Model selection and cross validation

FW 891

Click here to view presentation online

Christopher Cahill

8 November 2023

Quantitative Fisheries Center
MICHIGAN STATE UNIVERSITY

Purpose

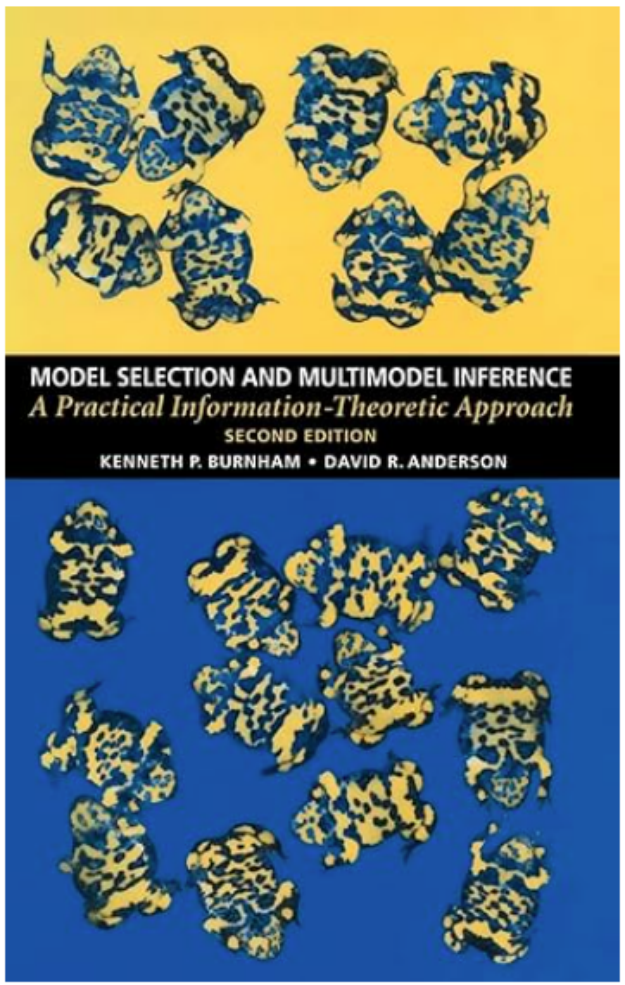
- Goal
- A philosophical preface to model selection
- k-fold and loo cross validation
- Performance criteria
- Challenges
- Approximate methods to loo cross validation
- R and Stan demo on how to implement these ideas

Useful reference on cross validation in Stan:

https://users.aalto.fi/~ave/CV-FAQ.html

TLDR: Model selection is hard and requires careful thought

Model selection: are we stuck between the devil and the deep blue sea?



Burnham and Anderson 1998; Navarro 2018; Bolker 2023

Between the devil and the deep blue sea?

- We can calculate how well we predict things

Navarro 2018; Bolker 2023

Between the devil and the deep blue sea?

- We can calculate how well we predict things
- The difference between prediction and inference

Navarro 2018; Bolker 2023

Between the devil and the deep blue sea?

- We can calculate how well we predict things
- The difference between prediction and inference
- If scientific reasoning takes place in a world where all our models are systematically wrong in some sense, what do we hope to achieve by selecting a model?

Navarro 2018; Bolker 2023

Between the devil and the deep blue sea?

- We can calculate how well we predict things
- The difference between prediction and inference
- If scientific reasoning takes place in a world where all our models are systematically wrong in some sense, what do we hope to achieve by selecting a model?
- *The devil*: statistical decision making

Navarro 2018; Bolker 2023

Between the devil and the deep blue sea?

- We can calculate how well we predict things
- The difference between prediction and inference
- If scientific reasoning takes place in a world where all our models are systematically wrong in some sense, what do we hope to achieve by selecting a model?
- *The devil*: statistical decision making
- *The deep blue sea*: addressing scientific questions

Navarro 2018; Bolker 2023

Between the devil and the deep blue sea?

- We can calculate how well we predict things
- The difference between prediction and inference
- If scientific reasoning takes place in a world where all our models are systematically wrong in some sense, what do we hope to achieve by selecting a model?
- *The devil*: statistical decision making
- *The deep blue sea*: addressing scientific questions
- A question well worth pondering that I have no intention of answering:

Navarro 2018; Bolker 2023

Between the devil and the deep blue sea?

- We can calculate how well we predict things
- The difference between prediction and inference
- If scientific reasoning takes place in a world where all our models are systematically wrong in some sense, what do we hope to achieve by selecting a model?
- *The devil*: statistical decision making
- *The deep blue sea*: addressing scientific questions
- A question well worth pondering that I have no intention of answering:
  - Are scientific model selection questions addressable with statistical tools?

Navarro 2018; Bolker 2023

With that in mind

Arthur Schopenhauer the philosophy Bunny peering into the inferential abyss

Introduction

- After fitting a Bayesian model, we often want to measure its predictive accuracy

Vehtari et al. 2019

Introduction

- After fitting a Bayesian model, we often want to measure its predictive accuracy
- We might do this to:

Vehtari et al. 2019

Introduction

- After fitting a Bayesian model, we often want to measure its predictive accuracy
- We might do this to:
  - For its own sake

Vehtari et al. 2019

Introduction

- After fitting a Bayesian model, we often want to measure its predictive accuracy
- We might do this to:
  - For its own sake
  - Compare models

Vehtari et al. 2019

Introduction

- After fitting a Bayesian model, we often want to measure its predictive accuracy
- We might do this to:
  - For its own sake
  - Compare models
  - Model selection

Vehtari et al. 2019

Introduction

- After fitting a Bayesian model, we often want to measure its predictive accuracy
- We might do this to:
  - For its own sake
  - Compare models
  - Model selection
  - Model averaging

Vehtari et al. 2019

What is cross validation?

- Cross validation is a family of techniques that try to estimate how well a model would predict previously unseen data

What is cross validation?

- Cross validation is a family of techniques that try to estimate how well a model would predict previously unseen data
  - Typically do this by fitting the model to some subset of the data, and then predicting the left out data

What is cross validation?

- Cross validation is a family of techniques that try to estimate how well a model would predict previously unseen data
  - Typically do this by fitting the model to some subset of the data, and then predicting the left out data
- Cross validation can be used to

What is cross validation?

- Cross validation is a family of techniques that try to estimate how well a model would predict previously unseen data
  - Typically do this by fitting the model to some subset of the data, and then predicting the left out data
- Cross validation can be used to
  - Assess the predictive performance of a single model

What is cross validation?

- Cross validation is a family of techniques that try to estimate how well a model would predict previously unseen data
  - Typically do this by fitting the model to some subset of the data, and then predicting the left out data
- Cross validation can be used to
  - Assess the predictive performance of a single model
  - Assess model misspecification

What is cross validation?

- Cross validation is a family of techniques that try to estimate how well a model would predict previously unseen data
  - Typically do this by fitting the model to some subset of the data, and then predicting the left out data
- Cross validation can be used to
  - Assess the predictive performance of a single model
  - Assess model misspecification
  - Compare multiple models

What is cross validation?

- Cross validation is a family of techniques that try to estimate how well a model would predict previously unseen data
  - Typically do this by fitting the model to some subset of the data, and then predicting the left out data
- Cross validation can be used to
  - Assess the predictive performance of a single model
  - Assess model misspecification
  - Compare multiple models
  - Select a single model from multiple candidates

What is cross validation?

- Cross validation is a family of techniques that try to estimate how well a model would predict previously unseen data
  - Typically do this by fitting the model to some subset of the data, and then predicting the left out data
- Cross validation can be used to
  - Assess the predictive performance of a single model
  - Assess model misspecification
  - Compare multiple models
  - Select a single model from multiple candidates
  - Combine the predictions of multiple models

K-fold and leave-one-out cross validation

- K-fold cross validation refers to splitting a dataset into K approximately equal sized chunks

K-fold and leave-one-out cross validation

- K-fold cross validation refers to splitting a dataset into K approximately equal sized chunks
  - Often K = 10

K-fold and leave-one-out cross validation

- K-fold cross validation refers to splitting a dataset into K approximately equal sized chunks
  - Often K = 10
- Procedure:
  - Estimate the model on K-1 of the chunks then predict the left out chunk

K-fold and leave-one-out cross validation

- K-fold cross validation refers to splitting a dataset into K approximately equal sized chunks
  - Often K = 10
- Procedure:
  - Estimate the model on K-1 of the chunks then predict the left out chunk
  - Repeat this process until we've shuffled through each chunk or fold of the data

K-fold and leave-one-out cross validation

- K-fold cross validation refers to splitting a dataset into K approximately equal sized chunks
  - Often K = 10
- Procedure:
  - Estimate the model on K-1 of the chunks then predict the left out chunk
  - Repeat this process until we've shuffled through each chunk or fold of the data
- Leave one out (LOO) cross validation represents the limit of K-fold cross validation, where K equals number of data points

Some measures of predictive accuracy

Mean Square Error (MSE)

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathrm{E}\left(y_i \mid \theta\right)\right)^2$$

- y_i is data point i
- θ represent fitted model parameters
- proportional to MSE if model is normal with constant variance
- Easy to compute and understand, but less appropriate for non-normal models

Gelman et al. 2014; Vehtari et al. 2016

Expected log pointwise predictive density (elpd)

- Consider data $y_1, \ldots, y_n$ modeled as independent given parameters θ
- Also suppose we have a prior distribution $p(\theta)$ yielding a posterior $p(\theta \mid y)$
- And a posterior predictive distribution $p(\tilde{y} \mid y) = \int p\left(\tilde{y}_i \mid \theta\right) p(\theta \mid y) d\theta$

Gelman et al. 2014; Vehtari et al. 2016

Expected log pointwise predictive density (elpd)

- Consider data $y_1, \ldots, y_n$ modeled as independent given parameters θ
- Also suppose we have a prior distribution $p(\theta)$ yielding a posterior $p(\theta \mid y)$
- And a posterior predictive distribution $p(\tilde{y} \mid y) = \int p\left(\tilde{y}_i \mid \theta\right) p(\theta \mid y) d\theta$
- We can then define a measure of predictive accuracy for the *n* data points as:

Gelman et al. 2014; Vehtari et al. 2016

Expected log pointwise predictive density (elpd)

- Consider data $y_1, \ldots, y_n$ modeled as independent given parameters θ
- Also suppose we have a prior distribution $p(\theta)$ yielding a posterior $p(\theta \mid y)$
- And a posterior predictive distribution $p(\tilde{y} \mid y) = \int p\left(\tilde{y}_i \mid \theta\right) p(\theta \mid y) d\theta$
- We can then define a measure of predictive accuracy for the *n* data points as:

$$\begin{aligned} \text{elpd} &= \text{expected log pointwise predictive density for a new dataset} \\ &= \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p\left(y_i \mid \theta^s\right) \right). \end{aligned}$$

Gelman et al. 2014; Vehtari et al. 2016

Expected log pointwise predictive density (elpd)

- Consider data $y_1, \ldots, y_n$ modeled as independent given parameters θ
- Also suppose we have a prior distribution $p(\theta)$ yielding a posterior $p(\theta \mid y)$
- And a posterior predictive distribution $p(\tilde{y} \mid y) = \int p\left(\tilde{y}_i \mid \theta\right) p(\theta \mid y) d\theta$
- We can then define a measure of predictive accuracy for the *n* data points as:

$$\begin{aligned} \text{elpd} &= \text{expected log pointwise predictive density for a new dataset} \\ &= \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p\left(y_i \mid \theta^s\right)\right). \end{aligned}$$

- where θ^s represent posterior simulations from $s = 1, \ldots, s$

Gelman et al. 2014; Vehtari et al. 2016

Some extensions to simple k-fold cross validation

- Often the data are subsetted randomly; however, this may not always represent the relevant prediction task

Roberts et al. 2017

Some extensions to simple k-fold cross validation

- Often the data are subsetted randomly; however, this may not always represent the relevant prediction task
- Ecological data are commonly correlated in space, time, groups, or even phylogenetic structure

Roberts et al. 2017

Some extensions to simple k-fold cross validation

- Often the data are subsetted randomly; however, this may not always represent the relevant prediction task
- Ecological data are commonly correlated in space, time, groups, or even phylogenetic structure
  - Dependency in groups, space, or time

Roberts et al. 2017

Some extensions to simple k-fold cross validation

- Often the data are subsetted randomly; however, this may not always represent the relevant prediction task
- Ecological data are commonly correlated in space, time, groups, or even phylogenetic structure
  - Dependency in groups, space, or time
- Many strategies we can use depending on our prediction task

Roberts et al. 2017

Cross validation and LOO have many limitations

Some known issues

- Computationally demanding

Gelman et al. 2013; Vehtari 2023

Some known issues

- Computationally demanding
- Methods run into problems with sparse data

Gelman et al. 2013; Vehtari 2023

Some known issues

- Computationally demanding
- Methods run into problems with sparse data
- When to not use cross-validation?
  - *In general, there is no need to do any model selection*

Gelman et al. 2013; Vehtari 2023

Some known issues

- Computationally demanding
- Methods run into problems with sparse data
- When to not use cross-validation?
  - *In general, there is no need to do any model selection*
  - Best approach is to build a rich model that includes all uncertainties, do model checking, and perhaps adjust that model if necessary

Gelman et al. 2013; Vehtari 2023

Some known issues

- Computationally demanding
- Methods run into problems with sparse data
- When to not use cross-validation?
  - *In general, there is no need to do any model selection*
  - Best approach is to build a rich model that includes all uncertainties, do model checking, and perhaps adjust that model if necessary
- Cross validation cannot directly answer the question “do the data provide evidence for some effect being non-zero?”

Gelman et al. 2013; Vehtari 2023

Some known issues

- Computationally demanding
- Methods run into problems with sparse data
- When to not use cross-validation?
  - *In general, there is no need to do any model selection*
  - Best approach is to build a rich model that includes all uncertainties, do model checking, and perhaps adjust that model if necessary
- Cross validation cannot directly answer the question “do the data provide evidence for some effect being non-zero?”
- What does cross validation tell you?

Gelman et al. 2013; Vehtari 2023

How do you view the world?

M-closed vs. M-open worlds

Navarro 2019

Approximate methods for calculating elpd (sneakery)

Approximate cross validation

- Vehtari et al. (2016; 2017) introduced a method that approximates the evaluations of leave-one-out cross validation inexpensively using only the data point log likelihoods of a single model fit

Vehtari et al. 2016; 2017

Approximate cross validation

- Vehtari et al. (2016; 2017) introduced a method that approximates the evaluations of leave-one-out cross validation inexpensively using only the data point log likelihoods of a single model fit
- Pareto-smoothed importance sampling (PSIS-LOO) allows us to compute an approximation to LOO without re-fitting the model many times

Vehtari et al. 2016; 2017

Importance sampling LOO

- Since we are Bayesian, we have samples from a posterior
- Approximate the likelihood our model would give some datum if we hadn't observed that datum:

$$\int p\left(y_1 \mid \theta\right) d\theta$$

- Since we are working with samples we move from an integral to an average over samples:

$$\frac{1}{S} \sum_s p\left(y_1 \mid \theta_s\right)$$

Vehtari et al. 2016; 2017

Importance sampling LOO

- Now we want to reweight the posterior samples as thought y_1 wasn't observed:

$$\frac{1}{\sum_s w_s} \sum_s w_s p\left(y_1 \mid \theta_s\right)$$

- The weighting we will use:

$$\frac{1}{p\left(y_1 \mid \theta_s\right)}$$

Vehtari et al. 2016; 2017

The Pareto part

- It turns out that importance sampling is very noisy, and the sampling weights have very heavy tails

Vehtari et al. 2016; 2017

The Pareto part

- It turns out that importance sampling is very noisy, and the sampling weights have very heavy tails
- We need to smooth out the tails so that a single datum doesn't dominate our adjusted posterior

Vehtari et al. 2016; 2017

The Pareto part

- It turns out that importance sampling is very noisy, and the sampling weights have very heavy tails
- We need to smooth out the tails so that a single datum doesn't dominate our adjusted posterior
- Turns out the upper tail of the importance weights fits a generalized Pareto distribution nicely and we can use this to smooth out our weights w_s

Vehtari et al. 2016; 2017

PSIS-LOO implementation

- If all of that overwhelms you…

Vehtari et al. 2016; 2017

PSIS-LOO implementation

- If all of that overwhelms you…
- There are packages and functions that help you do this

Vehtari et al. 2016; 2017

PSIS-LOO implementation

- If all of that overwhelms you…
- There are packages and functions that help you do this
- They have a lot of diagnostics to tell you when they think they are going wrong

Vehtari et al. 2016; 2017

To the R and Stan code

References

- Burnham and Anderson 1998. Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York, USA.
- Gelman , A., Hwang, J. & Vehtari , A. 2014. Understanding predictive information criteria for Bayesian models. Stat. Comput., 24, 997 1016.
- Navarro, D.J. 2019. Between the devil and the deep blue sea: tensions between scientific judgement and statistical model selection. Compuational Brain and Behavior. 2:28-34.
- Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. “Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC.” Statistics and Computing 27 (5): 1413–32.
- Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. 2019. Pareto smoothed importance sampling. preprint arXiv:1507.02646.
- Vehtari 2023. https://users.aalto.fi/~ave/CV-FAQ.html#1_What_is_cross-validation