A Review of Bayesian Variable Selection

By: Brian J Reich and Sujit Ghosh,
North Carolina State University

E-mail: reich@stat.ncsu.edu

Slides, references, and BUGS code are on my webpage,
http://www4.stat.ncsu.edu/~reich/.
Why Bayesian variable selection?

- Bayesian variable selection methods come equipped with natural measures of uncertainty, such as the posterior probability of each model and the marginal inclusion probabilities of the individual predictors.

- Stochastic variable search is extremely flexible and has been applied in a wide range of applications.

- Model uncertainty can be incorporated into prediction through model averaging, which usually improves prediction.

- Given the model (likelihood, prior, and loss function), there are formal justifications for choosing a particular model.

- Missing data is easily handled by MCMC.
Drawbacks of Bayesian methods

Computational issues:

▶ For the usual regression setting, the “BMA” (Bayesian Model Averaging) package in R can be used for Bayesian variable selection.

▶ WinBUGS is a free user-friendly software package that can be used for stochastic variable search.

Prior selection:

▶ Bayesian variable selection can be influenced by the prior.

▶ Priors can be chosen to mimic Frequentist solutions (Yuan and Lin).

▶ Priors can be derived to have large sample properties (Ishwaran and Rao).
Full model: \( y_i = \beta_0 + x_{1i} \beta_1 + ... + x_{pi} \beta_p + \epsilon_i \), where \( \epsilon_i \sim \text{iid} \, \mathcal{N}(0, \sigma^2) \).

The objective is to find a subset of the predictors that fits the data well.

Let \( \gamma_j = 1 \) if \( x_j \) is in the model and \( \gamma_j = 0 \) otherwise.

The vector of indicators \( \gamma = (\gamma_1, ..., \gamma_p) \) represents the subset of predictors in the model.

\( X_\gamma \) is the design matrix that includes only covariates with \( \gamma_j = 1 \).

The regression coefficients \( \beta_\gamma \) are often given the conjugate Zellner \( g \)-prior \( \beta_\gamma \sim \mathcal{N}(0, g \sigma^2 (X'_\gamma X_\gamma)^{-1}) \).
One way to compare models defined by $\gamma(1)$ and $\gamma(2)$ is with a Bayes factor.

\[
BF = \frac{p(\gamma(1)|y)/p(\gamma(2)|y)}{p(\gamma(1))/p(\gamma(2))},
\]

where

\[
p(\gamma(1)|y) \propto \int p(\gamma(1), \beta_{\gamma(1)}, \sigma^2|y) d\beta_{\gamma(1)} d\sigma^2
\]
is the posterior probability of the model specified by $\gamma(1)$.

$p(\gamma(1))$ is the prior probability of model specified by $\gamma(1)$. 

Bayesian variable selection
Schwarz (1978) showed that for large $n$,
\[-2\log(BF) \approx \Delta BIC = W - \log(n)(|\gamma(2)| - |\gamma(1)|),\]
where
\[
\Delta BIC \text{ is the change in } BIC \text{ from model 1 to model 2,}
\]
\[
W = -2\log \left[ \frac{\sup_{\gamma(1)} p(y|\beta_{\gamma(1)}\sigma^2)}{\sup_{\gamma(2)} p(y|\beta_{\gamma(2)}\sigma^2)} \right] \text{ is the usual likelihood ratio statistic,}
\]
and $|\gamma| = \sum_{j=1}^{p} \gamma_j$ is the number of parameters in the model defined by $\gamma$. 
The BIC approximation can be used to compute the posterior probability of each model.

This is implemented by the “BMA” package of Raftery et al.

The R command is $\text{bicreg}(x,y)$ where $x$ is the full $n \times p$ design matrix and $y$ is the vector of outcomes.

“BMA” also has functions for survival and generalized linear models.

For a list of papers and software for Bayesian model averaging, visit the BMA homepage, http://www.research.att.com/~volinsky/bma.html.

Example: subset of the Boston Housing data.
Boston Housing Data

Dependent variable \((n=100)\)
- MV: Median value of owner-occupied homes in $1000s

Independent variables \((p = 6)\)
- INDUS: Proportion of non-retail business acres per town.
- NOX: Nitric oxides concentration (parts per 10 million).
- RM: Average number of rooms per dwelling.
- TAX: Full-value property-tax rate per $10,000.
- PT: Pupil-teacher ratio.
- LSTAT: Proportion of population of lower status.
The results of usual linear regression are:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>SE</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INDUS</td>
<td>-0.31</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>NOX</td>
<td>-0.61</td>
<td>0.31</td>
<td>0.05</td>
</tr>
<tr>
<td>RM</td>
<td>3.35</td>
<td>0.33</td>
<td>0.00</td>
</tr>
<tr>
<td>TAX</td>
<td>-0.45</td>
<td>0.28</td>
<td>0.12</td>
</tr>
<tr>
<td>PT</td>
<td>-0.63</td>
<td>0.27</td>
<td>0.02</td>
</tr>
<tr>
<td>LSTAT</td>
<td>-1.77</td>
<td>0.35</td>
<td>0.00</td>
</tr>
</tbody>
</table>
The results of the R command `summary(bicreg(x,y))` are:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Post mean</th>
<th>Post sd.</th>
<th>Prob non-zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>INDUS</td>
<td>-0.13</td>
<td>0.25</td>
<td>0.28</td>
</tr>
<tr>
<td>NOX</td>
<td>-0.34</td>
<td>0.26</td>
<td>0.51</td>
</tr>
<tr>
<td>RM</td>
<td>3.45</td>
<td>0.41</td>
<td>1.00</td>
</tr>
<tr>
<td>TAX</td>
<td>-0.24</td>
<td>0.34</td>
<td>0.43</td>
</tr>
<tr>
<td>PT</td>
<td>-0.68</td>
<td>0.33</td>
<td>0.86</td>
</tr>
<tr>
<td>LSTAT</td>
<td>-1.93</td>
<td>0.38</td>
<td>1.00</td>
</tr>
</tbody>
</table>

The model with highest posterior probability (prob = 0.20) has RM, PT, and LSTAT.
Stochastic variable selection

- A limitation of the Bayes factor approach is that for large $p$, computing the BIC for all $2^p$ models is impossible.

- Madigan and Raftery (1994) limit the search to a class of models under consideration to models with a minimum level of posterior support and propose a greedy-search algorithm to find models with high posterior probability.

- Stochastic variable selection is an alternative.

- Here we use MCMC to draw samples from the model space so models with high posterior probability will be visited more often than low probability models.

- We’ll follow the notation and model of George and McCullough, JASA, 1993.
Likelihood: \( y_i = \beta_0 + x_{1i}\beta_1 + \ldots + x_{pi}\beta_p + \epsilon_i \), where \( \epsilon_i \overset{iid}{\sim} N(0,\sigma^2) \).

Mixture prior for \( \beta_j \): \( \beta_j \sim (1 - \gamma_j)N(0,\tau_j^2) + \gamma_jN(0,c_j^2\tau_j^2) \).

The constant \( \tau_j^2 \) is small, so that if \( \gamma_j = 0 \), \( \beta_j \) “could be ‘safely’ estimated by 0”.

The constant \( c_j^2 \) is large, so that if \( \gamma_j = 1 \), “a non-zero estimate of \( \beta_j \) should probably be included in the final model”.

One of the priors for the inclusion indicators they consider is \( \gamma_j \sim \text{Bernoulli}(0.5) \) prior.
A similar SVS model that has positive prior probably that $\beta_j = 0$ is the “spike and slab” prior of Mitchell and Beauchamp (1988).

One disadvantage of SVS when $p$ is large is that in 50,000 MCMC iterations, the model with highest posterior model may only be visited a handful of times.

SVS works well for computing the marginal inclusion probabilities of each covariate and for model averaging.

SVS is computationally convenient and extremely flexible.
# The likelihood:
for(i in 1:100){
    MV[i]~dnorm(mu[i],tau.e)
}

int~dnorm(0,0.01)  # int has prior variance 1/0.01.
tau.e~dgamma(0.1,0.1)  # tau.e is a precision (inverse variance)

# The priors:
for(j in 1:6){
    b[j]~dnorm(0,prec[j])
    prec[j]←1/var[j]
    var[j]←(1-gamma[j])*0.001 + gamma[j]*10
    gamma[j]~dbern(0.5)
}
The results of the SVS model from WinBUGS are:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Post mean</th>
<th>Post sd.</th>
<th>Prob non-zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>INDUS</td>
<td>-0.13</td>
<td>0.26</td>
<td>0.27</td>
</tr>
<tr>
<td>NOX</td>
<td>-0.34</td>
<td>0.41</td>
<td>0.48</td>
</tr>
<tr>
<td>RM</td>
<td>3.44</td>
<td>0.34</td>
<td>1.00</td>
</tr>
<tr>
<td>TAX</td>
<td>-0.22</td>
<td>0.32</td>
<td>0.39</td>
</tr>
<tr>
<td>PT</td>
<td>-0.64</td>
<td>0.40</td>
<td>0.81</td>
</tr>
<tr>
<td>LSTAT</td>
<td>-1.93</td>
<td>0.38</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Applications of SVS: Choosing the priors for $\beta$ and $\gamma$:

- Connection with the LASSO (Yuan and Li, JASA, 2005)

- Rescaled spike and slab priors of Ishwaran and Rao (Annals, 2005).

- Gene selection using microarray data.

- Nonparametric regression for complex computer models
Connection with penalized regression

- Yuan and Lin (JASA, 2005) show a connection between Bayesian variable selection and the LASSO penalized regression method.

- The LASSO solution is
  \[ \hat{\beta} = \arg\min_{\beta} \left( ||y - X\beta||^2 + \lambda \sum |\beta_j| \right) . \]

- The L1 penalty for the regression coefficients encourages sparsity.

- Yuan and Lin’s spike and slab prior for \( \beta_j \) is
  \[ \beta_j = (1 - \gamma_i)\delta(0) + \gamma_i DE(0, \lambda). \]

- \( \delta(0) \) is the point mass distribution centered at zero.

- \( DE(0, \lambda) \) is the double exponential distribution with density \( \lambda \exp(-\lambda|x|) \).
Connection with penalized regression

- The prior for the model indicators is
  \[ p(\gamma) \propto \pi|\gamma|(1 - \pi)^{p-|\gamma|}|X_{\gamma}X_{\gamma}|^{1/2}. \]

- \(|\gamma| = \sum_{i=1}^{p} \gamma_j\) is the number of predictors in the model.

- \(|X_{\gamma}X_{\gamma}|^{1/2}\) penalizes models with correlated predictors.

- Using a Laplace approximation, they show that the posterior mode of \(\gamma\) selects the same parameters as the LASSO.

- There are fast algorithms to find the LASSO solution, so this helps find models with high posterior probability.

- Yuan and Lin also propose an empirical Bayes estimate for the tuning parameter \(\lambda\).
Rescaled spike and slab model of Ishwaran and Rao

Under the usual SVS model, the effect of the prior vanishes as \( n \) increases and coefficients are rarely set to zero.

Ishwaran and Rao propose rescaled model that gives the prior a non-vanishing effect.

The covariates are standardized so \( \sum_{i=1}^{n} x_{ji} = 0 \) and \( \sum_{i=1}^{n} x_{ji}^2 = n \).

Let \( y_i^* = y_i / \sqrt{\hat{\sigma}^2 / n} \), where \( \hat{\sigma}^2 \) is the OLS estimate of \( \sigma^2 \).

The rescaled model is \( y_i^* = x_i' \beta + N(0, \sigma^2 \lambda_n) \) (say, \( \lambda_n = n \)).

\( \sigma^2 \) has an inverse gamma prior that does not depend on \( n \).

\( \beta_j \) has a continuous bimodal prior similar to George and McCullough.
Properties of the rescaled spike and slab model

- If $\lambda_n \to \infty$ and $\lambda_n/n \to 0$, the posterior mean of $\beta$ is consistent. While consistency is important, they choose $\lambda_n = n$ for variable selection.

- With $\lambda_n = n$, the posterior mean of $\beta$, $\hat{\beta}$, asymptotically maximizes the posterior and they recommend selecting the Zcut model that includes only variables with $|\hat{\beta}_j| > z_{\alpha/2}$.

- They compare this rule with OLS-hard rule that includes only variables with estimated z-scores more extreme than $|z_{\alpha/2}|$.

- They argue that since information of the posterior mean of $\beta$ is pooled across several models that the Zcut model is better than the OLS-hard model.

- Under certain conditions, they show that the Zcut model has an oracle property relative to the OLS-hard model, that is, the Zcut model includes less unimportant variables.
Gene selection

- SVS has been used for microarray data to identify genes that can distinguish between the two types of breast cancer. For example, Lee et al. (including Marina Vannucci), *Bioinformatics*, 2003.

- $y_i$ indicates the type of breast cancer for patient $i$.

- $x_{ji}$ measures the expression level of gene $j$ for patient $i$.

- Probit model: $y_i = \begin{cases} 1, & x_i \beta + \epsilon_i > 0; \\ 0, & x_i \beta + \epsilon_i \leq 0. \end{cases}$

- $\epsilon_i \sim N(0, 1)$.

- $\beta_\gamma \sim N \left( 0, c(X_\gamma' X_\gamma)^{-1} \right), \gamma_j \overset{iid}{\sim} Bern(\pi)$. 

Brian Reich

Bayesian variable selection
For complex computer models that take a long time to run, scientists often run a moderate number of simulations and use nonparametric regression to emulate the computer model.

Nonparametric regression: $y_i = f(x_{1i}, \ldots, x_{pi}) + \epsilon_i$

The goal is to estimate $f$.

A common model for $f$ is a Gaussian process prior with covariance function

$$\text{Cov} [f(x_{1i}, \ldots, x_{pi}), f(x_{1j}, \ldots, x_{pj})] = \tau^2 \exp(-\rho d_{ij}),$$

where $d_{ij} = \sqrt{(x_{1i} - x_{1j})^2 + \ldots + (x_{pi} - x_{pj})^2}$.

Should all the variables contribute equally to the distance measure in the covariance function?
One thing we’re (Drs. Bondell and Storlie) working on is an anisotropic covariance function that allows the variables to make different contributions to the distance measure, i.e.,

$$\tau^2 \exp \left( -\sqrt{\rho_1^2(x_{1i} - x_{1j})^2 + \ldots + \rho_p^2(x_{pi} - x_{pj})^2} \right)$$

A variable selection mixture prior for $\rho_j$ is

$$\rho_j = \gamma_j \text{Unif}(c_1, c_2) + (1 - \gamma_j)\delta(c_1),$$

where $c_1$ is chosen so that $c_1^2(x_{ji} - x_{ji'})^2 \approx 0$ for all $x_{ji}$ and $x_{ji'}$.

The posterior mean of $\gamma_j$ can be interpreted as the posterior probability that $x_j$ is in the model.

If $\gamma_1 = \ldots = \gamma_p = 0$, $f$ is essentially smoothed to zero for all $x$. 
Conclusions

- The Bayesian approach to variable selection has several attractive features, including straightforward quantification of variable importance and model uncertainty and the flexibility to handle missing data and non-Gaussian distributions.

- Outstanding issues include computational efficiency and concerns about priors.

- Hopefully we will see some interesting talks about these issues throughout the semester.

- THANKS!
Conclusions

▶ The Bayesian approach to variable selection has several attractive features, including straight-forward quantification of variable importance and model uncertainty and the flexibility to handle missing data and non-Gaussian distributions.

▶ Outstanding issues include computational efficiency and concerns about priors.

▶ Hopefully we will see some interesting talks about these issues throughout the semester.

▶ THANKS!