Quelques modèles de régression

Charles-Henri Kempeners

11 octobre 2010

Directeur de mémoire de M2 : Emmanuel Bacry
Directeur de stage au CREST : Alexandre Tsybakov

Résumé

Ce mémoire présente quelques méthodes statistiques de régression. Les approches introduites ici généralisent la régression linéaire des moindres carrés afin de l'adapter à des problèmes d'apprentissage et de prédiction plus complexes. Ces techniques jouent un rôle important dans de nombreux domaines tels que la biologie, la finance ou encore le traitement du signal.

Table des matières

Introduction

Soit $(X_1, ..., X_n)$ et $(Y_1, ..., Y_n)$ deux n-échantillons des variables réelles ou vectorielles X et Y, nous cherchons à construire le meilleur estimateur $\hat{f}$ de la fonction de régression f : $x \mapsto \mathbb{E}[Y|X = x]$ au sens d'une certaine fonction de perte. X constitue la variable d'entrée ou variable explicative et Y la variable de sortie ou variable expliquée. De nombreuses fonctions de pertes peuvent être utilisées. Citons par exemple l'erreur intégrée en moyenne quadratique $(MISE)$:

$$MISE(x) = \mathbb{E}[\int (f - \hat{f})^2] = \int b^2 + \int \sigma^2$$

où $b(x) = \mathbb{E}[\hat{f}(x)] - f(x)$ et $\sigma^2(x) = \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2]$. La décomposition précédente fait apparaître un terme de biais et un terme de variance. Tous les deux doivent être minimiser conjointement et constitue ce que l'on appelle l'équilibre biais-variance.

Le modèle de régression le plus connu et également le plus simple est l'estimateur linéaire des moindre carrés : la variable Y est expliquée par des combinaisons linéaires des composantes de X. Un tel modèle ne laisse que très peu de souplesse dans l'estimation de la fonction de régression. Les modèles réels sont généralement plus complexes et souvent non-linéaires. Il est donc nécessaire de développer des techniques statistiques pouvant reproduire cette complexité naturelle des données observées.

1 Lissage par noyau

Une classe de méthode de régression non-paramétrique consiste à effectuer au voisinage de chaque point une régression locale, celle-ci pouvant être linéaire (Nadaraya-Watson) ou polynomiale (polynômes locaux). La localisation est assurée par une fonction de poids K, appelée noyau, qui vérifie les deux conditions suivantes

$$K \geq 0 \text{ et } \int K = 1.$$

La flexibilité permise dans la détermination de la fonction de régression donne son intérêt à ce type de méthode.

1.1 Motivations

L' approche locale décrite précédemment peut être ammenée de plusieurs façons.

Supposons que nous observions un n-échantillon $(X_1, ..., X_n)$ d'une variable aléatoire X de densité g par rapport à la mesure de Lebesgue. Un estimateur naïf de la fonction de répartition G peut être donné par

$$\hat{G}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq n\}}.$$

On en déduit un estimateur de la densité en choisissant une fenêtre $h > 0$ et en posant

$$\hat{g}(x) = \frac{\hat{G}(x+h) - \hat{G}(x-h)}{2h} = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{X_i - x}{h})$$

où $K : t \mapsto \frac{1}{2}\mathbb{1}_{\{-1<t\leq 1\}}$ est le noyau rectangulaire.
Choisir un noyau K plus régulier peut permettre par exemple d'obtenir une densité plus lisse. Notons que la fenêtre h doit dépendre de n pour assurer la convergence de $\hat{g}$ vers g au sens $\mathbb{L}^2$.
En particulier h doit vérifier $nh_n \underset{n\to\infty}{\to} \infty$ et $h_n \underset{n\to\infty}{\to} 0$ afin que respectivement la variance et le biais tendent vers zéro.
Choisissons à présent un noyau K d'ordre au moins 1 (*i.e.* : $\int xK(x)dx = 0$).

$$\hat{f}(x) = \hat{\mathbb{E}}[Y|X = x] = \frac{\int y\hat{g}(x,y)dy}{\int \hat{g}(x,y)dy} \mathbb{1}_{\{\int \hat{g}(x,y)dy \neq 0\}} = \sum_{i=1}^{n} Y_i \frac{K(\frac{X_i - x}{h})}{\sum_{j=1}^{n} K(\frac{X_j - x}{h})} \mathbb{1}_{\{\sum_{j=1}^{n} K(\frac{X_j - x}{h}) \neq 0\}}$$

constitue un estimateur de la fonction de régression f, et est connu sous les nom d'estimateur de Nadaraya-Watson.

Cet estimateur peut également être introduit par généralisation de la méthode des K plus proches voisins. Supposons que l'on souhaite donner plus de souplesse à la régression linéaire simple, nous pouvons choisir d'effectuer une régression locale en tout point, en ne tenant compte que du K-voisinage immédiat. Cependant, plutôt que de donner un poids égal à chacun des points du voisinage, il peut être souhaitable de leur assigner des poids qui décroissent de manière lisse avec la distance. Si ces poids sont donnés par les valeurs d'un noyau, on retrouve l'expression de l'estimateur de Nadaraya-Watson.

1.2 Estimateur de Nadaraya-Watson

La discussion précédente se résume par la définition qui suit :

Définition. (estimateur de Nadaraya-Watson)
Soit $K : \mathbb{R} \to \mathbb{R}$ un noyau (i.e. $K \geq 0$ et $\int K = 1$), $h > 0$ une fenêtre, l'estimateur de Nadaraya-Watson de la régression de Y sur X est donné par

$$\hat{f}(x) = \sum_{i=1}^{n} Y_i W_{n,i}(x)$$

où $W_{n,i}(x) = \frac{K(\frac{X_i - x}{h})}{\sum_{j=1}^{n} K(\frac{X_j - x}{h})} \mathbb{1}_{\{\sum_{j=1}^{n} K(\frac{X_j - x}{h}) \neq 0\}}$

La fenêtre h est le paramètre d'intérêt du modèle. Elle contrôle la convergence de l'estimateur ainsi que l'équilibre biais-variance.
Une grande fenêtre s'accompagne d'une faible variance (la moyenne étant effectuée sur plus d'observations) mais d'un biais plus important (la fonction étant presque constante) et réciproquement.
Il n'est pas possible de déterminer directement la fenêtre h qui minimise l'erreur $\mathbb{L}^2$, celle-ci dépendrait en effet de la régularité de la vraie fonction f qui est inobservée. Plus formellement une telle fenêtre définit un oracle, qui peut être considéré comme un estimateur théorique optimal sur chaque classe de régularité de la fonction f. Fournir des estimateurs adaptatifs, c'est-à-dire trouver une méthode de sélection de la fenêtre h qui permet de converger "aussi vite" que l'oracle, est un des objectifs dans l'étude d'un modèle statistique.
Il est usuel de choisir pour méthode de sélection de h, la technique de validation croisée : il s'agit de minimiser sur h l'estimateur suivant de l'erreur $\mathbb{L}^2$

$$CV(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}^{(-i)}(X_i))^2$$

où $\hat{f}^{(-i)}$ est l 'estimateur de la régression qui ne tient pas compte de l' observation i.

1.3 Polynômes locaux

Se limiter à l'estimation linéaire de $\hat{f}$ (qui correspond à un développement de Taylor à l'ordre 1 autour de chaque point) peut s'avérer trop contraignant. Une généralisation naturelle de l'estimateur précédent consiste à effectuer des régressions localement polynômiales.

Définition. (estimateur par polynômes locaux)
Soit $K : \mathbb{R} \to \mathbb{R}$ un noyau, $h > 0$ une fenêtre et $l \geq 0$ un entier.
Le vecteur

$$\hat{\theta}_n(x) = \underset{\theta \in \mathbb{R}^{l+1}}{argmin} \sum_{i=1}^{n} (Y_i - \theta^T U(\frac{X_i - x}{h}))^2 K(\frac{X_i - x}{h})$$

où

$$U(u) = (1, u, \frac{u^2}{2}, ..., \frac{u^l}{l!})^T$$

est appelé estimateur localement polynomial d'ordre l de

$$\theta(n) = (f(x), f^{'}(x)h, f^{''}(x)h^2, ..., f^{(l)}(x)h^l)^T$$

. La statistique $\hat{f}_n(x) = U^T(x)\hat{\theta}_n(x)$ est appelé estimateur localement polynomial d'ordre l de $f(x)$.

Posons à présent $B_{n,x} = \frac{1}{nh}\sum_{i=1}^{n} U(\frac{X_i - x}{h})U^T(\frac{X_i - x}{h})K(\frac{X_i - x}{h})$.

Si $B_{n,x}$ est définie positive alors l'estimateur localement polynomial $\hat{f}_n(x)$ de $f(x)$ est un estimateur linéaire :

$$\hat{f}_n(x) = \sum_{i=1}^{n} Y_i W^*_{n,i}(x)$$

où $W^*_{n,i}(x) = \frac{1}{nh}U^T(0)B^{-1}_{n,x}U(\frac{X_i - x}{h})K(\frac{X_i - x}{h})$.

Les estimateurs de Nadaraya-Watson et par polynômes locaux peuvent être étendus directement à $\mathbb{R}^p$. Cependant, la vitesse de convergence se dégrade très rapidement avec la dimension. Cette propriété est connue sous le nom de fléau de la dimension.

2 Estimateurs par projection

2.1 Principe

Une autre manière de s'affranchir d'un modèle linéaire consiste à augmenter ou changer le vecteur d'entrée X. Les transformations utilisées forment un ensemble $\mathcal{D}$ de fonctions non linéaires, appelé dictionnaire.

$$f(x) = \sum_{h \in \mathcal{D}} \beta_h h(x).$$

La construction d'un dictionnaire se fait en général par l'une des trois approches suivantes :

- par restriction : des considérations contextuelles et des connaissances du modèle mènent à un choix naturel de cet ensemble de fonctions,
- par sélection : on construit un dictionnaire très vaste et l'on ne retient que les fonctions qui contribuent de manière significative à la régression,
- par régularisation : on construit un dictionnaire relativement vaste mais la régression s'effectue avec des contraintes sur les coefficients (*ridge* regression).

Notons que la taille M du dictionnaire joue un rôle analogue à celui de la fenêtre de l'estimateur à noyau : il s'agit d'un paramètre de lissage dont le choix est crucial pour établir l'équilibre biais-variance.

2.2 Quelques bases de projection usuelles

Les polynômes font partie des bases couramment utilisées. Cependant leur caractère global rend souvent l'estimation moins robuste : l'ajustement d'un coefficient dans une certaine zone peut dégrader fortement l'estimation dans d'autres régions.

La base trigonométrique se rencontre par exemple dans l'étude de phénomènes périodiques. Il s'agit de la base orthonormée de $\mathbb{L}^2[0,1]$ définie par

$$\begin{aligned} h_1(x) &= 1, \\ h_{2k}(x) &= \sqrt{2}cos(2\pi kx), \\ h_{2k+1}(x) &= \sqrt{2}sin(2\pi kx). \end{aligned}$$

Les bases d'ondelettes, quant à elles, sont couramment utilisées dans le cadre du traitement du signal.
Soit $\psi : \mathbb{R} \to \mathbb{R}$ une fonction à support compact. On définit

$$\phi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad j,k \in \mathbb{Z}.$$

Sous certaines conditions sur la fonction mère ψ, l'ensemble des fonctions filles ϕ forme une base orthonormée de $\mathbb{L}^2(\mathbb{R})$.
Les ondelettes localisent la fonction projetée simultanément en temps et en fréquence. Cette "double" localisation les rendent mieux adaptées pour la représentation de fonctions à grandes irrégularités locales.

Et enfin un dernier ensemble fréquemment utilisé : les splines. Ce qui suit décrit plus précisément ces fonctions.

2.3 Régression par splines

L'introduction des fonctions splines découle d'une volonté de partitionner l'espace des valeurs (ici la droite réelle) prises par la variable X. L'estimation d'une fonction constante par morceau peut se faire par exemple en choisissant

$$h_1(X) = \mathbb{1}_{X<k_1}, \quad h_2(x) = \mathbb{1}_{k_1 \leq X < k_2}, \quad ... \quad , h_M(x) = \mathbb{1}_{k_{M-1}<X}.$$

On peut généraliser cette approche en considérant des fonctions polynômiales par morceaux. Il est cependant souvent souhaitable d'obtenir un estimateur continu, ou même plus lisse encore. Les contraintes induites sur les coefficients peuvent être transférées à la base de fonctions. Cela amène aux splines, définies de la manière suivante :

$$\begin{aligned} h_j(X) &= X^{j-1}, \quad j = 1, ..., \tilde{M}, \\ h_{\tilde{M}+l} &= (X - k_l)_+^{\tilde{M}-1}, \quad l = 1, ..., L. \end{aligned}$$

On dénomme 'noeuds' les points $(k_1, ..., k_L)$ qui définissent la partition.

Les régressions obtenues exhibent cependant une grande variance au voisinage des noeuds extrêmes, ce qui rend l'extrapolation dangereuse. Imposer des contraintes supplémentaires dans ces régions permet de réduire ce comportement erratique. Les fonctions obtenues en imposant la linéarité au-delà des noeuds extrêmes forment ce que l'on appelle les splines naturelles. Dans le cas cubique on a la représentation suivante :

Définition. (splines cubiques naturelles)
Une base de fonction splines cubiques naturelles à L noeuds est donnée par l'ensemble des fonctions

$$\begin{aligned} h_1(X) &= 1, \\ h_2(X) &= X, \\ N_{2+l}(X) &= \frac{(X-k_l)_+^3 - (X-k_L)_+^3}{k_L - k_l} - \frac{(X-k_{L-1})_+^3 - (X-k_L)_+^3}{k_L - k_{L-1}}, \quad l = 1...L. \end{aligned}$$

Le choix des noeuds est parfois difficile et délicat. Utiliser des quantiles empiriques ou bien des méthodes adaptatives sont des approches possibles. Une manière de s'affranchir de ce choix consiste à considérer n noeuds correspondant aux valeurs de l'échantillon observé.
Soit le problème suivant : déterminer la fonction f à dérivées secondes continues qui minimise l'erreur $\mathbb{L}^2$ pénalisée suivante

$$E(f,\lambda) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int (f''(t))^2 dt$$

où λ est une constante positive. Le premier terme mesure la distance entre le modèle et les données observées alors que le second pénalise la courbure de la fonction. λ est un paramètre de lissage, il établit un équilibre entre ces deux termes. Notons que dans le cas $\lambda = 0$, f peut être n'importe quelle fonction qui interpole les données et dans le cas $\lambda = \infty$ on se ramène à une régression linéaire simple (la dérivée seconde étant forcée à zéro).
Un résultat remarquable qui justifie l'introduction des fonctions splines est le suivant : lorsque λ est fini non nul, cette erreur définie sur un espace de Sobolev de fonctions pour lesquelles le second terme est défini, admet une solution explicite, fini-dimensionnelle : une spline cubique naturelle à n noeuds localisés aux valeurs de l'échantillon.
Enfin, une validation croisée (comme celle présentée dans le cadre du lissage par noyau) permet de choisir le paramètre de lissage λ.
Notons aussi que cette approche de lissage avec pénalisation s'adapte au cadre des bases d'ondelettes.

2.4 Quelques remarques

L'approche qui vient d'être présentée offre un avantage important : la fonction de régression estimée est représentée dans une base relativement simple, ce qui facilite l'interprétation et la construction d'un modèle. Il s'agit cependant d'une méthode d'apprentissage, ce qui n'était pas le cas dans le cadre du lissage par noyau.
Enfin, ces estimateurs par projection s'étendent facilement au cadre multidimensionnel, cependant comme pour la méthode de lissage par noyaux, la vitesse de convergence se dégrade très rapidement avec la dimension.

3 Modèle de régression en grande dimension

Lorsque l'on dispose de plus de variables explicatives que de variables expliquées, les méthodes précédemment décrites deviennent inutilisables. Il s'agit d'un cas extrême, cependant même en dehors de ce scenario, lorsque le vecteur d'entrée X est multidimensionnel, se pose le problème du fléau de la dimension.
Un travail préalable de réduction de la dimension est alors nécessaire.
Nous présentons dans cette partie quelques méthodes qui rendent possible cette réduction, ainsi que deux nouveaux types de modèles de régression qui peuvent utiliser à bon escient ce traitement préalable.

3.1 Réduction de la dimension

De nombreuses méthodes permettent de réduire la dimension du vecteur d'entrée.

Des méthodes de sélection du type lasso ou *elastic net* permettent de contraindre un certain nombre de coefficients à être nuls (sparsité). Le lasso consiste à pénaliser la régression par la norme $\mathbb{L}^1$, ce qui s'écrit en terme de minimisation de la façon suivante :

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{d} \beta_j x_{i,j})^2 - \lambda \sum_{j=1}^{p} |\beta_j|$$

où p est la dimension de X et λ un réel positif. Il s'agit de la formulation Lagrangienne du problème des moindres carrés sous une contrainte de bornitude de la norme $\mathbb{L}^1$ de $\hat{\beta}$. Par la nature même de la contrainte, en choisissant λ suffisamment grand, un certain nombre de coefficients vaudront exactement zéro.

Une autre approche pour réduire la dimension du vecteur X consiste à ne sélectionner qu'un nombre resteint de composantes de l'espace vectoriel engendré par X. Ces directions sont choisies de façon à "expliquer" au mieux le vecteur observé. De manière plus formelle, il s'agit de rechercher le sous-espace vectoriel le plus proche de l'espace vectoriel engendré par X au sens d'une certaine distance. Il peut s'agir par exemple de chercher les vecteurs othogonaux (au sens $\mathbb{L}^2$) qui contiennent la plus grande part de la variance observée dans X, on parle alors d'analyse en composantes principales. On peut également chercher des vecteurs non pas orthogonaux mais indépendants (en minimisant la distance de Kullback-Leibler), on parle alors d'analyse en composantes indépendantes. Enfin des approches non linéaires peuvent également être envisagées.

3.2 Projection Pursuit Regression (Régression par Directions Révélatrices)

Le modèle par "Projection Pursuit Regression" (PPR) a la forme suivante :

$$Y = f(X) = \sum_{m=1}^{M} g_m(\omega_m^T X).$$

Ce modèle additif porte non pas sur les entrées X elles-mêmes mais sur M facteurs dérivés $V_m = \omega_m^T X$. Les fonctions g_m sont inconnues et estimées le long des directions portées par les ω_m par des méthodes de lissage.
La fonction $X \mapsto g_m(\omega_m^T X)$ est qualifiée de "ridge". Elle varie seulement dans la direction définie par ω_m. La variable réelle $V_m = \omega_m^T X$ est la projection de X sur le vecteur unitaire ω_m, et celui-ci est choisi de façon à minimiser une fonction de perte (l'erreur $\mathbb{L}^2$).
Le modèle PPR est très général et couvre une très large classe de fonction f de régression. Par exemple, le produit $X_1 X_2$ peut être écrit comme $\frac{1}{4}[(X_1 + X_2)^2 - (X_1 - X_2)^2]$ et les produits d'ordre plus élevé peuvent être représentés de manière similaire. En fait, si M est pris arbitrairement grand, pour un choix approprié des g_m, le modèle PPR peut approximer toute fonction continue de $\mathbb{R}^p \mapsto \mathbb{R}$ arbitrairement bien. Il s'agit donc d'un estimateur universel.
Pour trouver les g_m et ω_m il s'agit de minimiser l'erreur $\mathbb{L}^2$

$$\sum_{i=1}^{N} [y_i - \sum_{m=1}^{M} g_m(\omega_m^T x_i]^2.$$

Comme dans tous les problèmes de lissage il faut imposer des contraintes sur les fonctions g_m. Considérons un seul terme ($M = 1$), connaissant la direction ω on se ramène à un problème de lissage uni-dimensionnel en formant les variables $v_i = \omega^T x_i$. Dans l'autre sens, si l'on se donne g, nous devons minimiser l'erreur $\mathbb{L}^2$ sur ω. La méthode de Gauss-Newton se prête volontiers à ce problème.
Soit ω_c l'estimateur courant de ω,

$$g(\omega^T x_i) \approx g(\omega_c^T x_i) + g'(\omega_c^T x_i)(\omega - \omega_c)^T x_i,$$

et

$$\sum_{i=1}^{N} (y_i - g(\omega^T x_i))^2 \approx \sum_{i=1}^{N} g'(\omega_c^T x_i)^2 [(\omega_c^T x_i + \frac{y_i - g(\omega_c^T x_i)}{g'(\omega_c^T x_i)}) - \omega^T x_i]^2.$$

On actualise alors ω en minimisant le terme de droite (régression linéaire des moindres carrés avec poids).
Les deux étapes de minimisation sur g et ω sont itérées jusqu'à atteindre la convergence. Si l'on considère $M > 1$ on procède par l'ajout à chaque itération d'une paire (ω_m, g_m).
Quelques remarques sur l'implémentation de la méthode :

- bien que toute méthode de lissage peut potentiellement être utilisée, une méthode qui fournit des dérivées explicites sera préférée : c'est par exemple le cas des polynômes locaux et des splines.
- les ω_m ne sont pas nécessairement systématiquement actualisés afin d'éviter de trop lourds calculs.
- le nombre M est généralement choisi de façon à ce que l'ajout d'un terme supplémentaire à l'itération suivante n'améliore pas de manière apréciable la minimisation. Des méthodes de validation croisée peuvent également être utilisées.

3.3 Réseaux de neurones

Il existe une littérature abondante sur les réseaux de neurones. Le terme même de réseaux de neurones désigne parfois des modèles très différents, en particulier selon si l'on

s'y intéresse dans le cadre des statistiques ou dans le cadre de l'intelligence artificielle. Ici nous considérons les réseaux de neurones usuels, en tant que méthode de régression non-linéaire. Un tel réseau constitue une paramétrisation de la méthode PPR décrite précédemment.
La variable expliquée Y est modélisée par une combinaison linéaire des variables Z_m, $m = 1, ..., M$ définies par

$$Z_m = \sigma(\alpha_{0,m} + \alpha_m^T X)$$

où σ est une fonction d'activation souvent choisie comme étant la fonction sigmoïde : $\sigma(v) = \frac{1}{1+e^{-v}}$. La forme de cette fonction donne à $|\alpha_m|$ un rôle de taux d'activation. Lorsque cette norme est petite, Z_m intervient de façon quasi-linéaire dans l'estimation de Y.
Les Z_m sont en réalité des variables cachées, elles ne sont pas directement observables. Notons que l'on peut généraliser ce modèle en ajoutant des "couches" de variables cachées.

La calibration d'un réseau de neurones (minimisation de l'erreur $\mathbb{L}^2$) n'est pas toujours directe : le nombre de paramètres est souvent plus grand que le nombre d'observations. Il convient alors d'utiliser une méthode de régularisation : pénaliser l'estimation (décroissance exponentielle des coefficients par exemple) et déterminer le paramètre de pénalisation par validation croisée.

Références

[1] Ahamada I., Flachaire E., (2008). Econométrie non-paramétrique, *Economica*.

[2] Bishop C., (1995). Neural networks for pattern recognition, *Clarendon Press, Oxford*.

[3] Bunea F., Tsybakov A.B., Wegkamp M., (2007). Sparsity oracle inequalities for the lasso, *Electronic Journal of Statistics*.

[4] de Boor C., (1978). A practical guide to splines, *Springer*.

[5] Friedman J., Tukey J., (1974). A projection pursuit algorithm for exploratory data analysis, *IEEE Transactions on computer*, 23, 881-889.

[6] Gencay R., Selcuk F., Whitcher B.J., (2010). An introduction to wavelets and other filtering methods in finance and economics, *Academic Press Inc.*

[7] Hardle W., Kerkyacharian G., Picard D. Tsybakov A.B., (2000). Wavelets, approximation, and statistical applications, *Springer-Verlag New York Inc.*.

[8] Hastie T., Tibshirani R., Friedman J., (2009). The elements of statistical learning, *Springer Series in Statistics*.

[9] Mardia K., Kent J., Bibby J., (1979). Multivariate analysis, *Academic Press*.

[10] Neal R., Zhang J. (2006). High dimensional classification with bayesian neural networks and Dirichlet diffusion trees, *Features extraction, Fondations and Applications, Springer*, 265-296.

[11] Stone C.J., (1984). An asymptotically optimal window selection rule for kernel density estimates, *Annals of Statistics*, 12, 1285-1297.

[12] Tsybakov A.B., (2009). Introduction to nonparametric estimation, *Springer-Verlag New York Inc.*.

[13] Wand, M.P., Jones M.C., (1995). Kernel smoothing, *Chapman and Hall, London.*