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

profiles over 59 samples, it would be interesting to also show how the expression levels are related to the CGH measurements of $ABL1$ gene and the corresponding regression coefficients. Similarly, for Figure 5, it would be interesting to show how the CGH measurements of $BRYP$ clone are related to the expression levels of the three genes shown. It would also help to assess whether the linearity assumption holds. In contrast, Figure 4 is less interesting, as it simply pointed out that four probe sets of the same gene can be partitioned together into one group. I was also curious whether these transcripts and the CGH clones identified in Figures 3-5 can also be identified based on simple univariate analysis, e.g., whether these pairs rank on the top based on univariate analysis. This could be a great demonstration that simultaneous partitions of the responses and the covariates can indeed lead to something more substantial than simple univariate analysis. My last comment is that since the true copy numbers for clones are in fact discrete, I was wondering whether it might be a good idea to first estimate these copy numbers and then to link these copy numbers to the gene expression data, rather than directly using the CGH intensities as the covariates.

2 An alternative approach based on penalized likelihood

The model considered by MT is essentially a minor modification of the sparse regression mixture model (SRMM) introduced in Khalili and Chen (2007) and Li (2008). Using the same notation as in MT, let the data consist of N independent samples with p covariates, $\mathcal{X} = (X_1, \ldots, X_p)$ and q outcomes $\mathcal{Y} = (Y_1, \cdots, Y_q)$. Assume that each response Y_i can be assigned to one and only one of $M+1$ clusters. Define Z_j to be a random variable that follows a multinomial distribution $Mult(\pi)$ with $\pi = (\pi_0, \pi_1, \cdots, \pi_M)', \pi_m \geq 0$ and $\sum_{m=0}^{M} \pi_m = 1$. We assume that

$$\begin{aligned} Y_{ji}|Z_j = m &\sim \mathcal{N}(\alpha_j + \mu_{m,i}, \sigma_m^2), \\ \mu_{m,i} &= \sum_{r=1}^{p} \beta_{mr} X_{qi}, \end{aligned} \tag{1}$$

for $j = 1, \cdots, q$, $i = 1, \cdots, N$ and $m = 0, 1, \cdots, M$. We assume that $\beta_{0r} \equiv 0$ for $r = 1, \cdots, q$, which corresponds to the cluster that has no regression covariates associated with the response variable. If we appropriately center both the responses and the covariates, we can simply let $\alpha_j = 0$ for all $j = 1, \cdots, q$. MT consider the settings when p and q are large and the regression models are sparse and aim to simultaneously estimate the cluster membership Z_j and the corresponding covariates that can explain the cluster-specific variation of the responses. This is basically a variable selection problem, aiming to identify the non-zero elements of the vector $\beta_m = (\beta_{m1}, \cdots, \beta_{mp})'$ for each cluster or partition of the responses. We can impose the following sparsity constraints to model (1),

$$|\beta_m|_1 = \sum_{r=1}^{p} |\beta_{mr}| < s_m, \ m = 1, \cdots, M,$$

for some tuning parameters $s_m, m = 1, \cdots, M$. This is the SRMM considered in Li (2008).

One way to estimate the model parameters is through a penalized log-likelihood, which can be defined as

$$\tilde{l}(\beta, \pi, \sigma^2) = l(\beta, \pi, \sigma^2) + C_M \sum_{m=0}^{M} \log \pi_m - \sum_{m=1}^{M} \pi_m \lambda_m |\beta_m|_1, \tag{2}$$

where $\beta = (\beta_1, \cdots, \beta_M), \sigma^2 = (\sigma_0^2, \sigma_1^2, \cdots, \sigma_M^2)'$, $l(\beta, \pi, \sigma^2)$ is the standard log-likelihood function based on the mixture model with $M + 1$ groups, C_M and λ_m are the tuning parameters. For simplicity we can assume that $\lambda_1 = \cdots = \lambda_M = \lambda$. The first penalty function forces the estimated values of π_m away from 0 to prevent over-fitting with small values of mixing proportions (Chen and Kalbfleisch, 1996; Chen and Khalili, 2008). The second penalty function induces the sparse solutions and leads to cluster-specific variable selection, which serves the purpose of partitioning the covariates as in MT. Other penalty functions such as the SCAD (Fan and Li, 2001), bridge (Frank and Friedman, 1993) and the minimax concave penalty (MCP, Zhang, 2007) can also be used for the second penalty term.

For a given M, a modified EM-algorithm can be developed for maximizing the penalized log-likelihood (2) (Khalili and Chen, 2007; Li 2008), where the E-step is essentially the same as the standard finite mixture models and the M-step involves maximizing the conditional expectation of the complete data log-likelihood, which is given at the $(k+1)$th M-step as

$$\begin{aligned} Q(\Psi; \Psi^{(k)}) \quad = \quad & \sum_{j=1}^{q} \sum_{m=1}^{M} w_{jm}^{(k)} \log\{f(Y_j; \beta_m, \sigma_m^2)\} - \lambda \sum_{m=1}^{M} \pi_m |\beta_m|_1 \\ & + \sum_{j=1}^{q} \sum_{m=1}^{M} [w_{jm}^{(k)} + \frac{C_M}{N}] \log \pi_m, \end{aligned}$$

where $w_{jm}^{(k)}$ is the conditional probability of $Z_j = m$ given the data and the current estimate of the model parameters $\Psi^{(k)}$ and $f(Y_j; \beta_m, \sigma_m^2)$ is the normal density based on model (1). An efficient cyclical coordinate descent algorithm (Friedman *et al.*, 2007) can be used to find the β which maximizes $Q(\Psi; \Psi^{(k)})$.

We now briefly discuss the choice of the tuning parameter λ, C_M and also the number of clusters M. The results are expected to be not very sensitive to the value of C_M (see Chen and Khalili, 2008). For a given M, we can use a componentwise deviance-based GCV criterion for choosing $\lambda(M)$, which leads to estimate of the effective number of the parameters. We can then use the BIC to choose the number of clusters M.

It would be interesting to compare the performance of the stochastic partitioning method of MT and the regularization method for both simulated and NCI 60 cell line data sets. I was wondering whether MT can comment on the potential advantages of their Bayesan approach over the regularization-based approach outlined above.

Finally, for some genomic applications, it may make more sense to assume that some responses are clustered together if they are affected by the same predictors, but with possible different coefficients. I was wondering whether MT have any thoughts on whether it is possible to extend their stochastic partitioning method to such settings.

References

Brown, P.J., Vannucci, M. and Fearn, T. 1998. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Er. B*, 60: 627-641.

Chen, J. and Kalbfleisch, J.D. 1996. Penalized minimum-distance estimates in finite mixture models. *The Canadian Journal of Statistics*, 24: 167-175.

Chen, J. and Khalili, A. 2008. Order selection in finite mixture models. *Journal of American Statistical Association* 103: 1674-1683.

Fan, J. and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, 96: 1348-1360.

Frank, I.E. and Friedman, J.H. 1993. A statistical view of some chemometrics regression tools. *Technometics*, 35: 109-135.

Friedman, J., Hastie, T., Hoefling, H. and Tibshirani, R. 2007. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1, 302-332.

Khalili, A. and Chen, J. 2007. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association,* 102: 1025-1038.

Li, H. 2008. Statistical methods for inference of genetic networks and regulatory modules. *Analysis of Microarray Data: Network-based Approaches.* Edited by Emmert-Streid and Dehmer. Wiley VCH. pp 143-167.

Zhang, C.H. 2007. Penalized linear unbiased selection. Technical Report #2007-003, Department of Statistics and Biostatistics, Rutgers University, Piscataway, New Jersey.

Acknowledgments

This research was supported in part by NIH grant CA127334.