STOCHASTIC SEARCH
&
REGRESSION MODEL UNCERTAINTY
WITH VERY MANY PREDICTORS

Chris Hans and Mike West

Institute of Statistics and Decision Sciences

Duke University

ISBA World Meeting
Viña del Mar, Chile
23-27 May 2004
• Model averaging for prediction in large datasets

• Prior distributions to encourage sparse models

• Novel stochastic search procedures
  - intelligently search “interesting” model subspaces
  - find subsets of similar models (collinearity in data)

• Employ distributed (parallel) computing
We wish to use genomic data to model clinical outcomes in the context of cancer studies.

Possible outcomes include:
- survival times
- tumor recurrence/times
- tumor specific phenotypes
  (e.g. lymph node involvement in breast cancer patients)

⇒ Our goal is to build predictive models to be used in the clinical treatment setting, as well as identifying the genes that drive cancer activity.
We have
- observations \( y_i, i = 1, \ldots, n \), and
- predictor variables \( \mathbf{x}'_j = (x_{1j}, \ldots, x_{nj}), j = 1, \ldots, p \). \((p >> n)\)

Let
- \( \gamma \) be a \( p \times 1 \) indicator \( \iff \) a subset of the \( \mathbf{x}_j \),
- \( X\gamma \) 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:

\[
\mathbf{y} = X\gamma \beta_\gamma + \varepsilon, \quad \varepsilon \sim N_n(0, \psi I_n).
\]
We use the typical model selection prior

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

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

- **Common** inclusion probability: \( \Pr(\gamma_j = 1) = \pi \)
- Induces a binomial distribution over model size
- Must choose \( \pi \) *a priori* or treat it as an unknown parameter
**Incorporating Prior Beliefs about Sparsity**

**Sparsity** is key – we desire small models.

Allows for easy investigation of underlying biology in the genomic context.

Under the above prior, the expected model size is $p\pi$. Thus we set

$$\pi = m/p$$

(*m small*) to place most of the prior mass on sparse models, with average size $m$. 
Prior Distributions – Model Parameters

- Conjugate priors:

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

\[ \beta | \psi, \gamma \sim \text{N}_k(\mathbf{0}, \tau^{-1} \psi I_k) \]

where \( \tau \) is a scale parameter and \( \delta \) is a degrees of freedom parameter.

- Closed form calculation of the marginal likelihood, \( p(y|X\gamma) \).

- Consistency over models with respect to an encompassing model (see Dobra et al., 2004)
Exploring the Model Space

- Even though we focus on sparse models, the model space is prohibitively large (millions of billions of reasonably sized models for the example described below)

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

We describe a model search algorithm that

- quickly identifies interesting regions of the model space;
- considers incorporating each variable in the model at every iteration;
- moves intelligently across dimension; and
- utilizes distributed computing.
In general, a **Shotgun Stochastic Search** is one in which, given a current model, **multiple candidate models** are generated, “shooting out” proposed moves in various directions, with one of these candidates chosen as the new current model:

![Diagram of Shotgun Stochastic Search](image)

The key is that **many candidate models** are evaluated at each iteration, allowing **each variable** to be considered in the context of many different models, and allowing us to compile a running list of the **top models** we’ve evaluated.
Choosing the Proposal Models

Given a current model $\gamma$ of dimension $k$, we consider three types of proposal models:

- neighboring models $\gamma^-$ of dimension $k - 1$,
- neighboring models $\gamma^\circ$ of dimension $k$, and
- neighboring models $\gamma^+$ of dimension $k + 1$.

Together, these three sets of neighboring models make up the set of proposal models that we “shoot out” at each iteration.

A key to the Regression Model SSS approach is that this step can be done in parallel, using distributed computing.
Choosing the New Model

- There are typically many more models in $\gamma^\circ$ than in $\gamma^-$ and $\gamma^+$
- To balance this, we sample a new model in two steps

**Step 1** Sample three proposals:

\[
\begin{align*}
\gamma^- & \quad \text{from the set } \gamma^- \\
\gamma^\circ & \quad \text{from the set } \gamma^\circ \\
\gamma^+ & \quad \text{from the set } \gamma^+
\end{align*}
\]

**Step 2** Sample one of these three proposals as the new model

Sampling is done with probability proportional to the (unnormalized) posterior probability of each model, normalized within each of the three sets.
Regression Model SSS

One iteration in the algorithm:

\[
\gamma^t + f \gamma^t + g \gamma^{t+1}
\]

The blue ovals represent steps which are done in parallel. The red box represents the final three models from which we choose – this step determines the dimension of the new model.
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

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

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 \( \tau = 1, \delta = 3, m = 2 \)
- 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>&gt; 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 of inclusion</td>
<td>0.66</td>
<td>0.34</td>
<td>0.19</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Survival Study in Brain Cancer

Predicted survival probability vs. Months

- Case 1
- Case 2
- Case 3
CURRENT WORK

- Apply to a richer class of models:
  - binary regression already implemented
  - can be applied to censored survival analysis
- Develop prior distributions on the model space to account for collinearity
- Explore adaptive clustering methods
  - cluster variables (metagenes) “on the fly”
  - improve predictions (cleaner signal)
  - generate biological insights into genetic pathways
Acknowledgements

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

Citations

- Manuscript forthcoming for this work