Vitesses de concentration pour les procédures bayésiennes non-paramétriques

Vincent Rivoirard [1]

[1] *CEREMADE, Université Paris Dauphine,*
Place Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16.
Vincent.Rivoirard@dauphine.fr

Résumé. Dans cet exposé, je rappelle les résultats fondamentaux concernant les propriétés asymptotiques des distributions a posteriori dans le cadre de la statistique bayésienne non-paramétrique. Le but de mon exposé est de présenter les outils et d'expliquer les arguments clés qui permettent l'obtention des vitesses de concentration a posteriori.

Mots-clés. Loi a posteriori, tests asymptotiques, voisinage de Kullback, vitesse de concentration.

Abstract. In this talk, I recall the basic results about the asymptotic properties of posterior distributions in the Bayesian nonparametric setting. The goal of my talk is to present technical tools and key arguments to derive posterior concentration rates.

Keywords. Posterior distribution, asymptotic tests, Kullback neighborhood, concentration rate.

1 Introduction

L'approche non-paramétrique de la statistique bayésienne a pris une grande ampleur ces dernières années et ses applications ont essaimé dans divers champs disciplinaires comme la biostatistique, la physique, les neurosciences ou l'apprentissage. De nombreuses monographies, dont ce texte s'inspire, proposent des entrées en matière diverses qui permettent de comprendre les apports méthodologiques de la statistique bayésienne non-paramétrique ainsi que quelques illustrations applicatives (voir par exemple Dey *et al.* (1998), Ghosh et Ramamoorthi (2003), Hjort *et al.* (2010) ou Rasmussen et Williams (2006)). Pour accompagner l'élaboration de modèles plus sophistiqués, la compréhension des propriétés théoriques de ces modèles devient un enjeu crucial. Pour cela, on peut se placer, paradoxalement, dans un cadre fréquentiste qui consiste à postuler l'existence d'un vrai modèle et on étudie alors les propriétés asymptotiques des distributions a posteriori lorsque la taille des observations devient de plus en plus élevée.

Plaçons-nous dans le modèle statistique $(\Omega_n, \mathcal{B}, \mathbb{P}_\theta^{(n)}, \theta \in \Theta)$ dépendant d'un index n qui nous permettra l'étude de comportements asymptotiques. Nous noterons respectivement $\mathbb{E}_\theta^{(n)}$ et $\mathrm{Var}_\theta^{(n)}$ l'espérance et la variance associées à $\mathbb{P}_\theta^{(n)}$. On suppose que le modèle

est dominé par une mesure σ-finie $\mu^{(n)}$ sur Ω_n et on écrit $f_\theta^{(n)}$ la densité de $\mathbb{P}_\theta^{(n)}$ par rapport à $\mu^{(n)}$ et, pour un jeu d'observations X^n,

$$\ell_n(\theta) = \log f_\theta^{(n)}(X^n)$$

la log-vraisemblance associée. Nous munissons Θ d'une tribu $\mathcal{A}$ et d'une loi a priori Π. Alors la distribution a posteriori s'écrit, pour tout $A \in \mathcal{A}$,

$$\Pi(A|X^n) = \frac{\int_A f_\theta^{(n)}(X^n) d\Pi(\theta)}{\int_\Theta f_\theta^{(n)}(X^n) d\Pi(\theta)}.$$

Suivant l'approche fréquentiste, on note dans la suite $\theta_0 \in \Theta$ la vraie valeur du paramètre. On peut alors introduire la notion de vitesse de concentration (ou de contraction) de la loi a posteriori. Pour cela, on munit Θ d'une distance d et on suppose donnée $(\epsilon_n)_n$ une suite tendant vers 0 lorsque n tend vers $+\infty$.

Definition 1. *La distribution a posteriori $\Pi(\cdot|X^n)$ se concentre à la vitesse ϵ_n s'il existe une constante $M > 0$ telle que, lorsque n tend vers $+\infty$,*

$$\Pi\left\{\theta : \ d(\theta, \theta_0) > M\epsilon_n | X^n\right\} \to 0$$

en probabilité sous $\mathbb{P}_{\theta_0}^{(n)}$.

Bien évidemment, la vitesse de concentration ϵ_n dépendra du paramètre θ_0, en particulier lorsque ce dernier est fonctionnel mais également des propriétés de la loi a priori Π. Le calcul de vitesses de concentration offre plusieurs avantages. Tout d'abord, de manière évidente, il permet de mettre en valeur certains modèles a priori et d'en disqualifier d'autres. Par ailleurs, sous de faibles hypothèses, on peut déduire la convergence du risque fréquentiste à la vitesse ϵ_n des estimateurs bayésiens classiques comme la moyenne ou la médiane de la loi a posteriori (voir Ghosal *et al.* (2000)). Enfin, il permet de comprendre le comportement des ensembles de crédibilité bayésiens (voir Rousseau (2016)).

La section suivante présente les conditions permettant l'obtention de vitesses de concentration.

2 Résultats principaux

Dans toute la suite, on se donne $(\epsilon_n)_n$ tendant vers 0 telle que $n\epsilon_n^2 \to +\infty$. Commençons par présenter le résultat fondateur établi par Ghosal *et al.* (2000) et étendu ensuite par Ghosal et van der Vaart (2007). Pour cela, considérons le modèle de densité : On observe un n-échantillon $X_1, \ldots, X_n$ de densité f que l'on cherche à estimer pour la distance de Hellinger notée h. Pour tout ensemble de densités $\mathcal{F}$ et pour tout $\epsilon > 0$, on note $D(\epsilon, \mathcal{F})$ le nombre maximal de densités de $\mathcal{F}$ telles que pour chaque paire de cet ensemble leur distance de Hellinger est au moins égale à ϵ. On a alors le résultat suivant pour Π une loi a priori sur $\mathcal{F}$.

Théorème 1. *Soit $f_0 \in \mathcal{F}$. On suppose qu'il existe une constante $C > 0$ et pour tout n, un sous-ensemble de densités $\mathcal{F}_n \subset \mathcal{F}$ tels que*

$$\log D(\epsilon_n, \mathcal{F}_n) \leq n\epsilon_n^2, \tag{1}$$

$$\Pi(\mathcal{F} \setminus \mathcal{F}_n) \leq \exp(-(C+4)n\epsilon_n^2), \tag{2}$$

$$\Pi\left\{f : \int \log\left(\frac{f_0(x)}{f(x)}\right) f_0(x)dx \leq \epsilon_n^2, \int \log^2\left(\frac{f_0(x)}{f(x)}\right) f_0(x)dx \leq \epsilon_n^2\right\} \geq \exp(-Cn\epsilon_n^2). \tag{3}$$

Alors, pour M constante suffisamment grande, on a :

$$\Pi\left\{f : h(f, f_0) > M\epsilon_n | X_1, \ldots, X_n\right\} \to 0$$

en probabilité lorsque la densité des X_i est f_0.

Les conditions (1) et (3) sont les plus importantes. La condition (2) signifie que le support de Π est approximativement $\mathcal{F}_n$. L'hypothèse (3) exprime le fait que la masse des voisinages de f_0 sous la loi a priori n'est pas trop faible. Les voisinages sont définis à l'aide de la divergence de Kullback-Leibler qui apparaît naturellement dans une approche fondée sur la vraisemblance. La condition (1), que l'on peut exprimer en termes d'entropie, mesure la complexité du modèle $\mathcal{F}_n$. Elle permet de garantir l'existence de tests asymptotiques essentielle à la démonstration du résultat. Dans cet esprit, en reprenant les notations de la Section 1, on obtient la généralisation du théorème précédent (voir Rousseau (2016)).

Théorème 2. *Supposons qu'il existe une constante $c_1 > 0$, et pour tout n, une fonction test ϕ_n et $\Theta_n \subset \Theta$ avec les propriétés suivantes :*

- *L'ensemble Θ_n constitue une bonne approximation de Θ :*

$$\Pi(\Theta \setminus \Theta_n) = o(\exp(-(c_1+2)n\epsilon_n^2).$$

- *Le test ϕ_n constitue un test de séparation suffisamment puissant :*

$$\mathbb{E}_{\theta_0}^{(n)}[\phi_n] = o(1), \qquad \sup_{\theta \in \Theta_n,\ d(\theta_0,\theta) > M\epsilon_n} \mathbb{E}_{\theta}^{(n)}[1 - \phi_n] = o(\exp(-(c_1+2)n\epsilon_n^2).$$

- *La loi a priori "charge" suffisamment les voisinages de Kullback de θ_0 :*

$$\Pi\left\{\theta : KL(\theta_0, \theta) \leq n\epsilon_n^2,\ V_2(\theta_0, \theta) \leq n\epsilon_n^2\right\} \geq \exp(-c_1 n\epsilon_n^2),$$

avec

$$KL(\theta_0, \theta) = \mathbb{E}_{\theta_0}^{(n)}[\ell_n(\theta_0) - \ell_n(\theta)], \quad V_2(\theta_0, \theta) = Var_{\theta_0}^{(n)}(\ell_n(\theta_0) - \ell_n(\theta)).$$

Alors, lorsque n tend vers $+\infty$,

$$\Pi\left\{\theta:\ d(\theta,\theta_0) > M\epsilon_n | X^n\right\} \to 0$$

en probabilité sous $\mathbb{P}_{\theta_0}^{(n)}$.

L'exposé consistera à présenter la démonstration de ce résultat ou d'une de ses variantes comme celle que proposée par Rivoirard et Rousseau (2012). Si je dispose de suffisamment de temps, j'exposerai une version étendue de ce résultat à des processus de comptage (voir Donnet *et al.* (2015)) ou pour l'approche bayésienne empirique (voir Donnet *et al.* (2016)).

Références

[1] Dey D., Müller P. and Sinha, D. (editors) (1998) *Practical nonparametric and semiparametric Bayesian statistics*. Lecture Notes in Statistics, vol. 133. Springer.

[2] Donnet S., Rivoirard V., Rousseau J. and Scricciolo C. (2015) Posterior concentration rates for counting processes with Aalen multiplicative intensities. To appear in *Bayesian Analysis*.

[3] Donnet S., Rivoirard V., Rousseau J. and Scricciolo C. (2016) Posterior concentration rates for empirical Bayes procedures, with applications to Dirichlet Process mixtures. Submitted.

[4] Ghosal S., Ghosh J.K. and van der Vaart A.W. (2000) Convergence rates of posterior distributions, *The Annals of Statistics*, **28**(2), 500–531.

[5] Ghosal S. and van der Vaart A. (2007) Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, **35**(2), 697–723.

[6] Ghosh J.K. and Ramamoorthi R.V. (2003) *Bayesian Nonparametrics.* Springer-Verlag, New York.

[7] Hjort N.L., Holmes C., Müller P. and Walker S.G. (2010) *Bayesian Nonparametrics.* Cambridge University Press, Cambridge, UK.

[8] Rasmussen C.E. and Williams C.K.I. (2006) *Gaussian Processes for Machine Learning*. The MIT Press, Massachusetts Institute of Technology.

[9] Rivoirard V. and Rousseau J. (2012) Posterior concentration rates for infinite dimensional exponential families. *Bayesian Analysis*, **7**(2), 311–334.

[10] Rousseau J. (2016) On the frequentist properties of Bayesian nonparametric methods. To appear in *Annual Statistical Reviews.*