Test
DOI 10.1007/s11749-008-0106-x

DISCUSSION

Comments on: Augmenting the bootstrap to analyze high dimensional genomic data

Geoffrey J. McLachlan · K. Wang · S.K. Ng

© Sociedad de Estadística e Investigación Operativa 2008

1 Clustering of gene profiles

We congratulate the authors on their interesting article which addresses an important problem in the statistical analysis of high-dimensional data, namely how to estimate the inverse of the population covariance matrix. As the authors have explained, this estimation problem is very challenging with high-dimensional data, as the sample size is generally not large relative to the dimension of the data. Indeed, for example, with the analysis of microarray gene-expression data, the number of tissue samples is invariably very small relative to the number of genes under study. In the special case of diagonal covariance matrices, there is no problem, but this case corresponds to the use of the Euclidean metric, whereas the Mahalanobis distance corresponding to a general covariance matrix is usually more appropriate at least for clustering multivariate data.

In the subsequent discussion, we focus on the microarray example considered by the authors in Sect. 4.2, involving the expression levels on some $p = 321$ yeast genes measured across $n = 10$ time points. The authors used their proposed method to estimate the partial correlations between pairs of genes from their profiles sampled at ten time points in the so-called *stationary phase*. We decided to look at the detection of pairs of genes with a high degree of coexpression, and hence correlation, across the $n = 10$ time points by fitting a mixture of linear mixed models (LMMs) to the $p = 321$ gene profiles. We use the so-called EMMIX-WIRE (**EM**-based **MIX**ture

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-008-0098-6.

G.J. McLachlan (✉)
Department of Mathematics & Institute of Molecular Biosciences, University of Queensland
Brisbane, Queensland QLD 4019, Australia
e-mail: gjm@maths.uq.edu.au

analysis **W**ith **R**andom **E**ffects) developed by Ng et al. (2006) to handle the clustering of correlated data that may be replicated. They adopted conditionally a mixture of linear mixed models to specify the correlation structure between the variables and to allow for correlations among the observations.

To formulate this procedure, we consider the clustering of p gene profiles $\boldsymbol{y}_j$ ($j = 1, \ldots, p$), where we let $\boldsymbol{y}_j = (\boldsymbol{y}_{1j}^T, \ldots, \boldsymbol{y}_{mj}^T)^T$ contain the expression values for the jth gene profile, and $\boldsymbol{y}_{tj} = (y_{1tj}, \ldots, y_{r_t tj})^T$ ($t = 1, \ldots, m$) contains the r_t replicated values in the tth biological sample ($t = 1, \ldots, m$) on the jth gene. The dimension d of $\boldsymbol{y}_j$ is given by $n = \sum_{t=1}^{m} r_t$. With the EMMIX-WIRE procedure, the observed d-dimensional vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p$ are assumed to have come from a mixture of a finite number, say g, of components in some unknown proportions $\pi_1, \ldots, \pi_g$, which sum to one. Conditional on its membership of the ith component of the mixture, the profile vector $\boldsymbol{y}_j$ for the jth gene ($j = 1, \ldots, p$) follows the model

$$\boldsymbol{y}_j = \boldsymbol{X}\boldsymbol{\beta}_i + \boldsymbol{U}\boldsymbol{b}_{ij} + \boldsymbol{V}\boldsymbol{c}_i + \boldsymbol{\epsilon}_{ij}, \tag{1}$$

where the elements of $\boldsymbol{\beta}_i$ are fixed effects (unknown constants) modeling the conditional mean of $\boldsymbol{y}_j$ in the ith component ($i = 1, \ldots, g$). In (1), $\boldsymbol{b}_{ij}$ (a q_b-dimensional vector) and $\boldsymbol{c}_i$ (a q_c-dimensional vector) represent the unobservable gene- and tissue-specific random effects, respectively. These random effects represent the variation due to the heterogeneity of genes and samples (corresponding to $\boldsymbol{b}_i = (\boldsymbol{b}_{i1}^T, \ldots, \boldsymbol{b}_{ip}^T)^T$ and $\boldsymbol{c}_i$, respectively). The random effects $\boldsymbol{b}_i$ and $\boldsymbol{c}_i$, and the measurement error vector $(\boldsymbol{\epsilon}_{i1}^T, \ldots, \boldsymbol{\epsilon}_{ip}^T)^T$ are assumed to be mutually independent, where $\boldsymbol{X}$, $\boldsymbol{U}$, and $\boldsymbol{V}$ are known design matrices of the corresponding fixed or random effects, respectively. If the covariance matrix $\boldsymbol{H}_i$ is taken to be diagonal, then the expression levels on the jth gene in different biological samples are taken to be independent. The presence of the random effect $\boldsymbol{c}_i$ for the expression levels of genes in the ith component induces a correlation between the profiles of genes within the same cluster.

With the LMM, the distributions of $\boldsymbol{b}_{ij}$ and $\boldsymbol{c}_i$ are taken, respectively, to be multivariate normal $N_{q_b}(\boldsymbol{0}, \boldsymbol{H}_i)$ and $N_{q_c}(\boldsymbol{0}, \theta_{ci}\boldsymbol{I}_{q_c})$, where $\boldsymbol{H}_i$ is a $q_b \times q_b$ covariance matrix and $\boldsymbol{I}_{q_c}$ is the $q_c \times q_c$ identity matrix. The measurement error vector $\boldsymbol{\epsilon}_{ij}$ is also taken to be multivariate normal $N_n(\boldsymbol{0}, \boldsymbol{A}_i)$, where $\boldsymbol{A}_i = \mathrm{diag}(\boldsymbol{W}\boldsymbol{\xi}_i)$ is a diagonal matrix constructed from the vector $(\boldsymbol{W}\boldsymbol{\xi}_i)$ with $\boldsymbol{\xi}_i = (\sigma_{i1}^2, \ldots, \sigma_{iq_e}^2)^T$ and $\boldsymbol{W}$ a known $n \times q_e$ zero-one design matrix.

We applied this above procedure with $r_t = 1$, $\boldsymbol{X} = \boldsymbol{I}_{10}$, $\boldsymbol{U} = (1, 1, \ldots, 1)^T$, and $\boldsymbol{V} = \boldsymbol{I}_{10}$. It led to $g = 4$ clusters C_i ($i = 1, 2, 3, 4$) being obtained, containing 24, 139, 77, and 81 genes, respectively. The plots of the gene profiles are given in Fig. 1 for each of the four clusters. We can search for highly correlated genes by calculating the correlations between all pairs of genes within each cluster. For example, in Cluster C_1 which contains only $p_1 = 24$ genes, there is only one pair of genes with a correlation ρ greater than 0.9 (genes YMR019W and YOR303W with $\rho = 0.928$). Cluster C_2 with $p_2 = 139$ genes has 4 pairs with $\rho > 0.9$, the top pair being genes YAL022C and YJL139C with $\rho = 0.960$. The $p_3 = 77$ and $p_4 = 81$ gene profiles in Clusters C_3 and C_4 can be seen from Fig. 1 to be much more homogeneous over the $n = 10$ time points, and they have 38 and 151 pairs, respectively, with $\rho > 0.9$. The top pair of genes in C_3 are YEL010W and YMR326C with $\rho = 0.960$, and in C_4 are YAL022C and YJL139C with $\rho = 0.992$.

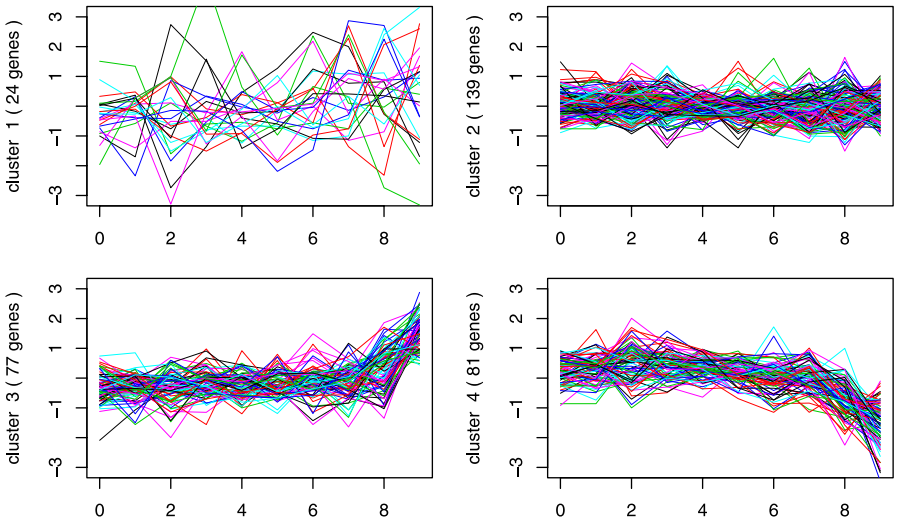

Fig. 1 Plots of gene profile clusters for 321 genes over 10 time points

2 Factor analysis model for dimension reduction

In the previous example, we could approach the problem of exploring for correlations between the gene profiles by proposing a LMM mixture model. If we were to attempt to work directly with the $n = 10$ tissue samples of $p = 321$ genes and try to estimate their 321×321 covariance matrix, then we obviously could not contemplate fitting a single normal or a normal mixture model to the $n = 10$ tissue samples. Even if n were much larger than 10 here, it would still not be possible to fit a normal model where the covariance matrix is unrestricted.

A global nonlinear approach to dimension reduction can be obtained by postulating a finite mixture of linear submodels for the distribution of the full observation vector $\boldsymbol{Y}_j$ given the (unobservable) factors. See Hinton et al. (1997) and McLachlan et al. (2007). Under this model, the ith component-covariance matrix $\boldsymbol{\Sigma}_i$ has the form $\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i$, where $\boldsymbol{B}_i$ is a $p \times q$ matrix of factor loadings and $\boldsymbol{D}_i$ is a diagonal matrix $(i = 1, \ldots, g)$. In cases where n is very small relative to p, McLachlan et al. (2002) have proposed a three-step procedure for the fitting of normal mixture models to cluster high-dimensional data. The first step considers the elimination of variables assessed to have little potential for clustering. On the second step, the retained variables (after appropriate scaling) are clustered into groups essentially using a soft-version of the k-means procedure. Then on the third step, the observations are clustered on the basis of representatives of the groups of variables (metavariables) using mixtures of factor analyzers if the number of metavariables is relatively high.

References

Hinton GE, Dayan P, Revow M (1997) Modeling the manifolds of images of handwritten digits. IEEE Trans Neural Netw 8:65–73

McLachlan GJ, Bean RW, Ben-Tovim Jones L (2007) Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. Comput Stat Data Anal 51:5327–5338

McLachlan GJ, Bean RW, Peel D (2002) A mixture model-based approach to the clustering of microarray expression data. Bioinformatics 18:413–422

Ng SK, McLachlan GJ, Wang K, Ben-Tovim Jones L, Ng SW (2006) A mixture model with random-effects components for clustering correlated gene-expression profiles. Bioinformatics 22(14):1745–1752