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

Måns Magnusson MANS.MAGNUSSON@AALTO.FI
Michael Riis Andersen MICHAEL.ANDERSEN@AALTO.FI
Aki Vehtari AKI.VEHTARI@AALTO.FI
Aalto University

Abstract

Model inference, such as model comparison and model checking, is an important part of model development. Leave-one-out cross-validation (loo) is a general approach for model inference that can asses the generalizability of a model, but unfortunately, loo does not scale well to large datasets. We propose a combination of using approximate inference and probability proportional to size sampling for loo model inference for large data sets.

Keywords: Approximate inference, Leave-one-out cross-validation, Subsampling

1. Introduction

Model inference, such as model comparison, checking, and choice, is an integral part of developing new models. After a model has been estimated, we want to study aspects of how well it *generalizes* to unseen data (Box, 1976; Vehtari et al., 2012), commonly measured through *expected log predictive density* (elpd) for a new dataset, defined as

$$\text{elpd}_M = \sum_{i=1}^{n} \int \log p_M(\tilde{y}_i|y) p_t(\tilde{y}_i) d\tilde{y}_i \,,$$

where $p_t(\tilde{y}_i)$ is the true probability distribution for the underlying dataset and $\log p_M(\tilde{y}_i|y)$ is the log predictive density for a new observation for the model M.

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

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

Model comparison and model evaluation are important for model development, but little focus has been put into finding ways of scaling loo to larger datasets. Naively, estimating elpd_{loo} would need the fitting of n models, something that is unfeasible for large datasets. Pareto smoothed importance sampling (PSIS) has been proposed as a method to estimate elpd_{loo} based on the full data posterior distribution, henceforth called posterior (Vehtari et al., 2017; Geweke, 1989). However, this requires (1) MCMC samples from the true posterior and (2) the estimation of the elpd_{loo} contributions for all observations. Both these requirements can be costly in a data-rich regime (Gelfand, 1996; Vehtari et al., 2017).

In this paper, we focus on the problems of fast model inference for large data sets. We use well-known posterior approximations and extend PSIS-loo to these posterior approximations. We also propose sampling with probability proportional to size to estimate elpd_{loo}. The limitations with the approach are the same as using PSIS-loo (Vehtari et al., 2017) and that the approximate posterior need to be sufficiently close to the true posterior that we can use the approximation as a proposal distribution using PSIS.

2. Bayesian leave-one-out cross-validation for large data sets

Laplace and variational approximations are attractive for fast model comparisons due to their computational speed and scalability. We can use the approximate posterior distribution, $q_M(\theta|y)$, as a proposal distribution in an importance sampling scheme where our target distribution, $p_M(\theta|y_{-i})$, is the loo posterior for observation y_i, and our expectation of interest is the $\text{elpd}_{\text{loo},i}$, the elpd contribution for observation y_i. The (self-normalized) importance sampling estimator for elpd_{loo} can be expressed as

$$\widehat{\text{elpd}}_{\text{loo},i} = \log\left(\frac{\frac{1}{S}\sum_{s=1}^{S} p_M(y_i|\theta_s) r(\theta_s)}{\frac{1}{S}\sum_{s=1}^{S} r(\theta_s)}\right)$$

where θ_s are draws from the posterior approximation,S is the number of samples from the approximate posterior and

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

where the last step is a well-known result of Gelfand (1996). The two-factor approach is needed since the approximation $q_M(\theta)$ will not factorize with regard to observations. The factorization in Eq. (1) 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 have, in general, lighter tails than the proposal distribution that will increase the variance of the estimate (Geweke, 1989; Gelfand, 1996).

Recently, Pareto-Smoothed importance sampling has been proposed as a way to stabilize the weights in Bayesian loo (Vehtari et al., 2015) and in evaluating variational inference approximations (Yao et al., 2018). Here we combine these two previous results and use PSIS to correct for both the posterior approximation and to estimate the loo posterior, enabling fast estimation of the $\text{elpd}_{\text{loo},i}$ contributions to elpd_{loo}, with the additional benefit that we can use $\hat{k}$, the shape parameter in the generalized Pareto distribution, to diagnose how well the estimation is working (Vehtari et al., 2015).

As the data size increases, we can expect approximations such as Laplace approximation and variational inference to become better approximations of the true posterior distribution. This would improve the stability of the IS estimates, a result that can be seen empirically in Yao et al. (2018). Second, as the data size increase, we can also expect the loo posterior to come closer to the full posterior and hence the variability in the importance weights $p_M(y_i|\theta^s)^{-1}$ would also decrease with the size of the data.

Using PSIS we can estimate each $\text{elpd}_{\text{loo},i}$ term and sum them to estimate elpd_{loo}. Estimating every $\text{elpd}_{\text{loo},i}$ can be costly, especially as n grows. In some situations, estimating

elpd_loo can take even longer than computing the posterior, due to the computational burden of estimating $\hat{k}$. To handle this problem we suggest using a sample to estimate elpd_loo. Estimating totals, such as elpd_loo, has a long tradition in sampling theory. If we have an auxiliary variable that is proportional to the $\text{elpd}_{\text{loo},i}$ and known for all observations, we can use this in a probability proportional to size (PPS) scheme to reduce the sampling variance in the estimate of elpd_loo using the Hansen-Hurwitz (HH) estimator (Hansen and Hurwitz, 1943). In our case we can easily compute $\log p_M(y_i|y)$ and sample observations proportional to $p_i \propto |\log p_M(y_i|y)|$ with the HH estimator

$$\widehat{\text{elpd}}_\text{loo} = \frac{1}{m}\sum_{i=1}^{m}\frac{\text{elpd}_{\text{loo},i}}{p_i}\,,$$

where m is the sample size. The estimator has the (unbiased) variance estimator

$$v(\widehat{\text{elpd}}_\text{loo}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{\text{elpd}_{\text{loo},i}}{p_i} - \widehat{\text{elpd}}_\text{loo}\right)^2 .$$

The variance estimator shows the benefit of the PPS sampling scheme. Since we can expect the $|\text{elpd}_{loo,i} - \log p_M(y_i|y)| \to 0$ as $n \to \infty$, it holds that $v(\widehat{\text{elpd}}_\text{loo}) \to 0$ as $n \to \infty$ irrespective of the sample size m. The HH estimator of $\widehat{\text{elpd}}_\text{loo}$ is not limited to approximation methods, but can be used with MCMC as well. PPS sampling also has the benefit that we can use Walker-Alias multinomial sampling (Walker, 1977) to sample observations in constant time. Hence we can continue to sample observations until we have sufficient precision in $\widehat{\text{elpd}}_\text{loo}$ for our model comparison purposes. In a similar way we can also produce unbiased estimates of $\text{SE}(\text{elpd}_\text{loo})$.

Our approach for large-scale loo can be summarized in the following steps:

1. Estimate the models of interest using any posterior approximation technique.
2. Check that the approximation is close enough to the true posterior using the ideas of Yao et al. (2018) by evaluating $\hat{k}$. If the posterior approximation performs poorly, improve the approximation or use other approximation techniques.
3. Compute $\log p(y_i|y)$ for all n observations.
4. Sample m observations using PPS sampling and compute $\text{elpd}_{\text{loo},i}$ using PSIS. Use $\hat{k}_i$ to diagnose the estimation of $\text{elpd}_{\text{loo},i}$.
5. Estimate $\widehat{\text{elpd}}_\text{loo}$ and $\text{SE}(\widehat{\text{elpd}_\text{loo}})$ using Hansen-Hurwitz estimator and compare model performance. Repeat step 4 and 5 until sufficient precision is reached.

3. Experiments

To study the characteristic of the approach we focus on three datasets, one simulated dataset used to fit a Bayesian linear regression model with 100 variables and 10 000 generated observations with correlated (c) and independent (i) covariates, and the roaches dataset of Gelman and Hill (2006), a dataset with only 262 observations. The latter data set is known to contain severe outliers when a Poisson regression model is estimated and is an example of a misspecified model. We use mean-field ADVI (Kucukelbir et al., 2017) and Laplace approximations implemented in Stan (Carpenter et al., 2017) as posterior approximations.

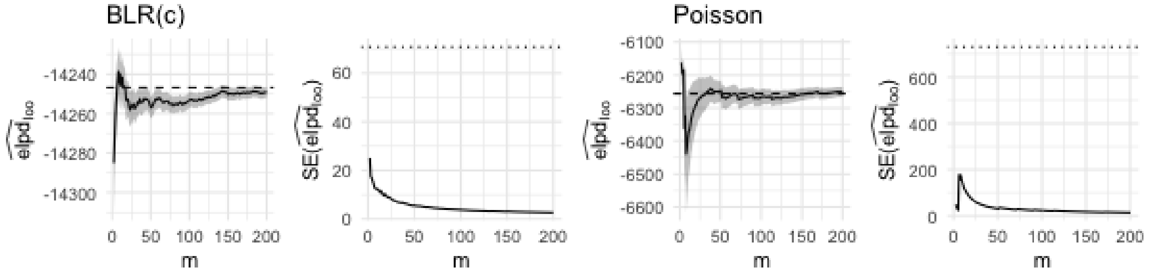
Here we study if the PSIS approach to estimate the elpd works correctly with and without (*) approximation correction and the elpd estimation using the HH estimator. Table 1 contains the estimation of elpd_{loo} for the different models with and without the corrections for the posterior approximation. From the table, we can see that using PSIS and posterior approximations to estimate elpd_{loo} is surprisingly robust.

Data	Estimate	MCMC	Laplace	Laplace(*)	ADVI	ADVI(*)
BLR(c)	elpd_{loo}	-14247	-14246	-14247	-14258	-14256
	$\text{SE}(\text{elpd}_{\text{loo}})$	70.7	70.7	71.4	70.5	70.7
	$\hat{k} > 0.7$ (%)	0.0	0.0	0.0	100.0	0.0
BLR(i)	elpd_{loo}	-14239	-14239	-14239	-14241	-14248
	$\text{SE}(\text{elpd}_{\text{loo}})$	71.2	71.1	71.9	71.2	71.1
	$\hat{k} > 0.7$ (%)	0.0	0.0	0.0	0.0	0.0
Poisson	elpd_{loo}	-6257	-6257	-6257	-6241	-6265
	$\text{SE}(\text{elpd}_{\text{loo}})$	730.7	730.6	730.3	726.5	728.9
	$\hat{k} > 0.7$ (%)	7.3	5.3	5.0	78.2	6.9

Table 1: elpd_{loo} and $\hat{k}$ diagnostics for different models and posterior approximations with and without (*) importance sampling correction of posterior approximation.

Figure 1 shows the HH estimate as function of subsample size m and it is clear that in this situation a relatively small sample size of $m = 50$ are sufficient to estimate elpd_{loo} with an uncertainty of the estimate that is only a fraction of the total $\text{SE}(\text{elpd}_{\text{loo}})$.

Figure 1: Estimation of elpd_{loo} using Hansen-Hurwitz estimator. The dashed line is the true elpd_{loo} value and the dotted line is the true $\text{SE}(\text{elpd}_{\text{loo}})$ value. The gray area show 1 $\text{SE}(\widehat{\text{elpd}}_{\text{loo}})$ from the point estimate (black).



4. Conclusions

This work shows promising improvements to enable fast and efficient model comparison and model evaluations for large datasets. The approach can be improved further to by adapting the posterior approximations to better suit the purpose of being used for importance sampling estimation of elpd_{loo}.

Acknowledgments

The research was funded by Academy of Finland (grants 298742, 313122).

References

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

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

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.

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.

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

Morris H. Hansen and William N. Hurwitz. On the theory of sampling from finite populations. *Ann. Math. Statist.*, 14(4):333–362, 12 1943. doi: 10.1214/aoms/1177731356. URL https://doi.org/10.1214/aoms/1177731356.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

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

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

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.

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

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5581–5590, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/yao18a.html.