# Datamining Introduction : le choix d'un modèle

M2 STD

September 10, 2015

# Data Mining: définition

## Définition de wikipédia

"L'exploration de données, connue aussi sous l'expression de fouille de données, forage de données, prospection de données, data mining, ou encore extraction de connaissances à partir de données, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques"

# Datamining: beaucoup de "branches"

Sous l'appélation un peu "fourre tout" on trouve énormément de méthodes (on ne verra pas tout cette année)

- **Les méthodes descriptives:** regroupent les cours d'add mais aussi d'autres méthodes (comme les méthodes de projection non linéaires qu'on étudira au prochain chapitre 2). Essentiellement Projection et classification non supervisée
- **Les méthodes predictives:** regroupent bien sur les méthodes de regression (cf cours de regression) et les méthodes de classification supervisées. On verra quelques une de ces méthodes dans les chapitre 3 et 4

# Datamining ou méthodes "classique"

Ce qui différencie (communément) le datamining des méthodes classiques (add regression).

- **"Classique"** hypothèse fortes sur le modèle (linéarité, bruit gausssien...) on peut alors avoir des résultats théoriques (asymptotiques) sur la "performance" du modèle (ex: inertie expliquée en acp, R2 et différents test en regression...)
- **"Datamining"** hypothèses plus faibles sur le modèle, le "prix a payer" est lourd: algorithmes plus compliqués qui dépendent de parametres a choisir, le plus souvent pas de résultats théoriques sur l'asymptotique.

# Le choix du parametre

L'objectif de ce cours est d'explorer quelques méthodes de choix de parametre

# Regression polynomiale

- **Problème :** On observe $X_1, \dots X_n$ iid et $Y_1, \dots, Y_n$. **On suppose** qu'il existe un modèle polynomial qui explique $Y$. i.e. que $Y = f(X) + \varepsilon$ avec $f$ un polynome de degrè $p$.
- **La méthode :** Si on connait le degré $p$ du polynome on estime les coefficients par la méthode des moindres carrés
- **Le paramètre** est donc le degré du polynome

# Regression polynomiale

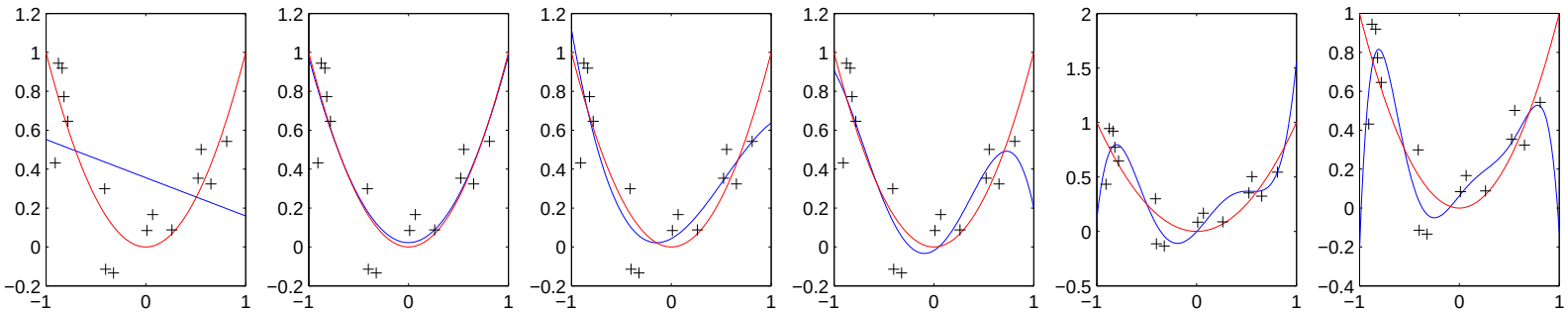


Figure: regression polynomiale: fonction estimée (bleue) vrai fonction (rouge) en fonction du degré (dans cet exemple $f(x) = x^2$)

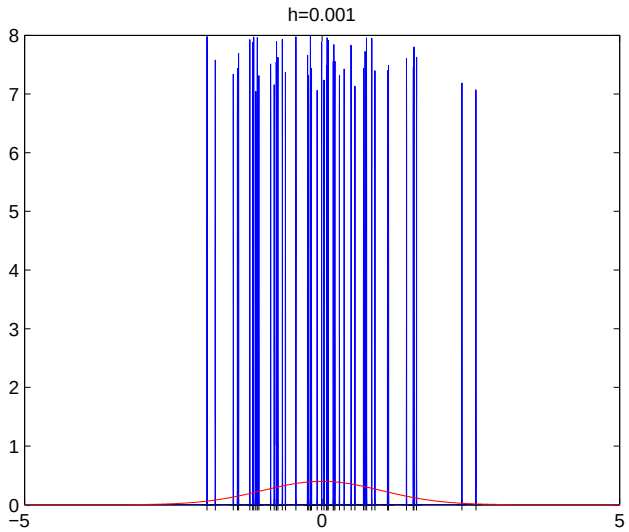
# Estimation de densité

- **Problème :** On observe $X_1, \dots X_n$ iid et on voudrait estimer la densité $f$ du tirage
- **La méthode :** Méthode de l'estimation a noyau: on choisit une fonction noyau $K$ (positive, paire, d'integrale 1) et un réel $h$. On estime la densité par la formule suivante:

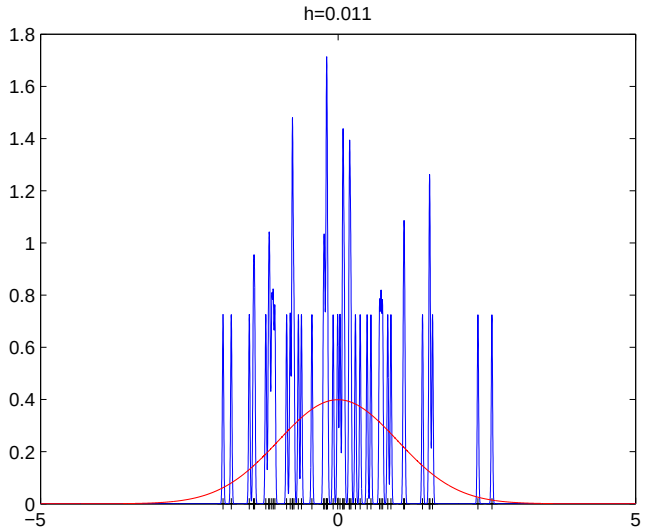
$$\hat{f}(x) = \frac{1}{nh} \sum_i K\left(\frac{x - X_i}{h}\right)$$

- **Les paramètre** $K$ et $h$
- **La théorie** le choix de $K$ est assez peu important (on prendra pour $K$ le noyau gaussien) en revanche celui de $h$ est fondamental la théorie dit que le meilleur $h$ donne la formule $h = n^{-1/5} C(f)$ (mais on ne connait pas $f$)

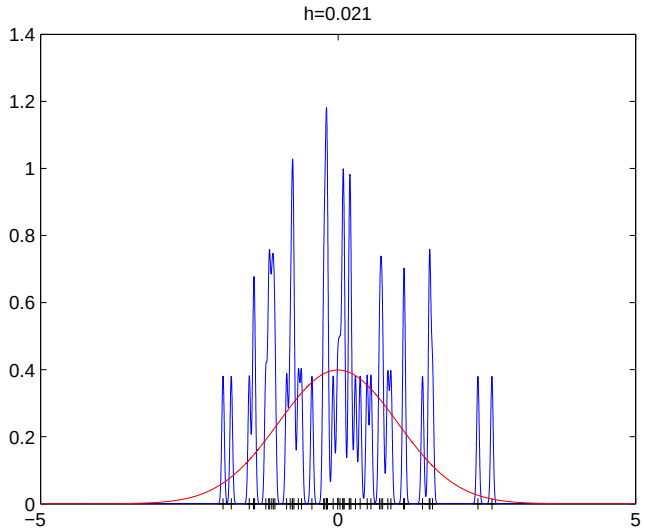
# Estimation de densité



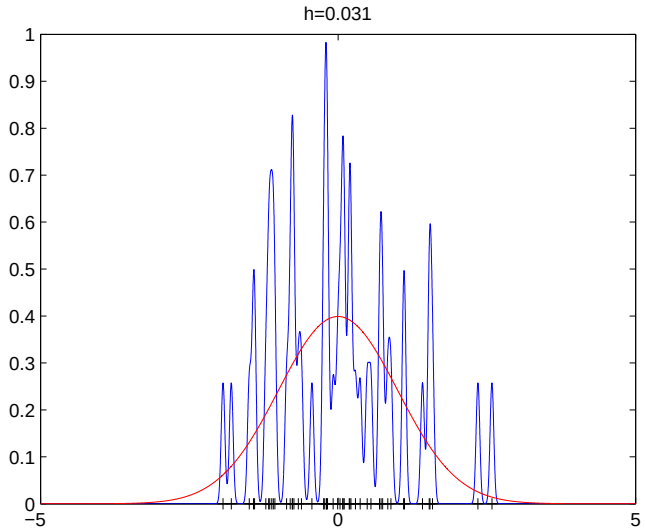
# Estimation de densité



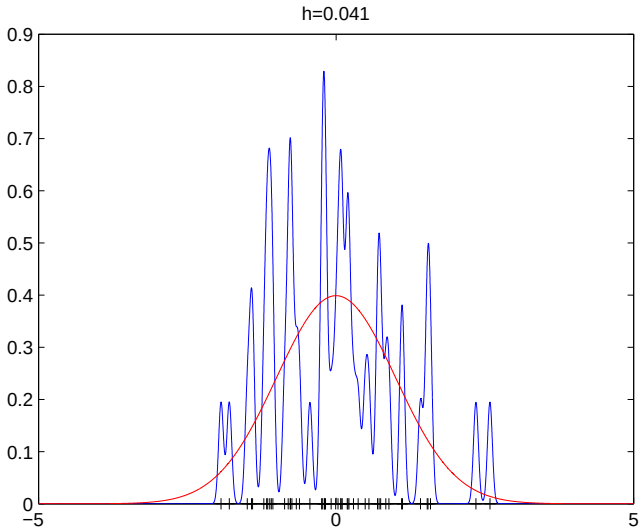
# Estimation de densité



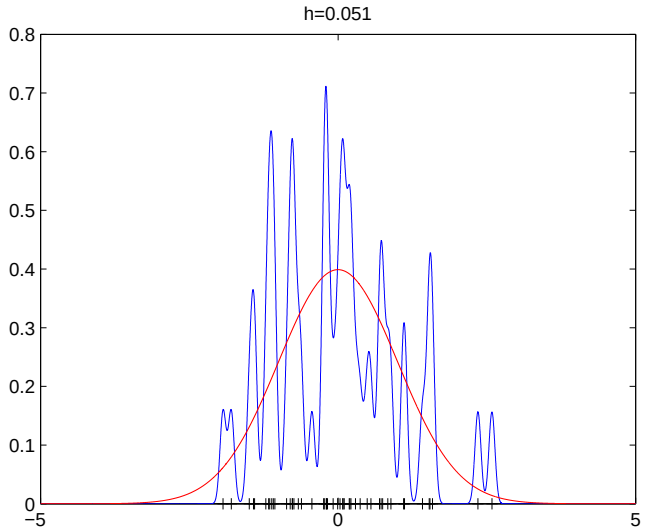
# Estimation de densité



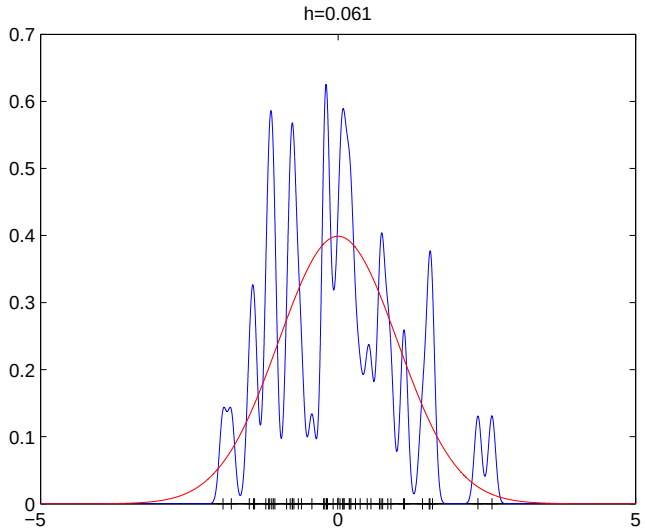
# Estimation de densité



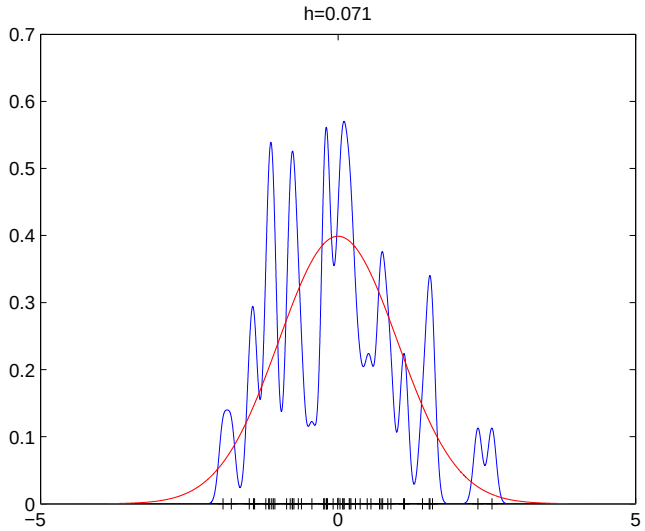
# Estimation de densité



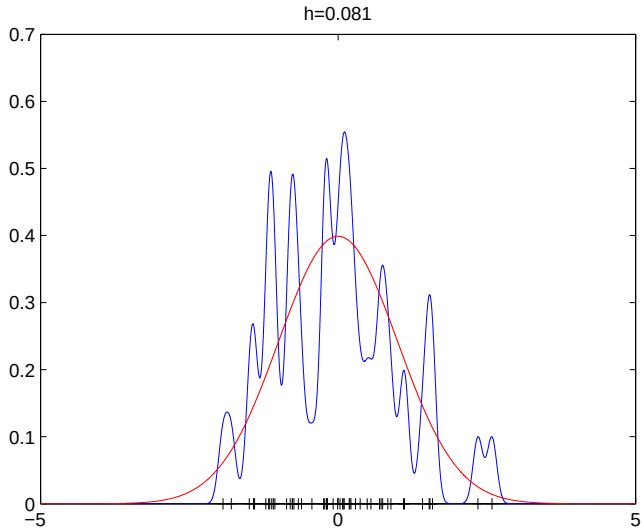
# Estimation de densité



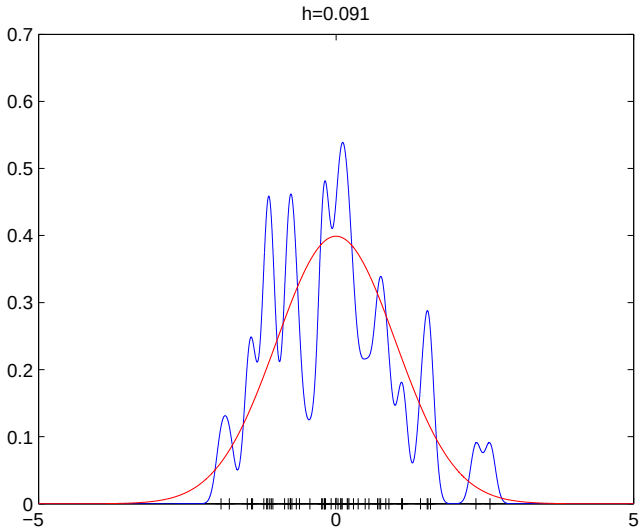
# Estimation de densité



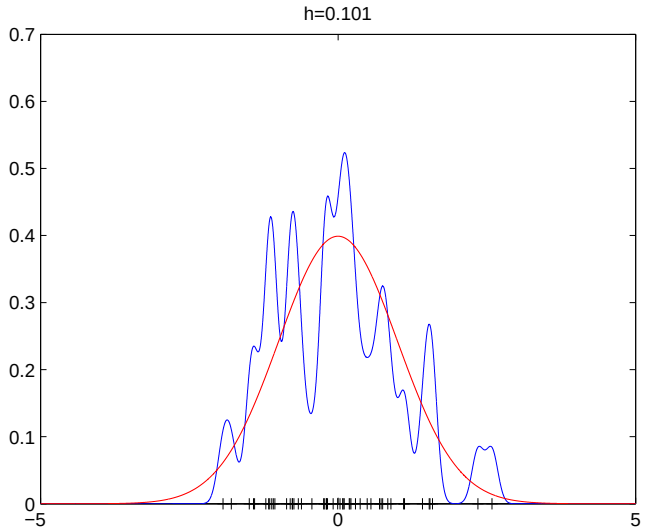
# Estimation de densité



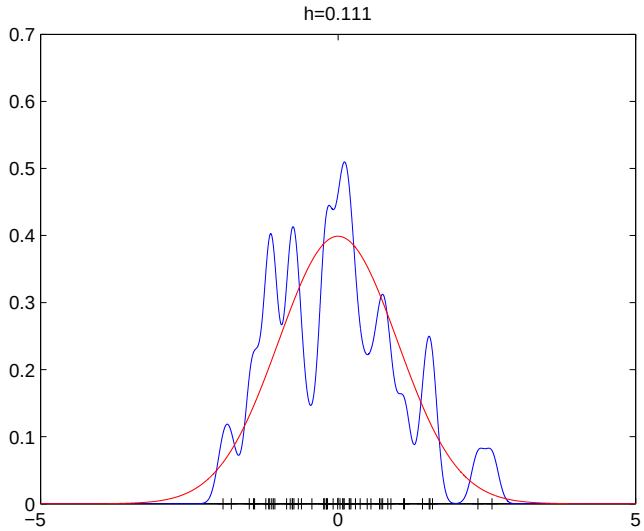
# Estimation de densité



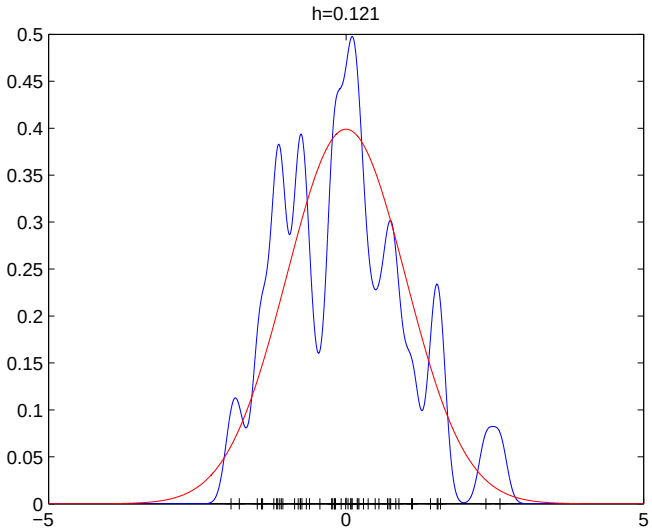
# Estimation de densité



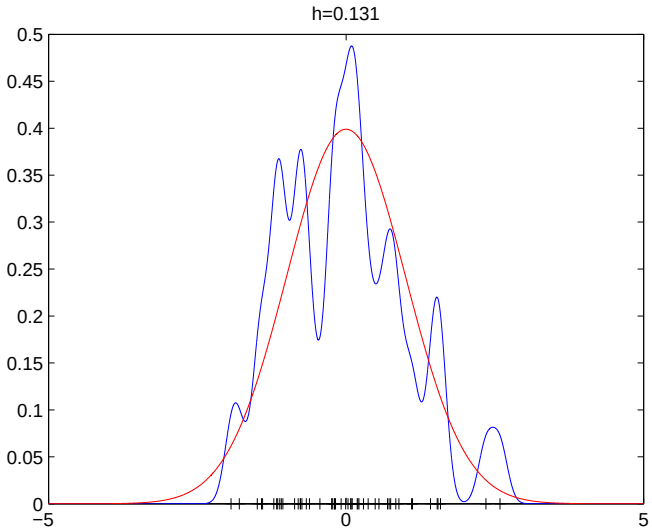
# Estimation de densité



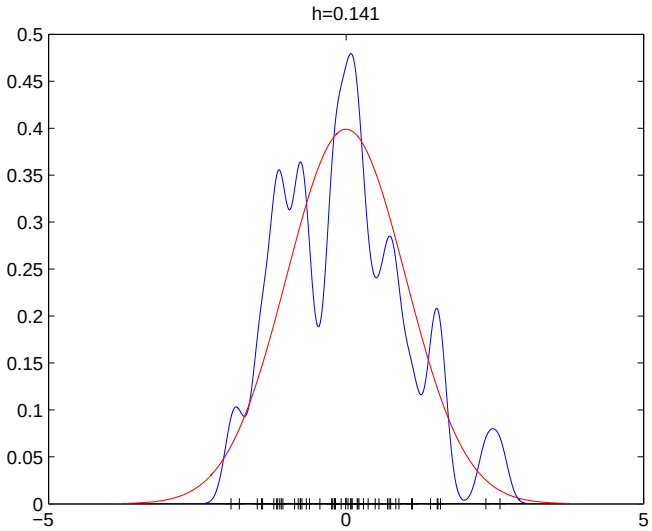
# Estimation de densité



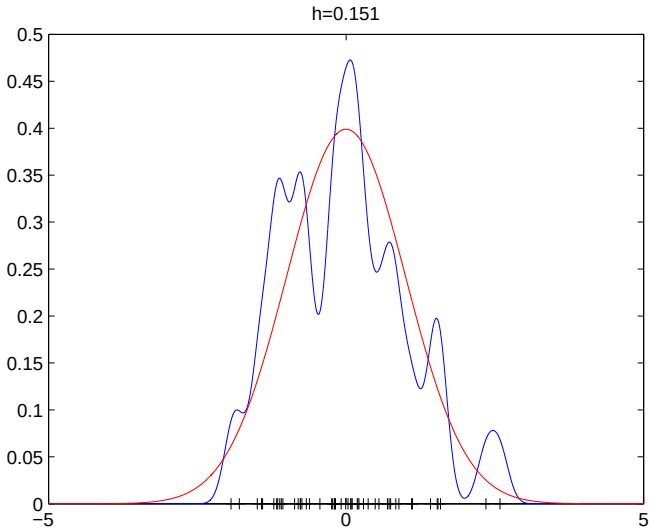
# Estimation de densité



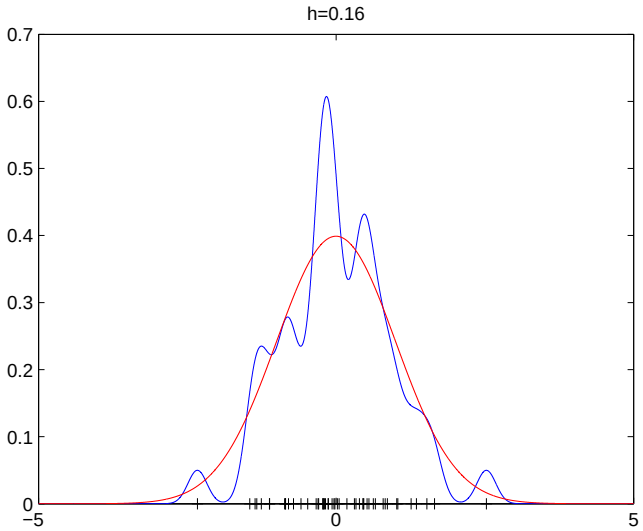
# Estimation de densité



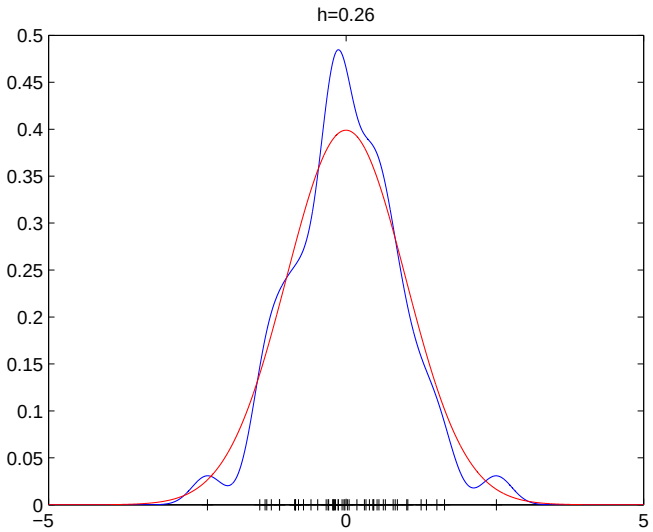
# Estimation de densité



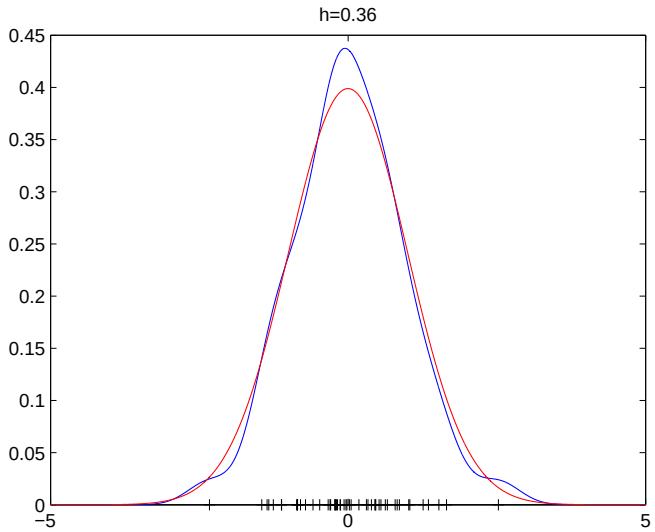
# Estimation de densité



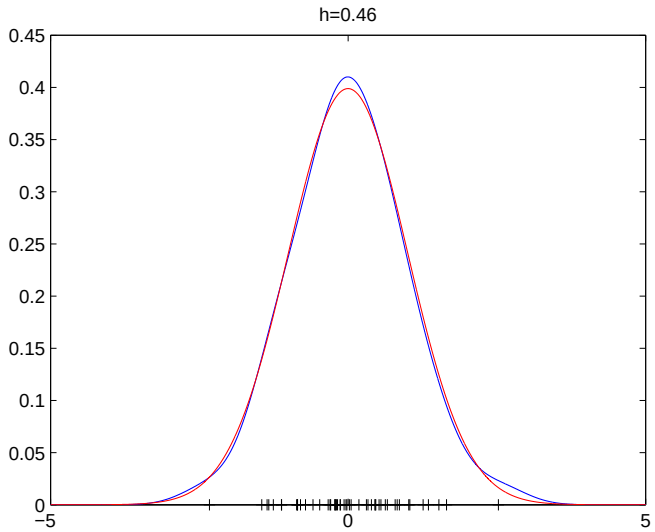
# Estimation de densité



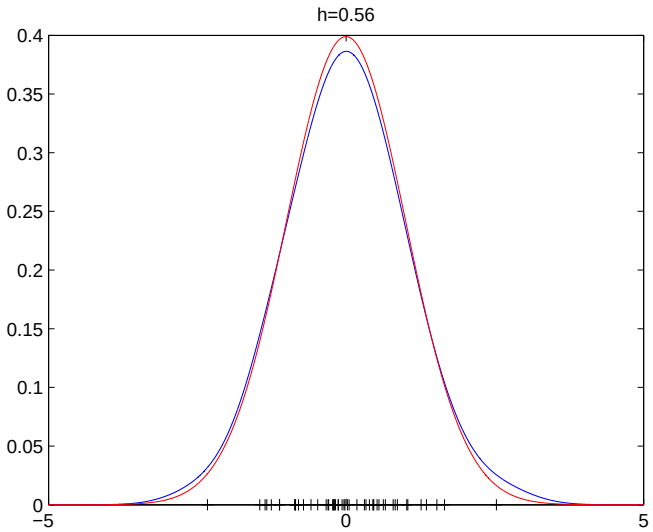
# Estimation de densité



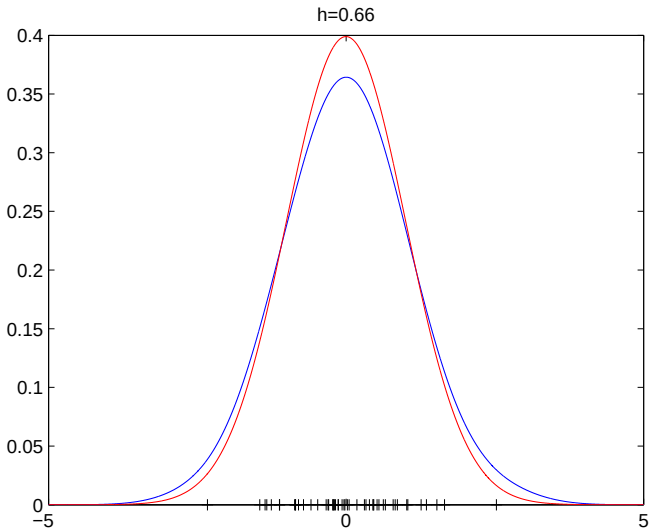
# Estimation de densité



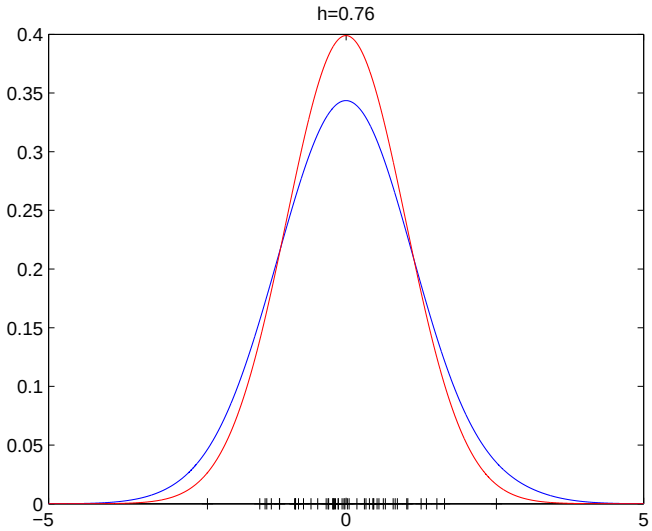
# Estimation de densité



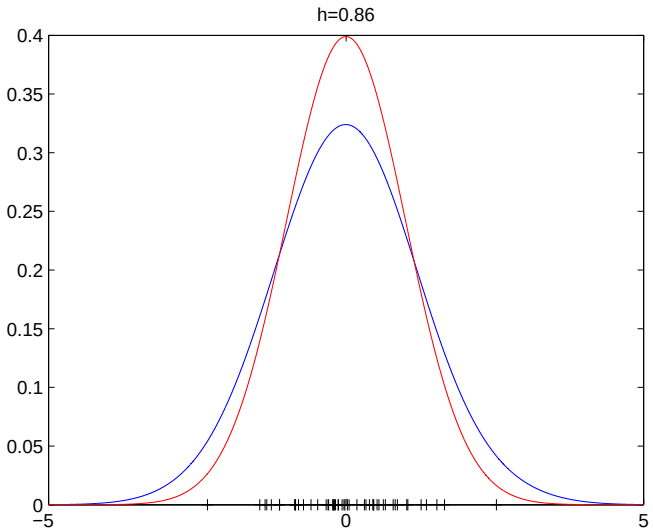
# Estimation de densité



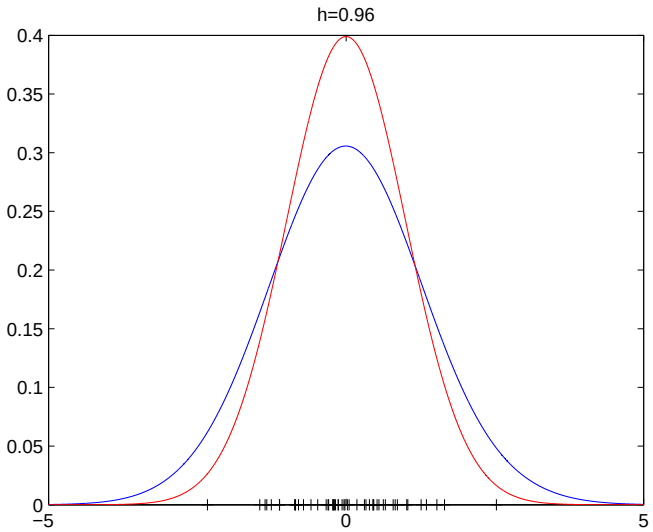
# Estimation de densité



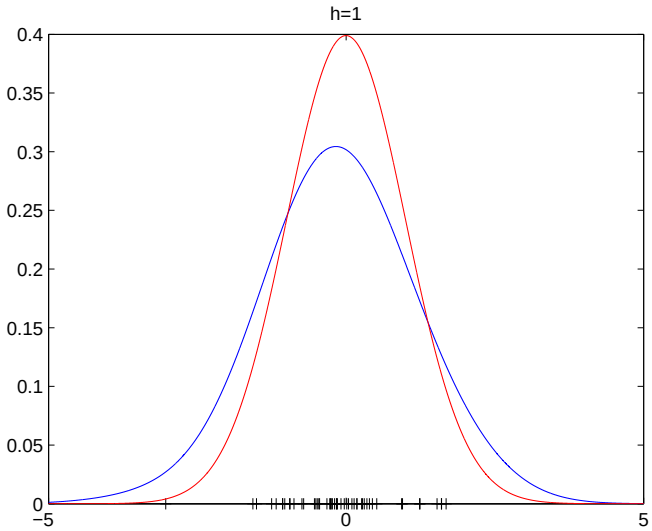
# Estimation de densité



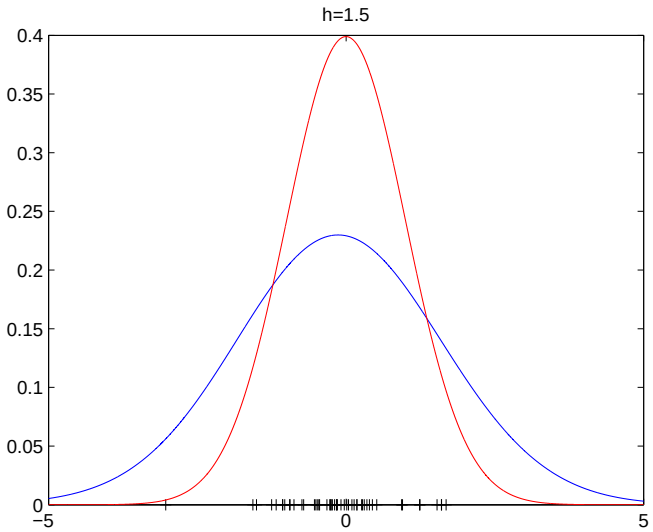
# Estimation de densité



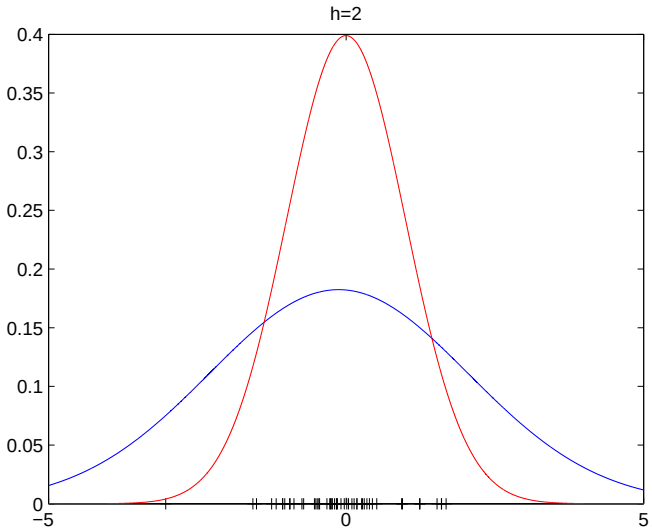
# Estimation de densité



# Estimation de densité



# Estimation de densité



# Définition d'une fonction "score"

On doit choisir une fonction score, en fonction du problème, et savoir si on veut plutot la minimiser ou la maximiser... C'est vague mais ca va devenir clair sur les exemples

# Un score pour la regression polynomiale

Pour un parametre (degré $p$) on estime des coefficients $a_{k,p}$ et une "erreur" $\varepsilon_{i,p}$ (par la méthode des moindres carrés) et le modèle (estimé) est :

$$Y_i = \sum_{k=0}^{p} a_{k,p}(X_i)^k + \varepsilon_{i,p}$$

(Rq attention aux notations puissances)
Un score possible est la somme des erreurs au carrées:

$$S(p) = \sum_i (\varepsilon_{i,p})^2$$

On veut avoir l'erreur la plus faible possible donc minimiser notre score

# Un score pour l'estimation de densité

Pour un parametre $h$ donné on obtient un estimateur de la densité $\hat{f}_h$ (c'est une fonction). On peut choisir (cf cours de stat de L3) le parametre qui maximise la vraisemblance $L$ des observations:

$$L(h) = \prod_{i=1}^{n} \hat{f}_h(X_i)$$

Le score est désormais la vraisemblance (c'est un choix possible et pas le seul) et on va chercher a la maximiser...

# Attention !!!!

Si on désormais on ne fait rien d'autre que de minimiser (ou maximiser) notre score directement on fait ce qui s'appelle du surapprentissage

- Dans le cas de la regression le degré du polynome sera $n-1$ et le score nul
- Dans le cas de l'estimation de densité $h$ sera 0 et la vraisemblance infinie

# note sur le vocabulaire

Ce nom de surappentissage viens de l'intelligence artificielle

# La notion de "généralisation" d'un modèle

Le langage de l'intelligence artificielle et de l'apprentissage.... on veut pouvoir généraliser notre modèle à de nouvelles données

# La validation : Le principe

On rappelle qu'on veut que le modèle selectionné se généralise bien à de nouvelles données... L'idée de la validation est très simple:

- On sépare les données en deux ensembles
  - une base "apprentissage" sur laquelle on apprend les parametres du modele (si besoin est)
  - une base "test" sur laquelle on évalue les performances de généralisation du modèle
- on choisit le parametre qui donne les meilleurs performances **de généralisation**
- on construit alors le modèle final (en corrigeant éventuellement le(s) parametre(s) pour prendre en compte le nouvel effectif) sur l'ensemble des données

# Validation : Exemple de la regression

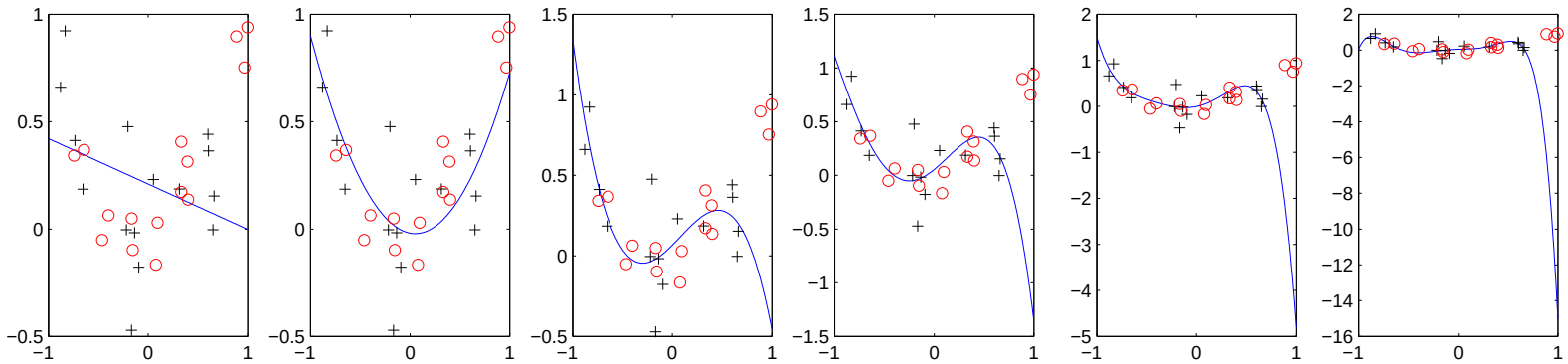


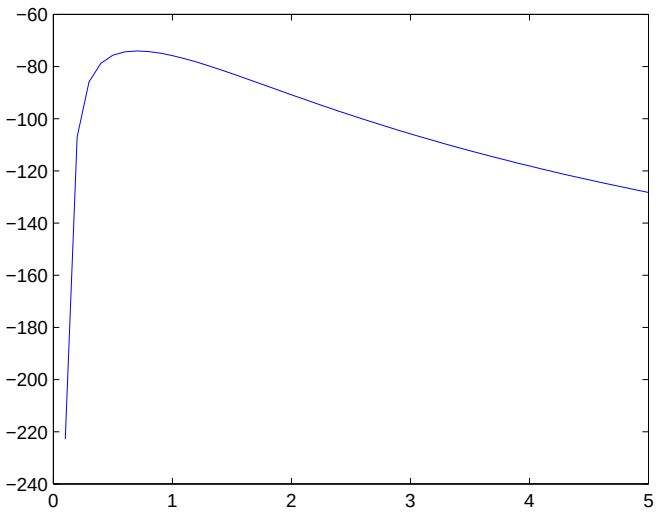
Figure: les erreurs de généralisation sont, 2.8, 0.5, 4.4, 11, 63, 450

On choisit $p = 2$ (ce qui est dans ce cas le "vrai" parametre)
L'étape d'après consiste a prendre toute la base et a estimer les coefficients de regression polynomiale avec $p = 2$.

# Validation : Exemple de l'estimation de densité

- On scinde la base en deux échantillons $X^1 = X_1^1, \ldots, X_{n_1}^1$ (apprentissage) et $X^2 = X_1^2, \ldots, X_{n_2}^2$ (test).
- On estime la densité sur la base apprentissage $\hat{f}_h^1(x) = \frac{1}{n_1 h} \sum K((X_i^1 - x)/h)$
- On évalue le modele en calculant la vraisemblance sur la base de test $L(h) = \prod \hat{f}_h^1(X_i^2)$ (et on cherche le meilleur $h$ appelons le $h_0$)
- On utilise notre connaissance de la théorie !!!: on sait que $h_{opt} = n^{1/5} C(f)$ donc au final on estime la densité sur toutes les données avec $h = (n_1/n)^{1/5} h_0$

# Validation : Exemple de l'estimation de densité



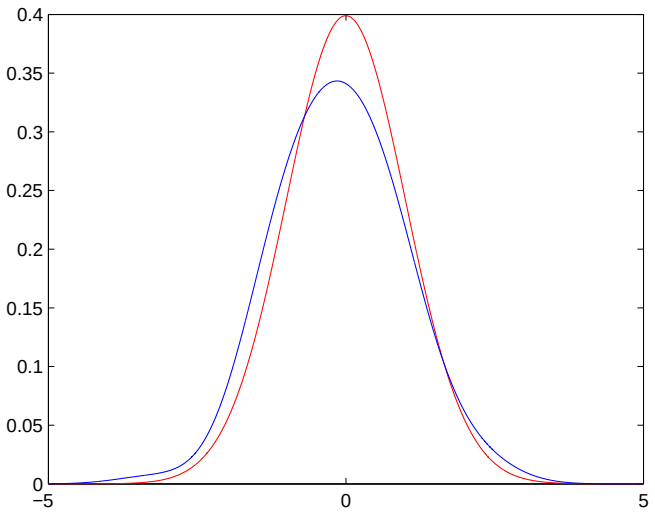




Figure: log de la vraisemblance de test et estimation de densité finale

# Le "Leave one out" : Le principe

On répète $n$ fois:

- apprentissage sur tout le monde sauf l'individu $i$
- mesure du score sur l'individu $i$

et on séléctionne le modèle ayant le "meilleur score global"

# Le "Leave one out" : Exemple de la regression

- appellons $\varepsilon_i^{*,p}$ l'erreur quadratique $Y_i - \hat{f}_p^i(X_i)$ ou $\hat{f}_p^i$ est le polynome de regression de degré $p$ estimé en prenant tout le monde sauf l'individu $i$
- on va chercher $p$ qui minimise $\sum_i (\varepsilon_i^{*,p})^2$.

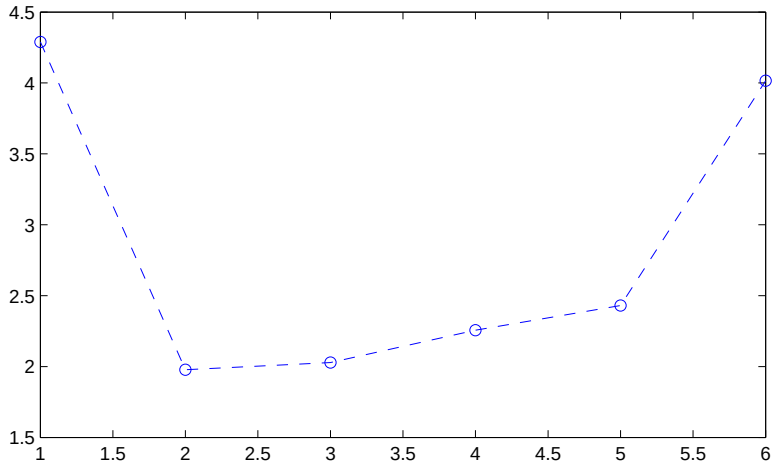



Figure: resultat dans notre exemple numérique

# Le "Leave one out" : Exemple de l'estimation de densité

$\hat{f}_{h,-i}(X_i) = \frac{1}{h(n-1)} \sum_{j \neq i} K((X_i - X_j)/h)$ et $L(h) = \prod \hat{f}_{h,-i}(X_i)$ On cherche a maximiser $L(h)$

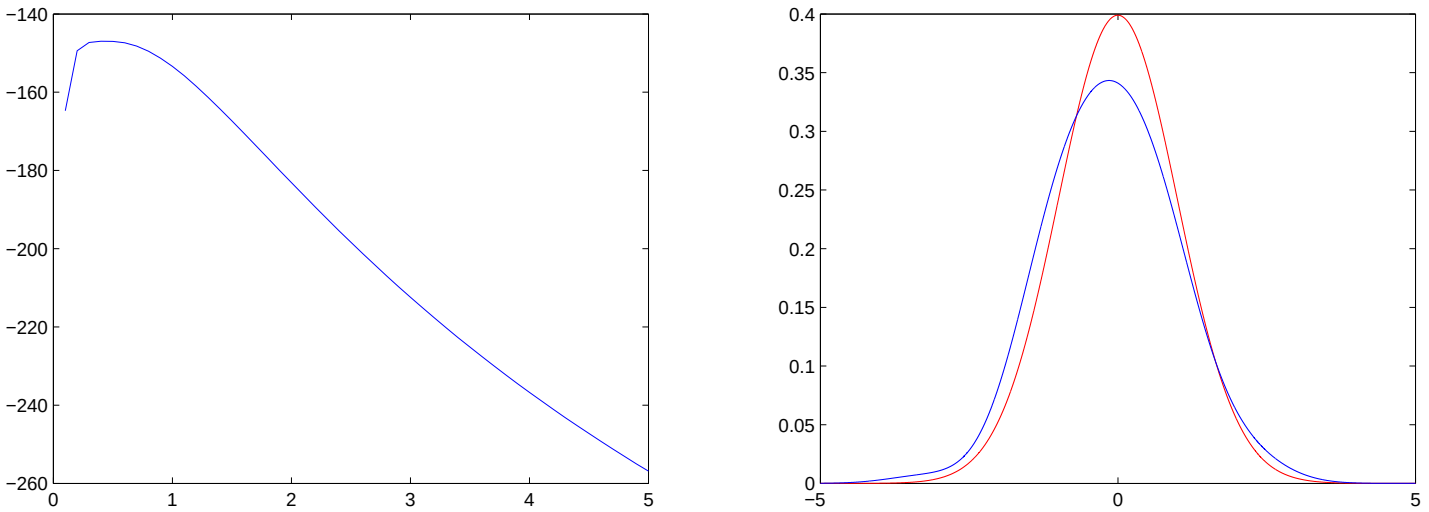



Figure: log de la vraisemblance de test et estimation de densité finale

**Remarque :** ici comme $n$ et $n-1$ ne sont pas tres différents on

# "Moralité"

On a la des "grands principe" qu'il faut savoir adapter a la méthode choisie !!

Celon vous quand prèfère t'on le leave one out ou la validation ? ou encore quels sont les avantages et inconvénients de chacunes de ces méthodes...

# Il existe d'autres méthodes qui reposent sur le même principe

principalement:

- répéter plusieurs validations pour avoir des résultats "moins sensibles"
- Le bootstrap (long et converge vers les résultats de la cross-validation)

# Le principe du controle de complexité

On rappelle qu'on cherche a éviter le "surapprentissage". Le surapprentissage peut etre vue comme le fait qu'on a utilisé un modèle ayant trop de parametres... donc trop complexe. L'idée du controle de complexité est la suivante:

- On travail sur tout l'echantillon et on calcule une erreur $E$
- On mesure une complexité du modele $C$
- On minimise $E + C$

# Le controle de complexité dans nos exemple

- dans la regression polynomiale l'erreur peut etre par exemple l'erreur quadratique et la complexité le degré du polynome
- dans l'estimation de densité le controle de complexité n'a pas vraiment de sens... tous les modèles ont la même complexité... Il existe des moyens d'éluder ce probleme en disant mesurant par exemple la regularité de la densité estimée (on dira qu'une densité est d'autant plus complexe qu'elle est "régulière")

# quelques exemples

- **Mallow :** $E + 2\frac{W}{N}\sigma^2$ ou $E$ est l'erreur quadratique moyenne, $W$ me nombre de parametres du modèle **linéaire**, $N$ la taille de l'échantillon et $\sigma^2$ une estimation de la variance du bruit
- **AIC :** $-2\log(L) + 2W$ ou $L$ est la vraisemblance
- **BIC :** $-2\log(L) + 2W\log(N)$

Vous verrez en détail et en série temporelles comment mettre en oeuvre ces méthodes qui ne sont justifiées théoriquement que dans le cas du modèle linéaire (et donc sont parfaitement adaptées au cours de série temporelles et moins a celui ci)