ non-overlapping slices $s_1, \ldots, s_{H_t}$. Denoting the hth slice weight (resp. mean) by $p_h = P(Y \in s_h)$ (resp. $m_h = \mathbb{E}(X|Y \in s_h)$), then the matrix Γ_t can be written as:

$$\Gamma_t = \sum_{h=1}^{H_t} p_h(m_h - \mu)(m_h - \mu)'. \tag{2}$$

Then it is straightforward to estimate the matrix Γ_t by substituting theoretical versions of the moments by their empirical counterparts. Let $\widehat{\Gamma}_t$ denote this estimator. One therefore obtains the estimated EDR direction $\widehat{u}_t$ as the eigenvector associated with the largest eigenvalue of $\widehat{\Sigma}^{-1}\widehat{\Gamma}_t$ where $\widehat{\Sigma}$ is an estimator of Σ. The vector $\widehat{u}_t$ is $\widehat{\Sigma}$-normalized. Let us define $\widehat{b}_t$ the I_p-normalized version of $\widehat{u}_t$ as $\widehat{b}_t = \widehat{u}_t/||\widehat{u}_t||$.

2.2 Population and sample versions of SIRds

In this section, we consider T sequentially arriving blocks of data. From each block t, we can obtain b_t, the I_p-normalized EDR direction with SIR as it has been described in the previous section. The question is now to combine these directions $b_1, \ldots, b_t, \ldots, b_T$ in order to provide an estimator of the EDR direction taking into account the available T blocks.

Averaging the vectors b_t is not such a good idea since only the direction of b_t is identifiable. Another way is to consider the following $p \times p$ matrix:

$$\widetilde{M}_T = \sum_{t=1}^{T} w_t b_t b_t', \tag{3}$$

where the w_t's are positive weights such that $\sum_{t=1}^{T} w_t = 1$. Assuming the linearity condition (LC) and model (1) in each block t, all the vectors b_t are collinear with β. Then the rank of the symmetric matrix $\widetilde{M}_T$ is one. The eigenvector $\tilde{v}_T$ associated with the non-null eigenvalue of $\widetilde{M}_T$ is also collinear with β: thus $\tilde{v}_T$ is an I_p-normalized EDR direction ($||\tilde{v}_T|| = 1$).

This approach of SIR for data stream is suitable but suffers from not being an adaptive one (if the parametric part of the underlying model evolves, for instance β is replaced by $\beta^* \neq \beta$ in the underlying model in block T). In the following, we introduce an adaptive version of $\widetilde{M}_T$ which will take into account the possible evolution of the parametric part of the underlying semiparametric model in each block.

Population version of SIRds. To give an adaptive SIR approach for data stream, we add in matrix $\widetilde{M}_T$ the weights $\cos^2(b_t, b_T) = \dfrac{(b_t' b_T)^2}{(b_t' b_t) \times (b_T' b_T)} = (b_t' b_T)^2$ since the b_t's are I_p-normalized. These weights will examine if the "new" block T provides the same information as the previous blocks, that is if the EDR direction obtained in block T is close to those of the previous blocks $t = 1, \ldots, T-1$. We consider the following matrix:

$$M_T = \sum_{t=1}^{T} w_t b_t b_t' \cos^2(b_t, b_T). \tag{4}$$

Theorem 1 *Assuming the linearity condition (LC) and model* (1) *in each block, the I_p-normalized eigenvector v_T associated with the largest eigenvalue of M_T is an EDR direction.*

Proof of Theorem 1. Under the assumptions of our model, the term $\cos^2(b_t, b_T)$ is equal to one since b_t and b_T are both collinear with β. Then similarly to matrix $\widetilde{M}_T$, it is also straightforward to show that the principal eigenvector v_T of M_T is collinear with β and is an EDR direction. □

Let us now exhibit the advantage of this adaptive SIRds version on a simple example. Let us assume that the underlying regression model is given in (1) for the first $T-1$ blocks. We also assume that for the last block T the parameter β is replaced by β^* in model (1) with $\beta^* \perp \beta$ for the usual inner product. Assuming the linearity condition, the SIRds approach provides an EDR direction collinear to β for the first $T-1$ blocks as it has been

mentioned above. When the block T arrives, from the population point of view, we have: $\cos^2(b_t, b_T) = 0$ for $t = 1, \ldots, T-1$ since each $b_t, t = 1, \ldots, T-1$ are collinear with β and b_T is collinear to β^*. Then $M_T = w_T b_T b_T'$ since $\cos^2(b_T, b_T) = 1$. Finally we obtain an EDR direction collinear with β^*. To conclude, the SIRds approach allows to detect an aberrant block (that is with a parametric part which differs from that of the previous block) and then to provide the "true" EDR direction of this specific block. A visualization of the weights $\cos^2(b_t, b_\mathcal{T})$ for $\mathcal{T} = 2, \ldots, T$ and $t = 1, \ldots, \mathcal{T}$ will be very useful for the user to detect if there is or not aberrant blocks or a drift in the data stream. Section 4 gives some graphical illustrations on various scenarios of data streams.

Let us remark that it is possible to reformulate this approach as an optimization problem.

Theorem 2 *Under linearity condition (LC) and model (1), the EDR direction v_T is the solution of the maximization problem*

$$\max_{v \in \mathbb{R}^p} \sum_{t=1}^{T} \omega_t \cos^2(b_t, v) \quad s.t.\ ||v|| = 1, \tag{5}$$

where $\omega_t = w_t \cos^2(b_t, b_T)$.

Proof of Theorem 2. Since $||b_t|| = 1$, we have:

$$\begin{aligned} \textstyle\sum_{t=1}^{T} \omega_t \cos^2(b_t, v) &= \textstyle\sum_{t=1}^{T} \omega_t < b_t, v >^2 = \sum_{t=1}^{T} \omega_t (b_t' v)^2 \\ &= \textstyle\sum_{t=1}^{T} \omega_t v' b_t b_t' v = v' \left(\sum_{t=1}^{T} \omega_t b_t b_t' \right) v \\ &= v' M_T v. \end{aligned}$$

Thus maximization problem (5) can be rewritten as

$$\max_{v \in \mathbb{R}^p} \frac{v' M_T v}{v' v}. \tag{6}$$

The solution of (6) is clearly the normalized principal eigenvector of M_T. □

Sample version of SIRds. For $t = 1, \ldots, T$, let us recall that $\widehat{b}_t$ is the I_p-normalized estimator of the EDR direction b_t calculated in each block t.

The estimator $\widehat{v}_T$ of the EDR direction v_T with the SIRds approach is the principal eigenvector of the $p \times p$ matrix defined as

$$\widehat{M}_T = \sum_{t=1}^{T} w_t \widehat{b}_t \widehat{b}_t' \cos^2(\widehat{b}_t, \widehat{b}_T). \tag{7}$$

One possible choice for the weights w_t can be $w_t = \frac{n_t}{\sum_{j=1}^{T} n_j}$ for $t = 1, \ldots, T$. In the next section, we give some asymptotic results for this estimator: $\sqrt{n}$-convergence and asymptotic normality.

2.3 Asymptotic results

The following assumptions are necessary to state our asymptotic result for a fixed number T of blocks and a sample size $n = \sum_{t=1}^{T}$ which tends to ∞. Let $n_{h,t}$ be the number of observations in the hth slice in the block t and let $n_t = \sum_{h=1}^{H_t} n_{h,t}$ be the number of observations in the block t.

- (A1) Each block t is a sample of independent observations from the single index model (1).
- (A2) For each block t, the support of Y is partitioned into a fixed number H_t of slices such that $p_h > 0, h = 1, \ldots, H_t$.
- (A3) For $t = 1, \ldots, T$ and $h = 1, \ldots, H_t$, $n_{h,t} \to \infty$ (and therefore $n_t \to \infty$) as $n \to \infty$.

Theorem 3 *Under linearity condition (LC) and (A1)-(A3), we have*

$$\widehat{v}_T = v_T + O_p(n^{-1/2}),$$

that is the estimated EDR direction converges in probability to the direction of β.

Proof of Theorem 3. For each block t and under the assumptions (LC), (A1)-(A3), from SIR theory of Li (1991) each estimated EDR direction $\widehat{b}_t$ converges to b_t at root n rate: that is, for $t = 1, \ldots, T$, $\widehat{b}_t = b_t + O_p(n^{-1/2})$. Since $\cos^2(\widehat{b}_t, \widehat{b}_T) = \cos^2(b_t, b_T) + O_p(n^{-1/2}) = 1 + O_p(n^{-1/2})$, we get $\widehat{M}_T = M_T + O_p(n^{-1/2})$. Therefore the principal eigenvector of $\widehat{M}_T$ converges to the corresponding one of M_T at the same rate: $\widehat{v}_T = v_T + O_p(n^{-1/2})$. Since v_T is collinear with β, then the estimated EDR direction $\widehat{v}_T$ converges to an EDR direction at root n rate. □

Theorem 4 *Under linearity condition (LC) and (A1)-(A3), we have*

$$\sqrt{n}(\widehat{v}_T - v_T) \longrightarrow_d W \sim \mathcal{N}(0, \Gamma_W),$$

where the expression of Γ_W is given in (12).

Proof of Theorem 4. Let $C_1 \otimes C_2$ denote the Kronecker product of the matrices C_1 and C_2 (see for instance Harville, 1999, for some useful properties of the Kronecker product). Let $C = [c_1, \ldots, c_q]$ be a $(p \times q)$ matrix, where the c_k's are p-dimensional column vectors. We note $\mathrm{vec}(C)$ the pq-dimensional column vector: $\mathrm{vec}(C) = \left(c_1', \ldots, c_q'\right)'$. We shall note

N^+ the Moore-Penrose generalized inverse of the square matrix N. In the sequel, we define the matrix $B = [b_1, \ldots, b_T]$ which contains all the EDR directions obtained from all the T blocks. Let us also define the matrix $\widehat{B} = [\widehat{b}_1, \ldots, \widehat{b}_T]$. The proof involves three steps.

Step 1: Asymptotic distribution of vec$(\widehat{B})$. Under (A1)-(A3), asymptotic theory of SIR gives us the following result for each block $t = 1, \ldots, T$: $\sqrt{n}(\widehat{b}_t - b_t) \longrightarrow_d U_t \sim \mathcal{N}(0, V_t)$, where the expression of V_t can be found in Saracco (1997) for instance. Then, we have

$$\sqrt{n}(\text{vec}(\widehat{B}) - \text{vec}(B)) \longrightarrow_d \text{vec} \begin{pmatrix} U_1 \\ \vdots \\ U_T \end{pmatrix} \sim \mathcal{N}(0, \Gamma_U) \text{ where } \Gamma_U = \begin{pmatrix} V_1 & & 0 \\ & \ddots & \\ 0 & & V_T \end{pmatrix} \quad (8)$$

Step 2: Asymptotic distribution of vec$(\widehat{M}_T)$. We have

$$\begin{aligned} \text{vec}(\widehat{M}_T) &= \sum_{t=1}^{T} w_t \text{vec}(\widehat{b}_t\widehat{b}_t')(\widehat{b}_t'\widehat{b}_T)^2 \quad (9) \\ &= f(\text{vec}(\widehat{B})) \quad (10) \end{aligned}$$

with $||\widehat{b}_t|| = 1, \ \forall t = 1, \ldots, T$ and where the function f is defined as:

$$\begin{aligned} f: \ \mathbb{R}^{pT} &\to \mathbb{R}^{p^2} \\ \text{vec}(B) &\mapsto \textstyle\sum_{t=1}^{T} w_t \text{vec}(b_t b_t')(b_t' b_T)^2. \end{aligned}$$

Then, the $(p^2 \times pT)$ Jacobian matrix $J = [J_1 | \ldots | J_T]$ associated to f is defined by the concatenation of the $p^2 \times p$ matrices J_t For $t = 1, \ldots, T-1$, we have:

$$\begin{aligned} J_t &= \frac{\partial f(\text{vec}(B))}{\partial b_t'} = \frac{\partial w_t \text{vec}(b_t b_t')(b_t' b_T)^2}{\partial b_t'} \\ &= w_t (K_{1,p} \otimes I_p)[b_t \otimes I_p + I_p \otimes b_t](b_t' b_T)^2 + w_t \text{vec}(b_t b_t') 2(b_t' b_T) b_T', \end{aligned}$$

where the vec-permutation matrix $K_{1,p}$ is equal to $K_{1,p} = \sum_{j=1}^{p}(E_{1j} \otimes E_{1j}')$ with $E_{1j} = e_{j,p}'$ and $e_{j,p}$ the jth column of I_p. The matrix J_T is defined by:

$$\begin{aligned} J_T &= \frac{\partial f(\text{vec}(B))}{\partial b_T'} = \frac{\partial \sum_{t=1}^{T} w_t \text{vec}(b_t b_t')(b_t' b_T)^2}{\partial b_t'} \\ &= \sum_{t=1}^{T-1} w_t \text{vec}(b_t b_t') 2(b_t' b_T) b_t' + \frac{\partial w_T \text{vec}(b_T b_T')(b_T' b_T)^2}{\partial b_T'} \\ &= \sum_{t=1}^{T-1} w_t \text{vec}(b_t b_t') 2(b_t' b_T) b_t' + w_T (K_{1,p} \otimes I_p)[b_T \otimes I_p + I_p \otimes b_T](b_T' b_T)^2 + w_T \text{vec}(b_T b_T') 4(b_T' b_T) b_T' \end{aligned}$$

Then using (8) and applying Delta-method, we obtain

$$\sqrt{n}(\text{vec}(\widehat{M}_T) - \text{vec}(M_T)) \longrightarrow_d V \sim \mathcal{N}(0, \Gamma_V = J \Gamma_U J'). \quad (11)$$

Step 3: Asymptotic distribution of $\widehat{b}$. The vector $\widehat{v}_T$ (resp. v_T) is the eigenvector associated to the largest eigenvalue $\widehat{\lambda}$ (resp. λ) of $\widehat{M_T}$ (resp. M_T). Since $\widehat{M_T} = M_T + O_p(n^{-1/2})$ and using (11), according to Lemma 1 of Saracco (1997), we obtain

$$\sqrt{n}(\widehat{v}_T - v_T) \longrightarrow_d W = (M_T - \lambda I_p)^+ V v_T \sim \mathcal{N}(0, \Gamma_W)$$

with

$$\Gamma_W = [v_T' \otimes (M_T - \lambda I_p)^+] \Gamma_V [v_T \otimes (M_T - \lambda I_p)^+]. \quad (12)$$

□

2.4 Computational complexity and data storage

Computational complexity. For sake of simplicity, let us assume that each block has the same sample size n^* and that $n^* >> p$. In such a case, the computational complexity of SIR computed on one block is of order $n^* p^2$ (denoted as $O(n^* p^2)$ hereafter). It corresponds to the cost of computing the empirical covariance matrix $\widehat{\Sigma}$. Our goal is to show that the SIRds approach performs faster than the sequential SIR method which consists in computing SIR on the union of the j first blocks for $j = 1, \ldots, T$. Clearly, the computational complexity of sequential SIR is $O(n^* p^2 + 2n^* p^2 + \cdots + Tn^* p^2) = O(T^2 n^* p^2)$ since it requires T computations of SIR on blocks of increasing sizes. The computational complexity of SIRds is $O(Tp^2(n^* + T))$, the additional term $T^2 p^2$ being due to the T computations of the matrix $\widehat{M_T}$. As a consequence, if $n^* >> \max(T, p)$, SIRds is $O(T)$ times faster than sequential SIR.

Data storage. Sequential SIR requires the storage of the whole matrix of regressors, its storage load is thus $O(Tn^* p^2)$. As a comparison, SIRds requires the storage of only one block of regressors and of the T EDR directions computed on the previous blocks, corresponding to a storage load $O(p(n^* + T))$. Under the previous assumptions, SIRds requires $O(T)$ less data storage than sequential SIR.

3 Extension to multiple indices model

We suggest in this section a possible extension of the proposed approach to the case of a multiple indices model. Let us first introduce the multiple indices model. The response variable Y is related to the p-dimensional quantitative regressor X (with $\mathbb{E}(X) = \mu$ and

$\mathbb{V}(X) = \Sigma$) only through the indices $X'\beta_k,\ k = 1, \ldots, K$:

$$Y = g(X'\beta_1, \ldots, X'\beta_K, \varepsilon). \tag{13}$$

As in the single index model, the error term ε is independent of X and the link function g is unknown. In other words, Y and X are independent conditionally on $(X'\beta_1, \ldots, X'\beta_K)$. In this multiple indices model, we search for a basis that spans the K-dimensional EDR space $E = \text{Span}(\beta_1, \ldots, \beta_K)$. As for the single index model, we shall seek with SIR for a basis of the EDR space for each block. In order to get theoretical results, we need to adapt the linearity condition and we now assume:

$$\text{(LC')}\ \ \mathbb{E}(X'v|X'\beta_1, \ldots, X'\beta_K)\ \text{ is linear in } X'\beta_1, \ldots, X'\beta_K\ \text{ for any } v \in \mathbb{R}^p.$$

To take into account the K-dimensional EDR space, we need to modify the matrix M_T defined in (4) for the single index model. We will replace b_t by an I_p-orthonormal basis $\mathbb{B}_t$ of the EDR space and the weights $\cos^2(b_t, b_T)$ by the following proximity measure between the linear subspaces spanned by $\mathbb{B}_t$ and $\mathbb{B}_T$ from the blocks t and T:

$$\frac{\text{Trace}(P_t P_T)}{K},$$

where $P_l = \mathbb{B}_l(\mathbb{B}_l'\mathbb{B}_l)^{-1}\mathbb{B}_l' = \mathbb{B}_l\mathbb{B}_l'$ is the I_p-orthogonal projector on the EDR space obtained from the block l (equal to t or T), that is $\text{Span}(\mathbb{B}_l)$. This measure takes its values in [0,1]. Note that $\frac{\text{Trace}(P_t P_T)}{K} = 1$ when $\text{Span}(B_t) = \text{Span}(B_T)$. The closer to one is this measure, the closer is the linear subspace $\text{Span}(B_t)$ to the the linear subspace $\text{Span}(B_T)$.

Population version. For each block t, assuming the linearity condition (LC') and model (13), the eigenvectors $u_{1,t}, \ldots, u_{K,t}$ associated with the largest K eigenvalues of the matrix $\Sigma^{-1}\Gamma_t$ are EDR directions, where the matrix Γ_t has been defined in (2). Note that the number H of slices for each block must be greater than K in order to avoid artificial dimension reduction. Let us define the matrix $\mathbb{U}_t = [u_{1,t}, \ldots, u_{K,t}]$ containing these EDR directions which form a Σ-orthogonal basis of E. Then the first K eigenvectors, $b_{1,T}, \ldots, b_{K,T}$ of the matrix $\mathbb{U}_t\mathbb{U}_t'$ form an I_p-orthonormal basis of E. We store them in the $p \times K$ matrix $\mathbb{B}_t = [b_{1,t}, \ldots, b_{K,t}]$.

We consider in SIRds approach the following matrix of interest

$$\mathbb{M}_T = \sum_{t=1}^{T} w_t \mathbb{B}_t \mathbb{B}_t' \frac{\text{Trace}(P_t P_T)}{K}, \tag{14}$$

where the w_t's are positive weights such that $\sum_{t=1}^{T} w_t = 1$. The K eigenvectors associated with the largest K eigenvalues of $\mathbb{M}_T$ will be denoted by $v_{1,T}, \ldots, v_{K,T}$ and we store them in the $p \times K$ matrix $\mathbb{V}_T = [v_{1,T}, \ldots, v_{K,T}]$.

Theorem 5 *Assuming the linearity condition (LC') and model* (13)*, the column vectors of* $\mathbb{V}_T$ *form an* I_p*-orthonormal basis of the EDR space* E*.*

Proof of Theorem 5. Since the column vectors of $\mathbb{B}_t$ form an I_p-orthonormal basis of E, we have $\text{Span}(\mathbb{B}_t) = E$ for each block t. Then the eigenvectors associated with the K largest eigenvalues of $\mathbb{B}_t\mathbb{B}_t'$ form an I_p-orthonormal basis of E. Since under the assumptions of the theorem, we have $\frac{\text{Trace}(P_t P_T)}{K} = 1$, then it straightforwardly follows that the eigenvectors associated with the K largest eigenvalues of $\mathbb{M}_T$ form an I_p-orthonormal basis of the EDR space E. □

Analogously to the single index model, it is also possible to rely this eigenvalue problem to an optimization problem. Let $\mathbb{A}$ be a $p \times K$ matrix containing K I_p-orthonormal vectors. Let P_* denote the I_p-orthogonal projector onto the linear subspace $\text{Span}(\mathbb{A})$. Let us introduce $Q(\mathbb{A}, \mathbb{B}_1, \ldots, \mathbb{B}_T)$ the following proximity measure between the linear subspace $\text{Span}(\mathbb{A})$ and the EDR spaces $\text{Span}(\mathbb{B}_1)$, …, $\text{Span}(\mathbb{B}_T)$ respectively obtained from the available T blocks:

$$Q(\mathbb{A}, \mathbb{B}_1, \ldots, \mathbb{B}_T) = \frac{1}{T} \sum_{t=1}^{T} \tilde{w}_t \text{Trace}(P_* P_t),$$

where the $\tilde{w}_t$'s are positive weights such that $\tilde{w}_t = w_t \frac{\text{Trace}(P_t P_T)}{K}$ and $\sum_{t=1}^{T} \tilde{w}_t \leq 1$ since $\forall t, w_t \geq 0$ and $\sum_{t=1}^{T} w_t = 1$. Note that this measure takes its values in [0,1]. We have $Q(\mathbb{A}, \mathbb{B}_1, \ldots, \mathbb{B}_T) = 1$ when $\text{Span}(\mathbb{A}) = \text{Span}(\mathbb{B}_1) = \cdots = \text{Span}(\mathbb{B}_T)$. The closer to one is this measure, the closer is the linear subspace $\text{Span}(\mathbb{A})$ to all the T linear subspaces $\text{Span}(\mathbb{B}_t)$, $t = 1, \ldots, T$. Remark that when $\mathbb{A} = \mathcal{B}$ (where $\mathcal{B} = [\beta_1, \ldots, \beta_K]$ spans the true EDR space E), under the assumptions of the model, we get $Q(\mathcal{B}, \mathbb{B}_1, \ldots, \mathbb{B}_T) = 1$.

Theorem 6 *Under linearity condition (LC') and model* (13)*, an* I_p*-orthonormal basis of the* K*-dimensional EDR space is solution of the following maximization problem:*

$$\max_{\mathbb{A}} Q(\mathbb{A}, \mathbb{B}_1, \ldots, \mathbb{B}_T). \tag{15}$$

PROOF OF THEOREM 6. Since the bases $\mathbb{A}$, $\mathbb{B}_1$, ..., $\mathbb{B}_T$ are assumed to be I_p-orthonormal, we have:

$$\begin{aligned} & T \times Q(\mathbb{A}, \mathbb{B}_1, \ldots, \mathbb{B}_T) \\ = & \sum_{t=1}^T \tilde{w}_t \text{Trace}(\mathbb{A}\mathbb{A}'\mathbb{B}_t\mathbb{B}_t') \\ = & \sum_{t=1}^T \tilde{w}_t \text{Trace}(\mathbb{A}'\mathbb{B}_t\mathbb{B}_t'\mathbb{A}) \\ = & \text{Trace}(\mathbb{A}'\{\sum_{t=1}^T \tilde{w}_t\mathbb{B}_t\mathbb{B}_t'\}\mathbb{A}) \\ = & \text{Trace}(\mathbb{A}'\mathbb{M}_T\mathbb{A}). \end{aligned}$$

Let $\mathbb{A}^* = \arg\max_{\mathbb{A}} Q(\mathbb{A}, \mathbb{B}_1, \ldots, \mathbb{B}_T)$. Hence since it is well known that $\mathbb{A}^*$ is given by the $p \times K$ matrix formed by the K eigenvectors $\mathbb{V}_T$ associated with the K largest eigenvalues of $\mathbb{M}_T$, the proof is complete. □

Sample version. We can now briefly describe the corresponding sample version. For each block t, using the corresponding sample, we first estimate a $\widehat{\Sigma}_t$-orthogonal basis of the EDR space via SIR: the first K eigenvectors of the matrix $\widehat{\Sigma}_t^{-1}\widehat{\Gamma}_t$. We store them in the matrix $\widehat{\mathbb{U}}_t$. Then we consider the first K eigenvectors of the matrix $\widehat{\mathbb{U}}_t\widehat{\mathbb{U}}_t'$ which form an I_p-orthogonal basis of the estimated EDR space and we store them in the matrix $\widehat{\mathbb{B}}_t$. Finally we construct the estimator of $\mathbb{M}_T$ as follows:

$$\widehat{\mathbb{M}}_T = \sum_{t=1}^T w_t\widehat{\mathbb{B}}_t\widehat{\mathbb{B}}_t'\frac{\text{Trace}(\widehat{P}_t\widehat{P}_T)}{K},$$

where $\widehat{P}_t = \widehat{\mathbb{B}}_t(\widehat{\mathbb{B}}_t'\widehat{\mathbb{B}}_t)^{-1}\widehat{\mathbb{B}}_t' = \widehat{\mathbb{B}}_t\widehat{\mathbb{B}}_t'$ is the I_p-orthogonal projector on the estimated EDR space $\text{Span}(\widehat{\mathbb{B}}_t)$ obtained from the block t. Then the K eigenvectors associated with the largest K eigenvalues of this matrix $\widehat{M}_T$, denoted by $\widehat{\mathbb{V}}_T = [\widehat{v}_{1,T}, \ldots, \widehat{v}_{K,T}]$, provide an I_p-basis of the estimated EDR space denoted by $\widehat{E}$.

Convergence in probability. Under the linearity condition (LC') and the assumptions (A1)-(A3) given in Section 2, as for single index model, we can show that the estimated EDR basis converges to an EDR basis at root n rate, that is the estimated EDR space $\widehat{E}$ converges in probability to the true EDR space E.

Theorem 7 *Under linearity condition (LC') and* ***(A1)-(A3)****, we have*

$$\widehat{v}_{k,T} = v_{k,T} + O_p(n^{-1/2}),\ k = 1, \ldots, K,$$

that is $\widehat{\mathbb{V}}_T = \mathbb{V}_T + O_p(n^{-1/2})$.

PROOF OF THEOREM 7. SIR theory provides $\widehat{\widetilde{\mathbb{B}}}_t = \widetilde{\mathbb{B}}_t + O_p(n^{-1/2})$ for each block t. Then the eigenvectors associated with the K largest eigenvalues of the matrix $\widehat{\mathbb{B}}_t\widehat{\mathbb{B}}_t'$ converge at same rate to the corresponding ones of $\mathbb{B}_t\mathbb{B}_t'$. Under the assumptions of the theorem, we have $\frac{\text{Trace}(\widehat{P}_t\widehat{P}_T)}{K} = 1 + O_p(n^{-1/2})$. Then we get $\widehat{\mathbb{M}}_T = \mathbb{M}_T + O_p(n^{-1/2})$, and finally we obtain $\widehat{v}_{k,T} = v_{k,T} + O_p(n^{-1/2}),\ k = 1, \ldots, K$. □

On asymptotic normality. As for the single index model, using Delta-method, asymptotic results of Tyler (1981) and Saracco (1997), the asymptotic normality of the eigenprojector onto the estimated EDR space can be obtained, as well as the asymptotic distribution of the estimated EDR directions, associated with eigenvalues assumed to be different.

Choice of dimension K. Since the beginning of this section, we assumed that the dimension K of the EDR space was known. However in most applications the number K of indices is a priori unknown and hence must be estimated from the data. From a practical point of view, we recommend to choose the dimension K using classical SIR in the first block. If a block appears to be aberrant, the user has to determine again the dimension in order to confirm that the true dimension of the whole EDR space is still K. Several approaches have been proposed in the literature for SIR. Others are based on hypothesis tests on the nullity of the last $(p - K)$ eigenvalues, see Li (1991), Schott (1994) or Barrios and Velilla (2007). Another approach relies on a quality measure based on the square trace correlation between the true EDR space E and its estimate $\widehat{E}$, see for instance Ferré (1998) or Liquet and Saracco (2008, 2012) for a graphical bootstrap based approach. In the application of Section 5 we used the graphical approach of Liquet and Saracco (2012) to determine the dimension K.

4 A simulation study

A simulation study is carried out to evaluate the numerical performance of the proposed method. First we recall the definition of the quality measure in Section 4.1. In Section 4.2, we present the single index model used in this simulation study and the estimation methods. In Section 4.3, we compare the numerical results obtained with SIRds with those provided by classical SIR approach. We also compare the mean computational times obtained with both approaches. Finally in Section 4.4, we study the behavior of SIRds considering various scenarios in which some blocks do not have the same EDR direction. Note that in the

real data application of Section 5, we shall consider and estimate a multiple indices model with $K = 2$.

4.1 Quality measure

Since in the simulation study we shall only consider a single index model, we shall use as quality measure the squared cosine of the angle formed by the true EDR direction β and its estimate $\widehat{\beta}$. This estimate can be obtained by SIRds or usual SIR for instance. Hence this quality measure defined by

$$\cos^2(\beta, \widehat{\beta}) = \frac{(\beta' \widehat{\beta})^2}{(\beta'\beta) \times (\widehat{\beta}'\widehat{\beta})} \tag{16}$$

belongs to $[0, 1]$: $\cos^2(\beta, \widehat{\beta}) = 0$ if $\widehat{\beta} \perp \beta$ and $\cos^2(\beta, \widehat{\beta}) = 1$ if $\text{Span}(\beta) = \text{Span}(\widehat{\beta})$. Therefore the closer this value is to one, the better is the estimation. Note that when the dimension of the EDR space is $K > 1$ (multiple indices model), the quality measure between the true EDR space and its estimate $\widehat{E}$ is calculated as follows: $\text{Trace}(P_E P_{\widehat{E}})/K$, where P_E (resp. $P_{\widehat{E}}$) is the I_p-orthogonal projector on E (resp. $\widehat{E}$). It also belongs to [0,1], equal to 0 (resp. 1) if $E \perp \widehat{E}$ (resp. $E = \widehat{E}$).

4.2 Simulated model and estimation methods

In this simulation study, we consider for each block of data the same following semiparametric regression model:

$$Y = (X'\beta)^3 + \epsilon, \tag{17}$$

where X follows the p-dimensional normal distribution $\mathcal{N}_p(0_p, \Sigma)$ with the covariance Σ arbitrarily chosen, ϵ follows the normal distribution $\mathcal{N}(0, \sigma^2)$ and is independent of X.

We set $p = 10$, $\beta = (1, -1, 2, -2, 0, \ldots, 0)'$ and $\sigma = 0.5$. In the following, we generate data streams of $T = 20$ blocks with $n^* = 200$.

For each data stream, we estimate the EDR direction as follows. At the arrival of each block t ($t = 1, \ldots, T$), we estimate the EDR direction with SIRds based on these first available t blocks. We also estimate the EDR direction with classical SIR approach based on the sample formed by the union of the first available t blocks, this approach is denoted by SIRu (for SIR on union of blocks) hereafter.

4.3 Numerical results and running times

First, we compare the numerical results obtained with SIRds and SIRu approaches using the quality measure defined in (16). Then we focus on the running time of these approaches to give priority to one of them.

Comparison of SIRds and SIRu approaches. Our aim is to compare the quality measure of the EDR directions estimated with SIRds and SIRu. We generate $B = 50$ data replications of data stream of size $T = 20$ blocks as it has been previously described. We represent in Figure 1 the boxplots of the quality measure of the corresponding estimated EDR directions for $T = 1, 5, 10, 15, 20$ blocks. Note that in the case $T = 1$ (only one block), the two approaches SIRds and SIRu are obviously equivalent to usual SIR.

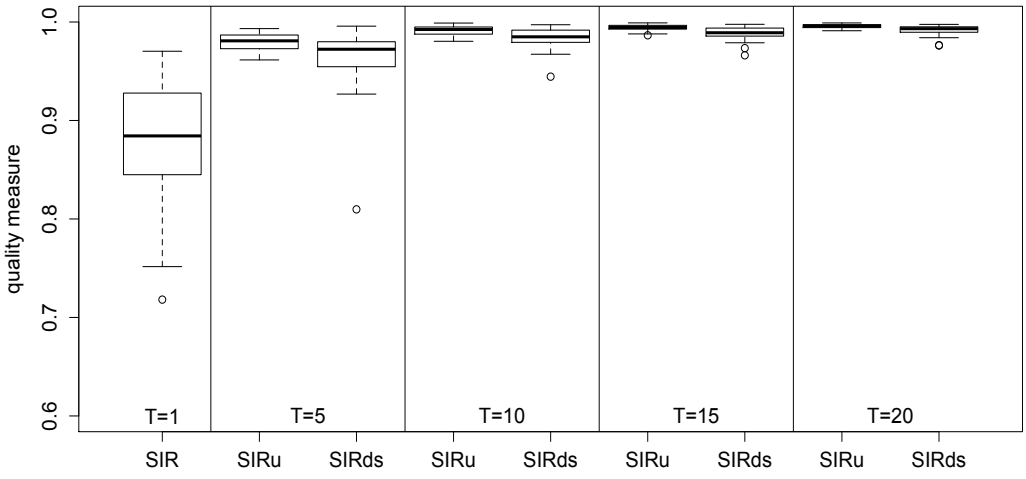


Figure 1: Boxplots of the squared cosines between the true EDR direction and the EDR directions estimated with SIRu and SIRds for different values of T.

The case $T = 1$ shows us the variability existing in each block of data. It gives an idea of the structure of the data, since each block is simulated according to the same model. Then we can notice that both methods, SIRds and SIRu, give reliable results with quality measures close to 1. Not surprisingly, the quality measure increases with the number of blocks. SIRu always provides slightly better results than SIRds as the EDR direction is estimated on the whole data set (when all blocks are collected and stored in a big dataset). But as mentioned previously, the disadvantage of SIRu is clearly the storage of all the blocks. SIRds has the advantage to keep only the estimated EDR directions of the previous blocks in memory which is an interesting gain in storage. The price to pay is a small lost of quality in the estimation of the EDR directions.

Running time. We compare the running time (in seconds) of our SIRds approach with sequential SIR (which has been defined in Section 2.4 as a sequential use of SIRu). More precisely, for data stream of T blocks, the running times have been calculated as follows:

- the running time of SIRds corresponds to the required time to compute an estimate of the EDR direction with SIR for the first block, plus the time necessary to compute an estimate of the EDR direction with SIRds for the first two blocks, ..., plus the time necessary to compute an estimate of the EDR direction with SIRds for the first T blocks ;

- the running time of sequential SIR corresponds to the time necessary to compute an estimate of the EDR direction with SIR for the first block, plus the time necessary to compute an estimate of the EDR direction with SIRu for the first two blocks, ..., plus the time necessary to compute an estimate of the EDR direction with SIRu for the first T blocks.

For various values of the dimension p of X, the size n^* of each block and the total number T of blocks in the data stream, we generate $\mathcal{B} = 50$ data streams and we evaluate the computational times for the two methods SIRds and sequential SIR.
Unsurprisingly one can observe in Figure 2 that the dimension p noticeably favors SIRds versus sequential SIR while the number T of blocks and the block size n^* hugely penalize sequential SIR approach in comparison with SIRds.

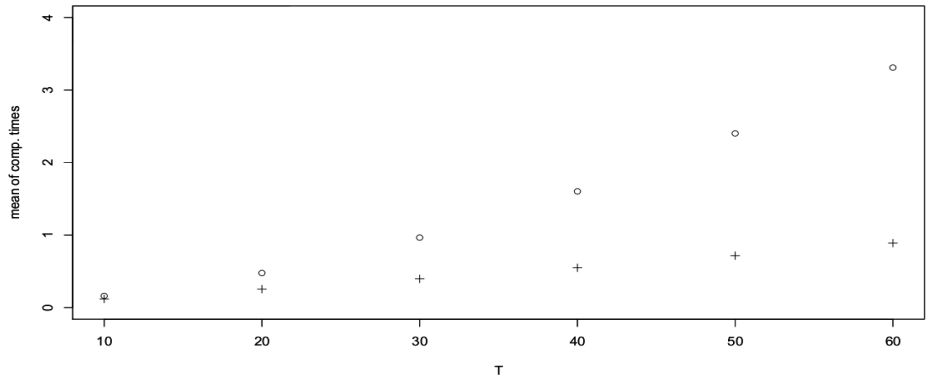
4.4 Adaptation to changes in the underlying model

In this section, we relax the assumption that the model is the same in all the blocks and the slope parameter β in model (17) is then indexed by t. In order to show the good behavior of SIRds in comparison with SIRu in such cases, we consider the following two scenarios. For each scenario, we generate $T = 20$ blocks as described below.

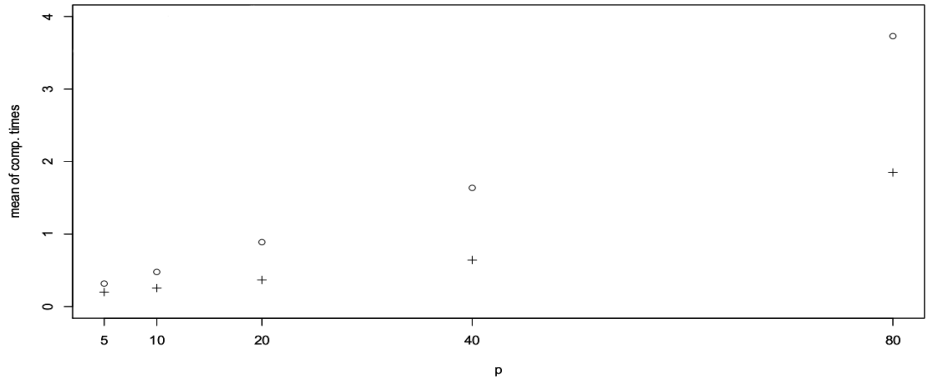
- Scenario 1: β_t is constant for $T - 1$ blocks and the 10th block is aberrant. We fix $\beta_t = (1, -1, 2, -2, 0, \ldots, 0)$ for each block t with $t \neq 10$ and we set $\beta_t = (1, 1, \ldots, 1)'$ for the 10th block.

- Scenario 2: $\beta_t = (1, -1, 2, -2, 0, \ldots, 0)'$ for the first 9 blocks ($t = 1, \ldots, 9$) and $\beta_t = (1, 1, \ldots, 1)'$ for the remaining ones ($t = 10, \ldots, 20$).

Mean of computational times according to T when $n^* = 200$ and $p = 10$



Mean of computational times according to p when $n^* = 200$ and $T = 20$



Mean of computational times according to n^* when $T = 20$ and $p = 10$

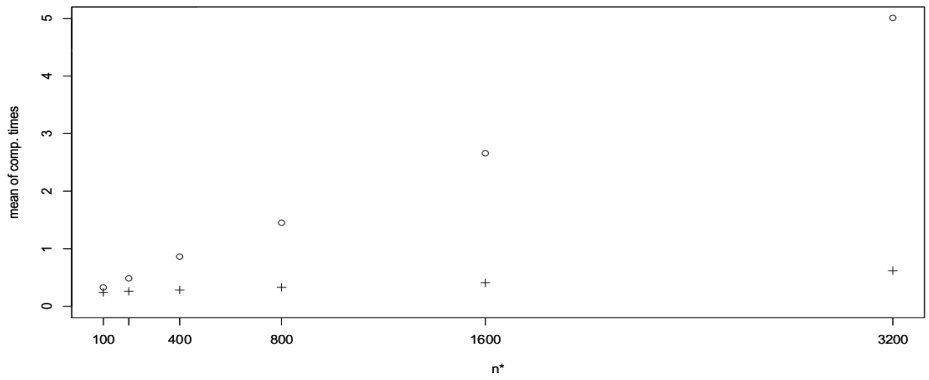


Figure 2: Running times (in seconds) of sequential SIR ($\circ$) and SIRds ($+$) for various values of p, n^* and T.

Then at each time t (i.e. when the first t blocks are available), we estimate the EDR direction with SIRds and SIRu approaches. We also estimate the EDR direction with classical SIR based only on the data of this block t.

For each scenario, we plot in Figures 3 and 4 the quality measure $\cos^2(\widehat{\beta}_t, \beta_t)$ of the estimator $\widehat{\beta}_t$ obtained with SIRds, SIRu or SIR estimators at each time t. We also represent in a color scaled image the weights $\cos^2(\widehat{b}_t, \widehat{b}_T)$ used in the computation of the SIRds estimator in equation (7). The lighter (yellow) is the color, the larger is the weight (close to 1). The darker (red) is the color, the lower is the corresponding weight (close to 0). This image will provide to the user an interesting graphic help in order to detect if aberrant blocks appear in the data stream or if a drift occurs for the underlying slope parameter.

Scenario 1: the 10th block is different

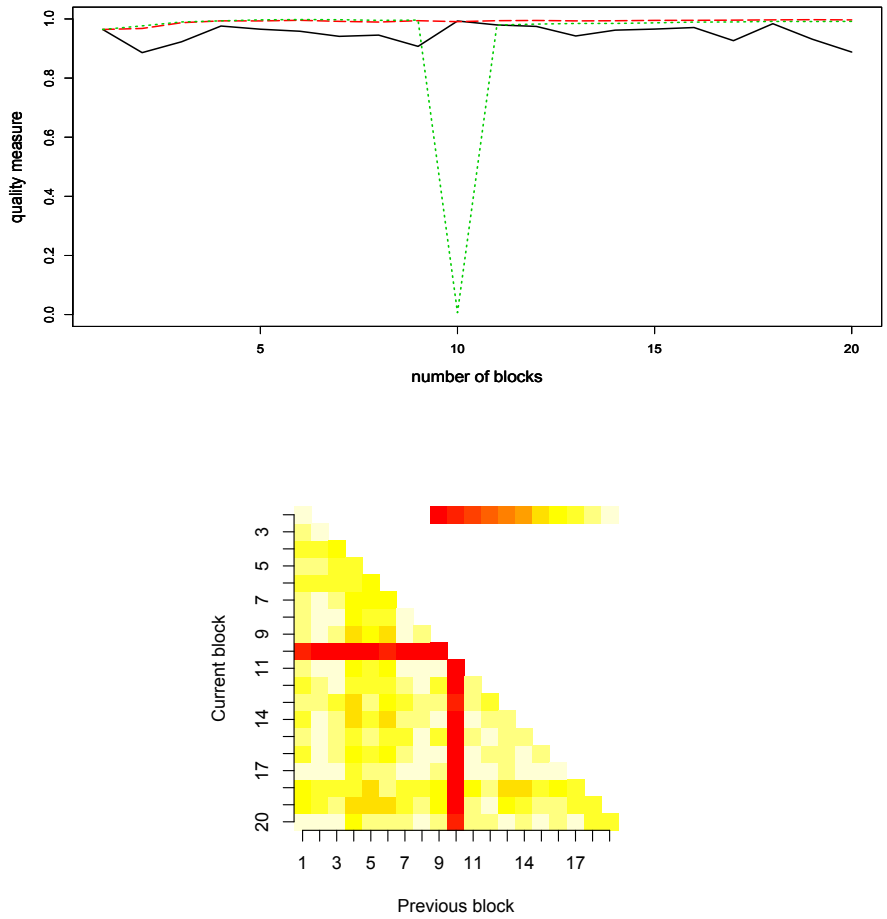


Figure 3: Numerical behavior of the SIRu and SIRds estimators for scenario 1. On the top, plot of the quality measure $\cos^2(\widehat{\beta}_t, \beta_t)$ versus the number t of blocks (dashed red line for SIRds on the first t blocks, dotted green line for SIRu on the first t blocks, solid block line for SIR on block t only). On the bottom: image of the weights $\cos^2(\widehat{b}_t, \widehat{b}_T)$ used in the computation of the SIRds estimator $\widehat{v}_T$.

For scenario 1 (see Figure 3), SIRds and SIRu perform well except for the 10th block for SIRu whereas SIRds recovers well the changes of direction in the 10th block. Moreover the image of the weights clearly indicates that this 10th block is aberrant. Note that classical SIR in each block provides a good estimated EDR direction. Take into account all the information of previous blocks allows SIRds to improve the estimation of the EDR direction from SIR in the current block.

Scenario 2: a drift occurs from the 10th block to the last block

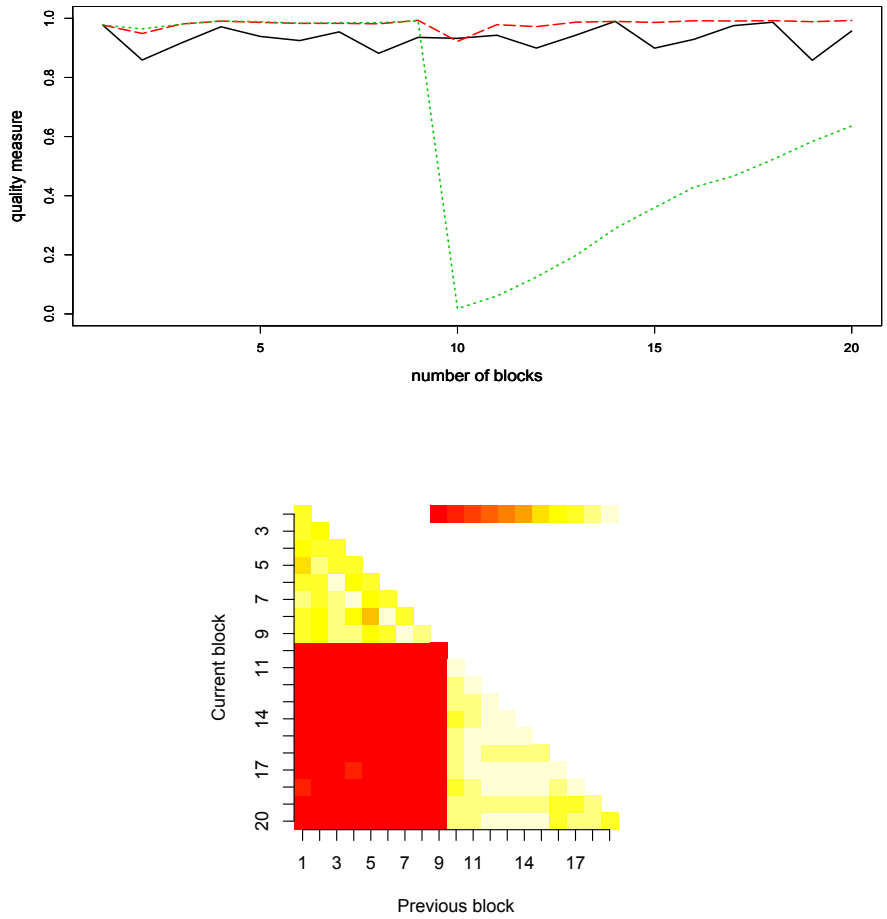


Figure 4: Numerical behavior of the SIRu and SIRds estimators for various scenario 2. On the top: plot of the quality measure $\cos^2(\widehat{\beta}_t, \beta_t)$ versus the number T of blocks (dashed red line for SIRds on the first t blocks, dotted green line for SIRu on the first t blocks, solid black line for SIR on block t only). On the bottom: image of the weights $\cos^2(\widehat{b}_t, \widehat{b}_T)$ used in the computation of the SIRds estimator $\widehat{v}_T$.

For scenario 2 (see Figure 4), the image of the weights clearly shows that there is a drift from the 10th block to the last one. The estimation of the true direction β_t for SIRds remains efficient after the 10th block whereas the estimation for SIRu falls after the 10th block and rises slowly after that. Here again the results obtained with SIRds are better than those

obtained by SIR based only on the current block because SIRds uses all the information of the first available blocks.

5 A real data illustration

As an illustration, we consider a nonlinear inverse problem in remote sensing. The goal is to estimate the physical properties of surface materials on the planet Mars from hyperspectral data. The method is based on the estimation of the functional relationship between some physical parameters Y and observed spectra X. For this purpose, a database of synthetic spectra is generated by a physical radiative transfer model. Bernard-Michel *et al.* (2009a) propose to reduce the high dimension of spectra ($p = 352$ wavelengths) with a regularized version of SIR. The need to regularize SIR in very high dimensions is well-known since the work of Zhong *et al.* (2005). Here, the empirical covariance matrix $\widehat{\Sigma}$ is replaced by $\widehat{\Sigma} + \lambda I_p$ where $\lambda > 0$, see Bernard-Michel *et al.* (2009b) for other types of regularization.

In practice, the database of synthetic spectra may be so large that it cannot be stored in a computer memory. Thus, a stream of smaller sub-databases is generated and we propose to apply our SIRds approach to this context.

Description of the data. We focus on an observation of the south pole of Mars at the end of summer, collected by the French imaging spectrometer OMEGA on board of 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 the generation of blocks of $n^* = 800$ synthetic spectra with the corresponding parameters. For the sake of simplicity, we limit ourselves to the study of the $T = 8$ first blocks. Besides, we focus on a terrain unit of strong CO_2 concentration determined by a classification method based on wavelets (Schmidt *et al.* (2007)), and the parameter of interest Y is the proportion of CO_2 ice.

Choice of the parameters. We choose $H = 19$ slices since it corresponds to the number of different values of Y simulated in the database. The regularization parameter is fixed to $\lambda = 0.00001$ thanks to a cross-validation procedure, see Bernard-Michel *et al.* (2009a) for further details.

The dimension K of the EDR space has been selected on the first block using the method

proposed in Liquet and Saracco (2012). The criterion used is the square trace correlation to study the closeness between two k-dimensional linear subspaces: the corresponding risk function is defined as

$$R_k = \mathbb{E}\left[\text{Trace}(P_k \widehat{P}_k)\right]/k, \tag{18}$$

where P_k denotes the orthogonal projector onto the space spanned by the first k basis vectors of E and $\widehat{P}_k$ is the orthogonal projector onto the space spanned by the first k vectors of $\widehat{E}$. This quantity R_k is only defined for any dimension k lower than or equal to the true dimension K of the EDR space. In our dimension reduction context, a value of R_k close to one indicates that the set of the k estimated linear combinations of $\mathbf{X}$ is close to the ideal set. So in terms of dimensionality, k is a feasible solution. On the other hand, a value of R_k perceptibly different from 1 means that this estimated set is slightly different from the ideal one, so the solution for the dimension is greater than k. Since R_K will converge to one as n tends to infinity (for the true dimension K), then, for a fixed n, a reasonable way to assess whether an EDR direction is available is to look at how much R_k departs from one. From a computational point of view, we require consistent estimates $\widehat{R}_k$ of R_k, so the feasible solution for the dimension can be obtained by computing the values of $\widehat{R}_k$ for $k = 1$ to p and observing how much it departs from one. Liquet and Saracco (2012) use a bootstrap estimator of this criterion. Note that in our application, the number of slices is fixed since the dependent variable Y is discrete ($H = 19$). Hence we slightly adapt here the criterion of Liquet and Saracco (2012) to select only the dimension K (and not to determine the couple (H, K) of parameters). Let $\mathcal{B}$ be the number of bootstrap replications of the data of the first block of size n_1. Let us consider $s^{(b)} = \left\{(\mathbf{X}_i^{(b)}, Y_i^{(b)}), i = 1, \ldots, n_1\right\}$ a non-parametric bootstrap sample replication. A naïve bootstrap estimate of the mean square risk function is defined by:

$$\widehat{R}_k = \frac{1}{\mathcal{B}} \sum_{b=1}^{\mathcal{B}} \widehat{R}_k^{(b)} \tag{19}$$

where $\widehat{R}_k^{(b)} = \text{Trace}\left(\widehat{P}_k \widehat{P}_k^{(b)}\right)/k$ and $\widehat{P}_k^{(b)}$ is the projector onto the subspace spanned by the first k eigenvectors of the matrix of interest, which is obtained from the bootstrap replication sample $s^{(b)}$. In practice, the criterion $\widehat{R}_k$ will be computed for all $k = 1 \ldots, p$ whereas from a theoretical point of view the R_k is only defined for $k = 1, \ldots, K$. The objective of the graphical method is to provide a practical choice of the dimension K of the model thanks to this bootstrap estimated version of the criterion. To do this, the proposed method consists

in evaluating the $\widehat{R}_k$ for all $k \in \{1, \ldots, p\}$ and then in observing how much it departs from one. Note that in Figure 5 (on the right), since p is large, we only plot $\widehat{R}_k$ versus k for $k \in \{1, \ldots, 15\}$. The best choice will be the value $\widehat{K}$ which gives a value of $\widehat{R}_k$ close to one, such that $\widehat{K} << p$. In practice, since there is no objective criterion to establish when a departure from one is close, a visual expertise of the plot of the $\widehat{R}_k$ versus k allows the best value to be chosen. It is also useful to provide, for each k, the boxplot of the $\widehat{R}_k^{(b)}$'s to have a look on the stability (or not) of the corresponding k-dimensional linear subspace. In Figure 5 (on the left), we clearly observe the one-dimensional and two-dimensional EDR spaces are stable, while the subspaces of greater dimension ($k \geq 3$) are more unstable. Following the results of Figure 5, it appears that $\widehat{K} = 2$ seems to be an appropriate choice in terms of stability of the estimated EDR space. This choice is also confirmed by the eigenvalues scree plot which presents a jump after $K = 2$, see Figure 6.

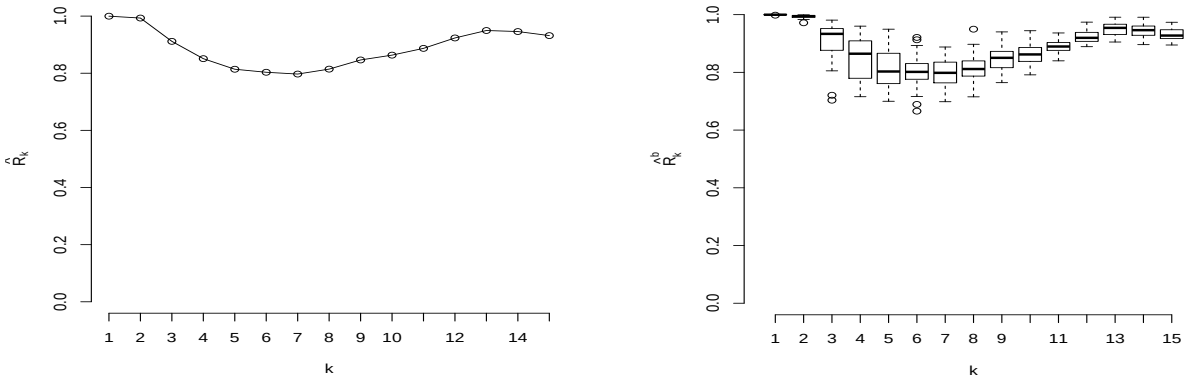


Figure 5: Choice of the dimension based on the stability of the estimated EDR space.

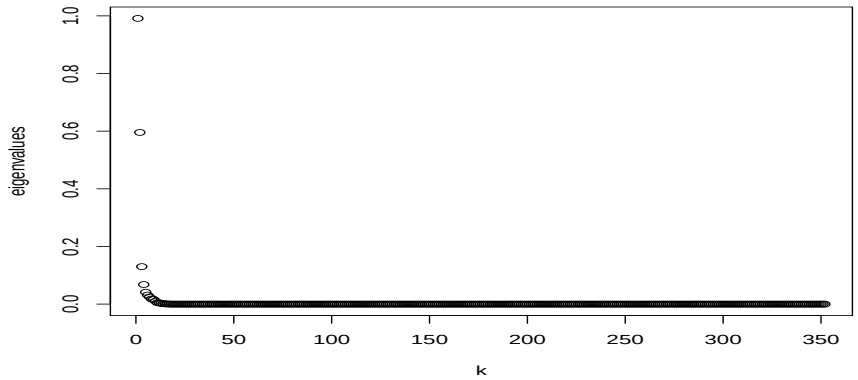


Figure 6: Eigenvalue scree plot of the first block.

The plots of the proportion of CO_2 ice Y versus the first and second EDR indices depicted on Figure 7 exhibit a nice structure. Similar results are obtained by plotting the dependent variable Y as a bivariate function of both first and second indices (see the graphic on the left of Figure 8). Let us also highlight that these structures are stable, *i.e.* they have also been observed on the other blocks. At the opposite, the plot of the proportion of CO_2 ice versus the third and fourth EDR indices (see the graphic on the right of Figure 8) does not exhibit any structure and is very different from one block to another. These graphical diagnostics therefore confirm the choice of $\widehat{K} = 2$.

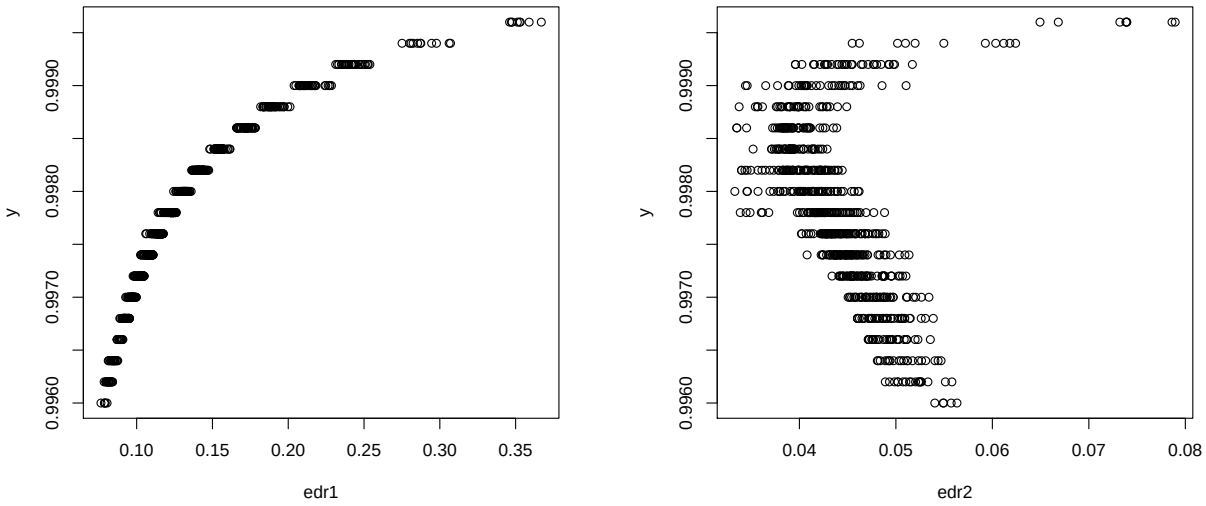


Figure 7: Plots of the dependent variable versus the first EDR index (on the left) and the second EDR index (on the right).

SIR data stream on all the blocks. Here, the true directions are unknown but it is still possible to assess the stability of the estimated EDR subspace by representing the weights $\cos^2(\widehat{b}_t, \widehat{b}_T)$ used in equation (7) on an image (see Figure 9 to visualize the weigths used in SIRds). The lighter is the color, the larger is the weight. It appears that all the squared cosines are larger than 0.98. The EDR subspace computed on the running block is very close to the EDR subspace computed on the previous ones.
Moreover, it can also be checked that SIRds and SIRu yield similar EDR subspaces with squared cosines larger than 0.999 at each step t=1,...,8. The plots of the proportion of CO_2 ice versus the first and second EDR indices computed on all the blocks are very similar to these of Figures 7 and 8. Finally, Figure 10 represents the coordinates of each of the two first EDR directions. It indicates which wavelengths are important (nonzero coordinates) for estimating the proportion of CO_2 ice.

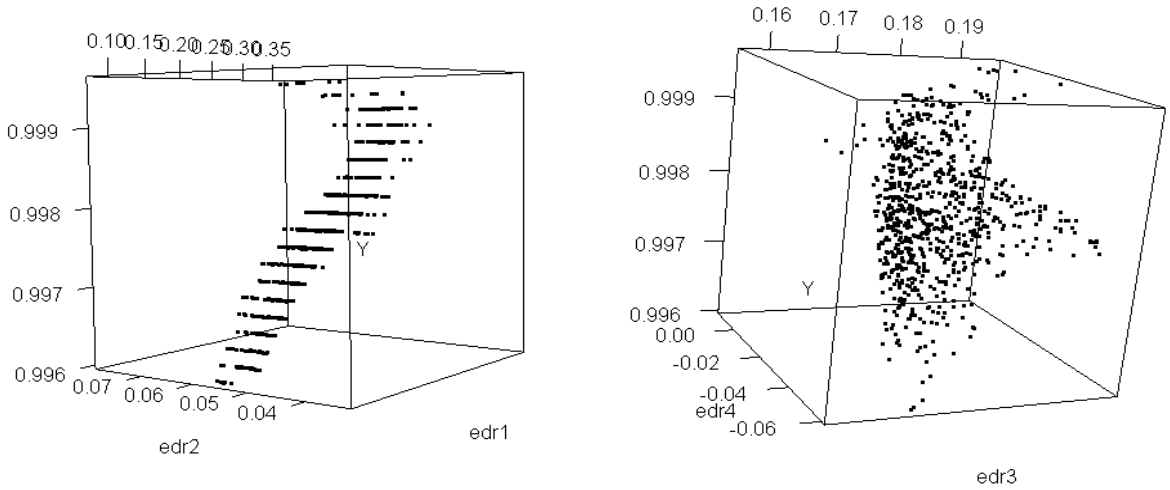


Figure 8: 3D-plot of the dependent variable versus the first two EDR indices (on the left); 3D-plot of the dependent variable versus the third and fourth EDR indices (on the right)

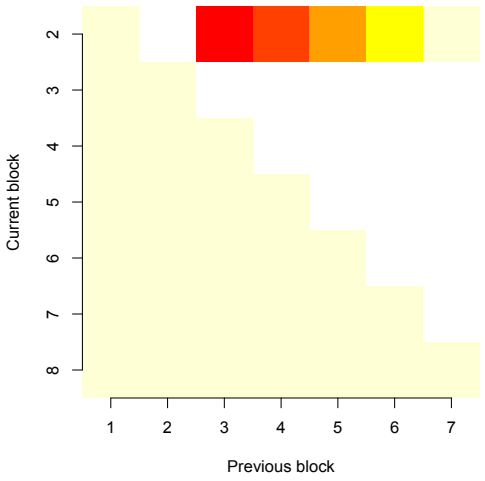


Figure 9: Weigths used in SIRds.

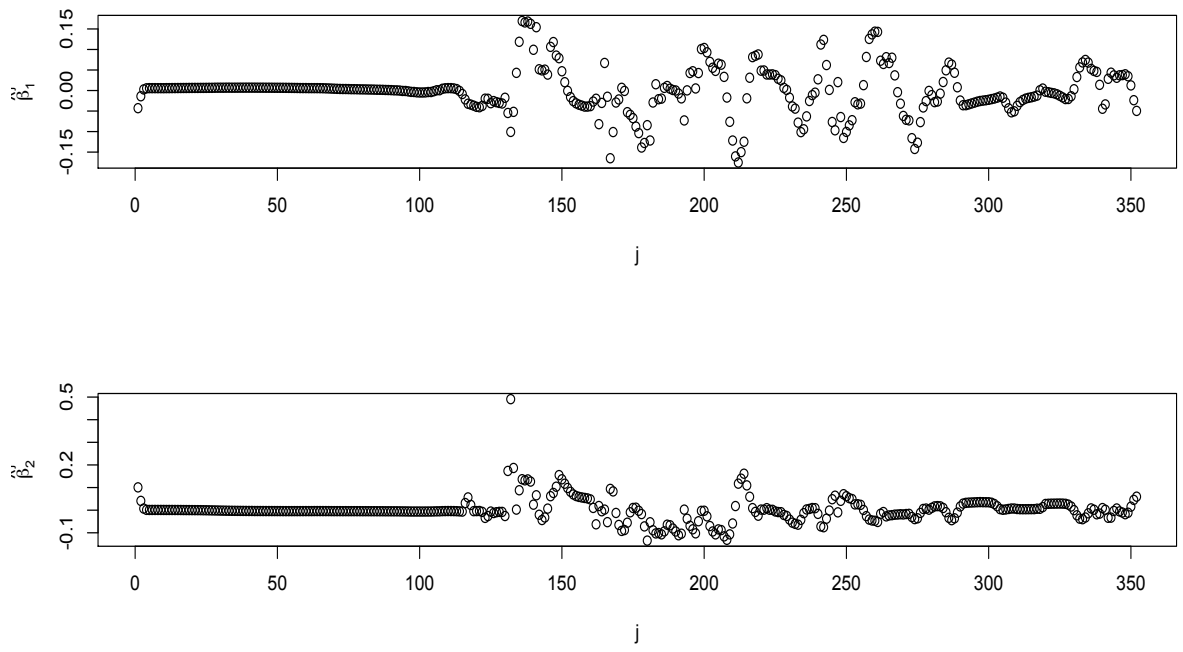


Figure 10: Final first and second EDR directions obtained with SIRds on all the blocks.

6 Concluding remarks

We present in this paper population and sample versions of SIRds based on usual SIR method for single index model or multiple indices model. We give some asymptotic results. The proposed approach SIRds performs well on simulated data and has been applied in the real data application. In this application, we use an extension of SIRds based on a regularized SIR version instead of usual SIR. It is also possible to use other alternative methods instead of SIR such as SIR-II, SAVE or SIR$_\alpha$ for example. These approaches are based on some properties of the conditional variance of X given $T(Y)$, see for instance Li (1991) or Shao *et al.* (2009). Another possible extension is to investigate the case of a multivariate response variable Y: the idea is then to use multivariate SIR approach instead of univariate SIR methods, see for instance Barreda *et al.* (2007), Saracco (2005) or Lue (2009).

References

Barreda, L., Gannoun, A., Saracco, J. (2007). Some extensions of multivariate SIR. *Journal of Statistical Computation and Simulation*, 77(1-2), 1-17.

Barrios, M.P.; Velilla, S. (2007). A bootstrap method for assessing the dimension of a general regression problem. *Statist. Probab. Lett.*, 77(3), 247-255.

Bernard-Michel, C., Douté, S., Fauvel, M., Gardes, L. and Girard, S. (2009a). Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse

Regression. *Journal of Geophysical Research - Planets*, 114, E06005.

Bernard-Michel, C., Gardes, L. and Girard, S. (2009b). Gaussian Regularized Sliced Inverse Regression, *Statistics and Computing*, 19, 85-98.

Chavent, M., Kuentz, V., Liquet, B. and Saracco, J. (2011). A sliced inverse regression approach for a stratified population. *Communications in statistics - Theory and methods*, 40, 1-22.

Chen, C-Hand Li, K-C (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica*, 8, no. 2, 289-316.

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

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

Ferré, L. (1998). Determining the dimension in sliced inverse regression and related methods. *J. Amer. Statist. Assoc.*, 93(441), 132-140.

Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, 21, 867-889.

Harville, D.A. (1999). *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag.

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

Liquet, B., Saracco, J. (2008). Application of the bootstrap approach to the choice of dimension and the α parameter in the SIR_α method. *Communications in Statistics - Simulation and Computation*, 37(6), 1198-1218.

Liquet, B. and Saracco, J. (2012). A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches. *Comput. Stat.*, 27, 103-125.

Lue, H-H. (2009). Sliced inverse regression for multivariate response regression. *J. Statist. Plann. Inference*, 139(8), 2656-2664.

Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Communications in Statistics - Theory and Methods*, 26(9), 2141-2171.

Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on SIR_α. *Journal of Multivariate Analysis*, 96, 117-135.

Schmidt, F., Douté, S. and Schmitt B. (2007). Wavanglet: An efficient supervised classifier for hyperspectral images. *Geoscience and Remote Sensing, IEEE Transactions*, **45**(5), 1374-1385.

Schott, J.R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, 89(425), 141-148.

Shao, Y., Cook, R. D.and Weisberg, S. (2009). Partial central subspace and sliced average variance estimation. *J. Statist. Plann. Inference*, 139(3), 952-961.

Tyler, D.E. (1981). Asymptotic inference for eigenvectors. *The Annals of Statistics*, 9(4), 725-736.

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

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