International Workshop on Statistical Modelling 2007

Barcelona, July 2-6, 2007

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

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:

$$\begin{aligned} Y &= \text{Dependent Variable} \\ \{X_1, ..., X_k\} &= k \text{ potential explanatory regressors} \end{aligned}$$

- Every model with regressors

$$\{X_{i_1}, ..., X_{i_q}\}$$

  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

  $H_0 : M =$ candidate model v. $H_A : M =$ reference model.

- For a Bayesian evaluation, the prior distribution should be
  - centered at each H_0.
  - specific to each null model M under consideration.

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

$H_0 : M =$ reduced model	This tests whether the
vs.	reduced model explains
$H_A : M =$ model with all predictors	significant variation

▷ Reduced Model $\in$ Full Model

Encompassing

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

$H_0 : M =$ reduced model
vs.
$H_A : M =$ 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

$H_0 : M =$ intercept-only model
vs.
$H_A : M =$ 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

$$\mathrm{BF} = \frac{m_1(x)}{m_2(x)}$$

Objective Probabilities

- Model Selection $\Rightarrow$
  - ▷ 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

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

  - $f_1(x|\theta_1)$ is nested in $f_2(x|\theta_2)$
  - $\pi_i^N(\theta_i)$ 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),

$$\pi_2^I(\theta_2|\theta_1) = \pi_2^N(\theta_2)\mathrm{E}^{M_2}_{x(\ell)|\theta_2}\left(\frac{f_1(x(\ell)|\theta_1)}{\int_{\Theta_2} f_2(x(\ell)|\theta_2)\pi_2^N(\theta_2)d\theta_2}\right)$$

$$\pi_1^I(\theta_1) = \pi_1^N(\theta_1)$$

- We average over all training samples

- No data dependence

Evaluating the Models - Encompassing from Above

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

Evaluating the Models - Encompassing from Above

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

- Submodels M_γ: $\mathbf{y} = \mathbf{X}\beta_\gamma, \quad \varepsilon \sim N_n(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}_n)$

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

Evaluating the Models - Encompassing from Above

- Full Model M_F: $\mathbf{y} = \mathbf{X}\alpha + \varepsilon, \quad \varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$
- Submodels M_γ: $\mathbf{y} = \mathbf{X}\beta_\gamma, \quad \varepsilon \sim N_n(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}_n)$

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

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

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

to measure the support for H_0.

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 α

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

- An elliptical multivariate distribution with mean β_γ.

▶ Centered at the null.

▹ Not typical among variable selection priors.

▶ Moments ≥ 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

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

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

where RSS = residual sum of squares.

Encompassing Details

- Encompassing from below

$$H_0 : \text{intercept only vs. } H_1 : \text{model } j$$

▷ Bayes Factor $= B_{j1}(\mathbf{y}, \mathbf{X})$

- Encompassing from above

$$H_0 : \text{model } j \text{ vs. } H_1 : \text{all regressors}$$

▷ Bayes Factor $= B_{kj}(\mathbf{y}, \mathbf{X})$

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.

$$\mathbf{2^{65} = 36,893,488,147,419,103,232}$$

- 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:
  - ▷ Choose candidate $M_{\gamma'} \sim g(\cdot)$, Independent
  - ▷ Calculate

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

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

$$\min\left\{e^{T_1 \times \text{MHRatio}}, 1\right\}$$

Hybrid Metropolis-Hastings

- At iteration t, second part:
- Choose candidate $M_{\gamma'} \sim$ Random Walk
  - $\triangleright$ Select variable at random:
  - $\triangleright$ Change $0 \longrightarrow 1$ or $1 \longrightarrow 0$
- $\text{MHRatio} = \log\left(\dfrac{P(M_{\gamma'}|\mathbf{y},\mathbf{X})}{P(M_{\gamma}|\mathbf{y},\mathbf{X})}\right)$
- 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 =Cold

- T_2 =Hot

Details

- Search Algorithm:
  - This is a reversible ergodic Markov chain
  - Stationary distribution $\propto P(M_{\boldsymbol{\gamma}}|\mathbf{y}, \mathbf{X})$.
- Convergence
  - Finite Sample Space - Uniformly Ergodic

Details

- Search Algorithm:
  - This is a reversible ergodic Markov chain
  - Stationary distribution $\propto P(M_{\boldsymbol{\gamma}}|\mathbf{y}, \mathbf{X})$.
- Convergence
  - Finite Sample Space - Uniformly Ergodic - HA HA!

Details

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

Details

- Finally, Metropolis is the only practical solution
  - Note that
$$P(M_\gamma|\mathbf{y},\mathbf{X}) = \frac{B_{\boldsymbol{\gamma}1}(\mathbf{y},\mathbf{X})}{1+\sum_{\gamma\in\Gamma,\gamma\neq\mathbf{1}} B_{\gamma1}(\mathbf{y},\mathbf{X})},$$
  - Denominator incalculable in large problems
  - But all probabilities have the same denominator.
  - Thus, it cancels out in
$$\frac{P(M_{\gamma'}|\mathbf{y},\mathbf{X})}{P(M_\gamma|\mathbf{y},\mathbf{X})}.$$
  - 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.

Variables	Posterior Probability
x_1, x_2	0.5224
x_1, x_4	0.1295
x_1, x_2, x_3	0.1225
x_1, x_2, x_4	0.1098
x_1, x_3, x_4	0.0925
x_2, x_3, x_4	0.0120
x_1, x_2, x_3, x_4	0.0095
x_3, x_4	0.0013

Examples - Hald Regression Data

- Comparison to Other Findings

Top Models

Intrinsic Prior	Berger/Pericchi	Draper/Smith
x_1, x_2	x_1, x_2	x_1, x_2
x_1, x_4	x_1, x_4	x_1, x_4
x_1, x_2, x_3	$---$	$---$
x_1, x_2, x_4	$---$	x_1, x_2, x_4
x_1, x_3, x_4	$---$	$---$
x_2, x_3, x_4	$----$	$---$
x_1, x_2, x_3, x_4	$---$	$---$
x_3, x_4	x_3, x_4	$---$

- 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

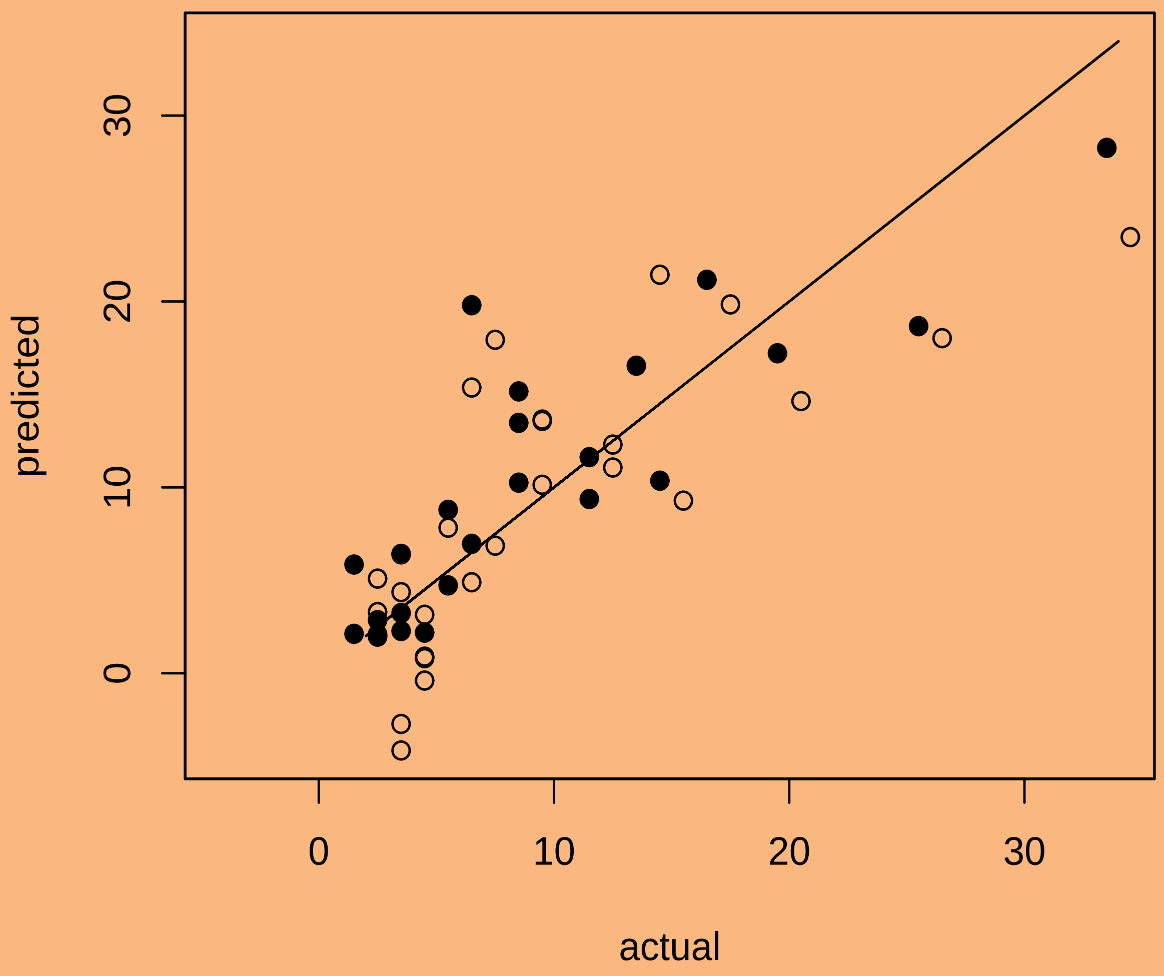
Ozone Data - Top Three Models

Variables	Post. Prob.	R^2	Avg. Pred. Error
$\{x_2, x_1^2, x_7^2, x_9^2, x_1x_5, x_2x_6, x_3x_7, x_4x_6, x_6x_8, x_6x_{10}\}$	0.214	0.758	0.873
$\{x_1x_9, x_1x_{10}, x_4x_6, x_5x_8, x_6x_7\}$	0.122	0.718	0.908
$\{x_6, x_5^2, x_7^2, x_9^2, x_1x_{10}, x_4x_7, x_4x_8, x_5x_{10}, x_6x_8\}$	0.114	0.748	0.818

▷ Prediction data not used in fitting

▷ All models improve on Breiman/Friedman

Examples - Ozone Data - Model Predictions



$\circ$ = B/F, $\bullet$ = our model

Better predictions
at smaller values

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

$$P(M_i|\mathbf{y}, \mathbf{X}) = \frac{m_i(\mathbf{y}, \mathbf{X})}{m_i(\mathbf{y}, \mathbf{X}) + m_j(\mathbf{y}, \mathbf{X})} = \frac{BF_{ij}}{1 + BF_{ij}},$$

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,
$$P(M_i|\mathbf{y}, \mathbf{X}) \rightarrow \begin{cases} 1 \text{ if } M_i \text{ is true} \\ 0 \text{ if } M_j \text{ is true} \end{cases},$$
as $n \rightarrow \infty$

- We want to extend this to the entire class of models.

Consistency in the Class of Models

- We compare all models $M_j \in \mathfrak{M}$ 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 $\mathfrak{M}$ with

$$P(M_j|\mathbf{y}, \mathbf{X}) = \frac{BF_{j1}}{1 + \sum_{j' \neq 1} BF_{j'1}}, \; M_j \in \mathfrak{M}.$$

Consistency in the Class of Models

- We have the following theorem.

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

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

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

whenever the model $M_j \neq M_T$.

Consistency in the Class of Models - Proof

- As $n \to \infty$, the ratio is approximated by

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

- Assuming $M_T \neq M_1$,

$$\frac{\mathcal{B}^n_{1T}}{\mathcal{B}^n_{1j}}|M_T \to \mathbf{c} < 1.$$

- Thus

$$\frac{P(M_j|\mathbf{y},\mathbf{X})}{P(M_T|\mathbf{y},\mathbf{X})} \to 0 \text{ for all } j \neq T$$

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
- $\mathbf{r}_p = (r_1, \ldots, r_p)$ the positions
- The sample density is

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

Changepoint Models

- Similar to before, we test

$$H_0 : M_0 \text{ vs. } H_1 : M_{\mathbf{r}_p},$$

where M_0 = the no change point model

- Here we need a prior distribution on $M_{\mathbf{r}_p}$
  - In Variable Selection we used Uniform on Models
  - In Changepoint, there are too many models to be totally uniform

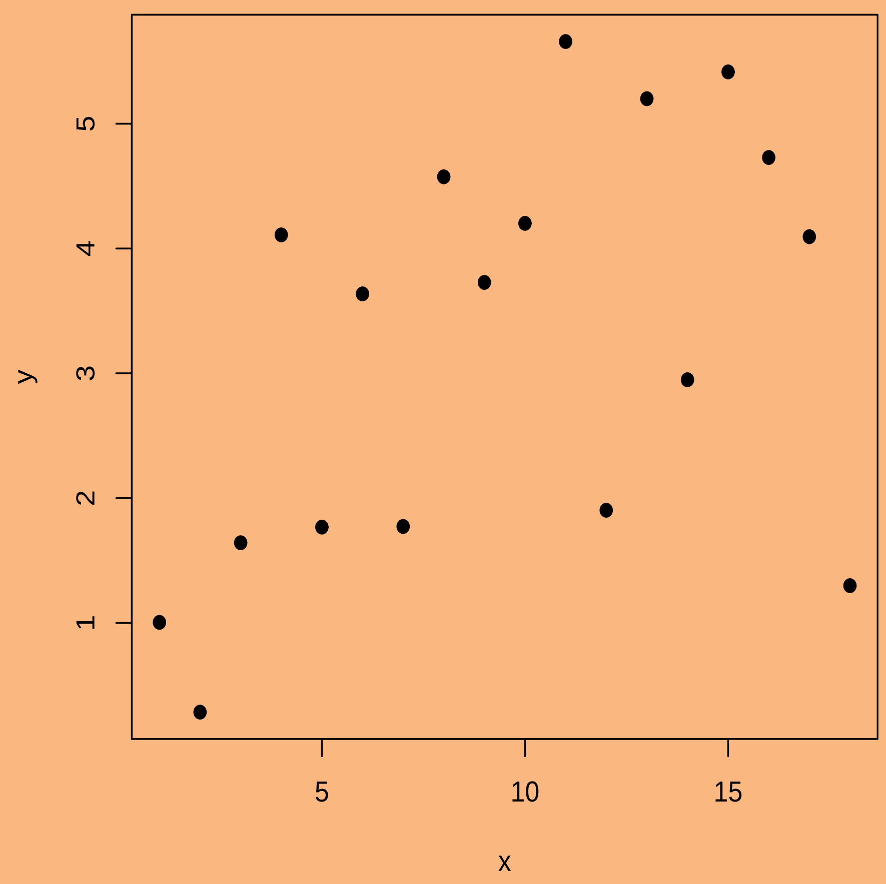
Changepoint Models

- To test $H_0 : M_0$ vs. $H_1 : M_{\mathbf{r}_p}$
- Model $M_{\mathbf{r}_p}$ has prior probability

$$\pi(\mathbf{r}, p) = \underbrace{\frac{1}{n}}_{\substack{\text{Uniform on Number} \\ \text{of Changepoints}}} \times \underbrace{\frac{1}{\binom{n-1}{p}}}_{\substack{\text{Uniform Given Number} \\ \text{of Changepoints}}}$$

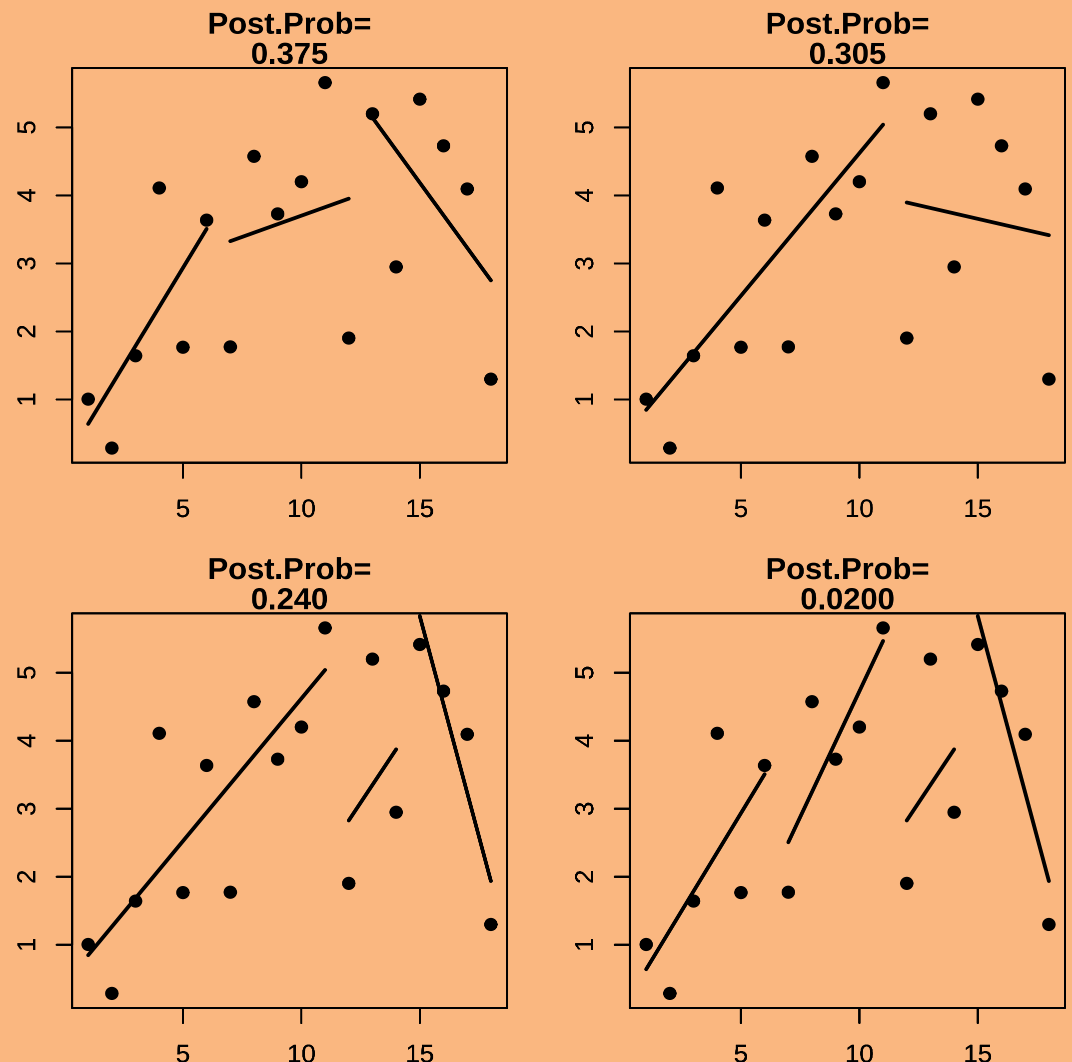
- And Rank Models by $P(M_{\mathbf{r}}|\mathbf{y})$.

Changepoint Models - Simulated Data



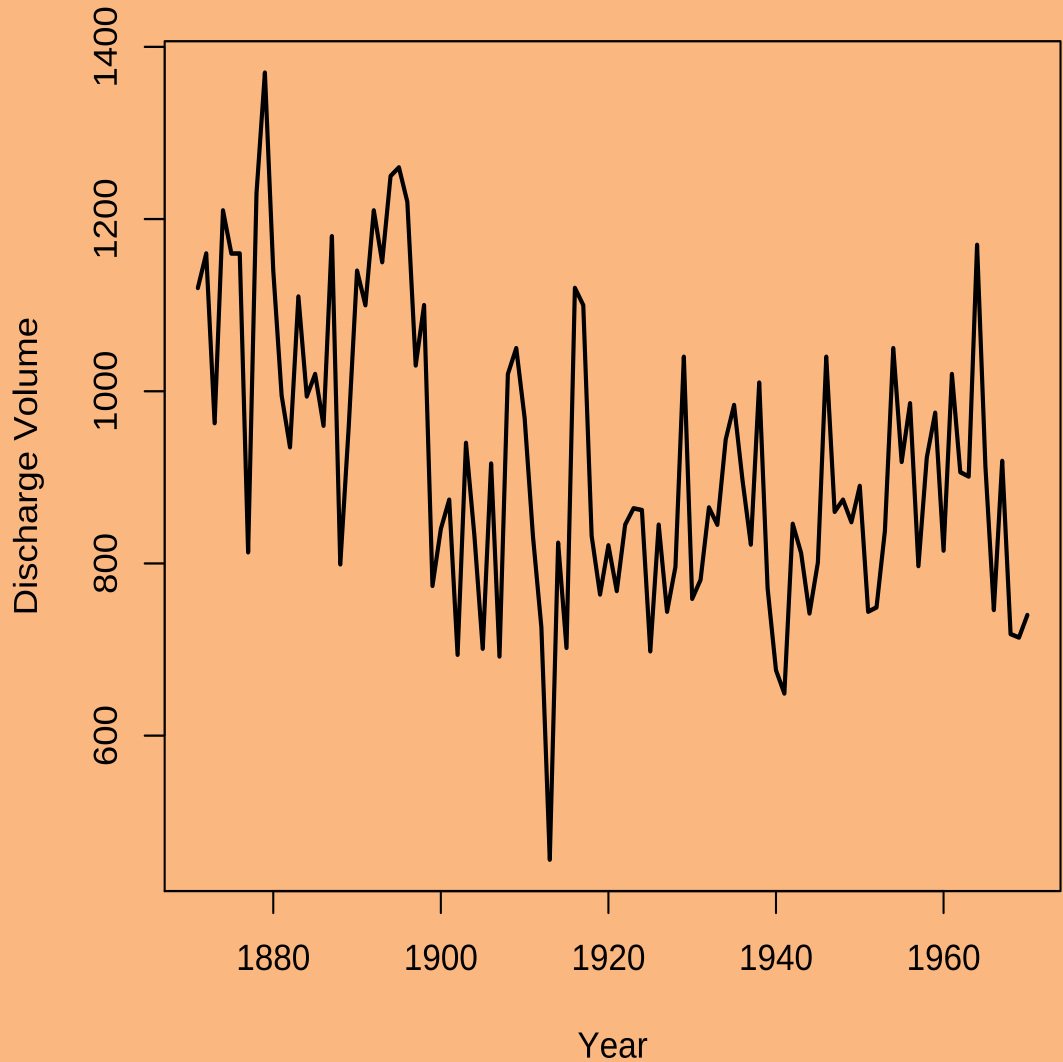
Do you see
the changepoints?

Changepoint Models - Simulated Data



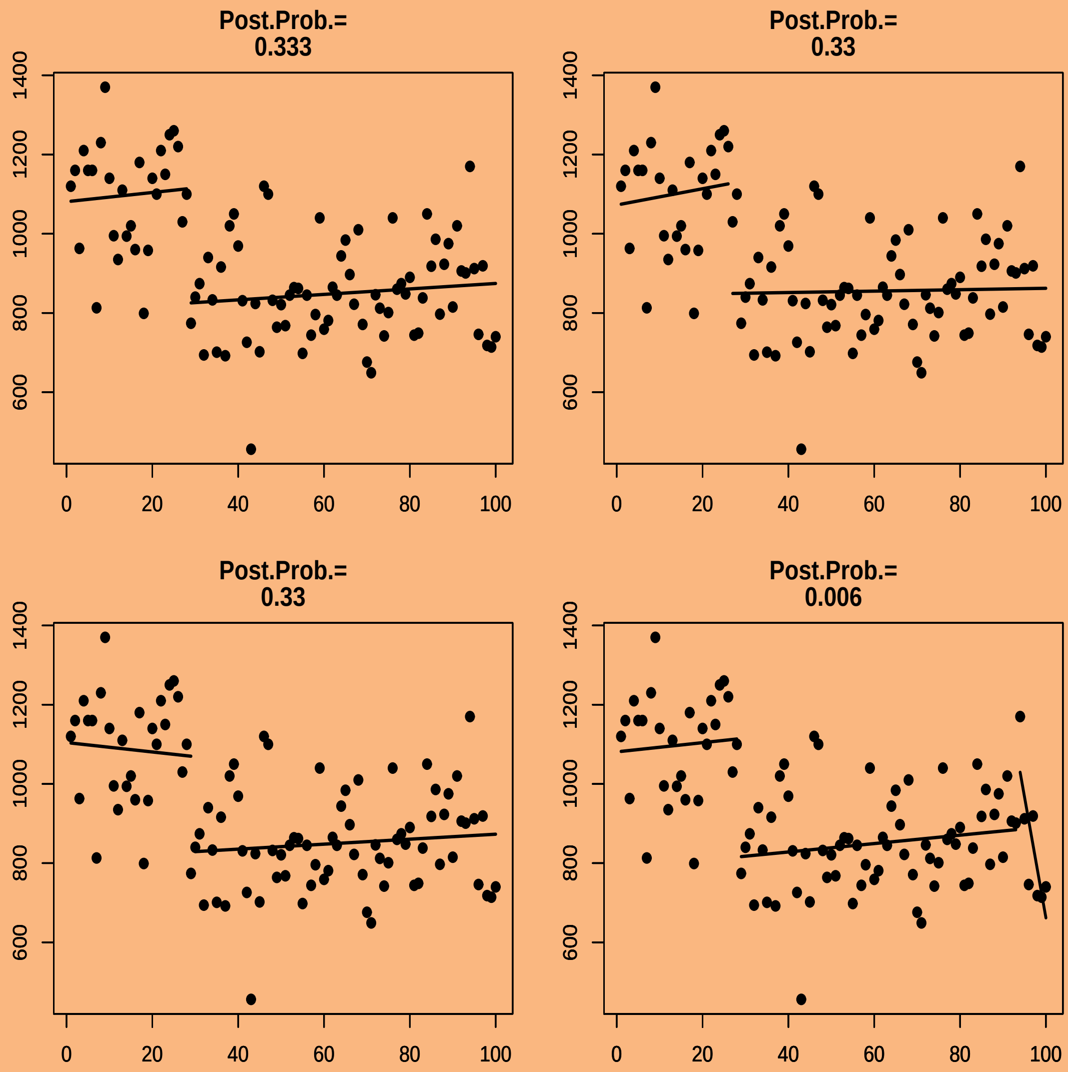
Changepoints at 6 and 12.

Changepoint Models - Nile River Data



Volume of Discharge 1871-1970.

Changepoint Models - Nile River Data



Historical and Statistica
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_{ω_p}
- Uniform: $\pi(\omega_p) = \frac{1}{n} \times \frac{1}{\mathcal{S}_{n,p}}$
  - $\mathcal{S}_{n,p} =$ 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. $H_1 : M_{\omega_p}$
  - $M_0 =$ the no cluster model

- $\pi(\omega_p|\lambda) = \frac{\Gamma(\lambda)}{\Gamma(n+\lambda)}\lambda^p \prod_{i=1}^{p} \Gamma(n_j)$
  - Crowley (1997 JASA)
  - Prior Expectation:

$$\mathrm{E}p = \lambda \sum_{i=0}^{n-1} \frac{1}{\lambda + i}$$

(Booth *et al.* 2006)

- Rank Models by $P(M_{\omega_p}|\mathbf{y})$.

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 $n = 8$, $k = 3$

$$\_\ \_\ \_\ \_\ \_\ \_\ \_\ \_ \qquad \text{Eight spaces}$$

$$\underline{1}\ \_\ \_\ \_\ \_\ \_\ \_\ \_ \qquad \text{Fix 1 in first space}$$

$$\underline{1}\ \underline{0}\ \underline{1}\ \underline{0}\ \underline{0}\ \underline{1}\ \underline{0}\ \underline{0}$$

Randomly distribute remaining 1s - Fill in 0s

- One cluster of size 2, Two clusters of size 3

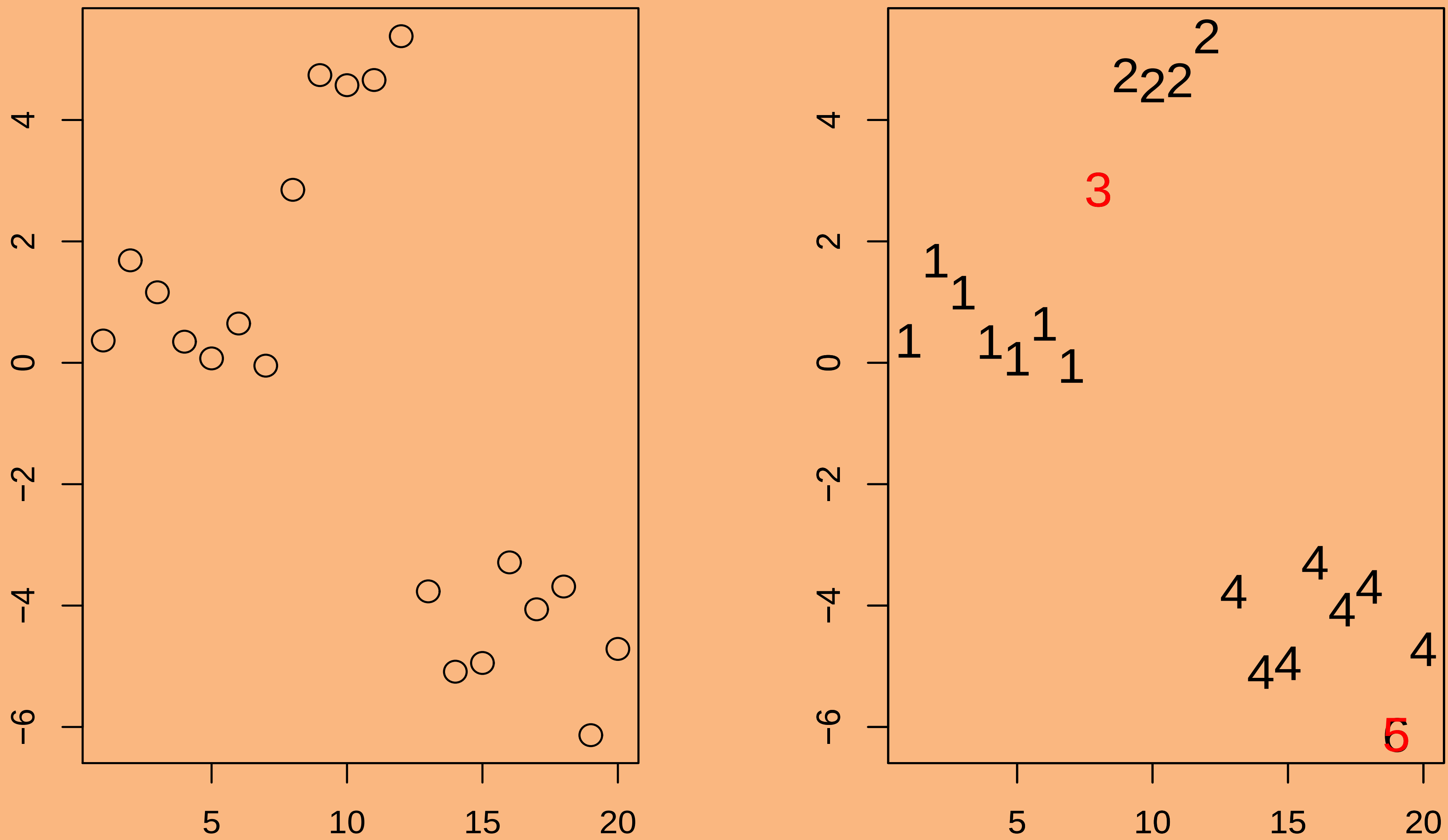
- The probability of a partition $\omega = \{n_1\ n_2\ \cdots\ n_k\}$ is

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

Hybrid Metropolis-Hastings - Variations

- Independent Jump
  - $p_k =$ posterior probability of partitions with k 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_{True}

$$\frac{P(M_j|\mathbf{y},\mathbf{X})}{P(M_{\text{True}}|\mathbf{y},\mathbf{X})} \to 0, \; ,$$

  whenever the model $M_j \neq M_{\text{True}}$

- Problem: Model space $\uparrow$ 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 $g(\omega) = \frac{1}{n}\frac{k!}{\binom{n-1}{k-1}\binom{n}{n_1\ n_2\ \cdots\ n_k}}$.

   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 `R` code $\Rightarrow$ Handle Large Problems
- Improve the `R` code $\Rightarrow$ `R` package

To Do

- Some Theory for Changepoint and Clustering Algorithms
- Improve the `R` code $\Rightarrow$ Handle Large Problems
- Improve the `R` code $\Rightarrow$ `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
  - Casella *et al.* (2006). Consistency of Bayesian Procedures for Variable Selection. Technical Report.
  - Girón *et al.* (2007) Objective Bayesian Analysis of Multiple Changepoints for Linear Models. *Bayesian Statistics 8*
- Tomorrow Never Knows
  - Clustering paper to be written
- Available at `http://www.stat.ufl.edu/~casella/Papers`

Thanks!