Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Mini Review
Volume 6 Issue 2 - April 2018
DOI: 10.19080/BBOAJ.2018.06.555682

Biostat Biometrics Open Acc J

Application of Penalized Mixed Model in Identification of Genes in Yeast Cell-Cycle Gene Expression Data

Mojtaba Ganjali[1,3]* and Taban Baghfalaki[2,3]

[1]*Department of Statistics, Shahid Beheshti University, Iran*

[2]*Department of Statistics, Tarbiat Modares University, Iran*

[3]*School of Biological Science, Institute for Research in Fundamental Sciences (IPM), Iran*

Submission: January 23, 2018; **Published:** April 18, 2018

***Corresponding author:** Mojtaba Ganjali, Department of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran, Email: m-ganjali@sbu.ac.ir

Abstract

High-dimensional time-course gene expression data refer to time course data with a large number of covariates. In this status, variable selection is a popular approach for selecting important variables. In this paper, we review penalized likelihood mixed effects model for variable selection in high-dimensional time-course data. Then, the approach is used for variable selection in yeast cell-cycle gene expression data

Keywords : Gene expression data; Mixed effects model; Penalty function; Penalized likelihood; Time-course data

Introduction

Linear mixed effects models have been used in a variety of study to analyze data with between-subject dependence [1]. For example, in analyzing longitudinal data, clustered data, repeated measurements and spatial statistics mixed effects models are often used. In this structure, the linear predictor contains a Gaussian zero-mean latent variable in addition to fixed effects. This latent variable is called random effects and this kind of models which contain fixed and random effects are called mixed effects models. These models are usually used for analyzing correlated outcomes in studies with small number of explanatory variables. But, the use of this model becomes a major problem in a high-dimensional dimensional setting or when the purpose of the study is variable selection. When the number of fixed and random variables increases, because of complexity of mixed effects model, the inference about the model become challenging. Therefore, the selection of fixed or random effects is a key problem in this status. There are many traditional approaches for variable selection. For example, AIC, conditional AIC, BIC, Bayesian variable selection and so on [2-5]. Most of these approaches are based on computing a chosen criterion and finding a subset of variables as the best subset based on the chosen criterion. Among these approaches conditional AIC [6] and Bayesian variable selection are commonly used for variable selection in mixed effects model. Another approach which proposed for variable selection in mixed effect models is the use of penalized likelihood approach. Although the use of penalized likelihood for the high-dimensional regression model (when $n \ll p$), which is famous to regularized regression method, is traditionally proposed [7,8]. But, nowadays the use of penalized likelihood for variable selection in mixed effects models is a popular approach [9-12]. In this paper, we review the ordinary penalized likelihood approach for variable selection in mixed effects model. Then, we use the approach for variable selection with lasso penalty for variable selection in yeast cell-cycle gene expression data set. This paper is organized as follows: in the next section penalized likelihood function for variable selection in mixed effects model is reviewed. The variable selection is given for yeast cell-cycle gene expression data in Section 3. The last Section includes some conclusions.

Penalized likelihood function for mixed effects model

In mixed effects model penalized likelihood function is usually used for both fixed and random effects. In the following, after introducing the notation used in this paper, at first penalized likelihood for fixed effects and then that for random effects are discussed.

Notation

Let there be N individuals in the study. Let y_{ij}, $i=1,\ldots,N,\ j=1,\ldots,n_i$ be the response variable for the i^{th} individual such that i^{th} the individual has n_i repeated measurements. Also, let x_{ij} and z_{ij} be $p\times 1$ and $q\times 1$ vector of covariates. Also, we define $n=\sum_{i=1}^{N} n_i$ to be the number of all observed responses in the study. In a matrix notation, we define $y_i=\left(y_{i1},y_{i2},\ldots,y_{in_i}\right)'$, , $X_i=\left(x_{i1},\ldots,x_{in_i}\right)'$ and $Z_i=\left(z_{i1},\ldots,z_{in_i}\right)'$. In linear mixed effects model, the model can be written as follows:

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i,$$

Where, β is a $p\times 1$ vector of regression coefficients, b_i is a $q\times 1$ vector of random effects with $b_i \sim N_q(0,D)$, $\varepsilon_i=\left(\varepsilon_{i1},\ldots,\varepsilon_{in_i}\right)'$ and $\varepsilon_{ij}\sim N\left(0,\sigma^2\right)$. Also, ε_i and $b_i\,(i=1,2,\ldots,N)$ are independent. In a matrix notation, let y,b,ε and X be matrices which obtained by stacking vectors of y_i,b_i,ε and X_i, respectively. Also, let $Z=diag\left(Z_1,\ldots,Z_N\right)$ and $\underline{m}\Delta=diag\left(D,\ldots,D\right)$ be a block-diagonal matrix. Then, the linear mixed effects model can be rewritten as

$$y = X\beta + Zb + \varepsilon. \qquad (1)$$

Where, $\varepsilon \sim N\left(0,\sigma^2 I\right)$ and $b\sim N\left(0,\Delta\right)$.

Selection of important fixed effects

The likelihood of marginal model (1) can be expressed as

$$L_n(\beta,D) \propto \exp\left\{\frac{-1}{2}(y-X\beta)'\,\Sigma\,(y-X\beta)\right\},$$

Where, dependence on sample size n is considered by adding the index n to $L(.,.)$, also, $\Sigma=\left(I+\sigma^{-2}Z\Delta Z'\right)^{-1}$. To select the important covariates, the use of the following penalized log-likelihood function is used:

$$\log\left(L_n(\beta,D)\right)-n\sum_{j=1}^{p_n}P_{\lambda_n}\left(\beta_j\right), \qquad (2)$$

Where, $P_{\lambda_n}(\cdot)$ is a penalty function with regularized parameter λ_n. In this notation, for showing dependence on sample size n the index n is used for λ. Maximizing (2) is equivalent to minimizing

$$\frac{1}{2}(y-X\beta)'\,\acute{O}(y-X\beta)+n\sum_{j=1}^{p_n}P_{\lambda_n}\left(\beta_j\right). \qquad (3)$$

As mentioned before, Σ depends on the unknown covariance matrix Δ and σ^2. Based on Theorem 1 of Fan & Li [12], the important fixed effects have oracle properties. The oracle property is that the asymptotic distribution of the estimator is the same as the asymptotic distribution of the MLE on only the true support. That is, the estimator adapts to knowing the true support without paying a price (in terms of the asymptotic distribution). In short, an oracle estimator must be consistent in parameter estimation and variable selection. Notice that an estimator that is consistent in variable selection is not necessarily consistent in parameter estimation [13].

Identifying important random effects

As mentioned by Fan & Li [12], the number of random effects q may be increased with sample size n so its dependency on n can be written by q_n. The estimation and therefore identifying of random effects are different from fixed effects. One of the most famous approaches in estimating random effects is the empirical Bayes approach [14]. But, this approach is not useful for selecting random effects. In the following, we review the proposed method of Fan & Li [12] for identifying important random effects. Consider $\acute{O}_x = I - X\left(X'X\right)^{-1}X'$ and a $n\times(n-p)$ matrix A such that $AA'=\acute{O}_x$. Let $w=A'y$, given the density function of w is given by

$$f_w\left(A'y\,|\,b\right)=\left(2\pi\sigma^2\right)^{\frac{-(n-p)}{2}}\exp\left\{\frac{-1}{2\sigma^2}(y-Zb)'\,\Sigma_x\,(y-Zb)\right\}. \qquad (4)$$

This conditional probability is independent of the fixed effects β and A As mentioned before $b\sim N(0,\Delta)$. Let Δ^+ be the Moore-Penrose generalized inverse of Δ Then, a group variable selection strategy is needed to identify true random effects. For this purpose, consider the following regularization problem:

$$\frac{1}{2}(y-Zb)'\,\acute{O}_x\,(y-Zb)+\frac{1}{2}\sigma^2 b'{\scriptstyle\Delta} b+n\sum_{k=1}^{q_n}P_{\lambda_n}\left(b_{0k}\right), \qquad (5)$$

Where $P_{\lambda_n}(.)$ is the penalty function with regularization parameter $\lambda_n>0$ and $b_{0k}=\left(\sum_{i=1}^{N}b_{ik}^2\right)^{1/2}$. Also, based on Theorem 2 of Fan & Li [12] the identified random effects is close to oracle estimator.

Tuning parameters selection

Different penalty function to achieve the purpose of variable selection in mixed effects models is proposed [10-17]. Some of the penalty function for variable selection in mixed effects models are lasso penalty function: $P(\gamma)=\lambda_1\sum_{j\in P}\gamma_j^2$, ridge penalty function: $P(\gamma)=\lambda_2\sum_{j\in P}\left|\gamma_j\right|$ and elastic net penalty function: $P(\gamma)=\lambda_1\sum_{j\in P}\gamma_j^2+\lambda_2\sum_{j\in P}\left|\gamma_j\right|$. In this paper, we use lasso penalty function and available lmmlasso in R for variable selection in time-course gene expression data [18]. One of the important stages in the used of the penalized likelihood function is the selection of the tuning parameter. The above-mentioned penalty function has some tuning parameter (λ_1 and λ_2). A popular approach for selecting tuning parameter is based on some criteria such as AIC and BIC. In this framework, the selected tuning parameter is that with minimal AIC or BIC.

Yeast cell-cycle gene expression data

In this section, we use yeast cell-cycle gene expression data which collected in yeast cell cycle analysis project by Spellman et al. [19]. The goal of the project is to identify all genes whose mRNA levels are regulated by the cell cycle. The experiment recorded genome-wide mRNA levels for 6178 yeast ORFs at 7-minute intervals for 119 minutes which covers two cell-cycle periods for a total of 18 time points. Transcription factors (TFs) have an important role in gene expression regulation. Transcription factors are proteins involved in the process of converting, or transcribing, DNA into RNA. Transcription factors include a wide number of proteins, excluding RNA polymerase, which initiate and regulate the transcription of genes. We have extracted the cell cycle gene expression data of 542 genes

How to cite this article: Mojtaba G, Taban B. Application of Penalized Mixed Model in Identification of Genes in Yeast Cell-Cycle Gene Expression Data. Biostat Biometrics Open Acc J. 2018; 6(2): 555682. DOI: 10.19080/BBOAJ.2018.06.555682.

from an α -factor based experiment. Each response variable corresponds to mRNA levels measured at every 7 minutes during 119 minutes (a total of 18 measurements). Also, we consider the ChIP-chip of the above-mentioned 106 TFs as explanatory variables. More information about this data set can be found in [20]. Figure 1 presents spaghetti plot for whole genes (panel penalized mixed model with 106 TFs as fixed effects with a random intercept and a random slop for time. Table 1 shows the 26 important covariates which selected using penalized likelihood mixed model by lasso penalty function. The tuning parameter is selected as $\lambda_1 = 3$. Also, none of the random effects

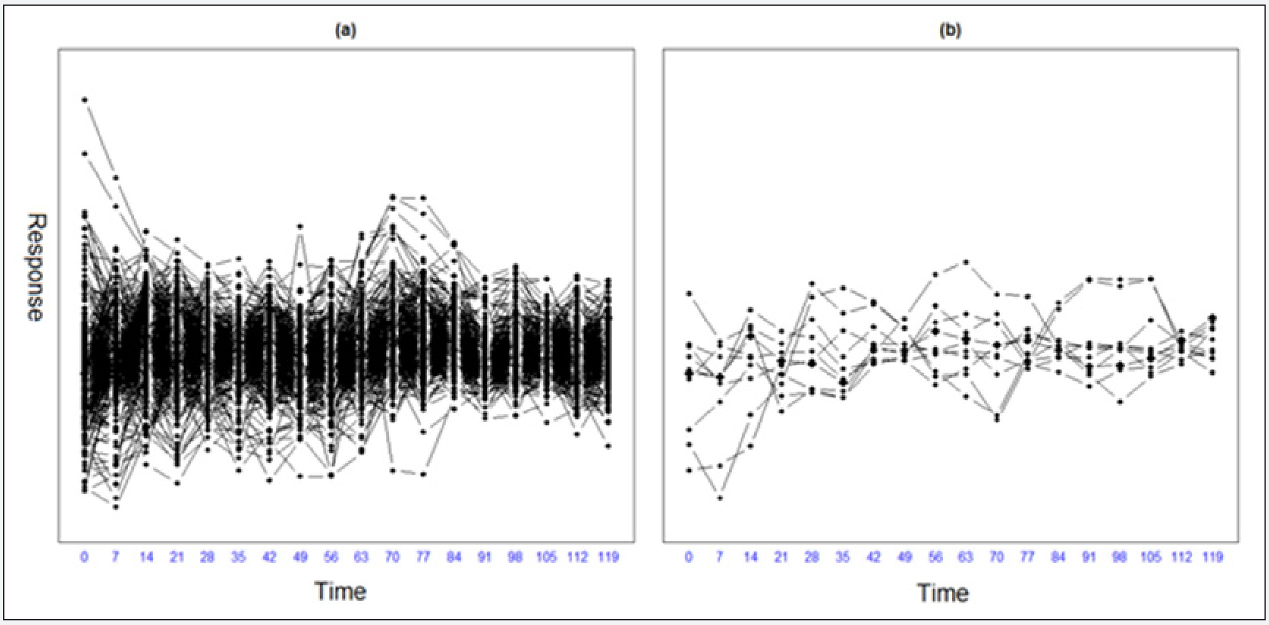

Figure 1 : spaghetti plot for response variable for whole genes (panel a) and 10 randomly selected genes (panel b).

Table 1: Selected TFs in yeast cell-cycle gene expression data using penalized mixed effects model with lasso penalty.

Name of TFs selected
ABF1, ARG80, ASH1, CHA4, CUP9, FHL1, GAT1, GAT3
GCR1, HIR2, HSF1, INO2, LEU3, MAC1, MAL13, MBP1
MCM1, PUT3, RLM1, SKN7, STE12, SUM1, SWI5, THI2
YAP5, ZMS1

Conclusion

In this paper, we review variable selection in mixed effects model using penalized likelihood approach. In this framework, we discussed how one can select fixed effects, random effects and tuning parameter. Also, in this paper, we consider lasso penalty function, also, we analyze a high-dimensional time course yeast gene expression data, where from 106 TFs, 26 of them were selected by the model to be important.

References

1. Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38(4): 963-974.
2. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. Selected Papers of Hirotugu Akaike 199-213.
3. Liang H, Wu H, Zou G (2008) A note on conditional AIC for linear mixed-effects models. Biometrika 95(3): 773-778.
4. Schwarz G (1978) Estimating the dimension of a model. The annals of statistics 6(2): 461-464.
5. Chen Z, Dunson DB (2003) Random effects selection in linear mixed models. Biometrics 59(4): 762-769.
6. Vaida F, Blanchard S (2005) Conditional Akaike information for mixed effects models. Biometrika 92: 351-370.
7. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 1: 267-288.
8. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(2): 301-320.
9. Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96(456): 1348-1360.
10. Bondell HD, Krishna A, Ghosh SK (2010) Joint Variable Selection for Fixed and Random Effects in Linear Mixed Effects Models. Biometrics 66(4): 1069-1077.

How to cite this article: Mojtaba G, Taban B. Application of Penalized Mixed Model in Identification of Genes in Yeast Cell-Cycle Gene Expression Data. Biostat Biometrics Open Acc J. 2018; 6(2): 555682. DOI: 10.19080/BBOAJ.2018.06.555682.

11. Ibrahim JG, Zhu H, Garcia RI, Guo R (2011) Fixed and Random Effects Selection in Mixed Effects Models. Biometrics 67: 495-503.
12. Fan Y, Li R (2012) Variable selection in linear mixed effects models. Annals of statistics 40(4): 2043.
13. Zou H (2006) The adaptive lasso and its oracle properties. Journal of the American statistical association 101(476): 1418-1429.
14. Verbeeke G, Molenberghs G (2000) Linear Mixed Models for Longitudinal Data. Springer, New York, USA.
15. Schelldorfer J, Bühlmann P, VAN S (2011) Estimation for High-Dimensional Linear Mixed-Effects Models Using l_1-Penalization. Scandinavian Journal of Statistics 38(2): 197-214.
16. Hui FK, Muller S, Welsh AH (2017) Joint selection in mixed models using regularized PQL. Journal of the American Statistical Association 112(519): 1323-1333.
17. Groll A (2011) glmmLasso: Variable selection for generalized linear mixed models by L1-penalized estimation. R package version 1(1).
18. Schelldorfer J (2011) lmmlasso: Linear mixed-effects models with Lasso. R package version 0.1-2.
19. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9(12): 3273-3297.
20. http://artax.karlin.mff.cuni.cz/r-help/library/spls/html/yeast.html

DOI: 10.19080/BBOAJ.2018.06.555682

How to cite this article: Mojtaba G, Taban B. Application of Penalized Mixed Model in Identification of Genes in Yeast Cell-Cycle Gene Expression Data. Biostat Biometrics Open Acc J. 2018; 6(2): 555682. DOI: 10.19080/BBOAJ.2018.06.555682.