Datamining 4: La regression Partie 1 si X est univariée

M2 STD

October 15, 2015

Problème

On a le modèle : $Y = f(X) + \varepsilon$ et on cherche a estimer f. Cette semaine on considère que X est de dimension 1 C'est un peu restrictif mais a néanmoins beaucoup d'applications:

- En temps que tel (modélisation)
- En séries chronologiques (recherche de tendance)
- Les données fonctionnelles (les individus sont caractérisés par des courbes)
  - Débruitage (la dimension est déjà infinie on peut avoir envie de travailler sur les courbes lissées pour réduire les problemes)
  - Avoir des observations comparables (si chaque courbe qui caractérise les individus n'est pas mesuré aux même points)

Quelques solutions

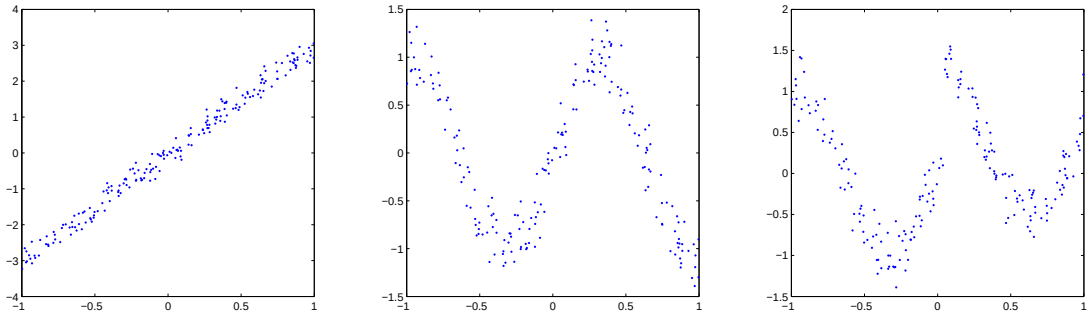
$Y = f(X) + \varepsilon$ et on cherche a estimer f
Dans le cas de la dimension 1 on peut (doit) en premier lieu :

Faire un dessin et calculer le coefficient de corrélation

- Si f est linéaire on fait une regression linéaire
- Toute fonction f continue peut etre approchée par un polynome ... on peut tenter une regression polynomiale (voir cours numero 1) Pb : ca sera difficile a généraliser au cas général ou X est de dimension f
- Les méthodes de lissage des séries chronologiques sont très bien aussi ex:
  - moyenne mobiles
  - Lissage a Noyaux

Méthodologie générale

Toujours faire un dessin car dans ce cas là on voit quelque chose



- Linéaire $\Rightarrow$ regression linéaire
- Continu $\Rightarrow$ regression a Noyaux
- Discontinu $\Rightarrow$ detection des ruptures et travail par morceaux

La regression a Noyau

On se donne une fonction noyau (cf estimation de densité):

- $K \geq 0$
- $\int K = 1$
- $\int uK(u) = 0$ (ou pour les profs dyslexiques $K(-u) = K(u)$)
- $\int u^2 K(u) < \infty$

La regression a Noyau

L'estimateur a Noyau de la regression est

$$\hat{f}(x) = \frac{\sum_i Y_i K((x - X_i)/h)}{\sum_i K((x - X_i)/h)}$$

Si on regarde bien on s'appercoit que les moyennes mobiles sont identiques a l'estimateur a Noyau

La théorie de l'estimateur a Noyau

Ecrivons l'estimateur a Noyau un peu différemment

$$\hat{f}_h(x) = \frac{\frac{1}{Nh}\sum_i Y_i K((x - X_i)/h)}{\frac{1}{Nh}\sum_i K((x - X_i)/h)}$$

On a déjà montré (dernière séance) que : si $h \to 0$ et $Nh \to \infty$ alors : $\frac{1}{Nh}\sum_i K((x - X_i)/h) \to \varphi_X(x)$ où φ_X est la densité de X.
On sait même que la vitesse optimale est pour $h = N^{-1/5}$
Il reste a montrer que si $Y_i = f(X_i) + \varepsilon_i$ avec $E(\varepsilon) = 0$ et f continue alors, on a $\frac{1}{Nh}\sum_i Y_i K((x - X_i)/h) \to \varphi_X(x) f(x)$ sous les mêmes condition pour h

La pratique de l'estimateur a Noyau

Comme d'habitude....

- Le vrai probleme est le choix de h
- La cross-validation est notre amie *i.e.* on va minimiser (en h) $S(h) = \sum(Y_i - \hat{f}_{-i,h}(X_i))^2$ où

$$\hat{f}_{-i,h}(x) = \frac{\sum_{j \neq i} Y_j K((x - X_j)/h)}{\sum_{j \neq i} K((x - X_j)/h)}$$

Exemple

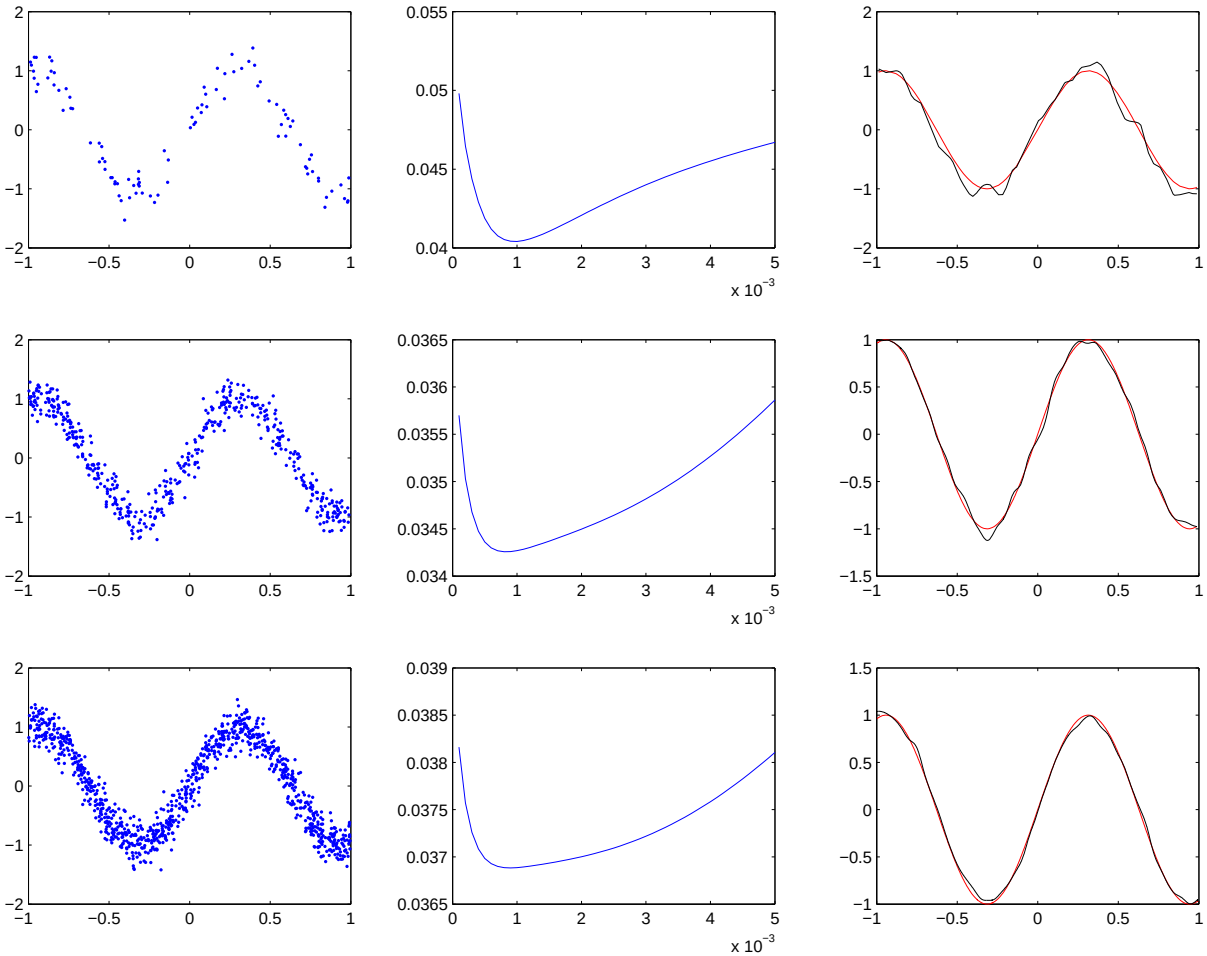


Figure: Joli non ?

Cas d'existence de ruptures

Demander a Pierre Bertrand ! Sinon faites ca a la main, en dimension 1 c'est possible Ou utiliser les noyaux pour detecter les ruptures

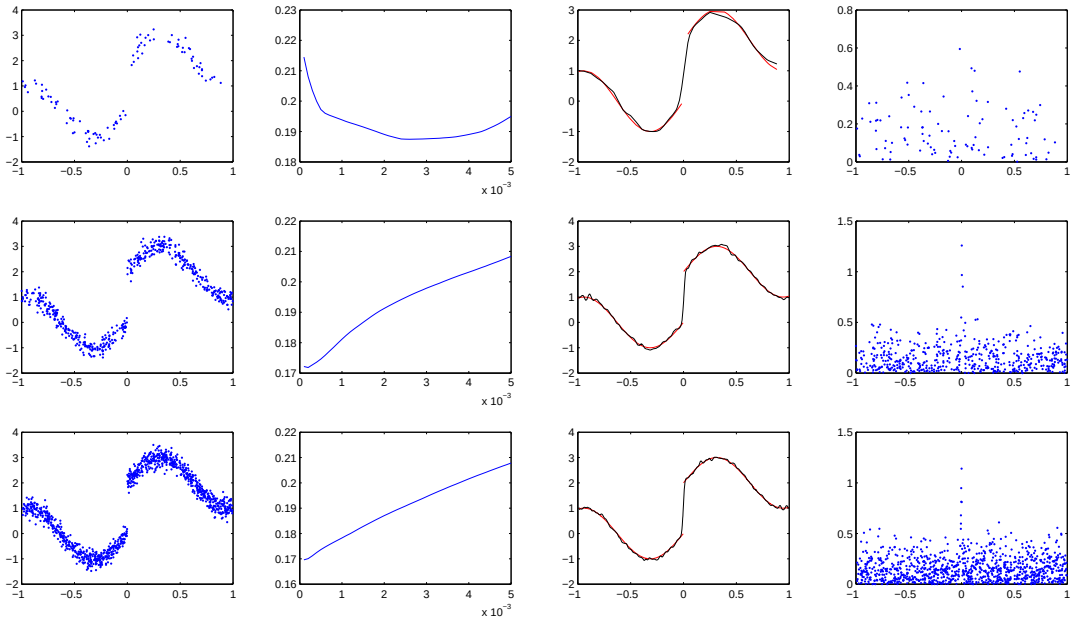


Figure: Au pire ca se passe pas si mal...