ESTIMATION NON PARAMÉTRIQUE DES DENSITÉS ET DES RÉGRESSIONS

Michel Delecroix

Les notes qui suivent ont pour but de présenter de façon très succinte une méthodologie déjà longuement répandue dans les centres de recherches en traitements de données, et appelée à devenir un des outils usuels de l'analyse exploratoire ou mathématique des grands ensembles de chiffres, c'est-à-dire une technologie utilisable au même titre que la régression linéaire standard, pour ne citer que cet exemple célèbre, dans l'analyse des enquêtes Insee : les méthodes dites "non paramétriques".

L'intégration de cette méthodologie dans les outils de base des statisticiens appliqués a été rendue possible par un formidable travail théorique effectué sur le sujet, depuis 30 ans, dans les centres de statistique mathématique (le papier fondateur de la méthode du "noyau" date de 1962). Des milliers d'articles ont été écrits sur ce thème, ce qui a permis d'arriver à un optimum pratique des méthodes utilisées.

Pour rendre plus accessible aux non initiés les deux exposés qui termineront la session, premiers témoignages d'une analyse non paramétrique de données Insee, on s'efforcera ici, simplement, d'introduire brièvement deux méthodes de base :

a) L'estimation de la densité commune de variables observées ;

b) L'estimation non paramétrique d'une courbe de régression.

Évidemment, le texte qui suit est très loin de correspondre au minimum "scolaire" jugé usuellement indispensable à une introduction sérieuse à ces sujets ! (20 heures environ...). Les lecteurs désireux de poursuivre cette première prise de contact peuvent se procurer une bibliographie auprès de l'auteur !

I Estimation de la densité

I-1 L'histogramme

Le problème posé est simple : le statisticien dispose des observations, notées $x_i,...x_n$ d'une certaine variable, par exemple la variable "revenus" sur un ensemble de ménages. En un premier temps, on suppose les x_i réels (série unidimensionnelle). La modélisation non paramétrique interprète alors ces chiffres x_i comme des réalisations de variables

aléatoires X_i, se distinguant par là de l'analyse des données dite "à la française" la plus classique, par exemple.

On suppose comme d'habitude (il existe bien sûr des solutions hors de ce cadre) que les X_i sont indépendantes et ont toutes la même loi caractérisée par une densité f. Le problème est de trouver f. Beaucoup de statisticiens admettent alors à ce niveau par routine, manque d'imagination, de logiciels ou de formation parfois, que cette densité appartient à une famille de lois connues avec un degré de variabilité dû à un paramètre qu'ils s'empressent d'estimer.

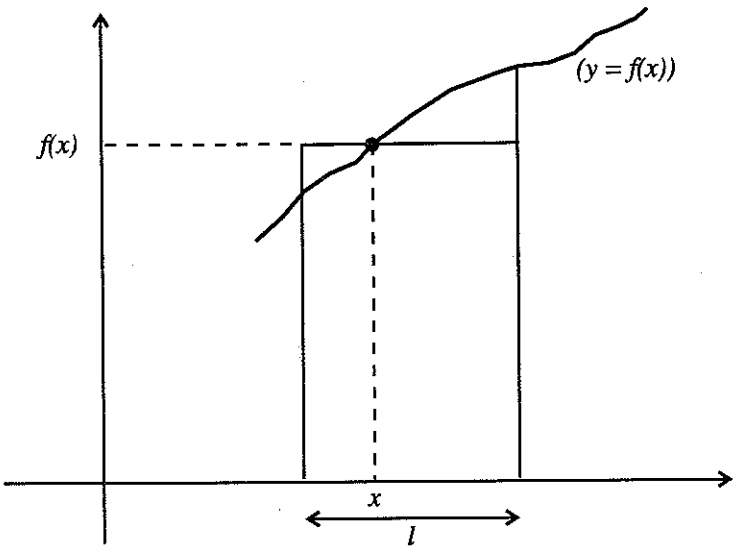
La statistique non paramétrique permet de s'affranchir de cette hypothèse en estimant directement la densité f en chaque point x. Ceci requiert évidemment un nombre suffisant d'observations bien réparties au long du domaine de valeurs des X_i, mais donne une plus grande variété de résultats possibles, qui "collent" bien aux données.

Comment faire en pratique ? On utilise deux idées très simples, l'une mathématique, l'autre statistique.

Pour que l'approximation mathématique utilisée marche, supposons f continue en x. Alors on a, si I est un intervalle de largeur l contenant x, et que l est assez petit :

$$f(x) \# \frac{1}{l} \int_I f(t)\,dt$$

C'est une conséquence du bon vieux théorème de la moyenne : la quantité de droite tend vers $f(x)$ si l tend vers 0, et cela s'interprète bien graphiquement :



Maintenant un vieux théorème de calcul des probabilités (la "loi des grands nombres") nous dit que si le nombre de variables observées est assez grand, on a, pour tout sous-ensemble A de $\mathbb{R}$

$$\frac{\textit{Nombre de } X_j \textit{, tels que } X_j \textit{ est dans } A}{n} \# P[X_i \in A]$$

C'est la classique interprétation "fréquentiste" des probabilités, longtemps définies comme la limite des fréquences dans la répétition d'épreuves indépendantes.

On a enfin (définition d'une densité) :

$$P[X_i \in A] = \int_A f(t)\, dt$$

ce que tout le monde utilise en écrivant, par exemple :

$$0{,}95 = \int_{-1{,}96}^{1{,}96} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{t^2}{2}}\, dt = P(|Y| < 1{,}96),$$

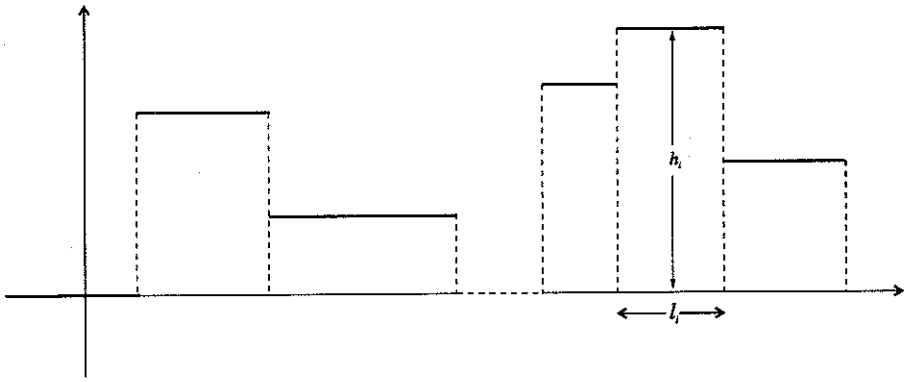
lorsque $Y \sim N(0,1)$

Alors en combinant les 2 approximations précédentes, on obtient pour n grand et l petit :

$$l \,.\, f(x) \# P[X_j \in I] \# \left\{ \frac{\textit{Nombre de } X_j \textit{ tels que } X_j \textit{ appartient à } I}{n} \right\}$$

Considérons alors un histogramme basé sur des classes $c_1, \ldots c_k$ de largeurs $l_1, \ldots l_k$ supposées assez petites. Si nous appelons $\hat{f}$ la fonction en escalier dont le graphe est constitué par les côtés supérieurs des rectangles, elle est définie, lorsque x est élément de la i^e classe, par définition de l'histogramme (surfaces des rectangles égales à leurs fréquences observées) par :

$$\hat{f}(x) = \frac{1}{l_i} \left\{ \frac{\textit{Nombre de } X_j \textit{ appartenant à } c_i}{n} \right\}$$



En effet, on a :

$$\begin{cases} (x \in C_i) \Rightarrow (\hat{f}(x) = h_i) \\ h_i \,.\, l_i = \dfrac{1}{n} \left| \textit{Nombre de } X_j \textit{ dans } C_i \right| \end{cases} \quad \text{(voir Fig.)}$$

La conclusion est que finalement "L'histogramme est un estimateur de la densité" si n est assez grand, l'échantillon bien réparti, les l_i assez petits. On a :

$$\hat{f}(x) \# f(x)$$

pour chacun des x repris dans une des classes de l'histogramme.

I-2 Fenêtre mobile et noyau

L'histogramme $\hat{f}$ redéfini ci-dessus est évidemment très sensible au choix des classes. En particulier en déplaçant légèrement l'extrémité d'une classe on peut beaucoup modifier $\hat{f}$ en un point x qui passerait d'une classe à fort effectif à une classe à faible effectif. Aussi a-t-on modifié une première fois l'estimateur en définissant l'estimateur de la "fenêtre mobile".

On se base sur les mêmes principes, mathématique et statistique, mais en calculant ici $\hat{f}(x)$ point par point. Pour un point x donné on peut en effet écrire, pour h assez petit ,

$$\begin{cases} 2h \,.\, f(x) \;\; \# \;\; \displaystyle\int_{x-h}^{x+h} f(t)\,dt \\ \displaystyle\int_{x-h}^{x+h} f(t)\,dt = P\,[X_j \in [x-h, x+h]] \;\; \# \;\; \frac{1}{n}\{\textit{Nombre de } X_j \textit{ dans } [x-h, x+h]\} \end{cases}$$

La seule idée est en quelque sorte de "centrer" la classe sur x. On obtient :

$$\hat{f}(x) = \frac{1}{2n \,.\, h}\,\{\textit{Nombre de } X_j \textit{ dans } [\,x-h, x+h\,]\}$$

Bien sûr le résultat au point x serait le même à partir d'un histogramme global qui contiendrait la classe $[x - h, x + h]$, mais l'estimateur de la fenêtre mobile se calcule point par point, et non sur des classes fixées *a priori*.

Au demeurant cet estimateur amène toujours à approcher f continue (base de la méthode) par une fonction qui ne peut prendre qu'un nombre fini de valeurs : le nombre de X_j dans $[x - h, x + h]$ vaut *0, 1, 2,...* au maximum n. Le résultat est toujours en escalier. Pour remédier à cet inconvénient, on va introduire une classe plus générale d'estimateurs dits "à noyau". L'idée est simple. On constate d'abord que $\hat{f}(x)$ peut s'écrire :

$$\frac{1}{n} \,.\, \frac{1}{h} \sum_{1}^{n} K_0\left(\frac{x - X_i}{h}\right)$$

où K_0 représente la fonction définie par :

$$K_0(x) = \frac{1}{2} I_{[-1,1]}(x)$$

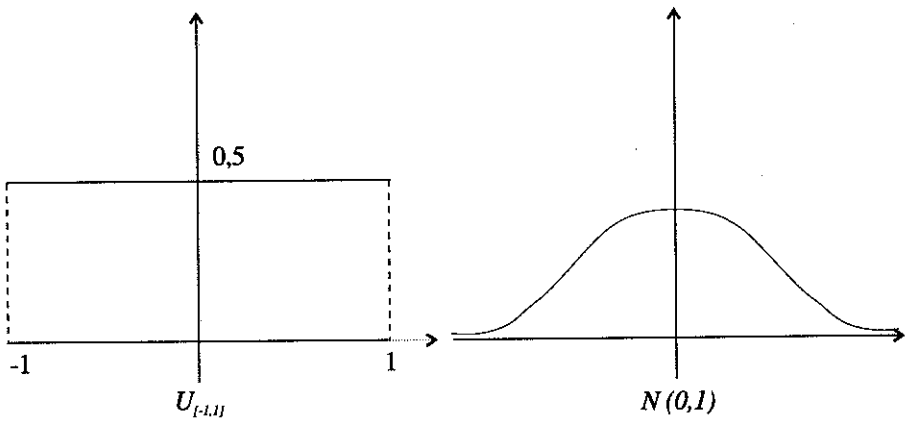
En effet :

$$X_i \in [x-h, x+h] \iff -1 \le \frac{X_i - x}{h} \le 1$$

$$\iff I_{[-1,1]}\left(\frac{x - X_i}{h}\right) = 1$$

avec la notation $I_A(x) = 1$ si x est dans A, 0 sinon. Dès lors le nombre de X_j apparaissant dans $[x - h, x + h]$est le nombre de fois où l'indicatrice vaut *1* dans la somme de ces indicatrices, les autres valant zéro ! Cela donne la formule en intégrant $\frac{1}{2}$ dans K_0.

Maintenant K_0 est une densité de probabilité, celle de la loi uniforme sur [-1,1]. Son rôle est de séparer les X_i de l'échantillon en ne gardant que ceux qui sont proches de x (X_i dans $[x - h, x + h]$). Pour remédier au caractère trop brutal de cette dichotomie, Parzen et Rosenblatt ont alors eu l'idée de remplacer K_0 par une densité de probabilité K continue et bornée, tendant assez vite vers 0 à l'infini : l'idée instinctive est celle d'une loi Gaussienne centrée réduite :



On voit que si on substitue K_0 par une fonction de ce type, on obtiendra une somme où les valeurs correspondant à des X_i loin de x sont infinitésimales, les autres de plus en plus fortes avec la proximité de x et X_i : on maintient l'effet initial cherché, mais en le "lissant". On définit ainsi une classe générale d'estimateurs de f en posant

$$\hat{f}(x) = \frac{1}{nh} \cdot \sum_{1}^{n} K\left(\frac{x - X_i}{h}\right)$$

K s'appelle un "noyau". L'intérêt de ses propriétés évoquées ci-dessus est qu'elles "passent " sur $\hat{f}$. Considérée comme fonction de x, $\hat{f}$ est elle-même une densité continue bornée, tendant vers 0 à l'infini c'est-à-dire "quelque chose" *a priori* plus ressemblant à f qu'un histogramme ou une fenêtre mobile !

Pour mettre en œuvre pratiquement la méthode il faut évidemment choisir, si possible en tenant compte des données, le noyau K et la "meilleure fenêtre" h. Le paragraphe qui suit donne quelques éléments de réponse à cette question fondamentale. Pour terminer celui-ci notons simplement que la méthode présentée se généralise directement au cas de données bivariées en multivariées.

I-3 Mise en œuvre pratique : K et h

Comment au départ décider qu'un estimateur construit à l'aide de K et h est "bon", de façon à guider notre choix ? On peut apporter deux types de réponses :

a) L'usage de simulations

On prend une loi de densité connue, qu'on sait tracer, on simule N échantillons de taille n de la loi. On estime f par la méthode précédente avec un choix particulier de K et h, et on regarde si en moyenne sur les N échantillons $\hat{f}$ est proche de f. En faisant varier K et h on essaie de juger de leur impact.

En fait on s'aperçoit que dès que $n = 100$, on reconstitue bien, avec $\hat{f}$, les densités usuelles. De plus on constate, ce qui sera confirmé par le 1er exposé qui suit, que le choix de K a, en première approximation, peu d'impact au niveau des résultats alors que celui de h est très sensible : les courbes représentatives de $\hat{f}$ sont très irrégulières si h est très petit et très plates quand h grandit.

Pour une fenêtre mobile, l'explication est claire : si h est très petit, l'approximation mathématique de f est fine, mais l'approximation statistique sera, elle, mauvaise : dans $[x - h, x + h]$ il y aura très peu de X_i (probabilité très faible d'y tomber), d'où l'irrégularité. Le principe inverse joue si h est grand. Pour mieux préciser cette idée et résoudre le problème du meilleur choix de h, à K fixe, il nous faut alors recourir au calcul.

b) Évaluation de l'écart

Pour mesurer l'écart entre $\hat{f}$ et f, on calcule en fait (choix *a priori*, non unique) la quantité :

$$\int_{-\infty}^{+\infty} \left(\hat{f}(x) - f(x)\right)^2 dx$$

En remplaçant $\hat{f}$ par sa valeur, on peut développer le calcul : le résultat dépend de façon compliquée de K, h, f (inconnue) et $X_1, ..., X_n$ qui déterminent $\hat{f}$.

Sur l'ensemble des échantillons $X_1, ..., X_n$ qu'on peut tirer, on peut calculer la moyenne théorique des résultats c'est-à-dire :

$$E\int_{-\infty}^{+\infty} \left(\hat{f}(x) - f(x)\right)^2 dx$$

Tous calculs faits on trouve : $\frac{A}{n \cdot h} + B\,h^4$ + (termes en h^5, h^6, ...)

On retrouve bien le fait qu'il faut prendre h petit (2e terme) mais pas trop (premier).

En négligeant devant h^4 les termes : h^5, h^6, ... on voit que h^*, minimisant en moyenne l'écart entre $\hat{f}$ et f, sera celui qui minimise

$$\frac{A}{n \cdot h} + B\,h^4$$

Un calcul élémentaire montre alors que :

$$h^* = \frac{C}{n^{1/5}}$$

C'est *a priori* logique ; on relie h à la taille de l'échantillon : plus n est grand, plus h doit être petit, mais sans trop l'être (ne pas prendre k/n ou k/n^2). En pratique, cependant, on bute sur la valeur de C, essentielle : le choix de $\frac{C}{n^{1/5}}$ peut passer de 0 à l'infini selon C ! En fait C a une valeur déterminée à partir de A et B : on peut tout y calculer sauf la quantité $\int_R (f''(x))^2\,dx$ évidemment inconnue, et qui apparaît au dénominateur. Plus cette quantité est faible, plus f est régulière, plate, moins il est nécessaire de capter les effets locaux et plus h^* peut être grand.

Que faire en pratique ? La solution retenue (c'est un des choix possibles) dans le premier papier est la suivante : on suppose que f appartient à une famille paramétrique de lois de densités $g_{\hat{\theta}}$. On cherche la densité ajustant le mieux f, $g_{\hat{\theta}}$ (même si elle en est loin...) et on calcule alors :

$$\int_R (g''_{\hat{\theta}})^2(x)\,dx$$

qu'on injecte dans C donc dans h^*. En pratique, ça ne marche pas mal (cf. simulations). Tous calculs faits, on obtient, avec K Gaussien :

$$\frac{S}{n^{1/5}}$$

où S est l'écart type de l'échantillon $x_1, \ldots x_n$, ce qui est très satisfaisant.

Finalement une autre approche est possible : évaluer h^* correspondant à l'échantillon observé en maximisant en h des quantités dépendant de cet échantillon et estimant l'écart entre $\hat{f}$ et f. Nous verrons cette méthode dite de "validation croisée" dans le cadre (plus facile sur ce point) de l'estimation d'une régression. Notons l'intérêt pratique de cette méthode : les programmes de calcul de h^* sont alors inclus dans les logiciels de détermination de $\hat{f}$, ce qui exclut tout effort.

II Estimation de la régression

On dispose ici de n observations (X_i, Y_i), supposées encore être des réalisations de vecteurs aléatoires indépendants de dimension 2 : Y_i est la variable qu'on cherche à expliquer par la valeur qu'a prise la variable X_i. L'hypothèse de base est que (X_i, Y_i) est un vecteur aléatoire de $\mathbb{R}^2$, de densité f.

Le "paramètre" fonctionnel du modèle qu'on cherche à estimer est la fonction m de régression

$$m : \mathbb{R}^2 \rightarrow \mathbb{R}$$

définie mathématiquement par le fait que $m(X_i)$ est la variable aléatoire fonction de X_i qui approxime le mieux Y_i en moyenne :

$$m = \arg \min E(Y_i - h(X_i))^2$$

On n'impose **pas** d'hypothèse paramétrique ni sur la loi de (X_i, Y_i), ni sur la loi conditionnelle, ni sur la forme de m, c'est-à-dire que pour être clair, on n'écrit **pas** *a priori* par exemple $m(x) = \theta.x$ (régression linéaire).

Toute l'étude sera faite avec une seule variable explicative X_i réelle, mais la méthode existe dans le cas d'un vecteur X_i de covariables de dimension supérieure ou égale à 2.

L'estimation de m se fera ici encore "point par point". On suppose au départ m continue, simplement. La conséquence directe de cette hypothèse est la suivante : si h est petit on a :

$$z \in [x-h, x+h] \Rightarrow (m(z) \# m(x)$$

ceci revient à dire que, sur $[x - h, x + h]$, m peut s'approximer par une fonction constante égale à $m(x)$. On pourrait d'ailleurs ici encore envisager l'approximation **globale** de f par une fonction en escalier dont $[x - h, x + h]$ serait une classe (principe des sommes de Riemann).

Soit alors J l'ensemble des indices des variables X_i appartenant à $[x - h, x + h]$. Avec les notations du **I- 1** on a :

$$i \in J \iff X_i \in [x-h, x+h] \iff K_0\left(\frac{x-X_i}{h}\right) = 1$$

Maintenant si $i \in J$, et si m est effectivement égale à une constante k sur $[x - h, x + h]$, on a :

$$Y_i = k + \varepsilon_i$$

D'où l'idée d'estimer k, par "moindres carrés locaux" c'est-à-dire de définir

$$\hat{k} = \arg \min \left\{ \sum_{i \in J} (Y_i - k)^2 \right\}$$

C'est logique : on cherche la fonction constante égale à k qui interpole le mieux le nuage des (X_i, Y_i) dans le voisinage $[x - h, x + h]$ de x ! $\hat{k}$ ne sera alors autre (calcul classique) que la moyenne arithmétique des Y_i tels que $X_i \in [x-h, x+h]$ c'est-à-dire :

$$\hat{k} = \frac{1}{Card\ J} \left(\sum_{i \in J} Y_i \right)$$

Avec les notations déjà introduites cela donne :

$$\begin{cases} \hat{k} = arg \min_k \sum_{i=1}^{n} (Y_i - k)^2 \ K_0\left(\frac{x-X_i}{h}\right) \\ \hat{k} = \sum_{1}^{n} Y_i \,.\, K_0\left(\frac{x-X_i}{h}\right) \Big/ \sum_{1}^{n} K_0\left(\frac{x-X_i}{h}\right) \end{cases}$$

Le tour est joué : on a vu que $m(x)$ peut être approchée par une constante k sur $[x - h, x + h]$, et on peut aussi dans le calcul précédent, remplacer K_0 par une autre densité K bornée tendant vers 0 à l'infini (voir 1er partie). Cela définit les estimateurs dits de Nadaraya-Watson, pour lesquels $\hat{m}(x)$ vaut :

$$\hat{m}(x) = \sum_{1}^{n} Y_i K\left(\frac{x-X_i}{h}\right) \Big/ \sum_{1}^{n} K\left(\frac{x-X_i}{h}\right)$$

En fait on peut généraliser le procédé et utiliser des approximations plus fines de m dans le voisinage $[x - h, x + h]$ de x. On a utilisé ci-dessus le fait que, si m est continue :

$$z \in [x-h, x-h] \Rightarrow m(z) \# m(x)$$

Il est tout aussi vrai que, si m est 2 fois différentiable

$$z \in [x-h, x-h] \Rightarrow m(z) \# m(x)+(z-x)\ m'(x)$$

ou encore $$m(z) \# m(x)+(z-x)\ m'(x) + \frac{(z-x)^2}{2}\ m''(x)$$

C'est le principe même des développements limités. Adopter la formule précédente, c'est écrire que sur $[x - h, x + h]$ on a pour certains coefficients a, b, c (qui dépendent de x, et changent avec le point).

$$m(z) \simeq a+b(z-x) + \frac{c}{2}\cdot(z-x)^2$$

On peut alors estimer ces coefficients par moindres carrés comme ci-dessus en écrivant que a, b, c doivent minimiser :

$$\sum_{i \in J} \left\{Y_i - [a+b(X_i - x) + \frac{c}{2}(X_i-x)^2]\right\}^2$$

On réalise un ajustement polynominal de degré 2 de m sur le nuage réduit aux (X_i, Y_i) tels que $X_i \in [x-h, x+h]$. Clairement :

$$\begin{pmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \end{pmatrix} \quad \text{estiment} \quad \begin{pmatrix} m(x) \\ m'(x) \\ m''(x) \end{pmatrix}$$

Formellement cela revient à écrire :

$$(\hat{m}(x), \hat{m}'(x), \hat{m}''(x)) = arg \min_{a,b,c} \sum_{i=1}^{n} \left\{Y_i - [a+b(X_i-x)+\frac{c}{2}(X_i-x)^2]\right\}^2 K_0\left(\frac{x-X_i}{h}\right)$$

et on remplacera K_0 par un noyau plus lisse (densité Gaussienne par exemple). C'est la méthode dite des **polynômes locaux** dont l'estimateur de Nadaraya-Watson est un

cas particulier (polynôme de degré 0) et qui peut être étendu à un degré polynomial quelconque selon le nombre de dérivées qu'ont veut estimer (et qu'on suppose exister). Si on dispose d'une formule explicite pour l'estimateur de Nadaraya-Watson, les formules donnant $\hat{a}, \hat{b}, \hat{c}$ dans le cas général sont beaucoup plus lourdes à écrire. Cependant la mise en œuvre pratique de la méthode ne pose aucun problème pratique : tous les programmes de calcul de M.C. pondérés peuvent servir à calculer $\hat{a}, \hat{b}, \hat{c}$ une fois choisis K et h. Cette méthode a été utilisée dans le 2e exposé qui suit à l'ordre 0 et à l'ordre 2.

L'importance du choix de h est ici encore très grande : pour $K = K_0$, si h est trop grand, on a toujours $K_0\left(\frac{x - x_i}{h}\right) = 1$ et on ajuste **globalement** m par un polynôme. En pratique, pour déterminer h, on peut adopter une méthode basée sur le calcul. Mais dans le 2e exposé qui suit on a en fait retenu l'approche "validation croisée" déjà évoquée et facile à justifier ici : pour un h donné on peut calculer ; pour tout x, un estimateur $\hat{m}_h^{-i}(x)$ en utilisant tout l'échantillon observé sauf X_i, à partir d'un noyau K et d'une fenêtre h. Alors $\hat{m}_h^{-i}(X_i)$ est un estimateur naturel de $m(Xi)$.

Dès lors, la somme des "résidus" du modèle est

$$\sum_{1}^{n} \left(Y_i - \hat{m}_h^{-i}(X_i)\right)^2$$

et ne dépend plus que de h, à K fixé. Il est alors naturel de choisir pour h la valeur h^* qui minimise en h l'expression précédente, c'est-à-dire telle que l'on "ajuste" au mieux le modèle ! Le programme de calcul détermine d'ailleurs automatiquement cette fenêtre optimale et l'injecte dans le calcul de $\hat{m}$, si on le désire. C'est ce qui a été utilisé dans les calculs exposés dans le 2e article présenté.

Cas des données bivariées : estimation d'une densité

Les données X_i sont ici des couples de réels : $Xi = (X_{i1}, X_{i2})$ (ex : des couples (âge-revenu)), réalisations de variables indépendantes de densité f :

$$P[(X_{i1}, X_{i2}) \in A] = \int\int_A f(x_1, x_2)\, dx_1\, dx_2$$

Comme dans le cas unidimensionnel, si n est grand on a, du point de vue statistique :

$$P[(X_{i1}, X_{i2}) \in A] \# \frac{1}{n} \{ Nombrede\ (X_{i1}, X_{i2})\ appartenant\ à\ A \}$$

Du point de vue mathématique, si h_1 et h_2 sont petits, f continue, on a :

$$f(x_1, x_2) \# \frac{1}{2h_1} \cdot \frac{1}{2h_2} \cdot \int_{x_1-h_1}^{x_1+h_1} \int_{x_2-h_2}^{x_2+h_2} f(u, v)\, du\, dv$$

À partir des 3 formules ci-dessus on obtient l'estimateur de base :

$$\hat{f}(x_1, x_2) = \frac{1}{2h_1} \cdot \frac{1}{2h_2} \cdot \frac{1}{n} \{ Nombrede\ (X_{i1}, X_{i2})\ dans\ [x_1 - h_1, x_1 + h_1] \times [x_2 - h_2, x_2 + h_2] \}$$

Il est facile de vérifier, comme dans le cas univarié, que le nombre de (X_{i1}, X_{i2}) dans $[x_1 - h_1, x_1 + h_1] \times [x_2 - h_2, x_2 + h_2]$ vaut :

$$\sum_{i=1}^{n} I_{[-1,1]} \left(\frac{x_1 - X_{i1}}{h_1} \right) I_{[-1,1]} \left(\frac{x_2 - X_{i2}}{h_2} \right)$$

donc

$$\hat{f}(x_1, x_2) = \frac{1}{n} \cdot \sum_{i=1}^{n} \left\{ \frac{1}{h_1} \cdot K_0 \left(\frac{x_1 - X_{i1}}{h_2} \right) \cdot \right\} \left\{ \frac{1}{h_1} \cdot K_0 \left(\frac{x_2 - X_{i2}}{h_2} \right) \right\}$$

$$\left((K_0(x) = \frac{1}{2} I_{[-1,1]}(x)) \right)$$

On lisse alors l'estimateur précédent en remplaçant K_0 par un noyau de Parzen-Rosenblatt K dans la formule précédente. Il est susceptible de diverses généralisations, et rien n'interdit de prendre $h_1 = h_2 = h$ dans la formule (mais rien n'y oblige). Finalement dans le cas de données univariées ou bivariées, le problème qui se pose en pratique est celui de "choisir" K et la (ou les) fenêtres h_1, h_2, h au mieux pour calculer concrètement l'estimateur.

On peut montrer cette fois que :

$$E\left[\int\int_{R^2} (\hat{f}\,(x_1, x_2)) - f((x_1, x_2)))^2\, dx_1\, dx_2\right] \# \frac{A}{nh_1 h_2} + B\,(h_1^2 + h_2^2)^2$$

et pour les mêmes raisons que ci-dessus on est conduit à :

$$h_i^* = \frac{S_i}{n^{1/6}}$$

S_i représentant l'écart-type des X_{ji} ($i = 1$ ou 2 selon les composantes).