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

Måns Magnusson Aalto University
Michael Riis Andersen Technical University of Denmark
Johan Jonasson Chalmers University of Technology and University of Gothenburg
Aki Vehtari Aalto University

Abstract

Recently, new methods for model assessment, based on subsampling and posterior approximations, have been proposed for scaling leave-one-out cross-validation (LOO) to large datasets. Although these methods work well for estimating predictive performance for individual models, they are less powerful in model comparison. We propose an efficient method for estimating differences in predictive performance by combining fast approximate LOO surrogates with exact LOO subsampling using the difference estimator and supply proofs with regards to scaling characteristics. The resulting approach can be orders of magnitude more efficient than previous approaches, as well as being better suited to model comparison.

1 INTRODUCTION

Model comparison is an important part of probabilistic machine learning. In many real-world domains, we are often confronted with multiple models and would like to choose the model that best generalizes to new, unseen data. This can be done in a large number of ways, but here we will restrict ourselves to choosing between models based on the *predictive* performance. Due to the growing data sizes over the last years, scaling model comparison methods to large data is an important problem.

One measure of predictive performance is the *expected log predictive density* (elpd) given by

$$\begin{aligned}\overline{\text{elpd}}_M &= \int \log p_M(\tilde{y}_i|y) p_t(\tilde{y}_i) d\tilde{y}_i \\ &= \int \log \left[\int p_M(\tilde{y}_i|\theta) p_M(\theta|y) d\theta \right] p_t(\tilde{y}_i) d\tilde{y}_i,\end{aligned} \tag{1}$$

where $\log p_M(\tilde{y}_i|y)$ is the log predictive density of model M for a new observation $\tilde{y}_i$, that has been generated by some true, unknown process, $p_t(\tilde{y}_i)$. The log predictive density, or the log score, has good theoretical properties in that it is both *local*, i.e., only depend on $\tilde{y}_i$, and *proper*, the expected reward is maximized by the true probability distribution (Bernardo, 1979; Bernardo and Smith, 1994; Gneiting et al., 2007; Vehtari and Ojanen, 2012). Although we focus on the log score in this paper, other scoring functions can be used.

1.1 Leave-one-out cross-validation

Leave-one-out cross-validation (LOO) is a method for estimating the elpd, or the generalization performance, of a model (Bernardo and Smith, 1994; Vehtari and Ojanen, 2012; Vehtari et al., 2017). This is done by training the model on all observations except observation y_i, and then predicting the hold-out observation y_i, something that is then repeated for all n observations. In this way we treat each observation y_i as a pseudo-Monte-Carlo sample from the true generating model p_t. We hence compute n leave-one-out (LOO) posterior distributions $p(\theta|y_{-i})$, where y_{-i} denotes the data with observation y_i removed. Using the LOO posteriors, we can estimate the elpd in Eq. (1) as

$$\begin{aligned}\overline{\text{elpd}}_{\text{loo}} &= \frac{1}{n} \sum_{i=1}^{n} \log p_M(y_i|y_{-i}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \log \int p_M(y_i|\theta) p_M(\theta|y_{-i}) d\theta \\ &= \frac{1}{n} \text{elpd}_{\text{loo}} \,,\end{aligned} \tag{2}$$

where $p_M(y_i|\theta)$ is the likelihood, and $p_M(\theta|y_{-i})$ is the posterior for θ where we hold out observation y_i.

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

Although the many good properties of LOO, scaling the approach to large data is a problem. The naive approach to LOO means that n posteriors need to be computed. In situations with large n, the cost of just computing one posterior may be large, hence leading to poor scaling.

1.2 Approximating LOO

A number of approximate techniques have recently been proposed to approximate exact LOO. Wang et al. (2018) and Giordano et al. (2019) propose LOO-approximations with very appealing error bounds for M-estimators. The main idea is to fit a model on the complete data set and then extrapolate to capture the effect of holding out individual observations using a second-order Taylor approximation. For some special classes of models, such as Gaussian processes, specialized LOO approximations have been proposed (Held et al., 2010; Vehtari et al., 2016).

In the Bayesian domain, similar ideas was introduced by Gelfand (1996) using self-normalized importance sampling (IS). The idea is to use the full posterior distribution as the proposal distribution in an importance sampling scheme with the LOO posterior as the target distribution. In this way, we only need to estimate the model once. Given S draws from the full posterior $p(\theta|y)$, we can estimate the individual elpd contributions as

$$\log \hat{p}(y_i|y_{-i}) = \log \left(\frac{\frac{1}{S}\sum_{s=1}^{S} p_M(y_i|\theta_s) r(\theta_s)}{\frac{1}{S}\sum_{s=1}^{S} r(\theta_s)} \right), \quad (3)$$

$$r(\theta_s) = \frac{p_M(\theta_s|y_{-i})}{p_M(\theta_s|y)} \propto \frac{1}{p_M(y_i|\theta_s)}, \quad (4)$$

and where the last step is the result for factorizable likelihoods. In case of highly influential observations, the proposal distribution, i.e., the full posterior distribution can be very different than the target LOO posterior, and importance sampling estimates may have large variance. Vehtari et al. (2019b) present Pareto-smoothed importance sampling (PSIS) to smooth the importance ratios $r(\theta_s)$, introducing a small bias, but reducing the overall mean-squared error. The PSIS approach also has the benefit that the estimated shape parameter k of the generalized Pareto distribution can diagnose when the importance sampling approach has too large (or infinite) variance (Vehtari et al., 2017).

LOO is closely related to the Watanabe-Akaike or widely applicable information criterion (WAIC, Watanabe, 2010). The elpd of a given model can be estimated using WAIC as

$$\begin{aligned} \text{elpd}_{\text{WAIC}} &= \sum_{i=1}^{n} \log p(y_i|y) - V_\theta(\log p(y_i|\theta)) \\ &= \sum_{i=1}^{n} \log p(y_i|y) - p_{i,\text{eff}}\,, \end{aligned} \quad (5)$$

where $V_\theta(\log p(y_i|\theta))$ is the variance of the log likelihood over the (full) posterior $p(\theta|y)$, often called the *effective number of parameters* or p_{eff}. It has been shown that WAIC and LOO are asymptotically equivalent (Watanabe, 2010), but LOO has been found to be more robust than WAIC in the finite data domain, especially in the case of outliers or weak priors. This is because the WAIC approximation ignores higher order terms and these may be non-negligible for finite data (Gelman et al., 2014; Vehtari et al., 2016, 2017). Importantly, both LOO and WAIC are consistent estimators of the true elpd under mild assumptions (Watanabe, 2010).

1.3 LOO for Large Data

Magnusson et al. (2019) address two problems with Bayesian PSIS-LOO for large data. First the results of Gelfand (1996) in Eq. (3) are extended to approximate inference methods such as variational Bayes (VB) and Laplace approximations, and second an efficient subsampling method using the Hansen-Hurwitz (HH, Hansen and Hurwitz, 1943) estimator is proposed. Magnusson et al. (2019) use the full log predictive density $\log p(y_i|y)$ (lpd) and the $\log p(y_i|\hat{\theta})$, the point log predictive density (plpd) as an auxiliary variable, $\tilde{\pi}$. The data is then subsampled proportionally $\tilde{\pi}$ to efficiently estimate the elpd as

$$\widehat{\text{elpd}}_{\text{HH}} = \frac{1}{m} \sum_{j \in \mathcal{S}} \frac{1}{\tilde{\pi}_j} \log \hat{p}(y_j|y_{-j})\,, \quad (6)$$

where m is the subsample size and $\mathcal{S}$ is the subsample. This approach works well for estimating the elpd of individual models and has good theoretical properties, but it has two problems when used for model comparison.

First, when comparing models we are often interested in the elpd for a set of different models. Since the auxiliary information is used in the subsampling step, this means that we would need to draw a new subsample for each estimate of interest, such as (1) the elpd of each model, (2) the elpd difference between models, and (3) the variance of each elpd estimate. Ideally, we would like to just draw one subsample and then based on that subsample compute all estimates of interest.

Second, using the $\log p(y_i|\hat{\theta})$ as the auxiliary variable misses the effect of the efficient number of parameters

in the model p_{eff}, i.e., the model complexity, as can be seen in Eq. (5). This means that we would need larger subsample sizes when estimating more complex models.

1.4 Contributions and Limitations

In this paper, we focus on methods for scaling Bayesian LOO methods for comparing models for large data. We show that using the difference estimator combined with simple random sampling without replacement is very well suited for model comparison purposes. Since model auxiliary information is not used in the sampling stage, but in the estimation, the approach is much better suited for the situation of model comparison.

We also show that incorporating estimates of p_{eff} improves the performance of the subsampling and propose fast methods to approximate p_{eff} for large data and propose computationally efficient approximations, $\tilde{\pi}$, that take p_{eff} into account.

We prove that the difference estimator will converge in mean to the true LOO ($\widehat{\text{elpd}}_{\text{loo}}$) for any LOO approximation $\tilde{\pi}$ that converge in mean to π, irrespective of subsample size and the number of draws from the posterior. We also prove that our proposed approximations will converge in mean to π.

Together this makes the approach well suited for generic large-data model inference, such as in probabilistic programming frameworks as Stan (Carpenter et al., 2017).

The limitations with the proposed approach are the same as using general PSIS-LOO (see Vehtari et al., 2017, for a detailed discussion), such as that the likelihood needs to be factorizable for Eq. (3) to hold.

2 LARGE DATA MODEL COMPARISON USING LOO

Let elpd_{A} and elpd_{B} be the elpd_{loo} for model A and model B, respectively. To compare models, we are interested in the difference in elpd between models, $\text{elpd}_{\text{D}} = \text{elpd}_{\text{A}} - \text{elpd}_{\text{B}}$ as well as $V(\text{elpd}_{\text{D}})$, the variability due to the data, where

$$V(\text{elpd}_{\text{D}}) = V(\text{elpd}_{\text{A}}) + V(\text{elpd}_{\text{B}}) - 2\,\text{Cov}(\text{elpd}_{\text{A}}, \text{elpd}_{\text{B}})\,.$$

To efficiently estimate $V(\text{elpd}_{\text{D}})$, we propose to use the *difference estimator* and simple random sampling without replacement (SRS), something that previously has been used to scale MCMC (Quiroz et al., 2019). We also propose to include p_{eff} in Eq. (5) for better approximations of $\log p(y_i|y_{-i})$. This makes it possible to better compare models by computing the full posterior distributions *once* and then compare models performance on *one* subsample of observations.

2.1 The Difference Estimator

Let $\pi_i = \log p(y_i|y_{-i})$ be our variable of interest where $\text{elpd}_{\text{loo}} = \sum_i^n \pi_i$. Then let $\tilde{\pi}_i$ be any approximation of $\log p(y_i|y_{-i})$. Given $\tilde{\pi}_i$ we can use the difference estimator, a special case of the regression estimator (Cochran, 1977), together with SRS. The elpd_{loo} can then be estimated as

$$\widehat{\text{elpd}}_{\text{diff,loo}} = \sum_{i=1}^{n} \tilde{\pi}_i + \frac{n}{m} \sum_{j \in \mathcal{S}} (\pi_j - \tilde{\pi}_j)\,, \tag{7}$$

where m is the subsampling size and $\mathcal{S}$ is the subsample. The (subsample) variance associated with the difference estimator is

$$V(\widehat{\text{elpd}}_{\text{diff,loo}}) = n^2 \left(1 - \frac{m}{n}\right) \frac{s_e^2}{m}\,, \tag{8}$$

where s_e^2 is the sample standard deviations of the approximation error $e_j = \pi_j - \tilde{\pi}_j$, i.e $s_e^2 = \frac{1}{m-1}\sum_j^m (e_j - \bar{e})^2$ and $\bar{e} = \frac{1}{m}\sum_j^m e_j$.

The proposed approach has two important properties. First, as the sequence of numbers $\tilde{\pi}_i \to \pi_i$, $V(\widehat{\text{elpd}}_{\text{diff,loo}}) \to 0$. Unlike the HH estimator in Eq. (6), we also have the property that as $\frac{m}{n} \to 1$, $V(\widehat{\text{elpd}}_{\text{diff,loo}}) \to 0$. This finite correction factor increases the efficiency also in smaller data, where LOO still can be costly.

Second, the main benefits of using the difference estimator is that we can use a sampling scheme that do not depend on the models. Instead, we use the *same* subsample to estimate all properties of interest, such as elpd_{loo} for all models. This reduce the computational cost, especially for model comparisons, since we can reuse the already computed values for the sample when computing the elpd_D. Similarly, for model comparison, we are also interested in estimating $V(\text{elpd}_{\text{loo}}) = \sigma^2_{\text{loo}}$, the variability of the elpd_{loo} and elpd_{D}, for comparing models. Using the difference estimator we estimate σ^2_{loo} as

$$\begin{aligned}\hat{\sigma}^2_{\text{diff,loo}} = &\sum_{i=1}^{n} \tilde{\pi}_i^2 + \frac{n}{m} \sum_{j \in \mathcal{S}} \left(\pi_j^2 - \tilde{\pi}_j^2\right) + \\ &\frac{1}{n}\left[\left(\frac{n}{m}\sum_{j\in\mathcal{S}}(\pi_j - \tilde{\pi}_j)\right)^2 - V(\widehat{\text{elpd}}_{\text{diff,loo}})\right] + \\ &\frac{1}{n}\left[2\left(\sum_{i=1}^{n}\tilde{\pi}_i\right)\widehat{\text{elpd}}_{\text{diff,loo}} - \left(\sum_{i=1}^{n}\tilde{\pi}_i\right)^2\right]\,.\end{aligned} \tag{9}$$

Eq. (9) shows that using the difference estimator, we only need to compute $\tilde{\pi}_i^2$ to estimate σ^2_{loo}, using the

same subsample. The difference estimator is hence better suited for the case of large data Bayesian model comparison. We conclude by noting that the difference estimator is unbiased.

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

Proof. See the supplementary material. □

Remark Note that $\hat{\sigma}^2_{\text{diff,loo}}$ is most often an optimistic estimate for the variability of elpd_{loo}, since no general unbiased estimator of the true variability exists (Bengio and Grandvalet, 2004).

2.2 Fast Approximate LOO Surrogates

For the difference estimator to have small variance, we need good approximations of the variable of interest. We start with the following definition.

Definition 1. *An approximation* $\tilde{\pi}_i$ *of* π_i *is said to converge in mean if* $\mathbb{E}|\pi_i - \tilde{\pi}_i| \to 0$ *as* $n \to \infty$.

When estimating elpd_{loo} we want the approximation $\tilde{\pi}_i$ to have the following three properties:

1. a good finite data approximation of π_i,
2. computationally cheap, and
3. converges in mean to π_i.

The last property is needed for Proposition 2 and 3, that shows favorable theoretical scaling characteristics of the estimator as $n \to \infty$.

The WAIC estimator in Eq. (5) indicates that using the plpd as $\tilde{\pi}_i$, such as in Magnusson et al. (2019), will essentially miss the effect of the effective number of parameters $p_{\text{eff}} = V_\theta(\log p(y_i|\theta))$ in approximating π_i. Since it has been shown by Watanabe (2010) that the WAIC and LOO are asymptotically equivalent, including p_{eff} will improve over the plpd, especially for more complex models. Using the WAIC as approximation we set $\tilde{\pi}_i = \log p(y_i|y) - V_\theta(\log p(y_i|\theta))$, where $\text{elpd}_{\text{WAIC}} = \sum_i^n \tilde{\pi}_i$ in Eq. (5).

A problem with this approximation is that for each observation we need to integrate over the posterior to compute $\tilde{\pi}_i$ based on the WAIC in Eq. (5) and hence the approximation is more costly than using the plpd. To reduce the cost of computing $\tilde{\pi}$, the simplest way is to reduce the number of draws to approximate $\tilde{\pi}$ to $S_{\tilde{\pi}}$ where $S_{\tilde{\pi}} < S$, but this also reduces the accuracy.

Another approach to approximate π_i more computationally efficient is to approximate $p_{i,\text{eff}}$ in Eq. (5) directly using a Taylor approximation:

$$\tilde{p}_{i,\text{eff}} = \nabla \log(p(y_i|\theta))^T \Sigma_\theta \nabla \log(p(y_i|\theta)) + \frac{1}{2} \operatorname{tr}(H_{i,\theta} \Sigma_\theta H_{i,\theta} \Sigma_\theta), \quad (10)$$

where $\nabla \log(p(y_i|\theta))$ and $H_{i,\theta}$ are the gradient and Hessian of $\log(p(y_i|\theta))$ with respect to θ, respectively. This gives us an approximation of the $p_{i,\text{eff}}$ without the need to compute $V_\theta(\log p(y_i|\theta))$ over all S draws. We can use the idea to produce three different approximations, Δ_2WAIC that uses Eq. (10), Δ_1WAIC that only use the first order (gradient) term and $\Delta_1\text{WAIC}_{\text{m}}$ that only uses the first order term and the marginal variances, i.e., using $\text{diag}(\Sigma_\theta)$ instead of Σ_θ in Eq. (10).

Another approach to approximate PSIS-LOO is to use truncated importance sampling (TIS, Ionides, 2008). TIS-LOO will increase the bias but is less computationally costly since we remove the cost of estimating Pareto-k and smoothing using the Pareto distribution. As has been shown in Vehtari et al. (2019b), TIS-LOO can approximate LOO better than WAIC at the same computational cost. As with WAIC, we can also use TIS-LOO with fewer posterior draws to compute computationally less costly approximations of $\tilde{\pi}$.

2.3 Summary of Approach

The difference estimator and fast LOO surrogates lead us to how we can compare models for large data.

1. Compute the posterior $p_A(\theta|y)$ and $p_B(\theta|y)$ for model A and B, respectively.
2. Compute $\tilde{\pi}$ for model A and B using an approximation that fulfill the properties in Sec. 2.2.
3. Compute the approximate differences as $\tilde{\pi}_{i,D} = \tilde{\pi}_{i,A} - \tilde{\pi}_{i,B}$ for all n.
4. Draw a subsample of size m and compute $\pi_{j,D} = \pi_{j,A} - \pi_{j,B}$ for all m.
5. Estimate elpd_{D} and $V(\text{elpd}_{\text{D}})$ using Eq. (7) and (8).

Depending on the accuracy of the chosen approximation $\tilde{\pi}$, we can easily increase the subsampling size m to reach the desired accuracy.

2.4 Asymptotic Properties

Here we study the asymptotic properties of using the difference estimator together with any approximation $\tilde{\pi}_i$ that converges in mean to π. Let $(y_1, y_2, \ldots, y_n)$, $y_i \in \mathcal{Y} \subseteq \mathbb{R}$ be drawn from a true density $p_t = p(\cdot|\theta_0)$ with the true parameter θ_0 that is assumed to be drawn from $p(\theta)$ on the parameter space Θ, an open and bounded subset of $\mathbb{R}^d$. To prove Proposition 2 and 3 we make the following assumptions:

(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) for all θ, $\int_{\mathcal{Y}}(-\log p(y|\theta))p(y|\theta)dy < \infty$.
(v) all assumptions needed in the Bernstein-von Mises Theorem (Walker, 1969), and
(vi) $p(\theta|y) > 0$ or all $\theta \in \Theta$

Here we also generalize the definition of $r(\theta_s)$ in Eq. (3) to handle arbitrary posterior approximations (see Magnusson et al., 2019, for an extended discussion). Hence, let

$$r(\theta_s) \propto \frac{1}{p(y_i|\theta_s)} \frac{p(\theta_s|y)}{q(\theta_s|y)}.$$

Now, let $\widehat{\overline{\text{elpd}}}_{\text{diff,loo}} = \frac{1}{n}\widehat{\text{elpd}}_{\text{diff,loo}}$, we then have the following propositions.

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

Proof. See the supplementary material. □

Proposition 3. *Let the subsampling size m and the number of posterior draws S be fixed at arbitrary integer numbers, let the data 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 mean for any of the following choices of $\tilde{\pi}_i$, $i = 1, \ldots, n$.

(a) $\tilde{\pi}_i = \log p(y_i|\hat{\theta}_q)$.
(b) $\tilde{\pi}_i = \log p(y_i|y) + V_{\theta\sim p(\cdot|y)}(\log p(y_i|\theta))$.
(c) $\tilde{\pi}_i = \log p(y_i|y) - \nabla \log p(y_i|\hat{\theta})^T \Sigma_\theta \nabla \log p(y_i|\hat{\theta})$ *for any given fixed $\hat{\theta}$ and where the covariance matrix is with respect to $\theta \sim p(\cdot|y)$.*
(d) $\tilde{\pi}_i = \log p(y_i|y) - \nabla \log p(y_i|\hat{\theta})^T \Sigma_\theta \nabla \log p(y_i|\hat{\theta}) - \frac{1}{2}\text{tr}(\text{H}_{\hat{\theta}}\Sigma_\theta \text{H}_{\hat{\theta}})\Sigma_\theta)$ *for any given fixed $\hat{\theta}$ and where the covariance matrix is as in (c)*
(e) $\tilde{\pi}_i = \log \hat{p}(y_i|y_{-i})$ *as defined in (3).*

Proof. See the supplementary material. □

Proposition 2 and 3 generalizes the scaling properties of Magnusson et al. (2019), namely that in the limit, we essentially only need a subsampling size of $m = 1$ and $S = 1$ draw from the posterior to estimate the $\overline{\text{elpd}}_{\text{loo}}$ exact using the difference estimator. Proposition 2 show that for any $\tilde{\pi}$ that converges in mean to π the difference estimator will converge in mean to the true $\overline{\text{elpd}}_{\text{loo}}$. Using Proposition 3 we also have that the favorable scaling properties holds also for WAIC, our proposed approximations of WAIC, and using importance sampling for any choice of S. The results also hold for posterior approximations, as long as consistent posterior approximations are used, such as variational inference and Laplace approximations for regular models. In the supplementary material we also extend Proposition 3 to additional choices of $\tilde{\pi}$.

$\tilde{\pi}$	Needs	Cost
plpd	$\hat{\theta}$	$O(nP)$
TIS_S	S draws from $p(\theta\|y)$	$O(nPS)$
WAIC_S	S draws from $p(\theta\|y)$	$O(nPS)$
$\Delta_1\text{WAIC}_\text{m}$	$\nabla \log p(y_i\|\theta)$, $\hat{\theta}$ and $\text{tr}(\Sigma_\theta)$	$O(nP)$
$\Delta_1\text{WAIC}$	$\nabla \log p(y_i\|\theta)$, $\hat{\theta}$ and Σ_θ	$O(nP^2)$
$\Delta_2\text{WAIC}$	$\nabla \log p(y_i\|\theta)$, H_θ, $\hat{\theta}$ and Σ_θ	$O(nP^3)$

Table 1: Computational costs for approximations of π.

M	Description
1	Full pooling
2	Partial pooling
3	No pooling
4	Variable intercept
5	Variable slope
6	Variable intercept and slope

Table 2: Radon models. For full model specification, see supplementary material.

2.5 Computational Cost

The computational cost of the proposed method will depend on the total number of observations for which we need to compute $\tilde{\pi}_i$ and hence, in most situations, computing $\tilde{\pi}$ will dominate. This makes it relevant to understand how these costs relate to the total number of parameters in the likelihood function P (not the total number of parameters in the model) and the total number of posterior draws S. The overall cost for the different approximations proposed is presented in Table 1. Computing the full PSIS-LOO has the cost of $O(nPS)$, given that the evaluation of the log-likelihood is linear in P, i.e., the same complexity as WAIC, but with larger constants. Different trade-offs can be made depending on the specific likelihood where the approximation cost range from the cheapest, the plpd, to the most costly, WAIC/TIS with a large number of posterior draws S. The plpd only computes the log-likelihood once, while the full WAIC/TIS approach needs to compute it S times.

3 EXPERIMENTS

We study the proposed method using both simulated and real data. We simulated datasets with 10^4, 10^5, and 10^6 observations to fit Bayesian linear regression (BLR) models. The simulated data is generated with

different signal-to-noise ratios resulting in R^2 values of approximately 0.1, 0.5 and 0.9. To simulate sparse regression, we generated another dataset with only one covariate with $\beta \neq 0$. See supplementary material for the details of simulations. As the first real data, we use the radon data of Lin et al. (1999). The dataset consists of roughly 12 000 radon level measurements in 400 counties (groups). Table 2 lists different non-hierarchical and hierarchical models used. The exact model specification with Stan code can be found in the supplementary material. To compare different logistic models with simple linear effects, interaction effects and splines, we use the arsenic wells data of Gelman and Hill (2006) with 3 020 observations. The datasets are big enough to demonstrate the most important properties, but small enough to be easily fit using MCMC as a gold standard.

We use Stan (Carpenter et al., 2017; Stan Development Team, 2018) for inference using 4 chains, a sample size of 2 000, a warmup of 1 000 iterations and a dynamic HMC algorithm (Hoffman and Gelman, 2014; Betancourt, 2017) and the `rstanarm` R package (Goodrich et al., 2018) to fit spline models. Convergence diagnostics were made using $\hat{R}$ diagnostic (Gelman et al., 2013) and HMC specific diagnostics (Betancourt, 2017). For simplicity, in this paper we limit the scope to MCMC, but the proposed approach is trivial to use with any consistent posterior approximation using the approximation correction of Magnusson et al. (2019). Our approach has been implemented based on the `loo` R package (Vehtari et al., 2019a) framework for Stan and is available at `https://github.com/stan-dev/loo` and at `https://cran.r-project.org/package=loo`.

With the empirical evaluations, we study the following research questions: (1) does using better approximations of π improve the empirical performance, (2) which approximation $\tilde{\pi}$ should be preferred, (3) how does the difference estimator compare with the HH approach, and (4) how well does the method scale for large-data model comparison?

3.1 Performance

Table 3 shows the estimator variance when including p_{eff} in the estimation, by comparing our proposed approach with the HH method proposed in Magnusson et al. (2019). From the results we can see the benefit of including p_{eff} in the estimation of elpd_{loo}. Including p_{eff} improves the estimation by orders of magnitude compared to using only the plpd, both for the HH as well as for the difference estimator, both in turn improve with orders of magnitude over simple random sampling. Table 4 shows similar results where we see that all approximations of $\tilde{\pi}$ that include p_{eff} improves over using just the plpd as approximate surrogate.

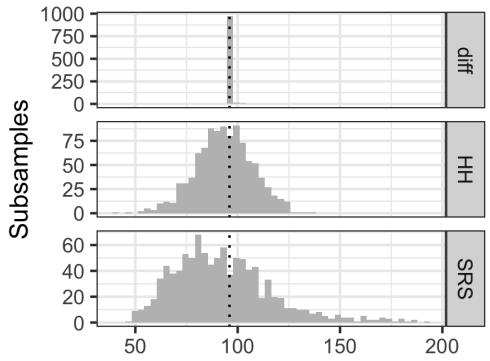

Figure 1: 1 000 estimates of σ_{loo} for Radon Model 6 using TIS_{2k} as $\tilde{\pi}$. True value is 96 (dotted line).

The downside of using the WAIC_{2k}, WAIC based on 2 000 draws, as the approximate surrogate variable is that it is computationally costly (see Table 1). Hence, it is of interest to study the performance of the different surrogate approximations shown in Table 4. We can see that any approximation is better than just using the plpd, as noted above. Using $\Delta_1\text{WAIC}_{\text{m}}$, that have the same computational complexity as the plpd, is better in all models, showing the benefit of including p_{eff} in $\tilde{\pi}_i$. Table 1 shows that as we improve the approximation we get better and better estimates of the elpd_{loo}, even though the benefit of the better approximations varies from model to model. TIS, computed by averaging 2 000 posterior draws, is the most accurate approach. Hence, we are confronted with the trade-off between sampling size and cost of computing $\tilde{\pi}$.

Table 3 contains the results of both the HH estimator and the difference estimator with WAIC_{2k} as $\tilde{\pi}$. The performance between the estimators are similar, but in most cases, the HH estimator is marginally better. This can be explained by the HH estimator being better at capturing heteroscedastic approximation errors (see Cochran, 1977, Ch. 7,9A). Since we expect the approximation to be worse for smaller values of π_i, this explains the difference in performance.

Figure 1 shows the benefit of the difference estimator in the estimation of σ_{loo}. It is quite clear that the difference estimator is much more efficient in estimating σ_{loo}, and that this efficiency comes from using both $\tilde{\pi}$ and $\tilde{\pi}^2$ as auxiliary variables in estimating σ_{loo}. To get the same efficiency for the HH estimator we would need to draw an additional subsample proportional to $\tilde{\pi}^2$ to reach a similar performance. Taken together, the HH estimator is marginally better in estimating elpd values for individual models. Although, the benefit of using $\tilde{\pi}_i$ in the estimation, instead of in the subsampling, is important in estimating σ^2_{loo} as well as when comparing models.

Data	M	$\text{SE}(\widehat{\text{elpd}}_{\text{diff,WAIC2k}})$	$\text{SE}(\widehat{\text{elpd}}_{\text{HH,plpd}})$	$\text{SE}(\widehat{\text{elpd}}_{\text{HH,WAIC2k}})$	$\text{SE}(\widehat{\text{elpd}}_{\text{SRS}})$	$\hat{\sigma}_{\text{loo}}$
Radon	1	0.002	0.7	0.001	949	88
	2	1.0	42	0.81	1012	94
	3	9.2	74	6.4	1012	94
	4	1.0	40	0.79	1013	94
	5	13	84	12	972	90
	6	10	97	15	1050	96
$R^2 = 0.9$	BLR	0.03	4.0	0.02	704	70
$R^2 = 0.5$		0.04	4.9	0.03	711	72
$R^2 = 0.1$		0.04	5.8	0.03	756	76

Table 3: The subsampling standard error (SE) for individual models of different subsampling approaches to estimate elpd_{loo} with $m = 100$. The results are averaged over 100 set of subsamples. $\hat{\sigma}_{\text{loo}}$ is the estimated standard deviation of the LOO-CV for comparison.

M	WAIC_{2k}	TIS_{2k}	WAIC_{100}	TIS_{100}	$\Delta_2\text{WAIC}$	$\Delta_1\text{WAIC}$	$\Delta_1\text{WAIC}_{\text{m}}$	plpd
1	0.0	0.0	1.6	1.6	0.5	0.5	0.6	1
2	1.0	0.2	21	20	30	31	31	53
3	9.2	1.7	29	29	45	56	56	87
4	1.0	0.3	22	22	29	30	30	51
5	13	9.8	26	36	32	39	40	81
6	10	7.5	34	42	50	57	90	107

Table 4: $\text{SE}(\text{elpd}_{\text{diff}})$ using different approximations $\tilde{\pi}$ for the Radon data with $m = 100$. The results are averaged over 100 set of subsamples.

n	M	$\tilde{\pi}$	$\widehat{\text{elpd}}_{\text{D}}$	$\text{SE}(\widehat{\text{elpd}}_{\text{D}})$
10^5	RHS vs. N	plpd	-47.6	5.2
	RHS vs. N	TIS_{10}	-52.3	0.04
10^6	RHS vs. N	plpd	-44.7	7.0
	RHS vs. N	TIS_{10}	-49.3	0.04

Table 5: Comparing models with Normal (N) and regularized horse-shoe (RHS) prior using a subsampling size of $m = 100$ and the difference estimator. The $\widehat{\text{elpd}}_D$ is the estimated difference in elpd_{loo} between models with the subsampling SE of the estimate. $\hat{\sigma}_D \approx 9$ for all.

3.2 Model Comparison

Table 5 shows a large-scale example of a Bayesian linear regression model with 100 covariates and 1 million and 100 000 simulated data points with only one $\beta \neq 0$ to compare between a normal prior and the regularized horseshoe shrinkage prior (Piironen and Vehtari, 2017). Using just the plpd we get a good approximation of π so a subsample of $m = 100$ is again sufficient to estimate the elpd with sufficient accuracy. Using a better LOO surrogate, such as TIS with only 10 posterior draws, increases the accuracy considerably for these large datasets, without much additional computational cost in computing $\tilde{\pi}$. Table 5 also shows the positive scaling characteristics, the size of subsample needed to compare models does not change much, even though the total number of folds in LOO is increased tenfold.

m	M	$\widehat{\text{elpd}}_{\text{D}}$	$\text{SE}(\widehat{\text{elpd}}_{\text{D}})$	$\hat{\sigma}_D$
100	6 vs. 4	-233	69	22
	6 vs. 2	-303	35	26
	6 vs. 3	-333	57	25
	6 vs. 5	-1451	32	51
	6 vs. 1	-1778	35	57
400	6 vs. 4	-236	22	24
	6 vs. 3	-294	20	26
	6 vs. 2	-298	16	27
	6 vs. 5	-1466	13	52
	6 vs. 1	-1780	13	58

Table 6: Comparing models using a subsample of size $m = 100, 400$, the difference estimator, and TIS_{100} as approximation. The $\widehat{\text{elpd}}_D$ is the estimated difference in elpd_{loo} between models with the subsampling SE of the estimate and $\hat{\sigma}_D$. $\hat{\sigma}_D$ is the estimated standard deviation of the elpd_D for comparison. The *naive* $\hat{\sigma}_D$ estimate is approximately 130 for all models

Table 6 shows an example of comparing the different models for the Radon data based on an individual subsample. Using only a subsample of size 100 we can roughly identify which models should be preferred. The estimated σ_{D} for difference in elpd compared to the reference model is much smaller than a naive approach where only the σ_{loo} of individual models are used, i.e., $\sigma^2_{\text{naive}} = V(\text{elpd}_A) + V(\text{elpd}_B)$. In this case, we use the best (and most complex) model as a reference with a subsample of size 100. Using a slightly bigger

Model	$\widehat{\text{elpd}}_{\text{D}}$	SE($\widehat{\text{elpd}}_{\text{D}}$)	$\hat{\sigma}_D$
GAM vs. interaction	-21	1.4	7.2
GAM vs. linear	-29	1.2	7.8

Table 7: Comparing arsenic models using $m = 300$ using the difference estimator and the TIS_{100} as approximation. The $\widehat{\text{elpd}}_D$ is the estimated difference in elpd_{loo} between models with the subsampling SE of the estimate. $\hat{\sigma}_D$ is the estimated standard deviation of the elpd_D.

Model	$\widehat{\text{elpd}}_{\text{D}}$	SE($\widehat{\text{elpd}}_{\text{D}}$)	$\hat{\sigma}_D$
$D = 100$ vs. $D = 101$	-0.5	0.03	0.6
$D = 100$ vs. $D = 110$	-4.3	0.04	3.3
$D = 100$ vs. $D = 99$	-9.7	0.04	5.1
$D = 100$ vs. $D = 90$	47.7	0.02	11.8

Table 8: Comparing BLR models with different number of included covariates for $R^2 \approx 0.1$, using a subsample of size $m = 100$ using the difference estimator and TIS_{100} as approximation. The $\widehat{\text{elpd}}_D$ is the estimated difference in elpd_{loo} between models with the subsampling SE of the estimate. $\hat{\sigma}_D$ is the estimated standard deviation of the elpd_D.

subsample size of 400, the subsampling uncertainty is reduced further and we can conclude that model 6 has the best predictive performance. Increasing the subsampling size is not very costly once $\tilde{\pi}$ has been computed.

With earlier approaches, these comparisons would be much more costly. Either we would need to compute π_i for all models and observations (full LOO), or if we use the HH estimator, we would need to draw a new set of observations for each model, each model comparison, and if we want an efficient estimate of σ_D, two sets of observations per model comparison.

In Table 7 we compare a generalized additive spline logistic model (GAM) with linear models with and without interactions. For these models, using TIS_{100} as approximation, we needed to increase the subsample size to 300 to compare the model performance. A small subsample size is sufficient as the model does not have a complex hierarchical structure and hence the approximation works well to compare even small differences in elpd_D.

Finally, Table 8 contains another simulated example with 100 covariates ($\beta = 1$ and $R^2 \approx 0.1$) and study the effect of adding irrelevant and removing relevant covariates for $n = 10\ 000$. Again we see that using only $m = 100$ we can get accurate estimates for elpd_D. The results of Table 8 also show the well-known inconsistency of LOO (Shao, 1993) in selecting the most parsimonious model, so for feature selection other approaches should be used (see Piironen et al., 2018).

4 CONCLUSIONS

Comparing different models is an important, but often overlooked, part of the process of predictive modeling. We propose a method for comparing and choosing between probabilistic models that are well suited to Bayesian model comparison for large data. First, using the difference estimator is much better suited to the setting of large-data model comparison. By using $\tilde{\pi}$ in the estimation rather than in sampling we reduce considerably the subsample size needed compared with approaches such as Magnusson et al. (2019), both when comparing models and estimating σ_{loo} and σ_{D}. Second, including the number of efficient parameters in the auxiliary variable when estimating the elpd improves with orders of magnitude over not using this information. But using the better surrogate approximations is costly and introduce an accuracy-computational cost trade-off. If the gradient of the likelihood with respect to likelihood parameters is available, $\Delta_1\text{WAIC}_{\text{m}}$ can be used to improve performance without an additional computational cost compared to plpd. In all, we recommend using plpd as approximation for simpler models, while TIS_{100} is recommended when comparing more complex hierarchical models.

We should not be too greedy and choose a very small subsample. A subsample that is too small, such as $m = 10$, may, due to randomness, miss observations for which the approximation is bad, but still are not too uncommon among the observations.

There are many additional possible improvements that we leave as future work. First, here we use the difference estimator with a simple random sample. We can use more or less any sampling strategy, such as stratifying the data based on the design matrix or by identified difficult observations. This can be further be used by adaptive optimal allocations between these strata. Second, we can further study other approaches to approximate LOO efficiently, here Wang et al. (2018); Giordano et al. (2019) are promising methods. As long as the approximation will converge in mean to π the theory holds, making the method general as well as highly tunable for the specific problem at hand.

In all, we propose a scalable method for fast model comparisons in case of large data.

Acknowledgements

We would like to thank the reviewers for their thoughtful comments and efforts in improving the quality of the paper. We would also thank participants at the Statistics Seminar at Stockholm University for valuable comments. The calculations presented above were partly performed using computer resources within the Aalto University School of Science "Science-IT" project. The research was funded by the Academy of Finland (grants 298742, 313122). Johan Jonasson was partly supported by WASP AI/Math, Måns Magnusson was partly funded by the Swedish Research Council (grants 201805170, 201806063), and Michael Riis Andersen was funded by Innovation Fund Denmark (grant number 8057-00036A).

References

Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5(Sep):1089–1105, 2004.

José M Bernardo. Expected information as expected utility. *The Annals of Statistics*, pages 686–690, 1979.

José M Bernardo and Adrian FM Smith. *Bayesian theory*. IOP Publishing, 1994.

Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

William G. Cochran. *Sampling Techniques, 3rd Edition*. John Wiley, 1977.

Alan E Gelfand. Model determination using sampling-based methods. *Markov chain Monte Carlo in practice*, pages 145–161, 1996.

Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, third edition, 2013.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and computing*, 24(6): 997–1016, 2014.

Ryan Giordano, William Stephenson, Runjing Liu, Michael Jordan, and Tamara Broderick. A Swiss Army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147, 2019.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

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.

Morris H. Hansen and William N. Hurwitz. On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362, 12 1943.

Leonhard Held, Birgit Schrödle, and Håvard Rue. Posterior and cross-validatory predictive checks: a comparison of mcmc and inla. In *Statistical modelling and regression structures*, pages 91–110. Springer, 2010.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

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

Chia-yu Lin, Andrew Gelman, Phillip N Price, and David H Krantz. Analysis of local decisions using hierarchical modeling, applied to home radon measurement and remediation. *Statistical Science*, pages 305–328, 1999.

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.

Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2): 5018–5051, 2017.

Juho Piironen, Markus Paasiniemi, and Aki Vehtari. Projective inference in high-dimensional problems: prediction and feature selection. *arXiv preprint arXiv:1810.02406*, 2018.

Matias Quiroz, Robert Kohn, Mattias Villani, and Minh-Ngoc Tran. Speeding up mcmc by efficient data subsampling. *Journal of the American Statistical Association*, 114(526):831–843, 2019.

Jun Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88 (422):486–494, 1993.

Stan Development Team. The Stan Core Library, 2018. URL `http://mc-stan.org/`. Version 2.18.0.

Aki Vehtari and Janne Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

Aki Vehtari, Tommi Mononen, Ville Tolvanen, Tuomas Sivula, and Ole Winther. Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *Journal of Machine Learning Research*, 17(1):3581–3618, 2016.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017.

Aki Vehtari, Jonah Gabry, Måns Magnusson, Yuling Yao, and Andrew Gelman. loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. *R package version 2.2.0*, 2019a. URL `https://mc-stan.org/loo`.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2019b.

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

Shuaiwen Wang, Wenda Zhou, Haihao Lu, Arian Maleki, and Vahab Mirrokni. Approximate leave-one-out for fast parameter tuning in high dimensions. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5228–5237. PMLR, 2018.

Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594, 2010.