Make cross-validation Bayes again

Yuling Yao
Flatiron Institute
New York, USA.
yyao@flatironinstitute.org

Aki Vehtari
Aalto University
Espoo, Finland.
aki.vehtari@aalto.fi

Abstract

There are two orthogonal paradigms for hyperparameter inference: either to make a joint estimation in a larger hierarchical Bayesian model or to optimize the tuning parameter with respect to cross-validation metrics. Both are limited: the "full Bayes" strategy is conceptually unjustified in misspecified models; The cross-validation strategy, besides its computation cost, typically results in a point estimate, ignoring the uncertainty in hyperparameters. To bridge the two extremes, we present a general paradigm: a full-Bayes model on top of the cross-validated log likelihood. This prediction-aware approach incorporates additional regularization during hyperparameter tuning, and facilities Bayesian workflow in many otherwise black-box learning algorithms. We develop theory justification and discuss its application in a model averaging example.

1 The problem

Consider a model with observations $y = (y_1, \dots, y_n)$, parameters β, and hyperparameters τ. We distinguish the parameters and hyperparameters such that y is conditionally independent of τ given β in the belief model. For brevity, we suppress notation dependency on covariates. We assume that data are conditionally exchangeable given parameters. Some familiar examples include multilevel models, the lasso and ridge regression, and Gaussian process regression. There are various ways to infer hyperparameters:

- Maximum-a-posteriori (MAP) estimate: the naive estimate seeks the joint mode of the posterior distribution $p(\beta, \tau|y) = p(y|\beta)p(\beta|\tau)p(\tau)$. It is often meaningless.
- Type-II MAP estimate: Instead of the meaningless joint mode, the marginal posterior mode $\hat{\tau} = \arg\max p(\tau|y)$ is one plausible point estimate.
- Full hierarchical Bayes. From a Bayesian point of view, MAP estimates are often criticized for ignoring uncertainty, and the full Bayes "gold" standard of hyper parameter inference is the posterior marginal distribution, $p_{\text{Bayes}}(\tau|y) = \int p(\tau, \beta|y)d\beta$.
- Cross-validation optimization: Yet another criticism against many MAP estimates is over-fitting. To replace empirical risks with some estimate of the excepted risk, we often rely on cross-validation (CV, Stone, 1974). For a fixed τ, we make inference on $p(\beta|y, \tau)$. With an additionally specified scoring rule $S(\cdot, p(\cdot|\tau))$ of the predictive distribution, we could evaluate the posterior prediction $\mathbb{E}_{y_{n+1}} S(y_{n+1}, p(y_{n+1}|\tau, y))$ by integrating out future unseen date y_{n+1}. To be concrete, consider the log score and leave-one-out (LOO) cross-validation. The expected log predictive density (elpd, Gelman et al., 2014) is estimated by

$$\text{elpd}_{\text{loo}}(y|\tau) \coloneqq \sum_{i=1}^{n} \log \int p(y_i|\beta, \tau)p(\beta|y_{-i}, \tau)d\beta, \tag{1}$$

  where y_{-i} stands for the leave-i-th-point-out dataset. In practice, we can choose the best parameter that maximizes the CV score. $\hat{\tau} = \arg\max_\tau \text{elpd}_{\text{loo}}(y|\tau)$.

NeurIPS 2021 Workshop Your Model Is Wrong: Robustness and Misspecification in Probabilistic Modeling.

Which approach is the best? The literature (Sundararajan and Keerthi, 2001; Qi et al., 2004; Bachoc, 2013; Bachoc et al., 2017) suggests that the answer often relies on the amount of model misspecification: The type-II MAP has a smaller variance, and integrating over the posterior can improve over a point estimate; But in case of model misspecification, the type-II MAP has a bigger bias while the CV-optimization is better. We outline the limitations of existing approaches as follows.

Decision theory can be noisy. The CV-optimization follows the Bayesian decision theory (Berger, 1985), where the action space is the support of τ, and we integrate out all remaining parameters β to obtain the expected utility. When using CV (1) to compute out-of-sample predictive utilities, the belief model for future data is not explicitly defined, but instead samples representing the distribution are assumed to be available, hence is an $\mathcal{M}$-open treatment in words of Vehtari and Ojanen (2012).

Despite the asymptotic optimality as ensured by the Bayesian decision theory, the point value of the hyperparameter that maximizes CV utility is noisy. Various heuristics have been proposed to handle the finite-sample noise, such as the "one standard error rule", while the standard error of CV itself is often infeasible to estimate in the first place (Bengio and Grandvalet, 2004; Sivula et al., 2020). Preferring a parsimonious model can be viewed as a prior regularization, but when encountering a high or infinite dimensional hyperparameter τ, it becomes cumbersome to generalize such heuristics to incorporate an appropriate amount of regularization. Lastly, the post-inference model evaluation becomes difficult. As we have already optimized the CV utility, the plug-in elpd at the optimum $\text{elpd}_{\text{loo}}(y|\hat{\tau})$ overestimates the expected out-of-sample predictive utility. We may use double cross-validation, but it squares the training time for each given τ.

Bayesian inference may be meaningless. The hierarchical Bayesian inference $p_{\text{Bayes}}(\tau|y)$ is nearly the opposite counterpart of cross-validation for it remedies all aforesaid defects: the Bayesian posterior seemingly summarizes the uncertainty of hyperparameters in a prima facie coherent way; It is straightforward to incorporate prior regularization $p(\tau)$ by Bayesian (non-)parametric techniques; The post-inference evaluation follows the vanilla Bayesian workflow (Gelman et al., 2020), in which we can make model-comparison, posterior predictive check and CV-based model evaluation, without the need to reinvent them on a case-by-case basis (e.g., Lee et al., 2016).

However, the standard Bayesian paradigm is to define a joint distribution of all observed and unobserved quantities. It is meaningless in a misspecified model (Gelman and Yao, 2020). That said, Bayesian inference on parameters can still be useful in a wrong model (Berk, 1966): as $n \to \infty$, the posterior distribution of parameters $p(\beta|y)$ is almost surely supported on the set on which the Kullback–Leibler (KL) divergence is minimized between the predictive distribution and the true data generating process: $\text{KL}\big(p_{\text{true}}(y)||p(y|\beta)\big)$.

Yet Bayesian inference on hyperparameter τ is even more meaningless in misspecified models. There is no such joint sampling distribution on hyperparameters; Even when $n \to \infty$, $p_{\text{Bayes}}(\tau|y)$ does not necessarily concentrate; When it does, it does not necessarily converge to an optimal point.

	whether to cross validate	
point or	MAP	CV optimization
probabilistic	"full" hierarchical Bayes	☺ the present paper

Table 1: *The main idea—To complement hierarchical Bayes and cross-validation (CV), we treat CV log predictive density as a log likelihood, and fold it into a Bayesian model.*

2 Cross-validated Bayesian inference

The idea. Rather than racing CV against Bayes, in this paper we combine both strategies. We view $\text{elpd}_{\text{loo}}(y|\tau)$ in Eq. 1 as a log likelihood function, and define a log posterior density on τ as:

$$\log \tilde{p}(\tau|y) \coloneqq \text{elpd}_{\text{loo}}(y|\tau) + \log p(\tau) + \text{constant}. \tag{2}$$

We call this $\tilde{p}(\tau|y)$ "cross-validated Bayesian inference", or CV-Bayes. Before we elaborate on the theory justification and practical implementation, we highlight that, as a bridge to complement the two composition elements, the benefit of our proposal is two-fold:

- Provide a "safety net" to regular Bayesian inference amid model misspecification.
- Enable regularization, uncertainty quantification, and post-inference evaluation in cross-validation.

2.1 Why is it a legitimate posterior inference

Cross-validation itself is a model: the distribution of next unseen future observation is modeled by the empirical CV distributions:

$$y_{n+1}|y \overset{d}{\approx} y_i|y_{-i}. \tag{3}$$

CV has been widely used during the model evaluation phase. The idea behind CV-Bayes is to augment the otherwise-always-wrong parametric belief model $p(y, \beta, \tau)$ with such an approximately-always-correct layer (3) during the inference phase.

To that end, we consider a two-stage training procedure. In the first stage, we only make inference on β conditional on τ and y using the parametric belief model. It follows the regular Bayesian posterior density $p(\beta|y, \tau)$.

In the second stage, we want to avoid using-data-twice. If there exist an extra hold-out dataset $\tilde{y}$ which is identically distributed and of the same sample size as y, then the resulting likelihood reads

$$p(\tilde{y}|\tau, y) = \int p(\tilde{y}|\beta, \tau)p(\beta|y, \tau)d\beta. \tag{4}$$

Now in lack of the hold-out dataset, we adopt a data augmentation (Meng and van Dyk, 1999) view, and integrate out $\tilde{y}$ in the likelihood (4) using the CV model (3). The expected log likelihood becomes

$$\mathbb{E}_{\tilde{y}} \log p(\tilde{y}|\tau, y) \approx \text{elpd}_{\text{loo}}(y|\tau). \tag{5}$$

The approximation is accurate when the sample size is large, because the leave-one-out log score is a consistent estimate of the expected out-of-sample log scoring rule (e.g., Le and Clarke, 2017).

The final inference on τ follows the usual Bayes rule. Plugin the right-hand side of data-augmented log likelihood (5) and a prior $p(\tau)$, we arrive at the CV-Bayes log posterior density (2).

As for a comparison, if we use the same belief model to integrate out $\tilde{y}$ in the second-stage augmented likelihood (4), then $\mathbb{E}_{\tilde{y}}\, p(\tau|y, \tilde{y}) = \int p(\tau|y, \tilde{y})p(\tilde{y}|y)d\tilde{y} = p_{\text{Bayes}}(\tau|y)$, such that data augmentation does not affect the regular Bayesian inference.

Parameter grouping. At the beginning of the paper, we divide parameters and hyperparameters per the routine of conditional independence. Such conditional independence is not necessary for our derivation of CV-Bayes, as we have deliberately kept the dependence on τ in the key identities (1) and (4). More generally, we can divide all parameters (including hyperparameters) into two groups: β and τ, then make CV-Bayes inference on τ using (2). If β is empty and all parameters are grouped into τ, then $\text{elpd}_{\text{loo}}(y|\tau) = p(y|\tau)$. Hence, the regular Bayesian statistics is recovered as a special case of the CV-Bayes framework. Dividing a model into sub-modules and making separate inferences are related to the idea of "cutting feedback" (Lunn et al., 2009; Jacob et al., 2017).

Scoring rules. Another extension is to consider other scoring rules. The log score, apart from being the only strictly local proper continuous scoring rule, is also critical in the data-augmentation justification (5). For a general scoring rule $S(\cdot, p(\cdot))$, the log CV-likelihood (1) is replaced by the CV scoring rule $\sum_{i=1}^{n} S(y_i, p(y_i|y_{-i}, \tau))$, with an additional power transformation γ for calibration. The CV-Bayes log posterior of τ is then defined by $\gamma \sum_{i=1}^{n} S(y_i, p(y_i|y_{-i}, \tau)) + \log p(\tau)$. The detailed discussion is beyond this paper.

It is not a new idea to connect a loss function with Bayesian inference. Bissiri et al. (2016) considered the "general Bayesian update", where the log likelihood is defined by loss functions of parameter estimates. CV-Bayes differs from this general Bayesian update in the additional CV step and the scoring rules of outcomes, both making the inference more aware of the out-of-sample prediction. From a predictivist Bayesian point of view, the main interest in statistical inference is about observable quantities such as the future observation, rather than parameter estimation.

2.2 Practical implementation

The largest obstacle in our approach is to evaluate the log likelihood (5), for it involves CV conditional on τ. The brute-force evaluation is only feasible if sample size n is small and τ is supported on a finite set. Nevertheless, there are two useful cases when closed-form likelihood is available.

First, many regression models, including Gaussian processes, admit a conjugate conditional model $\beta|\tau, y$, and a closed form conditional leave-one-out predictive density $p(y_i|y_{-i}, \tau)$. Hence, the log CV likelihood (5) is a closed form expression of τ.

Second, sometimes we either prefer or choose to conduct two stage estimation, such as in causal inference (McCandless et al., 2010), and model averaging (Section 3). In these models, the parameters are divided in such a way that $p(\beta|y, \tau)$ does not depend on τ in the belief model. Hence, we only need to train the model once and obtain $p(\beta|y)$, then we can use Pareto smoothed importance sampling (Vehtari et al., 2017, 2020) to make an efficient leave-one-out approximation of $p(y_i|y_{-i}, \tau)$.

In these two cases, $\text{elpd}_{\text{loo}}(y|\tau)$ is analytic. We use a generic MCMC sampler, such as Stan, to draw posterior samples from the CV-Bayes posterior $\tilde{p}(\tau|y)$, denoted by $(\tau_1, \ldots, \tau_S)$. Moreover, we can compute the post-inference out-of-sample log predictive density by another importance sampling. It only requires to compute the Pareto-smoothed importance ratios $r_{is} \approx \left(\exp(\text{elpd}_{\text{loo}}(y_i|\tau_s))\right)^{-1}$, then the out-of-sample log predictive density is approximated by importance sampling:

$$\sum_{i=1}^{n} \int \text{elpd}_{\text{loo}}(y_i|\tau)\tilde{p}(\tau|y_{-i})d\tau \approx \sum_{i=1}^{n} \log \frac{\sum_{i=1}^{S}(r_{is}\exp(\text{elpd}_{\text{loo}}(y_i|\tau_s)))}{\sum_{s=1}^{S} r_{is}}.$$

Although it is conceptually a double CV, in the whole process we only need to train the model once. This evaluation step helps select or average multiple priors $p(\tau)$ or different models.

Third, when there is no closed-form available but τ is discrete or has a low dimension, we can train $p(\beta|\tau, y)$ on a fixed grid $(\tau_1, \ldots \tau_S)$ and obtain the conditional CV, $\log p(y_i|\tau_s, y_{-i})$, by the same Pareto smoothed importance sampling approximation. In total, we train the model S times to evaluate $\tilde{p}(\tau|y)$ on the grid. The full posterior distribution is approximated by quadrature.

3 Application: what we have learned from model averaging

Parameter inference and model averaging/selection methods are connected in both directions. On one hand, if the parameter space is discrete such that each parameter value maps to a model, then making probabilistic inference is equivalent to model averaging: Bayesian inference becomes Bayesian model averaging (BMA, Hoeting et al., 1999) of parameters, while CV-Bayes becomes pseudo-BMA[1] (Yao et al., 2018). The present paper was motivated by the recent progress in model averaging methods—If BMA has been known to be an unsatisfactory tool for model averaging amid model misspecification, then shouldn't we do better than Bayesian parameter inference in general?

On the other hand, model averaging and selection methods can be derived from applying various inference paradigm to an encompassing model that includes individual models as special cases (Tab. 2). Given K models, each containing parameters $\theta_1, \ldots, \theta_K$, use our notation and let $\tau := (w_1, \ldots, w_K)$ be model weights and $\beta = (\theta_1, \ldots, \theta_K)$ be other model specific parameters. The larger encompassing model is $p(y|\beta, \tau) = \sum_k w_k p_k(y|\theta_k)$. If the support of τ is binary: $\{w_k \in \{0, 1\}, \sum_k w_k = 1\}$, and we apply Bayes/CV-Bayes to (τ, β) in the encompassing model, then we obtain BMA/peudso-BMA. If the support of τ is a simplex: $\{0 \le w_k \le 1, \sum_k w_k = 1\}$, and we apply CV-optimization/CV-Bayes to the encompassing model, then we get stacking/hierarchical stacking. In light of this derivation, CV-Bayes has already been successfully testified in the wild. The extensive simulations in Yao et al. (2018, 2021) on the comparison between pseudo-BMA and BMA, and between hierarchical stacking and no-pooling stacking, have indicated the advantage of CV-Bayes against hierarchical Bayes, and CV-MAP, respectively.

inference \ support	binary	simplex
MAP	model selection by marginal likelihood	stacking framed in Wolpert (1992)
CV-optimization	model selection by CV	stacking (Yao et al., 2018, 2020)
Bayes	Bayesian model averaging (BMA)	
CV-Bayes	pseudo-BMA	hierarchical stacking (Yao et al., 2021)

Table 2: *Various model selection and averaging methods can be viewed as the combinations of four inference paradigms and two settings of the parameter support.*

[1] One caveat: If τ is continuous, and we only evaluate the conditional CV on a discrete τ grid, then we should use quadrature as described in Sec. 2.2, rather than average theses discrete τ using pseudo-BMA.

References

Bachoc, F. (2013). Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66:55–69.

Bachoc, F., Lagnoux, A., and Nguyen, T. M. N. (2017). Cross-validation estimation of covariance parameters under fixed-domain asymptotics. *Journal of Multivariate Analysis*, 160:42–67.

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

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer, New York.

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, pages 51–58.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B*, 78(5):1103.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian workflow. *arXiv:2011.01808*.

Gelman, A. and Yao, Y. (2020). Holes in Bayesian statistics. *Journal of Physics G*, 48(1).

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–417.

Jacob, P. E., Murray, L. M., Holmes, C. C., and Robert, C. P. (2017). Better together? Statistical learning in models made of modules. *arXiv:1708.08719*.

Le, T. and Clarke, B. (2017). A Bayes interpretation of stacking for M-complete and M-open settings. *Bayesian Analysis*, 12(3):807–829.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., and Neuenschwander, B. (2009). Combining MCMC with 'sequential' PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics*, 36(1):19–38.

McCandless, L. C., Douglas, I. J., Evans, S. J., and Smeeth, L. (2010). Cutting feedback in Bayesian regression adjustment for the propensity score. *The International Journal of Biostatistics*, 6(2).

Meng, X.-L. and van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, 86:301–320.

Qi, Y., Minka, T. P., Picard, R. W., and Ghahramani, Z. (2004). Predictive automatic relevance determination by expectation propagation. In *International Conference on Machine Learning*.

Sivula, T., Magnusson, M., and Vehtari, A. (2020). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv:2008.10296*.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B*, 36(2):111–133.

Sundararajan, S. and Keerthi, S. S. (2001). Predictive approaches for choosing hyperparameters in Gaussian processes. *Neural Computation*, 13(5):1103–1118.

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

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

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2020). Pareto smoothed importance sampling. *arXiv:1507.02646*.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259.

Yao, Y., Pirš, G., Vehtari, A., and Gelman, A. (2021). Bayesian hierarchical stacking: Some models are (somewhere) useful. *Bayesian Analysis*, in press.

Yao, Y., Vehtari, A., and Gelman, A. (2020). Stacking for non-mixing bayesian computations: The curse and blessing of multimodal posteriors. *arXiv:2006.12335*.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1007.