

# Bayesian Model Selection and Model Averaging

## Bayesian Modeling for Socio-Environmental Data

N. Thompson Hobbs

July 27, 2016



# 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., 2015. The hidden costs of multimodel inference. *Journal of Wildlife Management*. In press.

# Often one model will do...

- 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 inferences conditional on one model.

# Often one model will do...

- 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., 2015. The hidden costs of multimodel inference. *Journal of Wildlife Management*, in press.
- 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.

# Multi-model inference

- Model selection: How do we decide which model is the “best” among a set of candidates?
- Model-averaging: How do we use multiple models as basis for inference by giving them weights?

## Model selection: How do we rank models that have been checked?

- Model validation
  - out-of-sample
  - within-sample
- Statistical regularization (introduction to the “ICs”)
- Likelihood based methods
  - Information theoretics and Akaike information criterion (AIC)
  - Bayesian information criterion (BIC)
- Bayesian methods
  - Deviance information criterion (DIC)
  - Wantanabe-Akaike information criterion (WAIC)
  - Posterior-predictive loss

# Model averaging, briefly

- The probability of the model and Bayes factors
- Indicator variable selection



# How do we compare “checked” models?

- One model is “better” than another one if it can make more accurate predictions.
- All approaches to model selection that we will discuss evaluate the predictive ability of a model.


# Model validation

- Works for maximum likelihood or Bayesian approaches
- Out-of-sample validation based on dataset withheld from model fitting.
- M-fold cross validation withholds sets with M elements from the data used for fitting.
- Both use a scoring function to evaluate model fit.
  - mean square prediction error (MSPE)
  - log predictive density (LPD)

# Out-of-sample validation

out-of-sample observation

prediction out-of-sample observation

$$\text{MSPE} = \sum_{i=1}^n \frac{(\vec{y}_{\text{oos},i} - \hat{y}_{\text{oos},i})^2}{n}$$


# To implement:

- Insert code into your JAGS model that makes a prediction for each out of sample data point.
- Calculate the squared difference between each model prediction and the out-of-sample data point.
- Sum the differences and divide by the number of withheld observations to obtain MSPE
- On the R side, find mean of MSPE.

## Code for MSPE, simple regression

```
for(i in 1:length(y.oos)){  
  y.hat[i]<-B0+B1*x.oos[i]  
  diff.sq<-(y.oos[i]-y.hat[i])^2  
}  
MSPE<-sum(diff.sq[i])/length(y.oos)
```

On R side, include MSPE in you variable list for  
JAGS or coda samples

# Out-of-sample validation

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

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

log predictive density (LPD) is the  
scoring function

# To implement:

- Insert code into your JAGS model that makes a prediction for each out of sample data point.
- Estimate 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  $[\mathbf{y}_{\text{oos}} | \mathbf{y}]$

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

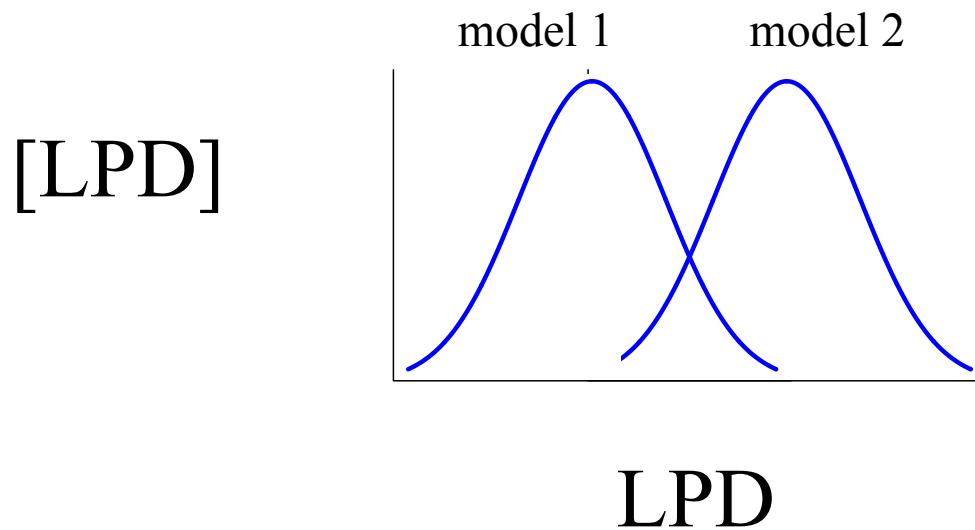
## Code for LPD, simple regression

```
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 you variable list for JAGS or coda samples. Take the log of the mean of PD to get LPD.



# Are scores “different?”



How would you obtain these distributions?

# 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>
- I have code for R function for leave-one-out if you are interested.

# 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.

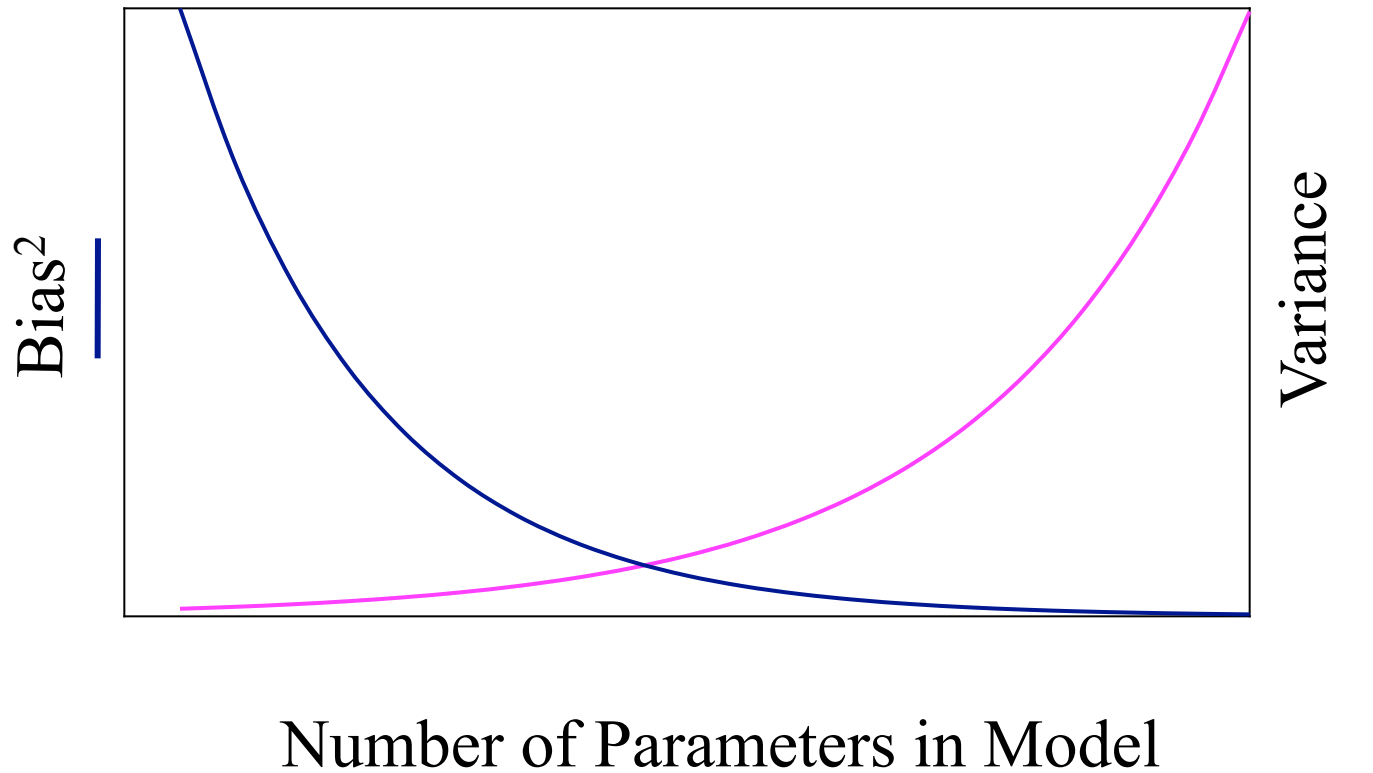
# Statistical regularization

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

# Examples

- The Bayesian prior  $\log [\theta|\mathbf{y}] \propto \log [\mathbf{y}|\theta] + \log [\theta]$
- Priors in MLE  $\log (L (\theta|\mathbf{y})) = \left( \sum_i \log [y_i|\theta] + \log [\theta] \right)$
- Ridge regression
- LASSO
- Information criteria

# 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

$$y = \beta_0 + \beta_1 x$$

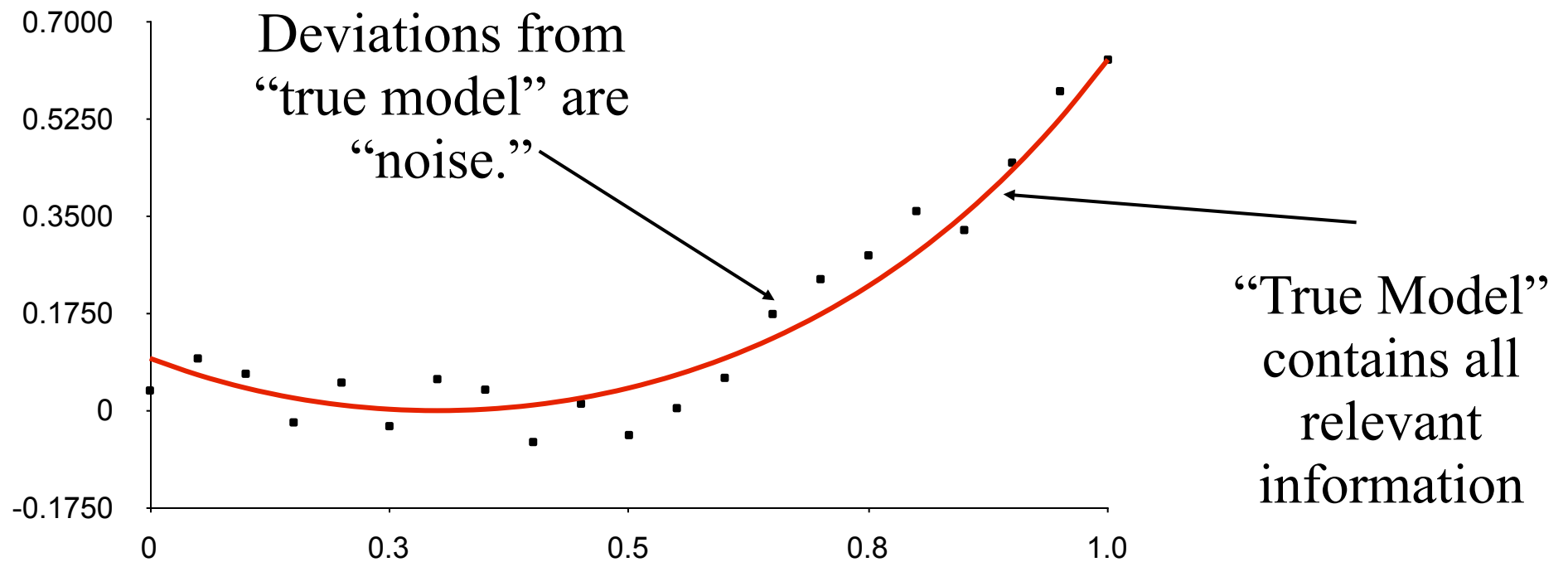
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

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

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

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

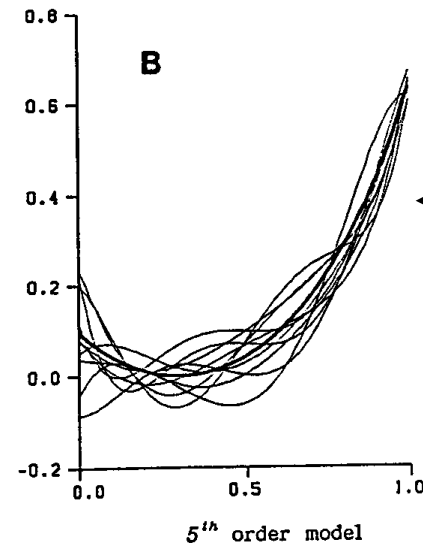
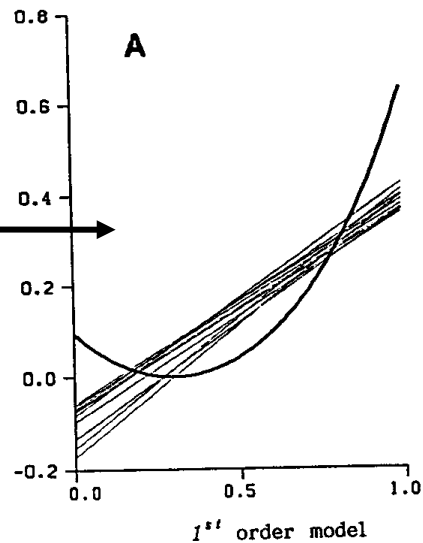
# What creates “noise” in models?



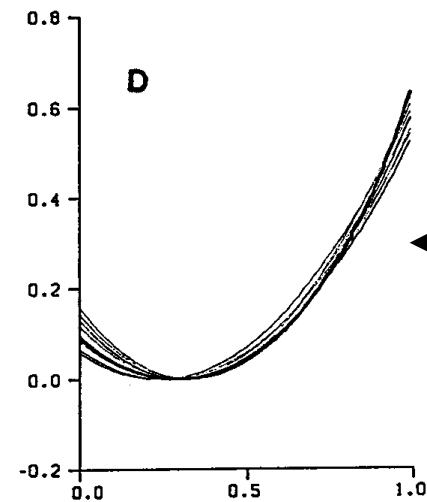
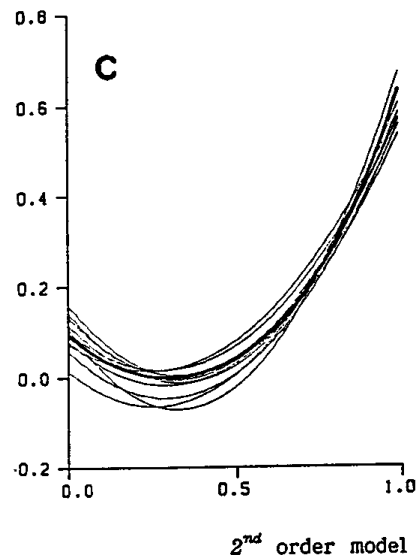


# Illustration of trade off

High bias

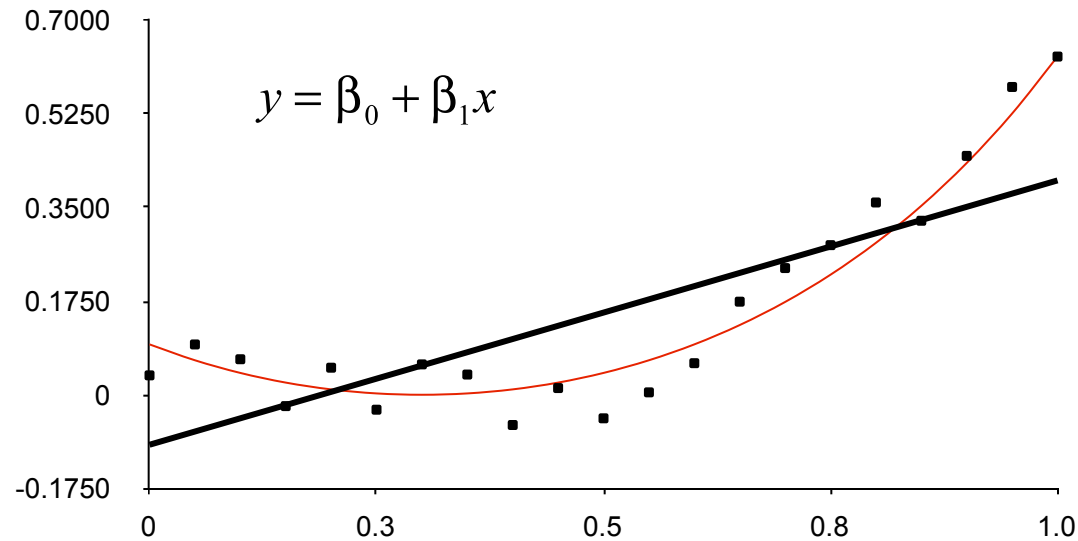


← High variance

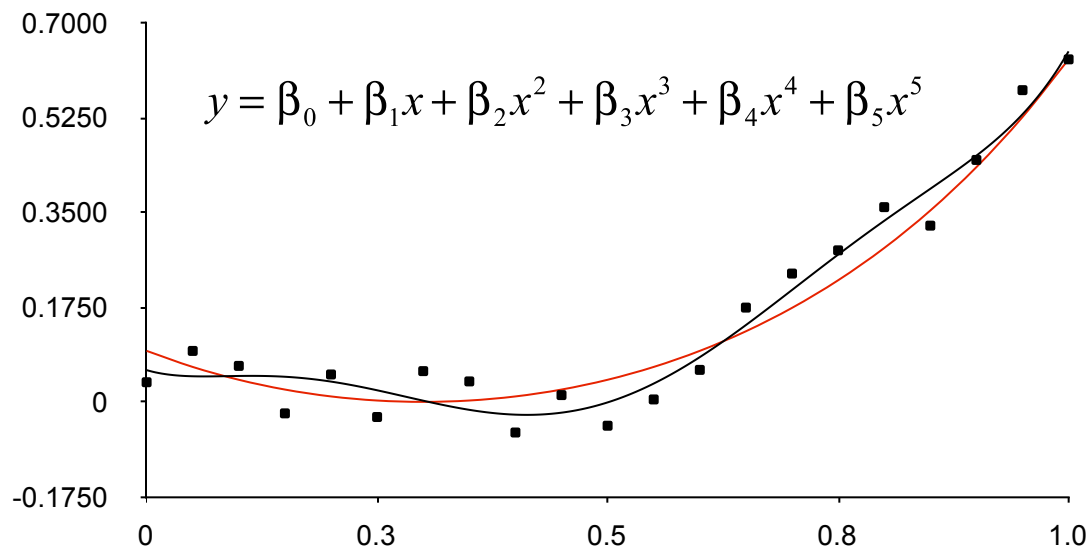


← Optimum, 2nd order with minimum at  $x = .$

03



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



Too many parameters--  
responds to “noise.”  
Variance is high.

# Statistical regularization

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

# Deviance

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

# Deviance in AIC

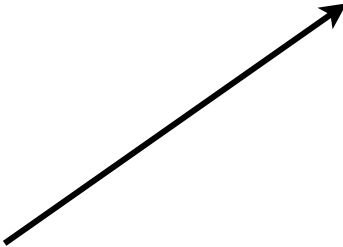
$$\begin{aligned}\text{AIC} &= \overbrace{-2 \log L(\hat{\boldsymbol{\theta}})}^{\text{deviance}} + 2K \\ &= -2 \log[\mathbf{y}|\hat{\boldsymbol{\theta}}] + 2K \\ &= -2 \log \left[ \mathbf{y} | g(\hat{\boldsymbol{\theta}}, \mathbf{x}), \sigma^2 \right] + 2K \\ &= -2 \log \prod_{i=1}^n \left[ y_i | g(\hat{\boldsymbol{\theta}},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

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?

# The deviance information criterion

$$\text{DIC} = D(\bar{\theta}) + 2\text{pd}$$



The deviance of the model evaluated at the means of the posterior distributions of the parameters.



The effective number of parameters

As with AIC, small values indicate higher predictive ability.

# Alternative Notation

$$\text{DIC} = D(\bar{\theta}) + 2pd$$

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

$$\text{DIC} = \overline{D(\theta)} + pd$$



If you have a simple Bayesian model that has a normal likelihood and uninformative priors, what is the relationship between AIC and DIC? Why?

# Some intuition for DIC

- The problem is parameters that are “free” to be influenced by noise in the data---think about the Sakamoto et al. example. How free are they?
- If a prior on a parameter is are very informative--the parameter is not “free” to respond to the data so it does not contribute to the effective number of parameters.
- If a prior is uninformative, the opposite is the case, it is free to respond and contributes to the effective number of parameters in the same way a in a likelihood analysis.
- If a parameter is part of a hierarchy, should it “count” the same way as a parameter that is part of a simple model?

# pD, effective number of parameters

$D(\bar{\theta})$  = deviance evaluated at mean of posterior

$\overline{D(\theta)}$  = posterior mean of the deviance

$\overline{D(\theta)} = E[D(\theta)]$

$$p_d = \overline{D(\theta)} - D(\bar{\theta})$$

You will also see  $D(\bar{\theta})$  notated as  $\hat{D}$

The effective number of parameters is given by the difference between the posterior mean of the deviance and the deviance evaluated at the mean of the posterior estimates of the parameters.

# Seeing DIC as regularization (remember small values indicate better models)

As the number of parameters increases, this gets larger.

$$\text{DIC} = \underbrace{D(\bar{\theta})}_{\substack{\text{As the number of parameters} \\ \text{increases, these get smaller.}}} + 2 \left( \overbrace{\overline{D(\theta)} - \underbrace{D(\bar{\theta})}}^{\substack{\text{As the number of parameters} \\ \text{increases, this gets larger.}}} \right)$$

# Requirements for DIC

- Posterior distributions are symmetric. Will not work for mixture models like zero inflation. Often fails when random variables are beta or gamma distributed.
- Number of observations must be much larger than effective number of parameters.

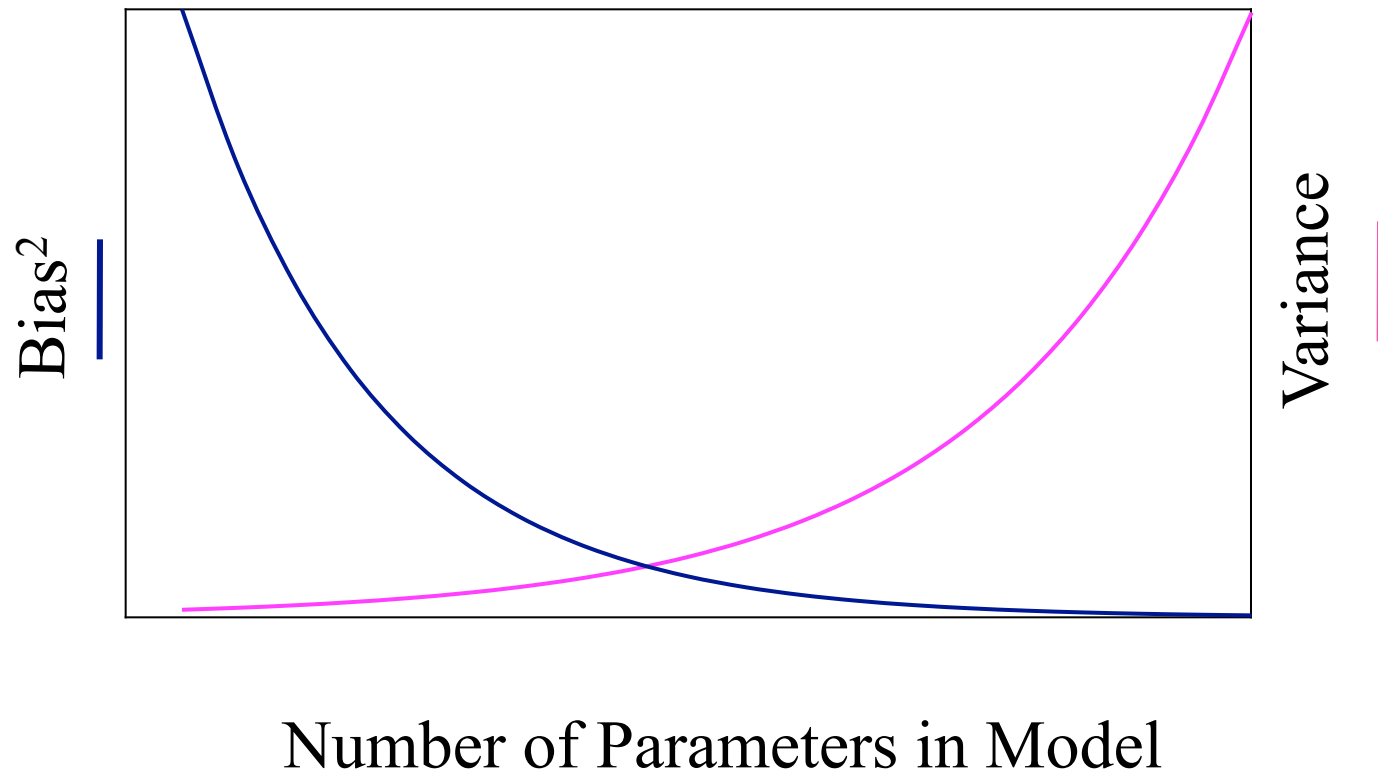
# Posterior predictive loss

$$D_{sel} = \underbrace{\sum_{i=1}^n (y_i - \mathbb{E}(y_i^{new} | \mathbf{y}))^2}_{\text{This decreases as the number of parameters increases.}} + \underbrace{\sum_{i=1}^n \text{Var}(y_i^{new} | \mathbf{y})}_{\text{This decreases then increases as the number of parameters increases.}}$$

This decreases as the number of parameters increases.

This decreases then increases as the number of parameters increases.

$$D_{sel} = \underbrace{\sum_{i=1}^n (y_i - E(y_i^{new} | \mathbf{y}))^2}_{\text{bias declines}} + \underbrace{\sum_{i=1}^n \text{Var}(y_i^{new} | \mathbf{y})}_{\text{variance increases}}$$



# Implementing posterior predictive loss

- Simulate a new observation (e.g. `y.new[i]`) for every data point by reversing the likelihood (See posterior predictive checks to remember how.)
- Include the vector `y.new` in your `jags.samples` variable list.
- On the R side, sum the squared differences between the *mean* of each `y.new[i]` and the corresponding data point.
- Sum the variances of the `y.new[i]` over `i` and subtract from the sum of the squared differences.



# Wantanabe-Akaike Information Criterion (WAIC)

- Truly Bayesian--based on posterior predictive distribution.
- Works for hierarchal models including mixture models (occupancy, mark-recapture, zero-inflation, etc.)
- Should *not* be used for data with structural dependence in data, i.e., spatial and dynamic models.

# Wantanabe-Akaike Information Criterion (WAIC)

$$\begin{aligned}\text{WAIC} &= \overbrace{-2 \sum_{i=1}^n \log \int [y_i | \boldsymbol{\theta}] [\boldsymbol{\theta} | \mathbf{y}] d\boldsymbol{\theta}}^{\text{pointwise predictive score}} + 2p_D \\ &\quad \underbrace{\hspace{10em}}_{\text{posterior predictive distribution}} \\ p_D &= \underbrace{\sum_{i=1}^n \text{Var}_{\boldsymbol{\theta} | \mathbf{y}} (\log [y_i | \boldsymbol{\theta}])}_{\text{effective number of parameters}}\end{aligned}$$

Notice that no new data are simulated here. We use the original data, which means the data *must* be independent.

# WAIC

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

# Implementing WAIC

- Insert a loop in JAGS code that calculates the probability density of each data point (the posterior predictive density, PPD) conditional on the model's prediction at that point. Also calculate the log of PPD.
- On the R side
  - Sum over the log of means of the posterior distribution of the PPD and multiply by -2.
  - Calculate pd as the variance of the log of PPD. Note -- do this by squaring the standard deviation.
  - Subtract 2pd from the sum calculated above.

# Guidance on Model Selection

- Out-of-sample validation: gold standard
- Cross-validation: when computation is feasible
- DIC : Simple Bayesian models in general linear modeling framework with symmetric posteriors
- Posterior-predictive loss: any Bayesian model
- WAIC--any Bayesian model lacking spatial or temporal dependence

# Model averaging, briefly

- The probability of the model and Bayes factors
- Indicator variable selection

# The probability of the model

Recall Bayes' theorem for discrete parameters:

$$P(\theta_i | \mathbf{y}) = \frac{P(\mathbf{y} | \theta_i) P(\theta_i)}{\sum_{j=1}^R P(\mathbf{y} | \theta_j) P(\theta_j)}$$

# The probability of a model

Given a set of  $i=1 \dots R$  models:

$$\mathbf{M} = \{M_1, M_2, \dots, M_R\}$$

The probability of the model (the model weight):

$$P(M_i | \mathbf{y}) = \frac{P(\mathbf{y} | M_i) P(M_i)}{\sum_{j=1}^R P(\mathbf{y} | M_j) P(M_j)}$$

Where model  $M$  is the function defined by the mathematical operations on  $\theta$  and covariates. Thus when we talk about model  $M$ , we are implicitly bringing along the parameters and a functional form. When we condition on  $\theta$ , we are also conditioning on the model.



# Multi-model inference

let  $w_i = P(M_i | \mathbf{y})$

We can incorporate model-selection uncertainty in predictions of our models by calculating a weighted average of the predictions,  $\mu$ :

$$\bar{\mu} = \mu_1 w_1 + \mu_2 w_2 + \mu_3 w_3 + \dots + \mu_r w_r$$

# Implementing probability of model

Reversible jump MCMC: Link, W. A., and R. J. Barker. 2010. Bayesian Inference with Ecological Applications. Academic Press. (Best)

# Indicator variable selection

Consider the basic regression model:

$$y_i \sim \text{normal}(B_0 + \mathbf{x}'_i(\mathbf{B}), \sigma^2)$$

for  $p$  predictor variables.

The coefficients  $(B_0, B_1, \dots, B_p)'$  are

$$B_j = \theta_j z_j, z_j = 0, 1.$$

We conclude that the predictor  $x_j$  is important if the mean of the posterior distribution of  $z_j$  is close to 1.

# Indicator variable selection

The challenge is specifying the prior.  
Independent priors:

$$z_j \sim \text{Bernoulli}(p)$$

$$\theta_j \sim \text{normal}(0, \zeta^2)$$

will not allow convergence.

# Indicator variable selection

So , we specify a joint prior:

$$[z_j, \theta_j] = [\theta_j | z_j][z_j]$$

$$z_j \sim \text{Bernoulli}(\phi)$$

$$\theta_j | z_j \sim z_j \cdot \text{normal}(0, \varsigma^2) + (1 - z_j) \cdot \text{normal}(\mu_{\text{tune}}, \sigma_{\text{tune}}^2)$$

The parameters  $\mu_{\text{tune}}, \sigma_{\text{tune}}^2$  are used to tune the MCMC to assure convergence.

See Hobbs and Hooten, section 9.1.4 for details.

“Multi-model inference is difficult, no matter whether one is a Bayesian or a frequentist. And why shouldn't it be? Choosing among models is central to science and it would be naive to think that the process could be automated, that all we need to do is collect data, gobs of it, and let our information criteria sort it out.”

Link and Barker 2010: 158