Model Selection

Models for Socio-Environmental Data

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

June 5, 2018

Readings

- Hooten, M. B., and N. T. Hobbs. 2015. A guide to Bayesian model selection for ecologists. Ecological Monographs 85:3-28.
- Hobbs, N. T. and Hooten M. B, Bayesian models: a statistical primer for ecologists. 2015. Princeton University Press. Chapter 9.
- Link, W. A., and R. J. Barker. 2006. Model weights and the foundations of multimodel inference. Ecology 87:2626-2635.
- 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

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? Today
- Model-averaging: How do we use multiple models as basis for inference by giving them weights? See readings.

Bayesian model selection

- Model validation
  - out of sample
  - cross validation
- Statistical regularization and the IC's
  - Deviance information criterion (DIC)
  - Wantanabe-Akaike information criterion (WAIC)
  - Posterior-predictive loss (D_{sel})

Out of sample validation: the gold standard

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

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

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 $\log([\mathbf{y}_{\text{oos}}|\mathbf{y}])$.

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)
```

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, $m = 1, ...M$.
- 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.

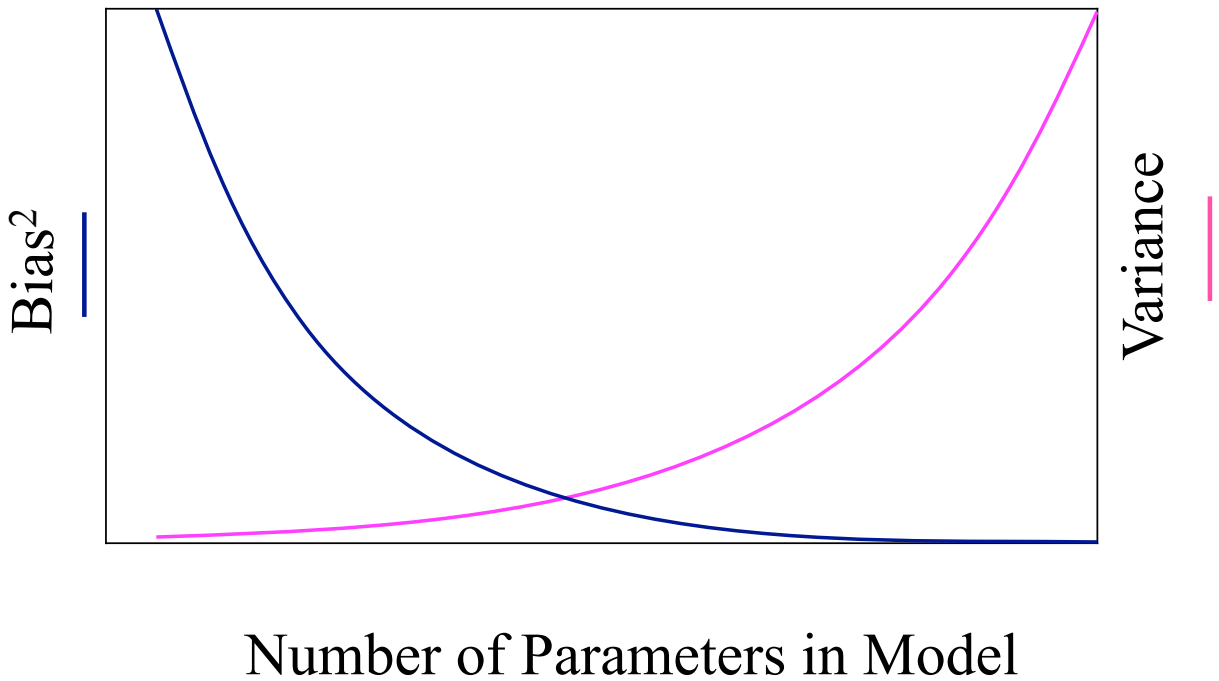
How to set up training and test datasets?

- 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?

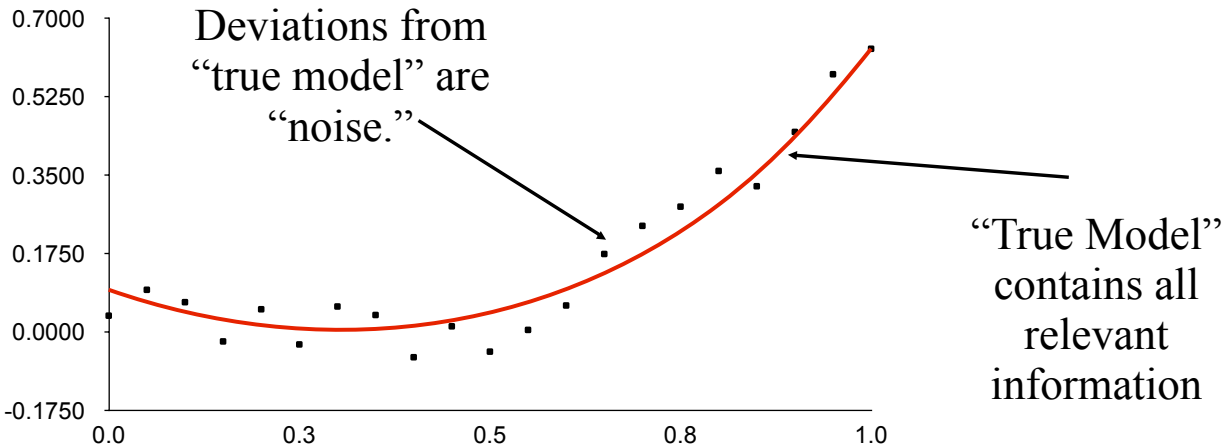
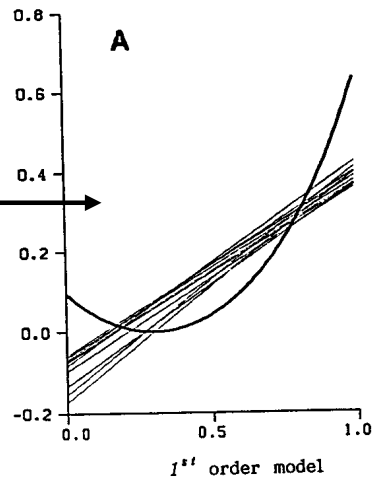
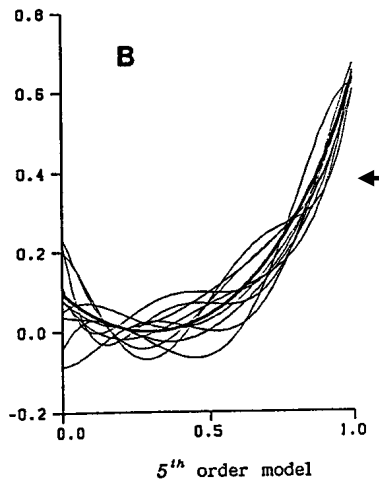


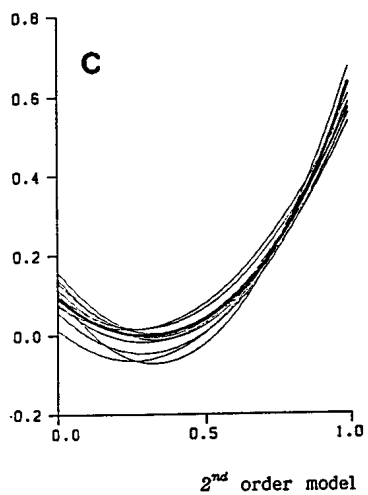
Illustration of trade off

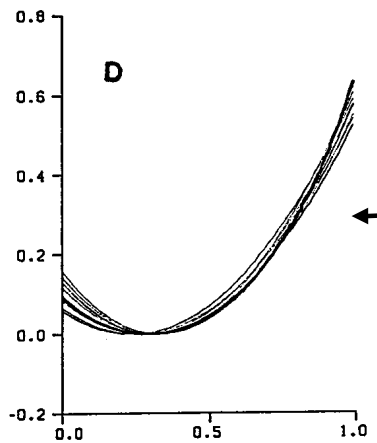
High bias →



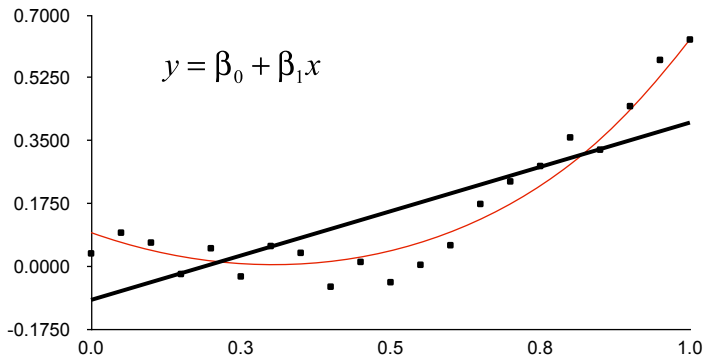


← High variance

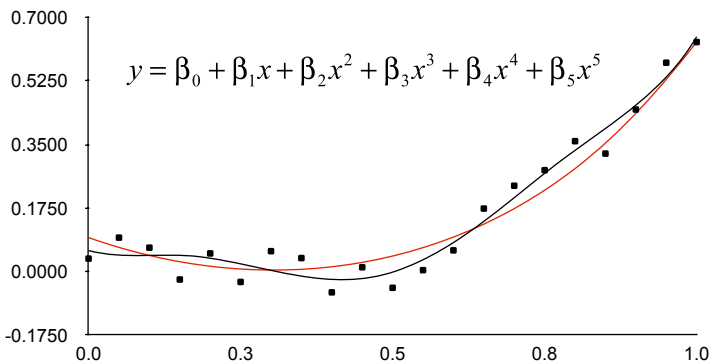




← Optimum, 2nd order with minimum at x=.03



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 $\log[\boldsymbol{\theta}|\mathbf{y}] \propto \log[\mathbf{y}|\boldsymbol{\theta}] + \log[\boldsymbol{\theta}]$
- Priors in penalized MLE $\log L[\boldsymbol{\theta}|\mathbf{y}] = \sum_{i=1}^{n} \log[y_i \mid \boldsymbol{\theta}] + \log(\boldsymbol{\theta})$
- 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)
- 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.

Point-wise predictive score

The Bayesian approach summarizes the posterior predictive distribution to predict new data. This would seem to be the logical basis for Bayesian model selection. However the posterior predictive distribution is not needed to compute DIC. Ideally, we would like

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

a scoring function indicating the ability of the model to predict new data. We could develop a single statistic indicating predictive ability using $\mathrm{E}_{\mathbf{y}^{new}}(\log[\mathbf{y}^{new}|\mathbf{y}])$,

$$\mathrm{E}_{\mathbf{y}^{new}}(\log[\mathbf{y}^{new}|\mathbf{y}]) = \int \log \int [\mathbf{y}^{new}|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta}[\mathbf{y}^{new}]d\mathbf{y}^{new} \ ,$$

but this is problematic because it requires knowing the the true distribution of $\mathbf{y}^{new}$, which is not known.

Point-wise predictive score

Instead of using $\mathrm{E}_{\mathbf{y}^{new}}(\log[\mathbf{y}^{new}|\mathbf{y}])$, we approximate the mean posterior predictive score using log point-wise predictive score where we evaluate the probability or the probability density of each data point using MCMC to compute the integral

$$\begin{aligned}\log\prod_{i=1}^{n}[y_i|\mathbf{y}] &= \sum_{i=1}^{n}\log\int[y_i|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta} \\ &\approx \sum_{i=1}^{n}\log\left(\underbrace{\frac{\sum_{k=1}^{K}\left[y_i|\boldsymbol{\theta}^{(k)}\right]}{K}}_{\text{mean over K interations}}\right).\end{aligned}$$

Effective number of parameters

We are using the same data to compute the scoring function as we use to fit the data, suggesting the need for a regulator

$$
\begin{aligned}
p_D &= \sum_{i=1}^{n} \mathrm{Var}_{\boldsymbol{\theta}|\mathbf{y}}(\log[y_i|\boldsymbol{\theta}]) \\
&\approx \underbrace{\mathrm{Var}(\log[y_i|\boldsymbol{\theta}^{(k)}])}_{\text{variance over K iterations}} .
\end{aligned}
$$

Watanabe-Akaike Information Criterion (WAIC) as a statistical regulator

$$\text{WAIC} = \overbrace{-2\sum_{i=1}^{n}\log\int[y_i|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta}}^{\text{decreases with more parameters}} + \overbrace{2\sum_{i=1}^{n}\text{Var}_{\boldsymbol{\theta}|\mathbf{y}}(\log[y_i|\boldsymbol{\theta}]}^{\text{increases with more parameters}}$$

$$\approx -2\left(\sum_{i=1}^{n}\log\left(\underbrace{\frac{\sum_{k=1}^{K}\left[y_i|\boldsymbol{\theta}^{(k)}\right]}{K}}_{\text{mean over K interations}}\right)\right) + \underbrace{2\text{Var}(\log[y_i|\boldsymbol{\theta}^{(k)}])}_{\text{variance over K iterations}}$$

Implementing WAIC

1. Compute the point-wise predictive score as probability or probability density of each data point conditional on the model $[y_i|\boldsymbol{\theta}]^{(k)}$ at each MCMC iteration. Compute two times the sum of the log of the mean of $[y_i|\boldsymbol{\theta}]^{(k)}$ across the $i = 1, .., n$ data points.
2. Compute $\log[y_i|\boldsymbol{\theta}]^{(k)}$ at each MCMC iteration. Compute p_D as the variance of $\log[y_i|\boldsymbol{\theta}]^{(k)}$

When to use WAIC

1. Truly Bayesian because it is based on posterior predictive distribution.
2. p_D is never negative (as it can be with DIC)
3. Works for mixture models
4. May not be used with temporally or spatially structured data because it depends on assumption that data are independent.

Posterior predictive loss

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

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.

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
- WAIC--any Bayesian model lacking spatial or temporal dependence
- Posterior-predictive loss: any Bayesian model