IUT de Paris - Rives de Seine
Université Paris Cité

Modélisation statistique avancée

Introduction à la statistique non paramétrique

IUT de Paris - Rives de Seine

BUT Science des Données, troisième année

Parcours : Exploration et Modélisation Statistique

Qu'est-ce que la statistique non paramétrique ?

Le but de la statistique non paramétrique est d'estimer des quantités d'intérêt (moyenne, variance, quantiles, fonction de répartition, densité, fonction de régression, courbe de survie, ...), sans supposer que la loi des variables aléatoires (ou la relation liant les variables) est connue à des paramètres près.

Pour les quantité numériques (moyenne, variance, quantiles, ...) on utilise le plus souvent l'estimateur empirique basé sur les observations (moyenne empirique, variance empirique, quantiles empiriques...) dont on peut établir les propriétés asymptotiques : consistance, vitesse de convergence, normalité asymptotique...

Dans ce cours, on va estimer des quantités plus complexes : la densité de probabilité ou la fonction de régression.

Contexte : on dispose d'observations $x_1, \ldots, x_n$ issues de variables aléatoires réelles iid $X_1, \ldots, X_n$ (par exemple obtenues par sondage aléatoire simple avec remise).

On suppose que les variables X_i possèdent une densité de probabilité f. On veut estimer f sur un intervalle $[a, b]$.

Rappel : f est une fonction de $\mathbb{R}$ dans $\mathbb{R}_+$ telle que, $\forall s \leq t$,

$$\mathbb{P}(X \in [s, t]) = \int_s^t f(x)dx\,.$$

Histogramme à pas régulier

Pour estimer f sur $[a, b]$ à partir des observations $x_1, \ldots, x_n$, une idée simple est d'utiliser un Histogramme à pas régulier (ou simplement "régulier"). Il est défini ainsi :

- On choisit une partition régulière de $[a, b]$ de taille m :

$$\{I_k\}_{k\in\{1,\ldots,m\}}, \quad \text{où} \quad I_k = \left[a + \frac{(k-1)(b-a)}{m}, a + \frac{k(b-a)}{m}\right[.$$

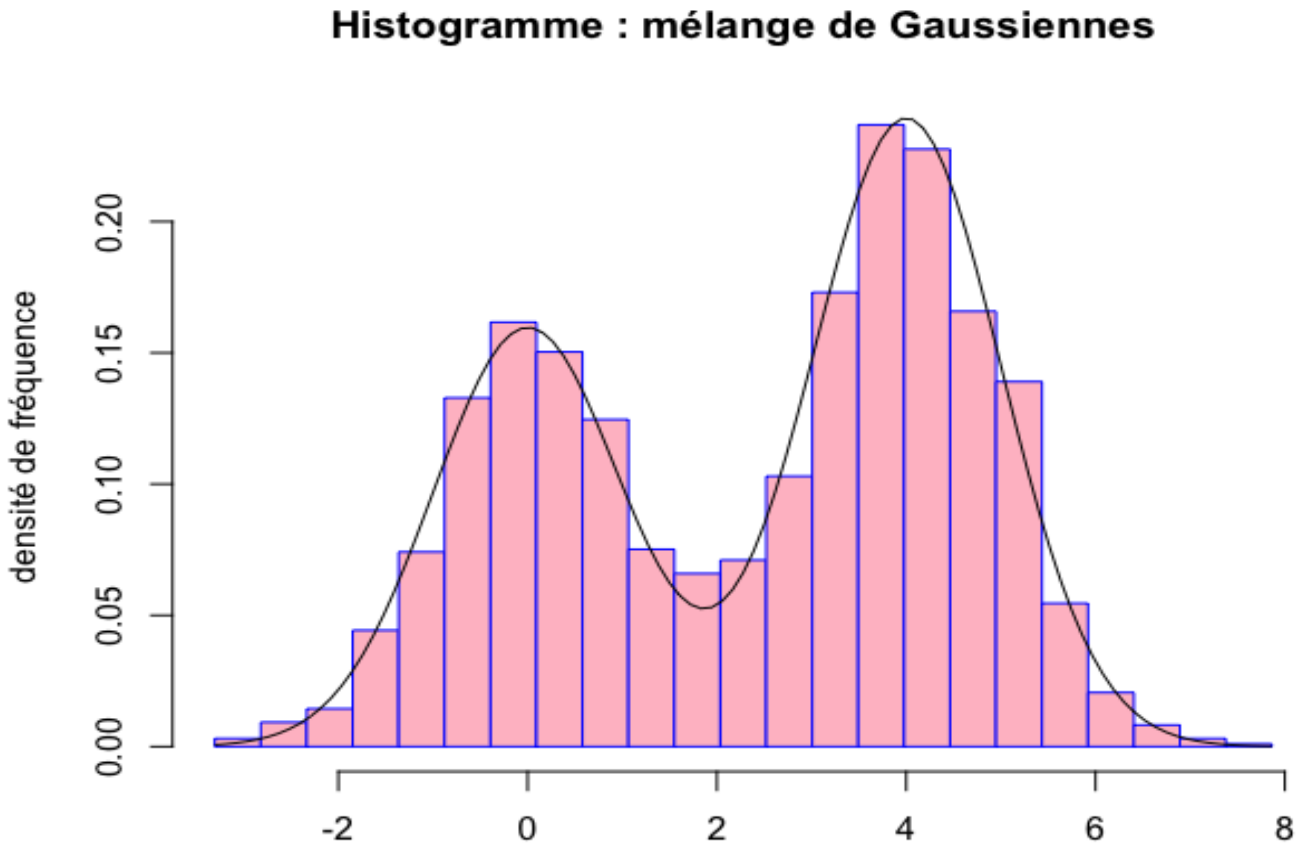
- On note $p_{n,k} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \in I_k}$ et $d_{n,k} = \frac{m p_{n,k}}{b-a}$.

  $d_{n,k}$ est la densité de fréquence de I_k.

- Pour $x \in [a, b]$, on définit l'histogramme régulier H_m:

$$H_m(x) = \sum_{k=1}^{m} d_{n,k} \mathbf{1}_{x \in I_k}.$$

Histogramme régulier : exemple



Exemple d'Histogramme à pas régulier, mélange de deux lois normales, données simulées, $n = 2000$, taille de la partition $m = 23$.

Histogramme régulier : risque quadratique intégré

On cherche à mesurer l'écart entre H_m et la densité f sur l'intervalle $[a, b]$. Pour des raison de facilité de calcul, on utilise souvent le risque quadratique intégré (RQI) :

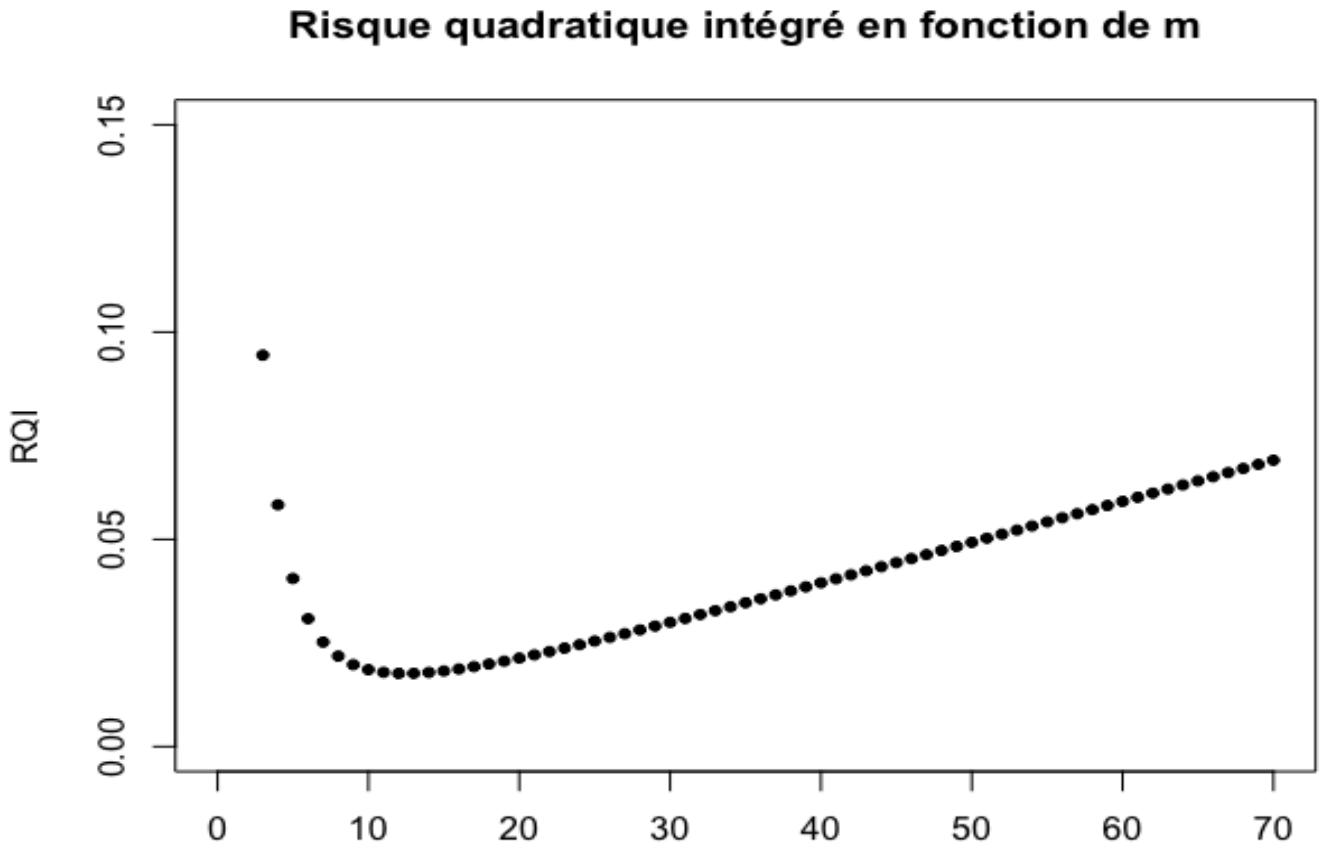
$$RQI(H_m, f) = \int_a^b \mathbb{E}\left((H_m(x) - f(x))^2\right) dx\,.$$

Soit m_n une suite d'entiers qui tend vers l'infini lorsque n tend vers l'infini. On dira que H_{m_n} est un estimateur consistant de f sur $[a, b]$ lorsque

$$\lim_{n\to\infty} RQI(H_{m_n}, f) = 0\,.$$

On peut montrer que, si f est continue sur $[a, b]$, alors H_{m_n} est un estimateur consistant de f pour toute suite m_n telle que $m_n \to \infty$ et $m_n/n \to 0$.

Histogramme régulier : exemple



Pour un exemple de loi $\beta(1.9, 1.9)$ (f connue, $n = 1000$), on peut calculer exactement $RQI(\hat{H}_m, f)$. Il est minimal pour $m = 12$.

Histogramme régulier : choix de m_n

La question qui se pose à présent est celle du choix de m_n. Y'a-t-il un choix naturel de m_n pour lequel on peut obtenir une vitesse de convergence intéressante pour le *RQI* ?

Pour répondre à cette question, on va faire une hypothèse de régularité sur f. On va supposer que f est dérivable et de dérivée continue sur $[a, b]$.

Rappel : par le théorème des accroissements finis, cela signifie que, pour tout $s, t \in [a, b]$,

$$|f(t) - f(s)| \leq C|t - s|,$$

où C est le maximum de $|f'(x)|$ pour $x \in [a, b]$.

Histogramme régulier : choix de m_n

Sous cette hypothèse de régularité de f, on va pouvoir trouver un m_n approprié. Pour simplifier les calculs, on suppose que $[a, b] = [0, 1]$, de sorte que $I_k = \left[\frac{k-1}{m}, \frac{k}{m}\right[$.

Notons d'abord que, pour tout $x \in [0, 1]$,

$$H_m(x) - f(x) = \sum_{k=1}^{m} (m p_{n,k} - f(x)) \mathbf{1}_{x \in I_k} .$$

Par conséquent

$$\begin{aligned} (H_m(x) - f(x))^2 &= \sum_{k=1}^{m} (m p_{n,k} - f(x))^2 \mathbf{1}_{x \in I_k} \\ \text{et donc } \mathbb{E}\left((H_m(x) - f(x))^2\right) &= \sum_{k=1}^{m} \mathbb{E}\left((m p_{n,k} - f(x))^2\right) \mathbf{1}_{x \in I_k} . \end{aligned}$$

Histogramme régulier : choix de m_n

Exercice

Notons $p_k = \mathbb{P}(X_1 \in I_k)$. Montrer que

$$\mathbb{E}\left((mp_{n,k} - f(x))^2\right) = \frac{m^2}{n} p_k(1-p_k) + (mp_k - f(x))^2 .$$

Puisque $p_k(1-p_k) \leq p_k$, on en déduit que

$$\mathbb{E}\left((H_m(x) - f(x))^2\right) \leq \sum_{k=1}^{m} \left(\frac{m^2}{n} p_k + (mp_k - f(x))^2 \right) \mathbf{1}_{x \in I_k} .$$

Histogramme régulier : choix de m_n

Par définition de p_k, on a que

$$mp_k = m \int_{(k-1)/m}^{k/m} f(t)dt \quad \text{et} \quad mp_k - f(x) = m \int_{(k-1)/m}^{k/m} (f(t) - f(x))dt\,.$$

Exercice

En utilisant l'hypothèse de régularité sur f, montrer que

$$(mp_k - f(x))^2 \mathbf{1}_{x \in I_k} \leq \frac{C^2}{m^2}\,.$$

Grâce aux calculs précédents, on a montré que

$$\mathbb{E}\left((H_m(x) - f(x))^2\right) \leq \sum_{k=1}^{m} \left(\frac{m^2}{n} p_k + \frac{C^2}{m^2}\right) \mathbf{1}_{x \in I_k} .$$

Par conséquent, puisque $\int_0^1 \mathbf{1}_{x \in I_k} dx = 1/m$,

$$\begin{aligned} RQI(H_m, f) &\leq \sum_{k=1}^{m} \left(\frac{m^2}{n} p_k + \frac{C^2}{m^2}\right) \int_0^1 \mathbf{1}_{x \in I_k} dx \\ &\leq \frac{m}{n} \left(\sum_{k=1}^{m} p_k\right) + \frac{C^2}{m^2} \\ &\leq \frac{m}{n} + \frac{C^2}{m^2} . \end{aligned}$$

Histogramme régulier : choix de m_n

Exercice

Calculer le minimum de la fonction $g : \mathbb{R}_+ \to \mathbb{R}$ définie par

$$g(m) = \frac{m}{n} + \frac{C^2}{m^2}$$

et déterminer l'unique $\tilde{m}_n$ pour lequel le minimum est atteint.

Conclusion : en choisissant $m_n = [n^{1/3}]$, on obtient que

$$RQI(H_{m_n}, f) = O\left(\frac{1}{n^{2/3}}\right) .$$

On dit que H_{m_n} converge vers f à la vitesse $n^{2/3}$ (pour le RQI).
On peut montrer que cette vitesse est optimale en un certain sens, sous les hypothèses de régularité de f que nous avons faites.

Histogramme régulier : choix de m à partir des données

Le résultat présenté dans le slide 12 ne permet pas de choisir m_n lorsque la densité f est moins régulière que ce que nous avons supposé. Même dans le cas ou f est dérivable, le "meilleur" choix de m_n dépend du max de $|f'(x)|$ qui est inconnu (voir exercice).

Les auteurs Castellan (2000) puis Birgé et Rozenholc (2006) ont mis au point une procédure pour choisir à partir des données un $m^* = m(X_1, \ldots, X_n)$ "proche" du meilleur m possible (inconnu).

Ce m^* est obtenu en maximisant $L_n(m) - pen(m)$ pour $1 \leq m \leq n/\log(n)$, où $pen(m) = m - 1 + (\log m)^{2.5}$ et

$$L_n(m) = \sum_{j=1}^{m} np_{n,k} \log(\max(1, np_{n,k})) + \log(m) \sum_{k=1}^{m} np_{n,k}$$

Cet Histogramme H_{m^*} sera étudié en TP.

Le contexte est le même qu'en partie 1 : on dispose d'observations $x_1, \ldots, x_n$ issues de variables aléatoires réelles iid $X_1, \ldots, X_n$.

On suppose que les variables X_i possèdent une densité de probabilité f. On veut estimer f en un point $x \in \mathbb{R}$. On supposera que f est bornée sur $\mathbb{R}$: $\sup_{t \in \mathbb{R}} f(t) = M_f < \infty$.

On va utiliser un noyau K, c'est à dire une fonction à valeurs réelles telle que $K \geq 0$, $\int_{\mathbb{R}} K(t)dt = 1$,

$$C_1(K) = \int_{\mathbb{R}} |t| K(t)dt < \infty\,, \quad C_2(K) = \int_{\mathbb{R}} (K(t))^2 dt < \infty\,.$$

Les noyaux usuels sont en général pairs : $K(t) = K(-t)$.

Estimation ponctuelle de la densité par Noyau

Soit $h > 0$. L'estimateur à noyau de la densité $\hat{f}_h$ est défini par

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

Pour estimer la densité au point x, on utilise le risque quadratique (RQ) :

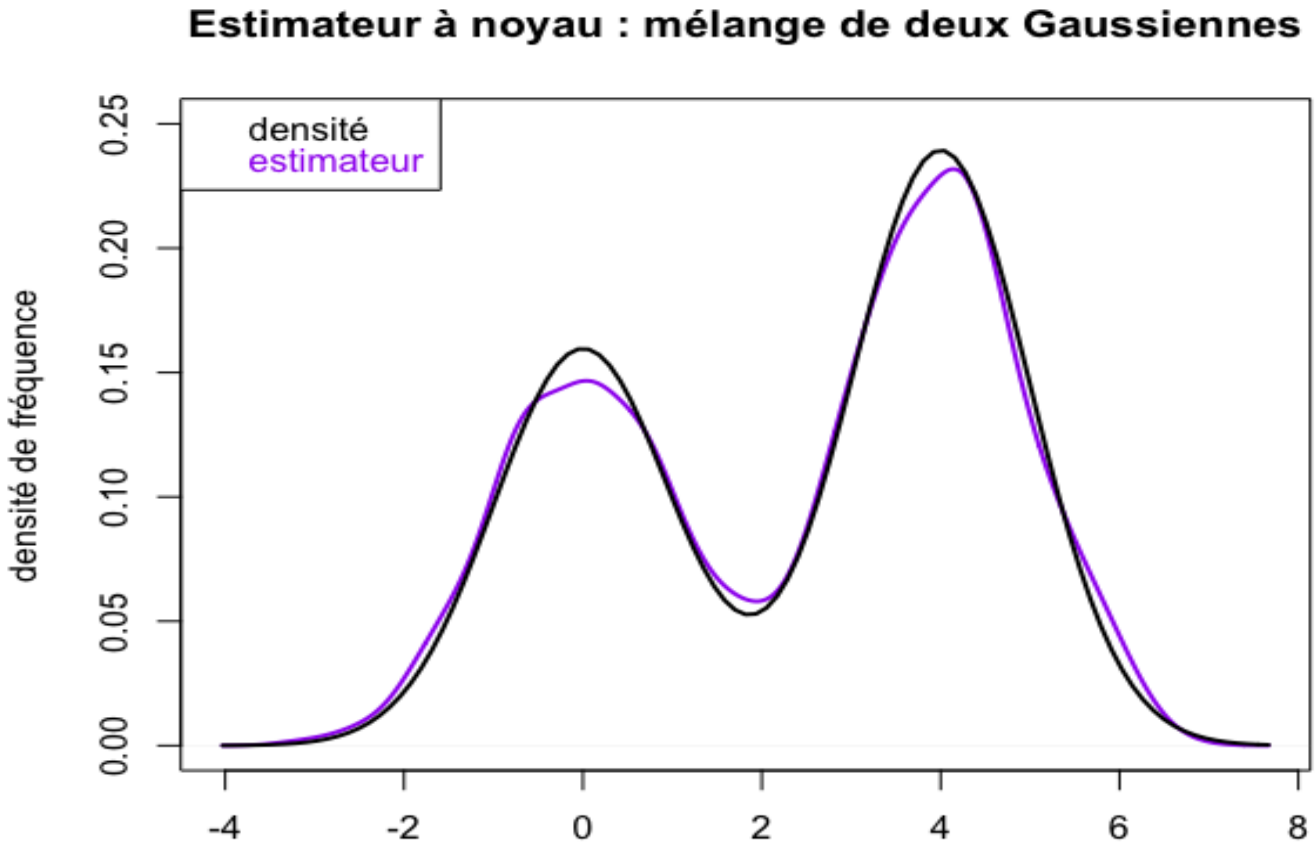
$$RQ(\hat{f}_h(x), f(x)) = \mathbb{E}\left((\hat{f}_h(x) - f(x))^2\right) .$$

Soit h_n une suite de réels qui tend vers 0 lorsque n tend vers l'infini. On dira que $\hat{f}_{h_n}(x)$ est un estimateur consistant de $f(x)$ lorsque

$$\lim_{n \to \infty} RQ(\hat{f}_{h_n}(x), f(x)) = 0 .$$

On peut montrer que, si f est continue au point x, alors $\hat{f}_{h_n}(x)$ est un estimateur consistant de f pour toute suite h_n telle que $h_n \to 0$ et $nh_n \to \infty$.

Estimateur à noyau : exemple



Exemple d'estimateur à noyau, mélange de deux lois normales, données simulées, $n = 2000$, $h = 0.2785$ (noyau Gaussien).

Estimateur à noyau : choix de h_n

La question qui se pose à présent est celle du choix de h_n. Y'a-t-il un choix naturel de h_n pour lequel on peut obtenir une vitesse de convergence intéressante ?

Pour répondre à cette question, on va faire une hypothèse de régularité sur f. On va supposer que f est dérivable et de dérivée continue sur un intervalle $[x-a, x+a]$, avec $a > 0$.

On a déjà vu que cela signifie que, pour tout $s, t \in [x-a, x+a]$,

$$|f(t) - f(s)| \leq C|t-s|,$$

où C est le maximum de $|f'(z)|$ pour $z \in [x-a, x+a]$.

Estimateur à noyau : choix de h_n

Sous cette hypothèse de régularité de f, on va pouvoir trouver un h_n approprié.

Exercice

On note $f_h(x) = \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x-t}{h}\right) f(t)dt$. Montrer que

$$RQ(\hat{f}_h(x), f(x)) \leq \frac{1}{n}\mathrm{Var}\left(\frac{1}{h}K\left(\frac{x-X_1}{h}\right)\right) + (f_h(x) - f(x))^2 .$$

Montrer ensuite que

$$\frac{1}{n}\mathrm{Var}\left(\frac{1}{h}K\left(\frac{x-X_1}{h}\right)\right) \leq \frac{M_f}{nh}\int_{\mathbb{R}} K^2(t)dt .$$

Estimateur à noyau : choix de h_n

On déduit de l'exercice précédent que

$$RQ(\hat{f}_h(x), f(x)) \leq \frac{M_f C_2(K)}{nh} + (f_h(x) - f(x))^2 .$$

Exercice

Rappel : pour tout $s, t \in [x-a, x+a]$, $|f(t) - f(s)| \leq C|t-s|$.
Montrer que

$$f_h(x) - f(x) = \int_{\mathbb{R}} K(y)(f(x-hy) - f(x))dy ,$$

puis que

$$(f_h(x) - f(x))^2 \leq h^2 \left(C_1(K)C + C_1(K)\frac{2M_f}{a} \right)^2 .$$

Estimateur à noyau : choix de h_n

Exercice

Calculer le minimum de la fonction $g : \mathbb{R}_+ \to \mathbb{R}$ définie par

$$g(h) = \frac{\kappa_1}{nh} + \kappa_2 h^2$$

et déterminer l'unique $\tilde{h}_n$ pour lequel le minimum est atteint.

Conclusion : en choisissant $h_n = \frac{1}{n^{1/3}}$, on obtient que

$$RQ(\hat{f}_{h_n}(x), f(x)) = O\left(\frac{1}{n^{2/3}}\right) .$$

On dit que $\hat{f}_{h_n}(x)$ converge vers f à la vitesse $n^{2/3}$ (pour le RQ).
On peut montrer que cette vitesse est optimale en un certain sens, sous les hypothèses de régularité de f que nous avons faites.

Estimateur à noyau : choix de h_n

Supposons à présent que f soit deux fois dérivable sur $\mathbb{R}$, de dérivée seconde continue et bornée par C_1. Si de plus

$$\int_{\mathbb{R}} tK(t)dt = 0 \quad \text{et} \quad C_3(K) = \int_{\mathbb{R}} t^2 K(t)dt < \infty\,,$$

alors on peut montrer que $(f_h(x) - f(x))^2 \leq \dfrac{h^4 C_1^2 C_3(K)^2}{4}$.

Exercice

Calculer le minimum de la fonction $g : \mathbb{R}_+ \to \mathbb{R}$ définie par

$$g(h) = \frac{\kappa_1}{nh} + \kappa_2 h^4$$

et déterminer l'unique $\tilde{h}_n$ pour lequel le minimum est atteint. En déduire la vitesse de convergence de $RQ(\hat{f}_{h_n}(x), f(x))$.

Estimateur à noyau : choix de h à partir des données

Le résultat présenté dans le slide 21 ne permet pas de choisir h_n lorsque la densité f est moins régulière que ce que nous avons supposé. Même dans le cas ou f est dérivable, le "meilleur" choix de h_n dépend de M_f et du max de $|f'(x)|$ qui sont inconnus.

Plusieurs méthodes pour choisir à partir des données un $h^* = h(X_1, \ldots, X_n)$ "proche" du meilleur h connu ont été développées.

Parmi ces méthodes, citons la validation croisée (nombreux articles entre 1975 et 1990) et la méthode de Goldenshluger et Lepski (2011).

Le principe de la validation croisée

On note $\hat{f}_h$ l'estimateur à noyau construit à partir de toutes les données, et $\hat{f}_h^{(-i)}$ l'estimateur à noyau construit à partir de toutes les données sauf la $i$ème. On construit ensuite

$$\hat{R}(h) = \int_{\mathbb{R}} \left(\hat{f}_h(x)\right)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_h^{(-i)}(X_i)\,.$$

on choisit ensuite h^* comme le plus petit h tel que $\hat{R}(h^*)$ soit minimal.

Il s'agit de la validation croisée leave one out. D'autres méthodes existent, basées sur les échantillons test et les échantillons d'apprentissage.

Contexte : on considère le modèle de régression suivant

$$Y_i = f(x_i) + \varepsilon_i \,,$$

où $(\varepsilon_i)_{1 \leq i \leq n}$ est une suite de variables réelles iid avec $\mathbb{E}(\varepsilon_i) = 0$ et $\mathbb{E}(\varepsilon_i^2) = \sigma^2$, et x_i est non aléatoire (dans un premier temps), à valeurs dans un intervalle $[a, b]$.

On observe $(Y_i, x_i)_{1 \leq i \leq n}$ et on veut estimer la fonction de régression f.

Régressogramme à pas régulier

Pour fixer les idées, on supposera que $x_i \in [0,1]$. Pour estimer la fonction de régression f, une première idée est d'utiliser un Régressogramme à pas régulier (ou simplement régulier). Il est défini ainsi :

- On choisit une partition régulière de $[0,1]$ de taille m :

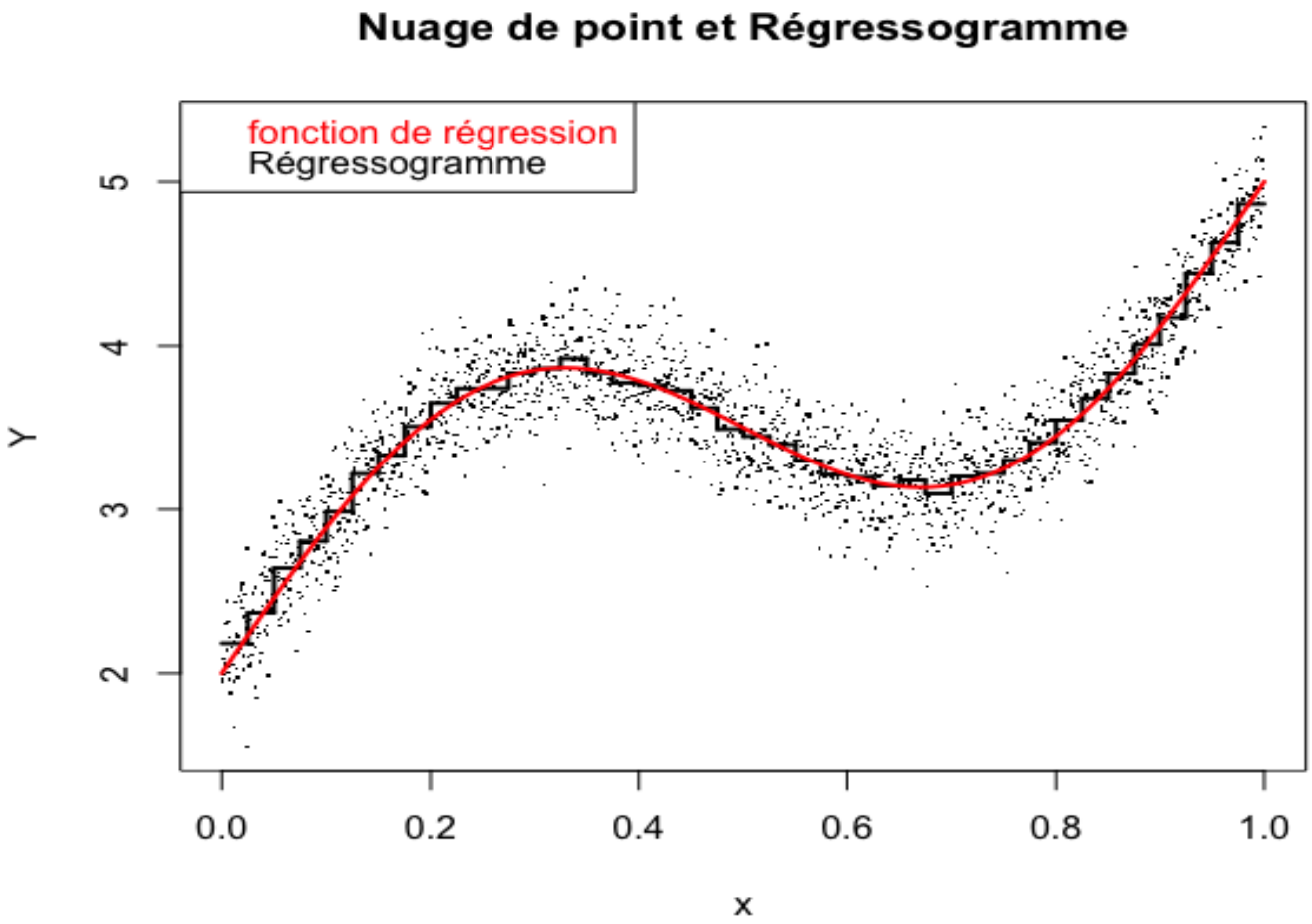
$$\{I_k\}_{k\in\{1,\dots,m\}}, \quad \text{où} \quad I_k = \left[\frac{(k-1)}{m}, \frac{k}{m}\right[.$$

- On note $\bar{Y}_k = \frac{1}{n_k}\sum_{i=1}^{n} Y_i \mathbf{1}_{x_i\in I_k}$ où $n_k = \sum_{i=1}^{n} \mathbf{1}_{x_i\in I_k}$.

- Pour $x \in [0,1]$, on définit $\hat{f}_m(x) = \sum_{k=1}^{m} \bar{Y}_k \mathbf{1}_{x\in I_k}$.

La fonction $\hat{f}_m$ est le régressogramme à pas régulier.

Régressogramme régulier : exemple



Fonction de régression $f(x) = 2 + 3x + \sin(2\pi x)$.
Exemple de Régressogramme à pas régulier, données simulées, $n = 2000$, taille de la partition $m = 40$.

Régressogramme régulier : écart quadratique moyen

On cherche à mesurer l'écart entre $\hat{f}_m$ et la fonction de régression f sur l'intervalle $[0,1]$. Pour des raison de facilité de calcul, on utilise souvent l'espérance de l'écart quadratique moyen (EQM) :

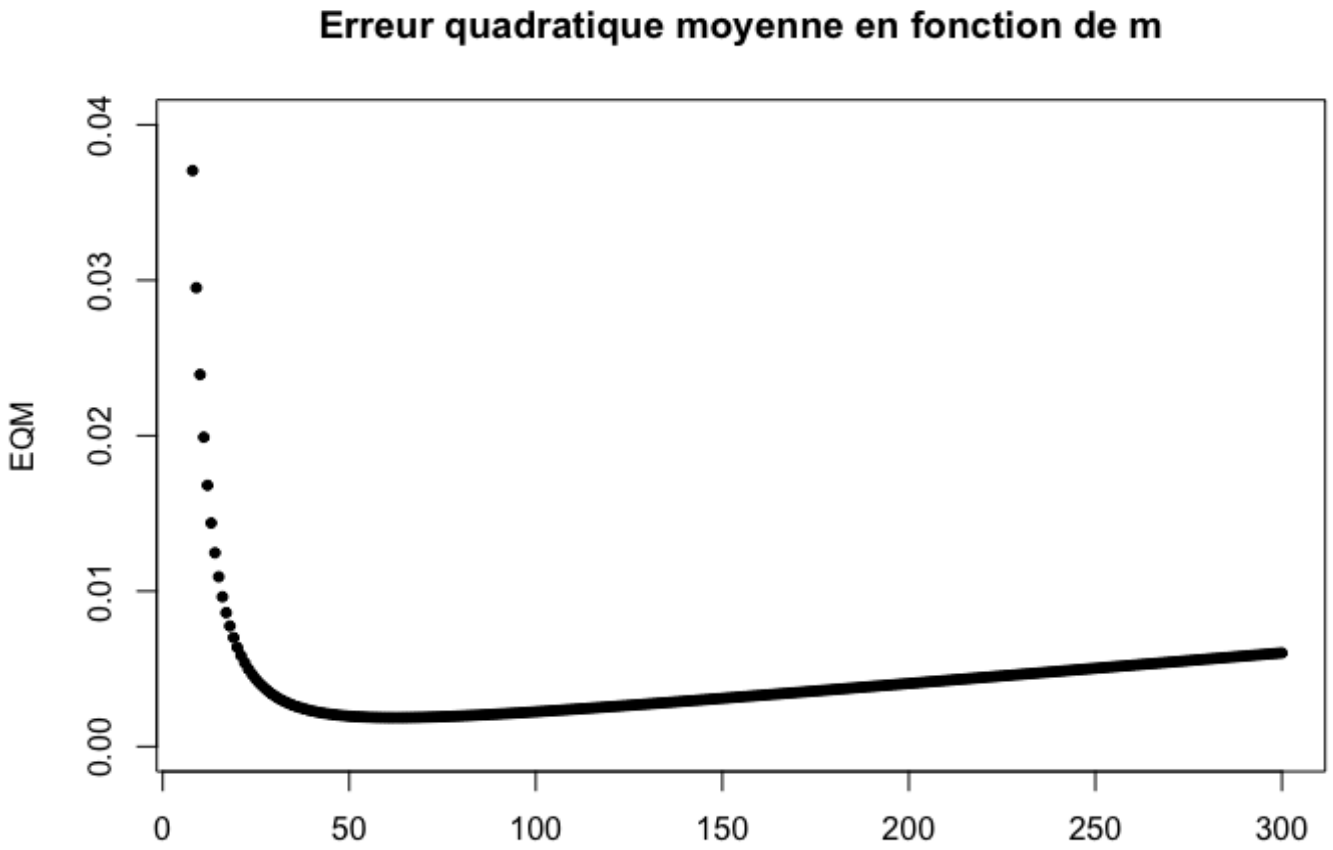
$$EQM(\hat{f}_m, f) = \mathbb{E}\left(\|\hat{f}_m - f\|_{2,n}^2\right) \text{ où } \|\hat{f}_m - f\|_{2,n}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_m(x_i) - f(x_i))^2.$$

Soit m_n une suite d'entiers qui tend vers l'infini lorsque n tend vers l'infini. On dira que $\hat{f}_{m_n}$ est un estimateur consistant de f lorsque

$$\lim_{n\to\infty} EQM(\hat{f}_{m_n}, f) = 0\,.$$

On peut montrer que, si f est continue sur $[0,1]$, alors $\hat{f}_{m_n}$ est un estimateur consistant de f pour toute suite m_n telle que $m_n \to \infty$ et $m_n/n \to 0$.

Régressogramme régulier : exemple



Pour l'exemple du slide 28, avec f connue, on peut calculer exactement $EQM(\hat{f}_m, f)$. Il est minimal pour $m = 62$.

Régressogramme régulier : choix de m_n

La question qui se pose à présent est celle du choix de m_n. Y'a-t-il un choix naturel de m_n pour lequel on peut obtenir une vitesse de convergence intéressante pour l'EQM?

Pour répondre à cette question, on va faire une hypothèse de régularité sur f. On va supposer que f est dérivable et de dérivée continue sur $[0, 1]$.

Comme dans la partie 1, on utilisera que, pour tout $s, t \in [0, 1]$

$$|f(t) - f(s)| \leq C|t - s|,$$

où C est le maximum de $|f'(x)|$ pour $x \in [0, 1]$.

Régressogramme régulier : choix de m_n

Sous cette hypothèse de régularité de f, on va pouvoir trouver un m_n approprié.

Notons d'abord que,

$$\|\hat{f}_m - f\|_{2,n}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m}(\bar{Y}_k - f(x_i))^2 \mathbf{1}_{x_i \in I_k}$$

et donc

$$\mathbb{E}\left(\|\hat{f}_m - f\|_{2,n}^2\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m}\mathbb{E}\left((\bar{Y}_k - f(x_i))^2\right)\mathbf{1}_{x_i \in I_k}\,.$$

Exercice

Soit $\bar{f}_k = \frac{1}{n_k} \sum_{i=1}^{n} f(x_i) \mathbf{1}_{x_i \in I_k}$. Rappelons que $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{E}(\varepsilon_i^2) = \sigma^2$.

Montrer que

$$\mathbb{E}\left((\bar{Y}_k - f(x_i))^2\right) = \frac{\sigma^2}{n_k} + (\bar{f}_k - f(x_i))^2 .$$

On en déduit que

$$\mathbb{E}\left(\|\hat{f}_m - f\|_{2,n}^2\right) = \frac{\sigma^2 m}{n} + \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m} (\bar{f}_k - f(x_i))^2 \mathbf{1}_{x_i \in I_k} .$$

Régressogramme régulier : choix de m_n

Exercice

En utilisant l'hypothèse de régularité sur f, montrer que

$$(\bar{f}_k - f(x_i))^2 \mathbf{1}_{x_i \in I_k} \leq \frac{C^2}{m^2} .$$

On en déduit que

$$EQM(\hat{f}_m, f) = \mathbb{E}\left(\|\hat{f}_m - f\|_{2,n}^2 \right) \leq \frac{\sigma^2 m}{n} + \frac{C^2}{m^2} .$$

Régressogramme régulier : choix de m_n

On a vu en partie 1 comment minimiser la fonction

$$h(m) = \frac{\sigma^2 m}{n} + \frac{C^2}{m^2}\,.$$

Conclusion : en choisissant $m_n = [n^{1/3}]$, on obtient que

$$EQM(\hat{f}_{m_n}, f) = O\left(\frac{1}{n^{2/3}}\right)\,.$$

On dit que $\hat{f}_{m_n}$ converge vers f à la vitesse $n^{2/3}$ (pour l'EQM).

Régressogramme régulier : choix de m à partir des données

Le résultat présenté dans le slide 34 ne permet pas de choisir m_n lorsque la densité f est moins régulière que ce que nous avons supposé. Même dans le cas ou f est dérivable, le "meilleur" choix de m_n dépend du max de $|f'(x)|$ et de σ^2 qui sont inconnus.

Baraud (2000) a mis au point une procédure pour choisir à partir des données un $m^* = m(X_1, \ldots, X_n)$ "proche" du meilleur m possible (inconnu), en supposant que $\mathbb{E}(|\varepsilon_i|^q) < \infty$ pour un $q > 6$.

Ce m^* est obtenu en minimisant, pour $1 \leq m \leq n$,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_m(x_i))^2 + pen(m) \quad \text{avec} \quad pen(m) = \frac{2m\sigma^2}{n}\,.$$

La penalisation $pen(m)$ est le C_p de Mallows (1973).

Régressogramme régulier : choix de m à partir des données

La procédure de Baraud (2000) fonctionne parfaitement si σ^2 est connu. S'il est inconnu, on peut essayer de l'estimer à partir des résidus d'un modèle de départ (voir encore Baraud (2000)).

Une autre façon de faire est de considérer une famille de pénalités, dépendant d'une constante κ, de la forme

$$pen_\kappa(m) = \frac{m\kappa}{n}$$

et d'essayer de trouver le "bon" κ, qui donnera le "bon" m. De façon un peu miraculeuse, cette procédure fonctionne. Un algorithme pour trouver κ est codé dans le package R `capushe`.

Régressogramme régulier : le cas du design aléatoire

Tout ce que l'on a vu dans les slides précédent se généralise au modèle de régression

$$Y_i = f(X_i) + \varepsilon_i \,,$$

où les variables $(X_i, \varepsilon_i)_{1 \leq i \leq n}$ sont iid, pourvu que X_i prenne ses valeurs dans $[a, b]$, que $\mathbb{E}(\varepsilon_i) = 0$ et $\mathbb{E}(\varepsilon_i^2) = \sigma^2$, et que X_i soit indépendante de ε_i.

Sous cette dernière hypothèse, on peut faire exactement les mêmes calculs conditionnellement aux $(X_i)_{1 \leq i \leq n}$. L'EQM, vaut alors

$$EQM(\hat{f}_m, f) = \mathbb{E}\left(\|\hat{f}_m - f\|_{2,n}^2\right) \text{ où } \|\hat{f}_m - f\|_{2,n}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_m(X_i) - f(X_i))^2.$$

Si f vérifie l'hypothèse de régularité, et $m_n = [n^{1/3}]$, on a encore que $\hat{f}_{m_n}$ converge vers f à la vitesse $n^{2/3}$ (pour l'EQM). De même, les règles de choix de m à partir des données fonctionnent encore.

Les tests de type Kolmogorov-Smirnov sont des tests basés sur les fonctions de répartition empiriques. Ils sont valables lorsque les variables sont continues.
On rappelle que pour une suite de variables aléatoires réelles iid $X_1, \ldots, X_n$, la fonction de répartition empirique F_n vaut

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leq t} \, .$$

Par la loi des grands nombres, on a que $F_n(t)$ converge en probabilité vers $F(t) = \mathbb{P}(X_1 \leq t)$ lorsque $n \to \infty$. En fait, on peut même montrer que

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \quad \text{converge en probabilité vers } 0.$$

C'est le théorème de Glivenko-Cantelli.

Le test d'adéquation à une loi connue F_0

On cherche à savoir si la fonction de répartition F des X_i est égale à une fonction de répartition connue et continue F_0. Les hypothèses de test sont donc

$$H_0 : F = F_0 \quad \text{contre} \quad H_1 : F \neq F_0.$$

On utilise la statistique de test

$$T_n = \sqrt{n} \sup_{t \in \mathbb{R}} |F_n(t) - F_0(t)| .$$

On va montrer qu'il s'agit d'une statistique libre, c'est à dire que sa loi sous H_0 ne dépend pas de F_0, et qu'elle peut donc être tabulée.

Le test d'adéquation à une loi connue F_0

Exercice

On suppose que F_0 est continue et strictement croissante. Montrer que

$$T_n = \sqrt{n} \sup_{s \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{F_0(X_i) \leq s} - s \right| .$$

Calculer la fonction de répartition de $F_0(X_1)$ sous H_0. En déduire que T_n est une statistique libre. Comment peut-on la tabuler ?

En fait le résultat reste vrai en supposant seulement que F_0 est continue.

Le test d'adéquation à une loi connue F_0

On peut aussi montrer que, sous H_0, la statistique T_n converge en loi vers une loi connue (la loi du sup du processus de Wiener). Ce résultat est dû à Donsker (1951).
On rejette H_0 au niveau de risque α lorsque $t_n > q_{n,1-\alpha}$, où $q_{n,1-\alpha}$ est tel que

$$\mathbb{P}_{H_0}(T_n > q_{n,1-\alpha}) \sim \alpha\,.$$

En pratique, on utilise les quantiles de la loi exacte de T_n sous H_0 lorsque n n'est pas trop grand, et les quantiles de la loi asymptotique lorsque n est grand.

Enfin, le théorème de Glivenko Cantelli montre que si $F \neq F_0$, alors T_n converge en probabilité vers l'infini lorsque $n \to \infty$, ce qui assure que la puissance du test converge vers 1 en tout point de H_1 (on dit que le test est consistant).

Le test de Kolmogorov-Smirnov à deux échantillons

On dispose de deux échantillons de variables aléatoires réelles iid $X_1, \ldots, X_{n_1}$ et $Y_1, \ldots, Y_{n_2}$. On note F_X la fonction de répartition des X_i et F_Y la fonction de répartition des Y_j.

On suppose que F_X et F_Y sont continues, et on veut tester

$$H_0 : F_X = F_Y \quad \text{contre} \quad H_1 : F_X \neq F_Y \, .$$

On note $F_{n_1,X}$ la fonction de répartition empirique des X_i, et $F_{n_2,Y}$ la fonction de répartition des Y_i. On utilise la statistique de test :

$$T_{n_1,n_2} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sup_{t \in \mathbb{R}} |F_{n_1,X}(t) - F_{n_2,Y}(t)| \, .$$

Le test de Kolmogorov-Smirnov à deux échantillons

Comme pour le test d'adéquation, on peut montrer que la statistique T_{n_1,n_2} est libre, c'est à dire que sa loi sous H_0 ne dépend pas de F_X, de sorte qu'elle peut-être tabulée.
On peut aussi montrer que, sous H_0, la statistique T_{n_1,n_2} converge en loi lorsque $n_1, n_2 \to \infty$ vers une loi connue (la loi du sup d'un processus Gaussien).
On rejette H_0 au niveau de risque α lorsque $t_{n_1,n_2} > q_{n_1,n_2,1-\alpha}$, où $q_{n_1,n_2,1-\alpha}$ est tel que

$$\mathbb{P}_{H_0}(T_{n_1,n_2} > q_{n_1,n_2,1-\alpha}) \sim \alpha\,.$$

En pratique, on utilise les quantiles de la loi exacte de T_{n_1,n_2} sous H_0 lorsque n_1, n_2 ne sont pas trop grands, et les quantiles de la loi asymptotique lorsque n_1, n_2 sont grands.

Enfin, le théorème de Glivenko Cantelli montre que si $F_X \neq F_Y$, alors T_{n_1,n_2} converge en probabilité vers l'infini lorsque $n_1, n_2 \to \infty$, ce qui assure que la puissance du test converge vers 1 en tout point de H_1.

Le test de Kolmogorov-Smirnov d'indépendance

On dispose d'un échantillon de couples de variables aléatoires iid $(X_1, Y_1), \ldots, (X_n, Y_n)$.

On suppose que F_X et F_Y sont continues, et on veut tester

H_0 : X_1 et Y_1 sont indépendantes, contre H_1 : X_1 et Y_1 ne le sont pas.

On note $F_{n,X}$ la fonction de répartition empirique des X_i, et $F_{n,Y}$ la fonction de répartition des Y_i. On note aussi $F_{n,X,Y}$ la fonction de répartition empirique bivariée

$$F_{n,X,Y}(s,t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leq t, Y_i \leq s}\,.$$

On utilise la statistique de test (Blum, Kiefer et Rosenblatt (1961)):

$$T_n = \sqrt{n} \sup_{t,s\in\mathbb{R}} |F_{n,X,Y}(t,s) - F_{n,X}(t)F_{n,Y}(s)|\,.$$

Le test de Kolmogorov-Smirnov d'indépendance

Comme pour le test d'adéquation, on peut montrer que la statistique T_n est libre, c'est à dire que sa loi sous H_0 ne dépend pas de F_X et F_Y, de sorte qu'elle peut-être tabulée.
On peut aussi montrer que, sous H_0, la statistique T_n converge en loi lorsque $n \to \infty$ vers une loi connue (la loi du sup d'un processus Gaussien).
On rejette H_0 au niveau de risque α lorsque $t_n > q_{n,1-\alpha}$, où $q_{n,1-\alpha}$ est tel que

$$\mathbb{P}_{H_0}(T_n > q_{n,1-\alpha}) \sim \alpha\,.$$

En pratique, on utilise les quantiles de la loi exacte de T_n sous H_0 lorsque n n'est pas trop grand, et les quantiles de la loi asymptotique lorsque n est grand.

Enfin, le théorème de Glivenko Cantelli (version bivarié) montre que si X_1 et Y_1 ne sont pas indépendantes, alors T_n converge en probabilité vers l'infini lorsque $n \to \infty$, ce qui assure que la puissance du test converge vers 1 en tout point de H_1.