An adaptive SIR method for block-wise evolving data streams

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

An adaptive SIR method for block-wise evolving data streams

Marie Chavent[1], Stéphane Girard[2], Vanessa Kuentz[3], Benoît Liquet[4], Thi Mong Ngoc Nguyen[1], and Jérôme Saracco[1]

[1] University of Bordeaux, Institut de Mathématiques de Bordeaux, INRIA Bordeaux Sud-Ouest, team CQFD, 351 cours de la libération, 33405 Talence Cedex, France (E-mail: {chavent,nguyen,saracco}@math.u-bordeaux1.fr)
[2] INRIA Rhône-Alpes, team Mistis, Inovallée, 655, avenue de l'Europe, Montbonnot, 38334 Saint-Ismier Cedex, France (E-mail: stephane.girard@inrialpes.fr)
[3] CEMAGREF, Unité ADBX "Aménités et Dynamiques des Espaces Ruraux", 50 avenue de Verdun - Gazinet, 33612 CESTAS Cedex (E-mail: vanessa.kuentz@cemagref.fr)
[4] INSERM U897, ISPED, Université Victor Segalen Bordeaux 2, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France (E-mail: benoit.liquet@isped.u-bordeaux2.fr)

Abstract. In this communication, we consider block-wise evolving data streams. When a semiparametric regression model involving a common dimension reduction direction β is assumed for each block, we propose an adaptive SIR (for sliced inverse regression) estimator of β. This estimator is faster than usual SIR applied to the union of all the blocks, both from computational complexity and running time points of view. We show the consistency of our estimator at the root-n rate. In a simulation, we illustrate the good numerical behaviour of the estimator. We also provide a graphical tool in order to detect if there exists a drift of the dimension reduction direction or some aberrant blocks of data. We illustrate our approach with various scenarios. Finally, possible extensions of this method are given.

Keywords: dimension reduction, sliced inverse regression (SIR), data stream.

1 Introduction

In the framework of high dimensional data, we consider the following semiparametric dimension reduction single index model proposed by Duan and Li [3]:

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

where the univariate response variable Y is linked with the p-dimensional regressor X (with expectation $E(X) = \mu$ and covariance matrix $V(X) = \Sigma$) only through the single index $X'\beta$. The error term ϵ is independent of X. The link function f and the vector β are unknown. Since β is not totally

identifiable in this model, we are interested in finding the linear subspace spanned by β, called the Effective Dimension Reduction (EDR) space.

In this paper we focus on data arriving sequentially by block in a stream. We assume that each data block t is composed of an independent and identically distributed (i.i.d.) sample $\{(X_i, Y_i), i = 1, \ldots, n_t\}$ available from model (1). Our goal is to estimate the EDR direction at each arrival of a new block of observations. A simple direct approach consists in pooling all the observed blocks and then estimate the EDR direction by the Sliced Inverse Regression method introduced by Li [5]. While SIR is a computationally simple and fast method, the drawback of pooling the data is the storage of the blocks since the size of the dataset considerably increases with the number of blocks. To avoid this, we propose an adaptive SIR method based on the SIR approach for a stratified population developped by Chavent *et al.* [2]. 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 regularizd SIR (see Bernard-Michel *et al.* [1] for details) and the proposed adaptive SIR method.

2 An adaptive SIR estimator

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. Moreover, in the presence of high-dimensional data, this condition is often approximately fulfilled, see Hall and Li [4] for details. Let us consider a monotone transformation $T(.)$ of Y. Under condition (C) and model (1), Li [5] showed that the centered inverse regression curve is contained in the one-dimensional linear subspace of $\Re^p$ spanned by $\Sigma\beta$. As a consequence, the eigenvector b_t of $\Sigma^{-1}\Gamma_t$ associated with the nonnull eigenvalue is an EDR direction (i.e. is collinear with β) where $\Gamma_t = V(E(X|T(Y)))$.

To obtain an estimator of Γ_t which can be easily used in practice, Li [5] 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

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

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

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 our model, the term $\cos^2(b_t, b_T)$ is equal to one since b_t and b_T are both colinear with β. Under the assumptions of our model, it is also straightforward to show that the principal eigenvector of M_T is colinear with β and then is an EDR direction.

Let us remark that it is possible to reformulate this approach as the following optimization problem:

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

Indeed since $||b_t|| = 1$, we have:

$$\sum_{t=1}^{T} w_t \cos^2(b_t, v) = \sum_{t=1}^{T} w_t < b_t, v >^2 = \sum_{t=1}^{T} w_t v' b_t b_t' v = v' \left(\sum_{t=1}^{T} w_t b_t b_t' \right) v = v' M_T v,$$

thus maximization problem (2) can be rewritten as

$$\max_{v \in \Re^p} \frac{v' M_T v}{v' v}. \tag{3}$$

The solution of (3) is clearly the normalized principal eigenvector of M_T. Let us denote by v_T this eigenvector.

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 with the SIRdatastream approach is the principal eigenvector of the $p \times p$ matrix defined as

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

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

Asymptotics. The following assumptions are necessary to state our asymptotic result for a fixed number T of blocks and a sample global 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 $h = 1, \ldots, H_t$, $n_{h,t} \to \infty$ (and therefore $n_t \to \infty$) as $n \to \infty$.

For each block t and under the assumptions (C), (A1)-(A3), from SIR theory of Li [5] each estimated EDR direction $\hat{b}_t$ converges to b_t at root n rate: that is, for $t = 1, \ldots, T$, $\hat{b}_t = b_t + O_p(n^{-1/2})$. Since $\cos^2(\hat{b}_t, \hat{b}_T) = \cos^2(b_t, b_T) + O_p(n^{-1/2}) = 1 + O_p(n^{-1/2})$, we get $\widehat{M}_T = M_T + O_p(n^{-1/2})$. Therefore the principal eigenvector of $\widehat{M}_T$ converges to the corresponding one of M_T at the same rate: $\widehat{v}_T = v_T + O_p(n^{-1/2})$. Since v_T is colinear with β, then the estimated EDR direction $\widehat{v}_T$ converges to an EDR direction at root n rate.

Computational complexity. For sake of simplicity, let us assume that each block has the same sample size n^*. One can show that the SIRdatastream approach performs faster than usual SIR based on the union of the first T blocks (called SIRglobal hereafter) provided that the sample size n^* is large enough: $n^* > 2(p+1)$. Moreover, when the total number T of blocks increases, the dataset used for SIR becomes larger and larger, and problem of data storage may appear. On the contrary, our SIRdatastream approach only needs the storage of the last block (necessary of the two last ones if a drift seems to occur) and the previous estimated EDR directions (which are only p-dimensional vectors).

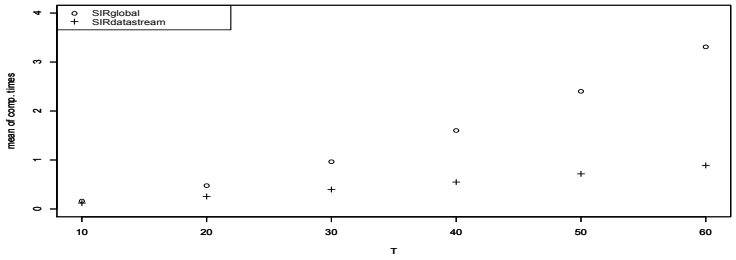
Running time. We compare the running time (in seconds) of our SIRdatastream approach with SIRglobal (based on the union of the first T blocks). For various values of the dimension p, the size n^* of each block and the total number T of blocks, we generate $\mathcal{B} = 20$ data streams and we evaluate the computational times for the two methods. Unsurprisingly one can observe in Figure 1 that the dimension p noticeably favours SIRdatastream versus usual SIR while the number T of blocks and the block size n^* hugely penalize the usual SIR approach in comparison with SIRdatastream.

3 A simulation study

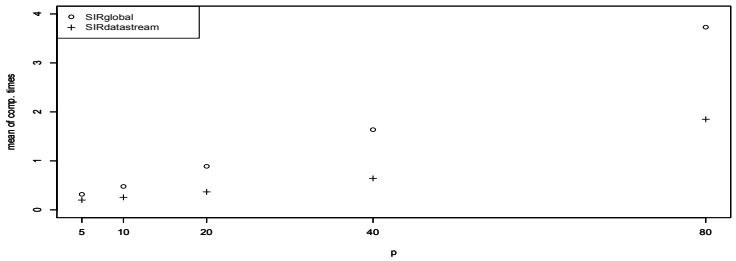
We consider for each block of data the same following semiparametric regression model:

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

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



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



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

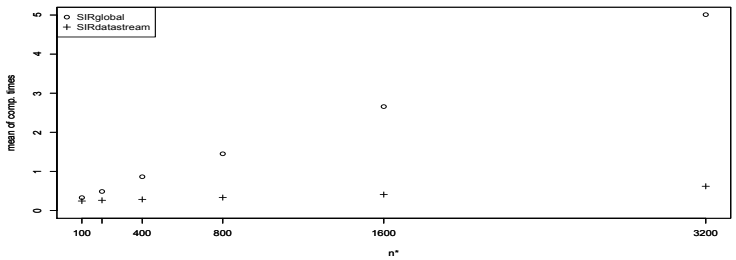


Fig. 1. Running times (in seconds) for various values of p, n^* and T.

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 $\mathcal{N}(0, 1)$ and is independent of X. For the slope parameter β, we consider various scenarios. For each scenario, we generate $T = 60$ blocks of size $n^* = 200$ with $p = 20$.

- Scenario 1: β is constant for all the T blocks. We fix $\beta = \beta_0$ with $\beta_0 = (1, -1, 2, -2, 0, \ldots, 0)'$.
- Scenario 2: β 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 $\beta_1 = (1, 1, \ldots, 1)'$.

- 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 β takes different values for the remaining 51 blocks. The 51 slope parameters β have been randomly generated as follows: each component β_j of β is randomly obtained from the normal distribution $\mathcal{N}(0, 1)$.

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.

For each scenario, we generate $T = 60$ blocks as described above. Then at each time t (i.e. when the first t blocks are available) we estimate the EDR direction with SIRdatastream and SIRglobal approaches. 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 plot the quality measure $\cos^2(\widehat{\beta}, \beta_0)$ (resp. $\cos^2(\widehat{\beta}, \beta_1)$) on Figure 2 (resp. 3) of the estimator $\widehat{\beta}$ (obtained with SIRdatastream, SIRglobal or SIR estimators at each time t). We also represent with an image the weights $\cos^2(\widehat{b}_t, \widehat{b}_T)$ used in equation (4). The lighter is the color, the larger is the weight. 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.

Some comments on Figures 2 and 3.

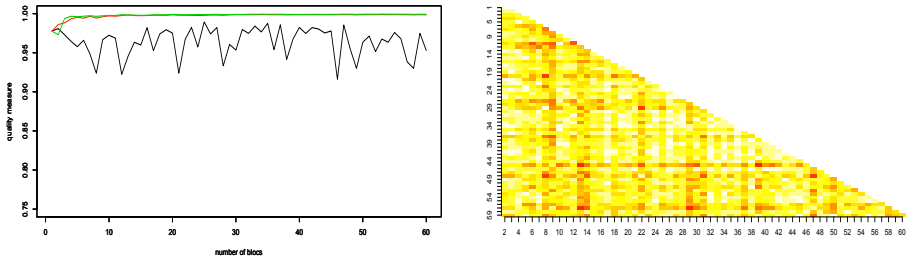
• SIRdatastream and SIRglobal perform well for scenario 1 (but keep in mind that SIRdatastream is an efficient method from running time and data storage points of view). The image of the weights does not exhibit any drift or aberrant block.

• For scenario 2, SIRdatastream and SIRglobal perform well except for the 10th block for SIRdatastream. But 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.

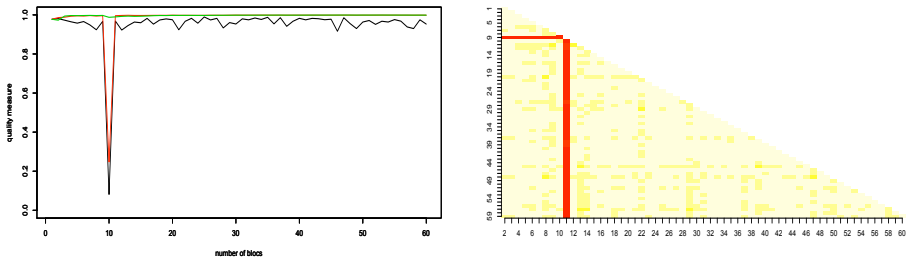
• For scenario 3, the image of the weights clearly shows that there is a drift from the 10th block to the last one. The estimation of the true direction β_0 is efficient for SIRdatastream and SIRglobal for the first 9 blocks and then becomes worse for the next blocks. Note that in Figure 3 one can see that SIRdatastream is efficient to estimate the true direction β_1 from the 10th block to the last one, this is not the case for SIRglobal.

• For scenario 4, 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 β_0 is efficient for SIRdatastream and SIRglobal for the first 9 blocks and then becomes worse for the next blocks. One can remark that SIRglobal always provides estimates close to the direction of β_0 after the 10th block even if there is no structure in this case, which may cause troubles in practical situations. This remark also remains valid for scenario 3 where

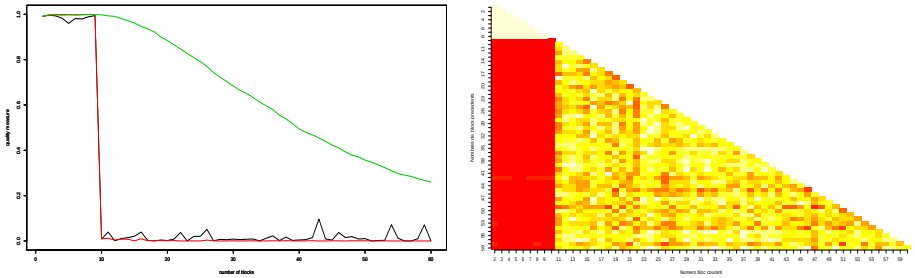
Scenario 1: a common direction in all the $T = 60$ blocks



Scenario 2: the 10th block is aberrant



Scenario 3: a drift occurs from the 10th block (β_0 *to* β_1)



Scenario 4: from the 10th block to the last one, there is no common direction β

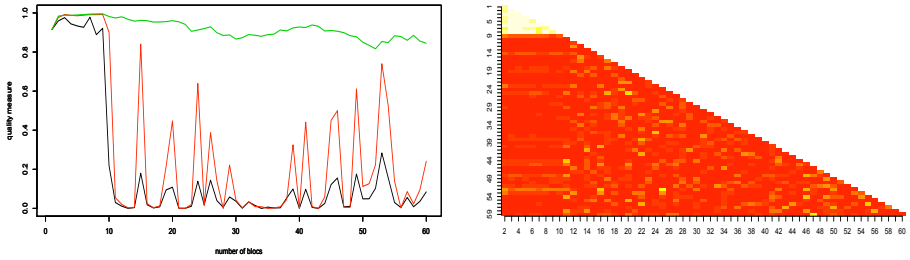


Fig. 2. Numerical behaviour of the SIRglobal and SIRdatastream estimators for various scenarios. On the left: plot of the quality measure $\cos^2(\widehat{\beta}, \beta_0)$ versus the number T of blocks (red for SIRdatastream, green for SIRglobal on the first t blocks, black for SIR on block t only). On the right: image of the weights $\cos^2(\widehat{b}_t, \widehat{b}_T)$ used in the computation of $\widehat{v}_T$.

SIRglobal provides estimates close to the direction of β_0 after the 10th block even if the new underlying direction is β_1. At the end of the 60th block in scenario 3, SIRglobal is still not able to provide a good estimate for β_1, see Figure 3.

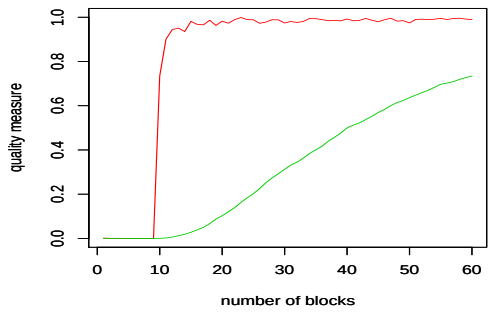


Fig. 3. Scenario 3: plot of the quality measure $\cos^2(\widehat{\beta}, \beta_1)$ versus the number T of blocks (red for SIRdatastream, green for SIRglobal)

4 Concluding remarks

The proposed approach performs well on simulated data. We present in this paper a single index version of the underlying model. It is possible to extend this approach to multiple indices model. In this expression of M_T, the vector b_t (resp. the squared cosine) will be replaced by a basis B_t of the EDR space (resp. a proximity measure between two K-dimensional EDR space, for instance the square trace correlation). It is also possible to use alternative methods instead of SIR (such as SIR-II, SAVE or SIR$_\alpha$ for example). In the next future, SIRdatastream will be applied on real data dealing with the estimation of Mars surface physical properties from hyperspectral images.

References

1.Bernard-Michel, C., Dout, S., Fauvel, M., Gardes, L. and Girard, S. Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression. *Journal of Geophysical Research - Planets*, 114, E06005, 2009.

2.Chavent, M., Kuentz, V., Liquet, B. and Saracco, J. A sliced inverse regression approach for a stratified population. To appear in *Communications in statistics - Theory and methods*, 2011.

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

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

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