Objective Bayesian Model Selection: Some Methods, Some Theory

George Casella
Department of Statistics
University of Florida
casella@ufl.edu

Elías Moreno
Department of Statistics
University of Granada

F. J. Girón
Department of Statistics
University of Malaga
Objective Bayes Model Selection: Some Methods and Some Theory

Overview

▶ Yesterday
  ▷ Variable Selection Methods
  ▷ Intrinsic Priors
  ▷ Stochastic Search Driven by Bayes Factors

▶ Things We Did Today
  ▷ Variable Selection Theory: Consistency
  ▷ Changepoint Problems

▶ Tomorrow Never Knows
  ▷ Clustering
  ▷ Conclusions
Part One: Yesterday
Variable Selection in Normal Regression Models

- The full model:

\[ Y = \text{Dependent Variable} \]
\[ \{X_1, \ldots, X_k\} = k \text{ potential explanatory regressors} \]

- Every model with regressors

\[ \{X_{i_1}, \ldots, X_{i_q}\} \]

is a priori a plausible model for \( Y \).

- \( 2^{k-1} \) potential models (intercept always included)
Model Selection as Multiple Hypothesis Testing

• Specify the hypotheses for each model evaluation.
• Evaluate model $M$ by
  \[ H_0 : M = \text{candidate model} \; \text{v.} \; H_A : M = \text{reference model}. \]
• For a Bayesian evaluation, the prior distribution should be
  o centered at each $H_0$.
  o specific to each null model $M$ under consideration.
Objective Bayes Model Selection: Some Methods and Some Theory

Reference Model

► The reference model encompasses all other models
  ▷ Models need to be nested
  ▷ Can do this in two ways

► Casella/Moreno (JASA 2006): Encompassing from Above

\[ H_0 : M = \text{reduced model} \quad \text{This tests whether the} \]
\[ H_A : M = \text{model with all predictors} \quad \text{reduced model explains} \]
\[ \text{vs.} \]
\[ \text{significant variation} \]

▷ Reduced Model \( \subseteq \) Full Model
Encompassing

Casella/Moreno (JASA 2006): Encompassing from Above

$H_0 : M = \text{reduced model}$
vs.
$H_A : M = \text{model with all predictors}$

This tests whether the reduced model explains significant variation

Candidate Model $\in$ Full Model

Girón et al. (2006): Encompassing from Below

$H_0 : M = \text{intercept-only model}$
vs.
$H_A : M = \text{reduced model}$

This tests whether the reduced model significantly improves on intercept-only

Null Model $\in$ Candidate Model
Bayes Factors

- Compare

\[ M_1 : \{ f_1(x|\theta_1), \pi_1(\theta_1) \} \text{ vs. } M_2 : \{ f_2(x|\theta_2), \pi_2(\theta_2) \} \]

- Marginal Distributions

\[ M_1 : m_1(x) = \int_{\Theta} f_1(x|\theta_1)\pi_1(\theta_1) \, d\theta \]
\[ M_2 : m_2(x) = \int_{\Theta} f_2(x|\theta_2)\pi_2(\theta_2) \, d\theta \]

- Bayes Factor

\[ BF = \frac{m_1(x)}{m_2(x)} \]
Objective Probabilities

- Model Selection ⇒
  - not confident about any given set of explanatory variables
  - little prior information on the regression coefficients

- Objective model choice approach is justified.

- Typical default priors are improper, and cannot be used.
  - The Bayes factor cannot be determined

- Intrinsic Priors (Berger/Pericchi 1996) address this problem
Intrinsic Priors

- Berger and Pericchi (1996)
  - Handle the *impropriety problem*
  - Provide *sensible objective proper priors*

- Moreno *et al.* (1998) develop *intrinsic priors* further
  - They show there is an entire class
  - They show which one to use
Intrinsic Priors - Details

► Compare

\[ M_1 : \{ f_1(x|\theta_1), \pi_1^N(\theta_1) \} \text{ vs. } M_2 : \{ f_2(x|\theta_2), \pi_2^N(\theta_2) \} \]

\[ \triangleright f_1(x|\theta_1) \text{ is nested in } f_2(x|\theta_2) \]

\[ \triangleright \pi_i^N(\theta_i) \text{ are the conventional (improper) priors.} \]

► We can use a training sample to convert \( \pi_i^N(\theta_i) \) into a proper posterior. That is,

\[
\pi_i^N(\theta_i|x(\ell)) = \frac{f_i(x(\ell)|\theta_i)\pi_i^N(\theta_i)}{m_i^N(x(\ell))}, \quad i = 1, 2.
\]
Intrinsic Priors - Details

- Actually, we use (Moreno 1997),

\[
\pi^I_2(\theta_2|\theta_1) = \pi^N_2(\theta_2) \mathbb{E}_{x(\ell)|\theta_2}^{M_2} \left( \frac{f_1(x(\ell)|\theta_1)}{\int_{\Theta_2} f_2(x(\ell)|\theta_2) \pi^N_2(\theta_2) d\theta_2} \right)
\]

\[
\pi^I_1(\theta_1) = \pi^N_1(\theta_1)
\]

- We average over all training samples

- No data dependence
Evaluating the Models - Encompassing from Above

- **Full Model** $M_F$: $y = X\alpha + \varepsilon$, $\varepsilon \sim N_n(0, \sigma^2 I_n)$
Evaluating the Models - Encompassing from Above

- **Full Model** $M_F$: $\mathbf{y} = \mathbf{X}\alpha + \mathbf{\varepsilon}$, $\mathbf{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$

- **Submodels** $M_\gamma$: $\mathbf{y} = \mathbf{X}\beta_\gamma$, $\mathbf{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2_\gamma\mathbf{I}_n)$

$\beta_\gamma = \alpha \cdot \gamma$, and $\gamma_i = \begin{cases} 
0, & \text{if } \alpha_i = 0, \\
1, & \text{otherwise}, 
\end{cases} \text{ for } i = 1, \ldots, k.$
Evaluating the Models - Encompassing from Above

**Full Model** $M_F$: $\mathbf{y} = \mathbf{X}\mathbf{\alpha} + \mathbf{\varepsilon}$, $\mathbf{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$

**Submodels** $M_\gamma$: $\mathbf{y} = \mathbf{X}\mathbf{\beta}_\gamma$, $\mathbf{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2_\gamma \mathbf{I}_n)$

$\mathbf{\beta}_\gamma = \mathbf{\alpha} \cdot \gamma$, and $\gamma_i = \begin{cases} 0, & \text{if } \alpha_i = 0, \\ 1, & \text{otherwise}, \end{cases}$ for $i = 1, \ldots, k$.

**Test** $H_0 : M = M_\gamma$ vs. $H_A : M = M_F$, using

$$P(M_\gamma|\mathbf{y}, \mathbf{X}) = \frac{m_\gamma(\mathbf{y}, \mathbf{X})}{m_1(\mathbf{y}, \mathbf{X}) + \sum_{\gamma \in \Gamma, \gamma \neq 1} m_\gamma(\mathbf{y}, \mathbf{X})},$$

to measure the support for $H_0$. 

George Casella, University of Florida 14
A Serious Discussion

Elías and Javier discussing appropriate model selection priors?
A Serious Discussion

Model selection priors?———-NO !!
A Serious Discussion

Elías and Javier discussing appropriate wines for dinner!
An Intrinsic Prior for \( \alpha \)

\[
\pi^I(\alpha | \beta, \sigma_\gamma) = \int N_k(\alpha | \beta, (\sigma_\gamma^2 + \sigma^2)W^{-1}) \frac{1}{\sigma_\gamma} \left( 1 + \frac{\sigma^2}{\sigma_\gamma^2} \right)^{-3/2} d\sigma
\]

- An elliptical multivariate distribution with mean \( \beta_\gamma \).
  
  - Centered at the null.
  
    - Not typical among variable selection priors.

- Moments \( \geq 2 \) do not exist \( \Rightarrow \) very heavy tails.
Encompassing Details

- Practically we have seen that the direction of encompassing makes little difference

- Computationally the formulas are very similar

- The Bayes factor to compare

\[
H_0: \text{model } i \text{ vs. } H_1: \text{model } j
\]

\[
B_{ji}(y, X) = \frac{2(j + 1)^{(j-i)/2}}{\pi} \times \int_0^{\pi/2} \frac{(\sin \phi)^{j-i} [n + (j + 1) \sin^2 \phi]^{(n-j)/2}}{[n \frac{\text{RSS}_j}{\text{RSS}_i} + (j + 1) \sin^2 \phi]^{(n-i)}} \, d\phi
\]

where RSS = residual sum of squares.
Encompassing Details

- **Encompassing from below**
  
  \[ H_0 : \text{intercept only} \ vs. \ H_1 : \text{model } j \]
  
  \[ ∆ \text{Bayes Factor} = B_{j1}(y, X) \]

- **Encompassing from above**
  
  \[ H_0 : \text{model } j \ vs. \ H_1 : \text{all regressors} \]
  
  \[ ∆ \text{Bayes Factor} = B_{kj}(y, X) \]
Implementation: Stochastic Search

► Why a Stochastic Search?
  ▶ Number of Models  ▶ Multiple Maxima

► What Drives the Search?
  ▶ Choice of Objective Function

► How to Search?
  ▶ Explore the entire space
  ▶ Hot/Cold Searching
  ▶ Metropolis is practical solution
Modern search algorithms

- Developed by George and McCulloch (1993)
  - Using the Gibbs sampler

- The stochastic search algorithm
  - ‘visits’ models having high probability
  - a ranking of models is obtained
  - can escape from local modes

- Models were not ranked according to any obvious criterion.

- We want a stochastic search with stationary distribution proportional to the model posterior probabilities.
Why a Stochastic Search?

- Predictors $x_1, x_2, x_3$, using squares and interactions, there are $2^{18}$ possible model,

  \[ 2^{18} = 262,144. \]

- We will see the Ozone data example, in which there are $2^{65}$ possible models.

  \[ 2^{65} = 36,893,488,147,419,103,232 \]

- A search algorithm is needed.
How to Search?

▷ Choice of Objective Function $\Rightarrow$ Stationary Distribution
  ▷ Use a Markov Chain (MCMC) with
  
  Stationary Distribution $\propto$ Model Posterior Probabilities

▷ Explore the entire space
  ▷ Don’t get trapped in local modes
  ▷ Visit models with high posterior probability
  ▷ Be sure to see everything
  ▷ Greedy algorithms can get “stuck”
How to Search?

- Metropolis-Hastings

- Have been through many incarnations

- We currently use a two-part *hybrid* algorithm
  - One part: Independent Jumps - Global Moves
  - Other part: Random Walk - Local Moves
Hybrid Metropolis-Hastings

At iteration $t$, **first part:**

$\triangleright$ Choose candidate $M_{\gamma'} \sim g(\cdot)$, Independent

$\triangleright$ Calculate

$$
MHRatio = \log \left( \frac{P(M_{\gamma'}|y, X)/g(M_{\gamma'})}{P(M_{\gamma}|y, X)/g(M_{\gamma})} \right)
$$

$\triangleright$ Accept candidate $M_{\gamma'} \sim g(\cdot)$ with probability

$$
\min \left\{ e^{T_1 \times MHRatio}, 1 \right\}
$$
Hybrid Metropolis-Hastings

At iteration $t$, second part:

Choose candidate $M_{\gamma'} \sim$ Random Walk
  - Select variable at random:
  - Change 0 $\rightarrow$ 1 or 1 $\rightarrow$ 0

$\text{MHRatio} = \log \left( \frac{P(M_{\gamma'}|\mathbf{y},\mathbf{X})}{P(M_{\gamma}|\mathbf{y},\mathbf{X})} \right)$

Accept candidate $M_{\gamma'}$ with probability

$$\min \left\{ e^{T_2 \times \text{MHRatio}}, 1 \right\}$$
Hybrid Metropolis-Hastings

- Tuning Parameters !!

- Acceptance Probabilities

  - Independent Jump: \( \min \left\{ e^{T_1 \times \text{MHRatio}}, 1 \right\} \)

  - Random Walk: \( \min \left\{ e^{T_2 \times \text{MHRatio}}, 1 \right\} \)

- \( T_1 = \text{Cold} \)

- \( T_2 = \text{Hot} \)
Details

- **Search Algorithm:**
  - This is a **reversible ergodic Markov chain**
  - Stationary distribution \( \propto P(M_{\gamma} | y, X) \).

- **Convergence**
  - Finite Sample Space - Uniformly Ergodic
Details

► Search Algorithm:
  ▶ This is a reversible ergodic Markov chain
  ▶ Stationary distribution $\propto P(M_{\gamma}|y, X)$.

► Convergence
  ▶ Finite Sample Space - Uniformly Ergodic - HA HA!
Details

► Search Algorithm:
  ▶ This is a reversible ergodic Markov chain
  ▶ Stationary distribution $\propto P(M|y, X)$.

► Convergence
  ▶ Finite Sample Space - Uniformly Ergodic - HA!

► Exploration
  ▶ Don’t have bound on convergence rate
  ▶ Close to stationary distribution?
  ▶ Probably do not see entire space
Finally, Metropolis is the only practical solution

Note that

\[ P(M_\gamma|y, X) = \frac{B_{\gamma_1}(y, X)}{1 + \sum_{\gamma \in \Gamma, \gamma \neq 1} B_{\gamma_1}(y, X)}, \]

Denominator incalculable in large problems

But all probabilities have the same denominator.

Thus, it cancels out in

\[ \frac{P(M_{\gamma'}|y, X)}{P(M_{\gamma}|y, X)}. \]

This is all we need for Metropolis.
Examples - Hald Regression Data

- Supports Intrinsic Prior/Encompassing from above
- Stochastic Search not needed
- An ancient and often-analyzed data set
  - Measure the effect of heat on cement
    - 13 observations on the dependent variable (heat)
    - 4 predictor variables
    - $2^4 = 16$ possible models
Examples - Hald Regression Data

- **Posterior probabilities for the best models.**

- **Other models had posterior probability** less than 0.00001.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Posterior Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1, x_2$</td>
<td>0.5224</td>
</tr>
<tr>
<td>$x_1, x_4$</td>
<td>0.1295</td>
</tr>
<tr>
<td>$x_1, x_2, x_3$</td>
<td>0.1225</td>
</tr>
<tr>
<td>$x_1, x_2, x_4$</td>
<td>0.1098</td>
</tr>
<tr>
<td>$x_1, x_3, x_4$</td>
<td>0.0925</td>
</tr>
<tr>
<td>$x_2, x_3, x_4$</td>
<td>0.0120</td>
</tr>
<tr>
<td>$x_1, x_2, x_3, x_4$</td>
<td>0.0095</td>
</tr>
<tr>
<td>$x_3, x_4$</td>
<td>0.0013</td>
</tr>
</tbody>
</table>
## Examples - Hald Regression Data

### Comparison to Other Findings

<table>
<thead>
<tr>
<th>Intrinsic Prior</th>
<th>Berger/Pericchi</th>
<th>Draper/Smith</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1, x_2$</td>
<td>$x_1, x_2$</td>
<td>$x_1, x_2$</td>
</tr>
<tr>
<td>$x_1, x_4$</td>
<td>$x_1, x_4$</td>
<td>$x_1, x_4$</td>
</tr>
<tr>
<td>$x_1, x_2, x_3$</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>$x_1, x_2, x_4$</td>
<td>$-$</td>
<td>$x_1, x_2, x_4$</td>
</tr>
<tr>
<td>$x_1, x_3, x_4$</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>$x_2, x_3, x_4$</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>$x_1, x_2, x_3, x_4$</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>$x_3, x_4$</td>
<td>$x_3, x_4$</td>
<td>$-$</td>
</tr>
</tbody>
</table>

### Berger/Pericchi: “...{$x_1, x_2$} is moderately preferred to {$x_1, x_4$} and quite strongly preferred to {$x_3, x_4$}”. 
Examples - Ozone Data

► First analyzed by Breiman and Friedman (1985)

► Breiman (2001) remarked that in the 1980s large linear regressions were run, using squares and interaction terms, with the goal of selecting a good prediction model.

► However, the project was not successful because the false-alarm rate was too high.

► We take the full model to be

  ▶ all linear, quadratic, and two-way interactions

  ▶ $10 + 10 + 45 = 65$ predictors and $2^{65}$ models
### Ozone Data - Top Three Models

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>${x_2, x_1^2, x_7^2, x_9^2, x_1x_5, x_2x_6, x_3x_7, x_4x_6, x_6x_8, x_6x_{10}}$</td>
<td>0.214</td>
<td>0.758</td>
<td>0.873</td>
</tr>
<tr>
<td>${x_1x_9, x_1x_{10}, x_4x_6, x_5x_8, x_6x_7}$</td>
<td>0.122</td>
<td>0.718</td>
<td>0.908</td>
</tr>
<tr>
<td>${x_6, x_5^2, x_7^2, x_9^2, x_1x_{10}, x_4x_7, x_4x_8, x_5x_{10}, x_6x_8}$</td>
<td>0.114</td>
<td>0.748</td>
<td>0.818</td>
</tr>
</tbody>
</table>

▷ Prediction data not used in fitting

▷ All models improve on Breiman/Friedman

George Casella, University of Florida  37
Examples - Ozone Data - Model Predictions

Better predictions at smaller values

○ = B/F, ● = our model
Part Two: Things We Did Today
It Does OK with Data, So...

- Testing a Procedure on Examples is Necessary
- But Examples Don’t Cover All Situations
- Can We Establish a Theoretical Property?
- We Go for the Minimum - Consistency
Pairwise Consistency

- To test the hypothesis

\[ H_0 : \text{Model } M_i \text{ vs. } H_A : \text{Model } M_j. \]

- \( M_i \) is nested in the model \( M_j \)

- The posterior probability of \( M_i \) is

\[
P(M_i|y, X) = \frac{m_i(y, X)}{m_i(y, X) + m_j(y, X)} = \frac{BF_{ij}}{1 + BF_{ij}},
\]
Pairwise Consistency

- For testing
  \[ H_0 : \text{Model } M_i \text{ vs. } H_A : \text{Model } M_j. \]

- It is well known that, under regularity conditions,
  \[ P(M_i|y, X) \to \begin{cases} 1 & \text{if } M_i \text{ is true} \\ 0 & \text{if } M_j \text{ is true} \end{cases}, \]
  as \( n \to \infty \)

- We want to extend this to the entire class of models.
Consistency in the Class of Models

- We compare all models $M_j \in \mathcal{M}$ through testing
  
  $H_0 : \text{Model } M_1 \text{ vs. } H_A : \text{Model } M_j,$

  where $M_1$ is the intercept only model.

- This gives an ordering in the space of all models $\mathcal{M}$ with
  
  $$P(M_j | y, X) = \frac{BF_{j1}}{1 + \sum_{j' \neq 1} BF_{j'1}}, \quad M_j \in \mathcal{M}.$$
Consistency in the Class of Models

We have the following theorem.

Suppose that $M_T \in \mathcal{M}$ is the true model.

**Theorem** In the class of linear models $\mathcal{M}$ with design matrices satisfying conditions . . . , the intrinsic Bayesian variable selection procedure is consistent. That is, when sampling from $M_T$ we have that

$$\frac{P(M_j | y, X)}{P(M_T | y, X)} \rightarrow 0,$$

whenever the model $M_j \neq M_T$. 
Consistency in the Class of Models - Proof

As \( n \to \infty \), the ratio is approximated by

\[
\frac{P(M_j|y, X)}{P(M_T|y, X)} \approx K \exp \left( \frac{T - j}{2} \log n + \frac{n}{2} \log \frac{\mathcal{B}_1^n}{\mathcal{B}_{1j}^n} \right).
\]

Assuming \( M_T \neq M_1 \),

\[
\frac{\mathcal{B}_1^n}{\mathcal{B}_{1j}^n}|M_T \to c < 1.
\]

Thus

\[
\frac{P(M_j|y, X)}{P(M_T|y, X)} \to 0 \text{ for all } j \neq T
\]
One Step Harder: Changepoints

- **Variable Selection**: $n$ observations, $k$ variables
  - Number of Models = $2^{k-1}$

- **Changepoint**: $n$ observations
  - Number of Models = $2^{n-1}$
Changepoint Formulation

- $p$, $1 \leq p \leq n - 1$, = the number of changepoints
- $r_p = (r_1, \ldots, r_p)$ the positions
- The sample density is

$$f(y|\theta_{p+1}, r_p, p) = \prod_{i=1}^{r_1} f(y_i|\theta_1) \prod_{i=r_1+1}^{r_2} f(y_i|\theta_2) \times \cdots \times \prod_{i=r_p+1}^{n} f(y_i|\theta_{p+1}),$$
Changepoint Models

Similar to before, we test

\[ H_0 : M_0 \text{ vs. } H_1 : M_{rp}, \]

where \( M_0 \) = the no change point model

Here we need a prior distribution on \( M_{rp} \)

▷ In Variable Selection we used Uniform on Models

▷ In Changepoint, there are too many models to be totally uniform
Changepoint Models

- To test $H_0: M_0$ vs. $H_1: M_{rp}$

- Model $M_{rp}$ has prior probability

$$\pi(r, p) = \frac{1}{n} \times \frac{1}{\binom{n-1}{p}}$$

  Uniform on Number of Changepoints
  Uniform Given Number of Changepoints

- And Rank Models by $P(M_r|y)$. 

George Casella, University of Florida 49
Changepoint Models - Simulated Data

Do you see the changepoints?
Changepoint Models - Simulated Data

Changepoints at 6 and 12.
Changepoint Models - Nile River Data

Changepoint Models - Nile River Data

Historical and Statistical Consensus is for change at \( x = 28 \) (1898).

Top Three Models have 1 changepoint at 28, 26, 29.
Fourth has 2: 28 and 93.
Part Three: Tomorrow Never Knows
One Step Harder: Clustering

- **Variable Selection:** $n$ observations, $k$ variables
  - Number of Models = $2^{k-1}$

- **Changepoint:** $n$ observations
  - Number of Models = $2^{n-1}$
  - $n = 20 \Rightarrow 524,288$ Models

- **Clustering:** $n$ observations
  - Number of Models = $\mathcal{B}_n$
  - $n = 20 \Rightarrow 51,724,158,235,372$ Models
Cluster Models

- Similar to before, we test $H_0 : M_0$ vs. $H_1 : M_{\omega p}$
  - $M_0 =$ the no cluster model

- Here we need a prior distribution on $M_{\omega p}$

- **Uniform:** $\pi(\omega_p) = \frac{1}{n} \times \frac{1}{S_{n,p}}$
  - $S_{n,p} =$ Stirling Number of the Second Kind
  - There are too many models to be totally uniform
  - Too much time in extreme models
Cluster Models

To test $H_0 : M_0$ vs. $H_1 : M_{\omega p}$

\[ M_0 = \text{the no cluster model} \]

\[ \pi(\omega_p|\lambda) = \frac{\Gamma(\lambda)}{\Gamma(n+\lambda)} \lambda^p \prod_{i=1}^p \Gamma(n_j) \]

\[ \text{Crowley (1997 JASA)} \]

\[ \text{Prior Expectation:} \]

\[ Ep = \lambda \sum_{i=0}^{n-1} \frac{1}{\lambda + i} \]

(Booth et al. 2006)

Rank Models by $P(M_{\omega p}|y)$. 

George Casella, University of Florida 57
Cluster Models - Stochastic Search

- **Mixes** Biased Random Walk and Independent Metropolis

- **Biased Random Walk:**
  - Randomly move object to another occupied cluster
  - Or start new cluster

- **Independent Metropolis**
  - Select partition size $p$ with probability $1/n$
  - Generate random partition with $p$ clusters
Generating Partitions of size $p$ from $n$

Example $n = 8$, $k = 3$


g(\omega) = \frac{k!}{(n-1)(n_1 n_2 \cdots n_k)}. 

George Casella, University of Florida
Hybrid Metropolis-Hastings - Variations

- Independent Jump
  - $p_k = \text{posterior probability of partitions with } k \text{ clusters}$
  - Choose $k \sim p_k$, then choose model
  - Need to estimate $p_k$

- Split-Merge Moves
  - With probability $p$: Merge two randomly chosen clusters
  - With probability $1 - p$: Randomly split a cluster

- Searching for good global moves
Cluster Models - Simulated Data
Changepoint and Cluster Models

**To Do:** Establish Consistency Results

Similar to variable selection,

▷ Show that when sampling from $M_{\text{True}}$

\[
\frac{P(M_j|y, X)}{P(M_{\text{True}}|y, X)} \to 0, \quad \text{whenever the model } M_j \neq M_{\text{True}}
\]

**Problem:** Model space ↑ with $n$
Conclusions- Model Selection

- Two distinct parts of a model selection method
  - Model selection criterion
  - Stochastic search
Conclusions- Model Selection

▶ Two distinct parts of a model selection method

Model selection criterion
Stochastic search

▶ Here

Model selection criterion  Intrinsic Post. Probabilities
Stochastic search  Driven by Criterion
Conclusions- Model Selection

- Two distinct parts of a model selection method
  - Model selection criterion
  - Stochastic search

- Here
  - Model selection criterion: Intrinsic Post. Probabilities
  - Stochastic search: Driven by Criterion

- Intrinsic posterior probabilities favor small models
Conclusions - Model Selection

- This strategy can be used in other settings
  - Can use other criteria to rank models
  - Can use other criteria drive search
Conclusions - Model Selection

► This strategy can be used in other settings
  ▶ Can use other criteria to rank models
  ▶ Can use other criteria drive search

► We use two “prior” distributions on model space
  1. Generate Independent Candidates More Diffuse
  2. Calibrate Bayes Factors Less Diffuse
Conclusions - Model Selection

- We use two “prior” distributions on model space
  1. Generate Independent Candidates More Diffuse
  2. Calibrate Bayes Factors Less Diffuse

- For example, in clustering

  1. Independent Candidates
     \[ g(\omega) = \frac{1}{n} \frac{k!}{(n-1)(n_1 n_2 \ldots n_k)} \]

  2. Calibrate Bayes Factors
     \[ \pi(\omega_p | \lambda) = \frac{\Gamma(\lambda)}{\Gamma(n+\lambda)} \lambda^p \prod_{i=1}^{p} \Gamma(n_j) \]
Conclusions - Stochastic Search

- The search algorithm is **Metropolis-Hastings**
  - Candidate from **mixture**

- **Important** to choose a good candidate distribution.

- The candidate must
  - **find states** having large values of the criterion
  - **escape from local modes** to better explore the space.

- The construction proposed here seems to do this.
To Do

- Some Theory for Changepoint and Clustering Algorithms
To Do

- Some Theory for Changepoint and Clustering Algorithms
- Improve the \texttt{R} code ⇒ Handle Large Problems
- Improve the \texttt{R} code ⇒ \texttt{R} package
To Do

- Some Theory for Changepoint and Clustering Algorithms
- Improve the \texttt{R} code \Rightarrow \textbf{Handle Large Problems}
- Improve the \texttt{R} code \Rightarrow \texttt{R package}
- Other Model Selections Problems
  - Mixed Models
  - GLM(M)
Details Can be Found In

▶ Yesterday
  ▶ Casella and Moreno (2006) Objective Bayes Variable Selection *JASA*

▶ Things We Did Today
  ▶ Girón *et al.* (2007) Objective Bayesian Analysis of Multiple Change-points for Linear Models. *Bayesian Statistics 8*

▶ Tomorrow Never Knows
  ▶ Clustering paper to be written

▶ Available at [http://www.stat.ufl.edu/~casella/Papers](http://www.stat.ufl.edu/~casella/Papers)
Thanks!