Bayesian Analysis (2013) **8**, Number 4, pp. 781–800

Bayesian Estimation of the Discrepancy with Misspecified Parametric Models

Pierpaolo De Blasi * and Stephen G. Walker †

Abstract. We study a Bayesian model where we have made specific requests about the parameter values to be estimated. The aim is to find the parameter of a parametric family which minimizes a distance to the data generating density and then to estimate the discrepancy using nonparametric methods. We illustrate how coherent updating can proceed given that the standard Bayesian posterior from an unidentifiable model is inappropriate. Our updating is performed using Markov Chain Monte Carlo methods and in particular a novel method for dealing with intractable normalizing constants is required. Illustrations using synthetic data are provided.

Keywords: Asymptotics, Bayesian nonparametrics, Semi–parametric density model, Gaussian process, Kullback–Leibler divergence, Posterior consistency

1 Introduction

We consider a semi–parametric density model which comprises two parts. One part is a parametric family of densities, $\{f_\theta(x),\ \theta \in \Theta\}$, which is assumed to be misspecified but is used as a possible approximation to the sampling density, labeled as $f_0(x)$. The other part is a nonparametric component which is used to model the discrepancy between $f_0(x)$ and the closest density, with respect to the Kullback–Leibler divergence, in the family $f_\theta(x)$. If θ_0 is the parameter value that identifies this density, i.e.

$$\theta_0 = \arg\min_{\theta\in\Theta} \left\{ -\int \log f_\theta(x)\, f_0(x)\, \mathrm{d}x \right\},$$

then the discrepancy of the parametric family can be measured via a divergence of the type $D(f_0, f_{\theta_0}) = \int f_{\theta_0}(x)\, g[f_0(x)/f_{\theta_0}(x)]\mathrm{d}x$, where g is a convex positive function such that $g(1) = 0$, see Liese and Vajda (2006). Such divergences can then be used to undertake model selection and adequacy. Therefore it is of interest to estimate the correction function

$$C_0(x) = f_0(x)/f_{\theta_0}(x),$$

and a convenient way of targeting this ratio is by perturbing $f_\theta(x)$ by a non-negative function $W(x)$,

$$f_{\theta,W}(x) = \frac{f_\theta(x)\, W(x)}{\int f_\theta(s)\, W(s)\mathrm{d}s}, \tag{1}$$

(provided the integral in the denominator exists) and to look at

$$C(x;\theta,W) = \frac{W(x)}{\int W(s)\, f_\theta(s)\, ds} \tag{2}$$

*University of Torino and Collegio Carlo Alberto, Torino, Italy, pierpaolo.deblasi@unito.it
†University of Texas at Austin, US, s.g.walker@math.utexas.edu

 DOI:10.1214/13-BA024

as the infinite dimensional parameter of interest. A graphical output of $C(x;\theta,W)$ conveys also information about the local fit of the parametric model, the general interpretation being that the closer the estimate of $C_0(x)$ is to a constant function, the better the fit.

Model (1) builds upon the Gaussian process prior of Leonard (1978); Lenk (1988) and its semi–parametric extension by Lenk (2003). Given a compact interval I on the real line, a density on I is defined by

$$f(x) = \frac{\mathrm{e}^{\mu(x)+Z(x)}}{\int_I \mathrm{e}^{\mu(s)+Z(s)}\mathrm{d}s}, \tag{3}$$

where $\mu(x)$ is a fixed continuous function and $Z(x)$ is a Gaussian process. Model (3) defines a prior distribution on the space of density functions on I with well known asymptotic properties such as posterior consistency and contraction rates. See Tokdar and Ghosh (2007), van der Vaart and van Zanten (2008, 2009) and De Blasi and Walker (2013). It is easy to see that (3) is an instance of (1) for $f_\mu(x) = \mathrm{e}^{\mu(x)} \Big/ \int_I \mathrm{e}^{\mu(s)}\mathrm{d}s$ perturbed by $W(x) = \mathrm{e}^{Z(x)}$. Therefore posterior consistency at $f_0(x)$ 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 $f_0(x)/f_\mu(x)$. An alternative way of building a nonparametric density model from a parametric one is by using the probability transform $f_\theta(x)g(F_\theta(x))$ for F_θ the cumulative distribution function of f_θ and g a density on $[0,1]$. A semi parametric model is obtained by placing a nonparametric prior on g, see Verdinelli and Wasserman (1998) and Rousseau (2008) where this model is used for goodness-of fit testing through Bayes factors. As noted by Tokdar (2007), model (1) can be recovered by taking $g(x) = \mathrm{e}^{Z(x)}/\int_0^1 \mathrm{e}^{Z(s)}\mathrm{d}s$ for $Z(x)$ a Gaussian process with covariance $\sigma(u,v)$ since then $f_\theta(x)g(F_\theta(x))$ takes form (1) with $W(x) = \mathrm{e}^{Z_\theta(x)}$ and $Z_\theta(x) = Z[F_\theta(x)]$ the Gaussian process with covariance $\sigma(F_\theta(x), F_\theta(y))$. In a frequentist setting, a construction similar to (1) has been considered by Hjort and Glad (1995), where an initial parametric density estimate of f_0 is multiplied with a nonparameteric correction factor estimated via kernel-type methods. The aim in Hjort and Glad (1995) is not model comparison, but rather showing that the this estimator of f_0 is more efficient than traditional kernel density estimators for f_0 in a broad neighborhood around the parametric family.

In this paper we discuss how the semi–parametric density $f_{\theta,W}(x)$ in (1) can be used to find a coherent and consistent estimation of $C(x;\theta,W)$ in (2) with Bayesian techniques. The problem to be faced is that (1) is an over–parametrized density whereas $C(x;\theta,W)$ targets the correction function $C_0(x)$ which is defined in terms of a particular value of θ, i.e. θ_0. In Section 2 we introduce and motivate an update for (θ,W) which deals with this problem by using the parametric model $\{f_\theta(x),\pi(\theta)\}$ to estimate a nonparametric functional. In Section 3 we derive an Markov Chain Monte Carlo(MCMC) sampling scheme which deals with the normalizing constant in (1) and we provide illustration of inference with synthetic data. In Section 4 we investigate the asymptotic behavior of the proposed update. Section 5 has some final concluding remarks.

2 Posterior Updating

Before we discuss how we update the parameters, so that we can properly explain the procedure, we consider a parametric model which is misspecified, and which can be exactly the $f_\theta(x)$ described in the semi–parametric model (1). So let us assume the observations are independent and identically distributed from some density function $f_0(x)$ which is not contained in the parametric family $f_\theta(x)$. Now talking about making inference on θ sounds hollow. There is a need to define exactly what parameter value of $\theta \in \Theta$ we are interested in learning about. Without such it would appear problematic to specify a $\pi(\theta)$ so that we can interpret the meaning of $\Pi(A) = \int_A \pi(\theta)\mathrm{d}\theta$ for all sets A. It would appear clear that we would want to learn about the best parameter value possible and this can adequately be defined in terms of the Kullback–Leibler divergence between the family $\{f_\theta(x) : \theta \in \Theta\}$ and $f_0(x)$. This is a real parameter value and hence it is possible to assign a probability (the prior) to such a value.

Moreover, it is also appropriate and foundationally sound to update the prior to the Bayes posterior in spite of the classification of $f_\theta(x)$ as misspecified. The justification can be derived from the representation theorem of symmetric densities obtained by de Finetti in the $\{0, 1\}$ setting and the more general representation by Hewitt and Savage (1955) in more general spaces of observations. The idea is that a sequence of guesses as to where the next observation is coming from, say $m_1(x)$ for the first observation, and subsequently, $m_n(x|x_1, \ldots, x_{n-1})$ for the nth, would form a joint density which would need to be symmetric, as the observations are independent and identically distributed. Every sequence of symmetric densities will adhere to the representation theorem and hence for all joint guesses $m(x_1, \ldots, x_n)$ this can be written as

$$m(x_1, \ldots, x_n) = \int_\Theta \prod_{i=1}^{n} f(x_i|\theta)\, \pi(\mathrm{d}\theta)$$

for some $\pi(\theta)$, even though one is acknowledging this sequence is simply guessing. The basic fact is that the representation theorem applies to all sequences of symmetric densities whether they be correct or misspecified models in the form of guesses. The less demanding version of Bayes theorem which applies to misspecified models is then to be found within the representation theorem. The fuller argument can be found in Walker (2013). Hence, with the data, and interest in the θ_0 which minimizes $-\int \log f_\theta(x)\, f_0(x)\, \mathrm{d}x$, the appropriate update of $\pi(\theta)$ is the Bayesian posterior; i.e.

$$\pi(\theta|x_1, \ldots, x_n) \propto \pi(\theta) \prod_{i=1}^{n} f_\theta(x_i). \tag{4}$$

It is also well known that this sequence of posteriors accumulates at θ_0 under general and mild regularity conditions, see Section 4. This fact provides an asymptotic validation that the Bayesian is indeed learning about θ_0 rather than any other parameter value.

We explain now how we update (θ, W) to estimate $C(x; \theta, W)$ in (2). It is clear that (1) is an over–parametrized density in the sense that for each θ there is a correction function $C(x; \theta, W)$, and hence a W, which yields the same density. However, for fixed

θ and a prior distribution $\pi(\mathrm{d}W)$ on the space $\mathbb{W}$ of possible perturbation functions W, the conditional posterior distribution

$$\pi(\mathrm{d}W|\theta, x_1, \ldots, x_n) \propto \pi(\mathrm{d}W) \prod_{i=1}^{n} f_{\theta,W}(x_i) \propto \pi(\mathrm{d}W) \prod_{i=1}^{n} C(x_i; \theta, W) \tag{5}$$

is a valid update for learning about the correction function $f_0(x)/f_\theta(x)$ via $C(x; \theta, W)$. Therefore, we keep the parametric model $\{f_\theta(x), \pi(\theta)\}$ as the working model as it is identifiable and hence it has a posterior distribution which is theoretically sound to construct and to interpret. Then we consider the posterior mean

$$C(x; \theta, x_1, \ldots, x_n) := \int_{\mathbb{W}} C(x; \theta, W)\, \pi(\mathrm{d}W|\theta, x_1, \ldots, x_n) \tag{6}$$

as a functional depending on both the data and θ, to be estimated by using the model $\{f_\theta(x), \pi(\theta)\}$. This estimation procedure would be valid whenever we are interested in a quantity $B(x; \theta, W)$ and the true $B_0(x)$ is strictly and only of the form $B_0(x) = B(x; \theta_0)$. Estimating the functional (6) in terms of the model $\{f_\theta(x), \pi(\theta)\}$ now means that our estimate of $C_0(x)$ is given by

$$\begin{aligned} C(x; x_1, \ldots, x_n) &:= \int_{\mathbb{W}} C(x; \theta, x_1, \ldots, x_n)\, \pi(\theta|x_1, \ldots, x_n)\mathrm{d}\theta \\ &= \int_{\Theta} \int_{\mathbb{W}} C(x; \theta, W)\, \pi(\mathrm{d}W|\theta, x_1, \ldots, x_n)\, \pi(\theta|x_1, \ldots, x_n)\mathrm{d}\theta. \end{aligned}$$

This now is effectively pursued by sampling $C(x; \theta, W)$ with respect to the joint distribution of (θ, W) given by equations (4) and (5), i.e.

$$\pi(\mathrm{d}W, \mathrm{d}\theta|x_1, \ldots, x_n) = \pi(\mathrm{d}W|\theta, x_1, \ldots, x_n)\pi(\theta|x_1, \ldots, x_n)\mathrm{d}\theta. \tag{7}$$

The mathematical justification of (7) is to be found in terms of estimation of the non-parametric functional (6) with the parametric model $\{f_\theta(x), \pi(\theta)\}$. In Section 4 we provide an asymptotic study of the proposed estimation procedure by proving that the joint distribution (7) accumulates in L_1-neighborhoods of $C_0(x)$, which in turns implies that $C(x; x_1, \ldots, x_n)$ provides a consistent estimator of $C_0(x)$.

On the other hand the use of a formal semi–parametric Bayesian model to update (θ, W) would be through the posterior

$$\tilde{\pi}(\mathrm{d}W, \mathrm{d}\theta|x_1, \ldots, x_n) \propto \pi(\mathrm{d}W)\pi(\theta)\mathrm{d}\theta \prod_{i=1}^{n} f_{\theta,W}(x_i). \tag{8}$$

However, while (8) is appropriate for learning about f_0; it is not so for learning about (θ_0, C_0) due to the lack of identifiability of (1). 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 $\int_{\mathbb{W}} \int_{\Theta} C(x; \theta, W)\, \tilde{\pi}(\mathrm{d}W, \theta|x_1, \ldots, x_n)\mathrm{d}\theta$

is not a valid estimator of $C_0(x)$ since the posterior (8) does not target any particular (θ, W).

An alternative ad hoc derivation of the joint distribution (7) is achieved by writing the formal semi–parametric posterior (8) as

$$\tilde{\pi}(\theta, W|x_1, \ldots, x_n) = \tilde{\pi}(\theta|W, x_1, \ldots, x_n)\tilde{\pi}(W|x_1, \ldots, x_n)$$

where

$$\begin{aligned}\tilde{\pi}(\theta|W, x_1, \ldots, x_n) &\propto \pi(\theta)\prod_{i=1}^n f_{\theta,W}(x_i) \\ &\propto \pi(\theta)\prod_{i=1}^n f_\theta(x_i) \times [\int W(s)f_\theta(s)\mathrm{d}s]^{-n}\end{aligned}$$

and then removing the term $[\int W(s)f_\theta(s)\mathrm{d}s]^{-n}$. This modification of the conditional of θ can be seen as a way of preventing estimation of θ to be confounded by estimation of W and it concurs with the notion that whether we are interested in W or not, beliefs about θ are unchanged. Within Bayesian analysis there is an increasing use of modifications to posterior distributions that do not strictly correspond to a full probability model. See Liu et al. (2008) for a review.

3 Inference via MCMC

The posterior update (4) for θ is a straightforward application of Bayes theorem to a parametric model and we do not believe taking space here to detail standard MCMC algorithms for parametric models is worthy. On the other hand, sampling from the conditional distribution (5) of W poses two problems. First, for any positive constant ϕ, $C(x; \theta, W) = C(x; \theta, \phi W)$, and so it is important to fix a scale through the prior distribution for W. Second, it is difficult to deal with the normalizing constant in posterior sampling. A way to deal with these problems is to fix the prior expectation of W and to constrain W to be bounded. Hence, rather than take $W(x) = e^{Z(x)}$ as in (3), which has been a standard practice in the literature, we take

$$W(x) = \frac{e^{Z(x)}}{1 + e^{Z(x)}}, \tag{9}$$

with $Z(x)$ a mean zero Gaussian process. Note that one can equally use other transformations to the unit interval, see Section 4 for a discussion. However we confine to the logistic link for the sake of illustration. In fact we do not believe the actual transformation is relevant as long as it is monotone and maps the real line onto the unit interval.

In this situation Walker (2011) describes a latent model which can deal with the intractable normalizing constant. This is based on the result that

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

suggesting that a suitable latent model which removes the normalizing constant is

$$p(x_1,\ldots,x_n,k,s_1,\ldots,s_k) = \left(\begin{array}{c} n+k-1 \\ k \end{array}\right) \prod_{i=1}^{n} W(x_i) \prod_{l=1}^{k} (1-W(s_l))\; f_\theta(s_l).$$

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

The remaining cause for concern is that $W(x)$ is an infinite dimensional object and so sampling using MCMC methods from this latent model is still not going to be simple. However, a common procedure here is to use a grid of points on I; we will in the illustrations be using $I = [0,1]$ and hence we split this I into intervals $I_j = ((j-1)/J, j/J]$ and define the process $W(x)$ as follows. We approximate the Gaussian process $Z(x)$ by the piecewise constant process $Z'(x) = Z_j$ whenever $(j-1)/J < x \leq j/J$ and $Z_j = Z((j-1/J))$. Hence we define $W(x) = W_j = \mathrm{e}^{Z_j}/(1+\mathrm{e}^{Z_j})$ for $(j-1)/J < x \leq j/J$.

1. Sampling the $(s_l)_{l=1}^k$ given k and $(W_j)_{j=1}^J$. With this set up it is possible, given k and Z, to sample the $\{s_l : l = 1,\ldots,k\}$ independently, as is required, from the density

$$f(s|\cdots) \propto \sum_{j=1}^{J} w_j\, f_{j,\theta}(s)$$

where

$$f_{j,\theta}(s) = \frac{f_\theta(s)\,\mathbf{1}(s \in I_j)}{F_\theta(I_j)}, \qquad w_j \propto (1-W_j) \times F_\theta(I_j)$$

with $F_\theta(I_j) = \int_{I_j} f_\theta(s)\,\mathrm{d}s$. Hence, the sampling of the s_l given k and Z is done in the usual way for sampling from a finite mixture model.

2. Sampling k given the $(s_l)_{l=1}^k$ and $(W_j)_{j=1}^J$. The sampling of k given the current values involves a reversible jump MCMC step (Green 1995; Godsill 2001) by using a technique detailed in Walker (2011). The idea is to complete an infinite sequence of $s = (s_1, s_2, \ldots)$ and to consider

$$p(k, s_1, \ldots, s_k, s_{k+1}, \ldots | W) \propto \left(\begin{array}{c} n+k-1 \\ k \end{array}\right) \left\{\prod_{l=1}^{k} (1-W(s_l))\, f_\theta(s_l)\right\} \times \prod_{l=k+1}^{\infty} f_\theta(s_l).$$

So when a proposal is made to move from k to $k-1$ or $k+1$, each with probability $1/2$ of being proposed, the accept probability is, for example when proposing to move to $k+1$, given by

$$\min\left\{1, \frac{p(k+1,s|W)}{p(k,s|W)}\right\} = \min\left\{1, \frac{n+k}{k+1}(1-W(s_{k+1}))\right\}$$

and s_{k+1} is sampled from $f_\theta(\cdot)$. Hence, due to the cancelation of terms, in order to implement this strategy we only need to actually sample $(s_1, \ldots, s_{k+1})$ at any iteration.

3. Sampling $(Z_j)_{j=1}^J$ given the $(s_l)_{l=1}^k$ and k. Finally, Z can be sampled as a J–multivariate Gaussian distribution. The exact form of the conditional distribution for Z_j is given by

$$\pi(Z_j|Z_{-j},\cdots) \propto \pi(Z_j|Z_{-j})\,\frac{e^{m_j\,Z_j}}{(1+e^{Z_j})^{l_j}},$$

where $m_j = \#\{x_i \in I_j\}$ and $l_j = \#\{x_i \in I_j\} + \#\{s_l \in I_j\}$. This can adequately be sampled using a rejection algorithm or a Metropolis–Hastings algorithm.

Thus, we can sample a (piecewise constant) $C(x;\theta,W)$ from the joint distribution (7) by, at each iteration of the MCMC, sampling a θ from the posterior (4) and then sample the corresponding $(k,s_1,\ldots,s_k,Z_1,\ldots,Z_J)$ and compute $C(x;\theta,W)$ in (2).

Here we present an illustration using simulated data; we take $n = 500$ independent and identically distributed samples from the density on $[0,1]$ given by $f_0(x) = 2(1-x)$, and we take

$$f_\theta(x) = \frac{\theta\,\exp(-x\theta)}{1-e^{-\theta}}.$$

It is then straightforward to compute that $\theta_0 \approx 2.15$. We first run the code with θ fixed at $\theta = 2$, that is we only update W according to (5), and at each iteration of the chain we compute $C(x;\theta,W)$ for such a fixed θ, with the integral being easy to perform as we are taking W to be piecewise constant. The correction function to be estimated is $f_0(x)/f_\theta(x) = (1-x)\,(1-e^{-2})\,\exp(2x)$. The code was run with a grid size of $1/50$ and a correlation of 0.9 between neighboring Z_j, and marginally each Z_j had a mean of 0 and a standard deviation of 1.5. For the examples we considered we did fix the correlation, and thus the covariance kernel of the Gaussian process, on the basis of the smoothness of the densities involved. For neighboring Z_j and $Z_{j'}$ we can anticipate high closeness and hence the appropriateness of the high correlation. We acknowledge in general that a hyper–parameter with associated prior would be needed if one is not sure that f_θ reflects the smoothness of f_0. But this is a straight forward procedure to achieve in practice and hence we felt it unnecessary to undertake on this occasion. The algorithm was run for 20,000 iterations and the estimate presented in Figure 1 is the average over all $C(x;\theta,W)$ from all iterations, which corresponds to $C(x;\theta,x_1,\ldots,x_n)$ in (6).

We then repeated the study by incorporating the prior for θ which is taken to be the improper prior on the positive reals proportional to $1/\theta$, and obtained posterior inference according to the proposed update (7). The estimate $C(x;x_1,\ldots,x_n)$ of $C_0(x)$ is presented in Figure 2, top panel. Samples from the posterior of $C(x;\theta,W)$ are given in lower panel of Figure 2. Note that the accuracy of estimating $C_0(x)$ is comparable with the case of fixed θ, an illustration of the asymptotic results of Remark 1 and Theorem 2. The shape of $C_0(x)$ illustrates clearly the lack of fit f_θ for x close to 1 as: in fact $f_0(x) \to 0$ while $f_{\theta_0}(1)$ remains bounded away from 0. As pointed out by an Associate Editor, a graphical output of $C_0(x)$ is not immediately informative about the discrepancy of the parametric model as measured by the divergence $D(f_0,f_{\theta_0}) = \int f_{\theta_0}(x)\,g[C_0(x)]\mathrm{d}x$. In fact, for $g(t) = t\log t$ (Kullback-Leibler divergence), $D(f_0,f_{\theta_0}) = \int f_0(x)\,\log C_0(x)\mathrm{d}x$, so departures of $C_0(x)$ from 1 are weighted differently according to the size of f_0. The

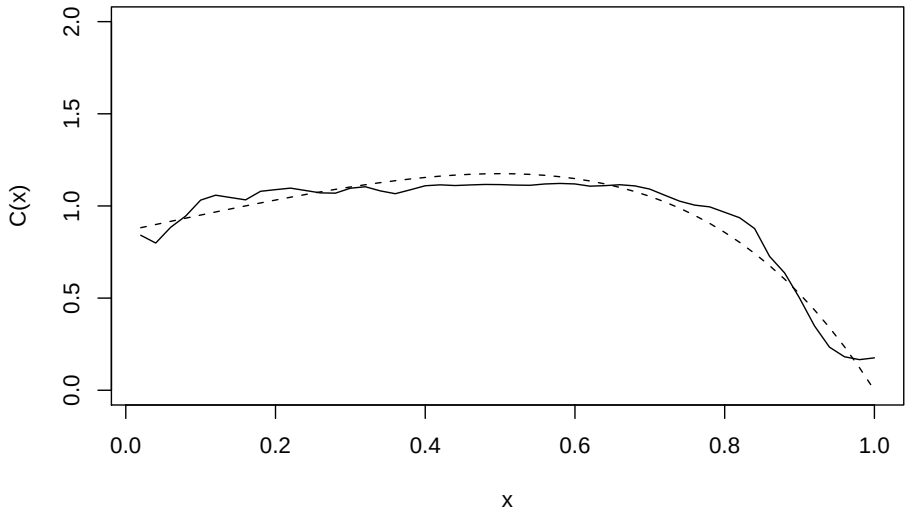

Figure 1: Estimated (bold) and true (dashed) function $f_0(x)/f_\theta(x)$ with θ fixed at 2.

histogram representation of the posterior (4) for θ is presented in Figure 3. The posterior mean for θ is given by 2.13, very close to the actual value of θ_0 as expected, since the posterior distribution for the parametric model is consistent at θ_0, see Theorem 1 below.

Finally, we compare update (7) with the formal semi–parametric update (8) using the same set up as before save this time we sample from the conditional posterior for θ given by

$$\tilde{\pi}(\theta|x_1,\ldots,x_n,k,s_1,\ldots,s_k) \propto \pi(\theta)\prod_{i=1}^{n} f_\theta(x_i)\prod_{l=1}^{k} f_\theta(s_l).$$

The estimated $C_0(x)$ is now presented in Figure 4 and the histogram representation of the formal semiparametric posterior for θ is presented in Figure 5. The posterior mean for θ in this case is given by 1.97. The plots indicate that both the formal semi–parametric update (8) and the proposed update (7) provide a suitable estimate for $C(x;\theta,W)$, which is not surprising given the flexibility of model (1) to estimate $C_0(x)$ with alternative values of θ. Yet the formal semi–parametric posterior for θ is less accurate than with the parametric posterior (4), a difference which did not disappear when we increased the sample size (indeed, $n = 500$ was already quite large).

To further explore the difference between the parametric and the formal semi–parametric

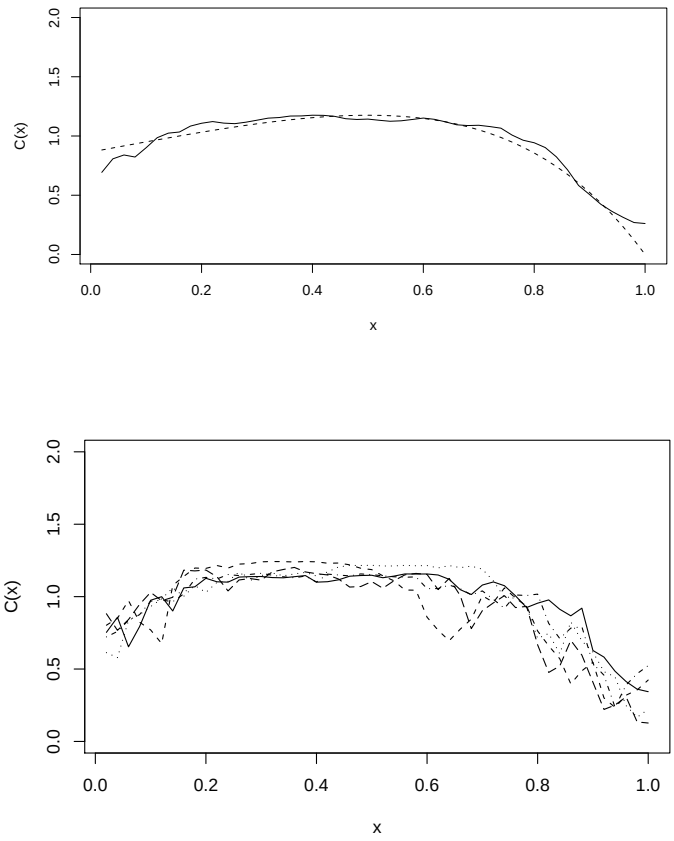

Figure 2: Estimated (bold) and true (dashed) functions of $C_0(x)$ with update (7) (top); samples of $C(x;\theta,W)$ (bottom).

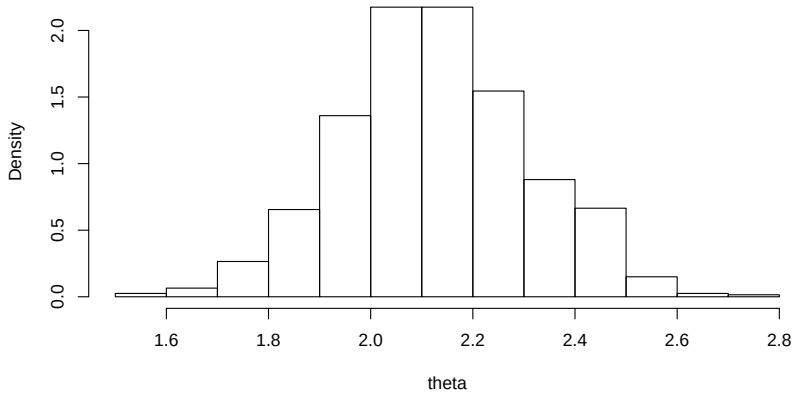

Figure 3: Posterior distribution of θ with update (4). Posterior mean is 2.13.

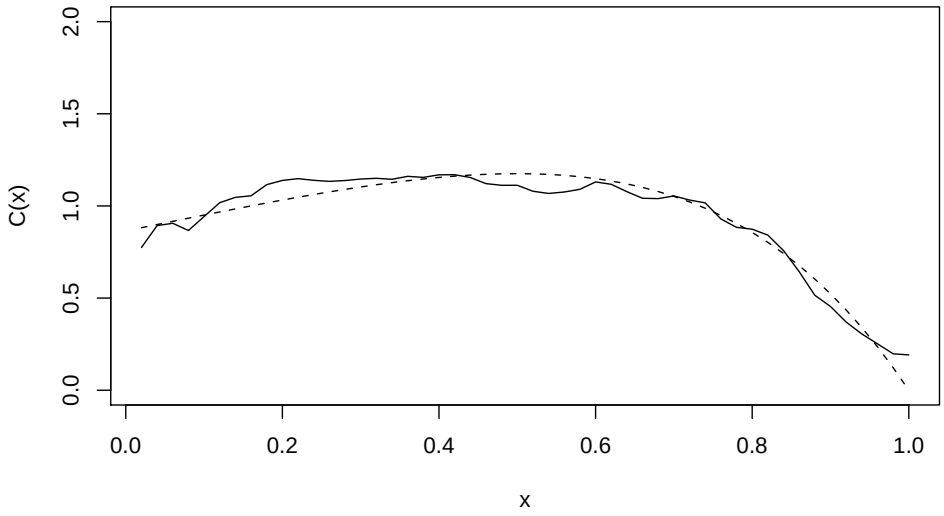

Figure 4: Estimated (bold) and true (dashed) functions of $C_0(x)$ using formal semi–parametric posterior (8).

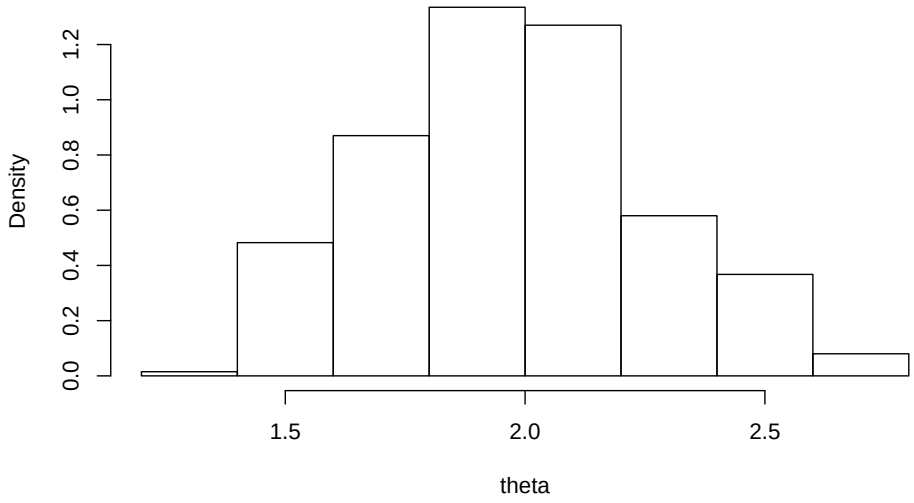

Figure 5: Posterior distribution of θ based on formal semi–parametric posterior update (8). Posterior mean is 1.97.

updates for θ we use the model

$$f_\theta(x) = \theta x^{\theta-1}$$

with $0 < \theta, x < 1$ and a uniform prior for θ. If $f_0(x) = 2x$ then $\theta_0 = 1$. Based on a sample of $n = 50$ we obtain the posterior distributions presented in Figure 6; the top is based on the parametric posterior update and is accumulating at 1, whereas the formal semiparametric update is accumulating at around 0.9. Furthermore for this example we draw the estimated $C_0(x)$ functions for both methods. This is in Figure 7 where it can clearly be seen that a better estimate of the true correction function $C_0(x) = 2x$ is provided by update (7).

4 Asymptotics

In this section we study the asymptotic behavior of the update (7) when the data $X_1, \ldots, X_n$ are i.i.d. from f_0. We recall the definition of θ_0 as the parameter value that minimizes $-\int \log f_\theta(x)\, f_0(x)\, \mathrm{d}x$ and of $C_0(x) = f_0(x)/f_{\theta_0}(x)$ as the correction function. Theorem 1 provides a rate of convergence of the posterior distribution (4) at θ_0, while Theorem 2 establishes that the conditional distribution of $C(x; \theta, W)$ corresponding to update (7) accumulates at C_0.

To set the notation, let $\mathbb{F}$ be the space of density functions on I and $\mathcal{C}(I)$ be the space of continuous functions on I. Also let F_0 denote the probability measure associated to f_0 and F_0^n stand for the associated n-fold product measure over the n-fold product space I^n. Integrals $\int_I g(x) f_0(x) \mathrm{d}x$ and $\int_{I^n} g(x_1, \ldots, x_n) \prod_{i=1}^n f_0(x_i) \mathrm{d}x_i$ are written as $F_0(g)$ and $F_0^n(g)$, respectively. For $f, g \in \mathbb{F}$, the Hellinger distance between f and g is $h(f,g) = [\int_I (\sqrt{f} - \sqrt{g})^2]^{1/2}$; the Kullback-Leibler divergence of g relative to f is $K(f,g) = \int_I f \log(f/g)$. Moreover, the sup norm and the L_1-norm of a real-valued function h are given by $\|h\|_\infty = \sup_{x \in I} |h(x)|$ and $\|h\|_1 = \int_I |h(x)| \mathrm{d}x$, respectively, while $\|\cdot\|$ stands for the Euclidean norm in Θ. Finally, the notation $\lesssim$ is used for "less than or equal to a constant times".

The following are regularity assumptions on f_θ.

A1) f_θ is continuous and bounded away from 0 on I for each $\theta \in \Theta$;

A2) $\theta \to \log f_\theta(X_1)$ is differentiable at θ_0 in F_0-probability with derivative $\dot{\ell}_{\theta_0}(X_1)$ and $F_0 \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}^T$ is invertible;

A3) there is an open neighborhood U of θ_0 such that for all $\theta_1, \theta_2 \in U$:

$$\|\log f_{\theta_1}/f_{\theta_2}\|_\infty \lesssim \|\theta_1 - \theta_2\|;$$

A4) $K(f_0, f_\theta)$ has a 2nd-order Taylor expansion around θ_0;

A5) $F_0(f_\theta/f_{\theta_0}) < \infty$ for all θ in a neighborhood of θ_0;

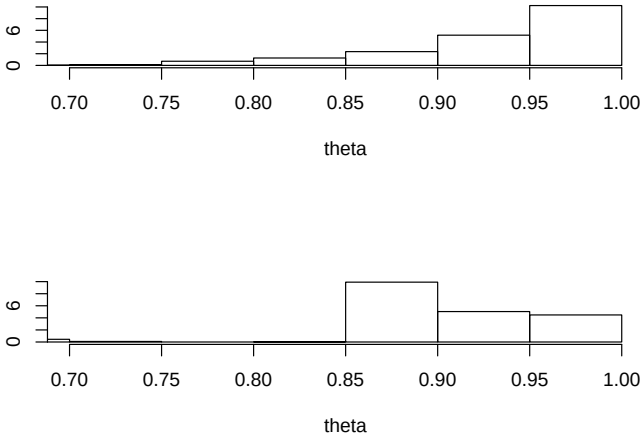

Figure 6: Posterior distributions of θ with update (4) (top) and formal semi–parametric update (8) (bottom).

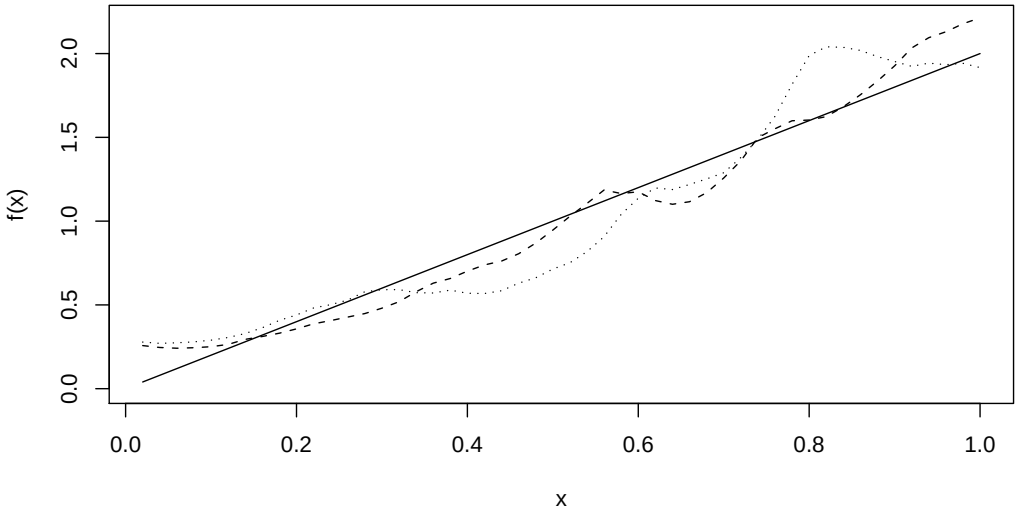

Figure 7: Posterior estimates of $C(x;\theta,W)$ based on (4) update (- - -) and formal semi–parametric update (8) (...), along side the true $C_0(x) = 2x$.

A6) for $Q_\theta(A) = \int_A f_0(f_\theta/f_{\theta_0})$, $Q_\theta(I) < \infty$ for all $\theta \in U$ and, for every $\epsilon > 0$, there exists a sequence of test (ϕ_n) such that

$$F_0^n \phi_n \to 0, \qquad \sup_{\{\theta:\|\theta-\theta_0\|>\epsilon\}} Q_\theta^n(1-\phi_n) \to 0$$

Assumption A1) is a technical condition needed for the proof of Proposition 1. Since I is a bounded interval of $\mathbb{R}$, it also implies that f_θ is bounded. Assumptions A2)–A6) reproduce the hypotheses of Theorem 3.1 of Kleijn and van der Vaart (2012), although the local property A3) is stronger than the corresponding one in Kleijn and van der Vaart (2012) and is later used in Theorem 2.

Theorem 1 (Kleijn and van der Vaart, 2012)**.** *Assume that $\pi(\theta)$ is continuous and positive in a neighborhood of θ_0 and that Assumptions A2)-A6) are satisfied. Then the posterior (4) converges at rate $1/\sqrt{n}$,*

$$\Pi\Big\{\theta : \|\theta-\theta_0\| > M_n n^{-1/2} | X_1, \ldots, X_n\Big\} \to 0, \tag{10}$$

in F_0^n-probability for any sequence $M_n \to \infty$.

We now discuss the choice of the prior distribution for W. We take $W(x)$ to be bounded by 1 by mapping a Gaussian process through a link function:

$$W(x) = \Psi(Z(x)),$$

where $\Psi(u)$ is a cumulative distribution function (cdf) on the real line and $Z(x)$ is a zero mean Gaussian process with covariance $\sigma(s,t)$. See Ghosal and Roy (2006) for a similar construction of priors for nonparametric binary regression. We further impose a Lipschitz condition on $\log \Psi(u)$ by assuming that Ψ is differentiable with bounded derivative $\psi(u)$ such that

$$\psi(u)/\Psi(u) < m, \text{ for any } u \in \mathbb{R} \tag{11}$$

for m a positive constant. Examples of $\Psi(u)$ satisfying (11) are the logistic cdf, the Laplace cdf and the Cauchy cdf. The probit link used in Ghosal and Roy (2006) violates this condition; indeed in Ghosal and Roy (2006) the Lipschitz condition is required directly on $\Psi(u)$.

Proposition 1 establishes posterior consistency of model (1) for fixed θ, that is with respect to the conditional posterior (5). The key condition is on $\mathcal{A}(\sigma)$, the reproducing kernel Hilbert space of the covariance kernel σ of Z, see van der Vaart and van Zanten (2008) for a formal definition. Let $\bar{\mathcal{A}}(\sigma)$ be the closure of $\mathcal{A}(\sigma)$ with respect to the sup norm.

Proposition 1. *Let $W(x) = \Psi(Z(x))$ for Ψ satisfying (11) and $\bar{\mathcal{A}}(\sigma) = \mathcal{C}(I)$. Let also f_θ satisfy assumption (A1) and $f_0(x)$ be continuous and positive on I. Then, for any $\epsilon > 0$, there is some $d > 0$ such that*

$$\Pi\left\{h(f_0, f_{\theta,W}) > \epsilon | \theta, X_1, \ldots, X_n\right\} \leq \mathrm{e}^{-dn},$$

in F_0^n-probability as $n \to \infty$.

Proof. The proof is based on an adaptation of the arguments used in van der Vaart and van Zanten (2008, Theorem 3.1) and amounts to (i) the derivation of inequalities (13) and (15) which relate the Hellinger distance and the Kullback-Leibler divergence with the sup distance on the space of real-valued functions $z(x)$; (ii) showing that the map $x \mapsto \Psi^{-1}\Big[f_0(x)/Mf_\theta(x)\Big]$ is a continuous function on I for M a large enough positive constant.

As for (i), let $p_z = f_{\theta,\Psi(z)}$. Using the Lipschitz condition (11),

$$\|\log \Psi(z_1) - \log \Psi(z_2)\|_\infty < m\|z_1 - z_2\|_\infty. \tag{12}$$

Since $\|\log \Psi(z_1) - \log \Psi(z_2)\|_\infty = \|\log f_\theta\Psi(z_1) - \log f_\theta\Psi(z_2)\|_\infty$ we can use the first inequality in van der Vaart and van Zanten (2008, Lemma 3.1) to conclude that

$$h(p_{z_1}, p_{z_2}) \leq m\|z_1 - z_2\|_\infty \mathrm{e}^{m\|z_1 - z_2\|_\infty/2}. \tag{13}$$

Moreover, (12) together with $\|\log p_{z_1}/p_{z_2}\|_\infty \leq 2\|\log f_\theta\Psi(z_1) - \log f_\theta\Psi(z_2)\|_\infty$, see last equation in the proof of Lemma 3.1 in van der Vaart and van Zanten (2008), implies

$$\|\log p_{z_1}/p_{z_2}\|_\infty < 2m\|z_1 - z_2\|_\infty \tag{14}$$

and an application of Lemma 8 in Ghosal and van der Vaart (2007) leads to

$$K(p_{z_1}, p_{z_2}) \lesssim m^2\|z_1 - z_2\|_\infty^2 \, \mathrm{e}^{m\|z_1 - z_2\|_\infty} \, (1 + 2m\|z_1 - z_2\|_\infty) \tag{15}$$

cf. the second inequality in van der Vaart and van Zanten (2008, Lemma 3.1). Following the lines of the proof of van der Vaart and van Zanten (2008, Theorem 3.1), we conclude that

$$\Pi\{Z : \ h(f_0, p_Z) > \epsilon | X_1, \ldots, X_n\} \to 0,$$

in F_0^n-probability as $n \to \infty$ provided that there exists $z(x) \in \mathcal{C}(I)$ such that $p_z = f_0$, that is there exists a constant M such that

$$z := \Psi^{-1}(f_0/Mf_\theta) \in \mathcal{C}(I).$$

Since $\Psi(u)$ is Lipschitz (as implied by (11)), the latter corresponds to continuity of f_0/Mf_θ and $0 < f_0/Mf_\theta < 1$. Continuity is implied by that of f_0 and f_θ. $f_0/Mf_\theta < 1$ holds whenever $M > \max_x f_0(x)/\min_x f_\theta(x)$, and existence of M is guaranteed by Assumption A1). Finally $f_0/Mf_\theta > 0$ is implied by the condition on $f_0(x)$ being positive on I. Finally, that convergence to 0 of the posterior probability can be made exponentially fast is then a side result of the way posterior consistency (and posterior contraction rate as well) is actually derived, see Choi and Ramamoorthi (2008). □

Remark 1. *It is well known that the Hellinger distance and the L_1-distance induce equivalent topologies on $\mathbb{F}$, therefore Proposition 1 can be also formulated in terms of L_1 neighborhood. Note that*

$$\begin{aligned}\|f_{\theta,W} - f_0\|_1 &= \int_I \Big|W(x)\Big/\textstyle\int_I f_\theta(s)W(s)\mathrm{d}s - f_0(x)/f_\theta(x)\Big| \, f_\theta(x)\mathrm{d}x \\ &\geq \min_{x\in I} f_\theta(x) \int_I |C(x;\theta,W) - f_0(x)/f_\theta(x)| \, \mathrm{d}x\end{aligned}$$

and $\min_{x\in I} f_\theta(x) > 0$ *by Assumption A1). Therefore, Proposition 1 implies that, for fixed* θ, $C(x;\theta,W)$ *in (2) consistently estimates the correction function* $f_0(x)/f_\theta(x)$ *in the* L_1*–metric.* □

Remark 2. *Here we discuss the case of* $f_0(x)$ *not being strictly positive. In this case one cannot rely on the continuity of the map* $x \mapsto \Psi^{-1}\left[f_0(x)/Mf_\theta(x)\right]$ *to conclude consistency by an application of van der Vaart and van Zanten (2008, Theorem 3.1). A possibility is to use a more general consistency result like Theorem 2 of Ghosal et al. (1999). We only mention here how to deal with the Kullback-Leibler condition, as the the entropy condition can be easily verified by using (13) along the lines of van der Vaart and van Zanten (2008, Theorem 2.1). Specifically, we use a technique laid down in Tokdar and Ghosh (2007, Theorem 4.6) which consists of constructing a strictly positive continuous density* f_1 *for which* $K(f_0,f_1)$ *is arbitrarily small. For fixed* $\epsilon > 0$, *define* $f_1(x) = (f_0(x)+\delta)/(1+\delta)$ *with* $\log(1+\delta) < \epsilon/2$. *Then*

$$K(f_0,p_z) = \log(1+\delta) + F_0\log[f_0/(f_0+\delta)] + F_0\log(f_1/p_z) \le \epsilon/2 + F_0\log(f_1/p_z).$$

Note that $f_1(x)$ *is strictly positive and bounded by* $(\max_x f_0(x)+\delta)/(1+\delta)$. *Therefore for* $M > \max_x f_1(x)/\min_x f_\theta(x) = [\max_x f_0(x)+\delta]/[(1+\delta)\min_x f_\theta(x)]$, $z_1 := \Psi^{-1}(f_1/Mf_\theta) \in \mathcal{C}(I)$ *by arguments similar those used in the proof of Proposition 1. Now, by the hypothesis made on the reproducing kernel Hilbert space* $\mathcal{A}(\sigma)$, $\Pi\{\|Z - z_1\|_\infty < \eta\} > 0$ *for any positive* η *and*

$$\begin{aligned}\Pi\{K(f_0,p_z) < \epsilon\} &\ge \Pi\{F_0\log(f_1/p_z) < \epsilon/2\} \ge \Pi\left\{\|\log(f_1/p_z)\|_\infty < \epsilon/2\right\}\\ &\ge \Pi\left\{\|Z - z_1\|_\infty < \epsilon/4m\right\} > 0\end{aligned}$$

where the last inequality follows from (14). □

We are now ready to state and prove the main result about the L_1-consistency of update (7) at $C_0(x)$.

Theorem 2. *Assume that the hypothesis of Theorem 1 and Proposition 1 are satisfied. Then, for every* $\epsilon > 0$,

$$\Pi\left\{\int_I |C(x;\theta,W) - C_0(x)|\mathrm{d}x > \epsilon | X_1,\ldots,X_n\right\} \to 0,$$

as $n \to \infty$ *in* F_0^n*-probability.*

Proof. The proof comprises two parts. We first show that, for any $\epsilon > 0$,

$$\Pi\{h(f_0, f_{\theta_0,W}) > \epsilon | X_1,\ldots,X_n\} \to 0. \tag{16}$$

Reasoning as in Remark 1, (16) implies that $\Pi\left\{\int_I |C(x;\theta_0,W) - C_0(x)|\mathrm{d}x > \epsilon | X_1,\ldots,X_n\right\} \to 0$, hence the thesis would follow by additionally showing that, for any $\epsilon > 0$,

$$\Pi\left\{\int_I |C(x;\theta,W) - C(x;\theta_0,W)|\mathrm{d}x > \epsilon | X_1,\ldots,X_n\right\} \to 0. \tag{17}$$

To simplify the notation, let $X_{1:n} = X_1, \ldots, X_n$, $A_{0,\epsilon} = \{W : \ h(f_0, f_{\theta_0,W}) > \epsilon\}$ and $I_{\theta,W} = \int_I f_\theta(x)W(x)\mathrm{d}x$. Split $\Pi(A_{0,\epsilon}|X_{1:n})$ as follows

$$\Pi(A_{0,\epsilon}|X_{1:n}) = \int_{\|\theta-\theta_0\|\leq M_n n^{-1/2}} \Pi(A_{0,\epsilon}|\theta, X_{1:n})\pi(\theta|X_{1:n})\mathrm{d}\theta + \int_{\|\theta-\theta_0\|> M_n n^{-1/2}} \Pi(A_{0,\epsilon}|\theta, X_{1:n})\pi(\theta|X_{1:n})\mathrm{d}\theta$$

where the second term in the right hand side vanishes to zero because of (10). As for the first term, we aim at establishing that

$$\sup_{\|\theta-\theta_0\|\leq M_n n^{-1/2}} \Pi(A_{0,\epsilon}|\theta, X_{1:n}) \to 0.$$

Using the notation $I_{\theta,W}$, we write

$$\Pi(A_{0,\epsilon}|\theta, X_{1:n}) = \frac{\int_{A_{0,\epsilon}} (I_{\theta_0,W}/I_{\theta,W})^n \prod_{i=1}^n f_{\theta_0,W}(x_i)\pi(W)\mathrm{d}W}{\int (I_{\theta_0,W}/I_{\theta,W})^n \prod_{i=1}^n f_{\theta_0,W}(x_i)\pi(W)\mathrm{d}W}.$$

It is easy to see that, for any W,

$$\exp\{-\|\log(f_{\theta_0}/f_\theta)\|_\infty\} \leq I_{\theta_0,W}/I_{\theta,W} \leq \exp\{\|\log(f_{\theta_0}/f_\theta)\|_\infty\} \tag{18}$$

so that, under Assumption A3),

$$\Pi(A_{0,\epsilon}|\theta, X_{1:n}) \leq \exp(2nc\|\theta_0 - \theta\|) \times \Pi(A_{0,\epsilon}|\theta_0, X_{1:n})$$

for some positive constant c. By Proposition 1, $\Pi(A_{0,\epsilon}|\theta_0, X_{1:n}) \leq \mathrm{e}^{-dn}$ for some positive d, hence

$$\sup_{\|\theta-\theta_0\|\leq M_n n^{-1/2}} \Pi(A_{0,\epsilon}|\theta, X_{1:n}) \leq \mathrm{e}^{2cM_n n^{1/2}}\mathrm{e}^{-dn}$$

and the right hand side goes to 0 F_0^∞- a.s. when $n \to \infty$ given the arbitrariness of the sequence M_n. This concludes the proof of (16).

As for (17), by (10) it is sufficient to show that, for any W,

$$\sup_{|\theta-\theta_0|\leq M_n n^{-1/2}} \int_I |C(x;\theta,W) - C(x;\theta_0,W)|\mathrm{d}x \to 0, \quad F_0^\infty\text{–a.s.} \tag{19}$$

Consider that

$$\int_I |C(x;\theta,W) - C(x;\theta_0,W)|\mathrm{d}x = \frac{1}{\mathbb{E}[f_{\theta_0}(S)]}\left|1 - \frac{\mathbb{E}[f_{\theta_0}(S)]}{\mathbb{E}[f_\theta(S)]}\right|$$

where $\mathbb{E}[f_\theta(S)]$ is the expected value of $f_\theta(S)$ for S distributed according to the density proportional to $W(s)$. Now, because of (18), Assumption A3) implies that, for any W,

$$\mathrm{e}^{-c\|\theta-\theta_0\|} \leq \frac{\mathbb{E}[f_{\theta_0}(S)]}{\mathbb{E}[f_\theta(S)]} \leq \mathrm{e}^{c\|\theta-\theta_0\|}$$

for some positive constant c. Also, $\mathbb{E}[f_{\theta_0}(S)] \geq \min_{x\in I} f_{\theta_0}(x)$ for any W. Finally, for $\|\theta - \theta_0\| \leq M_n n^{-1/2}$,

$$1 - \mathrm{e}^{cM_n n^{-1/2}} \leq \left| 1 - \frac{\mathbb{E}[f_{\theta_0}(S)]}{\mathbb{E}[f_\theta(S)]} \right| \leq 1 - \mathrm{e}^{-cM_n n^{-1/2}}$$

implies that

$$\int_I |C(x;\theta,W) - C(x;\theta_0,W)| \mathrm{d}x \leq \frac{\mathrm{e}^{cM_n n^{-1/2}} - 1}{\min_{x\in I} f_{\theta_0}(x)}$$

which goes to 0 as $n \to \infty$ by taking M_n diverging slow enough to infinity and using $\min_{x\in I} f_{\theta_0}(x) > 0$. From this result (19) follows and the proof is complete. □

5 Conclusions

We have discussed a popular form of semi–parametric density model. We argue that if interest is in density estimation then there is no need for a semi–parametric model. The semi–parametric model as we have set it up is specifically about estimating how good a particular parametric model is by examining the ratio between the true data generating density and the closest density within the parametric family to it with respect to the Kullback–Leibler divergence. We have also highlighted how this update and estimation can be done coherently and consistently which by necessity avoids the formal semi–parametric posterior update. Specifically, to learn about the parameter which minimizes the Kullback–Leibler divergence, the coherence and consistency of the parametric posterior update should have practical consequences and this is illustrated in density estimation examples.

References

Choi, T. and Ramamoorthi, R. V. (2008). "Remarks on consistency of posterior distributions." In Clarke, B. and Ghosal, S. (eds.), *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, volume 3, 170–186. Institute of Mathematical Statistics. 794

De Blasi, P. and Walker, S. G. (2013). "Bayesian asymptotics with misspecified models." *Statistica Sinica*, 23: 169–187. 782

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). "Posterior consistency of Dirichlet mixtures in density estimation." *The Annals of Statistics*, 27: 143–158. 795

Ghosal, S. and Roy, A. (2006). "Posterior consistency of Gaussian process prior for nonparametric binary regression." *The Annals of Statistics*, 34: 2413–2429. 793

Ghosal, S. and van der Vaart, A. W. (2007). "Posterior convergence rates of Dirichlet mixtures at smooth densities." *The Annals of Statistics*, 35: 697–723. 794

Godsill, S. J. (2001). "On the relationship between Markov chain Monte Carlo methods for model uncertainty." *Journal of Computational and Graphical Statistics*, 10: 230–248. 786

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82: 711–732. 786

Hewitt, E. and Savage, L. J. (1955). "Symmetric measures on Cartesian products." *Transaction of the American Mathematical Society*, 80: 470–501. 783

Hjort, N. L. and Glad, I. K. (1995). "Nonparametric density estimation with a parametric start." *The Annals of Statistics*, 23: 882–904. 782

Kleijn, B. J. K. and van der Vaart, A. W. (2012). "The Bernstein-Von Mises theorem under misspecification." *Electronic Journal of Statistics*, 6: 354–381. 793

Lenk, P. J. (1988). "The logistic normal distribution for Bayesian, nonparametric, predictive densities." *Journal of the American Statistical Association*, 83: 509–516. 782

— (2003). "Bayesian semiparametric density estimation and model verification using a logistic Gaussian process." *Journal of Computational and Graphical Statistics*, 12: 548–565. 782

Leonard, T. (1978). "Density estimation, stochastic processes and prior information (with discussion)." *Journal of the Royal Statistical Society - Series B*, 40: 113–146. 782

Liese, F. and Vajda, I. (2006). "On divergence and informations in statistics and information theory." *IEEE Transactions on Information Theory*, 52: 4394–4412. 781

Liu, F., Bayarri, M., and Berger, J. O. (2008). "Modularization in Bayesian analysis, with emphasis on analysis of computer models." *Bayesian Analysis*, 4: 119–150. 785

Rousseau, J. (2008). "Approximating interval hypothesis: p-values and Bayes factors." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics*, volume 8, 417–452. Oxford University Press. 782

Tokdar, S. (2007). "Towards a faster implementation of density estimation with logistic Gaussian process priors." *Journal of Computational and Graphical Statistics*, 16: 633–655. 782

Tokdar, S. and Ghosh, H. J. (2007). "Posterior consistency of Gaussian process priors in density estimation." *Journal of Statistical Planning and Inference*, 137: 34–42. 782, 795

van der Vaart, A. W. and van Zanten, J. H. (2008). "Rates of contraction of posterior distributions based on Gaussian process priors." *The Annals of Statistics*, 36: 1435–1463. 782, 793, 794, 795

— (2009). "Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth." *The Annals of Statistics*, 37: 2655–2675. 782

Verdinelli, I. and Wasserman, L. (1998). "Bayesian goodness–of–fit testing using infinite–dimensional exponential families." *The Annals of Statistics*, 26: 1215–1241. 782

Walker, S. G. (2011). "Posterior sampling when the normalizing constant is unknown." *Communications in Statistics - Simulation and Computation*, 40: 784–792. 785, 786

— (2013). "Bayesian inference with misspecified models." *Journal of Statistical Planning and Inference*. http://dx.doi.org/10.1016/j.jspi.2013.05.013i. 783

Acknowledgments

The authors are grateful for the comments of the Editor, an Associate Editor and two referees which have improved the paper. P. De Blasi is supported by the European Research Council (ERC) through StG "N-BNP" 306406. Support by Regione Piemonte is also acknowledged.