Estimation non-paramétrique dans un modèle de censure multiplicative avec un bruit symétrique

Charlotte Dion [1,2] & Fabienne Comte [2]

[1] *LJK, UMR CNRS 5224, Université Joseph Fourier Grenoble, charlotte.dion@imag.fr*
[2] *MAP5, UMR CNRS 8145, Université Paris Descartes Paris*

Résumé. Dans ce travail nous considérons le modèle : $Y_i = X_i U_i$, $i = 1, \ldots, n$, où les X_i, U_i et donc les Y_i sont tous indépendants et identiquement distribués. Les X_i ont pour densité f et sont les variables d'intérêt, les U_i sont des variables de bruit multiplicatif de densité uniforme sur l'intervalle $[1-a, 1+a]$, avec $0 < a < 1$, et ces deux séquences sont indépendantes. Cependant, seulement les Y_i sont observés. Pour la protection de l'information on peut utiliser un modèle de bruitage multiplicatif, uniforme. L'enjeu est alors de masquer assez les données tout en sachant retrouver l'information importante contenue dans les données d'origine. Nous nous intéressons à l'estimation non-paramétrique de la densité f puis de la fonction survie associée. Notre procédure conduit à un estimateur par projection d'une fonction auxiliaire, dont on déduit un estimateur de la fonction visée. On établit dans chaque cas une borne du risque quadratique intégré, qui montre que le compromis entre le paramètre de dimension et le pas d'estimation doit être fait. Nous proposons alors une méthode de sélection de modèle. Enfin nous prouvons une borne supérieure pour le risque des estimateurs finaux. Une étude sur simulation puis données réelles illustre notre propos et l'intérêt de ce modèle.

Mots-clés. Données censurée, Sélection de modèle, Bruit multiplicatif, Estimateur non-paramétrique

Abstract. In this work we consider the model $Y_i = X_i U_i$, $i = 1, \ldots, n$, where the X_i, the U_i and thus the Y_i are all independent and identically distributed. The X_i have density f and are the variables of interest, the U_i are multiplicative noise with uniform density on $[1-a, 1+a]$, for some $0 < a < 1$, and the two sequences are independent. However, only the Y_i are observed. A full multiplicative noise model is useful for information protection. The goal is to protect the data and simultaneously to be able to catch the main information from the noisy sample when the level of noise is known. We study nonparametric estimation of the density f and of the corresponding survival function. A projection estimator of an auxiliary function is built, from which estimator of the function of interest is deduced. Risk bounds in term of integrated squared error are provided, showing that the dimension parameter associated with the projection step has to perform a compromise. Thus, a model selection strategy is proposed. The resulting estimators are proven to reach the best possible risk bounds. Simulation experiments illustrate the good performances of the estimators and a real data example is described.

Keywords. Censored data, Model selection, Multiplicative noise, Nonparametric estimator.

1 Introduction

Considérons le modèle suivant :

$$Y_i = X_i U_i,\ i = 1, \ldots, n, \quad U_i \sim \mathcal{U}_{[1-a,1+a]},\ 0 < a < 1 \tag{1.1}$$

où $(X_i)_{\{i=1,\ldots,n\}}$ et $(U_i)_{\{i=1,\ldots,n\}}$ sont deux échantillons indépendants. Les U_i sont des variables aléatoires indépendantes et identiquement distribuées (*i.i.d.*) de densité uniforme sur l'intervalle $[1-a, 1+a]$ de $\mathbb{R}^+$ avec $0 < 1-a < 1+a$ où a est supposé connu. Les X_i sont *i.i.d.* selon une densité inconnue f sur $\mathbb{R}^+$. Seules les Y_i sont observés. Ils sont également *i.i.d.* de densité commune f_Y sur $\mathbb{R}^+$:

$$f_Y(y) = \frac{1}{2a} \int_{\frac{y}{1+a}}^{\frac{y}{1-a}} \frac{f(x)}{x} dx, \quad y \in]0, +\infty[, \tag{1.2}$$

Notre but est d'estimer de manière non-paramétrique la densité f des X_i à partir des observations Y_i, $i = 1, \ldots, n$.

L'équation (1.1) modélise une transmission approximative de l'information : les enregistrements Y_i correspondent à la valeur d'intérêt X_i à un erreur $\pm 100a\%$ près. Ce modèle peut d'une part représenter des situations classiques comme les résultats d'un sondage sur des quantités que les sondés connaissent approximativement : sur le poids, la taille, le salaire. D'autre part il peut être utilisé pour la protection de l'information. Par exemple [7] étudie ce modèle pour masquer des données de magnitudes et estime les quantiles de l'échantillon d'origine. Le cas où $U_i \sim \mathcal{U}([0, 1])$ a été le plus étudié dans la littérature, f et la fonction de survie associée sont estimées, par exemple dans [8], [2], [4], [5], [6].

Notre méthode d'estimation de f est en deux étapes. En effet l'inversion de la formule (1.2) n'est pas évidente, nous allons d'abord procéder à l'estimation d'une fonction auxiliaire par projection pour atteindre f dans un deuxième temps. Dans cette exposé nous présenterons cette méthode ainsi que l'extension à la fonction de survie. Puis nous illustrerons la méthode sur un jeu de données réelles.

2 Procédure d'estimation

2.1 Notations

Notons $\mathbb{L}^2(\mathbb{R}^+)$ l'espace des fonctions de carré intégrable sur les réels positifs. La norme associée est notée $\|t\|^2 = \int_{\mathbb{R}^+} |t(x)|^2 dx$. Et la norme infinie d'une fonction t bornée est notée $\|t\|_\infty = \sup_{x \in \mathbb{R}^+} |t(x)|$. La base de Laguerre sur laquelle nous allons nous appuyer est définie par :

$$\varphi_0(x) = \sqrt{2} e^{-x}, \quad \varphi_k(x) = \sqrt{2} L_k(2x) e^{-x} \text{ pour } k \geqslant 1, \quad x \geqslant 0,$$

avec L_k le $k^{\text{ème}}$ polynôme de Laguerre :

$$L_k(x) = \sum_{j=0}^{k} (-1)^j \binom{k}{j} \frac{x^j}{j!}.$$

C'est une base orthonormée : $\langle \varphi_j, \varphi_k \rangle = \delta_{j,k}$ où $\delta_{j,k}$ est le symbole de Kronecker valant 1 si $j = k$ et 0 sinon ; on a de plus la relation suivante sur les normes (*c.f.* [1]) :

$$\forall j \geqslant 0, \|\varphi_j\|_\infty \leqslant \sqrt{2}, \text{ and } \|\varphi_j'\|_\infty \leqslant 2\sqrt{2}(j+1),$$

où φ_j' est la dérivée de φ_j. Toute fonction de $\mathbb{L}^2(\mathbb{R}^+)$ peut être décomposée sur cette base.

2.2 Strategie d'estimation

Définissons g donnée par

$$g(x) := \frac{1}{2a}\left[f\left(\frac{x}{1+a}\right) - f\left(\frac{x}{1-a}\right)\right], \tag{2.1}$$

et considérons une fonction bornée t, dérivable de dérivée dans $\mathbb{L}^2(\mathbb{R}^+)$. Une intégration par partie et les propriétés de f_Y (notamment $\lim_{y\to 0} y f_Y(y) = 0$ et $\lim_{y\to+\infty} y f_Y(y) = 0$) impliquent l'égalité suivante

$$\begin{aligned} \mathbb{E}[t(Y_1) + Y_1 t'(Y_1)] &= \frac{1}{2a}\int_0^{+\infty} t(y)\left[f\left(\frac{y}{1+a}\right) - f\left(\frac{y}{1-a}\right)\right] dy \\ &= \langle t, g\rangle. \end{aligned} \tag{2.2}$$

En d'autres mots

$$\mathbb{E}[\psi_t(Y_1)] = \langle t, g\rangle \quad \text{avec} \quad \psi_t(y) := t(y) + y t'(y).$$

Notre stratégie est d'utiliser l'équation (2.2) pour construire un estimateur par projection de la fonction g, puis de chercher une façon d'inverser la formule (2.1) pour retrouver f. Plus précisément, l'équation (2.1) implique

$$f(x) - f\left(\left(\frac{1+a}{1-a}\right)\ x\right) = 2a\ g((1+a)x)$$

et en itérant ce procédé (en changeant x en $(1+a)x/(1-a)$, $x > 0$), on trouve

$$f(x) - f\left(\left(\frac{1+a}{1-a}\right)^N x\right) = 2a \sum_{k=0}^{N-1} g\left(\left(\frac{1+a}{1-a}\right)^k (1+a)x\right)$$

Donc on a construit une suite d'approximations de f, pour $x > 0$:

$$f_N(x) = 2a \sum_{k=0}^{N-1} g\left(\left(\frac{1+a}{1-a}\right)^k (1+a)x\right). \tag{2.3}$$

Puis, en utilisant que $f(x) - f_N(x) = f(((1+a)/(1-a))^N x)$, il est facile de vérifier que pour $f \in \mathbb{L}^2(\mathbb{R}^+)$, $\|f - f_N\|$ tend vers 0 quand N tend vers l'infini.

Si f est de carré intégrable alors g aussi et on peut écrire sa décomposition dans la base de Laguerre :

$$g(x) = \sum_{j=0}^{\infty} a_j(g)\varphi_j(x), \quad \text{avec } a_j(g) = \langle \varphi_j, g \rangle.$$

Rappelons que $g_m := \sum_{j=0}^{m-1} a_j(g)\varphi_j$ est la projection de g sur $\mathcal{S}_m$, nous avons $a_j(g) = \mathbb{E}[\varphi_j(Y_1) + Y_1\varphi_j'(Y_1)] = \langle \varphi_j, g \rangle$. Puis on estime la projection g_m de g sur $\mathcal{S}_m$ par

$$\widehat{g}_m = \sum_{j=0}^{m-1} \widehat{a}_j \varphi_j, \quad \widehat{a}_j = \frac{1}{n}\sum_{i=1}^{n} [Y_i\varphi_j'(Y_i) + \varphi_j(Y_i)] = n^{-1}\sum_{i=1}^{n} \psi_{\varphi_j}(Y_i), \tag{2.4}$$

avec m pris dans une collection $\mathcal{M}_n \subset \mathbb{N}$ que l'on précisera plus tard. Finalement, en remplaçant a_j par (2.4) et en remplaçant g dans (2.3), on obtient une collection d'estimateurs de f, pour $m \in \mathcal{M}_n$:

$$\widehat{f}_{N,m}(x) = 2a \sum_{k=0}^{N-1} \widehat{g}_m\left(\left(\frac{1+a}{1-a}\right)^k (1+a)x\right). \tag{2.5}$$

2.3 Borne du risque pour l'estimateur de la densité

Nous proposons une borne du risque quadratique intégré (MISE) de l'estimateur $\widehat{f}_{N,m}$ de f.

Proposition 1. *Supposons $f \in \mathbb{L}^2(\mathbb{R}^+)$ et $\mathbb{E}[X_1^2] < +\infty$.*
(i) L'estimateur $\widehat{g}_m$ de g défini par (2.4) vérifie

$$\mathbb{E}[\|\widehat{g}_m - g\|^2] \;\leqslant\; \|g - g_m\|^2 + c_1\frac{m}{n} + c_2\frac{m^3}{n}, \qquad c_1 = 4, \quad c_2 = 16\mathbb{E}[Y_1^2]. \tag{2.6}$$

(ii) L'estimateur $\widehat{f}_{N,m}$ de f défini par (2.5) vérifie

$$\mathbb{E}[\|\widehat{f}_{N,m} - f\|^2] \leqslant \frac{8a^2}{(\sqrt{1+a} - \sqrt{1-a})^2}\left(\|g - g_m\|^2 + c_1\frac{m}{n} + c_2\frac{m^3}{n}\right) + 2\left(\frac{1-a}{1+a}\right)^N \|f\|^2. \tag{2.7}$$

Chacune des bornes fait intervenir un terme de biais proportionnel à $\|g - g_m\|^2$ qui décroît quand m croît, et un terme de variance d'ordre principal m^3/n, qui croît avec m. Le dernier terme (2.7) est exponentiellement décroissant avec N. Comme la valeur de N est choisie par le statisticien, on peut choisir $N \geqslant \log(n)/|\log((1-a)/(1+a))|$ qui rend le terme négligeable (si $a = 0.5$, et $n = 1000$, la condition est $N \geqslant 8$).

On peut obtenir plus précisément une vitesse de convergence en supposant connue la régularité de f. Par exemple, si $f \in W^s(\mathbb{R}^+, L) := \{f : \mathbb{R}^+ \to \mathbb{R}, f \in \mathbb{L}^2(\mathbb{R}^+), \sum_{j\geqslant 0} j^s \langle f, \varphi_j\rangle^2 \leqslant L < +\infty\}$, un espace de Sobolev-Laguerre (voir [4]) on obtient $\mathbb{E}[\|\widehat{g}_{m_{\text{opt}}} - g\|^2] = O(n^{-s/(s+3)})$ pour $m_{\text{opt}} = Cn^{1/(s+3)}$ avec C une constante positive.

2.4 Sélection de modèle pour l'estimateur de la densité

Le choix $m = m_{\text{opt}}$ étant inaccessible, il nous faut une procédure de sélection de la dimension m dans la collection $\mathcal{M}_n = \{m \in [\![1, n]\!],\ m^3 \leqslant n\}$, réalisant le meilleur compromis biais-variance. On choisit le m minimisant le MISE de $\widehat{f}_{N,m}$. A partir de la borne (2.7), la valeur théorique est

$$m_{th} := \underset{m\in\mathcal{M}_n}{\operatorname{argmin}} \left\{ \|g - g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n} \right\} = \underset{m\in\mathcal{M}_n}{\operatorname{argmin}} \left\{ -\|g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n} \right\}.$$

Les projections g_m sont inconnues, on les remplace donc par des estimateurs. Alors on choisit m qui minimise la somme $-\|\widehat{g}_m\|^2 + \text{pen}(m)$ avec

$$\text{pen}(m) := \kappa_1 \frac{m}{n} + \kappa_2 \mathbb{E}[Y_1^2] \frac{m^3}{n} =: \text{pen}_1(m) + \text{pen}_2(m). \tag{2.8}$$

Les termes de penalité ont le même ordre que les termes de variance dans (2.6). La définition de $\mathcal{M}_n$ assure qu'ils restent bornés. Comme $\mathbb{E}[Y_1^2]$ est inconnu, finalement on propose de prendre son homologue empirique et il vient :

$$\widehat{m} = \underset{m\in\mathcal{M}_n}{\operatorname{argmin}} \{-\|\widehat{g}_m\|^2 + \widehat{\text{pen}}(m)\}, \tag{2.9}$$

avec

$$\widehat{\text{pen}}(m) = 2\kappa_1 \frac{m}{n} + 2\kappa_2 \widehat{C}_2 \frac{m^3}{n} := 2\text{pen}_1(m) + 2\widehat{\text{pen}}_2(m), \quad \widehat{C}_2 = \frac{1}{n}\sum_{k=1}^{n} Y_k^2.$$

Les constantes κ_1 et κ_2 sont calibrées par une étude sur simulations. Notons que $\|\widehat{g}_m\|^2 = \sum_{j=0}^{m-1} \widehat{a}_j^2$ avec $\widehat{a}_j$ donné par (2.4), est facile à calculer. Notre estimateur final est

$$\widehat{f}_{N,\widehat{m}}(x) = 2a \sum_{k=0}^{N-1} \widehat{g}_{\widehat{m}} \left(\left(\frac{1+a}{1-a} \right)^k (1+a)x \right). \tag{2.10}$$

On peut prouver le théorème suivant.

Théorème 2. *Supposons que $f \in \mathbb{L}^2(\mathbb{R}^+)$, f bornée et que $\mathbb{E}[X_1^8] < +\infty$. Pour l'estimateur final $\widehat{f}_{N,\widehat{m}}$ défini par (2.4), (2.9) et (2.10), il existe κ_0 tel que pour $\kappa_1, \kappa_2 \geqslant \kappa_0$,*

$$\mathbb{E}[\|\widehat{f}_{N,\widehat{m}} - f\|^2] \leqslant \frac{16a^2}{(\sqrt{1+a} - \sqrt{1-a})^2} \left(6 \inf_{m \in \mathcal{M}} \{\|g - g_m\|^2 + \text{pen}(m)\} + \frac{C_a}{n} \right) + \left(\frac{1-a}{1+a} \right)^N \|f\|^2,$$

avec pen *donnée par (2.8), et C_a une constante positive dépendant de a et $\|f\|_\infty$.*

Le Théorème 2 est un résultat non asymptotique et ne dépendant pas des données. Il montre que la méthode conduit à l'estimateur avec le plus petit risque parmi la collection.

La méthode présentée s'étend à l'estimation de la fonction de survie de X : $\overline{F}$. Une étude approfondie sur simulation permet de dire que cette procédure d'estimation de f et de $\overline{F}$ est rapide et fournit une bonne estimation. Lorsque a est petit : $a = 0.1$ par exemple, l'effet du bruit multiplicatif est très faible et on ne peut espérer approcher f mieux qu'avec un estimateur de f_Y. De plus si le but est de masquer les données, $a = 0.1$ n'a presque pas d'effet. En revanche le cas $a = 0.5$ est intéressant car alors la distribution des données d'origine est réellement perturbée par la multiplication et notre procédure d'estimation permet de la retrouver.

Pour des détails sur la procédure d'estimation de la fonction de survie et une étude numérique des estimateurs, on se réfèrera au preprint [3].

Bibliographie

[1] Abramowitz, M., Stegun, I. (1966) Handbook of mathematical functions with formulas, graphs, and mathematical tables. **55** *National Bureau of Standards Applied Mathematics Series.*

[2] Asgharian, M. , Carone, M. and Fakoor, V. (2012) Large-sample study of the kernel density estimators under multiplicative censoring. *The Annals of Statistics.* **40**(1) 159-187

[3] Comte, F. and Dion, C. (2016) Nonparametric Estimation in a Multiplicative Censoring Model with Symmetric Noise. *Preprint hal-01252780*

[4] Belomestny, D., Comte, F. and Genon-Catalot, V. (2015) Laguerre estimation for k-monotone densities observed with noise. *Preprint hal-01122847*

[5] Brunel, E., Comte, F. and Genon-Catalot, V. (2015) Nonparametric density and survival function estimation in the multiplicative censoring model.*Preprint hal-01122847 A paraitre dans TEST*

[6] van Es, B., Klaassen Chris A.J. and Oudshoorn, K. (2000) Survival analysis under cross-sectional sampling : length bias and multiplicative censoring. *Journal of Statistical Planning and Inference* **91**(2) 295-312

[7] Sinha, B., Nayak, T. and Zayatz, L. (2011) Privacy protection and quantile estimation from noise multiplied data. *Sankhya B.* **73**(2) 297-315

[8] Vardi, Y. and Zhang, Cun-H. (1992) Large sample study of empirical distributions in a random-multiplicative censoring model, The Annals of Statistics. **20**(2) 1022–1039