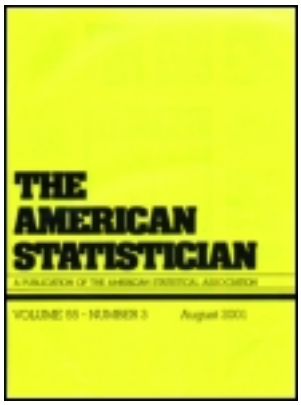


The American Statistician
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utas20

Closed Likelihood Ratio Testing Procedures to Assess Similarity of Covariance Matrices

Francesca Greselin [a] & Antonio Punzo [b]

[a] Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali , Università di Milano-Bicocca , Milano , Italy

[b] Dipartimento di Economia e Impresa , Università di Catania , Catania , Italy

Accepted author version posted online: 02 May 2013.

To cite this article: Francesca Greselin & Antonio Punzo (2013): Closed Likelihood Ratio Testing Procedures to Assess Similarity of Covariance Matrices, The American Statistician, DOI:10.1080/00031305.2013.791643

To link to this article: http://dx.doi.org/10.1080/00031305.2013.791643

Closed Likelihood Ratio Testing Procedures to Assess Similarity of Covariance Matrices

Francesca Greselin

Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali

Università di Milano-Bicocca, Milano, Italy

e-mail: francesca.greselin@unimib.it

Antonio Punzo

Dipartimento di Economia e Impresa

Università di Catania, Catania, Italy

e-mail: antonio.punzo@unict.it

Abstract

In this paper we introduce a multiple testing procedure to assess a common covariance structure between k groups. The new test allows for a choice among eight different patterns arising from the three-term eigen decomposition of the group covariances. It is based on the closed testing principle and adopts local likelihood ratio tests. The approach reveals richer information about the underlying data structure than classical methods, the most common one being only based on homo/heteroscedasticity. At the same time, it provides a more parsimonious parameterization, whenever the constrained model is suitable to describe the real data. The new inferential methodology is then applied to some well-known data sets chosen from the multivariate literature. Finally, simulation results are presented to investigate its performance in different situations representing gradual departures from homoscedasticity and to evaluate the reliability of using the asymptotic χ^2 to approximate the actual distribution of the local likelihood ratio test statistics.

Keywords: Closed Testing Procedures; Common Principal Components; Eigen Decomposition; Homoscedasticity; Likelihood Ratio Tests; Proportional Covariance Matrices.

1 Introduction

Let x_{ih} be the ith p-dimensional observation in group h, $i = 1, \ldots, n_h$, $h = 1, \ldots, k$. Suppose that data in each group are randomly drawn from a normal distribution with mean vector $\boldsymbol{\mu}_h$ and covariance matrix $\boldsymbol{\Sigma}_h$, $h = 1, \ldots, k$.

Unlike multivariate analysis of variance (MANOVA), where the interest is in comparing $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$, here the attention is focused on the detection of a common structure (or *pattern*) among the group covariances $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k$. The assessment of a common pattern provides a more informative estimation of the group conditional densities on the training set in normal discriminant analysis (see, e.g., Flury *et al.*, 1994 and Bagnato *et al.*, in press) and it is of intrinsic interest in such fields as psychometrics or genetics, where the similarity of genetic covariance structures among species is a fundamental subject of investigation. Moreover, discovering a common structure for group covariances allows one to reduce the number of estimated parameters (*principle of parsimony*).

To face the issue, this paper considers a very general framework given by the (three-term) eigen decomposition

$$\boldsymbol{\Sigma}_h = \lambda_h \boldsymbol{\Gamma}_h \boldsymbol{\Delta}_h \boldsymbol{\Gamma}_h', \qquad h = 1, \ldots, k, \tag{1}$$

where $\lambda_h = |\boldsymbol{\Sigma}_h|^{1/p}$, with $|\cdot|$ denoting the determinant, $\boldsymbol{\Delta}_h$ is the scaled ($|\boldsymbol{\Delta}_h| = 1$) diagonal matrix of the eigenvalues of $\boldsymbol{\Sigma}_h$ sorted in decreasing order, and $\boldsymbol{\Gamma}_h$ is a $p \times p$ orthogonal matrix whose columns are the normalized eigenvectors of $\boldsymbol{\Sigma}_h$, ordered according to their eigenvalues (note that $\boldsymbol{\Gamma}_h$ is not unique when eigenvalues have multiplicities greater than one). This decomposition is widely employed in the mixture framework (see Celeux and Govaert, 1995 and Bensmail and Celeux, 1996, among many others). Each component in the right side of (1) has a different geometric interpretation in terms of the hyper-ellipsoids of equal concentration (sections of the p-variate normal distribution having a given probability). The volume is determined by λ_h (more precisely, it is proportional to $\lambda_h^{p/2} = |\boldsymbol{\Sigma}_h|^{1/2}$; see Rencher and Christensen, 2012, Sections 3.11 and 4.1.3 for details). Shape and orientation are governed by $\boldsymbol{\Delta}_h$ and $\boldsymbol{\Gamma}_h$, respectively. Now, considering the

triplet ($\lambda_h, \mathbf{\Delta}_h, \mathbf{\Gamma}_h$), and allowing to its elements to be equal (E) or variable (V) among groups, we obtain the following family of parsimonious and easily interpretable models

$$\widetilde{\mathcal{M}} = \Big\{\text{EEE, VEE, EVE, EEV, VVE, VEV, EVV, VVV}\Big\}.$$

With this convention, writing $\lambda\mathbf{\Gamma}_h\mathbf{\Delta}\mathbf{\Gamma}_h'$, or EEV, means that we consider groups with equal volumes, equal shapes, and different orientations. The present approach is quite general because it includes many previously studied models as special cases. Table 1 schematizes all the models $M \in \widetilde{\mathcal{M}}$, giving their mapping with the covariance restrictions already known in the literature. Figure 1 exemplifies the models providing a graphical representation in the case $p = k = 2$.

To detect the best model in $\widetilde{\mathcal{M}}$, we use a closed multiple testing procedure. This is done according to a very familiar MANOVA paradigm, where three treatment groups are compared with a common control group, or a single treatment is evaluated against a single control on the basis of three different variables (see, e.g., Westfall and Wolfinger 2000 and Bretz *et al.* 2011, Chapter 4.1). Generalized likelihood ratio (LR) tests are considered as local tests in the procedure. The result is a "global procedure" in the sense that the model, for the observed data, is detected by *simultaneously* considering all the elements in $\widetilde{\mathcal{M}}$. Conversely, in the MANOVA framework (see, e.g. Christensen, 1996, p. 461) the user often follows a path that involves *a subset* of the models.

The paper is organized as follows. Section 2 shows some geometrical features on real data that suggested the development of the new approach. Section 3 gives basic notation about closed testing procedures, and Section 4 describes how to use LR tests as local tests. The new inferential procedure is then applied, in Section 5, to some well-known data in the multivariate literature to show how it works, in terms of improved information and parsimony. A simulation study is developed in Section 6 to assess the performances of the closed testing procedure under normal and heavy tailed densities and to evaluate the reliability of using the asymptotic χ^2 for the actual distribution of the local LR test statistics. Final considerations are given in Section 7.

2 Real data motivating inference on $\widetilde{\mathcal{M}}$

In this section two real data sets will be shown to motivate inference on $\widetilde{\mathcal{M}}$.

The first example refers to the crab data set of Campbell and Mahon (1974) on the genus *Leptograpsus*. Attention is focused on the sample of $n = 100$ blue crabs, there being $n_1 = 50$ males (group 1) and $n_2 = 50$ females (group 2), each specimen having $p = 2$ measurements (in millimeters) for the rear width (RW) and the length along the midline (CL) of the carapace. By Mardia's test, the two group-conditional distributions can be considered bivariate normal, while the LR test of homoscedasticity (versus heteroscedasticity) rejects the null hypothesis at any reasonable significance level. Figure 2(a) displays the scatter plot of RW versus CL. In both groups, the estimated ellipses of equal (95%) concentration (giving the graphical counterpart of the covariance matrices estimated under VVV) are superimposed to facilitate our understanding. Although the LR test points out heteroscedasticity, Figure 2(a) suggests that the two scatters for male and female crabs have approximately the same volume and the same shape (i.e., same proportion between the axes in the ellipse) but different orientations (directions of the main axes in the ellipses). This pattern is represented by model EEV. On the other hand, noting some difference in volumes in Figure 2(a), someone could lean toward VEV. As a consequence, we would appreciate a testing procedure to assess our conjectures with the aim of gaining information about the data and obtaining a more parsimonious model.

As a second example, we consider the turtle data set of Jolicoeur and Mosimann (1960) focusing on the variables carapace length (LH) and height (HT), measured in millimeters, on $n_1 = 24$ male and $n_2 = 24$ female painted turtles (*Chrysemys picta marginata*). Although homoscedasticity does not hold, Figure 2(b) shows some similarity in terms of shape for the two group scatters. If we were able to assess this conjecture, concluding that model VEV describes the covariance structure between groups, then we would obtain a gain in information and in parsimony, as in the previous example. Moreover, *shape* is a well-known and fundamental concept for the analysis of variation

in living organisms, allowing for significant classification (Jolicoeur and Mosimann, 1960, p. 339).

In this paper, we show how the new inferential method provides a rigorous systematization of these considerations.

3 Model assessment via closed testing procedures

In this section we provide a method for model assessment in $\widetilde{\mathcal{M}}$, in the framework of *multiple testing procedures* (MTPs; see e.g., Hochberg and Tamhane, 1987). Let us denote by H_0^M the null hypothesis related to the generic model $M \in \overline{\mathcal{M}} = \widetilde{\mathcal{M}} \setminus \{\text{VVV}\}$. Among the available alternatives for each null, we set H_1^{VVV} as the unique (benchmark) alternative hypothesis. This positions allows us to define seven *omnibus* tests, with VVV being the most general, unconstrained model. The seven null hypotheses are represented in Figure 3.

Now, let us denote by $\mathcal{H} = \{H_0^{\text{VVE}}, H_0^{\text{VEV}}, H_0^{\text{EVV}}\}$ the subfamily of *elementary* hypotheses in the bottom layer of Figure 3. They play a crucial role: depending on the true model in $\widetilde{\mathcal{M}}$, none, some, or all of the hypotheses in $\mathcal{H}$ may be the true null. Thus, for example, if the true model in $\widetilde{\mathcal{M}}$ is EEV, then H_0^{EVV} and H_0^{VEV} hold true, while H_0^{VVE} is false. Figure 3 is represented as a hierarchy where arrows indicate implications, because, for instance, H_0^{EEE} implies H_0^{VEE} (see Hochberg and Tamhane, 1987, p. 344); in other words, EEE is more restrictive than VEE.

The natural choice for our context is to adopt a *closed testing procedure* (CTP; Marcus *et al.*, 1976). It is the most powerful, among the available MTPs, that strongly controls the *familywise error rate* (FWER) at level α (as recently further corroborated via simulations by Giancristofaro Arboretti *et al.*, 2012). Controlling the FWER in a strong sense means to control the probability of committing at least one Type I error, under any partial configuration of true and false null hypotheses in $\mathcal{H}$. This is the only way to make inference on each hypothesis in $\mathcal{H}$. Operationally we reject, say, the elementary hypothesis H_0^{VEV} if and only if each test on the more restrictive hypotheses H_0^{VEE}, H_0^{EEV}, H_0^{EEE}, and also on H_0^{VEV} itself, yields a significant result. Denoting by p_{VEE}, p_{EEV}, p_{EEE}, and p_{VEV} the p-values for H_0^{VEE}, H_0^{EEV}, H_0^{EEE}, and H_0^{VEV}, respectively, we report the

adjusted p-value for H_0^{VEV} as $q_{\text{VEV}} = \max\{p_{\text{VEE}}, p_{\text{EEV}}, p_{\text{EEE}}, p_{\text{VEV}}\}$. An adjusted p-value represents the natural counterpart, in the multiple testing framework, of the classical p-value (see, e.g., Bretz *et al.*, 2011, p. 18).

4 Local likelihood ratio tests

To complete the definition of the closed testing procedure we still need to specify how to test elementary and non-elementary hypotheses.

For the elementary hypotheses, we adopt local α-level tests based on the LR statistic because they represent the standard method to compare k covariance matrices (Manly and Rayner, 1987).

For what concerns the non-elementary hypotheses, we first have to recall that every different method for testing them leads to a new and different CTP, and CTPs are best when these methods are as powerful as possible. The simplest approach consists of using MTPs such as, for example, the classical Bonferroni method or the slightly more complex, but much less conservative, method of Holm (1979). They are both applicable under arbitrary dependence among tests statistics. However, a non-multiple approach in this phase could guarantee a better performance, and this is the reason why we prefer to move towards local α-level tests, based on the LR test statistics, also for the non-elementary hypotheses.

For each $M \in \widetilde{\mathcal{M}}$, the joint likelihood function is

$$L_M\left(\left\{\boldsymbol{\mu}_h, \lambda_h^M, \boldsymbol{\Delta}_h^M, \boldsymbol{\Gamma}_h^M\right\}_{h=1}^k\right) = \prod_{h=1}^k \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n_h}(\boldsymbol{x}_{ih}-\boldsymbol{\mu}_h)'\left[\lambda_h^M\boldsymbol{\Gamma}_h^M\boldsymbol{\Delta}_h^M\left(\boldsymbol{\Gamma}_h^M\right)'\right]^{-1}(\boldsymbol{x}_{ih}-\boldsymbol{\mu}_h)\right\}}{(2\pi)^{\frac{1}{2}pn_h}\left|\lambda_h^M\boldsymbol{\Gamma}_h^M\boldsymbol{\Delta}_h^M\left(\boldsymbol{\Gamma}_h^M\right)\right|^{\frac{1}{2}n_h}}. \tag{2}$$

The corresponding joint log-likelihood function, $l_M = \ln L_M$, is given by

$$\begin{aligned} l_M\left(\left\{\boldsymbol{\mu}_h, \lambda_h^M, \boldsymbol{\Delta}_h^M, \boldsymbol{\Gamma}_h^M\right\}_{h=1}^k\right) = & \sum_{h=1}^k\left\{-\frac{pn_h}{2}\ln(2\pi) - \frac{n_h}{2}\ln\left|\lambda_h^M\boldsymbol{\Gamma}_h^M\boldsymbol{\Delta}_h^M\left(\boldsymbol{\Gamma}_h^M\right)'\right| + \right. \\ & \left. -\frac{1}{2}\sum_{i=1}^{n_h}(\boldsymbol{x}_{ih}-\boldsymbol{\mu}_h)'\left[\lambda_h^M\boldsymbol{\Gamma}_h^M\boldsymbol{\Delta}_h^M\left(\boldsymbol{\Gamma}_h^M\right)'\right]^{-1}(\boldsymbol{x}_{ih}-\boldsymbol{\mu}_h)\right\}. \end{aligned} \tag{3}$$

Regardless from $M \in \widetilde{\mathcal{M}}$, the maximum likelihood (ML) estimate of $\boldsymbol{\mu}_h$ is the sample mean $\overline{\boldsymbol{x}}_h$, and it is obtained as solution of

$$\frac{\partial l_M\left(\left\{\boldsymbol{\mu}_h, \lambda_h^M, \boldsymbol{\Delta}_h^M, \boldsymbol{\Gamma}_h^M\right\}_{h=1}^k\right)}{\partial \boldsymbol{\mu}_h} = \left[\lambda_h^M \boldsymbol{\Gamma}_h^M \boldsymbol{\Delta}_h^M \left(\boldsymbol{\Gamma}_h^M\right)'\right]^{-1} \sum_{i=1}^{n_h} (\boldsymbol{x}_{ih} - \boldsymbol{\mu}_h) = \mathbf{0}, \qquad h = 1, \ldots, k.$$

Since the maximizing value of $\boldsymbol{\mu}_h$ does not depend on λ_h^M, $\boldsymbol{\Delta}_h^M$, and $\boldsymbol{\Gamma}_h^M$, we can obtain the ML estimates for the latter three components by maximizing either L_M or l_M; then, through (1), we get the estimate of $\boldsymbol{\Sigma}_h^M$. The estimates of λ_h^{VVV}, $\boldsymbol{\Delta}_h^{\mathrm{VVV}}$ and $\boldsymbol{\Gamma}_h^{\mathrm{VVV}}$ will be defined in the following as *sample volume*, *sample shape-matrix* and *sample orientation-matrix*, respectively.

To maximize L_M (or l_M), iterative procedures are occasionally needed. The second column of Table 2 indicates if ML estimation can be achieved in a closed form (CF) or if an iterative procedure (IP) is needed. Details on the estimates can be found in Celeux and Govaert (1995) and Biernacki *et al.* (see 2008, pp. 22–24). Note that, contrarily to what described in the above cited works, the constraint of decreasing order of the diagonal elements in $\boldsymbol{\Delta}_h$, $h = 1, \ldots, k$, should be implemented for EVE and VVE.

Now, a natural way to compare each $M \in \overline{\mathcal{M}}$ with the diagnostic model VVV, consists of using the (generalized) likelihood ratio (LR) statistic

$$LR_M = -2 \ln \frac{L_M}{L_{\mathrm{VVV}}} \tag{4}$$

that, under H_0^M, is asymptotically distributed (when $\min_{h=1,\ldots,k} n_h \to \infty$) as a χ^2 with $\nu_M = \eta_{\mathrm{VVV}} - \eta_M$ degrees of freedom, where η_{VVV} and η_M denote the number of (free) parameters for VVV and M, respectively (see Anderson, 1984, pp. 405–406, for details on LR_{EEE}). The value of ν_M is the gain in parsimony that could be obtained, improving at the same time the information about the group covariance structures. Table 2 specifies the number of parameters η_M, and the degrees of freedom ν_M of the LR statistic, for each $M \in \widetilde{\mathcal{M}}$.

Finally, note that Box's test and the test discussed in Flury (1986) are already existing LR tests for the null hypotheses H_0^{EEE} and H_0^{VEE}, respectively. However, they are based on the unbiased

version of the maximum likelihood estimators for $\mathbf{\Sigma}_h^{\mathrm{EEE}}$ and $\mathbf{\Sigma}_h^{\mathrm{VEE}}$. For methodological uniformity, since we employ ML estimators of the covariance matrices for the other models in $\widetilde{\mathcal{M}}$, we prefer to do the same also for H_0^{EEE} and H_0^{VEE}.

5 Applications

In this section we will show how the proposed closed LR testing procedure acts on real data. To this end, some well-known examples in the multivariate literature, in addition to those already presented in Section 2, will be used. A nominal level of 0.05 is adopted hereafter for the FWER-control. We will compare our approach with the likelihood-based Information Criteria (IC) summarized in Table 3.

Note that a preliminary Mardia's test is performed on each set of sample data to check the underlying assumption of multivariate normality. The R-code (R Development Core Team, 2012) providing the ML-estimates in $\widetilde{\mathcal{M}}$, and implementing the closed LR testing procedure, with both adjusted and unadjusted p-values, is available at `http://www.economia.unict.it/punzo`. In particular, to maximize L_M in (2) for the models requiring an IP, we employed the general purpose optimizer `optim`, in the R-package `stats`.

Example 1: Iris data

This data set was made famous by Fisher (1936) illustrating discriminant analysis. The sample contains $n = 150$ observations on $n_h = 50$ flowers, $h = 1, 2, 3$, from each of $k = 3$ species of iris: *setosa*, *versicolor* and *virginica*. The $p = 4$ variables, measured in centimeters, are: *sepal length*, *sepal width*, *petal length*, and *petal width*. The matrix of scatter plots for the grouped-data is shown in Figure 4. Mardia's test is consistent with multivariate normality in each of the 3 groups.

Sample volumes, sample shape-matrices and sample orientation-matrices, for each group, are given in Table 4. A glance to Table 4 and Figure 4 indicates that the covariance structure among

groups is quite different, although some similarities in terms of shape are observed. The LR test for H_0^{EEE} is rejected ($p_{\text{EEE}} \approx 0$). Also, at the 0.05-level, heteroscedasticity has been detected for these data by Flury (1984, 1988, Example 4.4) after the use of the CPC, and by Greselin *et al.* (2011) via an augmentation multiple testing procedure conceived for the family of hypotheses composed by homoscedasticity, homometroscedasticity (EEV), and homotroposcedasticity (VVE).

Additional information about the degree of similarity between the three covariance matrices can be gained by employing the closed LR testing procedure. The diagram in Figure 5 lists unadjusted p-values, for all the hypotheses in the hierarchy, and adjusted p-values (in round brackets) for the elementary hypotheses. This way of proceeding, in line with Westfall and Young (1993, p. 10), highlights their disparity. The hierarchical arrangement of Figure 5 better illustrates how to compute the adjusted p-values. Thus, for example, the adjusted p-value for H_0^{EVV} is given by $q_{\text{EVV}} = \max\{p_{\text{EVV}}, p_{\text{EEV}}, p_{\text{EVE}}, p_{\text{EEE}}\} = \max\{5.20 \cdot 10^{-12}, 5.17 \cdot 10^{-11}, 0, 0\} = 5.17 \cdot 10^{-11}$.

At the 0.05-level, H_0^{VEV} is the only elementary hypothesis that is not rejected in $\mathcal{H}$ by the closed LR testing procedure (see Figure 5). Table 5 contains further details about the models, the results obtained by the LR testing procedure, and the adopted IC. Here all the considered IC agree with the closed LR testing procedure.

To summarize, the proposed inferential approach allows us to conclude that scatters in the three iris species differ in orientation and volume but are consistent with having the same shape. We have obtained a more parsimonious model with a gain of 6 parameters with respect to VVV. The analysis could be made even more accurate and parsimonious by applying the closed LR testing procedure to the two groups *versicolor* and *virginica* only, omitting *setosa* that appear well separated from them (see Flury, 1984, Example 1).

Example 2: Bank notes data

From Flury (1988, pp. 51–56), the $p = 2$ variables “width, measured on left side” (LEFT) and “width, measured on right side” (RIGHT), measured (in millimeters) on former Swiss 1000-franc

bank notes, will be considered. There are $k = 2$ groups of bills, genuine (group 1) and forged (group 2), each of them consisting of $n_h = 100$, $h = 1, 2$, observations. Mardia's test assesses bivariate normality for the group of forged bills, but it leaves some doubt with respect to the genuine one. However, we will proceed with the analysis, as Flury (1988) also does. The LR test of homoscedasticity rejects the null hypothesis providing a p-value $p_{\text{EEE}} = 0.00258$.

Table 6 summarizes the decomposition (1) in each group, and Figure 6 shows the scatter plots with the ellipses of equal (95%) concentration superimposed. The scatter plots for genuine and forged bank notes are separately displayed to avoid a large overlapping between sample points. Although the LR test for EEE points out heteroscedasticity, the data in Table 6 suggest some similarity in orientation and shape, while some difference in their volume emerges for the two ellipses in Figure 6(a) and Figure 6(b). This conjecture is statistically corroborated by Flury (1988, p. 154): within his hierarchy of similarity between the k covariance matrices, he leans towards PCM, that coincides with VEE.

Results obtained by the closed LR testing procedure, given in Figure 7 and Table 7, confirm the validity of VEE. Some concern arises when noting how different IC can lead to different choices. In particular, the AIC leads to VEV, while the remaining ones choose VEE. However, this disagreement may be partly illuminated by the differences in the complexity penalizations of the criteria (e.g., the AIC features a rather weak penalization whereas the penalization of the BIC is more stringent).

Example 3: Crab data

The crab data set introduced in Section 2, and graphically represented in Figure 2(a), will be considered. For sake of brevity, we present only the final findings by means of Table 8 that corroborate the considerations we made previously; in particular, at the 0.05-level, H_0^{EEV} is not rejected because its components H_0^{EVV} and H_0^{VEV} are also not rejected. Moreover, comparing q_{VEV} with q_{EVV}, we note that the former provides stronger evidence of similarity between groups in terms of shape.

On the contrary, the AIC and the AIC_3 lean towards the less parsimonious VEV (as VEV adds one parameter with respect to EEV).

Example 4: Turtle data

The turtle data set introduced in Section 2, and graphically represented in Figure 2(b), will be taken into account. The overall results are given in Table 9. We can see that H_0^{VEV} is the only elementary hypothesis that is not rejected at the 0.05-level, meaning that VEV is able to describe the covariance structure between groups with a slight gain of one parameter with respect to heteroschedasticity (VVV) assessed by the LR test of homoscedasticity.

6 A simulation study

We present here the results of a simulation study, implemented in R, developed with a twofold aim: to show the performance of the proposed closed LR testing procedure in different situations of gradual departure from homoscedasticity (EEE), and to ascertain that the actual distribution of the test statistic LR_M, $M \in \overline{\mathcal{M}}$, can be well approximated by the expected χ^2 with ν_M degrees of freedom.

6.1 Performance evaluation of the closed LR testing procedure

We measure how effective is the proposed testing procedure in assessing the correct structure between groups by measuring the acceptance rates in different settings. As a lot of factors come into play (the number of groups k and their size, the dimension p of the observed variables, the overall sample size n, the nominal level α of the test, the volume, shape, and orientation components of the eigen decomposition), some of them have been necessarily considered fixed.

Three different scenarios have been taken into account. We work with bivariate data ($p = 2$), and $k = 2$ groups of equal size ($n_1 = n_2$). We have considered overall sample sizes of $n = 100$ and

$n = 200$.

In the first scenario, data have been drawn from the bivariate normal distribution, with mean vectors $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Each $\boldsymbol{\Sigma}_h$, according to the eigen decomposition (1), can be written as

$$\boldsymbol{\Sigma}_h = \lambda_h \boldsymbol{\Gamma}_h \boldsymbol{\Delta}_h \boldsymbol{\Gamma}_h' = \lambda_h \boldsymbol{R}(\vartheta_h) \begin{pmatrix} 1/\delta_h & 0 \\ 0 & \delta_h \end{pmatrix} \boldsymbol{R}(\vartheta_h)', \tag{5}$$

where

$$\boldsymbol{R}(\vartheta_h) = \begin{pmatrix} \cos\vartheta_h & -\sin\vartheta_h \\ \sin\vartheta_h & \cos\vartheta_h \end{pmatrix}$$

is the rotation matrix of angle ϑ_h, and $\delta_h \in (0, 1]$. The elements in the shape matrix arise from the constraint: $|\boldsymbol{\Delta}_h| = 1$. Hence, we have a single parameter for each eigen decomposition term: λ_h is the volume parameter, δ_h is the shape parameter, and ϑ_h is the orientation parameter. In the case $p = 2$, this is an alternative way to explain why $\boldsymbol{\Sigma}_h$, as well as its eigen decomposition, has three free parameters, $h = 1, \ldots, k$ (analogous considerations hold true for $p > 2$).

To simulate different ways of gradual departure from homoscedasticity, after setting $\boldsymbol{\Sigma}_1$ by $\lambda_1 = 1$, $\delta_1 = 0.7$, and $\vartheta_1 = 0$, we have considered some values for λ_2, δ_2, and ϑ_2 to specify $\boldsymbol{\Sigma}_2$: three values for λ_2 (1, 2, and 3), two values for δ_2 (0.7 and 0.3), and three values for ϑ_2 (0, $\pi/4$, and $\pi/2$ in radiants, i.e., 0°, 45°, and 90° in degrees). All the 18 combinations of these three parameters have been taken into account in the simulations.

Table 10 displays the simulated acceptance rates for this first setting. The rates refer to a nominal level of 0.05 and are calculated simulating 1000 samples for each of the 18 possible settings of $(\lambda_2, \delta_2, \vartheta_2)$ and for both the values of n. Bold numbers highlight the true model in each setting. As expected, the proposed procedure appears to be consistent since its power increases with n. To give an example, for model VVE given by $(\lambda_2 = 2, \delta_2 = 0.3, \vartheta_2 = 0)$, the simulated acceptance rate raises from 0.783 to 0.948 when $n = 100$ moves to $n = 200$. Moreover, given n, the acceptance rates for the true model gradually increase in line with its strength; thus, for example,

with reference to the case $n = 100$, model VEE is better assessed for $\lambda_2 = 3$ (simulated acceptance rate of 0.951), than for $\lambda_2 = 2$ (simulated acceptance rate of 0.769). Analogous considerations hold for model VEV in the case $\lambda_2 = 3$, for $n = 200$; here, the acceptance rate increases from 0.826 to 0.907 when ϑ_2 increases from $\pi/4$ to $\pi/2$. It is also interesting to note that, with reference to the complete null (EEE), the procedure is able to maintain the size in the case $n = 100$; in the case $n = 200$, in accordance with the classical MTPs, the procedure appears to be slightly conservative (simulated acceptance rate of 0.967).

To verify the robustness of the procedure with respect to departures from normality, we have considered two further scenarios, where data have been drawn from two bivariate *t*-distributions with mean vectors $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$, scale matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, and degrees of freedom $\nu_1 = \nu_2 = 10$, in the first case, and $\nu_1 = \nu_2 = 5$ in the second one. We recall that the covariance matrix of a multivariate *t* distribution is proportional to the scale matrix $\boldsymbol{\Sigma}$ via the relation $\nu/(\nu - 2)\boldsymbol{\Sigma}$. For simplicity, the same scheme above described for the normal case has been applied to the scale matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ when drawing samples from the *t* distribution.

Table 11 displays the simulated acceptance rates for the second scenario. As before, the rates refer to a nominal level of 0.05 and are calculated simulating 1000 samples for each of the 18 possible settings of $(\lambda_2, \delta_2, \vartheta_2)$ and for both values of *n*; bold numbers are in correspondence to the true model in each setting. The performance of the test worsens (compare Table 10 with Table 11) when considering samples drawn from the *t* distribution with 10 degrees of freedom. Only as an example, compare the simulated acceptance rates under the complete null in both tables. When the degrees of freedom decrease to 5, as expected, the simulations give even worse results, here omitted for sake of brevity. This attests the importance, for the proposed procedure, of the underlying assumption of multivariate normality in each group.

6.2 Asymptotic χ^2 distributions of the local LR statistics

We are interested in comparing the simulated distribution function (SDF) of the test statistic LR_M, $M \in \overline{\mathcal{M}}$, with its asymptotic χ^2-distribution with ν_M degrees of freedom. According to the simulation scheme of Section 6.1, we have considered: sample sizes of $n = 100$ and $n = 200$, bivariate data ($p = 2$), and $k = 2$ groups of equal size ($n_1 = n_2$). As before, data in group 1 are generated with parameters $\lambda_1 = 1$, $\delta_1 = 0.7$, and $\vartheta_1 = 0$, while departures in volume, shape, and orientation are obtained by specifying $\lambda_2 = 3$, $\delta_2 = 0.3$, and $\vartheta_2 = \pi/2$, respectively. Thus, for example, data under H_0^{VEV} have been randomly generated from two bivariate normal distributions having zero-mean and covariance matrices respectively defined by $(\lambda_1 = 1, \delta_1 = 0.7, \vartheta_1 = 0)$ and $(\lambda_2 = 3, \delta_2 = 0.7, \vartheta_2 = \pi/2)$.

Since the obtained results are very similar between models, we report only those referred to EVV (see Figure 8). A quick look at the plots confirms that the discrepancy between the curves is negligible; moreover, as expected, the increase of n seems to improve this agreement. These results confirms the validity of the theoretical asymptotic χ^2 distribution in our context.

7 Concluding remarks and discussion

This paper shows how some relevant configurations of similarity between covariance matrices $\mathbf{\Sigma}_h$, referring to k normal groups, can be described by considering the three-term eigen decomposition $\mathbf{\Sigma}_h = \lambda_h \mathbf{\Gamma}_h \mathbf{\Delta}_h \mathbf{\Gamma}_h'$. Each of these terms denotes specific geometric features (volume, shape, and orientation). This approach leads to eight different models by allowing each of the terms to be common or not between groups. However, no statistical test to assess the "correct" model among them exists and, still today, the *omnibus* Box test of homoscedasticity (versus heteroscedasticity) is widely used; unfortunately, being *omnibus*, after a rejection of the null hypothesis it leaves the practitioner without any further information. In this paper, such a gap has been addressed by providing a closed multiple testing procedure, using local likelihood ratio tests, to assess the choice

between the eight models. Although, in principle, an information criterion could be employed, a large number of these criteria has been proposed in literature (possibly leading to different choices) and practitioners tend to use one of them almost randomly or routinely. On the other hand, the closed LR test proposed here offers a straightforward assessment of covariance configurations and is based on only one subjective element, the significance level α, whose meaning is clear to everyone. Moreover, the adjusted *p*-values provide also a measure of "how significant" the test result is. We have evaluated, through simulations, the ability of the proposed procedure to detect the true underlying structure. Further, the real data sets analyzed in this paper have shown the gain in information and in parsimony that can be obtained via the new inferential method.

A further remark refers to the field of application that is not restricted to the (completely) supervised paradigm, as it could be inferred by the underlying assumptions. Our proposal provides indeed a suitable way to select a common covariance structure on labelled data, hence it can be naturally translated into the framework of model-based classification (see, e.g., Bagnato *et al.*, 2013). The estimation of the underlying mixture model could be also improved by employing this procedure in semi-supervised mode.

A direct application of this test can be devised in the field of biology, where a researcher may be interested in studying isometric or allometric scaling with respect to groups induced by species, gender, age, and so on. Isometric scaling occurs when changes in size (during growth or over evolutionary time) do not lead to changes in proportion. An example is found in frogs: aside from a brief period during the few weeks after metamorphosis, they grow isometrically (see Emerson, 1978). Instead, allometric scaling is any change that deviates from isometry. A classic example is the skeleton of mammals, that becomes much more robust and massive, relative to the size of the body, as the body grows (Schmidt-Nielsen, 1984). Three of the four data sets we employed come from biology. However, our procedure is far more widely applicable, as illustrated by the Swiss bank notes example herein.

The likelihood ratio tests for patterned covariance matrices, used as local α-level tests in our

procedure, are however sensitive to violations of the normality assumption (as also noted in the simulation results of Section 6). The problem of their generalization to wider classes of distributions has generated a huge amount of literature (see, e.g., Hallin and Paindaveine, 2009, and the references therein, for EEE, VEE and EVV; and Boente and Orellana, 2004, for VEE). Future work will consider extending these results to the remaining local α-level tests in our hierarchy to robustify the multiple testing procedure.

References

Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akademiai Kiado.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, 2nd edition.

Bagnato, L., Greselin, F., and Punzo, A. (2013). On the spectral decomposition in normal discriminant analysis. *Communications in Statistics - Simulation and Computation*. DOI: 10.1080/03610918.2012.735318.

Bensmail, H. and Celeux, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, **91**(436), 1743–1748.

Biernacki, C., Celeux, G., Govaert, G., Langrognet, F., Noulin, G., and Vernaz, Y. (2008). *MIXMOD - Statistical Documentation*. downloadable from `http://www.mixmod.org/IMG/pdf/statdoc_2_1_1.pdf`.

Boente, G. and Orellana, L. (2004). Robust plug-in estimators in proportional scatter models. *Journal of Statistical Planning and Inference*, **122**(1–2), 95–110.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**(3), 345–370.

Bozdogan, H. (1994). Theory & methodology of time series analysis. In *Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach*, volume 1, Dordrecht. Kluwer Academic Publishers.

Bretz, F., Hothorn, T., and Westfall, P. (2011). *Multiple Comparisons Using R.* Chapman & Hall, London.

Campbell, N. A. and Mahon, R. J. (1974). A multivariate study of variation in two species of rock crab of genus Leptograpsus. *Australian Journal of Zoology*, **22**(3), 417–425.

Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics & Probability Letters*, **44**(4), 333–344.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, **28**(5), 781–793.

Christensen, R. (1996). *Analysis of Variance, Design and Regression: Applied Statistical Methods*. Chapman & Hall, London.

Emerson, S. (1978). Allometry and jumping in frogs: helping the twain to meet. *Evolution*, **32**(3), 551–564.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**(2), 179–188.

Flury, B. N. (1984). Common principal components in *k* groups. *Journal of the American Statistical Association*, **79**(388), 892–898.

Flury, B. N. (1986). Proportionality of *k* covariance matrices. *Statistics & Probability Letters*, **4**(4), 29–33.

Flury, B. N. (1988). *Common Principal Components and Related Multivariate Models*. Wiley, New York.

Flury, B. N., Schmid, M., and Narayanan, A. (1994). Error rates in quadratic discrimination with constraints on the covariance matrices. *Journal of Classification*, **11**(1), 101–120.

Giancristofaro Arboretti, R., Bolzan, M., Bonnini, S., Corain, L., and Solmi, F. (2012). Advantages of the closed testing method in multiple comparisons procedures. *Communications in Statistics - Simulation and Computation*, **41**(6), 746–763.

Greselin, F., Ingrassia, S., and Punzo, A. (2011). Assessing the pattern of covariance matrices via an augmentation multiple testing procedure. *Statistical Methods & Applications*, **20**(2), 141–170.

Hallin, M. and Paindaveine, D. (2009). Optimal tests for homogeneity of covariance, scale, and shape. *Journal of Multivariate Analysis*, **100**(3), 422–444.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**(2), 65–70.

Jolicoeur, P. and Mosimann, J. (1960). Size and shape variation in the painted turtle: A principal component analysis. *Growth*, **24**(4), 339–354.

Manly, B. F. J. and Rayner, J. C. W. (1987). The comparison of sample covariance matrices using likelihood ratio tests. *Biometrika*, **74**(4), 841–847.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**(3), 655–660.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rencher, A. C. and Christensen, W. F. (2012). *Methods of Multivariate Analysis*. Wiley, New York, 3rd edition.

Schmidt-Nielsen, K. (1984). *Scaling, why is animal size so important?* Cambridge University Press, Cambridge.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Westfall, P. and Wolfinger, R. (2000). *Closed multiple testing procedures and PROC MULTTEST*. Available at `http://support.sas.com/kb/22/addl/fusion22950_1_multtest.pdf`.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*. Wiley, New York.

Table 1: Models in $\widetilde{\mathcal{M}}$ described by their covariance restrictions

M	Volume λ_h	Shape $\mathbf{\Delta}_h$	Orientation $\mathbf{\Gamma}_h$	Decomposition of $\mathbf{\Sigma_h}$	Relationship with existing models
EEE	Equal	Equal	Equal	$\lambda\mathbf{\Gamma\Delta\Gamma}'$	Homoscedasticity
VEE	Variable	Equal	Equal	$\lambda_h\mathbf{\Gamma\Delta\Gamma}'$	Proportional Covariance Matrices (PCM; Flury, 1986)
EVE	Equal	Variable	Equal	$\lambda\mathbf{\Gamma\Delta}_h\mathbf{\Gamma}'$	-
EEV	Equal	Equal	Variable	$\lambda\mathbf{\Gamma}_h\mathbf{\Delta\Gamma}_h'$	Homometroscedasticity (Greselin *et al.*, 2011)
VVE	Variable	Variable	Equal	$\lambda_h\mathbf{\Gamma\Delta}_h\mathbf{\Gamma}'$	Common Principal Components[a] (CPC; Flury, 1984) Homotroposcedasticity (Greselin *et al.*, 2011)
VEV	Variable	Equal	Variable	$\lambda_h\mathbf{\Gamma}_h\mathbf{\Delta\Gamma}_h'$	-
EVV	Equal	Variable	Variable	$\lambda\mathbf{\Gamma}_h\mathbf{\Delta}_h\mathbf{\Gamma}_h'$	-
VVV	Variable	Variable	Variable	$\lambda_h\mathbf{\Gamma}_h\mathbf{\Delta}_h\mathbf{\Gamma}_h'$	Heteroscedasticity

[a] If the constraint on the decreasing order of the diagonal elements of $\mathbf{\Delta}_h$ is relaxed, $h = 1, \ldots, k$.

Table 2: Scheme of computation of ν_M, for the asymptotic χ^2-approximation of LR_M, starting from η_{VVV} and η_M, for each model $M \in \widetilde{\mathcal{M}}$.

M	ML	η_{VVV}		η_M		ν_M
EEE	CF	$k\frac{p(p+1)}{2}$	$-$	$\frac{p(p+1)}{2}$	$=$	$(k-1)\frac{p(p+1)}{2}$
VEE	IP	$k\frac{p(p+1)}{2}$	$-$	$\frac{p(p+1)}{2}+(k-1)$	$=$	$(k-1)\left(\frac{p(p+1)}{2}-1\right)$
EVE	IP	$k\frac{p(p+1)}{2}$	$-$	$\frac{p(p+1)}{2}+(k-1)(p-1)$	$=$	$(k-1)\left(\frac{p(p-1)}{2}+1\right)$
EEV	CF	$k\frac{p(p+1)}{2}$	$-$	$k\frac{p(p+1)}{2}-(k-1)p$	$=$	$(k-1)p$
VVE	IP	$k\frac{p(p+1)}{2}$	$-$	$\frac{p(p+1)}{2}+(k-1)p$	$=$	$(k-1)\frac{p(p-1)}{2}$
VEV	IP	$k\frac{p(p+1)}{2}$	$-$	$k\frac{p(p+1)}{2}-(k-1)(p-1)$	$=$	$(k-1)(p-1)$
EVV	CF	$k\frac{p(p+1)}{2}$	$-$	$k\frac{p(p+1)}{2}-(k-1)$	$=$	$(k-1)$
VVV	CF	$k\frac{p(p+1)}{2}$	$-$	$k\frac{p(p+1)}{2}$	$=$	0

Table 3: Definitions and references for the adopted likelihood-based information criteria.

Information Criterion	Definition	Reference
AIC	$-2l_M + 2\eta_M$	Akaike (1973)
AIC_3	$-2l_M + 3\eta_M$	Bozdogan (1994), Cavanaugh (1999)
BIC	$-2l_M + \eta_M \ln n$	Schwarz (1978)
CAIC	$-2l_M + \eta_M (1 + \ln n)$	Bozdogan (1987)

Table 4: Three-term decomposition of $\boldsymbol{\Sigma}_h^{\text{VVV}}$, $h = 1, 2, 3$, in the Fisher iris data.

Group	Volume	Shape	Orientation
setosa ($n_1 = 50$)	$\lambda_1^{\text{VVV}} = 0.037$	$\boldsymbol{\Delta}_1^{\text{VVV}} = \begin{bmatrix} 6.202 & 0 & 0 & 0 \\ 0 & 0.968 & 0 & 0 \\ 0 & 0 & 0.703 & 0 \\ 0 & 0 & 0 & 0.237 \end{bmatrix}$	$\boldsymbol{\Gamma}_1^{\text{VVV}} = \begin{bmatrix} 0.669 & -0.599 & -0.440 & 0.036 \\ 0.734 & 0.621 & 0.275 & 0.020 \\ 0.097 & -0.490 & 0.832 & 0.240 \\ 0.064 & -0.131 & 0.195 & -0.970 \end{bmatrix}$
versicolor ($n_2 = 50$)	$\lambda_2^{\text{VVV}} = 0.065$	$\boldsymbol{\Delta}_2^{\text{VVV}} = \begin{bmatrix} 7.396 & 0 & 0 & 0 \\ 0 & 1.097 & 0 & 0 \\ 0 & 0 & 0.830 & 0 \\ 0 & 0 & 0 & 0.148 \end{bmatrix}$	$\boldsymbol{\Gamma}_2^{\text{VVV}} = \begin{bmatrix} 0.687 & 0.669 & -0.265 & 0.102 \\ 0.305 & -0.567 & -0.730 & -0.229 \\ 0.624 & -0.343 & 0.627 & -0.316 \\ 0.215 & -0.335 & 0.064 & 0.915 \end{bmatrix}$
virginica ($n_3 = 50$)	$\lambda_3^{\text{VVV}} = 0.105$	$\boldsymbol{\Delta}_3^{\text{VVV}} = \begin{bmatrix} 6.477 & 0 & 0 & 0 \\ 0 & 0.993 & 0 & 0 \\ 0 & 0 & 0.487 & 0 \\ 0 & 0 & 0 & 0.319 \end{bmatrix}$	$\boldsymbol{\Gamma}_3^{\text{VVV}} = \begin{bmatrix} 0.741 & -0.165 & 0.534 & 0.371 \\ 0.203 & 0.749 & 0.325 & -0.541 \\ 0.628 & -0.169 & -0.652 & -0.391 \\ 0.124 & 0.619 & -0.429 & 0.646 \end{bmatrix}$

Table 5: Fisher iris data. Details on: models $M \in \widetilde{\mathcal{M}}$, closed LR testing procedure, and likelihood-based IC. Bold numbers refer to the "not rejected" hypotheses in $\mathcal{H}$ at the 0.05-level (column q_M) and to the best model (minimum column value) for each IC.

M	η_M	LR_M	ν_M	p_M	q_M	$-2l_M$	AIC	AIC_3	BIC	CAIC
EEE	10	149.66	20	≈ 0		196.82	216.82	226.82	246.93	256.93
VEE	12	114.61	18	≈ 0		161.78	185.78	197.78	221.91	233.91
EVE	16	107.24	14	≈ 0		154.41	186.41	202.41	234.58	250.58
EEV	22	64.85	8	$5.17 \cdot 10^{-11}$		112.02	156.02	178.02	222.25	244.25
VVE	18	66.31	12	$1.56 \cdot 10^{-9}$	$1.56 \cdot 10^{-9}$	113.47	149.47	167.47	203.67	221.67
VEV	24	11.34	6	0.07831	**0.07831**	58.51	**106.51**	**130.51**	**178.77**	**202.77**
EVV	28	51.96	2	$5.20 \cdot 10^{-12}$	$5.17 \cdot 10^{-11}$	99.13	155.13	183.13	239.43	267.43
VVV	30					47.17	107.17	137.17	197.49	227.49

Table 6: Three-term decomposition of $\mathbf{\Sigma}_h^{\text{VVV}}$, $h = 1, 2$, in the Bank notes data.

Group	Volume	Shape	Orientation
genuine ($n_1 = 100$)	$\lambda_1^{\text{VVV}} = 0.096$	$\mathbf{\Delta}_1^{\text{VVV}} = \begin{bmatrix} 2.226 & 0 \\ 0 & 0.450 \end{bmatrix}$	$\mathbf{\Gamma}_1^{\text{VVV}} = \begin{bmatrix} 0.720 & -0.694 \\ 0.694 & 0.720 \end{bmatrix}$
forged ($n_2 = 100$)	$\lambda_2^{\text{VVV}} = 0.059$	$\mathbf{\Delta}_2^{\text{VVV}} = \begin{bmatrix} 2.088 & 0 \\ 0 & 0.479 \end{bmatrix}$	$\mathbf{\Gamma}_2^{\text{VVV}} = \begin{bmatrix} 0.613 & -0.790 \\ 0.790 & 0.613 \end{bmatrix}$

Table 7: Bank notes data. Details on: models $M \in \widetilde{\mathcal{M}}$, closed LR testing procedure, and likelihood-based IC. Bold numbers refer to the "not rejected" hypotheses in $\mathcal{H}$ at the 0.05-level (column q_M) and to the best model (minimum column value) for each IC.

M	η_M	LR_M	ν_M	p_M	q_M	$-2l_M$	AIC	AIC_3	BIC	CAIC
EEE	3	14.25	3	0.00258		115.53	121.53	124.53	131.43	134.43
VEE	4	3.10	2	0.21221		104.38	112.38	**116.38**	**125.58**	**129.58**
EVE	4	13.98	2	0.00092		115.26	123.26	127.26	136.45	140.45
EEV	4	11.51	2	0.00316		112.79	120.79	124.79	133.99	137.99
VVE	5	2.88	1	0.08946	**0.21221**	104.17	114.17	119.17	130.66	135.66
VEV	5	0.20	1	0.65122	**0.65122**	101.49	**111.49**	116.49	127.98	132.98
EVV	5	11.32	1	0.00077	0.00316	112.60	122.60	127.60	139.09	144.09
VVV	6					101.28	113.28	119.28	133.07	139.07

Table 8: crab data. Details on: models $M \in \widetilde{\mathcal{M}}$, closed LR testing procedure, and likelihood-based IC. Bold numbers refer to the "not rejected" hypotheses in $\mathcal{H}$ at the 0.05-level (column q_M) and to the best model (minimum column value) for each IC.

M	η_M	LR_M	ν_M	p_M	q_M	$-2l_M$	AIC	AIC_3	BIC	CAIC
EEE	3	69.09	3	$6.66 \cdot 10^{-15}$		834.61	840.61	843.61	848.43	851.43
VEE	4	67.42	2	$2.33 \cdot 10^{-15}$		832.95	840.95	844.95	851.37	855.37
EVE	4	67.89	2	$1.78 \cdot 10^{-15}$		833.42	841.42	845.42	851.84	855.84
EEV	4	3.25	2	0.19724		768.77	776.77	780.77	**787.19**	**791.19**
VVE	5	67.29	1	≈ 0	$6.66 \cdot 10^{-15}$	832.81	842.81	847.81	855.84	860.84
VEV	5	0.01	1	0.93579	**0.93579**	765.53	**775.53**	**780.53**	788.55	793.55
EVV	5	3.24	1	0.07185	**0.19724**	768.76	778.76	783.76	791.79	796.79
VVV	6					765.52	777.52	783.52	793.15	799.15

Table 9: turtle data. Details on: models $M \in \widetilde{\mathcal{M}}$, closed LR testing procedure, and likelihood-based IC. Bold numbers refer to the "not rejected" hypotheses in $\mathcal{H}$ at the 0.05-level (column q_M) and to the best model (minimum column value) for each IC.

M	η_M	LR_M	ν_M	p_M	q_M	$-2l_M$	AIC	AIC_3	BIC	CAIC
EEE	3	25.27	3	$1.36 \cdot 10^{-6}$		593.70	599.70	602.70	605.31	608.31
VEE	4	13.66	2	0.00108		582.09	590.09	594.09	597.58	601.58
EVE	4	23.97	2	$6.22 \cdot 10^{-6}$		592.41	600.41	604.41	607.89	611.89
EEV	4	14.97	2	0.00056		583.41	591.41	595.41	598.89	602.89
VVE	5	13.48	1	0.00024	0.00108	581.92	591.92	596.92	601.27	606.27
VEV	5	0.11	1	0.73504	**0.73504**	568.55	**578.55**	**583.55**	**587.90**	**592.90**
EVV	5	14.89	1	0.00011	0.00056	583.32	593.32	598.32	602.68	607.68
VVV	6					568.43	580.43	586.43	591.66	597.66

Table 10: Simulated acceptance rates from the closed LR testing procedure. Samples are drawn with sizes $n = 100$ and $n = 200$ by $k = 2$ groups of bivariate normal, varying the volume (λ_2), the shape (δ_2), and the orientation (ϑ_2) of the covariance matrix $\mathbf{\Sigma}_2$ with respect to $\mathbf{\Sigma}_1$ ($\lambda_1 = 1$, $\delta_1 = 0.7$, and $\vartheta_1 = 0$). Rates refer to 1000 replications and to a nominal level of 0.05. Bold numbers highlight the true model in each setting.

			$n = 100$								$n = 200$							
λ_2	δ_2	θ_2	EEE	VEE	EVE	EEV	VVE	VEV	EVV	VVV	EEE	VEE	EVE	EEV	VVE	VEV	EVV	VVV
1	0.7	0	**0.950**	0.020	0.012	0.018	0.000	0.000	0.000	0.000	**0.967**	0.014	0.010	0.007	0.002	0.000	0.000	0.000
		$\pi/4$	0.621	0.011	0.011	**0.351**	0.001	0.004	0.001	0.000	0.229	0.004	0.014	**0.727**	0.001	0.017	0.007	0.001
		$\pi/2$	0.479	0.008	0.016	**0.476**	0.002	0.019	0.000	0.000	0.126	0.001	0.017	**0.825**	0.000	0.023	0.007	0.001
	0.3	0	0.049	0.003	**0.889**	0.004	0.028	0.000	0.026	0.001	0.000	0.000	**0.946**	0.000	0.026	0.000	0.024	0.004
		$\pi/4$	0.000	0.000	0.340	0.044	0.010	0.000	**0.565**	0.041	0.000	0.000	0.085	0.000	0.002	0.000	**0.847**	0.066
		$\pi/2$	0.000	0.000	0.304	0.031	0.006	0.000	**0.620**	0.039	0.000	0.000	0.073	0.000	0.002	0.000	**0.877**	0.048
2	0.7	0	0.180	**0.769**	0.003	0.008	0.016	0.024	0.000	0.000	0.014	**0.942**	0.000	0.001	0.027	0.016	0.000	0.000
		$\pi/4$	0.073	0.431	0.002	0.078	0.035	**0.374**	0.000	0.007	0.001	0.113	0.001	0.005	0.019	**0.841**	0.000	0.020
		$\pi/2$	0.048	0.270	0.004	0.109	0.031	**0.536**	0.000	0.002	0.000	0.045	0.000	0.010	0.020	**0.901**	0.000	0.024
	0.3	0	0.007	0.033	0.126	0.000	**0.783**	0.001	0.002	0.048	0.000	0.000	0.007	0.000	**0.948**	0.000	0.000	0.045
		$\pi/4$	0.000	0.000	0.021	0.004	0.216	0.016	0.045	**0.698**	0.000	0.000	0.000	0.000	0.035	0.000	0.003	**0.962**
		$\pi/2$	0.000	0.000	0.024	0.002	0.193	0.014	0.069	**0.698**	0.000	0.000	0.000	0.000	0.036	0.000	0.003	**0.961**
3	0.7	0	0.001	**0.951**	0.000	0.001	0.028	0.019	0.000	0.000	0.000	**0.956**	0.000	0.000	0.022	0.021	0.000	0.001
		$\pi/4$	0.000	0.485	0.000	0.001	0.026	**0.479**	0.000	0.009	0.000	0.123	0.000	0.000	0.026	**0.826**	0.000	0.025
		$\pi/2$	0.001	0.297	0.000	0.000	0.041	**0.656**	0.000	0.005	0.000	0.035	0.000	0.000	0.022	**0.907**	0.000	0.036
	0.3	0	0.000	0.026	0.001	0.000	**0.914**	0.000	0.000	0.059	0.000	0.000	0.000	0.000	**0.945**	0.000	0.000	0.055
		$\pi/4$	0.000	0.000	0.001	0.000	0.245	0.015	0.000	**0.739**	0.000	0.000	0.000	0.000	0.044	0.000	0.000	**0.956**
		$\pi/2$	0.000	0.000	0.000	0.000	0.215	0.024	0.000	**0.761**	0.000	0.000	0.000	0.000	0.037	0.000	0.000	**0.963**

Table 11: Simulated acceptance rates from the closed LR testing procedure. Samples are drawn with sizes $n = 100$ and $n = 200$ by $k = 2$ groups of bivariate t with 10 degrees of freedom, varying the volume (λ_2), the shape (δ_2), and the orientation (ϑ_2) of the scalar matrix $\mathbf{\Sigma}_2$ with respect to $\mathbf{\Sigma}_1$ ($\lambda_1 = 1$, $\delta_1 = 0.7$, and $\vartheta_1 = 0$). Rates refer to 1000 replications and to a nominal level of 0.05. Bold numbers highlight the true model in each setting.

			$n = 100$								$n = 200$							
λ_2	δ_2	θ_2	EEE	VEE	EVE	EEV	VVE	VEV	EVV	VVV	EEE	VEE	EVE	EEV	VVE	VEV	EVV	VVV
1	0.7	0	**0.894**	0.057	0.023	0.019	0.005	0.002	0.000	0.000	**0.891**	0.046	0.027	0.025	0.008	0.002	0.000	0.001
		$\pi/4$	0.546	0.034	0.033	**0.355**	0.002	0.028	0.001	0.001	0.209	0.013	0.022	**0.677**	0.004	0.047	0.020	0.008
		$\pi/2$	0.464	0.024	0.032	**0.426**	0.003	0.048	0.002	0.001	0.125	0.006	0.022	**0.741**	0.001	0.074	0.028	0.003
	0.3	0	0.069	0.006	**0.799**	0.006	0.068	0.000	0.041	0.011	0.002	0.000	**0.861**	0.000	0.085	0.000	0.043	0.009
		$\pi/4$	0.001	0.000	0.322	0.049	0.023	0.004	**0.511**	0.090	0.000	0.000	0.091	0.000	0.007	0.001	**0.785**	0.116
		$\pi/2$	0.001	0.000	0.281	0.046	0.024	0.005	**0.553**	0.090	0.000	0.000	0.072	0.001	0.004	0.000	**0.818**	0.105
2	0.7	0	0.207	**0.701**	0.008	0.004	0.034	0.044	0.001	0.001	0.034	**0.878**	0.002	0.003	0.036	0.046	0.000	0.001
		$\pi/4$	0.103	0.385	0.012	0.089	0.035	**0.364**	0.001	0.011	0.005	0.155	0.001	0.018	0.033	**0.738**	0.001	0.049
		$\pi/2$	0.061	0.250	0.009	0.108	0.051	**0.505**	0.001	0.015	0.002	0.064	0.000	0.017	0.021	**0.856**	0.000	0.040
	0.3	0	0.014	0.037	0.161	0.001	**0.693**	0.008	0.004	0.082	0.000	0.000	0.029	0.000	**0.875**	0.000	0.000	0.096
		$\pi/4$	0.000	0.000	0.051	0.004	0.185	0.026	0.091	**0.643**	0.000	0.000	0.001	0.000	0.056	0.000	0.017	**0.926**
		$\pi/2$	0.000	0.000	0.044	0.005	0.193	0.030	0.086	**0.642**	0.000	0.000	0.001	0.000	0.052	0.000	0.015	**0.932**
3	0.7	0	0.008	**0.907**	0.002	0.000	0.039	0.044	0.000	0.000	0.104	**0.795**	0.000	0.000	0.055	0.045	0.000	0.001
		$\pi/4$	0.003	0.491	0.000	0.004	0.056	**0.440**	0.000	0.006	0.101	0.140	0.000	0.000	0.038	**0.692**	0.000	0.029
		$\pi/2$	0.003	0.321	0.000	0.004	0.060	**0.601**	0.000	0.011	0.095	0.045	0.000	0.000	0.028	**0.785**	0.000	0.047
	0.3	0	0.001	0.030	0.006	0.000	**0.869**	0.009	0.000	0.085	0.102	0.001	0.000	0.000	**0.815**	0.000	0.000	0.082
		$\pi/4$	0.000	0.000	0.004	0.000	0.273	0.029	0.001	**0.693**	0.093	0.000	0.000	0.000	0.059	0.000	0.000	**0.848**
		$\pi/2$	0.000	0.000	0.000	0.000	0.246	0.021	0.003	**0.730**	0.089	0.000	0.000	0.000	0.047	0.000	0.000	**0.864**

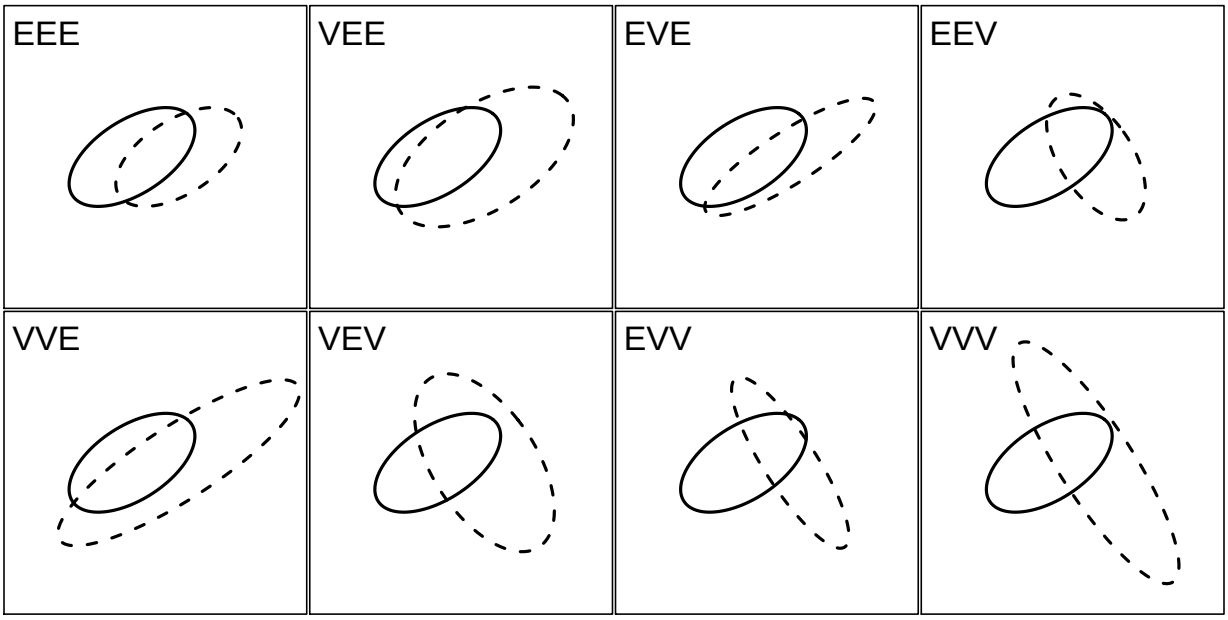
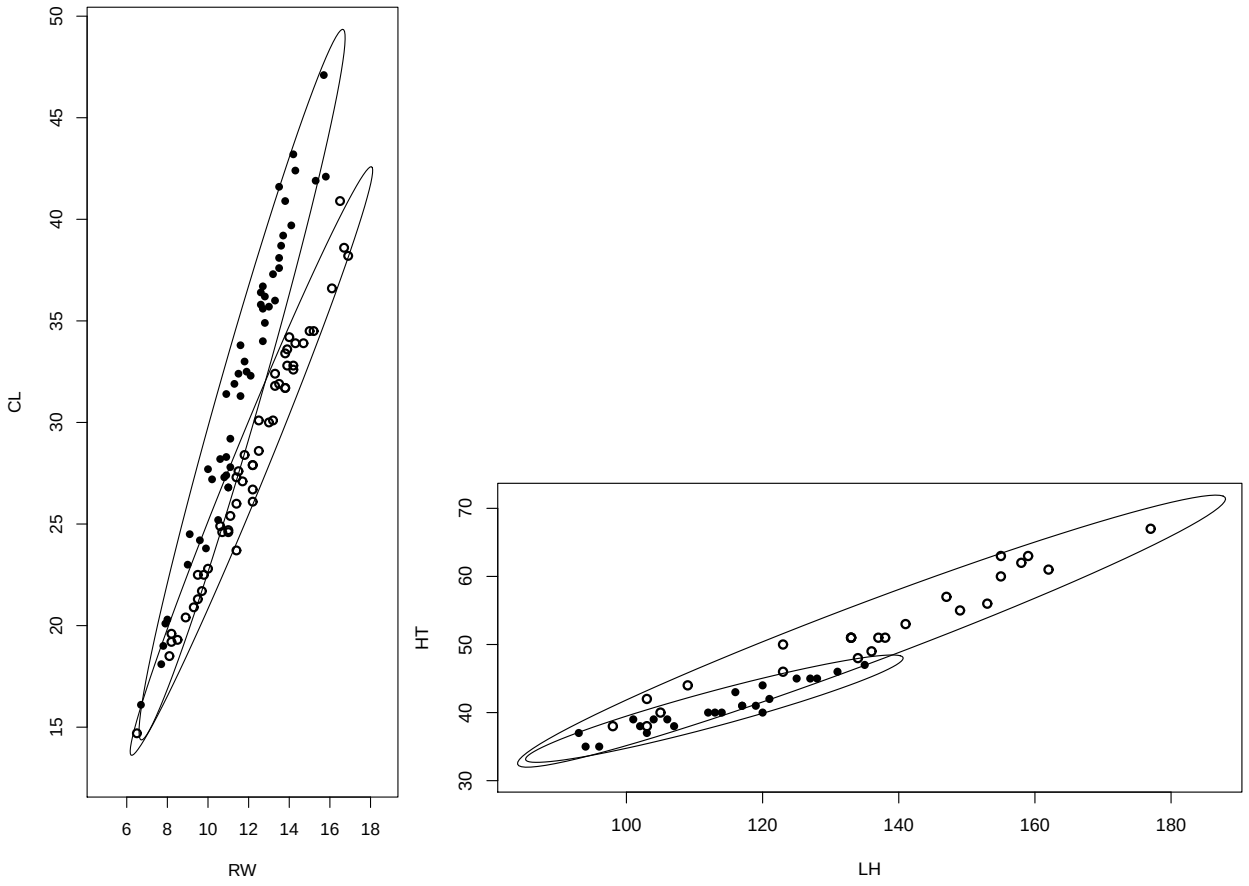


Figure 1: Examples of the models in $\widetilde{\mathcal{M}}$. The bivariate case ($p = 2$), and $k = 2$ groups, are considered.



(a) crab data (b) turtle data

Figure 2: Scatter plots, and related ellipses of equal (95%) concentration, for crab and turtle data (in both cases, ∘ denotes female and • male).

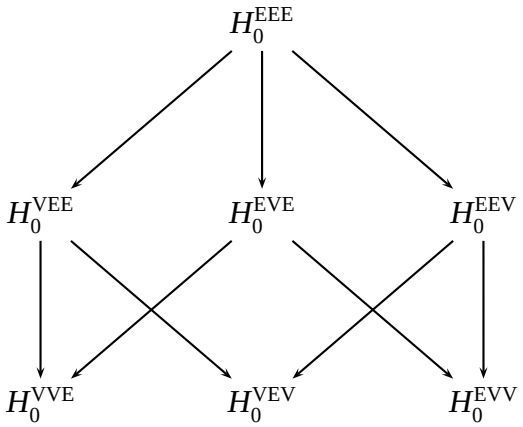


Figure 3: Graph of the hierarchy of relationships between the null hypotheses.

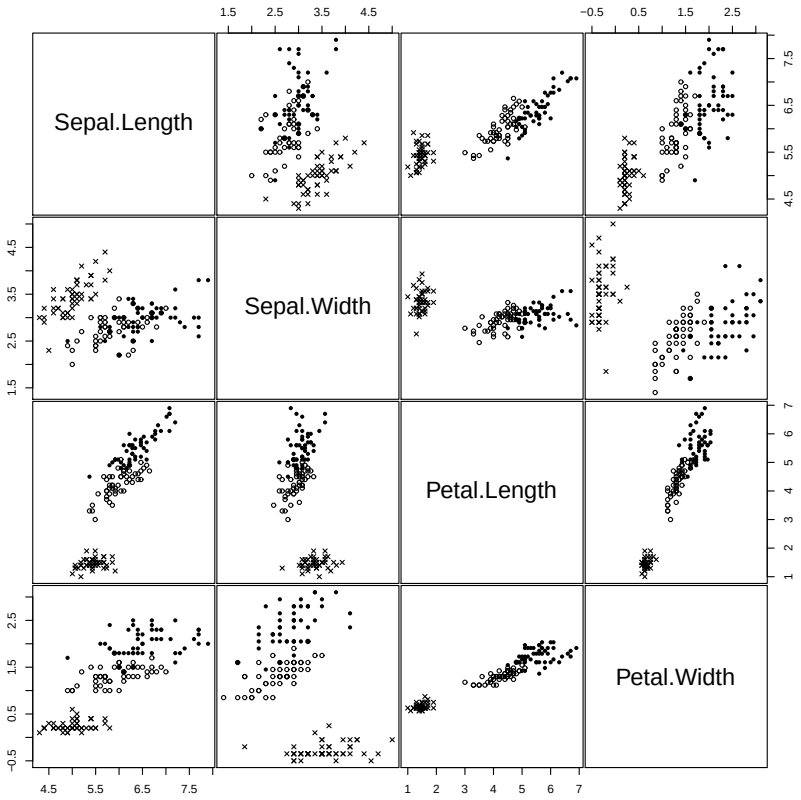


Figure 4: Matrix of scatter plots of Fisher's iris data (× denotes setosa, ∘ denotes versicolor, and • denotes virginica)

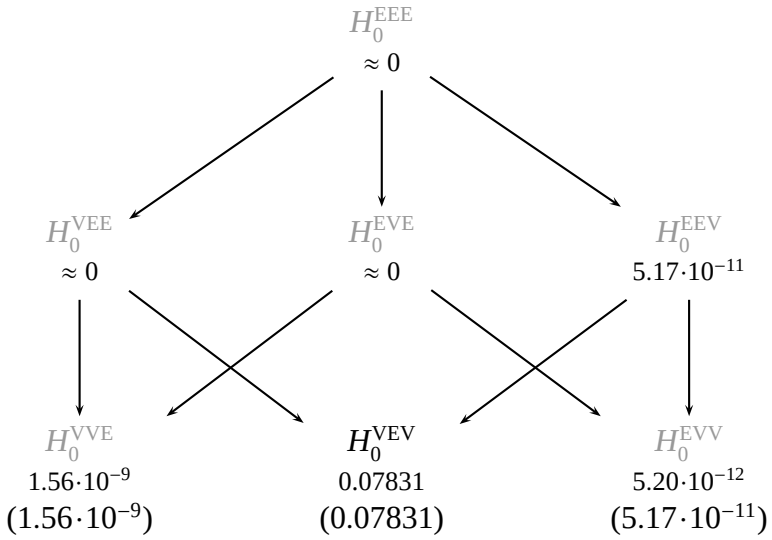
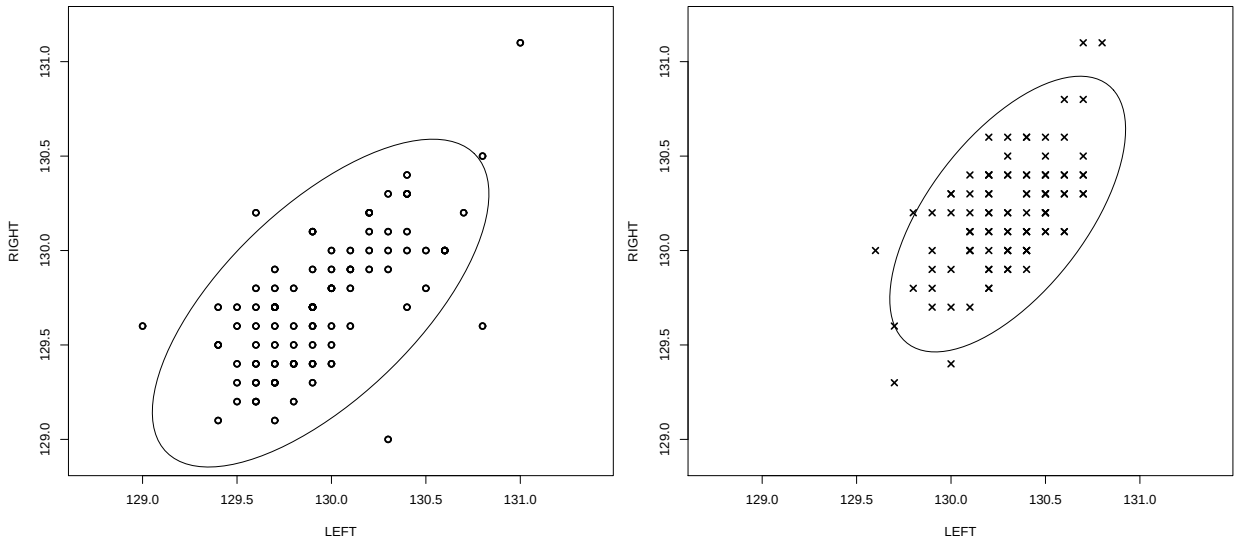


Figure 5: Unadjusted and adjusted *p*-values (in round brackets) related to the closed LR testing procedure applied to the Iris data set. Rejected hypotheses, at the 0.05-level, are displayed in gray.



(a) Genuine (b) Forged

Figure 6: Scatter plots, and related ellipses of equal (95%) concentration, of variables LEFT and RIGHT in two groups of Swiss bank notes. Coinciding points are marked by a single symbol only.

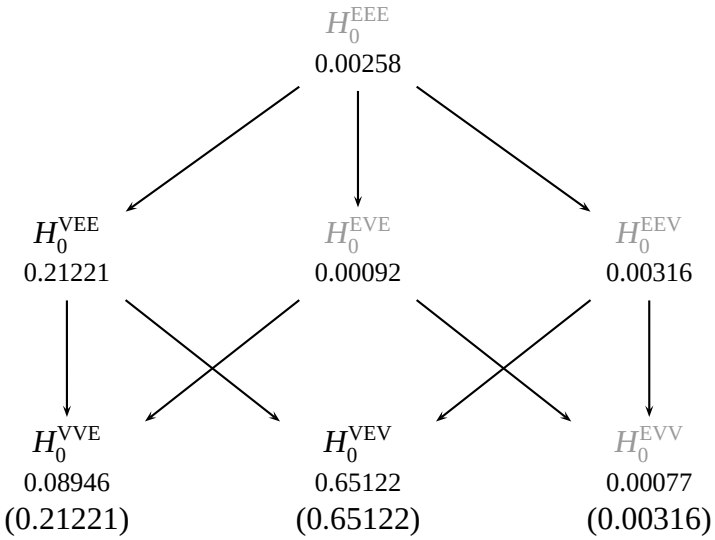
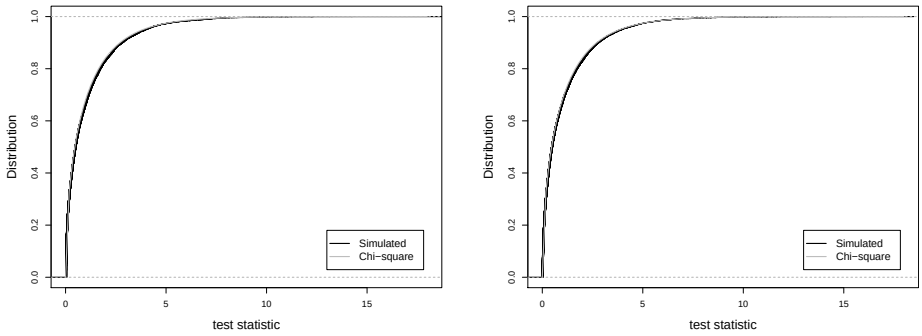


Figure 7: Unadjusted and adjusted *p*-values (in round brackets) related to the closed LR testing procedure applied to the bank notes data. Rejected hypotheses, at the 0.05-level, are displayed in gray.



(a) $n = 100$ (b) $n = 200$

Figure 8: Model EVV: SDF of the LR test statistic compared with the asymptotic χ^2 distribution, from 10000 replications.