Regularization methods for Sliced Inverse Regression

Stéphane Girard

Team Mistis, INRIA Rhône-Alpes, France
http ://mistis.inrialpes.fr/~girard

January 2010

Joint work with Caroline Bernard-Michel, Sylvain Douté, Mathieu Fauvel and Laurent Gardes

Outline

Outline

Multivariate regression

- Given two r.v. $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$, estimate $G : \mathbb{R}^p \to \mathbb{R}$ such that

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

- When p is large, curse of dimensionality.
- **Sufficient dimension reduction** aims at replacing X by its projection onto a subspace of smaller dimension without loss of information on the distribution of Y given X.
- The **central subspace** is the smallest subspace S such that, conditionally on the projection of X on S, Y and X are independent.

Dimension reduction principle

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

  $$Y = g(b^t X) + \xi \quad \text{where } \xi \text{ is independent of } X.$$

- The estimation of a $p-$ variate function G is replaced by the estimation of a univariate function g and of an axis b.
- **Goal of SIR** [Li, 1991] : to estimate a basis of the central subspace (*i.e.* b *in this case*).

SIR : Basic principle

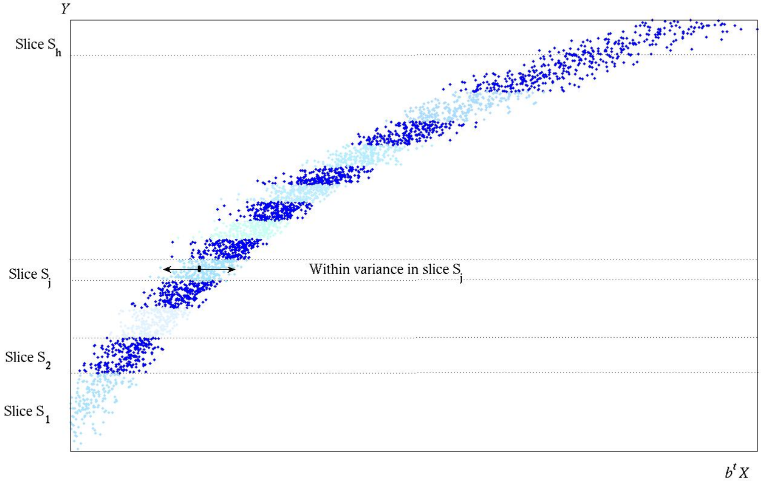
Idea :

- 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 range of Y is partitioned into h slices S_j.
- Minimize the within slice variance of $b^t X$ under the normalization constraint var$(b^t X) = 1$.
- Equivalent to maximizing the between slice variance under the same constraint.

SIR : Illustration



SIR : Estimation procedure

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

$$\hat{b} = \underset{b}{\operatorname{argmax}}\, b^t \hat{\Gamma} b \ \text{ u.c. } \ b^t \hat{\Sigma} b = 1. \tag{1}$$

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

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

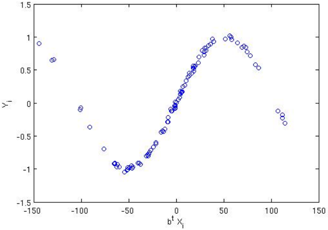
with n_j is proportion of observations in slice S_j. The optimization problem (1) has an explicit solution : $\hat{b}$ is the eigenvector of $\hat{\Sigma}^{-1}\hat{\Gamma}$ associated to its largest eigenvalue.

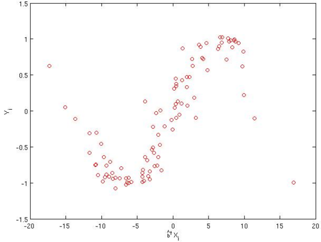
SIR : Illustration

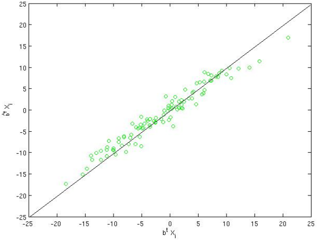
Experimental set-up.

- A sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of size $n = 100$ where $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$, for $i = 1, \ldots, n$.
- $X_i \sim \mathcal{N}_p(0, \Sigma)$ with $\Sigma = Q\Delta Q^t$ where
  - $\Delta =$diag$(p^\theta, \ldots, 2^\theta, 1^\theta)$,
  - θ tunes the eigenvalue scree,
  - Q is a 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 $b = 5^{-1/2} Q(1, 1, 1, 1, 1, 0, \ldots, 0)^t$,
  - $\xi \sim \mathcal{N}_1(0, 9.10^{-4})$

Results with $\theta = 2$ and $p = 10$

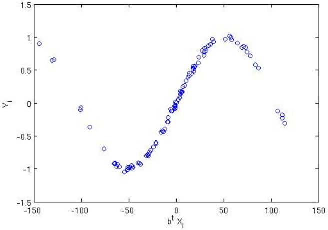


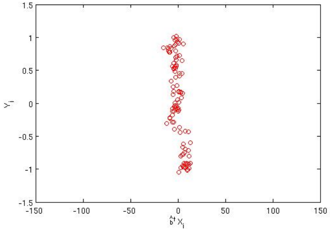


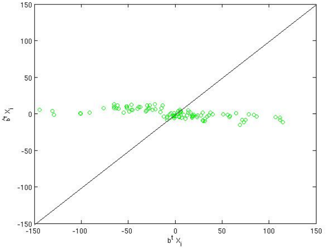


Blue : Projections $b^t X_i$ on the true direction b versus Y_i,
Red : Projections $\hat{b}^t X_i$ on the estimated direction $\hat{b}$ versus Y_i,
Green : $b^t X_i$ versus $\hat{b}^t X_i$.

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 $\hat{b}^t X_i$.

SIR : Limitations

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

- Since rank$(\hat{\Sigma}) \leq \min(n-1, p)$, if $n \leq p$ then $\hat{\Sigma}$ is singular.
- Even when n and p are of the same order, $\hat{\Sigma}$ is ill-conditioned, and its inversion introduces numerical instabilities in the estimation of the central subspace.
- Similar phenomena occur when the coordinates of X are highly correlated.

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

Outline

Single-index inverse regression model

Model introduced in [Cook, 2007].

$$X = \mu + c(Y)Vb + \varepsilon, \qquad (2)$$

where

- μ and b are non-random $\mathbb{R}^p-$ vectors,
- $\varepsilon \sim \mathcal{N}_p(0, V)$, independent of Y,
- $c : \mathbb{R} \to \mathbb{R}$ is a nonrandom coordinate function.

Consequence : The conditional expectation of $X - \mu$ given Y is a degenerated random vector located in the direction Vb.

Maximum Likelihood estimation (1/3)

- **Projection estimator of the coordinate function.** $c(.)$ is expanded as a linear combination of h basis functions $s_j(.)$,

$$c(.) = \sum_{j=1}^{h} c_j s_j(.) = s^t(.)c,$$

where $c = (c_1, \ldots, c_h)^t$ is unknown and $s(.) = (s_1(.), \ldots, s_h(.))^t$. Model (2) can be rewritten as

$$X = \mu + s^t(Y)cVb + \varepsilon, \;\; \varepsilon \sim \mathcal{N}_p(0, V),$$

- Definition : **Signal to Noise Ratio in the direction** b.

$$\rho = \frac{b^t \Sigma b - b^t V b}{b^t V b},$$

where $\Sigma =$cov(X).

Maximum Likelihood estimation (2/3)

Notations

- W : the $h \times h$ empirical covariance matrix of $s(Y)$ defined by

$$W = \frac{1}{n}\sum_{i=1}^{n}(s(Y_i) - \bar{s})(s(Y_i) - \bar{s})^t \ \text{ with } \ \bar{s} = \frac{1}{n}\sum_{i=1}^{n} s(Y_i).$$

- M : the $h \times p$ matrix defined by

$$M = \frac{1}{n}\sum_{i=1}^{n}(s(Y_i) - \bar{s})(X_i - \bar{X})^t,$$

Maximum Likelihood estimation (3/3)

If W and $\hat{\Sigma}$ are regular, then the ML estimators are :

- **Direction** : $\hat{b}$ is the eigenvector associated to the largest eigenvalue $\hat{\lambda}$ of $\hat{\Sigma}^{-1}M^tW^{-1}M$,
- **Coordinate** : $\hat{c} = W^{-1}M\hat{b}/\hat{b}^t\hat{V}\hat{b}$,
- **Location parameter** : $\hat{\mu} = \bar{X} - \bar{s}^t\hat{c}\hat{V}\hat{b}$,
- **Covariance matrix** : $\hat{V} = \hat{\Sigma} - \hat{\lambda}\hat{\Sigma}\hat{b}\hat{b}^t\hat{\Sigma}/\hat{b}^t\hat{\Sigma}\hat{b}$,
- **Signal to Noise Ratio** : $\hat{\rho} = \hat{\lambda}/(1 - \hat{\lambda})$.

The inversion of $\hat{\Sigma}$ is still necessary.

SIR : A particular case

In the particular case of **piecewise constant basis functions**

$$s_j(.) = \mathbb{I}\{. \in S_j\}, \;\; j = 1, \dots, h,$$

standard calculations show that

$$M^t W^{-1} M = \hat{\Gamma}$$

and thus the ML estimator $\hat{b}$ of b is the eigenvector associated to the largest eigenvalue of $\hat{\Sigma}^{-1}\hat{\Gamma}$.

$\Longrightarrow$ SIR method.

Outline

Gaussian prior

Introduction of a prior information on the projection of X on b appearing in the inverse regression model

$$(1+\rho)^{-1/2}\,(s(Y)-\bar{s})^t cb \sim \mathcal{N}(0,\Omega).$$

- $(1+\rho)^{-1/2}$ is introduced for normalization purposes, permitting to preserve the interpretation of the eigenvalue in terms of signal to noise ratio.
- Ω describes which directions in $\mathbb{R}^p$ are the most likely to contain b.

Gaussian regularized estimators

If W and $\Omega\hat{\Sigma} + I_p$ are regular, the ML estimators are

- **Direction :** $\hat{b}$ is the eigenvector associated to the largest eigenvalue $\hat{\lambda}$ of $(\Omega\hat{\Sigma} + I_p)^{-1}\Omega M^t W^{-1} M$,
- **Coordinate :** $\hat{c} = W^{-1}M\hat{b}/((1+\eta(\hat{b}))\hat{b}^t\hat{V}\hat{b})$, with $\eta(\hat{b}) = \hat{b}^t\Omega^{-1}\hat{b}/\hat{b}^t\hat{\Sigma}\hat{b}$,
- $\hat{\mu}$, $\hat{V}$ and $\hat{\rho}$ are unchanged.

$\Longrightarrow$The inversion of $\hat{\Sigma}$ is replaced by the inversion of $\Omega\hat{\Sigma} + I_p$.
$\Longrightarrow$ For a properly chosen prior matrix Ω, the numerical instabilities in the estimation of b disappear.

Gaussian regularized SIR (1/2)

GRSIR : In the particular case of piecewise constant basis functions, the ML estimator $\hat{b}$ of b is the eigenvector associated to the largest eigenvalue of $(\Omega\hat{\Sigma} + I_p)^{-1}\Omega\hat{\Gamma}$.

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

Gaussian regularized SIR (2/2)

Three new methods

- PCA+ridge :

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

  No privileged direction in the d-dimensional eigenspace.

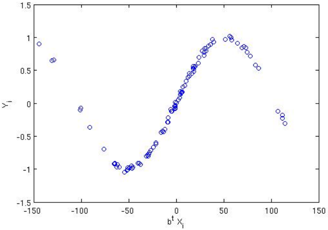
- Tikhonov : $\Omega = \tau^{-1} \hat{\Sigma}$. Directions with large variance are most likely.
- PCA+Tikhonov :

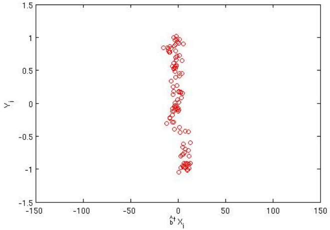
$$\Omega = \frac{1}{\tau} \sum_{j=1}^{d} \hat{\delta}_j \hat{q}_j \hat{q}_j^t.$$

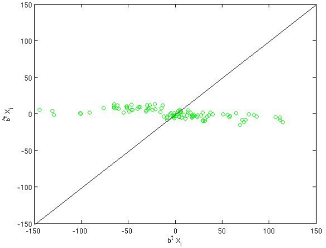
  In the d-dimensional eigenspace, directions with large variance are most likely.

Outline

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

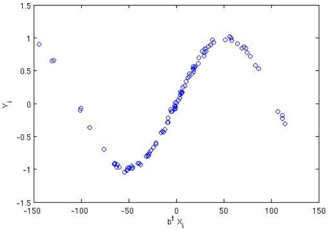


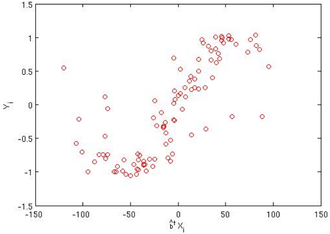


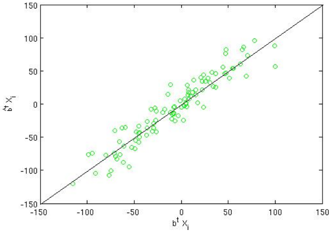


Blue : Projections $b^t X_i$ on the true direction b versus Y_i,
Red : Projections $\hat{b}^t X_i$ on the estimated direction $\hat{b}$ versus Y_i,
Green : $b^t X_i$ versus $\hat{b}^t X_i$.

GRSIR 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 $\hat{b}^t X_i$.

Validation on simulations

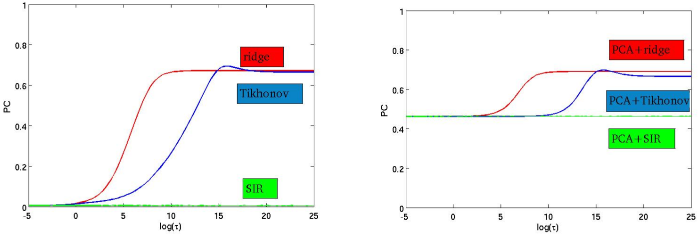
Proximity criterion between the true direction b and the estimated ones $\hat{b}^{(r)}$ on $N = 100$ replications :

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

- $0 \leq \text{PC} \leq 1$,
- 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.

Influence of the regularization parameter

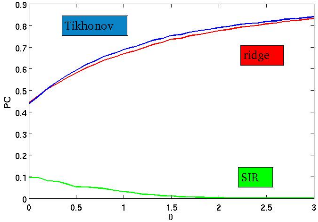
$\log \tau$ versus PC. The "cut-off" dimension and the condition number are fixed ($d = 20$ and $\theta = 2$).

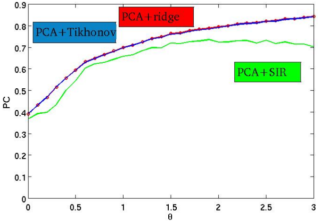


- Ridge and Tikhonov : significant improvement if τ is large,
- PCA+SIR : reasonable results compared to SIR,
- PCA+ridge and PCA+Tikhonov : small sensitivity to τ.

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.

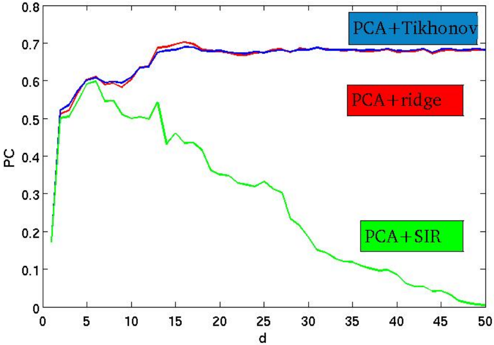




- Only SIR is very sensitive to the ill-conditioning,
- ridge and Tikhonov : similar results,
- PCA+ridge and PCA+Tikhonov : similar results.

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.

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.

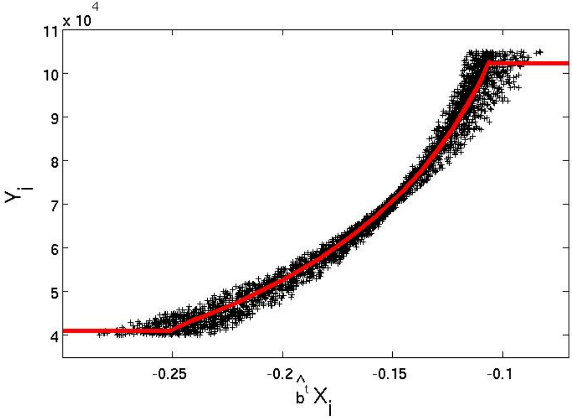
- Physical modeling of individual spectra with a surface reflectance model.
- Starting from a physical parameter Y, simulate $X = F(Y)$.
- Generation of $n = 12,000$ synthetic spectra with the corresponding parameters.

$\Longrightarrow$ Learning database.

Inverse problem.

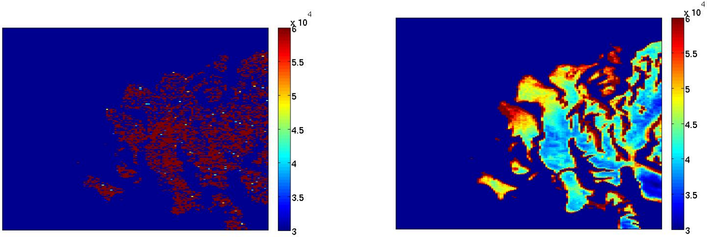
- Estimate the fonctional relationship $Y = G(X)$.
- Dimension reduction assumption $G(X) = g(b^t X)$.
- b is estimated by SIR/GRSIR, g is estimated by a nonparametric one-dimensional regression.

Estimated functional relationship



Functional relationship between reduced spectra $\hat{b}^t X$ on the first GRSIR (PCA+ridge prior) direction and Y, the grain size of CO_2.

Estimated CO_2 maps



Grain size of CO_2 estimated by SIR (left) and GRSIR (right) on an hyperspectral image observed on Mars during orbit 61.

SIR references

- [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.

References on this work

- Bernard-Michel, C., Douté, S., Fauvel, M., Gardes, L. et 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. `http ://hal.inria.fr/inria-00276116/fr`
- Bernard-Michel, C., Gardes, L. et Girard, S. (2009). Gaussian Regularized Sliced Inverse Regression, *Statistics and Computing*, **19**, 85–98. `http ://hal.inria.fr/inria-00180458/fr`
- Bernard-Michel, C., Gardes, L. et Girard, S. (2008). A Note on Sliced Inverse Regression with Regularizations, *Biometrics*, **64**, 982–986. `http ://hal.inria.fr/inria-00180496/fr`