Regression Model Search and Uncertainty
With Many Predictors

Chris Hans
and
Mike West

{hans, mw}@isds.duke.edu

Institute of Statistics and Decision Sciences

Duke University

JSM — Minneapolis, MN
7 August 2005
Regression Model Uncertainty

- Many Possible Predictors
- Model/Variable Selection
- Model Averaging
- Regression Model Space Exploration
- SSS Methods
Large "p" Regression Uncertainty

Model Search Strategies

• **Stepwise Methods**
  – Forward/Backward Selection
  – “Leaps and Bounds” [Furnival & Wilson, 1974]

• **MCMC Methods**
  – Gibbs sampling
  – Metropolis-Hastings
Bayesian Regression Methods

• **Model Space Posterior** [Bayes’ Theorem]

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

• **Marginal Likelihood**

\[ p(y|M) = \int p(y|\theta, M)p(\theta|M)d\theta \]

• **Specification of Prior Distribution**
Sparsity

• Focus on “sparse” models

• $\gamma$ is an indicator vector → regression model
  – sparsity → many zeros
  – “dimension”

• Sparsity encoded in part via the prior

$$Pr(\gamma_j = 1) = \pi \rightarrow p(\gamma) = \pi^k (1 - \pi)^{p-k}$$
 Shotgun Stochastic Search

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

- Current Model
- Parallel Computing Step
- Proposals
- New Model
SSS Output and Posterior Summarization

- As the search progresses:
  - Maintain list $\Gamma^*$ of the best models evaluated
  - Based on a score function $[\log p(y | \gamma) + \log p(\gamma)]$
  - Use these models to summarize the posterior

- Condition on $\Gamma^*$:  
  \[ C = \sum_{\gamma \in \Gamma^*} p(y | \gamma) p(\gamma) \]

  \[ \tilde{p}(\gamma | y) = \frac{p(y | \gamma) p(\gamma)}{C} \]
  \[ \tilde{p}(\gamma_j = 1 | y) = C^{-1} \sum_{\gamma \in \Gamma^*} 1(\gamma_j = 1) p(y | \gamma) p(\gamma) \]
Relationship to MCMC

• Metropolis-Hastings: \[ P(x) = Q(x)/Z \]

  - Use \( P(x) \) restricted to a neighborhood \( B(x) \) as the proposal distribution

\[
T(x_{t+1}; x_t) = \frac{Q(x_{t+1})1(x_{t+1} \in B(x_t))}{\sum_{s \in B(x_t)} Q(s)}
\]

  - Acceptance probability:

\[
\alpha = \min \left\{ 1, \frac{Q(B(x_{t+1}))}{Q(B(x_t))} \right\} \quad \text{vs.} \quad \alpha^* = \min \left\{ 1, \frac{p(y|\gamma')p(\gamma')}{p(y|\gamma[t])p(\gamma[t])} \right\}
\]
Large "p" Regression Uncertainty

Comparison to MCMC

- 40,000 SSS iterations
- SSS: 11 hrs. 53 min.
- 29,163 Gibbs iterations
- Gibbs: 75.41%

- 1,137,195,208 model evaluations
- 135,252 Gibbs iterations
- Gibbs: 55 hrs. 13 min.
- Gibbs: 97.49%
Example: Lung Cancer Survival

- **Weibull Regression**
  - $n = 91$ patients, $d = 45$ observed survival times
  - $p = 2,717$ probesets (genes)
Concluding Remarks

• Linear, Binary, Weibull Regression…

• Connections to MCMC
  – Advantageous Use of Parallel Computing

• Code Available
  – http://www.isds.duke.edu/research/software/

  hans@stat.ohio-state.edu