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

Johan Jonasson [1] **Måns Magnusson** [2] **Michael Riis Andersen** [2] **Aki Vehtari** [2]

1. Proof of Proposition 1

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 $\mathbb{R}^d$.

A number of 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.
- We can (and will) without loss of generality assume that $M = 1/2$ is sufficient in (iii), for if not then simply transform data and consider $z_i = 2My_i$ instead of y_i.

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

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

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 $\overline{elpd}_{loo}$ is given by*

$$\overline{elpd}_{loo} = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|y_{-i}), \tag{1}$$

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

To estimate $\overline{\text{elpd}}_{loo}$ in turn, we use importance sampling and the Hansen-Hurwitz estimator. Definitions follow.

Definition 3. *The Hansen-Hurwitz estimator is given by*

$$\widehat{\overline{elpd}}_{loo}(m, q) = \frac{1}{m}\frac{1}{n}\sum_{j=1}^{m} \frac{1}{\tilde{\pi}_j} \log \hat{p}(y_j|y_{-j})$$

*Equal contribution [1]Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Sweden [2]Department of Computer Science, Aalto University, Finland. Correspondence to: Måns Magnusson <mans.magnusson@aalto.fi>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

where $\tilde{\pi}_i$ is the probability of subsampling observation i, $\log \hat{p}(y_i|y_{-i})$ is the (self-normalized) importance sampling estimate of $\log p(y_i|y_{-i})$ defined as

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

where

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

and where $q(\theta|y)$ is an approximation of the posterior distribution, θ_s is a sample from the approximate posterior distribution $q(\theta|y)$ and S is the total posterior sample size.

Proposition 1. *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*

$$|\widehat{\overline{elpd}}_{loo}(m, q) - \overline{elpd}_{loo}| \to 0$$

in probability as $n \to \infty$ for any of the following choices of π_i, $i = 1, \ldots, n$.

(a) $\pi_i = -\log p(y_i|y)$,

(b) $\pi_i = -\mathbb{E}_y[\log p(y_i|y)]$,

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

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

(e) $\pi_i = -\log p(y_i|\hat{\theta}_q)$.

Remark. By the variational BvM Theorems of Wang and Blei, (Wang & Blei, 2018), 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 1, provided that one adopts the mild conditions in their paper.

The proof of Proposition 1 will be focused on proving (a) and then (b)-(e) will follow easily. We begin with the following key lemma.

Lemma 2. *With all quantities as defined above,*

$$\mathbb{E}_{y\sim p_t} |\pi_i - \log p(y_i|\theta_0)| \to 0, \tag{2}$$

with any of the definitions (a)-(e) of π_i of Proposition 1. Furthermore,

$$\mathbb{E}_{y\sim p_t} |\log p(y_i|y_{-i}) - \log p(y_i|\theta_0)| \to 0, \tag{3}$$

and

$$\mathbb{E}_{y\sim p_t} |\log \hat{p}(y_i|y) - \log p(y_i|\theta_0)| \to 0. \tag{4}$$

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 (3) and an identical argument proves (2) for $\pi_i = p(y_i|y)$.

For $\pi_i = -\mathbb{E}_y[\log p(y_i|y)]$, 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 (3) and (a). Hence we can replace $\pi_i = -\mathbb{E}[\log p(y_i|y)]$ with $\pi_i = -\mathbb{E}[\log p(y_i|y_{-i})]$ 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 (3) and the reverse inequality is analogous.

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

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

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 (2) for $\pi_i = p(y_i|\hat{\theta}_q)$ is established.

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

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

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.

Finally for $\pi_i = -\mathbb{E}_{\theta\sim q}[\log p(y_i|\theta)]$,

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

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.

For (4), 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.

□

For convenience we will write $\hat{e} := \hat{e}_{m,q} = \widehat{\text{elpd}}_{loo}$, which for our purposes is more usefully expressed as

$$\hat{e} = \frac{1}{n}\frac{1}{m}\sum_{i=1}^n\sum_{j=1}^m I_{ij}\frac{1}{\bar{\pi}_i}\log\hat{p}(y_i|y_{-i}),$$

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

$$e = \frac{1}{n}\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m} I_{ij}\frac{1}{\bar{\pi}_i}\log p(y_i|y_{-i}).$$

In other words, e is the HH estimator with $\hat{p}$ replaced with p.

Lemma 3. *With the notation as just defined and* $\pi_i = -\log p(y_i|y)$,

$$\mathbb{E}|\hat{e} - e| \to 0$$

as $n \to \infty$.

Proof. We have, with expectations with respect to all sources of randomness involved in $\hat{e}$ and e

$$\begin{aligned}
&\mathbb{E}|\hat{e} - e| \\
&\leq \frac{1}{m}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}\left[\mathbb{E}\left[I_{ij}\frac{1}{\bar{\pi}_i}|\log\hat{p}(y_i|y_{-i}) - \log p(y_i|y_{-i})|\Big|y\right]\right] \\
&= \mathbb{E}\left[\frac{1}{n}\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}|\log\hat{p}(y_i|y_{-i}) - \log p(y_i|y_{-i})|\right] \\
&= \mathbb{E}|\log\hat{p}(y_i|y_{-i}) - \log p(y_i|y_{-i})|.
\end{aligned}$$

The result now follows from (3), (4) and the triangle inequality. □

Proof of Proposition 1. As stated before, we start with a focus on (a), which means that for now we have $\pi_i = -\log p(y_i|y)$ By Lemma 3, it suffices to prove that $|e - \overline{\text{elpd}}_{loo}| \to 0$ in probability with π_i chosen according to any of (a)-(e). The variance of a HH estimator is well known and some easy manipulation then tells us that the conditional variance of e given y is given by

$$V(e) = \mathbb{V}\text{ar}(e|y) = \frac{1}{n^2}\frac{1}{m}(S_\pi S_2 - S_p^2),$$

where $S_p = \sum_{i=1}^n p_i$, $S_\pi = \sum_{i=1}^n \pi_i$ and $S_2 = \sum_{i=1}^n (p_i^2/\pi_i)$. We claim that for any $\delta > 0$, for n sufficiently large, $\mathbb{P}_y(V(e) < \delta) > 1 - \delta$. To this end, observe first that

$$\begin{aligned}
&\mathbb{E}_y[-\log p(y_i|y)] \\
&\leq \mathbb{E}_y[-\log p(y_i|\theta_0)] + \mathbb{E}_y|\log p(y_i|y) - \log p(y_i|\theta_0)| \\
&\leq \mathbb{E}_y[-\log p(y_i|\theta_0)] + \delta < \infty
\end{aligned}$$

for sufficiently large n, since the first term is finite by condition (v). Let $A = A_n = \mathbb{E}_y[-\log p(y_i|y)]$.

Now,

$$\mathbb{E}_y\left[\frac{1}{n}|S_p - S_\pi|\right] = \mathbb{E}_y\left[\frac{1}{n}\left|\sum_{i=1}^n \pi_i - \sum_{i=1}^n p_i\right|\right] \to 0$$

as $n \to \infty$ by (2) and (3). Hence for arbitrary $\alpha > 0$, $\mathbb{P}_y(|S_p - S_\pi| < \alpha^2 n) > 1 - \alpha$ for n large enough. Also

$$\frac{p_i^2}{\pi_i} \leq \frac{(\pi_i + |p_i - \pi_i|)^2}{\pi_i} < \pi_i + 4|\pi_i - p_i|$$

(the last inequality using condition (iii): $\pi_i \geq -\log(1/2) > 1/2$), so $n^{-1}\mathbb{E}_y|S_\pi - S_2| \to 0$ and so $\mathbb{P}_y(|S_p - S_2| < \alpha^2 n) > 1 - \alpha$ for sufficiently large n. Hence with probability exceeding $1 - 2\alpha$, y will be such that for sufficiently large n,

$$\begin{aligned}
V(e) &\leq \frac{1}{n^2}\frac{1}{m}\left((S_p + \alpha^2 n)^2 - S_p^2\right) \\
&= \frac{1}{n^2}\frac{1}{m}(2\alpha^2 n S_p + \alpha^4 n^2).
\end{aligned}$$

We had $\mathbb{E}_y[S_p] = An$ and Markov's inequality thus entails that $\mathbb{P}_y(S_p < An/\alpha) > 1 - \alpha$. Adding this piece of information to the above, we get that with probability larger than $1 - 3\alpha$, y will for sufficiently large n be such that

$$V(e) \leq (2\alpha + \alpha^4)n^2 < 3\alpha.$$

For such y, Chebyshev's inequality gives

$$\mathbb{P}(|e - \mathbb{E}[e|y]| > \alpha^{1/2}|y) < 3\alpha^{1/2}.$$

The HH estimator is unbiased, so $\mathbb{E}[e|y] = \overline{\text{elpd}}_{loo}$. We get for arbitrary $\epsilon > 0$ on taking α sufficiently small and n correspondingly large, taking all randomness into account

$$\mathbb{P}(|e - \overline{\text{elpd}}_{loo}| > \epsilon) < 1 - \epsilon$$

which entails that $|e - \overline{\text{elpd}}_{loo}| \to 0$ in probability. As observed above, this proves (a).

For the remaining parts, write e_p when taking π_i in e according to statement (p) in the proposition. By (2), $\mathbb{E}|e_p - e_a| \to 0$ for $p = b, c, d, e$ and we are done.

□

References

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

Wang, Y. and Blei, D. M. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, (just-accepted):1–85, 2018.

2. Unbiasness of Using the Hansen-Hurwitz Estimator

2.1. On the Hansen-Hurwitz Estimator

Let $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$ be a set of non-negative observations, $y_i > 0$ and let $\pi = \{\pi_1, \pi_2, \ldots, \pi_N\}$ be a probability vector s.t. $\sum \pi_j = 1$. Furthermore, let $a_k \in \{1, 2, \ldots, N\}$ be i.i.d. samples from a multinomial distribution with probabilities π, i.e. $a_k \overset{iid}{\sim} \text{Multinomial}(\pi)$.

We want to estimate the total

$$\tau = \sum_{n=1}^{N} y_i \tag{5}$$

using the Hansen-Hurwitz estimator given by

$$\hat{\tau} = \frac{1}{M} \sum_{m=1}^{M} \frac{x_m}{p_m}, \tag{6}$$

where $x_m \equiv y_{a_m}$, $p_m \equiv \pi_{a_m}$, and $a_m \sim \text{Multinomial}(\pi)$.

We can decompose x_m and p_m as follows

$$x_m \equiv y_{a_m} = \sum_{j=1}^{N} \mathbb{I}[a_m = j]\, y_j \tag{7}$$

$$p_m \equiv p_{a_m} = \sum_{j=1}^{N} \mathbb{I}[a_m = j]\, \pi_j \tag{8}$$

2.2. The Hansen-Hurwitz Estimator is Unbiased

First, we will show that the HH estimator, $\hat{\tau}$, is unbiased. We have,

$$\mathbb{E}[\hat{\tau}] = \mathbb{E}\left[\frac{1}{M} \sum_{m=1}^{M} \frac{x_m}{p_m}\right] = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\left[\frac{x_m}{p_m}\right] \tag{9}$$

Using the definitions in eq. (7) and (8) yields

$$\begin{aligned}
\mathbb{E}[\hat{\tau}] &= \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\left[\frac{\sum_{j=1}^{N} \mathbb{I}[a_m = j]\, y_j}{\sum_{j=1}^{N} \mathbb{I}[a_m = j]\, \pi_j}\right] \\
&= \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\left[\sum_{j=1}^{N} \frac{y_j}{\pi_j} \mathbb{I}[a_m = j]\right] \\
&= \frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{N} \frac{y_j}{\pi_j} \mathbb{E}[\mathbb{I}[a_m = j]] \\
&= \frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{N} \frac{y_j}{\pi_j} \pi_j
\end{aligned} \tag{10}$$

since $\pi_j = \mathbb{P}[a_m = j] = \mathbb{E}[\mathbb{I}[a_m = j]]$.

Now it follows that

$$\mathbb{E}[\hat{\tau}] = \frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{N} y_j = \sum_{j=1}^{N} y_j = \tau. \tag{11}$$

2.3. An Unbiased Estimator of σ^2_{loo}

We also want to estimate the variance of the population $\mathcal{Y}$, i.e.

$$\sigma_y^2 = \frac{1}{N} \sum_{n=1}^{N} (y_n - \bar{y})^2, \tag{12}$$

where $\bar{y} = \frac{1}{N} \sum y_n$.

First, we decompose the above as follows

$$\sigma_y^2 = \frac{1}{N} \sum_{n=1}^{N} y_n^2 - \bar{y}^2. \tag{13}$$

We will consider estimators for the two terms, $\frac{1}{N} \sum_{n=1}^{N} y_n^2$ (1) and $\bar{y}^2$ (2), separately. First, we will show that the following is an unbiased estimate of the first term,

$$T_1 = \frac{1}{NM} \sum_{m=1}^{M} \frac{x_m^2}{p_m}. \tag{14}$$

We have

$$\mathbb{E}[T_1] = \mathbb{E}\left[\frac{1}{NM} \sum_{m=1}^{M} \frac{x_m^2}{p_m}\right] = \frac{1}{NM} \sum_{m=1}^{M} \mathbb{E}\left[\frac{x_m^2}{p_m}\right] \tag{15}$$

Again, we use the representations in eq. (7) and (8) to get

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{NM} \sum_{m=1}^{M} \frac{x_m^2}{p_m}\right] &= \frac{1}{NM} \sum_{m=1}^{M} \mathbb{E}\left[\frac{\sum_{j=1}^{N} \mathbb{I}[a_m = j]\, y_j^2}{\sum_{j=1}^{N} \mathbb{I}[a_m = j]\, \pi_j}\right] \\
&= \frac{1}{NM} \sum_{m=1}^{M} \mathbb{E}\left[\sum_{j=1}^{N} \mathbb{I}[a_m = j]\, \frac{y_j^2}{\pi_j}\right] \\
&= \frac{1}{NM} \sum_{m=1}^{M} \sum_{j=1}^{N} \frac{y_j^2}{\pi_j} \mathbb{E}[\mathbb{I}[a_m = j]] \\
&= \frac{1}{NM} \sum_{m=1}^{M} \sum_{j=1}^{N} \frac{y_j^2}{\pi_j} \pi_j \\
&= \frac{1}{N} \sum_{j=1}^{N} y_j^2.
\end{aligned} \tag{16}$$

This completes the proof of for the first term.

For the second term, we use the estimator T_2 given by

$$T_2 = \frac{1}{M(M-1)} \sum_{m=1}^{M} \left[\frac{x_m}{Np_m} - \frac{1}{N} \sum_{k=1}^{M} \frac{x_k}{Mp_k} \right]^2 - \left[\frac{1}{N} \sum_{k=1}^{M} \frac{x_k}{Mp_k} \right]^2. \tag{17}$$

We have

$$\frac{1}{M(M-1)} \sum_{m=1}^{M} \left[\frac{x_m}{Np_m} - \sum_{k=1}^{M} \frac{x_k}{NMp_k} \right]^2 - \left[\sum_{k=1}^{M} \frac{x_k}{NMp_k} \right]^2 = \frac{1}{N^2M(M-1)} \sum_{m=1}^{M} \frac{x_m^2}{p_m^2} - \frac{1}{N^2M(M-1)} \left[\sum_{k=1}^{M} \frac{x_k}{p_k} \right]^2 \tag{18}$$

We consider now the expectation of the first term in the equation above

$$\begin{aligned}
\mathbb{E}\left[\sum_{m=1}^{M} \frac{x_m^2}{p_m^2} \right] &= \sum_{m=1}^{M} \mathbb{E}\left[\frac{x_m^2}{p_m^2} \right] \\
&= \sum_{m=1}^{M} \mathbb{E}\left[\frac{\sum_{j=1}^{N} \mathbb{I}\left[a_m = j\right] y_j^2}{\sum_{j=1}^{N} \mathbb{I}\left[a_m = j\right] \pi_j^2} \right] \\
&= \sum_{m=1}^{M} \mathbb{E}\left[\sum_{j=1}^{N} \mathbb{I}\left[a_m = j\right] \frac{y_j^2}{\pi_j^2} \right] \\
&= \sum_{m=1}^{M} \sum_{j=1}^{N} \mathbb{E}\left[\mathbb{I}\left[a_m = j\right] \right] \frac{y_j^2}{\pi_j^2} \\
&= M \sum_{j=1}^{N} \frac{y_j^2}{\pi_j}
\end{aligned} \tag{19}$$

and the second term

$$\begin{aligned}
\mathbb{E}\left[\left[\sum_{k=1}^{M} \frac{x_k}{p_k} \right]^2 \right] &= \mathbb{E}\left[\sum_{k=1}^{M} \sum_{j=1}^{M} \frac{x_k}{p_k} \frac{x_j}{p_j} \right] \\
&= \sum_{k=1}^{M} \sum_{j=1}^{M} \mathbb{E}\left[\frac{x_k}{p_k} \frac{x_j}{p_j} \right] \\
&= \sum_{j \neq k}^{M} \mathbb{E}\left[\frac{x_k}{p_k} \frac{x_j}{p_j} \right] + \sum_{k=1}^{M} \mathbb{E}\left[\frac{x_k^2}{p_k^2} \right] \\
&= \sum_{j \neq k}^{M} \mathbb{E}\left[\frac{x_k}{p_k} \right] \mathbb{E}\left[\frac{x_j}{p_j} \right] + \sum_{k=1}^{M} \sum_{j=1}^{N} \frac{y_j^2}{\pi_j} \\
&= \sum_{j \neq k}^{M} \mathbb{E}\left[\frac{x_k}{p_k} \right] \mathbb{E}\left[\frac{x_j}{p_j} \right] + M \sum_{j=1}^{N} \frac{y_j^2}{\pi_j} \\
&= M(M-1)\tau^2 + M \sum_{j=1}^{N} \frac{y_j^2}{\pi_j}.
\end{aligned} \tag{20}$$

Substituting back, we get

$$\begin{aligned}
&\frac{1}{M(M-1)} \sum_{m=1}^{M} \left[\frac{x_m}{Np_m} - \frac{1}{N} \sum_{k=1}^{M} \frac{x_k}{Mp_k} \right]^2 - \left[\frac{1}{N} \sum_{k=1}^{M} \frac{x_k}{Mp_k} \right]^2 \\
&= \frac{1}{N^2M(M-1)} M \sum_{j=1}^{N} \frac{y_j^2}{\pi_j} - \\
&\quad \frac{1}{N^2M(M-1)} \left[M(M-1)\tau^2 + M \sum_{j=1}^{N} \frac{y_j^2}{\pi_j} \right] \\
&= \frac{1}{N^2(M-1)} \sum_{j=1}^{N} \frac{y_j^2}{\pi_j} - \\
&\quad \frac{1}{N^2(M-1)} \left[(M-1)\tau^2 + \sum_{j=1}^{N} \frac{y_j^2}{\pi_j} \right] \\
&= - \frac{1}{N^2(M-1)} (M-1)\tau^2 \\
&= - \frac{\tau^2}{N^2} \\
&= - \bar{y}^2.
\end{aligned} \tag{21}$$

Combining the two estimators T_1 and T_2 we have:

$$\begin{aligned}
\mathbb{E}(T_1 + T_2) &= \frac{1}{N} \sum_{j=1}^{N} y_j^2 - \bar{y}^2 \\
&= \sigma_y^2
\end{aligned}$$

Hence, we have shown that the estimator of σ_y^2 is unbiased using the sum of the estimators T_1 in Eq. 14 and T_2 in Eq. 18.

3. Hierarchical Models for the Radon Dataset

We compare seven different models of predicting the radon levels in individual houses (indexed by i) by county (indexed by j). First we fit a pooled model (model 1)

$$\begin{aligned} y_{ij} &= \alpha + x_{ij}\beta + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma_y) \\ \alpha, \beta &\sim N(0, 10) \\ \sigma_y &\sim N^+(0, 1)\,, \end{aligned}$$

where y_{ij} is the log radon level in house i in county j, x_{ij} is the floor measurement and ϵ_{ij} is $N^+(0, 1)$ is a truncated Normal distribution at the positive real line. We compare this to a non-pooled model (model 2),

$$\begin{aligned} y_{ij} &= \alpha_j + x_{ij}\beta + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma_y) \\ \alpha_j, \beta &\sim N(0, 10) \\ \sigma_y &\sim N^+(0, 1)\,, \end{aligned}$$

a partially pooled model (model 3),

$$\begin{aligned} y_{ij} &= \alpha_j + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma_y) \\ \alpha_j &\sim N(\mu_\alpha, \sigma_\alpha) \\ \mu_\alpha &\sim N(0, 10) \\ \sigma_y, \sigma_\alpha &\sim N^+(0, 1)\,, \end{aligned}$$

a variable intercept model (model 4),

$$\begin{aligned} y_{ij} &= \alpha_j + x_{ij}\beta + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma_y) \\ \alpha_j &\sim N(\mu_\alpha, \sigma_\alpha) \\ \mu_\alpha, \beta &\sim N(0, 10) \\ \sigma_y, \sigma_\alpha &\sim N^+(0, 1)\,, \end{aligned}$$

a variable slope model (model 5),

$$\begin{aligned} y_{ij} &= \alpha + x_{ij}\beta_j + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma_y) \\ \beta_j &\sim N(\mu_\beta, \sigma_\beta) \\ \mu_\beta, \alpha &\sim N(0, 10) \\ \sigma_y, \sigma_\beta &\sim N^+(0, 1)\,, \end{aligned}$$

a variable intercept and slope model (model 6),

$$\begin{aligned} y_{ij} &= \alpha_j + x_{ij}\beta_j + \epsilon_{ij} \\ \alpha_j &\sim N(\mu_\alpha, \sigma_\alpha) \\ \beta_j &\sim N(\mu_\beta, \sigma_\beta) \\ \mu_\alpha, \mu_\beta &\sim N(0, 10) \\ \sigma_y, \sigma_\alpha, \sigma_\beta &\sim N^+(0, 1)\,, \end{aligned}$$

and finally a model with county level covariates and county level intercepts

$$\begin{aligned} y_{ij} &= \alpha_j + x_{ij}\beta_1 + u_j\beta_2 + \epsilon_{ij} \\ \alpha_j &\sim N(\mu_\alpha, \sigma_\alpha) \\ \beta, \mu_\alpha &\sim N(0, 10) \\ \sigma_y, \sigma_\alpha &\sim N^+(0, 1)\,, \end{aligned}$$

where u_j is the log uranium level in the county. The Stan code used can be found below.

4. Stan models

4.1. Linear Regression Model

```
data {
  int <lower=0> N;
  int <lower=0> D;
  matrix [N,D] x ;
  vector [N] y;
}
parameters {
  vector [D] b;
  real <lower=0> sigma;
}
model {
  target += normal_lpdf(y | x * b, sigma);
  target += normal_lpdf(b | 0, 1);
}


generated quantities{
  real  log_joint_density_unconstrained;
  vector[N] log_lik;
  // Compute the log likelihoods for loo
  for (n in 1:N) {
    log_lik[n] =
      normal_lpdf(y[n] | x[n,] * b, sigma);
  }
}
```

4.2. Radon pooled model (1)

```
data {
  int<lower=0> N;
  vector[N] x;
  vector[N] y;
  int<lower=0,upper=1> holdout[N];
}

parameters {
  vector[2] beta;
  real<lower=0> sigma_y;
}

model {
  vector[N] mu;

  // priors
  sigma_y ~ normal(0,1);
  beta ~ normal(0,10);

  // likelihood
  mu = beta[1] + beta[2] * x;
  for(n in 1:N){
    if(holdout[n] == 0){
      target +=
        normal_lpdf(y[n]|mu[n],sigma_y);
    }
  }
}
```

4.3. Radon pooled model (2)

```
data {
  int<lower=0> N;
  int<lower=0> J;
  int<lower=1,upper=J> county[N];
  vector[N] x;
  vector[N] y;
  int<lower=0,upper=1> holdout[N];
}

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

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

  // Likelihood
  for(n in 1:N){
    mu[n] = beta*x[n] + a[county[n]];
    if(holdout[n] == 0){
      target +=
        normal_lpdf(y[n]|mu[n],sigma_y);
    }
  }
}
```

4.4. Radon partially pooled model (3)

```
data {
  int<lower=0> N;
  int<lower=0> J;
  int<lower=1,upper=J> county[N];
  vector[N] y;
  int<lower=0,upper=1> holdout[N];
}
parameters {
  vector[J] a;
  real mu_a;
  real<lower=0> sigma_a;
  real<lower=0> sigma_y;
}

model {
  vector[N] mu;

  // priors
  sigma_y ~ normal(0,1);
  sigma_a ~ normal(0,1);
  mu_a ~ normal(0,10);

  // likelihood
  a ~ normal (mu_a, sigma_a);
  for(n in 1:N){
    mu[n] = a[county[n]];
    if(holdout[n] == 0){
      target +=
        normal_lpdf(y[n]|mu[n],sigma_y);
    }
  }
}
```

4.5. Variable intercept model (4)

```
data {
  int<lower=0> J;
  int<lower=0> N;
  int<lower=1,upper=J> county[N];
  vector[N] x;
  vector[N] y;
  int<lower=0,upper=1> holdout[N];
}
parameters {
  vector[J] a;
  real beta;
  real mu_a;
  real<lower=0> sigma_a;
  real<lower=0> sigma_y;
}

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

  a ~ normal (mu_a, sigma_a);
  for(n in 1:N){
    mu[n] = a[county[n]] + x[n]*beta;
    if(holdout[n] == 0){
      target +=
        normal_lpdf(y[n]|mu[n],sigma_y);
    }
  }
}
```

4.6. Variable slope model (5)

```
data {
  int<lower=0> J;
  int<lower=0> N;
  int<lower=1,upper=J> county[N];
  vector[N] x;
  vector[N] y;
  int<lower=0,upper=1> holdout[N];
}
parameters {
  real a;
  vector[J] beta;
  real mu_beta;
  real<lower=0> sigma_beta;
  real<lower=0> sigma_y;
}

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

  beta ~ normal(mu_beta,sigma_beta);
  for(n in 1:N){
    mu[n] = a + x[n] * beta[county[n]];
    if(holdout[n] == 0){
      target +=
        normal_lpdf(y[n]|mu[n],sigma_y);
    }
  }
}
```

4.7. Variable intercept and slope model (6)

```
data {
  int<lower=0> N;
  int<lower=0> J;
  vector[N] y;
  vector[N] x;
  int county[N];
  int<lower=0,upper=1> holdout[N];
}
parameters {
  real<lower=0> sigma_y;
  real<lower=0> sigma_a;
  real<lower=0> sigma_beta;
  vector[J] a;
  vector[J] beta;
  real mu_a;
  real mu_beta;
}

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

  a ~ normal(mu_a, sigma_a);
  beta ~ normal(mu_beta, sigma_beta);
  for(n in 1:N){
    mu[n] = a[county[n]] + x[n]*beta[county[n]];
    if(holdout[n] == 0){
      target +=
        normal_lpdf(y[n]|mu[n],sigma_y);
    }
  }

}
```

4.8. Hierarchical intercept model (7)

```
data {
  int<lower=0> J;
  int<lower=0> N;
  int<lower=1,upper=J> county[N];
  vector[N] u;
  vector[N] x;
  vector[N] y;
  int<lower=0,upper=1> holdout[N];
}
parameters {
  vector[J] a;
  vector[2] beta;
  real mu_a;
  real<lower=0> sigma_a;
  real<lower=0> sigma_y;
}
transformed parameters {
}

model {
  vector[N] mu;
  vector[N] m;

  sigma_a ~ normal(0, 1);
  sigma_y ~ normal(0, 1);
  mu_a ~ normal(0, 10);
  beta ~ normal(0, 10);

  a ~ normal(mu_a, sigma_a);
  for(n in 1:N){
    m[n] = a[county[n]] + u[n] * beta[1];
    mu[n] = m[n] + x[n] * beta[2];
    if(holdout[n] == 0){
      target += normal_lpdf(y[n] | mu[n], sigma_y);
    }
  }
}
```

5. R package

The functions are implemented based upon the `loo` package structure as the functions `quick_loo()`, `approx_psis()` and `psis_approximate_posterior()`. An example how to run the code can be found in the documentation for `quick_loo()`. No changes to author lists, versions or date has been changed to preserve anonymity. If accepted, the code will be published open source.