Utilisation de la méthode S.I.R. pour la reconstruction de paramètres physiques

Laurent GARDES

Objectif

Etablir une liaison fonctionnelle entre :

- Les spectres $x \in \mathbb{R}^p$ (p=184 ou 256) fournis par la mission Mars Express,
- des paramètres physiques $y \in \mathbb{R}^q$ ($q \approx 5$) (taille des grains de CO_2, H_2O, poussière, etc ...). On suppose pour simplifier que $q = 1$.

Il s'agit donc de construire une fonction

$$f : \mathbb{R}^p \to \mathbb{R}.$$

Pour ce faire, on dispose d'un échantillon d'apprentissage $(x_i, y_i), i = 1, \ldots, n$. (généré par le LPG qui utilise un modèle physique permettant de construire un spectre a partir des paramètres).

Difficulté

- Il faut construire une fonction de p variables avec p grand : problème du **fléau de la dimension**.

Reconstruction des paramètres physiques

- Il faut au préalable **réduire la dimension de** x. Autrement dit, il faut trouver un (ou plusieurs) axe $a \in \mathbb{R}^p$ sur lequel on projette x.
- On estime ensuite la fonction g : $\mathbb{R} \to \mathbb{R}$

$$y = g(< a, x >).$$

Réduction de la dimension

- **ACP sur** x

Cette méthode consiste à trouver la projection maximisant la variance de x. Le premier axe de projection de l'ACP est défini par :

$$\hat{a}_{ACP} = \arg \max_{\|a\|=1} \mathrm{Var}(< a, x >) = \arg \max_{\|a\|=1} a' \Sigma_x a,$$

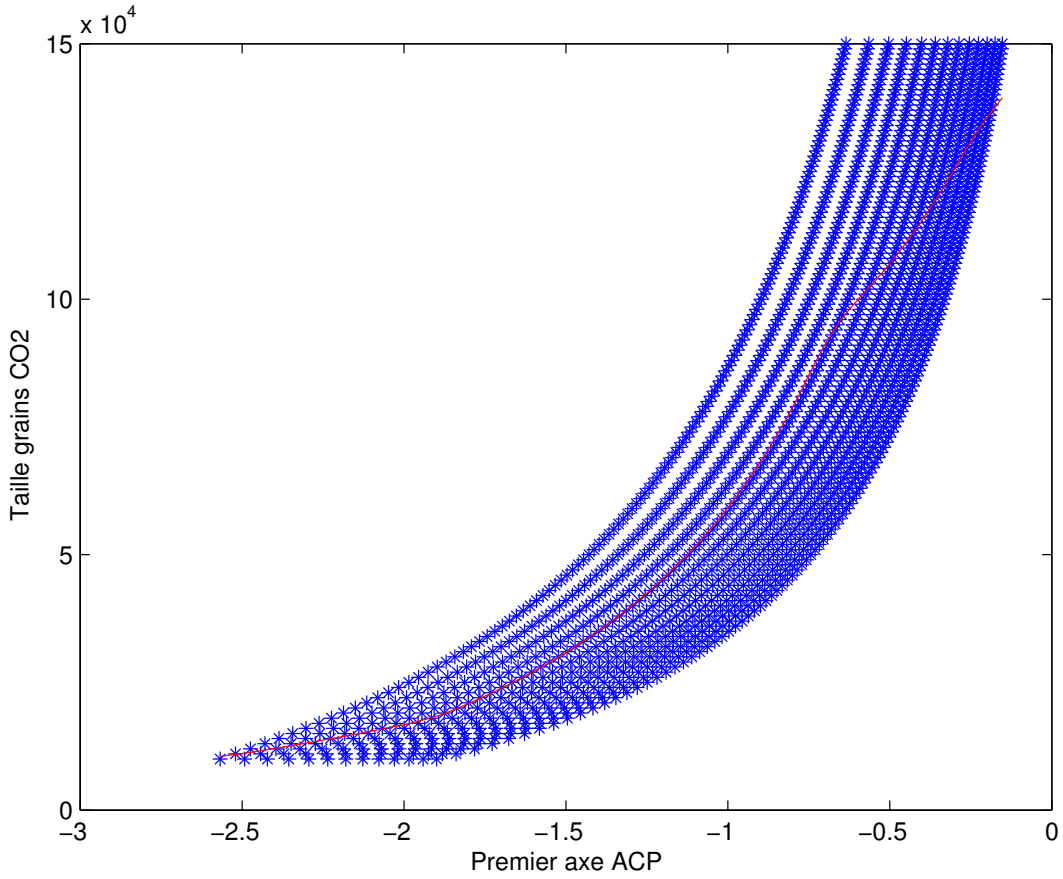
où Σ_x est la matrice de variance-covariance de x.

Cela revient plus simplement à chercher le vecteur propre associé à la plus grande valeur propre de Σ_x.

Inconvénient de l'ACP : Cette méthode ne tient pas compte de y. Elle n'est donc pas adaptée pour ce problème.

Illustration

On utilise la méthode ACP sur un jeu de données de taille $n = 1833$ avec des spectres x de dimension $p = 256$. Ces spectres ont été généré à partir de plusieurs paramètres physiques. On ne considère ici que le paramètre y "taille des grains de CO_2".


x 10^4
15
10
5
0
−3
−2.5
−2
−1.5
−1
−0.5
0
Taille grains CO2
Premier axe ACP

• **SIR univarié (Sliced Inverse Regression)**

Cette méthode s'appuie principalement sur deux hypothèses :

1) Il existe une fonction g telle que

$$y = g(< a_1, x >, \ldots, < a_K, x >, \epsilon),$$

où ϵ est une erreur indépendante de x et $K < p$ (Pour simplifier, $K = 1$).

2) La distribution de x est elliptique. (Loi Gaussienne).

Idée de la méthode SIR : Trouver l'axe a_{SIR} qui maximise la variance de la courbe de régression inverse.

La méthode SIR est donc une ACP sur la courbe de régression inverse.

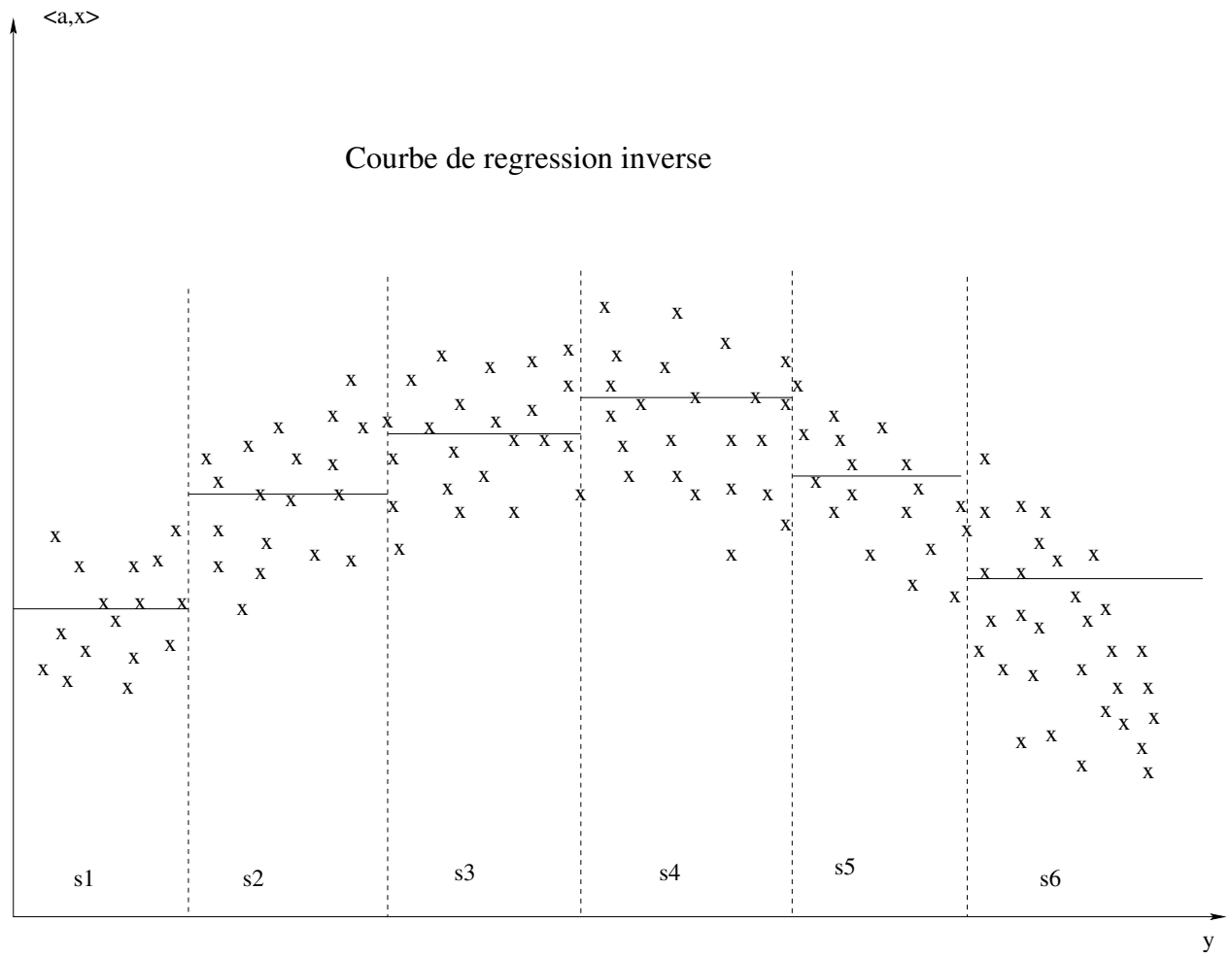
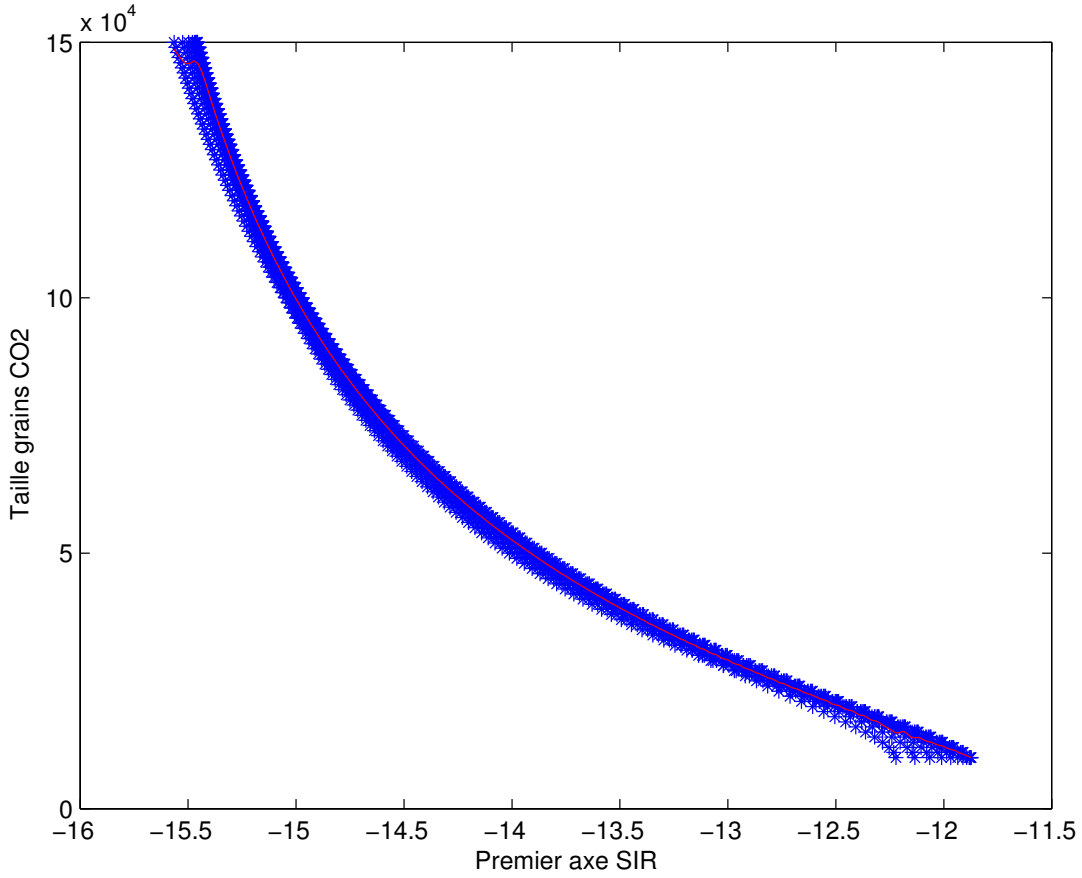

<a,x>
Courbe de regression inverse
s1
s2
s3
s4
s5
s6
y

Illustration de la méthode SIR sur le même jeu de données.



Estimation de la fonction de régression

On utilise l'estimateur à noyau défini par

$$\hat{g}(x_0) = \frac{\sum_{i=1}^{n} y_i K((x_i - x_0)/h)}{\sum_{i=1}^{n} K((x_i - x_0)/h)}.$$

Exemple :

$$K(x) = \cos^2(\pi x), \text{ sur l'intervalle [-1/2,1/2]}$$

Le choix du paramètre de lissage h peut se faire par validation croisée.

Perspectives

- Utiliser plusieurs paramètres (SIR multivarié).
- Garder plusieurs axes pour SIR ($K > 1$).
- La méthode SIR necessite l'inversion de la matrice Σ_x. Une autre conséquence de la grande dimension x est la (presque) singularité de Σ_x. La solution que l'on a adopté pour remédier à cela est de faire au préalable une ACP sur x pour réduire sa dimension. Il faudrait explorer d'autres pistes.