creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

이학석사 학위논문

Protein Significance Analysis of Multiple Reaction Monitoring (MRM) Data via Generalized Linear Mixed Effect Models

일반화 선형혼합모형을 통한 다중반응관측 자료의 단백질 연관성 분석

2016년 2월

서울대학교 대학원

통계학과

전 종 수

Protein Significance Analysis of Multiple Reaction Monitoring (MRM) Data via Generalized Linear Mixed Effect Models

지도 교수 박 태 성

이 논문을 이학석사 학위논문으로 제출함
2015년 10월

서울대학교 대학원
통계학과
전 종 수

전종수의 이학석사 학위논문을 인준함
2015년 12월

위 원 장 _____ 오 희 석 _____ (인)

부위원장 _____ 박 태 성 _____ (인)

위 원 _____ 원 중 호 _____ (인)

Abstract

Protein Significance Analysis of Multiple Reaction Monitoring (MRM) Data via Generalized Linear Mixed Effect Models

Jongsu Jun

Department of Statistics

The Graduate School

Seoul National University

Discovering protein biomarkers is one of the current important issues in biomedical research. The enzyme-linked immunosorbent assay (ELISA) is one of traditional protein quantitation techniques. As many novel proteins being studied, some issues of using ELISA for protein biomarker were emerged. The multiple reaction monitoring (MRM) mass spectrometry is a method for targeted protein quantification as well as an alternative of ELISA and has been widely utilized recently. However, development of statistical methods for this MRM data was not significant. In early analysis for MRM data, basic statistical methods such as two

sample or paired t-test and linear models were employed. In 2012, statistical methods for protein significance analysis using linear mixed model (LMM) called MSstats was proposed and it has been widely used since then. This LMM approach is implemented on Skyline program and R programming language. The resultant protein significant *p*-value form this LMM approach is diversified for same data set depending on the model setting which could provide many false positives and many true negatives. Furthermore, there is a loss of power of this previously proposed LMM approach if some features behave oppositely from the others. These characteristics motivated us to develop a model that is robust on the true effect type and the behaviour of features among the groups so that the model provides robust *p*-values for protein significance analysis. We proposed variance component test of generalized linear mixed model (GLMM) approach for protein significance analysis. Through various simulation studies, we observed that the proposed GLMM approach is more robust on the type of effect and more powerful when there is the interaction effect between features and groups. Moreover, there is no loss of power of proposed GLMM approach when there are oppositely behaving features while LMM approach performed poorly. In real data analysis, we observed cases that the previously proposed LMM approach hardly detects while the GLMM approach provided significant *p*-values. Consequently, not only previously proposed LMM approach but proposed GLMM approach should be considered for protein significance analysis using MRM data.

Keywords: Protein significance analysis, MRM, Linear mixed model, MSstats, Loss of power, Variance component test, GLMM,

***Student Number*: 2014-20297**

Contents

List of Figures

List of Tables

1. Introduction

1.1 Background

Discovering protein biomarkers is one of the current primary issues in biomedical research [1]. The enzyme-linked immunosorbent assay (ELISA) is one of traditional protein quantitation techniques that provide high sensitivity and throughput [2]. The ELISA approach is called the “gold standard” for targeted protein quantification [3]. However, recent studies discovered many novel proteins, yet the availability of highly qualified ELISAs for those protein markers is limited [4]. The development of a high quality ELISA assay consumes a plenty of time and resources [5]. These have created a need for different of targeted protein quantitation technique [6]. Quantitative mass spectrometry of proteins has advanced significantly over the last decade and several methods were developed for relative quantification of targeted proteins using this technique [7]. The multiple reaction monitoring (MRM) mass spectrometry is a method used in tandem mass spectrometry for systematic development of targeted protein assays which can be an alternative of ELISA [8]. The MRM assay targets sequence-specific tandem mass spectrometer fragmentations of peptides that can provide highly selective measurements for specific proteins [9]. Without enrichment or fractionation approaches, MRM assay covers almost complete dynamic range of abundance of cellular proteome [10]. Development time of MRM assay is relatively shorter than that of ELISA [11] and there is no cost of antibody development for MRM assays which can be another advantage of MRM assays compare to ELISA.

1.2 Purpose

MRM assays are gradually used in systems biology and clinical investigations at present [12-15]. The development of statistical methods to analyze significant proteins with MRM assay has not received enough attention compared to the improvement of the assay itself. In early use of the MRM assays, two sample t-test or paired t-test was applied to identify proteins that change in abundance between two groups [16-18]. To test with multiple groups, one-way ANOVA was employed [19, 20]. Advanced statistical model using linear mixed model (LMM) for protein significance analysis was proposed in 2012 [21] and it has been widely used since then [22]. The proposed LMM treats either or both subject and run effect as random or fixed. However, we observed that the proposed LMM approach provides diversified *p*-values for same data depending on which effect was treated as random. Moreover, existence of oppositely behaving features also affected the *p*-value. If intensity patterns of features of a protein were different from each other, there is a loss of power likewise the Cochran–Mantel–Haenszel statistics [23]. These characteristics motivated us to develop a model that is robust on the true effect type and the intensity patterns of features so that the model provides robust *p*-values for protein significance analysis under various effect type and various data structure. We applied variance component test of generalized linear mixed model (GLMM) [24] to detect proteins that change in abundance between groups.

2. Materials and Methods

2.1 Experimental Design of the Multiple Reaction Monitoring Mass Spectrometry

For a single MS run, all targeted features from an individual are examined. In terms of normalization, synthetic features that are identical to each targeted features but labeled, are included in every MS run. The MRM data structure is depicted in Table 1. Although these synthetic features give us relative intensity values, we need to handle carefully when other factors, such as, groups and subjects in Table 1, are included in a model.

Table 1. The table of the experimental design for MRM data.

The first subscript of $\boldsymbol{y}$ stands for the group, the second subscript stands for the subject that nested in a group, the third subscript stands for the feature and the last subscript stands for the run.

		Group 1			Group 2		
		Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6
		Run 1	Run 2	Run 3	Run 4	Run 5	Run 6
$\log_2(intensity)$ of endogenous feature	Feature 1	$y_{1,1(1),1,1}$	$y_{1,2(1),1,2}$	$y_{1,3(1),1,3}$	$y_{2,4(2),1,4}$	$y_{2,5(2),1,5}$	$y_{2,6(2),1,6}$
	Feature 2	$y_{1,1(1),2,1}$	$y_{1,2(1),2,2}$	$y_{1,3(1),2,3}$	$y_{2,4(2),2,4}$	$y_{2,5(2),2,5}$	$y_{2,6(2),2,6}$
	Feature 3	$y_{1,1(1),3,1}$	$y_{1,2(1),3,2}$	$y_{1,3(1),3,3}$	$y_{2,4(2),3,4}$	$y_{2,5(2),3,5}$	$y_{2,6(2),3,6}$
		Group 0					
		Subject 0					
		Run 1	Run 2	Run 3	Run 4	Run 5	Run 6
$\log_2(intensity)$ of synthetic feature	Feature 1	$y_{0,0(0),1,1}$	$y_{0,0(0),1,2}$	$y_{0,0(0),1,3}$	$y_{0,0(0),1,4}$	$y_{0,0(0),1,5}$	$y_{0,0(0),1,6}$
	Feature 2	$y_{0,0(0),2,1}$	$y_{0,0(0),2,2}$	$y_{0,0(0),2,3}$	$y_{0,0(0),2,4}$	$y_{0,0(0),2,5}$	$y_{0,0(0),2,6}$
	Feature 3	$y_{0,0(0),3,1}$	$y_{0,0(0),3,2}$	$y_{0,0(0),3,3}$	$y_{0,0(0),3,4}$	$y_{0,0(0),3,5}$	$y_{0,0(0),3,6}$

2.2 Linear Mixed Model (LMM) Approach

Proposed LMM approach at [21] is as follows

$$y_{i,j(i),k,l} = \mu + G_i + S(G)_{j(i)} + F_k + R_l + (G \times F)_{i,k} + (F \times R)_{k,l} + \epsilon_{i,j(i),k,l} \quad (1)$$

with restrictions $\sum_{i=0}^{2} G_i = 0$, $\sum_{j=0}^{J} S(G)_{j(i)} = 0$, $\sum_{k=1}^{K} F_k = 0$, $\sum_{l=1}^{L} R_l = 0$, $\sum_{i=0}^{2}(G \times F)_{i,k} = 0$, $\sum_{k=1}^{K}(G \times F)_{i,k} = 0$, $\sum_{k=1}^{K}(F \times R)_{k,l} = 0$ and $\sum_{l=1}^{L}(F \times R)_{k,l} = 0$ and $\epsilon_{i,j(i),k,l} \sim N(0, \sigma_\epsilon^2)$, where $y_{i,j(i),k,l}$ denotes $\log_2(intensity)$ value of the *j*-th subject nested in the *i*-th group of the *k*-th feature and the *l*-th run, μ is a global mean, G_i stands for the *i*-th group effect, $S(G)_{j(i)}$ for the *j*-th subject effect nested in *i*-th group, F_k for the *k*-th feature effect, R_l for the *l*-th run effect, $(G \times F)_{ik}$ for interaction effect of the *i*-th group and the *k*-th feature and $(F \times R)_{kl}$ for interaction effect of the *k*-th feature and the *l*-th run. In the above model, it was suggested to treat subject effects and run effects as random so that the restrictions of $S(G)_{j(i)}$, R_l and $(F \times R)_{kl}$ are replaced to $S(G)_{j(i)} \sim N(0, \sigma_S^2)$, $R_l \sim N(0, \sigma_R^2)$ and $(F \times R)_{kl} \sim N(0, \sigma_{F\times R}^2)$, respectively. The equivalent linear regression model formula is

$$y_i = \beta_0 + g_i'\beta_G + s_i'\beta_S + f_i'\beta_F + r_i'\beta_R + (g \times f)_i'\beta_{G\times F} + (r \times f)_i'\beta_{R\times F} + \epsilon_i$$

In the above equation, y_i is a $\log_2(intensity)$ value of i^{th} sample, g_i is a $(G \times 1)$ group dummy variable, where G stands for the number of groups except the reference group, S_i is a $(N \times 1)$ subject dummy variable, where N stands for

the number of subjects except the reference sample, F_i is a $(K-1\times 1)$ feature dummy variable, where K stands for the number of features, R_i is a $(R-1\times 1)$ run dummy variable, where R stands for the number of MS runs, $(G\times F)_i$ is a interaction of group and feature dummy variable, $(R\times F)_i$ is a interaction of run and feature dummy variable and ϵ_i is error that follows normal distribution mean 0 and variance σ_ϵ^2. β_S, β_R and $\beta_{R\times F}$ are coefficients of subject, run and interaction of run and feature, respectively. Here, that each effect can be treated either as fixed effect or as random effect. Testing null hypothesis

$$H_0: K\left(\beta_{G(1)}-\beta_{G(2)}\right)+\left(\beta_{G(1)\times F(2)}+\cdots+\beta_{G(1)\times F(K)}\right) -\left(\beta_{G(2)\times F(2)}+\cdots+\beta_{G(2)\times F(K)}\right)=0 \quad (2)$$

could provide a protein significance *p*-value, where $\beta_{G(a)}$ is the coefficient of the group a and $\beta_{G(a)\times F(b)}$ is the interaction coefficient of group a and feature b, is the objective of the model. Since $\beta_{G(a)}$ refers to $\mu_{G(a)\times F(1)}$, the mean of $\log_2(intensity)$ of group a and feature 1, and $\beta_{G(a)\times F(b)}$ refers to $\mu_{G(a)\times F(b)}-\mu_{G(a)\times F(1)}$, the difference of the mean of $\log_2(intensity)$ of group a and feature b and the mean of $\log_2(intensity)$ of group a and feature 1, the null hypothesis is equivalent to

$$H_0: \sum_{k=1}^{K}\mu_{G(1)\times F(k)}=\sum_{k=1}^{K}\mu_{G(2)\times F(k)}$$

2.3 Generalized Linear Mixed Model (GLMM) Approach

Consider a GLM as follows

$$logit\big(P(Z_l = 1)\big) = \alpha + \beta_1 y_{l1} + \cdots + \beta_K y_{lK}$$

In the above equation, Z_l is a group indicator of the *l*-th individual, $l = 1, \ldots, n$, α is an intercept, y_{lk} and β_k is the $\log_2(intensity)$ value of the k^{th} feature and the corresponding coefficient, respectively. The null hypothesis,

$$H_0: \beta_1 = \cdots = \beta_K = 0,$$

could provide a *p*-value for protein significant test. If we assume that β_k follows normal distribution with mean 0 and variance $w_j\tau$, where w_j is a prior weight, then the GLM is expanded to a GLMM and testing the hypothesis, $H_0: \beta_1 = \cdots = \beta_K = 0$, is equivalent to testing

$$H_0: \tau = 0.$$

Since the hypothesis $H_0: \tau = 0$ is on the boundary of the parameter space, the variance-component score test can be considered. The variance-component score test statistic for $\tau = 0$ is

$$Q = (\boldsymbol{Z} - \widehat{\boldsymbol{\mu}}_0)' \boldsymbol{K} (\boldsymbol{Z} - \widehat{\boldsymbol{\mu}}_0)$$

where $\widehat{\boldsymbol{\mu}}_0$ is estimated probability under H_0 and $\boldsymbol{K} = \boldsymbol{YWY}'$ with $\boldsymbol{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1K} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nK} \end{bmatrix}$ and $\boldsymbol{W} = \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_K \end{bmatrix}$. It is known that, under the normality assumption, Q follows mixture of chi-square distribution $\sum_{k=1}^{K} \lambda_k \chi_{1,k}{}^2$, where $\chi_{1,k}{}^2$'s are independent chi-square distributions with 1 degree of freedom, λ_k is the *k*-th eigenvalue of $\boldsymbol{P^{1/2} K P^{1/2}}$, $\boldsymbol{P} = \widehat{\boldsymbol{V}}^{-1} - \widehat{\boldsymbol{V}}^{-1}\boldsymbol{1}\big(\boldsymbol{1}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{1}\big)\boldsymbol{1}'\widehat{\boldsymbol{V}}^{-1}$ and $\widehat{\boldsymbol{V}} = \begin{bmatrix} \hat{\mu}_{01}(1-\hat{\mu}_{01}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\mu}_{0n}(1-\hat{\mu}_{0n}) \end{bmatrix}$

3. Simulations

3.1 Simulation Settings

We performed simulation studies to investigate whether or not LMM approach and GLMM approach preserved type I error rates and to investigate their power when various effect type of interaction of groups and features were presented. Furthermore, we compared the performance that identifies significant protein set when there are non-significant proteins mixed in a data set. We considered two types of effects of model (1) for type I error and power simulations as random effects and fixed effects. When we generated an effect as random, we controlled a variance of the effect such that the effect follows independent and identically distributed normal distribution with mean 0 and the specific variance. On the other hand, for fixed effect, we generated equally spaced sequence such that the average of the sequence is 0 and its squared average is same with the value of the variance that we specified to generate random effect. For a fixed interaction effect between features and groups, we consider two types of effect. (i) First is where the difference of a feature effect between two groups are same, $(G \times F)_{2,1} - (G \times F)_{1,1} = \cdots = (G \times F)_{2,K} - (G \times F)_{1,K}$. (ii) The second is where the difference of a feature effect between two groups are different, $(G \times F)_{2,1} - (G \times F)_{1,1} > \cdots > (G \times F)_{2,K} - (G \times F)_{1,K}$. In case of the first fixed interaction type, we also considered the different direction of the interaction effects when the number of features was 4, $(G \times F)_{2,1} - (G \times F)_{1,1} = (G \times F)_{2,2} - (G \times F)_{1,2} = (G \times F)_{1,3} - (G \times F)_{2,3} = (G \times F)_{1,4} - (G \times F)_{2,4}$. For the type I error estimation and for the empirical power estimation, the significance level, α, was set as 0.05. For the prediction performance

comparison, 0.05/10 significance level was used under the Bonferroni correction. All simulation data sets were generated from (1).

3.1.1 Settings for Type I Error Estimation

In (1), global mean, μ, was set as 15 and σ_ϵ^2, the variance of $\epsilon_{i,j(i),k,l}$, was set as 0.5 through this simulations. $S(G)_{j(i)}$, F_k, R_l and $(F \times R)_{kl}$ are nuisance factors to test hypothesis (2). Therefore we did not considered a simulation setting to observe their effects. These effect sizes were not varied through simulations and their variances or squared average were set as 0.25, 0.1, 0.25 and 0.1 for $S(G)_{j(i)}$, F_k, R_l and $(F \times R)_{kl}$, respectively. G_i and $(G \times F)_{ik}$ were set to 0 for all $i = 0, 1, 2$ and $k = 1, \dots, K$. Various number of features, 2 to 5, were considered to observe whether the number of features affects type I error. Various sample sizes, 20, 50 and 100, were also considered with 1:1 ratio of case and control.

3.1.2 Settings for Empirical Power Comparison

A simulation structure of global mean, variance of error, nuisance effects and sample sizes and their ratio of case and control were identical to those of structure that explained in section 3.1.1. We considered 5 scenarios for group effect such that $(G_0, G_1, G_2) \in \{(0,0,0), (0,0,0.25), (0,0,0.5), (0,0,0.75), (0,0,1)\}$ and considered 5 scenarios for interaction effect such that the variance or squared average of $(G \times F)_{ik} \in \{0, 0.05, 0.1, 0.15, 0.2\}$. Total 24 scenarios, all 0 case was neglected, for G_i and $(G \times F)_{ik}$ were set to acquire empirical power of LMM and

GLMM approach. Since the interaction effect between groups and features was considered, we examine the various scenarios of the interaction effect. In case of the number of features was 5, 1 to 5 number of causal features that different between groups were contemplated. In case of the number of features was 4, 2 number of antithetical features that present opposite direction of effect between groups compares to the other features were also contemplated, only in fixed effects (i).

3.1.3 Settings for Prediction Performance Comparison

Moreover we compare the performance of LMM approach, proposed GLMM approach and group LASSO approach to identify protein set that change in abundance. 10 protein sets were generated independently. To reduce computational burden, we narrowed down simulation scenarios form those of explained in section 3.1.2. For each protein, the structure of global mean, variance of error, nuisance effects and sample sizes and their ratio of case and control were identical to those of structure that explained in section 3.1.1. Here, $K = 3$ scenarios were covered. Effects, except global mean and group effect, were randomly generated from a normal distribution with their specified variance. 2 and 5 significant proteins were considered among total 10 proteins and their group effects were set to $(G_0, G_1, G_2) = (0, 0, 1)$. Among 2 and 5 significant proteins, 1 and 2 proteins, respectively, were considered to have the interaction effect of groups and features with their variance as 0.2 with all features has same interaction effect, (i), between groups.

3.2 Results of Simulation

3.2.1 Results for Type 1 Error Estimation

Figure 1 shows the empirical type I error for LMM approaches and the GLMM approach for randomly generated effect. LMM approaches that treat subject as random effect, LMM(RF) and LMM(RR), and the GLMM approach were well controlled type I error while LMM approaches that treat subject as fixed effect, LMM(FF) and LMM(FR), were not. Since the AIC value of LMM(FF) have tendency to have smallest among 4 LMMs, the best LMM, LMM(best), was dominated by LMM(FF) and its behavior was very similar with LMM(FF). Interestingly, as the number of features increases, type I error of LMM(FF) and LMM(FR) tend to increase and consequently that of LMM(best) increases. When effects were set to be fixed, for both type of features and groups interaction, the pattern of the empirical type I error of LMM approaches and the GLMM approach shows antithetic results to those of results when effects were randomly generated. Figure 2 and Figure 3 shows the empirical type I error for LMM approaches and a GLMM approach for fixedly generated effect with same interaction size and different interaction size, respectively. LMM approach that treat both subject and run effect as fixed effect, LMM(FF), were controlled type I error well. LMM approaches that treat subject as random effect, LMM(RF) and LMM(RR), controlled type I error rate conservatively. The GLMM approach also controlled type I error rate more conservatively than those of LMM approaches that treat subject effect as fixed and less conservatively than those of LMM approaches that treat subject effect as random. Since the AIC value of LMM(FF) was tend to be smallest among 4 LMMs the best LMM, LMM(best), was dominated by LMM(FF) and its behavior was very similar

with LMM(FF), much alike the simulation settings that effects were randomly generated. Interestingly, as the number of features increases, type I error of LMM(FF) and LMM(FR) tend to increase and consequently that of LMM(best) increases.

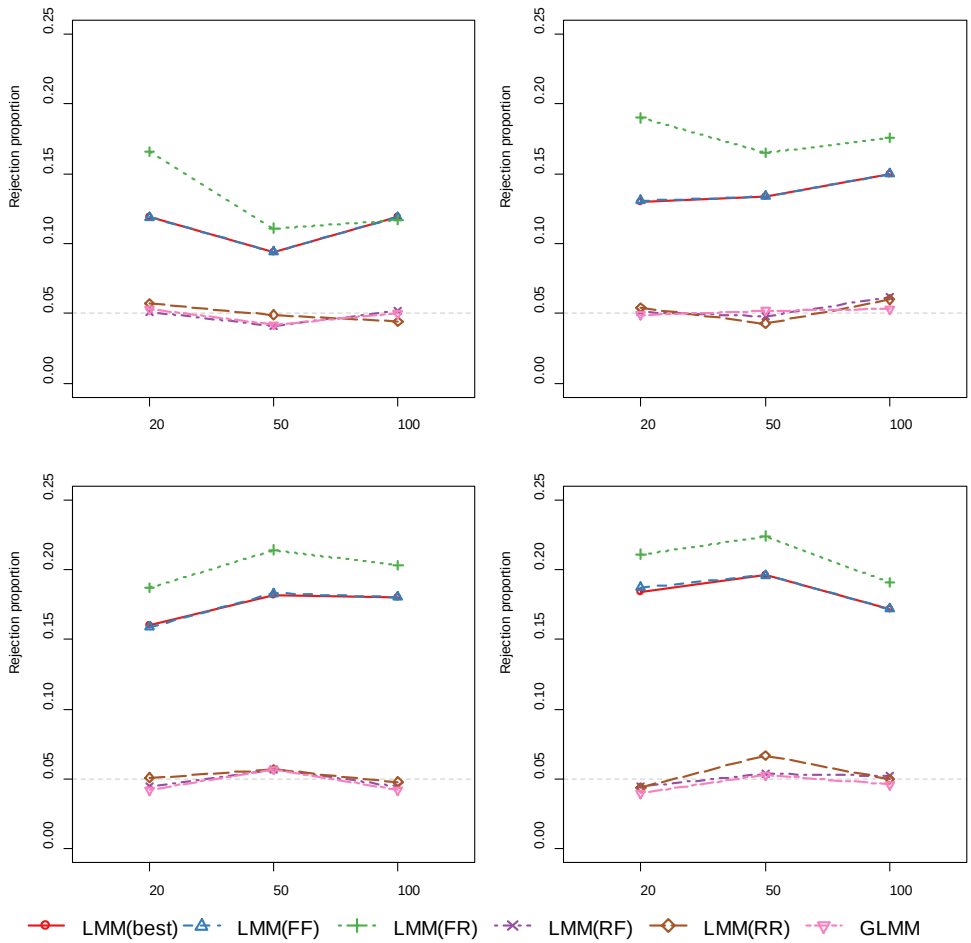

Figure 1. The empirical type I error of LMM approaches and the GLMM approach in different sample size for randomly generated effects.

The number of features was 2, 3, 4 and 5 for the top left panel, the top right panel, the bottom left panel and the bottom right panel, respectively. The x-axis and y-axis of each panels represent the number of samples and empirical type I error, respectively. Gray dotted horizontal line of each panels represents the significant level. LMM(FF), LMM(FR), LMM(RF) and LMM(RR) denoted in the legend stand for the view of subject and run effects in model (1) as random, R, or fixed, F, in respective order. The best LMM approach, LMM(best), was selected by AIC among 4 LMM approaches for each 1000 simulation data sets.

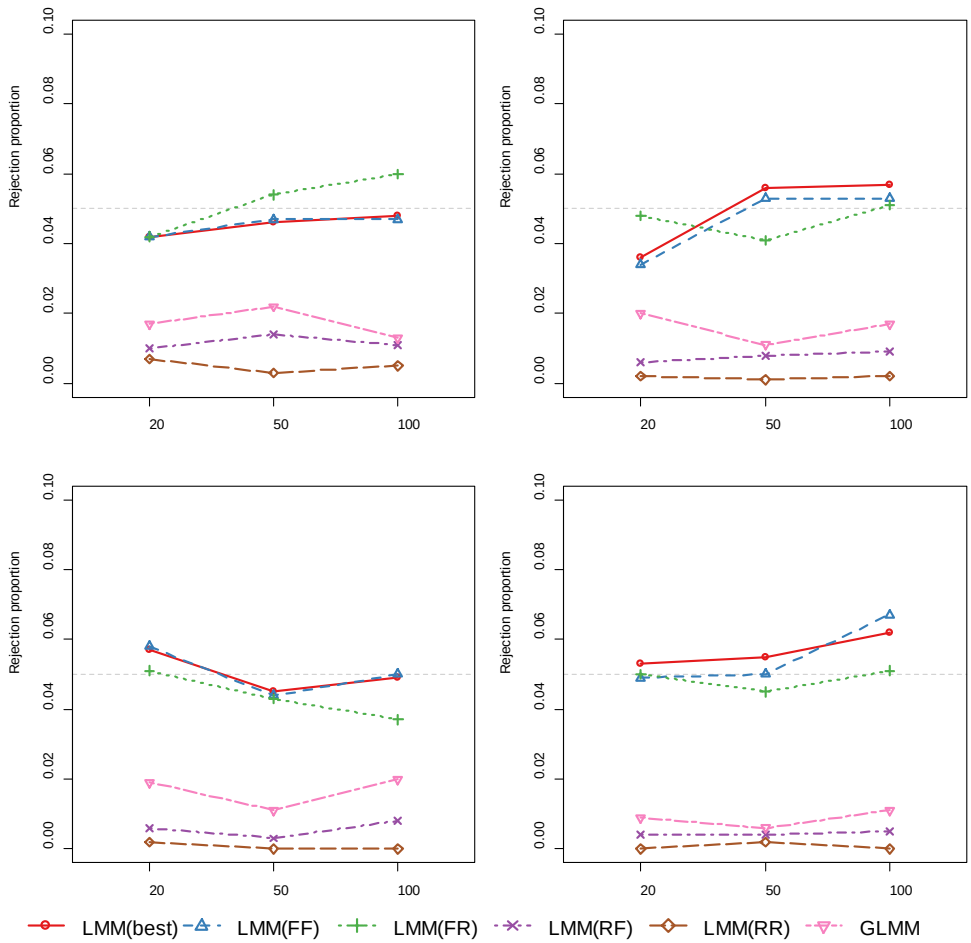

Figure 2. The empirical type I error of LMM approaches and the GLMM approach in different sample size for fixedly generated effects with same size of interaction effect.

The number of features was 2, 3, 4 and 5 for the top left panel, the top right panel, the bottom left panel and the bottom right panel, respectively. The x-axis and y-axis of each panels represent the number of samples and empirical type I error, respectively. Gray dotted horizontal line of each panels represents the significant level. LMM(FF), LMM(FR), LMM(RF) and LMM(RR) denoted in the legend stand for the view of subject and run effects in model (1) s random, R, or fixed, F, in respective order. The best LMM approach, LMM(best), was selected by AIC among 4 LMM approaches for each 1000 simulation data sets.

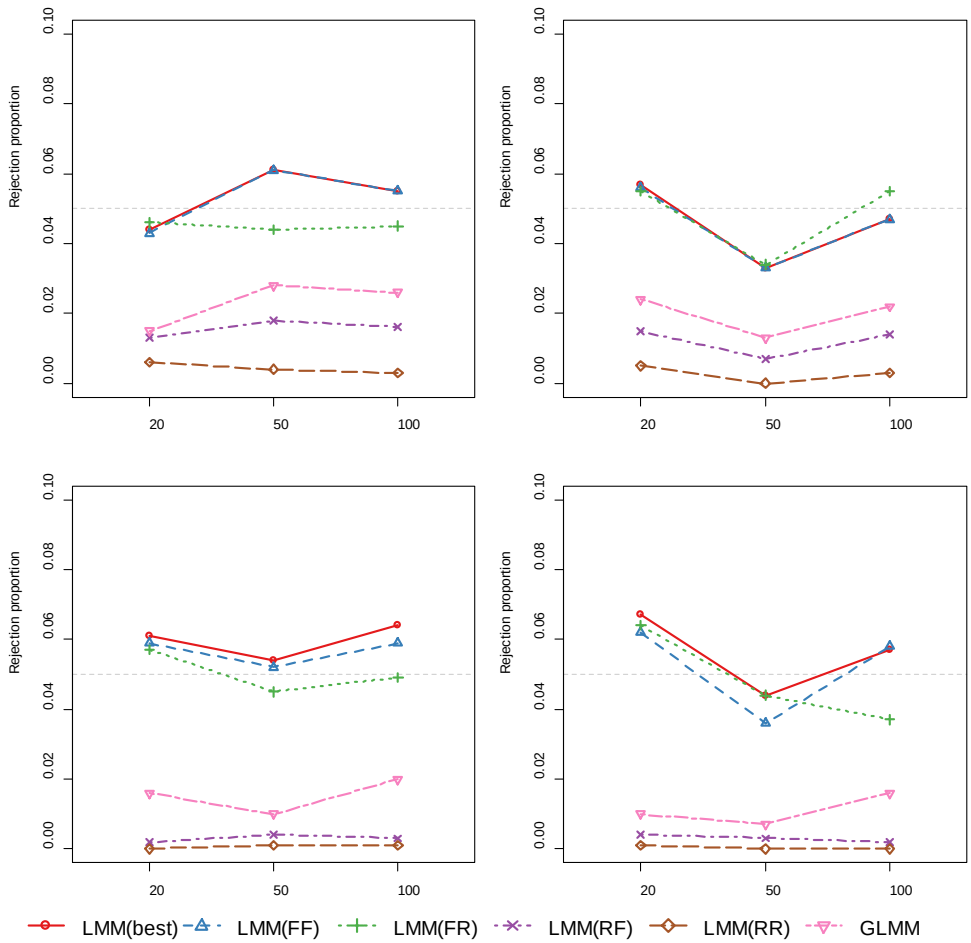

Figure 3. The empirical type I error of LMM approaches and a GLMM approach in different sample size for fixedly generated effects with different size of interaction effect.

The number of features was 2, 3, 4 and 5 for the top left panel, the top right panel, the bottom left panel and the bottom right panel, respectively. The x-axis and y-axis of each panels represent the number of samples and empirical type I error, respectively. Gray dotted horizontal line of each panels represents the significant level. LMM(FF), LMM(FR), LMM(RF) and LMM(RR) denoted in the legend stand for the view of subject and run effects in model (1) as random, R, or fixed, F, in respective order. The best LMM approach, LMM(best), was selected by AIC among 4 LMM approaches for each 1000 simulation data sets.

3.2.2 Results for Empirical Power Comparison

Figure 4 shows empirical power for LMM approaches and the GLMM approach when effects were randomly generated. Since both group effect and interaction effect is 0 for the top left panel, it represents type I error. As interaction effect gets greater, the power of GLMM approach rapidly increases compared to those of LMM approaches. On the other hand, as group effect gets greater, the power of LMM approaches rapidly increases compared to that of the GLMM approach. Once the interaction effect between features and groups presented, the power of the GLMM approach increase faster than those of LMM approaches as the sample size increased. Although the LMM(FF) and LMM(FR) were not well controlled type I error, the power of the GLMM approach exceeds those of LMM approaches when there was enough size of interaction effect presented. The pattern of the best LMM approach behaved similar with scenario for the empirical type I error estimation. Figure 5 shows empirical power for LMM approaches and the GLMM approach when effects were fixedly generated effect with same interaction size. The top left panel illustrates type I error as previously explained. As group effect gets greater, the power of LMM approaches and the GLMM approach increased. As interaction effect between features and groups gets greater, the power of LMM approaches and the GLMM approach increased faster than those of cases when effects were randomly generated. As sample size gets larger, increasing power of all competed models were bigger than those of cases when effects were randomly generated. Although the LMM(FR) shows the highest power among all competed models for all cases, its AIC was not the smallest one among four LMM approaches while the LMM(FF) present smallest AIC like previous simulation studies. The GLMM approach tend to have the lowest power among all competed models. Figure 6 shows the power for

five competed approaches for the different size of the interaction effects cases. The overall patterns of competed approaches were similar with the cases for fixedly generated with same interaction size. However, it is noticeable that once the interaction effect was presented, the power of the GLMM approach exceeds those of LMM approaches that treat subject effect as random. The number of causal features between groups affected the power of all competed models, obviously. However the amount of influence of the number of causal features was different. As the number of causal features decrease, power decrement of the GLMM approach was smaller than those of LMM approaches that treat subject effect as random. Although the power of any LMM approaches was higher than that of GLMM approach when all 5 features were causal, once the number of causal features decline 4, the power of the GLMM approach exceeded those of LMM approaches that treat subject effect as random as depicted in the left panel of Figure 7. Additionally, as depicted in the right panel of Figure 7, when 2 features were behave oppositely between groups among 4 features with same interaction size, the power of all LMM approaches was not increased as sample size increases. This oppositely behaving features counterbalanced the statistic for testing hypothesis (2) which causes LMM approaches could not detect the difference between groups.

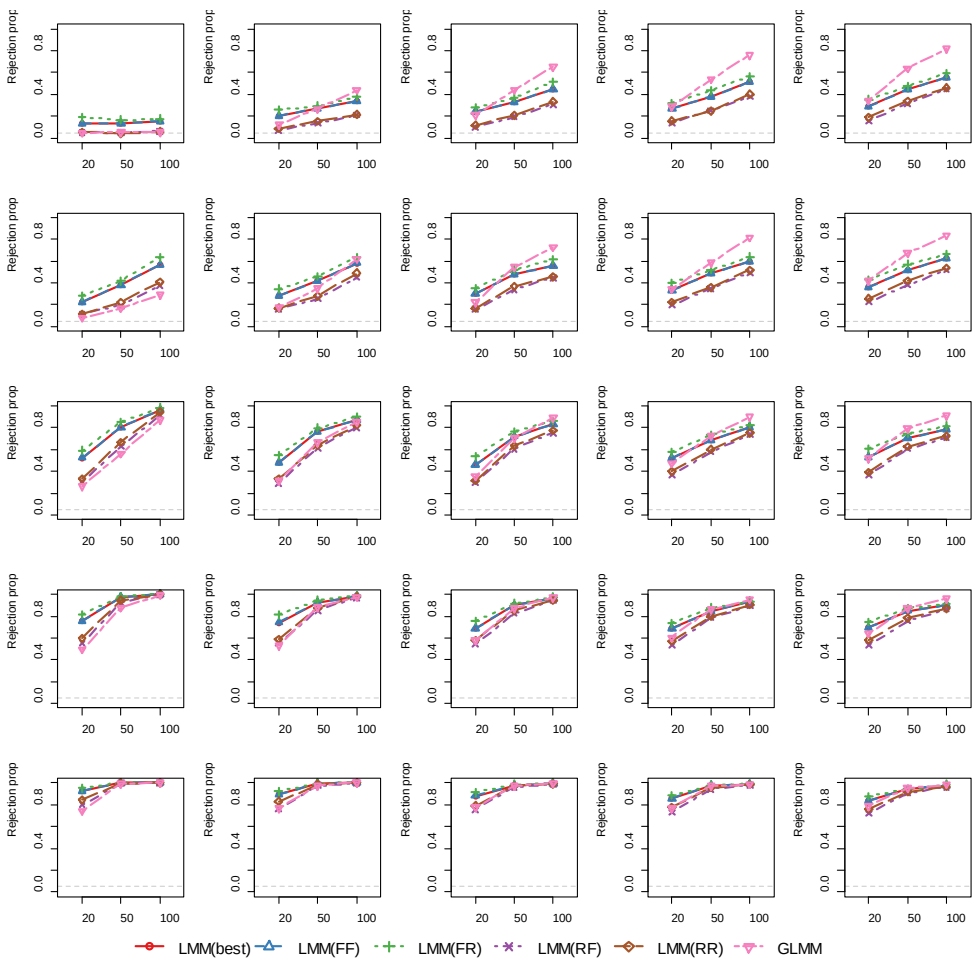

Figure 4. The empirical power of LMM approaches and the GLMM approach in different sample size for randomly generated effects.

The columns represents the variance of $(\boldsymbol{G} \times \boldsymbol{F})_{\boldsymbol{i},\boldsymbol{k}}$ as 0, 0.05, 0.1, 0.15 and 0.2, respectively. The rows represents the group effect as (0, 0, 0), (0, 0, 0.25), (0, 0, 0.5), (0, 0, 0.75) and (0, 0, 1), respectively. The x-axis and y-axis of each panels represent the number of samples and empirical power, respectively. Gray dotted horizontal line of each panels represents the significant level. LMM(FF), LMM(FR), LMM(RF) and LMM(RR) denoted in the legend stand for the view of subject and run effects in model (1) as random, R, or fixed, F, in respective order. The best LMM approach, LMM(best), was selected by AIC among 4 LMM approaches for each 1000 simulation data sets.

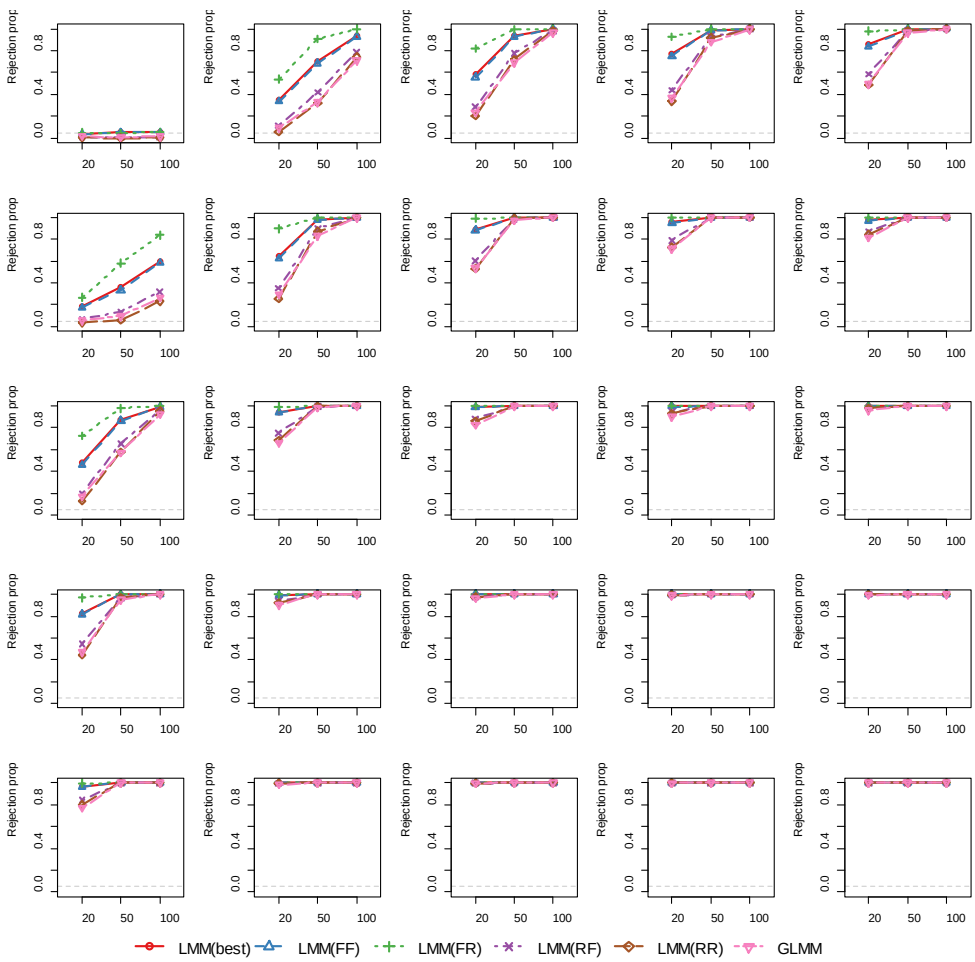

Figure 5. The empirical power of LMM approaches and the GLMM approach in different sample size for fixedly generated effects with same size of interaction effect.

The columns represents the variance of $(\boldsymbol{G} \times \boldsymbol{F})_{\boldsymbol{i,k}}$ as 0, 0.05, 0.1, 0.15 and 0.2, respectively. The rows represents the group effect as (0, 0, 0), (0, 0, 0.25), (0, 0, 0.5), (0, 0, 0.75) and (0, 0, 1), respectively. The x-axis and y-axis of each panels represent the number of samples and empirical power, respectively. Gray dotted horizontal line of each panels represents the significant level. LMM(FF), LMM(FR), LMM(RF) and LMM(RR) denoted in the legend stand for the view of subject and run effects in model (1) as random, R, or fixed, F, in respective order. The best LMM approach, LMM(best), was selected by AIC among 4 LMM approaches for each 1000 simulation data sets.

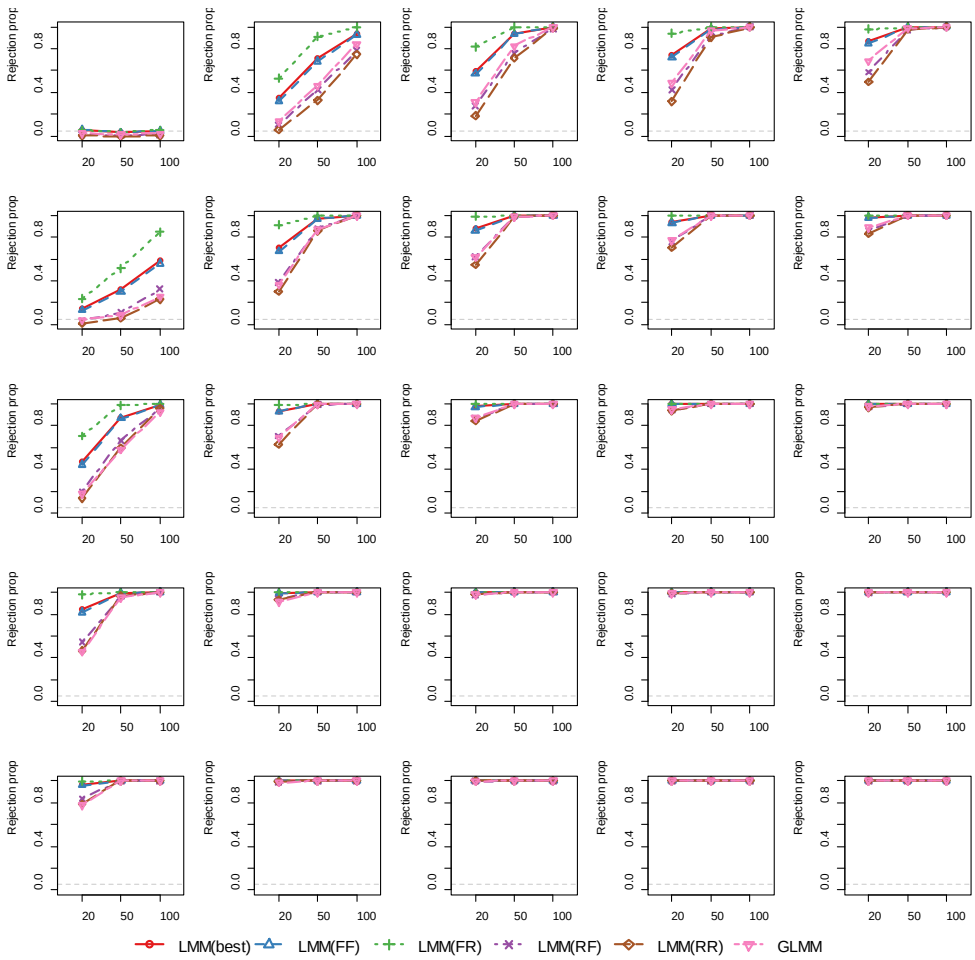

Figure 6. The empirical power of LMM approaches and the GLMM approach in different sample size for fixedly generated effects with different size of interaction effect.

The columns represents the variance of $(\boldsymbol{G} \times \boldsymbol{F})_{\boldsymbol{i,k}}$ as 0, 0.05, 0.1, 0.15 and 0.2, respectively. The rows represents the group effect as (0, 0, 0), (0, 0, 0.25), (0, 0, 0.5), (0, 0, 0.75) and (0, 0, 1), respectively. The x-axis and y-axis of each panels represent the number of samples and empirical power, respectively. Gray dotted horizontal line of each panels represents the significant level. LMM(FF), LMM(FR), LMM(RF) and LMM(RR) denoted in the legend stand for the view of subject and run effects in model (1) as random, R, or fixed, F, in respective order. The best LMM approach, LMM(best), was selected by AIC among 4 LMM approaches for each 1000 simulation data sets.

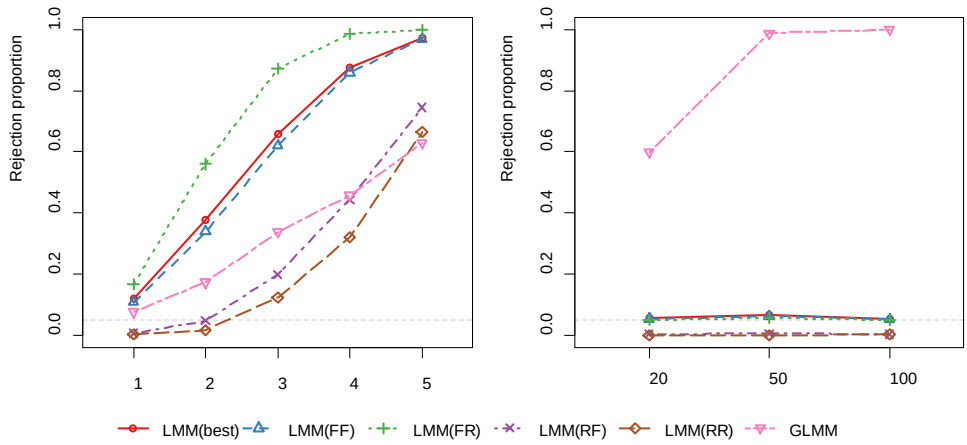

Figure 7. The empirical power of LMM approaches and the GLMM approach for fixedly generated effects when the number of features was 5 and 4.

Left panel: x-axis represents the number of causal features when the number of features was 5. Right panel: x-axis represents the number of sample size when the interaction effect 2 features behave oppositely.

3.1.3 Results for Prediction Performance Comparison

We compared the performance of the best LMM approach, the GLMM approach and the group LASSO approach that identifying significant proteins when there were non-significant proteins mixed in the data set. When the number of significant proteins were 5, best LMM shows the highest proportion that exactly identify significant proteins among three approaches for 20 individuals while the GLMM approach shows the lowest proportion. When the number of individuals increase to 50 the exactly identified proportion of the GLMM approach exceed that of group LASSO approach. For 100 individual case, the GLMM approach shows the highest proportion among three approaches while group LASSO approach shows the

lowest proportion. If we define selection proportion that detected proteins includes significant proteins, then best LMM and group LASSO approach show significantly increased selection proportion while that of the GLMM approach does not increase significantly. When the number of significant proteins were 2, group LASSO approach shows the highest proportion that exactly identify significant proteins among three approaches for 20 individuals while the GLMM approach shows the lowest proportion. When the number of individuals increase to 50 the exactly identified proportion of the best LMM exceed that of group LASSO approach. For 100 individual case, the GLMM approach shows the highest proportion among three approaches while the best LMM shows the lowest proportion. Similar to the case of the number of significant proteins was 5, if we define selection proportion that detected protein set includes significant proteins, then the best LMM and group LASSO approach show significantly increased selection proportion while that of the GLMM approach does not increase significantly.

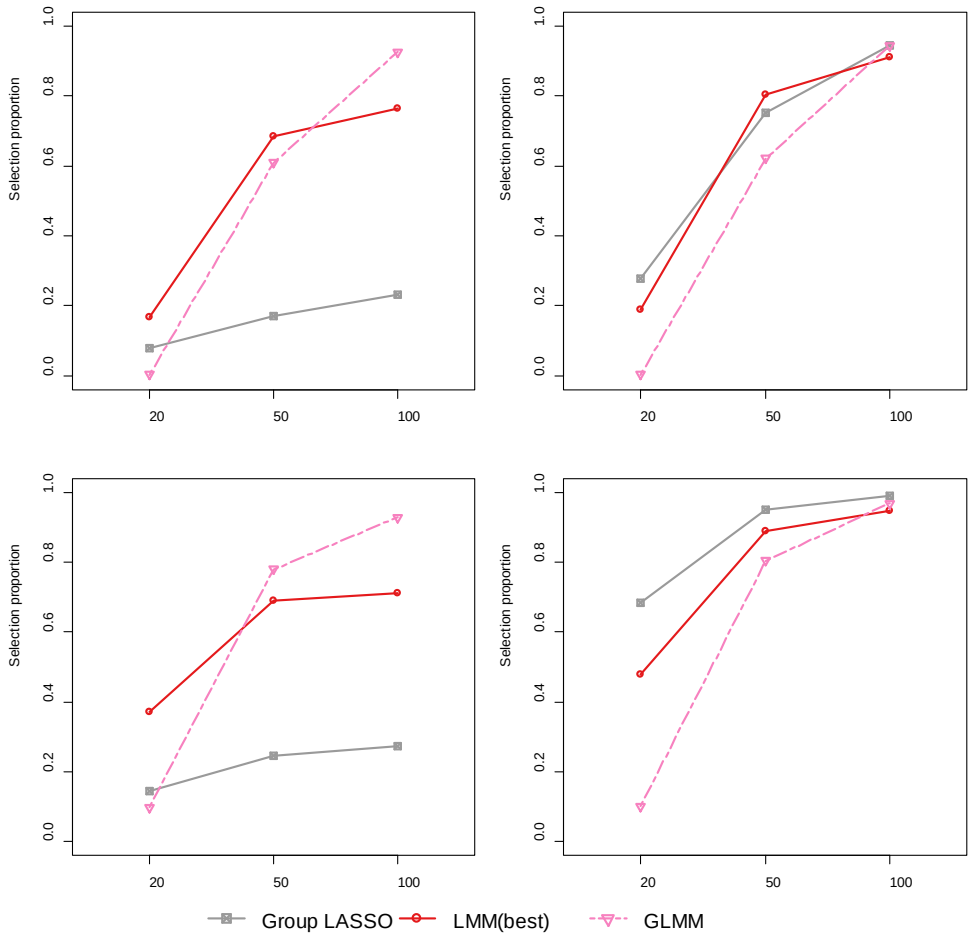

Figure 8. The prediction performance comparison for the group LASSO approach, the best LMM approach and the GLMM approach.

Top panel: Among 10 proteins, 5 proteins were significant, Bottom panel: Among 10 proteins, 2 proteins were significant. Left panel: The proportion that significant proteins detected exactly. Right panel: The proportion that significant proteins were included in detected proteins. The best LMM approach, LMM(best), was selected by AIC among 4 LMM approaches for each 1000 simulation data sets.

4. Application to Real Data

4.1 Sorafenib Drug Response MRM Data

Serum samples were block randomized and immunodepletion on 6 multiple affinity removal system (MARS) columns, denaturation, trypsin digestion, and desalting, followed by reversed-phase liquid chromatography (LC) and MRM mass spectrometry analysis of the obtained peptide samples. All MRM mass spectrometry analysis were performed on an Agilent 6490 triple quadrupole (QqQ) mass spectrometer with Jetstream electrospray source coupled with a 1260 Capillary LC system (Agilent Technologies, Santa Clara, CA). Peptides were separated on a reversed phase analytical column (150mm × 0.5mm i.d., 3.5μm particle size, Agilent Zorbax SB-C18) with mobile phases A: water 0.1% (v/v) formic acid and B: acetonitrile 0.1% (v/v) formic acid. 5 μL of tryptic digest were injected into a column and separated using a binary gradient. The Skyline software [25] was used to import and align all MRM raw data files and perform feature quantitation. Candidate protein biomarkers were chosen based on a LiverAtlas database (http://liveratlas.hupo.org.cn). The following criteria were used to identify candidate biomarkers proteins of sorafenib response: The LiverAtlas database from the 50,265 proteins, known for hepatic disease-associated proteins were downloaded. We focused on proteins that high-throughput experiments and highly focused biochemical studies from the 50,208 proteins. Finally selected 1,683 proteins that secreted or might be secreted proteins on the basis of UniProt Knowledgebase (UniProtKB, http://www.uniprot.org/) database. Then, 930 proteins were filtered with tandem mass spectrometry (MS/MS) spectrum from National Institute of Standards and Technology (NIST) MS/MS library for empirical evidence of mass

spectrometry spectrum detectability. From these proteins, 124 interference free proteins were left after reproducibility check. 76 proteins that contain multiple peptides were analyzed in 50 serum samples using the MRM-based platform. The response of sample were measured by the modified evaluation criteria in solid tumors (mRECIST) guideline. A response of subject that complete response, partial response and stable disease was classified to responder. A response of subject that progressive disease was classified to non-responder. The previously used LMM approach and the proposed GLMM approach were used to analyze all 76 proteins.

4.2 Results

As we illustrated at simulation section, LMM approaches that treat subject effect as fixed show lower *p*-value compared to the other methods in most cases. Proposed GLMM approach identified 26 proteins that change in abundance between responder and non-responder groups with 0.05 significance level. Among 26 proteins that identified by GLMM approach, there were 2 proteins, A2GL and NRP1, that the GLMM approach shows lower *p*-value than four LMM approaches and there were 6 proteins, ANT3, APOA1, BTD, CO7, FETUA and PROS, that the GLMM approach shows lower *p*-value than two LMM approaches that treat subject effect as random. The *p*-values for 8 proteins are depicted in Table 2. The estimates of interaction coefficients between features and groups for A2GL and NRP1 are depicted in Table 3. As we observed in simulation studies, different direction of interaction effects, ENQLEVLEVSWLHGLK / non-responder and VAAGAFQGLR / non-responder, made the *p*-values of the GLMM approach

smaller than 0.05 while those of LMM approaches were not. The estimated interaction coefficients between features and groups for 6 proteins that the GLMM approach performed better than LMM approaches with random subject effect are depicted in Table 4. As we observed in simulation studies, different interaction effect sizes made loss of power of LMM approaches so that LMM approaches with random subject effect has small *p*-values but greater than that of the GLMM approach.

Table 2. *p*-values of 8 proteins that the GLMM approach shows lower *p*-value.

LMM(FF), LMM(FR), LMM(RF) and LMM(RR) represent same model that described in Figure 1. If a *p*-value was smaller than 0.0005, it was represented as "<0.0005".

protein	Model				
	LMM(FF)	**LMM(FR)**	**LMM(RF)**	**LMM(RR)**	**GLMM**
A2GL	0.4748	0.6909	0.5641	0.6951	0.0245
ANT3	<0.0005	0.001	0.0409	0.049	0.0333
APOA1	<0.0005	<0.0005	0.018	0.0162	0.0159
BTD	0.0023	0.0028	0.1007	0.0679	0.0478
CO7	<0.0005	<0.0005	0.0142	0.0165	0.0076
FETUA	<0.0005	0.0012	0.0353	0.0348	0.0244
NRP1	0.327	0.5571	0.292	0.4409	0.002
PROS	0.0179	0.0177	0.2015	0.1826	0.0488

Table 3. Estimated interaction coefficients for A2GL and NRP1 proteins.

LMM(FF), LMM(FR), LMM(RF) and LMM(RR) represent same model that described in Figure 1. The estimated values were rounded up in the third decimal point. Interaction coefficients of feature and groups are represented as "feature / group".

protein	Interaction	Model			
		LMM(FF)	**LMM(FR)**	**LMM(RF)**	**LMM(RR)**
A2GL	ENQLEVLEVSWLHGLK / non-responder	-0.55	-0.51	-0.55	-0.52
	VAAGAFQGLR / non-responder	0.64	0.69	0.64	0.69
	ENQLEVLEVSWLHGLK / responder	0.49	0.44	0.49	0.44
	VAAGAFQGLR / responder	0.45	0.36	0.45	0.37
NRP1	LYQVIFEGEIGK / non-responder	0.8	0.81	0.8	0.81
	LYQVIFEGEIGK / responder	-0.06	-0.08	-0.06	-0.08

Table 4. Estimated interaction coefficients for ANT3, APOA1, BTD, CO7M FETUA and PROS proteins.

LMM(FF), LMM(FR), LMM(RF) and LMM(RR) represent same model that described in Figure 1. The estimated values were rounded up in the third decimal point. Interaction coefficients of feature and groups are represented as "feature / group"

protein	Interaction	Model			
		LMM(FF)	LMM(FR)	LMM(RF)	LMM(RR)
ANT3	TSDQIHFFFAK / non-responder	1.25	1.22	1.25	1.22
	TSDQIHFFFAK / responder	0.42	0.47	0.42	0.47
APOA1	ATEHLSTLSEK / non-responder	4.24	4.26	4.24	4.26
	DLATVYVDVLK / non-responder	1.49	1.64	1.49	1.63
	DYVSQFEGSALGK / non-responder	2.16	2.23	2.16	2.22
	EQLGPVTQEFWDNLEK / non-responder	6.42	6.55	6.42	6.54
	LLDNWDSVTSTFSK / non-responder	3.88	3.86	3.88	3.87
	THLAPYSDELR / non-responder	3.21	3.25	3.21	3.25
	ATEHLSTLSEK / responder	3.19	3.14	3.19	3.14
	DLATVYVDVLK / responder	1.31	1.07	1.31	1.09
	DYVSQFEGSALGK / responder	1.95	1.84	1.95	1.85
	EQLGPVTQEFWDNLEK / responder	6.64	6.42	6.64	6.44
	LLDNWDSVTSTFSK / responder	2.79	2.82	2.79	2.82
	THLAPYSDELR / responder	2.34	2.28	2.34	2.28
BTD	LSSGLVTAALYGR / non-responder	1.08	1.06	1.08	1.06
	LSSGLVTAALYGR / responder	0.31	0.34	0.31	0.34
CO7	LSGNVLSYTFQVK / non-responder	0.61	0.66	0.61	0.65
	VLFYVDSEK / non-responder	-1.34	-1.3	-1.34	-1.3
	LSGNVLSYTFQVK / responder	1.14	1.07	1.14	1.07
	VLFYVDSEK / responder	0.13	0.06	0.13	0.06
FETUA	FSVVYAK / non-responder	-0.09	-0.12	-0.09	-0.12
	FSVVYAK / responder	-0.9	-0.85	-0.9	-0.85
PROS	NNLELSTPLK / non-responder	0.13	0.09	0.13	0.09
	SFQTGLFTAAR / non-responder	1.94	1.92	1.94	1.92
	NNLELSTPLK / responder	-0.16	-0.1	-0.16	-0.1
	SFQTGLFTAAR / responder	0.31	0.35	0.31	0.35

5. Discussions

Although LMM is a powerful model to detect significant proteins, it provides diversified *p*-values for a same data depending on the type of effects, random or fixed. It is well known properties of LMM that underestimate the variance when fixed effect model was applied while the true effect was random and vice versa [26]. Implemented LMM approaches in Skyline program and R programming language let user specify the type, random or fixed, of subject and run effects and these generate diversified *p*-values. However, MSstats does not provide log-likelihood or AIC to do model selection. These various *p*-values and no criterion for model selection made user difficult to determine which model should be used and interpreted. Even if one performs model selection with AIC, it is not useful since the likelihood of the LMM(FR) was dominating among 4 LMM approaches. We proposed the GLMM approach to detect proteins that change in abundance. The proposed GLMM approach was more robust on the type of effects, random or fixed, than LMM approaches. Moreover, GLMM can detect special interaction effect that LMM approaches could not detect. In real data analysis, there were 8 proteins that the GLMM approach presents significant and smaller *p*-values than those of LMM approaches with random subject effect. In this case, either the direction of the interaction effect or the size of the interaction effect was different from what we observed in simulation study. Since representing features was selected under the assumption that the features well represent the protein, it is hard to say that the special interaction effect, LMM approaches poorly detect, is general. However, we identified proteins that belong to the case that LMM approaches could not detect. With all things taken together, when analyzing a MRM data, we propose that not

only LMM approaches but the GLMM approach should be considered.

Bibliography

1. Frantzi, M., A. Bhat, and A. Latosinska, *Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development.* Clin Transl Med, 2014. **3**(1): p. 7.
2. Shi, T., et al., *Advancing the sensitivity of selected reaction monitoring-based targeted quantitative proteomics.* Proteomics, 2012. **12**(8): p. 1074-1092.
3. Pan, S., et al., *Mass spectrometry based targeted protein quantification: methods and applications.* Journal of proteome research, 2008. **8**(2): p. 787-797.
4. Shi, T., et al., *Antibody-free, targeted mass-spectrometric approach for quantification of proteins at low picogram per milliliter levels in human plasma/serum.* Proceedings of the National Academy of Sciences, 2012. **109**(38): p. 15395-15400.
5. Haab, B.B., et al., *A Reagent Resource to Identify Proteins and Peptides of Interest for the Cancer Community A WORKSHOP REPORT.* Molecular & Cellular Proteomics, 2006. **5**(10): p. 1996-2007.
6. Rifai, N., M.A. Gillette, and S.A. Carr, *Protein biomarker discovery and validation: the long and uncertain path to clinical utility.* Nature biotechnology, 2006. **24**(8): p. 971-983.
7. Mesri, M., *Advances in Proteomic Technologies and Its Contribution to the Field of Cancer.* Advances in Medicine, 2014. **2014**.
8. Lin, D., et al., *Comparison of protein immunoprecipitation-multiple reaction monitoring with ELISA for assay of biomarker candidates in plasma.* Journal of proteome research, 2013. **12**(12): p. 5996-6003.
9. Liebler, D.C. and L.J. Zimmerman, *Targeted quantitation of proteins*

by mass spectrometry. Biochemistry, 2013. **52**(22): p. 3797-3806.

10. Barnidge, D.R., et al., *Absolute quantification of the model biomarker prostate-specific antigen in serum by LC-Ms/MS using protein cleavage and isotope dilution mass spectrometry.* Journal of proteome research, 2004. **3**(3): p. 644-652.
11. Han, B. and R.E. Higgs, *Proteomics: from hypothesis to quantitative assay on a single platform. Guidelines for developing MRM assays using ion trap mass spectrometers.* Briefings in functional genomics & proteomics, 2008. **7**(5): p. 340-354.
12. Hale, J.E., *Advantageous uses of mass spectrometry for the quantification of proteins.* International journal of proteomics, 2013. **2013**.
13. Carr, S.A., et al., *Targeted peptide measurements in biology and medicine: best practices for mass spectrometry-based assay development using a fit-for-purpose approach.* Molecular & Cellular Proteomics, 2014. **13**(3): p. 907-917.
14. Picotti, P. and R. Aebersold, *Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions.* Nature methods, 2012. **9**(6): p. 555-566.
15. Grebe, S.K. and R.J. Singh, *LC-MS/MS in the Clinical laboratory–Where to from here?* The Clinical Biochemist Reviews, 2011. **32**(1): p. 5.
16. Mani, D., S.E. Abbatiello, and S.A. Carr, *Statistical characterization of multiple-reaction monitoring mass spectrometry (MRM-MS) assays for quantitative proteomics.* BMC bioinformatics, 2012. **13**(Suppl 16): p. S9.
17. Colangelo, C.M., et al., *Review of software tools for design and analysis of large scale MRM proteomic datasets.* Methods, 2013. **61**(3): p. 287-298.
18. Freue, G.V.C. and C.H. Borchers, *Multiple Reaction Monitoring*

(MRM) Principles and Application to Coronary Artery Disease. Circulation: Cardiovascular Genetics, 2012. **5**(3): p. 378-378.

19. Yassine, H.N., et al., *The application of multiple reaction monitoring to assess ApoA-I methionine oxidations in diabetes and cardiovascular disease.* Translational proteomics, 2014. **4**: p. 18-24.
20. Zhang, P., et al., *Multiple reaction monitoring to identify site-specific troponin I phosphorylated residues in the failing human heart.* Circulation, 2012: p. CIRCULATIONAHA. 112.096388.
21. Chang, C.-Y., et al., *Protein significance analysis in selected reaction monitoring (SRM) measurements.* Molecular & Cellular Proteomics, 2012. **11**(4): p. M111. 014662.
22. Pursiheimo, A., et al., *Optimization of statistical methods impact on quantitative proteomics data.* Journal of proteome research, 2015. **14**(10): p. 4118-4126.
23. McDonald, J.H., *Handbook of biological statistics*. Vol. 2. 2009: Sparky House Publishing Baltimore, MD.
24. Lin, X., *Variance component testing in generalised linear models with random effects.* Biometrika, 1997. **84**(2): p. 309-326.
25. MacLean, B., et al., *Skyline: an open source document editor for creating and analyzing targeted proteomics experiments.* Bioinformatics, 2010. **26**(7): p. 966-968.
26. Chung, Y., S. Rabe-Hesketh, and I.H. Choi, *Avoiding zero between-study variance estimates in random-effects meta-analysis.* Statistics in medicine, 2013. **32**(23): p. 4071-4089.

초 록

단백질 바이오마커의 발굴은 현재 생물의학 연구의 중요한 쟁점 중 하나이다. 엘라이자(enzyme-linked immunosorbent assay, ELISA) 는 전통적인 단백질 정량 방법의 하나다. 많은 수의 새로운 단백질들이 연구됨에 따라 엘라이자를 이용한 단백질 바이오마커 발굴에 있어 새로운 쟁점들이 드러났다. 다중반응관측(multiple reaction monitoring, MRM) 질량분석 방법은 엘라이자를 대체할 수 있는 특정 단백질 정량 방법이며 최근 더욱 활용되고 있다. 그러나 이러한 다중반응관측 자료를 이용한 통계적인 단백질 연관성 분석 방법의 개발은 크게 주목받지 못하였다. 초기에는 다중반응관측 자료를 이용하여 두 집단의 평균에 차이를 검증하기 위해 t 검정 혹은 쌍체 t 검정이 이루어 졌고 여러 집단의 평균 차이를 검증하기 위해 선형모형을 이용한 방법이 적용되었다. 2012년에 MSstats 이라 불리는 선형혼합모형을 이용한 단백질 연관성 분석 방법이 제안되었고 이후 널리 사용되고 있다. 이 선형혼합모형을 이용한 방법은 Skyline 프로그램과 R 프로그래밍 언어를 통해서 사용할 수 있다. 선형혼합모형을 통해 계산된 단백질 연관성 p 값은 모형 설정에 따라 많은 변화가 있고 이로 말미암아 많은 거짓 양성 혹은 참 음성이 생길 수 있다. 더욱이 이전에 제안된 선형혼합모형을 이용한 방법은 단백질을 대표하는 특징요인들의 집단 간 발현 양상이 서로 다르게 되면 검정력에 손실이 생기게 된다. 이러한 요인들이 임의효과인지 고정효과인지 에 대해 더욱 강건하며 특징요인들의 집단 간 발현 양상에도 더욱 강건한 모형을 제안하는

동기가 되었다. 우리는 일반화 선형혼합모형의 분산성분 검정방법을 이용한 단백질 연관성 분석을 제안하였다. 우리는 일반화 선형혼합모형 방법이 이전에 제안되었던 선형혼합모형 방법보다 요인들의 효과의 유형에 대해 더욱 강건하다는 것을 다양한 모의실험을 통해 관찰하였고 더욱 검정력이 좋은 경우를 다양한 모의실험을 통해 관찰하였다. 그뿐만 아니라 특징요인들의 집단 간 발현 양상이 서로 다를 때에 선형혼합모형 방법은 저조한 검정력을 보이지만 새롭게 제안한 일반화 선형혼합모형을 이용한 방법은 검정력에 손실이 없었다. 새롭게 제안한 일반화 선형혼합모형 방법이 유의미한 p 값을 제시하고 이전에 제안되었던 선형혼합모형 방법이 유의미하지 않은 p 값을 제시하는 경우를 실제 자료 분석을 통해 관찰하였다. 따라서 다중반응관측 자료를 이용해 단백질 연관성 분석을 할 때에는 이전에 제안되었던 선형혼합모형을 이용한 방법뿐만 아니라 새롭게 제안한 일반화 선형혼합모형을 이용한 방법도 같이 사용되어야 한다.

주요어: 단백질 연관성 분석, 선형혼합모형, 검정력 손실, 분산성분 검정, 일반화 선형혼합모형

학 번: 2014-20297