Bayesian Variable Selection in GLMs

• Suppose we start with a vector of $p$ candidate predictors, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$

• In variable selection, our goal is to select an subset of $q \leq p$ predictors to be included in the model.

• Potentially, one can fit different possibilities and then compare using a criterion-based method, such as AIC or BIC.

• Alternative, a frequentist stepwise selection approach could be used
• A formal Bayesian alternative is to choose a prior on the set of possible models.

• Letting the linear predictor in the full model be $\eta_i = x_i'\beta$, a common choice is:

$$\pi(\beta) = \prod_{j=1}^{p} \delta_0(\beta_h)p_{0h} + (1 - p_{0h})N(\beta_h; 0, c_h^2),$$

with $p_{0h} = p_0 = 0.5$ and $c_h = c$ as a common convention.

• This is a mixture prior consisting of a point mass at 0 (with probability $p_0$) and a normal prior.
• Most of the attention in the Bayesian literature has focused on the linear regression case.

• For linear regression, the variable selection mixture prior is conjugate.

• This is easily seen, by calculating the full conditional posterior distribution for $\beta_h$. 
• The case in which $\beta_h = 0$ corresponds to the $h$th predictor being excluded from the model.

• We can calculate the marginal probability of $\Pr(\beta_h = 0 \mid \beta_{-h}, \text{data})$ in closed form for normal linear regression.

• For other GLMs, these marginal model probabilities are not available in closed form.
• The typical convention is either to approximate the marginal probabilities (e.g., using Taylor series or BIC-based approximations) or use data augmentation.

• In particular, for underlying normal probit models, we have conditional-conjugacy after data augmentation.

• Hence, in these special cases we can apply a stochastic search Gibbs sampling algorithm in which we sequentially sample from the conditional posterior distributions of $\beta$ to explore the model space.
Stochastic Search Variable Selection in Linear Regression

- **Likelihood:**
  \[
  \pi(y \mid X, \beta, \sigma^2) = \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2\right\}.
  \]

- **Priors:**
  \[
  \pi(\beta) = \prod_{j=1}^{p} \delta_0(\beta_j)p_{0j} + (1 - p_{0j})N(\beta_j; 0, c_j^2),
  \]
  \[
  \pi(\sigma^{-2}) = G(\sigma^{-2}; a, b),
  \]
  where \(G(x; a, b) = b^a/\Gamma(a)x^{a-1}\exp(-bx)\) is the gamma density.
• Conditional Posterior Distribution of Regression Coefficients:
\[
\pi(\beta_j | \beta_{(-j)}, y, X, \sigma^2) = \hat{p}_j \delta_0(\beta_j) + (1 - \hat{p}_0) N(\beta_j; E_j, V_j),
\]
where \( V_j = (c_j^{-2} + \sigma^{-2} X'X)^{-1} \), \( E_j = V_j X'y \), and
\[
\hat{p}_j = \frac{p_{0j}}{p_{0j} + (1 - p_{0j}) \frac{N(0; 0, c_j^2)}{N(0; E_j, V_j)}}
\]
is the conditional probability of \( \beta_j = 0 \) (i.e., we exclude the \( j \)th predictor).

• This conditional can be calculated by straightforward algebra by multiplying the priors by the likelihood, factoring out terms not involving \( \beta_j \), and normalizing.
• The conditional posterior distribution of $\sigma^{-2}$ also has a simple form, which can be calculated following similar algebraic routes:

$$
\pi(\sigma^{-2} | \beta, y, X) = \mathcal{G}\left(a + \frac{n}{2}, b + \frac{1}{2} \sum_{i=1}^{n} (y_i - x'_i \beta)^2\right).
$$

• The stochastic search variable selection Gibbs sampler alternately samples from the conditional posterior distributions of the $\beta_j$'s and $\sigma^2$.

• After convergence, this algorithm generates samples of models, corresponding to subsets of the set of $p$ candidate predictors, from the posterior distribution.
• Based on a large number of iterations of the stochastic search Gibbs sampler, we can estimate posterior probabilities for each of the models.

• These models are subsets of the original $p$ predictors.

• For example, the full model may appear in 10% of the samples collected after convergence, so that model would be assigned posterior probability of 0.10.
• Different posterior probabilities will be assigned to each of the possible models

• To summarize, one can present a table of the top 10 or 100 models

• It is also useful to calculate marginal probabilities that a given predictor is present across the different models.