Stat Methods Appl (2017) 26:273–292
DOI 10.1007/s10260-016-0369-4

ORIGINAL PAPER

High dimensional extension of the growth curve model and its application in genetics

Sayantee Jana[1] · Narayanaswamy Balakrishnan[1] · Dietrich von Rosen[2,3] · Jemila Seid Hamid[1,4,5]

Accepted: 9 September 2016 / Published online: 19 September 2016

Abstract Recent advances in technology have allowed researchers to collect large scale complex biological data, simultaneously, often in matrix format. In genomic studies, for instance, measurements from tens to hundreds of thousands of genes are taken from individuals across several experimental groups. In time course microarray experiments, gene expression is measured at several time points for each individual across the whole genome resulting in a high-dimensional matrix for each gene. In such experiments, researchers are faced with high-dimensional longitudinal data. Unfortunately, traditional methods for longitudinal data are not appropriate for high-dimensional situations. In this paper, we use the growth curve model and introduce test useful for high-dimensional longitudinal data and evaluate its performance using simulations. We also show how our approach can be used to filter genes in time course genomic experiments. We illustrate this using publicly available genomic data, involving experiments comparing normal human lung tissue with vanadium pentoxide treated human lung tissue, designed with the aim of understanding the susceptibility of individuals working in petro-chemical factories to airway re-modelling. Using our method, we were able to filter out 1053 (about 5 %) genes as non-noise genes from a pool of 22,277. Although our focus is on hypothesis testing, we also provided modi-

✉ Jemila Seid Hamid
jhamid@mcmaster.ca

[1] Department of Mathematics and Statistics, McMaster University, Hamilton, Canada

[2] Department of Energy and Technology, Swedish Agricultural University, Uppsala, Sweden

[3] Department of Mathematics, Linköping University, Linköping, Sweden

[4] Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, ON, Canada

[5] Department Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada

fied maximum likelihood estimator for the mean parameter of the growth curve model and assessed its performance through bias and mean squared error.

Keywords Growth curve model · Longitudinal data · High-dimensional data · Time course data · Genomic data · Gene filtering

1 Introduction

Technological advances within the last couple of decades have allowed statisticians to have acquisition of large scale and complex biological data, where large number of variables are measured simultaneously from often very small number of experimental units. One such example is the field of genomics, where measurements from tens to hundreds of thousands of genes are taken simultaneously from a given individual (Jonhstone and Titterington 2009; Smyth 2004; Lönnstedt and Speed 2002). However, with this access to abundant data, statisticians are also faced with enormous methodological and computational challenges, where traditional approaches, often designed under the assumption of large n and small p, fail under high-dimensional situations.

In time course genomic experiments, data become even more complex, where measurements for the tens of thousands of genes are taken from individuals at several time points. Adding to this layer of complexity, sample size is often smaller than the number of time points, leading to singularity problems. Moreover, given very few replications and several thousands of genes, the covariance matrices are poorly estimated (Yuan and Kendziorksi 2006; Ma 2006; Tai and Speed 2006, 2009; Hamid and Beyene 2009). Time course gene expression data allow researchers to explore the temporal expressions of tens of thousands of genes simultaneously and researchers are interested in identifying genes that are relevant to the study at hand (gene filtering); that is, genes with change in expression profiles over time as well as genes that have different expression profiles across groups.

A few methods useful in the analysis of time course genomic experiments have been proposed in recent years, among them are a moderated Hotelling's T^2 statistic (Tai and Speed 2006) and moderated likelihood ratio statistic under the MANOVA model (moderated Wilks' lambda) (Tai and Speed 2009). These approaches are based on the empirical Bayes method, where the authors used a prior on the covariance matrix. Moderated sample variance-covariance matrix is then provided using the posterior distribution. The parameters of the prior distribution are estimated from data across the genes. Although both methods address the problem of high-dimensionality while accounting for the correlation across time points, they rely on MANOVA models and hence do not account for temporal ordering. Moreover, the models do not account for time dependency and the actual time points do not enter the model. Other methods for high dimensional problems that have been proposed in the recent past include two-sample test for high-dimensional data (Chen and Qin 2010) and regularized Hotelling's T^2 test statistic (Chen et al. 2012).

In most longitudinal studies, in general, and time course genomic studies, in particular, the mean is often structured. Gene expression over time is a biologically continuous process and can thus be represented by a continuous function (Hamid and Beyene

2009; Ma 2006). Furthermore, it is indicated that individual genes often share similar expression patterns, although the shape of the functions might be different. Methods that overcome the challenges of high dimensionality, incorporate correlation across time points (within individuals) as well as model the gene expression profile over time, are more optimal. Using the growth curve model (GCM), which is a generalized multivariate analysis of variance (GMANOVA) model, Hamid and Beyene (2009) proposed a method for analysing time-course microarry experiments. They used gene specific GCM to model expression values for a given gene, where the covariance matrix is assumed to be distributed as an inverted Wishart random variable similar to Tai and Speed (2006). The authors then applied Potthoff and Roy's approach to transform the model into MANOVA, where the moderated sample covariance matrix using empirical Bayes approach, is used as the arbitrary matrix **G** in the transformation. Genes were then ranked using the resulting test statistic. The method takes the correlation into account and also allows structured mean, and hence allowing the relationship across time to be modelled. However, the limitation with this approach is that data are first transformed, where each group is represented by polynomials over time. The test statistic is then calculated using the transformed data. The transformation might lead to loss of relevant information and hence result in low power. Moreover, the distribution of the test statistic under the null hypothesis is difficult to obtain, as a result it is not possible to get a critical value for the test statistic, allowing only gene ranking to be performed. For some other approaches concerning MANOVA modelling under high-dimensionality, see Läuter (2009), where additional references are also given.

In this paper, we consider the trace test for the GCM model proposed by Hamid et al. (2011) and provide a moderated version of the test, using the Moore–Penrose generalized inverse. We perform extensive simulations to show the performance of the proposed test statistic. We illustrate application of the method for gene-filtering using publicly available time course genetic data.

2 Trace test for the growth curve model and its high-dimensional extension

Suppose we have k groups where p repeated observations at different time points are taken from individuals in each group. Time dependency is assumed to be a polynomial of degree $q-1 (p \geq q)$. The expected value of the ith group can, therefore, be described as (Pan and Fang 2002; Hamid et al. 2011)

$$b_{0,i} + b_{1,i}t + \cdots + b_{q-1,i}t^{q-1}, \quad i = 1, 2, \ldots, k. \tag{1}$$

In matrix notation, the model can be represented as the following multivariate bilinear setup, which is referred to as the Growth Curve Model (GCM) (Pan and Fang 2002; Kollo and Rosen 2005)

$$\boldsymbol{Y} = \boldsymbol{ZBX} + \boldsymbol{E}, \tag{2}$$

where $\boldsymbol{Y} : p \times n$ is the observation or response matrix, $\boldsymbol{B} : q \times k$ is the parameter matrix, $\boldsymbol{Z} : p \times q$ is the within individual design matrix, $\boldsymbol{X} : k \times n$ is the between individual design matrix. Columns of the error matrix $\boldsymbol{E}$ are assumed to follow the multivariate normal distribution with mean zero and positive definite variance-covariance matrix, $\boldsymbol{\Sigma}$ i.e. $\boldsymbol{E} \sim N_{p\times n}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I})$ and $q \leq p$, $rank(\boldsymbol{X}) + p \leq n$ (Potthoff and Roy 1964; Khatri 1966; Rosen 1989; Pan and Fang 2002). The within individual design matrix $\boldsymbol{Z}$ takes care of the time dependency within the individuals, whereas the between individual design matrix $\boldsymbol{X}$ models the group differences. Note that when $\boldsymbol{Z} = \boldsymbol{I}$, the GCM reduces to the classical MANOVA model, indicating that linear restriction on the MANOVA models results in the GCM. Note also that the GCM can be viewed as a generalization of MANOVA, and hence the name Generalized Multivariate Analysis of Variance (GMANOVA) model.

Suppose now that the GCM is fitted for a given longitudinal data and we are interested in the hypotheses $H_0 : \boldsymbol{B} = \mathbf{0}$ against $H_1 : \boldsymbol{B} \neq \mathbf{0}$. Consider the likelihood function for the GCM under the assumption of multivariate normal distribution, which can be written as

$$L = \alpha|\boldsymbol{\Sigma}|^{-\frac{n}{2}}e^{-\frac{1}{2}tr\{\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{ZBX})(\mathbf{Y}-\mathbf{ZBX})'\}},$$

where $\alpha = (2\pi)^{-\frac{1}{2}np}$. The likelihood function can be decomposed into two independent components as

$$L = L_1 * L_2, \tag{3}$$

where

$$\begin{aligned} L_1 &= \alpha e^{-\frac{1}{2}tr\{\boldsymbol{\Sigma}^{-1}(\mathbf{YX}'(\mathbf{XX}')^{-}\mathbf{X}-\mathbf{ZBX})(\mathbf{YX}'(\mathbf{XX}')^{-}\mathbf{X}-\mathbf{ZBX})'\}}, \\ L_2 &= |\boldsymbol{\Sigma}|^{-\frac{n}{2}}e^{-\frac{1}{2}tr\{\boldsymbol{\Sigma}^{-1}\mathbf{S}\}}, \end{aligned}$$

where,

$$\mathbf{S} = \mathbf{Y}(\mathbf{I} - \mathbf{X}'(\mathbf{XX}')^{-}\mathbf{X})\mathbf{Y}'.$$

Note that the first and the second components of the likelihood, respectively, are functions of the sample mean and sample covariance matrix of a multivariate normal distribution and hence are independent. In our next step, we maximize the L_2 with respect to $\boldsymbol{\Sigma}$, which results in

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{S}.$$

We then replace $\boldsymbol{\Sigma}$ by its estimator to get the estimated likelihood

$$EL = \alpha_1|\mathbf{S}|^{-\frac{n}{2}}e^{-\frac{1}{2}ntr\{\mathbf{S}^{-1}(\mathbf{YX}'(\mathbf{XX}')^{-}\mathbf{X}-\mathbf{ZBX})(\mathbf{YX}'(\mathbf{XX}')^{-}\mathbf{X}-\mathbf{ZBX})'\}}, \tag{4}$$

where $\alpha_1 = n^{\frac{n}{2}}(2\pi)^{-\frac{1}{2}np}e^{-\frac{1}{2}pn}$. The ratio of the maximum of the estimated likelihood under H_o and $H_o \cup H_1$ is then taken to obtain the trace test

$$\phi(\mathbf{Y}) = tr\{\mathbf{S}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{S}^{-1}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{S}^{-1}\mathbf{YX}'(\mathbf{XX}')^{-}\mathbf{XY}'\}. \tag{5}$$

The null hypothesis is rejected when $\phi(\mathbf{Y}) > c$, where c is calculated such that $P_{H_o}(\phi(\mathbf{Y}) > c) = \alpha$, where α is the desired level for the test.

Suppose now that the covariance matrix $\boldsymbol{\Sigma}$ is known, then the ratio of the likelihood function of the GCM under H_o and $H_o \cup H_1$ results in

$$\phi(\mathbf{Y}) = tr\{\boldsymbol{\Sigma}^{-1}\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z})^{-}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-}\mathbf{X}\mathbf{Y}'\}. \tag{6}$$

That is, the test statistic reduces to (6), when $\boldsymbol{\Sigma}$ is known. If we compare the two test statistics provided in (5) and (6), it appears that the covariance matrix $\boldsymbol{\Sigma}$ in (6) is replaced by the sample covariance matrix $\boldsymbol{S}$ in (5), when $\boldsymbol{\Sigma}$ is unknown. Now, let us consider high-dimensional longitudinal data, where the sample covariance matrix $\boldsymbol{S}$ is singular and (5) is no longer a valid test. Different regularization methods can be considered to overcome this challenge, however, in this paper we consider a simple approach, where the inverse of the unknown covariance matrix, $\boldsymbol{\Sigma}^{-1}$ is replaced by the Moore–Penrose generalized inverse of the sample covariance matrix, $\boldsymbol{S}$. For details on the Moore–Penrose generalized inverse, we refer the reader to Moore (1920), Penrose (1955) and Rao and Mitra (1972). We refer to this test as the moderated trace test and it can be written as

$$\widetilde{\phi} = tr\{\mathbf{S}^{-}\mathbf{Z}(\mathbf{Z}'\mathbf{S}^{-}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{S}^{-}\mathbf{Y}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-}\mathbf{X}\mathbf{Y}'\}. \tag{7}$$

Now consider the original trace test provided in (5), it has been shown that it has the same distribution as (Hamid et al. 2011)

$$tr\{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{W}^{-1}(\boldsymbol{I}:\mathbf{0})'((\boldsymbol{I}:\mathbf{0})\boldsymbol{W}^{-1}(\boldsymbol{I}:\mathbf{0})')^{-}(\boldsymbol{I}:\mathbf{0})\boldsymbol{W}^{-1}\},$$

where $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}$, and $\boldsymbol{U} = \boldsymbol{Y}\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-}\boldsymbol{Z}\boldsymbol{Y}'$. Note that $\boldsymbol{W}$ and $\boldsymbol{U}$ are independent and are distributed as Wishart random variables, both free of the unknown covariance matrix $\boldsymbol{\Sigma}$. Let us now re-write the trace test as (by substituting $\boldsymbol{W}^{-1} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}^{1/2}$)

$$\begin{aligned} &tr\{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}S^{-1}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{I}:\mathbf{0})'((\boldsymbol{I}:\mathbf{0})\boldsymbol{\Sigma}^{1/2}\mathbf{S}^{-1}\boldsymbol{\Sigma}^{1/2} \\ &\quad(\boldsymbol{I}:\mathbf{0})')^{-}(\boldsymbol{I}:\mathbf{0})\boldsymbol{\Sigma}^{1/2}\mathbf{S}^{-1}\boldsymbol{\Sigma}^{1/2}\}. \end{aligned} \tag{8}$$

Observe now that, in the moderated test, where $\mathbf{S}$ is singular due to high-dimensionality, we substitute $\mathbf{S}^{\mathbf{-1}}$ by the generalized inverse $\mathbf{S}^{-}$. Equation (8), therefore, reduces to

$$tr\{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}S^{-}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{I}:\mathbf{0})'((\boldsymbol{I}:\mathbf{0})\boldsymbol{\Sigma}^{1/2}S^{-}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{I}:\mathbf{0})')^{-}(\boldsymbol{I}:\mathbf{0})\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}^{-}\boldsymbol{\Sigma}^{1/2}\}.$$

Now consider $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}$. It is shown above that its distribution is free of $\boldsymbol{\Sigma}$. Using the property of a generalized inverse, we can write $\boldsymbol{S}$ as $\boldsymbol{S}\boldsymbol{S}^{-}\boldsymbol{S}$, so that $\boldsymbol{W}$ can be re-written as

$$\boldsymbol{W} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{S}^{-}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2},$$

which can be further re-written as

$$\boldsymbol{W} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}^{-}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{W}\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}^{-}\boldsymbol{\Sigma}^{1/2}\boldsymbol{W},$$

indicating that $\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}^{-}\boldsymbol{\Sigma}^{1/2}$ is a reflexive generalized inverse of $\boldsymbol{W}$. However, in order the $\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}^{-}\boldsymbol{\Sigma}^{1/2}$ to be equivalent to the Moore–Penrose generalized inverse of $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}$, we need to assume that $\boldsymbol{S}$ is $\boldsymbol{\Sigma}$ invariant, i.e $C(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) = C(\boldsymbol{S})$, where $C(.)$, is the column space of a matrix. Consequently, the distribution of the moderated trace test is free of $\boldsymbol{\Sigma}$, only in some special cases. As a result, in practical applications involving high-dimensional data, we recommend re-sampling approaches (such as the bootstrap) to empirically generate the null distribution.

Note that we are testing statistical hypothesis for a matrix parameter, and so the difference between the values under the null and alternative hypotheses is also a matrix, and investigating some properties of the power curve (eg. unbiasedness, monotonicity) is not obvious. For simplicity, we used the vector function to change the matrix into a long vector and calculated Euclidean distance between the two vectors, i.e. $d =$ Euclidean Distance $(vec(\boldsymbol{B}_o), vec(\boldsymbol{B}_1))$, where $vec(\boldsymbol{B}_o)$ and $vec(\boldsymbol{B}_1)$ are the vectorized forms of parameter matrices under the null and alternative hypotheses, respectively. Note that this distance measure is equivalent to the Euclidean norm used to calculate distance between two matrices, which is also often referred to as Frobenius norm or the Hilbert-Schimdt norm (Horn and Johnson 1985). When assessing symmetry of the power curve, we used the matrix form where we compared power curves for $\boldsymbol{B}_1$ and $-\boldsymbol{B}_1$, that is matrices of same distance from zero but in opposite directions.

3 Multivariate bias and mean squared error (MSE)

Although the main focus of this paper is testing linear hypothesis in high-dimensional longitudinal situations, we also investigated the maximum likelihood estimator of the parameter matrix $\boldsymbol{B}$, mainly because this estimator is also a function of $\boldsymbol{S}^{-1}$ and hence undefined in high-dimensional situations.

Consider the maximum likelihood estimator (MLE) for $\boldsymbol{B}$, which is derived by (Khatri 1966). For high-dimensional data, where $\boldsymbol{S}$ in singular, we use the Moore–Penrose generalized inverse and examine the performance of this estimator through simulation studies. To avoid confusion with MLE, we will refer to this estimator as the moderated maximum likelihood estimator (MMLE).

$$\widehat{\boldsymbol{B}} = (\boldsymbol{Z}'\boldsymbol{S}^{-}\boldsymbol{Z})^{-}\boldsymbol{Z}'\boldsymbol{S}^{-}\boldsymbol{Y}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-}. \tag{9}$$

Investigation of performance of estimators is often done using bias and mean squared error. Since our parameter of interest, $\boldsymbol{B}$, is a matrix, we define the bias matrix as the expected difference of the estimator and the true parameter matrix, keeping it analogous to its univariate counterpart, where bias is defined as $E_\theta(\widehat{\theta} - \theta)$. However, comparison of bias matrices among different scenarios (for different parameter matrices across different sample sizes) is not straight forward. It is necessary to reduce the bias matrix to a scalar quantity. To that effect, we first converted the bias matrix for each parameter matrix into a vector format and calculated the Euclidean distance of the bias vectors from a vector of zeros. Scenarios are, therefore, compared using the Euclidean distance of the bias vectors. This is again equivalent to the Frobenius norm measuring the distance between two matrices under the L^2 norm defining the distance between the bias matrix and the unbiased matrix (a matrix of all zeroes), thus giving us an intuitive measure for Bias.

Investigating only bias of a given estimator can be misleading since decrease in bias might be at the expense of increase in variance (decrease in precision). Bias investigation, therefore, should be accompanied by investigation of mean squared error (MSE). The mean squared error for an estimator of a scalar parameter is defined as $E_\theta(\widehat{\theta} - \theta)^2$ (Casella and Berger 2002). Analogous to the definition of the scalar MSE, we define matrix MSE for our estimator as the expectation of the inner product of the difference between the estimator and the true parameter matrix which is

$$\begin{aligned}
MSE &= E[(\widehat{\boldsymbol{B}} - \boldsymbol{B})'(\widehat{\boldsymbol{B}} - \boldsymbol{B})] \\
&= E[(\widehat{\boldsymbol{B}} - E(\widehat{\boldsymbol{B}}) + E(\widehat{\boldsymbol{B}}) - \boldsymbol{B})'(\widehat{\boldsymbol{B}} - E(\widehat{\boldsymbol{B}}) + E(\widehat{\boldsymbol{B}}) - \boldsymbol{B})] \\
&= E[(\widehat{\boldsymbol{B}} - E(\widehat{\boldsymbol{B}}))'(\widehat{\boldsymbol{B}} - E(\widehat{\boldsymbol{B}})) - (\widehat{\boldsymbol{B}} - E(\widehat{\boldsymbol{B}}))'(\boldsymbol{B} - E(\widehat{\boldsymbol{B}})) \\
&\quad + (E(\widehat{\boldsymbol{B}}) - \boldsymbol{B})'(\widehat{\boldsymbol{B}} - E(\widehat{\boldsymbol{B}})) + (E(\widehat{\boldsymbol{B}}) - \boldsymbol{B})'(E(\widehat{\boldsymbol{B}}) - \boldsymbol{B})] \\
&= Cov(\widehat{\boldsymbol{B}}) - (E(\widehat{\boldsymbol{B}}) - E(\widehat{\boldsymbol{B}}))'(\widehat{\boldsymbol{B}} - E(\boldsymbol{B})) \\
&\quad + (E(\widehat{\boldsymbol{B}}) - \boldsymbol{B})'(E(\widehat{\boldsymbol{B}}) - E(\widehat{\boldsymbol{B}})) + Bias'Bias \\
&= Cov(\widehat{\boldsymbol{B}}) + Bias'Bias,
\end{aligned}$$

where $Cov(\widehat{\boldsymbol{B}})$ matrix in MSE is the variance-covariance matrix. Unlike the univariate case, where $MSE = bias^2 + variance$, our matrix MSE involves covariance terms. However, covariances of estimators may not be associated with precision of estimators, and hence we also considered another form of the matrix MSE by taking just the diagonal of the variance covariance matrix. In comparing MSE among the different scenarios, we first converted the matrix MSE to vector form and calculated the Euclidean distance between this vector and a vector of zeroes.

4 Simulations

We performed extensive simulations to investigate the performance of the proposed moderated trace test. This was done by considering many plausible values of the parameter of interest, $\boldsymbol{B}$, and several covariance structures for the longitudinal data. We calculated empirical level and power of the test for all the scenarios included. We also investigated power for different sample sizes and for different number of time points, and investigated power curves for increasing sample size and increasing $\boldsymbol{B}$ with respect to the Euclidean distance. For the estimation problem, we calculated empirical bias and MSE for several scenarios. The description as well as the results of the simulations are presented in this section. Table 1 shows the range of the parameters used in our simulation.

Table 1 Range of parameters considered in the simulation study

Number of groups, k	2
Average growth curve	Linear
Sample size, n	$n = 5, 10, 15, 20, 25, 30, 35, 40, 50, 60$ and $n < p$
Number of time-points, p	$p = 25, 40, 100$
Variance covariance matrix	Generated from the Wishart distribution
Euclidean distance of the parameter matrices $\boldsymbol{B}$ from the null	0–1.8, increment by 0.1
Distribution of the outcome variable	Multivariate normal

Description of simulation

For each simulation scenario, the Growth Curve Model (GCM) is fitted, where a combination of parameters were generated according to the values provided in Table 1. The design matrices $\boldsymbol{Z}$ and $\boldsymbol{X}$ are

$$\mathbf{Z} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & p \end{pmatrix} \text{ and } \mathbf{X} = \begin{pmatrix} 1_{n_1} & 0_{n_2} \\ 0_{n_1} & 1_{n_2} \end{pmatrix}.$$

The hypothesis of $\boldsymbol{B} = \mathbf{0}$ vs $\boldsymbol{B} \neq \mathbf{0}$ is investigated and empirical level and power of the proposed test are calculated for each scenario. The methods presented and hence the simulation results do not depend on (are invariant to) the number of groups k. Therefore, we have, without loss of generality, assumed the number of groups to be 2. Similarly, the test statistic does not depend on the dimension of $\boldsymbol{B}$ and hence the average growth curve for each of the groups is assumed to be linear. However, we have also performed simulations with quadratic growth curves and the findings are similar.

Simulation results

The null distributions for the moderated trace test under different scenarios are presented in the histograms in Fig. 1, where the empirical critical values are indicated using arrows. As can be seen from the figure, the distribution is positively skewed. We have investigated the null distribution for many different values of p and n and the results are consistent, where the distributions for all the scenarios are positively skewed.

In all the scenarios considered, the empirical level of the moderated test averaged over 10,000 simulations is very close to the nominal level of significance $\alpha = 0.05$ (Fig. 2a). When we take a closer look at the figure (enlarged figure is also provided), we see that empirical level are randomly distributed around the nominal level of significance regardless of sample size (Fig. 2b). Most of the level values lie within the 95 % confidence bands. The results are consistent for different values of p.

In evaluating the power of the moderated trace test, we considered 38 different parameter matrices under the alternative hypothesis. Out of the 38 $\boldsymbol{B}$ matrices considered, 19 have positive entries and are thus plotted on the right hand side of the null hypothesis ($\boldsymbol{B} = 0$) and the remaining 19 have exactly the same magnitude but had negative

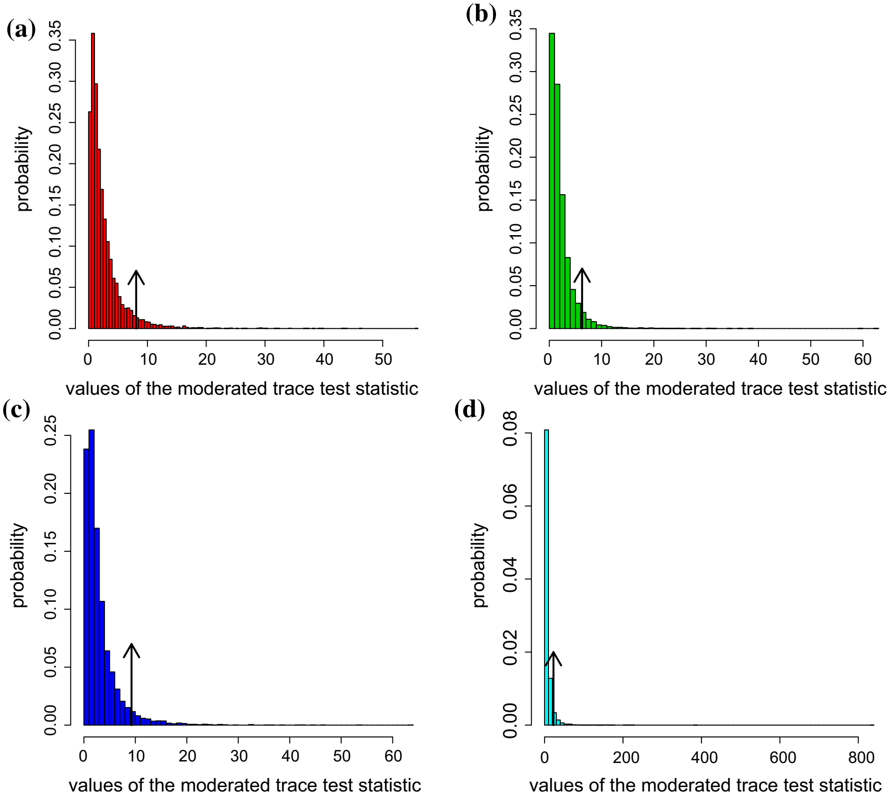


Fig. 1 Histogram of the null distribution of the moderated trace test when $p = 25$ for sample sizes **a** 5, **b** 10, **c** 15 and **d** 20. **a** $n = 5$. **b** $n = 10$. **c** $n = 15$. **d** $n = 20$

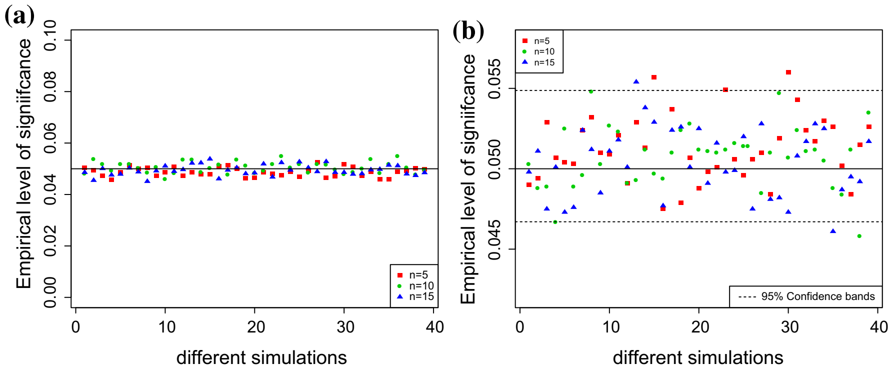


Fig. 2 Empirical level of the moderated test averaged over 10,000 simulations, three different sample sizes are considered and $p = 25$. **a** a. **b** b

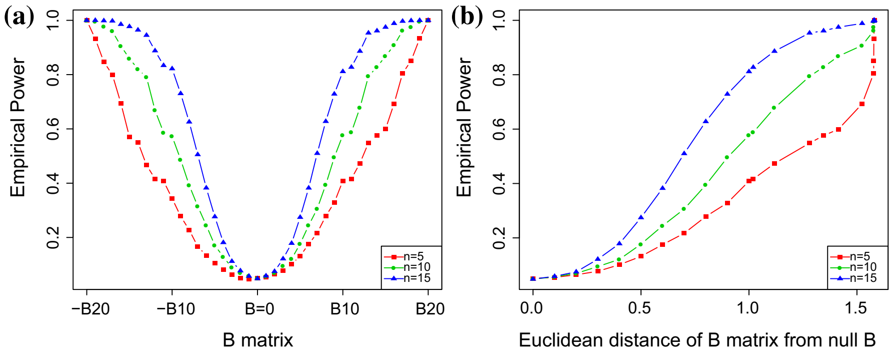


Fig. 3 Empirical power of the moderated trace test for $p = 25$ and different $\boldsymbol{B}$ across $n = 5, 10, 15$. **a** Using ordinal scale where each value on the x-axis represents a $\boldsymbol{B}$, **b** continuous scale with Euclidean distance of $\boldsymbol{B}$ from the null. **a** Ordinal scale. **b** Continuous scale

elements, and hence are plotted on the left of the null (Fig. 3a). The results show that our test statistic is symmetric with respect to the parameter $\boldsymbol{B}$, where the power of the test depends only on distance from the null hypothesis (does not depend on direction). Moreover, it can be seen, from Fig. 3 and Table 2, that the test is unbiased, where the power of the test is always greater than the level of the test. The results also indicate that the moderated test statistic is monotone with respect to the parameter matrix $\boldsymbol{B}$ (with respect to the Euclidean distance) as well as sample size n.

We also investigated power for different covariance matrices. We generated the variance covariance matrices from the Wishart distribution, where weak, moderate and strong correlations were considered. The results indicate that the power curve for weak correlation is uniformly higher than the power curves for moderate and strong correlations, and power for moderate correlations is slightly higher than that of strong correlations. This indicates that the power of the moderated trace test decreases for higher correlations.

Closer look at the results from our simulation revealed that the monotonicity seems to be disrupted as the sample size gets closer to p, where power values remain the same or at times decrease when we approach singularity (when $n \approx p$). We refer to this interval as “near singularity”, and in order to further investigate the behaviour of our test statistic in the near singularity zone, we performed simulations for fixed $\boldsymbol{B}$ values across many sample sizes. Figure 4, for instance, shows power curve for 3 different $\boldsymbol{B}$ values, $p = 40$ and across many sample sizes ranging from $n = 5$ to $n = 30$. Note that, here we used larger p value to allow consideration of more scenarios with different sample sizes. However, the results (conclusions) are consistent for other values of p. The results indeed confirmed that power starts to decrease in the near-singularity zone, when sample size approaches p. This is consistent for all the scenarios we considered (for different values of p, n, $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$). Further investigation revealed that the area, where power starts to decline, may be described as a function of the n/p ratio regardless of the correlation structure. However, more extensive simulations with special focus on the near-singularity zone is required to formulate and evaluate this hypothesis.

Table 2 Empirical power of the moderated trace test with respect to Euclidean distance of the alternative from null for different values of n and p

Euclidean distance	$p = 25$			$p = 40$		
	n = 5	n = 10	n = 15	n = 10	n = 20	n = 30
0	0.051	0.048	0.048	0.044	0.054	0.051
0.10	0.047	0.058	0.058	0.054	0.056	0.054
0.20	0.050	0.070	0.077	0.065	0.076	0.078
0.30	0.062	0.089	0.112	0.085	0.111	0.113
0.40	0.081	0.128	0.181	0.121	0.170	0.173
0.50	0.105	0.171	0.277	0.157	0.252	0.268
0.60	0.134	0.245	0.383	0.209	0.367	0.373
0.70	0.167	0.315	0.505	0.277	0.469	0.497
0.80	0.228	0.393	0.626	0.356	0.588	0.620
0.90	0.278	0.487	0.730	0.429	0.707	0.736
1.00	0.344	0.573	0.822	0.513	0.794	0.814
1.02	0.408	0.587	0.834	0.527	0.817	0.826
1.12	0.416	0.668	0.888	0.615	0.882	0.894
1.28	0.467	0.791	0.945	0.728	0.952	0.958
1.35	0.551	0.821	0.964	0.775	0.964	0.970
1.41	0.571	0.859	0.977	0.811	0.981	0.979
1.5	0.612	1.000	1.000	1.000	1.000	1.000

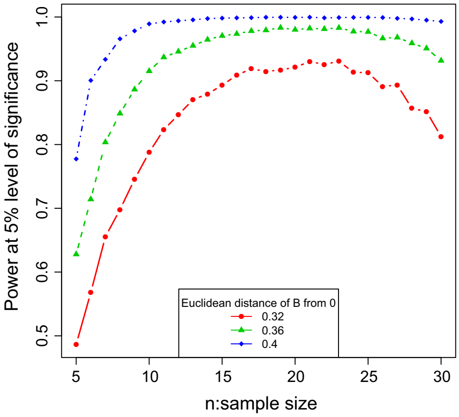


Fig. 4 Empirical power against increasing sample size across three different parameter matrices for $p = 40$ and a fixed covariance matrix, $\boldsymbol{\Sigma}$

In the simulations considered so far, we did not distinguish between the slope and intercept parameters of the growth curves. However, while creating the different $\boldsymbol{B}$ values for alternative hypotheses, during the investigation of power, we noticed that the behaviour of the power curve behaved slightly differently for values of slopes and intercepts. We, therefore, performed more simulations, where we fixed slopes to

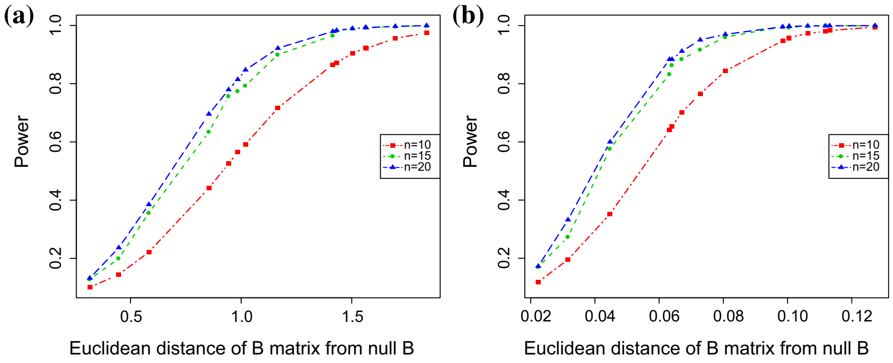


Fig. 5 Empirical power for **a** increasing intercept with fixed slope and **b** increasing slope with fixed intercept, for $p = 25$ and different values of n. **a** Intercept. **b** Slope

Table 3 Power of the moderated test for increasing intercept, with slope being fixed

Euclidean distance	$p = 25$			$p = 40$		
	n = 10	n = 15	n = 20	n = 10	n = 20	n = 30
0.85	0.442	0.634	0.696	0.388	0.652	0.678
0.94	0.526	0.757	0.780	0.467	0.734	0.765
1.02	0.592	0.794	0.847	0.522	0.814	0.831
1.17	0.716	0.900	0.923	0.653	0.907	0.910
1.41	0.866	0.965	0.980	0.809	0.980	0.978
1.50	0.905	0.989	0.990	0.853	0.988	0.989
1.56	0.923	0.989	0.993	0.888	0.992	0.994
1.70	0.956	0.997	0.997	0.923	0.997	0.997
1.84	0.974	0.998	1.000	0.958	0.999	0.999

investigate power for intercepts and vice versa. The results for these simulations are presented in Fig. 5a, b and Tables 3 and 4.

The results show that power curves have similar general properties for slopes and intercepts. For instance, the unbiasedness, symmetry as well as monotonicity with respect to n and Euclidean distance of $\boldsymbol{B}$, are satisfied in both. However, the results indicate that the moderated trace test is slightly more sensitive (more powerful) in detecting departure of the slope parameter from zero than that of the intercept. Nevertheless, the power of the moderated trace test for intercepts is also quite good, although slightly smaller than the power for the slopes. Although the power curves in Fig. 5a, b, placed side by side, look similar, it is important to observe the difference in range of the Euclidean distances presented on the x-axis. Initially the same Euclidean distance was considered for both the slope and intercept, but the resulting power values for the slope were all 1. As a results, smaller values of the slop.

Finally, we would like to highlight that we have considered higher values of p along with relatively smaller sample sizes in our simulations and the results are consistent,

Table 4 Power of the moderated test for increasing slope, with intercept being fixed

Euclidean distance	$p = 25$			$p = 40$		
	n = 10	n = 15	n = 20	n = 10	n = 20	n = 30
0.02	0.118	0.170	0.172	0.211	0.349	0.345
0.03	0.195	0.273	0.332	0.379	0.639	0.633
0.04	0.352	0.577	0.600	0.679	0.914	0.914
0.06	0.654	0.864	0.884	0.930	0.997	0.997
0.07	0.765	0.917	0.951	0.975	0.999	0.999
0.08	0.844	0.961	0.970	0.990	1.000	1.000
0.10	0.958	0.994	0.997	0.999	1.000	1.000
0.11	0.983	0.999	0.999	1.000	1.000	1.000
0.13	0.995	1.000	1.000	1.000	1.000	1.000

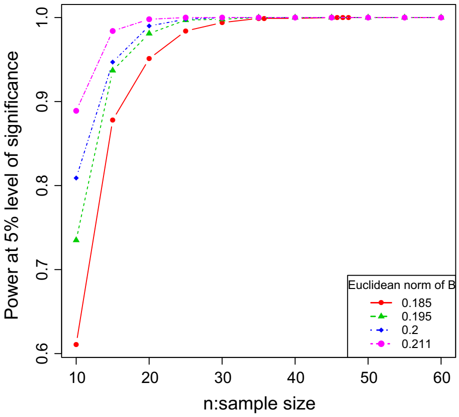


Fig. 6 Empirical power against increasing sample size across 4 different parameter matrices for $p = 100$

where the test maintains its good properties: monotone, unbiased and symmetric. The test has a level close to the nominal and has a reasonably good power for detecting a relatively small departure from the null hypothesis. The power curves for p=100 are presented in Fig. 6.

Simulation results for the MMLE

In this section, we illustrate the results of the empirical bias and MSE for the moderated maximum likelihood estimator (MMLE) provided in (9). The parameter matrices, $\boldsymbol{B}$ and t $\boldsymbol{\Sigma}$, considered in these simulations were the same as those considered in the previous section. We considered $p = 25$ and $p = 40$, and the sample sizes 5, 10, 15, 20, 25, 35. Since we are investigating performance in high-dimensional situations, we only considered scenarios where $n < p$. The results of the simulation are similar for the different values of p and different covariance structures. We, therefore, provide results for only $p = 25$ across multiple sample sizes. The bias and the MSE for the MMLE are presented in Figs. 7 and 8, respectively.

Our results show that bias is randomly distributed around zero within a very small range (0–0.015) of Euclidean distance. This indicates unbiasedness of the MMLE,

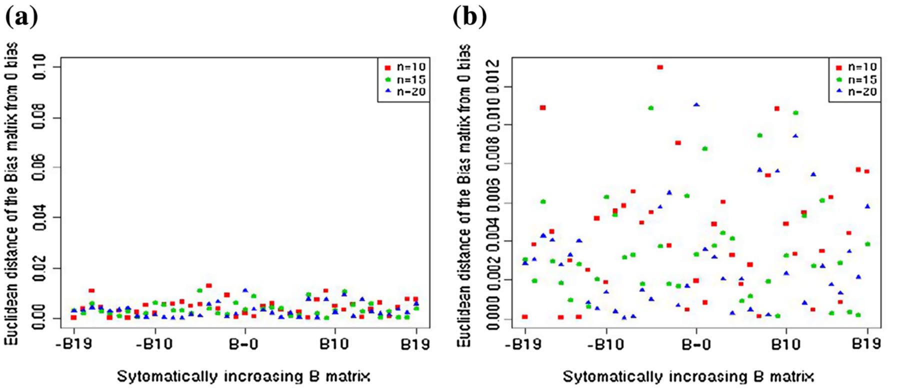
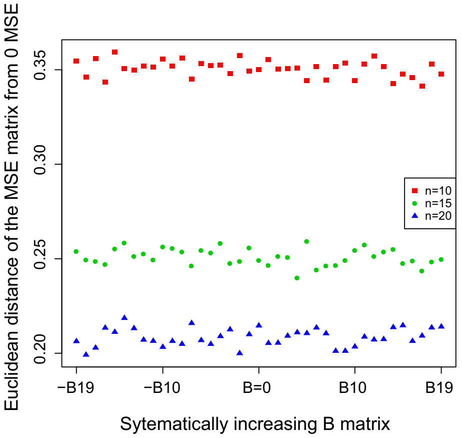


Fig. 7 Euclidean distance of the empirical Bias matrix for $p = 25$, many scenarios of $\boldsymbol{B}$ and three different sample sizes

Fig. 8 Euclidean distance of the empirical MSE matrix for $p = 25$, many scenarios of $\boldsymbol{B}$ and three different sample sizes



similar to what is previously established for the MLE in the non high-dimensional cases. When we magnify the bias figure (as indicated in Fig. 7b), we can see that there does not seem to be any systematic difference with the values of $\boldsymbol{B}$ or n. However, the rate of fluctuation (variability) of bias seems to decrease slightly with increase in sample size, but the difference is negligible. In Fig. 8, the Euclidean distance (from zero) of the MSE matrix for the MMLE is provided. Similar to bias, the results show that, for a fixed sample size, MSE is randomly distributed around zero and does not seem to be affected by the magnitude or direction of the true parameter $\boldsymbol{B}$. However, unlike bias, MSE decreases monotonically with increasing sample size, indicating that MMLE is a consistent estimator of $\boldsymbol{B}$, similar to what is previously established for the MLE in non high-dimensional cases.

5 Gene filtering using the moderated trace test

In gene expression microarry experiments, it has been indicated that the majority of the genes (95 %) are housekeeping (or noise) genes and only about 5 % of the genes are often relevant. For cross-sectional gene expression data, traditionally, variance or Inter quartile range (IQR) is used to filter noise genes, with the aim of identifying relevant genes and reducing false discovery rates due to multiplicity. Genes that have low variability (or IQR), below an adhoc cutoff value, are considered as noises and removed from further analysis. However, in time course microarray experiments, measurements are taken at different time points, consequently the observations are correlated among each other. As a result univariate measures such as variance or IQR may not be appropriate. Here, we use gene specific growth curve model to represent expression profiles for each gene and test the hypothesis, $H_0 : \boldsymbol{B} = \mathbf{0}$ vs $H_1 : \boldsymbol{B} \neq \mathbf{0}$ to identify noise genes. We used the moderated trace test proposed in the paper.

We use a publicly available data obtained from Gene Expression Omnibus (GEO) under a series name GSE5339. The objective of the original study was to investigate if exposure to chemicals, for individuals working in the petrochemical industry, is susceptible to bronchitis and eventually to lung cancer. It was already established in previous studies that exposure to V_2O_5 causes occupational bronchitis and occupational asthma among workers in the petrochemical industries (Levy 1984; Ingram 2003, 2007). The study was conducted to identify gene expression profiles among cultured lung fibroblasts exposed to V_2O_5 and compare their expression with healthy (normal) lung tissue. Data consists of 22,277 genes and gene expression measurements taken at five time points ($p = 5$: 1, 4, 8, 12, 24 h) from $n = 6$ individuals (3 in each group).

We assumed gene specific growth curve model (GCM) for all the 22,277 genes, where the GCM was fitted to each of the 22,277 genes separately. For each gene, we have an $n = 6$ by $p = 5$ matrix of longitudinal data representing the $p = 5$ repeated measurements taken from $n = 6$ (3 from each group) individuals. The within and between individual design matrices $\boldsymbol{Z}$ and $\boldsymbol{X}$ (same for all genes because of the nature of the design) are as follows:

$$\mathbf{Z}' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 4 & 8 & 12 & 24 \\ 1 & 16 & 64 & 144 & 576 \end{pmatrix} \text{ and } \mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Note that, the within individual design matrix $\boldsymbol{Z}$ above shows that we assumed quadratic curve to represent the gene specific growth profile. This choice was based on previous clinical as well as statistical knowledge (Hamid and Beyene 2009). Moreover, we also performed preliminary graphical investigation, using profile plots at gene level for some of the genes, to see how the gene expressions change over time. It is worth mentioning, though, that we also worked with linear GCM for this data, to be consistent with the simulations performed in the earlier sections. However, both the quadratic and linear models yielded similar results with respect to gene ranking and filtering indicating that the results are insensitive to model mis-specification.

We calculated the moderated trace test statistic for all the genes, and genes were ranked according to the value of the test statistic. Since we are performing tens of thousands of tests simultaneously, it is necessary to consider correction for multiple testing in calculating the required critical (cut-off) value. Since the purpose here is illustration of the proposed test statistic, we used the Bonferroni correction to adjust

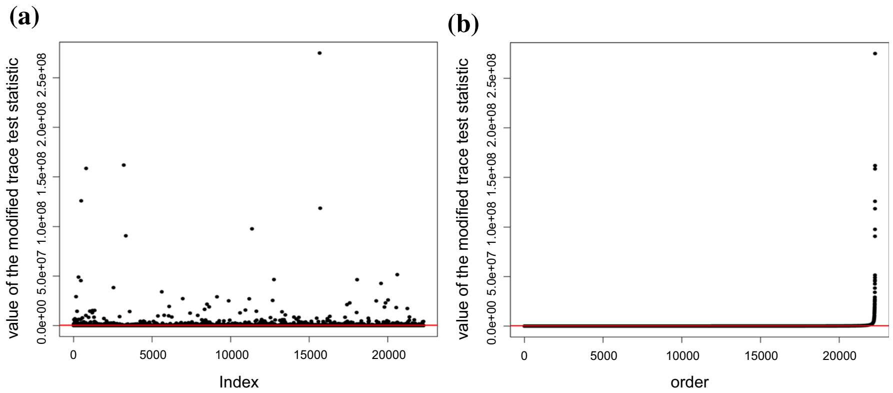


Fig. 9 Scatter plot of the test statistic value for all the genes from the real data, where the *red line* represents the Bonferroni corrected critical value **a** scatter plot, **b** ordered test statistic

the p values. However, this approach may be too conservative in practice. Since the purpose of gene filtering is mostly dimension reduction, researchers may want to relax the level so that they do not miss genes that may be potentially relevant. They may therefore use less conservative methods such as the false discovery rate (FDR). The scatter plot and ordered scatter plot of the test statistic, with the Bonferroni cutoff value, are provided in Fig. 9.

As we can see from the scatter plot in Fig. 9, most of the genes have similar test statistic values with only a small proportion of the genes standing out of the bulk of the data. This can be seen clearly in the plot of the ordered test statistic where the top 7 genes can be seen to have a value which is much larger than the rest of the genes. We were able to identify 1053 (4.73 %) significant genes using the moderated trace test, indicating that the majority of the genes were noises. This is in agreement with biological and medical literature indicating that, in genomic experiments where tens of thousands of genes are assayed at once, only a small fraction of the genes (often less than 5 %) are expressed and hence relevant to a given study. We annotated the top 7 ranked genes and the results of the annotation are as presented in Table 5. The annotation is done using the gene annotation web site, GeneAnnot (Chalifa-Caspi 2004; Ferrari 2007).

6 Discussion

In this study, we proposed a moderated test statistic for GMANOVA models that is especially useful for analysing high-dimensional longitudinal data, i.e., when the sample size n is smaller or equal to the number of time points p. The test is particularly useful in the analysis of time course genomic experiments. We considered an existing trace test proposed recently by Hamid et al. (2011) and modified the test using the Moore–Penrose generalized inverse of the sample variance-covariance matrix to overcome the singularity problem caused by the high-dimensional nature of the data.

Table 5 Description of the top 7 ranked genes, ordered according to their ranks

Gene code/Name	Description	Location	Comments	References
RPS12	Ribosomal protein S12	Chromosome 6	Differentially expressed in lung squamous cell carcinomas	Liu (2007)
PSMB7	Proteasome (prosome, macropain) subunit, beta type, 7	Chromosome 9	Differentially expressed in small-cell lung cancer cells	Yan, (2012)
FZD5	Frizzled family receptors	Chromosome 2	Highly expressed gene in Esophageal and other cancers	Kimchi (2005)
MRPS14	Mitochondrial ribosomal protein S14	Chromosome 1	Up-regulated in lung cancers	Jia (2011)
RPLP0	Ribosomal protein, large P0	Chromosome 12	Expressed in non small cell lung cancer cell lines	Liu et al. (2005)
Z21967	Partial cDNA for homologue of mPOU homeobox protein	Chromosome 12	Expressed in lung, heart skeletal muscle and brain	Wey et al. (1994)
SDCBP	Syndecan binding protein (Syntenin)	Chromosome 8	Expressed in gastric, colon and breast carcinomas	Koo (2002)

Empirical investigation using extensive simulations show that our proposed test statistic performs very well, with the test yielding high power in detecting small differences. The level of the test is always close to the nominal level α, where estimates of the level are randomly distributed within a very tight interval around the nominal level. Furthermore, the results indicate that the moderated trace test has many desirable properties. The power of the test is always greater than the level of the test indicating unbiasedness. The test is also symmetric with respect to the matrix parameter $\boldsymbol{B}$, i.e., the test was able to detect both positive and negative deviations from the null hypotheses with equal power. It is also shown to be monotone with respect to both sample size and the true value of the parameter in the alternative hypothesis, the latter monotonicity is established with respect to the Euclidean distance.

However, we would like to note that the moderated trace test does not perform very well in the near singularity zone, when the number of time points p is close to n. In fact, we performed further investigations and showed that the power initially increases with n (where $n << p$) and starts to decrease as the sample size approaches p ($n \approx p$), and recovers back again when $n > p$, in the non-high-dimensional situations. We observed that power drops in the near singularity zone regardless of $p, n, \boldsymbol{B}$ and $\boldsymbol{\Sigma}$ and the results show that the point, where this drop in power occurs, might be a function of the n/p ratio, the parameter matrix, $\boldsymbol{B}$ and the covariance matrix $\boldsymbol{\Sigma}$. For instance, the drop point starts early for strong correlations compared to weak correlations. The general optimal properties of the moderated test (such as unbiasedness, monotonicity with respect to $\boldsymbol{B}$ and symmetry) are maintained in the near singularity zone. The level of the test is also close to the nominal value. Further investigation is needed to understand the property of the test in the near singularity zone and to see if something is possible to improve power. We also hope to be able to define the near singularity zone as a function of the correlation matrix, the parameter matrix, $\boldsymbol{B}$ and the n/p ratio to be able to make sample size recommendations that may be useful in practical applications.

Although the focus here is on testing hypothesis in high-dimensional longitudinal problems using the GCM, we also considered the maximum likelihood estimates (MLE) for the GCM and provided a moderated estimator, MMLE, for the matrix parameter $\boldsymbol{B}$, by using the Moore–Penrose generalized inverse. Extensive simulations demonstrate that the Moore–Penrose generalized inverse is a very good alternative, when the sample covariance matrix is singular, where the moderated estimator using the Moore–Penrose inverse is shown to be empirically unbiased and consistent estimator for the matrix parameter $\boldsymbol{B}$. However, similar to the power analysis, an increase in MSE is observed in the near singularity zone. Further investigation revealed that bias remained low for any p and n, however, precision of the estimators increased near singularity zone leading to an increase in MSE. Another improved approach is needed to improve performance when $p \approx n$.

We would like to mention that we have also considered Stein's shrinkage estimator, where we considered several shrinkage parameters and target covariance matrices recommended by previous literature. However, the results from our simulation were unstable, where power fluctuated highly. This was true for all the targets considered.

Although we considered the hypotheses $H_0 : \boldsymbol{B} = \boldsymbol{0}\ vs\ H_1 : \boldsymbol{B} \neq 0$, our approach can be extended to include the general linear hypothesis ($H_0 : \mathbf{GBF} = 0$ $vs\ H_1 : \mathbf{GBF} \neq 0$, where $\boldsymbol{G}$ and $\boldsymbol{F}$ are matrices of zeroes and ones). In this work, we were interested in identifying noise genes that do not express in situations where a time-course gene expression measurements are taken from individuals in one or more groups. After removing noise genes and reducing the dimension of data, researchers may be interested in identifying genes that have expression values that change over time (slope different from zero) or identify differentially expressed genes (genes with expression profiles that are different for different groups). Such hypotheses can be formally expressed using the general linear hypothesis and a moderated test statistic can be defined using a moderated version of the more general trace test. Work in these directions is currently in progress and we hope to report these findings in a future paper.

An important limitation of using the Moore–Penrose inverse is that the null distribution is free of the unknown covariance matrix $\boldsymbol{\Sigma}$, provided that $\boldsymbol{S}$ is $\boldsymbol{\Sigma}$ invariant, a property which is difficult to show in practical applications. In high-dimensional situations where this property is not satisfied, re-sampling approaches such as bootstrapping might be a simple alternative. However, this may be computationally intensive in studies requiring identification of differentially expressed genes, where statistical tests on tens to hundreds of thousands of genes are make simultaneously. For gene ranking and filtering, however, the computational requirements in implementing our method in practice is reasonably small. In our simulations, for instance, where we considered more than 500 scenarios to establish properties of the test and evaluate performance. These analyses took more than an estimated 15 h. However, analysis of a single high-dimensional data took under 2 min (107.3 s); and this is what is required in real data analysis. For the genetic data we provided as illustration in our paper, generating the null distribution and calculating the critical value from a common variance-covariance matrix took a little longer than 1 min (66 s) and an additional 16.6 s were required to calculate the observed test statistic form all the 22,777 genes. Ranking the genes, therefore, required under 2 min, indicating that gene filtering using our method requires a very small computational time. We used the R statistical package and a computer with Intel Core$^{TM}i3$ CPU^{M}350@2.27 GHZ RAM 3GB. Nevertheless, we would like to highlight that calculating the empirical p values based using re-sampling and determining statistical significance for such genomic data, involving tens of thousands of genes, will require a much longer computational time. For instance, for the lung cancer data we used (22,777 genes), it will require us an additional 107 h (17 s for each gene). Our approach is, therefore, recommended for gene ranking and selection based on ordering of genes according to the test statistics. Further research is needed to alleviate the financial burden required in identifying differentially expressed genes as well as improve performance in the near singularity zone.

References

Casella G, Berger RL (2002) Statistical inference, 2nd edn. Thompson Learning, Duxbury Press, Belmont, CA

Chalifa-Caspi V et al (2004) GeneAnnot: comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes. Bioinformatics 20(9):1457–1458

Chen SX, Qin YL (2010) A two-sample test for high-dimensional data with applications to gene-set testing. Ann Stat 38(2):808–835

Chen LS, Paul D, Prentice RL, Wang P (2012) A regularized Hotellings T2 test for pathway analysis in proteomic studies. J Am Stat Assoc 106(496):1345–1360

Ferrari F et al (2007) Novel definition files for human GeneChips based on GeneAnnot. BMC Bioinform 8:446–451

Hamid JS, Beyene J (2009) A multivariate growth curve model for ranking genes in replicated time-course microarray data. Stat Appl Genet Mol Biol 8(1):1–26

Hamid JS, Beyene J, Rosen DV (2011) A novel trace test for the mean parameters in a multivariate growth curve model. J Multivar Anal 102(2):238–251

Horn RA, Johnson CR (1985) Matrix analysis. Cambridge University Press, Cambridge

Ingram JL et al (2003) Vanadium-induced HB-EGF expression in human lung fibroblasts is oxidant-dependent and requires MAP kinases. Am J Physiol Lung Cell Mol Physiol 284(5):774–782

Ingram JL et al (2007) Genomic analysis of human lung fibroblasts exposed to vanadium pentoxide to identify candidate genes for occupational bronchitis. Respir Res 8(34.10):1186–1198

Jia D et al (2011) Genome-wide copy number analyses identified novel cancer genes in hepatocellular carcinoma. Hepatology 54(4):1227–1236

Jonhstone IM, Titterington DM (2009) Statistical challenges of high-dimensional data. Philos Trans R Soc Lond A Math Phys Eng Sci 367(1906):4237–4253

Khatri CG (1966) A note on a MANOVA model applied to problems in growth curve. Ann Inst Stat Math 18(1):75–86

Kimchi ET et al (2005) Progression of Barrett's metaplasia to adenocarcinoma is associated with the suppression of the transcriptional programs of epidermal differentiation. Cancer Res 65(8):3146–3154

Kollo T, von Rosen D (2005) Advanced multivariate statistics with matrices. Springer, The Netherlands

Koo TH (2002) Syntenin is overexpressed and promotes cell migration in metastatic human breast and gastric cancer cell lines. Oncogene 21(26):4080–4088

Läuter J (2009) High-dimensional data analysis: selection of variables, data compression and graphics—application to gene expression. Biometr J 51(2):235–251

Levy BS et al (1984) Boilermakers' bronchitis: respiratory tract irritation associated with vanadium pentoxide exposure during oil-to-coal conversion of a power plant. J Occup Environ Med 26(6):567–570

Liu DW, Chen ST, Liu HP (2005) Choice of endogenous control for gene expression in non small cell lung cancer. Eur Respir J 20:1002–1008

Liu Y et al (2007) Identification of genes differentailly expressed in human primary lung squamous cell carcinoma. Lung Cancer 56(3):307–317

Lönnstedt I, Speed T (2002) Replicated microarray data. Stat Sin 12(1):31–46

Ma P et al (2006) A data-driven clustering method for time course gene expression data. Nucleic Acids Res 34(4):1261–1269

Moore EH (1920) On the reciprocal of the general algebraic matrix. Bull Am Math Soc 26:294–295

Pan JX, Fang KT (2002) Growth curve models and statistical diagnostics. Springer, New York

Penrose R (1955) A generalized inverse for matrices. Proc Camb Philos Soc 51(3):406–413

Potthoff RF, Roy SN (1964) A generalized multivariate analysis of variance model useful especially for growth curve model. Biometrika 51(3–4):313–326

Rao CR, Mitra SK (1972) Generalized inverse of a matrix and its applications. In: Proceedings of the 6th Berkeley symposium on mathematical statistics and probability, vol 1, pp 601–620

Smyth GK (2004) Linear models and empirical Bayes methods for assessing diferential expression in microarray experiments. Stat Appl Genet Mol Biol 3(1):1–25

Tai YC, Speed TP (2006) A multivariate empirical Bayes statistic for replicated microarray time course data. Ann Stat 34(5):2387–2412

Tai YC, Speed TP (2009) On gene ranking using replicated microarray time course data. Biometrics 65(1):40–51

von Rosen D (1989) Maximum likelihood estimators in multivariate linear normal models. J Multivar Anal 31(2):187–200

Wey E, Lyons GE, Schafer BW (1994) A human POU domain gene, mPOU, is expressed in developing brain and specific adult tissues. Eur J Biochem 220(3):753–762

Yan X et al (2012) External Qi of Yan Xin Qigong induces cell death and gene expression alterations promoting apoptosis and inhibiting proliferation, migration and glucose metabolism in small-cell lung cancer cells. Mol Cell Biochem 363(1–2):245–255

Yuan M, Kendziorksi C (2006) Hidden Markov models for microarray time course data in multiple biological conditions. J Am Stat Assoc 101(476):1323–1332