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

Bayesian leave-one-out cross-validation for large data

Published in:
36th International Conference on Machine Learning, ICML 2019

Published: 01/01/2019

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Magnusson, M., Andersen, M., Jonasson, J., & Vehtari, A. (2019). Bayesian leave-one-out cross-validation for large data. In *36th International Conference on Machine Learning, ICML 2019* (Proceedings of Machine Learning Research; Vol. 97). JMLR. http://proceedings.mlr.press/v97/magnusson19a.html

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

Måns Magnusson [1] **Michael Riis Andersen** [1 2] **Johan Jonasson** [3] **Aki Vehtari** [1]

Abstract

Model inference, such as model comparison, model checking, and model selection, is an important part of model development. Leave-one-out cross-validation (LOO-CV) is a general approach for assessing the generalizability of a model, but unfortunately, LOO-CV does not scale well to large datasets. We propose a combination of using approximate inference techniques and probability-proportional-to-size-sampling (PPS) for fast LOO-CV model evaluation for large data. We provide both theoretical and empirical results showing good properties for large data.

1. Introduction

Model inference, such as model comparison, checking, and selection, is an integral part of developing new models. From a Bayesian decision-theoretic point of view (see Vehtari & Ojanen, 2012) we want to make a choice $a \in \mathcal{A}$, in our case a model p_M, that maximize our *expected utility* for a utility function $u(a, \cdot)$ as

$$\bar{u}(a) = \int u(a, \tilde{y}_i) p_t(\tilde{y}_i) d\tilde{y}_i \,,$$

where $p_t(\tilde{y}_i)$ is the true probability distribution generating observation $\tilde{y}_i$.

A common scenario is to study how well a model *generalizes* to unseen data (Box, 1976; Vehtari & Ojanen, 2012; Vehtari et al., 2017). A popular utility function u with good theoretical properties for probabilistic models is the log score function (Bernardo, 1979; Robert, 1996; Vehtari & Ojanen, 2012). The log score function give rise to using the *expected log predictive density* (elpd) for model inference, defined as

$$\overline{\text{elpd}}_M = \int \log p_M(\tilde{y}_i|y) p_t(\tilde{y}_i) d\tilde{y}_i \,,$$

where $\log p_M(\tilde{y}_i|y)$ is the log predictive density for a new observation for the model M.

Leave-one-out cross-validation (LOO-CV) is one approach to estimate the elpd for a given model, and is the method of focus in this paper (Bernardo & Smith, 1994; Vehtari & Ojanen, 2012; Vehtari et al., 2017). Using LOO-CV we can treat our observations as pseudo-Monte Carlo samples from $p_t(\tilde{y}_i)$ and estimate the $\overline{\text{elpd}}_{\text{loo}}$ as

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

where n is the number of observations (that may be very large), $p_M(y_i|\theta)$ is the likelihood, and $p_M(\theta|y_{-i})$ is the posterior for θ where we hold out observation i. This will henceforth be called the leave-one-out (LOO) posterior and $p_M(\theta|y)$ will be referred to as the full posterior. In this paper both $\overline{\text{elpd}}_{\text{loo}}$ and elpd_{loo} will be quantities of interest, depending on the situation.

Bayesian LOO-CV has many appealing theoretical properties compared to other common model evaluation techniques. The popular k-fold cross-validation is, in general, a biased estimator of elpd_M, since each model is only trained using a subset of the full data (Vehtari & Ojanen, 2012). The LOO-CV is, just as the Watanabe-Akaike Information criteria (WAIC), a consistent estimate of the true elpd_M for regular and singular models (Watanabe, 2010). A model is regular if the map taking the parameters to the probability distribution is one-to-one and the Fisher information is positive-definitive. If a model is not regular, then the model is singular (Watanabe, 2010). Since many models, such as neural networks, normal mixture models, hidden Markov models, and topic models, are singular, we need consistent methods to estimate the elpd for singular models (Watanabe, 2010). Although the WAIC and LOO-CV have the same asymptotic properties, recent research has shown that the LOO-CV is more robust than WAIC in the finite data domain (Vehtari et al., 2017).

In addition to the theoretical properties, the LOO-CV also gives an intuitive framework for evaluating models where the user easily can use different utility functions of interest

[1]Department of Computer Science, Aalto University, Finland [2]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark [3]Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Sweden. Correspondence to: Måns Magnusson <mans.magnusson@aalto.fi>.

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

as well as easily taking hierarchical data structures into account by using leave-one-group-out or leave-one-cluster out cross-validation (see Merkle et al., 2018). Taken together, LOO-CV has many very good properties, both empirical and theoretical. In this paper, we will focus on LOO-CV as a way of evaluating models.

Modern probabilistic machine learning techniques need to scale to massive data. In a data-rich regime, we often want complex models, such as hierarchical and non-linear models. Model comparison and model evaluation are important for model development, but little focus has been put into finding ways of scaling LOO-CV to larger data. The main problem is that a straight-forward implementation means that n models need to be estimated. Even if this problem is solved, for example using importance sampling (see below), we still have two problems.

First, many posterior approximation techniques, such as Markov Chain Monte Carlo (MCMC), does not scale well for large n. Second, computing elpd_{loo} still needs to be computed over n observations. If it is costly to estimate individual contributions (i.e. $\log p_M(y_i|y_{-i})$), computing the total elpd_{loo} may be very costly for very large models.

1.1. Pareto-Smoothed Importance Sampling

If we would implement LOO-CV naively, inference needs to be repeated n times for each model. Gelfand (1996) propose the use of importance sampling to solve this problem. The idea is to estimate $p_M(y_i|y_{-i})$ in Eq. (1) using the importance sampling approximation

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

where θ_s are $s \in 1, ..., S$ draws from the full posterior $p(\theta|y)$, and

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

where the last step is a well-known result of Gelfand (1996). In this way, only the full posterior is needed.[1] The ratios $r(\theta_s)$ can be unstable due to a long right tail, but this can be resolved using Pareto-smoothed importance sampling (PSIS) (Vehtari et al., 2015). When using PSIS to estimate $\log \hat{p}(y_i|y_{-i})$ (PSIS-LOO) we fit a generalized Pareto distribution to the largest weights $r(\theta_s)$ and replace the largest importance sample ratios with order statistics from the estimated generalized Pareto distribution, decreasing the variance by introducing a small bias. PSIS also has the benefit that we can use the estimated shape parameter $\hat{k}$ from the generalized Pareto distribution to determine the reliability of the estimate. For data-points with $\hat{k} > 0.7$ the estimates of $\log p(y_i|y_{-i})$ *can* be unreliable and hence $\hat{k}$ can be used as a diagnostic (Vehtari et al., 2017).

However, PSIS-LOO has the same scaling problem as LOO-CV in general since it requires (1) samples from the true posterior (e.g. using MCMC) and (2) the estimation of the elpd_{loo} contributions from all observations (Gelfand, 1996; Vehtari et al., 2017). Both of these requirements can be costly in a data-rich regime and are the main problems we address in this paper.

1.2. Contributions and Limitations

In this paper, we focus on the problems of LOO-CV for large datasets and our contributions are three-fold. First, we extend the method of Gelfand (1996) to posterior approximations by including a correction term to the importance sampling weights. In this way, we only need to estimate the posterior once to estimate the elpd_{loo} using Laplace and variational inference. Second, we propose sampling individual elpd_{loo} components with probability-proportional-to-size sampling (PPS) to estimate elpd_{loo}. Third, we show theoretically that these contributions have very favorable asymptotic properties as $n \to \infty$. We show that the proposed estimator for elpd_{loo} is consistent for any consistent posterior approximation q (such as Laplace approximations, mean-field, and full-rank variational inference posterior approximations). We also show that the variance due to subsampling will decrease as the number of observations n grows. In the limit, and given the assumptions in Section 2.3, we only need one subsampled observation, and one draw from the full posterior, to estimate $\overline{\text{elpd}}_{\text{loo}}$ with zero variance. Taken together this introduces a new, fast, and theoretically motivated approach to model evaluation for large datasets.

The limitations of our approach are the same as using PSIS-LOO (Vehtari et al., 2017) as well as the requirement that the approximate posterior needs to be sufficiently close to the true posterior (see Yao et al., 2018, for a discussion).

2. Bayesian Leave-One-Out Cross-Validation for Large Data Sets

Leave-one-out cross-validation (LOO-CV) has very good theoretical and practical properties. This makes it relevant to develop tools to scale LOO-CV. We solve this problem using scalable posterior approximations, such as Laplace and variational approximations and using probability-proportional-to-log-predictive-density subsampling inspired by Hansen & Hurwitz (1943).

[1] For certain models we can do efficient computations of LOO directly. See Vehtari et al. (2016) for an example using Gaussian processes.

2.1. Estimating the elpd_{loo} Using Posterior Approximations

Laplace and variational posterior approximations are attractive for fast model comparisons due to their computational scalability. Laplace (Lap) approximation approximates the posterior distribution with a multivariate normal distribution $q_{Lap}(\theta|y)$ with the mean being the mode of the posterior and the covariance the inverse Hessian at the mode (Azevedo-Filho & Shachter, 1994).

In variational inference, we minimize the Kullback-Leibler (KL) divergence between an approximate family $\mathcal{Q}$ of densities and the true posterior $p(\theta|y)$ (Jordan et al., 1999; Blei et al., 2017). Hence we find the approximation q that is closest to the true posterior in a KL divergence sense. Here we let $\mathcal{Q}$ be a family of multivariate normal distributions with a diagonal covariance structure (mean-field) or a full covariance structure (full-rank). Hence we will work with a mean-field (MF) variational approximation $q_{MF}(\theta|y)$ and a full-rank (FR) variational approximation $q_{FR}(\theta|y)$.

Although, all these posterior approximations, $q_{Lap}(\theta|y), q_{MF}(\theta|y)$, and $q_{FR}(\theta|y)$, will, in general, be different than the true posterior distribution. Although, we can use them as a proposal distribution in an importance sampling scheme. In this scheme we use a posterior approximation $q_M(\theta|y)$ for a model M as the proposal distribution and $p_M(\theta|y_{-i})$, the LOO posterior, as our target distribution. The expectation of interest is the same as in the standard PSIS-LOO given by Eq. (2), but we also propose to correct for the posterior approximation error. Hence we change $r(\theta)$ to

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

The factorization in Eq. (3) shows that the importance correction contains two parts, the correction from the full posterior to the LOO posterior and the correction from the full approximate distribution to the full posterior. Both components often have lighter tailed proposal distribution than the corresponding target distribution which can increase the variance of the importance sampling estimate (Geweke, 1989; Gelfand, 1996).

Pareto-smoothed importance sampling can be used to both stabilize the weights in estimating the contributions to the elpd_{loo} and in evaluating variational inference approximations using $\hat{k}$ as a diagnostic (Vehtari et al., 2015; Yao et al., 2018). Hence we use PSIS to stabilize the weights with the added benefit that we can use $\hat{k}$, the shape parameter in the generalized Pareto distribution, to diagnose how well the approximation is working (Vehtari et al., 2015). Together this gives us a tool for evaluating models using LOO-CV posterior but with the need of only computing one posterior approximation.

2.2. Probability-Proportional-to-Size Subsampling and Hansen-Hurwitz Estimation

Using PSIS we can estimate each $\log \hat{p}(y_i|y_{-i})$ term and sum them to estimate elpd_{loo}. Estimating every $\log \hat{p}(y_i|y_{-i})$ can be costly, especially as n grows. In some situations using PSIS-LOO, estimating elpd_{loo} can take even longer than computing the full posterior once, due to the computational burden of computing $\log \hat{p}(y_i|y_{-i})$, estimating $\hat{k}$ and using the generalized Pareto distribution to stabilize the weights for each individual observation. To handle this problem we suggest using a sample of the elpd_{loo} components to estimate elpd_{loo}.

Estimating totals, such as elpd_{loo}, has a long tradition in sampling theory (see Cochran, 1977). If we have auxiliary variables that are a good approximation of our variable of interest, we can use a probability-proportional-to-size (PPS) sampling scheme to reduce the sampling variance in the estimate of elpd_{loo} using the *unbiased* Hansen-Hurwitz (HH) estimator (Hansen & Hurwitz, 1943). When evaluating models, we can often easily compute $\log p_M(y_i|y)$, the full posterior log predictive density, for all observations. We then sample $m < n$ observations proportional to $\tilde{\pi}_i \propto \pi_i = -\log p_M(y_i|y) = -\log \int p_M(y_i|\theta) p_M(\theta|y) d\theta$. We here assume that all $\log p_M(y_i|y) < 0$, but this assumption is only for convenience.

In the case of regular models and large n, we can also approximate $\log p_M(y_i|y) \approx \log p_M(y_i|\hat{\theta})$ where $\hat{\theta}$ can be a Laplace posterior mean estimate $\hat{\theta}_q$ or a VI mean estimate $\mathbb{E}_{\theta\sim q}[\theta]$. In the case of VI and Laplace approximations, this further speeds up the computation of the $\tilde{\pi}_i$s since we do not need to integrate over θ for all n observations. Using a sampling with probability-proportional-to-size scheme, the estimator for elpd_{loo} can be formulated as

$$\widehat{\text{elpd}}_{\text{loo},q} = \frac{1}{n}\frac{1}{m}\sum_{i}^{m} \frac{1}{\tilde{\pi}_i} \log \hat{p}(y_i|y_{-i}) , \tag{4}$$

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 $\log p(y_i|y_{-i})$ given by Eq. (2) and (3), and m is the subsample size. The variance estimator can be expressed as (see Cochran, 1977, Theorem 9A.2.)

$$v(\widehat{\overline{\text{elpd}}}_{\text{loo},q}) = \frac{1}{n^2 m(m-1)} \sum_{i=1}^{m} \left(\frac{\log \hat{p}(y_i|y_{-i})}{\tilde{\pi}_i} - n\widehat{\overline{\text{elpd}}}_{\text{loo}} \right)^2. \quad (5)$$

The benefits of the HH estimator are many. First, if the probabilities are proportional to the variable of interest ($\log \hat{p}(y_i|y_{-i})$ here), the variance in Eq. (5) will go to zero, a property of use in the asymptotic analysis in Section 2.3. Second, the estimator of $\overline{\text{elpd}}_{\text{loo}}$ is not limited to posterior approximation methods, but can also be used with MCMC, but without the importance sampling correction factor. Third, PPS sampling has the benefit that we can use Walker-Alias multinomial sampling (Walker, 1977). By building an Alias table in $O(n)$ time we can then sample a new observation in $O(1)$ time. This means that can continue to sample observations until we have sufficient precision in $\overline{\text{elpd}}_{\text{loo}}$ for our model comparison purposes, independent of the number of observations n. Fourth, the estimator is unbiased for all $\tilde{\pi}_i$. So by using $\log p(y_i|\hat{\theta})$ instead of $\log p(y_i|y)$ we would expect a small increase in variance since, for finite n, $\log p(y_i|y)$ would be a better approximation of $\log p(y_i|y_{-i})$ than $\log p(y_i|\hat{\theta})$, but at a greater computational cost.

To compare models, we are often also interested in the variance of $\overline{\text{elpd}}_{\text{loo}}$, or for the dataset, henceforth called σ^2_{loo}. To estimate σ^2_{loo} we can use the same observations as sampled previously, as

$$\hat{\sigma}^2_{\text{loo}} = \frac{1}{nm} \sum_{i}^{m} \frac{\hat{p}_i^2}{\tilde{\pi}_i} + \frac{1}{n^2 m(m-1)} \sum_{i}^{m} \left(\frac{\hat{p}_i}{\tilde{\pi}_i} - \frac{1}{m} \sum_{i}^{m} \frac{\hat{p}_i}{\tilde{\pi}_i} \right)^2 - \left(\frac{1}{nm} \sum_{i}^{m} \frac{\hat{p}_i}{\tilde{\pi}_i} \right)^2 \quad (6)$$

where $\hat{p}_i = \log \hat{p}(y_i|y_{-i})$. For a proof of unbiasedness of the $\hat{\sigma}^2_{\text{loo}}$ estimator for σ^2_{loo} in Eq. (6), see the supplementary material. Also, note that here $\sigma^2_{\text{loo}} = \frac{1}{n}\sum_i^n (\hat{p}_i^2 - (\frac{1}{n}\sum_i^n \hat{p}_i^2)^2)$, which in itself is not an unbiased estimate for the true σ^2_{loo} (Bengio & Grandvalet, 2004).

Although the variance estimator is unbiased, it is not as efficient as the estimator of elpd_{loo}. This is partly due to the fact that $\tilde{\pi}_i$ is not proportional to $\hat{p}_i^2$ in the first line in Eq. (6). This can be solved by sampling in two steps both proportional to $\hat{p}_i$ and $\hat{p}_i^2$.

2.3. Asymptotic Properties

For larger data sets the asymptotic properties of the method are crucial and we derive asymptotic properties for the methods as follows. We consider a generic Bayesian model; 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, $\int_{\mathcal{Y}}(-\log p(y|\theta))p(y|\theta)dy < \infty$.

Of these assumptions, (i) and (iv) are the most restrictive. The assumption that the parameter space is bounded is not very restrictive in practice since we can approximate any proper prior arbitrarily well with a truncated approximation.

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

Proof. See the supplementary material. □

This proposition has three main points. First, the estimator of the $\widehat{\overline{\text{elpd}}}_{\text{loo}}$ is consistent for any consistent posterior approximation. In the limit, the mean-field variational approximation will also estimate the true $\overline{\text{elpd}}_{\text{loo}}$. Second, the estimator is also consistent irrespective of the sub-sampling

size m and the number of draws, S, from the posterior. This is a very good scaling characteristic. Third, the estimator is consistent also if we approximate $\tilde{\pi}_i$ with $\log p(y_i|\hat{\theta}_q)$. This means that for larger data we can plug in point estimates to quickly compute $\tilde{\pi}_i$ and still have the consistency property.

The main limitations with Proposition 1 are that it is based on the consistency of the posterior approximations and the proposition does only hold for regular models for which q are consistent. This is mainly due to the fact that Laplace and variational inference are not, in general, consistent for singular models.

2.4. Computational Complexity

In the large n domain it is also of interest to study the computational complexity of our approach. Assuming that the additional cost of computing $p(y_i|y_{-i})$ compared to the point log predictive density (lpd) at $\hat{\theta}$, $\log p(y_i|\hat{\theta})$, is $O(S)$, where S is the number of samples from the full posterior. Then the cost of computing the full elpd_{loo} is

$$O(nS)\,.$$

If we instead use our proposed method we would have the complexity

$$O(n + mS)\,,$$

where m is the subsampling size. Using the proposed approach, we get an *unbiased* estimate of elpd_{loo} together with the variance $v(\widehat{\text{elpd}}_{\text{loo}})$ of that estimate, giving us information on the precision of the method for a given m.

Finally, we could, for large n just use the same lpd as an approximation with complexity

$$O(n)\,.$$

This estimate is though *biased* for all finite n, and we have no diagnostic indicating how good or bad the approximation is.

This shows the large-scale characteristic of our proposed approach. By adding a small cost, mS, we will have a good estimate of the true elpd_{loo} at the same cost as computing just the lpd. If using the point lpd is a good approximation we would need less m. On the other hand, if the point lpd would be a bad approximation, we would need a larger m. The variance estimator in Eq. (5) would in these situations serve as an indicator, with a higher variance estimate.

2.5. Method Summary

We have presented a method for estimating the elpd efficiently using posterior approximation and PPS subsampling. One of the attractive properties of the method is that we can diagnose if the method is working. Using PSIS-LOO we can diagnose the estimation of each individual $\log \hat{p}(y_i|y_{-i})$ as well as the overall posterior approximation using the $\hat{k}$ diagnostic. Then the variance of the HH estimator in Eq. (5) captures the effect of the subsampling in the finite n case. Our approach for large-scale LOO-CV can be summarized in the following steps.

1. Estimate the models of interest using any consistent posterior approximation technique.
2. Compute the $\hat{k}$ diagnostic for the posterior to asses the general overall posterior approximation (see Yao et al., 2018).
3. Compute $\tilde{\pi}_i \propto -\log p(y_i|y)$ for all n observations. For regular models this can be approximated with $\tilde{\pi}_i \propto -\log p(y_i|\hat{\theta})$ for large data.
4. Sample m observations using PPS sampling and compute $\log \hat{p}(y_i|y_{-i})$ using Eq. (2) and (3). Use $\hat{k}$ to diagnose the estimation of each individual $\log \hat{p}(y_i|y_{-i})$.
5. Estimate $\widehat{\text{elpd}}_{\text{loo}}$, $v(\widehat{\text{elpd}}_{\text{loo}})$, and $\hat{\sigma}^2_{\text{loo}}$ using Eq. (4), (5), and (6) to compare model predictive performance.
6. Repeat step 3 and 4 until sufficient precision is reached.

The downside is that the $\hat{k}$ diagnostic can be too conservative for our purpose. In the case of a correlated posterior and mean-field variational inference, $\hat{k}$ may indicate a poor approximation even though the estimation of elpd_{loo} is still consistent and may work well. In this situation, we would get a result indicating that all $\log \hat{p}(y_i|y_{-i})$ are problematic, even though the estimation actually work well, something we will see in the experiments. Also note that setting m too small and then repeating step 3 and 4 multiple times may create a multiple comparison problem.

3. Experiments

To study the characteristic of the proposed approach we study multiple models and datasets. We use simulated datasets used to fit a Bayesian linear regression model with D variables and N observations. The data is generated such that so we get either a correlated (c) or an independent (i) posterior for the regression parameters by construction. This will enable us to study the effect of the mean-field assumptions in variational posterior approximations. In addition, we use data from the radon example of Lin et al. (1999) to show performance on a larger dataset with multiple models.

All posterior computations use Stan 2.18 (Carpenter et al., 2017; Stan Development Team, 2018) and all models used can be found in the supplementary material. The methods have been implemented using the `loo` R package (Vehtari et al., 2018) framework for Stan and are available as supplementary material. We use mean-field and full-rank

Automatic Differentiation Variational Inference (ADVI, Kucukelbir et al., 2017) and Laplace approximations as implemented in Stan. ADVI automatically handles constrained variables and uses stochastic variational inference. We used 100 000 iterations for ADVI and 1000 warmup iterations and 2000 samples from 4 chains for the MCMC.

3.1. Estimating elpd_{loo} Using Posterior Approximations

Table 1 contains the estimated values of elpd_{loo} for different posterior approximations. From the table, we can see that using PSIS-LOO and posterior approximations to estimate elpd_{loo} works well or diagnostic correctly indicates the failure. As we would expect, the mean-field approximation for the correlated posterior does not approximate the true posterior very well when the posterior has correlated parameters and the $\hat{k}$ values are too high for all observations. In spite of the high $\hat{k}$ values, the estimate of the elpd_{loo} is not very far from the (gold-standard) MCMC estimate, showing the consistency result in Prop. 1 for mean-field approximations - even when the true posterior covariance structure is not in the variational family.

The second result is that the full-rank VI approximation has a poor fit for a large number of parameters ($D = 100$). This comes from that the full rank ADVI needs to approximate the full posterior covariance structure (with 5 000 parameters) based on stochastic gradients. The increased perturbation in the estimate of the covariance matrix has the effect of increasing the overall $\hat{k}$, especially for larger dimensions indicating a less good approximation of the posterior.

3.2. Subsampling Using Probability-Proportional-to-Size Sampling

Table 2 show empirical results on the effect of using subsampling proportional to the predictive density compared to simple random sampling (SRS). The results are much in line with what we would expect from the theory presented in Section 2.3. We can see that the proposed method, sampling proportional to $-\log(p(y_i|y))$ (PPS(1)) and sampling proportional to $-\log(p(y_i|\hat{\theta}))$ (PPS(2)) outperforms SRS with orders of magnitude. Using just a sample size of $m = 10$ observations using our proposed method is much more precise than using $m = 1000$ observations with SRS, although we can see that the estimate $\hat{\sigma}_{\text{loo}}$ is less reliable for such small sample sizes. This results can be explained by the sampling probabilities used in the subsampling procedure. Figure 3.2 show the distribution of sampling probabilities where we can see that the probabilities are highly skewed, indicating the reason for the inefficiency of the SRS compared to the HH approaches. Table 2 also show that sampling with $\tilde{\pi} \propto -\log(p(y_i|\hat{\theta}))$ is marginally more accurate. In many situations with larger data we also would expect that using a point estimate, $\hat{\theta}$, of the parameters in computing the likelihood values would be much faster than computing $-\log(p(y_i|y))$.

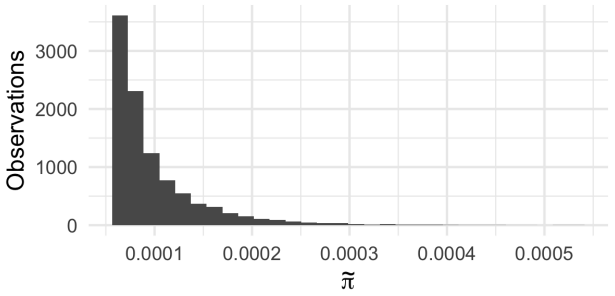

Figure 1. Sampling probabilities ($\tilde{\pi}$) for the LR(cor) 100D data and $\pi_i = -\log p(y_i|y)$. The results are very similar for the other LR data and for $\pi_i = -\log p(y_i|\hat{\theta})$.

Table 3 shows empirical results on the scaling characteristics of the proposed method. We can see that for the PPS estimator, as the size of the data, n, increases, the variance of the estimator is more or less constant. Using a SRS scheme, on the other hand, clearly show that to estimate the total elpd_{loo}, we would need to increase the sample size, m, as the number of observations, n, increases.

3.3. Hierarchical Models for Radon Measurements

As an example of how the proposed method can be used, we exemplify with the dataset of (Lin et al., 1999), used as an example of hierarchical modeling by Gelman & Hill (2006).[2] The data make up a total of 12 573 home radon measurements in a total of 386 counties with a different number of observations per county. This example is enlightening for a number of reasons. First, it is large enough to actually take some computing time to analyze but is still small enough so we can use MCMC to compute the full data elpd_{loo} as a gold standard. The models used here are also both regular and singular, showing the usability in a broader class of models. Finally, this example also shows how we can mix different approximation techniques for different models when doing model comparisons.

We compare seven different linear models of predicting the log radon levels in individual houses based on floor measurements and county uranium levels. The seven models are a pooled simple linear model (model 1), a non-pooled model with one intercept estimated per county (model 2), a partially pooled model with a hierarchical mean parameter per county (model 3), a variable intercept model per county (model 4), a variable slope model per county (model 5), a

[2]We base our example and data on the Stan case study by Chris Fonnesbeck at https://mc-stan.org/users/documentation/case-studies/radon.html

Data		ADVI(FR)	ADVI(MF)	Laplace	MCMC
LR(c) 100D	elpd_{loo}	-14249	-14267	-14247	-14247
	$\hat{k} > 0.7$ (%)	100	100	0	0
LR(c) 10D	elpd_{loo}	-14271	-14271	-14272	-14272
	$\hat{k} > 0.7$ (%)	0	100	0	0
LR(i) 100D	elpd_{loo}	-14193	-14239	-14238	-14239
	$\hat{k} > 0.7$ (%)	100	0	0	0
LR(i) 10D	elpd_{loo}	-14202	-14202	-14202	-14203
	$\hat{k} > 0.7$ (%)	0	0	0	0

Table 1. Estimation of elpd_{loo} using posterior approximations. For all models and posterior approximations, $\sigma_{\text{LOO}} \approx 70$. No subsampling is used and MCMC is gold standard.

Data	m	Method	$\widehat{\text{elpd}}_{\text{loo}}$	$\text{SE}(\widehat{\text{elpd}}_{\text{loo}})$	$\hat{\sigma}_{\text{loo}}$
LR(c)	-	**True**	**-14245**	**0**	**71**
100D	10	PPS(1)	-14231	21.6	46
		PPS(2)	-14236	9.1	63
		SRS	-14309	2316.1	73
	100	PPS(1)	-14240	7.1	67
		PPS(2)	-14242	2.3	76
		SRS	-14798	678.6	68
	1000	PPS(1)	-14245	2.4	72
		PPS(2)	-14246	0.7	72
		SRS	-13897	181.8	61
LR(c)	-	**True**	**-14272**	**0**	**71**
10D	10	PPS(1)	-14277	5.5	98
		PPS(2)	-14273	2.9	98
		SRS	-11437	775.3	25
	100	PPS(1)	-14272	1.0	69
		PPS(2)	-14272	0.5	71
		SRS	-14083	637.5	64
	1000	PPS(1)	-14271	0.3	69
		PPS(2)	-14272	0.2	69
		SRS	-14591	236.9	79

Table 2. Effect of subsampling proportional to log predictive density. The result are based on MCMC draws and $\hat{\theta}$ is the posterior mean for the parameters. PPS(1) is subsampling proportional to $-\log(p(y_i|y))$, PPS(2) is subsampling proportional to $-\log(p(y_i|\hat{\theta}))$, and SRS is simple random sampling.

m	n	PPS	SRS	σ_{loo}
10	100	4.3	16	8
	1000	2.3	360	22
	10000	5.7	9351	72
	100000	30	19715	225
100	100	1.3	9.7	8
	1000	1.2	83.7	22
	10000	2.2	1144	72
	100000	11.3	5894	225

Table 3. Standard errors, $\text{SE}(\widehat{\text{elpd}}_{\text{loo}})$, for PPS and SRS subsampling in relationship with σ_{loo}. The result are based on MCMC draws and $\hat{\theta}$ is the posterior mean for the parameters. PPS is subsampling proportional to $-\log(p(y_i|\hat{\theta}))$.

M	Laplace	ADVI(FR)	ADVI(MF)
1	0.24	0.23	0.34
2	0.93	5.11	0.45
3	1.44	4.19	0.28
4	2.05	6.99	0.71
5	-	5.62	1.04
6	-	10.99	2.98
7	1.70	7.39	0.89

Table 4. Posterior $\hat{k}$ values for the different radon models and approximations. Laplace was not possible for model 5 and 6.

variable intercept and slope model (model 6), and finally a model with a county level features and county level intercepts using the log uranium level in the county. We use vague priors based on the Stan prior choice recommendations[3] with $N(0, 10)$ priors on regression coefficients and intercepts and half-$N(0, 1)$ for variance parameters.

[3]See https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Table 4 show the $\hat{k}$ values for different approximations for the different models. We can see that for the simplest model (1) we get a good posterior approximation with just Laplace approximation, but as the models become more complex (and singular) we need better approximation techniques such as ADVI. We can see that ADVI(MF) in general perform well and can be used for inference in many models, even though more complex models (such as model 4-7) is not approximated sufficiently well using the mean-field approx-

M	Method	$\widehat{\text{elpd}}_{\text{loo}}$	SE	elpd_{loo}	$\text{elpd}_{\text{10fcv}}$
1	Laplace	-18560	0.3	-18560	-18561
	ADVI(FR)	-18562	1.0	-18559	-18560
	ADVI(MF)	-18564	2.1	-18559	-18558
	MCMC	-18560	0.4	-18560	-18561
2	Laplace	-17058	31.9	-17049	-17142
	ADVI(MF)	-17068	50.1	-17059	-17105
	MCMC	-17069	29.6	-17067	-17137
3	Laplace	-17035	33.0	-17017	-17117
	ADVI(MF)	-17097	20.7	-17090	-17090
	MCMC	-17096	18.2	-17086	-17112
4	Laplace	-17003	66.6	-16866	-17057
	ADVI(MF)	-16990	19.8	-17013	-17034
	MCMC	-17013	17.1	-17021	-17049
5	ADVI(MF)	-18223	37.6	-18225	-18285
	MCMC	-18200	16.7	-18259	-18288
6	ADVI(MF)	-16656	90.8	-16603	-16869
	MCMC	-16758	29.2	-16801	-16873
7	Laplace	-17096	45.8	-17063	-17140
	ADVI(MF)	-16996	25.4	-16957	-17035
	MCMC	-17130	33.2	-17138	-17060

Table 5. The estimated $\widehat{\text{elpd}}_{\text{loo}}$ using a subsample of size $m = 500$ and its standard error (SE). The full elpd_{loo} based on all observations is also included as well as $\text{elpd}_{\text{10fcv}}$, an estimation of the elpd using 10-fold cross-validation. The $\sigma_{\text{loo}} \approx 90$ for all approximations and models. Less than 1% of the observation have problematic $\hat{k}$ using MCMC, making it a good gold standard.

imation. ADVI(FR), again, have problems due to the larger number of parameters in the more complex models.

Based on these approximate posteriors we can analyze the elpd_{loo} for the different models. Table 5 shows the elpd_{loo} and an estimate, $\widehat{\text{elpd}}_{\text{loo}}$, based on a subsample of size 500. As a comparison we also compute $\text{elpd}_{\text{10fcv}}$, computing an estimate of elpd using a 10-fold cross-validation scheme, without bias correction. For the simple baseline model 1, we can use Laplace approximation and a subsample to estimate the elpd_{loo} in roughly 2 seconds with a sufficient precision for most purposes. Using MCMC and computing the full elpd take roughly 35 seconds for this medium-sized dataset.

Table 5 also shows that ADVI(MF) work well both for regular and singular models. Using ADVI(MF) for the singular models 3 and 4, where the $\hat{k}$ values indicating a good posterior approximation, the approach works really well. We can also see that the $\hat{k}$ diagnostic works well as an indicator. The Laplace and ADVI(MF) approximations with high $\hat{k}$ values can be quite off, see model 7 for an example.

The results of Table 5 also give us an idea of how the subsampling can be used. By comparing the SE of our estimates with σ_{loo}, that is roughly 90 for all models, we see how far a subsample with 500 observations takes us. For most models, our SE is small enough to help us decide between models, while for the more complex models. The precision needed depends on the specific use case and if we need better precision we can simply add more subsamples to get the precision needed.

If we study Table 5 we see that using ADVI(MF), Laplace and a subsample of size 500 we can get quite far comparing these models. We could quickly rule out model 1 and 5, but where we would need to use MCMC for model 5, due to the high $\hat{k}$ for the ADVI approximations. Model 4 and 6 are the most promising but we need to estimate the models using MCMC due to the high $\hat{k}$ values for the ADVI approximations. Although, based on just the subsample, we can see that model 6, the variable intercept and slope model, seem to be the most promising model for this data. Comparing the fully computed elpd_{loo} for the different models we could compute the difference in elpd_{loo} between model 6 and 4 to 220 with a standard error of 26, clearly indicating that model 6 is the one to prefer in this situation. Using 10-fold cross-validation (10fcv, see $\text{elpd}_{\text{10fcv}}$), we arrive at a similar result, but at the cost of re-estimating the model 10 times.

4. Conclusions

In this work we solve the two major hurdles for using LOO-CV for large data, namely using posterior approximations to estimate the elpd_{loo} for individual observations and efficient subsampling. We prove the consistency in n and also show that for regular models we have consistency also for common posterior approximations such as Laplace and ADVI, even for mean-field ADVI in situations with correlated posteriors, making the results promising for large-scale model evaluations. Finally, our proposed method also comes with diagnostics to assess if the quality of the subsampling and posterior approximations. We can use the $\hat{k}$ diagnostic to asses the posterior approximations and $v(\widehat{\text{elpd}}_{\text{loo}})$, the variance of the HH estimator, to give us a good measure of the uncertainty due to subsampling.

Acknowledgments

The research was funded by the Academy of Finland (grants 298742, 313122). We would like to thank the reviewers for their thoughtful comments and efforts in improving the quality of the paper. The calculations presented above were partly performed using computer resources within the Aalto University School of Science "Science-IT" project. Johan Jonasson was partly supported by WASP AI/Math.

References

Azevedo-Filho, A. and Shachter, R. D. Laplace's method approximations for probabilistic inference in belief networks with continuous variables. In *Uncertainty Proceedings 1994*, pp. 28–36. Elsevier, 1994.

Bengio, Y. and Grandvalet, Y. No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 5(Sep):1089–1105, 2004.

Bernardo, J. M. Expected information as expected utility. *the Annals of Statistics*, pp. 686–690, 1979.

Bernardo, J. M. and Smith, A. F. *Bayesian theory*. IOP Publishing, 1994.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Box, G. E. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

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

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

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

Geweke, J. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society*, pp. 1317–1339, 1989.

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

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1): 430–474, 2017.

Lin, C.-y., Gelman, A., Price, P. N., and Krantz, D. H. Analysis of local decisions using hierarchical modeling, applied to home radon measurement and remediation. *Statistical Science*, pp. 305–328, 1999.

Merkle, E., Furr, D., and Rabe-Hesketh, S. Bayesian model assessment: Use of conditional vs marginal likelihoods. *arXiv preprint arXiv:1802.04452*, 2018.

Robert, C. P. Intrinsic losses. *Theory and decision*, 40(2): 191–214, 1996.

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

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

Vehtari, A., Gelman, A., and Gabry, J. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., and Winther, O. Bayesian leave-one-out cross-validation approximations for gaussian latent variable models. *The Journal of Machine Learning Research*, 17(1):3581–3618, 2016.

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

Vehtari, A., Gelman, A., Gabry, J., Yao, Y., Piironen, J., and Goodrich, B. loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. *R package version 2.0.0*, 2018.

Walker, A. J. An efficient method for generating discrete random variables with general distributions. *ACM Transactions on Mathematical Software (TOMS)*, 3(3):253–256, 1977.

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

Watanabe, S. 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.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. Yes, but did it work?: Evaluating variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5581–5590, 2018.