

Bayesian Consistency and Smoothing

Stephen Walker
University of Texas at Austin
Leiden June 2019
Joint work with Minwoo Chae

Introduction

A prior with Kullback–Leibler support for p^* ; a sequence of posterior distributions can fail to be L_1 consistent if every so often the posterior puts mass on densities spiking at the data.

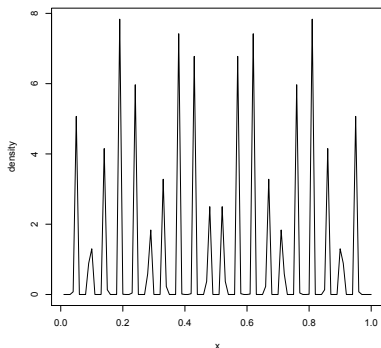


Figure: Spiking density at data points

Smoothing the spikes via a kernel is possible: in this case a uniform kernel on a small interval.

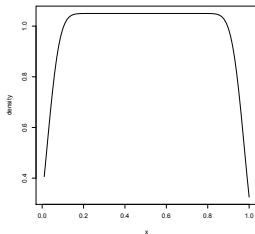


Figure: Smoothed spiking density

Assume working on bounded interval Ω ; e.g. $\Omega = (0, 1)$.
Suppose that

$$d_W(p_n, p^*) \rightarrow 0$$

where p_n is a predictive density and p^* a true density (data from which is constructing p_n), and

$$d_W(f, g) = \int |F(x) - G(x)| dx$$

with f and g densities and F and G corresponding distribution functions.

This is the Wasserstein 1 metric and metrizes weak convergence on Ω .

If the prior provides Kullback–Leibler support for p^* ; i.e.

$$\Pi \left(p : \int p^* \log(p^*/p) < \epsilon \right) > 0$$

for all $\epsilon > 0$, then

$$d_W(p_n, p^*) \rightarrow 0 \quad \text{a.s.}$$

where p_n is the predictive density

$$p_n(x) = \int p(x) \Pi(dp | x_1, \dots, x_n).$$

One smoothed version of p_n which is L_1 consistent is the Cèsaro average;

$$\bar{p}_N(x) = \frac{1}{N} \sum_{n=1}^N p_n(x).$$

That is

$$\bar{p}_N \rightarrow p^*$$

a.s. with respect to L_1 .

This is a bit complicated to compute, e.g. for normal mixture model, a Markov chain needs to run over all N predictives and moreover the $\bar{p}_N$ depends on an ordering of the data.

Define a smoothed version of p_n by

$$\tilde{p}_n(x) = \frac{1}{2\eta_n} [P_n(x + \eta_n) - P_n(x - \eta_n)],$$

where P_n is the distribution function of p_n and $\eta_n \downarrow 0$.

This is a density function and is a smoothed version of p_n ;

$$\tilde{p}_n(x) = \frac{1}{2\eta_n} \int_{x-\eta_n}^{x+\eta_n} p_n(y) dy.$$

Kernel smoothed with a uniform kernel.

If

$$d_W(p_n, p^*) \rightarrow 0$$

then

$$d_L(\tilde{p}_n, p^*) \rightarrow 0,$$

where d_L denotes the L_1 distance.

Now

$$d_L(\tilde{p}_n, p^*) \leq d_L(\tilde{p}_n, \tilde{p}^*) + d_L(\tilde{p}^*, p^*)$$

where

$$\tilde{p}^*(x) = \frac{1}{2\eta_n} [P^*(x + \eta_n) - P^*(x - \eta_n)]$$

is the corresponding smoothed version of p^* .

Now

$$|\tilde{p}_n(x) - \tilde{p}^*(x)| = \frac{1}{2\eta_n} |P_n(x + \eta_n) - P^*(x + \eta_n) - P_n(x - \eta_n) + P^*(x - \eta_n)|$$

so

$$d_L(\tilde{p}_n, \tilde{p}^*) \leq \frac{1}{\eta_n} d_W(p_n, p^*).$$

Also

$$|\tilde{p}^*(x) - p^*(x)| = \left| \frac{1}{2\eta_n} (P^*(x + \eta_n) - P^*(x - \eta_n)) - p^*(x) \right|,$$

can write as

$$\frac{\eta_n^2}{6} |p^{*''}(x_-) - p^{*''}(x_+)|$$

where $x - \eta_n < x_- < x$ and $x < x_+ < x + \eta_n$.

So if

$$\int |p^{*''}(x)| dx < \infty$$

then $d_L(\tilde{p}_n, p^*)$ converges to 0 at rate

$$\frac{\epsilon_n}{\eta_n} + A\eta_n^2;$$

$$i.e. \quad \max \left\{ \frac{\epsilon_n}{\eta_n}, \eta_n^2 \right\}$$

where $\epsilon_n \downarrow 0$ is the rate at which $d_W(p_n, p^*)$ converges to 0.

Hence, select $\eta_n = \epsilon_n^{1/3}$ yielding $d_L(\tilde{p}_n, p^*)$ converging to 0 at a rate of $\epsilon_n^{2/3}$.

If, on the other hand, Ω is not bounded, e.g. $\Omega = (-\infty, +\infty)$, then there is no guarantee that $d_W(p_n, p^*)$ converges to 0.

Would need extra condition that

$$\int |x| dP_n \rightarrow \int |x| dP^*.$$

However, as shown in Chae and Walker (2017), if

$$p^*(x) \text{ bounded and } \int |x| dP^*(x) < \infty$$

then

$$d_L(\tilde{p}_n, \tilde{p}^*) \leq \frac{d_P(p_n, p^*)}{2\eta_n}.$$

Here d_P represents the Lévy–Prokhorov distance;

$$d_P(p, q) = \inf_{\epsilon \geq 0} \{P(A) \leq Q(A^\epsilon) + \epsilon\}.$$

Hence, in this case can take

$$\eta_n = \epsilon_n^{1/3}$$

where $d_P(p_n, p^*)$ converges to 0 at a rate ϵ_n .

So as before $d_L(\tilde{p}_n, p^*)$ converges to 0 at a rate of $\epsilon_n^{2/3}$.

Can use prior $\tilde{\Pi}$ based on Π – a mixture of normals,

$$p(x) = \int N(x|\mu, \sigma^2) dQ(\mu);$$

change to

$$\tilde{p}(x) = \int \frac{1}{2\eta_n} \left[\Phi\left(\frac{x - \mu + \eta_n}{\sigma}\right) - \Phi\left(\frac{x - \mu - \eta_n}{\sigma}\right) \right] dQ(\mu).$$

So

$$\begin{aligned} \tilde{\Pi}(d_L(\tilde{p}, p^*) > \delta | \text{data}) &< \tilde{\Pi}(d_L(\tilde{p}, \tilde{p}^*) + d_L(\tilde{p}^*, p^*) > \delta | \text{data}) \\ &< \tilde{\Pi}(d_P(p, p^*)/\eta_n + C\eta_n^2 > \delta | \text{data}). \end{aligned}$$

Then

$$\tilde{\Pi}(d_L(\tilde{p}, p^*) > \delta | \text{data}) < \tilde{\Pi}(d_P(p, p^*) > \eta_n(\delta - C\eta_n^2) | \text{data}).$$

We can work out ϵ_n : suppose

$$\Pi\{p : K(p, p^*) < \epsilon_n^2\} > e^{-C n \epsilon_n^2},$$

then

$$E K(\bar{p}_n, p^*) \leq (1 + C) \epsilon_n^2,$$

where

$$\bar{p}_n(x) = \frac{1}{n} \sum_{i=1}^n p_i(x).$$

Therefore

$$E d_L(\bar{p}_n, p^*) \leq A \epsilon_n$$

and on a bounded interval

$$E d_W(\bar{P}_n, P^*) \leq B \epsilon_n.$$

Given that

$$d_W(P_n, P^*) \rightarrow 0 \quad \text{a.s.}$$

it follows that

$$\mathbb{E} d_W(P_n, P^*) \leq D \epsilon_n,$$

caveat: $\mathbb{E} d_W(P_n, P^*)$ decreasing.

On a bounded interval; if

$$\Pi\{p : K(p, p^*) < \epsilon_n^2\} > e^{-C n \epsilon_n^2},$$

and

$$\int |p^{*''}(x)| dx < \infty,$$

then

$$d_L(\tilde{p}_n, p^*) \rightarrow 0$$

in probability at a rate of $\epsilon_n^{2/3}$.

Comments:

Sample size dependent prior which provides strong consistency with just the support condition.

If Ω is unbounded can exchange Wasserstein distance for the Lévy–Prokhorov.

Can get better rates if assume higher derivatives exist.