BIOINFORMATICS LETTER TO THE EDITOR Vol. 27 no. 8 2011, page 1194
doi:10.1093/bioinformatics/btr071

Gene expression

Advance Access publication February 11, 2011

Comments on 'Bayesian variable selection for disease classification using gene expression data'

Meïli C. Baragatti[1,2,*] and Denys Pommeret[2]

[1]Ipsogen SA, Luminy Biotech Entreprises, Case 923 and [2]Institute of Mathematics of Luminy (IML) CNRS Marseille, Case 907, Campus de Luminy 13288 Marseille Cedex 9, France

Associate Editor: Olga Troyanskaya

Contact: baragatt@iml.univ-mrs.fr

Received on October 8, 2010; revised on February 3, 2011; accepted on February 4, 2011

In their paper in *Bioinformatics*, Yang and Song (2010a) proposed a Bayesian probit regression model for gene selection. Their model is an extension of Lee *et al.* (2000) specially adapted to overcome a problem of singular matrix which is frequently encountered under situations with large number of covariates and (relatively) small number of observations. Yang and Song (2010a) showed that their method permits simple simulations with flexible initial parametrization. Besides, their algorithm proved to be efficient through two well-known datasets. In the case of singularity of the covariance matrix of the *g*-prior (Zellner, 1986), the computation of the posterior distributions proposed by Yang and Song (2010a) has a technical issue. In this comment, we highlight a specific formula that yield intractable computation, in some cases, and we offer a solution.

Following notations of Yang and Song, the matrix of selected covariates is denoted by X_γ. The matrix $X'_\gamma X_\gamma$ coincides with the covariance matrix of the *g*-prior distribution for the coefficient parameter β_γ. There are two standard cases in which the matrix $X'_\gamma X_\gamma$ is singular, and where conventional approaches do not work:

(1) If the number of observations is lower than the number of selected variables, $n < p_\gamma$.

(2) If the matrix X_γ is not of full rank, which is the case if some variables are linear combinations of others.

In these cases, Yang and Song (2010a) proposed to replace the matrix $(X'_\gamma X_\gamma)^{-1}$ by its Moore–Penrose generalized inverse $(X'_\gamma X_\gamma)^{+}$. They obtained a modified form of the *g*-prior, namely the *gsg*-prior (see West, 2000). They proposed to use a collapsed Gibbs sampler (see for instance van Dyk and Park, 2008) by integrating β_γ out from the joint posterior distribution. Details are given in the supplementary material (Yang and Song, 2010b). In their calculus for the distribution of β_γ, the inverse of the following matrix

$$A = X'_\gamma\{(I_n + h\mathbf{1}\mathbf{1}')^{-1} + c^{-1}I_n\}X_\gamma,$$

is used in their Equation (A 7) given by

$$-\frac{\beta'_\gamma A\beta_\gamma - 2\beta'_\gamma B}{2} - \frac{Z'(I_n + h11')^{-1}Z}{2}$$
$$= -\frac{(\beta_\gamma - A^{-1}B)'A(\beta_\gamma - A^{-1}B)}{2} - \frac{Z'(I_n + h11')^{-1}Z - B'A^{-1}B}{2}. \quad \text{(A 7)}$$

However, A is not invertible if $X'_\gamma X_\gamma$ is singular. An idea should be to use A^+ instead of A^{-1} to adapt (A 7). But even using A^+, it is not clear to us that we can recover a Gaussian density for β_γ since we have

$$\beta'_\gamma A\beta_\gamma - 2\beta'_\gamma B \neq (\beta_\gamma - A^+B)'A(\beta_\gamma - A^+B) - B'A^+B.$$

Therefore it seems intractable to express the distribution of β_γ and to integrate out this parameter.

The method proposed by Yang and Song (2010a) can be clearly applied when $X'_\gamma X_\gamma$ is invertible and it has demonstrated good performances. Note that in the invertible case the *gsg*-prior coincides with the *g*-prior [as it is underlined by Yang and Song (2010a)], and the proposed model can be viewed as an extension of that of Lee *et al.* (2000). To avoid the case where $n < p_\gamma$, Yang and Song suggested the use of small prior values for π_i, restricting the number of genes. Another solution is to fix the number of selected covariates, as in Baragatti (2010). It appears computationally advantageous and it reduces the effect of the variable selection coefficient c used in the *g*-prior. Eventually, concerning the case where the X_γ matrix is not of full rank, it would be of interest to consider an alternative prior for β_γ, by combining the approach of Baragatti (2010) with the concept of ridge regression (Baragatti and Pommeret, 2011).

Conflict of Interest: none declared.

REFERENCES

Baragatti,M. (2010) Bayesian variable selection for probit mixed models applied to gene selection. Accepted in Bayesian Analysis, arXiv:1101.4577.

Baragatti,M. and Pommeret,D. (2011) Ridge parameter for *g*-prior distribution in Probit mixed model. arXiv:1102.0470

Lee,K.E. *et al.* (2000) Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **1**, 90–97.

van Dyk,D. and Park,T. (2008) Partially collapsed gibbs samplers: theory and methods. *J. Am. Stat. Assoc.*, **103**, 790–796.

West,M. (2000) Bayesian factor regression models in the large p small n paradigm. In Bernardo,J.M. *et al.* (eds) *Bayesian Statistics* 7. Oxford University Press, Oxford, pp. 733–742.

Yang,A.J. and Song,X.Y. (2010a) Bayesian variable selection for disease classification using gene expression data. *Bioinformatics*, **2**, 215–222.

Yang,A.J. and Song,X.Y. (2010b) Supplementary material to Bayesian. selection for disease classification using gene expression data. *Bioinformatics*, **26**, 215–222.

Zellner,A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Goel,P. and Zellner,A. (eds) *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*. Elsevier Science, North-Holland, Amsterdam, pp. 233–243.

*To whom correspondence should be addressed.