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

Andersen, Michael Riis; Magnusson, Mans; Jonasson, Johan; Vehtari, Aki

Published in: Proceedings of the 36th International Conference on Machine Learning

Publication date: 2019

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

Link back to DTU Orbit

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 \( a \in A \), 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 gives rise to using the expected log predictive density (elpd) for model inference, defined as

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

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 \( \text{elpd}_{\text{loo}} \) as

\[
\text{elpd}_{\text{loo}} = \frac{1}{n} \sum_{i=1}^n \log p_M(y_i|y_{-i})
= \frac{1}{n} \sum_{i=1}^n \log \int p_M(y_i|\theta)p_M(\theta|y_{-i})d\theta
= \frac{1}{n} \text{elpd}_{\text{loo}},
\]

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 \( \theta \) 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 \( \text{elpd}_{\text{loo}} \) and \( \text{elpd}_{\text{loo}} \) 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 \( \text{elpd}_M \), 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 \( \text{elpd}_M \) 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-definite. 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.

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

Bayesian LOO-CV for Large Data

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 straightforward 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

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

(2)

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

\[ r(\theta_s) = \frac{p_M(\theta_s|y_{-i})}{p_M(\theta_s|y)} \]

\[ \propto \frac{1}{p_M(\theta_s|y_{-i})}, \]

where the last step is a well-known result of Gelfand (1996). In this way, only the full posterior is needed. 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\textsubscript{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\textsubscript{loo} using Laplace and variational inference. Second, we propose sampling individual elpd\textsubscript{loo} components with probability-proportional-to-size sampling (PPS) to estimate elpd\textsubscript{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\textsubscript{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 elpd\textsubscript{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).
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 \( q_{\text{Lap}}(\theta | y) \) 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 \( Q \) 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 \( Q \) 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 \( q_{\text{MF}}(\theta | y) \) and a full-rank (FR) variational approximation \( q_{\text{FR}}(\theta | y) \).

Although, all these posterior approximations, \( q_{\text{Lap}}(\theta | y) \), \( q_{\text{MF}}(\theta | y) \), and \( q_{\text{FR}}(\theta | y) \), 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

\[
r(\theta) = \frac{p_{M}(\theta_{i} | y_{-i})}{q_{M}(\theta_{i} | y)} = \frac{p_{M}(\theta_{i} | y_{-i}) p_{M}(\theta_{i} | y)}{p_{M}(\theta_{i} | y)} q_{M}(\theta_{i} | y) \approx \frac{1}{p_{M}(y_{i} | \theta)} q_{M}(\theta_{i} | y). \tag{3}
\]

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 \( \pi_{i} \propto \pi_{i} = -\log p_{M}(y_{i} | y) = -\log \int p_{M}(y | \theta)p_{M}(\theta | y)d\theta \). 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 \( \log p_{M}(y_{i}|y) \approx \log p_{M}(y_{i}|\theta) \) where \( \theta \) can be a Laplace posterior mean estimate \( \hat{\theta}_{q} \) or a VI mean estimate \( E_{\theta \sim q}[\theta] \). In the case of VI and Laplace approximations, this further speeds up the computation of the \( \pi_{i}s \) since we do not need to integrate over \( \theta \) for all \( n \) observations. Using a sampling with probability-proportional-to-size scheme, the estimator for elpd_{loo} can be formulated as

\[
\hat{\text{elpd}}_{\text{loo}, q} = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\pi_{i}} \log \hat{p}(y_{i}|y_{-i}), \tag{4}
\]

where \( \pi_{i} \) 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.)
Although the variance estimator is unbiased, it is not as 
would expect a small increase in variance since, for finite 
\( n \) (Eq. (6)). This can be solved by sampling in two steps both 
sampled previously, as 
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 \( \hat{\text{elpd}}_{\text{loo}} \) for our model comparison purposes, independent of the number of observations \( n \). Fourth, the estimator is unbiased for all \( \pi_i \). So by using \( \log p(y_i|\theta) \) 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 \( \log p(y_i|\theta) \), but at a greater computational cost.

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

\[
\tilde{\sigma}^2_{\text{loo}} = \frac{1}{nm} \sum_{i=1}^{m} \left( \hat{p}_i - \frac{1}{nm} \right)^2 \frac{\hat{p}_i}{\hat{\pi}_i} \sum_{i=1}^{m} \left( \hat{p}_i - \frac{1}{nm} \right)^2 \frac{\hat{p}_i}{\hat{\pi}_i} \sum_{i=1}^{m} \left( \hat{p}_i - \frac{1}{nm} \right)^2 \frac{\hat{p}_i}{\hat{\pi}_i} \]  \tag{6}

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

Although the variance estimator is unbiased, it is not as efficient as the estimator of \( \hat{\text{elpd}}_{\text{loo}} \). This is partly due to the fact that \( \hat{\pi}_i \) 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 \( \theta_0 \). The parameter \( \theta_0 \) is assumed to be drawn from a prior \( p(\theta) \) on the parameter space \( \Theta \), 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 \( C : \mathcal{Y} \rightarrow \mathbb{R}_+ \), such that \( \mathbb{E}_{y \sim p_t}[C(y)^2] < \infty \) and such that for all \( \theta_1 \) and \( \theta_2 \), \( |p(y|\theta_1) - p(y|\theta_2)| \leq C(y)p(y|\theta_2)||\theta_1 - \theta_2|| \).

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

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

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

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

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 \( \theta_0 \). Then

\[
|\hat{\text{elpd}}_{\text{loo}}(m, q) - \hat{\text{elpd}}_{\text{loo}}(q)| \rightarrow 0
\]

in probability as \( n \rightarrow \infty \) for any of the following choices of \( \pi_i, i = 1, \ldots, n \).

(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 \( \hat{\text{elpd}}_{\text{loo}} \) is consistent for any consistent posterior approximation. In the limit, the mean-field variational approximation will also estimate the true \( \hat{\text{elpd}}_{\text{loo}} \). Second, the estimator is also consistent irrespective of the sub-sampling
We have presented a method for estimating the elpd efficiently using posterior approximation and PPS subsampling. This is a very good scaling characteristic. Third, the estimator is consistent also if we approximate \( \hat{\pi}_i \) with \( \log p(y_i|\theta_q) \). This means that for larger data we can plug in point estimates to quickly compute \( \hat{\pi}_i \) 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\(_{\text{loo}}\) 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\(_{\text{loo}}\) together with the variance \( \nu(\text{elpd}_{\text{loo}}) \) 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\(_{\text{loo}}\) 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 assess the general overall posterior approximation (see Yao et al., 2018).
3. Compute \( \hat{\pi}_i \propto -\log p(y_i|y) \) for all \( n \) observations. For regular models this can be approximated with \( \hat{\pi}_i \propto -\log p(y_i|\hat{\theta}) \) 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 \( \hat{\text{elpd}_{\text{loo}}}, \nu(\text{elpd}_{\text{loo}}) \), and \( \hat{\sigma}_{\text{loo}}^2 \) 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\(_{\text{loo}}\) 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 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 \texttt{loo} R package (Vehântari 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 \(-\log(p(y_i|\theta))\) (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 \( \bar{\pi} \propto -\log(p(y_i|\theta)) \) 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 (\( \bar{\pi} \)) 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 \( \pi_i = -\log p(y_i|\theta) \).](image)

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). 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...
Bayesian LOO-CV for Large Data

<table>
<thead>
<tr>
<th>Data</th>
<th>Method</th>
<th>elpd_{loo}</th>
<th>SE(elpd_{loo})</th>
<th>σ_{loo}</th>
<th>k &gt; 0.7 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR(c) 100D</td>
<td>True</td>
<td>-14245</td>
<td>0</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPS(1)</td>
<td>-14231</td>
<td>21.6</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPS(2)</td>
<td>-14236</td>
<td>9.1</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SRS</td>
<td>-14309</td>
<td>2316.1</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>PPS(1)</td>
<td>-14240</td>
<td>7.1</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PPS(2)</td>
<td>-14242</td>
<td>2.3</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SRS</td>
<td>-14798</td>
<td>678.6</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>PPS(1)</td>
<td>-14245</td>
<td>2.4</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PPS(2)</td>
<td>-14246</td>
<td>0.7</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SRS</td>
<td>-13897</td>
<td>181.8</td>
<td>61</td>
</tr>
</tbody>
</table>

Table 1. Estimation of elpd_{loo} using posterior approximations. For all models and posterior approximations, σ_{loo} ≈ 70. No subsampling is used and MCMC is gold standard.

<table>
<thead>
<tr>
<th>Data</th>
<th>Method</th>
<th>m</th>
<th>n</th>
<th>PPS</th>
<th>SRS</th>
<th>σ_{loo}</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR(c) 100D</td>
<td></td>
<td>10</td>
<td>100</td>
<td>4.3</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>1000</td>
<td>2.3</td>
<td>360</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10000</td>
<td>10000</td>
<td>5.7</td>
<td>9351</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100000</td>
<td>100000</td>
<td>30</td>
<td>19715</td>
<td>225</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>100</td>
<td>1.3</td>
<td>9.7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>1000</td>
<td>1.2</td>
<td>83.7</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10000</td>
<td>10000</td>
<td>2.2</td>
<td>1144</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100000</td>
<td>100000</td>
<td>11.3</td>
<td>5894</td>
<td>225</td>
</tr>
</tbody>
</table>

Table 3. Standard errors, SE(elpd_{loo}), for PPS and SRS subsampling in relationship with σ_{loo}. The result are based on MCMC draws and θ is the posterior mean for the parameters. PPS is subsampling proportional to −log(p(y_i|θ)).

<table>
<thead>
<tr>
<th>Data</th>
<th>Method</th>
<th>M</th>
<th>Laplace</th>
<th>ADVI(FR)</th>
<th>ADVI(MF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR(c) 100D</td>
<td></td>
<td>1</td>
<td>0.24</td>
<td>0.23</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0.93</td>
<td>5.11</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>1.44</td>
<td>4.19</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>2.05</td>
<td>6.99</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td>5.62</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td></td>
<td>10.99</td>
<td>2.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>1.70</td>
<td>7.39</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Table 4. Posterior k̂ 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](https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations)
### Table 5.
The estimated $\text{elpd}_{\text{loo}}$ using a subsample of size $m = 500$ and its standard error (SE). The full $\text{elpd}_{\text{loo}}$ based on all observations is also included as well as $\text{elpd}_{10\text{fcv}}$, an estimation of the $\text{elpd}$ using 10-fold cross-validation. The $\sigma_{\text{loo}} \approx 90$ for all approximations and models. Less than 1% of the observations have problematic $\hat{k}$ using MCMC, making it a good gold standard.

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

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 $\sigma_{\text{loo}}$, 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 $\text{elpd}_{\text{loo}}$ for the different models we could compute the difference in $\text{elpd}_{\text{loo}}$ 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 ($10\text{fcv}$, see $\text{elpd}_{10\text{fcv}}$), 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 $\text{elpd}_{\text{loo}}$ 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 assess the posterior approximations and $v(\text{elpd}_{\text{loo}})$, 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


