A sliced inverse regression approach for block-wise evolving data streams

(Régression inverse par tranches sur un flux de données)

Jérôme Saracco[1,2]

en collaboration avec

M. Chavent[1,2], **S. Girard**[3], **V. Kuentz**[4], **B. Liquet**[5], **T.M.N. Nguyen**[6]

[1] Institut de Mathématiques de Bordeaux (IMB), Université de Bordeaux - IPB
[2] Equipe CQFD, INRIA Bordeaux Sud-Ouest
[3] Equipe MISTIS, INRIA Rhône Alpes
[4] IRISTEA, Cestas
[5] ISPED, Université de Bordeaux
[6] IRMA, Université de Strasbourg

JdS 2012 - Bruxelles - Mai 2012

Outline

Outline

Introduction : a SIR method for data stream

Initial motivation of this work : an applied problem based on real data dealing with the estimation of Mars surface physical properties from hyperspectral images.

The goal of the study : estimate the link between some physical parameters Y and observed spectra X

- via sliced inverse regression approach for reducing high the dimension of spectra ($p = 352$ wavelengths) $\rightarrow$ Gardes, Girard, ...
- Moreover, the **database of synthetic spectra** may be so large that it **cannot be stored in a computer memory**.

  Thus, a stream of T smaller sub-databases is generated.

  $\Longrightarrow$ **Sliced inverse regression (SIR) approach for block-wise evolving data streams**

Our basic model : semi-parametric single index model proposed by Duan and Li (1991) :

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

where

- the response variable Y is univariate,
- the regressor X is p-dimensional (with expectation $E(X) = \mu$ and covariance matrix $V(X) = \Sigma$),
- the error term ϵ is independent of X,
- the link function f and the vector β are unknown.

Since f is unknown, β is not totally identifiable in this model. Then we are interested in finding the linear subspace spanned by β, called the **Effective Dimension Reduction (EDR) space**.

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

In this communication, we focus on data arriving sequentially by block in a stream.

Let us consider T blocks.

We assume that each data block t is composed of an i.i.d. sample $\{(X_i, Y_i), i = 1, \ldots, n_t\}$ available from model (1).

Our goal : estimate the EDR direction at each arrival of a new block of observations.

A simple and direct approach :

- pool all the observed blocks (union of the blocks)
- estimate the EDR direction by the **Sliced Inverse Regression (SIR)** method introduced by Li (1991).

While SIR is a computationally simple method, the drawbacks of pooling the data are

- the storage of the blocks since the size of the dataset considerably increases with the number of blocks,
- the running time for high dimensional data.

To avoid these drawbacks,
we propose the **SIRdatastream** approach.

Recall on SIR in block t

The population version SIR relies on the following linear condition :

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

which is fulfilled when X is elliptically distributed and almost surely fulfilled in the presence of high-dimensional data, see Hall and Li (1993) for details.

Let us consider a monotone transformation $T(.)$ of Y.
Under condition (C) and model (1), Li (1991) showed that the principal eigenvector b_t of

$$\Sigma^{-1}\Gamma_t \qquad \text{where } \Gamma_t = V(E(X|T(Y)))$$

is **an EDR direction** (i.e. is collinear with β).

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

To obtain an estimator of Γ_t which can be easily estimated and 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. $m_h = E(X|Y \in s_h)$), then the matrix Γ_t can be written as

$$\Gamma_t = V(E(X|T(Y))) = \sum_{h=1}^{H_t} p_h (m_h - \mu)(m_h - \mu)'.$$

To estimate the matrix Γ_t : substitute theoretical versions of the moments by their empirical counterparts.

The **estimated EDR direction** $\widehat{b}_t$ is the principal eigenvector of $\widehat{\Sigma}^{-1}\widehat{\Gamma}_t$ where $\widehat{\Gamma}_t$ and $\widehat{\Sigma}$ are estimators of Γ_t and Σ.

Outline

Population version of SIRdatastream

For $t = 1, \ldots, T$, let us denote by b_t the EDR direction obtained in the block t.

How to combine them ?

We can consider the following matrix

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

where the w_t's are positive weights such that $\sum_{t=1}^{T} w_t = 1$.
Under the assumptions of the model,

- each b_t is colinear to β ;
- the principal eigenvector of $\widetilde{M}_T$ is colinear with β and then is an EDR direction.

This version is a non "adaptive" one (if the model evolves in terms of β).

To provide an "adaptive" version (if the model evolves in terms of β) of our approach,
we can consider the follwing matrix :

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

Under the assumptions of the model (if it does not evolve),

- the weight $\cos^2(b_t, b_T)$ is equal to one since b_t and b_T are both colinear with β ;
- the principal eigenvector v_T of M_T is colinear with β and then is an EDR direction.

Reformulation of this approach as an optimization problem :

$$\max_{v \in \Re^p} \frac{v' M_T v}{v' v}. \tag{2}$$

The solution is clearly v_T, the normalized principal eigenvector of M_T.

Since $||b_t|| = 1$, we can show that : $\sum_{t=1}^{T} \omega_t \cos^2(b_t, v) = v' M_T v$
where $\omega_t = w_t \cos^2(b_t, b_T)$.
Thus maximization problem (2) can be rewritten as

$$\max_{v \in \Re^p} \sum_{t=1}^{T} \omega_t \cos^2(b_t, v) \quad \text{s.t. } ||v|| = 1. \tag{3}$$

Sample version of SIRdatastream

For $t = 1, \ldots, T$, let us denote by $\hat{b}_t$ the estimator of the EDR direction calculated on each block t.

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

$$\widehat{M}_T = \sum_{t=1}^{T} w_t \hat{b}_t \hat{b}_t' \cos^2(\hat{b}_t, \hat{b}_T) \qquad (4)$$

where $w_t = \frac{n_t}{\sum_{j=1}^{T} n_j}$ (for instance) and $\cos^2(\hat{b}_t, \hat{b}_T) = \frac{(\hat{b}_t' \hat{b}_T)^2}{(\hat{b}_t' \hat{b}_t) \times (\hat{b}_T' \hat{b}_T)}$.

Asymptotic results

Theorem ($\sqrt{n}$-convergence of the estimated EDR direction)

Under the assumptions (C), (A1)-(A3), we have

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

Since v_T is colinear with β, then the estimated EDR direction $\widehat{v}_T$ converges to an EDR direction at $\sqrt{n}$-rate.

Theorem (Asymptotic normality)

Under the assumptions (C), (A1)-(A3), we have

$$\sqrt{n}\,(\widehat{v}_T - v_T) \longrightarrow_d \mathcal{N}_p(0_p, \Sigma_V).$$

Computational complexity

For sake of simplicity, 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

$$O(Tn^* p^2 + T^2 p^2) = O(Tp^2(n^* + T)),$$

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 $O(Tn^*p)$.

As a comparison, **SIRds** requires the storage of only one block of regressors and of the EDR directions computed on the previous blocks, corresponding to a storage load $O(pn^* + pT)$.

Under the previous assumptions, **SIRds** requires $O(T)$ less data storage than **sequential SIR**.

Running time (on simulations)

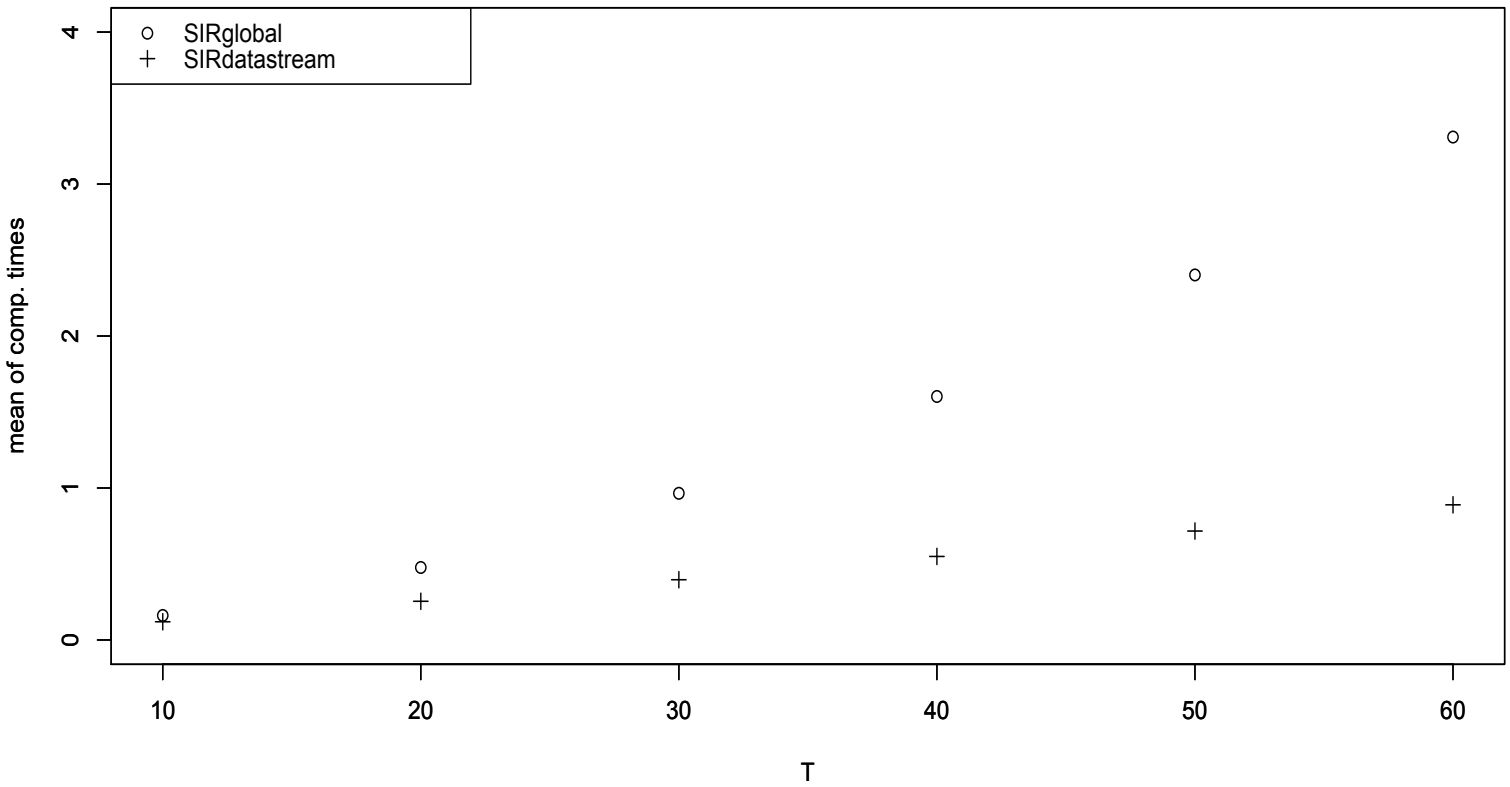
We compare the running time (in seconds) of our **SIR datastream** approach with **sequential SIR**.

We evaluate the computational time for these two methods :

- for various values of the total number T of blocks,
- for various values of the dimension p of the covariable X,
- for various values of the size n^* of each block.

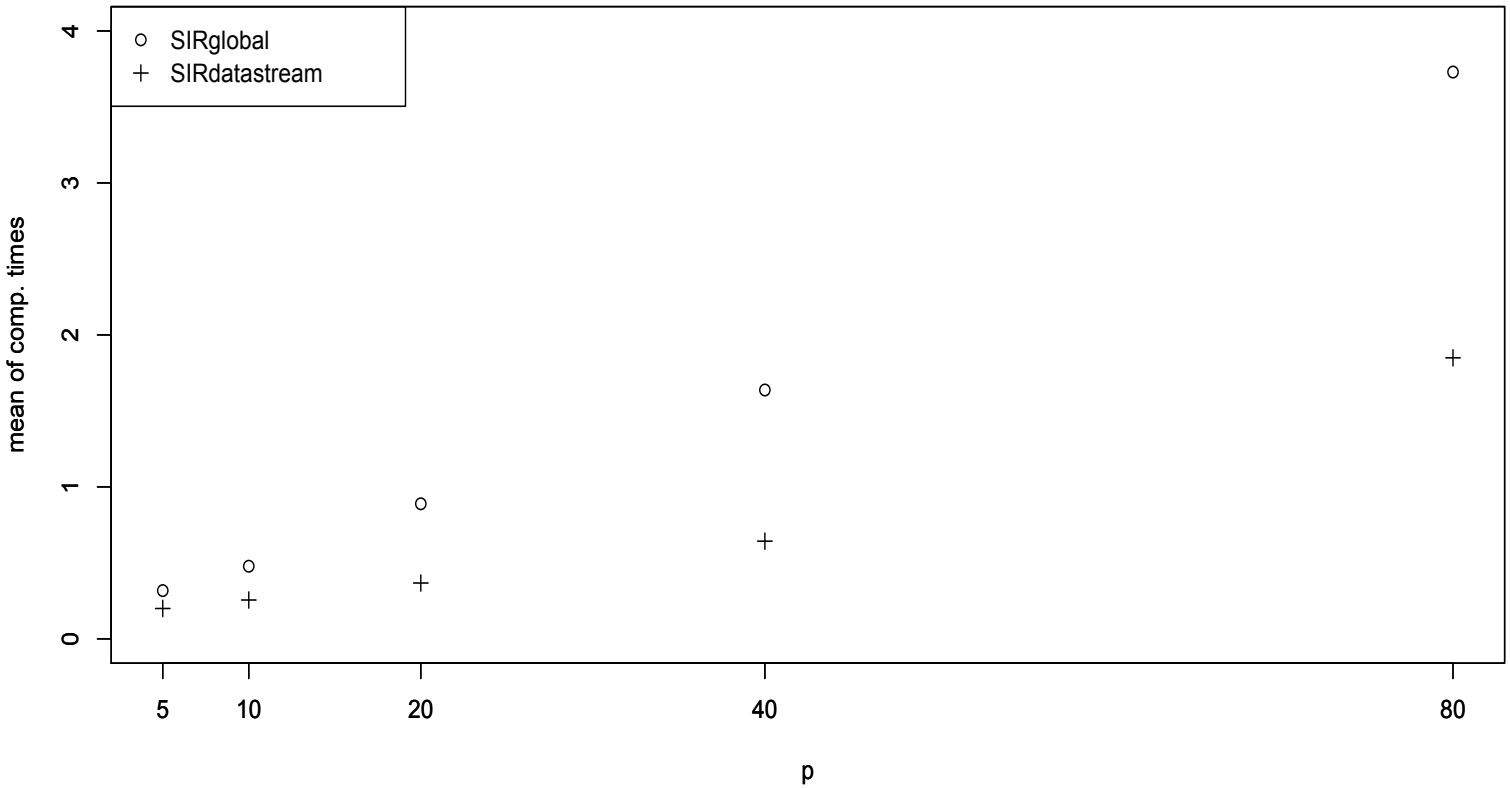
We generate $\mathcal{B} = 50$ data streams for each (T, p, n^*) and we calculate the mean of running times.

Mean of running times (in seconds) **according to** T
when $n^* = 200$ and $p = 10$



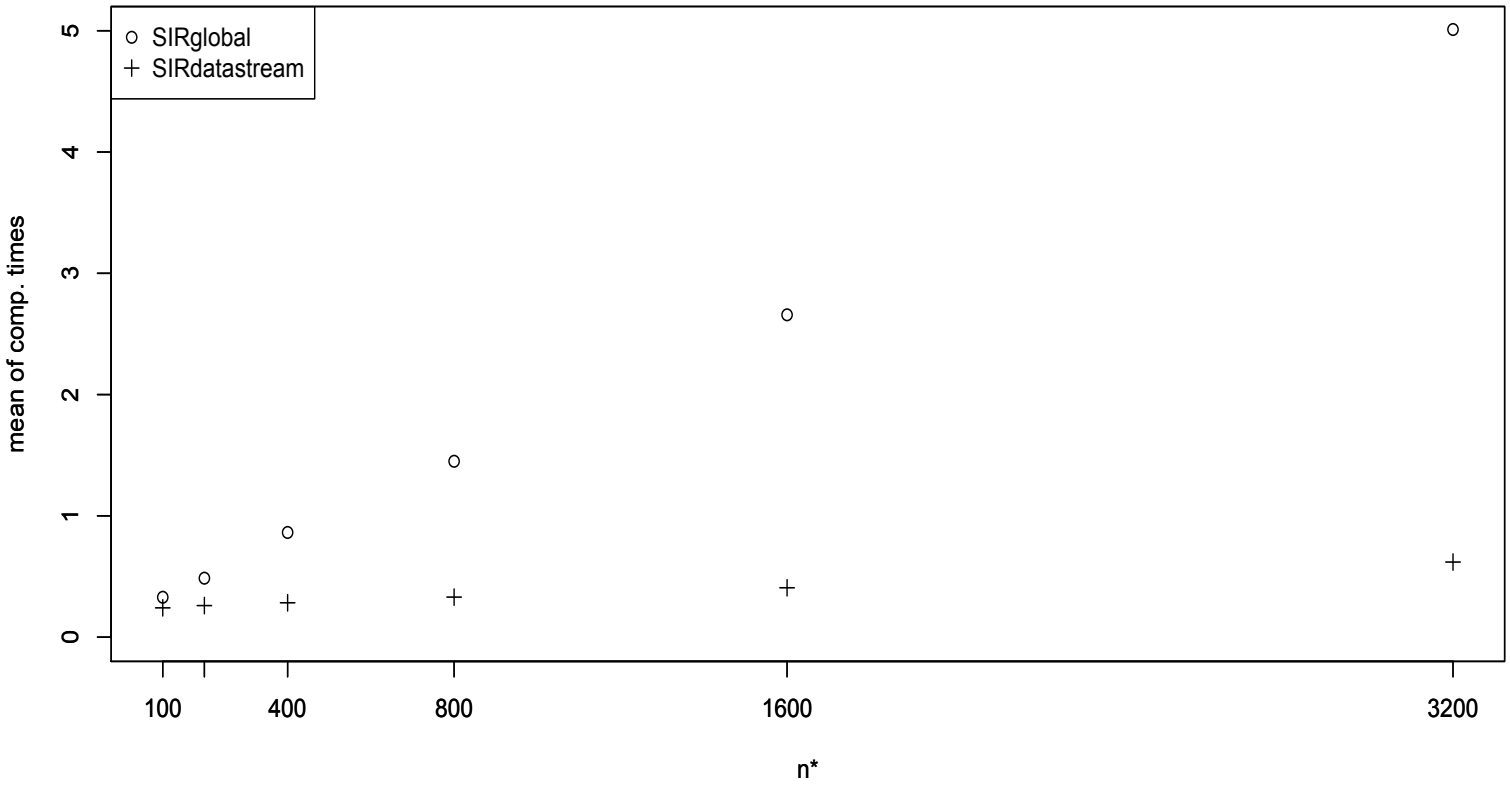
The number T of blocks hugely penalizes the "sequential SIR" approach by comparison with "SIR datastream".

Mean of running times (in seconds) **according to** p
when $n^* = 200$ and $T = 20$



The dimension p noticeably favours "SIR datastream" versus "sequential SIR".

Mean of running times (in seconds) **according to** n^*
when $T = 20$ and $p = 10$



The block size n^* hugely penalizes the "sequential SIR" approach in comparison with "SIR datastream".

Outline

We consider for each block of data the same following semiparametric regression model :

$$Y = (X'\beta)^3 + \epsilon, \tag{5}$$

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 $\mathcal{N}(0, 0.5)$ and is independent of X.

We generate $T = 20$ blocks of size $n^* = 200$ with $p = 20$.

For each data stream, at the arrival of each block t ($t = 1, \ldots, T$), we estimate the EDR direction

- with **SIRds** (for SIRdatastream) based on these first available t blocks ;
- with **SIRu** (for SIR on union of blocks), i.e. classical SIR approach based on the sample formed by the union of the first available t blocks.

Quality measure of the estimated EDR direction.

We use the following quality measure for any estimator (denoted by $\widehat{\beta}$) of the direction β :

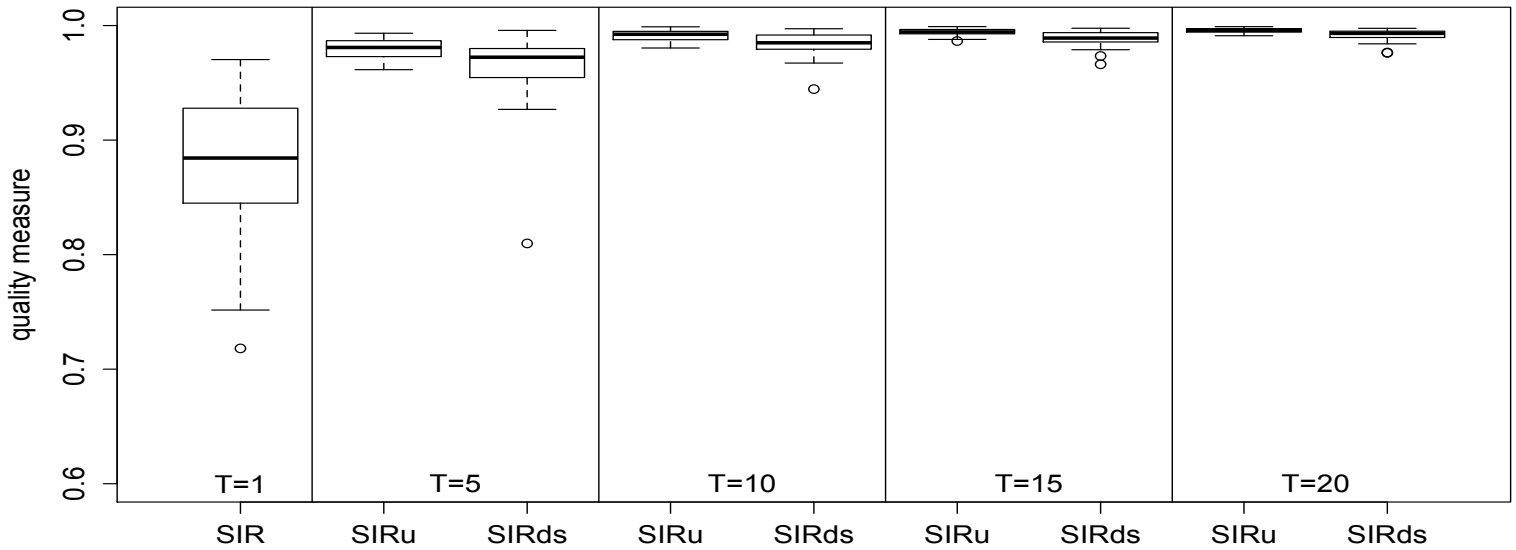
$$\cos^2(\widehat{\beta}, \beta) = \frac{(\widehat{\beta}'\beta)^2}{(\widehat{\beta}'\widehat{\beta}) \times (\beta'\beta)}.$$

The closer to one is this measure, the better is the estimate.

Comparison of SIRds and SIRu.

Our aim : compare the quality measure of the EDR directions estimated with **SIRds** and **SIRu**.

We generate $\mathcal{B} = 50$ data replications of data stream of size $T = 20$ blocks as it has been previously described.



When $T = 1$ (only one block), approaches **SIRds** and **SIRu** are obviously equivalent to usual SIR.

Adaptation to changes in the underlying model

In this simulation, we relax the assumption that the model is the same in all the blocks and the slope parameter β in model (5) is then indexed by t :

$$Y = (X'\beta_t)^3 + \epsilon. \tag{6}$$

In order to show the good behavior of **SIRds** in comparison with **SIRu** in such cases, we consider two scenarios.

For each scenario, we generate $T = 20$ blocks as described in the next slide.

- **Scenario 1 : the 10th block is different (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 : a drift occurs from the 10th block to the last block.**

  We fix $\beta_t = (1, -1, 2, -2, 0, \ldots, 0)'$ for the first 9 blocks $(t = 1, \ldots, 9)$.
  And we set $\beta_t = (1, 1, \ldots, 1)'$ for the remaining ones $(t = 10, \ldots, 20)$.

Then, at each time t (i.e. when the first t blocks are available), we estimate the EDR direction

- with **SIRds** approach,
- with **SIRu** approach,
- with **classical SIR** based only on the data of this block t.

For each scenario, we give two graphics :

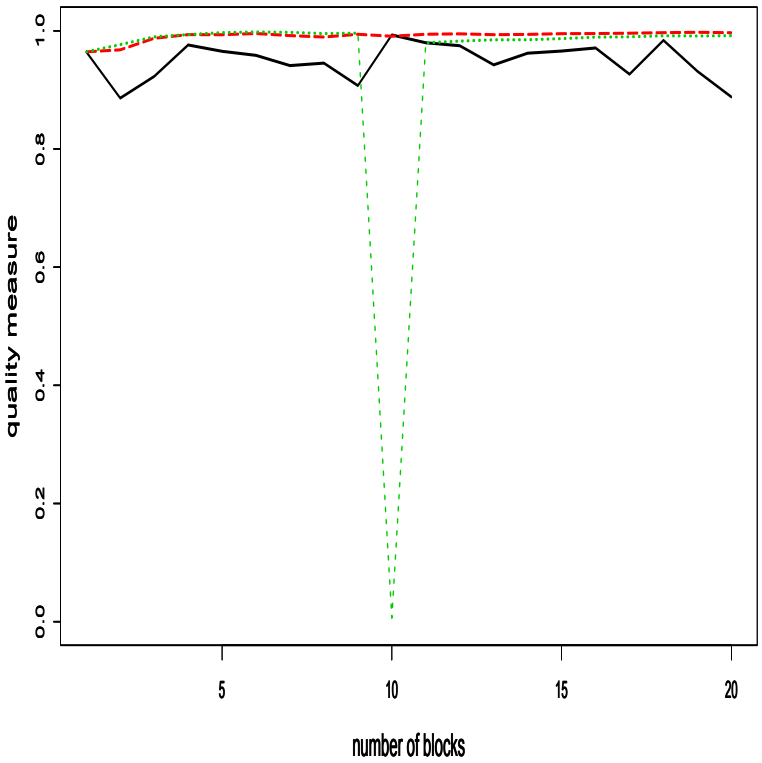
- a plot of the quality measures $\cos^2(\widehat{\beta}_t, \beta_t)$ versus t, for the estimates $\widehat{\beta}_t$ obtained with **SIRds**, **SIRu** or **SIR** estimators at each time t.

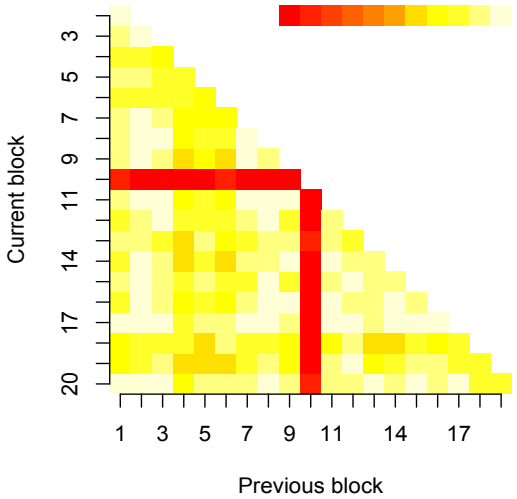
- a color scaled image the weights $\cos^2(\widehat{b}_t, \widehat{b}_T)$ used in the computation of the **SIRds** estimator in equation (4).

  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

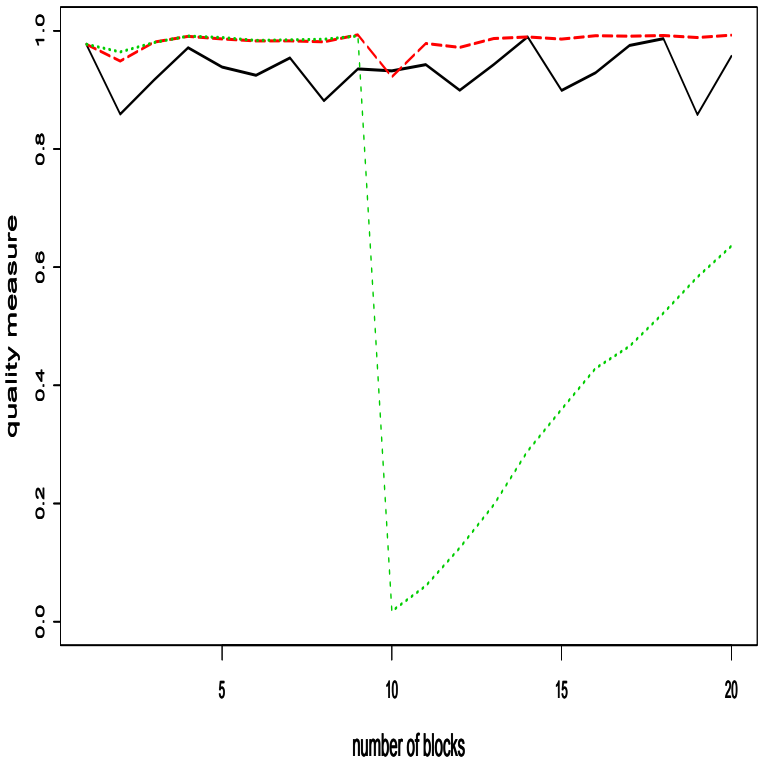


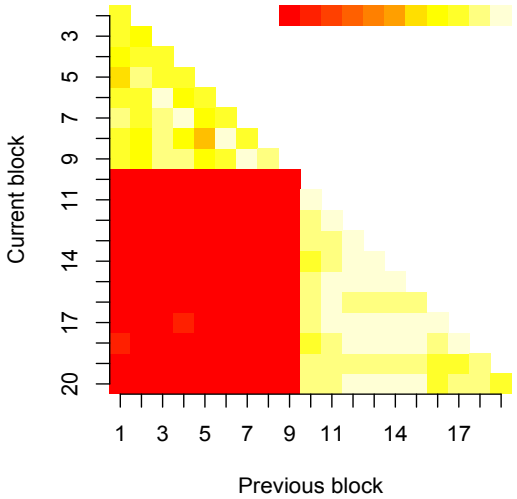


On the left : plot of the quality measure $\cos^2(\widehat{\beta}_t, \beta_t)$ versus the number T of blocks (for **SIRds** on , **SIRu** and **SIR** on block t only).

On the right : image of the weights $\cos^2(\widehat{b}_t, \widehat{b}_T)$ used in the computation of the SIRds estimator.

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





On the left : plot of the quality measure $\cos^2(\widehat{\beta}_t, \beta_t)$ versus the number T of blocks (for **SIRds** on , **SIRu** and **SIR** on block t only).

On the right : image of the weights $\cos^2(\widehat{b}_t, \widehat{b}_T)$ used in the computation of the SIRds estimator.

Outline

Concluding remarks

- The proposed approach performs well on simulated data and works on a real dataset.
- It is possible to extend this approach to multiple indices models (it has be done in the application) :
  - b_t will be replaced by a basis B_t of the EDR space ;
  - the squared cosine will be replaced by a proximity measure between two K-dimensional EDR spaces, for instance the square trace correlation.
- It is also possible to use alternative SIR methods (such as MSIR, SIR-II, SAVE, SIR_α, ...) or multivariate SIR.

Thank you for your attention.

Assume that **model has evolved in block** $T-1$:

$$\beta \text{ becomes } \beta^* \text{ (with } \beta^* \perp \beta).$$

We have : $\cos^2(b_t, b_{T-1}) = 0, \forall t \neq T-1$ and $\cos^2(b_{T-1}, b_{T-1}) = 1$.

Under the assumptions of the model,

- We have : $M_{T-1} = \sum_{t=1}^{T-1} w_t b_t b_t' \cos^2(b_t, b_{T-1}) = w_{T-1} b_{T-1} b_{T-1}'$.

  The principal eigenvector v_{T-1} of M_{T-1} is colinear with β^* and then is **the EDR direction of the current block**.

- We have : $M_T = \sum_{t=1}^{T} w_t b_t b_t' \cos^2(b_t, b_{T-1}) = \sum_{t \neq T-1} w_t b_t b_t'$.

  the principal eigenvector v_T of M_T is colinear with β and then is **the EDR direction of the current block (taking into account the other blocks having the same EDR direction)**.

Asymptotic results

Assumptions :

We consider a fixed number T of blocks and a total sample size n 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 $p_h \neq 0, h = 1, \ldots, H_t$.
- (A3) For $t = 1, \ldots, T$ and $h = 1, \ldots, H_t$,
$$n_{h,t} \to \infty \text{ (and therefore } n_t \to \infty) \text{ as } n \to \infty.$$

Back on the application...

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 ($p = 352$ wavelengths).

$$Y = f(X'\beta_1, \ldots, X'\beta_K, \alpha)$$

$\hookrightarrow$ we focus on the EDR space $E = \text{Span}(\beta_1, \ldots, \beta_K)$.

The parameter of interest Y is the proportion of CO_2 ice.

Following Bernard-Michel *et al.* (2009a), we propose to reduce the high dimension ($p = 352$) of spectra with a regularized version of SIR.

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.

The **need to regularize SIR** in very high dimensions is well-known. 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.

Choice of the regularized SIR 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.

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.

In practice, the **database of synthetic spectra** may be so large that it **cannot be stored in a computer memory**.

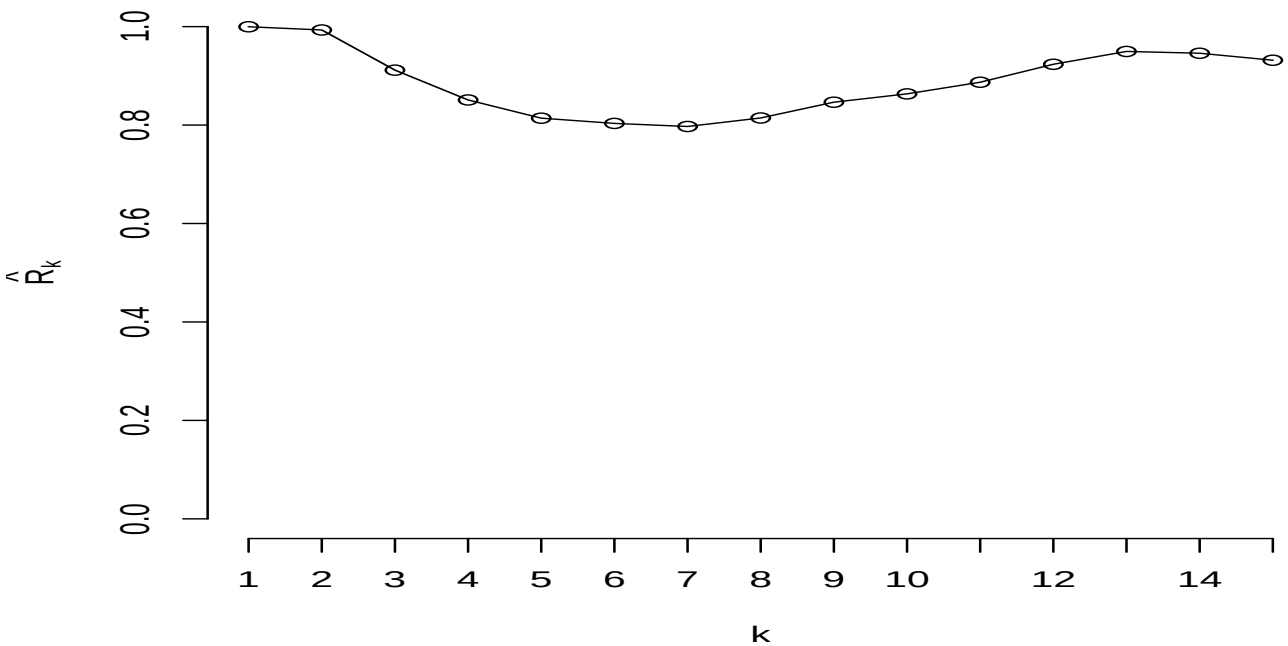
Thus, a stream of smaller sub-databases is generated and we propose to apply our SIRds approach to this context.

Here we will only consider $T = 8$ sub-databases (blocks).

Back on the application...

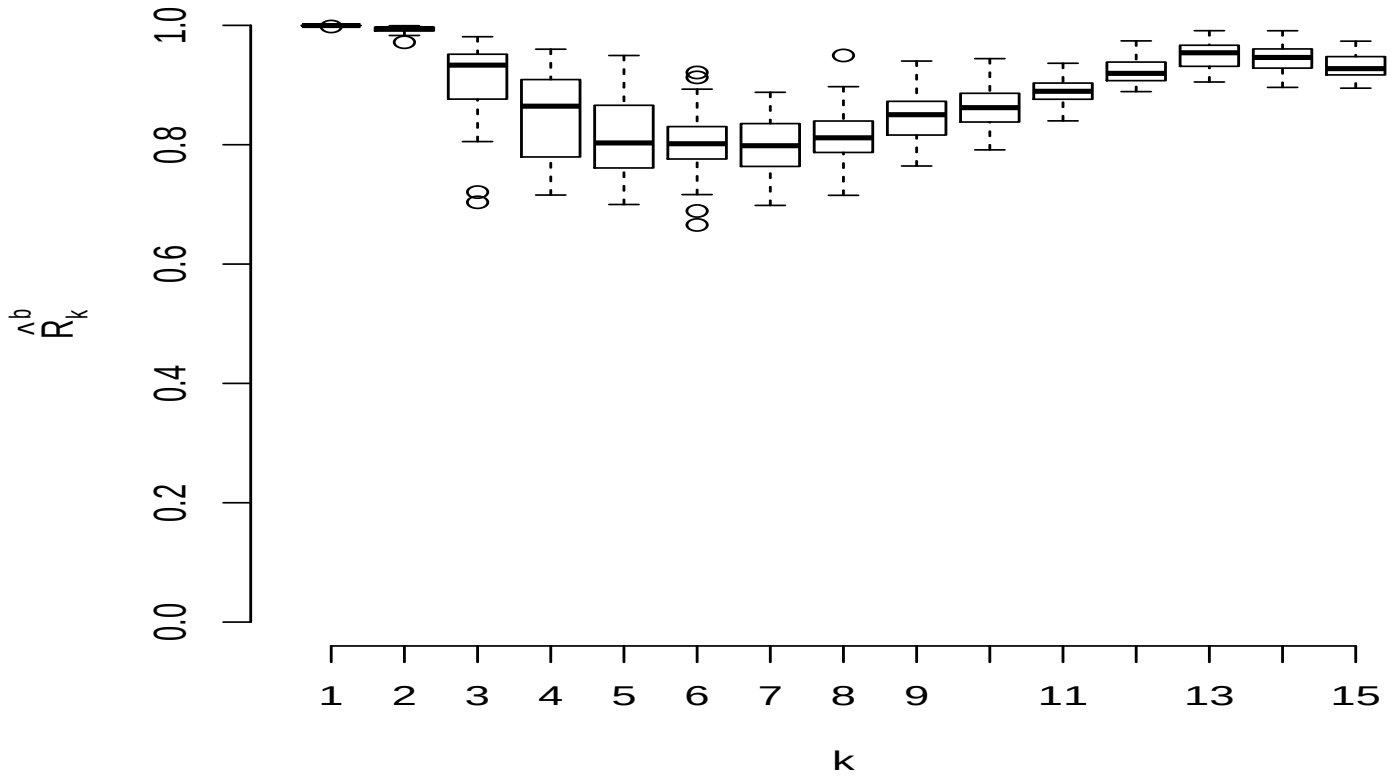
Choice of the dimension K on the first block

$\hookrightarrow$ use of the "square trace correlation criterion" (see Liquet and Saracco (2012) for details) and package R **edrGraphicalTools**.

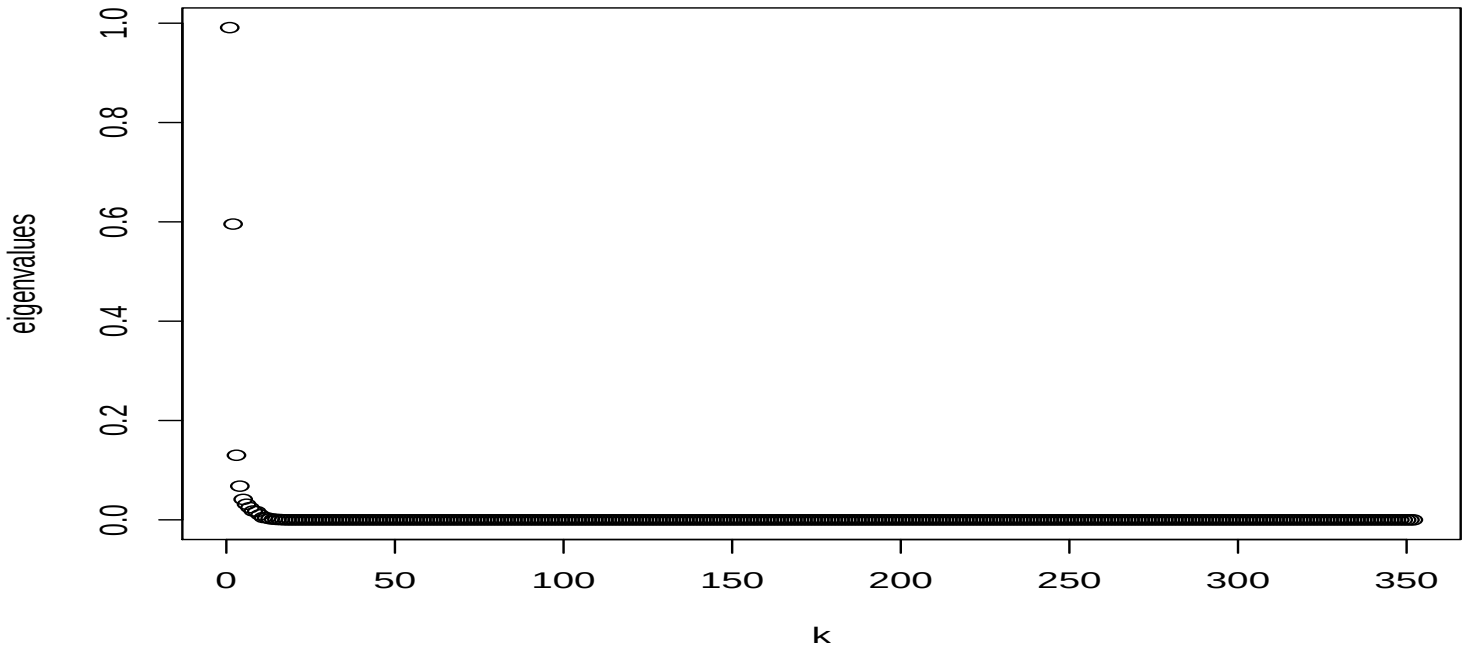


Liquet, B. and Saracco, J. (2012). A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches. *Comput. Stat.*, 27, 103-125.

Choice of the dimension K on the first block

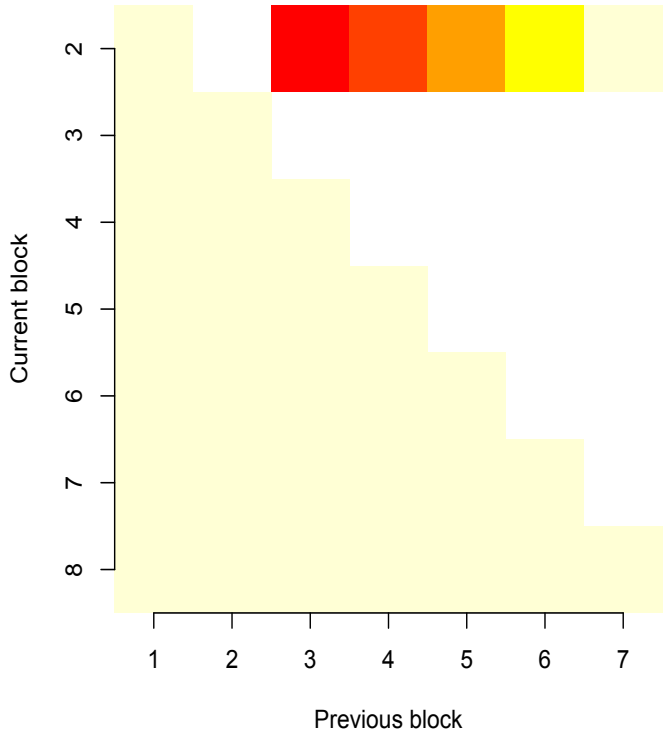


Choice of the dimension K on the first block

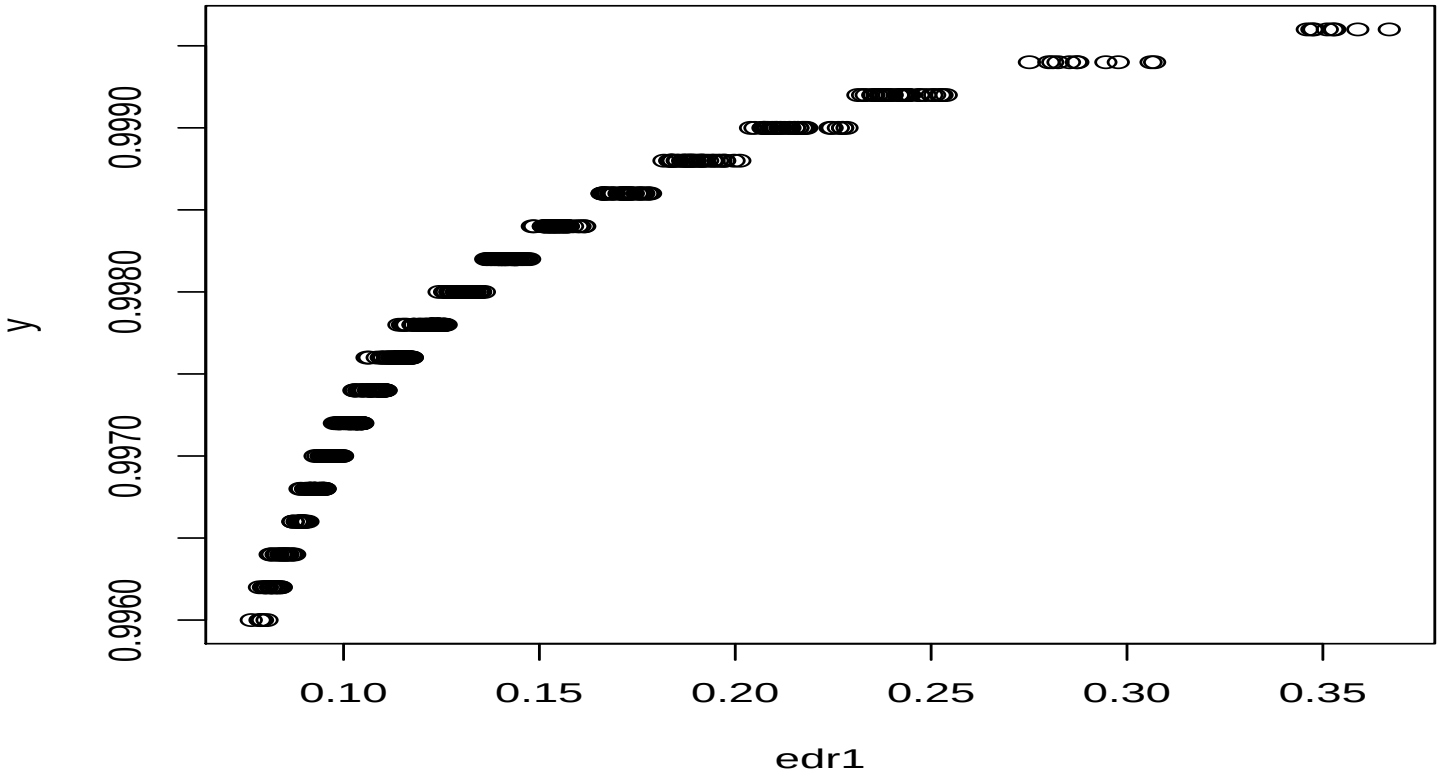


$\hookrightarrow$ **Conclusion :** $\widehat{K} = 2$

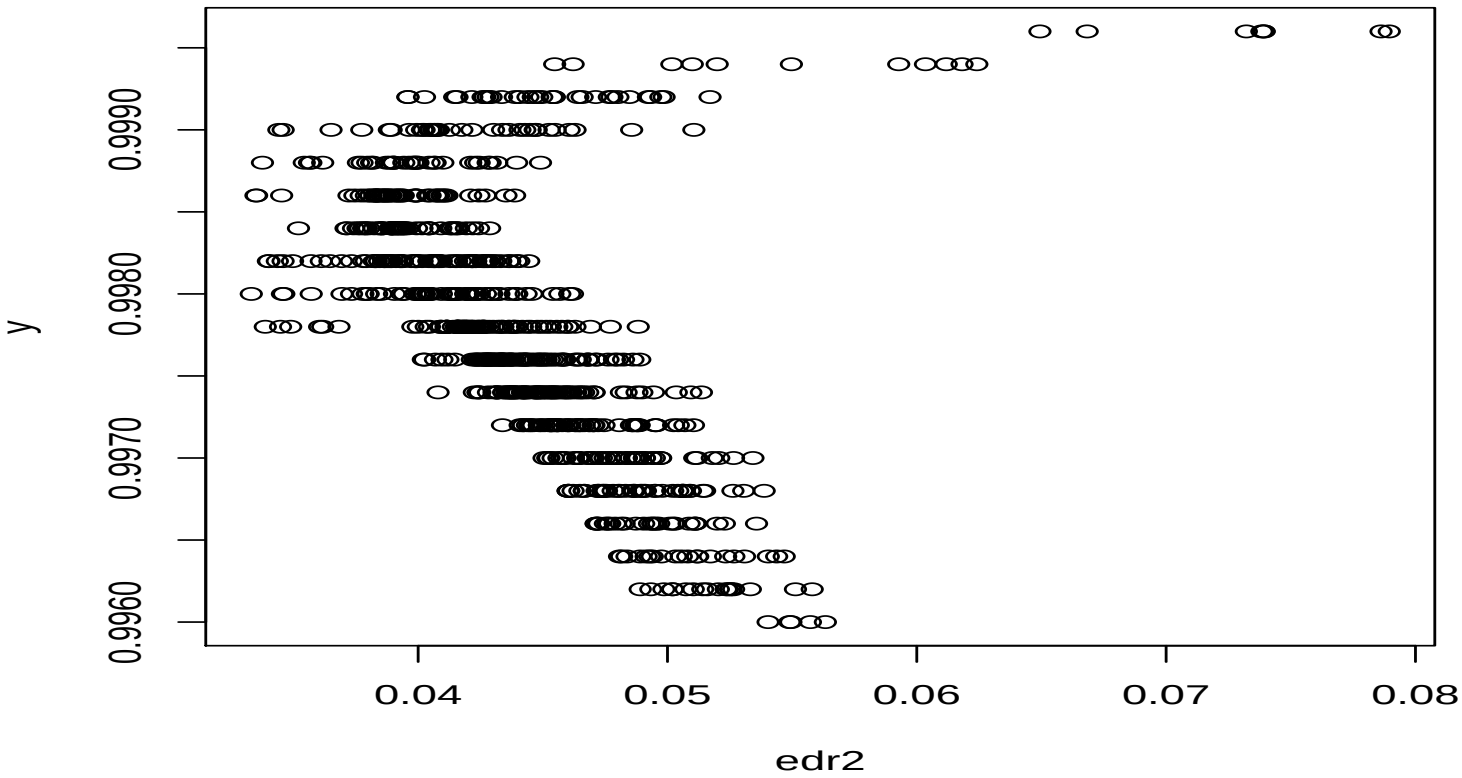
Image of the weights for the first 8 blocks



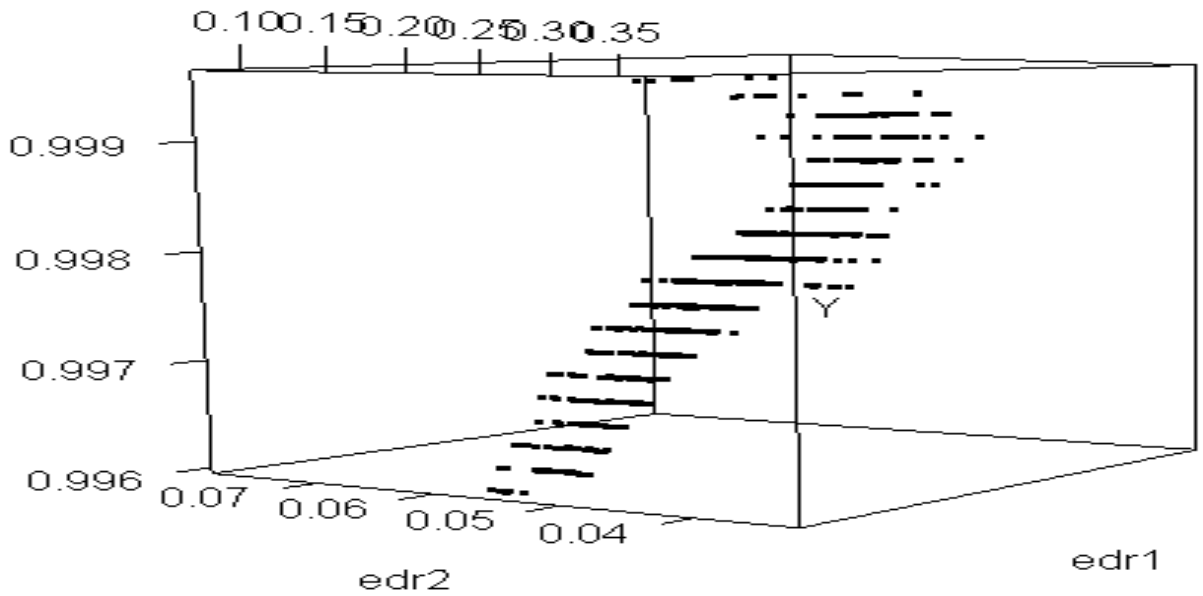
Visualization of the data



Visualization of the data

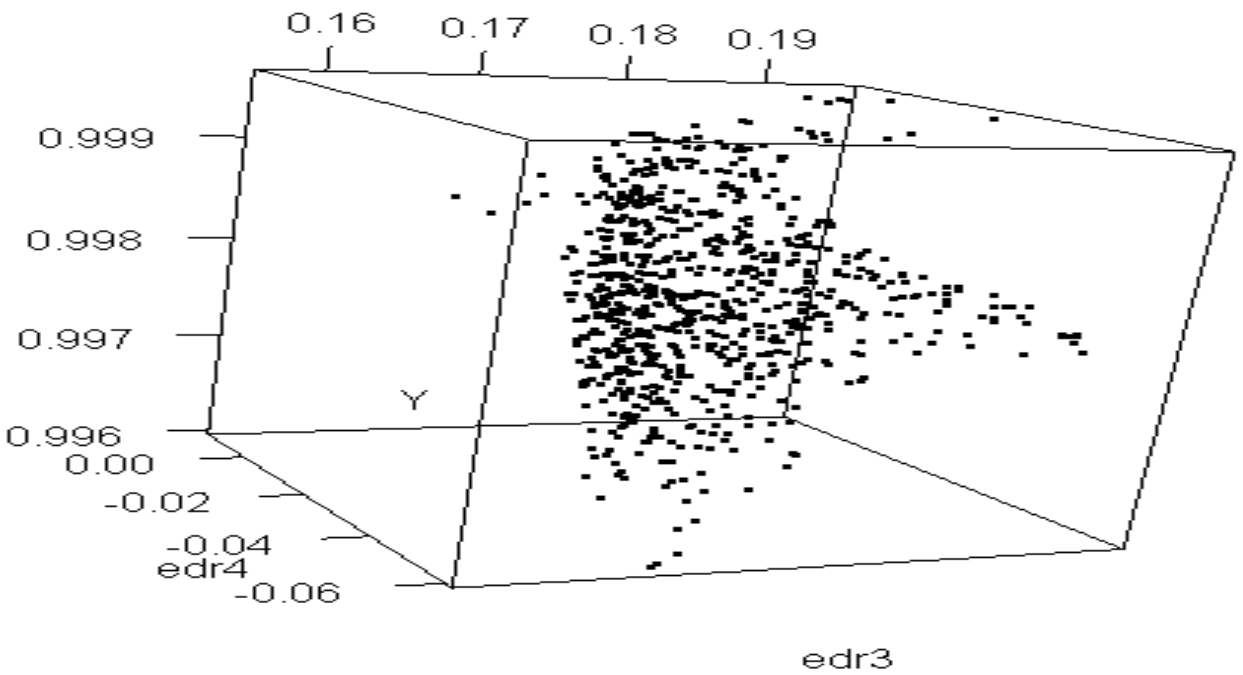


Visualization of the data



.

Visualization of the data



.

"Plot" of the estimated EDR direction