Régularisation en régression inverse par tranches

Caroline Bernard-Michel, Laurent Gardes, et Stéphane Girard

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

Novembre 2008

Plan

Plan

Régression multivariée

Soient $Y \in \mathbb{R}$ et $X \in \mathbb{R}^p$. On souhaite estimer $G : \mathbb{R}^p \to \mathbb{R}$ telle que

$$Y = G(X) + \xi \quad \text{avec } \xi \text{ indépendant de } X.$$

- Irréaliste quand p est grand, fléau de la dimension (*curse of dimensionality*).
- **Réduction de dimension** : remplacer X par sa projection sur un sous-espace de dimension inférieure sans perte d'information sur la loi de Y sachant X.
- **Sous-espace central** : plus petit sous-espace S tel que, conditionnellement à la projection de X sur S, Y et X sont indépendants.

Principe de la réduction de dimension

- On suppose $\dim(S) = 1$ pour simplifier *i.e.* $S =$span(b), avec $b \in \mathbb{R}^p \Longrightarrow$ **Single index model :**

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

où ξ est indépendant of X.

- L'estimation d'une fonction G de p variables est remplacée par l'estimation d'une fonction g d'une seule variable et d'un axe b.
- **But de SIR** [Li, 1991] : estimer une base du sous-espace central. (*i.e.* b *dans ce cas précis.*)

Principe de SIR

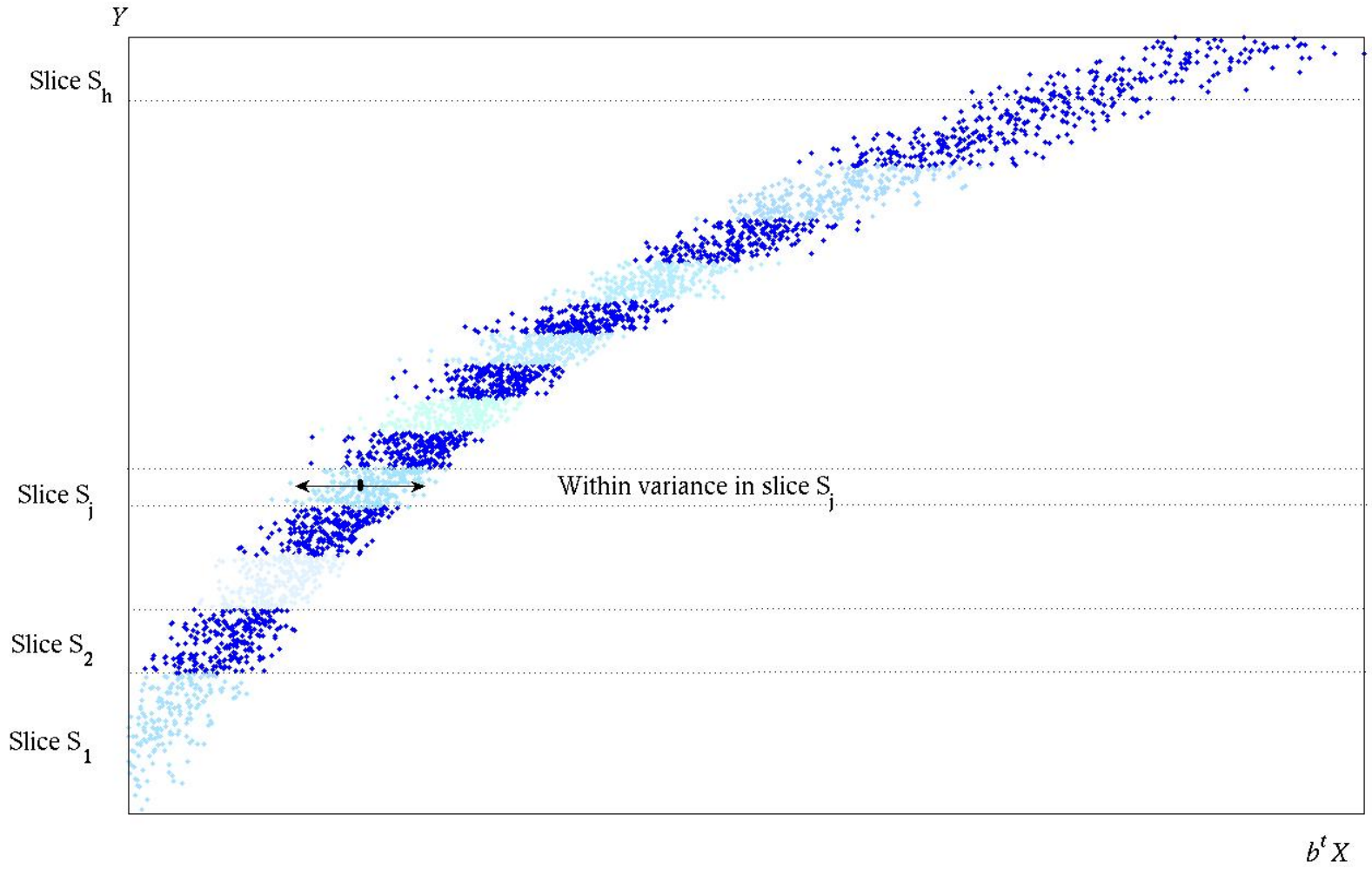
Idée :

- Trouver l'axe b tel que $b^t X$ explique au mieux Y.
- Inversement, à Y fixé, $b^t X$ doit peu varier.
- Trouver l'axe b minimisant les variations de $b^t X$ sachant Y.

En pratique :

- Le support de Y est découpé en h tranches S_j (*slices*).
- Minimisation de la variance intra-tranche $b^t X$ sous la contrainte $\text{var}(b^t X) = 1$.
- Equivalent à maximiser la variance inter-tranches sous la même contrainte.

Illustration



Procédure d'estimation

Etant donné un échantillon $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, l'axe b est estimé par

$$\hat{b} = \underset{b}{\operatorname{argmax}}\, b^t \hat{\Gamma} b \text{ sous la contrainte } b^t \hat{\Sigma} b = 1. \qquad (1)$$

où $\hat{\Sigma}$ est la matrice de covariance empirique et $\hat{\Gamma}$ est la matrice de covariance inter-tranches définie par

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

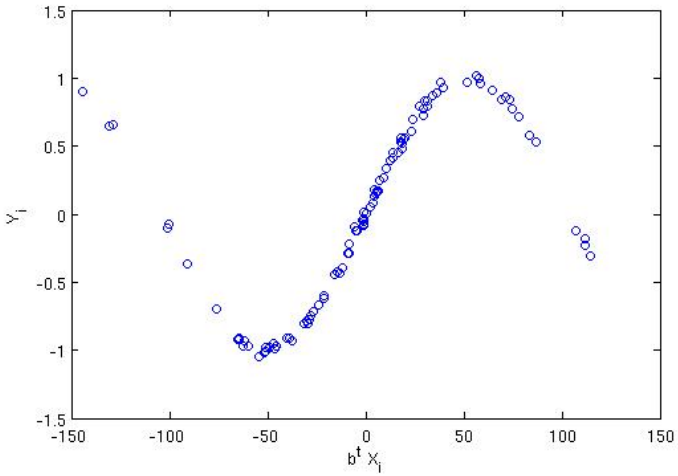
avec n_j la proportion d'observations dans la tranche S_j.
Le problème d'optimisation (1) a une solution explicite : $\hat{b}$ est le vecteur propre de $\hat{\Sigma}^{-1}\hat{\Gamma}$ associé à la plus grande valeur propre.

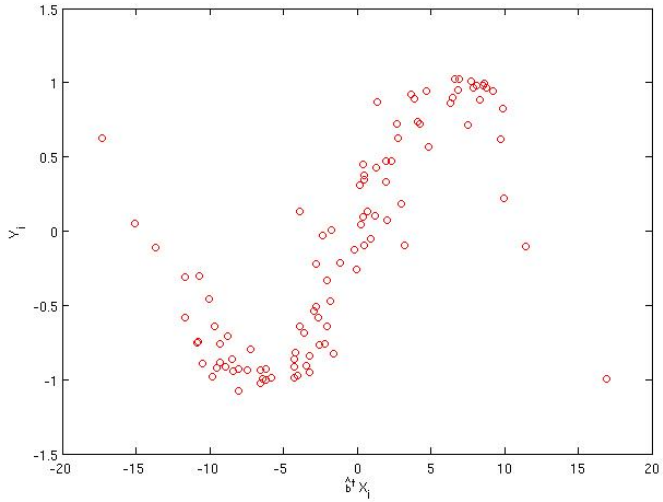
Illustration

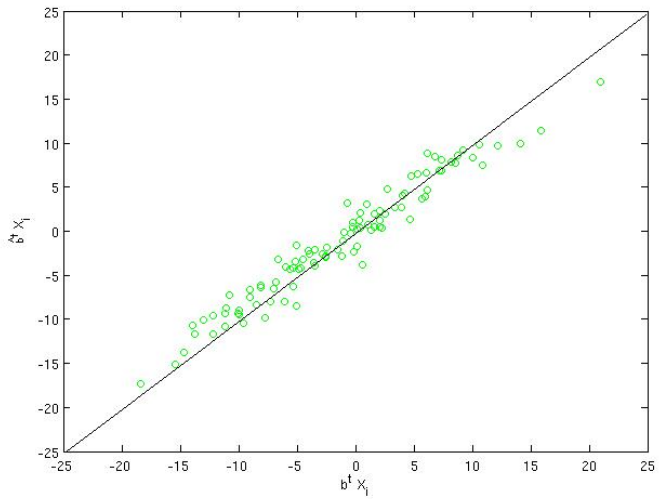
Données simulées.

- Echantillon $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ de taille $n = 100$ avec $X_i \in \mathbb{R}^p$ et $Y_i \in \mathbb{R}$, $i = 1, \ldots, n$.
- $X_i \sim \mathcal{N}_p(0, \Sigma)$ avec $\Sigma = Q\Delta Q^t$ où
  - $\Delta = \text{diag}(p^\theta, \ldots, 2^\theta, 1^\theta)$,
  - θ contrôle la décroissance des valeurs propres,
  - Q est une matrice d'orientation tirée selon la loi uniforme sur l'ensemble des matrices orthogonales.
- $Y_i = g(b^t X_i) + \xi$ où
  - g est la fonction de lien $g(t) = \sin(\pi t/2)$,
  - b est le vrai axe $b = 5^{-1/2} Q(1, 1, 1, 1, 1, 0, \ldots, 0)^t$,
  - $\xi \sim \mathcal{N}_1(0, 9.10^{-4})$

Résultats avec $\theta = 2$, dimension $p = 10$

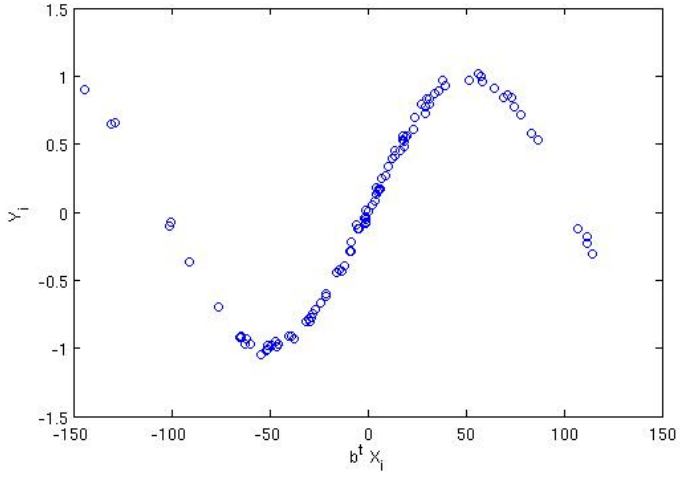


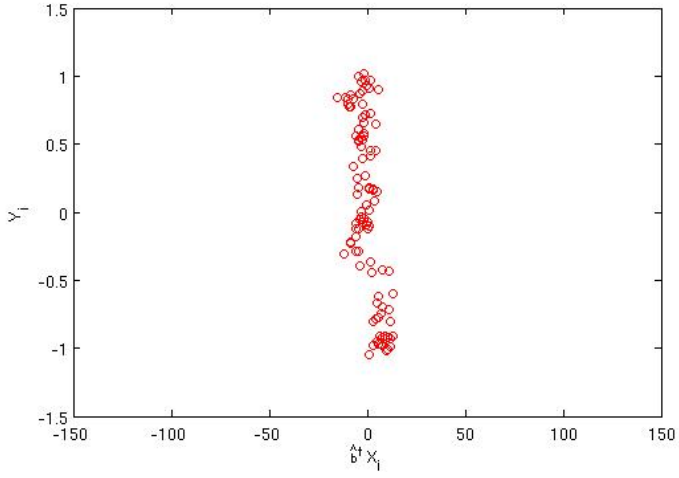


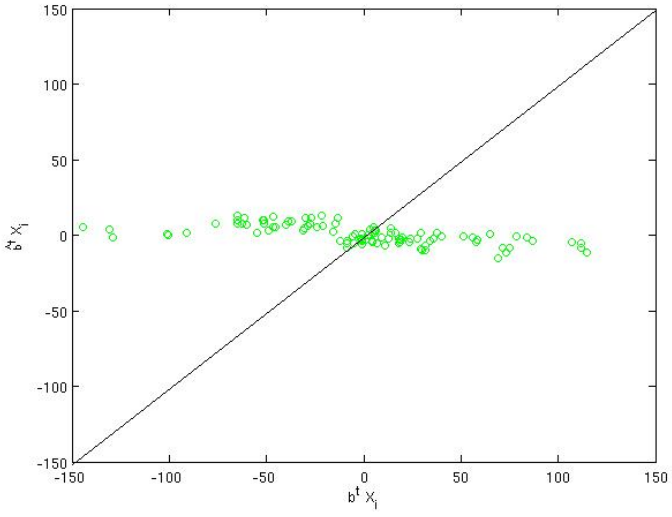


Bleu : Y_i en fonction des projections $b^t X_i$ sur le vrai axe b,
Rouge : Y_i en fonction des projections $\hat{b}^t X_i$ sur l'axe estimé $\hat{b}$,
Vert : $\hat{b}^t X_i$ en fonction de $b^t X_i$.

Résultats avec $\theta = 2$, dimension $p = 50$







Bleu : Y_i en fonction des projections $b^t X_i$ sur le vrai axe b,
Rouge : Y_i en fonction des projections $\hat{b}^t X_i$ sur l'axe estimé $\hat{b}$,
Vert : $\hat{b}^t X_i$ en fonction de $b^t X_i$.

Explication

Problème : $\hat{\Sigma}$ peut être singulière ou au moins mal-conditionnée dans plusieurs situations.

- Sachant que $\text{rank}(\hat{\Sigma}) \leq \min(n-1, p)$, si $n \leq p$ alors $\hat{\Sigma}$ est singulière.
- Même si n et p sont du même ordre, $\hat{\Sigma}$ est mal-conditionnée, et son inversion introduit des instabilités numériques dans l'estimation du sous-espace central.
- Le même phénomène se produit si les coordonnées de X sont très corrélées.

Dans l'exemple précédent, le conditionnement de Σ est p^{θ}.

Plan

Modèle de régression inverse

Modèle introduit par [Cook, 2007].

$$X = \mu + c(Y)Vb + \varepsilon, \tag{2}$$

avec

- μ et b des vecteurs de $\mathbb{R}^p$,
- $\varepsilon \sim \mathcal{N}_p(0, V)$, indépendant de Y,
- $c : \mathbb{R} \to \mathbb{R}$ une fonction coordonnée.

Conséquence : L'espérance conditionnelle de $X - \mu$ sachant Y est un vecteur aléatoire colinéaire à Vb.

Estimation par maximum de vraisemblance (1/3)

- **Estimation par projection de la fonction coordonnée.** $c(.)$ est décomposée comme une combinaison linéaire de h fonctions de base $s_j(.)$,

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

où $c = (c_1, \ldots, c_h)^t$ est inconnu et $s(.) = (s_1(.), \ldots, s_h(.))^t$. Le modèle (2) se réécrit alors

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

- Définition : **Rapport signal sur bruit dans la direction** b.

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

où $\Sigma =$cov(X).

Estimation par maximum de vraisemblance (2/3)

Notations

- W : la matrice de covariance empirique $h \times h$ de $s(Y)$ définie par

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

- M : la matrice $h \times p$ définie par

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

Estimation par maximum de vraisemblance (3/3)

Si W et $\hat{\Sigma}$ sont inversibles, alors les estimateurs de maximum de vraisemblance sont :

- **Axe** : $\hat{b}$ est le vecteur propre associé à la plus grande valeur propre $\hat{\lambda}$ de $\hat{\Sigma}^{-1} M^t W^{-1} M$,
- **Coordonnée** : $\hat{c} = W^{-1} M \hat{b} / \hat{b}^t \hat{V} \hat{b}$,
- **Paramètre de position** : $\hat{\mu} = \bar{X} - \bar{s}^t \hat{c} \hat{V} \hat{b}$,
- **Matrice de covariance** : $\hat{V} = \hat{\Sigma} - \hat{\lambda} \hat{\Sigma} \hat{b} \hat{b}^t \hat{\Sigma} / \hat{b}^t \hat{\Sigma} \hat{b}$,
- **Rapport signal sur bruit** : $\hat{\rho} = \hat{\lambda} / (1 - \hat{\lambda})$.

L'inversion de $\hat{\Sigma}$ est encore nécessaire.

SIR : un cas particulier

Dans le cas particulier de **fonctions de base constantes par morceaux**

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

on montre que

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

et l'estimateur de maximum de vraisemblance $\hat{b}$ de b est le vecteur propre associé à la plus grande valeur propre de $\hat{\Sigma}^{-1}\hat{\Gamma}$.

$\Longrightarrow$ Méthode SIR.

Plan

A priori gaussien

Introduction d'une information *a priori* sur la projection de X sur b apparaissant dans le modèle de régression inverse :

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

- $(1+\rho)^{-1/2}$ est un facteur de normalisation permettant de préserver l'interprétation de la valeur propre en termes de rapport signal sur bruit.
- Ω décrit quelles directions dans $\mathbb{R}^p$ sont les plus probables pour b.

Estimateurs régularisés

Si W et $\Omega\hat{\Sigma} + I_p$ sont inversibles, les estimateurs sont les suivants :

- **Axe :** $\hat{b}$ est le vecteur propre associé à la plus grande valeur propre $\hat{\lambda}$ de $(\Omega\hat{\Sigma} + I_p)^{-1}\Omega M^t W^{-1} M$,
- **Coordonnée :** $\hat{c} = W^{-1}M\hat{b}/((1 + \eta(\hat{b}))\hat{b}^t\hat{V}\hat{b})$, avec $\eta(\hat{b}) = \hat{b}^t\Omega^{-1}\hat{b}/\hat{b}^t\hat{\Sigma}\hat{b}$,
- $\hat{\mu}$, $\hat{V}$ et $\hat{\rho}$ sont inchangés.

$\Longrightarrow$L'inversion de $\hat{\Sigma}$ est remplacée par l'inversion de $\Omega\hat{\Sigma} + I_p$.
$\Longrightarrow$ Pour une matrice *a priori* Ω bien choisie, les problèmes d'instabilité disparaissent.

Gaussian regularized SIR (1/2)

GRSIR : Dans le cas particulier de fonctions de base constantes par tranches, l'estimateur régularisé $\hat{b}$ de b est le vecteur propre associé à la plus grande valeur propre de $(\Omega\hat{\Sigma} + I_p)^{-1}\Omega\hat{\Gamma}$.

Liens avec les méthodes existantes

- Ridge [Zhong et al, 2005] : $\Omega = \tau^{-1} I_p$. Pas de direction privilégiée pour b dans $\mathbb{R}^p$. $\tau > 0$ est un paramètre de régularisation.
- ACP+SIR [Chiaromonte et al, 2002] :

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

où $d \in \{1, \ldots, p\}$ est fixée, $\hat{\delta}_1 \geq \cdots \geq \hat{\delta}_d$ sont les d plus grandes valeurs propres de $\hat{\Sigma}$ et $\hat{q}_1, \ldots, \hat{q}_d$ sont les vecteurs propres associés.

Gaussian regularized SIR (2/2)

Trois nouvelles méthodes

- ACP+ridge :

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

  Dans l'espace propre de dimension d, toutes les directions sont *a priori* équivalentes.

- Tikhonov : $\Omega = \tau^{-1} \hat{\Sigma}$. Les directions de grande variance sont les plus probables.

- ACP+Tikhonov :

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

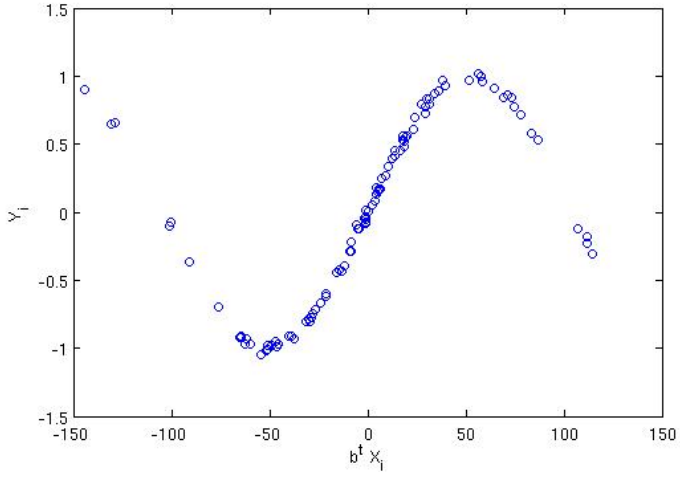
  Dans l'espace propre de dimension d, les directions de grande variance sont les plus probables.

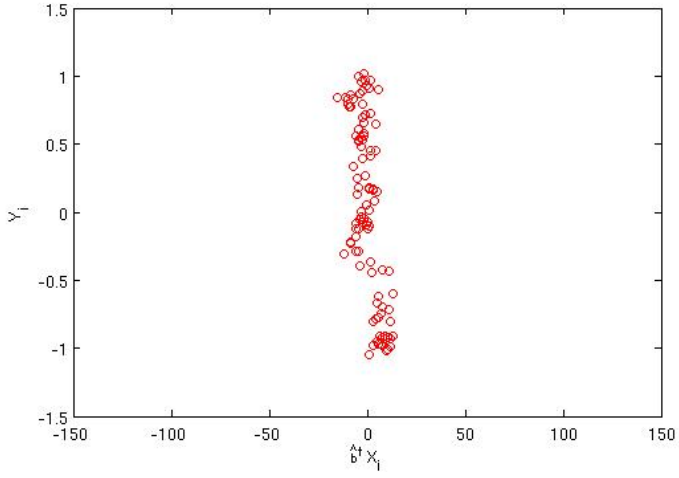
Plan

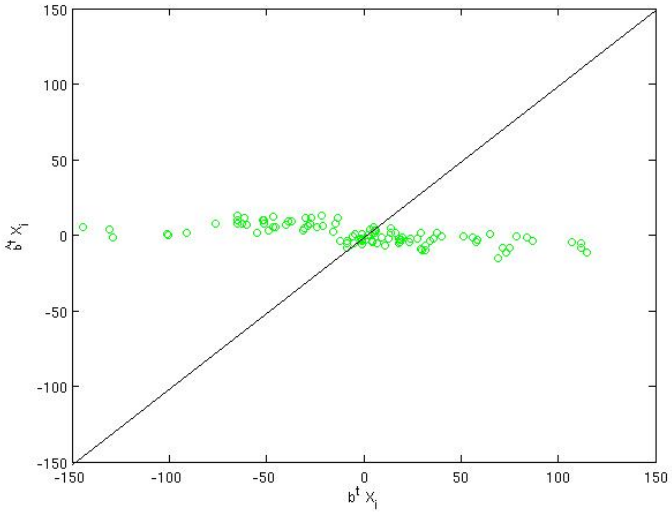
Rappel : données simulées

- Echantillon $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ de taille $n = 100$ avec $X_i \in \mathbb{R}^p$, $p = 50$, et $Y_i \in \mathbb{R}$, $i = 1, \ldots, n$.
- $X_i \sim \mathcal{N}_p(0, \Sigma)$ avec $\Sigma = Q\Delta Q^t$ où
  - $\Delta =$diag$(p^\theta, \ldots, 2^\theta, 1^\theta)$,
  - θ contrôle la décroissance des valeurs propres,
  - Q est une matrice d'orientation tirée selon la loi uniforme sur l'ensemble des matrices orthogonales.
- $Y_i = g(b^t X_i) + \xi$ où
  - g est la fonction de lien $g(t) = \sin(\pi t/2)$,
  - b est le vrai axe $b = 5^{-1/2} Q(1, 1, 1, 1, 1, 0, \ldots, 0)^t$,
  - $\xi \sim \mathcal{N}_1(0, 9.10^{-4})$

Rappel : résultats de SIR, $\theta = 2$, dimension $p = 50$

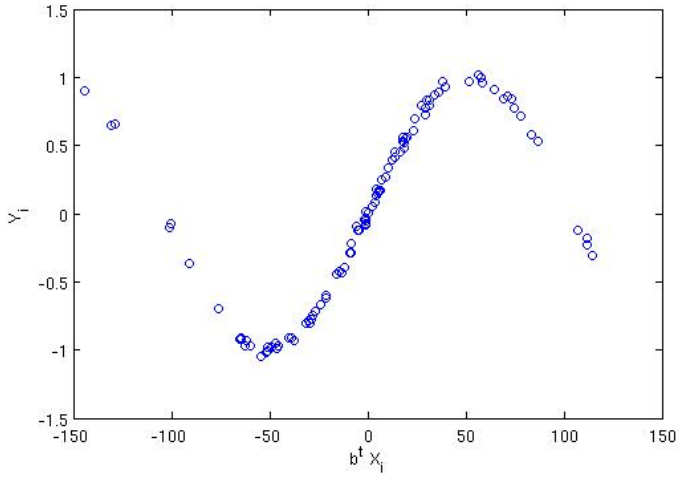


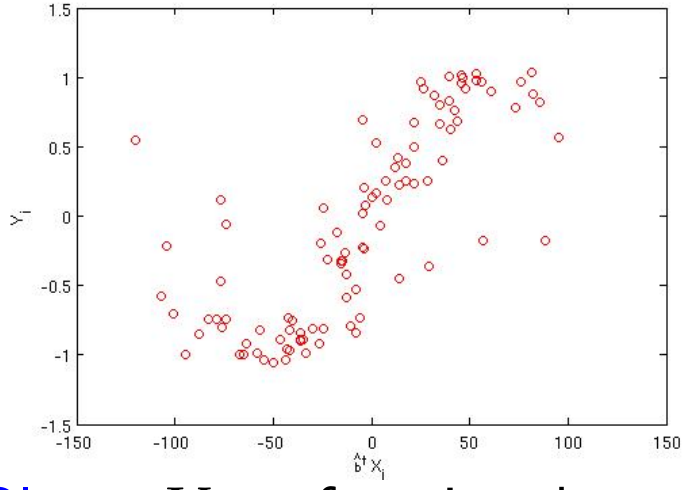


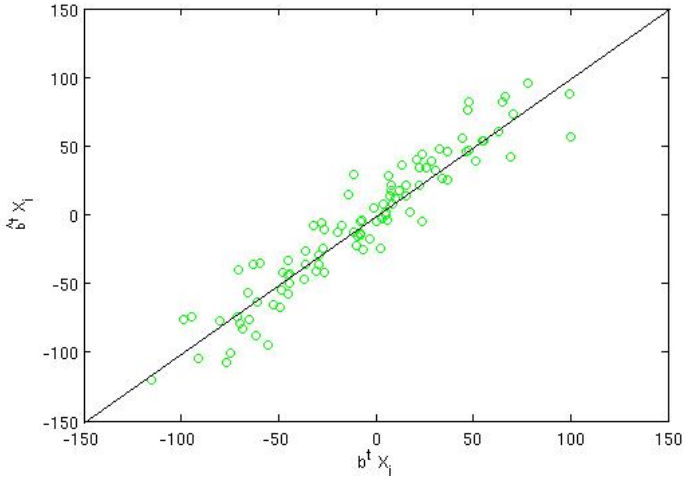


Bleu : Y_i en fonction des projections $b^t X_i$ sur le vrai axe b,
Rouge : Y_i en fonction des projections $\hat{b}^t X_i$ sur l'axe estimé $\hat{b}$,
Vert : $\hat{b}^t X_i$ en fonction de $b^t X_i$.

Résultats de GRSIR (ACP+Ridge)







Bleu : Y_i en fonction des projections $b^t X_i$ sur le vrai axe b,
Rouge : Y_i en fonction des projections $\hat{b}^t X_i$ sur l'axe estimé $\hat{b}$,
Vert : $\hat{b}^t X_i$ en fonction de $b^t X_i$.

Validation sur simulations

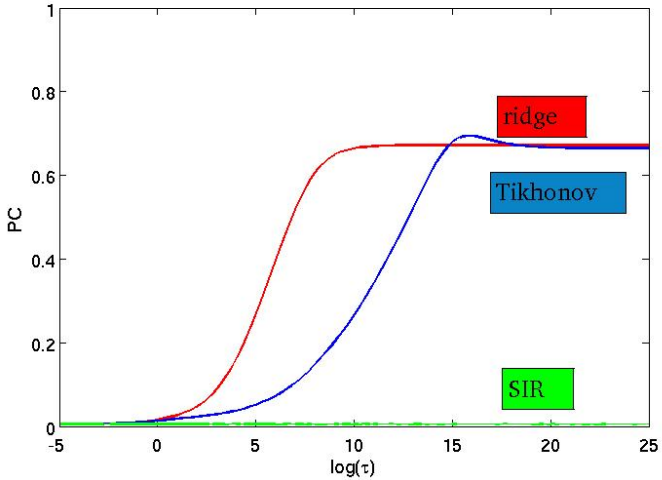
Critère de proximité entre le vrai axe b et son estimation $\hat{b}^{(r)}$ sur $N = 100$ réplications :

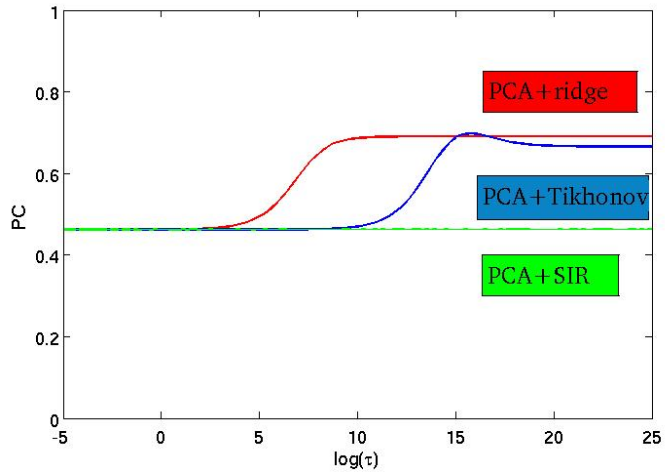
$$\text{PC} = \frac{1}{N} \sum_{r=1}^{N} (b^t \hat{b}^{(r)})^2$$

- $0 \leq$ PC≤ 1,
- Une valeur proche de 0 signifie une faible proximité : les estimations $\hat{b}^{(r)}$ sont presque orthogonales à b,
- Une valeur proche de 1 signifie une forte proximité : les estimations $\hat{b}^{(r)}$ sont presque colinéaires avec b.

Influence du paramètre de régularisation

PC en fonction de $\log \tau$. La dimension du sous-espace propre et le conditionnement sont fixés ($d = 20$ et $\theta = 2$).

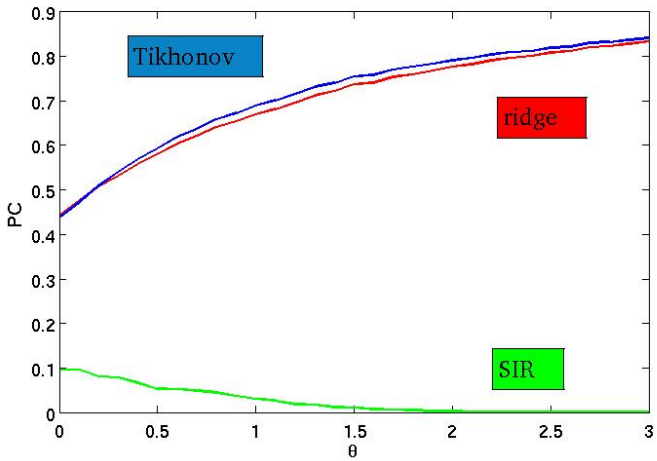


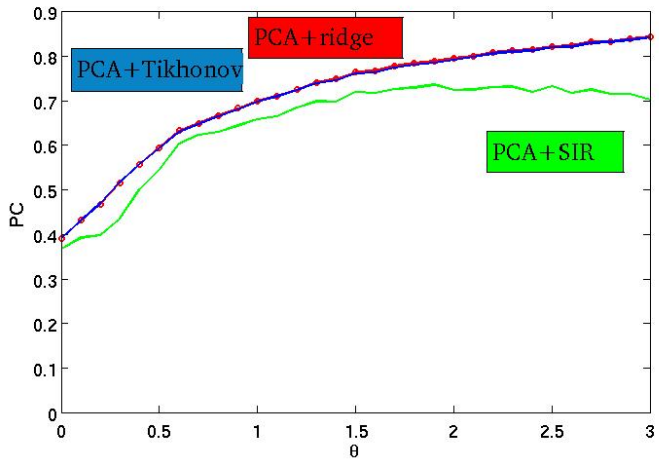


- Ridge et Tikhonov : bons résultats si τ est grand,
- ACP+SIR : résultats corrects comparés à SIR,
- ACP+ridge et ACP+Tikhonov : faible sensibilité à τ.

Sensibilité au conditionnement de la matrice de covariance

PC en fonction de θ. La dimension du sous-espace propre est fixée à $d = 20$. La valeur optimale du paramètre de régularisation est calculée pour chaque θ.

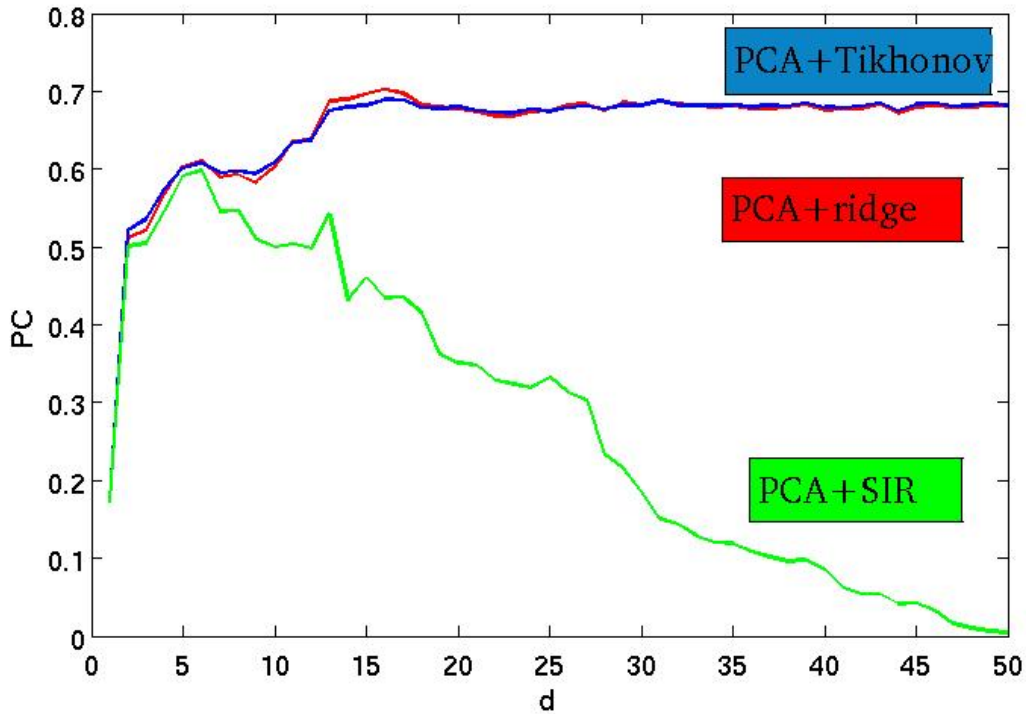




- Seule SIR est très sensible au mauvais conditionnement,
- ridge et Tikhonov : résultats similaires,
- ACP+ridge et ACP+Tikhonov : résultats similaires.

Sensibilité à la dimension du sous-espace propre

PC en fonction de d. Le conditionnement est fixé ($\theta = 2$). La valeur optimale du paramètre de régularisation est calculée pour chaque d.



- ACP+SIR : très sensible à d.
- ACP+ridge et ACP+Tikhonov : stables lorsque d augmente.

Plan

Estimation des propriétés physiques de la surface de Mars à partir d'images hyperspectrales

Contexte : ANR, Masse de données et connaissances.

- Observation du pôle sud de Mars à la fin de l'été, orbite 61 du spectromètre OMEGA à bord de la mission Mars Express.
- Cette partie de Mars contient essentiellement de la glace, du CO_2 et de la poussière.
- Image 3D : En chaque pixel, un spectre de $p = 184$ longueurs d'ondes est enregistré.

But : Pour chaque spectre $X \in \mathbb{R}^p$, estimer les paramètres physiques correspondants $Y \in \mathbb{R}$ (ici, la taille des grains de CO_2).

Un problème inverse

Problème direct.

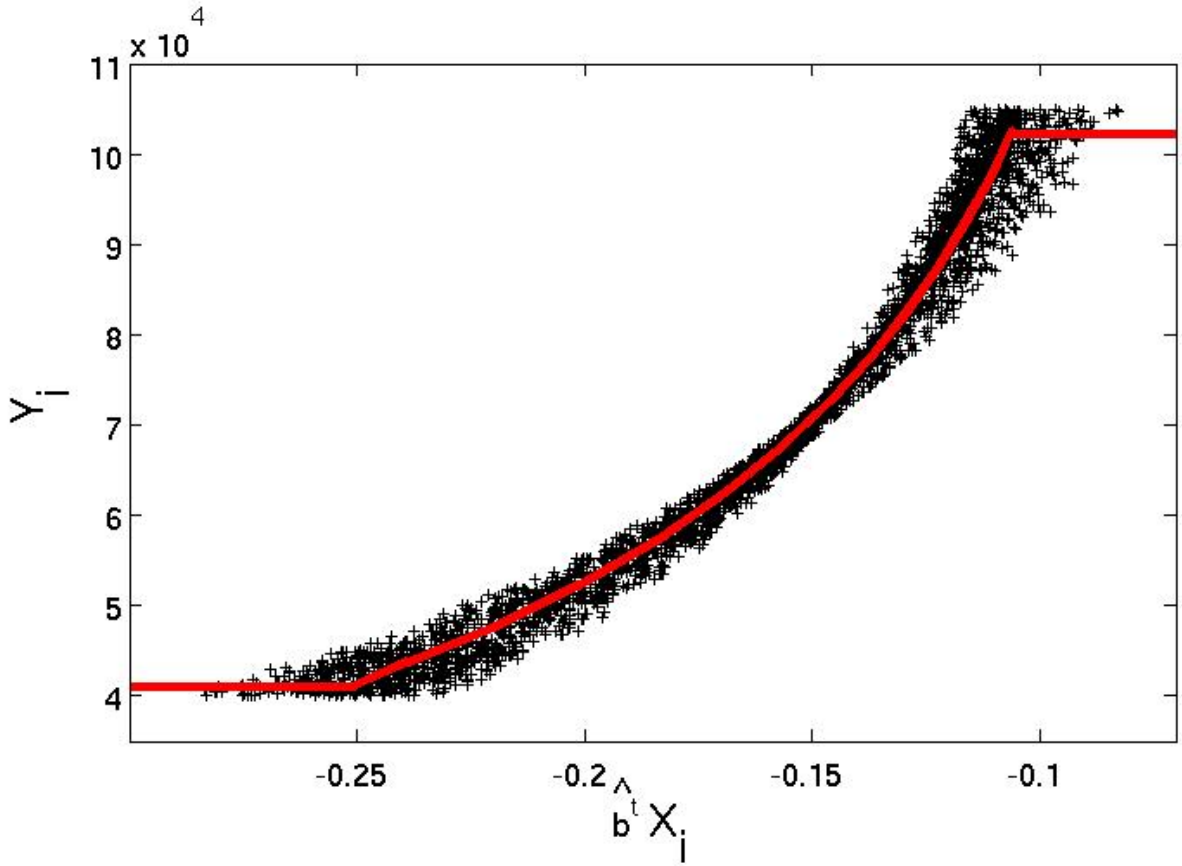
- Modèle physique de la formation d'un spectre (modèle de réflectance de surface).
- Pour un paramètre physique Y, on sait donc simuler le spectre correspondant $X = F(Y)$.
- Génération de $n = 12.000$ spectres synthétiques à partir des paramètres physiques correspondants.

$\Longrightarrow$ Base d'apprentissage.

Problème inverse.

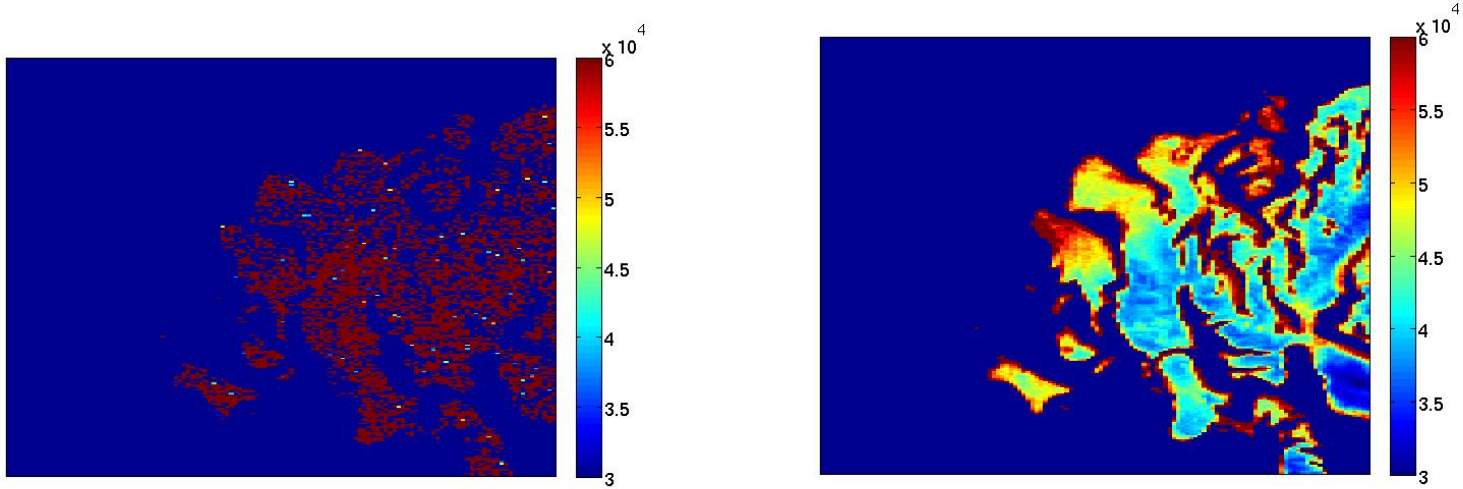
- Estimer la relation fonctionnelle inverse $Y = G(X)$.
- Réduction de dimension $G(X) = g(b^t X)$.
- b est estimé par SIR/GRSIR, g est estimée par une régression non-paramétrique univariée.

Relation fonctionnelle estimée



Relation estimée entre les spectres projetés $\hat{b}^t X$ sur le premier axe de GRSIR (ACP+ridge) et Y, la taille des grains de CO_2.

Cartes de CO_2 estimées



Taille des grains de CO_2 estimées par SIR (gauche) et GRSIR (droite) sur une image hyperspectrale observée sur Mars.

Références sur ce travail

- Bernard-Michel, C., Douté, S., Fauvel, M., Gardes, L. et Girard, S. (2008). Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression.
  `http ://hal.inria.fr/inria-00276116/fr`
- Bernard-Michel, C., Gardes, L. et Girard, S. (2008). Gaussian Regularized Sliced Inverse Regression, *Statistics and Computing*, à paraître.
  `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`

Références sur 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.