REGRESSION MODEL UNCERTAINTY
WITH VERY MANY PREDICTORS

Chris Hans and Mike West
Institute of Statistics and Decision Sciences

2nd IMS-ISBA Joint Meeting
Bormio, Italy
12-14 January 2005
Variable Selection and Model Uncertainty

- Model averaging for large datasets with many variables
  - (tens of) thousands of possible predictors
- Regression and prediction
- Novel stochastic search procedures
  - intelligently search “interesting” model subspaces
  - identify key variables for prediction
- Integrate stochastic search with parallel computing
We have
- observations $y_i$, $i = 1, \ldots, n$, and
- predictor variables $x'_j = (x_{1j}, \ldots, x_{nj})$, $j = 1, \ldots, p$.  \hspace{1cm} (p >> n)

Let
- $\gamma$ be a $p \times 1$ indicator $\iff$ a subset of the $x_j$,
- $X\gamma$ be the design matrix under model $\gamma$, and
- $\beta_\gamma$ be the corresponding coefficient vector.

We assume the normal linear model framework for a given model:

$$y = X\gamma \beta_\gamma + \varepsilon, \quad \varepsilon \sim N_n(0, \psi I_n).$$
Prior Distributions

Models:

\[ p(\gamma) = \pi^k(1 - \pi)^{p-k} \]

\( k = \sum_{j=1}^{p} \gamma_j \) is the number of variables in model \( \gamma \).

- “Sparsity” is determined by \( \pi \)
  - Can fix \( \pi \) a priori (prior expected model size)
  - Can model \( \pi \) as a parameter
  - Can estimate \( \pi \) (Empirical Bayes type approach)

Parameters:

\[ \psi^{-1}|\gamma \sim \text{Gamma}\left(\frac{\delta + k}{2}, \frac{\tau}{2}\right) \]

\[ \beta|\psi, \gamma \sim N_k(0, \tau^{-1}\psi I_k) \]

- Consistency (via conditioning) with respect to an encompassing model
- Closed form calculation of marginal likelihood \( p(y|X_\gamma) \)
Exploring the Model Space

- The model space is prohibitively large
- MCMC methods of model space exploration would need to be run excessively long to explore the space adequately

**Shotgun Stochastic Search Method**

- “Shoot out” many proposal models based on the current model
- Evaluate them in parallel
- Sample a new model based on relative posterior probabilities
Proposal Models:

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

\[ t \]

Parallel computing step

Sample a neighboring model
Inference and Model Uncertainty

Using the list of top models discovered by the SSS, we

- condition on the list, creating a posterior distribution of models,
- sample from this posterior distribution,
- implement leave-one-out cross validation via importance sampling, and
- compute model averaged predictions or fitted values for the outcomes.
Survival Study in Brain Cancer

Keck Center for Neurooncogenomics at Duke University

- $n = 41$ glioblastoma patients, with survival times $y_i$
- $p = 8408$ expression levels for genes in tumor tissue
- expression levels are standardized

We assume that

$$\log y \sim N(X'\gamma, \beta \gamma, \psi I_n)$$

and aim to both identify the important genes with respect to survival and build a predictive model for survival times.
Survival Study in Brain Cancer

- **Priors**  \( \tau = 1, \delta = 3, \pi = 2/p \)
- Saved top 10000 models from 20000 iterations
- 2,862 genes appear in this list \( (p = 8408) \)
- Mix of models of small dimension:

<table>
<thead>
<tr>
<th>Dimension:</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td># of models</td>
<td>6493</td>
<td>1578</td>
<td>1912</td>
<td>17</td>
</tr>
<tr>
<td>posterior probability</td>
<td>0.55</td>
<td>0.35</td>
<td>0.10</td>
<td>&lt; 0.01</td>
</tr>
</tbody>
</table>

- Several key genes identified:

<table>
<thead>
<tr>
<th>Gene</th>
<th>sparc</th>
<th>semaphorin</th>
<th>doublecortin</th>
<th>doublecortex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posterior Probability</td>
<td>0.66</td>
<td>0.34</td>
<td>0.19</td>
<td>0.09</td>
</tr>
<tr>
<td>of inclusion</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Survival Study in Brain Cancer

Observations vs Predicted Values

Predicted values vs Log survival times
Survival Study in Brain Cancer

metagene

sparc doublecortex

Regression Model Uncertainty
**CURRENT WORK – INFERENCE ON π**

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

We have the first of the two inner summands, and we can estimate the second. If \(x'_ix_j = 0\), \(y\) is independent of \(x_j\), and \(X_\gamma \sim N(0, I)\), then

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

![Graph showing the distribution of \(p(y|\pi)\)]
ACKNOWLEDGEMENTS

- Adrian Dobra (CAGP – Duke)
- Beatrix Jones (Massey University, NZ)
- Keck Center for Neurooncogenomics at Duke University

CITATIONS

- Manuscript forthcoming for this work