Objective Bayesian
Variable Selection

George Casella
University of Florida

Elías Moreno
University of Granada

casella@stat.ufl.edu

December 2, 2002
Overview

• **Introduction**
  - Model Selection

• **Evaluating the Models**
  - Objective Bayesian Solution

• **Evaluation of Posterior Probabilities**
  - Simulated and Real Data

• **Implementation**
  - MCMC Stochastic Search

• **Evaluation of Search Algorithm**
  - Simulated and Real Data

• **Conclusions**
Introduction

• Variable Selection in Normal Regression Models

• A dependent random variable $Y$ and a set \{${X_1, ..., X_k}$\} of $k$ potential explanatory regressors

• Every model with regressors

\[ \{X_{i1}, ..., X_{iq}\} \]

is a priori a plausible model for $Y$.

• $2^{k-1}$ potential models (intercept always included).
Introduction

• Interest here is in **model selection**.

• If interest is in **prediction**:
  ◦ The prediction can be through model averaging
  ◦ The selection problem seems to be avoided.
  ◦ But it may be impossible to compute every model.
Introduction

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

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

- Before model averaging we must select models to average.
- So prediction will be preceded by model selection.
Two Aspects of Model Selection

- The selection mechanism to be criterion-based and fully automatic
  - Criterion-based selection
    - clear understanding of the properties of the selected models
  - Fully automatic algorithms
    - no tuning parameters, hyperparameters, etc. to estimate
    - easy to implement
    - no sensitivity analysis needed
Model Selection is Multiple Hypothesis Testing

- must exactly specify the hypotheses for each model evaluation.
- the evaluation of model $M$ should be
  \[ H_0 : M = \text{reduced model} \]
  vs.
  \[ H_A : M = \text{model with all predictor variables}. \]
- The full model comes from the subject-matter, and is the correct reference.
Model Selection

• We assume that all predictors have some importance, and examine if a smaller subset is adequate.

• For a Bayesian evaluation, the prior distribution should be
  ○ centered at each $H_0$.
  ○ specific to each null model $M$ under consideration.
Objective Probabilities

• Since we are not confident about any given set of explanatory variables, little prior information on their regression coefficients could be expected.

• If we were confident about a particular model, there would be no model selection problem!
Objective Probabilities

• With little prior information, an objective model choice approach is justified.
• Since typical default priors for normal regression are improper, they cannot be used.
Subjective Bayesian Variable Selection

- **History:**
  - Atkinson (1978)
  - Smith and Spiegelhalter (1980)
  - Pericchi (1984)
  - Poirier (1985)
  - Box and Meyer (1986)
  - Clyde, DeSimone and Parmigiani (1996)
  - Geweke (1996)
  - Smith and Kohn (1996)
  - and others.
Subjective Bayesian Variable Selection

• The prior distributions are typically
  ○ conjugate priors
  ○ some closely related distribution

• Also,
  ○ typical to center the priors at zero
  ○ the null hypothesis is the model with no regressors
Objective Model Selection

- Mitchell and Beauchamp (1988)
  - regression coefficients \( a \text{ priori } \) iid
  - prior distribution that concentrates some probability mass on zero and distributes the rest uniformly on a compact set.
  - variable selection problem is essentially an estimation problem
Objective Model Selection

• Spiegelhalter and Smith (1982)
  ◦ used *conventional improper priors* for the regression coefficients
  ◦ analysis based on a *formal* rather than an *actual* Bayes factor
  ◦ calibrated with subjective information
Intrinsic Bayes Factors

• A fully automatic analysis for model comparison in regression was given in Berger and Pericchi (1996).

• They use
  ◦ encompassing model approach
  ◦ empirical measure for model comparison, the intrinsic Bayes factor
Evaluating the Models

• Full Model:
  \[ y = X\alpha + \varepsilon, \quad \varepsilon \sim N_n(0, \sigma^2 I_n) \]

• Submodels:
  \[ y = X\beta \gamma, \quad \varepsilon \sim N_n(0, \sigma^2 \gamma I_n) \]
  where
  \[ \beta \gamma = \alpha \cdot \gamma, \]
  and
  \[ \gamma_i = \begin{cases} 
    0, & \text{if } \alpha_i = 0, \\
    1, & \text{otherwise,} 
  \end{cases} \]
  for \( i = 1, \ldots, k \).
Prior Distributions

- Complete model specification:
  \[ M_\gamma : \left\{ N_n(\mathbf{y} | \mathbf{X}\beta_\gamma, \sigma_\gamma^2\mathbf{I}_n), \pi(\beta_\gamma, \sigma_\gamma), \gamma \in \Gamma \right\}. \]

- Default prior on the set of models
  \[ P(M_\gamma) = 2^{-(k-1)}, \quad \{ M_\gamma, \gamma \in \Gamma \}. \]
Hypothesis Tests

• Test

\[ H_0 : M = M\gamma \text{ vs. } H_A : M = M_1, \]

using

\[ P(M\gamma|y, X) = \frac{m\gamma(y, X)}{m_1(y, X) + \sum_{\gamma \in \Gamma, \gamma \neq 1} m\gamma(y, X)}, \]

to measure the support for \( H_0 \).
Hypothesis Tests

- Note that

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

so every posterior probability has the same denominator. This will be important in later calculations.
Default Priors

• We want a default or “automatic” prior
  ◦ To remove subjectivity from the choice of $\pi(\beta, \gamma, \sigma \gamma)$
  ◦ to make our procedure automatic
Default Priors

• The standard default prior is improper
  ◦ The integral of the marginal is infinite
  ◦ The Bayes factor can only be computed up to an arbitrary positive constant that cannot be determined
Intrinsic Priors

- Berger and Pericchi (1996)
  - Fix the impropriety problem
  - Provide sensible objective proper priors
- Moreno et al. (1998) develop intrinsic priors further and show
  - there is an entire class
  - which one to use
An Intrinsic Prior for Model Selection

Lemma 1 The intrinsic prior for \( \alpha \) conditional on a fixed point \((\beta_\gamma, \sigma_\gamma)\) is

\[
\pi^I(\alpha, \sigma|\beta_\gamma, \sigma_\gamma) = 
\]

\[
N_k(\alpha|\beta_\gamma, (\sigma_\gamma^2 + \sigma^2)W^{-1}) \frac{1}{\sigma_\gamma} \left(1 + \frac{\sigma^2}{\sigma_\gamma^2}\right)^{-3/2},
\]

where

\begin{itemize}
  \item \( W = Z^tZ \)
  \item \( Z_{(k+1)\times k} \) is a theoretical design matrix
\end{itemize}
The prior of $\alpha$

$$\pi^I(\alpha|\beta, \gamma, \sigma) = \int \mathcal{N}_k(\alpha|\beta, \gamma, (\sigma^2 + \sigma^2)\mathbf{W}^{-1}) \frac{1}{\sigma \gamma} \left(1 + \frac{\sigma^2}{\sigma^2 \gamma}\right)^{-3/2} d\sigma$$

- An elliptical multivariate distribution with mean $\beta \gamma$. 

The prior of $\alpha$

- The intrinsic prior for $\alpha$ is centered at the null.
- This property is not shared by many other variable selection priors.
- Moments $\geq 2$ do not exist. This implies that the intrinsic prior has very heavy tails, as expected for a default prior.
Performance of the Intrinsic Posterior Probabilities

• Are the posterior probabilities a reasonable tool for finding the true model?

• Example: Full Model

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \epsilon, \]

where \( \epsilon \sim N(0, \sigma^2) \).
Performance of the Intrinsic Posterior Probabilities

• The $x_i$ values are generated uniformly in the interval $(0, 10)$

• We simulated 1000 data sets, with $n = 10$ and true model $\{1, 1, 1, 0, 0\}$:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon. \]
Example: Hald Regression Data

- An ancient and often-analyzed data set
- Measure the effect of heat on the composition of cement
  - 13 observations on the dependent variable (heat)
  - 4 predictor variables (which relate to the composition of the cement)
  - $2^4 = 16$ possible models
Example: Hald Regression Data

- Posterior probabilities for the models of the Hald data.
- All 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>
Example: Hald Regression Data

- Comparison to Other Findings

## Top Models

<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 $\{x_3, x_4\}$.”
Example: Hald Regression Data

- Comparison to Other Findings
- Stochastic search of George and McCulloch (1993)
  - visited $\{x_1, x_2\}$ less than 7% of the time.
  - selected as the best model the intercept-only model
  - possibly a consequence of using the no-regressors-model (not even an intercept) as the null model
Calculating the Posterior Probabilities

- Computation is relatively easy.
- The matrix $W^{-1}$ is

$$W^{-1} = \frac{1}{L} \sum_{\ell=1}^{L} (Z^t(\ell)Z(\ell))^{-1},$$

where $\{Z(\ell), \ell = 1, \ldots, L\}$ is the set of all submatrices of $X$ of order $(k + 1) \times k$ of rank $k$, a training sample of minimal size.
Calculating the Posterior Probabilities

- Using a polar transformation, the Bayes factor can be written

\[ B_{\gamma_1}(y, X) = \left( |X_1^T X_1\gamma|^{1/2}(y^T (I_n - H\gamma)y)^{(n-k_n+1)/2} I_{\gamma} \right)^{-1} \]
Calculating the Posterior Probabilities

\[ H_\gamma = X_1^\gamma (X_1^\gamma X_1^\gamma)^{-1} X_t^\gamma, \]

\[ I_\gamma = \int_0^{\pi/2} \frac{|B(\varphi)|^{1/2} d\varphi}{|A_\gamma(\varphi)|^{1/2} E_\gamma(\varphi) \frac{n-k_\gamma+1}{2}} \]

\[ B(\varphi) = [(\sin^2 \varphi)I_n + X W^{-1} X^t]^{-1}, \]

\[ A_\gamma(\varphi) = X_t^\gamma B(\varphi) X_1^\gamma \]

\[ E_\gamma(\varphi) = y^t \left( B(\varphi) - B(\varphi) X_1^\gamma A_\gamma^{-1}(\varphi) X_1^\gamma B(\varphi) \right) y \]

The important point is that there is only one integral!
Implementation

• We can now rank the models by their posterior probabilities.

• However, calculating all posterior probabilities is only possible in small problems.
  
  ◦ Example: Predictors $x_1, x_2, x_3$, using squares and interactions, there are

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

  models.

  ◦ A search algorithm is needed.
Modern search algorithms

• First 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 are not ranked according to any obvious criterion.

• Here, we want a stochastic search with a stationary distribution proportional to the model posterior probabilities.
Stochastic Search

• **Best**: calculate all of the posterior probabilities

• **Second Best**: draw independent samples from a distribution

\[ P(M_{\gamma}|y, X) \propto \text{posterior probability} \]

• **Can’t do either** - needs exhaustive calculation of all of the posterior probabilities
Stochastic Search

• **Third Best**: Construct an MCMC algorithm with

\[ P(M_\gamma | y, X) \propto \text{posterior probability} \]

as the stationary distribution.

- visits every model
- visits the better models more often
- frequency of visits \( \propto \) posterior probabilities.
Metropolis-Hastings

• In theory, construction of the algorithm is easy.
  ◦ With the chain is in model $M_\gamma$, draw a candidate model $M_{\gamma'}$.
  ◦ Move to this new model with probability
    \[
    \min \left\{ 1, \frac{P(M_{\gamma'}|y, X)}{P(M_\gamma|y, X)} \right\}.
    \]

• This is a reversible ergodic Markov chain with stationary distribution $P(M_\gamma|y, X)$. 
Metropolis-Hastings

- Recall the denominator of
  \[ P(M\gamma|y, X) \]
  is the same for all \( \gamma \)

- Thus, it cancels out in
  \[
  \min \left\{ 1, \frac{P(M\gamma'|y, X)}{P(M\gamma|y, X)} \right\}.
  \]

  This is good.

  In large problems the denominator sum is not calculable
Candidate Distribution

• We want our candidate distribution to
  ◦ adequately explore the entire space
  ◦ not get trapped in local modes
  ◦ visit models with high posterior probability

• We construct the candidate distribution in two parts
Candidate Distribution

• Write the models as
  \[ B = \bigcup_i B_i \]
  \[ B_i = \{ M_\gamma : \gamma = \{1, \gamma'\} \} \]
  where \( \gamma' \) has \( i \) components equal to 1.

• At iteration \( t \), choose the subset \( B_i \) with probability
  \[ \hat{P}_i \propto \frac{c}{\log(t + 1)} + \sum_{j \in B_i} p_{ij} / \sum_{i,j} p_{ij} \]
  \( p_{ij} = \) posterior probability

• Update the posterior probabilities.
Candidate Distribution

- Two Pieces:

\[
\hat{P}_i \propto \frac{c}{\log(t + 1)} + \frac{\sum_{j \in B_i} p_{ij}}{\sum_{i,j} p_{ij}}
\]

- Insures Mixing
- Proportional to Bayes Factor
Stochastic Search

- At iteration $t$, choose a candidate model $M_{\gamma'}$
  - by first selecting $B_i$ according to $\hat{P}_i$
  - then selecting $\gamma'$ at random from $B_i$
- With probability
  \[
  \min \left\{ 1, \frac{P(M_{\gamma'}|y, X)}{P(M_{\gamma}|y, X)} \right\}
  \]
  move to $M_{\gamma'}$
Effectiveness of the Stochastic Search

- 10-predictor model

\[ y = \beta_0 + \sum_{i=1}^{3} \beta_i x_i + \sum_{i=1}^{3} \tau_i x_i^2 + \sum_{i>j} \eta_{ij} x_i x_j + \eta_{ijk} x_i x_j x_k + \varepsilon, \]

where \( x_i \) are Uniform (0, 10), \( \varepsilon \sim N(0, \sigma^2) \)

- True model is \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \)

- There are \( 2^{10} = 1024 \) candidate models

- To check, we calculated all posterior probabilities.
Ozone Data

• First analyzed by Breiman and Friedman (1985)

• Using the ACE algorithm, they identified a set of four predictors \( \{x_7, x_8, x_9, x_{10}\} \)

  • We use 10 predictor variables
    o Only linear terms
    o \(2^{10} = 1024\) models

• exhaustive calculation of posterior probabilities
# Ozone Data Linear Predictors

<table>
<thead>
<tr>
<th>Variables</th>
<th>Posterior Probability</th>
<th>$R^2$</th>
<th>Avg. Pred. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_6, x_7, x_8$</td>
<td>0.491</td>
<td>.686</td>
<td>0.992</td>
</tr>
<tr>
<td>$x_1, x_6, x_7, x_8, x_{10}$</td>
<td>0.156</td>
<td>.699</td>
<td>0.974</td>
</tr>
<tr>
<td>$x_1, x_6, x_7, x_8, x_9$</td>
<td>0.041</td>
<td>.696</td>
<td>0.972</td>
</tr>
<tr>
<td>$x_1, x_6, x_7, x_8$</td>
<td>0.028</td>
<td>.691</td>
<td>0.964</td>
</tr>
<tr>
<td>$x_1, x_4, x_6, x_7, x_8$</td>
<td>0.027</td>
<td>.694</td>
<td>0.968</td>
</tr>
<tr>
<td>$x_7, x_8, x_9, x_{10}$</td>
<td>&lt; .00001</td>
<td>.669</td>
<td>1.056</td>
</tr>
</tbody>
</table>

- 25 observations held out of the fitting set to compute prediction error.
- Breiman/Friedman identified $x_7$ as most important
Ozone Data - All Predictors

- 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
- Search ran for 30,000 iterations.
## Ozone Data - All Predictors

<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>

• Top three models
Ozone Data - All Predictors

- Other models visited with frequencies .02—.10
- The search found a very simple model
- The models tend to use $x_7 - x_{10}$ more often.
- Somewhat (but not totally) alleviates the problem of overprediction.
Conclusions

• Two distinct parts of a model selection method
  ○ Model selection criterion: intrinsic posterior probabilities
  ○ The model selection criterion was used to direct a stochastic search
Conclusions

• The two parts function well together
  ○ Intrinsic posterior probabilities is a good criterion
  ○ The stochastic search algorithm finds the good models

• We note the intrinsic posterior probabilities tend to favor small models.
Conclusions

• Either part of our method can be used in other settings
  ◦ For example, we can use other priors to calculate the posterior probabilities for model selection
  ◦ and can use other criteria (for example, $R^2$) to direct the stochastic search
Conclusions

• The search algorithm is straightforward Metropolis-Hastings
• The difficulty is 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.