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[5,6], Thi Mong Ngoc Nguyen[7] 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,jerome.saracco}@math.u-bordeaux1.fr

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

[3] Inria Grenoble Rhône-Alpes & LJK, 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] Univ. Bordeaux, ISPED, centre INSERM U-897-Epidémiologie-Biostatistique,
Bordeaux, F-33000, FRANCE

[6] INSERM, ISPED, centre INSERM U-897-Epidémiologie-Biostatistique, Bordeaux, F-33000, FRANCE
e-mail: Benoit.Liquet@isped.u-bordeaux2.fr

[7] 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 blocks 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 of pooling all the observed blocks and estimating the EDR direction by the SIR (Sliced Inverse Regression) method. But in practice, some disadvantages appear such as the storage of the blocks and the running time for large sample sizes. To overcome these drawbacks, we propose an adaptive SIR estimator of β based on the optimization of a quality measure. The corresponding approach is faster both in terms of computational complexity and running time, and provides data storage benefits. The consistency of our estimator is established and its asymptotic distribution is given. An extension to multiple indices model is proposed. A graphical tool is also provided in order to detect changes in the underlying model, i.e., drift in the EDR direction or aberrant blocks in the data stream. A simulation study illustrates the numerical behavior of our estimator. Finally, an application to real data concerning the estimation of physical properties of the Mars surface is presented.

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

1 Introduction

Regression models are used to highlight the relationship between one response variable Y and a p-dimensional regressor X. When p is large, parametric as well as nonparametric regression methods are faced with the so-called curse of dimensionality. One way to overcome this problem is to use dimension reduction techniques which aim at replacing X with a projection onto a smaller dimension subspace. For the sake of simplicity, let us first consider a semiparametric single index model.

More precisely, let $(Y, X) \in \mathbb{R} \times \mathbb{R}^p$ with $\mathbb{E}(X) = \mu$, covariance matrix $\mathbb{V}(X) = \Sigma$ and $\mathbb{E}((X'X)^2) < \infty$. Let ε be a real random error independent of X. It is assumed that Y depends on X only through $X'\beta$ according to the following dimension reduction model:

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

where the real-valued link function f and the parameter $\beta \in \mathbb{R}^p$ are unknown. Since the function f is unspecified, only the linear subspace spanned by β may be identified. This subspace is usually called Dimension Reduction (DR) subspace. Clearly, there are many DR subspaces for a fixed model. This smallest DR subspace, which is the intersection of all DR subspaces, is called the Effective Dimension Reduction (EDR) subspace (see Duan and Li, 1991) or the central subspace (see Cook, 2007, section 8.3 for details and discussion). The former term EDR is adopted in this paper, in order to be consistent with the closest references to our work. To estimate the EDR subspace, Duan and Li (1991) introduced a link-free and distribution-free method called SIR (Sliced Inverse Regression). The basic principle of SIR is to reverse the role of Y and X and to study the geometric property of the first inverse moment $\mathbb{E}(X|Y)$; see for instance Li (1991), Chen and Li (1998), Zhu *et al.* (2007) among others. 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 expectation.

In this paper we focus on data streams, that is, data arriving sequentially by blocks in a stream. Since an increasing number of data sets are not fixed, but evolve over time, the study of dimension reduction model in this case appears to be very useful. It is assumed 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 the model in (1). A first simple approach to estimate the EDR direction consists of 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 lies in the storage of the blocks since the size of

the dataset increases considerably with the number of blocks.

To avoid this, we propose an adaptive SIR method based on the optimization of a criterion which assesses the proximity between the EDR directions of each block and the "global" direction of the overall data stream. This optimization problem is equivalent to the eigen-decomposition of a symmetric $p \times p$ matrix. It involves weighting terms which evaluate the squared cosines between the current direction of the block and the previous ones. This leads to an adaptive procedure which can detect aberrant blocks and which can also recover some possible changes of EDR direction in the data. Another main advantage of this method is in terms of storage since it is not necessary to stock 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 streams (named SIRds). Both population and sample versions are described. This approach is extended to multiple indices models in Section 3. A simulation study is carried out in Section 4 in order to illustrate the behavior of our estimator and to compare it to classical SIR applied to all blocks. The superiority of SIRds is evaluated in terms of computational complexity and running time. SIRds adaptivity is investigated for recovering potential changes of direction in some blocks. In Section 5, the proposed adaptive SIRds method is used to study the physical properties of surface materials on planet Mars from hyperspectral images. Concluding remarks are given in Section 6.

Finally, let us note that a short french summary of this work (without proofs, real data application and with only few simulation results) can be found in the proceedings of the "Journées de Statistique" conference, see Chavent *et al.* (2012).

2 An adaptive SIR estimator for data stream: SIRds

Let us first recall in Section 2.1 the population and sample versions of SIR based on a single block. Then, the population and sample versions of our adaptive method for a data stream (of blocks) are presented in Section 2.2. Asymptotic results for the SIRds estimator are given in Section 2.3. Finally, Section 2.4 provides some comments on computational complexity and data storage for SIRds and the usual SIR applied to the union of all the blocks.

2.1 Recall on SIR in block t

In this section, let us focus on a single block t. We first present the population version of SIR and then derive its sample version.

The population version of SIR relies on the following linear condition:

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

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), Duan and 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 $\Sigma^{-1}\Gamma$ associated with the non-null eigenvalue is an EDR direction (i.e., is collinear with β) where $\Gamma = \mathbb{V}(\mathbb{E}(X|T(Y)))$. 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 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 can be written as:

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

Using the n_t observations (X_i, Y_i) within block t, it is straightforward to estimate the matrix Γ by substituting theoretical versions of the moments by their empirical counterparts. Let $\widehat{\Gamma}$ denote this estimator. Therefore, one obtains the estimated EDR direction $\widehat{u}_t$ as the eigenvector associated with the largest eigenvalue of $\widehat{\Sigma}^{-1}\widehat{\Gamma}$ where $\widehat{\Sigma}$ is an estimator of Σ. Let us highlight that $\widehat{\Sigma}$ is assumed to be invertible which implies that $n_t > p$. 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, let us consider T sequentially arriving blocks of data. From each block t, we can obtain b_t, the I_p-normalized EDR direction computed with SIR as 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 T available blocks. Averaging the vectors b_t is not ideal since only the direction of b_t is identifiable: even if the

vectors b_t have unit length and are collinear, their mean can be zero if they do not share the same orientation.

Another way is to recover the direction "most collinear" with the vectors $b_1, \ldots, b_T$. The collinearity between two unit vectors a and b is measured by $m(a,b) = \cos^2(a,b) = (a'b)^2$. The following optimization problem is then considered:

$$\max_{a \in \mathbb{R}^p} \sum_{t=1}^{T} w_t m(b_t, a) \quad \text{s.t. } ||a|| = 1, \tag{3}$$

where the w_t's are positive weights such that $\sum_{t=1}^{T} w_t = 1$. These weights provide more flexibility in the application: for instance, they allow the algorithm to take into account different block sizes. This approach is suitable but suffers from not being adaptive with respect to changes in the parametric part of the model (i.e., β moves to $\beta^* \neq \beta$ in the underlying model in block T). In the following, an adaptive version of the maximization problem (3) is introduced in order to take into account the possible evolution of the parametric part of the semiparametric model in each block.

Population version of SIRds. To give an adaptive SIR approach for data streams, let us add in (3) the weights $m(b_t, b_T)$. These weights will examine if the "new" block T provides the same information as the previous blocks, that is if the EDR direction b_T obtained in block T is close to the directions $b_t,\ t = 1, \ldots, T-1$ of the previous blocks. The following optimization problem is thus considered:

$$\max_{a \in \mathbb{R}^p} \sum_{t=1}^{T} w_t m(b_t, b_T) m(b_t, a) \quad \text{s.t. } ||a|| = 1. \tag{4}$$

From now on, let us define $Q(a, b_1, \ldots, b_T) := \sum_{t=1}^{T} w_t m(b_t, b_T) m(b_t, a)$ for all $a \in \mathbb{R}^p$.

Theorem 1

(i) The solution $v_T \in \mathbb{R}^p$ of the maximization problem (4) is the normalized principal eigenvector of

$$M_T = \sum_{t=1}^{T} w_t b_t b_t' m(b_t, b_T) \tag{5}$$

associated with the largest eigenvalue $Q(v_T, b_1, \ldots, b_T)$.

(ii) Under linearity condition (LC) and model (1), v_T is an EDR direction.

The proof is provided in the Appendix. Under (LC) and model (1), note that $Q(v_T, b_1, \ldots, b_T) = 1$, and otherwise $Q(v_T, b_1, \ldots, b_T) \in [0, 1]$. The closer to one this measure is, the closer the

linear subspace $\text{Span}(v_t)$ is to all the T linear subspaces $\text{Span}(b_t)$, $t = 1, \ldots, T$. Let us also highlight that, ignoring the adaptive term $m(b_t, b_T)$ in (5), one obtains the matrix of interest used in the SIR approach for a stratified population developed by Chavent *et al.* (2011).

Let us now explain the advantage of this adaptive SIRds version on a simple example. Let us assume that the underlying regression model for the first $T-1$ blocks is given in (1). It also assumed 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 with β for the first $T-1$ blocks as mentioned above, which is the "true" direction of the underlying model for these blocks. When block T arrives, from the population point of view, we have: $m(b_t, b_T) = 0$ for $t = 1, \ldots, T-1$ (since each b_t, for $t = 1, \ldots, T-1$, is collinear with β and b_T is collinear with β^*). Then $M_T = w_T b_T b_T'$ since $m(b_T, b_T) = 1$. Finally, an EDR direction collinear with β^* is obtained, which is the "true" direction of the current block. To conclude, the SIRds approach allows

- either 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,
- or to use the available information from all blocks with a common direction, and then to provide the common EDR direction.

A visualization of the weights $m(b_t, b_{\mathcal{T}})$ for $t = 1, \ldots, \mathcal{T}$ and $\mathcal{T} = 2, \ldots, T$ will be very useful for the user to detect whether or not aberrant blocks or drifts are present in the data stream. Section 4 gives some graphical illustrations on various scenarios of data streams.

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

$$\widehat{M}_T = \sum_{t=1}^{T} w_t \widehat{b}_t \widehat{b}_t' \, m(\widehat{b}_t, \widehat{b}_T). \tag{6}$$

One possible choice for the weights w_t can be $w_t = \frac{n_t}{\sum_{j=1}^{T} n_j}$ for $t = 1, \ldots, T$, that is, the relative size of the block t. The next section provides some asymptotic results for this estimator: weak consistency and asymptotic normality.

2.3 Asymptotic results

The following assumptions are necessary to state our asymptotic results on the convergence of the estimated EDR direction by SIRds. Let us consider a fixed number T of blocks and a sample size $n = \sum_{t=1}^{T} n_t$ which tends to ∞. Let H_t be the number of slices in block t and let $n_{h,t}$ be the number of observations in the hth slice such that $\sum_{h=1}^{H_t} n_{h,t} = n_t$ is the total number of observations in 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 $h = 1, \ldots, H_t$, $n_{h,t} \to \infty$ (therefore $n_t \to \infty$ and $n \to \infty$).

Theorem 2 *Under model (1), linearity condition (LC) and assumptions (A1)-(A3), one has:*

(*i*) $\widehat{v}_T = v_T + O_p(\underline{n}^{-1/2})$ *where* $\underline{n} := \min(n_t,\ t = 1, \ldots, T)$.

(*ii*) *If, moreover,* $n_t = n/T =: \overline{n}$ *for all* $t = 1 \ldots, T$, *then*

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

where the expression for Γ_W *is given in* (19).

The proof is postponed to the Appendix.

2.4 Computational complexity and data storage

Computational complexity. For the sake of simplicity, let us assume that each block has the same sample size $\overline{n}$. In such a case, the computational complexity of SIR computed on one block is of order $p^2(\overline{n} + p)$ (denoted as $O(p^2(\overline{n} + p))$ hereafter). The first term $(\overline{n}p^2)$ corresponds to the cost of computing the empirical covariance matrix $\widehat{\Sigma}$, the second term (p^3) is the cost for computing the matrix $\widehat{\Sigma}^{-1}\widehat{\Gamma}$ and its eigendecomposition.

Our goal is to show that the SIRds approach performs faster than the sequential SIR method which consists of computing SIR on the union of the j first blocks for $j = 1, \ldots, T$. Clearly, the computational complexity of sequential SIR is

$$O((\overline{n}p^2 + p^3) + (2\overline{n}p^2 + p^3) + \cdots + (T\overline{n}p^2 + p^3)) = O(Tp^2(\overline{n}T + p)), \tag{7}$$

since it requires T computations of SIR on blocks of increasing sizes. The computational complexity of SIRds is

$$O((\overline{n}p^2 + p^3) + (\overline{n}p^2 + 2p^2 + p^3) + \cdots + (\overline{n}p^2 + Tp^2 + p^3)) = O(Tp^2(\overline{n} + T + p)). \quad (8)$$

The terms $O(\overline{n}p^2)$ correspond to the calculation of the covariance matrices $\widehat{\Sigma}$ in each block t. The terms $O(\mathcal{T}p^2)$ represent the calculation of the matrix $\widehat{M}_{\mathcal{T}}$ when $\mathcal{T} > 1$ blocks are available. Finally the terms $O(p^3)$ stand for the cost of the eigendecompositions.

From (7) and (8), it appears that the complexity of sequential SIR depends on the product $\overline{n}T$ while the complexity of SIRds depends on the sum $\overline{n}+T$. From the computational complexity point of view, SIRds outperforms sequential SIR as soon as $\frac{1}{T} + \frac{1}{\overline{n}} \ll 1$, which is often the case in practical situations.

Data storage. Sequential SIR requires the storage of the whole matrix of regressors (i.e. T blocks of size $\overline{n}p$), its storage load is thus $O(T\overline{n}p)$. 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 $O((\overline{n}+T)p)$. Similarly, for the data storage comparison, SIRds has better performance than sequential SIR as soon as $\frac{1}{T} + \frac{1}{\overline{n}} \ll 1$, since the product $\overline{n}T$ is replaced by the sum $\overline{n} + T$.

3 Extension to multiple indices model

In this section, an extension of SIRds approach to the case of a multiple indices model is proposed. Let us first recall the corresponding model introduced by Li (1991).

3.1 Recall on the multiple indices SIR on block t

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 $X'\beta_k,\ k = 1, \ldots, K$:

$$Y = g(X'\beta_1, \ldots, X'\beta_K, \varepsilon). \quad (9)$$

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 subspace $E = \text{Span}(\beta_1, \ldots, \beta_K)$. As for the single index model, SIR is used to seek for a

basis of the EDR subspace for each block. In order to get theoretical results, the linearity condition (LC) is replaced by:

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

For the block t, assuming the linearity condition (LC') and model (9), the eigenvectors $u_{1,t},\ldots,u_{K,t}$ associated with the largest K eigenvalues of the matrix $\Sigma^{-1}\Gamma$ are EDR directions, where the matrix Γ has been defined in (2). Note that the number H of slices 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. They are stored in the $p\times K$ matrix $\mathbb{B}_t=[b_{1,t},\ldots,b_{K,t}]$.

3.2 Population and sample versions of SIRds

Since the dimension K of the EDR space is greater than one, the optimization problem (4) can be adapted as follows. Direction b_t is replaced by an I_p-orthonormal basis $\mathbb{B}_t$ of the EDR subspace 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:

$$m(\mathbb{B}_t,\mathbb{B}_T)=\frac{\text{Trace}(P_tP_T)}{K},$$

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 onto $\text{Span}(\mathbb{B}_l)$, the EDR subspace obtained from the block l (equal to t or T). This measure takes its values in [0,1]. Note that $m(\mathbb{B}_t,\mathbb{B}_T)=1$ when $\text{Span}(\mathbb{B}_t)=\text{Span}(\mathbb{B}_T)$. The closer this measure is to one, the closer the linear subspace $\text{Span}(\mathbb{B}_t)$ is to the linear subspace $\text{Span}(\mathbb{B}_T)$.

Let $\mathbb{A}$ be a $p\times K$ matrix such that $\mathbb{A}'\mathbb{A}=I_K$. 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 subspaces $\text{Span}(\mathbb{B}_1)$, …, $\text{Span}(\mathbb{B}_T)$ respectively obtained from the T available blocks:

$$Q(\mathbb{A},\mathbb{B}_1,\ldots,\mathbb{B}_T)=\sum_{t=1}^{T}w_tm(\mathbb{B}_t,\mathbb{B}_T)m(\mathbb{A},\mathbb{B}_t),$$

where $\forall t, w_t\geq 0$ 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 this measure is to one, the closer the linear subspace $\text{Span}(\mathbb{A})$ is to all the T linear subspaces $\text{Span}(\mathbb{B}_t),\ t=1,\ldots,T$.

Population version of SIRds. Similarly to the single index case, let us now deal with the maximization problem:

$$\max_{\mathbb{A}} Q(\mathbb{A}, \mathbb{B}_1, \ldots, \mathbb{B}_T) \quad \text{s.t. } \mathbb{A}'\mathbb{A} = I_K. \tag{10}$$

The following theorem provides the solution of (10) and establishes the link with the EDR subspace.

Theorem 3

(i) The solution $\mathbb{V}_T = [v_{1,T}, \ldots, v_{K,T}]$ of the maximization problem (10) is an I_p-orthonormal basis of the K-dimensional eigenspace associated with the K largest eigenvalues $\lambda_1, \ldots, \lambda_K$ of the $p \times p$ matrix

$$\mathbb{M}_T = \sum_{t=1}^{T} w_t \frac{\mathbb{B}_t \mathbb{B}_t'}{K} m(\mathbb{B}_t, \mathbb{B}_T). \tag{11}$$

Moreover $Q(\mathbb{V}_T, \mathbb{B}_1, \ldots, \mathbb{B}_T) = \lambda_1 + \cdots + \lambda_K$.

(ii) Under linearity condition (LC') and model (9), the column vectors of $\mathbb{V}_T$ form an I_p-orthonormal basis of the EDR subspace E.

The proof can be found in the Appendix. Under (LC') and model (9), note that $Q(\mathbb{V}_T, \mathbb{B}_1, \ldots, \mathbb{B}_T) = 1$, and otherwise, $Q(\mathbb{V}_T, \mathbb{B}_1, \ldots, \mathbb{B}_T) \in [0, 1]$. The closer this measure is to one, the closer the linear subspace $\text{Span}(\mathbb{V}_t)$ is to all the T linear subspaces $\text{Span}(\mathbb{B}_t),\ t = 1, \ldots, T$.

Sample version of SIRds. The corresponding sample version can now be briefly described. For each block t, using the corresponding sample, a $\widehat{\Sigma}$-orthogonal basis of the EDR subspace is first estimated with SIR. The basis vectors are stored in the matrix $\widehat{\mathbb{U}}_t$. Then the first K eigenvectors of the matrix $\widehat{\mathbb{U}}_t \widehat{\mathbb{U}}_t'$ are computed and stored in the matrix $\widehat{\mathbb{B}}_t$. They form an I_p-orthogonal basis of the estimated EDR subspace. Finally the estimator of $\mathbb{M}_T$ is constructed as follows:

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

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 subspace denoted by $\widehat{E}$.

3.3 Asymptotics

Under the linearity condition (LC') and the assumptions (A1)-(A3) given in Section 2, for single index model, it can be shown that the estimated EDR basis converges to an EDR basis at root $\underline{n}$ rate, that is the estimated EDR subspace $\widehat{E}$ converges to the true EDR subspace:

Theorem 4 *Assume that* $\lambda_1 > \lambda_2 > \cdots > \lambda_K > 0$. *Under linearity condition (LC') and assumptions (A1)-(A3), we have*

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

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

The proof can be found in the Appendix. Note that, as for the single index model, using the Delta-method, and asymptotic results of Tyler (1981) and Saracco (1997), the asymptotic normality of the eigenprojector onto the estimated EDR subspace can be obtained, as well as the asymptotic distribution of the estimated EDR directions, associated with eigenvalues assumed to be different.

3.4 Discussion on the choice of dimension K

Since the beginning of this section, the dimension K of the EDR subspace was assumed to be known. However in most applications the number K of indices is unknown a priori and, hence, must be estimated from the data. Several approaches have been proposed in the literature for SIR. Some are based on hypothesis tests of 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 squared trace correlation between the true EDR subspace 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 datastream context, under assumption (LC'), the dimension K is common to all the blocks since it is assumed that the underlying model in each block relies on the same EDR subspace E. From the theoretical point of view, it can thus be estimated from any block or from any combinations of blocks. From the practical point of view, we recommend choosing the dimension K using classical SIR in the first block. If a block appears to be aberrant, the user has to again determine the dimension in order to confirm that the true dimension of the whole EDR subspace is still K. In the example given in Section 5, the

graphical approach of Liquet and Saracco (2012) is described and used to determine the suitable dimension K.

4 A simulation study

A simulation study is carried out to evaluate the numerical performance of the proposed method. First, Section 4.1 describes the single index model used in this simulation study and the estimation methods (note that in the real data application of Section 5, we shall consider and estimate a multiple indices model with $K = 2$). In Section 4.2, the numerical results obtained with SIRds are compared with those provided by classical SIR approach. The mean computational times obtained with both approaches are also exhibited. Finally Section 4.3 illustrates the behavior of SIRds on various scenarios in which some blocks do not have the same EDR direction.

4.1 Simulated model and estimation methods

In this simulation study, two semiparametric regression models are considered:

$$Y = \frac{3}{10}(X'\beta)^3 + \epsilon, \tag{12}$$

and

$$Y = \sin(X'\beta) + |X'\beta|\epsilon, \tag{13}$$

where X follows the p-dimensional normal distribution $\mathcal{N}_p(0_p, \Sigma)$ with a covariance Σ arbitrarily chosen as follows: a matrix A is randomly filled using the uniform distribution on $[-1, 1]$, then $\Sigma = AA' + I_p$ in order to avoid possible problems of inversion of Σ. The error term ϵ follows the normal distribution $\mathcal{N}(0, \sigma^2)$ and is independent of X. Model (12) is homoscedastic while model (13) is a heteroscedastic one. We set $p = 10$, $\beta = (1, -1, 2, -2, 0, \ldots, 0)'/\sqrt{10}$ and $\sigma = 0.5$. In the following, for each model and a fixed covariance matrix Σ, data streams of $T = 20$ blocks are generated with $\overline{n} = 200$. One can observe in Figure 1 the scatterplots of the true index $X'\beta$ versus Y for models (12) and (13) using data from one block. Note that the corresponding empirical mean (over 500 replications) of signal-to-noise ratio in model (12) (resp. model (13)) is equal to 6.48 (resp. 2.78).

For each model and for each data stream, the EDR direction is estimated as follows. At the arrival of each block t ($t = 1, \ldots, T$), the EDR direction is computed with SIRds based on these first available t blocks. The EDR direction is also calculated with the classical

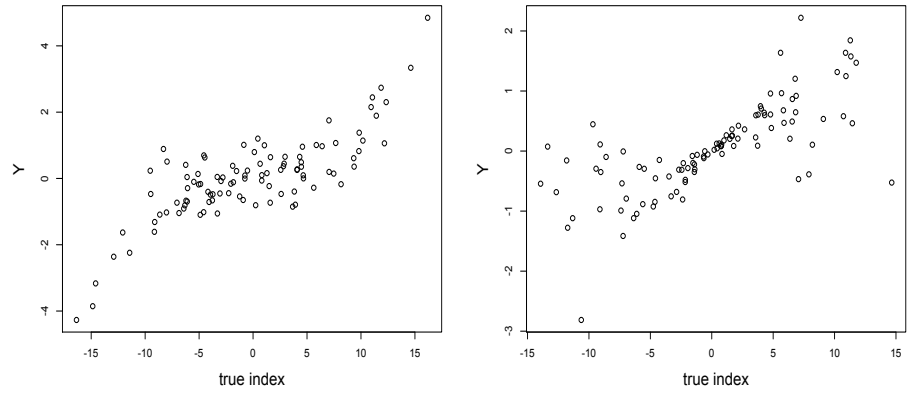

Figure 1: Scatterplots of the true index $X'\beta$ versus Y for model (12), on the left, and (13), on the right, using data of one block.

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.

4.2 Numerical results and running times

First, for each model, numerical results obtained with SIRds and SIRu approaches are compared using the quality measure m, that is the squared cosine between the true EDR direction and the estimated one. Then, we focus on the running time of these approaches to determine which is fastest. The experiments have been conducted using the R language on a laptop with a 2.53GHz processor.

Comparison of SIRds and SIRu approaches. Here our aim is to compare quality measures of EDR directions estimated with SIRds and SIRu. For each model, $B = 500$ data replications of a data stream of size $T = 20$ blocks are generated as previously. Figure 2 shows boxplots of the quality measure of the corresponding estimated EDR directions for $T = 1, 5, 10, 15, 20$ blocks. Let us first remark that both models provide very similar numerical results. Note that in the case $T = 1$ (only one block), the two approaches SIRds and SIRu are obviously equivalent to the usual SIR.

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. 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

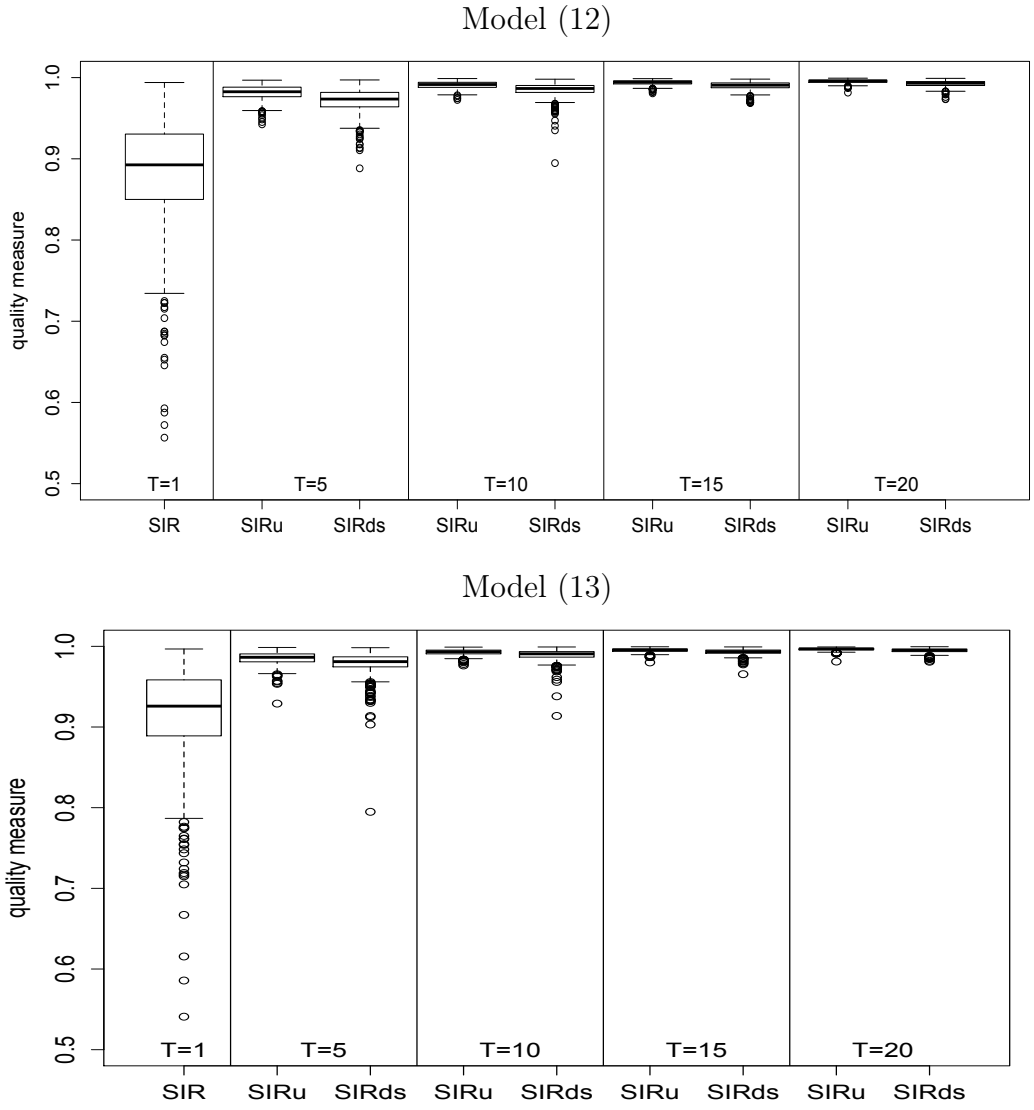

Figure 2: Boxplots of the squared cosines between the true EDR direction and the EDR directions estimated with SIRu and SIRds for different values of T.

mentioned previously, the disadvantage of SIRu is clearly the storage of all the blocks. SIRds is advantageous in keeping 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 loss of quality in the estimation of the EDR directions.

Running time. The running time (in seconds) of SIRds approach are now compared with sequential SIR (which has been defined in Section 2.4 as a sequential use of SIRu). More precisely, for a data stream of T blocks, the running times were 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.

From model (12), $B = 500$ data streams are generated for various values of the dimension p of X, the size $\overline{n}$ of each block and the total number T of blocks in the data stream. Then the computational times are measured for both methods, SIRds and sequential SIR. Unsurprisingly one can observe in Figure 3 that the dimension p noticeably favors SIRds versus sequential SIR while the number T of blocks and the block size $\overline{n}$ hugely penalize the sequential SIR approach in comparison with SIRds.

4.3 Adaptation to changes in the underlying model

In this section, the assumption that the model is the same in all the blocks is relaxed. From now on, the slope parameter β in model (12) is thus indexed with t. In order to illustrate the adaptivity of SIRds in comparison with SIRu in such cases, the following two scenarios are considered.

For each scenario, $T = 20$ blocks are generated as described below:

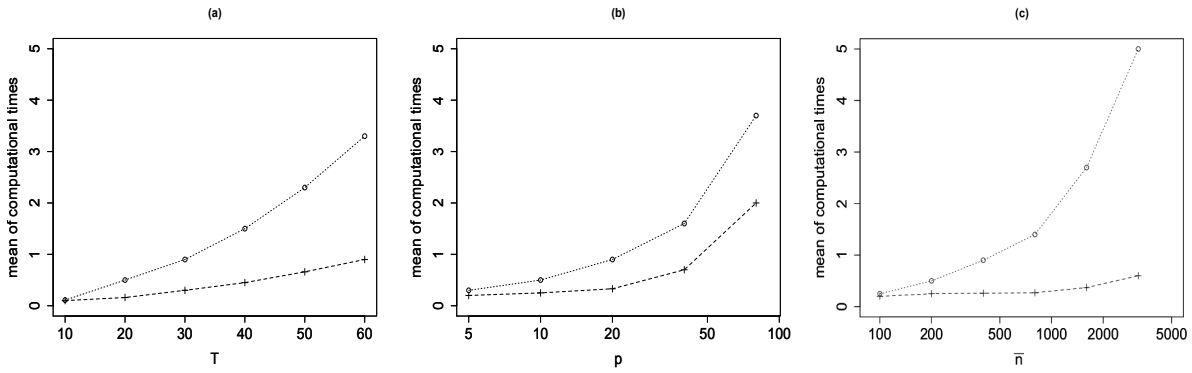

Figure 3: Mean of computational times (in seconds) of sequential SIR ($\circ$) and SIRds ($+$) for various values of T, p and $\overline{n}$. (a) $T \in \{10, 20, 30, 40, 50, 60\}$ for fixed $\overline{n} = 200$ and $p = 10$. (b) $p \in \{5, 10, 20, 40, 80\}$ (plotted in log-scale) for fixed $\overline{n} = 200$ and $T = 20$. (c) $\overline{n} \in \{100, 200, 400, 800, 1600, 3200\}$ (plotted in log-scale) for fixed $T = 20$ and $p = 10$. For the sake of readability, the points have been joined up with dotted and dashed lines.

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

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

At each time t (i.e., when the first t blocks are available), the corresponding true direction β_t is estimated with the SIRds and SIRu approaches. This direction is also computed with classical SIR using only the data from block t. The quality measure of the estimator $\widehat{\beta}_t$ obtained with SIRds, SIRu or SIR estimators is thus $\cos^2(\widehat{\beta}_t, \beta_t)$. It is plotted for each scenario in Figures 4 and 5. The weights $\cos^2(\widehat{b}_t, \widehat{b}_T)$ used in the computation of the SIRds estimator in equation (6) are represented in a color scaled image. The lighter the shade of yellow is, the larger the weight is (close to 1). The darker the shade of red is, the lower the corresponding weight is (close to 0). This image will provide the user an interesting chart to help detect if aberrant blocks appear in the data stream or if the underlying slope parameter has drifted.

For scenario 1 (see Figure 4), SIRds and SIRu perform well on the first nine blocks. For the aberrant 10th block, SIRds, in contrast to SIRu, is able to detect the change of direction and allows to estimate the "true" EDR direction, i.e., the direction of the current block. Moreover, the image of the weights clearly indicates that this 10th block is aberrant. Note that the classical SIR in each block provides a good estimation of the EDR direction. Taking

Scenario 1: the 10th block is different

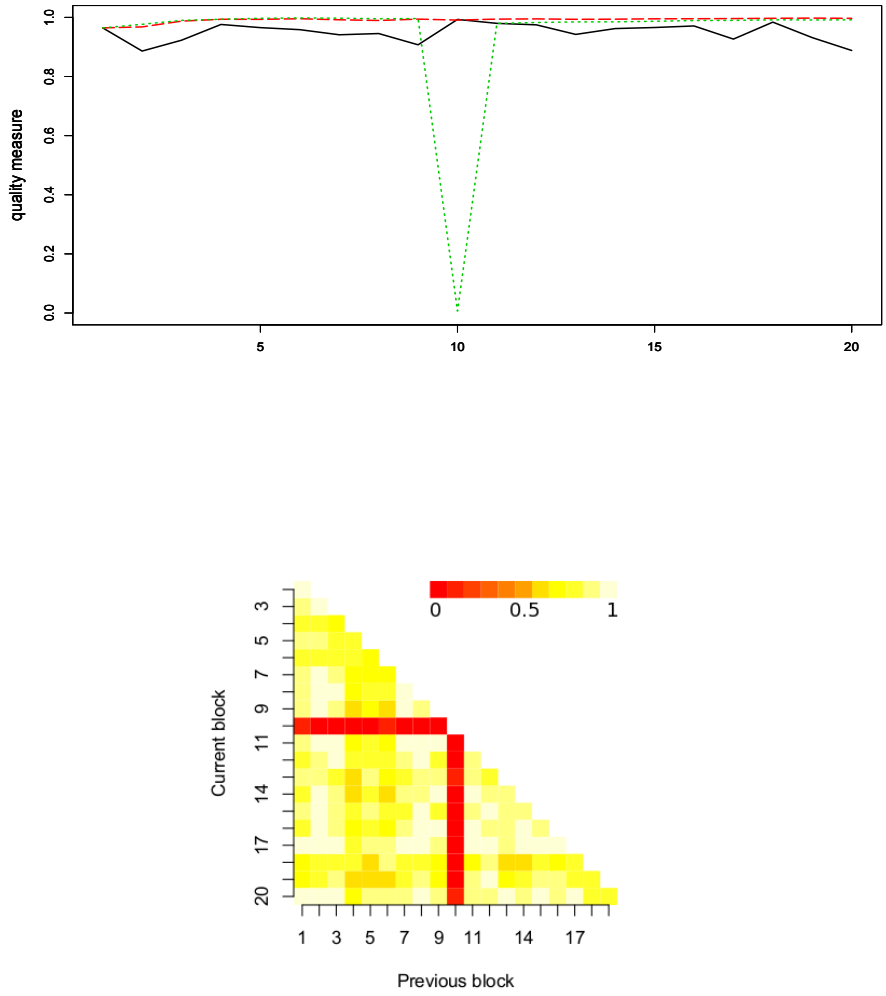

Figure 4: Numerical behavior of the SIRu and SIRds estimators for scenario 1. Top: plot of the quality measure $m(\widehat{\beta}_t, \beta_t)$ versus the number t of blocks (dashed red line for SIRds on the first t blocks, dotted green line for SIRu on the first t blocks, solid block line for SIR on block t only). Bottom: image of the weights $\cos^2(\widehat{b}_t, \widehat{b}_T)$ used in the computation of the SIRds estimator $\widehat{v}_T$.

into account all the information in previous blocks allows SIRds to improve the estimation of the EDR direction from SIR in the current block.
For scenario 2 (see Figure 5), 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 accuracy of the estimation from SIRu falls after the 10th block and rises slowly after that. Again, the results obtained with SIRds are better than those obtained with SIR based only on the current block because SIRds uses all the information of the previous available blocks sharing the same EDR direction.

As a conclusion, in both experiments, SIRds showed an interesting adaptive behavior with respect to aberrant blocks and drift of the EDR direction.

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 planet Mars from hyperspectral data. Specifically, the aim is to estimate 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 as shown by 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) or Scrucca (2007) 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 SIRds approach is applied in 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 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 $\overline{n} = 800$ synthetic spectra with the corresponding parameters. For the sake of simplicity, we limit ourselves to the study of the first $T = 8$ blocks. Also, let us focus on a terrain unit of strong CO_2 concentration determined by a classification method

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

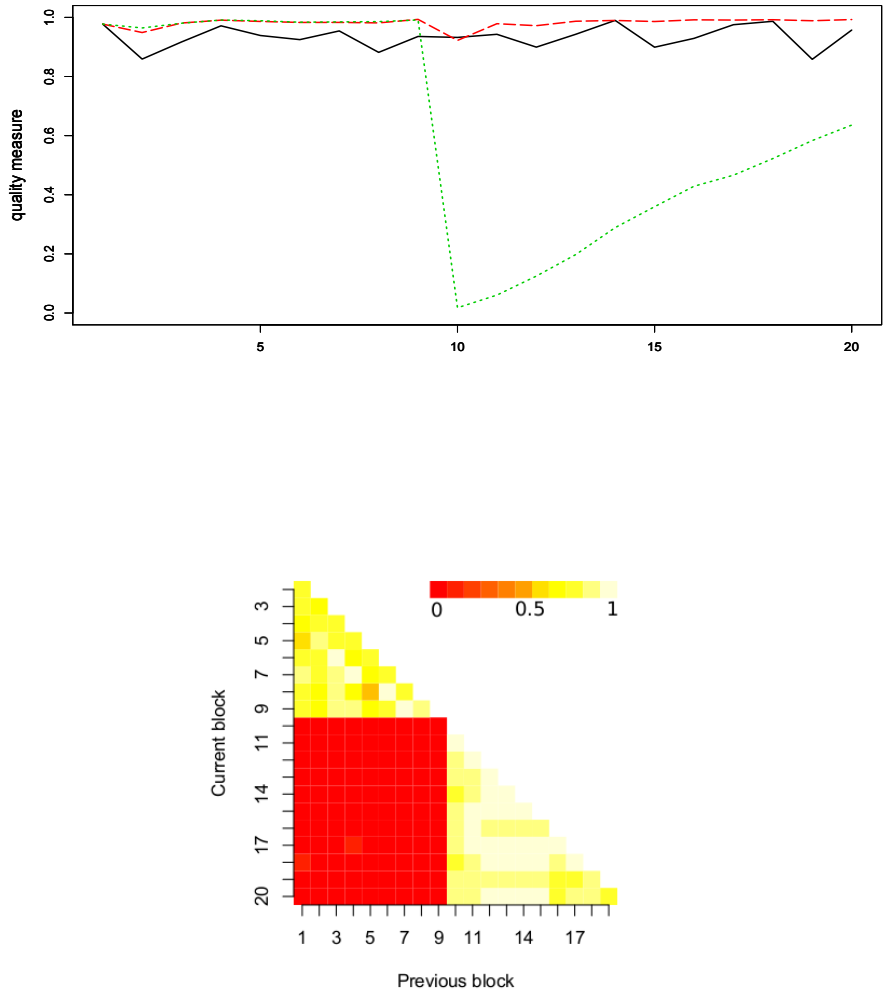

Figure 5: Numerical behavior of the SIRu and SIRds estimators for scenario 2. Top: plot of the quality measure $m(\widehat{\beta}_t, \beta_t)$ versus the number T of blocks (dashed red line for SIRds on the first t blocks, dotted green line for SIRu on the first t blocks, solid black line for SIR on block t only). Bottom: image of the weights $\cos^2(\widehat{b}_t, \widehat{b}_T)$ used in the computation of the SIRds estimator $\widehat{v}_T$.

based on wavelets (Schmidt *et al.* (2007)). The parameter of interest, Y, is the proportion of CO_2 ice.

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

The dimension K of the EDR subspace E was selected on the first block using the method proposed in Liquet and Saracco (2012). The criterion used is the squared trace correlation to study the closeness between two k-dimensional linear subspaces: the corresponding risk function is defined as

$$R_k = \mathbb{E}\left[\text{Trace}(P_k \widehat{P}_k)\right]/k, \tag{14}$$

where P_k denotes the orthogonal projector onto the subspace spanned by the first k basis vectors of E and $\widehat{P}_k$ is the orthogonal projector onto the subspace spanned by the first k vectors of $\widehat{E}$. This quantity R_k is only defined for any dimension k less than or equal to the true dimension K of the EDR subspace. In our dimension reduction context, a value of R_k close to one indicates that the set of the k estimated linear combinations of 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 the deviation of R_k from one. From a computational point of view, consistent estimates $\widehat{R}_k$ of R_k are required, so a 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 deviates 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, here, we adapt slightly the criterion proposed by 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 from the first block of size $\overline{n}$. Let us consider $s^{(b)} = \left\{(\mathbf{X}_i^{(b)}, Y_i^{(b)}), i = 1, \ldots, \overline{n}\right\}$ a non-parametric bootstrap sample replication. A naïve bootstrap estimate of the mean

squared risk function is defined by:

$$\widehat{R}_k = \frac{1}{\mathcal{B}} \sum_{b=1}^{\mathcal{B}} \widehat{R}_k^{(b)} \tag{15}$$

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, as 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 the bootstrap estimated version of the criterion. To do this, the method consists of evaluating the $\widehat{R}_k$ for all $k \in \{1, \ldots, p\}$ and then in observing how much they depart from one. Note that in Figure 6 (on the right), since p is large, $\widehat{R}_k$ versus k is plotted only 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 $\widehat{K} \ll p$. In practice, since there is no objective criterion to establish when a departure from one is small, a visual inspection 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 inspect the stability (or not) of the corresponding k-dimensional linear subspace. In Figure 6 (on the left), it clearly appears that the one- and two-dimensional EDR subspaces are stable, while the subspaces of greater dimension ($k \geq 3$) are more unstable. Following inspection of Figure 6, it appears that $\widehat{K} = 2$ seems to be an appropriate choice in terms of stability of the estimated EDR subspace. This choice is also confirmed by the eigenvalues scree plot which presents a jump after $K = 2$, see Figure 7.

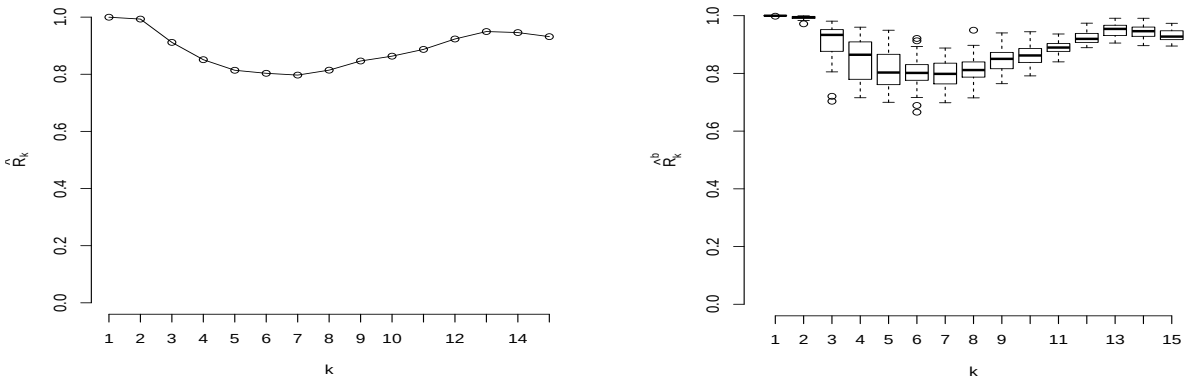

Figure 6: Choice of the dimension based on the stability of the estimated EDR subspace.

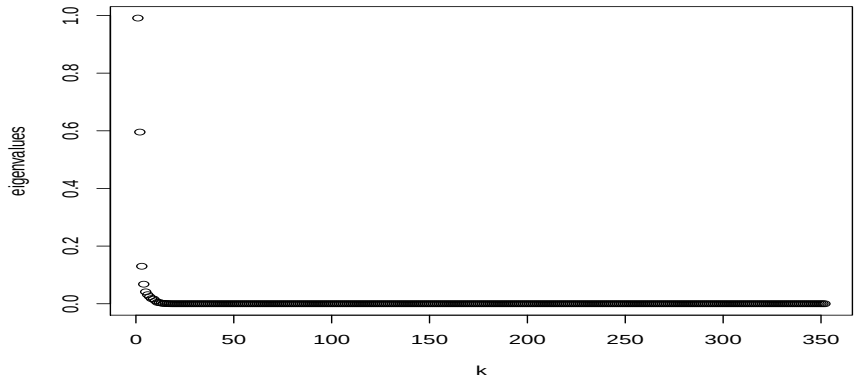

Figure 7: Eigenvalue scree plot of the first block.

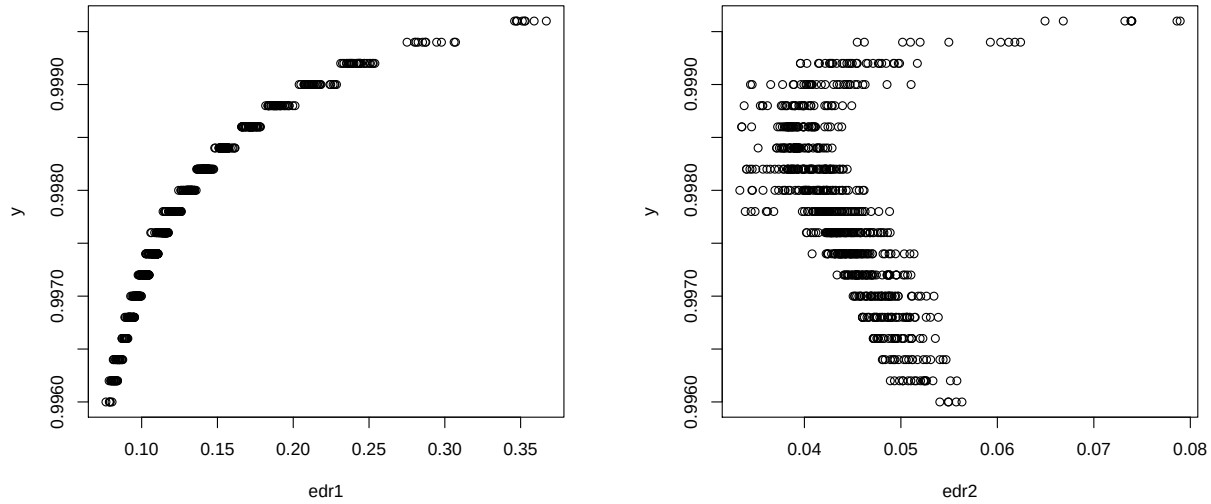

Figure 8: Plots of the dependent variable versus the first EDR index (on the left) and the second EDR index (on the right).

The plots of the proportion of CO_2 ice (Y) versus the first and second EDR indices depicted on Figure 8 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 left panel of Figure 9). Let us highlight that these structures are stable, i.e., they have also been observed on the other blocks. In contrast, the plot of the proportion of CO_2 ice versus the third and fourth EDR indices (see the graphic on the right of Figure 9) does not exhibit any structure and is very different from one block to another. These graphical diagnostics therefore confirm the choice $\widehat{K} = 2$.

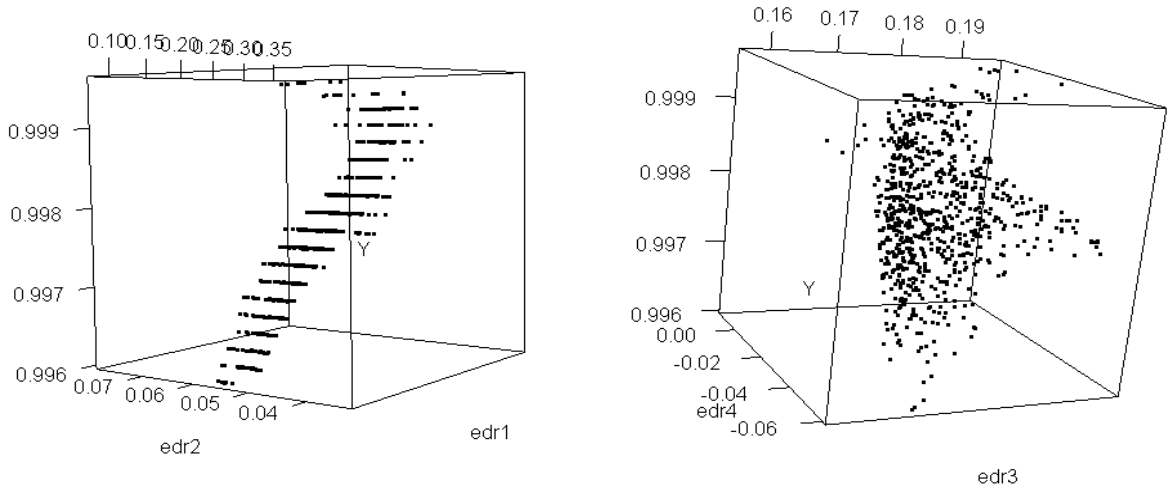

Figure 9: 3D-plot of the dependent variable versus the first two EDR indices (on the left); 3D-plot of the dependent variable versus the third and fourth EDR indices (on the right)

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 (6) graphically. However, since all the squared cosines are larger than 0.98, we do not provide the corresponding figure to visualize the weigths used in SIRds because only one color (light yellow) would be used. Therefore, the EDR subspace computed on the current block is very close to the EDR subspace computed on the previous blocks. Thus there exists an unique underlying EDR direction in this data stream.

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 those in Figures 8 and 9. Finally, Figure 10 represents the coordinates of each of the first two EDR directions. It indicates which wavelengths are important (nonzero coordinates) for estimating the proportion of CO_2 ice.

6 Concluding remarks

We present in this paper population and sample versions of SIRds for a single index or multiple indices model. The proposed SIRds approach performs well on simulated data as long as the EDR could be accurately estimated in each block. This implies that the size n_t of

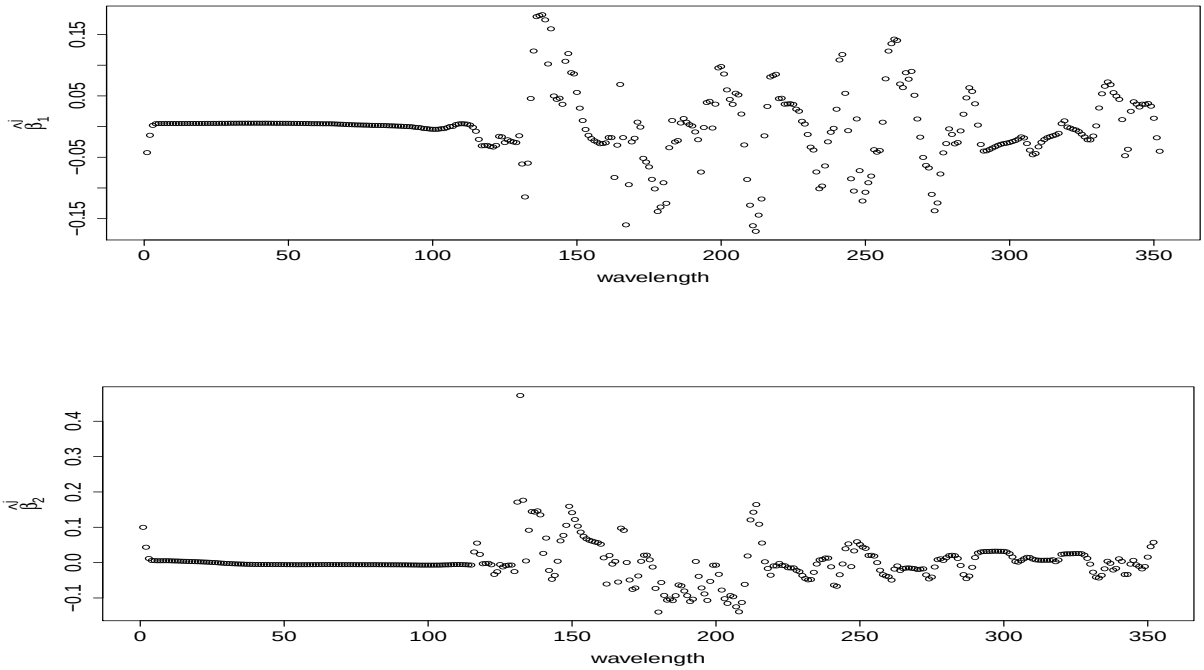

Figure 10: Final first and second EDR directions obtained with SIRds on all the blocks.

each block should be large enough in regards to the dimension p. Our approach has been also applied to real data concerning the estimation of physical properties of the surface of Mars. In this application, we use an extension of SIRds based on a regularized SIR version instead of the usual SIR. It is also possible to use alternative methods instead of SIR, such as SIR-II, SAVE or SIR$_\alpha$ for example. These approaches are based on 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 would be to use a multivariate SIR approach instead of univariate SIR methods, see for instance Barreda *et al.* (2007), Saracco (2005) or Lue (2009).

Acknowledgments

The authors thank Sylvain Douté for his contribution to the data. They are grateful to the anonymous referees for contributing to the improvement of this paper through their useful remarks and detailed comments.

Appendix: Proofs

PROOF OF THEOREM 1. (i) Since $||b_t|| = 1$ and introducing $\alpha_t = w_t m(b_t, b_T)$, we have:

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

Thus the maximization problem (4) can be rewritten as

$$\max_{v \in \mathbb{R}^p} \frac{v' M_T v}{v' v}. \tag{16}$$

The solution of (16) is clearly the normalized principal eigenvector of M_T.

(ii) Assuming the linearity condition (LC) and model (1) for each block t, all the vectors b_t are collinear with β. The rank of the symmetric matrix M_T is therefore one. The eigenvector v_T associated with the non-null eigenvalue of M_T is also collinear with β: thus v_T is a normalized EDR direction ($||v_T|| = 1$). □

PROOF OF THEOREM 2. (i) For each block t and under the assumptions (LC), (A1)-(A3), from the SIR theory of Li (1991) each estimated EDR direction $\widehat{b}_t$ converges to b_t at root n_t rate: that is, for $t = 1, \ldots, T$, $\widehat{b}_t = b_t + O_p(n_t^{-1/2})$. It can be shown that $\cos^2(\widehat{b}_t, \widehat{b}_T) = \cos^2(b_t, b_T) + O_p(\underline{n}^{-1/2}) = 1 + O_p(\underline{n}^{-1/2})$, and thus $\widehat{M}_T = M_T + O_p(\underline{n}^{-1/2})$. Therefore the principal eigenvector of $\widehat{M}_T$ converges to that corresponding to M_T at the same rate: $\widehat{v}_T = v_T + O_p(\underline{n}^{-1/2})$. Since v_T is collinear with β, the estimated EDR direction $\widehat{v}_T$ converges to an EDR direction at root $\underline{n}$ rate.

(ii) 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. Let $\text{vec}(C)$ denote the pq-dimensional column vector: $\text{vec}(C) = \left(c_1', \ldots, c_q'\right)'$. We shall denote by N^+ the Moore-Penrose generalized inverse of the square matrix N. In the sequel, let $B = [b_1, \ldots, b_T]$ be the matrix which contains all the EDR directions obtained from all 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, it follows that:

$$\sqrt{\overline{n}}(\text{vec}(\widehat{B}) - \text{vec}(B)) \longrightarrow_d \text{vec}\begin{pmatrix} U_1 \\ \vdots \\ U_T \end{pmatrix} \sim \mathcal{N}(0, \Gamma_U) \text{ where } \Gamma_U = \begin{pmatrix} V_1 & & 0 \\ & \ddots & \\ 0 & & V_T \end{pmatrix} \quad (17)$$

Step 2: Asymptotic distribution of vec$(\widehat{M_T})$. Standard properties of "vec" operator yield:

$$\text{vec}(\widehat{M_T}) = \sum_{t=1}^{T} w_t \text{vec}(\widehat{b}_t\widehat{b}_t')(\widehat{b}_t'\widehat{b}_T)^2 = f(\text{vec}(\widehat{B})),$$

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

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

Let $K_{1,p}$ be the vec-permutation matrix given by $K_{1,p} = \sum_{j=1}^{p}(E_{1j} \otimes E_{1j}')$ with $E_{1j} = e_{j,p}'$ and $e_{j,p}$ is the jth column of I_p. The $(p^2 \times pT)$ Jacobian matrix $J = [J_1|\ldots|J_T]$ associated with f is defined by the concatenation of the $p^2 \times p$ matrices J_t, where, for $t = 1, \ldots, T-1$,

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

and 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 (17) and applying Delta-method entail

$$\sqrt{\overline{n}}(\text{vec}(\widehat{M_T}) - \text{vec}(M_T)) \longrightarrow_d V \sim \mathcal{N}(0, \Gamma_V = J\Gamma_U J'). \quad (18)$$

Step 3: Asymptotic distribution of $\widehat{b}$. The vector $\widehat{v}_T$ (resp. v_T) is the eigenvector associated with the largest eigenvalue $\widehat{\lambda}$ (resp. λ) of $\widehat{M_T}$ (resp. M_T). Since $\widehat{M_T} = M_T + O_p(1/\sqrt{\overline{n}})$ and using (18), Lemma 1 of Saracco (1997) yields:

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

with

$$\Gamma_W = [v_T' \otimes (M_T - \lambda I_p)^+]\Gamma_V[v_T \otimes (M_T - \lambda I_p)^+]. \tag{19}$$

□

PROOF OF THEOREM 3. (*i*) Since the bases $\mathbb{A}, \mathbb{B}_1, \ldots, \mathbb{B}_T$ are assumed to be I_p-orthonormal and introducing $\alpha_t = w_t m(\mathbb{B}_t, \mathbb{B}_T)$, we have:

$$\begin{aligned}
Q(\mathbb{A}, \mathbb{B}_1, \ldots, \mathbb{B}_T) &= \textstyle\sum_{t=1}^T \alpha_t m(\mathbb{A}, \mathbb{B}_t) \\
&= \textstyle\sum_{t=1}^T \alpha_t \mathrm{Trace}(\mathbb{A}\mathbb{A}'\mathbb{B}_t\mathbb{B}_t')/K \\
&= \textstyle\sum_{t=1}^T \alpha_t \mathrm{Trace}(\mathbb{A}'\mathbb{B}_t\mathbb{B}_t'\mathbb{A})/K \\
&= \mathrm{Trace}(\mathbb{A}'\{\textstyle\sum_{t=1}^T \alpha_t \mathbb{B}_t\mathbb{B}_t'\}\mathbb{A})/K \\
&= \mathrm{Trace}(\mathbb{A}'\{\textstyle\sum_{t=1}^T w_t \frac{\mathbb{B}_t\mathbb{B}_t'}{K} m(\mathbb{B}_t, \mathbb{B}_T)\}\mathbb{A}) \\
&= \mathrm{Trace}(\mathbb{A}'\mathbb{M}_T\mathbb{A}).
\end{aligned}$$

Let $\mathbb{V}_T = \arg\max_{\mathbb{A}} Q(\mathbb{A}, \mathbb{B}_1, \ldots, \mathbb{B}_T)$. Since it is well known that $\mathbb{V}_T$ 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.

(*ii*) Since the column vectors of $\mathbb{B}_t$ form an I_p-orthonormal basis of E, we have $\mathrm{Span}(\mathbb{B}_t) = E$ 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. The assumptions of the theorem imply that $m(\mathbb{B}_t, \mathbb{B}_T) = 1$. Then it follows that the eigenvectors associated with the K largest eigenvalues of $\mathbb{M}_T$ form an I_p-orthonormal basis of the EDR subspace E. □

PROOF OF THEOREM 4. From SIR theory, one can derive $\widehat{\mathbb{B}}_t = \mathbb{B}_t + O_p(n_t^{-1/2})$ 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 the same rate to the corresponding eigenvectors associated with the K non-null eigenvalues of $\mathbb{B}_t\mathbb{B}_t'$. Under the assumptions of the theorem, we have $m(\widehat{\mathbb{B}}_t, \widehat{\mathbb{B}}_T) = 1 + O_p(\underline{n}^{-1/2})$. As a consequence $\widehat{\mathbb{M}}_T = \mathbb{M}_T + O_p(\underline{n}^{-1/2})$, and finally $\widehat{v}_{k,T} = v_{k,T} + O_p(\underline{n}^{-1/2})$, $k = 1, \ldots, K$, which completes the proof. □

References

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

Barrios, M.P. and 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.

Chavent, M., Girard, S., Kuentz, V., Liquet, B., Nguyen, T. M. N. and Saracco J. (2012) Régression inverse par tranches sur flux de données. In: *44èmes Journées de Statistique (SFdS)*, Brussels, Belgium, `http://hal.archives-ouvertes.fr/hal-00736584` (in french).

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

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression (with discussion). *Statistical Science*, 22, 1-26.

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

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$_\alpha$. *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.

Scrucca, L. (2007.) Class prediction and gene selection for DNA microarrays using regularized sliced inverse regression. *Computational Statistics and Data Analysis*, 52, 438-451.

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.