Convergence rate and consistency at zero for decreasing densities on $[0, 1]$

J-B. Salomond

CREST, Université Paris Dauphine

Abstract

In this paper, we consider the well known problem of estimating a density function defined on a compact set under qualitative assumption. More precisely, we study monotone non increasing densities in a Bayesian setting, and derive convergence rate for the posterior distribution for two types of priors, namely, a Dirichlet process and finite mixture prior. We prove that the posterior distribution based on both prior concentrates at the rate $(n/\log(n))^{-1/3}$, which is the minimax rate of estimation up to a $\log(n)$ factor. We also prove that the posterior probability of the continuous limit of the density at 0 is consistent, which is of particular interest as the classical non parametric maximum likelihood estimator is known to be inconsistent at this point.

1 Introduction

The non parametric problem of estimating monotone non increasing curve, and monotone densities in particular, has been well studied in the literature, both form a theoretical and applied perspective. For instance, Groeneboom (1985) and more recently, Balabdaoui and Wellner (2007) studied very precisely the asymptotic properties of the non parametric maximum likelihood estimator (NPMLE) , namely, the Grenander estimator, which is proved to be consistent and to converge at the minimax rate $n^{-1/3}$ when the support of the distribution is compact. Monotone non increasing densities arise naturally when considering survival analysis, where it is natural to assume that the uncensored survival time have a monotone non increasing density.

In this paper, we study the asymptotic properties of a Bayesian estimator of a monotone non increasing density function with a compact support. Following Khazaei et al. (2010) we study two families of non parametric priors on $\mathcal{F}$, the class of monotone non increasing densities with support on $[0, 1]$. We obtained the two families of prior a concentration rate of order $(n/\log(n))^{-1/3}$. Interestingly, the NPMLE is not consistent at 0, (see Sun and Woodroofe (1996) and Balabdaoui and Wellner (2007) for instance). However, we prove that the posterior distribution of $f(0)$ is still concentrate around the true value a the rate $(n/\log(n))^{-1/3}$. The non parametric prior are constructed from the mixture representation of monotone non increasing density. It is known since Williamson (1956) that any monotone non increasing density on $\mathbb{R}^+$ has a mixture repre-

sentation

$$f(x) = \int_0^\infty \frac{\mathbb{I}_{[0,\theta]}(x)}{\theta} dP(\theta) \tag{1}$$

where P is a probability distribution on $\mathbb{R}^+$, namely a mixing distribution. In order to indicate the dependence on P, we shall note f_P the density function admiting representation (1). Let $\mathcal{F}$ be the set

$$\mathcal{F} = \left\{ f \text{ s.t. } f > 0,\ f \searrow \int_0^L f = 1 \right\}$$

Mixtures models have been well studied in a Bayesian framework and various types of prior have been considered for the mixing distribution. The most popular non parametric mixture prior model is the Dirichlet Process prior (DP) introduced by Ferguson (1983) and in their article, Wu and Ghosal (2008) studied properties of general mixtures models. Using the mixture representation od monotone non increasing densities (1) we construct non parametric priors on the set $\mathcal{F}$ by considering a prior on the mixing distribution P. We thus fall in the set up of non parametric mixture prior models. We consider two types of prior

Type 1 : Dirichlet Process prior $P \sim DP(A, \alpha)$ where A is a positive constant and α a probability distribution on $[0, L]$

Type 2 : Finite mixture $P = \sum_{j=1}^K p_j \delta_{x_j}$ with K a non zero integer and δ_x the dirac distribution on x. We define a prior on P by a distribution Q on K and given K, distributions $\pi_{x,K}$ on $(x_1, \ldots, x_K)$ and $\pi_{p,K}$ on $(p_1, \ldots, p_K)$

Each of these are prior on the set of probability distributions. For $\mathbf{X^n} = (X_1, \ldots, X_n)$, a sample of n independent and identically distributes random variables with a common probability distribution function f_0 in $\mathcal{F}$ with respect to the Lebesgue measure, we denote $\Pi(\cdot|\mathbf{X^n})$ the posterior probability measure associated with the prior Π. Most of the results on the convergence rates of posterior distribution requires that the prior puts enough mass on Kullback Leiber neighborhood of the true density

$$\Pi(\{P, KL(f_P, f_0) \leq \epsilon\}) > 0 \tag{2}$$

where $KL(f_1, f2)$ is the Kullback Leiber divergence between f_1 and f_2 defined by $KL(f_1, f_2) = \int f_1 \log(f_1/f_2)$. In particular, Wu and Ghosal (2008) proved that this condition is satisfied by the Type 1 prior in the case of monotone non increasing densities, under mild conditions on f_0. The study of the Type 2 prior is more delicate as it does not satisfy (2) in general. Moreover, it has been proved (see Khazaei et al. (2010)) that for this choice of prior, $\Pi(\{P, KL(f_P, f_0) = \infty\}) = 1$. However, we achieve to prove that the posterior based on a Type 2 prior concentrate at the same rate without the Kullback Leiber property.

The paper is organised as follow. the main results are given in section 2, where conditions on the priors are discussed. The proof are given afterward.

2 Main results

Convergence rate of the posterior distribution have been well studied in the literature and some general results links the rate to the prior (see Ghosal et al. (2000)). We deduce from these results the optimality of the Bayesian estimator (up to a $\log(n)$ factor).The following theorem gives general conditions on the prior to achieve a posterior convergence rate of $(n/\log(n))^{-1/3}$.

Theorem 1. *Let $\mathbf{X^n} = (X_1, \dots, X_n)$ be a iid sample with a common probability distribution function $f_0 \in \mathcal{F}$; and let α be a positive probability density on $[0,1]$ with respect to the Lebesgue measure that satisfy for θ close to 0, and $t > 1$*

$$\alpha(\theta) \lesssim \theta^t \tag{3a}$$

Define also Q a probability distribution on $\mathbb{N}$ and $\pi_{p,K}$ a probability distribution on the simplex of $\mathbb{R}^K$ satisfying for some positive constants $C_1, C_2, a_1, \dots, a_K, c$

$$e^{-C_1 K \log(K)} \geq Q(K) \geq e^{-C_2 K \log(K)} \tag{3b}$$

$$\pi_{p,k}(p_1, \dots, p_K) \geq K^{-K} c^K p_1^{a_1} \dots p_K^{a_K} \tag{3c}$$

and finally, let $(x_i)_i$ be the order statistics of K iid random variables from α. If d is either the L^1 or Hellinger distance, then for Π a Type 1 or Type 2 prior, there exists a positive constant C such that

$$\Pi\left(f, d(f, f_0) \geq C\left(\frac{n}{\log(n)}\right)^{-1/3} |\mathbf{X^n}\right) \to 0, \qquad P_0\, a.s. \tag{4}$$

when n goes to infinity.

The proof of this theorem is given in section 3. Condition (3) are roughly the same than in Khazaei et al. (2010). This Theorem is thus an extension of their result when the density's support is compact. Given that f_0 is in $\mathcal{F}$, we can uniquely define $f_0(0)$ by considering the left limit of $f_0(x)$ when $x \searrow 0$. Let $f_0(0) = \lim_{x \searrow 0} f_0(x)$, under some mild conditions on the first derivative of f, we get the consistency and the convergence rate of the posterior distribution and of the Bayesian estimator associated with the absolute loss, namely, the posterior median.

Theorem 2. *Let $f_0 \in \mathcal{F}$ such that, $f'(0^+) < 0$. Let $X_i \overset{iid}{\sim} f_0$ for $i = 1 \dots n$, and let Π be a prior satisfying (3), then, if $\epsilon_n = C(n/\log(n))^{-1/3}$*

$$P^\pi\big(|f_P(0) - f_0(0)| > \epsilon_n |\mathbf{X^n}\big) \to 0 \tag{5}$$

Consider the posterior median $\hat{f}_n(0) = \inf\{x, P^\pi\big[f_p(0) \leq x\big] > 1/2\}$ thus

$$P_0\big(|\hat{f}_n(0) - f_0(0)| > \epsilon_n |\mathbf{X^n}\big) \to 0 \tag{6}$$

Thus the Bayesian approach is yield a consistent estimator of $f_0(0)$, which is not the case of the maximum likelihood approach. It is known that integrating the parameter as done in Bayesian approaches induces a penalization. This is particularly useful in testing or model choice problems but can also be effective in estimation problems, see for instance Rousseau and Mengersen (2011). The

problem of estimating $f_0(0)$ under monotonicity constraints is another example of the effectiveness of penalization induced by integration on the parameters. However, contrariwise to Rousseau and Mengersen (2011), we have not clearly identified how the penalisation act, and only observe that it leads to a consistent posterior distribution and a consistent estimator. Furthermore, we only studied here the consistency in probability and it is not clear whether or not stronger consistency results, such as almost sure consistency, should be expected.

3 Proofs

We present here the proof of the two main theorem. The proof of Theorem 1 adapted from Ghosal et al. (2000), and is based on a piecewise constant approximation of f_0 in the sense of Kullback Leiber divergence. We therefore detail the construction of this approximation in the following section. Note that this is not the construction proposed in Khazaei et al. (2010) which was adapted to the L^1 distance but not to the Kullback Leiber divergence. This construction is adapted from the one proposed by van der Vaart and Wellner (1996) in the proof of theorem 2.7.5.

3.1 Proof of Theorem 1

The piecewise constant approximation of f_0 is base on a sequential subdivision of the interval $[0, L]$ with more refined subdivision where f_0 is less regular. We then identify the piecewise constant density by a mixture of uniformes. The following Lemma gives the form of the probability distribution P such that f_P is in the Kullback Leiber neighbourhood of f_0.

Lemma 3. *Let $f \in \mathcal{F}$ be such that $f(0) \leq M < +\infty$. For all $0 < \epsilon < 1$ there exists $m \lesssim M^{2/3}\epsilon^{-1}$, $p = (p_1, \ldots, p_m)$ and $x = (x_1, \ldots, x_m)$ such that $P = \sum_{i=1}^{m} \delta_{x_i} p_i$ satisfy*

$$KL(f, f_P) \lesssim \epsilon^2, \ \int f \log\left(\frac{f}{f_P}\right)^2 \lesssim \epsilon^2 \tag{7}$$

where f_P is defined as in (1) and KL denote the Kullback Leiber divergence.

Proof. For a fixed ϵ, let f be in $\mathcal{F}$. Consider $\mathcal{P}_0$ the coarsest partition :

$$0 = x_0^0 < x_1^0 = L$$

at the i^{th} step, let $\mathcal{P}_i$ be the partition

$$0 = x_0^i < x_1^i < \cdots < x_{n_i}^i = L$$

and define

$$\varepsilon_i = \max_j \left\{ (f(x_{j-1}^i) - f(x_j^i))(x_j^i - x_{j-1}^i)^{1/2} \right\}$$

For each $j \leq 1$, if $(f(x_{j-1}^i) - f(x_j^i))(x_j^i - x_{j-1}^i)^{1/2} \geq \frac{\varepsilon_i}{\sqrt{2}}$ we split the interval $[x_{j-1}, x_j]$ into two subsets of equal length. We then get a new partition $\mathcal{P}_{i+1}$. We continue the partitioning until the first k such that $\varepsilon_k^2 \leq \epsilon^3$. At each step i, let n_i be the number of intervals in $\mathcal{P}_i$, s_i the number of interval in $\mathcal{P}_i$

that have been divided to obtain $\mathcal{P}_{i+1}$, and $c = 1/\sqrt{2}$. Thus, it is clear that $\varepsilon_{i+1} \leq 2^{-1/2}\varepsilon_i = c\varepsilon_i$

$$\begin{aligned} s_i(c\varepsilon_i)^{2/3} &\leq \sum_j (f(x_{j-1}^i) - f(x_j^i))^{2/3}(x_j^i - x_{j-1}^i)^{1/3} \\ &\leq \left(\sum_j f(x_{j-1}^i) - f(x_j^i)\right)^{2/3} \left(\sum_j x_j^i - x_{j-1}^i\right)^{1/3} \leq M^{2/3}L^{1/3} \end{aligned}$$

using Hölder inequality. We then deduce that

$$\begin{aligned} \sum_{j=1}^k n_j = k + \sum_{j=1}^k j s_{k-j} \leq 2\sum_{j=1}^k j s_{k-j} &\leq 2\sum_{j=1}^k j M^{2/3}L^{1/3}(c\varepsilon_{k-j})^{-2/3} \\ &\leq 2M^{2/3}L^{1/3}\varepsilon_k^{-2/3}2^{1/3}\sum_{j=1}^k j 2^{-j/3} \\ &\leq K_0 M^{2/3}L^{1/3}\varepsilon_k^{-2/3} \end{aligned}$$

where K_0 is a constant. Thus

$$n_k \leq K_0 M^{2/3}L^{1/3}\epsilon^{-1} \tag{8}$$

Now, let f be in $\mathcal{F}$, we prove that there exists a stepwise density with less than $C\frac{1}{\epsilon}$ pieces, where C is a constant depending on f, such that

$$KL(f,h) \leq \mathcal{O}(\epsilon^2) \text{ and } \int f \log(\frac{f_0}{f_P})^2(x)dx \leq \mathcal{O}(\epsilon^2) \tag{9}$$

In order to lighten notations, we define

$$x_i = x_i^k, \quad l_i = x_i - x_{i-1}, \quad g_i = f(x_{i-1})^{1/2}$$

Thus consider the relation defined above with $f^{1/2}$ which is also monotone nonincreasing based on $g = \sum \mathbb{I}_{[x_{i-1},x_i]}g_i$

$$\begin{aligned} ||f^{1/2} - g||_2^2 = \int (f^{1/2} - g)^2(x)dx &= \sum_{i=1}^{n_k} \int_{I_i} (f^{1/2} - g)^2(x)dx \\ &\leq \sum_{i=1}^{n_k} \int_{I_i} (f^{1/2}(x_{i-1}^k) - f^{1/2}(x_i^k))^2 dx \\ &\leq \sum_{i=1}^{n_k} (x_i^k - x_{i-1}^k)(f^{1/2}(x_{i-1}^k) - f^{1/2}(x_i^k))^2 \\ &\leq n_k\varepsilon_k^2 \leq K_0 M^{2/3}\epsilon^2 \end{aligned}$$

We then define $h = \frac{g^2}{\int g^2}$ and have

$$\begin{aligned}\int g^2 dx &= \int (g^2 - f)(x)dx + 1 \\ &= \int (g - \sqrt{f})(g + \sqrt{f})(x)dx + 1 \\ &= 1 + \mathcal{O}(\epsilon)\end{aligned}$$

and deduce that $(\int g^2)^{1/2} = 1 + \mathcal{O}(\varepsilon)$. Let H be the Hellinger distance

$$\begin{aligned}H^2(f,h) &= H^2\left(f, \frac{g^2}{\int g^2}\right) \\ &\leq H^2(f, g^2) + H^2(g^2, \frac{g^2}{\int g^2}) \\ &\leq \epsilon^2 + \int (g - \frac{g}{(\int g^2)^{1/2}})^2(x)dx \leq \mathcal{O}(\epsilon^2)\end{aligned}$$

Since $||h/f||_\infty = ||g/f||_\infty (\int g^2)^{-1} \leq (\int g^2)^{-1}$ together with the above bound on $H(f,h)$ and the following lemma 8 from Ghosal and van der Vaart (2007), we obtain the required result.

Let P be a probability distribution defined by

$$P = \sum_{i=1}^{n_k} p_i \delta(x_i^k) \quad p_i = (h_{i-1} - h_i)x_i^k \quad p_{n_k} = h_{n_k} x_{n_k}^k = h_{n_k} L$$

thus $f_P = h$ and given the previous result, lemma 3 is proved. □

The proof of Theorem 1 is based on theorem 2.1 of Ghosal et al. (2000). It consists in obtaining a lower bound on the prior mass of Kullback Leiber neighbourhoods of any density in $\mathcal{F}$. An interesting feature of mixtures distributions having a parameter dependent support is that in many cases, the prior mass of sets the sets $\{f, KL(f_0, f) = +\infty\}$ is 1 for most $f_0 \in \mathcal{F}$. Hence we cannot apply the result of Ghosal et al. (2000). We thus extended the approach used in Khazaei et al. (2010) to the convergence rate framework and get similar results as those presented in Ghosal et al. (2000). Thus we only have to bound from below, for a sequence θ_n such that $F_0(\theta_n)^n \in [1-\epsilon, 1-\epsilon/2]$, the prior mass of sets

$$S_n(\epsilon, \theta_n) = \left\{ f, KL(f_n, f_{0,n}) \leq \epsilon^2, \int f_{0,n}(x) \log\left(\frac{f(x)}{f_0(x)}\right)^2 dx \leq \epsilon^2 \right\} \tag{10}$$

where

$$f_n(\cdot) = \frac{f(\cdot)\mathbb{I}_{[0,\theta_n]}(\cdot)}{F(\theta_n)}, \; f_{0,n}(\cdot) = \frac{f_0(\cdot)\mathbb{I}_{[0,\theta_n]}(\cdot)}{F_0(\theta_n)} \tag{11}$$

Lemma 4. *Let* Π *either a Type 1 or Type 2 prior on* $\mathcal{F}$ *satisfying* (3) *and let* $S_n(\epsilon, \theta_n)$ *be a set as in* (10)*, then*

$$\Pi(S_n(\epsilon, \theta_n) \gtrsim e^{C_1 \epsilon^{-1} \log(\epsilon)} \tag{12}$$

Consider $\epsilon > 0$ and θ_n such that $F_0(\theta_n)^n \in [1-\epsilon, 1-\epsilon/2]$, thus $\theta_n \leq 1 - \epsilon(2nf_0(0))^{-1}$ and define f_{Pn} and $f_{0,n}$ as in (11). Using lemma 3 with $L = \theta_n$, we obtain that there exists a distribution $P = \sum_{i=1}^{n_k} \delta_{x_i} p_i$ such that for some constants $C_0, C_0' > 0$ such that

$$KL(f_{0,n}, f_{Pn}) \leq C_0\epsilon^2, \text{ and } \int f_{0,n} \log\left(\frac{f_{0,n}}{f_{Pn}}\right)^2 \leq C_0'\epsilon^2$$

Note that f_P has support $[0, \theta_n]$ and put a positive mass on θ_n. Now, set $m = n_k$ and consider P' the mixing distribution associated with $\{m, x_1', \ldots, x_m', p_1' \ldots, p_m'\}$ with $\sum_{i=1}^m p_i' = 1$. Define for $1 \leq j \leq m-1$ the set $U_i = [0 \vee (x_i - \epsilon^3, x_i + \epsilon^3]$ and $U_m = (\theta_n, \theta_n + \epsilon(1-\theta_n) \wedge \epsilon^3]$. Construct P' such that $x_i' \in U_i$ and $|P'(U_i) - p_i| \leq \epsilon^2 n^{-1}$

References

Balabdaoui, F. and Wellner, J. A. (2007). Estimation of a k-monotone density: limit distribution theory and the spline connection. *Ann. Statist.*, 35:2536–2564.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. pages 287–302.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.

Ghosal, S. and van der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723.

Groeneboom, P. (1985). Estimating a monotone density. pages 539–555.

Khazaei, S., Rousseau, J., and Balabdaoui, F. (2010). Bayesian nonparametric inference of decreasing densities. In *42èmes Journées de Statistique*, Marseille, France France.

Rivoirard, V. and Rousseau, J. (2009). On the bernstein von mises theorem for linear functionals of the density. Technical report, Université Paris Dauphine.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710.

Sun, J. and Woodroofe, M. (1996). Adaptive smoothing for a penalized NPMLE of a non-increasing density. *J. Statist. Plann. Inference*, 52(2):143–159.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

Williamson, R. E. (1956). Multiply monotone functions and their laplace transforms. *Duke Mathematical Journal*, pages 189–207.

Wu, Y. and Ghosal, S. (2008). Kullback leibler property of kernel mixture priors in bayesian density estimation. *Electron. J. Stat.*, 2:298–331.