Estimation adaptative de densité dans un modèle de transformation nonlinéaire

Tabea Rebafka & Fabienne Comte

LPMA, Université Paris 6, UPMC, 4 place Jussieu, 75252 Paris cedex 05
& MAP5, Université Paris Descartes, 45 rue des Saints-Pères, 75270 Paris cedex 06

Résumé. Nous considérons le problème d'estimation nonparamétrique de densité pour un problème inverse motivé par une application en fluorescence résolue dans le temps. La densité à estimer est perturbée, d'une part, par un bruit additif et, d'autre part, par une distorsion nonlinéaire. Nous proposons un estimateur qui prend en compte ces deux difficultés. Par ailleurs, le problème de la sélection de modèle est étudié. Nous proposons un estimateur adaptatif, et une inégalité d'oracle pour le risque quadratique est établie. La performance de l'estimateur est évalué sur une série de mesures de fluorescence.
Abstract. The problem of nonparametric density estimation in an inverse problem is considered motivated by an application in time-resolved fluorescence. The target density is disturbed, on the one hand, by some additive noise, on the other hand, by a nonlinear distortion. An estimator is proposed that takes into account both difficulties. Moreover, the problem of model selection is addressed. An adaptive estimator is developed and an oracle inequality for the square loss is established. The performance of the estimator is evaluated on fluorescence lifetime measurements.
Mots-clés. Modèles semi et non-paramétriques ; Biostatistique.

1 Introduction

Nous considérons un problème d'estimation nonparamétrique d'une densité dans un problème inverse caractérisé, d'une part, par un bruit additif et, d'autre part, par une distorsion nonlinéaire. Notons par f_Y la densité (par rapport à la mesure de Lebesgue) à estimer, et par F_Y sa fonction de répartition. Nous supposons que la loi est positive.

Soit f_η la densité connue d'un bruit positif. Soient $Y_1, \dots, Y_n$ un échantillon i.i.d. de f_Y, et $\eta_1, \dots, \eta_n$ un échantillon i.i.d. de η, indépendant du premier échantillon. Nous définissions les variables aléatoires $X_i = Y_i + \eta_i$ de densité $f_X = f_Y \otimes f_\eta$.

Dans le *modèle* dit *de transformation nonlinéaire* on observe une loi G qui résulte d'une distorsion nonlinéaire d'une fonction de répartition F par $G = 1 - M \circ \bar{F}$, où $\bar{F} = 1 - F$ est la fonction de survie de F [8, 6]. La fonction M, ici supposée connue, est bijective et croissante telle que $\bar{G} = M \circ \bar{F}$ est une fonction de survie pour toute fonction de survie $\bar{F}$. Nous supposons en plus que M soit deux fois continûment dérivable et ses dérivées $\dot{M}$ et $\ddot{M}$ sont bornée par des constantes $0 < a < b < +\infty$ telles que $a < \dot{M} < b$ et $a < \ddot{M} < b$.

Dans ce papier, nous observons un modèle de transformation nonlinéaire appliqué à la loi convoluée $F_X = F_Y \otimes F_\eta$. Autrement dit, on dispose d'un échantillon i.i.d. $Z_1, \dots, Z_n$

de la loi $G = 1 - M \circ (\overline{F_Y \otimes F_\eta})$. Le but est d'estimer la densité f_Y. Nous présentons un estimateur nonparamétrique adaptatif de la densité f_Y, qui a été développé dans [1], et nous évaluons sa performance sur une série de données réelles en fluorescence.

Modèle d'empilement Pour donner un exemple, nous considérons le *modèle* dit *d'empilement*. Dans ce modèle on observe des réalisations indépendantes de la variable aléatoire Z définie comme le minimum d'un nombre aléatoire de variables

$$Z = \min\{Y_1 + \eta_1, \ldots, Y_N + \eta_N\} \ ,$$

où les $(Y_i)_i$ et les $(\eta_i)_i$ sont deux échantillons indépendants de la loi f_Y et f_η, respectivement, et N est une variable aléatoire à valeurs dans $\{1, 2, \ldots\}$, indépendante des $(Y_i)_i$ et $(\eta_i)_i$. On peut montrer que la loi de Z est donnée par un modèle de transformation nonlinéaire avec $F = F_Y \otimes F_\eta$. La fonction M est en effet la fonction génératrice de la loi de N donnée par $M(u) = \mathbb{E}[u^N]$ pour tout $u \in [0, 1]$.

Le modèle d'empilement est rencontré entre autres en fluorescence résolue dans le temps [2, 7], où les Y_i représentent les durées de vies associées à différentes molécules et η_i sont des durées aléatoires en provenance de l'appareil de mesure. Par suite, les $X_i + \eta_i$ représentent des temps d'arrivée de différents photons de fluorescence sur un capteur. Pour des raisons techniques, on ne peut observer que le tout premier photon qui arrive sur le capteur, car celui-ci aveugle l'appareil de mesure. Typiquement, on suppose une loi de Poisson $\mathcal{P}(\lambda)$ pour le nombre N de photons par mesure [3]. Plus précisément, c'est une loi de Poisson renormalisée après avoir écarté l'événement $\{N = 0\}$, ce qui veut dire que $\mathbb{P}(N = n) = \lambda^n / n! / (\mathrm{e}^\lambda - 1)$ pour tout $n = 1, 2, \ldots$. Nous remarquons qu'en fluorescence il est possible d'estimer le paramètre λ à part avec grand précision. Donc, on peut considérer λ connu, ainsi que la fonction génératrice $M(u) = (\mathrm{e}^{\lambda u} - 1)/(\mathrm{e}^\lambda - 1)$. Dans ce contexte, d'après [5], la densité f_Z s'écrit

$$f_Z(z) = \lambda (f_Y \otimes f_\eta)(z) \ \mathrm{e}^{-\lambda (F_Y \otimes F_\eta)(z)} \ , \quad z > 0 \ .$$

Apparemment, l'inversion de cette formule pour obtenir f_Y n'est pas évidente et nécessite un peu d'ingéniosité.

2 Estimateur non paramétrique de densité

Nous proposons un estimateur de la densité f_Y qui est d'une part basée sur la transformée de Fourier et d'autre part sur la propriété suivante du modèle de transformation nonlinéaire.

Sous les hypothèses ci-dessus, la fonction M d'un modèle de transformation nonlinéaire est inversible, et par conséquent, on a $\bar{F} = M^{-1} \circ \bar{G}$ pour toute fonction de survie $\bar{F}$. Pour les densités f et g associées, il découle la relation $f = g(M^{-1})' M \circ \bar{G} = g / \dot{M} \circ M^{-1} \circ \bar{G}$.

Cela implique pour toute fonction intégrable h une relation intéressante entre les moments de la loi de F et de G. En effet, on a

$$\mathbb{E}[h(X)] = \mathbb{E}[h(Z)w \circ G(Z)] \,, \tag{1}$$

où X suit la loi F, Z suit la loi G et $w(u) = 1/\dot{M} \circ M^{-1}(1-u)$. Cette propriété permet de construire des estimateurs des moments de la loi de F en utilisant des observations de la loi de G. Pour cela on remplace l'espérance à droite par la moyenne empirique et la fonction de répartition G par sa version empirique $\hat{G}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Z_i \leq t\}$, ce qui nous mène à un estimateur de $\mathbb{E}[h(X)]$ donné par

$$\frac{1}{n}\sum_{j=1}^n h(Z_j)w \circ (\hat{G}_n(Z_j)) = \frac{1}{n}\sum_{j=1}^n h(Z_{(j)})w(j/n) \,,$$

où on utilise que les statistiques d'ordre $Z_{(1)} \leq Z_{(2)} \leq \ldots Z_{(n)}$ vérifient $\hat{G}_n(Z_{(j)}) = j/n$. Remarquons que l'estimateur est bien défini car la fonction w est connue, car M est connue. Pour plus de détails sur cette relation entre les moments nous référons à [4].

On dénote par $t^*(u) = \int t(x)\mathrm{e}^{-iux}\mathrm{d}x$ la transformée de Fourier d'une fonction intégrable t. Dans notre modèle on a $f_Y^* = f_X^*/f_\eta^*$. D'après la propriété (1), on obtient la relation $f_X^*(u) = \mathbb{E}\left[\mathrm{e}^{-iuX}\right] = \mathbb{E}\left[\mathrm{e}^{-iuZ}w \circ G(Z)\right]$, où X suit la loi $F_Y \otimes F_\eta$ et Z suit la loi $G = 1 - M \circ (\overline{F_Y \otimes F_\eta})$. On obtient donc un estimateur de f_X^* basé sur des observations $Z_1, \ldots, Z_n$ d'un modèle de transformation nonlinéaire de loi G suivant l'astuce décrite ci-dessus en définissant

$$\bar{f}_X^*(u) = \frac{1}{n}\sum_{j=1}^n \mathrm{e}^{-iuZ_j}w \circ (\hat{G}_n(Z_j)) = \frac{1}{n}\sum_{j=1}^n w(j/n)\mathrm{e}^{-iuZ_{(j)}} \,.$$

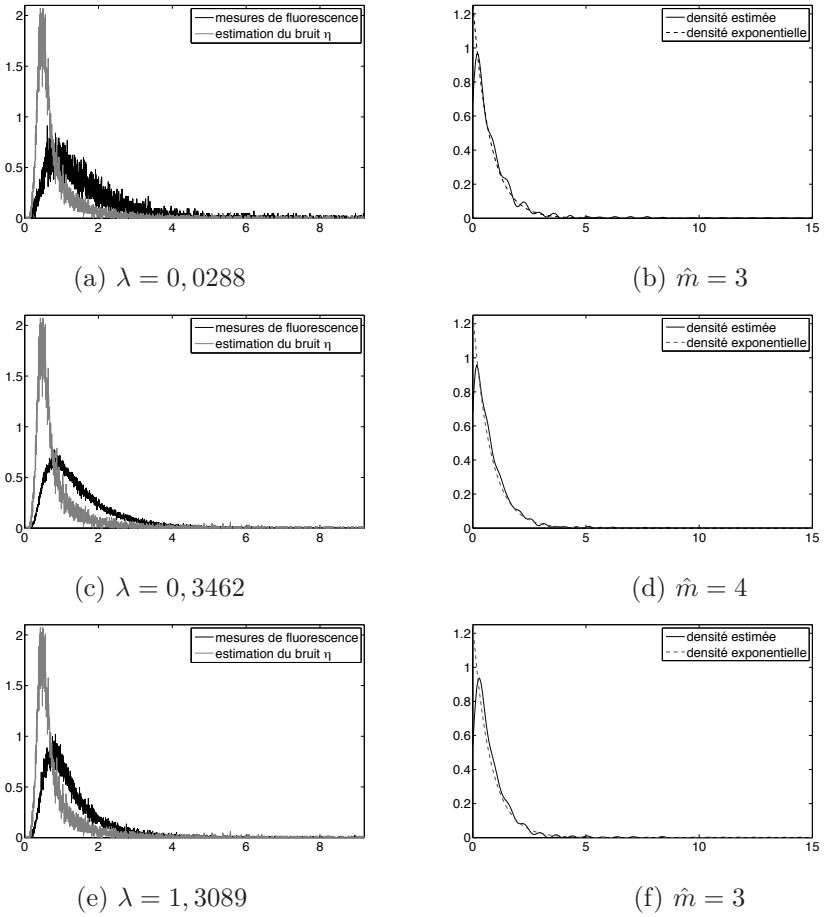
Cette estimateur est maintenant injecté dans la transformée de Fourier inverse, $f_Y(x) = (2\pi)^{-1}(f_X^*/f_\eta^*)^*(-x)$, pour arriver à l'estimateur suivant de la densité f_Y

$$\hat{f}_m(x) = \frac{1}{2\pi}\int_{-m\pi}^{m\pi} \mathrm{e}^{iux}\frac{\bar{f}_X^*(u)}{f_\eta^*(u)}\mathrm{d}u \,.$$

L'intégrale est tronquée afin d'éviter des éventuels problèmes d'intégrabilité liés au fait que $f_\eta^*(u)$ tend vers 0 lorsque $|u|$ tend vers $+\infty$. Le paramètre m est donc un paramètre de troncature, ayant un impact sur le biais et la variance de l'estimateur. En effet, le biais dépend de la régularité de f_Y, alors que la variance est proportionnelle à

$$\Delta_\eta(m) = \frac{1}{n}\int_{-m\pi}^{m\pi} \frac{1}{|f_\eta^*(u)|^2}\mathrm{d}u \,.$$

Le théorème suivant donne une borne pour le risque quadratique de l'estimateur $\hat{f}_m$.



(a) $\lambda = 0,0288$ (b) $\hat{m} = 3$

(c) $\lambda = 0,3462$ (d) $\hat{m} = 4$

(e) $\lambda = 1,3089$ (f) $\hat{m} = 3$

FIGURE 1 – (a), (c) et (e) présentent les mesures de fluorescence pour des différentes valeurs de λ et des observations indépendantes du bruit η. Les figures (b), (d) et (f) montrent les densités estimées correspondantes en comparaison avec une densité exponentielle de paramètre $0,76$.

Théorème 1 *Soit w une fonction Lipschitzienne qui vérifie $0 < w_0 \leq w(u) \leq w_1 < \infty$ pour tout $u \in [0,1]$. Notons par $f_{Y,m}$ la fonction qui vérifie $f^*_{Y,m} = f^*_Y \mathbb{1}_{[-\pi m, \pi m]}$. Alors*

$$\mathbb{E}(\|\hat{f}_m - f_Y\|^2) \leq \|f_Y - f_{Y,m}\|^2 + C\frac{\Delta_\eta(m)}{n} ,$$

où C dépend de $\int_0^1 w^2(u)\mathrm{d}u$ et de la constante Lipschitzienne c_w de w.

3 Estimation adaptative

Afin de construire un estimateur adaptatif, où le choix du paramètre m est basé sur les données, il est utile de constater que l'estimateur $\hat{f}_m$ peut être dérivé autrement. En

effet, $\hat{f}_m$ minimise le contraste $\gamma_n(h)$ donné par

$$\gamma_n(h) = \|h\|^2 - \frac{1}{\pi}\int h^*(-u)\frac{\widehat{f_X^*}(u)}{f_\eta^*(u)}\mathrm{d}u$$

sur l'ensemble des fonctions $S_m = \{h,\ \ \mathrm{support}(h^*) \subset [-\pi m, \pi m]\}$. L'approche générale pour la sélection de modèle consiste à trouver une pénalité pen(.) telle que le modèle

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n}(\gamma_n(\hat{f}_m) + \mathrm{pen}(m)) \tag{2}$$

atteint le compromis biais-variance sur une collection de modèle $\mathcal{M}_n$. Pour un choix approprié de pénalité, nous montrons une inégalité d'oracle. En revanche, en comparaison avec des résultats classiques, il apparaît un terme supplémentaire d'ordre $\ln(n)$.

Théorème 2 *Soient f_Y de carré intégrable sur $\mathbb{R}$ et η tel que $c_0(1+u^2)^{-\gamma} \leq |f_\eta^*(u)|^2 \leq C_0(1+u^2)^{-\gamma}$. Soit w une fonction Lipschitzienne qui vérifie $0 < w_0 \leq w(u) \leq w_1 < \infty$ pour tout $u \in [0,1]$. Soit $\hat{m}$ l'estimateur définie par (2) avec*

$$\mathrm{pen}(m) = \kappa\left(\int_0^1 w^2(u)\mathrm{d}u + \kappa' c_w^2 \ln(n)\right)\frac{\Delta_\eta(m)}{n}\ ,$$

où $\kappa > 0$ et $\kappa' > 0$. Soit la collection de modèles décrite par $\mathcal{M}_n = \{m \in \mathbb{N}, \Delta_\eta(m) \leq n\}$. Alors, il existe des constantes κ et κ' telles que

$$\mathbb{E}\left(\|\hat{f}_{\hat{m}} - f_Y\|^2\right) \leq C\left(\inf_{m\in\mathcal{M}_n}\|f_Y - f_{Y,m}\|^2 + \mathrm{pen}(m)\right) + C'\frac{\ln(n)}{n}\ ,$$

où $C > 0$ est une constante et $C' > 0$ dépend de c_w et des bornes de w.

4 Application numérique

Nous appliquons l'estimateur adaptatif $\hat{f}_{\hat{m}}$ à des mesures de fluorescence. Le modèle sous-jacent est un modèle d'empilement où N suit une loi de Poisson renormalisée $\mathcal{P}(\lambda)$. Nous disposons de trois échantillons de mesures de fluorescence pour des valeurs différentes de λ. Plus λ est élevé, plus la distorsion nonlinéaire est forte. Pour $\lambda = 0,0288$, on dispose de 13.648 observations, pour $\lambda = 0,3462$ on a 140.449 observations et pour $\lambda = 1,3089$ la taille d'échantillon est 58.392. Dans cette application, nous disposons d'un échantillon $\eta_1, \ldots, \eta_p$ de la loi du bruit η qui est indépendant des mesures de fluorescence et qui sert à estimer la densité f_η. La taille de cet échantillon est de 21.197.

Après calibration sur des données simulées, l'estimateur adaptatif $\hat{f}_{\hat{m}}$ est appliqué avec des constantes $\kappa = 50$ et $\kappa' = 0.01$. Pour le premier et le dernier échantillon, le modèle sélectionné est $\hat{m} = 3$, alors que dans le deuxième cas on a $\hat{m} = 4$. Ceci est cohérent

avec le fait que le deuxième échantillon est le plus grand. Ici, la densité f_Y est la même pour tous les trois échantillons. D'après les physiciens qui ont effectué l'expérience, f_Y serait une loi exponentielle d'un paramètre environ $0,76$. La Figure 1 présente, d'une part, les mesures et, d'autre part, les densités estimées en comparaison avec une densité exponentielle. On voit que les trois densités estimées se ressemblent beaucoup malgré les différentes valeurs de λ. Par ailleurs, dans les trois cas, la densité estimée approche bien la densité exponentielle, bien qu'un effet de bord près de zéro est observé.

En conclusion, le nouvel estimateur $\hat{f}_{\hat{m}}$ exhibe une très bonne performance sur des données réelles en fluorescence. Plus particulièrement, la valeur de λ ne semble pas influencer les résultats d'estimation.

Références

[1] F. Comte and T. Rebafka. Adaptive density estimation in the pile-up model involving measurement errors. article soumis, disponible à `http://arxiv.org/abs/1011.0592`, 2010.

[2] J. R. Lakowicz. *Principles of Fluorescence Spectroscopy*. Academic/Plenum, New York, 1999.

[3] D. V. O'Connor and D. Phillips. *Time-correlated single photon counting*. Academic Press, London, 1984.

[4] T. Rebafka, F. Roueff, and A. Souloumiac. A corrected likelihood approach for the pile-up model with application to fluorescence lifetime measurements using exponential mixtures. *The International Journal of Biostatistics*, 6(1), 2010.

[5] T. Rebafka, F. Roueff, and A. Souloumiac. Information bounds and MCMC parameter estimation for the pile-up model. *Journal of Statistical Planning and Inference*, 141(1) :1–16, 2011.

[6] A. Tsodikov. A generalized self-consistency approach. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(3) :759–774, 2003.

[7] Bernard Valeur. *Molecular Fluorescence*. Wiley-VCH, Weinheim, 2002.

[8] A. Yakovlev and A. Tsodikov. *Stochastic Models of Tumor Latency and their Biostatistical Applications*. World Scientific, 1996.