SIR for datastreams

Stéphane Girard

Mistis team, INRIA Grenoble Rhône-Alpes.
http://mistis.inrialpes.fr/~girard

Joint work with CQFD team, INRIA Bordeaux Sud-Ouest and Institut de Planétologie et d'Astrophysique de Grenoble (IPAG).

Outline

Outline

Multivariate regression

Let $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$. The goal is to estimate $G : \mathbb{R}^p \to \mathbb{R}$ such that

$$Y = G(X) + \xi \quad \text{where } \xi \text{ is independent of } X.$$

- Unrealistic when p is large (*curse of dimensionality*).
- **Dimension reduction** : Replace X by its projection on a subspace of lower dimension without loss of information on the distribution of Y given X.
- **Central subspace** : smallest subspace S such that, conditionally on the projection of X on S, Y and X are independent.

Dimension reduction

- Assume (for the sake of simplicity) that $\dim(S) = 1$ *i.e.* $S =$span(b), with $b \in \mathbb{R}^p \Longrightarrow$ **Single index model :**

$$Y = g(< b, X >) + \xi$$

  where ξ is independent of X.
- The estimation of the p-variate function G is replaced by the estimation of the univariate function g and of the direction b.
- **Goal of SIR** [Li, 1991] : Estimate a basis of the central subspace. (*i.e.* b *in this particular case.*)

SIR

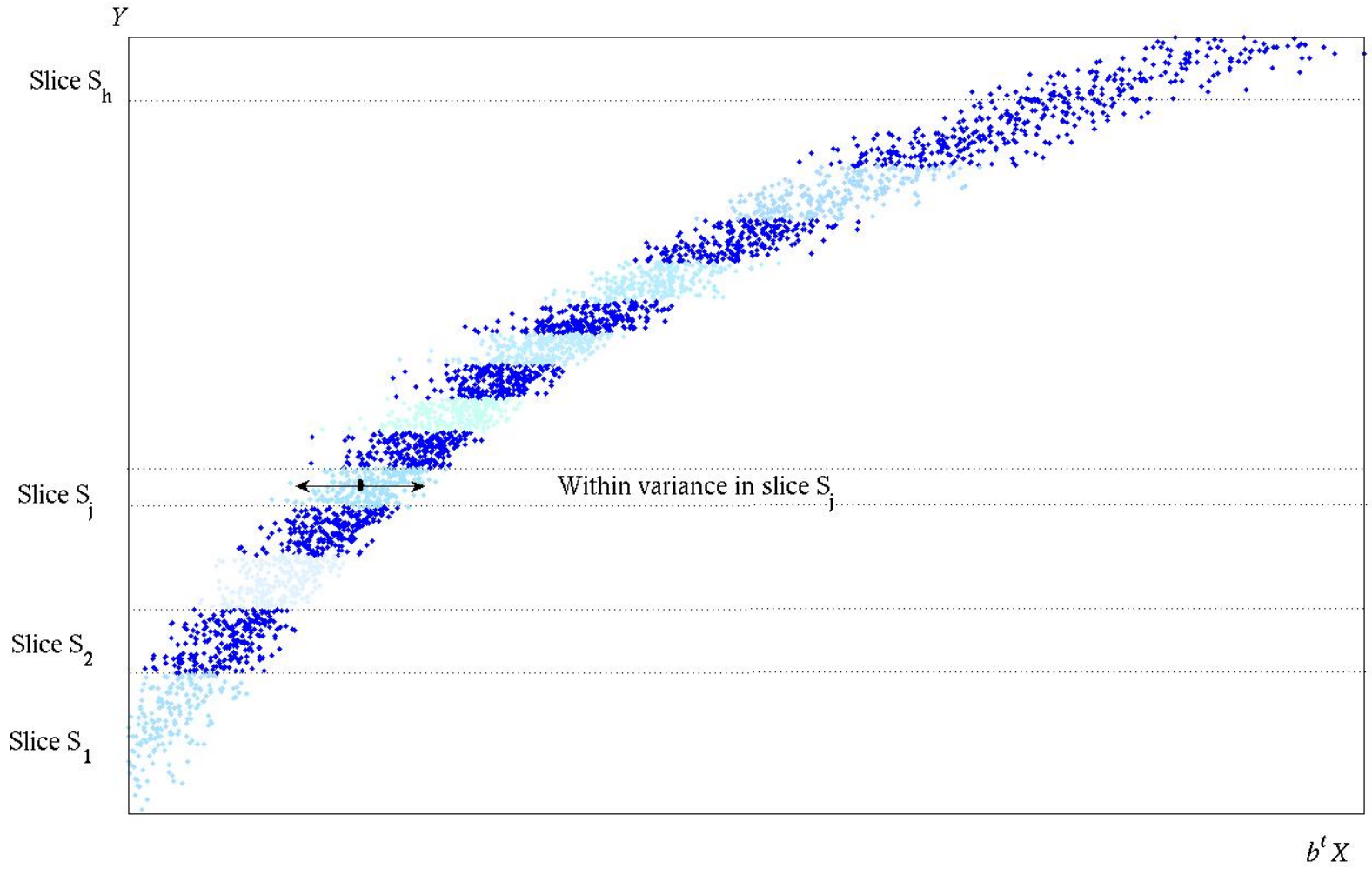
Idea :

- Find the direction b such that $< b, X >$ best explains Y.
- Conversely, when Y is fixed, $< b, X >$ should not vary.
- Find the direction b minimizing the variations of $< b, X >$ given Y.

In practice :

- The support of Y is divided into h slices S_j.
- Minimization of the within-slice variance of $< b, X >$ under the constraint var$(< b, X >) = 1$.
- Equivalent to maximizing the between-slice variance under the same constraint.

Illustration



Estimation procedure

Given a sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, the direction b is estimated by

$$\hat{b} = \underset{b}{\operatorname{argmax}}\, b'\hat{\Gamma}b \text{ such that } b'\hat{\Sigma}b = 1. \tag{1}$$

where $\hat{\Sigma}$ is the empirical covariance matrix and $\hat{\Gamma}$ is the between-slice covariance matrix defined by

$$\hat{\Gamma} = \sum_{j=1}^{h} \frac{n_j}{n}(\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})', \quad \bar{X}_j = \frac{1}{n_j} \sum_{Y_i \in S_j} X_i,$$

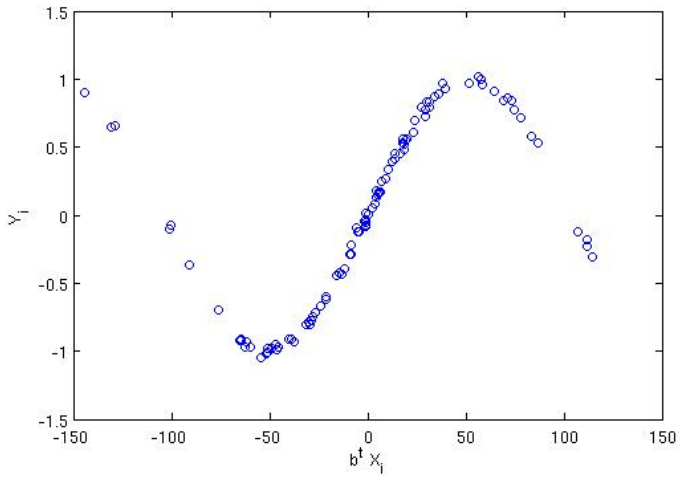
where n_j is the number of observations in the slice S_j.
The optimization problem (1) has a closed-form solution : $\hat{b}$ is the eigenvector of $\hat{\Sigma}^{-1}\hat{\Gamma}$ associated to the largest eigenvalue.

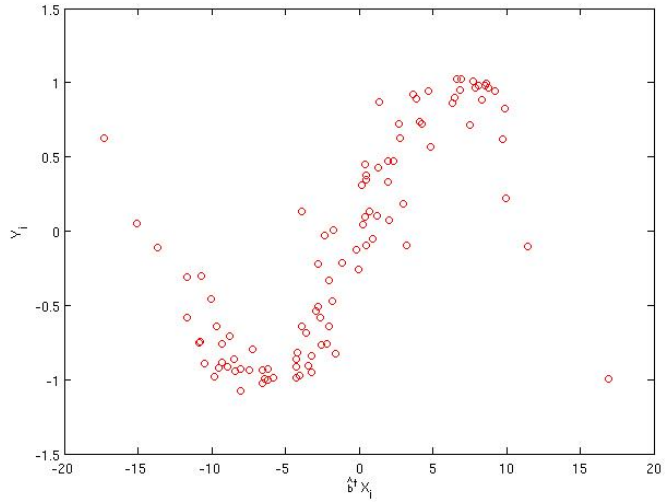
Illustration

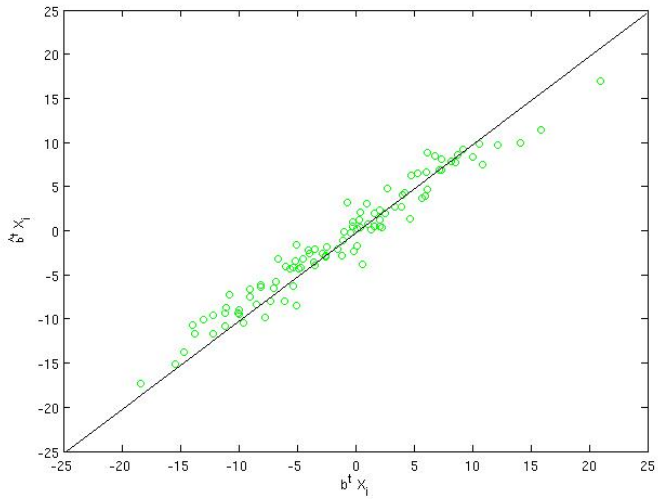
Simulated data.

- Sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of size $n = 100$ with $X_i \in \mathbb{R}^p$, dimension $p = 10$ and $Y_i \in \mathbb{R}$, $i = 1, \ldots, n$.
- $X_i \sim \mathcal{N}_p(0, \Sigma)$ where $\Sigma = Q\Delta Q'$ with
  - $\Delta =$diag$(p^2, \ldots, 2^2, 1^2)$,
  - Q is an orientation matrix drawn from the uniform distribution on the set of orthogonal matrices.
- $Y_i = g(< b, X_i >) + \xi_i$ where
  - g is the link function $g(t) = \sin(\pi t/2)$,
  - b is the true direction $b = 5^{-1/2}Q(1, 1, 1, 1, 1, 0, \ldots, 0)'$,
  - $\xi \sim \mathcal{N}_1(0, 9.10^{-4})$

Results







Blue : Y_i versus the projections $< b, X_i >$ on the true direction b,

Red : Y_i versus the projections $< \hat{b}, X_i >$ on the estimated direction $\hat{b}$,

Green : $< \hat{b}, X_i >$ versus $< b, X_i >$.

Outline

Context

- We consider data arriving sequentially by blocks in a stream.
- Each data block $t = 1, \ldots, T$ is an i.i.d. sample (X_i, Y_i), $i = 1, \ldots, n$ from the regression model $Y = g(< b, X >) + \xi$.
- **Goal** : Update the estimation of the direction b at each arrival of a new block of observations.

Method

- Compute the **individual directions** $\hat{b}_t$ on each block $t = 1, \ldots, T$ using SIR.
- Compute a **common direction** as

$$\hat{b} = \underset{||b||=1}{\operatorname{argmax}} \sum_{t=1}^{T} \cos^2(\hat{b}_t, b) \cos^2(\hat{b}_t, \hat{b}_T).$$

Idea : If $\hat{b}_t$ is close to $\hat{b}_T$ then $\hat{b}$ should be close to $\hat{b}_t$.
Explicit solution : $\hat{b}$ is the eigenvector associated to the largest eigenvalue of

$$M_T = \sum_{t=1}^{T} \hat{b}_t \hat{b}_t' \cos^2(\hat{b}_t, \hat{b}_T).$$

Advantages of SIRdatastream

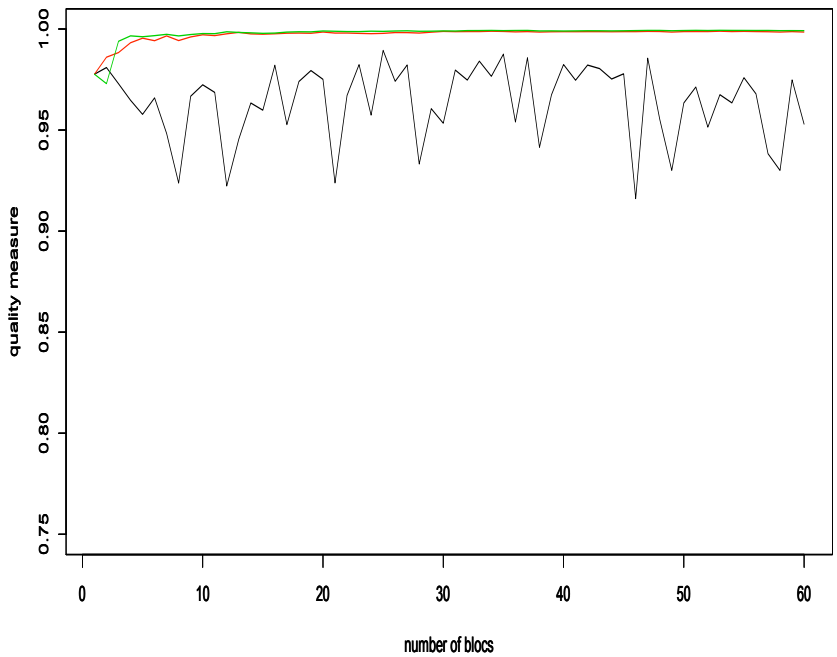
- Computational complexity $O(Tnp^2)$ v.s. $O(T^2np^2)$ for the brute-force method which would consist in applying regularized SIR on the union of the t first blocks for $t = 1, \ldots, T$.
- Data storage $O(np)$ v.s. $O(Tnp)$ for the brute-force method.

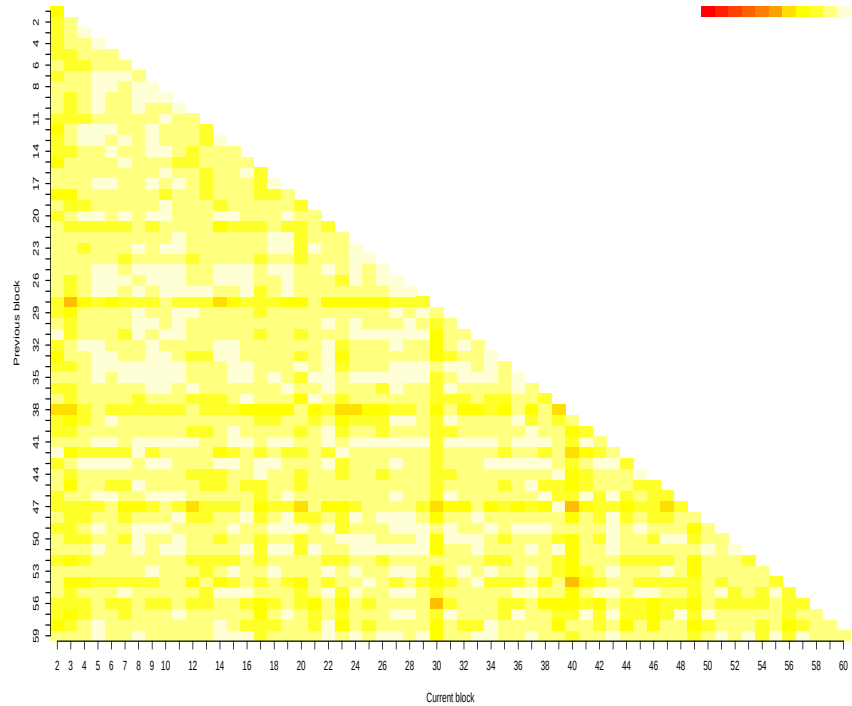
(under the assumption $n >> \max(T, p)$).

- Interpretation of the weights $\cos^2(\hat{b}_t, \hat{b}_T)$.

Illustration on simulations

Scenario 1 : A common direction in all the 60 blocks.

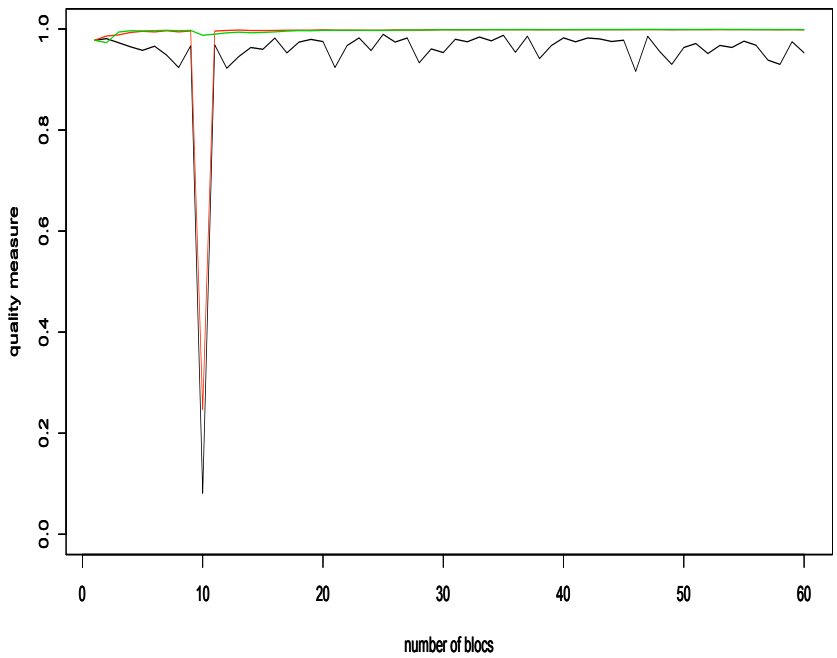


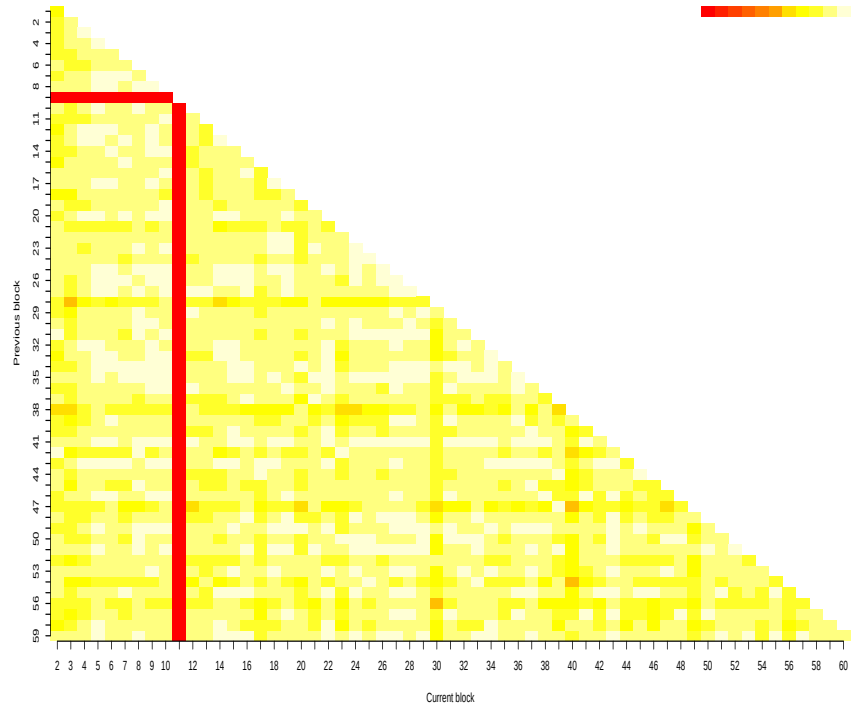


Left : $\cos^2(\hat{b}, b)$ for SIRdatastream, SIR brute-force and SIR estimators at each time t. *Right :* $\cos^2(\hat{b}_t, \hat{b}_T)$. The lighter (yellow) is the color, the larger is the weight. Red color stands for very small squared cosines.

Illustration on simulations

Scenario 2 : The 10th block is different from the other ones.

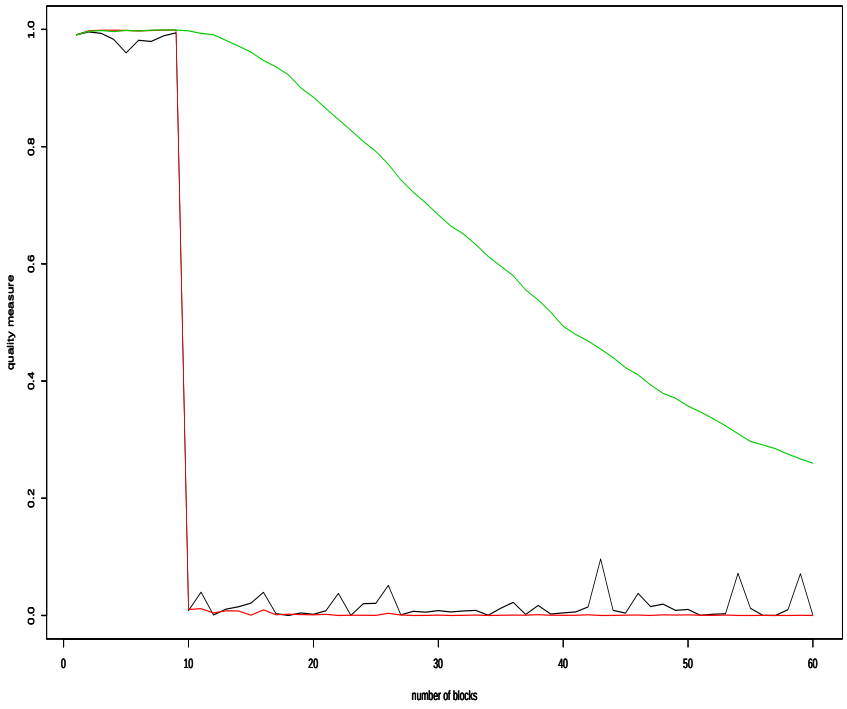


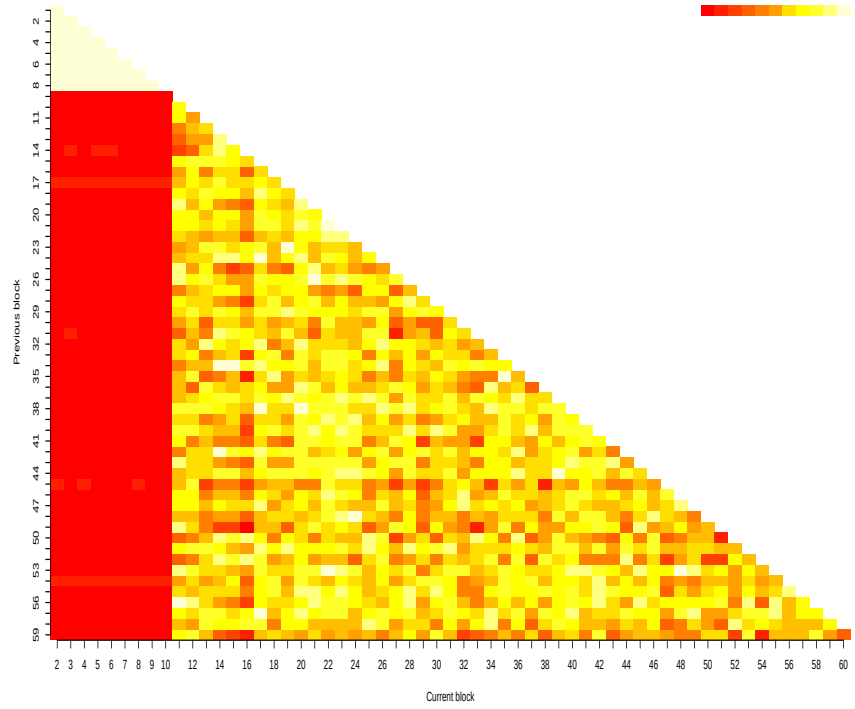


Left : $\cos^2(\hat{b}, b)$ for SIRdatastream, SIR brute-force and SIR estimators at each time t. *Right :* $\cos^2(\hat{b}_t, \hat{b}_T)$. The lighter (yellow) is the color, the larger is the weight. Red color stands for very small squared cosines.

Illustration on simulations

Scenario 3 : A drift occurs from the 10th block (b to $\tilde{b}$)

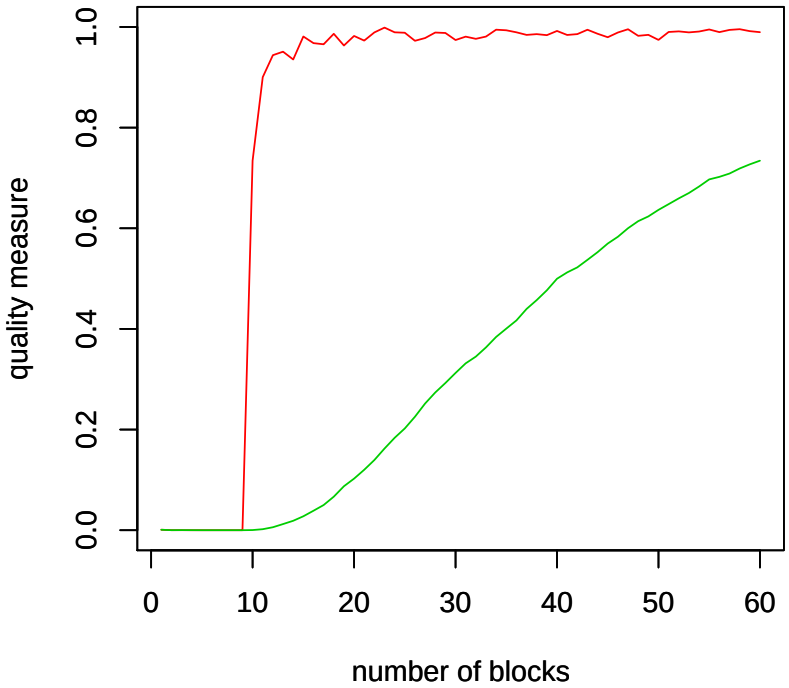


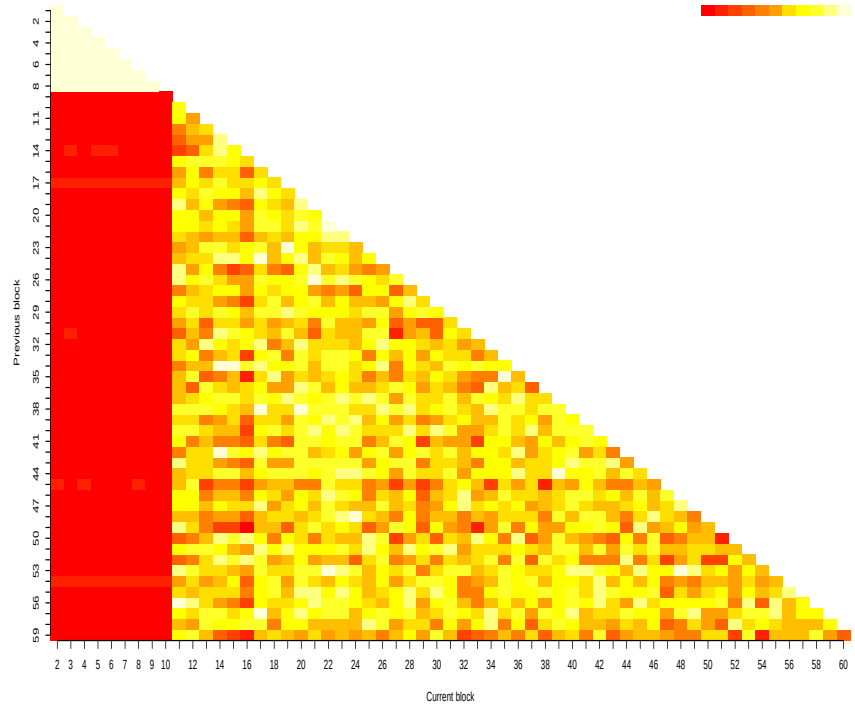


Left : $\cos^2(\hat{b}, b)$ for SIRdatastream, SIR brute-force and SIR estimators at each time t. *Right :* $\cos^2(\hat{b}_t, \hat{b}_T)$. The lighter (yellow) is the color, the larger is the weight. Red color stands for very small squared cosines.

Illustration on simulations

Scenario 3 (cont'd) :A drift occurs from the 10th block (b to $\tilde{b}$)

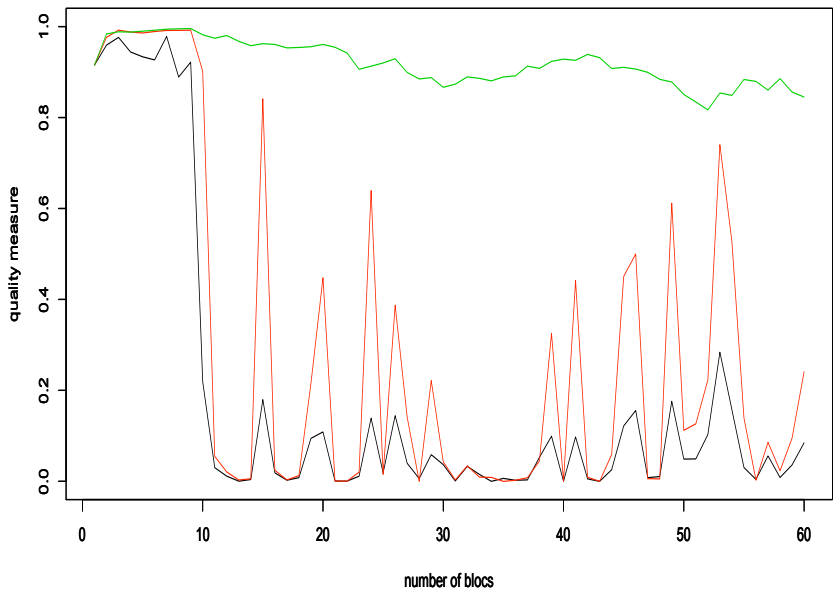


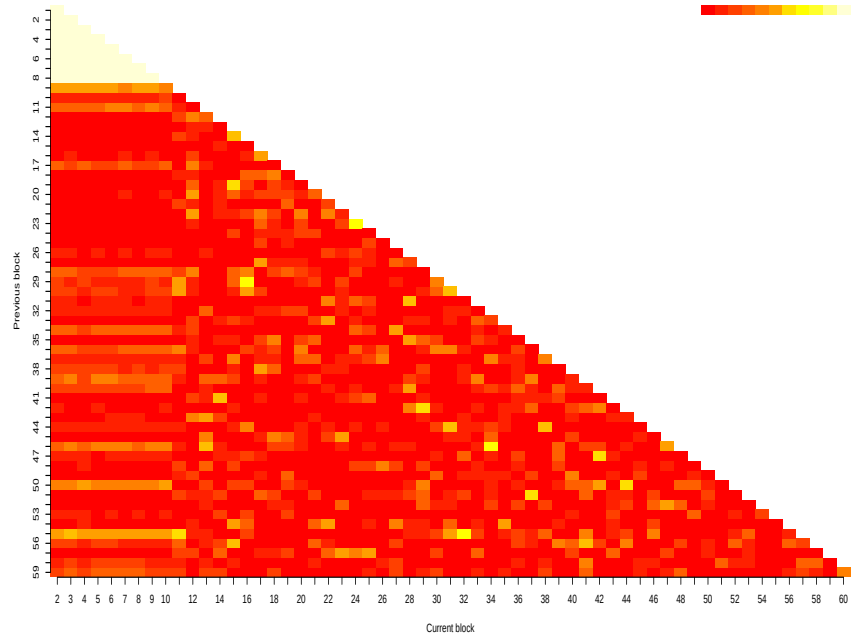


Left : $\cos^2(\hat{b}, \tilde{b})$ for SIRdatastream and SIR brute-force. *Right :* $\cos^2(\hat{b}, \tilde{b})$

Illustration on simulations

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





Left : $\cos^2(\hat{b}, b)$ for SIRdatastream, SIR brute-force and SIR estimators at each time t. *Right :* $\cos^2(\hat{b}_t, \hat{b}_T)$. The lighter (yellow) is the color, the larger is the weight. Red color stands for very small squared cosines.

Outline

Estimation of Mars surface physical properties from hyperspectral images

Context :

- Observation of the south pole of Mars at the end of summer, collected during orbit 61 by the French imaging spectrometer OMEGA on board Mars Express Mission.
- 3D image : On each pixel, a spectra containing $p = 184$ wavelengths is recorded.
- This portion of Mars mainly contains water ice, CO_2 and dust.

Goal : For each spectra $X \in \mathbb{R}^p$, estimate the corresponding physical parameter $Y \in \mathbb{R}$ (grain size of CO_2).

An inverse problem

Forward problem.

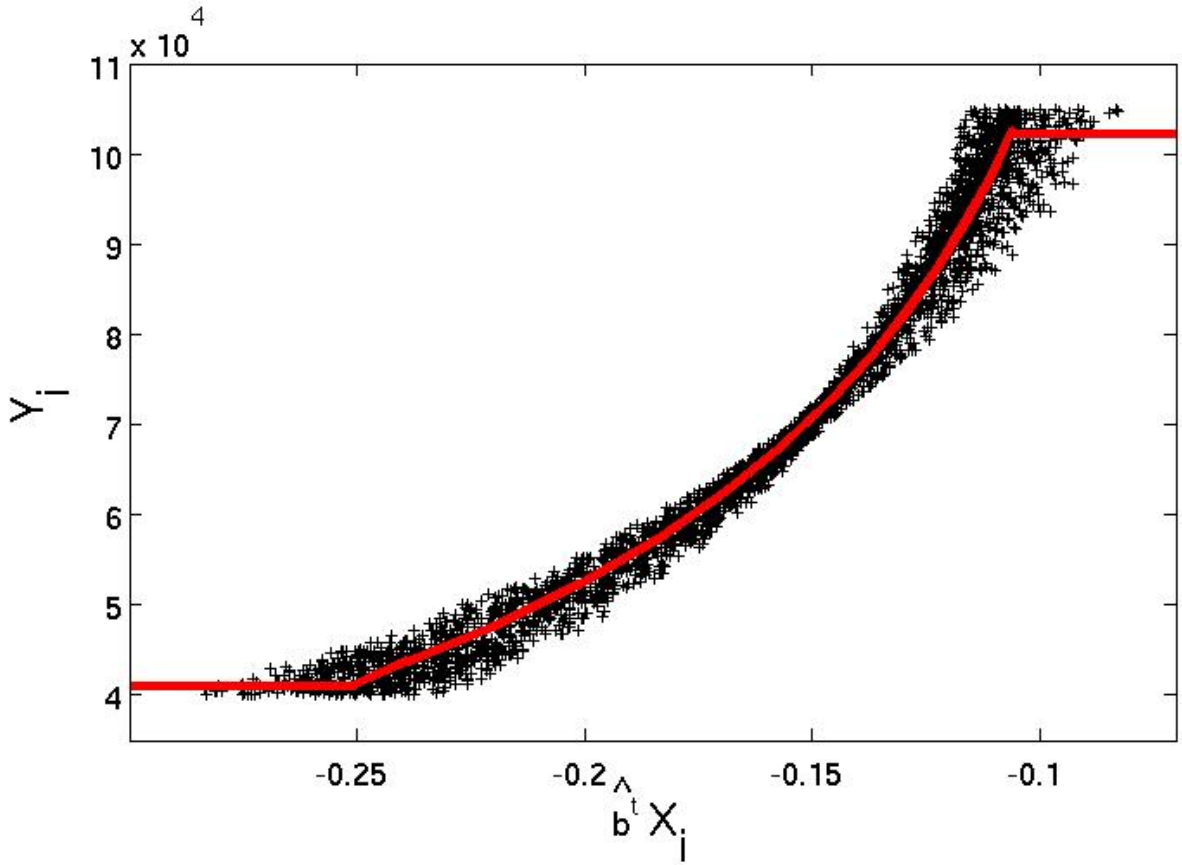
- Physical modeling of individual spectra with a surface reflectance model.
- Starting from a physical parameter Y, simulate $X = F(Y)$.
- Generation of $n = 12,000$ synthetic spectra with the corresponding parameters.

$\Longrightarrow$ Learning database.

Inverse problem.

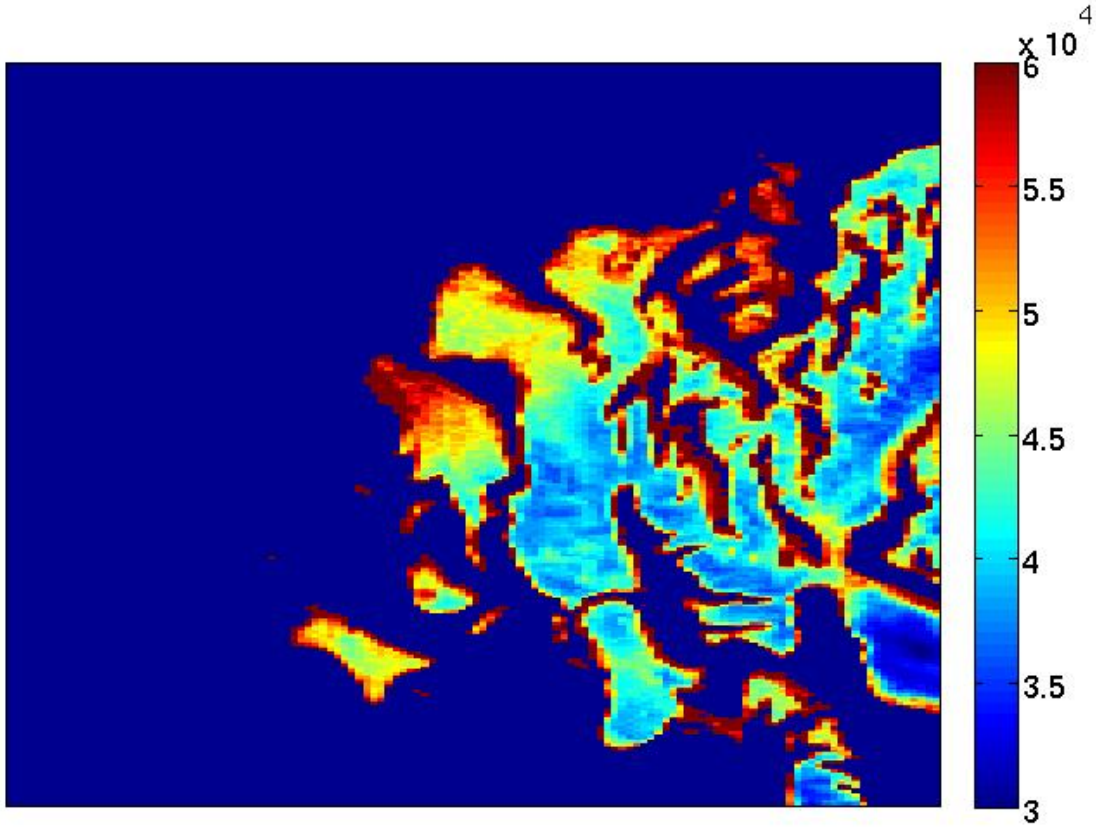
- Estimate the functional relationship $Y = G(X)$.
- Dimension reduction assumption $G(X) = g(< b, X >)$.
- b is estimated by SIR, g is estimated by a nonparametric one-dimensional regression.

Estimated function g



Estimated function g between the projected spectra $< \hat{b}, X >$ on the first axis of SIR and Y, the grain size of CO_2.

Estimated CO_2 maps



Grain size of CO_2 estimated with SIR on a hyperspectral image of Mars.

References

- [Li, 1991] Li, K.C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316–327.
- Bernard-Michel, C., Douté, S., Fauvel, M., Gardes, L. & Girard, S. (2009). Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression. *Journal of Geophysical Research - Planets*, **114**, E06005.
- M. Chavent, S. Girard, V. Kuentz, B. Liquet, T.M.N. Nguyen & J. Saracco. A sliced inverse regression approach for data stream, *Computational Statistics*, **29**, 1129–1152.