Fast cross-validation for Bayesian inference using proposals on a linear subspace

Mohamed Tarek [1,2], Manu Francis [3], Anastasios Panagiotelis [2]

1) Pumas-AI Inc., USA; 2) Business School, University of Sydney, Australia; 3) Formerly at Pumas-AI Inc., USA

MOTIVATION

In the Bayesian workflow [1], it is common to evaluate and compare models using their predictive power for out-of-sample data. Reviews of Bayesian model validation [2,3] as well as the literature on scoring rules [4] motivate using the expected log predictive density (ELPD) for model evaluation. Estimates of the ELPD can be obtained by running Markov chain Monte Carlo (MCMC) multiple times in cross validation schemes. One downside in the Bayesian setting is that MCMC is computationally intensive and must be repeated every for every fold of cross validation. This poster proposes the use of subspace inference as a means for reducing the computational demands of Bayesian cross-validation (CV).

Let M be a model with p parameters $\theta \in \Theta$ that describes the data generating process of the observed data y. Let y_i be the i^{th} of N observations, y_{-i} be all data excluding y_i, and $\theta_{-i}^{[j]}$ be the j^{th} of M draws from the posterior $\mathrm{p}(\theta \mid y = y_{-i}, M)$. The leave-one-out (loo) cross-validation (CV) estimate of the ELPD is given by:

$$ELPD = \frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{M}\sum_{j=1}^{M} p\left(y_i \mid \theta = \theta_{-i}^{[j]}, M\right)$$

The pointwise term inside the outer summation is often called the pointwise ELPD.

Evaluating the above expression exactly is expensive since one needs to draw samples from N different posteriors. Typically this will be done by Markov chain Monte Carlo Methods (MCMC), e.g. the Hamiltonian Monte Carlo (HMC) [5] methods, which are computationally expensive. One approach to overcome this difficulty, is the Pareto-smoothed importance sampling method for leave-one-out, cross-validation (PSIS-LOO-CV), proposed by [6]. In PSIS-LOO-CV, MCMC is run once on the full data. The same samples are then re-used in each iteration of CV but using different weights. Additionally, the shape parameter of the Pareto distribution fitted to the raw weights is an important diagnostic that can be used to find influential data points and indicate when the PSIS-LOO-CV estimate of the ELPD is unreliable. If the shape parameter exceeds 0.7, [6] recommends not trusting the ELPD estimates and instead re-sampling from the posterior. Unfortunately, in practice, the shape parameters in PSIS-LOO-CV can often be greater than 0.7 especially when there are only a few data points or many parameters that are sensitive to specific points.

PROPOSED ALGORITHM

We propose an alternative cross-validation method based on the idea of subspace inference [7,8]. In the proposed methodology, we first run MCMC on the full dataset. We then perform dimension reduction via principal components analysis (PCA) on the MCMC samples. This finds a low-dimensional subspace where most of the variance of the posterior lies. Assume the sample covariance matrix of the posterior samples $\theta^{[1]} \ldots \theta^{[M]}$ is Σ where $\theta^{[j]} \sim p(\theta \mid y, M)$. Let the eigenvalue decomposition of the covariance matrix be:

$$\Sigma = U\,\Lambda\,U^T$$
$$U = \begin{bmatrix} U_{\parallel} & U_{\perp} \end{bmatrix}$$

where U is the $p \times p$ matrix of eigenvectors, p is the number of parameters in θ, Λ is the diagonal matrix of eigenvalues sorted in descending order, $U_{\parallel}$ is the first $d < p$ columns of U, and $U_{\perp}$ is the remaining $p - d$ columns of U. We then re-parameterise the model in terms of ξ such that:

$$\xi = U^T \cdot (\theta - \hat{\theta})$$
$$\begin{bmatrix} \xi_{\parallel} \\ \xi_{\perp} \end{bmatrix} = \begin{bmatrix} U_{\parallel}^T \cdot (\theta - \hat{\theta}) \\ U_{\perp}^T \cdot (\theta - \hat{\theta}) \end{bmatrix}$$

where $\hat{\theta}$ is the posterior mean approximated from the MCMC samples. Each time an observation is left out in cross-validation, MCMC sampling is re-run with the subset of data. However, rather than drawing samples from the full posterior during each MCMC run, we draw samples from a low dimensional approximation of the posterior.

This is achieved by making proposals of $\xi_{\parallel}$, which are constrained to lie in the subspace spanned by $U_{\parallel}$ rather than the full-dimensional parameter space. Prior and likelihood calculations needed for inference can be made through back-transforming from $\xi_{\parallel}$ to θ by computing:

$$\theta = f(\xi_{\parallel}) = U_{\parallel} \cdot \xi_{\parallel} + \hat{\theta}$$

The low dimensional approximation is denoted by $\tilde{p}(\theta \mid y = y_{-i})$ and the full procedure is:

1. Draw samples $\theta^{[1]} \ldots \theta^{[M]}$ from the posterior given all the data, $\theta^{[j]} \sim p(\theta \mid y, M)$.
2. Compute the mean estimate $\hat{\theta}$ and covariance matrix estimate Σ.
3. Find the eigenvectors U of the covariance matrix estimate and define $U_{\parallel}$.
4. Partition the data into k training-validation splits.
5. For each split, perform subspace MCMC to draw samples from the posterior $\tilde{p}(\theta \mid y = y_{-i})$ given the training data. Proposals are made in the d-dimensional linear subspace of $\xi_{\parallel}$ and then transformed to a proposal in the p-dimensional vector space of θ using $\theta = f(\xi_{\parallel})$.
6. Estimate the ELPD using the new posterior samples for each training-validation split.

One advantage of the proposed algorithm is that the model and sampler do not need to be altered. The only requirement is that the argument of the log probability function is written in terms of $\xi_{\parallel}$ rather than θ, the functional form of the likelihood and prior remain unchanged. While the derivation above assumed loo CV, the same algorithm can be applied to leave-future-out or leave-k-out CV as we demonstrate in the experiments.

RESULTS AND DISCUSSION

The experiments were run using the Pumas software. The test model used was a pharmacokinetic (PK) model with 2 depot compartments, 1 central compartment and linear absorption and clearance. The model had 13 fixed effects including between subject variability and 3 random effects per subject. A synthetic population of 12 subjects and 11 observations per subject was simulated and used for the inference. Leave-future-1-observation CV was run using the re-running CV, PSIS CV, and the subspace CV methods. No less than 4 observations per subject was allowed in each CV run. The following table shows the maximum mean discrepancy (MMD) comparing the pointwise ELPD using the subspace CV and PSIS CV methods to the “re-running inference” CV method. A subspace size of 5 was used in the subspace CV method. The running time is also shown. As shown in the table below, the proposed subspace CV method could achieve more than 500 times lower MMD than PSIS CV using approximately 1% of the running time of the re-running method.

Table 1: Comparison of proposed and existing CV methods.

	Re-running Inference CV	PSIS CV	Subspace CV
MMD	0	1.32061	0.00178
Time (sec)	1158.40	1.01	12.47

REFERENCES

[1] Gelman, A., Vehtari, A., Simpson, D., Margossian, C.C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., Modrák, M.: Bayesian Workflow. arXiv:2011.01808 [stat] (2020). arXiv: 2011.01808.

[2] Vehtari, A., Ojanen, J.: A survey of Bayesian predictive methods for model assessment, selection and comparison (2012)

[3] Piironen, J., Vehtari, A.: Comparison of Bayesian predictive methods for model selection. Statistics and Computing 27, 711–735 (2017)

[4] Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102(477), 359–378 (2007)

[5] Neal, R.M., et al.: MCMC using Hamiltonian dynamics. Handbook of Markov chain Monte Carlo 2(11), 2 (2011)

[6] Vehtari, A., Simpson, D., Gelman, A., Yao, Y., Gabry, J.: Pareto Smoothed Importance Sampling. arXiv:1507.02646 [stat] (2021). arXiv: 1507.02646.

[7] Izmailov, P., Maddox, W.J., Kirichenko, P., Garipov, T., Vetrov, D., Wilson, A.G.: Subspace inference for bayesian deep learning. In: Uncertainty in Artificial Intelligence, pp. 1169–1179 (2020). PMLR

[8] Manu Francis, Vijay Ivaturi, Mohamed Tarek: Subspace MCMC algorithm for Bayesian parameter estimation of hierarchical PK/PD models in Pumas. In: Population Approach Group in Europe (2021).