Bayesian Analysis (2009) **4**, Number 3, pp. 449–452

Comment on Article by Monni and Tadesse

Hongzhe Li[*]

I congratulate Dr. Monni and Dr. Tadesse (MT) on an elegant Bayesian implementation of an important problem of linking two types of high-dimensional genomic data in small sample size settings. This type of data appears frequently in genomic research. MT demonstrated their methods using the gene expression and array CGH data on NCI 60 cell lines samples. Other potential applications include identifying the SNPs that are associated with gene expression variations (e.g., in the context of eQTL analysis) and identifying the epigenomic features that are associated with genomic features. The methods of MT represent a major methodological development in the area of stochastic partitioning and Bayesian variable selection and will find many applications in these areas. My discussion consists of two parts: (1) some comments on simulations and application to NCI60 cancer cell line data set; and (2) an alternative approach to the same problem based on penalized likelihood and regularization.

1 Comments on simulations and real data analysis

I suspect that the very high signal-to-noise ratios (SNR) used for the first set of simulations have led to almost perfect performance of the proposed procedure, as represented in Figure 1 and Figure 2 of the paper. It is not surprising that the method of MT performed better than the multivariate method of Brown *et al.* (1998) for the simulated scenario since the later method allows for possible different regression coefficients for the same set of covariates over different responses in the same partition. I was wondering how the univariate stochastic search variable selection (SSVS) algorithm, when applied to each response separately, performs in such high SNR settings. I therefore would put more weights on the results presented in Section 4.1.6 when the regression coefficients were sampled in the range [-1.5,-0.5] and [0.5,1.5]. I was wondering whether the authors have similar plots as Figure 1 and Figure 2 for this set of simulations. I would explain the better performance of the proposed method over the SSVS by the implicit increases in sample sizes when the correct partitions of the responses are identified since the same mean models are assumed for all the responses in the same partition. I was wondering whether MT have checked what would happen if different responses in the same partition depend on the same set of the covariates but with different coefficients.

The results from analysis of aCGH and gene expression profiles based on the NCI 60 cell lines are interesting and provide certain insights on how copy number changes affect the gene expressions. For example, the deletion of the $c-abl$ oncogene 1 ($ABL1$), a receptor tyrosine kinase, in leukemia cell lines was found to be related to increased transcript abundance in four genes involved in hematopoietic development and lymphocyte proliferation. While Figure 3 shows that the four genes have similar expression

[*]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA mailto:hongzhe@upenn.edu

DOI:10.1214/09-BA416C