Datamining 5: La grande dimension
La regression regression sparse (parcimonieuse)

M2 STD

November 26, 2014

L'idée de base

X est de (très) grande dimension et Y de dimension 1.
On a le modèle linéaire habituel : $Y = X.A + b + \varepsilon$
Mais on suppose que seul un petit nombre des variables de X rentrent en compte dans le modele. c'est a dire qu'on suppose qu'un grand nombre des a_i est nul

L'idée de base

Comme X est de (très) grande dimension il y a beaucoup de liaisons dans X, imaginons par exemple le cas caricatural où $X_1 = X_2$ et ou X_1 et X_2 sont indépendant de Y (ne jouent pas dans la regression) alors tout modele de regression $Y = aX_1 - aX_2 + \ldots$ est équivalent. On peut donc obtenir des coefficients a arbitrairement grands sans nuire a la qualité prédictive du modele (mais on aua des pb de généralization, d'instabilité et on ne pourra pas les interpréter). En pratique de par l'inversion d'une matrice tres singulière il va apparaitre des coefficients très grands dans ce cas la

Si X est juste de "grande" dimension

Rappelons (cours de regression) les critères BIC et AIC On va minimiser: $\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 + \lambda \sharp\{a_i, i \neq 0\}$ avec :

1. $\lambda = \sigma^2/n$ (AIC)
2. $\lambda = \sigma^2 \ln n/(2n)$ (BIC)

Le cout algorithmique est enorme ! et il est déraisonabme d'utiliser un tel critère lorsque p est trop grand.

Solution(s)

Ridge

Résoudre le probleme :
minimiser $\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$ sachant $\sum_{i=1}^{p} a_i^2 < t$
Equivalent (lagrangien) minimiser : $\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 + \lambda \sum_{i=1}^{p} a_i^2$

LASSO (lasso dans matlab)

Résoudre le probleme :
minimiser $\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$ sachant $\sum_{i=1}^{p} |a_i| < t$
Equivalent (lagrangien) minimiser : $\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 + \lambda \sum_{i=1}^{p} |a_i|$

Remarque : (ridge dans matlab)

Dans les deux cas on introduit un parametre λ qu'il va falloir regler soigneusement (validation)

Les résultats théoriques de ces deux méthodes sont très bons (mais très compliqués a établir) les gens interressés peuvent aller voir le très bon cours : de franck Picard a Lyon (cours Lasso)

Un exemple (sous matlab)

60 individus qui sont différentes essences (de voitures) sur lesquelle on a effectuer une spectrometrie i.e. chaque essence est caractérisé par une "courbe" de 401 point (401 variables). On cherche a estimer le taux d'octane. On regarde les résultats d'une pénalisation LASSO sous matlab (mieux faite que ridge, entre autre la cross validation est intégrée)

Choix de la méthode de selection de λ

On n'a que 60 individus on va faire de la cross validation

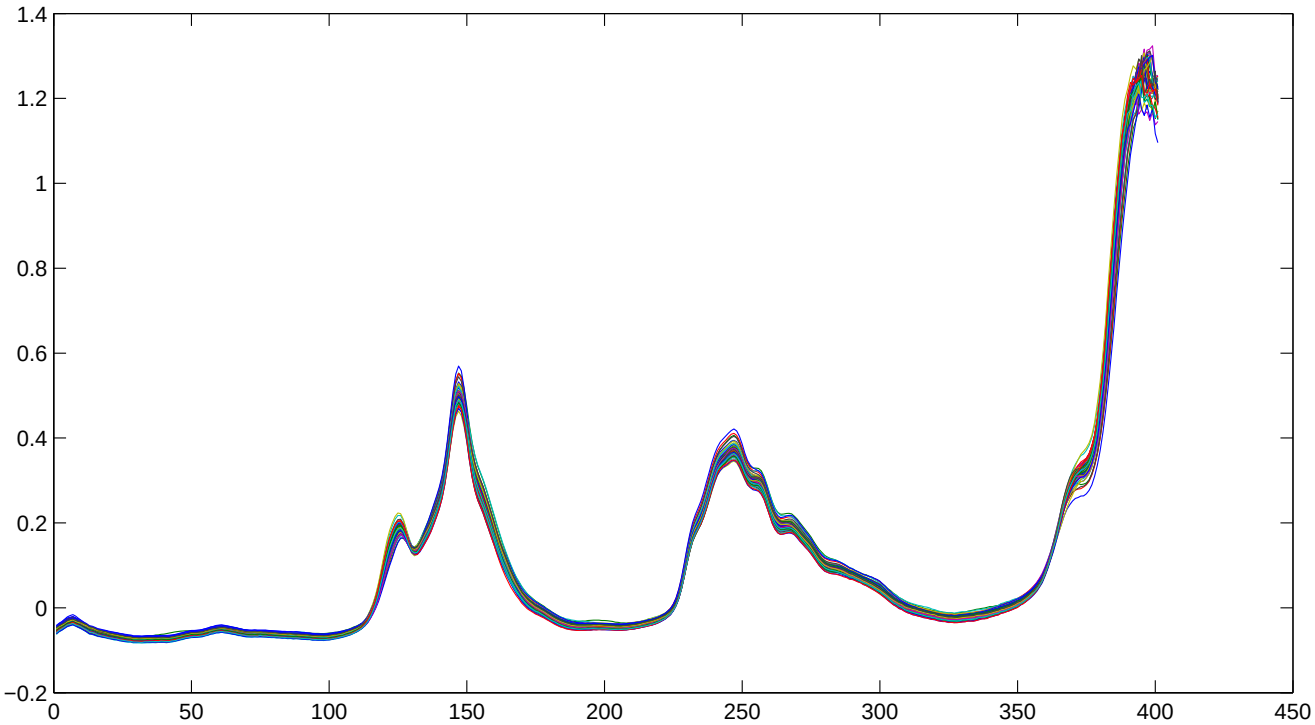

Figure: Les données

Selection du λ

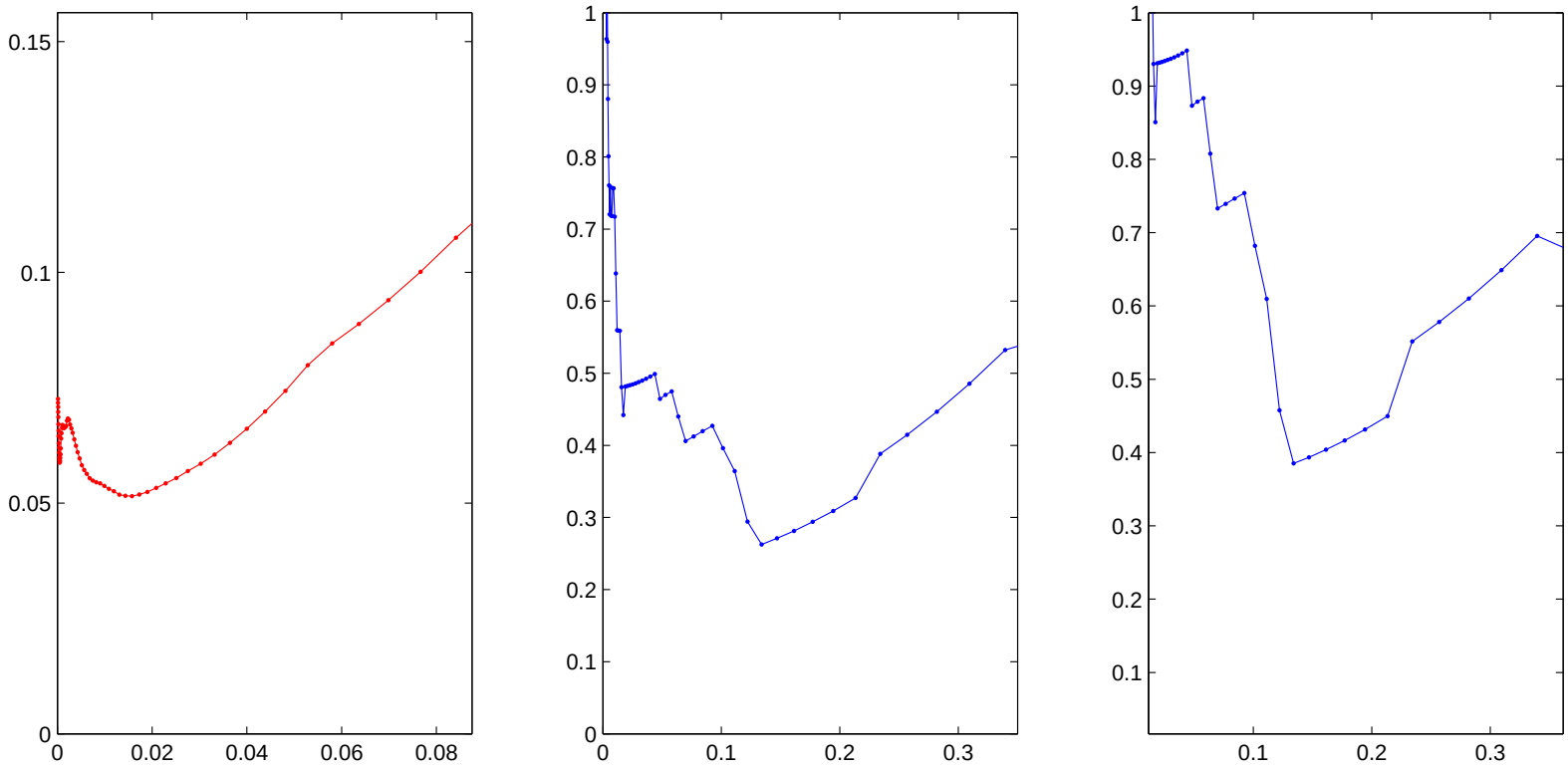

Figure: Pour chaque λ on obtient un certain nombre de coefficients non nuls... rien ne nous empeche de faire du AIC ou BIC a posteriori

Résultat

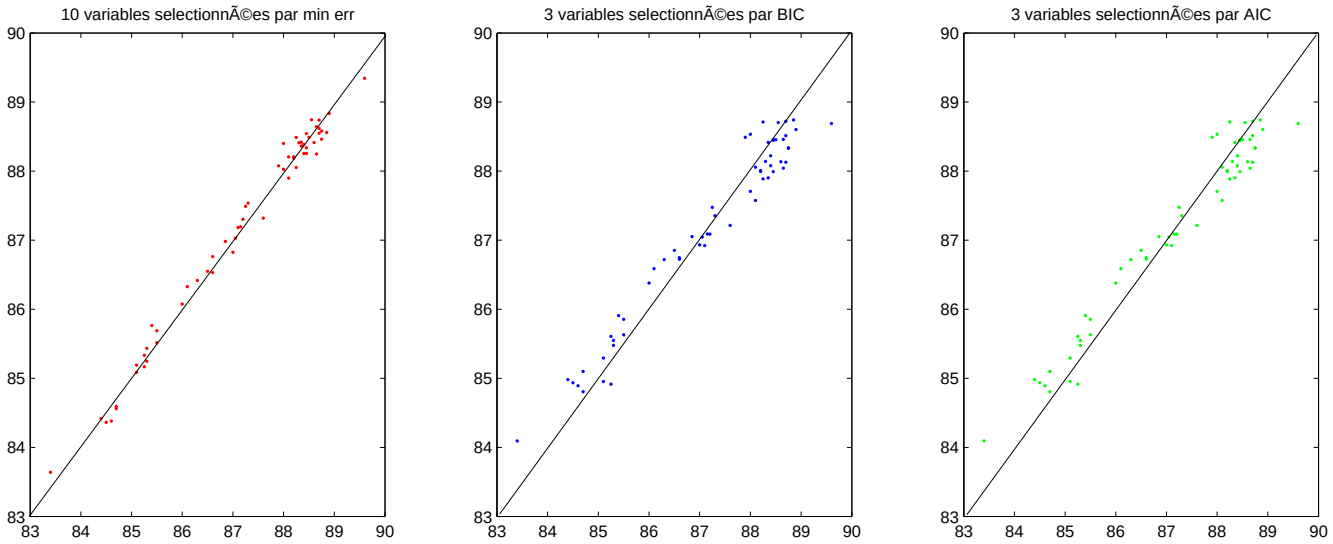

Figure: Les données

remarque La regression pénalisée peut etre un “point de départ”

Noux venons en fait de selectionner trois variables (sur 401) qui expliquent très bien notre probleme et d’essayer de voir si on ne peut pas améliorer les choses avec du non linéaire... (dans notre exemple non)