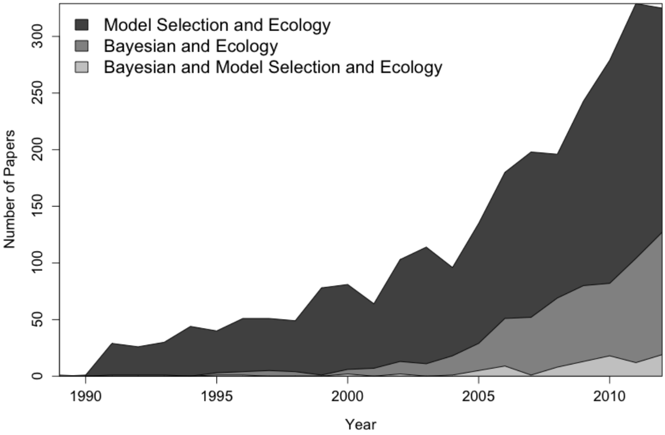
A Guide to Bayesian Model Selection

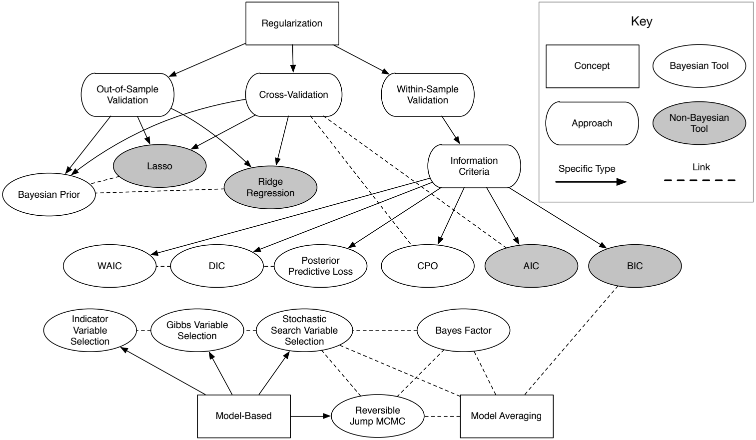
Mevin Hooten

Colorado Cooperative Fish and Wildlife Research Unit
U.S. Geological Survey
—
Department of Fish, Wildlife, and Conservation Biology
Department of Statistics
Colorado State University

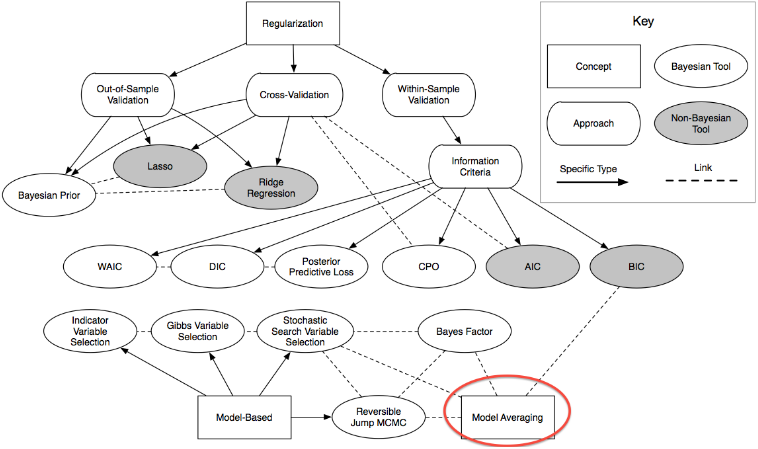
Bayes and Model Selection



Overview


Regularization
Out-of-Sample Validation
Cross-Validation
Within-Sample Validation
Lasso
Ridge Regression
Bayesian Prior
Information Criteria
WAIC
DIC
Posterior Predictive Loss
CPO
AIC
BIC
Indicator Variable Selection
Gibbs Variable Selection
Stochastic Search Variable Selection
Bayes Factor
Model-Based
Reversible Jump MCMC
Model Averaging
Key
Concept
Bayesian Tool
Approach
Non-Bayesian Tool
Specific Type
Link

Bayesian Model Averaging



Average Posterior Distribution

$$[g|\mathbf{y}] = \sum_{l=1}^{L} [g|\mathbf{y}, M_l] P(M_l|\mathbf{y})$$

- Models: $M_1, \ldots, M_l, \ldots, M_L$.
- Quantity of Interest: $g \equiv g(\boldsymbol{\theta}, \tilde{\mathbf{y}})$, function of parameters or data.

Marginal Data Distribution

- Bayes Rule:

$$[\boldsymbol{\theta}|\mathbf{y}] = \frac{[\mathbf{y}|\boldsymbol{\theta}][\boldsymbol{\theta}]}{[\mathbf{y}]}$$

- Evidence:

$$[\mathbf{y}] \equiv [\mathbf{y}|M_l] = \int [\mathbf{y}, \boldsymbol{\theta}|M_l] d\boldsymbol{\theta}$$

Posterior Model Probabilities

$$P(M_l|\mathbf{y}) = \frac{[\mathbf{y}|M_l]P(M_l)}{\sum_{l=1}^{L}[\mathbf{y}|M_l]P(M_l)}$$

- Prior Model Probabilities:

$$P(M_l) \text{ for } l = 1, \ldots, L$$

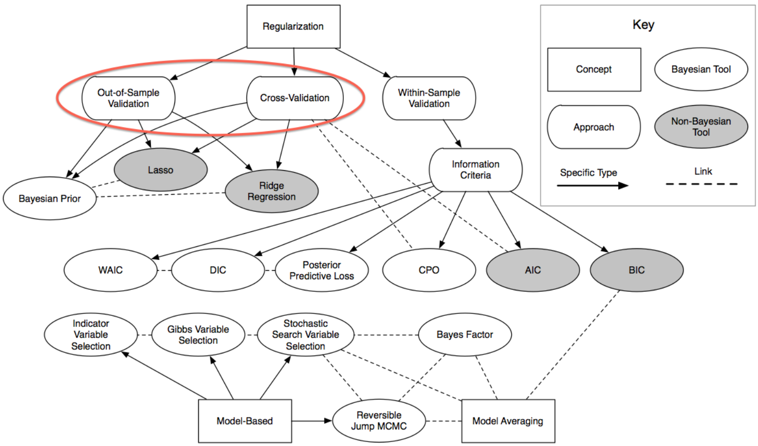
- Bayes Factors:

$$B_{l,l'} = \frac{[\mathbf{y}|M_l]}{[\mathbf{y}|M_{l'}]}$$

Bayesian Model Averaging

- Advantages:
  - Natural and rigorous framework for model averaging.
  - Averaged quantities are less biased with higher precision.
  - Can use prior model probabilities.
- Disadvantages:
  - The marginal data distribution is hard to calculate!
  - Must have proper priors (and a few other caveats).
  - Must choose prior probabilities (can't be lazy).

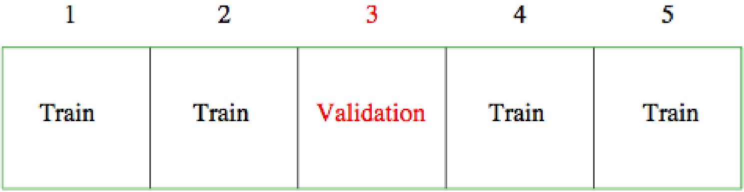
Out-of-Sample Validation



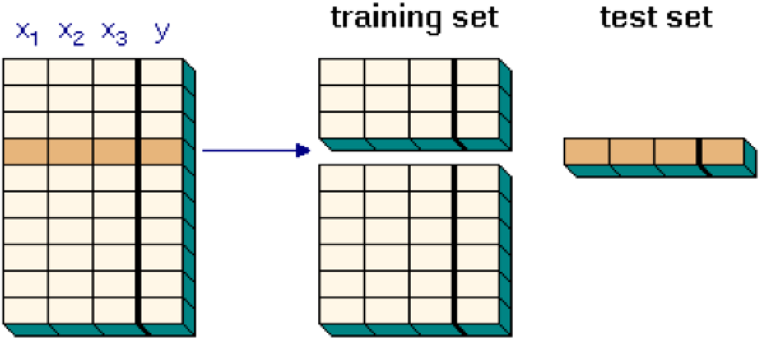
Prediction

- Validation: Out-of-sample data (training and test sets).
- Cross-Validation: Cycle through training and test sets.

Validation



Validation (regression)



Scoring Rules

- Out-of-sample deviance:

$$D(\mathbf{y}_{\mathsf{oos}}, \boldsymbol{\theta}, M_l) = -2\log[\mathbf{y}_{\mathsf{oos}}|\boldsymbol{\theta}, M_l]$$

- Posterior mean of deviance:

$$\bar{D}(\mathbf{y}_{\mathsf{oos}}, M_l) = \int -2\log[\mathbf{y}_{\mathsf{oos}}|\boldsymbol{\theta}, M_l][\boldsymbol{\theta}|\mathbf{y}, M_l]d\boldsymbol{\theta}$$

Posterior Predictive Score

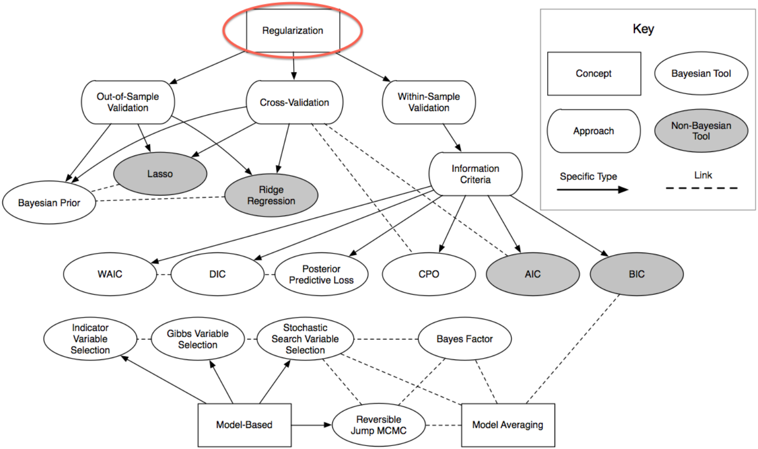
- Log Posterior Predictive Score:

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

- Using cross-validation:

$$\sum_{k=1}^{K}\log\left(\frac{\sum_{t=1}^{T}[\mathbf{y}_k|\mathbf{y}_{-k},\boldsymbol{\theta}^{(t)}]}{T}\right)$$

Regularization



Traditional Regularization

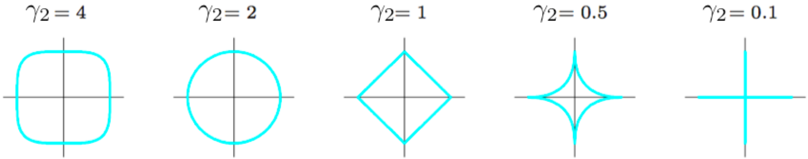
- Linear Regression Model:

$$y_i \sim \mathsf{N}(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$$

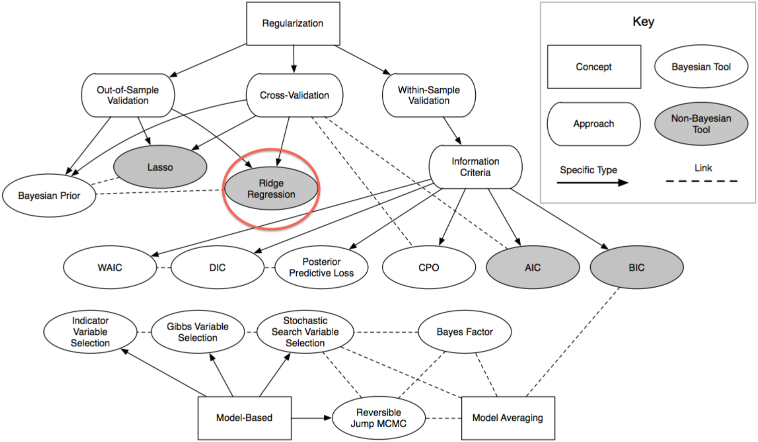
- Add Penalty to Likelihood:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \gamma_1 \sum_{j=1}^{p} |\beta_j|^{\gamma_2}$$

Regulators



Ridge Regression

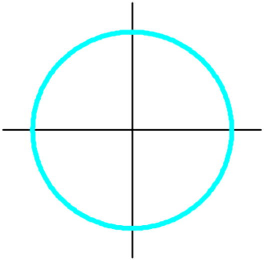


Ridge Regression

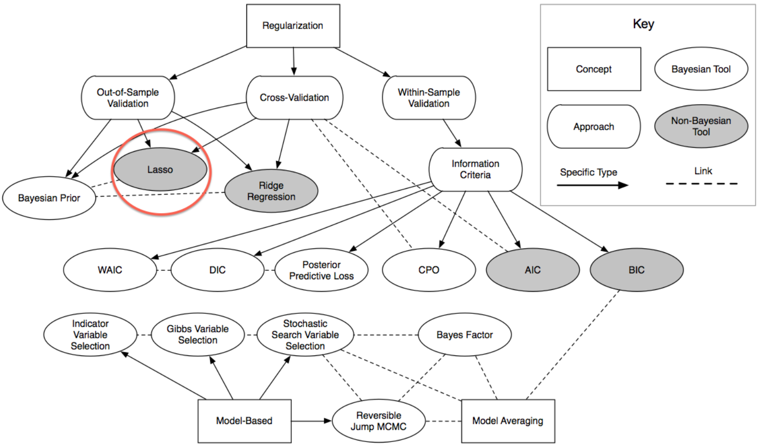
- Let $\gamma_2 = 2$.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \gamma_1 \sum_{j=1}^{p} \beta_j^2$$

- Notice: as $\gamma_1 \to \infty$, the constraint gets stronger, and $\boldsymbol{\beta} \to \mathbf{0}$



Lasso

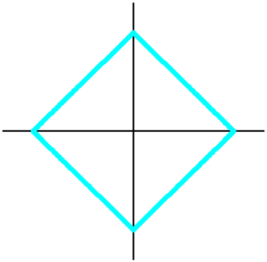

Regularization
Out-of-Sample Validation
Cross-Validation
Within-Sample Validation
Lasso
Bayesian Prior
Ridge Regression
Information Criteria
WAIC
DIC
Posterior Predictive Loss
CPO
AIC
BIC
Indicator Variable Selection
Gibbs Variable Selection
Stochastic Search Variable Selection
Bayes Factor
Model-Based
Reversible Jump MCMC
Model Averaging
Key
Concept
Bayesian Tool
Approach
Non-Bayesian Tool
Specific Type
Link

Lasso

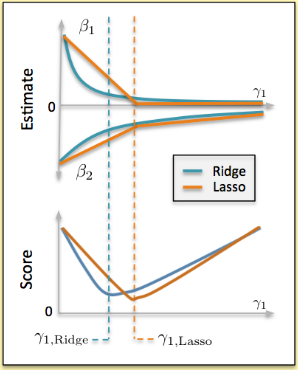
- Let $\gamma_2 = 1$.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \gamma_1 \sum_{j=1}^{p} |\beta_j|$$

- Notice: as $\gamma_1 \to \infty$, the constraint gets stronger, and $\boldsymbol{\beta} \to \mathbf{0}$

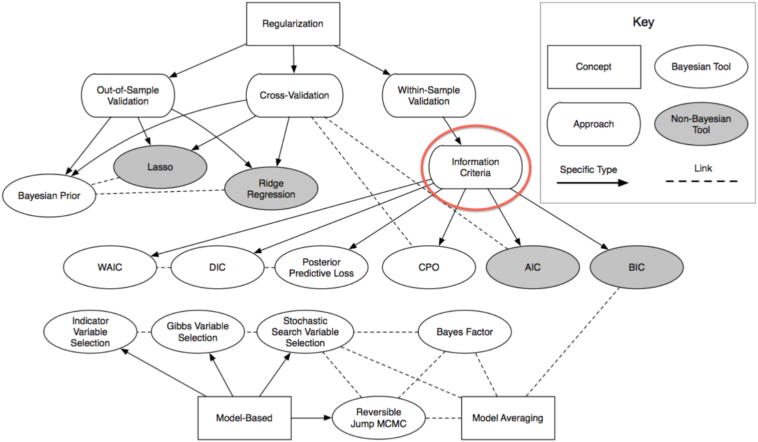


How to get γ_1?



- Parameter estimates are found for a range of shrinkage parameter values (i.e., penalties). In the example at left, there are two parameters in the model.
- Ridge and Lasso provide different shrinkage trajectories due to their different penalty functions.
- Lasso estimates shrink to zero exactly at higher penalties; ridge estimates are asymptotic.
- Out-of-sample predictions are obtained for the model fit at each shrinkage value.
- The parameter estimates at the best predictive "score" are retained for inference.
- Scores are typically presented in terms of deviance, where smaller values are better.

Information Criteria



No out-of-sample data?

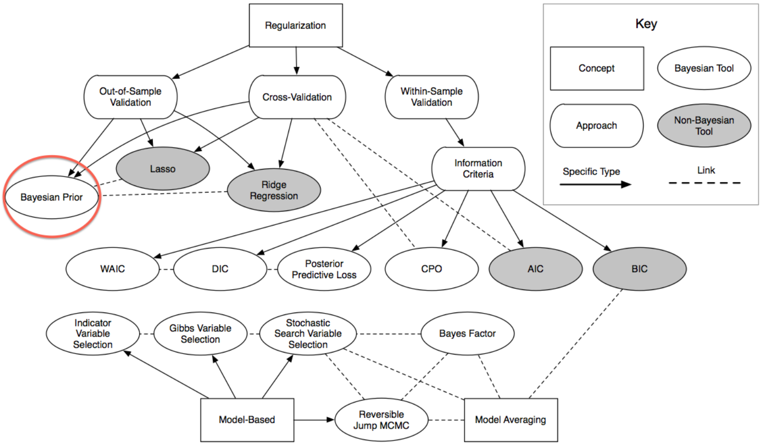
- AIC: $\gamma_1 = 2$ and $\gamma_2 = 0$

$$\text{penalty} = 2 \sum_{j=1}^{p} |\beta_j|^0$$

- BIC: $\gamma_1 = \log(n)$ and $\gamma_2 = 0$

$$\text{penalty} = \log(n) \sum_{j=1}^{p} |\beta_j|^0$$

Bayesian Regularization



How is this Bayesian?

- Linear Regression Model:

$$
\begin{aligned}
y_i &\sim \mathsf{N}(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}, \sigma^2) \\
\beta_0 &\sim \mathsf{N}(\mu_0, \sigma_0^2) \\
\boldsymbol{\beta} &\sim \mathsf{N}(\mathbf{0}, \sigma_\beta^2 \mathbf{I}) \\
\sigma^2 &\sim \mathsf{IG}(q, r)
\end{aligned}
$$

- Posterior:

$$
\begin{aligned}
[\beta_0, \boldsymbol{\beta}, \sigma^2|\mathbf{y}] &\propto [\mathbf{y}|\beta_0, \boldsymbol{\beta}, \sigma^2][\beta_0][\boldsymbol{\beta}][\sigma^2] \\
&\propto \prod_{i=1}^{n} \mathsf{N}(y_i|\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)\mathsf{N}(\beta_0|\mu_0, \sigma_0^2) \prod_{j=1}^{p} \mathsf{N}(\beta_j|\mu_j, \sigma_\beta^2)\mathsf{IG}(\sigma^2|q, r)
\end{aligned}
$$

Bayesian Regularization

- Full-Conditional for $\boldsymbol{\beta}$:

$$[\boldsymbol{\beta}|\cdot] \propto \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \frac{\sigma^2}{\sigma_\beta^2}\sum_{j=1}^{p}|\beta_j|^2\right)\right)$$

Bayesian Regularization

- Full-Conditional for $\boldsymbol{\beta}$:

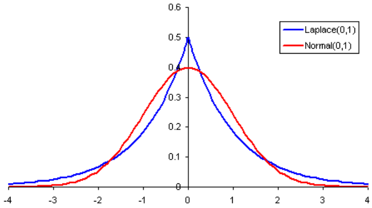
$$[\boldsymbol{\beta}|\cdot] \propto \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \frac{\sigma^2}{\sigma_\beta^2}\sum_{j=1}^{p}|\beta_j|^2\right)\right)$$

- $\gamma_1 = \sigma^2/\sigma_\beta^2$
- $\gamma_2 = 2$

Don't like the penalty?

- Lasso:

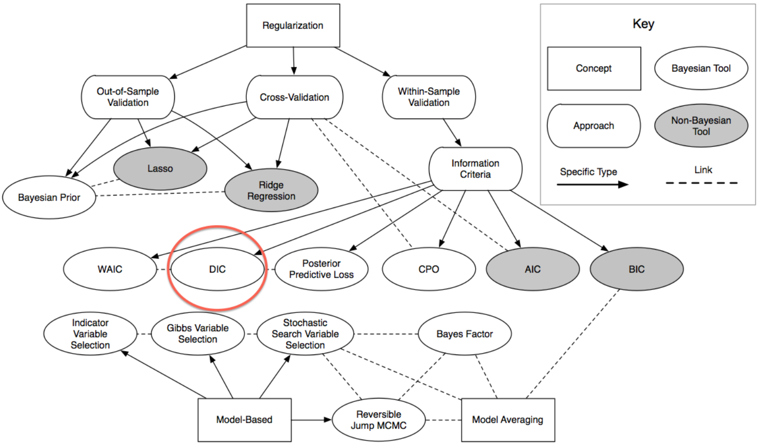
$$\gamma_2 = 1$$



- Implies the prior:

$$\beta_j \sim \text{Laplace}(\mu = 0, \sigma_\beta^2) \propto \exp\left(-\frac{|\beta_j|}{\sqrt{\sigma_\beta^2}}\right)$$

DIC


Regularization
Out-of-Sample Validation
Cross-Validation
Within-Sample Validation
Lasso
Bayesian Prior
Ridge Regression
Information Criteria
WAIC
DIC
Posterior Predictive Loss
CPO
AIC
BIC
Indicator Variable Selection
Gibbs Variable Selection
Stochastic Search Variable Selection
Bayes Factor
Model-Based
Reversible Jump MCMC
Model Averaging
Key
Concept
Bayesian Tool
Approach
Non-Bayesian Tool
Specific Type
Link

A “Bayesian” information criterion

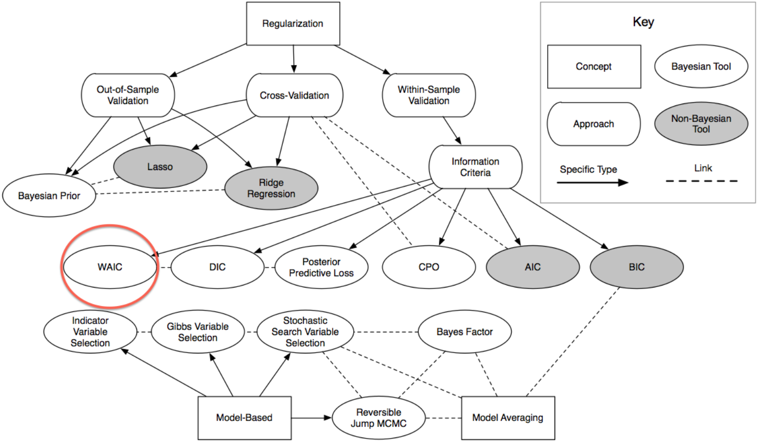
$$\begin{aligned}\text{DIC} &= -2\log[\mathbf{y}|E(\boldsymbol{\theta}|\mathbf{y})] + 2p_D \\ &= \hat{D} + 2p_D\end{aligned}$$

- $p_D = \bar{D} - \hat{D}$
- $\bar{D} = E_{\boldsymbol{\theta}|\mathbf{y}}(-2\log[\mathbf{y}|\boldsymbol{\theta}])$

Notes on DIC

- $p_D < 0$ may occur when the mean does not represent the posterior.
- DIC is not consistent (like AIC).
- DIC is not good when $p_D > n$.
- DIC seems to be ok for the same models AIC works with.
- No theoretical basis for use with BMA.
- Doesn't use the posterior predictive distribution.

WAIC


Regularization
Out-of-Sample Validation
Cross-Validation
Within-Sample Validation
Lasso
Ridge Regression
Bayesian Prior
Information Criteria
WAIC
DIC
Posterior Predictive Loss
CPO
AIC
BIC
Indicator Variable Selection
Gibbs Variable Selection
Stochastic Search Variable Selection
Bayes Factor
Model-Based
Reversible Jump MCMC
Model Averaging
Key
Concept
Bayesian Tool
Approach
Non-Bayesian Tool
Specific Type
Link

A real Bayesian information criterion

$$\text{WAIC} = -2\sum_{i=1}^{n}\log\int[y_i|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta} + 2p_D$$

- $p_D = \sum_{i=1}^{n} \text{var}_{\boldsymbol{\theta}|\mathbf{y}}(\log[y_i|\boldsymbol{\theta}])$
- $\sum_{i=1}^{n}\log\int[y_i|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta} = \log\prod_{i=1}^{n}[y_i|\mathbf{y}]$

Computing WAIC using MCMC

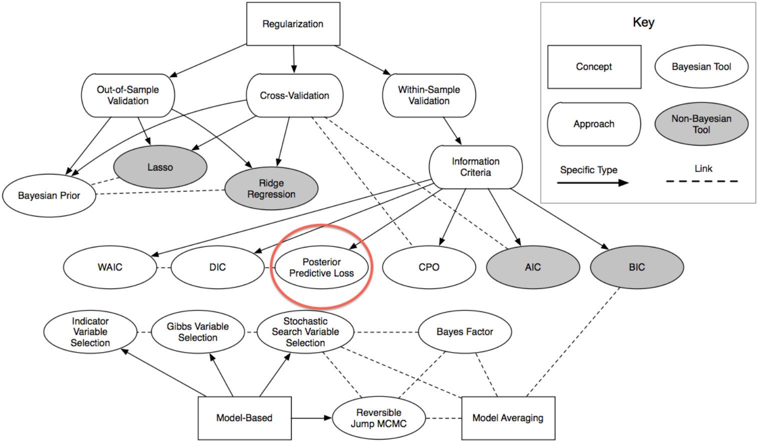
$$\sum_{i=1}^{n} \log \int [y_i|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta} \approx \sum_{i=1}^{n} \log \frac{\sum_{t=1}^{T}[y_i|\boldsymbol{\theta}^{(t)}]}{T}$$

$$\sum_{i=1}^{n} \mathsf{var}_{\boldsymbol{\theta}|\mathbf{y}}(\log[y_i|\boldsymbol{\theta}]) \approx \sum_{i=1}^{n} \frac{\sum_{t=1}^{T}(\log[y_i|\boldsymbol{\theta}^{(t)}] - \sum_{t=1}^{T} \log[y_i|\boldsymbol{\theta}^{(t)}]/T)^2}{T}$$

Notes on WAIC

- Based on posterior predictive distribution.
- $p_D > 0$.
- Works for hierarchical models.
- Product PPD assumes independence.

PPL


Regularization
Out-of-Sample Validation
Cross-Validation
Within-Sample Validation
Lasso
Bayesian Prior
Ridge Regression
Information Criteria
WAIC
DIC
Posterior Predictive Loss
CPO
AIC
BIC
Indicator Variable Selection
Gibbs Variable Selection
Stochastic Search Variable Selection
Bayes Factor
Model-Based
Reversible Jump MCMC
Model Averaging
Key
Concept
Bayesian Tool
Approach
Non-Bayesian Tool
Specific Type
Link

Posterior predictive risk

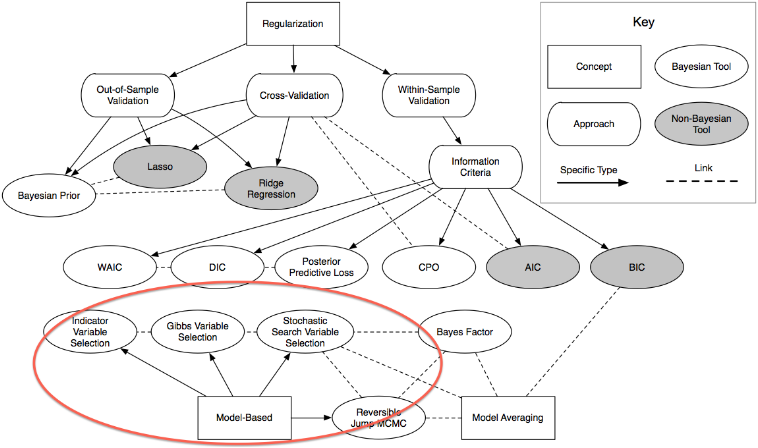
- General:

$$D_w = \sum_{i=1}^{n} \min_{\hat{y}_i} \int (L(\tilde{y}_i, \hat{y}_i) + wL(y_i, \hat{y}_i))[\tilde{y}_i|\mathbf{y}]d\tilde{y}_i$$

- Using squared error loss and $w \to \infty$:

$$D_{\infty,\mathsf{sel}} = \sum_{i=1}^{n}(y_i - \mathsf{E}(\tilde{y}_i|\mathbf{y}))^2 + \sum_{i=1}^{n} \mathsf{Var}(\tilde{y}_i|\mathbf{y})$$

MBMS


Regularization
Out-of-Sample Validation
Cross-Validation
Within-Sample Validation
Lasso
Ridge Regression
Bayesian Prior
Information Criteria
WAIC
DIC
Posterior Predictive Loss
CPO
AIC
BIC
Indicator Variable Selection
Gibbs Variable Selection
Stochastic Search Variable Selection
Bayes Factor
Model-Based
Reversible Jump MCMC
Model Averaging
Key
Concept
Bayesian Tool
Approach
Non-Bayesian Tool
Specific Type
Link

Automatic model selection

- Indicator Variable Selection
- Gibbs Variable Selection
- Stochastic Search Variable Selection
- Reversible-Jump MCMC

Indicator Variable Selection

$$y_i \sim \mathsf{N}(\beta_0 + \mathbf{x}_i' \boldsymbol{\beta}, \sigma^2)$$

- $\beta_j = z_j \cdot \theta_j$ for $j = 1, \ldots, p$.

$$z_j \sim \mathsf{Bern}(\phi)$$
$$\theta_j \sim \mathsf{N}(0, \tau^2)$$

Indicator Variable Selection

$$y_i \sim \mathsf{N}(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$$

- $\beta_j = z_j \cdot \theta_j$ for $j = 1, \ldots, p$.

$$z_j \sim \mathsf{Bern}(\phi)$$
$$\theta_j \sim \mathsf{N}(0, \tau^2)$$

- When $z_j = 0$ in MCMC, sample θ_j from its prior.
- Future $z_j = 1$ will be unlikely if τ^2 is large (try GVS or SSVS).

Reversible-Jump MCMC

- For model M_l, we have parameters $\boldsymbol{\beta}_l$, with varying dimensions p_l.
- RJMCMC puts prior on model index l or model dimension p_l.

$$[\boldsymbol{\theta}_l|\mathbf{y}] \propto [\mathbf{y}|\boldsymbol{\beta}_l, l][\boldsymbol{\beta}_l|l][l]$$

Reversible-Jump MCMC

- For model M_l, we have parameters $\boldsymbol{\beta}_l$, with varying dimensions p_l.
- RJMCMC puts prior on model index l or model dimension p_l.

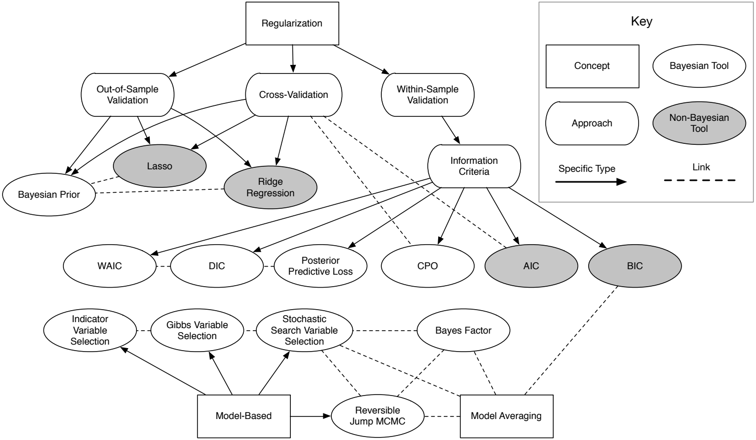
$$[\boldsymbol{\theta}_l|\mathbf{y}] \propto [\mathbf{y}|\boldsymbol{\beta}_l, l][\boldsymbol{\beta}_l|l][l]$$

- MCMC is complicated because the model dimension p_l changes on each iteration.
- RJMCMC is “reversible” because the M-H ratio is modified to allow for moves back to certain model dimensions.

RJMCMC and $P(M_l|\mathbf{y})$

- $P(M_l|\mathbf{y})$ proportional to number of visits to each model in the RJMCMC algorithm.
- RJMCMC can be tricky to program.
- Gibbs and stochastic search variable selection are related but sidestep the transdimensional issue.
- Barker and Link (2013) describe a method that yields RJMCMC results using a two stage process:
  1. Fit models individually.
  2. $P(M_l|\mathbf{y})$ using a second MCMC algorithm and results from individual model fits.

Summary


Regularization
Out-of-Sample Validation
Cross-Validation
Within-Sample Validation
Lasso
Ridge Regression
Bayesian Prior
Information Criteria
WAIC
DIC
Posterior Predictive Loss
CPO
AIC
BIC
Indicator Variable Selection
Gibbs Variable Selection
Stochastic Search Variable Selection
Bayes Factor
Model-Based
Reversible Jump MCMC
Model Averaging
Key
Concept
Bayesian Tool
Approach
Non-Bayesian Tool
Specific Type
Link

Planning a new study?

- Collect two sets of data:
  1. Training.
  2. Validation.

- When prediction is important, there is no substitute for out-of-sample data.

- Time for a paradigm shift in study design?

Historical data set?

- If n is large and you have plenty of time:
  - K-fold cross-validation.
  - Try parallel computing.
- If n is small:
  - Leave-one-out cross-validation.
  - All methods have problems when $n \to 0$.

Want to predict, but not much time?

- If non-hierarchical, consider DIC:
  - DIC is similar to AIC, but for Bayesian models.
  - DIC is not good for multimodal posteriors.
  - $p_D << n$.
- If hierarchical, consider WAIC:
  - WAIC is similar to DIC and AIC for Bayesian models.
  - WAIC works with multimodal posteriors.
  - If data are dependent, try posterior predictive loss or ask for an extension (then do cross-validation).

Want to do model averaging?

- Compute Bayes factors directly:
  - Can be computationally difficult.
  - Allows you to specify $P(M_l)$.
  - Watch out for collinearity and improper priors on parameters (Cade, 2015).
- Use BIC:
  - Only if using the posterior mode with uniform priors on parameters.
  - Assumes prior model probabilities are equal.
- Use RJMCMC:
  - Assumes prior model probabilities are equal.
  - Good luck with the programming!
  - Could try Barker and Link (2013) method.

Want automatic procedure?

- Indicator variable selection:
  - Independent priors require no tuning.
  - MCMC mixing could be poor.
- Gibbs variable selection:
  - Requires tuning, but improved mixing.
  - Tuning doesn't influence posterior.
- Stochastic search variable selection:
  - Requires tuning.
  - Tuning influences posterior, but MCMC may mix better.

References

- 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 M.B. Hooten (2015). Bayesian Models: A Statistical Primer for Ecologists. Princeton University Press.