



Profile Analysis in High Dimensions

Cigdem Cengiz^{1,2} · Dietrich von Rosen^{1,2}  · Martin Singull²

Accepted: 2 December 2020 / Published online: 22 December 2020
© The Author(s) 2020

Abstract

The three tests in profile analysis: test of parallelism, test of level and test of flatness are modified so that high-dimensional data can be analysed. Using specific scores, dimension reduction is performed and the exact null distributions are derived for the three hypotheses.

Keywords High-dimensional data · Hypothesis testing · Linear scores · Multivariate analysis · Profile analysis · Spherical distributions

1 Introduction

In this article, we are going to consider profile analysis from a high-dimensional perspective, i.e. there are so many parameters in the model that there are not enough degrees of freedom to test those hypotheses which are part of the analysis. Profile analysis consists of three tests, and the tests are carried out in a certain order. The tests are: (1) the test of parallelism; (2) the test of equal level; and (3) the test of flatness. Later we will be more specific how the tests are performed. The approach differs somewhat from the usual likelihood ratio testing procedure since, in particular, the hypotheses are chosen in a certain order.

There are two possible distinct scenarios in the analysis of profiles which both imply a need to extend the classical theory to cover high-dimensional models:

This article is part of the topical collection “Celebrating the Centenary of Professor C. R. Rao” guest edited by, Ravi Khattree, Sreenivasa Rao Jammalamadaka, and M. B. Rao.

✉ Dietrich von Rosen
Dietrich.von.Rosen@slu.se

¹ Department of Energy and Technology, Swedish University of Agricultural Sciences, SE-750 07 Uppsala, Sweden

² Department of Mathematics, Linköping University, SE-581 83 Linköping, Sweden

1. The same variable is observed for each subject over many time points (repeated measurements) which often are tightly distributed along a finite interval. There can be more repeated measurements than the number of independent subjects.
2. For each subject many variables should be analysed simultaneously. There can be more variables than the number of independent observations.

In (1) there is a natural ordering of the repeated measurements. For example, it can be a growth curve or in general any stream of observations which is generated from some measurement device. In (2) there does not exist any ordering between variables. Instead one can measure hundreds of characteristics on a subject, for example components from a vehicle.

Over the years profile analysis has been studied by many authors. One of the first contributions was published by Greenhouse and Geisser [4]. Originally, mean values were compared and modelled. Many years after profile analysis had been introduced, Srivastava [17] derived likelihood ratio-based test statistics together with their distributions. Illuminating chapters on profile analysis can be found in the books by Srivastava and Carter [19] and Srivastava [18]. A more sophisticated approach was suggested by Ohlson and Srivastava [13] who considered profile analysis of several groups, where the groups of subjects only partly had a common profile. In von Rosen [22] an overview of classical profile analysis has been given.

Fujikoshi [3], followed by Seo et al. [16], derived likelihood ratio tests for the parallelism, level and flatness hypothesis, respectively, when analysing growth curve data. For a parallel profile model, i.e. assuming parallel profiles, it has also been proposed to consider different covariance structures. In Yokoyama and Fujikoshi [25] and Yokoyama [24] the random-effect covariance structure was considered and some tests for the random-effect and flatness were derived. Later, Srivastava and Singull [20] constructed likelihood ratio tests in profile analysis, without any restrictions on the parameter space, for testing the covariance matrix for random-effect structure or sphericity.

Moreover, profile analysis has also been discussed under more general models. Okamoto et al. [14] studied the asymptotic expansions of the distributions of some test statistics considering elliptical distributions. Others have extended this model and discuss the asymptotic expansions for the null distribution of test statistics for profile analysis under non-normality, e.g. see Maruyama [12] and Harrar and Xu [6].

Our ideas about analysing high-dimensional profile data stem from works by Läuter [8, 9] and Läuter et al. [10, 11] where random scores were utilized in MANOVA models. Scores have for a long time been used in statistics and mostly they constitute of known linear combinations of random vectors/matrices. The idea with the random scores is that they should be applied to test statistics which to some extent are robust, i.e. instead of having test statistics which are based on normally distributed vectors they can be elliptically distributed and still the test statistics follow the same distribution as when the observed variables are normally distributed.

In Sect. 2 profile analysis is introduced and necessary background information for the rest of the paper is presented. Thereafter, in Sect. 3 the high-dimensional

approach of this paper is described and in Sect. 4 the usual test statistics are modified so that high-dimensional data can be handled. Section 5 comprises some concluding remarks.

Concerning notation bold upper case letters denote matrices and bold lower case vectors. Other notations are defined when they first appear.

2 Profile Analysis

Assume that there are q groups which should be compared with n_i p -dimensional random vectors \mathbf{x}_{ij} , $j \in \{1, \dots, n_i\}$, which are independently normally distributed as $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i \in \{1, \dots, q\}$, where $\boldsymbol{\mu}_i = (\mu_{1,i}, \dots, \mu_{p,i})'$ and $\boldsymbol{\Sigma}$ is an unknown positive definite dispersion matrix. As mentioned in the introduction there are three different hypotheses which are commonly considered in profile analysis:

1. *Parallelism hypothesis*

$H_1 : \boldsymbol{\mu}_i - \boldsymbol{\mu}_q = \gamma_i \mathbf{1}_p, i \in \{1, \dots, q - 1\}$ and $A_1 \neq H_1$, meaning that $\boldsymbol{\mu}_i - \boldsymbol{\mu}_q \neq \gamma_i \mathbf{1}_p$, $i \in \{1, \dots, q - 1\}$, where A_1 stands for alternative hypothesis, the parameters γ_i are unknown scalars and $\mathbf{1}_p$ is a p -dimensional vector of ones;

2. *Level hypothesis*

$H_2|H_1 : \gamma_i = 0, i \in \{1, \dots, q - 1\}$ and $A_2 \neq H_2|H_1$, implying that $\gamma_i \neq 0$, $i \in \{1, \dots, q - 1\}$, where $H_2|H_1$ means H_2 under the assumption that H_1 is true;

3. *Flatness hypothesis*

$H_3|H_1 : \boldsymbol{\mu}_i = \psi_i \mathbf{1}_p, i \in \{1, \dots, q\}$ and $A_3 \neq H_3|H_1$, where $H_3|H_1$ means H_3 under the assumption that H_1 is true, and the parameters ψ_i are unknown scalars.

One can note that instead of $H_3|H_1$ the strategy can be to test $H_3|H_2$, when we have $\psi_1 = \dots = \psi_q$. In this way, profile analysis can be built up around a chain of tests.

Profile analysis can also be reformulated with the help of matrices and the MANOVA and growth curve model (GMANOVA) as well as the extended growth curve model which all belong to the class of bilinear models (see von Rosen [23]). Moreover, for technical details we refer to the report by Cengiz and von Rosen [1].

Let the observation matrix be matrix normally distributed, i.e. $\mathbf{X} \sim N_{p,N}(\mathbf{BC}, \boldsymbol{\Sigma}, \mathbf{I})$, $\mathbf{B}: p \times q$ and $\boldsymbol{\Sigma}: p \times p$, consist of the unknown parameters and $\mathbf{C}: q \times N$, is the design matrix describing the q groups. In this article, to simplify presentation, \mathbf{C} is supposed to be of full rank. It can be noted that $\mathbf{B} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_q)$ and one choice of \mathbf{C} is

$$\mathbf{C} = \begin{pmatrix} \mathbf{1}'_{n_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}'_{n_2} & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1}'_{n_q} \end{pmatrix}.$$

The null hypothesis and the alternative hypothesis for parallelism can be written

$$\begin{aligned}
 H_1 : E[X] &= BC, \quad FBG = \mathbf{0}, \\
 A_1 : E[X] &= BC, \quad \text{no restrictions on } B,
 \end{aligned}
 \tag{1}$$

where F and G are contrast matrices given by

$$F = \begin{pmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix}, \quad G = \begin{pmatrix} 1 & 0 & \dots & 0 \\ -1 & 1 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & -1 \end{pmatrix}.
 \tag{2}$$

It can be noted that F is of size $(p - 1) \times p$ and G is of size $q \times (q - 1)$, respectively, of ranks $p - 1$ and $q - 1$.

If $N - 2$ is larger than p , the likelihood ratio statistic for the parallelism hypothesis can be given as

$$\lambda_p = \frac{|FSF'|}{|FSF' + FXP_{C'(CC')^{-1}G}X'F'|},
 \tag{3}$$

where $|\cdot|$ stands for the determinant, the projection $P_M = M(M'M)^{-1}M'$, for any matrix expression M of full rank and $S = X(I - P_C)X'$. The notation of a projection will frequently be used in this article. Let $r(M)$ denote the rank of a matrix M . Moreover,

$$\begin{aligned}
 FSF' &\sim W_{p-1}(F\Sigma F', N - r(C)), \\
 FXP_{C'(CC')^{-1}G}X'F' &\sim W_{p-1}(F\Sigma F', r(G)),
 \end{aligned}$$

are two expressions which are independently distributed, where $W_p(\Psi, n)$ denotes the Wishart distribution with scale parameter Ψ and n degrees of freedom. Note that $r(C) = q$ and $r(G) = q - 1$. If $p = 1$ then a Wishart variable is proportional to a chi-square variable. Furthermore, if $U \sim W_p(\Sigma, n)$ is independent of $V \sim W_p(\Sigma, m)$ then

$$\frac{|U|}{|U + V|} \sim \Lambda(p, m, n),$$

which is known as Wilk’s lambda distribution. Hence, the distribution for the likelihood ratio statistic given in (3) is

$$\lambda_p \sim \Lambda(p - 1, r(G), N - r(C)).$$

If the profiles are parallel, we can say that there is no interaction between the responses and the groups. Given that the parallelism hypothesis holds, the next step is to proceed with testing the second hypothesis, H_2 , which indicates that there is no group effect. Moreover, if the first hypothesis holds, one may also want to test the third hypothesis, H_3 , meaning that the response is constant “over time”. Note that failing to reject H_1 , as always, does not mean that the hypothesis is true but in profile analysis it is used as a strategy for analysing data.

The level hypothesis in matrix form equals

$$\begin{aligned} H_2|H_1 : E[X] &= BC, & BG &= \mathbf{0}, \\ A_2|H_1 : E[X] &= BC, & FBG &= \mathbf{0}, \end{aligned} \tag{4}$$

where F and G are defined in (2). Let M° denote any matrix of full rank generating the orthogonal complement to the column space of M . The corresponding likelihood ratio test statistic for the level hypothesis can be expressed as

$$\lambda_L = \frac{((F')^\circ)'S^{-1}(F')^\circ)^{-1}}{((F')^\circ)'S^{-1}(F')^\circ)^{-1} + HXC'(CC')^{-1}GQ^{-1}G'(CC')^{-1}CX'H'}, \tag{5}$$

where

$$\begin{aligned} H &= ((F')^\circ)'S^{-1}(F')^\circ)^{-1}(F')^\circ)'S^{-1}, \\ Q &= G'(CC')^{-1}G + G'(CC')^{-1}CX'F'(FSF')^{-1}FXC'(CC')^{-1}G. \end{aligned}$$

Note that $r((F')^\circ) = 1$ since $r(F) = p - 1$. Then, it can be shown that

$$\begin{aligned} ((F')^\circ)'S^{-1}(F')^\circ)^{-1} &\sim W_1(((F')^\circ)' \Sigma^{-1} (F')^\circ)^{-1}, N - r(C) - p + 1), \\ HXC'(CC')^{-1}GQ^{-1}G'(CC')^{-1}CX'H' &\sim W_1(((F')^\circ)' \Sigma^{-1} (F')^\circ)^{-1}, r(G)), \end{aligned}$$

which are independently distributed. Moreover, the Wilk’s lambda distribution for (5) is given by

$$\lambda_L \sim \Lambda(1, r(G), N - r(C) - p + 1)$$

which equals a beta-distribution, with $(N - r(C) - p + 1)/2$ and $r(G)/2$ degrees of freedom. Furthermore, by a one-to-one transformation λ_L can be converted into an F -statistic.

Assuming that the profiles are parallel, a test can be constructed to see if the profiles are flat, i.e.

$$\begin{aligned} H_3|H_1 : E[X] &= BC, & FB &= \mathbf{0}, \\ A_3|H_1 : E(X) &= BC, & FBG &= \mathbf{0}, \end{aligned} \tag{6}$$

where F and G are defined in (2). The likelihood ratio statistic for the flatness hypothesis equals

$$\lambda_F = \frac{|FSF' + FXP_{C'(CC')^{-1}G}X'F'|}{|FSF' + FXP_{C'(CC')^{-1}G}X'F' + FXP_{C'G}X'F'|}, \tag{7}$$

where there are two independently distributed Wishart matrices:

$$\begin{aligned} FXP_{C'G}X'F' &\sim W_{p-1}(F \Sigma F', r(C'G^\circ)), \\ FSF' + FXP_{C'(CC')^{-1}G}X'F' &\sim W_{p-1}(F \Sigma F', N - r(C) + r(G)). \end{aligned}$$

Thus,

$$\lambda_F \sim \Lambda(p-1, r(\mathbf{C}'\mathbf{G}^\circ), N - r(\mathbf{C}) + r(\mathbf{G}))$$

and $r(\mathbf{C}'\mathbf{G}^\circ) = r(\mathbf{C}) - r(\mathbf{G}) = 1$.

To find an exact significance level for the level hypothesis test and the flatness hypothesis test is still an open question. The main reason for this fact is that it is a difficult problem to solve since the different “conditional” test statistics are dependent. If the significance level is important, a Bonferroni approach can be used. Since the focus in this article is to construct test statistics in “high dimensions”, we leave it to the future to compare the classical profile analysis approach with an approach based on “unconditional tests”.

3 High-Dimensional Setting

The focus in this article is on high-dimensional profile analysis. Several authors have approached the analysis of high-dimensional profiles. Onozawa et al. [15] and Harrar and Kong [5] (see also Hyodo [7]) derived test statistics for high-dimensional profile analysis with unequal covariance matrices. Takahashi and Shutoh [21] proposed new test statistics in profile analysis with high-dimensional data by applying the Cauchy–Schwarz inequality. The above-mentioned authors introduce different high-dimensional asymptotic frameworks and derive the test statistics in profile analysis under these frameworks. The approach in this article is different since we will not focus on the asymptotic distributions of the test statistics. Instead a fixed p (number of repeated measurements) and n (number of observations) are of interest with a p which can be much larger than n .

The method adopted in this article is mainly based on ideas put forward by Läuter [8, 9] and Läuter et al. [10, 11] who proposed a scoring method for dealing with high-dimensional problems in MANOVA. The method is more advanced than principal component analysis and tests based on these scores are exact. However, for the level test in profile analysis this article presents a completely new approach. The tests which arise from Läuter’s [8, 9] and Läuter et al.’s [10, 11] approach are based on linear scores which are constructed with the help of sums of products matrices. These scores are linear combinations of the repeated measurements. The approach implies that high-dimensional observations are compressed into low-dimensional observations and then these are used in the analysis instead of the original data. Note that we are very briefly mentioning the choice of scores and only one explicit expression of the scores is given in this work. However, there exist different kinds of scores and for details it can be referred to Läuter et al. [10] where several examples are presented.

Now it is started with a brief introductory mathematical presentation of the theory. Suppose

$$\mathbf{X} \sim N_{p,n}(\boldsymbol{\mu}\mathbf{1}'_n, \boldsymbol{\Sigma}, \mathbf{I}_n)$$

and consider a single score

$$z' = (z_1, z_2, \dots, z_n) = (d_1, d_2, \dots, d_p)X = d'X,$$

where d is the vector of weights and z_j 's, $j \in \{1, \dots, n\}$, are the individual scores. Suppose that the null hypothesis $\mu = \mathbf{0}$ is of interest. In this case one can choose the vector d to be a unique function of XX' which is the total sums of product matrix of size $p \times p$. Then, with $\bar{z} = \frac{1}{n}z' \mathbf{1}_n$ and $s_z^2 = \frac{1}{n-1}(z'z - n\bar{z}^2)$,

$$t = \frac{\sqrt{n}\bar{z}}{s_z}, \tag{8}$$

is t -distributed with $n - 1$ degrees of freedom. Note that the vector of random scores z is not normally distributed. The result on this type of ‘‘robustness’’ follows from a general result stated in the next lemma.

Lemma 3.1 (Fang and Zhang [2], Theorem 2.5.8). *Let Γz and z have the same distribution for all orthogonal matrices Γ and suppose that $P(z = \mathbf{0}) = 0$ which defines a class of spherical distributions, say Φ_+ . The distribution of a statistic $t(z)$, where the distribution of z belongs to Φ_+ , is the same for each member of Φ_+ , if for all $\alpha > 0$ the statistic $t(\alpha z)$ has the same distribution as $t(z)$.*

The lemma is useful if the distribution of the statistic mentioned in Lemma 3.1 can be derived for one member of Φ_+ . In particular if it holds for $u \sim N_n(\mathbf{0}, I)$ because u belongs to Φ_+ and then the distribution has been obtained for all members of the class Φ_+ . These facts, among others, establish why (8) is true for all spherical distributions.

Wilk’s Λ statistic is frequently used in this article and Theorem 1 in Läuter et al. [11] implies the following theorem:

Theorem 3.1 *Let $V \sim W_p(\Sigma, m)$ and $W \sim W_p(\Sigma, n)$ be independently Wishart distributed and put*

$$\lambda = \frac{|D'WD|}{|D'(W + V)D|},$$

where $D: s \times p$, $s \leq p$ is a function of $W + V$ and the rank of $D'(W + V)D$ equals s with probability 1. Then $\lambda \sim \Lambda(s, m, n)$.

In the theorem there is Σ involved in $V \sim W_p(\Sigma, m)$ and $W \sim W_p(\Sigma, n)$ but the distribution of λ is the same for all Σ . It is only crucial that the same Σ is included in the distributions for W and V . One way of constructing the weights D is to use the so-called principal component method (see Läuter et al. [11]), where the weights are determined by solving the eigenvalue problem

$$(W + V)D = D\Psi, \quad D'D = I, \tag{9}$$

and Ψ is a diagonal matrix with the positive eigenvalues.

The next corollary of the theorem is what is needed in this article.

Corollary 3.1 Let $V \sim W_p(\Sigma, m)$ and $W \sim W_p(\Sigma, n)$ be independently Wishart distributed and

$$\beta = \frac{d'Vd}{d'Wd + d'Vd},$$

where $d \in \mathcal{R}^p$ is a function of $W + V$ such that $d'(W + d'V)d > 0$ with probability 1. Then β is β -distributed with $m/2$ and $n/2$ degrees of freedom.

Proof By definition of a Wishart distribution there exist $Y \sim N_{p,m}(\mathbf{0}, \Sigma, I_m)$ and $Z \sim N_{p,n}(\mathbf{0}, \Sigma, I_n)$ such that $V = YY'$ and $W = ZZ'$. Put $X = (Y : Z)$ and then $V + W = XX' = YY' + ZZ'$. Since $d = g(V + W)$ for some function $g(\bullet)$, $d'X = g'(XX')X$. Moreover, for any orthogonal matrix Γ , $g(XX') = g(X\Gamma\Gamma'X')$ and since $X\Gamma$ has the same distribution as X , the score $d'X = g'(XX')X$ is spherically distributed. Note that β can be written

$$\beta = \frac{d'X(I : \mathbf{0})'(I : \mathbf{0})X'd}{d'XX'd}$$

and it follows from Lemma 3.1 that since the statement instead of $d'X$ also is true for a normally distributed variable with a dispersion matrix equal to I , β is indeed β -distributed with $m/2$ and $n/2$ degrees of freedom. □

Corresponding to (9) one alternative way to determine the weight d in Corollary 3.1 is by solving the eigenvalue/eigenvector problem

$$(W + V)d = \psi d, \quad d'd = 1, \quad \psi > 0.$$

4 Main Results

In Sect. 2 the three hypothesis in classical profile analysis was presented, i.e. when $N > p + q$. Now it will be focused on the high-dimensional setting when $p > N - q$. Consider, for example, the test statistic given in (3). When p is large the problem is that S is singular and the determinant in (3) equals 0.

Läuter [8, 9] and Läuter et al. [10, 11] directly applied a vector d to the observation matrix X . In this article, since there is a bilinear testing situation, it is proposed to apply d to FX where F is given in (2). Hence, d is of size $p - 1$.

Let $Y = FX$, then the following test statistic is proposed to test the hypothesis of parallel profiles which is based on the formulation in (1):

Proposition 4.1 Let the parallelism hypothesis be defined via (1) and let $Y = FX$. A test statistic for testing the hypothesis is given by

$$\lambda_{Ph} = \frac{d'Y(I - P_{C'})Y'd}{d'Y(I - P_{C'})Y'd + d'YP_{C'(CC')^{-1}G}Y'd}. \tag{10}$$

It can be noted that $Y(I - P_{C'})Y' \sim W_{p-1}(F\Sigma F', N - r(C))$ is independent of $YP_{C'(CC')^{-1}G}Y' \sim W_{p-1}(F\Sigma F', r(G))$ and if d is a function of $Y(I - P_{C'})Y' + YP_{C'(CC')^{-1}G}Y' = Y(I - P_{C'G'})Y'$ Corollary 3.1 establishes the next theorem.

Theorem 4.1 *If d is a nonzero function (with probability 1) of $Y(I - P_{C'G'})Y'$ the test statistic for testing for parallel profiles in high dimensions, given in (10), follows a β -distribution with parameters $(N - r(C))/2$ and $r(G)/2$.*

In (5) the likelihood ratio was presented for the level hypothesis defined in (4). As it can be seen from the expression in (5) problems occur in high dimensions because S^{-1} does not exist when $p > N - r(C) + 1$. Therefore, $((F')^\circ S^{-1}(F')^\circ)^{-1}$ and $((F')^\circ S^{-1}(F')^\circ)^{-1}(F')^\circ S^{-1}$ have to be studied. It can be noted that if $p < N - r(C) + 1$,

$$((F')^\circ S^{-1}(F')^\circ)^{-1} \sim W_1(((F')^\circ \Sigma^{-1}(F')^\circ)^{-1}, N - r(C) - p + 1). \tag{11}$$

Thus, it can be seen that if $p > N - r(C) + 1$ negative degrees of freedom appear which of course is impossible. Thus, in order to test the level hypothesis in high dimensions the statistic in (5) has to be modified significantly. In this article the idea will be to modify the statistic in (5) as little as possible but so much that high-dimensional statistical analyses can be carried out.

A couple of ideas will bring us to a proposition where a test statistic and its distribution are given. The first idea will be to prove (11) in detail and see if anything can be modified so that when p is large reasonable expressions appear. To simplify notation $(A'S^{-1}A)^{-1}$ will be discussed where, for example, $A = (F')^\circ$ and the inverse is supposed to exist. The following chain of equalities shows some interesting structure:

$$\begin{aligned} (A'S^{-1}A)^{-1} &= (A'\Sigma^{-1}A)^{-1}A'\Sigma^{-1}A(A'S^{-1}A)^{-1}A'\Sigma^{-1}A(A'\Sigma^{-1}A)^{-1} \\ &= (A'\Sigma^{-1}A)^{-1}A'\Sigma^{-1}(S - SA^\circ(A^\circ SA^\circ)^{-1}A^\circ S)\Sigma^{-1}A(A'\Sigma^{-1}A)^{-1} \\ &= (A'\Sigma^{-1}A)^{-1}A'\Sigma^{-1}X P X' \Sigma^{-1}A(A'\Sigma^{-1}A)^{-1}, \end{aligned} \tag{12}$$

where

$$P = I - P_{C'} - (I - P_{C'})X'A^\circ(A^\circ X(I - P_{C'})X'A^\circ)^{-1}A^\circ X(I - P_{C'}). \tag{13}$$

Since $A'\Sigma^{-1}X$ is independent of $A^\circ X$ it is also independent of P . Moreover, P is idempotent and $r(P) = N - r(C) - p + r(A)$. Thus, for any choice of A° , conditionally on $A^\circ X$, the expression in (12) is Wishart distributed but it also appears that this distribution is independent of $A^\circ X$ and therefore (11) is established for the specific choice $A = (F')^\circ$.

The critical point in the high-dimensional setting is that $r(P)$ will approach 0, even if the inverse in (13) is replaced by a g-inverse, and therefore, it is proposed that P is modified in such a way that instead of this projection (expressed in F' instead of A°)

$$\tilde{P} = I - P_{C'} - (I - P_{C'})X'F'd(d'FX(I - P_{C'})X'F'd)^{-1}d'FX(I - P_{C'}) \quad (14)$$

will be used, where d is a function in FX . Summarizing these calculations will give a quantity which will be used as nominator in a test statistic for testing the level hypothesis:

$$U = ((F')^\circ \Sigma^{-1} (F')^\circ)^{-1} (F')^\circ \Sigma^{-1} X \tilde{P} X' \Sigma^{-1} (F')^\circ ((F')^\circ \Sigma^{-1} (F')^\circ)^{-1}. \quad (15)$$

How to choose d in (14) is not clear. On the one hand the distribution for U does not depend on d but if explicit expressions are to be calculated this will be a function d where now data has replaced X . From (5) it also follows that $(A'S^{-1}A)^{-1}A'S^{-1}$ has to be considered. Similar calculations to those in (12) yield

$$\begin{aligned} (A'S^{-1}A)^{-1}A'S^{-1}X &= (A'\Sigma^{-1}A)^{-1}A'\Sigma^{-1}A(A'S^{-1}A)^{-1}A'S^{-1}X \\ &= (A'\Sigma^{-1}A)^{-1}A'\Sigma^{-1}X \\ &\quad \times (I - (I - P_{C'})X'(A')^\circ((A')^\circ X(I - P_{C'})X'(A')^\circ)^{-1}(A')^\circ X). \end{aligned} \quad (16)$$

Replacing A by $(F')^\circ$ and then using the same d vector as in (14) leads to

$$h = P_1' X' \Sigma^{-1} (F')^\circ ((F')^\circ \Sigma^{-1} (F')^\circ)^{-1}, \quad (17)$$

where

$$P_1 = I - (I - P_{C'})X'F'd(d'FX(I - P_{C'})X'F'd)^{-1}d'FX \quad (18)$$

is an idempotent matrix. Moreover, let

$$\begin{aligned} \tilde{Q} &= G'(CC')^{-1}G + G'(CC')^{-1}CX'F'd \\ &\quad \times (d'FX(I - P_{C'})X'F'd)^{-1}d'FXC'(CC')^{-1}G \end{aligned} \quad (19)$$

and put

$$V = h'XC'(CC')^{-1}G\tilde{Q}^{-1}G'(CC')^{-1}CX'h. \quad (20)$$

Corresponding to λ_L in (5) the next quantity is proposed to test the level hypothesis:

$$\lambda_{Lh,prel} = \frac{U}{U + V}, \quad (21)$$

where U and V are defined in (15) and (20), respectively. However, it is not possible to use $\lambda_{Lh,prel}$ in (21) because both U and V include the unknown dispersion matrix Σ so a few more results have to be established. It follows immediately that

$$U \sim W_1(((F')^\circ \Sigma^{-1} (F')^\circ)^{-1}, N - r(C) - 1), \quad (22)$$

$$V \sim W_1(((F')^\circ \Sigma^{-1} (F')^\circ)^{-1}, r(G)). \quad (23)$$

If multiplying the nominator and denominator in (21) by $((\mathbf{F}')^\circ)' \boldsymbol{\Sigma}^{-1} (\mathbf{F}')^\circ)^{1/2}$ it follows since the distribution of

$$((\mathbf{F}')^\circ)' \boldsymbol{\Sigma}^{-1} (\mathbf{F}')^\circ)^{-1/2} ((\mathbf{F}')^\circ)' \boldsymbol{\Sigma}^{-1} \mathbf{X}$$

is independent of $\boldsymbol{\Sigma}$ that the distribution of $\lambda_{Lh,prel}$ is independent of $\boldsymbol{\Sigma}$. Thus, in order to have a test statistic which is functionally independent of the dispersion matrix $\boldsymbol{\Sigma} = \mathbf{I}$ is chosen. In the next proposition the test statistic for the level hypothesis is stated.

Proposition 4.2 *Let the level hypothesis testing problem be defined via (4). A test statistic for testing the level hypothesis is given by*

$$\lambda_{Lh} = \frac{\tilde{U}}{\tilde{U} + \tilde{V}},$$

where

$$\begin{aligned} \tilde{U} &= ((\mathbf{F}')^\circ)' (\mathbf{F}')^\circ)^{-1} (\mathbf{F}')^\circ)' \tilde{\mathbf{X}} \tilde{\mathbf{P}} \tilde{\mathbf{X}}' (\mathbf{F}')^\circ)' ((\mathbf{F}')^\circ)' (\mathbf{F}')^\circ)^{-1}, \\ \tilde{V} &= \tilde{\mathbf{h}}' \tilde{\mathbf{X}} \mathbf{C}' (\mathbf{C} \mathbf{C}')^{-1} \tilde{\mathbf{G}} \tilde{\mathbf{Q}}^{-1} \tilde{\mathbf{G}}' (\mathbf{C} \mathbf{C}')^{-1} \tilde{\mathbf{C}} \tilde{\mathbf{X}}' \tilde{\mathbf{h}}, \\ \tilde{\mathbf{h}} &= \tilde{\mathbf{P}}_1' \tilde{\mathbf{X}}' (\mathbf{F}')^\circ)' ((\mathbf{F}')^\circ)' (\mathbf{F}')^\circ)^{-1} \end{aligned}$$

with $\tilde{\mathbf{P}}$, $\tilde{\mathbf{P}}_1$ and $\tilde{\mathbf{Q}}$ defined in (14), (18) and (19), respectively.

How to choose \mathbf{d} in (14) and (18) is not clear. On the one side the distribution for λ_{Lh} does not depend on the choice of \mathbf{d} whereas on the other side when calculating an explicit value of λ_{Lh} the choice of \mathbf{d} will have an impact. This type of phenomena has been discussed very rarely. (We only know about it when a singular Gauss-Markov model has been considered.)

Based on the results in (22) and (23) the next theorem can be formulated.

Theorem 4.2 *Let \mathbf{d} in (14) be a nonzero function (with probability 1) of $\mathbf{F}\mathbf{X}$. The test statistic in Proposition 4.2 for testing the level hypothesis in (4) follows a β -distribution with parameters $(N - r(\mathbf{C}) - 1)/2$ and $r(\mathbf{G})/2$.*

The third hypothesis of flatness was stated in (6). Since the approach for creating a test statistic is the same as for the parallelism hypothesis, the results are stated without any proofs.

Proposition 4.3 *Let the flatness hypothesis be defined via (6) and put $\mathbf{Y} = \mathbf{F}\mathbf{X}$. A test statistic for testing the hypothesis is given by*

$$\lambda_{Fh} = \frac{\mathbf{d}' \mathbf{Y} (\mathbf{I} - \mathbf{P}_{C'}) \mathbf{Y}' \mathbf{d} + \mathbf{d}' \mathbf{Y} \mathbf{P}_{C'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{G}} \mathbf{Y}' \mathbf{d}}{\mathbf{d}' \mathbf{Y} (\mathbf{I} - \mathbf{P}_{C'}) \mathbf{Y}' \mathbf{d} + \mathbf{d}' \mathbf{Y} \mathbf{P}_{C'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{G}} \mathbf{Y}' \mathbf{d} + \mathbf{d}' \mathbf{Y} \mathbf{P}_{C'\mathbf{G}^{\circ}} \mathbf{Y}' \mathbf{d}}, \tag{24}$$

where $\mathbf{d} \in \mathfrak{R}^{p-1}$, with probability 1, is a nonzero function in $\mathbf{Y}\mathbf{Y}'$.

The motivation that \mathbf{d} is a function in $\mathbf{Y}\mathbf{Y}'$ follows from the fact that

$$\mathbf{Y}\mathbf{Y}' = \mathbf{Y}(\mathbf{I} - \mathbf{P}_{C'})\mathbf{Y}' + \mathbf{Y}\mathbf{P}_{C'(CC')^{-1}G}\mathbf{Y}' + \mathbf{Y}\mathbf{P}_{C'G^\circ}\mathbf{Y}'.$$

Moreover,

$$\begin{aligned} \mathbf{Y}(\mathbf{I} - \mathbf{P}_{C'})\mathbf{Y}' + \mathbf{Y}\mathbf{P}_{C'(CC')^{-1}G}\mathbf{Y}' &\sim W_{p-1}(\mathbf{F}\Sigma\mathbf{F}', N - r(\mathbf{C}'\mathbf{G}^\circ)), \\ \mathbf{Y}\mathbf{P}_{C'G^\circ}\mathbf{Y}' &\sim W_{p-1}(\mathbf{F}\Sigma\mathbf{F}', r(\mathbf{C}'\mathbf{G}^\circ)). \end{aligned}$$

These results imply that the next theorem can be established.

Theorem 4.3 *The test statistic for testing for flatness in high dimensions, given in Proposition 4.3, follows a β -distribution with parameters $(N - r(\mathbf{C}'\mathbf{G}^\circ))/2$ and $r(\mathbf{C}'\mathbf{G}^\circ)/2$.*

5 Concluding Remarks

In this article the three well-known test statistics in profile analysis have been modified so that high-dimensional data can be handled in a non-asymptotic approach. The test for parallelism and flatness was derived following ideas given by Läuter [8, 9] and Läuter et al. [10, 11] which originally was developed for handling MANOVA problems. Concerning the level testing a completely new approach is proposed. Here we modify the degrees of freedom and an exact test is derived.

The vector \mathbf{d} , which is utilized in our approach when testing for parallelism and flatness, is a function of some sums of squares has only briefly been considered in this article. Instead it is referred to Läuter et al. [10], Section 4, where several different alternatives for determining \mathbf{d} are proposed. Furthermore, a generalization of the presented approach in this article will be to apply a matrix \mathbf{D} , i.e. study several linear scores, instead of a vector \mathbf{d} which only give one score.

There is another important problem (observed by one of the reviewers) that the choice of \mathbf{F} can have an effect on the choice of \mathbf{d} and thus the test statistic depends in fact on the choice of \mathbf{F} . It is important to continue this work and establish restrictions on the choice of \mathbf{d} so that the vector only depends on the space generated by the columns in \mathbf{F} , i.e. instead of using \mathbf{F} using the projection $\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'$.

Acknowledgements We would like to thank both reviewers who delivered many insightful comments and at the same time recognized many “small” errors. Dietrich von Rosen is supported by the Swedish Research Council (2017-03003).

Funding Open Access funding provided by Swedish University of Agricultural Sciences.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Cengiz C, von Rosen D (2020) High-dimensional profile analysis. Report LiTH-MAT-R–2020/07–SE, Department of Mathematics, Linköping University, Linköping, Sweden
2. Fang KT, Zhang YT (1990) Generalized multivariate analysis. Springer, Berlin
3. Fujikoshi Y (2009) Statistical inference for parallelism hypothesis in growth curve model. *SUT J Math* 45:137–148
4. Greenhouse SW, Geisser S (1959) On the methods in the analysis of profile data. *Psychometrika* 24:95–112
5. Harrar SW, Kong X (2016) High-dimensional multivariate repeated measures analysis with unequal covariance matrices. *J Multivar Anal* 145:1–21
6. Harrar SW, Xu J (2008) Asymptotic expansion of the null distributions of test statistics for profile analysis under general conditions. In: International conference on multivariate statistical modelling and high dimensional data mining, HDM-2008, Kayseri, Turkey
7. Hyodo M (2017) Tests for the parallelism and flatness hypotheses of multi-group profile analysis for high-dimensional elliptical populations. *J Multivar Anal* 162:82–92
8. Läuter J (1996) Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics* 52:964–970
9. Läuter J (2016) *Multivariate Statistik - drei Manuskripte*. Shaker Verlag, Aachen
10. Läuter J, Glimm E, Kropf S (1996) New multivariate tests for data with an inherent structure. *Biom J* 38:5–23
11. Läuter J, Glimm E, Kropf S (1998) Multivariate tests based on left-spherically distributed linear scores. *Ann Stat* 26:1972–1988
12. Maruyama Y (2007) Asymptotic expansion of the null distributions of some test statistics for profile analysis in general conditions. *J Stat Plan Inference* 137:506–526
13. Ohlson M, Srivastava MS (2010) Profile analysis for a growth curve model. *J Jpn Stat Soc* 40:1–21
14. Okamoto N, Miura N, Seo T (2006) On the distributions of some test statistics for profile analysis in elliptical populations. *Am J Math Manag Sci* 26:1–31
15. Onozawa M, Nishiyama T, Seo T (2016) On test statistics in profile analysis with high-dimensional data. *Commun Stat Simul Comput* 45:3716–3743
16. Seo T, Sakurai T, Fujikoshi Y (2011) LR tests for two hypotheses in profile analysis of growth curve data. *SUT J Math* 47:105–118
17. Srivastava MS (1987) Profile analysis of several groups. *Commun Stat Theory Methods* 16:909–926
18. Srivastava MS (2002) *Methods of multivariate statistics*. Wiley, New York
19. Srivastava MS, Carter EM (1983) *An introduction to applied multivariate statistics*. North Holland, New York
20. Srivastava MS, Singull M (2012) Profile analysis with random-effects covariance structure. *J Jpn Stat Soc* 42:145–164
21. Takahashi S, Shutoh N (2016) Tests for parallelism and flatness hypotheses of two mean vectors in high-dimensional settings. *J Stat Comput Simul* 86:1150–1165
22. von Rosen D (2016) Profile analysis. Wiley StatsRef. <https://doi.org/10.1002/9781118445112.stat02458.pub2>
23. von Rosen D (2018) *Bilinear regression analysis: an introduction*. Lecture notes in statistics, vol 220, Springer, New York

24. Yokoyama T (1995) LR test for random-effects covariance structure in a parallel profile model. *Ann Inst Stat Math* 47:309–320
25. Yokoyama T, Fujikoshi Y (1993) A parallel profile model with random-effects covariance structure. *J Jpn Stat Soc* 23:83–89

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.