Regression Model Uncertainty
With Many Predictors

Chris Hans
(Joint work with Mike West)

Institute of Statistics and Decision Sciences

Duke University

17 January 2005
Universitat de València
Variable Selection & Model Uncertainty

• Regression Modeling
  – Many possible predictor variables
  – Prediction and Variable Identification

• Model “selection” understates uncertainty

• Dimension reduction often undesirable

• Identify “best” models in a large space
Motivating Example

Gene Expression Data
- Tissue samples
- Interested in genomic “activity”
- Continuous measure

- (Tens of) Thousands of genes

- Possible Outcomes
  - Survival time (continuous)
  - Recur/non-recur (binary)
  - LN +/-
  - etc…
Motivating Example

Keck Center for Neurooncogenomics at Duke

Survival Study in Brain Cancer

- $n = 41$ glioblastoma patients $y_i, i = 1, \ldots, n$
- survival times
- $p = 8408$ variables (genes), $x_i$

- Build predictive models
- Identify important predictors
Linear Model Framework

Notation for Regression Subsets

- $\gamma$ is a $p \times 1$ indicator vector:
  $\gamma_j = 1$ if $x_j$ is in the model
  $\gamma_j = 0$ if $x_j$ is NOT in the model

For a given model:

$$y = X\gamma \beta_\gamma + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

Which subsets of variables?
What size subsets?
Bayesian Framework

Posterior distribution of a model:

\[ p(\gamma | y) \propto p(y | \gamma) p(\gamma) \]

\[ = \left[ \int p(y | \theta, \gamma) p(\theta | \gamma) d\theta \right] p(\gamma) \]

Left to specify the prior distribution:

\[ p(\theta, \gamma) = p(\theta | \gamma) p(\gamma) \]
Prior Distributions – Model Space

Sparsity is key

• Average over many small models

• Why?
  – Parsimony
  – (application specific) intuition / prior knowledge

• \( \Pr(\gamma_j = 1) = \pi \)

\[
p(\gamma) = \pi^k (1 - \pi)^{p-k}, \quad k = \sum_{j=1}^{p} \gamma_j
\]
Sparsity via $p(\gamma)$

**Sparsity** is enforced through $\pi$
- Can fix $\pi$ *a priori*
- Can estimate $\pi$ (EB type approach)
- Can model $\pi$ as a parameter

**Binomial Distribution on Model Size**
Prior Distributions – Parameter Space

- **Conjugate Priors**

\[
\sigma^2 | \gamma \sim IG \left( \frac{\delta + k}{2}, \frac{\tau}{2} \right) \\
\beta | \sigma^2, \gamma \sim N(0, \tau^{-1} \sigma^2 I_k)
\]

- **Closed form calculation of** \( p(y|\gamma) \)

- **Consistency** with an encompassing model
  - \((y_i, x_i')' \sim N(0, \Sigma)\)
  - Prior induced on regression coefficients under \( p(y|X\gamma, \Sigma) \)
Model Space Exploration

About $3.5 \times 10^{17}$ models w/up to 5 variables

Common Strategies:

- **MCMC**: estimate $p(\gamma_j \mid y)$ and $\hat{Y}$ based on relative frequency of models in Markov chain

- **Stochastic Search**: use as a tool to create a list of “best” models, $\Gamma^* \subset \Gamma$. Condition on this list and do model averaging.

**Key**: how do we determine $\Gamma^*$?
Stochastic Search

\[ \gamma^{(1)} \rightarrow \gamma^{(2)} \rightarrow \gamma^{(3)} \rightarrow \ldots \]

At each iteration:

1. Move across dimension effectively
2. Allow each variable to be considered
3. Quickly identify “similar” models
Shotgun Stochastic Search

- Shoot out many proposals based on current model
- Evaluate them in parallel
- Sample a new model from the proposals
Regression Model SSS

Choosing the Proposal Models

**Current model $\gamma$ of size $k$**

Consider three types of proposals:

- neighboring models $\gamma^{-}$ of dimension $k - 1$
- neighboring models $\gamma^{o}$ of dimension $k$
- neighboring models $\gamma^{+}$ of dimension $k + 1$

These are the models “shot out” *(in parallel)*
Neighborhood Example

\[ \gamma = \{ x_1, x_3, x_4 \} \text{ with } p = 5 \]

- \( \gamma^- = \{ \{ x_1, x_3 \}, \{ x_1, x_4 \}, \{ x_3, x_4 \} \} \) deletion set
- \( \gamma^o = \left\{ \{ x_j, x_3, x_4 \}, \{ x_1, x_j, x_4 \}, \{ x_1, x_3, x_j \} \right\}_{j \in \{2,5\}} \) swap set
- \( \gamma^+ = \{ \{ x_1, x_2, x_3, x_4 \}, \{ x_1, x_3, x_4, x_5 \} \} \) addition set

- Note
  - \( |\gamma^-| = k \) if \( k > 2 \) \( (\gamma^- = \emptyset \) if \( k = 1 \)
  - \( |\gamma^o| = k (p - k) \)
  - \( |\gamma^+| = p - k \)
Regression Model SSS

Choosing the New Model

- Would like to balance dimension

STEP 1  Sample three proposals: \[
\begin{align*}
\gamma^-_* & \quad \text{from the set} \quad \gamma^- & \quad [k] \\
\gamma^0_\ast & \quad \text{from the set} \quad \gamma^0 & \quad [k(p - k)] \\
\gamma^+_\ast & \quad \text{from the set} \quad \gamma^+ & \quad [p - k]
\end{align*}
\]

STEP 2  Sample one proposal as the new model

Sample with probability proportional to posterior probability
Regression Model SSS

- Current Model
- Parallel Computing Step
- Three proposals
- New Model
Brain Cancer Survival Study

Keck Center for Neurooncogenomics at Duke

- $n = 41$ glioblastoma patients
- $p = 8408$ genes for each patient
- Affymetrix Human U133A GeneChip (>22,500 genes)
- expression intensities estimated using RMA (Bioconductor)
- expression levels are standardized

$$\log \text{ survival } \sim \mathcal{N}(X_\gamma \beta_\gamma, \sigma^2 I_n)$$

- Priors: $\tau = 1$, $\delta = 3$, $\pi = 10/p$
Posterior Summary

- 1,000,000 models from 20,000 iterations

- Mix of models of small dimension:

<table>
<thead>
<tr>
<th>Dimension</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td># of models</td>
<td>1</td>
<td>8408</td>
<td>146,290</td>
<td>123,155</td>
<td>287,172</td>
<td>433,453</td>
</tr>
<tr>
<td>Relative Probability</td>
<td>0.03</td>
<td>0.09</td>
<td>0.07</td>
<td>0.17</td>
<td>0.43</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Key Genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>sparcs</th>
<th>semaphorin</th>
<th>double cortex</th>
<th>double cortin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Probability</td>
<td>0.81</td>
<td>0.64</td>
<td>0.35</td>
<td>0.13</td>
</tr>
</tbody>
</table>
Model Averaged Fit
Stratifying Samples
Marginal Likelihood for $\pi$

$$p(y|\pi) = \sum_{\gamma \in \Gamma} p(y, \gamma|\pi) = \sum_{\gamma \in \Gamma} p(y|\gamma)p(\gamma|\pi)$$

$$= \sum_{k=0}^{p} p(\gamma|\pi) \sum_{\gamma \in \Gamma_k} p(y|\gamma)$$

$$= \sum_{k=0}^{p} \pi^k (1 - \pi)^{p-k} \left[ \sum_{\gamma \in SSS_k} p(y|\gamma) + \sum_{\gamma \notin SSS_k} p(y|\gamma) \right]$$

Estimate average value of $p(y|\gamma)$ for models in

Randomly sample models within each dimension
Log Marginal Likelihood

- $k = 1$
- $k = 2$
- $k = 3$
- $k = 4$
- $k = 5$
- $k = 6$
Approaches

• Could compute **empirical averages**

• Could make **distributional assumptions**
  – Lower bound?
  – Parametric approximation?
Lower Bound on $p(y \mid \gamma)$

Assuming that $y$ and the $x_j$ are standardized and each $X_{\gamma}$ is full rank, then

$$p(y \mid \gamma) = \frac{\Gamma \left( \frac{n+\delta+k}{2} \right) / \Gamma \left( \frac{\delta+k}{2} \right)}{\pi^{n/2} \tau^{(n-k)/2} |M|^{1/2} \{1 + q/\tau\}^{(n+\delta+k)/2}}$$

$$M = \tau I_k + X'X, \text{ and } q = y'y - y'XM^{-1}X'y.$$
Parametric Approach

- Shift by lower bound
  \[ \log p(y | \gamma) - \log p^*(y | \gamma) \]

- Fit Gamma distribution within each dimension
  - MOM estimates
Gamma Approximation to $\log p(y \mid \gamma)$
Approximate Distribution

Assume: $y$ is independent of $x_j$, and $X_\gamma \sim N(0, I)$:

$$\log \left( \frac{p(y|\gamma)}{p^*(y|\gamma)} \right) \approx \text{Gamma} \left( \frac{k}{2}, \frac{(\tau+n-1)^2}{(n-1)(n+\delta+k)} \right)$$

Can estimate sum of missing marginal likelihoods:

$$\sum_{\gamma \notin SSS_k} p(y|\gamma)$$
Marginal Likelihood for $p(y | p)$
Marginal Posterior $p(\pi \mid y)$

$$p(\pi) = \text{Beta}(k', p) \quad \mathbb{E}[\pi] \approx \frac{k'}{p} \quad \text{and} \quad \mathbb{V}[\pi] \approx \frac{k'}{p^2}$$
Current and Future Work

- Regression Model SSS for
  - binary outcomes (done)
  - censored survival data (Weibull regression, done)

- Priors on $\pi$

- Assessing evidence against null hypothesis

- Extensions to other applications