Estimateur récursif de la fonction de lien dans un modèle semi-paramétrique

Thi Mong Ngoc NGUYEN[1,2]
Bernard BERCU[1,2] **et Jérôme SARACCO** [1,2,3]

[1] IMB, UMR CNRS 5251, Université Bordeaux 1
[2]Equipe CQFD, INRIA Bordeaux Sud-Ouest, France
[3] GREThA, UMR CNRS 5113, Université Montesquieu Bordeaux 4

9èmes Colloque "Jeunes Probabilistes et Statisticiens" - Mai 2010

Plan

Modèle de régression

- **Objectif :** Modéliser la liaison entre une variable à expliquer y et une variable explicative x.
- **Applications :** Nombreux domaines tels que l'économie, la biostatistique, les sciences de l'environnement, ...
- Deux grandes classes de modèles de régression sont omniprésentes : les modèles paramétriques et les modèles non paramétriques.

Modèle paramétrique : $y = f_\theta(x) + \varepsilon$

- **Objectif :** Estimer le paramètre θ.
- **Technique d'estimation :** Méthode du maximum de vraisemblance, méthode des moindres carrés, ...
- **Avantages spécifiques :** Ils permettent une interprétation claire de l'impact de la variable explicative sur la variable à expliquer.
- **Défauts spécifiques :**
  - Le choix d'un bon modèle paramétrique au vu des données n'est pas toujours évident.
  - Le modèle paramétrique choisi peut ne pas être en adéquation avec les données et peut donc parfois être très "éloigné" de la réalité de données $\Rightarrow$ les conclusions en découlant peuvent alors être erronées.

Modèle non paramétrique : $y = f(x) + \varepsilon$

- **Objectif :** Estimer la fonction de lien f.
- **Technique d'estimation :** Méthodes des estimateurs à noyau, des splines de lissage, des ondelettes, ...
- **Avantages spécifiques :** Ils offrent davantage de flexibilité (aucune hypothèse paramétrique n'est imposée dans ce modèle, seules des hypothèses de régularité sur f sont imposées).
- **Défauts spécifiques :**
  - Il faut estimer la fonction de lien le plus souvent au moyen de procédure de calculs intensifs en particulier en ce qui concerne la recherche des paramètres de lissage, ce qui est lourd en temps de calcul.
  - L'interprétation de la fonction de lien n'est pas toujours évidente.

Modèle semi-paramétrique : $y \in \mathbb{R}$, $x \in \mathbb{R}^p$,

$$Y_{n+1} = f(\theta' X_n) + \varepsilon_{n+1} \tag{1}$$

où : (i) le paramètre $\theta \in \mathbb{R}^p$, inconnu ;
(ii) le bruit $\varepsilon \perp x$, aucune hypothèse sur la distribution de ε ;
(iii) la fonction de lien f inconnue.

- **Objectif :** Estimer le paramètre θ et la fonction de lien f.
- **Technique d'estimation :**
  - Méthode **SIR** (Sliced Inverse Regression) permet d'estimer la partie paramétrique θ du modèle (1) sans avoir à estimer la fonction f.
  - Ensuite, la fonction de lien f peut être estimée via une méthode non paramétrique.

Les méthodes récursives d'estimation n'ont jamais été développées dans le cadre de ce modèle semi-paramétrique.

Notre Objectif :

- Proposer un estimateur récursif
  - de la direction θ dans (1) en adaptant au cadre récursif la méthode SIR ;
  - de la fonction lien f dans (1) en combinant l'estimateur de Nadaraya-Watson récursif de f à l'estimateur récursif de θ estimé par la méthode SIR récursive.

- Proposer quelques propriétés asymptotiques associées à nos estimateurs récusifs.

L'avantage des méthodes récursives

- Prendre en compte l'arrivée temporelle des informations et affiner ainsi au fil du temps les algorithmes d'estimation mis en œuvre.
- Il n'est pas nécessaire de relancer tous les calculs d'estimation des paramètres du modèle à chaque fois que la base de données est complétée par de nouvelles observations.
- Idée : utiliser les estimations calculées sur la base de données initiale et les remettre à jour en tenant uniquement compte des nouvelles données arrivant dans la base.
- Le gain en terme de temps de calcul peut être très intéressant et les applications d'une telle appproche sont nombreuses.

Méthode SIR

SIR = Slice Inverse Regression (Régression inverse par tranches)

- Sliced $\longrightarrow$ discrétisation (ou "tranchage") de y
  - va permettre de simplifier l'estimation des moments intervenant dans les propriétés géométriques,
  - ne modifie pas la partie paramétrique du modèle (1).
- Inverse $\longrightarrow$ utilisation de propriétés géométriques des moments "inverse" de x sachant y : $\mathbb{E}[x \mid y]$ et $\mathbb{V}[x \mid y]$.

$\Rightarrow$ avantage : la dimension du problème a été réduite ;
$\Rightarrow$ coût à payer : rajouter une hypothèse :

(H) *La variable explicative x possède une distribution de probabilité non dégénérée telle que, $\forall b \in \mathbb{R}^p$, $\mathbb{E}[b'x \mid \theta'x]$ est linéaire en $\theta'x$.*
(vérifiée lorsque x suit une distribution elliptique).

Estimateur récursif de la direction de θ

- Remarque : Le paramètre θ n'est pas totalement identifiable, seule la direction de θ est d'identifiable
  $\Rightarrow$ direction EDR (Effective Dimension Reduction).
- Vecteur propre $\tilde{\theta}$ associé à la valeur propre non nulle de $\Sigma^{-1}\Gamma$ est colinéaire à $\theta \Rightarrow \tilde{\theta}$ est une direction EDR
  (où : $\Sigma = \mathbb{V}(x)$ et $\Gamma = \mathbb{V}(\mathbb{E}[x \mid \mathbb{T}(y)])$).
- Échantillon : $\{(x_i, y_i), i = 1, \ldots, n\}$ de v.a iid (x, y) issues du (1).
- Scinder cet échantillon en 2 parties : le sous-échantillon $\{(x_i, y_i), i = 1, \ldots, n-1\}$ et une nouvelle observation (x_n, y_n).
- Discrétisation de y en 2 tranches distinctes s_1 et s_2,
  supposons que (x_n, y_n) est telle que $y_n \in s_{h^*}$ avec $h^* = 1$ ou 2.

Estimateur récursif de la direction de θ

Estimateur récursif $\hat{\theta}_n$ de $\tilde{\theta}$:

$$\begin{aligned}\hat{\theta}_n &= \frac{n}{n-1}\hat{\theta}_{n-1} - \frac{n}{(n-1)(n+\rho_n)}\Sigma_{n-1}^{-1}\Psi_n\Psi_n'\hat{\theta}_{n-1} \\ &- \frac{(-1)^{h^*}n}{(n_{h^*,n-1}+1)(n-1)}\left(\Sigma_{n-1}^{-1} - \frac{1}{n+\rho_n}\Sigma_{n-1}^{-1}\Psi_n\Psi_n'\Sigma_{n-1}^{-1}\right)\Psi_{h^*,n}.\end{aligned}$$

où : $\Psi_n = x_n - \bar{x}_{n-1}$;

$\Psi_{h^*,n} = x_n - m_{h^*,n-1}$;

$\rho_n = \Psi_n'\Sigma_{n-1}^{-1}\Psi_n$.

Estimateur de Nadaraya-Watson récursif

$$\text{Si } \sum_{i=1}^{n} \frac{1}{h_{i-1}} K\left(\frac{x - X_{i-1}}{h_{i-1}}\right) \neq 0,$$

$$\hat{f}_n(x) = \frac{1}{\sum_{i=1}^{n} \frac{1}{h_{i-1}} K\left(\frac{x - X_{i-1}}{h_{i-1}}\right)} \sum_{i=1}^{n} \frac{1}{h_{i-1}} K\left(\frac{x - X_{i-1}}{h_{i-1}}\right) Y_i$$

Autre écriture :

$$\hat{f}_{n+1}(x) = \hat{f}_n(x) + \frac{1}{\sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - X_i}{h_i}\right)} \frac{1}{h_n} K\left(\frac{x - X_n}{h_n}\right)(Y_{n+1} - \hat{f}_n(x)).$$

Posons $\Phi_n = \theta' X_n$,

$\hat{\theta}_n$: estimateur récursif de θ

$\Rightarrow \hat{\Phi}_n = \hat{\theta}'_n X_n$: prédicteur de Φ_n.

En combinant l'estimateur de Nadaraya-Watson récursif de f à l'estimateur récursif de θ, nous avons : $\forall z \in \mathbb{R}$,

$$\hat{f}_n(z) = \frac{1}{\sum_{i=1}^{n} \frac{1}{h_{i-1}} K\left(\frac{z - \hat{\Phi}_{i-1}}{h_{i-1}}\right)} \sum_{i=1}^{n} \frac{1}{h_{i-1}} K\left(\frac{z - \hat{\Phi}_{i-1}}{h_{i-1}}\right) Y_i.$$

Autre écriture :

$$\hat{f}_{n+1}(z) = \hat{f}_n(z) + \frac{1}{\sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{z - \hat{\Phi}_i}{h_i}\right)} \frac{1}{h_n} K\left(\frac{z - \hat{\Phi}_n}{h_n}\right)(Y_{n+1} - \hat{f}_n(z)).$$

Résultats asymptotiques pour $\hat{\theta}_n$

- **Hypothèses :**

(A1) Les observations $(x_i, y_i), i = 1, ..., n$, sont échantillonnées de manière indépendante à partir du modèle (1).

(A2) Le support de y est partionné en deux tranches fixes s_1 et s_2 telles que $\mathbb{P}(y \in s_h) \neq 0$ pour $h = 1, 2$.

- **Résultats de convergence :**

Théorème (Convergence presque sûrement)

Sous les hypothèses **(H)**, *(A1) et (A2), nous avons*

$$|| \hat{\theta}_n - \tilde{\theta} ||= \mathcal{O}\left(\sqrt{\frac{\log(\log n)}{n}}\right) \qquad p.s.,$$

où le vecteur $\tilde{\theta}$ *est colinéaire à* θ.

Théorème (Convergence en loi)

Sous les hypothèses **(H)**, $(A1)$ *et* $(A2)$, *nous avons :*

$$\sqrt{n}(\hat{\theta}_n - \tilde{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma^{-1}\Delta_3\Sigma^{-1}),$$

où Δ_3 *peut être calculée explicitement.*

Résultats asymptotiques pour $\hat{f}_n$

- **Hypothèses :**
  - $(H.1)$ Fenêtre $h_n = n^{-\alpha}$ est positive telle que :

$$\lim_{n\to\infty} h_n = 0 \qquad et \qquad \lim_{n\to\infty} nh_n = \infty.$$

  - $(H.2)$ Noyau K est un noyau à support compact, mesurable, positif et borné satisfaisant :

$$\int_{\mathbb{R}} K(x)dx = 1, \int_{\mathbb{R}} \mid x \mid K(x)dx < +\infty, \int_{\mathbb{R}} K^2(x)dx = \tau^2$$

  - $(H.3)$ Fonction de lien f est une fonction Lipschitzienne, bornée et deux fois continûment dérivable sur $\mathbb{R}$.

- **Résultats de convergence :** En utilisant le résultat de convergence presque sûrement de $\hat{\theta}_n$, nous avons :

Théorème (Convergence presque sûrement)

Sous les hypothèses $(H.1) - (H.3)$, *nous avons quand* $n \to \infty$ *:*

$$|| \hat{f}_n(z) - f(z) || = \mathcal{O}\left(n^{2\alpha} \sqrt{\frac{\log(\log n)}{n}} \right) \quad p.s..$$

Théorème (Convergence en loi)

Sous les hypothèses $(H.1) - (H.3)$, *supposons que* $\mathbb{E}(Y^2) < \infty$, $\forall \alpha \in]1/3, 1/2[$ *et* $h(\Phi) > 0$, $\forall z \in \mathbb{R}$, *nous avons quand* $n \to \infty$ *:*

$$\sqrt{nh_n} \left(\hat{f}_n(z) - f(z) \right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2 \tau^2}{h(\Phi)(1+\alpha)} \right).$$

où $\sigma^2 = \mathbb{E}[\varepsilon_n^2 \mid \mathbb{F}_{n-1}]$ *et* $h(\Phi)$ *est la densité de* (Φ_n).

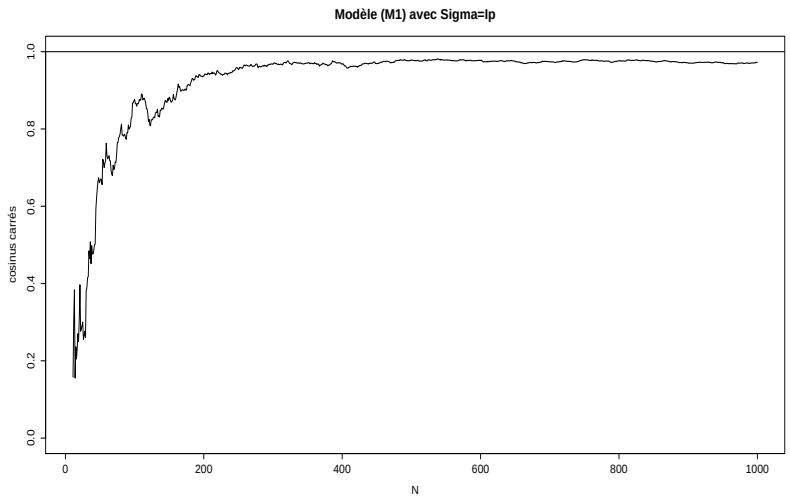
Résultats de simulation pour $\hat{\theta}_n$

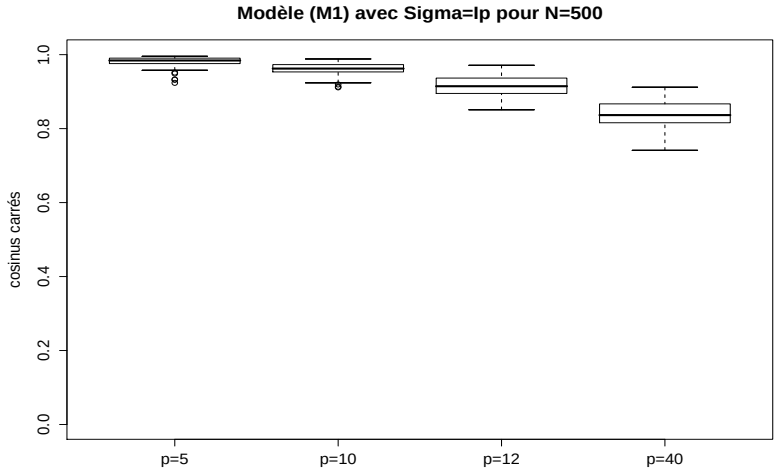
- **Objectif :** Étudier le comportement numérique de l'estimateur récursif $\hat{\theta}_n$ (la convergence de $\hat{\theta}_n$ vers la vraie direction θ du modèle).
- **Modèle simulé :** (M1) : $y = (\theta' x)^3 + \varepsilon$
  avec $x \sim \mathcal{N}_p(0, I_p)$, $\theta = (1, -1, 0, \ldots, 0) \in \mathbb{R}^p$, $\varepsilon \sim \mathcal{N}(0, 1)$.
- **Motivations :**
  - Montrer l'évolution, en fonction la taille de l'échantillon n, de la qualité de l' estimateur récursif, et l'effet de la dimension p de x sur la qualité de l'estimation.
  - Illustrer la normalité asymptotique de l'estimateur récursif.

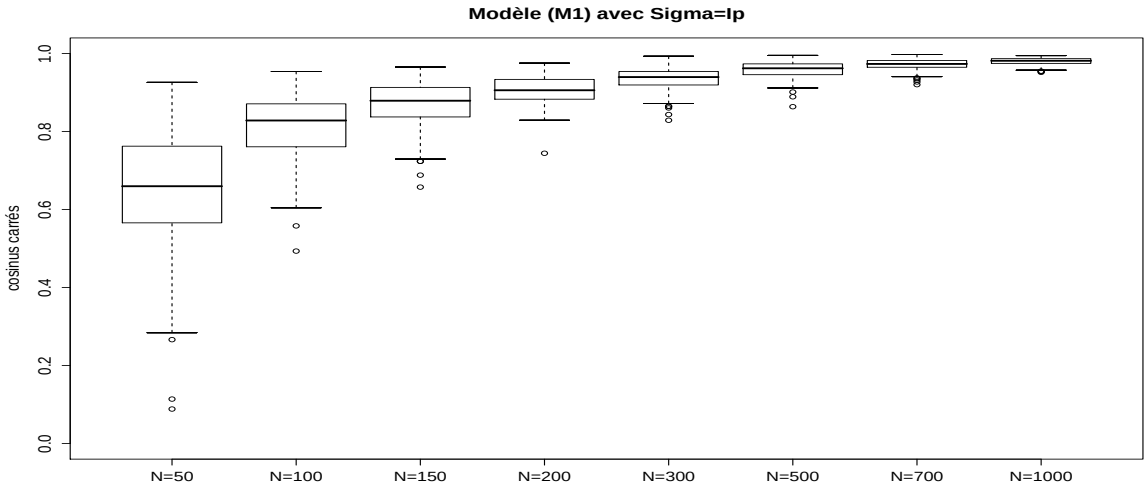
La qualité de l'estimation sera mesurée par

$$\cos^2(\hat{\theta}_n, \theta) = \frac{(< \hat{\theta}_n, \theta >)^2}{|| \hat{\theta}_n || \times || \theta ||}.$$

Plus $\cos^2(\hat{\theta}_n, \theta)$ est proche de 1, meilleure est la qualité de l'estimation.







Évolution de la qualité de l'estimateur de $\hat{\theta}_n$, en fonction de n et en fonction de p, sur un échantillon et sur 500 échantillons.

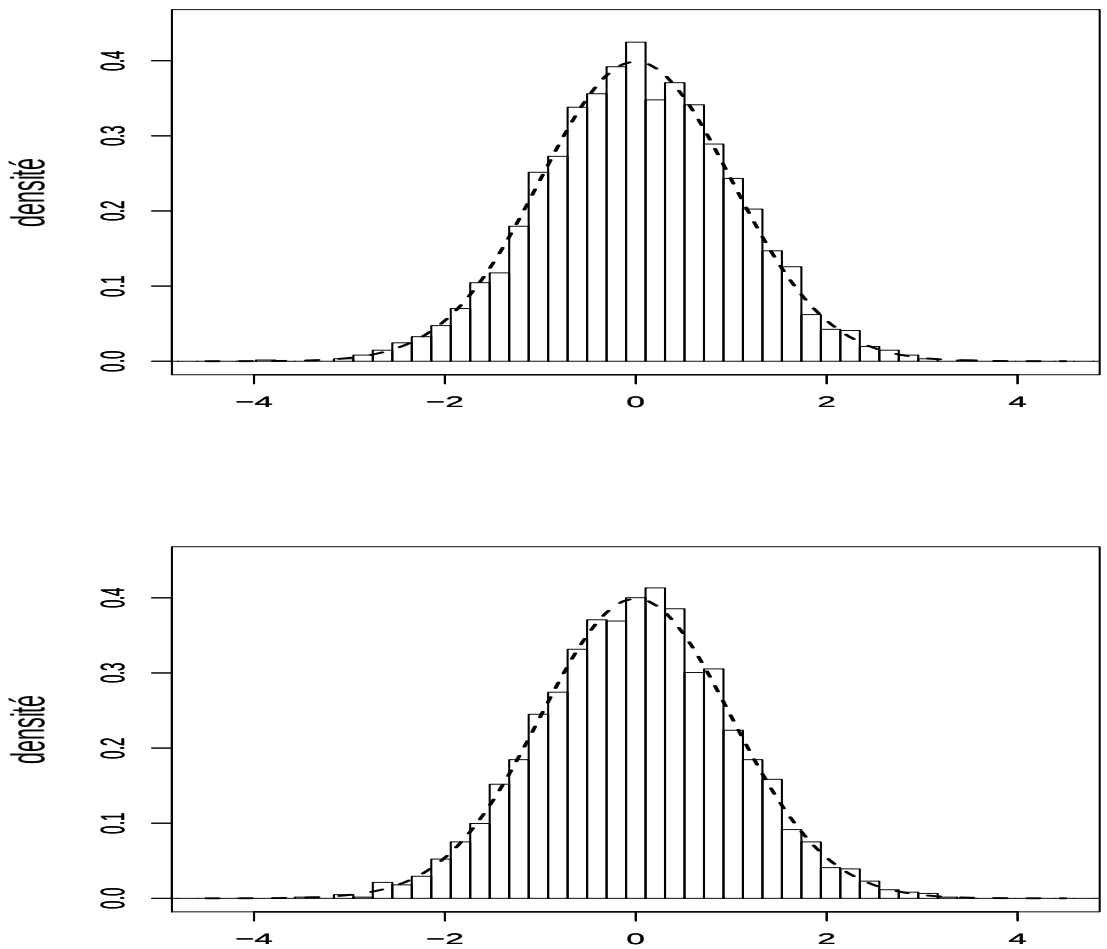


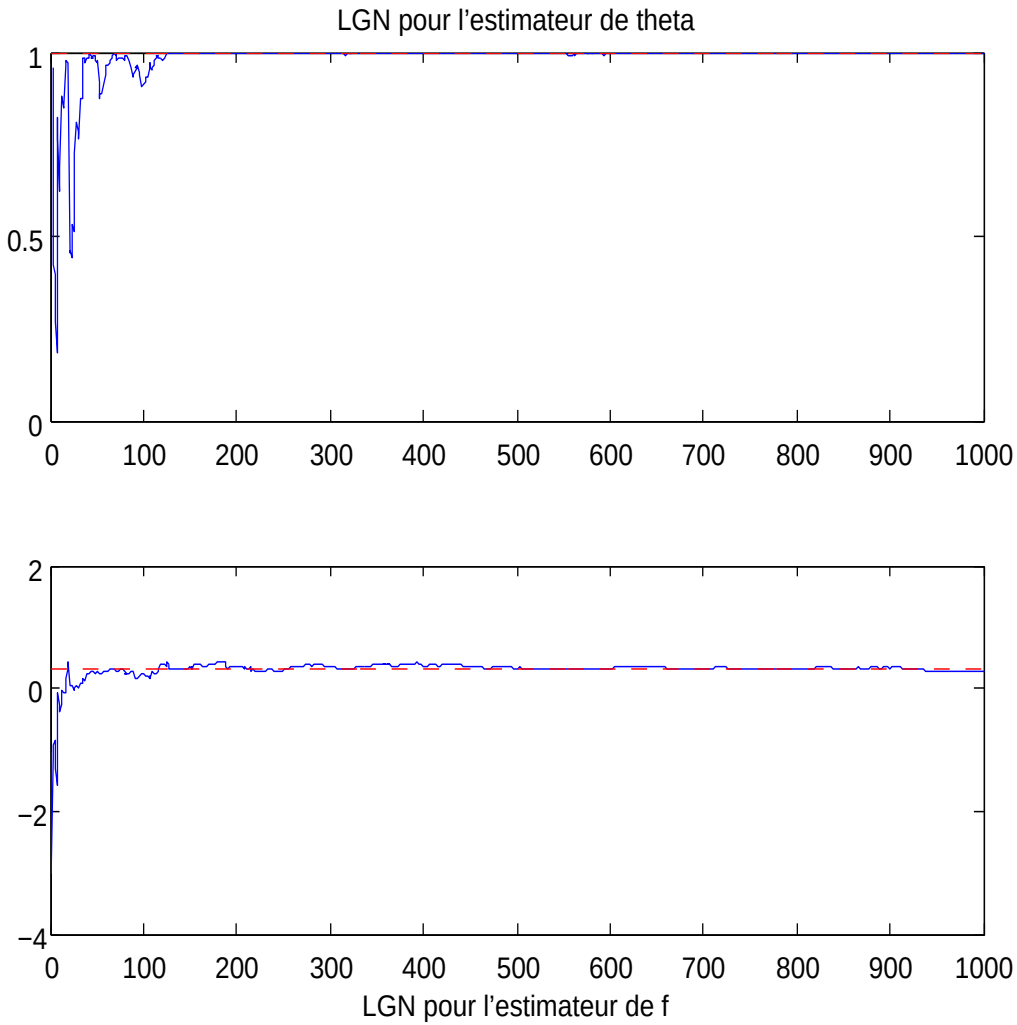
Illustration de la normalité asymtotique pour 2 composantes de $\hat{\theta}_n$
(le graphe de la densité de la loi $\mathcal{N}(0,1)$, en pointillé, est superposé à l'histogramme).

Résultats de simulation pour $\hat{f}_n$

- **Objectif :** Étudier le comportement numérique de l'estimateur récursif $\hat{f}_n$ en combinant l'estimateur de Nadaraya-Watson récursif de f à l'estimateur récursif $\hat{\theta}_n$.

- **Modèle simulé :** (M2) : $y = (\theta' x) exp(-\theta' x) + \varepsilon$ avec $x \sim \mathcal{N}(m, \sigma^2)$, $\theta \in [-10; 10]) \in \mathbb{R}^p$, $\varepsilon \sim \mathcal{N}(0, 1)$.

- **Motivations :**
  - Montrer l'évolution, en fonction la taille de l'échantillon n, de la qualité de l' estimateur récursif.
  - Illustrer la normalité asymptotique de l'estimateur récursif.



Évolution,en fonction de n, de la qualité de l'estimateur de $\hat{\theta}_n$ et de $\hat{f}_n$.

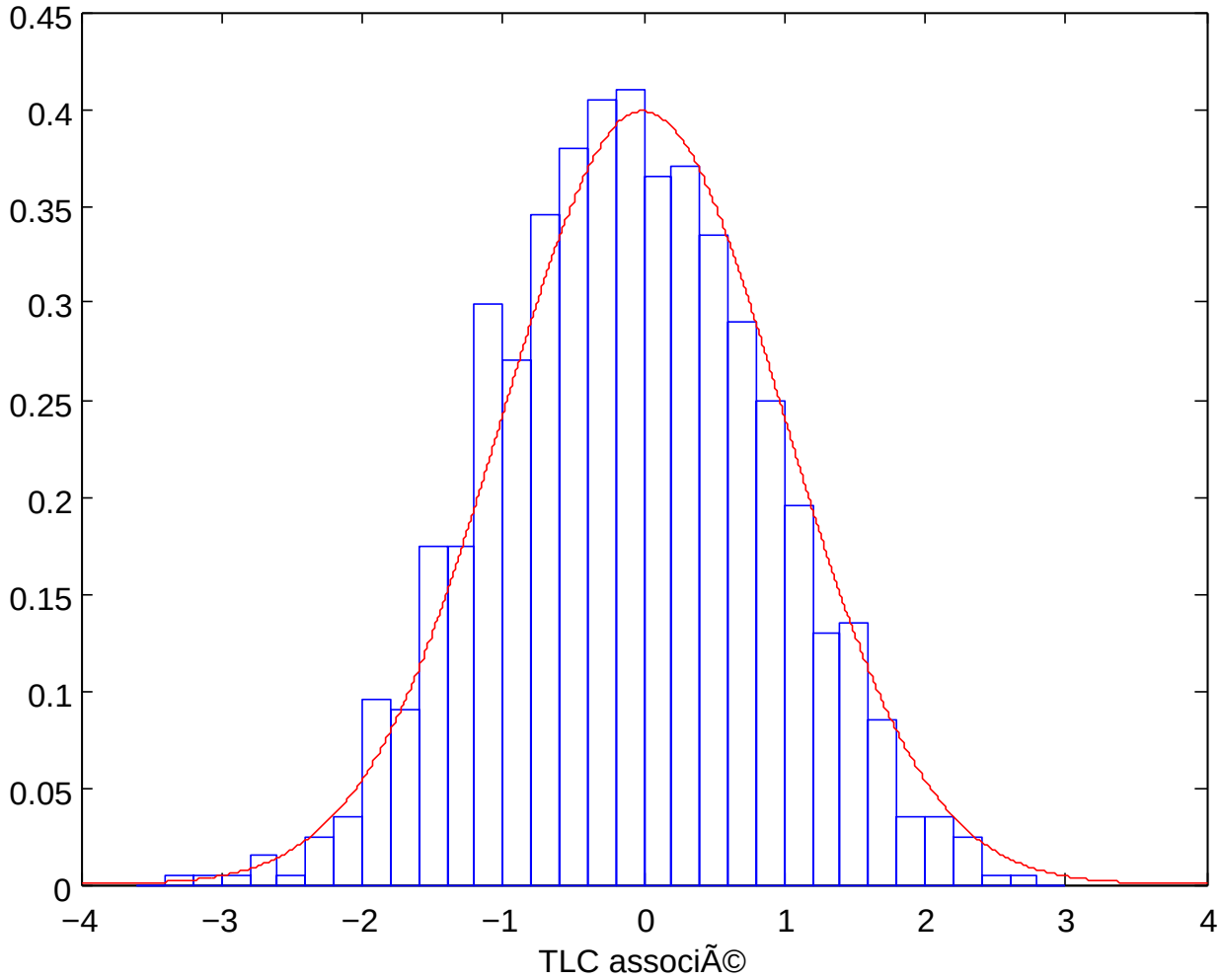


Illustration de la normalité asymtotique de $\hat{f}_n(z)$
dans le modèle (M2) avec le choix de noyau Gaussien et de fenêtre $n^{-0.45}$.

- Conclusion
  - Les estimateurs récursifs proposés semblent bien fonctionner numériquement pour des tailles d'échantillons raisonnables et même lorsque la dimension de la covariable x est importante.
  - Nous obtenons bien la normalité asymptotique des estimateurs proposés.
- Perspective

  Le choix de la fenêtre $h_n = n^{-\alpha}$ est crucial. Nous continuons à travailler sur la partie théorique afin d'élargir l'intervalle $\alpha \in]1/3; 1/2[$.

MERCI DE VOTRE ATTENTION !