Bayesian estimation of the discrepancy with misspecified parametric models

Pierpaolo De Blasi

University of Torino & Collegio Carlo Alberto

Bayesian Nonparametrics workshop

ICERM, 17-21 September 2012

Joint work with S. Walker

Outline

Semiparametric density estimation

Asymptotics and illustration

References

BNP density estimation

- Let $X_1, \ldots, X_n$ be exchangeable (i.e. conditionally iid) observations from an unknown density f on the real line.
- If $\mathcal{F}$ is the density space and $\Pi(\mathrm{d}f)$ the prior, via Bayes theorem

$$\Pi(A|X_1, \ldots, X_n) = \frac{\int_A \prod_{i=1}^n f(X_i)\, \Pi(\mathrm{d}f)}{\int_{\mathcal{F}} \prod_{i=1}^n f(X_i)\, \Pi(\mathrm{d}f)}$$

- Wealth of Bayesian nonparametric (BNP) models
  - Dirichlet process mixtures of continuos densities;
  - log spline models;
  - Bernstein polynomials;
  - log Gaussian processes.
- All with well studied asymptotic properties, e.g. posterior concentration rates

$$\Pi(f : d(f, f_0) > M\epsilon_n | X_1, \ldots, X_n) \overset{n\to\infty}{\to} 0,$$

  when $X_1, X_2, \ldots$ are iid from some "true" f_0.

Discrepancy from a parametric model

- Suppose now we have a favorite parametric family

$$f_\theta(x),\ \theta \in \Theta \subset \mathbb{R}^p.$$

likely to be misspecified: *there is no θ such that $f_0 = f_\theta$.*

- We want to learn about the best parameter value θ_0 which minimizes the Kullback-Leibler divergence from true f_0:

$$\theta_0 = \arg\min_{\Theta} \int f_0 \log(f_0/f_\theta)$$

- A nonparametric component W is introduced to model the discrepancy between f_0 and the closest density f_{θ_0}:

$$f_{\theta,W}(x) \propto f_\theta(x)\, W(x),$$

so that

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

is designed to estimate $C_0(x) = f_0(x)/f_{\theta_0}(x)$.

Related works - Frequentist

Hjort and Glad (1995)

- Start with a parametric density estimate $f_{\hat{\theta}}(x)$, $\hat{\theta}$ being, e.g., the MLE of θ with respect to the likelihood $\prod_{i=1}^{n} \log f_\theta(x_i)$.
- Then multiply it with a nonparametric kernel-type of the correction function $r(x) = f_0(x)/f_{\hat{\theta}}(x)$:

$$\hat{f}(x) = f_{\hat{\theta}}(x)\hat{r}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)\frac{f_{\hat{\theta}}(x)}{f_{\hat{\theta}}(x_i)}$$

  in a *two-stage sequential analysis*.
- $\hat{f}$ is shown to be more precise than traditional kernel density estimator in a broad neighborhood around the parametric family, while losing little when the f_0 is far from the parametric family.

Related works - Bayes

Nonparametric prior built around a parametric model via

$$f(x) = f_\theta(x) g(F_\theta(x)),$$

where F_θ is the cdf of f_θ and g is a density on $[0,1]$ with prior Π.

- Verdinelli and Wasserman (1999): Π as an infinite exponential family. Application to goodness of fit testing.
- Rousseau (2008): Π as a mixtures of betas. Application to goodness of fit testing.
- Tokdar (2007): Π as a log Gaussian process prior. Application to posterior inference for densities with unbounded support.

  For $g(x) = \mathrm{e}^{Z(x)} / \int_0^1 \mathrm{e}^{Z(s)} \mathrm{d}s$ and Z Gaussian process with covariance $\sigma(\cdot,\cdot)$, $f(x)$ can be written

$$f(x) \propto f_\theta(x) \underbrace{\mathrm{e}^{\tilde{Z}(x)}}_{W(x)}$$

  for $\tilde{Z}$ Gaussian process with covariance $\sigma(F_\theta(\cdot), F_\theta(\cdot))$.

Posterior updating

$$f_{\theta,W}(x) \propto f_\theta(x)\, W(x), \quad C(x) := \frac{W(x)}{\int W(s)\, f_\theta(s)\, ds}.$$

- Truly semi–parametric: aim is at learning about the best parameter θ_0, then at seeing how close f_{θ_0} is to f_0 via $C(x) = W(x)/\int W(s)\, f_\theta(s)\, \mathrm{d}s$.
- Situation in which the updating process from prior to posterior may be seen as problematic:

$$\underline{\textit{the model } f_{\theta,W} \textit{ is intrinsically non identified in } (\theta, C)}$$

- The full Bayesian update

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

  is appropriate for learning about f_0; it is not so for learning about (θ_0, C_0).
- The marginal posterior $\tilde{\pi}(\theta|x_1, \ldots, x_n) = \int \tilde{\pi}(\theta, W|x_1, \ldots, x_n)\mathrm{d}W$ has no interpretation: it is not identified what parameter value this $\tilde{\pi}$ is targeting.

Posterior updating

- What removes us from the formal Bayes set–up is the desire to specifically learn about θ_0.
- θ_0 defined without any reference to W, or C. Whether we are interested in learning about C_0 or not, our beliefs about θ_0 should not change.
- Hence, the appropriate update for θ is the parametric one:

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

- We keep updating W according to the semi–parametric model,

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

so our updating scheme is

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

non-full Bayesian update

Posterior updating

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

- (θ, W) are estimated *sequentially*, with W reflecting additional uncertainty on θ.
- Marginalization of the posterior over W is *well defined*,

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

  since $\pi(\theta|x_1, \ldots, x_n)$ describes the beliefs about the real parameter θ_0.
- Coherence is about properly defining the quantities of interest and showing that Bayesian updates provide learning about these quantities and this is checked by what is yielded asymptotically.
- Hence we seek frequentist validation: we show that

the posterior of (θ, C) converges to a point mass at (θ_0, C_0).

Lenk (2003)

- Let I be a compact interval on the real line and Z a Gaussian process. Lenk (2003) considers the semi–parametric density model

$$f(x) = \frac{f_\theta(x)\, e^{Z(x)}}{\int_I f_\theta(s)\, e^{Z(s)} ds}$$

  for $f_\theta(x)$ member of the exponential family.
- In the Loève expansion of $Z(x)$, the orthogonal basis is chosen so that the sample paths integrate to zero.
- Further assumption for identification: the orthogonal basis does not contain any of the canonical statistics of $f_\theta(x)$.
- Estimation based on truncation of the series expansion or by imputation of the Gaussian process at a fixed grid of points, see Tokdar (2007).

Bounded $W(x)$

- Building upon Lenk (2003), we keep working with Gaussian processes and consider

$$f_{\theta,W}(x) = \frac{f_\theta(x)\, W(x)}{\int_I f_\theta(s)\, W(s)\mathrm{d}s}, \quad W(x) = \Psi(Z(x))$$

  where $\Psi(u)$ is a cdf having a smooth unimodal symmetric density $\psi(u)$ on the real line.
- With an additional condition on $\Psi(u)$, we can show that $W(x)$ preserves the asymptotic properties of log Gaussian process prior.
- On the other hand, with $W(x) \leq 1$, Walker (2011) describes a latent model which can deal with the intractable normalizing constant. It is based on

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

Link function $\Psi(u)$

- Lipschitz condition on $\log \Psi(u)$:

$$\psi(u)/\Psi(u) \leq m \quad \text{uniformly on } \mathbb{R}$$

satisfied by the standard Laplace cdf, standard logistic cdf or standard Cauchy cdf, but not by the standard normal cdf.

- For fixed θ, write $p_z = f_{\theta,\Psi(z)}$. It can be shown that, when $\|z_1 - z_2\|_\infty < \epsilon$,

$$\begin{cases} h(p_{z_1}, p_{z_2}) & \leq m\epsilon\, \mathrm{e}^{m\epsilon/2} \\ K(p_{z_1}, p_{z_2}) & \lesssim m^2\epsilon^2\, \mathrm{e}^{m\epsilon}\,(1 + m\epsilon) \end{cases}$$

- Posterior asymptotic results of van der Vaart and van Zanten (2008) carries over to this setting:

If $\Psi^{-1}(f_0/f_\theta)$ is contained in the support of Z, then

$$\Pi\{p_z : \ h(p_z, f_0) > \epsilon | X_1, \ldots, X_n\} \to 0, \ F_0^\infty - \text{a.s.}$$

Results on posterior contraction rate can be also derived.

Conditional posterior of W

(A) Lipschitz condition on $\log \Psi(u)$;

(B) $f_\theta(x)$ is continuous and bounded away from zero;

(C) the support of Z contains the space $C(I)$ of continuous densities on I.

Theorem 1. Under assumptions (A), (B) and (C), the conditional posterior of W given θ is *exponentially* consistent at all $f_0 \in C(I)$, i.e. for any $\epsilon > 0$,

$$\tilde{\pi}\{W : \ h(f_{\theta,W}, f_0) > \epsilon | \theta, X_1, \ldots, X_n\} \leq \mathrm{e}^{-dn}, \quad F_0^\infty - \text{a.s.}$$

for some $d > 0$ as $n \to \infty$.

- As corollary, for fixed θ, the posterior of $C(x) = W(x)/\int_I f_\theta(s) W(s) \mathrm{d}s$ consistently estimates the discrepancy $f_0(x)/f_\theta(x)$.
- The exponential convergence to 0 is a by-product of standard techniques for proving posterior consistency.

Marginal posterior of θ

- For given f_0, let θ_0 be the parameter value that minimize $\int_I f_0 \log(f_0/f_\theta)$:

$$\theta_0 = \arg\min_{\Theta} \int f_0 \log(f_0/f_\theta)$$

- Under some regularity condition on the family f_θ and on the prior at θ_0, the posterior accumulates at θ_0 with rate $\sqrt{n}$:

$$\pi\{|\theta - \theta_0| > M_n n^{-1/2} | X_1, \ldots, X_n\} \to 0, \; F_0^\infty - \text{a.s.}$$

  see Kleijn and van der Vaart (2012).

- One of the key regularity conditions on f_θ is the existence of an open neighborhood of θ_0 and a square-integrable function $m_{\theta_0}(x)$ such that, for all $\theta_1, \theta_2 \in U$,

$$|\log(f_{\theta_1}/f_{\theta_2})| \leq m_{\theta_0} |\theta_1 - \theta_2|, \quad P_0 - \text{a.s.}$$

- For our purposes, we focus on a different local property: there exist $\alpha > 0$ and an open neighborhood U of θ_0 such that for all $\theta_1, \theta_2 \in U$:

$$\|\log f_{\theta_1}/f_{\theta_2}\|_\infty \lesssim |\theta_1 - \theta_2|^\alpha \qquad \text{(D)}$$

Marginal posterior of W

Assume the regularity conditions on f_θ and $\pi(\theta)$ are satisfied for $f_0 \in C(I)$. Recall the definition of the marginal posterior of W,

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

Theorem 2. Under assumptions (A), (B), (C) and (D), the marginal posterior of $W(x)$ satisfies

$$\pi\left\{W:\ h(f_{\theta_0,W},f_0) > \epsilon|X_1,\ldots,X_n\right\} \to 0, \quad F_0^\infty - \text{a.s.}$$

as $n \to \infty$.

- The marginal posterior of W is evaluated outside a neighborhood defined in terms of θ_0. Clearly, if $\pi(\theta)$ is degenerate at θ_0, the result follows directly from Theorem 1.
- *Hint of the proof*: sufficient to consider the posterior when the prior is restricted on $|\theta-\theta_0| \le M_n n^{-1/2}$. We then manipulate numerator and denominator by using (D) together with the inequalities

$$\exp\{-\|\log(f_{\theta_0}/f_\theta)\|_\infty\} \le \frac{\int_I f_{\theta_0}(x)W(x)\mathrm{d}x}{\int_I f_\theta(x)W(x)\mathrm{d}x} \le \exp\{\|\log(f_{\theta_0}/f_\theta)\|_\infty\}$$

Marginal posterior of C

Recall the definition

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

which is designed to estimate $C_0(x) = f_0(x)/f_{\theta_0}(x)$.

Corollary. Under the hypotheses of Theorem 2, as $n \to \infty$,

$$\pi\left\{\int_I |C - C_0| > \epsilon | X_1, \ldots, X_n\right\} \to 0, \quad F_0^\infty\text{–a.s.}$$

- Together with $\pi\{|\theta - \theta_0| > M_n n^{-1/2} | X_1, \ldots, X_n\} \to 0$, we conclude that

  the posterior of (θ, C) *converges to* (θ_0, C_0).

- *Hint of the proof*: Theorem 2 implies that $\int_I |C_{\theta_0,W} - C_0|$ goes to 0. By triangular inequality, it is sufficient to show that, uniformly over $|\theta - \theta_0| \le M n^{-1/2}$, $\int_I |C_{\theta,W} - C_{\theta_0,W}| \to 0$. This we show by using (D).

Illustration 1

$n = 500$ observations from $f_0(x) = 2(1 - x)$;
$f_\theta(x) = \theta e^{-\theta x}/(1 - e^{-\theta})$ with improper prior $\pi(\theta) \propto 1/\theta$.

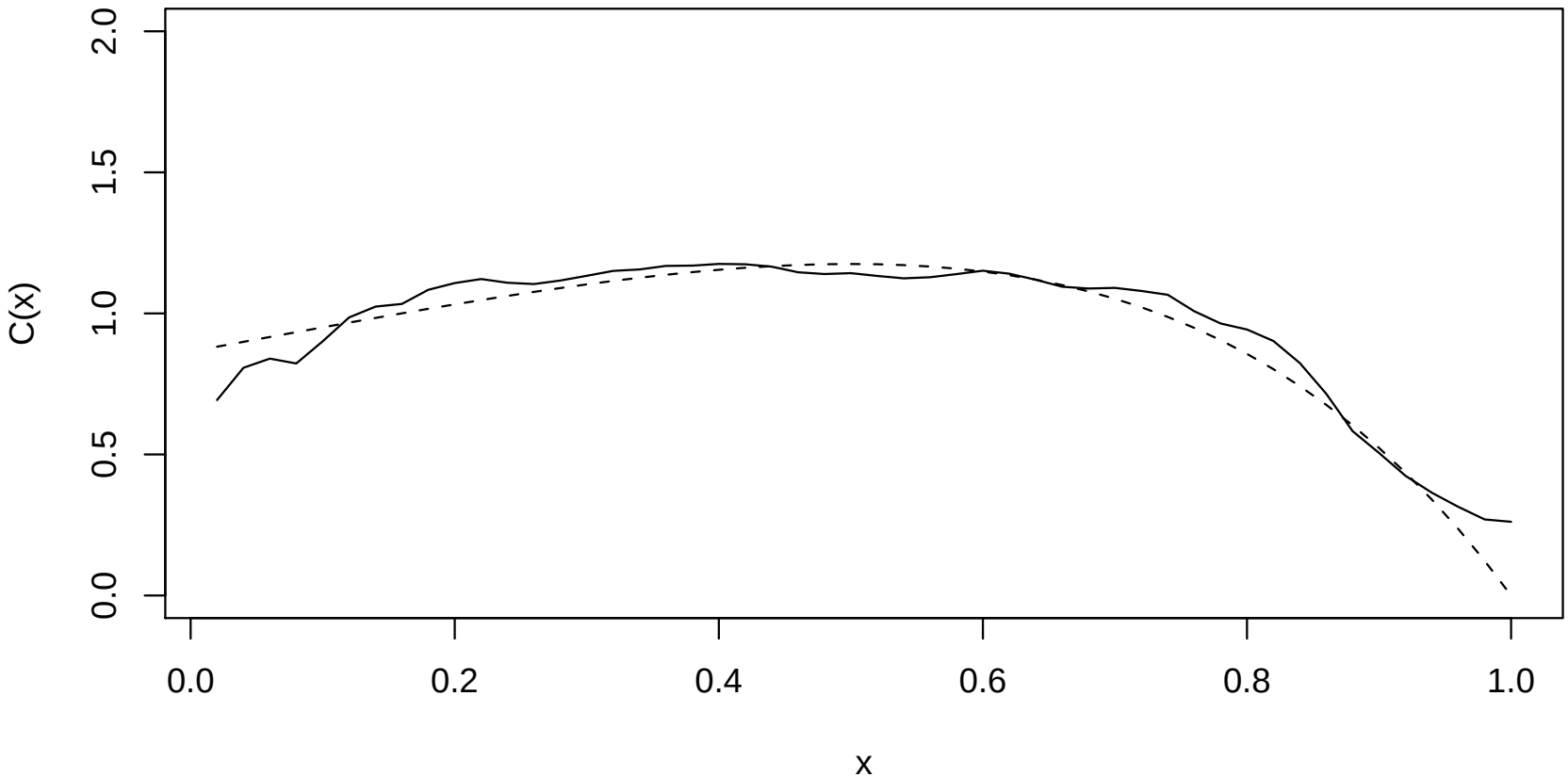

Figure: Estimated (bold) and true (dashed) functions of $C_0(x)$

Parametric Bayes update

Simulation from the posterior of θ. The minimum K-L parameter value is $\theta_0 = 2.15$.

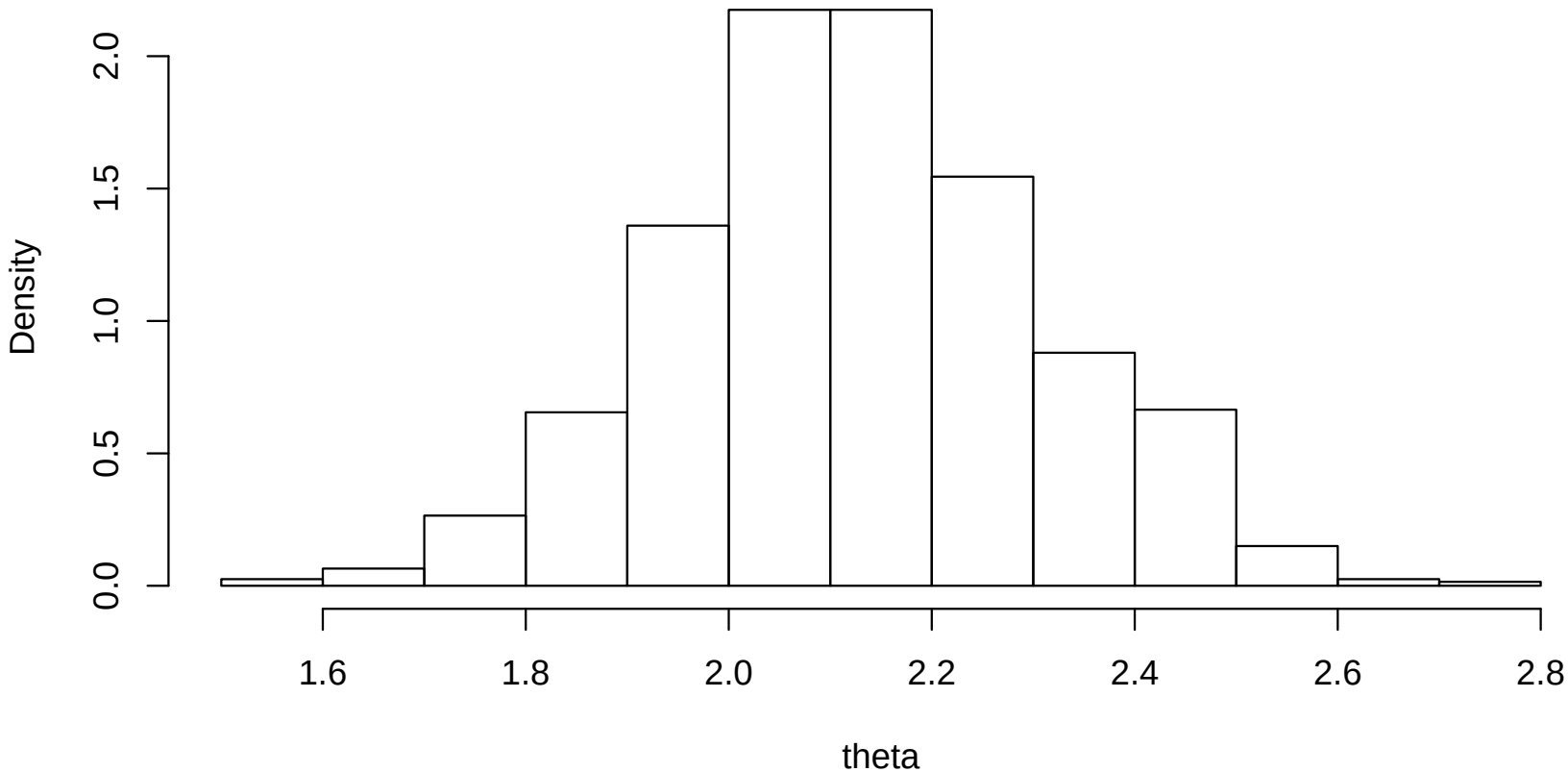

Figure: Posterior distribution of θ with parametric Bayes update. Posterior mean 2.13.

Full Bayes update

Using the proper conditional posterior $\tilde{\pi}(\theta|W, x_1, \ldots, x_n) \propto \pi(\theta) \prod_i f_{\theta,W}(x_i)$

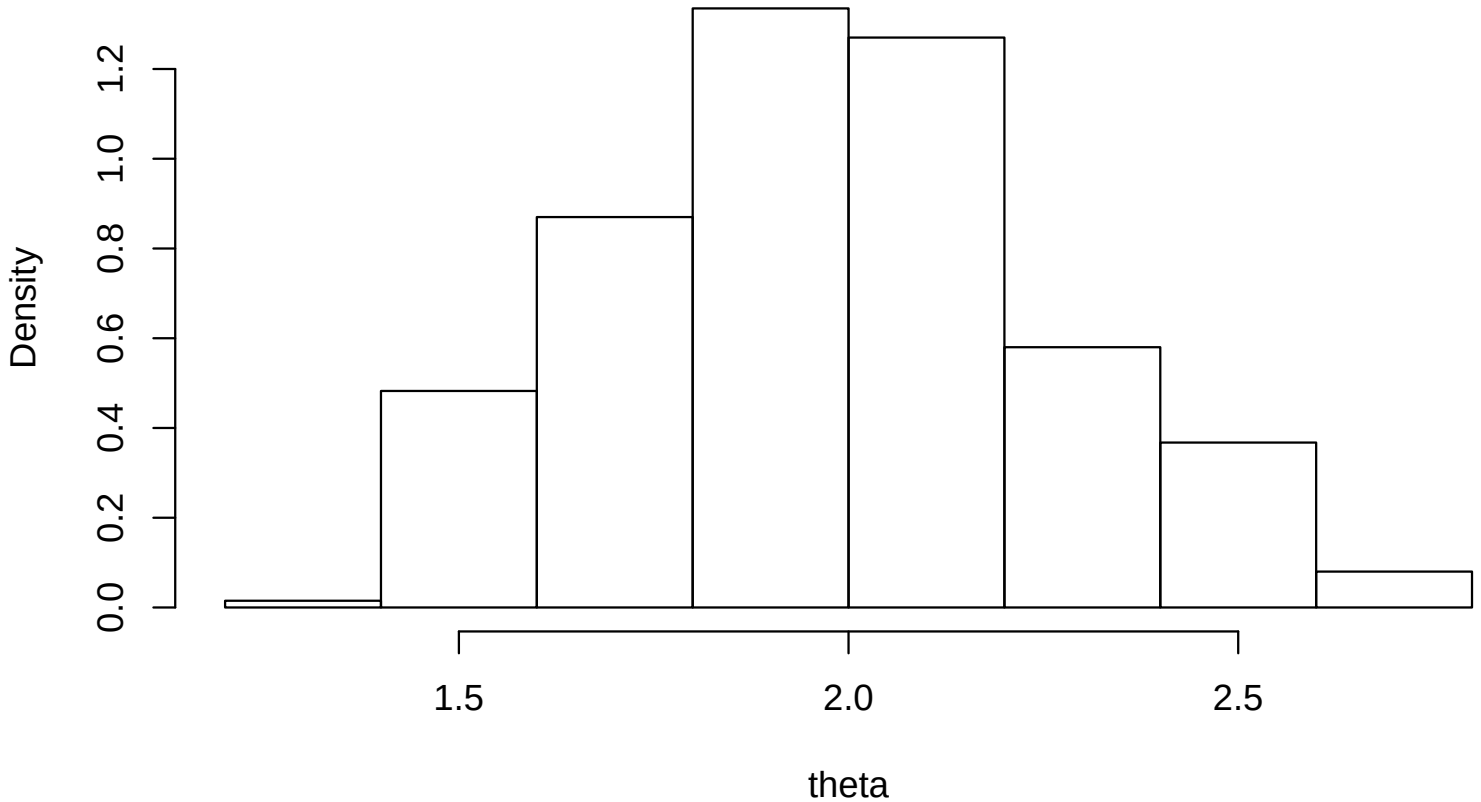

Figure: Posterior distribution of θ with formal Bayes update. Posterior mean 1.97.

Illustration 2

$n = 500$ observations from $f_0(x) = 2x$;
$f_\theta(x) = \theta x^{\theta-1}$ with $0 < \theta < 1$ and uniform prior for θ. $\theta_0 = 1$.

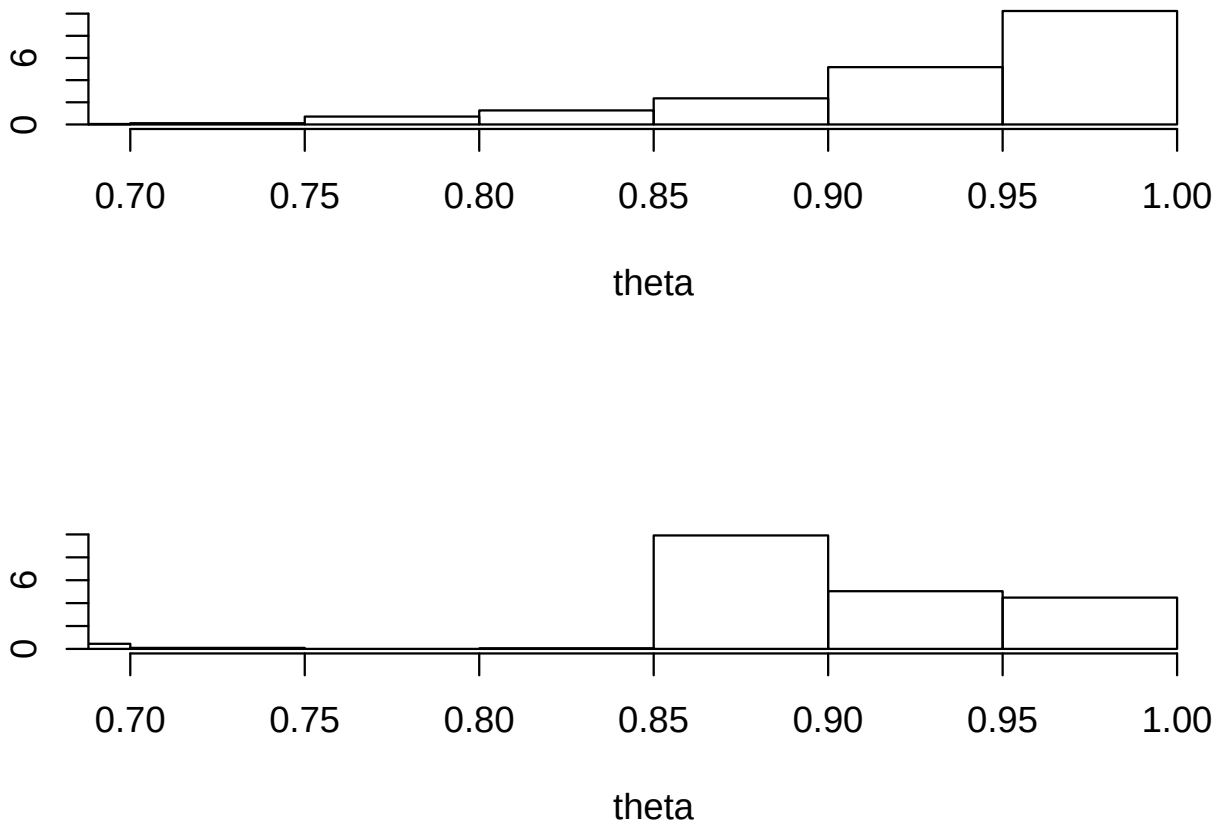

Figure: Posterior distributions of θ with parametric Bayes update (top) and formal Bayes update (bottom).

Discussion

- Both the proposed update and formal Bayes update seem to provide a suitable estimate for C, which is not surprising given its flexibility of estimating C_0 with alternative values of θ. Yet the posterior for θ is more accurate for the parametric Bayesian posterior.
- It shows that semiparametric models need to be thought about carefully: the parametric part needs to define which θ has been targeted.

Future work will deal with

- f_θ with unbounded support.
- Extension to posterior contraction rates.
- Connections with asymptotic properties of empirical Bayes.
- Use the $C(x)$ function for model selection

References

• De Blasi & Walker (2012). Bayesian estimation of the discrepancy with misspecified parametric models. *Tech. Rep.*, submitted.

• Hjort & Glad (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.* **23**, 882-904.

• Kleijn & van der Vaart (2012). The Bernstein-Von Mises theorem under misspecification. *Electron. J. Stat.* **6**, 354-381.

• Lenk (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.* **83**, 509-516.

• Rousseau (2008). Approximating interval hypothesis: p-values and Bayes factors. In *Bayesian Statistics 8*, 417-452.

• Tokdar (2007). Towards a faster implementation of density estimation with logistic Gaussian process priors. *J. Comp. Graph. Statist.* **16**, 633-655.

• van der Vaart & van Zanten (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36**, 1435-1463.

• Verdinelli & Wasserman (1998). Bayesuan goodness–of–fit testing using infinite–dimensional exponential families. *Ann. Statist.* **26**, 1215-1241.

• Walker (2011). Posterior sampling when the normalizing constant is unknown. *Comm. Statist. Simulation Comput.* **40**, 784-792.