TUTORIAL ON MODEL ASSESSMENT, SELECTION AND INFERENCE AFTER SELECTION

Aki Vehtari

Department of Computer Science,
Aalto University, Finland
aki.vehtari@aalto.fi
Basics of predictive performance estimation

When cross-validation is not needed
- Simple model we trust

When cross-validation is useful
- We don’t trust the model
- Complex model with posterior dependencies

On accuracy of cross-validation

Cross-validation and hierarchical models

When cross-validation is not enough
- large number of models

loo 2.0
LOO vs. WAIC

- LOO and WAIC estimate the same predictive performance criterion and are asymptotically equal
  - some of the discussion holds for WAIC, too
  - WAIC doesn’t have as good diagnostics and fails earlier than PSIS-LOO used in loo package.

See more in
LOO and WAIC estimate the same predictive performance criterion and are asymptotically equal

- some of the discussion holds for WAIC, too
- WAIC doesn’t have as good diagnostics and fails earlier than PSIS-LOO used in loo package.

See more in

Ideal predictive performance with log score

$$\text{elpd} = \int p_t(\tilde{y}) \log p(\tilde{y} | D, M_k) d\tilde{y},$$

where $p_t(\tilde{y})$ is unknown true future distribution.
Following Bernardo & Smith (1994), there are three different approaches for dealing with the unknown $p_t$:

- $M$-open
- $M$-closed
- $M$-completed
Explicit specification of $p_t(\tilde{y})$ in

$$
\int p_t(\tilde{y}) \log p(\tilde{y} \mid D, M_k) d\tilde{y},
$$

is avoided by re-using the observed data $D$ as a pseudo Monte Carlo samples from the distribution of future data.

Bayesian leave-one-out cross-validation

$$
\widehat{\text{elpd}}_{\text{LOO}} = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i \mid x_i, D_{-i}, M_k)
$$

almost unbiased estimate for a single model
Naïve computation requires computation of $n$ posteriors

Less computation with

- analytic solutions and approximations available for some models
- importance sampling using the full posterior as the proposal (easy to use with Stan)
- $K$-fold cross-validation
  - most robust
Special case is if we leave only one data point out (LOO-CV)

LOO predictive density evaluated at \( y_i \)

\[
p(y_i|x_i, D_{-i}) = \int p(y_i|x_i, \theta)p(\theta|D_{-i})d\theta,
\]

where \( D_{-i} \) is all the data except \((y_i, x_i)\)

- leave-one-out posterior \( p(\theta|D_{-i}) \) is close to full posterior \( p(\theta|D) \), but we still avoid the double use of data
- naïve implementation requires to do the posterior inference \( n \) times
Importance sampling

Having samples $\theta^s$ from $p(\theta^s|D)$

$$p(\tilde{y}_i|x_i, D_{-i}) \approx \frac{\sum_{s=1}^{S} p(\tilde{y}_i|\theta^s)w^s_i}{\sum_{s=1}^{S} w^s_i},$$

where $w^s_i$ are importance weights and

$$w^s_i = \frac{p(\theta^s|x_i, D_{-i})}{p(\theta^s|D)} \propto \frac{1}{p(y_i|\theta^s)}.$$
... model {
  vector[N] eta;
  eta <- beta0 + z*beta;
  beta ~ normal(0, phi);
  phi ~ double_exponential(0, 10);
  y ~ bernoulli_logit(eta);
}

generated quantities {
  vector[N] log_lik;
  vector[N] eta;
  eta <- beta0 + z*beta;
  for (n in 1:N)
    log_lik[n] <- bernoulli_logit_lpdf(y[n], eta[n]);
}
Pareto smoothed importance sampling

\[ p(\tilde{y}_i | x_i, D_{-i}) \approx \frac{\sum_{s=1}^{S} p(\tilde{y}_i | \theta^s) w_i^s}{\sum_{s=1}^{S} w_i^s} \]

- The variance of the importance weights in IS-LOO can be large or even infinite.
- By fitting a generalized Pareto distribution to the tail of the weight distribution:
  - Obtain an estimate of the shape parameter \( k \)
  - If \( k < \frac{1}{2} \) variance is finite, the central limit theorem holds.
  - If \( \frac{1}{2} \leq k < 1 \) variance is infinite but mean exists, the generalized central limit theorem holds.
  - If \( k \geq 1 \) variance and mean do not exist, the truncated estimate will have a finite variance but considerable bias.
- Variance of the IS estimate can be reduced by Pareto smoothing the weights → PSIS-LOO.
- For \( k < 0.7 \) finite sample convergence rates practical.
loo package in CRAN implements PSIS-LOO
  - loo 2.0 is using new version of Pareto smoothing
rstanarm has integrated support
References
Model comparison with LOO

- Pairwise comparison of individual elpd’s

\[
\hat{\text{elpd}}_{\text{diff}} = \frac{1}{n} \sum_{i=1}^{n} \left[ \hat{\text{elpd}}_{\text{LOO},i,M_2} - \hat{\text{elpd}}_{\text{LOO},i,M_1} \right]
\]

- Compute also se for accuracy of the comparison
Example where cross-validation is not needed

- Simple model: we can look at the posterior directly
  - treatment effect of beta-blockers on mortality – betablockers.Rmd
Ideal predictive performance with log score

$$\text{elpd} = \int p_t(\tilde{y}) \log p(\tilde{y} | D, M_k) d\tilde{y},$$

Reference predictive approach

$$\widehat{\text{elpd}}_{\text{ref}} = \int p(\tilde{y} | D, M^*_k) \log p(\tilde{y} | D, M_k) d\tilde{y},$$

where $M^*_k$ is a reference model we trust

- using a model decreases variance, but may introduce bias
- smaller error more useful than unbiasedness, but need to be careful as bias can be very large
Examples where cross-validation is useful

- We don’t trust the model: possible model misspecification
  - treatment effect on number of roaches – roaches.Rmd
- Complex model with posterior dependencies: difficult to analyse posterior
  - colinearity in covariates – colinear.Rmd
On Accuracy of cross-validation

- se for elpd_loo is slightly underestimated and the distribution is often highly skewed
- se for elpd_diff is also underestimated and true distribution might be skewed
- If using elpd_diff and se to compute the probability that one model is better than other ($\Phi(0|\text{elpd\_diff, se})$), these probabilities are not calibrated
  - be cautious when interpreting or reporting these
- We know how to slightly improve the calibration, and we’ll report results on the effects of miscalibration later this year
Compute leave-one-out posterior exactly for those observations for which \( \hat{k} > 0.7 \)

- For rstanarm models: `loo(rstanarmfit, k_threshold=0.7)`
Instead of leaving one observation out, leave a block of observations.

When data is divided in $K$ blocks the approach is called $K$-fold-CV.

If, for example, $K = 10$, then 90% of data is used to form the posterior, which often produces similar posterior as full data.

$k$-fold-CV should be used if PSIS-LOO diagnostics indicate problems with importance sampling and PSIS-LOO+ would compute many more than $K$ posteriors.

For rstanarm models: kfold(rstanarmfit, K = 10)

if the prediction task is for groups
Hierarchical models

1) Predicting new $y_{ij}$ given an existing group $j \in (1, \ldots, J)$
   - LOO or randomized/stratified $K$-fold-CV
2) Predicting new $y_{ij}$ given a new group $j = J + 1$
   - grouped $K$-fold-CV
Importance sampling is less likely to work, as a group of observations is likely to be more influential than just one (and thus full posterior and loo posterior can be different).
Importance sampling is less likely to work, as a group of observations is likely to be more influential than just one (and thus full posterior and loo posterior can be different).

Marginalization in style of Rabe-Hesketh and Furr (Invited talk Wednesday morning) can be used (currently with quadrature implemented in additional software).
Hierarchical models

- Hierarchical model for polling results in different states
  - Predicting new $y_{ij}$ given an existing group $j \in (1, \ldots, J)$
Cross-validation for hierarchical models

- rstanarm support for leave-one-out-group cross-validation in progress
- Hierarchical model comparison examples in progress
Selection induced bias in LOO-CV

- same data is used to assess the performance and make the selection
- the selected model fits more to the data
- the LOO-CV estimate for the selected model is biased
- recognised already, e.g., by Stone (1974)
Selection induced bias in LOO-CV
- same data is used to assess the performance and make the selection
- the selected model fits more to the data
- the LOO-CV estimate for the selected model is biased
- recognised already, e.g., by Stone (1974)

Performance of the selection process itself can be assessed using two level cross-validation, but it does not help choosing better models
Selection induced bias in LOO-CV
- same data is used to assess the performance and make the selection
- the selected model fits more to the data
- the LOO-CV estimate for the selected model is biased
- recognised already, e.g., by Stone (1974)

Performance of the selection process itself can be assessed using two level cross-validation, but it does not help choosing better models

Bigger problem if there is a large number of models as in covariate selection
Selection induced bias in variable selection

![Graphs showing selection induced bias for different sample sizes (n = 20, n = 50, n = 100). The graphs illustrate the relationship between variable selection and the induced bias, with axes indicating the number of variables and bias values.](image-url)
Selection induced bias in variable selection

Piironen & Vehtari (2017)
Selection induced bias in variable selection

Piironen & Vehtari (2017)
Selection induced bias in variable selection

Sonar

CV-10 / IS-LOO-CV

WAIC

DIC

MPP

BMA-ref

BMA-proj

Piironen & Vehtari (2017)

28 / 44
Originally proposed for generalized linear models by Goutis and Robert (1998), Dupuis and Robert (2003) (the decision theoretic idea of using the full model can be tracked to Lindley (1968), see also many related references in Vehtari and Ojanen (2012))

Performs well in practice in comparison to many other methods (Piironen and Vehtari 2016)

- has low variance
- able to preserve information from the full model
Projection predictive method, general idea

- Originally proposed for generalized linear models by Goutis and Robert (1998), Dupuis and Robert (2003) (the decision theoretic idea of using the full model can be tracked to Lindley (1968), see also many related references in Vehtari and Ojanen (2012))

- Performs well in practice in comparison to many other methods (Piironen and Vehtari 2016)
  - has low variance
  - able to preserve information from the full model

- General idea
  - Fit the full encompassing model (with all the inputs) with best possible prior information
Originally proposed for generalized linear models by Goutis and Robert (1998), Dupuis and Robert (2003) (the decision theoretic idea of using the full model can be tracked to Lindley (1968), see also many related references in Vehtari and Ojanen (2012))

Performs well in practice in comparison to many other methods (Piironen and Vehtari 2016)

- has low variance
- able to preserve information from the full model

General idea

- Fit the full encompassing model (with all the inputs) with best possible prior information
- Any submodel (reduced number of inputs) is trained by minimizing predictive Kullback-Leibler (KL) divergence to the full model (= projection)
  - For a given number of variables, choose the model with minimal projection discrepancy
The full model predictive distribution represents our best knowledge about future $\tilde{y}$

$$p(\tilde{y}|D) = \int p(\tilde{y}|\theta)p(\theta|D)d\theta,$$

where $\theta = (\beta, \sigma^2)$ and $\beta$ is in general non-sparse (all $\beta_j \neq 0$)
The full model predictive distribution represents our best knowledge about future $\tilde{y}$

$$p(\tilde{y} | D) = \int p(\tilde{y} | \theta) p(\theta | D) d\theta,$$

where $\theta = (\beta, \sigma^2)$ and $\beta$ is in general non-sparse (all $\beta_j \neq 0$)

What is the best distribution $q_{\perp}(\theta)$ given a constraint that only selected covariates have nonzero coefficient
Projective predictive covariate selection, idea

- The full model predictive distribution represents our best knowledge about future $\tilde{y}$

$$p(\tilde{y} | D) = \int p(\tilde{y} | \theta)p(\theta | D)d\theta,$$

where $\theta = (\beta, \sigma^2))$ and $\beta$ is in general non-sparse (all $\beta_j \neq 0$)

- What is the best distribution $q_\perp(\theta)$ given a constraint that only selected covariates have nonzero coefficient

- Optimization problem:

$$q_\perp = \arg \min_q \frac{1}{n} \sum_{i=1}^{n} KL \left( p(\tilde{y}_i | D) \| \int p(\tilde{y}_i | \theta)q(\theta)d\theta \right)$$
The full model predictive distribution represents our best knowledge about future $\tilde{y}$

$$p(\tilde{y}|D) = \int p(\tilde{y}|\theta)p(\theta|D)d\theta,$$

where $\theta = (\beta, \sigma^2)$ and $\beta$ is in general non-sparse (all $\beta_j \neq 0$)

What is the best distribution $q_{\perp}(\theta)$ given a constraint that only selected covariates have nonzero coefficient

Optimization problem:

$$q_{\perp} = \arg \min_q \frac{1}{n} \sum_{i=1}^{n} KL\left(p(\tilde{y}_i | D) \parallel \int p(\tilde{y}_i | \theta)q(\theta)d\theta\right)$$

Optimal projection from the full posterior to a sparse posterior (with minimal predictive loss)
We have posterior draws \( \{ \theta^s \}_{s=1}^S \), for the full model \( \theta = (\beta, \sigma^2) \) and \( \beta \) is in general non-sparse (all \( \beta_j \neq 0 \)).
We have posterior draws \( \{ \theta^s \}_{s=1}^S \), for the full model \( \theta = (\beta, \sigma^2) \) and \( \beta \) is in general non-sparse (all \( \beta_j \neq 0 \)).

The predictive distribution \( p(\tilde{y} \mid D) \approx \frac{1}{S} \sum_s p(\tilde{y} \mid \theta^s) \) represents our best knowledge about future \( \tilde{y} \).
We have posterior draws $\{\theta^s\}_{s=1}^{S}$, for the full model $\theta = (\beta, \sigma^2)$ and $\beta$ is in general non-sparse (all $\beta_j \neq 0$)

The predictive distribution $p(\tilde{y} \mid D) \approx \frac{1}{S} \sum_s p(\tilde{y} \mid \theta^s)$ represents our best knowledge about future $\tilde{y}$

Easier optimization problem by changing the order of integration and optimization (Goutis & Robert, 1998):

$$\theta^s_{\perp} = \arg\min_{\hat{\theta}} \frac{1}{n} \sum_{i=1}^{n} \text{KL}\left( p(\tilde{y}_i \mid \theta^s) \parallel p(\tilde{y}_i \mid \hat{\theta}) \right)$$
We have posterior draws $\{\theta^s\}_{s=1}^S$, for the full model $(\theta = (\beta, \sigma^2))$ and $\beta$ is in general non-sparse (all $\beta_j \neq 0$).

The predictive distribution $p(\tilde{y} \mid D) \approx \frac{1}{S} \sum_s p(\tilde{y} \mid \theta^s)$ represents our best knowledge about future $\tilde{y}$.

Easier optimization problem by changing the order of integration and optimization (Goutis & Robert, 1998):

$$\theta^s_s = \arg\min_{\hat{\theta}} \frac{1}{n} \sum_{i=1}^n KL\left( p(\tilde{y}_i \mid \theta^s) \parallel p(\tilde{y}_i \mid \hat{\theta}) \right)$$

$\theta^s_s$ are now (approximate) draws from the projected distribution.
Projection by draws

- Projection of one Monte Carlo sample can be solved
  - Gaussian case: analytically

\[ w_\perp = (X_\perp^T X_\perp)^{-1} X_\perp^T f \]

\[ \sigma^2_\perp = \sigma^2 + \frac{1}{n} (f - f_\perp)^T (f - f_\perp) \]
Projection of one Monte Carlo sample can be solved

- **Gaussian case:** analytically
  
  \[ w_\perp = (X_\perp^T X_\perp)^{-1} X_\perp^T f \]
  
  \[ \sigma^2_\perp = \sigma^2 + \frac{1}{n} (f - f_\perp)^T (f - f_\perp) \]

- **Exponential family case:** equivalent to finding the maximum likelihood parameters for the submodel with the observations replaced by the fit of the reference model (Goutis & Robert, 1998; Dupuis & Robert, 2003)
Example

The full model

A projected model (with variables ordered in relevance)

rstanarm + projpred + bayesplot
In variable selection usually not feasible to go through all variable combinations
In variable selection usually not feasible to go through all variable combinations

Use e.g. forward search to explore promising combinations

- start from the empty model, at each step add the variable that reduces the objective the most
- stop when the performance similar to the full model
  - can use PSIS-LOO to estimate the performance and to choose the model size
Selection induced bias in variable selection

$\text{CV-10}$

$\text{WAIC}$

$\text{DIC}$

$\text{MPP}$

$\text{BMA-ref}$

$\text{BMA-proj}$

$n = 100$

$n = 200$

$n = 400$

Piironen & Vehtari (2017)
Selection induced bias in variable selection

![Graphs showing selection induced bias in variable selection with n = 100, n = 200, and n = 400 for different criteria: CV-10, WAIC, DIC, MPP, BMA-ref, BMA-proj.](image)

Piironen & Vehtari (2017)
Simulated example

\( n = 80, \ p = 200, \) only 7 features are relevant
Simulated example

\( n = 80, p = 200, \) only 7 features are relevant

- Lasso-path when \( \lambda \) is varied, optimal model size by cross-validation (dotted) vertical axis shows the test error
Simulated example

MSE vs Number of features

- Lasso-path (black)
Simulated example

- Lasso-path (black), full model Bayes with HS-prior (dashed)
Lasso-path (black), full model Bayes with HS-prior (dashed) and the $L_1$-projection (blue)
Predicting disease risk with logistic regression – diabetes.Rmd
Not yet in CRAN, but hopefully this spring
Improved PSIS-LOO
- Improved Pareto diagnostics and smoothing
- Effective sample size and se estimates

Model averaging
- Pseudo-BMA+ weights
- Stacking weights

Helper functions for $K$-fold-CV


