On Bayesian Nonparametric Estimation of Discontinuous Densities*

Olivier Binette†

October 31, 2018

Abstract

We consider the Bayesian nonparametric estimation of a density f_0 supported on a compact metric space. Conditions on sieve priors are given as to ensure posterior consistency at any bounded density. No continuity assumptions are made on f_0.

1 Introduction

We consider the problem of estimating a density f_0 supported on a finitely measured compact metric space $(\mathbb{M}, d)$ with independent observations $X_i \sim f_0$, $i = 1, 2, \ldots, k$. This is of particular interest in the context of circular statistics, directional statistics and of statistics on manifolds, where $\mathbb{M}$ is correspondingly the circle, the sphere or some compact Riemannian surface. Applications range from the analysis of seasonal and angular measurements to the statistics of shapes and configurations (Jammalamadaka and SenGupta, 2001; Bhattacharya and Bhattacharya, 2012). In bioinformatics, for instance, an important problem is that of using the chemical composition of a protein to predict the conformational angles of its backbone (Al-Lazikani et al., 2001). Bayesian nonparametric methods, accounting for the wrapping of angular data through an adequate metric structure, have been successfully applied in this context (Lennox et al., 2009, 2010).

In the case where $\mathbb{M} = [0, 1]$, Petrone (1999) used the distribution of a random polynomial density as a prior on the space of all densities. Strong posterior consistency was shown by Petrone and Wasserman (2002) under a continuity assumption on f_0, while adaptative convergence rates under Holder continuity assumptions have been obtained by Kruijer and van der Vaart (2008). The latter show that with a slightly modified prior, the posterior contraction rates are minimax optimal (up to a log factor) over the class of α-smooth densities ($0 \leqslant \alpha \leqslant 1$).

More recently, Bhattacharya and Dunson (2012) studied the strong L^1 posterior consistency of Dirichlet process kernel mixtures on compact metric spaces. They require the true density f_0 to be continuous as a fundamental assumption in the development of their theory. This continuity assumption is also present in the approach of Wu and Ghosal (2008) to the study of the Kullback-Leibler support of sieve priors and it is prevalent in the Bayesian literature on posterior consistency.

*Notes for a talk presented at Texas A&M on November 2, 2018. Large parts of this manuscript have been adapted from Sections 3.2 and 3.3 of (Binette and Guillotte, 2018).

†Université du Québec à Montréal. Email: olivier.binette@gmail.com

This contrasts with kernel estimators that are known to be L^1 consistent at discontinuous densities since (Devroye and Wagner, 1979).

1.1 Contributions

This manuscript shows how to construct sieve priors for which strong posterior consistency holds at any bounded, not necessarily continuous, density. Hence while posterior convergence may be arbitrarily slow when f_0 is non-regular, consistency still holds in many cases. In particular, in the context of the random polynomial priors of Petrone and Wasserman (2002), we show that continuity of f_0 is not necessary for posterior consistency; boundedness is sufficient.

The construction relies on the use of finite dimensional sieves with particular shape-preserving approximation properties. This allows us to apply a "reverse Pinsker inequality" (Sason and Verdú, 2016; Binette, 2018), used to ensure that that any bounded density is in the Kullback-Leibler support of the prior.

1.2 Background and Notations

Let $\mathbb{F}$ be a space of densities and let Π be a prior on $\mathbb{F}$. Let $\mathrm{K}(f_0, f) = \int_{\{f_0>0\}} f_0 \log f_0/f \, d\mu$ be the Kullback-Leibler divergence between the densities f_0 and f, and denote $\mathrm{B}_{\mathrm{KL}}(f_0, \varepsilon) := \{f \in \mathbb{F} : \mathrm{K}(f_0, f) < \varepsilon\}$. The *Kullback-Leibler support* of Π is the set of all densities f_0 such that $\Pi(\mathrm{B}_{\mathrm{KL}}(f_0, \varepsilon)) > 0$ for all $\varepsilon > 0$.

Here, strong posterior consistency at $f_0 \in \mathbb{F}$ means that if $X_1, \dots, X_n$ are independent random variables and identically distributed according to the probability distribution P_{f_0} with density f_0, denoted $(X_i)_{i\geqslant 1} \sim P_{f_0}^{(\infty)}$, then for all $\varepsilon > 0$,

$$\Pi\left(\left\{f \in \mathbb{F} : \int |f - f_0| < \varepsilon\right\} \mid (X_i)_{i=1}^n\right) \to 1, \quad \mathrm{P}_{f_0}^{(\infty)}\text{-a.s.} \tag{1.1}$$

2 The general framework

Let $(\mathbb{M}, d)$ be a compact metric space together with a finite measure μ defined on its σ-algebra $\mathfrak{B}_{\mathbb{M}}$ and let $\mathbb{F}$ be the space of all bounded densities with respect to μ. We consider a sequence of linear operators $T_n : L^1(\mathbb{M}) \to L^1(\mathbb{M})$, $n \in \mathbb{N}$, which maps densities to densities and $T_n(\mathbb{F}) \subset \mathbb{F}$, so that we obtain a model $\mathcal{C} = \bigcup_{n\in\mathbb{N}} \mathcal{C}_n$ with $\mathcal{C}_n := T_n(\mathbb{F})$.

Now let $\mathfrak{B}$ be the Borel σ-algebra of $\mathbb{F}$ for the L^1 metric and let $\mathfrak{B}_n$ be the restriction of $\mathfrak{B}$ to $\mathcal{C}_n$, $n \geqslant 0$. A sieve prior Π on $\mathbb{F}$ can be specified through priors Π_n on $(\mathcal{C}_n, \mathfrak{B}_n)$ and a distribution ρ on $n \in \{0, 1, 2, \dots\}$ as

$$\Pi(B) = \sum_{n\geqslant 0} \rho(n)\Pi_n(B \cap \mathcal{C}_n), \quad B \in \mathfrak{B}. \tag{2.1}$$

Theorem 1 below gives simple conditions on T_n, Π_n and ρ, in this framework, as to ensure that any $f_0 \in \mathbb{F}$ is in the Kullback-Leibler support of Π.

Theorem 2.1 (See Binette and Guillotte (2018))**.** *Let $\mathbb{F}$, Π_n, Π and T_n be as above. Suppose that $T_n(\mathbb{F}) \subset \mathbb{F}$ is of finite dimensions and also that $\|T_n f - f\|_\infty \to 0$, as $n \to \infty$, for every continuous function f on $\mathbb{M}$. If $\rho(n) > 0$ and if Π_n has support $T_n(\mathbb{F})$, then any $f_0 \in \mathbb{F}$ is in the Kullback-Leibler support of Π.*

Together with a prior complexity constraint, the Kullback-Leibler support condition entails strong L^1 posterior consistency (see e.g. Xing and Ranneby (2009)).

Corollary 2.2. *Under the hypotheses of Theorem 2.1, suppose further that* $\rho(n) < ce^{-Cd_n}$, *for some* $c > 0$, $C > 0$ *and for an increasing sequence* d_n *such that* $d_n \geqslant \dim T_n(\mathbb{F})$. *Then the posterior distribution of* Π *is strongly consistent at every* $f_0 \in \mathbb{F}$.

Remark 2.3. *The results still hold when the space* $\mathbb{F}$ *is constrained such as being some convex subset of bounded densities containing at least one density that is bounded away from zero or a star-shaped subset around such a density (e.g.* $\mathbb{F}$ *may be a set of bounded unimodal densities or a set of multivariate copula densities).*

The idea of the proof of Theorem 2.1 is largely summarized by the following lemma and its proof. The reader is referred to Binette and Guillotte (2018) for more details.

Lemma 2.4. *Let* T_n *satisfy the hypotheses of Theorem 2.1 and let* f *be a density on* $(\mathbb{M}, \mathfrak{B}_{\mathbb{M}}, \mu)$ *such that* $\inf f > 0$ *and* $\sup f < \infty$. *Then* $\mathrm{K}(f, T_n f) \to 0$ *as* $n \to \infty$.

Proof. We first show that $\|h - T_n h\|_1 \to 0$ as $n \to \infty$ for every $h \in L^1(\mathbb{M})$. Let $\varepsilon > 0$ be arbitrary. We can find g continuous with $\|h - g\|_1 < \varepsilon/3$; this is because $\mathbb{M}$ is compact so the set of continuous functions on $\mathbb{M}$ is dense in $L^1(\mathbb{M})$. By assumption there exists $N \geqslant 0$ such that $n > N$ implies $\|T_n g - g\|_\infty < \varepsilon/(3\mu(\mathbb{M}))$ and $\|T_n g - g\|_1 \leqslant \mu(\mathbb{M})\|T_n g - g\|_\infty < \varepsilon/3$. The fact that T_n is linear and maps densities to densities implies that it is monotonous and $\|T_n(f-g)\|_1 \leqslant \|f - g\|_1$. Hence $\|T_n f - f\|_1 \leqslant \|T_n(f-g)\|_1 + \|T_n g - g\|_1 + \|g - f\|_1 < \varepsilon$ whenever $n > N$.

Now to finish the proof, notice that

$$\mathrm{K}(f, T_n f) \leqslant \|f/T_n f\|_\infty \|f - T_n f\|_1 \leqslant \frac{\sup f}{\inf T_n f} \|f - T_n f\|_1. \tag{2.2}$$

By monotony of T_n, $\inf T_n f \geqslant \inf T_n(\inf f) \to \inf f > 0$. This shows $\frac{\sup f}{\inf T_n f} = \mathcal{O}(1)$. It follows from the preceding paragraph that $\mathrm{K}(f, T_n f) \to 0$ as $n \to \infty$. □

Remark 2.5. *The first inequality in (2.2) is a simple version of a reverse Pinsker inequality. It shows that, provided non-trivial lower and upper bounds on the likelihood ratio, the Kullback-Leibler divergence is equivalent to the total variation distance. The best possible upper bound on* K *by a multiple constant of the total variation distance under these conditions is provided in Binette (2018).*

2.1 Discussion of the assumptions

The preceding results are based on an abstraction of the submodels $\mathcal{C}_n$ to the operators T_n which are such that $T_n(\mathbb{F}) = \mathcal{C}_n$. Three types of assumptions are made on the sequence of operators.

1. *Shape preservation*: it is assumed that $T_n(\mathbb{F}) \subset \mathbb{F}$, so that the support of Π lies within $\mathbb{F}$. Furthermore, we assume that T_n maps densities to densities. This ensures that T_n is a monotonous operator of norm 1. Using the approximation hypothesis below, the fact that $\|T_n 1 - 1\|_\infty \to 0$ shows $T_n 1 \approx 1$. This positivity and (near) reproduction of constants is fundamental in the proof of Theorem 2.1.

2. *Approximation*: we assume that $\|T_n f - f\|_\infty \to 0$, for every continuous function f on $\mathbb{M}$. Together with (1), the finiteness of μ and the boundedness of f_0, this is used to show that $\mathrm{K}(f_0, T_n f_0) \to 0$ whenever $f_0 \in \mathbb{F}$ and $\inf f_0 > 0$.

3. *Finite dimensionality*: The assumption $\dim T_n(\mathbb{F}) < \infty$ is used to obtain a comparison of the L^1 and L^∞ norms on the submodels $\mathcal{C}_n$, ensuring that the Kullback-Leibler approximation properties of T_n also carry to sets of positive prior probability.

2.2 Linear Operators and Kernel Mixtures

By the Riesz representation theorem, any positive linear operator T_n mapping densities to densities and such that $T_n(1) = 1$ takes the form $T_n f(x) = \mathbb{E}[f(Y_n(x))]$ for some families $\{Y_n(x) \mid x \in \mathbb{M}\}$, $n \in \mathbb{N}$, of random variables with values in $\mathbb{M}$. If $Y_n(x)$ has a density $K_n(x, \cdot)$ with respect to μ, then also $T_n f(x) = \int K_n(x,t) f(t)\, dt$. Integrating the kernel $K_n(x, \cdot)$ with respect to a Dirichlet process rather than f and letting n be random yields a Dirichlet Process Mixture to which our results may be applied. An example is given in the following section.

3 Application to random polynomial priors

Let T_n be the Bernstein-Kantorovich operator on $L^1([0,1])$ defined by

$$T_n f(x) = (n+1) \sum_{i=0}^{n} \int_{\frac{i}{n+1}}^{\frac{i+1}{n+1}} f(t)\, dt\, p_{i,n}(x), \quad x \in [0,1],$$

where $p_{i,n}(u) = \binom{n}{i} u^i (1-u)^{n-i}$ is a Bernstein polynomial of degree n. It may be extended to act on a probability measure P by letting

$$T_n P = (n+1) \sum_{i=0}^{n} P\left(\left[\frac{i}{n+1}, \frac{i+1}{n+1}\right)\right) p_{i,n}(x).$$

Clearly, if P has a density f, then $T_n P = T_n f$.

The Bernstein-Dirichlet prior of Petrone (1999) on the space of densities supported on $[0,1]$ can be seen as the Dirichlet Process Mixture induced by the random density $T_N \mathcal{D}$, where N has some distribution ρ and is independent from a Dirichlet Process $\mathcal{D}$. If for instance $\mathcal{D}$ has base Lesbegue measure on $[0,1]$, and if N is Poisson or Geometric, then Theorem 2.1 may be readily applied as to ensure posterior consistency at any bounded density.

Indeed, it is well known that $\|T_n f - f\|_\infty \to 0$ for any continuous f; see e.g. DeVore and Lorentz (1993). Clearly T_n maps densities to densities and its image is of dimension n. The condition on the base measure of $\mathcal{D}$ ensures that the conditional distributions $T_n(\mathcal{D})$, $n \in \mathbb{N}$, have full support on $T_n(\mathbb{F})$. Hence by Corollary 2.2, if $0 < \rho(n) < ce^{-Cn}$ for some $c, C > 0$, then the posterior distribution of this Dirichlet process mixture is strongly consistent at any bounded density. This is satisfied if ρ is a Poisson or Geometric distribution.

As stated in the introduction, this extends Theorem 3 of Petrone and Wasserman (2002) by showing that a boundedness assumption on f_0 is sufficient for posterior consistency; continuity is not necessary.

References

Al-Lazikani, B., J. Jung, Z. Xiang, and B. Honig (2001). Protein structure prediction. *Current Opinion in Chemical Biology 5*(1), 51–56.

Bhattacharya, A. and R. Bhattacharya (2012). *Nonparametric inference on manifolds*, Volume 2 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press, Cambridge. With applications to shape spaces.

Bhattacharya, A. and D. B. Dunson (2012). Strong consistency of nonparametric Bayes density estimation on compact metric spaces with applications to specific manifolds. *Ann. Inst. Statist. Math. 64*(4), 687–714.

Binette, O. (2018). Note on reverse pinsker inequalities. *arXiv:1805.05135*.

Binette, O. and S. Guillotte (2018). Bayesian nonparametrics for directional statistics. *arXiv:1807.00305*.

DeVore, R. A. and G. G. Lorentz (1993). *Constructive approximation*, Volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.

Devroye, L. P. and T. J. Wagner (1979). The l1 convergence of kernel density estimates. *The Annals of Statistics 7*(5), 1136–1139.

Jammalamadaka, S. R. and A. SenGupta (2001). *Topics in circular statistics*, Volume 5 of *Series on Multivariate Analysis*. World Scientific Publishing Co., Inc., River Edge, NJ.

Kruijer, W. and A. van der Vaart (2008). Posterior convergence rates for dirichlet mixtures of beta densities. *Journal of Statistical Planning and Inference 138*(7), 1981 – 1992.

Lennox, K. P., D. B. Dahl, M. Vannucci, R. Day, and J. W. Tsai (2010, 06). A dirichlet process mixture of hidden markov models for protein structure prediction. *Ann. Appl. Stat. 4*(2), 916–942.

Lennox, K. P., D. B. Dahl, M. Vannucci, and J. W. Tsai (2009). Density Estimation for Protein Conformation Angles Using a Bivariate von Mises Distribution and Bayesian Nonparametrics. *Journal of the American Statistical Association 104*(486), 586–596.

Petrone, S. (1999). Random bernstein polynomials. *Scandinavian Journal of Statistics 26*(3), 373–393.

Petrone, S. and L. Wasserman (2002). Consistency of Bernstein polynomial posteriors. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 64*(1), 79–100.

Sason, I. and S. Verdú (2016, Nov). f -divergence inequalities. *IEEE Transactions on Information Theory 62*(11), 5973–6006.

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

Xing, Y. and B. Ranneby (2009). Sufficient conditions for Bayesian consistency. *Journal of Statistical Planning and Inference 139*(7), 2479–2489.