HAL
archives-ouvertes.fr

Détection de rupture dans un modèle exponentiel

Olivier Lopez, Vladimir Spokoiny

▶ **To cite this version:**

Olivier Lopez, Vladimir Spokoiny. Détection de rupture dans un modèle exponentiel. 42èmes Journées de Statistique, 2010, Marseille, France, France. 2010. <inria-00494806>

HAL Id: inria-00494806
https://hal.inria.fr/inria-00494806

Submitted on 24 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Détection de rupture dans un modèle exponentiel

Olivier Lopez & Vladimir Spokoiny

Université Paris VI
Laboratoire de Statistique Théorique et Appliquée
175 rue du Chevaleret
75013 Paris
e-mail : olivier.lopez0@upmc.fr
&
Weierstrass Institute für Angewandte Analysis und Stochastik
et Humboldt Universität Berlin
Mohrenstrasse, 39
10117 Berlin, Allemagne
e-mail : spokoiny@wias-berlin.de

1 Modèle et Exemples

Soit le modèle exponentiel canonique de densité par rapport à une mesure dominante μ donne par

$$\frac{d\mathbf{P}_f(y)}{d\mu} = p(y)\exp(fy - d(f)),$$

avec p densité de probabilité, et d une fonction convexe C^2. Sans perte de généralité, nous supposons que la variable de densité p est d'espérance nulle.

Nous considérons un modèle de rupture du paramètre f dans le modèle exponentiel ci-dessus. Dans cette situation, les observations sont constituées de variables $(Y_i)_{1\leq i\leq n}$ de loi $\mathbf{P}_{f_i}$, avec

$$f_i = f_i(\theta_0) = a_0\mathbf{1}_{i\leq\tau_0},$$

en définissant $\theta_0 = (a_0, \tau_0) \in \Theta = (\mathbf{R}^+ \times \{1, ..., n\}) \cup (0, 0)$ est le paramètre à estimer. Ce modèle contient le cas limite où $a_0 = 0$ et $\tau_0 = 0$ qui correspond au cas de variables i.i.d., i.e. d'absence de rupture.

Exemples.

1. Cas gaussien : on considère le cas où $Y_i \sim \mathcal{N}(m_0, \sigma^2)$ pour $i > \tau_0$, et $Y_i \sim \mathcal{N}(m_0 + m, \sigma^2)$ pour $i \leq \tau_0$, avec $m \geq 0$ inconnu. Ce cas rentre dans le modèle ci-dessus, avec $a_0 = m\sigma^2/2$, et $d(m) = m^2\sigma^2/2$.

2. Cas poissonnien : on considère le cas où $Y_i \sim \mathcal{P}(\lambda)$ pour $i > \tau_0$, et $Y_i \sim \mathcal{P}(\mu)$ avec $\mu \geq \lambda$ inconnu pour $i \leq \tau_0$. Dans ce cas, $a_0 = \log(\mu/\lambda)$, et $d(a) = \lambda(\exp(a) - 1)$.

Nous considérons l'estimation du paramètre θ par maximum de vraisemblance. Plus précisément, nous nous intéressons à l'estimateur $\hat{\theta}$ qui maximise la log-vraisemblance

$$L(\theta) = a \sum_{i=1}^{\tau} Y_i - \tau d(a).$$

En pratique, la maximisation de ce critère ne pose pas de problème numérique. En effet, pour tout τ, $L(a, \tau)$ est maximum pour

$$\hat{a}(\tau) = d'^{-1}\left(\frac{1}{\tau}\sum_{i=1}^{\tau} Y_i\right).$$

L'estimateur de τ_0 est obtenu comme la valeur de τ réalisant le maximum de $L(\hat{a}(\tau), \tau)$, ce qui s'obtient sans difficulté puisque τ_0 appartient à un ensemble discret et fini.

Ce type de problème a été souvent considéré dans la littérature, voir par exemple Csörgo et Horváth (1997). Néanmoins, l'essentiel des résultats existant ne sont qu'asymptotiques. Ils ne permettent pas de cernner le comportement de ces procédures juste après la rupture, et notamment lorsque celle-ci a une amplitude a petite.

L'apport de ce travail consiste à fournir des résultats à distance finie pour $\hat{\theta}$. L'approche développée est similaire à celle utilisée par Golubev et Spokoiny (2009), qui repose sur l'utilisation de nouvelles bornes exponentielles. En revanche, Golubev et Spokoiny (2009) ne traitaient que le cas d'une rupture gaussienne et ne se focalisaient que sur l'estimation de τ_0. Nous produisons des bornes exponentielles pour l'estimation de a_0 et de τ_0.

Par ailleurs, nous considérons également le cas où le modèle est potentiellement mal spécifié, dans le sens où $f_i \neq f_i(\theta_0)$. Dans cette situation, θ_0 est défini comme

$$\theta_0 = \arg\max_{\theta \in \Theta} E\left[L(\theta)\right].$$

Nous montrons que la méthode reste convergente, sous des conditions sur l'écart entre f_i et $f_i(\theta_0)$.

2 Bornes exponentielles

Considérons $L(\theta', \theta) = L(\theta') - L(\theta)$, et définissons la log-transformée de Laplace comme

$$\mathcal{M}(\mu, \theta, \theta') = -\log E\left[\exp(\mu L(\theta, \theta'))\right],$$

pour $0 < \mu < 1$, où l'espérance est calculée dans le cas du modèle correctement spécifié. La fonction $\theta \to \mathcal{M}(\mu, \theta, \theta_0)$ va jouer le rôle de distance entre θ et θ_0.

Par définition, on a

$$E\left[\exp(\mu L(\theta,\theta_0)+\mathcal{M}(\mu,\theta,\theta_0))\right]=1,$$

pour tout θ. L'obtention de nos résultats à distance finie est basée sur l'extension de cette identité en la rendant uniforme en θ. En effet, l'idée consiste à montrer que, pour $0<\alpha<1$, on a

$$E\left[\exp(\mu L(\theta,\theta_0)+\alpha\mathcal{M}(\mu,\theta,\theta_0))\right]\leq C_n. \tag{1}$$

Ce résultat se généralise au cas d'un modèle mal spécifié, avec une borne un peu différente.

On déduit notamment de (1) que

$$E\left[\exp(\alpha\mathcal{M}(\mu,\theta,\theta_0))\right]\leq C_n,$$

puisque, par définition de $\hat{\theta}$ on a $L(\hat{\theta},\theta_0)\geq 0$. En appliquant l'inégalité de Jensen, on en déduit

$$E\left[\alpha\mathcal{M}(\mu,\theta,\theta_0)\right]\leq \log C_n.$$

Par ailleurs, en utilisant (1) et l'inégalité de Chernoff, on peut utiliser ce résultat pour obtenir des régions de confiance pour θ_0 du type $\{L(\hat{\theta},\theta_0)\geq z\}$.

Remarquons que, si l'on définit $\mu^*(\theta)=\arg\max_{0<\mu<1}\mathcal{M}(\mu,\theta,\theta_0)$, on a, dans certains cas (cas gaussien notamment), $\mu^*(\theta)=\mu^*$ ($1/2$ dans le cas gaussien). De sorte que l'on peut encore affiner les résultats ci-dessus en les appliquant à $\mathcal{M}(\mu^*,\theta,\theta_0)$.

3 Résultats

Nous distinguons deux cas. Dans un premier temps, nous supposons l'amplitude a_0 de la rupture connue. Dans ce cas, sous certaines hypothèses concernant la fonction d, nous démontrons que

$$E\left[a_0^2|\hat{\tau}-\tau_0|\right]\leq C(a_0).$$

Sous certaines hypothèses supplèmentaires (vérifiées notamment dans le cas gaussien et poissonnien), cette constante peut être majorée indépendamment de a_0. Dans tous les cas, cette constante ne dépend pas de n.

Dans le cas où l'amplitude n'est pas connue, cette borne devient

$$E\left[(\hat{a}-a_0)^2\{\hat{\tau}\mathbf{1}_{\tau_0\leq\hat{\tau}}+\mathbf{1}_{\tau_0>\hat{\tau}}\}+|\hat{\tau}-\tau_0|\{a_0^2\mathbf{1}_{\hat{\tau}\leq\tau_0}+\hat{a}^2\mathbf{1}_{\hat{\tau}>\tau_0}\}\right]\leq C(a_0)\log\log n.$$

Si cette constante ne dépend pas de a_0, on obtient donc une borne pour le risque minimax associé à une fonction de perte particulière.

Cette borne supérieure peut être comparée à une borne inférieure, que nous obtenons par une version améliorée du Lemme de Fano, voir Birgé (2001). Cette borne inférieure fait elle aussi apparaître un terme $\log\log n$.

Bibliographie

[1] Birgé, L. (2001) A New Look at an Old Result: Fanos Lemma. *Lab. de Probabilits de lUniversit Paris VI, Paris.*
[2] Csörgo, M., Horváth., L. (1997) *Limith Theorems in Change-Point analysis,* Wiley, New York.
[3] Golubev, Y., Spokoiny, V. (2009) Exponential bounds for minimum contrast estimators. *Electronic Journal of Statistics*, 3, 712–746.