Model Selection

ESS 575 Models for Ecological Data

N. Thompson Hobbs

April 11, 2017

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

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

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

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

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

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 $[\mathbf{y}_{\text{oos}} \mid \mathbf{y}]$

- On the R side, take the log of the mean of $[\mathbf{y}_{\text{oos}} \mid \mathbf{y}]$
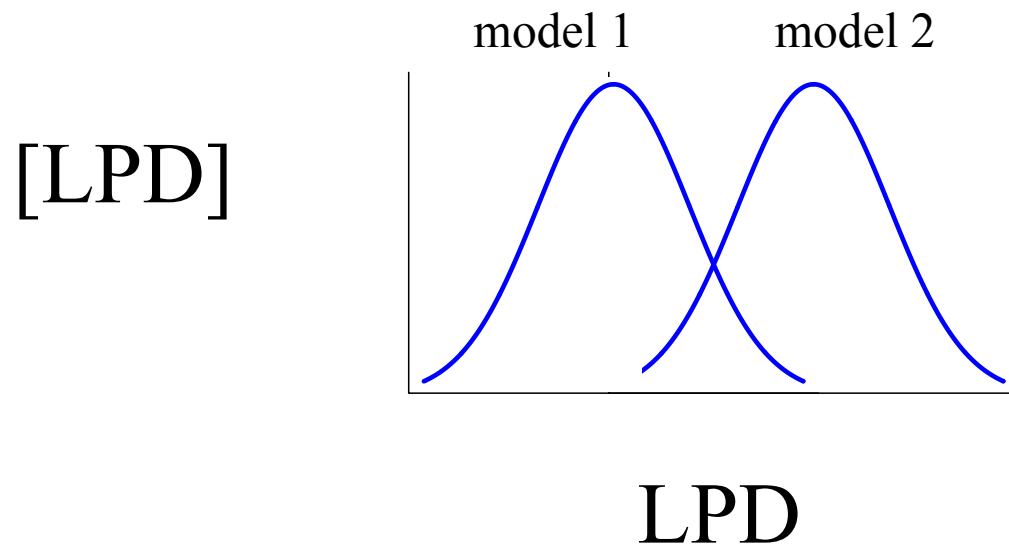
$$\log[\mathbf{y}_{\text{oos}} \mid \mathbf{y}] \approx \log \left(\frac{\sum_{k=1}^{K} [\mathbf{y}_{\text{oos}} \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}]}{K} \right)$$

Code for LPD, simple regression

```r
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.

Are scores "different?"

[LPD]

model 1   model 2

LPD

How would you obtain
these distributions?

M-fold cross validation

- Group the data into M groups.

- Fit model leaving out the observations for each group.

- Calculate predictive score based on ability of model to predict the withheld observations for each group.

- Store the mean of the predictive density for each model fit. Sum the logs of the means.

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

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 an 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.
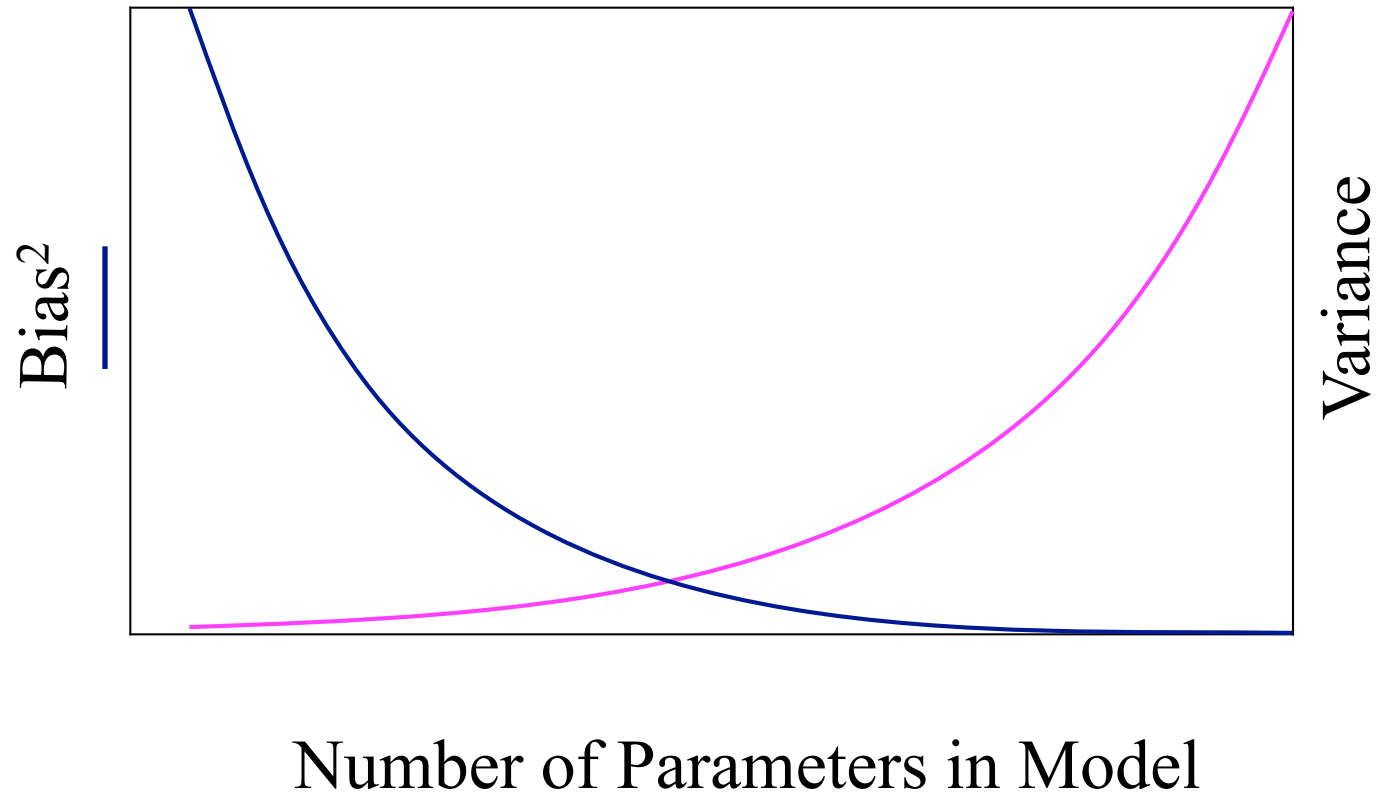
Statistical regularization

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

Examples

- The Bayesian prior $\quad \log\left[\theta|\mathbf{y}\right] \propto \log\left[\mathbf{y}|\theta\right] + \log\left[\theta\right]$

- Priors in MLE $\quad \log\left(L\left(\theta|\mathbf{y}\right)\right) = \left(\sum_i \log\left[y_i|\theta\right] + \log\left[\theta\right]\right)$

- Ridge regression

- LASSO

- Information criteria

LASSO = least absolute shrinkage and selection operator

Statistical Regularization

Bias^2 Variance

Number of Parameters in Model

Sakamoto et al. 1986

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

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

Fit 5 approximating models to the 10 data sets

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

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

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

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

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5$$
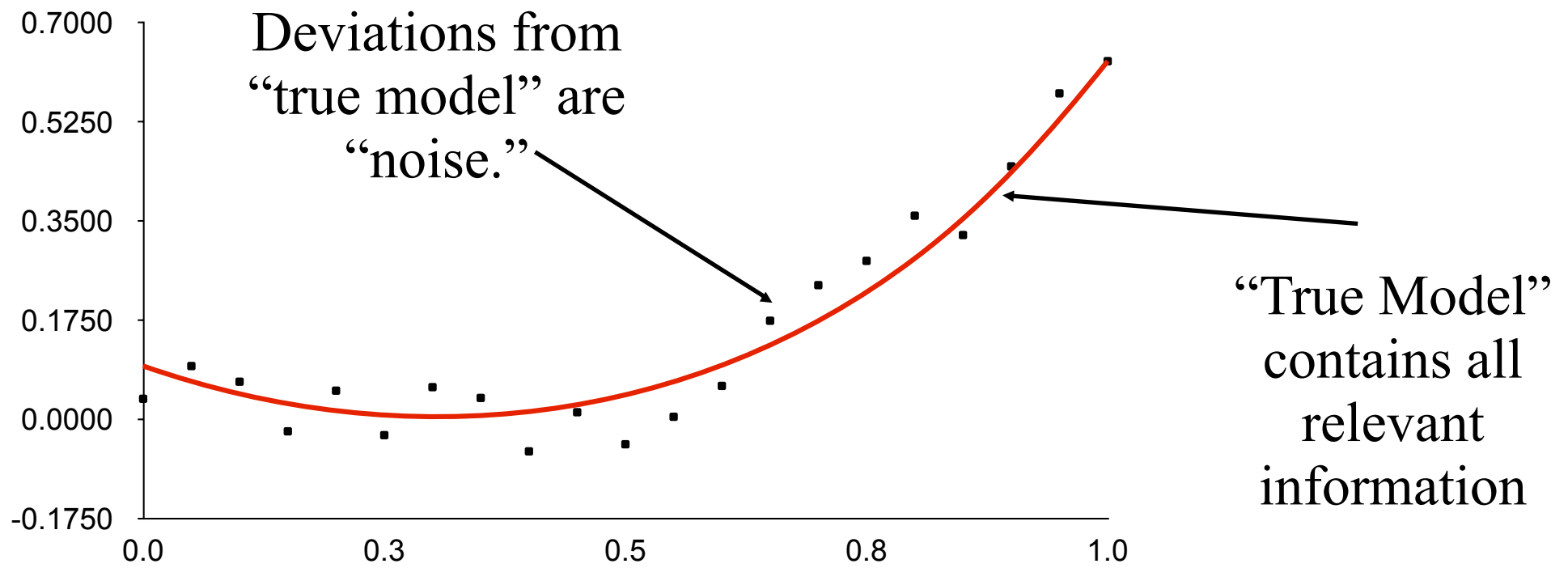
What creates "noise" in models?
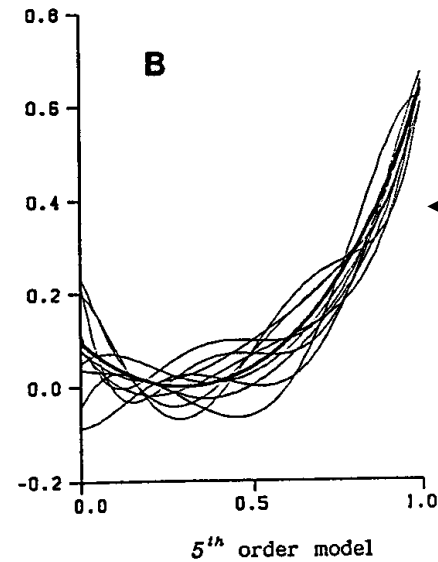

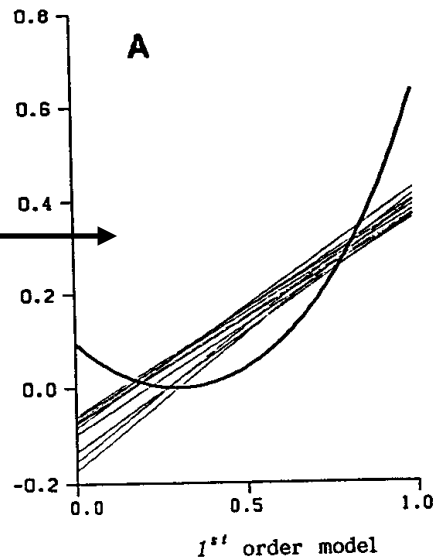Deviations from "true model" are "noise."

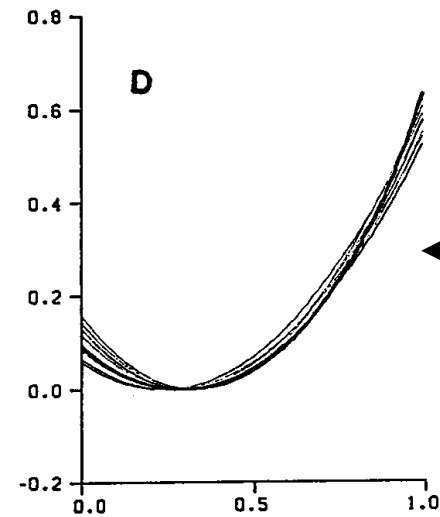"True Model" contains all relevant information

Illustration of trade off

High bias →

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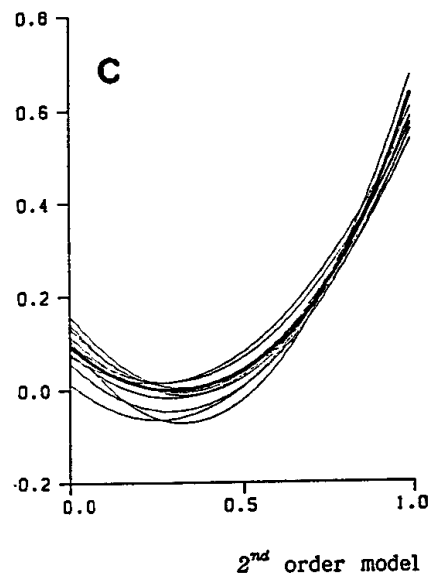
Optimum, 2nd order with minimum at x=.03

$$y = \beta_0 + \beta_1 x$$
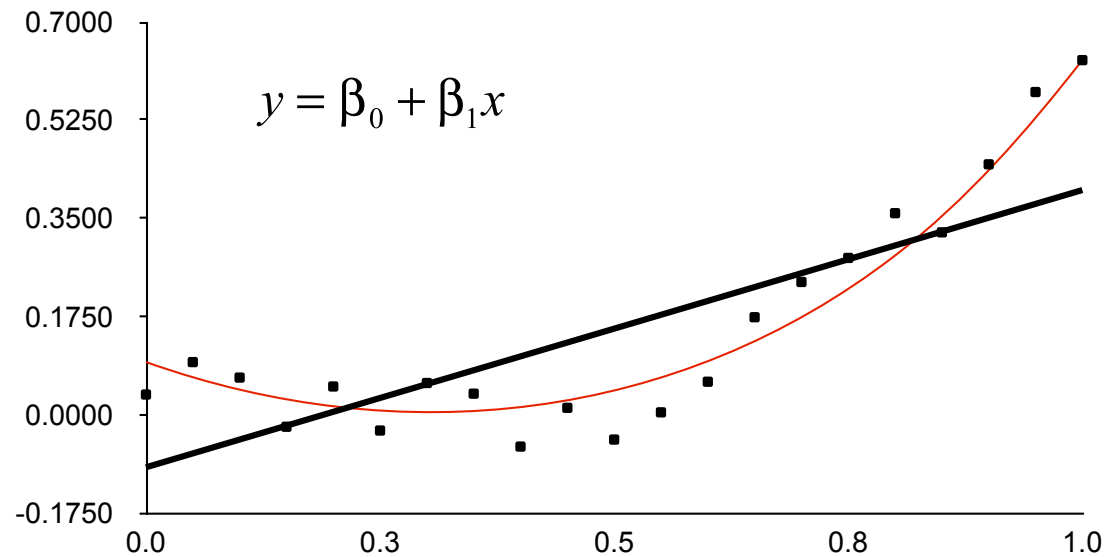
Two few parameters-- fails to respond to information. Bias is high.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5$$
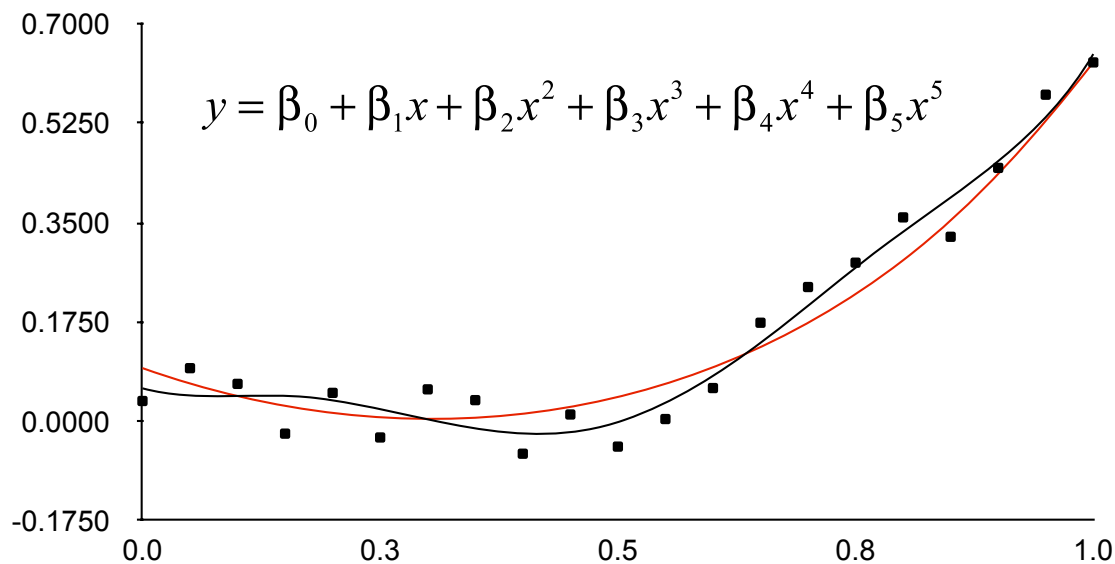
Too many parameters-- responds to "noise." Variance is high.

Deviance

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

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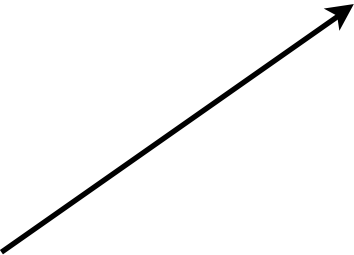
$$
\begin{aligned}
\text{AIC} \quad &= \quad \overbrace{-2\log \text{L}(\hat{\boldsymbol{\theta}})}^{\text{deviance}} + 2K \\
&= \quad -2\log[\boldsymbol{y}|\hat{\boldsymbol{\theta}}] + 2K \\
&= \quad -2\log\left[\mathbf{y}|g(\hat{\boldsymbol{\theta}}, \mathbf{x}), \sigma^2\right] + 2K \\
&= \quad -2\log\prod_{i=1}^{n}\left[y_i|g(\hat{\boldsymbol{\theta}},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

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

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

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

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

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

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

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

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

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

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

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

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

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.

$$\mathrm{DIC} = \underbrace{D\left(\bar{\theta}\right)} + 2\left(\overline{D(\theta)} - \underbrace{D\left(\bar{\theta}\right)}\right)$$

As the number of parameters increases, these get smaller.

Predicting a new observation

Analytically, we need to integrate

$$[y_4^{new}|\mathbf{y}] = \int_{\beta_o} \int_{\beta_1} \int_{\sigma^2} \left[y_4^{new}|\beta_0, \beta_1, \sigma^2\right] \left[\beta_0, \beta_1, \sigma^2|\mathbf{y}\right] d\beta_0 \, d\beta_1 \, d\sigma^2,$$

which is called the *posterior predictive distribution* of y_4. To approximate this formidable integral, we make a draw from

$$y_4^{new(k)} \sim \text{normal}(\beta_0^{(k)} + \beta_1^{(k)} x_4, \sigma^{2(k)})$$

at each MCMC iteration. The posterior predictive distribution of y^{new} is approximated by the elements in its row, a process called *composition sampling.* This is the basis for posterior predictive checks, a topic we will cover soon.
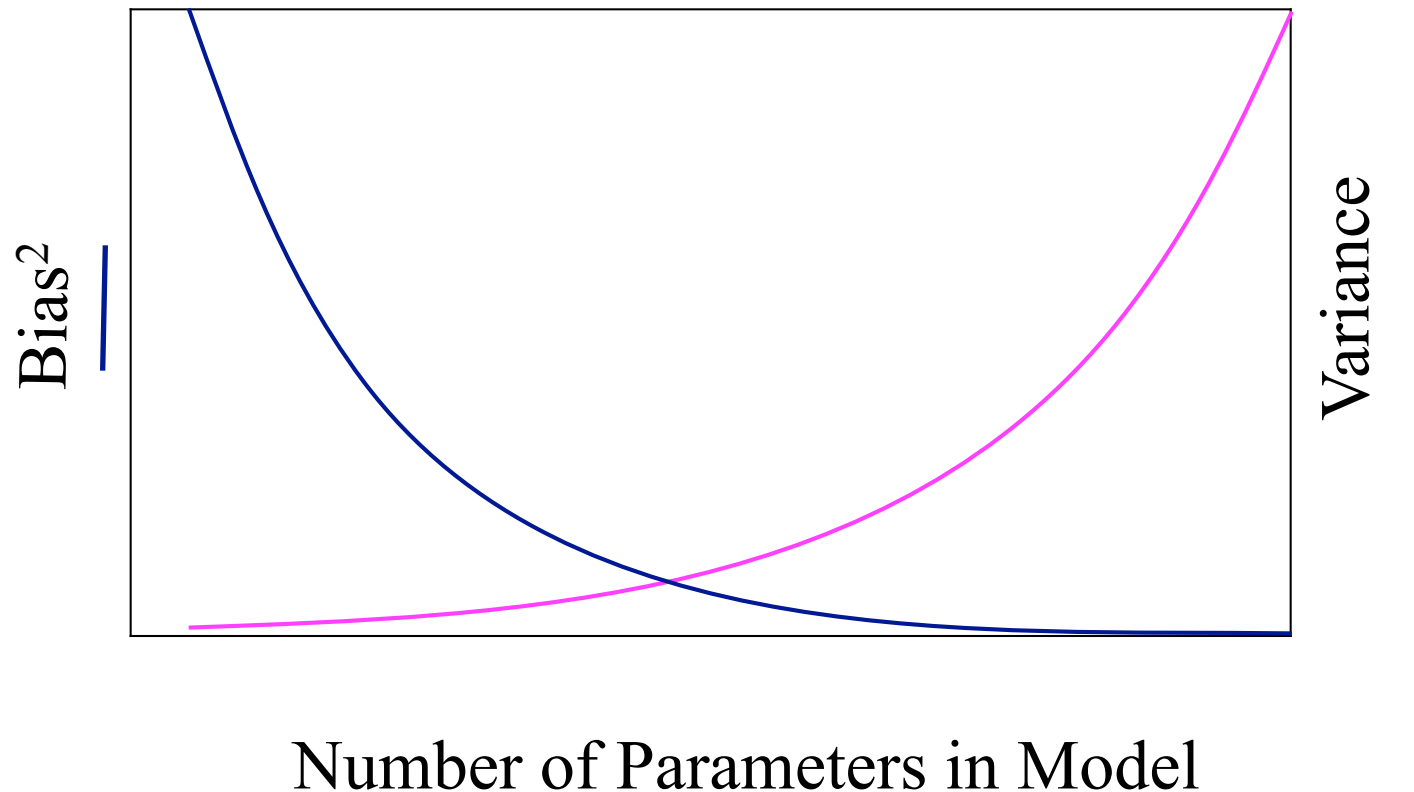
Posterior predictive loss

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

This decreases as the number of parameters increases.

This increases as the number of parameters increases.

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

Number of Parameters in Model

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)

$$\text{WAIC} = \overbrace{-2\sum_{i=1}^{n}\log\underbrace{\int [y_i|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta}}_{\text{posterior predictive distribution}}}^{\text{pointwise predictive score}} + 2p_D$$

$$p_D = \underbrace{\sum_{i=1}^{n}\text{Var}_{\boldsymbol{\theta}|\mathbf{y}}(\log[y_i|\boldsymbol{\theta}])}_{\text{effective number of parameters}}$$

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

WAIC

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

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 if you are using the JAGS object summary function.

  - 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 Bayesisan 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:

$$[\theta|y] = \frac{[y|\theta_i]\,[\theta_i]}{\sum_{\theta_i \in \{\Theta\}} [y|\theta_i]\,[\theta_i]}$$

The probability of a model

Given a set of i=1...R models:

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

The probability of the model (the model weight):

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

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

Multi-model inference

let $w_i = P\left(M_i \mid \mathbf{y}\right)$

We can incorporate model-selection uncertainty in predictions of our models by calcuating a weighted average of the predictions, μ:

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

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 $(B_0, B_1, \ldots B_p)'$ 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:

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

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

will not allow convergence.

Indicator variable selection

So, we specify a joint prior:

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

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

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

The parameters $\mu_{\text{tune}}, \sigma^2_{\text{tune}}$ are used to a tweek 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 should'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