STA 605 Homework 3

Due on 11:59 pm CT, November 17, 2021 at Canvas

Name:

INSTRUCTIONS:

- **Show all work**, clearly and in order, if you want to receive full credit. When you use your calculator, explain all relevant mathematics. I reserve the right to take off points if I cannot see how you arrived at your answer (even if your final answer is correct).
- Circle or otherwise indicate your final answers.
- **Answer all the questions in the space provided. You may attach additional sheets if necessary.**
- This test has 1 problem and is worth 100 points. It is your responsibility to make sure that you have all of the problems.
- **Good luck!**

Prob. No.	Max Points	Earned Pts.
1	100	

TOTAL: ___________

Question 1. (100 pts.) Conduct a replicated study to assess the performance of variational Bayes and Markov chain Monte Carlo methods (both spike and slab and continuous shrinkage priors) for Bayesian variable selection. Let $n \in \{100, 200\}$ and $p \in \{200, 500\}$. For each of the four (n, p) combinations, generate the design matrix X by drawing the subject-specific vector of covariates x_i from a mean-zero Gaussian distribution, $x_i \overset{ind.}{\sim} \mathcal{N}_p(0, \Sigma)$ for $i = 1, \ldots, n$. Consider $\Sigma_{jj'} = \rho + (1-\rho)\mathbb{1}(j = j')$ with $\rho = 0.5$ (correlated design with a compound symmetry covariance structure). Column-standardize the Gaussian design matrix and continue to denote it by X. Then generate the response vector Y be setting

$$Y = F + \sigma_0\, \varepsilon, \quad F = X\beta_0, \quad \varepsilon \sim \mathcal{N}(0, I_n),$$

where β_0 is an k_0-sparse vector with the non-zero coordinates all equaling one. Ler $k_0 = 10$. Set the residual variance σ_0^2 to different values to control the signal-to-noise (SNR) ratio. Specifically, vary SNR $\in \{2, 4\}$, and set

$$\sigma_0^2 = \frac{\mathrm{var}_n(F)}{\mathrm{SNR} \times \mathrm{var}_n(\varepsilon)},$$

with $\mathrm{var}_n(z) = n^{-1}\sum_{i=1}^n (z_i - \bar{z})^2$ for $z \in \mathbb{R}^n$.

The above combinations led to a total 8 different simulation settings. For each setting, generate 50 independent simulation replicates and implement the variational Bayes method (using the `varbvs` package) and the MCMC algorithm outlined in [1] involving single and double flip updates and the Horseshoe. Use the hyperparameters recommended in the paper. For variable selection, use the median probability model (variables with marginal inclusion probability ≥ 0.5) for both the methods.

For each method (use 2 means algorithm on the absolute value of the coefficients to separate out the signals and the noise coefficients in the MCMC samples for Horseshoe), report the following three summary measures:

1. zero-one error, $\sum_{j=1}^p [\mathbb{1}(\hat{\beta}_j = 0, \beta_{0j} \neq 0) + \mathbb{1}(\hat{\beta}_j \neq 0, \beta_{0j} = 0)]$,
2. support size, $\sum_{j=1}^p [\mathbb{1}(\hat{\beta}_j \neq 0)]$.
3. CPU run-time of the algorithm in seconds.

References

[1] Yun Yang, Martin J Wainwright, and Michael I Jordan. On the computational complexity of high-dimensional bayesian variable selection. *The Annals of Statistics*, 44(6):2497–2532, 2016.