

# An adaptive SIR method for block-wise evolving data streams

**M. Chavent, S. Girard, V. Kuentz, B. Liquet,  
T.M.N.Nguyen, J. Saracco**

INRIA Bordeaux Sud Ouest & Institut de Mathématiques de Bordeaux  
University of Bordeaux & CEMAGREF & INRIA Rhône-Alpes  
France

ASMDA 2011 - Rome

# Outline

- 1 Introduction
- 2 Presentation of SIRdatastream : an adaptive SIR estimator
  - Population version of SIRdatastream
  - Sample version of SIRdatastream
  - An asymptotic result
  - Computational complexity and running time
- 3 A simulation study
- 4 Concluding remarks

# Outline

- 1 Introduction
- 2 Presentation of SIRdatastream : an adaptive SIR estimator
  - Population version of SIRdatastream
  - Sample version of SIRdatastream
  - An asymptotic result
  - Computational complexity and running time
- 3 A simulation study
- 4 Concluding remarks

**Initial motivation of this work : an applied problem.** Our approach SIRdatastream will be applied 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  $X$ .

To this end, a **stream** of synthetic spectra is generated by a physical radiative transfer model. The high dimension of spectra ( $p = 184$  wavelengths) will be reduced using SIR method (regularized version) and the proposed adaptive SIR method for data stream.

**In this communication :**

↔ Presentation of SIRdatastream from a theoretical point of view

↔ Illustration on simulated data.

↔ BUT not enough time for presenting the application (and not enough expert...)

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

$$Y = f(X'\beta, \epsilon) \quad (1)$$

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  $\epsilon$  is independent of  $X$ ,
- the link function  $f$  and the vector  $\beta$  are unknown.

Since  $f$  is unknown,  $\beta$  is not totally identifiable in this model.

Then we are interested in finding the linear subspace spanned by  $\beta$ , called the **Effective Dimension Reduction (EDR) space**.

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  $\{(X_i, Y_i), i = 1, \dots, n_t\}$  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 **an adaptive SIR method**, called **SIRdatastream**.

## Recall on SIR in block $t$

The population version SIR relies on the following linear condition :

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

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  $T(\cdot)$  of  $Y$ .

Under condition (C) and model (1), Li (1991) showed that the **principal eigenvector**  $b_t$  of

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

is **an EDR direction** (i.e. is collinear with  $\beta$ ).



To obtain an estimator of  $\Gamma_t$  which can be easily estimated and used in practice, Li (1991) proposed for  $T(\cdot)$  a slicing into  $H_t \geq 2$  non-overlapping slices  $s_1, \dots, s_{H_t}$ .

Denoting the  $h$ th slice weight (resp. mean) by  $p_h = P(Y \in s_h)$  (resp.  $m_h = E(X|Y \in s_h)$ ), then the matrix  $\Gamma_t$  can be written as

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

To estimate the matrix  $\Gamma_t$  : substitute theoretical versions of the moments by their empirical counterparts.

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

# Outline

- 1 Introduction
- 2 Presentation of SIRdatastream : an adaptive SIR estimator
  - Population version of SIRdatastream
  - Sample version of SIRdatastream
  - An asymptotic result
  - Computational complexity and running time
- 3 A simulation study
- 4 Concluding remarks

## Population version of SIRdatastream

Let us denote by  $b_t$  the EDR direction obtained in the block  $t$ .  
We consider the matrix

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

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

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

Reformulation of this approach as an **optimization problem** :

$$\max_{v \in \mathbb{R}^p} \frac{v' M_T v}{v' v}. \quad (2)$$

The **solution** is clearly the **normalized principal eigenvector** of  $M_T$ , denoted by  $v_T$  hereafter.

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

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

## Sample version of SIRdatastream

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

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

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

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

# An asymptotic result

## Assumptions :

We consider a fixed number  $T$  of blocks and a total sample size  $n$  which tends to  $\infty$ .

Let  $n_{h,t}$  be the number of observations in the  $h$ th 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  $p_h \neq 0$ ,  $h = 1, \dots, H_t$ .
- (A3) For  $t = 1, \dots, T$  and  $h = 1, \dots, H_t$ ,  
 $n_{h,t} \rightarrow \infty$  (and therefore  $n_t \rightarrow \infty$ ) as  $n \rightarrow \infty$ .

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

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

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

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

## Computational complexity

For sake of simplicity, let us assume that **each block  $t$  has the same sample size  $n^*$** .

Let us denote by **SIRglobal** the usual SIR approach based on the union of the  $T$  blocks.

- **SIRdatastream approach performs faster than SIRglobal provided that the sample size  $n^*$  is large enough :**

$$n^* > 2(p + 1).$$

- When the total number  $T$  of blocks increases,  
**some problems of data storage may appear for SIRglobal** (the dataset used becomes larger and larger) ;  
**SIRdatastream approach only needs the storage of the last block and of the previous estimated EDR directions (which are only  $p$ -dimensional vectors).**



## Running time

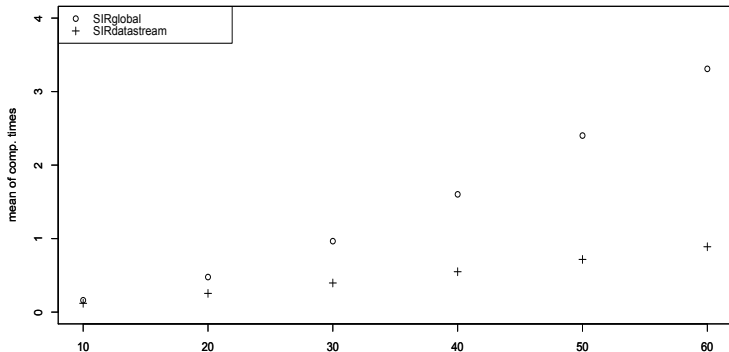
We compare the running time (in seconds) of our **SIRdatastream** approach with **SIRglobal** (based on the union of the first  $T$  blocks).

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.

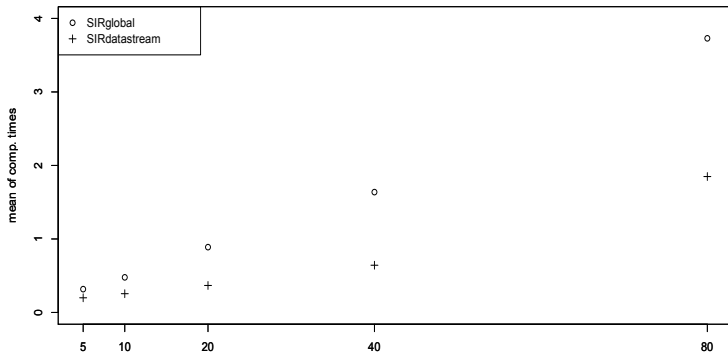
We generate  $\mathcal{B} = 20$  data streams for each  $(T, p, n^*)$  and we calculate the mean of running times.

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



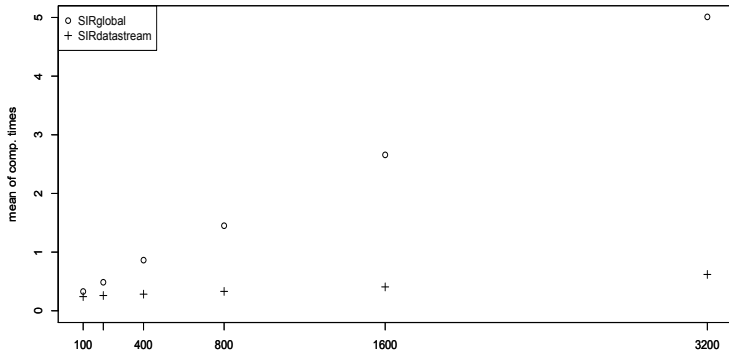
The number  $T$  of blocks hugely penalizes the SIRglobal approach by comparison with SIRdatastream.

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



The dimension  $p$  noticeably favours SIRdatastream versus SIRglobal.

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



The block size  $n^*$  hugely penalizes the SIRglobal approach in comparison with SIRdatastream.

# Outline

- 1 Introduction
- 2 Presentation of SIRdatastream : an adaptive SIR estimator
  - Population version of SIRdatastream
  - Sample version of SIRdatastream
  - An asymptotic result
  - Computational complexity and running time
- 3 A simulation study
- 4 Concluding remarks

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

$$Y = (X'\beta)^3 + \epsilon, \quad (5)$$

where  $X$  follows the  $p$ -dimensional normal distribution  $\mathcal{N}_p(0_p, \Sigma)$  with the covariance  $\Sigma$  arbitrarily chosen,  $\epsilon$  follows the normal distribution  $\mathcal{N}(0, 1)$  and is independent of  $X$ .

For the slope parameter  $\beta$ , we consider **various scenarios**.

For each scenario, we generate  $T = 60$  blocks of size  $n^* = 200$   
with  $p = 20$ .

- **Scenario 1 :  $\beta$  is constant for all the  $T$  blocks.** We fix  $\beta = \beta_0$  with  $\beta_0 = (1, -1, 2, -2, 0, \dots, 0)'$ .
- **Scenario 2 :  $\beta$  is constant for  $T - 1$  blocks and the 10th block is aberrant.** We fix  $\beta = \beta_0$  for each block  $t$  with  $t \neq 10$  and we set  $\beta = \beta_1$  for the 10th block with  $\beta_1 = (1, 1, \dots, 1)'$ .
- **Scenario 3 :  $\beta = \beta_0$  for the first 9 blocks and  $\beta = \beta_1$  for the remaining 51 ones.**
- **Scenario 4 :  $\beta = \beta_0$  for the first 9 blocks and  $\beta$  takes different values for the remaining 51 blocks.** The 51 slope parameters  $\beta$  have been randomly generated.

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

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

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

At each time  $t$  (i.e. when the first  $t$  blocks are available), we estimate the EDR direction with **SIRdatastream** and **SIRglobal**. We also estimate the EDR direction with usual **SIR based only on the data of this block  $t$** .



In the following, for each scenario, we show

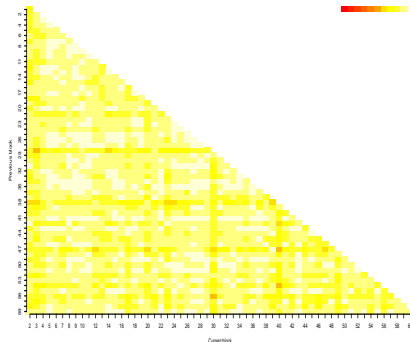
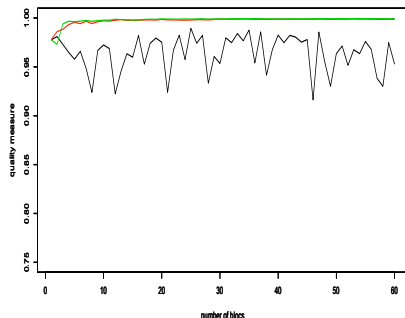
- the plot of the quality measure  $\cos^2(\hat{\beta}, \beta_0)$  of the estimation  $\hat{\beta}$  obtained with **SIRdatastream**, **SIRglobal** and **SIR estimators at each time t**.
- an image of the weights  $\cos^2(\hat{b}_t, \hat{b}_T)$  used in the computation of  $\hat{v}_T$ .

The lighter (yellow) is the color, the larger is the weight.

↔ Red color stands for very small square cosines

*This image will provide to the user an interesting graphic in order to **detect if a drift occurs or if aberrant blocks appear in the data stream.***

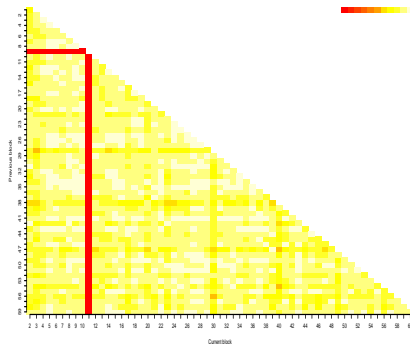
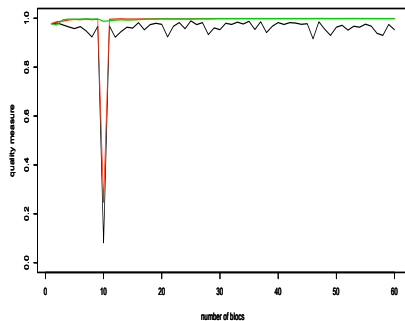
## Scenario 1 : a common direction in all the 60 blocks



SIRdatastream and SIRglobal perform well (but SIRdatastream is an efficient method from running time and data storage points of view).

The image of the weights does not exhibit any drift or aberrant block.

## Scenario 2 : the 10th block is aberrant

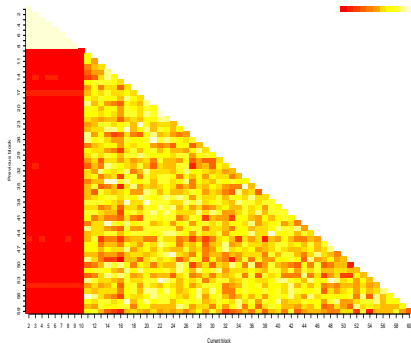
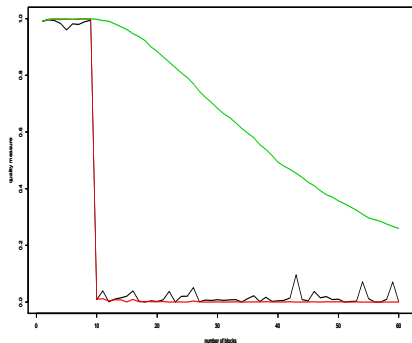


SIRdatastream performs well except for the 10th block, while SIRglobal always works well.

The image of the weights clearly indicates that this block is aberrant and then the effect of this block on the SIRdatastream estimator disappears when the new blocks are available.

NB :  $\cos^2(\hat{\beta}_{\text{datastream}}, \beta_1) \simeq 0.95$ .

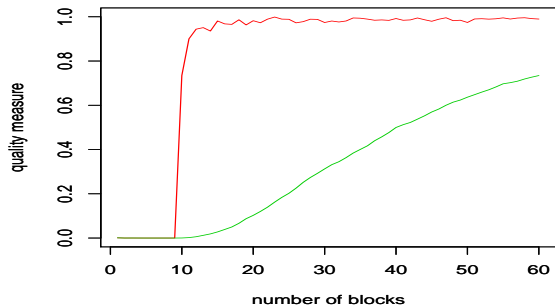
## Scenario 3 : a drift occurs from the 10th block ( $\beta_0$ to $\beta_1$ )



The image of the weights clearly shows that there is a drift from the 10th block to the last one.

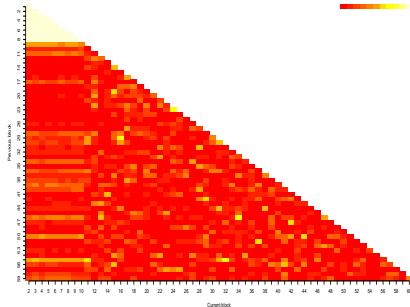
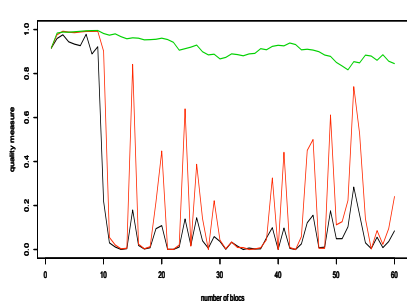
The estimation of the true direction  $\beta_0$  is efficient for SIRdatastream and SIRglobal for the first 9 blocks and then becomes worse for the next blocks.

**Scenario 3 (next)** : plot of the quality measure  $\cos^2(\hat{\beta}, \beta_1)$  versus the number  $T$  of blocks



SIRdatastream is efficient to estimate the true direction  $\beta_1$  from the 10th block to the last one, whereas this is not the case for SIRglobal.

## Scenario 4 : from the 10th block to the last one, there is no common direction $\beta$



The image of the weights clearly indicates that there is no common structure from the 10th block to the last one.

The estimation of the true direction  $\beta_0$  is efficient for SIRdatastream for the first 9 blocks and then becomes worse for the next blocks.

SIRglobal still provides estimates close to the direction of  $\beta_0$  after the 10th block even if there is no structure in this case, which may cause troubles in practical situations.

# Outline






- 1 Introduction
- 2 Presentation of SIRdatastream : an adaptive SIR estimator
  - Population version of SIRdatastream
  - Sample version of SIRdatastream
  - An asymptotic result
  - Computational complexity and running time
- 3 A simulation study
- 4 Concluding remarks

- The proposed approach performs well on **simulated data**.
- It is possible to extend this approach to **multiple indices models** :
  - $b_t$  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 SIR-II, SAVE,  $SIR_\alpha$  or multivariate SIR).



# Thanks for your attention.

The proposed adaptive SIR method will be used to evaluate the physical properties of surface materials on the planet Mars from hyperspectral images. Our goal is to estimate the function  $G$  between some physical parameters  $Y$  and observed spectra  $X$ . To this end, a stream of synthetic spectra is generated by a physical radiative transfer model. The high dimension of spectra ( $p = 184$  wavelengths) will be reduced using regularized SIR (see Bernard-Michel *et al.* (2009) for details) and the proposed adaptive SIR method.

-  R. Coudret, S. Girard, & J. Saracco. "A new sliced inverse regression method for multivariate response", Computational Statistics and Data Analysis, 77, 285–299, 2014.
-  M. Chavent, S. Girard, V. Kuentz, B. Lique, T.M.N. Nguyen & J. Saracco. "A sliced inverse regression approach for data stream", Computational Statistics, 29, 1129–1152, 2014.
-  C. Bernard-Michel, S. Douté, M. Fauvel, L. Gardes & S. Girard. "Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression", Journal of Geophysical Research - Planets, 114, E06005, 2009.
-  C. Bernard-Michel, L. Gardes & S. Girard. "A Note on Sliced Inverse Regression with Regularizations", Biometrics, 64, 982–986, 2008.
-  C. Bernard-Michel, L. Gardes & S. Girard. "Gaussian Regularized Sliced Inverse Regression", Statistics and Computing, 19, 85–98, 2009.