Model Selection

Models for Socio-Environmental Data

Chris Che-Castaldo, Mary B. Collins, N. Thompson Hobbs

May 29, 2019

Learning objective

- Consider when you need to use multimodel inference and when inference from a single model is sufficient.
- Appreciate different methods for multi-model inference in the Bayesian framework.
- Be able to write code for alternative model selection and model averaging methods.

Often one model is all you need.

- When parameters are based on well established mechanism and we want to estimate them and evaluate their importance.
- When form of model is dictated by objectives.
- Whenever we can make inference conditional on a single model.

Often one model is all you need.

- Hobbs, N. T., H. Andren, J. Persson, M. Aronsson, and G. Chapron. 2012. Native predators reduce harvest of reindeer by Sami pastoralists. Ecological Applications 22:1640-1654.
- Ver Hoef, J. M. and P. L. Boveng. 2015. Iterating on a single model is a viable alternative to multimodel inference. The Journal of Wildlife Management, 79(5):719–729
- Gelman, A., and D. B. Rubin. 1995. Avoiding model selection in Bayesian social research. Pages 165-173 Sociological Methodology 1995, Vol 25.

"Model selection and model averaging are deep waters, mathematically, and no consensus has emerged in the substantial literature on a single approach. Indeed, our only criticism of the wide use of AIC weights in wildlife and ecological statistics is with their uncritical acceptance and the view that this challenging problem has been simply resolved."

Link, W. A., and R. J. Barker. 2006. Model weights and the foundations of multimodel inference. Ecology 87:2626-2635.

The candidate set of models

We have a set of L alternative models differing in the number of parameters they contain, in their functional forms, or both. Call this set of models $\mathscr{M} = \{M_1, \ldots, M_l, \ldots, M_L\}$. Assume that all of these models have been chosen thoughtfully by the researcher and have passed posterior predictive checks. Model checking first, then model selection if needed.

Multi-model inference

- Model selection: How do we decide which model is the "best" among a set of candidates?
  - Model validation
  - Deviance information criterion (DIC)
  - Posterior predictive loss (Dsel)
  - Not covered: Wantanabe information criterion (WAIC): limited by assumptions on spatial and temporal dependence in data, otherwise an excellent Bayesian method. See Hobbs and Hooten, chapter 9 and suggestions for further study.
- Model-averaging: How do we use multiple models as basis for inference by giving them weights?
  - Indicator variable selection
  - Not covered: Probability of the model and Bayes factors. Requires reversible jump MCMC, which is tough to code. See Link, W. A., and R. J. Barker. 2010. Bayesian Inference with ecological applications. Academic Press, chapter 7 and BMA package in R.

Out of sample validation: the gold standard

$$[\mathbf{y}_{\text{oos}}|\mathbf{y}] = \int \dots \int [\mathbf{y}_{\text{oos}}|\mathbf{y}, \boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}] d\theta_1, \dots d\theta_p \ ,$$

$$\text{log predictive density, LPD} = \log([\mathbf{y}_{\text{oos}}|\mathbf{y}])$$

which evaluates to a scalar after the data are collected. Larger LPD indicates greater predictive ability.

Approximated by

$$\log[\mathbf{y}_{\text{oos}}|\mathbf{y}] \approx \log\left(\frac{\sum_{k=1}^{K}[\mathbf{y}_{\text{oos}}|\mathbf{y}, \boldsymbol{\theta}^{(k)}]}{K}\right) \ ,$$

Implementing out-of-sample validation

- Insert code into your JAGS model that makes a prediction for each out of sample data point.
- Compute the probability density of each of the out of sample observations conditional on the model prediction of the observation.
- Take the product of the probability densities across all observation-prediction pairs to obtain $[\mathbf{y}_{\text{oos}}|\mathbf{y}]$.
- On the R side, take the log of the mean of $[\mathbf{y}_{\text{oos}}|\mathbf{y}]$.

Example code

```
for(i in 1:length(y.oos)){
    y.hat[i]<-B0+B1*x.oos[i]
    density[i]<-
    dnorm(y.oos[i],y.hat[i],tau[i])
    }
    PD<-
    prod(density) ##may need to do this as sum of logs
```

On R side, include PD in your variable list for JAGS or coda samples. Take the log of the mean of PD to get LPD.

Show differences in JAGS density functions on the board

M-fold cross validation: the next best to OOS validation

- Group the data into M groups, $m = 1, \ldots M$.
- Fit model leaving out the observations for each of the M groups.
- Calculate predictive score at each MCMC iteration based on ability of model to predict the withheld observations for each group, $[\mathbf{y}_m | \mathbf{y}_{-m}, \boldsymbol{\theta}^{(k)}]$.
- Store the mean of the predictive density for each model fit. Sum the logs of the means:

$$\sum_{m=1}^{M} \log \left(\frac{\sum_{k=1}^{K} [\mathbf{y}_m | \mathbf{y}_{-m}, \boldsymbol{\theta}^{(k)}]}{K} \right) .$$

Example: leave one out cross validation

1. Create M data sets, each of which omits a single observation.
2. Fit candidate model to each dataset and calculate the probability (or probability density) of the left-out observation conditional on the model's prediction of that observation.
3. Compute the mean of the probability or probability density across the K MCMC iterations for each of the M left out datasets and sum the log of those means. Larger values indicate greater predictive ability.

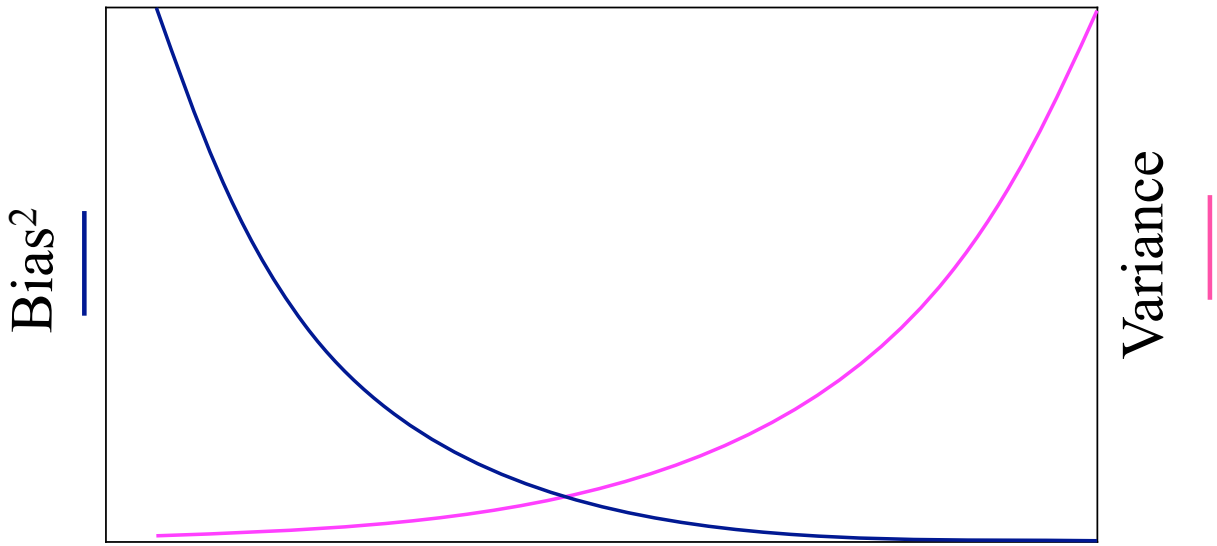
How to set up training and test datasets?

- http://stats.stackexchange.com/questions/61090/how-to-split-a-data-set-to-do-10-fold-cross-validation

Information criteria: the IC's

- Cross validation has a large computational cost.
- There may not be sufficient data for out-of-sample validation.
- Information criteria attempt to obtain same inference as validation procedures by calculating a single statistic from data that are used for model fitting. All are based on the idea of statistical regularization.

Statistical Regularization



Sakamoto et al. 1986

"True model:" $y = e^{(x-0.3)^2} - 1 + \varepsilon,$

Generated 10 data sets sampling from normal distribution with mean = 0 and variance = .01

Fit 5 approximating models to the 10 data sets

$y = \beta_0 + \beta_1 x$

$y = \beta_0 + \beta_1 x + \beta_2 x^2$

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5$

What creates "noise" in models?

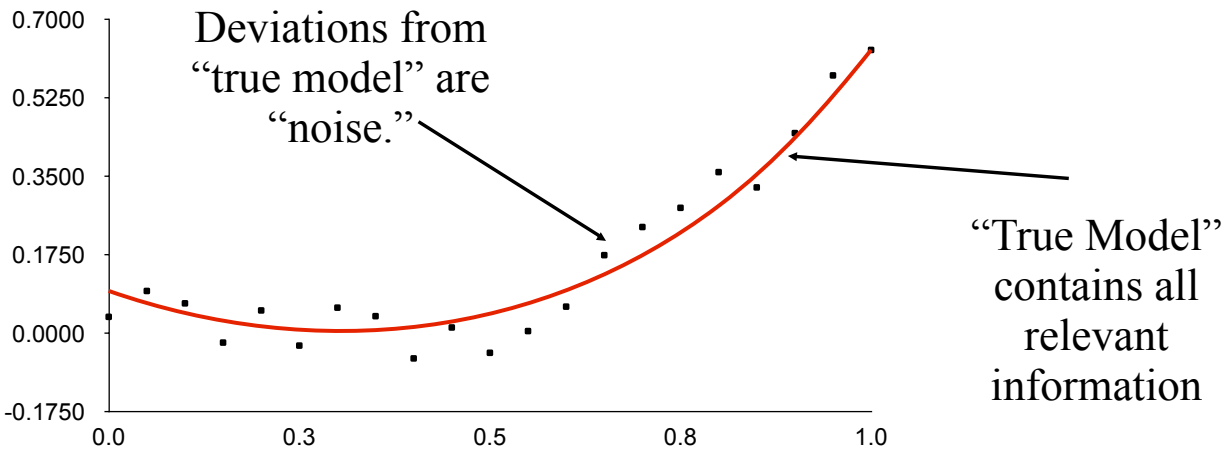
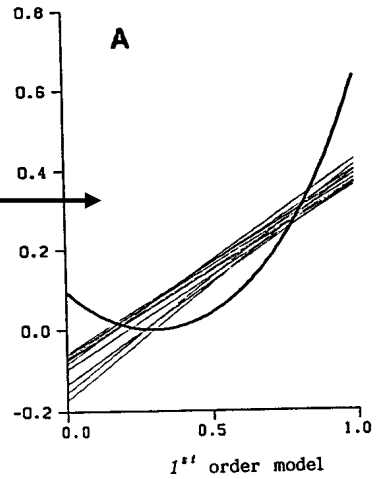
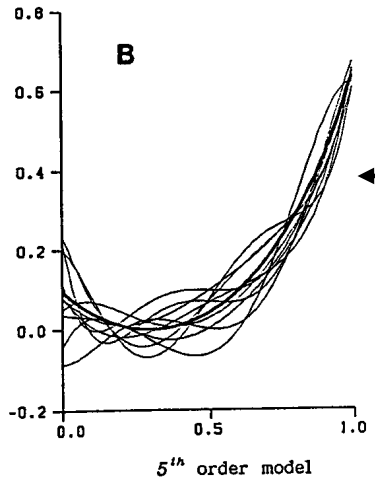


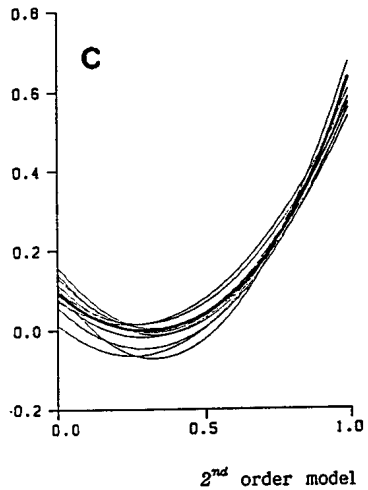
Illustration of trade off

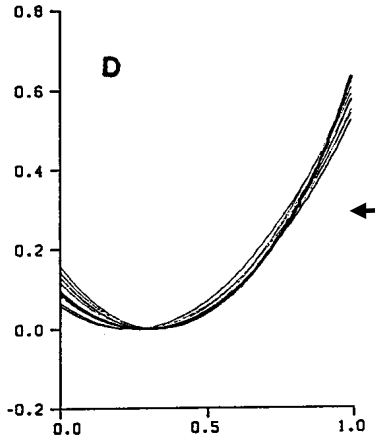
High bias →



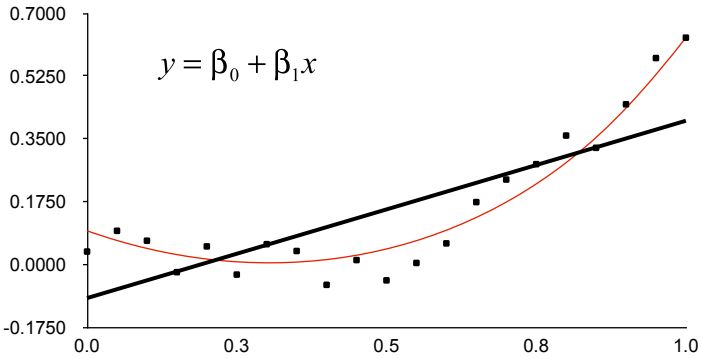


← High variance

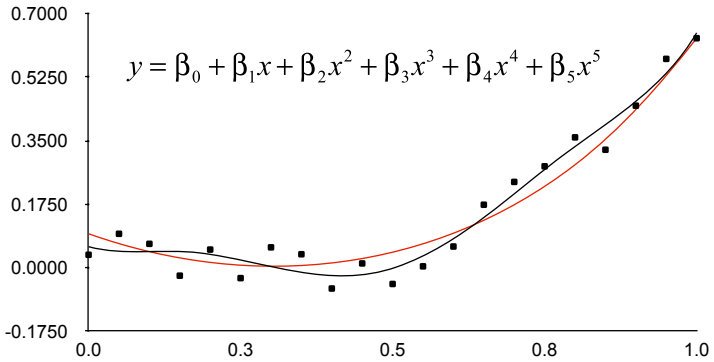




← Optimum, 2nd order with minimum at x=.03



Two few parameters--fails to respond to information. Bias is high.



Too many parameters--responds to "noise." Variance is high.

Statistical regularization

$$\underbrace{\mathscr{L}(\mathbf{y}, \boldsymbol{\theta})}_{\text{loss function}} + \underbrace{r(\boldsymbol{\theta}, \boldsymbol{\gamma})}_{\text{regulator}}$$

The term "regularization" comes from the use of a function that regulates an optimization. Can shrink variance of estimates or increase accuracy of predictions or both.[1]

[1] Do not confuse $\mathscr{L}(\mathbf{y}, \boldsymbol{\theta})$ with a likelihood.

Examples of statistical regularization

- The Bayesian prior $\log[\theta|\mathbf{y}] \propto \log[\mathbf{y}|\theta] + \log[\theta]$
- Priors in penalized MLE $\log L[\theta|\mathbf{y}] = \sum_{i=1}^{n} \log[y_i \mid \theta] + \log(\theta)$
- Ridge regression
- LASSO[2]
- Information criteria (AIC, BIC, DIC, WAIC)
- Posterior predictive loss

[2] LASSO = least absolute shrinkage and selection operator

Deviance

$$
\begin{aligned}
D(\boldsymbol{\theta}) &= \overbrace{-2\log\left[\mathbf{y}|\boldsymbol{\theta}\right]}^{\text{Deviance}} \\
&= -2\log\left[\mathbf{y}|g(\boldsymbol{\theta},\mathbf{x}),\sigma^2\right] \\
&= -2\log\prod_{i=1}^{n}\left[y_i|g(\boldsymbol{\theta},x_i),\sigma^2\right]
\end{aligned}
$$

Predictive models have small (more negative) deviance.

Exercise

Write an expression for the deviance of a simple linear regression with 20 continuous, strictly non-negative, data points.

Deviance in AIC

$$
\begin{aligned}
\text{AIC} &= \overbrace{-2\log \mathrm{L}(\hat{\boldsymbol{\theta}})}^{\text{deviance}} + 2K \\
&= -2\log[\boldsymbol{y}|\hat{\boldsymbol{\theta}}] + 2K \\
&= -2\log\left[\mathbf{y}|g(\hat{\boldsymbol{\theta}}, \mathbf{x}), \sigma^2\right] + 2K \\
&= -2\log\prod_{i=1}^{n}\left[y_i|g(\hat{\boldsymbol{\theta}}, x_i), \sigma^2\right] + 2K
\end{aligned}
$$

Note that deviance does not involve prediction. No new values of y are produced and evaluated relative to the data.

What is the interpretation of counting parameters in a Bayesian or a likelihood- based model with informative priors?

DIC, the deviance information criterion

$$\text{DIC} = \hat{D} + 2p_D$$

$\hat{D} = -2\log[\mathbf{y}|\text{E}(\boldsymbol{\theta}|\mathbf{y})]$ = deviance of model evaluated at the means of the parameters

$p_D = \bar{D} - \hat{D}$ =effective number of parameters

$\bar{D}$ =posterior mean of the deviance

$$\begin{aligned} \bar{D} &= \text{E}_{\boldsymbol{\theta}|\mathbf{y}}(-2\log[\mathbf{y}|\boldsymbol{\theta}]) \\ &= \int -2\log[\mathbf{y}|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta} \ . \end{aligned}$$

DIC cannot be interpreted directly. Models with greater predictive ability have lower DIC values.

DIC as a statistical regulator

Smaller DIC values indicate better prediction.

With increasing number of parameters:

$$\text{DIC} = \overbrace{\hat{D}}^{\text{smaller}} + 2(\overbrace{\bar{D} - \underbrace{\hat{D}}_{\text{smaller}}}^{\text{larger}})$$

Implementing DIC

Compute $\bar{D}$ by calculating the model deviance at each iteration of the MCMC algorithm,

$$D^{(k)} = -2\log\left[\mathbf{y}|\boldsymbol{\theta}^{(k)}\right] \tag{1}$$

and finding the mean of D across all of the iterations,
$\bar{D} = \frac{\sum_{k=1}^{K} D^{(k)}}{K}$.
We estimate $\hat{D}$ by calculating the model deviance using the means of the posterior distributions of each of the parameters,
$\hat{D} = -2\log[\mathbf{y} \mid \bar{\boldsymbol{\theta}}]$.

$$\text{DIC} = \overbrace{\hat{D}}^{\text{smaller}} + 2(\overbrace{\bar{D} - \underbrace{\hat{D}}_{\text{smaller}}}^{\text{larger}})$$

When to use DIC

- p_D must be much smaller than n
- Symmetric, unimodal posteriors (no mixture models unless integrated)
- May not be used for "model" weights as is often done for AIC (with little theoretical basis as probabilities)
- Look into the issue of "focus of prediction" when using with hierarchical models.
- Do not be seduced by convenience.

Integrated DIC

Model sans priors

$$y_i \sim \begin{cases} 0 & , \ z_i = 0 \\ \text{binomial}(n_i, p) & , \ z_i = 1 \end{cases}$$

$$z_i \sim \text{Bernoulli}(z_i)$$

Likelihood:

$$[y_i|z_i, p, \boldsymbol{\psi}] = [\mathsf{y}_i|p, \mathsf{n}_i]^{z_i} 1^{1-z_i}_{\{\mathsf{y}_i=0\}} [z_i|\boldsymbol{\psi}]$$

Keeping in mind that when $z_i = 0$, y_i must also equal 0. This simplifies the integration (via sum):

$$[y_i|p, \boldsymbol{\psi}] = \sum_{z_i=0}^{1} [\mathsf{y}_i|p, \mathsf{n}_i]^{z_i} 1^{1-z_i}_{\{\mathsf{y}_i=0\}} [z_i|\boldsymbol{\psi}].$$

Use same code as for regular DIC except substitute this for likelihood.

Posterior predictive loss

$$[y^{new} \mid \mathbf{y}] = \int [y^{new} \mid \boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]\, d\boldsymbol{\theta}$$

$$D_{sel} = \overbrace{\sum_{i=1}^{n} (y_i - \mathsf{E}(y_i^{new}|\mathbf{y}))^2}^{\text{decreases with more parameters}} + \overbrace{\sum_{i=1}^{n} \mathsf{Var}(y_i^{new}|\mathbf{y})}^{\text{increases with more parameters}}$$

Implementing D_{sel}

1. Simulate a new dataset ($\mathbf{y}^{new}$) at each MCMC iteration (in JAGS)
2. Compute the sum of the squared difference between the mean of the y_i^{new} and the y_i (in R).
3. Compute the sum of the the variances of the y_i^{new} across all of the MCMC iterations (in R).
4. Subtract the result in 3 from the result in 2.

When to use D_{sel}

- The Swiss Army knife of Bayesian model selection - works for any model.
- Again, truly Bayesian because it depends on posterior predictive distribution.
- Style points.

Indicator variable selection

Consider a model in the general linear modeling family,

$$\mu_i = g(\beta_0 + \mathbf{x}_i \boldsymbol{\beta})$$
$$y_i \sim [y_i \mid \mu_i, \sigma^2]$$

where $\boldsymbol{\beta}$ is vector of covariates of length p. Evaluating all possible combinations of the coefficients using model selection criteria becomes tedious if there are many coefficients. (Personally, I *hate* this style of modeling.) What is an alternative?

Indicator variable selection

Recast the parameter vector to become

$$\beta_j = \theta_j z_j,\ j = 1, ..., p,$$

where z_j is a zero or one indicator variable and

$$\theta_j \sim \text{normal}(0, 100000)$$
$$z_j \sim \text{Bernoulli}(\phi)$$
$$\phi \sim \text{uniform}(0, 1)$$

The mean of the z_j across all MCMC samples can be interpreted as the relative weight of the j^{th} coefficient, that is, the probability of including the parameter in the model (with standardized covariates). Values close to one indicate that the j^{th} coefficient is an important predictor of the response. Values close zero indicate it is not important. This wraps model selection, model averaging, and variable importance together in a single tidy package.

Indicator variable selection

- But there is a catch. The independent priors on the $\boldsymbol{\beta}$ and ϕ above often lead to MCMC samples that fail to converge if the prior for the θ_j is too vague, although this is not always the case.
- Remedy: Stochastic search variable selection[3].
  - Use a joint prior for θ_j and z_j, $[z_j, \theta_j] = [\theta_j \mid z_j][z_j]$

$$\theta_j | z_j \sim z_j \text{normal}(0, c\varsigma^2) + (1 - z_j)\text{normal}(0, \varsigma^2).$$

  - Tune c and ς^2 so that ς^2 is small creating a spike at zero and $c\varsigma^2$ is larger creating a slab around zero. The slab provide the prior for θ_j when β_j is in the model (i.e. when $z_j = 1$. See example code.

[3] E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88(423):881–889, 1993.

Example JAGS code[4]

```
model{
    alpha ~ dnorm(0, 1) #vague because x's standardized
    sd_y ~ dunif(0, 10)
    tau_y <- pow(sd_y, -2)
    ####### ssvs priors
    sd_bet ~ dunif(0, 10)     #tune as needed
    tau_in <- pow(sd_bet, -2)
   c = 1000 #tune as needed
    tau[1] <- tau_in        # coef effectively zero
    tau[2] <- tau_in / c #nonzero coef
    p_ind[1] <- 1/2
    p_ind[2] <- 1 - p_ind[1]
    for (j in 1:n.coef){  #for each coefficient
      indA[j] ~ dcat(p_ind[]) # returns 1 or 2
      z[j] <- indA[j] - 1   # returns 0 or 1
      beta[j] ~ dnorm(0, tau[indA[j]])
    }
    ###### likelihood
    for (i in 1:nobs){
      Y[i] ~ dnorm(alpha + X[i ,] %*% beta[], tau_y)
    }
    } #end of model
```

[4] https://mbjoseph.github.io/posts/2018-12-27-stochastic-search-variable-selection-in-jags/

Guidance

- Out-of-sample validation: gold standard
- Cross-validation: when computation is feasible
- DIC : Simple Bayesisan models in general linear modeling framework with symmetric posteriors
- Indicator variable selection: When you seek to combine model averaging with model selection and understand relative importance of coefficients.
- Posterior-predictive loss: any Bayesian model

Further study

- Model selection laboratory exercise (optional) in repository
- Folder on indicator variable selection in repository
- BMA package in R
- Hobbs, N. T. and Hooten M. B, Bayesian models: a statistical primer for ecologists. 2015. Princeton University Press. Chapter 9.
- Hooten, M. B., and N. T. Hobbs. 2015. A guide to Bayesian model selection for ecologists. Ecological Monographs 85:3-28.
- Link, W. A., and R. J. Barker. 2006. Model weights and the foundations of multimodel inference. Ecology 87:2626-2635.
- See Link, W. A., and R. J. Barker. 2010. Bayesian Inference with ecological applications. Academic Press, chapter 7
- Ver Hoef, J. M. and P. L. Boveng. 2015. Iterating on a single model is a viable alternative to multimodel inference. The Journal of Wildlife Management, 79(5):719–729