Bayesian nonparametric estimation and asymptotics with misspecified density models

Pierpaolo De Blasi and Stephen G. Walker

Abstract In this paper we summarize some recent findings concerning Bayesian misspecified density models. We first discuss a prior summability condition for the posterior to accumulate around the densities in the model closest in the Kullback–Leibler sense to the data generating density. This condition is shown to be satisfied by popular nonparametric priors such as infinite mixtures of normal densities and Gaussian process priors. In smooth parametric models, the posterior shrinks at a $\sqrt{n}$-rate of convergence around the parameter value minimizing the Kulback–Leibler divergence. In this setting we show how Gaussian process priors can be used to consistently estimate the discrepancy of the parametric model from the data generating density. A novel Monte Carlo Markov Chain methods is devised for dealing with intractable normalizing constants.

Abstract *In questo lavoro verranno richiamati alcuni recenti risultati sulla stima Bayesiana per modelli di densità mal specificati. Si discuterà una condizione di sommabilità sulla distribuzione a priori sufficiente a garantire che la massa di probabilità a posteriori si concentri attorno a densità vicine, rispetto alla divergenza di Kullback-Leibler, a quella che genera i dati . Tale condizione è soddisfatta da ben note distribuzioni a priori non parametriche quali misture infinite di densità normali e processi Gaussiani. Nel caso di modelli parametrici che soddisfano opportune ipotesi di differenziabilià, la distribuzione a posteriori converge al tasso $\sqrt{n}$ attorno al valore del parametro che minimizza la divergenza di Kullback-Leibler. In questo contesto si mostrerà come usare processi Gaussiani per stimare la discrepanza del modello parametrico dalla densità che genera i dati.*

Key words: Asymptotics, Bayesian nonparametrics, Gaussian process, Kullback–Leibler divergence, misspecified model

---

Pierpaolo De Blasi (pierpaolo.deblasi@unito.it)
University of Torino, corso Unione Sovietica 218/bis, Torino 10134, Italy

Stephen G. Walker (s.g.walker@math.utexas.edu)
University of Texas at Austin, 1 University Station C1200 Austin, TX 78712, USA

1 Introduction

In this paper we sketch results that are extensively presented and proved in [1, 2] about posterior inference in misspecified density models. To set the notation, let P be a probability distribution on $\mathbb{R}$ dominated by the Lebesgue measure and p denote the corresponding density. The expectation of a random variable f with respect to a probability measure P is denoted $Pf = \int f(x)p(x)\mathrm{d}x$. We assume that $X_1, X_2, \ldots$ are i.i.d. observations, each distributed according to a probability measure P_0. Given a prior Π, supported on a set $\mathscr{P}$ of dominated probability measures, the posterior mass of a measurable subset $A \subset \mathscr{P}$ is

$$\Pi_n(A) = \int_A \prod_{i=1}^n p(X_i)\mathrm{d}\Pi(P) \Big/ \int_{\mathscr{P}} \prod_{i=1}^n p(X_i)\mathrm{d}\Pi(P).$$

The model is well specified when $P_0 \in \mathscr{P}$; in this case posterior consistency means that $\Pi_n\{P \in \mathscr{P} : \ d(P_0, P) > \varepsilon\} \to 0$, P_0–a.s., for every $\varepsilon > 0$, where d is a metric on $\mathscr{P}$. The model is misspecified when P_0 is at a positive distance from $\mathscr{P}$ according to $\inf_{P\in\mathscr{P}} D(P_0, P) = \delta_1 > 0$, where $D(P_0, P) = P_0 \log(p_0/p)$ is the Kullback-Leibler (KL) divergence. In Section 2 we do not assume the existence of a unique minimizer of the KL divergence and focus rather on a set of densities $\mathscr{P}_1$ associated with the minimum divergence δ_1. Interest is then in investigating whether the posterior accumulates around $\mathscr{P}_1$ according to

$$\Pi_n\{P \in \mathscr{P} : \ d(\mathscr{P}_1, P) > \varepsilon\} \to 0, \quad P_0\text{-a.s.} \tag{1}$$

where $d(\mathscr{P}_1, P) = \inf_{P_1\in\mathscr{P}_1} d(P_1, P)$.

Provided there exists a unique $P_1 \in \mathscr{P}$ such that $D(P_0, P_1) = \delta_1$, we define $C_0(x) = p_0(x)/p_1(x)$ as the correction function which measures the discrepancy of the model $\mathscr{P}$ from P_0. In Section 3 we consider Bayesian estimation of $C_0(x)$. Specifically, we consider a parametric model, $\mathscr{P} = \{P_\theta, \theta \in \Theta\}$ with density p_θ supported on a bounded interval I of $\mathbb{R}$, and a prior π on θ. Under mild regularity condition, the posterior $\pi(\theta|X_1, \ldots, X_n)$ accumulates in $n^{-1/2}$ neighborhoods of $\theta_0 = \arg\min_\theta D(P_0, P_\theta)$. A nonparametric envelope around $\mathscr{P}$ is defined as

$$p_{\theta,W}(x) = \frac{p_\theta(x)W(x)}{\int_I p_\theta(y)W(y)\mathrm{d}y} \tag{2}$$

for a non negative perturbation function $W(x)$. The infinite dimensional parameter of interest is then $C(x; \theta, W) = W(x)/\int_I p_\theta(y)W(y)\mathrm{d}y$. We specify a prior for W via Gaussian process priors and illustrate how coherent updating can proceed given that the standard Bayesian posterior from an unidentified model is inappropriate. For estimation, we describe Monte Carlo Markov Chain methods that deal with intractable normalizing constants in (2). An asymptotic validation is given in terms of accumulation of the posterior in L_1–neighborhoods of $C_0(x) = p_0(x)/p_{\theta_0}(x)$,

$$\Pi_n\left\{\int |C(x; \theta, W) - C_0(x)|\mathrm{d}x > \varepsilon\right\} \to 0, \quad P_0\text{–a.s.} \tag{3}$$

2 Misspecification in nonparametric models

We start by introducing some additional notation. The Hellinger distance between P and P' is denoted $H(P,P') = [\int (p(x)^{1/2} - p'(x)^{1/2})^2 \mathrm{d}x]^{1/2}$. The space of densities p, also denoted by $\mathscr{P}$, endowed with the Hellinger metric is complete and separable. For $\alpha \in (0,1)$, we consider the Hellinger integral $h_\alpha(p,p') = \int p^{1-\alpha} p'^{\alpha}$ and the family of divergences $D_\alpha(P,P') = \alpha^{-1}[1 - h_\alpha(p,p')]$, see [3] and references therein. We recall that $\lim_{\alpha\to 0} D_\alpha(P,P') = D(P,P')$ and $D_{1/2}(P,P') = H(P,P')^2$. Finally, upon definition of $\delta_1 = \inf_{P\in\mathscr{P}} D(P_0,P)$, we define

$$\mathscr{P}_1 = \{P \in \overline{\mathscr{P}} :\ D(P_0,P) \leq \delta_1\}$$

where $\overline{\mathscr{P}}$ is the Hellinger closure of $\mathscr{P}$. In order to establish (1), instead of defining neighborhoods of $p_1 \in \mathscr{P}_1$, see e.g. [6], we work with neighborhoods of p_0:

$$A_{\alpha,\varepsilon} = \{P :\ D_\alpha(P_0,P) > \delta_1 + \varepsilon/\alpha\}, \quad A_\alpha = \{P :\ D_\alpha(P_0,P) > \delta_1 + \alpha\}.$$

A_α^c is monotonically decreasing in α to $\mathscr{P}_1$,

$$\textstyle\bigcap_\alpha \overline{A_\alpha^c} \subseteq \mathscr{P}_1. \tag{4}$$

see Lemma 2 in [1]. Moreover, A_α is recovered by $A_{\alpha,\varepsilon}$ as $\alpha \to 0$ by letting ε depends on α such that $\varepsilon_\alpha/\alpha \to 0$. Without loss of generality, we take $\varepsilon = \alpha^2$, so that (1) is equivalent to $\Pi_n(A_{\alpha,\alpha^2}) \to 0$ for all α sufficiently small. As it is customary in Bayesian asymptotics, we deal with the numerator and denominator of $\Pi_n(A_{\alpha,\alpha^2})$ separately. As for the denominator, we extend the usual prior positivity condition on KL neighborhoods of P_0 to the misspecified case as follows,

$$\Pi(P \in \mathscr{P} :\ D(P_0,P) \leq \delta_1 + \eta) > 0 \tag{5}$$

for any $\eta > 0$, cfr. Theorem 2.1 in [6]. As for the numerator, the key condition can be stated, similar to [10], in terms of summability of powers of prior probabilities. Let $(B_{j,\varepsilon})_{j\geq 1}$ forms a Hellinger covering of $\mathscr{P}$ in terms of balls of size $\varepsilon > 0$. The main result can be stated as follows, see [1] for details.

Theorem 1. *For a given model $\mathscr{P}$ and prior Π, assume that* (5) *holds and that for some $\alpha \leq 1/2$ and $\varepsilon = 2(\varepsilon'/2)^{1/2\alpha}$,*

$$\sum_{j\geq 1} \Pi(B_{j,\varepsilon})^\alpha < \infty, \tag{6}$$

Then $\Pi_n(A_{\alpha,\varepsilon'}) \to 0$ P_0–a.s.

A corollary to Theorem 1 and (4) provides the sufficient condition for accumulation of the posterior at $\mathscr{P}_1$.

Corollary 1. *Assume that Π satisfies* (5) *and* (6) *for any $\alpha \leq 1/2$ and $\varepsilon = 2(\alpha^2/2)^{1/2\alpha}$. Then $\Pi_n\{P \in \mathscr{P} :\ d(\mathscr{P}_1,P) > \varepsilon\} \to 0$ P_0–a.s.*

Consistency in the well specified case is recovered from Theorem 1. In fact $\Pi_n(A_{\alpha,\varepsilon}) \to 0$ for $\alpha = 1/2$ and any $\varepsilon > 0$ corresponds to Hellinger consistency when $\delta_1 = 0$. It turns out that in the well specified case it is sufficient that the prior summability condition (6) is satisfied by an arbitrary power of α.

Corollary 2. *Let $\delta_1 = 0$. Assume that Π satisfies* (5) *and* (6) *for some $\alpha \in (0,1)$. Then $\Pi_n\{P \in \mathscr{P} : d(P_0,P) > \varepsilon\} \to 0$, P_0–a.s.*

Corollaries 1 and 2 show clearly how sufficient conditions for consistency in the well specified case ($\delta_1 = 0$) are weaker than in the misspecified case ($\delta_1 > 0$). We conclude this section by considering two popular nonparametric density models and illustrate the prior summability condition (6). The first example is given by infinite mixtures of normal densities,

$$p_{\sigma,F}(x) = \int \tfrac{1}{\sigma}\phi\left(\tfrac{x-\theta}{\sigma}\right) \mathrm{d}F(\theta)$$

with prior π on the scale parameter σ and a prior on the space of mixing distribution F with prior guess F^*. According to [5], (6) holds under $\pi(\sigma < 1/k) \le \mathrm{e}^{-\gamma k}$ for all $\gamma > 0$ and $F^*([-a,a]^c) = O(a^{-(1+r)})$ for $r > 1/\alpha - 1$. It can be shown that in the misspecified case one needs to restrict further the tail behavior of F^* to $F^*([-a,a]^c) \le \mathrm{e}^{-\eta a}$ for some $\eta > 0$. The second example is given by Gaussian process priors

$$p(x) = \frac{\mathrm{e}^{\mu(x)+Z(x)}}{\int_I \mathrm{e}^{\mu(s)+Z(s)}\mathrm{d}s}, \tag{7}$$

where $\mu(x)$ is a fixed continuous function and $Z(x)$ is a Gaussian process with covariance kernel $\sigma(x,y) = \sigma_0(\lambda x, \lambda y)$. Here σ_0 is a fixed covariance kernel and $\lambda \ge 0$ is assigned a prior $\pi(\lambda)$. See [8]. For smooth σ_0, it can be shown that the summability condition (6) boils down to the tail behavior of the prior on the scale parameter λ, in particular the key requirement is $\pi(\lambda > t) \sim \mathrm{e}^{-\gamma t^2}$ for some $\gamma > 0$ for both the well and misspecified cases.

3 Discrepancy in misspecified parametric models

In this section we turn to parametric models $\{p_\theta, \theta \in \Theta\}$, and assume there is no $\theta \in \Theta$ such that $p_0 = p_\theta$. Define θ_0 as the parameter value θ minimizing $p \mapsto P_0 \log(p_0/p_\theta)$, provided it exists and is unique. The discrepancy of the parametric model can be measured by a divergence of the type $P_{\theta_0} g(p_0/p_{\theta_0})$ for a convex function $g : \mathbb{R}_+ \to \mathbb{R}$ such that $g(1) = 0$, see [3]. Specifically, $g(u) = u\log u$ and $g(u) = (\alpha-1)^{-1}(u^\alpha - 1)$ yield the KL and the α-divergence $D_\alpha(P_0, P_{\theta_0})$, respectively. Such divergences can be used to undertake model selection and evaluate model adequacy. Therefore it is of interest to estimate the correction function

$$C_0(x) = p_0(x)/p_{\theta_0}(x)$$

$C_0(x)$ also coveys information about the local fit of the model $\{p_\theta, \theta \in \Theta\}$, the closer C_0 is to a constant function, the better the fit.

Let π be a prior measure on the parameter set Θ. Under mild regularity conditions, see Theorem 3.1 in [7],

$$\pi(|\theta-\theta_0|>M_n n^{-1/2}|X_1,\ldots,X_n)\to 0 \tag{8}$$

for every sequence $M_n\to\infty$. In order to estimate $C_0(x)$, consider model (2), obtained by perturbing p_θ by a nonnegative function W and Π a prior on W. (2) builds upon the Gaussian process prior (7) and its semi–parametric extension by [4]. In fact, (7) is an instance of (2) for $p_\mu(x)=\mathrm{e}^{\mu(x)}/\int_I \mathrm{e}^{\mu(s)}\mathrm{d}s$ perturbed by $W(x)=\mathrm{e}^{Z(x)}$. Therefore posterior consistency at P_0 implies that the posterior distribution of $\mathrm{e}^{Z(x)}/\int_I \mathrm{e}^{\mu(s)+Z(s)}\mathrm{d}s$ accumulates around the correction function $p_0(x)/p_\mu(x)$. In the present setting, we need to take into account the fact that model (2) is not identified. Note however that the conditional posterior distribution

$$\Pi(\mathrm{d}W|\theta,X_1,\ldots,X_n)\propto\Pi(\mathrm{d}W)\prod_{i=1}^n C(X_i;\theta,W) \tag{9}$$

is a valid update for learning about the correction function $p_0(x)/p_\theta(x)$ via $C(x;\theta,W)$. Hence, we consider the posterior mean $C_n(x;\theta)=\int C(x;\theta,W)\Pi(\mathrm{d}W|\theta,X_1,\ldots,X_n)$ as a functional depending on both the data and θ, to be estimated by using the parametric model $\{p_\theta,\theta\in\Theta\}$ in view of the asymptotic behavior in (8). To estimate $C_0(x)=p_0(x)/p_{\theta_1}(x)$ we then average $C_n(x;\theta)$ with respect to the parametric posterior, $\int C_n(x;\theta)\pi(\theta|X_1,\ldots,X_n)\mathrm{d}\theta$. This now is effectively pursued by sampling $C(x;\theta,W)$ with respect to the joint distribution of (θ,W) given by

$$\Pi_n(\mathrm{d}W,\mathrm{d}\theta):=\Pi(\mathrm{d}W|\theta,X_1,\ldots,X_n)\pi(\theta|X_1,\ldots,X_n)\mathrm{d}\theta \tag{10}$$

noting however that (10) is not a proper posterior distribution. The use of a formal semi–parametric Bayesian model to update (θ,W) would be through the posterior

$$\widetilde{\Pi}_n(\mathrm{d}W,\mathrm{d}\theta)\propto\Pi(\mathrm{d}W)\pi(\theta)\mathrm{d}\theta\prod_{i=1}^n p_{\theta,W}(X_i). \tag{11}$$

However, while (11) is appropriate for learning about p_0, it is not so for learning about (θ_0,C_0) due to the lack of identifiability of (2). A practical consequence is that the marginalized $\tilde{\Pi}(\theta|X_1,\ldots,X_n)=\int\tilde{\Pi}(\theta,W|X_1,\ldots,X_n)\mathrm{d}W$ has no interpretation, since it is not clear what parameter value this $\tilde{\Pi}$ is targeting. That is $\tilde{\Pi}(\theta\in A|X_1,\ldots,X_n)$ is meaningless as it is no longer clear what is the real parameter value the prior $\pi(\theta)$ is specifying beliefs on. Moreover, the posterior mean of $C(x;\theta,W)$ with respect to $\tilde{\Pi}_n(\mathrm{d}W,\mathrm{d}\theta)$ is not a valid estimator of $C_0(x)$ since the posterior (11) does not target any particular (θ,W).

As for sampling from (9), in order to tackle with the normalizing constant $\int W(y)\,p_\theta(y)\mathrm{d}y$, we take W to be bounded by 1 by mapping the Gaussian process Z through a logistic link,

$$W(x)=\frac{\mathrm{e}^{Z(x)}}{1+\mathrm{e}^{Z(x)}}. \tag{12}$$

In this setting, we can adapt techniques laid in [11]: based on

$$\sum_{k=0}^{\infty}\binom{n+k-1}{k}\left[\int p_\theta(y)\,(1-W(y))\,\mathrm{d}y\right]^k=\left(\frac{1}{\int W(y)\,p_\theta(y)\mathrm{d}y}\right)^n,$$

a suitable latent model which removes the normalizing constant is

$$p(k,s_1,\ldots,s_k,W|\theta,X_1,\ldots,X_n)=\binom{n+k-1}{k}\prod_{i=1}^{n}W(X_i)\prod_{l=1}^{k}(1-W(s_l))\;p_\theta(s_l).$$

Hence, in any MCMC algorithm for now estimating the posterior, or drawing samples from it, would need to sample the variables $(k,s_1,\ldots,s_k,W)$. See [2] for details.

Finally, we investigate the asymptotic behavior of the sequence of (pseudo) posteriors (10). Proposition 1 establishes posterior consistency of model (2) for fixed θ, that is with respect to the conditional posterior (9). The key condition is on $\mathscr{A}(\sigma)$, the reproducing kernel Hilbert space of the covariance kernel σ of Z, see [9] for a formal definition. Let $\bar{\mathscr{A}}(\sigma)$ be the closure of $\mathscr{A}(\sigma)$ with respect to the sup norm.

Proposition 1. *Let $W(x)$ be defined in* (12) *and p_θ be continuous and bounded away from* 0 *on I. Assume that $p_0(x)\in\bar{\mathscr{A}}(\sigma)$. Then there is some $d>0$ such that $\Pi\left\{H(p_0,p_{\theta,W})>\varepsilon|\theta,X_1,\ldots,X_n\right\}\le \mathrm{e}^{-dn}$, P_0–a.s., for every $\varepsilon>0$.*

Using Proposition 1, one can prove L_1–consistency at $C_0(x)$. See [2] for details.

Theorem 2. *Assume that the hypotheses of Theorem 3.1 in [7] and Proposition 1 are satisfied. Π_n is defined as in* (10). *Then $\Pi_n\left\{\int_I|C(x;\theta,W)-C_0(x)|\mathrm{d}x>\varepsilon\right\}\to 0$, P_0–a.s., for every $\varepsilon>0$.*

References

1. De Blasi, P., Walker S.G.: Bayesian asymptotics with misspecified models. Statistica Sinica **23**, 1299–1322 (2013a)
2. De Blasi, P., Walker S.G.: Bayesian estimation of the discrepancy with misspecified parametric models. Bayesian Analysis **8**, 781–800 (2013b)
3. Leise, F., Vajda, I.: On divergence and informations in statistics and information theory. IEEE Trans. Inform. Theory **52**, 4394–4412 (2006)
4. Lenk, P.J.: Bayesian semiparametric density estimation and model verification using a logistic Gaussian process. J. Comp. Graph. Statist. **12**, 548–565 (2003)
5. Lijoi, A., Prünster, I., Walker, S.G.: On consistency of nonparametric normal mixtures for Bayesian density density estimation. J. Amer. Statist. Assoc. **100**, 1292–1296 (2005)
6. Kleijn, B.J.K., van der Vaart, A.W.: Misspecification in infinite–dimensional Bayesian statistics. Ann. Statist. **34**, 837–887 (2006)
7. Kleijn, B.J.K., van der Vaart, A.W.: The Bernstein-von Mises theorem under misspecification. Electron. J. Stat. **6**, 354–381 (2006)
8. Tokdar, S., Ghosh, J.: Posterior consistency of Gaussian process priorsin density estimation. J. Statist. Plann. Inference **137**, 34–42 (2006)
9. van der Vaart, A.W., van Zanten, J.H.: Rates of contraction of posterior distributions based on Gaussian process priors. Ann. Statist. **36**, 1435–1463 (2008)
10. Walker, S.G.: New approaches to Bayesian consistency. Ann. Statist. **32**, 2028–2043 (2004)
11. Walker, S.G.: Posterior sampling when the normalizing constant is unknown. Comm. Statist. Simulation Comput. **40**, 784-792 (2011)