An introduction to SIR: A statistical method for dimension reduction in multivariate regression

Stéphane Girard

An introduction to SIR: A statistical method for dimension reduction in multivariate regression

Stéphane Girard

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

1 Sliced Inverse Regression (SIR)

1.1 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

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

- 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.

1.2 Dimension reduction

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

  $$Y = g(b^t X) + \xi$$

  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.*)

1.3 Reminder

Let $X_1, \dots X_n$ be n points in $\mathbb{R}^p$ divided into h classes C_j, $j = 1, \dots, h$.

- **Empirical covariance matrix**

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^t, \text{ where } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- **Within-class covariance matrix** "mean of covariances"

$$\hat{W} = \sum_{j=1}^{h} \frac{n_j}{n} \hat{\Sigma}_j,$$

  where $\hat{\Sigma}_j$ is the empirical covariance matrix of class j and n_j =card(C_j).

- **Between-class covariance matrix** "covariance of means"

$$\hat{B} = \sum_{i=1}^{n} \frac{n_j}{n}(\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^t, \text{ where } \bar{X}_j = \frac{1}{n_j}\sum_{X_i \in C_j} X_i.$$

- $\hat{\Sigma} = \hat{B} + \hat{W}$
- Let $b^t X$ the projection of the random vector on the axis b. Then, $\text{var}(b^t X) = b^t \text{cov}(X) b$.

1.4 SIR

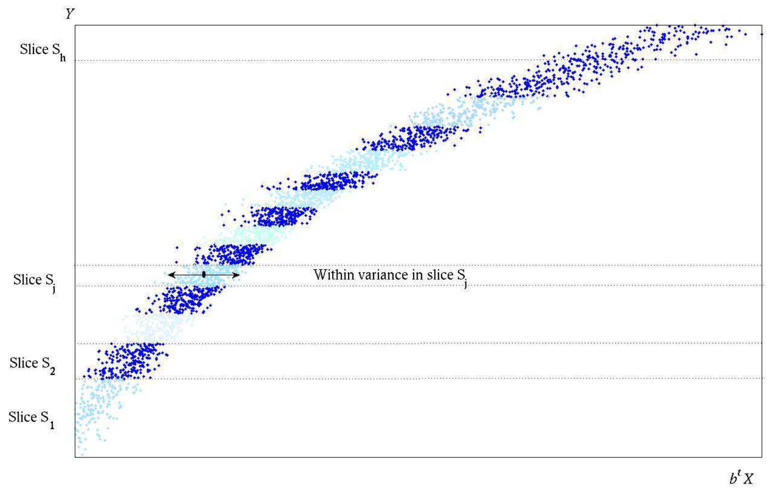
Idea:

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

In practice:

- The support of Y is divided into h slices S_j.
- Minimization of the within-slice variance of $b^t X$ under the constraint $\text{var}(b^t X) = 1$.
- Equivalent to maximizing the between-slice variance under the same constraint.

1.5 Illustration



1.6 Estimation procedure

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

$$\hat{b} = \underset{b}{\operatorname{argmax}}\, b^t \hat{\Gamma} b \;\text{ such that }\; b^t \hat{\Sigma} b = 1. \qquad (1)$$

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

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

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.

1.7 Illustration

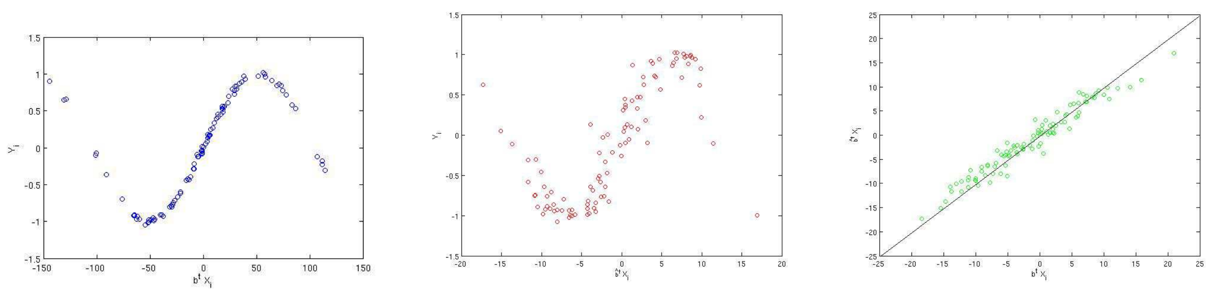
Simulated data.

- Sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of size $n = 100$ with $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$, $i = 1, \ldots, n$.
- $X_i \sim \mathcal{N}_p(0, \Sigma)$ where $\Sigma = Q \Delta Q^t$ with
  - Δ =diag$(p^\theta, \ldots, 2^\theta, 1^\theta)$,
  - θ controls the decreasing rate of the eigenvalue screeplot,

- Q is an orientation matrix drawn from the uniform distribution on the set of orthogonal matrices.

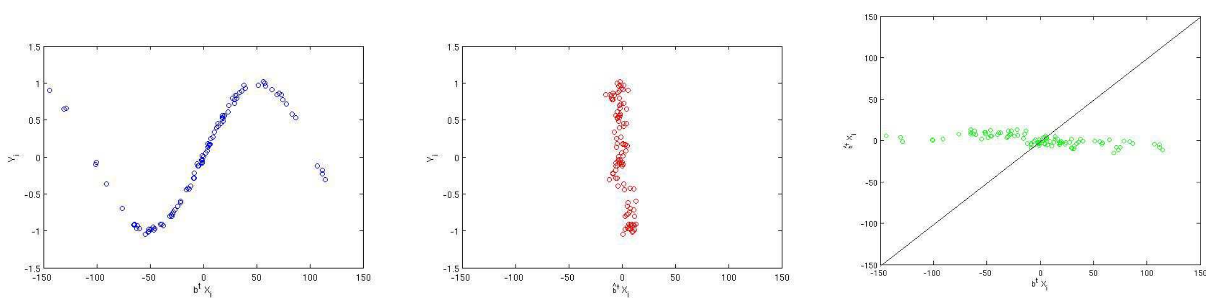
- $Y_i = g(b^t X_i) + \xi$ where
  - g is the link function $g(t) = \sin(\pi t/2)$,
  - b is the true direction $b = 5^{-1/2} Q(1, 1, 1, 1, 1, 0, \ldots, 0)^t$,
  - $\xi \sim \mathcal{N}_1(0, 9.10^{-4})$

1.8 Results with $\theta = 2$, dimension $p = 10$



Blue: Y_i versus the projections $b^t X_i$ on the true direction b,
Red: Y_i versus the projections $\hat{b}^t X_i$ on the estimated direction $\hat{b}$,
Green: $\hat{b}^t X_i$ versus $b^t X_i$.

1.9 Results with $\theta = 2$, dimension $p = 50$



Blue: Y_i versus the projections $b^t X_i$ on the true direction b,
Red: Y_i versus the projections $\hat{b}^t X_i$ on the estimated direction $\hat{b}$,
Green: $\hat{b}^t X_i$ versus $b^t X_i$.

1.10 Explanation

Problem : $\hat{\Sigma}$ may be singular or at least ill-conditioned in several situations.

- Since $\text{rank}(\hat{\Sigma}) \leq \min(n-1, p)$, if $n \leq p$ then $\hat{\Sigma}$ is singular.

- Even if n and p are of the same order, $\hat{\Sigma}$ is ill-conditioned, and its inversion yields numerical problems in the estimation of the central subspace.
- The same phenomenon occurs if the coordinates of X are strongly correlated.

In the previous example, the condition number of Σ was p^θ.

2 Regularization of SIR

2.1 Regularized SIR

- We propose to compute $\hat{b}$ as the eigenvector associated to the largest eigenvalue of $(\Omega\hat{\Sigma} + I_p)^{-1}\Omega\hat{\Gamma}$.
- Ω describes which directions in $\mathbb{R}^p$ are more likely to contain b.

$\Longrightarrow$ The inversion of $\hat{\Sigma}$ is replaced by the inversion of $\Omega\hat{\Sigma} + I_p$.
$\Longrightarrow$ For a well-chosen *a priori* matrix Ω, numerical problems disappear.

2.2 Links with existing methods

- Ridge [Zhong et al, 2005]: $\Omega = \tau^{-1} I_p$. No privileged direction for b in $\mathbb{R}^p$. $\tau > 0$ is a regularization parameter.
- PCA+SIR [Chiaromonte et al, 2002]:

$$\Omega = \sum_{j=1}^{d} \frac{1}{\hat{\delta}_j} \hat{q}_j \hat{q}_j^t,$$

  where $d \in \{1, \dots, p\}$ is fixed, $\hat{\delta}_1 \geq \dots \geq \hat{\delta}_d$ are the d largest eigenvalues of $\hat{\Sigma}$ and $\hat{q}_1, \dots, \hat{q}_d$ are the associated eigenvectors.

2.3 Three new methods

- PCA+ridge:

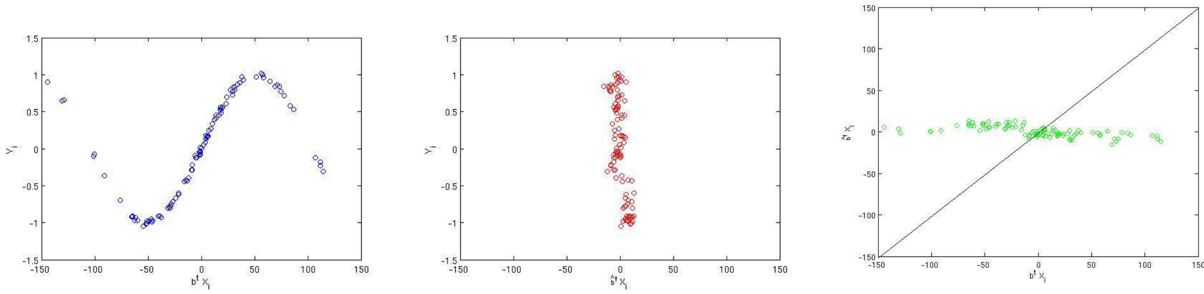
$$\Omega = \frac{1}{\tau} \sum_{j=1}^{d} \hat{q}_j \hat{q}_j^t.$$

  In the eigenspace of dimension d, all the directions are *a priori* equivalent.
- Tikhonov: $\Omega = \tau^{-1}\hat{\Sigma}$. The directions with large variance are the most likely to contain b.
- PCA+Tikhonov:

$$\Omega = \frac{1}{\tau} \sum_{j=1}^{d} \hat{\delta}_j \hat{q}_j \hat{q}_j^t.$$

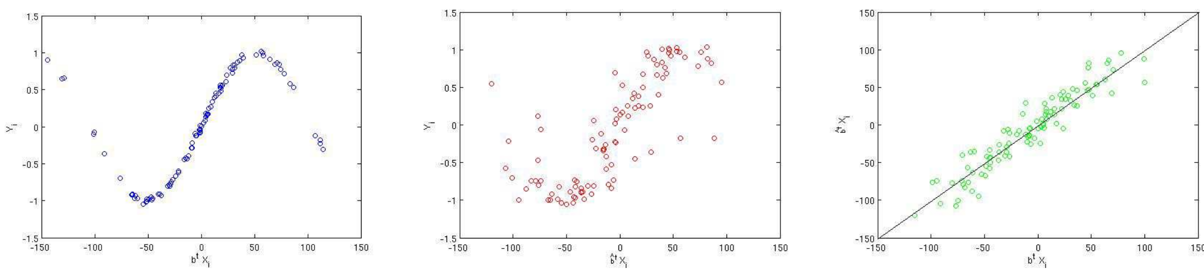
  In the eigenspace of dimension d, the directions with large variance are the most likely to contain b.

2.4 Recall of SIR results with $\theta = 2$ and $p = 50$



Blue: Projections $b^t X_i$ on the true direction b versus Y_i,
Red: Projections $\hat{b}^t X_i$ on the estimated direction $\hat{b}$ versus Y_i,
Green: $b^t X_i$ versus $\hat{b}^t X_i$.

2.5 Regularized SIR results (PCA+Ridge)



Blue: Projections $b^t X_i$ on the true direction b versus Y_i,
Red: Projections $\hat{b}^t X_i$ on the estimated direction $\hat{b}$ versus Y_i,
Green: $b^t X_i$ versus $\hat{b}^t X_i$.

2.6 Validation on simulations

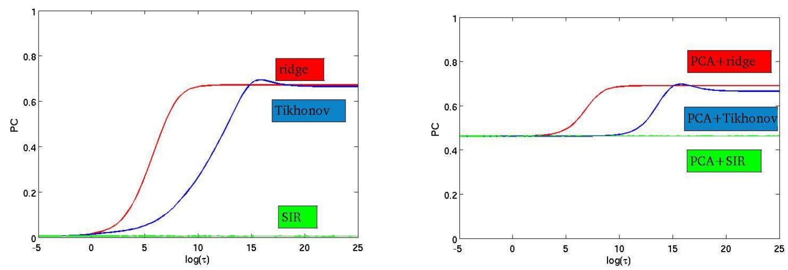
Proximity criterion between the true direction b and the estimated ones $\hat{b}^{(r)}$ on $N = 100$ replications:

$$\text{PC} = \frac{1}{N} \sum_{r=1}^{N} \cos^2(b, \hat{b}^{(r)})$$

- $0 \leq \text{PC} \leq 1$,
- a value close to 0 implies a low proximity: The $\hat{b}^{(r)}$ are nearly orthogonal to b,
- a value close to 1 implies a high proximity: The $\hat{b}^{(r)}$ are approximately collinear with b.

2.7 Influence of the regularization parameter

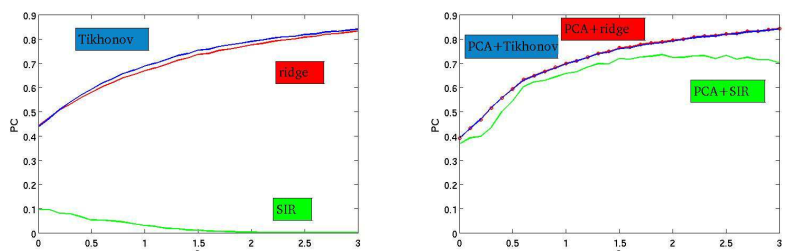
$\log \tau$ versus PC. The "cut-off" dimension and the condition number are fixed ($d = 20$ and $\theta = 2$).



- Ridge and Tikhonov: significant improvement if τ is large,
- PCA+SIR: reasonable results compared to SIR,
- PCA+ridge and PCA+Tikhonov: small sensitivity to τ.

2.8 Sensitivity with respect to the condition number of the covariance matrix

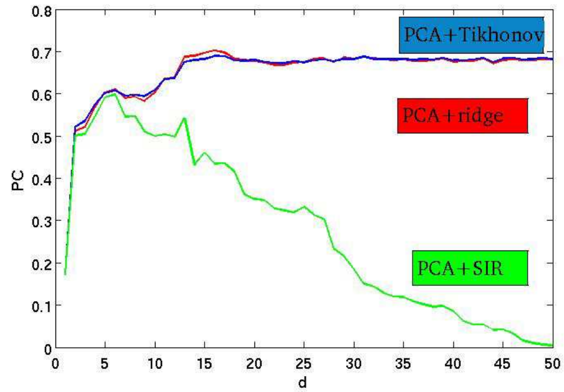
θ versus PC. The "cut-off" dimension is fixed to $d = 20$. The optimal regularization parameter is used for each value of θ.



- Only SIR is very sensitive to the ill-conditioning,
- ridge and Tikhonov: similar results,
- PCA+ridge and PCA+Tikhonov: similar results.

2.9 Sensitivity with respect to the "cut-off" dimension

d versus PC. The condition number is fixed ($\theta = 2$) The optimal regularization parameter is used for each value of d.



- PCA+SIR: very sensitive to d.
- PCA+ridge and PCA+Tikhonov: stable as d increases.

3 Application to real data

3.1 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).

3.2 An inverse problem

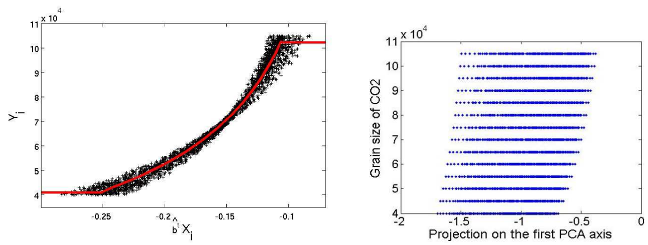
Forward problem.

- 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.

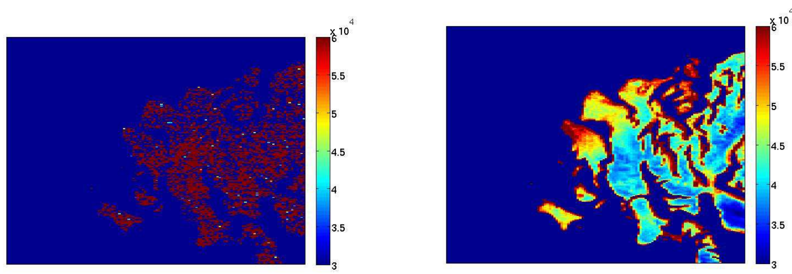
- Estimate the functional relationship $Y = G(X)$.
- Dimension reduction assumption $G(X) = g(b^t X)$.
- b is estimated by (regularized) SIR, g is estimated by a nonparametric one-dimensional regression.

3.3 Estimated function g



Estimated function g between the projected spectra $\hat{b}^t X$ on the first axis of regularized SIR (PCA+ridge) and Y, the grain size of CO_2.

3.4 Estimated CO_2 maps



Grain size of CO_2 estimated with SIR (left) and regularized SIR (right) on a hyperspectral image of Mars.

3.5 Extensions

- **Kernel SIR**. The usual dot product $b^t X$ is replaced by a kernel.
  Wu, H. M. (2008). Kernel Sliced Inverse Regression with Applications to Classification, *Journal of Computational and Graphical Statistics*, **17**(3), 590–610.

  `http://www.hmwu.idv.tw/KSIR/`

- **Sparse SIR**. Introduction of a L_1 penalty on b to obtain sparse axes.
  Li, L. and Nachtsheim, C. (2006). Sparse Sliced Inverse Regression, *Technometrics*, **48**(4), 503–510.

3.6 References on this work

- Bernard-Michel, C., Gardes, L. and Girard, S. (2008). A Note on Sliced Inverse Regression with Regularizations. *Biometrics*, **64**, 982–986.

- Bernard-Michel, C., Douté, S., Fauvel, M., Gardes, L. and 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

- Bernard-Michel, C., Gardes, L. and Girard, S. (2009). Gaussian Regularized Sliced Inverse Regression, *Statistics and Computing*, **19**, 85–98.

- Coudret, R., Girard, S. and Saracco, J. (2014). A new sliced inverse regression method for multivariate response. *Computational Statistics and Data Analysis*, **77**, 285–299.

- Chavent, M., Girard, S., Kuentz, V., Liquet, B., Nguyen, T.M.N. and Saracco, J. (2014). A sliced inverse regression approach for data stream. *Computational Statistics*, to appear.

3.7 References on SIR

- [Li, 1991] Li, K.C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316–327.

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

- [Zhong et al, 2005]: 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.

- [Chiaromonte et al, 2002]: Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences*, **176**, 123–144.