Bayesian Leave-One-Out Cross-Validation

Chap. 6 describes an implementation shortcut for leave-one-out cross-validation for linear smoothers that requires running through the training data only once. This section describes Bayesian analogues to this idea that are not restricted to linear smoothers. Relevant background literature is Rubin (1988), Smith and Gelfand (1992), Gelfand (1996), Albert (2009) and Gianola and Schön (2016).

In a general setting the leave-one-out validating mean squared error is equal to

$$MSE_v = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widehat{y}_i\right)^2, \tag{1}$$

where N is the total number of records and y_i is the ith observed record.

A first approach that accounts for posterior uncertainty about the parameters of the Bayesian model and for sampling variation of an individual record, is to draw a predicted value $\hat{y}_i$ from the posterior predictive distribution with density

$$p\left(y_i|y_{-i}\right) = \int p\left(y_i|\theta, y_{-i}\right) p\left(\theta|y_{-i}\right) d\theta, \tag{2}$$

where often, $p\left(y_i|\theta, y_{-i}\right) = p\left(y_i|\theta\right)$ assuming that given θ, y_i is independent of y_{-i}. In the expression above, θ is the vector of parameters of the Bayesian model and the vector of complete data y is $y = (y_i, y_{-i})$, where y_{-i} is y excluding the ith datum. From a Bayesian perspective inferences are conditional on the observed data y and therefore the only stochastic term in (1) is $\widehat{y}_i$. Since MSE_v is a transformation of $\widehat{y}_i$, (1) defines a posterior predictive distribution. An estimate of this distribution can be obtained using McMC: once the Markov chain has converged, vector $\left(\widehat{y}_1^{(t)}, \widehat{y}_2^{(t)}, \ldots, \widehat{y}_N^{(t)}\right)$, $t = 1, 2, \ldots, T$, drawn from the posterior predictive distribution at cycle t of a Markov chain of length T is used to compute $MSE_v^{(t)}$. This single draw $MSE_v^{(t)}$ constitutes one approximate sample from the marginal posterior distribution of MSE_v. A Monte Carlo estimate of the complete marginal posterior distribution of MSE_v is constructed using the collection of T samples.

The computing protocol to obtain a sample from the posterior predictive distribution of (1) is straightforward; at cycle t of an McMC sampler that has reached convergence to its stationary distribution, for datum i,

- Draw $\theta^{(t)}$ from $[\theta|y_{-i}]$,
- Given $\theta^{(t)}$, draw $\widehat{y}_i^{(t)}$ from $\left[y_i|\theta^{(t)}\right]$ if y_i is continuous or generate $\widehat{y}_i^{(t)}$ from $\Pr\left(Y_i = y_i|\theta^{(t)}\right)$ if Y is binary,
- Compute $MSE_v^{(t)} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widehat{y}_i^{(t)}\right)^2$.

This is repeated for $i = 1, 2, \ldots, N$. When the Markov chain has converged to the posterior distribution, each vector $\left(\theta^{(t)}, \widehat{y}_i^{(t)}\right)$ is an approximate draw from $[\theta, y_i|y_{-i}]$ and the coordinates are approximate draws from $[\theta|y_{-i}]$ and $[y_i|y_{-i}]$. Further, $MSE_v^{(t)}$ is a draw from the posterior predictive distribution of MSE_v. The computation of this estimator of the leave-one-out validating mean squared error is shown in the R-code below, `sumloo` and `mselooBayes`.

A second approach is to construct a Bayesian leave-one-out validating mean squared error that incorporates posterior uncertainty about the parameters of the model. This is achieved by defining the predictor $\hat{y}_i^{(t)}$ as a function of $\theta^{(t)}$ after the first step in the protocol. For example, in the Gaussian linear regression model where $\theta = (b, \sigma^2)$, $\hat{y}_i^{(t)} = x_i' b^{(t)}$, where b is a vector of unknown regression parameters and x_i is an observed vector of covariates. In a probit model for the analysis of binary records, $\theta = b$ and $\hat{y}_i^{(t)}$ can be constructed using Bayes rule $I\left[\Phi\left(x_i' b^{(t)}\right) > 0.5\right]$, where $I[\cdot]$ is the indicator function that takes the value 1 if the argument is satisfied, and 0 otherwise, and $\Phi(\cdot)$ is the standard normal integral. The computation of this estimator of the leave-one-out validating mean squared error is shown in the R-code below, `sumXb` and `mselooXb`.

These approaches can be applied to a range of models of different complexity, including hierarchical models for analysis of binary and categorical data. *The challenge is to obtain draws from the distribution* $[\theta|y_{-i}]$, $i = 1, 2, \ldots, N$, *running through the data only once rather than N times.* A useful option is to use importance sampling (see Chap. 4). In order to obtain approximate draws from the distribution $[\theta|y_{-i}]$ using the available collection of samples $\theta^* = \left(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}\right)$, each from the distribution $[\theta|y]$, first calculate the importance ratio

$$w_{ij} = \frac{p\left(\theta^{(j)}|y_{-i}\right)}{p\left(\theta^{(j)}|y\right)},\ i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, T, \tag{3}$$

and then resample with replacement from the vector θ^* with probabilities proportional to w_{ij}. This is a weighted bootstrap where samples with replacement from θ^* are drawn with unequal probabilities $w_{i1}, w_{i2}, \ldots, w_{iT}$. The resulting vector is an approximate sample from the distribution $[\theta|y_{-i}]$ (Gelfand, 1996). The importance ratio (3) takes the following

simple form:

$$\begin{aligned}
\frac{p\left(\theta^{(j)}|y_{-i}\right)}{p\left(\theta^{(j)}|y\right)} &= \frac{p\left(y_{-i}|\theta^{(j)}\right)p\left(\theta^{(j)}\right)\Big/\,p\left(y_{-i}\right)}{p\left(y|\theta^{(j)}\right)p\left(\theta^{(j)}\right)\Big/\,p\left(y\right)} \\
&= \frac{p\left(y_{-i}|\theta^{(j)}\right)}{p\left(y_i, y_{-i}|\theta^{(j)}\right)}\frac{p\left(y\right)}{p\left(y_{-i}\right)} \\
&= \frac{1}{p\left(y_i|y_{-i}, \theta^{(j)}\right)}\frac{p\left(y\right)}{p\left(y_{-i}\right)} \\
&\propto \frac{1}{p\left(y_i|\theta^{(j)}\right)} \qquad (4)
\end{aligned}$$

that holds also for probability mass functions associated with discrete data. The final step assumes that y_i and y_{-i} are conditionally independent, given θ. A vector with T elements representing probabilities (that sum to 1) is obtained by scaling (4); element (ij) is equal to

$$w_{ij} = \frac{p^{-1}\left(y_i|\theta^{(j)}\right)}{\sum_{j=1}^{T} p^{-1}\left(y_i|\theta^{(j)}\right)},\ i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, T. \qquad (5)$$

A third Bayesian approach, developed by Gianola and Schön (2016), is a leave-one-out cross-validation point estimator of validating mean squared error (a point mass) for Bayesian linear models of the form $E\left(y_i|b, x_i\right) = x_i'b$, where x_i' is an observed row vector of covariates and b is an unobserved column vector of regression parameters. The point estimator holds irrespective of the prior adopted for b. The expression for the validating mean squared error is given by (1) with $\hat{y}_i = x_i'\hat{E}(b|y_{-i})$, where

$$\hat{E}(b|y_{-i}) = \sum_{j=1}^{T} w_{ij}b^{(j)},\ i = 1, 2 \ldots, N, \qquad (6)$$

w_{ij} is given by (5) and $b^{(j)}$ is the jth draw ($j = 1, 2, \ldots, T$) from a Markov chain that at convergence represents an approximate sample from the posterior distribution $[b|y]$ based on the complete data y. The computation of the point estimator of the leave-one-out validating mean squared error based on (6) is shown in the `R`-code below, `sumdg` and `mseloodg`.

An alternative to (6) is to calculate

$$\tilde{E}(b|y_{-i}) = \frac{1}{T}\sum_{j=1}^{T} b_{-i}^{(j)}, \qquad (7)$$

where $b_{-i}^{(j)}$ is the jth McMC draw ($j = 1, 2, \ldots, T$) from the approximation to $[b|y_{-i}]$ from the weighted bootstrap. A point estimator of (1) is obtained using $\hat{y}_i = x_i' \tilde{E}(b|y_{-i})$. The computation of the point estimator based on (7) is shown in the R-code below, `sumavrb` and `mselooavrb`.

These sampling-based approaches are sensitive to the properties of the distribution of the weights w_{ij}. The presence of extreme large values will lead to noisy estimates and some form of smoothing of the weights may be required. An entry to the literature on the subject can be found in Gelman et al. (2014).

The methods are easy to implement in an McMC environment. Imagine that at cycle i, $i = 1, 2, \ldots, rep$, of a Monte Carlo Markov chain, draws of the vector of the conditional variance are stored in a vector `ve` of length *rep*, draws of the vector $\theta = \mu + Zb$ are stored in a matrix `storeMean` of dimension ($rep \times nindiv$), where *nindiv* is the length of the vector of data y, μ is an overall mean, Z is a matrix of observed covariates of dimension ($nindiv \times p$) and b is a column vector with p regression parameters. The vector θ could also represent draws from probability functions $\Phi(\mu + Zb)$ in the case of binary data analysed with a hierarchical probit model. Then the kernel of the R-code to compute Bayesian leave-one-out validating mean squared errors for a linear model is as follows:

```
################### COMPUTE LOOCV BAYES ####################################
sumloo <- 0
sumdg <- 0
sumavrb <- 0
sumXb <- 0
for(i in 1:nindiv){
 w <- 1/dnorm(y.train[i],storeMean[,i],sqrt(ve))
 w <- w/sum(w)
 index <-sample(1:rep,rep,replace=T,prob=w)
 yhatminusi <- rnorm(rep,storeMean[index,i],sqrt(ve))
 sumloo <- sumloo + (y.train[i]-yhatminusi)^2
 expvalueXbminusi <- w%*%storeMean[,i]
 sumdg <- sumdg + (y.train[i]-expvalueXbminusi)^2
 sumavrb <- sumavrb + (y.train[i]-mean(storeMean[index,i]))^2 # (y - Xb^hat)^2;
 #    Xb^hat: post mean of b
 sumXb <- sumXb + (y.train[i]-storeMean[index,i])^2 # (y - Xb)^2; b~[b|y_i]

}
mselooBayes <- sumloo/length(y.train) # estimate based on y^hat = y^*~[y|y_t]
summary(mselooBayes)
quantile(mselooBayes,c(0.025,0.975))
mselooXb <- sumXb/length(y.train) # estimate based on y^hat = xb, b~[b|y_t]
summary(mselooXb)
quantile(mselooXb,c(0.025,0.975))


mseloodg <- sumdg/(length(y.train)) # estimate of mseloodg (Gianola et al, 2016)
mseloodg
mselooavrb <- sumavrb/length(y.train) # McMC estimate of mseloodg (Gianola et al. 2016)
mselooavrb
```

Example: Bayesian Leave-One-Out Cross-Validation

This example illustrates applications of Bayesian leave-one-out cross-validation (BLOOCV) in a simplified genomic context. Two models are fitted to two sets of data consisting

of $2,500$ phenotypic records y_i, $i = 1, 2, \ldots, 2\,500$, and $5,000$ observed genetic marker covariates X_{ij}, $i = 1, 2, \ldots, 2\,500$, $j = 1, 2, \ldots, 5\,000$: a spike and slab model and a genomic model that is a form of ridge regression. Details of these two fitted models and of their McMC implementation are in Chap. 5, page 227, and Chap. 7, page 321. Each covariate X_{ij} is independently drawn from a binomial distribution $Bi\,(X_{ij}|n = 2, p = 0.5)$ and can take the value 0, 1 or 2 with probabilities $(1-p)^2$, $2p\,(1-p)$ and p^2. The elements of the resulting $(2\,500 \times 5\,000)$ matrix of marker covariates are subsequently centred and scaled.

The true models that give rise to the two sets of data are as follows. For both models, observations are the result of the additive effects of a genetic component and of an environmental component. In the first model, the genetic component is due to the additive and equal effects of 25 QTL and in the second is due to 500 additive QTL. The QTL are randomly chosen from the $5,000$ covariates. The total additive genetic variance contributed by these QTL is the same in both scenarios and is equal to 10 squared units. The additive genetic effect of each of the 25 QTL in the first data set is equal to 0.14 phenotypic standard deviations of y. In the second data set, for the model of 500 QTL, this figure is equal to 0.03. The environmental variance in both sets of data is set equal to 30 squared units and the heritability of the quantitative trait is therefore equal to 0.25.

The data available to the analyst consist of two sets of $2,500$ observed phenotypic records and $5,000$ observed genetic marker covariates. In the first data set with 25 QTL (information unknown to the analyst), the spike and slab model is expected to lead to better predictions than the ridge regression. This is so provided that a significant proportion of the 25 QTL are detected and the redundant QTL are partly eliminated. This results in a decline in the variance of the predictor relative to the variance expected from the ridge regression model, that incorporates all the $5,000$ covariates in the construction of the predictor. A smaller variance of the predictor leads to a smaller mean squared error, on average. On the other hand, with 500 QTL each emitting weak signals that are difficult to detect by the spike and slab model, both models are likely to perform similarly.

The performance of the three measures of Bayesian leave-one-out cross-validation computed using the spike and slab model and the ridge regression model, fitted to data set 1 generated assuming 25 QTL and to data set 2 generated assuming 500 QTL, is summarised in Table 1. The first row (LOO_{y^*}) corresponds to the BLOOCV based on the predictor drawn from the posterior predictive distribution (2). The corresponding validating mean squared error (MSE_v) incorporates the posterior uncertainty of the parameters of the Bayesian model θ and sampling variation of individual records. The second row (LOO_θ) corresponds to the BLOOCV where the predictor has the form $\hat{y}_i = x_i' b^*$, where b^* is an extraction from the posterior distribution $[b|y]$. The corresponding MSE_v incorporates posterior uncertainty of the parameters of the Bayesian model. The third row (LOO) shows the point estimates of BLOOCV with entries u/w, where u corresponds to the BLOOCV where the predictor has the form $\hat{y}_i = x_i' \hat{E}(b|y_{-i})$ and $\hat{E}(b|y_{-i})$ is given by (6), and w corresponds to the BLOOCV where the predictor has the form $\hat{y}_i = x_i' \tilde{E}(b|y_{-i})$ and $\tilde{E}(b|y_{-i})$ is given by (7). The first four rows of the table display the McMC estimates of the posterior means and underneath, the 95% posterior intervals (in brackets) of LOO_{y^*} and of LOO_θ.

	Spike and Slab		Ridge	
QTL	25	500	25	500
$LOO_{y^\star}$	61.0	76.1	75.8	75.3
	(58.1;63.8)	(72.5;79.7)	(72.0;79.3)	(71.7;79.0)
LOO_θ	32.0	47.4	46.2	47.2
	(31.5;32.6)	(46.1;48.6)	(44.7;47.6)	(45.7;48.8)
LOO	30.5/30.5	38.0/38.0	37.9/37.9	37.7/37.6

Table 1: McMC estimates of validating mean squared errors (1) (first four rows: estimated means and 95% posterior intervals of marginal posterior distributions in brackets) derived from three forms of Bayesian leave-one-out cross-validation (BLOOCV) using a spike and slab model and a ridge regression model (genomic model). Details of the two models are in Chap. 5 and Chap. 7. The two models were fitted to two simulated data sets. Data set 1 is generated using 25 additive QTL and data set 2 using 500 QTL. $LOO_{y^\star}$: predictor $\hat{y}_i$ drawn from the posterior predictive distribution (2); LOO_θ: predictor of the form $\hat{y}_i = x_i' b^*$, $b^* \sim [b|y]$; LOO: point estimate of the BLOOCV. The row has entries u/w; u corresponds to the form $\hat{y}_i = x_i' \hat{E}(b|y_{-i})$ and $\hat{E}(b|y_{-i})$ is given by (6); w corresponds to the form $\hat{y}_i = x_i' \tilde{E}(b|y_{-i})$ and $\tilde{E}(b|y_{-i})$ is given by (7).

As anticipated, the spike and slab model outperforms the genomic model in the case of data set 1 with 25 QTL, and both models give similar results when fitted to data set 2 with 500 QTL. Within each column, the size of the MSE_v increases (moving from the bottom of the column) as more sources of variation are incorporated into the calculations.

The results indicate that the difference in predictive performance between the two fitted models is consistent within each of the three measures of BLOOCV. If there is interest in a preliminary comparison of model predictive performance, ignoring uncertainty, then the point predictor of BLOOCV is an option. If there is interest in learning how models will perform to predict an average value of y, given a particular value of the covariates, accounting for the fact that the predictor is constructed from a finite set of data, one would consider using LOO_θ. If there is interest in the ability of models to predict new individual records y_i, given a particular value of the covariates, the choice could fall on LOO_{y^*}.

As a further example of an output from an McMC implementation, Figure 1 shows the estimated marginal posterior distributions of BLOOCV based on LOO_{y^*} when the spike and slab model (left subfigure) and the ridge regression model (right subfigure) are fitted to data set 1, with 25 QTL. The information provided by the uncertainty of the BLOOCV in the form of the estimates of marginal posterior distributions enriches inferences. In the case of data set 1, a simple eye-ball comparison of the non-overlapping supports of the estimated posterior distributions of BLOOCV in Figure 1, or a glance at the corresponding posterior intervals of Table 1, leaves little doubt concerning which is the model with best likely predictive performance.

Figure 3 displays the results of a GWAS on data set 1 with 25 QTL (first row, first column) and on data set 2 with 500 QTL (second row, first column) and of the spike

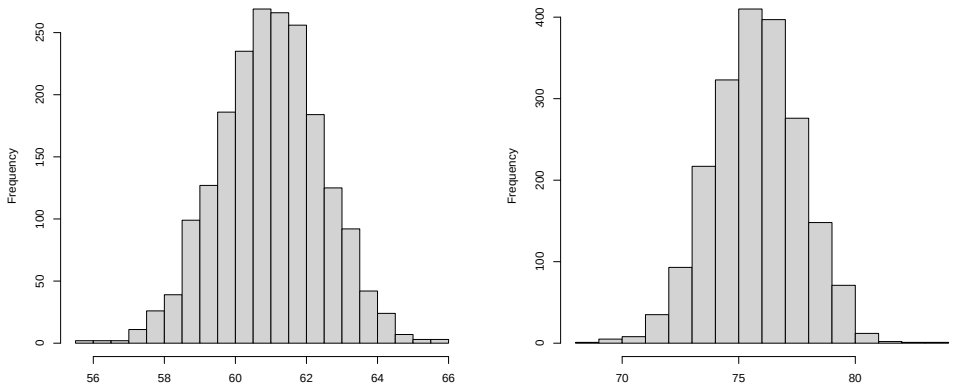

Figure 1: Estimated marginal posterior distributions of validating mean squared errors obtained using the Bayesian leave-one-out cross-validation estimator $LOO_{y^\star}$ (see text for details) derived from a spike and slab model (left subfigure) and from the ridge regression model (right subfigure). Both models fitted to simulated data set 1, generated using 25 QTL.

and slab model (second column). The $y-$axes represent the Bonferroni adjusted $p-$values for the GWAS analysis, and the posterior probability that a given genetic marker has an effect on the trait, for all $5,000$ markers, for the study based on the spike and slab model. The analysis of data set 1 reveals that GWAS detects 15 out of the 25 QTL, with 0 false positives whereas the spike and slab model detects all the 25 QTL, and includes extra 2 false positive results (given the arbitrarily chosen threshold of the posterior probability for declaring a result as "significant" equal to 0.5). On the other hand, with 500 QTL, GWAS fails to detect any QTL, and the spike and slab model detects only 8 QTL including one false positive (given the arbitrarily chosen threshold of the posterior probability equal to 0.5). There is a large proportion of the 500 QTL that are assigned similar posterior probabilities as the non-QTL. Therefore the spike and slab model behaves similarly to the ridge regression; the latter allocates similar values to all the $5,000$ marker loci. This results in approximately equal estimated predictive performance for both fitted models.

The `R`-code `CODE1003` fits the spike and slab model and computes BLOOCV. The ridge regression model is fitted using `R`-code `CODE1313`; the extra lines required to compute BLOOCV are not part of this second code. These two `R`-codes can be found at https://github.com/SorensenS/SLGDS/Codes.

References

Albert, J. (2009). *Bayesian Computation with R.* Springer.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 145–161. Chapman and Hall.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014). *Bayesian Data Analysis*. Chapman and Hall.

Gianola, D. and C. C. Schön (2016). Cross-validation without doing cross-validation in genome-enabled prediction. *G3 6*, 3107–3128.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions (with discussion). In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics 3*, pp. 395–402. Oxford University Press.

Smith, A. F. M. and A. E. Gelfand (1992). Bayesian statistics without tears: A sampling–resampling perspective. *The American Statistician 46*, 84–88.

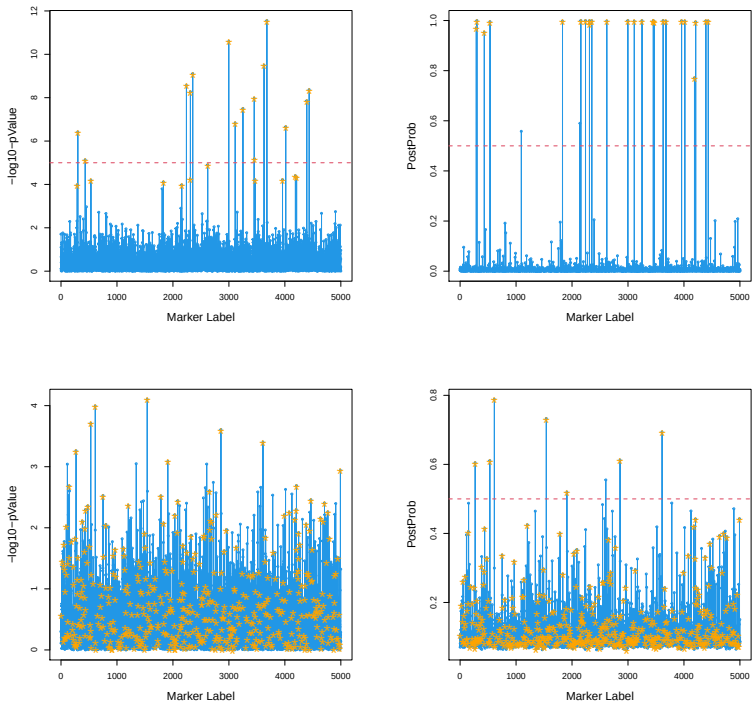

Figure 2: bla

Figure 3: Top row: Data set 1 with 25 QTL; bottom row: Data set 2 with 500 QTL. Top row, left subfigure: $-log_{10}\, p-$values for testing the null hypothesis that the gene has no effect on the trait using GWAS with a Bonferroni adjustment. The horizontal red line is the Bonferroni threshold corresponding to $-log_{10}(0.05/5,000)$. Top row, right subfigure: posterior probability that the gene has an effect on the trait obtained using a Bayesian, Markov chain Monte Carlo implementation of a spike and slab model, for each of the $5,000$ genetic marker loci. Horizontal red line is the arbitrarily chosen threshold corresponding to a posterior probability of 0.5. Bottom row: similar description as top row, except that the models are fitted to data set 2, with 500 QTL. Orange stars represent the 25 QTL in the top subfigures and the 500 QTL in the bottom subfigures.