Leave-One-Out Cross-Validation for Bayesian Model Comparison in Large Data - Supplementary Material

Johan Jonasson, Måns Magnusson, Michael Riis Andersen, Aki Vehtari

Proofs

The main quantity of interest is the mean expected log pointwise predictive density, which we want to use for model evaluation and comparison.

Definition 1 ($\overline{elpd}$)**.** *The* mean expected log pointwise predictive density *for a model p is defined as*

$$\overline{\text{elpd}} = \int p_t(x) \log p(x)\, dx$$

where $p_t(x) = p(x|\theta_0)$ is the true *density at a new unseen observation x and $\log p(x)$ is the log predictive density for observation x.*

We estimate $\overline{\text{elpd}}$ using *leave-one-out cross-validation (loo).*

Definition 2 (Leave-one-out cross-validation)**.** *The loo estimator $\overline{\text{elpd}}_{\text{loo}}$ is given by*

$$\overline{\text{elpd}}_{\text{loo}} = \frac{1}{n} \sum_{i=1}^{n} \pi_i, \tag{1}$$

where $\pi_i = \log p(y_i|y_{-i}) = \int \log p(y_i|\theta) p(\theta|y_{-i}) d\theta$.

To estimate $\overline{elpd}_{\text{loo}}$ in turn, we use difference estimator. Definitions follow.

Definition 3. *Let $\tilde{\pi}_i$ be any approximation of π_i. The difference estimator of $\overline{elpd}_{\text{loo}}$ based on $\tilde{\pi}_i$ is given by*

$$\widehat{\overline{\text{elpd}}}_{\text{loo,diff}} = \frac{1}{n} \left(\sum_{i=1}^{n} \tilde{\pi}_i + \frac{n}{m} \sum_{j \in \mathcal{S}} (\pi_j - \tilde{\pi}_j) \right),$$

where $\mathcal{S}$ is the subsample set, m is the subsampling size, and the probability of subsampling observation i is $1/n$, i.e. the subsample is uniform with replacement.

One important estimator of π_i among others is the importance sampling estimator

$$\log \hat{p}(y_i|y_{-i}) = \log\left(\frac{\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta_s) r(\theta_s)}{\frac{1}{S}\sum_{s=1}^{S} r(\theta_s)}\right), \tag{2}$$

where $r(\theta)$ is any suitable weight function such that $0 < r(\theta) < \infty$ for all $\theta \in \Theta$ and $(\theta_1, \ldots, \theta_S)$ is a sample from a suitable approximation of the posterior $p(\theta|y)$. We are in particular interested in the weight function

$$\begin{aligned} r(\theta_s) =& \frac{p(\theta_s|y_{-i})}{p(\theta_s|y)} \frac{p(\theta_s|y)}{q(\theta_s|y)} \\ \propto& \frac{1}{p(y_i|\theta_s)} \frac{p(\theta_s|y)}{q(\theta_s|y)} \end{aligned} \tag{3}$$

and where $q(\cdot|y)$ is an approximation of the posterior distribution that satisfies for each y that $q(\theta|y)$ iff $\theta \in \Theta$, θ_s is a sample point from q and S is the total posterior sample size. (The condition on q makes sure that $0 < r(\theta) < \infty$ for all θ.)

In the case of truncated importance sampling, we instead truncate these weights and replace r with r_τ given by

$$r_\tau(\theta_s) = \min(r(\theta_s), \tau), \tag{4}$$

where $\tau > 0$ is the weight truncation [see Ionides, 2008, for a more elaborate discussion on the choice of τ].

Proof of Proposition 1

Proposition 1. *The estimators* $\widehat{\text{elpd}}_{\text{diff}}$ *and* $\hat{\sigma}^2_{\text{loo}}$ *are unbiased with regard to* $\text{elpd}_{\text{diff}}$ *and* σ^2_{loo}.

Proof. We start out by proving unbiasedness for the general estimator. Write the difference estimator as

$$\widehat{\text{elpd}}_{\text{loo,diff}} = \sum_{i=1}^{n} \tilde{\pi}_i + \frac{n}{m} \sum_{i=1}^{n} \sum_{j \in \mathcal{S}} I_{ij}(\pi_j - \tilde{\pi}_j),$$

where I_{ij} is the indicator that data point i is chosen as the j'th point of the subsample. Since $\mathbb{E}[I_{ij}] = 1/n$, the expectation of the double sum is $\sum_i (\pi_i - \tilde{\pi}_i)$ and $\mathbb{E}[\widehat{\text{elpd}}_{\text{loo,diff}}] = \sum_i \pi_i$ as desired.

Next we prove unbiasedess of $\hat{\sigma}^2_{\text{loo,diff}}$. We are interested in estimating the finite sampling variance using the difference estimator. This can be done as

$$\sigma^2_{\text{loo}} = \frac{1}{n}\sum_{i=1}^{n}(\pi_i - \bar{\pi})^2 \tag{5}$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\pi_i^2}_{a} - \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\pi_i\right)^2}_{b} \tag{6}$$

We can estimate a and b separately as follows. The first part can be estimated using the difference estimator with $\tilde{\pi}_i^2$ as auxiliary variable. Let $t_\epsilon = \sum_i^n \epsilon_i = \sum_i^n \pi_i^2 - \tilde{\pi}_i^2 = t_{\pi^2} - t_{\tilde{\pi}^2}$, the we can estimate a as

$$\hat{a} = \frac{1}{n}(t_{\tilde{\pi}^2} + \hat{t}_\epsilon)\,,$$

where

$$\hat{t}_\epsilon = \frac{n}{m}\sum_{j\in\mathcal{S}}\left(\pi_j^2 - \tilde{\pi}_j^2\right)\,.$$

From the previous section, it follows directly that

$$E(\hat{a}) = \frac{1}{n}t_{\pi^2} = \frac{1}{n}\sum_{i=1}^{n}\pi_i^2,$$

The second part, b, can then be estimated as

$$\hat{b} = \frac{1}{n^2}\left[\hat{t}_e^2 - v(\hat{t}_e) + 2t_{\tilde{\pi}}\hat{t}_\pi - t_{\tilde{\pi}}^2\right]\,, \tag{7}$$

with the expectation

$$E(\hat{b}) = \frac{1}{n^2}\left[E(\hat{t}_e^2) - E(v(\hat{t}_e)) + 2t_{\tilde{\pi}}E(\hat{t}_\pi) - t_{\tilde{\pi}}^2\right] \tag{8}$$

$$= \frac{1}{n^2}\left[V(\hat{t}_e) + E(\hat{t}_e)^2 - V(\hat{t}_e) + 2t_{\tilde{\pi}}t_\pi - t_{\tilde{\pi}}^2\right] \tag{9}$$

$$= \frac{1}{n^2}\left[t_e^2 + 2t_{\tilde{\pi}}t_\pi - t_{\tilde{\pi}}^2\right] \tag{10}$$

$$= \frac{1}{n^2}\left[(t_\pi - t_{\tilde{\pi}})^2 + 2t_{\tilde{\pi}}t_\pi - t_{\tilde{\pi}}^2\right] \tag{11}$$

$$= \frac{1}{n^2}t_\pi^2 = \left(\frac{1}{n}\sum_i^n \pi_i\right)^2 \tag{12}$$

Using that

$$E(v(\hat{t}_e)) = n^2\left(1 - \frac{m}{n}\right)\frac{E(s_e^2)}{m} = n^2\left(1 - \frac{m}{n}\right)\frac{S_e^2}{m} = V(\hat{t}_e)\,. \tag{13}$$

Combining the results we have that

$$E(\hat{a} - \hat{b}) = \frac{1}{n}\sum_{i=1}^{n}\pi_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\pi_i\right)^2 = \sigma_{\text{loo}}^2\,. \tag{14}$$

$\square$

Remark. We believe this has probably been proven before, and hence this is probably not a new theoretical result.

Proof of Proposition 2 and 3

The proof follows, in general, the proof of Magnusson et al. [2019]. A generic Bayesian model is considered; a sample $(y_1, y_2, \ldots, y_n)$, $y_i \in \mathcal{Y} \subseteq \mathbb{R}$, is drawn from a true density $p_t = p(\cdot|\theta_0)$ for some true parameter θ_0. The parameter θ_0 is assumed to be drawn from a prior $p(\theta)$ on the parameter space Θ, which we assume to be an open and bounded subset of $\mathbb{R}^d$.

Several conditions are used. They are as follows.

(i) the likelihood $p(y|\theta)$ satisfies that there is a function $C : \mathcal{Y} \to \mathbb{R}_+$, such that $\mathbb{E}_{y\sim p_t}[C(y)^2] < \infty$ and such that for all θ_1 and θ_2, $|p(y|\theta_1) - p(y|\theta_2)| \leq C(y)p(y|\theta_2)\|\theta_1 - \theta_2\|$.

(ii) $p(y|\theta) > 0$ for all $(y, \theta) \in \mathcal{Y} \times \Theta$,

(iii) There is a constant $M < \infty$ such that $p(y|\theta) < M$ for all (y, θ),

(iv) all assumptions needed in the Bernstein-von Mises (BvM) Theorem [Walker, 1969],

(v) for all θ, $\int_{\mathcal{Y}}(-\log p(y|\theta))p(y|\theta)dy < \infty$.

Remarks.

- There are alternatives or relaxations to (i) that also work. One is to assume that there is an $\alpha > 0$ and C with $\mathbb{E}_y[C(y)^2] < \infty$ such that $|p(y|\theta_1) - p(y|\theta_2)| \leq C(y)p(y|\theta_2)\|\theta_1 - \theta_2\|^\alpha$. There are many examples when (i) holds, e.g. when y is normal, Laplace distributed or Cauchy distributed with θ as a one-dimensional location parameter.

- The assumption that Θ is bounded will be used solely to draw the conclusion that $\mathbb{E}_{y,\theta}\|\theta - \theta_0\| \to 0$ as $n \to \infty$, where y is the sample and θ is either distributed according to the true posterior (which is consistent by BvM) or according to a consistent approximate posterior. The conclusion is valid by the definition of consistency and the fact that the boundedness of Θ makes $\|\theta - \theta_0\|$ a bounded function of θ. If it can be shown by other means for special cases that $\mathbb{E}_{y,\theta}\|\theta - \theta_0\| \to 0$ despite Θ being unbounded, then our results also hold.

Proposition 2. *For any approximation $\tilde{\pi}_i$ that converges in L^1 to π_i, we have that $\widehat{\overline{\mathrm{elpd}}}_{\mathrm{loo,diff}}$ converges in L^1 to $\overline{\mathrm{elpd}}_{\mathrm{loo}}$.*

Proof. For convenience we will write $\hat{e} := \widehat{\overline{\mathrm{elpd}}}_{\mathrm{loo,diff}}$, which for our purposes is more usefully expressed as

$$\hat{e} = \frac{1}{n}\left(\sum_{i=1}^{n} \log \tilde{\pi}_i + \frac{n}{m}\sum_{i=1}^{n}\sum_{j=1}^{m} I_{ij}(\pi_i - \tilde{\pi}_i)\right),$$

where I_{ij} is the indicator that sample point y_i is chosen in draw j for the subsample used in $\hat{e}$.

We then get, with respect to all randomness involved (i.e. the randomness in generating y and the randomness in choosing the subsample in $\hat{e}$)

$$\begin{aligned}
\mathbb{E}|\hat{e} - \overline{\mathrm{elpd}}_{\mathrm{loo}}| &\leq \frac{1}{n}\mathbb{E}\left[\sum_{1}^{n} |\tilde{\pi}_i - \pi_i| + \frac{n}{m}\sum_{i=1}^{n}\sum_{j=1}^{m} I_{ij}|\pi_i - \tilde{\pi}_i|\right] \\
&= \mathbb{E}|\log \tilde{\pi}_i - \pi_i| + \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{n}\mathbb{E}|\pi_i - \tilde{\pi}_i| \\
&= 2\mathbb{E}|\tilde{\pi}_i - \pi_i| \\
&\to 0.
\end{aligned}$$

□

Proposition 3. *Let the subsampling size m and the number of posterior draws S be fixed at arbitrary integer numbers, let the sample size n grow, assume that (i)-(vi) hold and let $q = q_n(\cdot|y)$ be any consistent approximate posterior. Write $\hat{\theta}_q = \arg\max\{q(\theta) : \theta \in \Theta\}$ and assume further that $\hat{\theta}_q$ is a consistent estimator of θ_0. Then*

$$\tilde{\pi}_i \to \pi_i$$

in L^1 for any of the following choices of π_i, $i = 1, \ldots, n$.

(a) $\tilde{\pi}_i = \log p(y_i|y)$,

(b) $\tilde{\pi}_i = \mathbb{E}_y[\log p(y_i|y)]$,

(c) $\tilde{\pi}_i = \mathbb{E}_{\theta\sim q}[\log p(y_i|\theta)]$,

(d) $\tilde{\pi}_i = \log p(y_i|\mathbb{E}_{\theta\sim q}[\theta])$,

(e) $\tilde{\pi}_i = \log p(y_i|\hat{\theta}_q)$.

(f) $\tilde{\pi}_i = \log p(y_i|y) + V_{\theta\sim p(\cdot|y)}(\log p(y_i|\theta))$.

(g) $\tilde{\pi}_i = \log p(y_i|y) - \nabla \log p(y_i|\hat{\theta})^T \Sigma_\theta \nabla \log p(y_i|\hat{\theta})$ *for any given fixed $\hat{\theta}$ and where the covariance matrix is with respect to $\theta \sim p(\cdot|y)$.*

(h) $\tilde{\pi}_i = \log p(y_i|y) - \nabla \log p(y_i|\hat{\theta})^T \Sigma_\theta \nabla \log p(y_i|\hat{\theta}) - \frac{1}{2}\mathrm{tr}(\mathrm{H}_{\hat{\theta}}\Sigma_\theta \mathrm{H}_{\hat{\theta}})\Sigma_\theta)$ *for any given fixed* $\hat{\theta}$ *and where the covariance matrix is as in (g)*

(i) $\tilde{\pi}_i = \log p(y_i|\hat{\theta}_q) - \nabla \log p(y_i|\hat{\theta})^T \Sigma_\theta \nabla \log p(y_i|\hat{\theta})$ *for any given fixed* $\hat{\theta}$ *and where the covariance matrix is as in (g)*

(j) $\tilde{\pi}_i = \log p(y_i|y) - \nabla \log p(y_i|\hat{\theta})^T \Sigma_\theta \nabla \log p(y_i|\hat{\theta}) - \frac{1}{2}\mathrm{tr}(\mathrm{H}_{\hat{\theta}}\Sigma_\theta \mathrm{H}_{\hat{\theta}})\Sigma_\theta)$ *for any given fixed* $\hat{\theta}$ *and where the covariance matrix is as in (g)*

(k) $\tilde{\pi}_i = \log \hat{p}(y_i|y_{-i})$ *as defined in (2) for any weight function* r *such that* $r(\theta) > 0$ *for all* $\theta \in \Theta$*.*

Note. Part (k) holds in particular for the weight functions (3) and (4).

Remark. By the variational BvM Theorems of Wang and Blei [2019], q can be taken to be either q_{Lap}, q_{MF} or q_{FR}, i.e. the approximate posteriors of the Laplace, mean-field or full-rank variational families respectively in Proposition 3, provided that one adopts the mild conditions in their paper.

The proof of Proposition 3 will be focused on proving (a) and then (b)-(e) will follow easily and (f)-(l) with only a few simple observations on the posterior variance of θ. Note that parts (a)-(e) are contained in Magnusson et al. [2019] and the proof of them is identical to that. Proposition 3 follows immediately from the following lemma.

Lemma 4. *With all quantities as defined above,*

$$\mathbb{E}_{y\sim p_t}|\pi_i - \log p(y_i|\theta_0)| \to 0, \tag{15}$$

with any of the definitions (a)-(e) of π_i *of Proposition 3. Furthermore,*

$$\mathbb{E}_{y\sim p_t}|\log p(y_i|y_{-i}) - \log p(y_i|\theta_0)| \to 0, \tag{16}$$

as $n \to \infty$*.*

Proof. To avoid burdening the notation unnecessarily, we write throughout the proof $\mathbb{E}_y$ for $\mathbb{E}_{y\sim p_t}$. For now, we also write $\mathbb{E}_\theta$ as shorthand for $\mathbb{E}_{\theta\sim p(\cdot|y_{-i})}$. Recall that $x_+ = \max(x, 0) = ReLU(x)$.

Hence

$$\begin{aligned}
\mathbb{E}_y\left[\left(\log\frac{p(y_i|y_{-i})}{p(y_i|\theta_0)}\right)_+\right] &= \mathbb{E}_y\left[\left(\log\frac{\mathbb{E}_\theta[p(y_i|\theta)]}{p(y_i|\theta_0)}\right)_+\right] \\
&\leq \mathbb{E}_y\left[\log\left(1+\frac{\mathbb{E}_\theta\left[C(y_i)p(y_i|\theta_0)\|\theta-\theta_0\|\right]}{p(y_i|\theta_0)}\right)\right] \\
&\leq \mathbb{E}_{y,\theta}[C(y_i)\|\theta-\theta_0\|] \\
&\leq \left(\mathbb{E}_{y_i}[C(y_i)^2]\mathbb{E}_{y,\theta}\left[\|\theta-\theta_0\|^2\right]\right)^{1/2} \\
&\to 0 \text{ as } n\to\infty.
\end{aligned}$$

Here the first inequality follows from condition (i) and the second inequality from the fact that $\log(1+x) < x$ for $x \geq 0$. The third inequality is Schwarz inequality. The limit conclusion follows from the consistency of the posterior $p(\cdot|y_{-i})$ and the definition of weak convergence, since $\|\theta - \theta_0\|^2$ is a continuous bounded function of θ (recall that Θ is bounded) and that the first factor is finite by condition (i).

For the reverse inequality,

$$\begin{aligned}
\mathbb{E}_y\left[\left(\log\frac{p(y_i|\theta_0)}{p(y_i|y_{-i})}\right)_+\right] &= \mathbb{E}_y\left[\left(\log \mathbb{E}_\theta\left[\frac{p(y_i|\theta_0)]}{p(y_i|\theta)}\right]\right)_+\right] \\
&\leq \mathbb{E}_y\left[\log\left(1+\mathbb{E}_\theta\left[\frac{C(y_i)p(y_i|\theta)\|\theta-\theta_0\|}{p(y_i|\theta)}\right]\right)\right] \\
&\leq \left(\mathbb{E}_{y_i}[C(y_i)^2]\mathbb{E}_{y,\theta}\left[\|\theta-\theta_0\|^2\right]\right)^{1/2} \\
&\to 0 \text{ as } n \to \infty.
\end{aligned}$$

This proves (16) and an identical argument (now letting $\mathbb{E}_\theta$ stand for $\mathbb{E}_{\theta\sim p(\cdot|y)}$) proves (15) for $\tilde{\pi}_i = p(y_i|y)$.

For $\tilde{\pi}_i = -\mathbb{E}_y[\log p(y_i|y)]$, note first that

$$\begin{aligned}
\mathbb{E}_y\left|\mathbb{E}_y[\log p(y_i|y)] - \mathbb{E}_y[\log p(y_i|y_{-i})]\right| &= \left|\mathbb{E}_y[\log p(y_i|y) - \log p(y_i|y_{-i})]\right| \\
&\leq \mathbb{E}_y\left|\log p(y_i|y) - \log p(y_i|y_{-i})\right|
\end{aligned}$$

which goes to 0 by (16) and (a). Hence we can replace $\tilde{\pi}_i = -\mathbb{E}[\log p(y_i|y)]$ with $\tilde{\pi}_i = -\mathbb{E}[\log p(y_i|y_{-i})]$ when proving (b). To that end, observe that

$$\begin{aligned}
\left(\mathbb{E}_y[\log p(y_i|y_{-i})] - \log p(y_i|\theta_0)\right)_+ &= \left(\mathbb{E}_{y_i}\left[\mathbb{E}_{y_{-i}}\left[\log\frac{p(y_i|y_{-i})}{p(y_i|\theta_0)}\right]\right]\right)_+ \\
&\leq \mathbb{E}_y\left[\left(\log\frac{p(y_i|y_{-i})}{p(y_i|\theta_0)}\right)_+\right].
\end{aligned}$$

where the inequality is Jensen's inequality used twice on the convex function $x \to x_+$. Now everything is identical to the proof of (16) and the reverse inequality is analogous.

The other choices of $\tilde{\pi}_i$ follow along very similar lines. For $\tilde{\pi}_i = -\log p(y_i|\hat{\theta}_q)$, we have on mimicking the above that

$$\mathbb{E}_y\left[\left(\log\frac{p(y_i|\hat{\theta}_q)}{p(y_i|\theta_0)}\right)_+\right] \leq \left(\mathbb{E}_{y_i}[C(y_i)^2]\mathbb{E}_y\left[\|\hat{\theta}_q-\theta_0\|^2\right]\right)^{1/2}$$

and $\mathbb{E}_y[\|\hat{\theta}_q - \theta_0\|^2] \to 0$ as $n \to \infty$ by the assumed consistency of $\hat{\theta}_q$. The reverse inequality is analogous and (15) for $\pi_i = p(y_i|\hat{\theta}_q)$ is established.

For the case $\tilde{\pi}_i = -\log p(y_i|\mathbb{E}_{\theta\sim q}\theta)$, the analogous analysis gives

$$\mathbb{E}_y\left[\left(\log\frac{p(y_i|\mathbb{E}_{\theta\sim q}\theta)}{p(y_i|\theta_0)}\right)_+\right] \leq \mathbb{E}_{y_i}[C(y_i)^2]\mathbb{E}_y[\|\mathbb{E}_{\theta\sim q}\theta - \theta_0\|^2].$$

Since $x \to \|x-\theta_0\|^2$ is convex, the second factor on the right hand side is bounded by $\mathbb{E}_{y,\theta\sim q}[\|\theta-\theta_0\|^2]$ which goes to 0 by the consistency of q and the boundedness of Θ. The reverse inequality is again analogous.

For $\tilde{\pi}_i = -\mathbb{E}_{\theta\sim q}[\log p(y_i|\theta)]$,

$$
\begin{aligned}
\mathbb{E}_y\left[(\mathbb{E}_{\theta\sim q}[\log p(y_i|\theta)] - \log p(y_i|\theta_0))_+\right] &= \mathbb{E}_y\left[\left(\mathbb{E}_{\theta\sim q}\left[\log\frac{p(y_i|\theta)}{p(y_i|\theta_0)}\right]\right)_+\right] \\
&\le \mathbb{E}_{y,\theta\sim q}\left[\left(\log\frac{p(y_i|\theta)}{p(y_i|\theta_0)}\right)_+\right] \\
&\le \left(\mathbb{E}_{y_i}[C(y_i)^2]\mathbb{E}_{y,\theta\sim q}[\|\theta-\theta_0\|^2]\right)^{1/2} \to 0
\end{aligned}
$$

as $n\to\infty$ by the consistency of q. Here the first inequality is Jensen's inequality applied to $x \to x_+$ and the second inequality follows along the same lines as before.

To prove (f) it suffices by the triangle inequality to prove that $\mathbb{E}_y[V_{\theta\sim p(\cdot|y)}(\log p(y_i|\theta))] \to 0$ as $n\to\infty$. This follows from

$$
\begin{aligned}
\mathbb{E}_y\left[\mathbb{E}_{\theta\sim p(\cdot|y)}\left[(\log p(y_i|\theta) - \log p(y_i|\theta_0))_+^2\right]\right] &\le \mathbb{E}_{y,\theta}\left[\log\left(1+\frac{C(y_i)p(y_i|\theta)\|\theta-\theta_0\|}{p(y_i|\theta_0)}\right)^2\right] \\
&\le \mathbb{E}_{y,\theta}[2C(y_i)\|\theta-\theta_0\|] \\
&\le 2\mathbb{E}_{y,\theta}[C(y_i)^2]^{1/2}\mathbb{E}_{y,\theta}[\|\theta-\theta_0\|^2]^{1/2} \to 0.
\end{aligned}
$$

To prove that $\mathbb{E}_y[\mathbb{E}_{\theta\sim p(\cdot|y)}\left[(\log p(y_i|\theta_0) - \log p(y_i|\theta))_+^2\right] \to 0$ is analogous.

For (g) and (h) it suffices to observe that $\max_{i,j}|\mathbb{C}\text{ov}(\theta(i),\theta(j))| \to 0$. However

$$
\begin{aligned}
|\max_{i,j}\mathbb{C}\text{ov}(\theta(i),\theta(j))| &= \max_i V(\theta(j)) \\
&\le \max_i \mathbb{E}[|\theta(i)-\theta_0(i)|^2] \\
&\to 0
\end{aligned}
$$

where the final conclusion follows from the consistency of $\theta \sim p(\cdot|y)$ and the boundedness of Θ. Hence (g) and (h) are established. Similarly (g2) and (h2) follows from (g), (h) and (e).

For (k), write $r'(\theta_s) = r(\theta_s)/\sum_{j=1}^S r(\theta_j)$ for the random weights given to the individual θ_s:s in the expression for $\hat{p}(y_i|y_{-i})$. Then we have, with $\theta = (\theta_1,\ldots,\theta_S)$

chosen according to q,

$$
\begin{aligned}
\mathbb{E}_y\left[\left(\log\frac{\hat{p}(y_i|y_{-i})}{p(y_i|\theta_0)}\right)_+\right] &= \mathbb{E}_{y,\theta}\left[\left(\log\frac{\sum_{s=1}^S r'(\theta_s)p(y_i|\theta_s)}{p(y_i|\theta_0)}\right)_+\right] \\
&\leq \mathbb{E}_{y,\theta}\left[\log\left(1+\frac{\sum_{s=1}^S r'(\theta_s)|p(y_i|\theta_s)-p(y_i|\theta_0)|}{p(y_i|\theta_0)}\right)\right] \\
&\leq \mathbb{E}_{y,\theta}\left[\log\left(1+C(y_i)\sum_{s=1}^S r'(\theta_s)\|\theta_s-\theta_0\|\right)\right] \\
&\leq \mathbb{E}_{y,\theta}\left[\log\left(1+C(y_i)\sum_{s=1}^S \|\theta_s-\theta_0\|\right)\right] \\
&\leq \mathbb{E}_{y,\theta}\left[C(y_i)\sum_{s=1}^S \|\theta_s-\theta_0\|\right] \\
&\leq \left(\mathbb{E}_{y_i}[C(y_i)^2]\mathbb{E}_{y,\theta}\left[\left(\sum_{s=1}^S \|\theta_s-\theta_0\|\right)^2\right]\right)^{1/2},
\end{aligned}
$$

where the second inequality is condition (i) and the limit conclusion follows from the consistency of q. For the reverse inequality to go through analogously, observe that

$$
\begin{aligned}
\frac{|p(y_i|\theta_0)-\sum_s r'(\theta_s)p(y_i|\theta_s)|}{\sum_s r'(\theta_s)p(y_i|\theta_s)} &\leq \frac{\sum_s r'(\theta_s)|p(y_i|\theta_s)-p(y_i|\theta_0)|}{\sum_s r'(\theta_s)p(y_i|\theta_s)} \\
&\leq \frac{\sum_s r'(\theta_s)p(y_i|\theta_s)\|\theta_s-\theta_0\|}{\sum_s r'(\theta_s)p(y_i|\theta_s)} \\
&\leq \max_s \|\theta_s-\theta_0\| \\
&\leq \sum_s \|\theta_s-\theta_0\|.
\end{aligned}
$$

Equipped with this observation, mimic the above.

□

Reproducing results

The arsenic data

For the spline model comparison we use the `rstanarm` R package [Goodrich et al., 2018] with the following R script.

```
#' **Load data**
url <-
  "http://stat.columbia.edu/~gelman/arm/examples/arsenic/wells.dat"
wells <- read.table(url)
```

```r
wells$dist100 <- with(wells, dist / 100)
wells$y <- wells$switch

#' **Centering the input variables**
wells$c_dist100 <- wells$dist100 - mean(wells$dist100)
wells$c_arsenic <- wells$arsenic - mean(wells$arsenic)
wells$c_educ4 <- wells$educ/4 - mean(wells$educ/4)

#* **Latent linear model no interactions**
fit_1 <- stan_glm(y ~ c_dist100 + c_arsenic + c_educ4,
                  family = binomial(link="logit"),
                  data = wells,
                  iter = 1500,
                  warmup = 1000,
                  chains = 4)

#* **Latent linear model**
fit_2 <- stan_glm(y ~ c_dist100 + c_arsenic + c_educ4 +
                      c_dist100:c_educ4 + c_arsenic:c_educ4,
                  family = binomial(link="logit"),
                  data = wells,
                  iter = 1500,
                  warmup = 1000,
                  chains = 4)

#* **Latent GAM**
fit_3 <- stan_gamm4(y ~ s(dist100) + s(arsenic) + s(dist100, c_educ4),
                    family = binomial(link="logit"),
                    data = wells,
                    iter = 1500,
                    warmup = 1000,
                    chains = 4)
```

Generating data and fitting regularized horse-shoe and normal model

```r
library(arm)
library(rstanarm)

n <- 1e6

set.seed(1656)
x <- rnorm(n)
xn <- matrix(rnorm(n*99),nrow=n)
a <- 2
b <- 3
sigma <- 10
y <- a + b*x + sigma*rnorm(n)
fake <- data.frame(x, xn, y)

fit1 <- stan_glm(y ~ ., data=fake,
                 mean_PPD=FALSE,
                 refresh=0,
                 seed=SEED,
                 chains = 4,
```

```
                 warmup = 1000,
                 iter = 1500)

fit2 <- stan_glm(y ~ ., prior=hs(), data=fake,
                 mean_PPD=FALSE,
                 refresh=0,
                 seed=SEED,
                 chains = 4,
                 warmup = 1000,
                 iter = 1500)
```

Models

Stan Models

Bayesian linear regression (BLR)

```
data {
  int <lower=0> N;
  int <lower=0> D;
  matrix [N, D] X;
  vector [N] y;
}

parameters {
  vector [D] beta;
  real <lower=0> sigma;
}

model {
  // prior
  target += normal_lpdf(beta | 0, 10);
  target += normal_lpdf(sigma | 0, 1);
  // likelihood
  target += normal_lpdf(y | X * beta, sigma);
}
```

Pooled model (1)

```
data {
  int<lower=0> N;
  vector[N] floor_measure;
  vector[N] log_radon;
}

parameters {
  real alpha;
  real beta;
  real<lower=0> sigma_y;
}

model {
  vector[N] mu;

  // priors
  sigma_y ~ normal(0, 1);
```

```
  alpha ~ normal(0, 10);
  beta ~ normal(0, 10);

  // likelihood
  mu = alpha + beta * floor_measure;
  for(n in 1:N){
      target += normal_lpdf(log_radon[n]| mu[n], sigma_y);
  }
}
```

Partially pooled model (2)

```
data {
  int<lower=0> N;
  int<lower=0> J;
  int<lower=1,upper=J> county_idx[N];
  vector[N] log_radon;
}
parameters {
  vector[J] alpha_raw;
  real mu_alpha;
  real<lower=0> sigma_alpha;
  real<lower=0> sigma_y;
}

transformed parameters {
  vector[J] alpha;
  // implies: alpha ~ normal(mu_alpha, sigma_alpha);
  alpha = mu_alpha + sigma_alpha * alpha_raw;
}

model {
  vector[N] mu;

  // priors
  sigma_y ~ normal(0,1);
  sigma_alpha ~ normal(0,1);
  mu_alpha ~ normal(0,10);
  alpha_raw ~ normal(0, 1);

  // likelihood
  for(n in 1:N){
    mu[n] = alpha[county_idx[n]];
    target += normal_lpdf(log_radon[n] | mu[n], sigma_y);
  }
}
```

No pooled model (3)

```
data {
  int<lower=0> N;
  int<lower=0> J;
  int<lower=1,upper=J> county_idx[N];
  vector[N] floor_measure;
  vector[N] log_radon;
}
```

```
parameters {
  vector[J] alpha;
  real beta;
  real<lower=0> sigma_y;
}

model {
  vector[N] mu;
  // Prior
  sigma_y ~ normal(0, 1);
  alpha ~ normal(0, 10);
  beta ~ normal(0, 10);

  // Likelihood
  for(n in 1:N){
    mu[n] = alpha[county_idx[n]] + beta * floor_measure[n];
    target += normal_lpdf(log_radon[n] | mu[n], sigma_y);
  }
}
```

Variable intercept model (4)

```
data {
  int<lower=0> J;
  int<lower=0> N;
  int<lower=1,upper=J> county_idx[N];
  vector[N] floor_measure;
  vector[N] log_radon;
}
parameters {
  vector[J] alpha_raw;
  real beta;
  real mu_alpha;
  real<lower=0> sigma_alpha;
  real<lower=0> sigma_y;
}

transformed parameters {
  vector[J] alpha;
  // implies: alpha ~ normal(mu_alpha, sigma_alpha);
  alpha = mu_alpha + sigma_alpha * alpha_raw;
}

model {
  vector[N] mu;

  // Prior
  sigma_y ~ normal(0,1);
  sigma_alpha ~ normal(0,1);
  mu_alpha ~ normal(0,10);
  beta ~ normal(0,10);
  alpha_raw ~ normal(0, 1);

  for(n in 1:N){
    mu[n] = alpha[county_idx[n]] + floor_measure[n]*beta;
    target += normal_lpdf(log_radon[n]|mu[n],sigma_y);
  }
}
```

Variable slope model (5)

```
data {
  int<lower=0> J;
  int<lower=0> N;
  int<lower=1,upper=J> county_idx[N];
  vector[N] floor_measure;
  vector[N] log_radon;
}

parameters {
  real alpha;
  vector[J] beta_raw;
  real mu_beta;
  real<lower=0> sigma_beta;
  real<lower=0> sigma_y;
}

transformed parameters {
  vector[J] beta;
  // implies: beta ~ normal(mu_beta, sigma_beta);
  beta = mu_beta + sigma_beta * beta_raw;
}

model {
  vector[N] mu;
  // Prior
  alpha ~ normal(0,10);
  sigma_y ~ normal(0,1);
  sigma_beta ~ normal(0,1);
  mu_beta ~ normal(0,10);
  beta_raw ~ normal(0, 1);

  for(n in 1:N){
    mu[n] = alpha + floor_measure[n] * beta[county_idx[n]];
    target += normal_lpdf(log_radon[n]|mu[n],sigma_y);
  }
}
```

Variable intercept and slope model (6)

```
data {
  int<lower=0> N;
  int<lower=0> J;
  int<lower=1,upper=J> county_idx[N];
  vector[N] floor_measure;
  vector[N] log_radon;
}

parameters {
  real<lower=0> sigma_y;
  real<lower=0> sigma_alpha;
  real<lower=0> sigma_beta;
  vector[J] alpha_raw;
  vector[J] beta_raw;
  real mu_alpha;
  real mu_beta;
```

```
}
transformed parameters {
  vector[J] alpha;
  vector[J] beta;
  // implies: alpha ~ normal(mu_alpha, sigma_alpha);
  alpha = mu_alpha + sigma_alpha * alpha_raw;
  // implies: beta ~ normal(mu_beta, sigma_beta);
  beta = mu_beta + sigma_beta * beta_raw;
}

model {
  vector[N] mu;
  // Prior
  sigma_y ~ normal(0,1);
  sigma_beta ~ normal(0,1);
  sigma_alpha ~ normal(0,1);
  mu_alpha ~ normal(0,10);
  mu_beta ~ normal(0,10);
  alpha_raw ~ normal(0, 1);
  beta_raw ~ normal(0, 1);

  // Likelihood
  for(n in 1:N){
    mu[n] = alpha[county_idx[n]] + floor_measure[n] * beta[county_idx[n]];
    target += normal_lpdf(log_radon[n] | mu[n], sigma_y);
  }
}
```

References

Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2018. URL `http://mc-stan.org/`. R package version 2.17.4.

Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

Måns Magnusson, Michael Andersen, Johan Jonasson, and Aki Vehtari. Bayesian leave-one-out cross-validation for large data. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4244–4253. PMLR, 2019.

Andrew M Walker. On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 80–88, 1969.

Yixin Wang and David M Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.