Colloquium Biometricum 42
2012, 103–109

APPLICATION OF LINEAR MIXED MODELS IN THE SELECTION OF GENES FROM MICROARRAY EXPERIMENTS WITH REPEATED MEASUREMENTS

Alicja Szabelska, Joanna Zyprych-Walczak, Idzi Siatkowski

Department of Mathematical and Statistical Methods
Poznan University of Life Sciences
Wojska Polskiego 28, 60-637 Poznań, Poland
e-mail: aszab@up.poznan.pl, zjoanna@up.poznan.pl, idzi@up.poznan.pl

Summary

This paper focuses on the application of linear mixed models to microarray experiments. The main focus is on experimental design with biological as well as technical replicates. The results suggest that, depending on the considered number of top genes, different tests for linear fixed model or linear mixed model show better outcomes. In particular, cross validation revealed that the fixed model with parametric tests along with the mixed model with permutational tests based on residuals attained the lowest classification errors. On the other hand, ROC curve analysis implied that parametric tests for fixed as well as mixed model return the highest values for performance effectiveness.

Keywords and phrases: linear fixed model**,** linear mixed model, R software, selection, permutational F test, classification

Classification AMS 2000: 62-07, 62-09

1. Introduction

In the past ten years the technology of microarrays has become a widely used tool for the simultaneous investigation of thousands of genes. Improvements in the quality and precision of this technique make it necessary to apply accurate statistical analysis of the microarray data. Following Smyth (2004), standard procedures for obtaining differential genes require linear models. There are several tests available for linear fixed model, such as the *t* test, the Mann–Whitney test for two groups or F test, and the Kruskal–Wallis test for three or more groups. In this case, assumptions include independence among all observations and only one source of random variation.

A fixed model is widely used in many types of microarray experiments. However, it allows only one source of variation. In addition, it requires the assumption of independence of the observations. In particular, when research involves biological as well as technical replicates, statistical analysis would need to apply modifications of these methods. The most natural way to include several types of random variation is to use a linear mixed model. In this paper we present applications of several tests based on a linear mixed model in a microarray experiment and compare them with tests for linear fixed model. We would like to note that all the computations were performed using the R platform, version 2.12.1 (R Development Core Team, 2010).

2. Materials and Methods

The results (presented in the Results section below) are based on two datasets. The first one (the 'Mouse' dataset) was obtained from the microarray experiment described by Wu et al. (2011). It consists of 18 Affymetrix microarrays. There were three mouse strains, AJ, B6 and their F1 offspring considered in the experiment with three biological replicates each and two technical replicates for each individual. Each microarray contains the expression levels of the 500 genes that were investigated. The second dataset was produced by the Institute of Bioorganic Chemistry of the Polish Academy of Sciences and was kindly provided for the present analysis. It consists of 52 microarrays dedicated to acute myeloid leukemia (the 'Leukemia' dataset). The arrays include biological material of 13 patients with diagnosed acute myeloid leukemia and 13 healthy patients as a control. In total the dataset consists of thirteen biological replicates each with two technical replicates for each individual. Each microarray contains the expression levels of the 919 genes that were investigated.

One of the main aims of the microarray experiments is to find the set of genes that are differentially expressed with respect to several interesting features. If it is expected that the relationship between considered features and the expression level of genes is linear, the model can be written as:

$$y_{kij} = \beta_{ki} + e_{kij},$$

where y_{kij} is the expression level of gene *k* in the *i*-th group in the *j*-th observation, *k=1,…,G, i=1,…,m, j=1,…,* m_i , *G* is the number of genes, *m* is the number of groups, *m*$_i$ is the number of observations in the *i*-th group, β_{ki} is the mean expression level of gene *k* in the *i*-th group, and the errors e_{kij} are assumed to be independently distributed as $N(0,\sigma_{ki}^2)$ random variables of the model.

When it is assumed that there is an additional source of random variation such as technical micromatrix replications, the model can be written as:

$$y_{kij} = \beta_{ki} + b_{kj} + e_{kij},$$

where the notation is the same as above and b_{kj} is a random variable representing the deviation coming from the *j*-th replicate.

To verify the most differential genes, several tests were performed for each gene. The p-values from each test were obtained and were corrected using the FDR correction based on the procedure introduced by Benjamini and Hochberg (1995). The obtained adjusted p-values represent the level of differentiation of the particular gene. Next, genes were ranked with respect to corrected p-values, and 50, 100, 150 and 200 of the most differential genes were selected respectively. The chosen sets of genes were subjected to three prediction methods: the naive Bayesian method (NB), the k nearest neighbor method (KNN), and the support vector machine method (SVM). Cross validation (leave-one-out cross validation) was performed for the classifier obtained by the use of one of these methods. This procedure is repeated for every data point in the set. At each step of the calculations an error was determined that identifies whether the remaining data point was correctly classified. As a result the number of misclassified samples based on the chosen classifier was obtained. The errors of prediction were compared for every test mentioned above and for three prediction methods. Additionally, the area under the ROC curves (Receiver Operating Characteristic curves) was investigated for the model. The ROC curves were created using stacked regression according to Wolpert (1992). By this means the effectiveness of the methods was verified. All the calculations for cross validation were performed using the MLInterfaces package, and the analysis of ROC curves was based on the pROC package (Robin et al. 2011).

3. Results

In the first step of the analysis it was verified how many jointly differential genes each pair of tests contains. Analysis of normality in the groups revealed that 14% and 35% of genes do not fulfill the assumption of normality of the data in the case of the ‘Mouse’ and ‘Leukemia’ datasets, respectively. Hence for the differential analysis, parametric as well as nonparametric tests were applied. Considering the design of the experiment there were considered 5 groups of tests: the parametric F and t tests for linear fixed model (denoted as ‘fp’), the non-parametric Kruskal–Wallis test and Wilcoxon test for linear fixed model (denoted as ‘fn’), parametric tests F and t for the mixed model (denoted as ‘mp’), permutational tests based on residual sampling for the mixed model (denoted as ‘mnr’) and permutational tests based on sample sampling for the mixed model (denoted as ‘mns’). Differential analysis resulted in sets of the 50, 100, 150 and 200 most differential genes for each considered model and test. In total this gave 5x4=20 results for each dataset. These genes were used to determine cross validation of the methods.

Table 1. Number of misclassified samples, where fp: linear fixed model with parametric test, fn: linear fixed model with nonparametric test, mp: linear mixed model with parametric test, mnr: linear mixed model with nonparametric test 1 (permutation based on residuals), mns: linear mixed model with nonparametric test 2 (permutation based on samples).

LEUKEMIA DATASET					MOUSE DATASET				
Test	**Number of genes**				**Test**	**Number of genes**			
	50	**100**	**150**	**200**		**50**	**100**	**150**	**200**
fp	1.00	1.33	8.33	7.33	**fp**	3.,67	1.67	1.33	2.33
fn	1.67	1.67	5.67	7.67	**fn**	3.67	2.67	3.67	1.67
mp	2.00	2.00	5.33	8.00	**mp**	3.33	2.00	1.33	2.67
mnr	2.33	3.33	4.67	6.67	**mnr**	3.00	1.67	1.67	2.00
mns	2.67	3.00	5.00	7.00	**mns**	2.33	2.00	1.67	1.67

Table 1 presents a comparison of the average classification errors based on the three classifiers applied to each considered test and for each chosen set of differential genes for each considered dataset.

Furthermore, for each dataset and each considered number of informative genes, the performances of the five test statistics were ranked. The 95% confidence interval (CI) of the mean rank for each test was obtained using these ranks. This information is summarized in Figure 1.

From Figure 1 we can observe that the lowest results were obtained for the fixed model with parametric tests along with the mixed model with permutational test based on residuals. The highest averaged error was given by the fixed model with nonparametric tests.

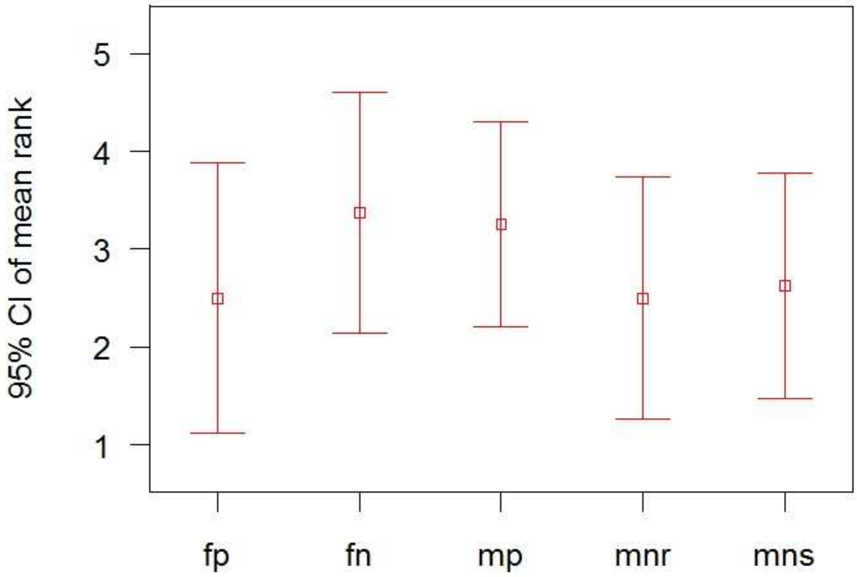

Fig. 1. The performance of five considered tests based on the average errors of ranks.

We can obtain additional information about the effectiveness of the methods based on the ROC curves. It is known that a more accurate predictor can be found by combining a set of single ones (Krzyśko et al. 2008). In this paper we use stacked regression to improve prediction accuracy. So far it is possible to visualize the ROC curves only for two classes. For this reason we present these results only for the Leukemia data (Figure 2).

Table 2. The values of AUC for each selection method for the ‘Leukemia’ and ‘Mouse’ datasets.

LEUKEMIA DATASET					**MOUSE DATASET**				
Test	**Number of genes**				**Test**	**Number of genes**			
	50	**100**	**150**	**200**		**50**	**100**	**150**	**200**
fp	1.,000	0.999	0.997	0.996	**fp**	0.910	0.886	0.877	0.864
fn	1.000	1.000	0.993	0.988	**fn**	0.861	0.873	0.927	0.914
mp	0.994	1.000	1.000	0.994	**mp**	0.910	0.901	0.941	0.890
mnr	0.999	1.000	1.,000	0.994	**mnr**	0.889	0.906	0.912	0.895
mns	0.999	0.997	0.997	0.994	**mns**	0.889	0.892	0.863	0.880

As we can see from Figure 2, for different numbers of chosen genes the considered tests give different results. To gain a better overview of the ROC analysis, the AUC values (Area Under the Curve) were also calculated. The outcomes for the ‘Leukemia’ and ‘Mouse’ data sets are presented in Table 2.

From the results we can observe that the AUC values for the ‘Leukemia’ dataset are higher than for ‘Mouse’. This can be explained by the different number of classes in the two datasets. The range of the AUC values is (0.988; 1) and (0.861; 0.941) for ‘Leukemia’ and ‘Mouse’ respectively. In the case of the ‘Leukemia’ dataset, the t test for fixed model, t test for mixed model as well as the permutational test based on residual sampling for mixed model resulted in the highest AUC values. However for the ‘Mouse’ dataset the best result was obtained by the mixed model with F test.

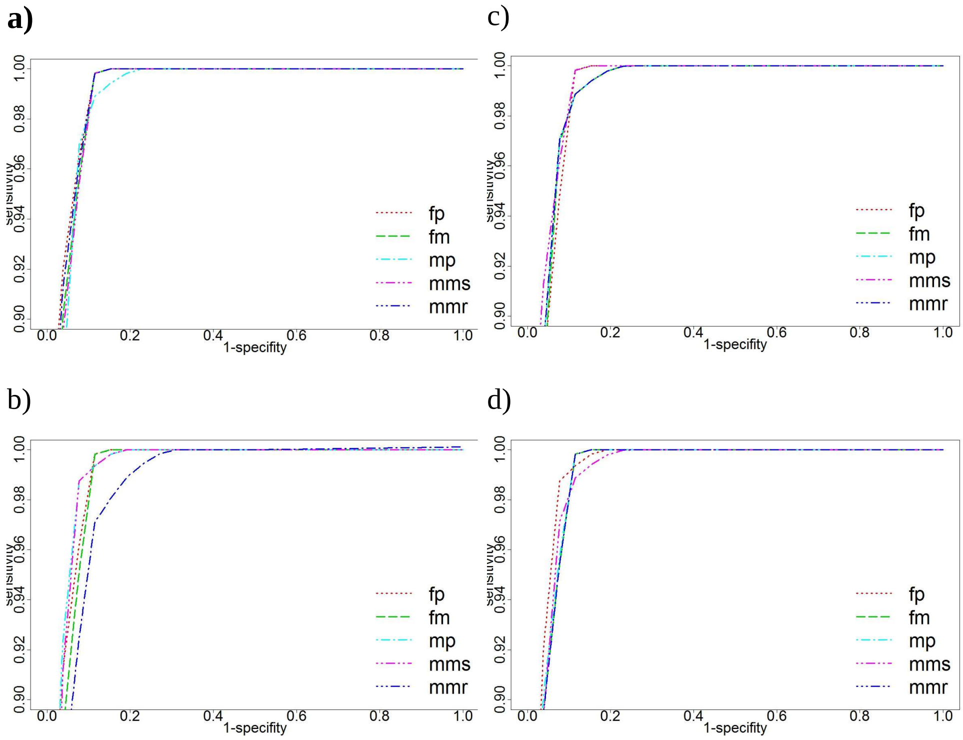

Fig. 2. The ROC curves for each method of selection and for the Leukemia dataset, where a)-d) are the ROC curves based on 50, 100, 150 and 200 differentially expressed genes, respectively**.**

4. Conclusions

The results of the analysis are unambiguous and suggest that, depending on the number of differential genes considered, different methods return the lowest values of misclassified genes. Even though the permutational tests resulted in similar sets of differentially expressed genes, the effectiveness of these methods is substantially different. In particular, cross validation revealed that the fixed model with parametric tests along with the mixed model with permutational tests based on residuals attained the lowest classification errors. Moreover the ROC curve analysis suggested that parametric tests for fixed as well as mixed model return the highest values for performance effectiveness. In our view, the results indicate that in the case of experimental designs with additional sources of variance (i.e. technical replicates) researchers should investigate methods based on fixed and mixed models. Depending on the comparison of these techniques, along with biological reasoning and the assumptions of the experiment, specific models should be chosen for particular study.

Acknowledgement

The authors would like to thank to Luiza Handschuh for the access to dataset that was used in the analysis. Thank the Secretary and an anonymous reviewer by constructive comments and suggestions on an early version of the manuscript.

References

Benjamini Y., Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Statistical Society Ser. B* 57, 289–300.

Krzyśko M., Wołyński W., Górecki T., Skorzybut M. (2008). *Systemy uczące się. Rozpoznawanie wzorców. Analiza skupień i redukcja wymiarowości.* WNT, Warszawa.

Robin X., Turck N., Hainard A., Tiberti N., Lisacek F., Sanchez J.Ch., Műller M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 7, 220. DOI: 10.1186/1471–2105–12–77.

R Development Core Team (2010). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r–project.org.

Smyth G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3, No. 1, Article 3. http://www.bepress.com/sagmb/vol3/iss1/art3.

Wolpert D.H. (1992). Stacked generalization. *Neural Networks* 5(5), 41–259.

Wu H., Yang H., Churchill G.A. (2011). R/MAANOVA: An extensive R environment for the Analysis of Microarray Experiments. http://cran.za.r-roject.org/web/packages/maanova/vignettes/maanova.pdf.