Roman Zmyślony

Institute of Mathematics,

University of Zielona Góra

Zielona Góra, January 28, 2022

# Referee report on habilitation thesis "Estimation and Testing in Multivariate Linear Models" by RNDr. Daniel Klein, PhD.

## 1. General characteristics of results of the habilitation dissertation and organizing scientific and teaching activity ..

The habilitation thesis consists of a series of 12 papers published in journals with an international dimension. In his own reports, they were marked with the symbols P1-P12 and were published in the years 2009-2018. Among them, 6 works were published with 1 co-author and 5 works with 2 co-authors and 1 with 3 co-authors. RNDr. Daniel Klein has published a total of 74 papers, monographs or conference abstracts, including 32 papers published in international journals. Moreover, some works are intended for printing. The Impact Factor of journals in which papers are published have high index and the number of citations according to Web of Science is 170. RNDr. Daniel Klein presented his results in 24 international conferences and two of them as invited speaker. Moreover, he has jointed international workshops and presented 16 lecture. He was also a member of organizing committee in several conferences. He was very active in teaching students and supervised several diplomas thesis.

## 2. The main results and evaluation of the works included in the dissertation.

The dissertation consists of four papers P1-P4 concerning the estimation of the parameters in so called growth curve model, while in papers P5 and P9 problem of maximum likelihood estimation, in so called generalized growth curve model, is considered. Papers P6, P7, P8, P11 and P12 deal with problem of testing hypotheses connected with structure of covariance operator in multivariate normal models. Paper P10 is an algebraic paper and contains definition and properties of partial trace and block trace operators of partitioned matrices and the applications of this results in statistics.

In **paper P1** there is considered normally distributed multivariate model with the following structure

$$\boldsymbol{Y} = \boldsymbol{XBZ} + \mathbf{e}, \qquad \boldsymbol{E e} = 0, \quad \mathrm{Var}(vec\ \boldsymbol{e}) = \boldsymbol{\Sigma} \otimes \boldsymbol{I},$$

where the covariance matrix $\boldsymbol{\Sigma}$ has the serial correlation structure. It means that matrix $\boldsymbol{\Sigma}$ has two parameters. In Theorem 4. the MLE for fixed parameter matrix are given in explicit form while estimators of two parameter of covariance are given as a solution of two nonlinear equation.

In **paper P2** the assumption about structure of expectation is the same as in paper P1 with additional assumption that matrix $\boldsymbol{X}$ is an ANOVA design matrix, while covariance of $\boldsymbol{\Sigma}$ have the following linear structure $\theta_1 G + \theta_2 \omega\omega'$ with known positive definite matrix $G$ and known vector $\omega$. The main value of the paper is how to use the orthogonal decompositions to get estimators, distribution and variances of estimators.

**Paper P3** mainly contains overview of the results in linear models of growth-curve-type with 52 citations.

In **Paper P4** confidence region for 2 parameters (variance and corelation coefficient) in a model with uniform correlation structure are constructed on the base of joint density, which is given in Lemma 1.1.. Confidence interval for each parameter are also given. There are also simulation study and some suggestions for application of those result.

**Paper P5** deals with the following so called extended growth curve model (EGCM)

$$\boldsymbol{Y} = \sum_{i=1}^{k} \boldsymbol{X}_i \boldsymbol{B}_i \boldsymbol{Z}_i' + \boldsymbol{\mathcal{E}}, \quad \text{vec}\, \boldsymbol{\mathcal{E}} \sim \boldsymbol{N}(0, \boldsymbol{\Sigma} \otimes \boldsymbol{I}).$$

In Theorem 2, under assumption, that spaces generated by columns of matrices $\boldsymbol{Z}_i$ decrease if $i$ increase, the set of all MLE estimators for $\boldsymbol{B}_i$ is presented in the explicit form (in general the MLE for $\boldsymbol{B}_i$ are not uniquely defined). Moreover, and explicit MLE estimator of $\boldsymbol{\Sigma}$ is given. Theorem 3 states that MLE estimators for $\boldsymbol{B}_i$ are uniquely defined iff the matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are of full rank, while MLE of $\boldsymbol{\Sigma}$ is the same as in Theorem 2.

Estimators of $\boldsymbol{B}_i$ given in Theorem 3 are unbiased and their variances are calculated (see Theorem 7). Moreover, the expected value MLE of $\boldsymbol{\Sigma}$ is calculated and presented as Theorem 8. **Let me mention that the proofs of the above two theorems are very complicated.**

In **Paper P6** a new test statistic for testing equality of mean vectors is proposed. It is assumed that two vectors are correlated and have different variance matrices and both have block compound structure (BCS). That why it was possible to find the covariance matrix of the difference of vectors. After Helmert's transformation of the differences they have BCS covariance structure. and test statistic is represented as a sum of two independent quadratic forms of the best linear and the best quadratic unbiased estimators of mean and covariance, respectively. Finally, it is proved that this statistic under null hypotheses is distributed as the convolution of two independent Hotelling's $T^2$ or as sum of two independent F distributed variable with different degrees of freedom for denominator. Moreover, there are calculated $p$-values for two sets of real data for the new test

In the next two papers **P7 and P8** the problem of comparison of MLE test with Rao score test (RST) for testing hypotheses about separability of covariance structure $\Omega$ is presented. The RST test advantage is that it is especially recommended for small data analysis

**In P7** the following test for separability

$$H_0: \Omega = \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \quad \boldsymbol{\Psi} \text{ is CS} \quad \text{vs.} \quad H_{A1}: \Omega \text{ US}$$

is considered. .Here CS and US means complete symmetric and unstructured, respectively.

In **P8** the following two hypotheses

$$H_{01}: \Omega = \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \quad \boldsymbol{\Psi} \text{ is AR(1)} \quad \text{vs.} \quad H_{A1}: \Omega \text{ US},$$

$$H_{02}: \Omega = \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \quad \boldsymbol{\Psi} \text{ is AR(1)} \quad \text{vs.} \quad H_{A2}: \Omega \text{ US},$$

are considered. Matrix $\boldsymbol{\Sigma}$ is assumed to be unstructured in all three above hypothesis. For all 3 cases of testing hypothesis, Rao's score statistics (RST) are presented and they are function of MLEs calculated under null hypothesis. Moreover, empirical histograms and limiting distribution for LRT statistics and for RST statistics for small and large sample sizes are presented. Resalts for some real data set are compare using both tests LRT and RST. **In my opinion the results presented in the above two papers are very interesting and useful from theoretical and practical point of view. Moreover, these proofs and simulations are not trivial.**

In paper **P9** estimation of parameters under generalized growth curve normal distributed model is considered. The observed data are presented as multi-index matrix (tensor). If there are 3 indexes then presentation of data can be matricized in three ways (see Definition 2). In this paper there are used two of them for convenience and because vectorization of these matrices have the same structure of means and covariance matrices. Namely, vector of means equals $(\mathbf{C}\otimes\mathbf{B}\otimes\mathbf{A})vec\boldsymbol{X}$, where matrices **A, B** and **C** are known matrices, while $\boldsymbol{X}$ is a matrix of unknown parameters. **The great value of this paper** is that MLEs of parameters are obtained for the following 3 cases of covariance matrices:

(i) $\boldsymbol{\Psi}\otimes\boldsymbol{\Sigma}\otimes\mathbf{I}$, where $\boldsymbol{\Psi}$ is known

(ii) $\boldsymbol{\Psi}\otimes\boldsymbol{\Sigma}\otimes\mathbf{I}$, where $\boldsymbol{\Psi}$ is unknown

(iii) $\boldsymbol{\Omega}\otimes\mathbf{I}$.

The applications of the results are illustrated on the set of real data.

Finally, the results are extended to tensors of order great than 3.

In paper **P10** two linear operators (defined on partitional matrices) the partial trace operator and the block trace operator are defined. Main part of this paper deal with properties of the partial trace operator and the block trace operator. This properties are given in 15th lemmas. The wide applications in statistics of these operators are presented.**The wide applications in statistics of these operators are presented.**

In paper **P11** Rao's score test statistic is derived for testing hypothesis that the covariance matrix have block compound symmetric (BCS) structure, i.e.

$H_0$: $\Omega$ =BCS covariance matrix $\Gamma$ vs. $H_1$: $\Omega$ = unstructured, positive definite,

where $\Gamma = \mathrm{I_u}\otimes(\boldsymbol{U_0} - \boldsymbol{U_1}) + \mathrm{J_u}\otimes\boldsymbol{U_1}$ Here $\mathrm{I_u}$ is the identity matrix, $\mathrm{J_u} = 1_u 1_u'$ , where $1_u$ is the $u \times 1$ vector of ones, while the matrices $\boldsymbol{U_0}$, $\boldsymbol{U_1}$ are $m$ ×m symmetric matrices of parameters.

It is assumed that mean $um \times 1$ vector $\mu$ is unknown. In Theorem 3.1 MLEs for unknown parameters under null hypothesis are given. For calculation RTS statistic, denoted by RS, the MLEs under null hypotheses are plugin into score vector and into Fisher information matrix. Theorem 3.3 is very important, because it says that the distribution of RS is independent of parameters under null hypothesis. Moreover, RS depends on MLE of $\Omega$ under $H_1$and it is independent of its inverse. Note that MLE of $\Omega$ under $H_1$ is singular for $n \leq u\ m$ this implies that LRT is not definite. But if $m < n$ then RTS statistics is well defined. Moreover, empirical histograms and limiting distributions for LRT statistics and for RST statistics for small and large sample sizes are presented. For small real data with $m = 5$, $u = 2, n = 8$ the illustration how to apply results are discussed. In this case, RST statistics is well definite but MLT statistics can't be calculated. The paper contains six table, where empirical critical value are given by simulation study for making an use of them for small and big data sets. Moreover, for random directions and for four point in this direction the powers of two tests are compared.. **In my opinion the results presented in this paper are very interesting, especially for applications.**

The last paper P12 is continuation of the problems given in paper P6 in the following meaning. The covariance have BCS structure as in paper P6 or it is defined in my report about paper P11 (one page up), while the mean vector is unstructured. In this paper simply modification of null hypotheses about mean vector is presented, namely .testing hypotheses that mean vector is equal null vector in paper P6 is change for mean vector is equal to a given vector. But there is no important generalization. Int his paper P12 it is called $BT^2$ statistic and if null hypothesis holds then the statistic $BT^2 \sim T^2_{q,n-1} \oplus T^2_{q,(n-1)(p-1)}$, where $\oplus$ denotes the convolution operation and $T^2$ denotes Hotelling $T^2$ distribution with appropriate degree of freedom. In this paper the new test statistic is proposed and it is so called $D^2$ based on different transformation of data set. $D^2$ is obtain with different orthogonal transformations than $BT^2$ and two important relation are proved. First, that the distribution of $D^2$ is the convolution of two LH-trace distribution i.e. $D^2 \sim T^2_0(q; 1, n-1) \oplus T^2_0(q; p-1, (n-1), (p-1))$. Since $T^2_0(q; 1, n-1)$ is equal to usual $T^2_{q,n-1}$ then $BT^2$and $D^2$ are different by distribution of the second terms. Second difference is that statistic for all Helmert's orthogonal transformation with first column proportional to vector of one's , which define test statistic $BT^2$, the following inequality hols $BT^2 \leq D^2$. .and equality holds only if $p = 2$, (for any $q$) $or\ q = 1$, (for any $q$)

In model with exchangeable mean structure for testing hypothesis that mean vector is equal to a given vector, the Hotelling T2 statistic is obtain in similar way.

The next part of the paper is connected with testing hypotheses for means vector for two independent samples with equal covariance BCS structure. As it is expected, the similar results are given like for one sample. Estimator for covariance matrix is computed from two samples

The results of simulations study for comparison the powers of two tests and control of significant levels are also presented

**3. Final conclusions.**

The work contributes significantly to the theory of estimation and testing hypothesis in multivariate normal distributions. The author obtained new theoretical results in the field of testing hypotheses on parameters and in selection correct model among several matrix structures. The models are rightly selected based on minimal sufficient statistics. These statistics are also an estimators obtained by the maximum likelihood method in a model with any positively determined unknown covariance matrix. In obtaining the above-mentioned results, fluent knowledge and application of the formulas for differentiating vector and matrix functions in statistical problems was important. Moreover, the author of the work presented the use of the results of the work for real data. Based on the above considerations and taking into account the high substantive evaluation of the work, I am asking for it to be distinguished.

Roman Zmyślony