Stochastic search and model averaging

- So far we have selected from a small pool of models.

- What about linear regression with \( p \) predictors and thus \( 2^p \) possible models?

- We surely can’t compute DIC for each model if \( p \) is large.

- Stochastic search variable selection is viable alternative.

- In addition to computational problems, linear regression with large \( p \) also makes it unlikely that any one model will have high posterior probability.

- Therefore, rather than pick a single model, we can improve prediction by averaging predictions across models.
Stochastic search variable selection (SSVS)

- Consider the usual regression model $Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i$.

- If $p$ is large we might want to select a small subset of the predictors.

- If $\delta_j = 1$ covariate $j$ is included; if $\delta_j = 0$ it is excluded.

- Let $\delta = (\delta_1, \ldots, \delta_p)$ indicate the subset of covariates included in the model.

- For example, $\delta = (1, 0, 1, 0, \ldots, 0)$ denotes the model $E(Y_i) = \beta_0 + X_{i1}\beta_1 + X_{i3}\beta_3$.

- Variable selection is the search for the optimal $\delta$.

- Options:
SSVS

- SSVS treats $\delta$ as an unknown parameter to be estimated and updated in MCMC.
- A common prior the inclusion indicators is:

- We also need a prior for the regression coefficients for the variables that are included:

- An equivalent model is:

- All parameters have conjugate full conditionals!
SSVS

- An equivalent formulation (actually originally proposed by George and McCulloch) is:

$Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij} \beta_j + \sum_{k<l}^{p} X_{ik} X_{il} \beta_{kl} + \epsilon_i.$

- This general idea has many applications. For example, consider the model for the regression with interactions

- How would you specify a prior so that interactions are included only if both main effects are included?
SSVS

Three ideas for summarizing the posterior

1. Model probabilities, $p(\delta|y)$.

2. Marginal inclusion probabilities, $\text{Prob}(\delta_j = 1|y)$.

3. Marginal 95% intervals for $\beta_j|y$. 
Bayesian model averaging (BMA)

- For prediction of $Y_{new}|X_{new}$, we could pick a single model, refit, and use the posterior predictive distribution.
- Let $\hat{Y}_j = X_{new}\hat{\beta}_j$ be the prediction under model $j$.
- Rather than pick one model, we could average predictions across models

$$\hat{Y} = \sum_{j=1}^{k} w_j \hat{Y}_j.$$  

- A natural choice for weights is:

- This is the Bayes rule under squared error loss.
- BMA is easy to compute using SSVS.
- You simply make a draw of $Y_{new}$ at each MCMC iteration from the current model, and this averages predictions appropriately over models.
- Code is available at

  http://www4.stat.ncsu.edu/~reich/ST740/code/SSVS_BMA.R.