Assessing the pattern of covariance matrices via an augmentation multiple testing procedure

**Francesca Greselin · Salvatore Ingrassia ·
Antonio Punzo**

Abstract This paper extends the scedasticity comparison among several groups of observations, usually complying with the homoscedastic and the heteroscedastic cases, in order to deal with data sets laying in an intermediate situation. As is well known, homoscedasticity corresponds to equality in orientation, shape and size of the group scatters. Here our attention is focused on two weaker requirements: scatters with the same orientation, but with different shape and size, or scatters with the same shape and size but different orientation. We introduce a multiple testing procedure that takes into account each of the above conditions. This approach discloses a richer information on the data underlying structure than the classical method only based on homo/heteroscedasticity. At the same time, it allows a more parsimonious parametrization, whenever the patterned model is appropriate to describe the real data. The new inferential methodology is then applied to some well-known data sets, chosen in the multivariate literature, to show the real gain in using this more informative approach. Finally, a wide simulation study illustrates and compares the performance of the proposal using data sets with gradual departure from homoscedasticity.

F. Greselin (✉)
Dipartimento di Metodi Quantitativi per le Scienze Economiche Aziendali,
Università di Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano, Italy
e-mail: francesca.greselin@unimib.it

S. Ingrassia · A. Punzo
Dipartimento di Impresa, Culture e Società,
Università di Catania, Corso Italia, 55, 95129 Catania, Italy
e-mail: s.ingrassia@unict.it

A. Punzo
e-mail: antonio.punzo@unict.it

1 Introduction

In multivariate analysis it is customary to deal with covariance matrices referred to several groups as different (*heteroscedastic*) if an overall test of equality (*homoscedasticity test*) states that they are not identical: in this case an unbiased version of the likelihood ratio (LR) test, known as Box's M-test, is commonly used, providing both χ^2- and F-approximations for the distribution of the LR statistic M under the assumption of multivariate normality (see, e.g., Rencher 1998, pp. 138–140).

However, in many applications, data groups escape from matching with homoscedasticity and one can still observe some sort of common basic structure (the same *pattern*) among covariance matrices. In contrast with the univariate case, in the multivariate setting the scatters may depart from equality in several different ways. Following the suggestions of Flury (1988), these different "ways" can be described by using the *eigenvalues* and *eigenvectors* representation of the covariances, according to the principal component analysis. Flury motivated his approach by recalling that, for example, male and female turtles (Jolicoeur and Mosimann 1960), human and animal bones (Jolicoeur 1963), genuine and forged bank notes (Flury and Riedwyl 1983), are clear examples in which some sort of equality between eigenvectors is quite plausible even if the usual homoscedasticity test suggests that the underlying population covariance matrices are not identical in all groups. The same author defined the common principal components (CPC) to summarize this situation.

On the other hand, Flury did not focus his attention on equality between eigenvalues, a condition which could also have interesting practical implications, as underlined in the following example. Clustering male and female blue crabs (Campbell and Mahon 1974) by a mixture model-based approach, Peel and McLachlan (2000) assumed that the two group-conditional distributions were bivariate normal with a common covariance matrix (see also McLachlan and Peel 2000, pp. 90–92) on the basis of the simultaneous test for multivariate normality and homoscedasticity proposed by Hawkins (1981). As pointed out by the authors, in this case the homoscedasticity assumption has a marked impact on the implied clustering of the data: indeed, the choice of a homoscedastic model produces a larger misallocation rate than the unconstrained model. Greselin and Ingrassia (2009, 2010) resolved this apparent contradiction by weakening the constraints and requiring only covariance matrices with the same set of (ordered) eigenvalues. From a geometrical point of view, this assumption means that covariance matrices are related to ellipsoids of equal concentration with the same size and shape. This particular condition is summarized by the authors through the introduction of the concept of *weak homoscedasticity*.

Patterned covariance matrices were firstly introduced by Murtagh and Raftery (1984) and systematically employed for Gaussian parsimonious clustering models in Celeux and Govaert (1995). Here, Flury's proposal and the above remarks are jointly considered with the aim of providing a unified framework for a more informative scedasticity comparison among groups.

The rest of the paper is organized as follows. In Sect. 2 the concepts of *homo-metroscedasticity*, denoting scatters with the same size and shape, and *homotropo-scedasticity*, meaning scatters with the same orientation, are introduced. Their logical intersection allows for a redefinition of homoscedasticity. In order to assess each of these new conditions, two statistical procedures are introduced in Sect. 3. All these "ingredients" are used in Sect. 4 to define an MTP (abbreviation of multiple testing procedure, a statistical method that is designed to take into account and properly control the multiplicity effect through some combined joint measure of erroneous inferences; Hochberg and Tamhane 1987, p. 2) that is more informative than the classical Box M-test. Here, we propose a union-intersection MTP with control of the false discovery exceedance (FDX; Van der Laan et al. 2004; Genovese and Wasserman 2006) by augmentation and we will motivate our choice. In Sect. 5, some well-known data sets in the multivariate literature are considered, to show real cases of homometroscedasticity and homotroposcedasticity. The new inferential method is hence applied to the same data, pointing out that it provides more information on the data underlying structure than the previous approach only based on homo/heteroscedasticity. In Sect. 6, we present the results of a wide simulation study which has been carried out in order to evaluate the performance of the proposed augmentation MTP. Concluding remarks are finally given in Sect. 7.

2 Basic definitions

Suppose p variables are measured on statistical units arising from $k \geq 2$ different groups and let $\boldsymbol{x}_1^{(h)}, \ldots, \boldsymbol{x}_{n_h}^{(h)}$ denote n_h independent observations, for the hth group, drawn from a normal distribution with mean vector $\boldsymbol{\mu}_h$ and covariance matrix $\boldsymbol{\Sigma}_h$, $h = 1, \ldots, k$. The null hypothesis of homoscedasticity

$$H_0 : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}, \tag{1}$$

versus the alternative hypothesis of heteroscedasticity (at least two $\boldsymbol{\Sigma}_h$'s are unequal), can be tested on the basis of the k independent samples, where $\boldsymbol{\Sigma}$ denotes the unknown common covariance matrix.

Let $\boldsymbol{\Sigma}_h = \boldsymbol{\Gamma}_h \boldsymbol{\Lambda}_h \boldsymbol{\Gamma}_h'$ be the spectral decomposition of the matrix $\boldsymbol{\Sigma}_h$, $h = 1, \ldots, k$, where $\boldsymbol{\Lambda}_h = \operatorname{diag}(\lambda_1^{(h)}, \ldots, \lambda_p^{(h)})$ is the diagonal matrix of the eigenvalues of $\boldsymbol{\Sigma}_h$ sorted in non-increasing order, i.e. $\lambda_1^{(h)} \geq \cdots \geq \lambda_p^{(h)} > 0$, and $\boldsymbol{\Gamma}_h$ is the $p \times p$ orthogonal matrix whose columns $\boldsymbol{\gamma}_1^{(h)}, \ldots, \boldsymbol{\gamma}_p^{(h)}$ are the normalized eigenvectors of $\boldsymbol{\Sigma}_h$ ordered according to their eigenvalues; here, and in what follows, the prime denotes the transpose of a matrix. The representation is unique (and we can call it *canonical*), provided that we have distinct eigenvalues and non-null elements on the diagonal of $\boldsymbol{\Lambda}_h$. Note also that the matrix $\boldsymbol{\Gamma}_h$ is the rotation in p-space that realizes the transformation of the data into their principal components. In line with Banfield and Raftery (1993), the greatest eigenvalue $\lambda_1^{(h)}$ denotes the *size* of the ellipsoids of equal concentration, while $\boldsymbol{\Lambda}_h/\lambda_1^{(h)}$ is their *shape*.

The spectral decomposition induces the following consideration: if k covariance matrices $\mathbf{\Sigma}_h$, $h = 1, \ldots, k$, are equal to some $\mathbf{\Sigma}$, as in (1), then they share both the matrix of eigenvalues $\mathbf{\Lambda}$, that is $\mathbf{\Lambda}_1 = \cdots = \mathbf{\Lambda}_k = \mathbf{\Lambda}$, and the matrix of orthonormalized eigenvectors $\mathbf{\Gamma}$, that is $\mathbf{\Gamma}_1 = \cdots = \mathbf{\Gamma}_k = \mathbf{\Gamma}$.

The condition $\mathbf{\Lambda}_1 = \cdots = \mathbf{\Lambda}_k = \mathbf{\Lambda}$ means that the ellipsoids of equal concentration referred to the k groups are equivalent to each other, up to some convenient rotation. In other words, they have the same shape and the same size. In an attempt to conform to the existing nomenclature, we will combine the three Greek words "*hòmoios*" (equal), "*mètron*" (measure) and "*skèdasis*" (dispersion) in the following way:

Definition 1 (*Homometroscedasticity*) Let $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_k$ be k covariance matrices, each with canonical spectral decomposition $\mathbf{\Sigma}_h = \mathbf{\Gamma}_h \mathbf{\Lambda}_h \mathbf{\Gamma}'_h$. If $\mathbf{\Lambda}_1 = \cdots = \mathbf{\Lambda}_k = \mathbf{\Lambda}$ the k covariance matrices are said to be **homometroscedastic**.

Note that homoscedasticity is homometroscedasticity in the univariate case. Moreover, in case of spherical scatters, homometroscedasticity is not wholly meaningful and it reduces only to equality in size.

On the other hand, k covariance matrices which share the same matrix of orthonormalized eigenvectors $\mathbf{\Gamma}$ have ellipsoids of equal concentration with the same axes *orientation* in p-space. In other words, they own the same ordered set of principal components and hence their ellipsoids are congruent up to a suitable dilation/contraction (alteration in size) and/or "deformation" (alteration in shape). We will denote this situation by using the Greek term "*tròpos*" (direction) in order to coin the following notion:

Definition 2 (*Homotroposcedasticity*) Let $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_k$ be k covariance matrices, each with canonical spectral decomposition $\mathbf{\Sigma}_h = \mathbf{\Gamma}_h \mathbf{\Lambda}_h \mathbf{\Gamma}'_h$. If $\mathbf{\Gamma}_1 = \cdots = \mathbf{\Gamma}_k = \mathbf{\Gamma}$ the k covariance matrices are said to be **homotroposcedastic**.

In Fig. 1 examples of covariance matrices having different shape, size and orientation are given in terms of ellipsoids of equal concentration. In more detail, both Fig. 1a and b represent covariance matrices that are homotroposcedastic but not homometroscedastic while Fig. 1c shows covariance matrices that are homometroscedastic but not homotroposcedastic. Naturally, such considerations allow us to re-formulate homoscedasticity, in terms of matrices for which Definition 1 and Definition 2 simultaneously hold. At the same time, such definitions represent two weaker conditions of similarity among k covariance matrices, that is, intermediate cases between homoscedasticity and heteroscedasticity.

3 Statistical tests for scedasticity analysis

In this section, in line with the more detailed approach to "scedasticity" given in Sect. 2, some multiple statistical tests are proposed. Consider the null hypothesis of homometroscedasticity

$$H_0^{\mathbf{\Lambda}} : \mathbf{\Lambda}_1 = \cdots = \mathbf{\Lambda}_k = \mathbf{\Lambda}, \tag{2}$$
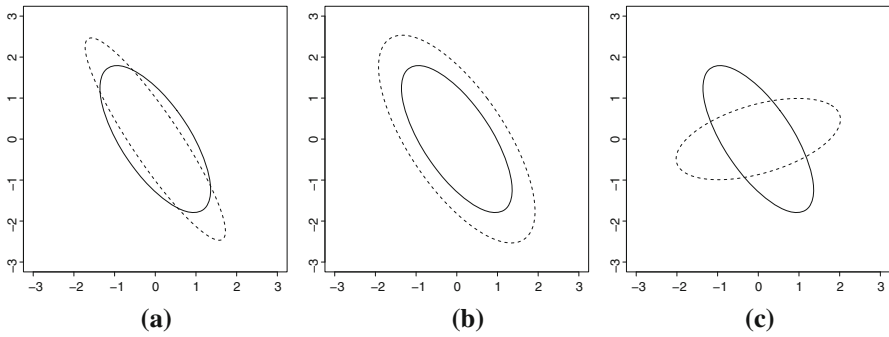
Fig. 1 Three different kinds of similarity between covariance matrices in terms of ellipses of equal concentration. The first two graphics are related to homotroposcedasticity while the last is connect to homometroscedasticity. **a** Different shape, **b** different size, **c** different orientation

versus the alternative

$$H_1^{\boldsymbol{\Lambda}} : \boldsymbol{\Lambda}_h \neq \boldsymbol{\Lambda}_l \quad \text{for some } h, l \in \{1, \ldots, k\}, \quad h \neq l. \tag{3}$$

Afterward, consider the null hypothesis of homotroposcedasticity

$$H_0^{\boldsymbol{\Gamma}} : \boldsymbol{\Gamma}_1 = \cdots = \boldsymbol{\Gamma}_k = \boldsymbol{\Gamma}, \tag{4}$$

versus the alternative

$$H_1^{\boldsymbol{\Gamma}} : \boldsymbol{\Gamma}_h \neq \boldsymbol{\Gamma}_l \quad \text{for some } h, l \in \{1, \ldots, k\}, \quad h \neq l. \tag{5}$$

Thus, the homoscedasticity test in (1) can be re-expressed as follows

$$H_0 : H_0^{\boldsymbol{\Lambda}} \cap H_0^{\boldsymbol{\Gamma}} \quad \text{versus} \quad H_1 : H_1^{\boldsymbol{\Lambda}} \cup H_1^{\boldsymbol{\Gamma}}. \tag{6}$$

Thus, H_0 is accepted against the alternative if, and only if, both $H_0^{\boldsymbol{\Lambda}}$ and $H_0^{\boldsymbol{\Gamma}}$ are accepted. Differently from a usual "direct" test of homoscedasticity, the present approach shows its profitability whenever the null hypothesis in (6) is rejected; indeed, the results for $H_0^{\boldsymbol{\Lambda}}$ and $H_0^{\boldsymbol{\Gamma}}$ permit us to investigate the nature of the departure from H_0.

3.1 Testing homometroscedasticity

Let $\overline{\boldsymbol{x}}_h$ and $\boldsymbol{S}_h$ be respectively the sample mean vector and the unbiased sample covariance matrix in the hth group, $h = 1, \ldots, k$. Moreover, let $\boldsymbol{G}_h$ be the $p \times p$ orthogonal matrix whose columns are the normalized eigenvectors of $\boldsymbol{S}_h$ ordered by the non-increasing sequence of the eigenvalues of $\boldsymbol{S}_h$, $h = 1, \ldots, k$. According to the *principal component (linear) transformation*

$$x_i^{(h)} \to y_i^{(h)} = G'_h \left(x_i^{(h)} - \overline{x}_h \right), \quad i = 1, \ldots, n_h,$$

the data $y_1^{(h)}, \ldots, y_{n_h}^{(h)}$, $h = 1, \ldots, k$, result uncorrelated, and their covariance matrix L_h is the diagonal matrix containing the non-negative eigenvalues of S_h. Since the assumption of multivariate normality holds, these components are also independent and normally distributed. Based on these results, the null hypothesis in (2) can be re-expressed as follows

$$H_0^{\mathbf{\Lambda}} : \bigcap_{j=1}^{p} H_0^{\lambda_j}, \tag{7}$$

where

$$H_0^{\lambda_j} : \lambda_j^{(1)} = \cdots = \lambda_j^{(k)} = \lambda_j,$$

and λ_j is the unknown jth eigenvalue, common to the k groups. Hence, the multivariate test for homometroscedasticity has been reduced to an equivalent set of p simultaneous Bartlett's tests of univariate homoscedasticity (see Bartlett 1937) among the k groups. Note that p simultaneous well known F-tests can be considered for the special case $k = 2$. All these techniques are implemented in most statistical software like the R libraries.

3.2 Testing homotroposcedasticity

Consider the spectral decomposition of $\mathbf{\Sigma}_h$, i.e. $\mathbf{\Sigma}_h = \mathbf{\Gamma}_h \mathbf{\Lambda}_h \mathbf{\Gamma}'_h$. Under the null hypothesis $H_0^{\mathbf{\Gamma}}$ given in (4), the $p \times p$ matrix $\mathbf{\Gamma}$ simultaneously diagonalizes all covariance matrices and generates the eigenvalue matrices $\mathbf{\Lambda}_h$, $h = 1, \ldots, k$. Thus the null hypothesis (4) can be equivalently re-formulated as follows

$$H_0^{\mathbf{\Gamma}} : \mathbf{\Gamma} \mathbf{\Lambda}_h \mathbf{\Gamma}' = \mathbf{\Sigma}_h, \quad h = 1, \ldots, k. \tag{8}$$

Under the assumption of multivariate normality, Flury (1984) derived the log-LR statistics for testing the weaker null hypothesis of common principal components

$$H_0^{\mathrm{CPC}} : \underset{\sim}{\mathbf{\Gamma}} \underset{\sim}{\mathbf{\Lambda}}_h \underset{\sim}{\mathbf{\Gamma}}' = \mathbf{\Sigma}_h, \quad h = 1, \ldots, k, \tag{9}$$

where $\underset{\sim}{\mathbf{\Gamma}}$ is a $p \times p$ orthonormalized matrix that simultaneously diagonalizes all covariance matrices, and $\underset{\sim}{\mathbf{\Lambda}}_h$ is one of the possible $p!$ diagonal matrices of eigenvalues in the hth group, $h = 1, \ldots, k$. Note that, in contrast with the formulation (8), no canonical ordering of the columns of $\underset{\sim}{\mathbf{\Gamma}}$ is specified here, since the rank order of the diagonal elements of the $\underset{\sim}{\mathbf{\Lambda}}_h$ is not necessarily the same for all groups. In order to apply Flury's proposal, the *F–G algorithm*, developed by Flury and Gautschi (1986) and coded in
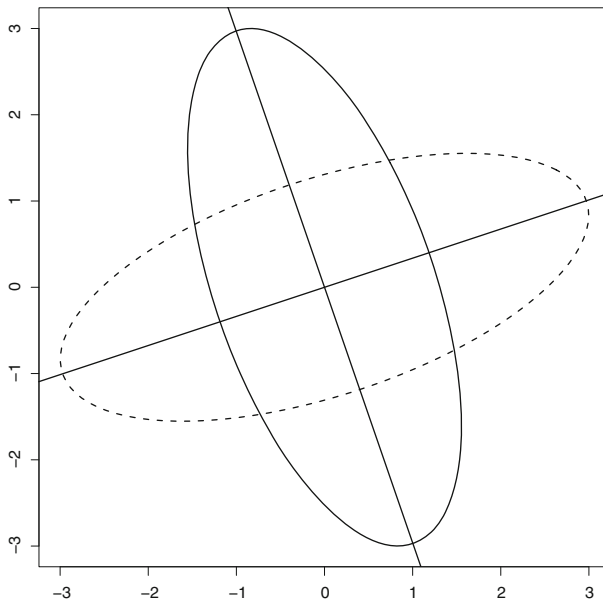
Fig. 2 Ellipses of equal concentration of two bivariate normal distributions, with the same principal axes but with different orientation

FORTRAN by Flury and Constantine (1985), can be considered to estimate the sample counterpart $\underset{\sim}{G}$ of $\underset{\sim}{\Gamma}$ (see Flury 1984). Under H_0^{CPC}, the following transformation

$$x_i^{(h)} \rightarrow \underset{\sim}{y}_i^{(h)} = \underset{\sim}{G}' \left(x_i^{(h)} - \overline{x}_h\right), \quad i = 1, \ldots, n_h,$$

holds, where the data $\underset{\sim}{y}_1^{(h)}, \ldots, \underset{\sim}{y}_{n_h}^{(h)}$, $h = 1, \ldots, k$, are uncorrelated with diagonal covariance matrix $\underset{\sim}{L}_h$ (the sample counterpart of Λ_h). Under multivariate normality, these components are also independent and normally distributed.

From a geometrical point of view, under H_0^{Γ}, the k ellipsoids of equal concentration have the same orientation in p-space while, under H_0^{CPC}, they have only the same principal axes. Figure 2 gives an example of ellipses of equal concentration with different orientation, but with the same principal axes.

In the light of these considerations, and aiming at testing H_0^{Γ}, we require a further statistical test to evaluate the null hypothesis

$$H_0^{\text{R}} : \left\{ \begin{array}{l} \text{the ranking of the eigenvalues } \underset{\sim}{\lambda}_1^{(h)}, \ldots, \underset{\sim}{\lambda}_p^{(h)} \\ \text{of } \Lambda_h, h = 1, \ldots, k, \text{ is the same for all the } k \text{ groups} \end{array} \right\}, \quad (10)$$

provided that H_0^{CPC} holds. In order to test H_0^{R}, a rank correlation test between k sets of rankings (of length p) induced by the sequences $\underset{\sim}{\lambda}_1^{(h)}, \ldots, \underset{\sim}{\lambda}_p^{(h)}$, $h = 1, \ldots, k$, appears to be a natural way to assess perfect monotonic relationship between sets of rankings. In practice, this solution is not good for two reasons. Firstly, information

is sacrificed when interval/ratio data (eigenvalues could be considered as ratio data since their positivity) are transformed into ordinal (rank-ordered) data. Secondly, the null hypothesis in all rank correlation tests is "uncorrelation between sets of rankings" (see, e.g., Sheskin 2000) since the distribution of the test statistic is known only in this case.

In order to overcome these issues, let

$$
A = \begin{pmatrix}
\lambda_1^{(1)} & \cdots & \lambda_j^{(1)} & \cdots & \lambda_p^{(1)} \\
\vdots & & \vdots & & \vdots \\
\lambda_1^{(h)} & \cdots & \lambda_j^{(h)} & \cdots & \lambda_p^{(h)} \\
\vdots & & \vdots & & \vdots \\
\lambda_1^{(k)} & \cdots & \lambda_j^{(k)} & \cdots & \lambda_p^{(k)}
\end{pmatrix}
$$

be the $k \times p$ matrix composed by the k sequences of p eigenvalues, arranged by row. Let $\lambda_{[1]}^{(1)}, \ldots, \lambda_{[p]}^{(1)}$ be the sequence of the entries on the first row of A in non-increasing order, that is, $\lambda_{[1]}^{(1)} \geq \cdots \geq \lambda_{[p]}^{(1)}$. Permuting the columns of A according to the order induced by $\lambda_{[1]}^{(1)}, \ldots, \lambda_{[p]}^{(1)}$, we obtain

$$
B = \begin{pmatrix}
\lambda_{[1]}^{(1)} & \cdots & \lambda_{[j]}^{(1)} & \cdots & \lambda_{[p]}^{(1)} \\
\vdots & & \vdots & & \vdots \\
\lambda_{[1]}^{(h)} & \cdots & \lambda_{[j]}^{(h)} & \cdots & \lambda_{[p]}^{(h)} \\
\vdots & & \vdots & & \vdots \\
\lambda_{[1]}^{(k)} & \cdots & \lambda_{[j]}^{(k)} & \cdots & \lambda_{[p]}^{(k)}
\end{pmatrix}.
$$

Thus, the null hypothesis H_0^R in (10) can be re-formulated as

$$
H_0^R : \lambda_{[1]}^{(l)} \geq \cdots \geq \lambda_{[p]}^{(l)}, \quad l = 2, \ldots, k. \tag{11}
$$

For a fixed $l \in \{2, \ldots, k\}$, the problem can be approached by means of $p - 1$ F-tests for the null hypothesis

$$
H_0^{(l),[j]} : \lambda_{[j-1]}^{(l)} \geq \lambda_{[j]}^{(l)}, \quad j = 2, \ldots, p, \tag{12}
$$

that, simultaneously considered, lead to

$$
H_0^{(l)} : \bigcap_{j=2}^{p} H_0^{(l),[j]}. \tag{13}
$$

Logically, the $p - 1$ tests in (12) require independence between the p variables in the lth group, $l = 2, \ldots, k$. Thus, they make sense only under H_0^{CPC} and multivariate

normality in the lth group, $l = 2, \ldots, k$. Also, the null hypothesis in (11) can be further re-formulated as

$$H_0^{\text{R}} : \bigcap_{l=2}^{k} H_0^{(l)}. \tag{14}$$

Really, with the aim of testing $H_0^{\boldsymbol{\Gamma}}$, we note again that all the hypotheses (10)–(14) have been developed under H_0^{CPC}, and we can express this assumption by explicitly rewriting the hypotheses (12) and (13) respectively as follows:

$$*H_0^{(l),[j]} : H_0^{\text{CPC}} \cap H_0^{(l),[j]} \tag{15}$$

$$*H_0^{(l)} : H_0^{\text{CPC}} \cap H_0^{(l)} = \bigcap_{j=2}^{p} *H_0^{(l),[j]}. \tag{16}$$

Although it can be considered as trivial, it is important to note that, while $H_0^{(l),[j]}$ and $H_0^{(l)}$ have to be (conditionally) interpreted as "given" H_0^{CPC}, $*H_0^{(l),[j]}$ and $*H_0^{(l)}$ have to be meant, respectively, as $H_0^{(l),[j]}$ and $H_0^{(l)}$ "jointly to" H_0^{CPC}. Thus, we can express the null hypothesis $H_0^{\boldsymbol{\Gamma}}$ as

$$H_0^{\boldsymbol{\Gamma}} : H_0^{\text{CPC}} \cap H_0^{\text{R}} = \bigcap_{l=2}^{k} *H_0^{(l)}.$$

In Sect. 4.4 we will provide details for testing the hypotheses $*H_0^{(l),[j]}$, $l = 2, \ldots, k$ and $j = 2, \ldots, p$, that are the building blocks to test $H_0^{\boldsymbol{\Gamma}}$.

4 Scedasticity tests in the framework of multiple testing procedures

The reciprocal relationships among the null hypotheses introduced in the above section are represented in Fig. 3, which describes also their hierarchical structure. Our purpose is to make inference on H_0, hence the decision should depend on some joint measure of erroneous statements in the given set of inferences. According to the classical terminology for MTPs (see, e.g., Hochberg and Tamhane 1987; Shaffer 1995), we can jointly consider all the hypotheses in the following set

$$\mathcal{H} = \left\{ H_0, H_0^{\boldsymbol{\Lambda}}, H_0^{\boldsymbol{\Gamma}}, H_0^{\text{R}}, \left\{ H_0^{\lambda_j} \right\}_{j=1,\ldots,p}, \left\{ *H_0^{(l)} \right\}_{l=2,\ldots,k}, \left\{ *H_0^{(l),[j]} \right\}_{\substack{l=2,\ldots,k \\ j=2,\ldots,p}} \right\},$$

as a finite *family* of inferences.

We said above that $\mathcal{H}$ is a *hierarchy*; this is true because at least an implication relation holds between the hypotheses as, for instance, between H_0 and $H_0^{\boldsymbol{\Lambda}}$, where $H_0 \Rightarrow H_0^{\boldsymbol{\Lambda}}$ means $H_0^{\boldsymbol{\Lambda}} \subseteq H_0$ (see Hochberg and Tamhane 1987, p. 344). In this
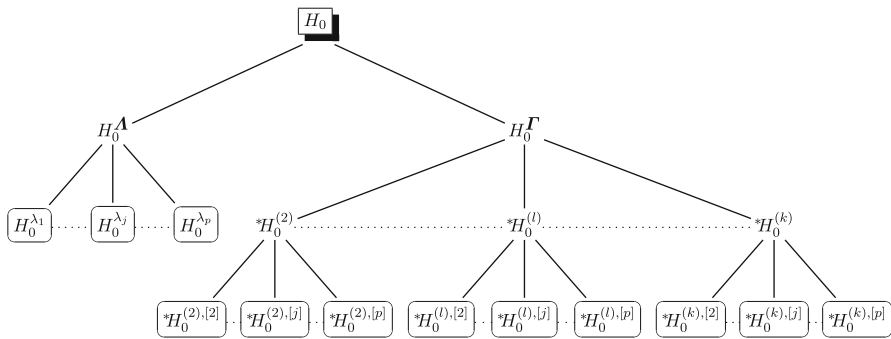
Fig. 3 Hierarchy among the null hypotheses in $\mathcal{H}$

case $H_0^{\boldsymbol{\Lambda}}$ is said to be a *component* of H_0 (Gabriel 1969). All the implications among the null hypotheses are shown by moving from top to bottom of the tree in Fig. 3. Hypotheses that do not have any component are referred to as *minimal* (represented in oval boxes); all other hypotheses are referred to as *non-minimal*. In particular, the hypothesis which is the intersection of all the hypotheses in the family (displayed in the shadow box) is said to be the *overall hypothesis*.

4.1 Coherence and consonance for hierarchical families

Any MTP for a hierarchical family of hypotheses is generally required to possess two logical consistency properties: coherence and consonance (Gabriel 1969). We recall that a *coherent* MTP avoids the inconsistency of rejecting a hypothesis without also rejecting all hypotheses implying it, and it is said to be *consonant* if, whenever any non-minimal hypothesis is rejected, at least one of its components is also rejected. The lack of consonance (referred to as *dissonance*), according to Hochberg and Tamhane (1987), is not as serious a drawback as the lack of coherence (because of the nonsymmetry in the interpretation of rejection and acceptance decisions in classical significance testing). In our case, however, we believe that the practitioner cannot, for example, find himself in the situation of rejecting H_0 without having rejected at least one between $H_0^{\boldsymbol{\Lambda}}$ or $H_0^{\boldsymbol{\Gamma}}$.

Motivated by these considerations, we adopted the union-intersection (UI) procedure. Indeed, it is the only one that guarantees both properties among the single-step MTPs (for further details, see Hochberg and Tamhane 1987, pp. 47–50). The underlying idea of a UI procedure for a hierarchical family, is to test all minimal hypotheses and then to step up through the hierarchy of non-minimal hypotheses guaranteeing coherence and consonance, without performing further tests.

4.2 The union-intersection procedure, for hierarchical families, via adjusted *p*-values

UI procedures can be described either by means of test statistics (see, e.g., Hochberg and Tamhane 1987, pp. 47–50), or by *adjusted p-values* (Rosenthal and Rubin 1983),

Table 1 Outcomes in testing m hypotheses

| | H_0 | | |
	Not rejected	Rejected	Total
H_0			
True	$N_{0|0}$	$N_{1|0}$	M_0
False	$N_{0|1}$	$N_{1|1}$	M_1
Total	$m - R$	R	m

which are the natural counterpart, in the multiple testing framework, of the classical p-values. The reasons for preferring adjusted p-values are the same as those for using ordinary (unadjusted) p-values. In detail, they provide information about whether an individual statistical hypothesis is significant or not (an adjusted p-value can be compared directly with any chosen significance level α and if it is less than or equal to α, the individual hypothesis is rejected), they indicate "how significant" the result is (the smaller the adjusted p-value, the stronger the evidence against the individual null hypothesis), they are interpretable on the same scale as those for tests of individual hypotheses, making comparison with single hypothesis testing easier (for further details on their chief advantages, see Wright 1992; Shaffer 1995, pp. 567–568; Westfall and Young 1993, pp. 11–12).

Roughly speaking, in the UI procedure for hierarchical families, p-values are initially computed for the minimal hypotheses and then they are modified (adjusted) to take into account the multiplicity effect. The adjusted p-value for a non-minimal hypothesis is then obtained as the smallest adjusted p-value associated with its minimal components. It is straightforward to show that this method guarantees a coherent and consonant MTP.

4.3 Choosing the error rate to control

To complete the description of the UI procedure to test H_0, we need now to specify how to obtain adjusted p-values of the minimal hypotheses. Such a choice is directly connected with the error rate to control in a given MTP and this latter issue has generated a wide discussion in the literature. To give a brief description of the issue, we have to compare what happens in the single test framework with respect to MTPs.

Consider a situation in which m tests have to be performed; in our case, $m = k(p-1) + 1$ is the number of minimal hypotheses to test. Suppose M_0 of the m hypotheses to be true, and M_1 false. Table 1 shows the possible outcomes in this situation: R is the number of rejections, while $N_{1|0}$ and $N_{0|1}$ denote the exact (unknown) number of committed errors.

In the usual (single) test setting, where $m = 1$ and $N_{1|0} \in \{0, 1\}$, one controls the probability of a false rejection (Type I error) while looking for a procedure that possibly minimizes the probability of observing a false negative (Type II error). Unfortunately, in the multiple case, as m increases the number of false positives can explode. Thus, adjustment of p-values arises as one has to look for other specific Type I error measures, still based on $N_{1|0}$, that can serve as possible generalizations of the probability of Type I error in the multiple testing context.

All Type I error measures refer to probabilities that are conditioned on which minimal hypotheses are true. Thus, the control of the error rate can be categorized as *weak* if all the minimal hypotheses are supposed true (the so called "complete null"), and *strong* when the Type I error measure is controlled for any of the 2^m possible true/false configurations among the m minimal hypotheses. Logically, the results of a weak controlling procedure can be used only to imply that there is some false hypothesis among the m, but do not allow the researcher to say which hypotheses are false. We choose strong control on H_0 to derive inferences on each hypothesis in $\mathcal{H}$, in particular on $H_0^{\boldsymbol{\Lambda}}$ and $H_0^{\boldsymbol{\Gamma}}$, with the aim of obtaining a more informative MTP for H_0 than the standard approach given by the Box test.

The most classical multiple Type I error measure is the familywise error rate (FWER), defined as $P\left(N_{1|0} > 0\right)$. Nevertheless, while FWER controlling methods are preferable in many situations, they can have low power in certain large m applications. In response to the need of an error measure that would allow for good power, in particular with large m, many modern Type I error rates are based on the false discovery proportion (FDP), defined as $N_{1|0}/R$ if $R > 0$, and 0 if $R = 0$ (see Benjamini 2010, for a brief discussion on the success of the FDP-based error measures). Among them, the most important ones are the false discovery rate (FDR), proposed by Benjamini and Hochberg (1995) and defined as the expectation of *FDP*, say $E(FDP)$, and the more recent false discovery exceedance (FDX), independently proposed by Van der Laan et al. (2004) and Genovese and Wasserman (2006), and defined as $P(FDP > c)$, $c \in (0, 1)$. It is worth recalling that FDX-control at level α ensures also FDR-control at level $\alpha + (1 - \alpha) c$ (see Genovese and Wasserman 2006), as both FDX and FDR are functionals of the same random variable FDP. To find out more on the relations between these two error measures see Farcomeni (2008, p. 351).

Roughly speaking, in FDX-control interest is taken in the tails of the distribution of the FDP rather than in its central part. Hence FDX-control can be preferred when the random variable FDP has a weak concentration on its mean, the FDR: this situation, for instance, can be due to a high variance of the FDP and can be further affected by dependence among the test statistics.

In our case, apart from the subset $\left\{H_0^{\lambda_j}\right\}_{j=1,\dots,p} \subset \mathcal{H}$ where independence between the related test statistics holds, we do not know the dependence structure for the test statistics of the other minimal hypotheses. Thus, we choose to control FDX in order to prevent any situation of dependence.

4.4 FDX-control by augmentation MTPs

Among the procedures controlling FDX (for a detailed review on the different strategies to control this and the other error rates above mentioned, see Farcomeni 2008), we decided to adopt *augmentation* due to Van der Laan et al. (2004). This method relies on the idea that relaxing FWER control will result in the rejection of at least the same hypotheses. For this reason, the procedure starts, in a first step, by controlling the FWER and then, in a second step, it augments the number of the previously

rejected hypotheses, denoted by r_{FWER}, by an opportune additional number $a(c, \alpha) \in \{0, 1, \ldots, m - r_{\text{FWER}}\}$ (for more details, see Dudoit and van der Laan 2008, Chapter 6). The choice of this procedure is primarily justified by its robustness to dependence if the FWER, in the first step, is controlled under general dependence and, secondly, by its clear idea and its computational low cost. In these terms, there are two main procedures controlling FWER under arbitrary dependence: the classical method of Bonferroni (1936), and the slightly more complex, but much less conservative, method of Holm (1979) that we have chosen to adopt.

In Holm's procedure, the unadjusted p-values $p_1, \ldots, p_i, \ldots, p_m$ are ordered so that $p_{(1)} \leq \cdots \leq p_{(i)} \leq \cdots \leq p_{(m)}$, and each $p_{(i)}$ is compared to $\alpha/(m - i + 1)$ rather than to α/m as in the Bonferroni method. In other words, $(m - i + 1)p_{(i)}$, $i = 1, \ldots, m$ is compared to α. Note that these values are not necessarily the adjusted p-values for Holm's procedure, since a logical property of "comonotonicity" with $p_{(1)}, \ldots, p_{(i)}, \ldots, p_{(m)}$ has to be respected (see Wright 1992, that exemplifies this issue). Thus, the ordered adjusted p-values $p_{(i)}^{\text{Holm}}$ for the Holm procedure can be obtained as

$$p_{(i)}^{\text{Holm}} = \max_{j=1,\ldots,i} \left\{ (m - j + 1) \, p_{(j)} \right\}, \quad i = 1, \ldots, m. \tag{17}$$

These adjusted p-values are intentionally not truncated to 1 to emphasize the degree of conservatism of the procedure. Note that Holm's method is also adopted to obtain the adjusted p-value $^*p_{(l),[j]}$ for each minimal hypothesis $^*H_0^{(l),[j]}$, $l = 2, \ldots, k$, $j = 2, \ldots, p$; here, the procedure is applied to p_{CPC} and $p_{(l),[j]}$ which are, respectively, the p-values for H_0^{CPC} and $H_0^{(l),[j]}$, in order to compute the adjusted Holm's p-values $p_{\text{CPC}}^{\text{Holm}}$ and $p_{(l),[j]}^{\text{Holm}}$ according to (17). Then $^*p_{(l),[j]} = \min \left\{ p_{\text{CPC}}^{\text{Holm}}, p_{(l),[j]}^{\text{Holm}} \right\}$.

Once r_{FWER} hypotheses have been rejected by the Holm procedure, for a given bound $c \in (0, 1)$, they can be augmented by the additional number of rejections

$$a(c, \alpha) = \max_{j=0,1,\ldots,m-r_{\text{FWER}}} \left\{ j : \frac{j}{j + r_{\text{FWER}}} \leq c \right\}. \tag{18}$$

Logically, the $a(c, \alpha)$ hypotheses not yet rejected that have to be selected, are those with most significant adjusted p-values in (17). Moreover, when no hypothesis is rejected at the first step (something not uncommon for Holm's procedure if m is large) none will be at the second step for any c. On the other hand, if all the m hypotheses are rejected in the first step, the augmentation procedure does not make sense. It is interesting to note from (18) that, fixed the value of c, the greater is r_{FWER}, the greater is $P(a(c, \alpha) > 0)$.

From Eq. (18), the ordered adjusted p-values $p_{(i)}^{\text{Aug}}$ for the augmentation FDX-controlling procedure are given by

$$p_{(i)}^{\text{Aug}} = p_{(\lceil (1-c)i \rceil)}^{\text{Holm}}, \quad i = 1, \ldots, m, \tag{19}$$

where the ceiling function $\lceil x \rceil$ denotes the least integer greater than or equal to x (see Dudoit and van der Laan 2008, pp. 250–251). Thus, the ordered adjusted p-values

in (19) are simply ic-shifted versions (up to a ceiling integer transformation) of the ordered adjusted p-values in (17). Regarding the selection of an appropriate value of c, it is generally chosen as a value in the interval $(0, \alpha)$ but, if one wants to extend control also to FDR, the optimal value of c should be close to zero (see Farcomeni 2008, pp. 369–371).

5 Some applications to real data sets

This section pursues the double intent to exemplify and to analyze the augmentation UI procedure described above. To this aim, some well-known examples in the multivariate literature will be used. The value $c = 0.01$ and a nominal level of 0.05 are used hereafter for FDX-control, hence FDR-control at level 0.0595 is also obtained.

Calculations have been carried out using R; the R functions necessary to implement all the proposed tests, the F–G algorithm, and the augmentation UI procedure with both adjusted and unadjusted p-values are available from the authors upon request. Finally, note that a preliminary Mardia's test (Mardia 1985) is performed on each set of data to check the underlying assumption of multivariate normality.

Example 1: Fisher's iris data

This data set, undoubtedly the most popular in the multivariate literature, was first published by Anderson (1935) and used by Fisher (1936) as an example of discriminant analysis. The sample is composed of $n = 150$ 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 scatterplots referred to the grouped-data is shown in Fig. 4. Based on Mardia's test, there is statistical evidence for multivariate normality in each of the 3 groups.

A graphical inspection of Fig. 4 underlines that the covariance structure among groups is different so that both homometroscedasticity and homotroposcedasticity (and hence homoscedasticity) appears to be too strong for these data. To ascertain this conjecture, the Box M-test with F-approximation has been applied; at any reasonable significance level the null hypothesis of common covariance matrices is rejected, since the p-value resulted lower than 10^{-30}.

To investigate the nature of the deviation of the data from homoscedasticity, the augmentation UI procedure has been performed. The testing hierarchy for the iris data, along with the adjusted p-values for each hypothesis, is shown in Fig. 5. For the minimal hypotheses we have chosen to show both the unadjusted p-values (in round brackets) and the adjusted p-values computed according to (19); this way of proceeding, in line with Westfall and Young (1993, p. 10), highlights their disparity. For each non-minimal hypothesis, the adjusted p-values have been obtained according to the UI procedure, as specified at the end of Sect. 4.2.

In detail, in Fig. 5 we have $m = 10$ minimal hypotheses of which $p = 4$ referred to $H_0^{\mathbf{\Lambda}}$ and $(k - 1)(p - 1) = 6$ referred to $H_0^{\mathbf{\Gamma}}$; consequently, we have to compute the
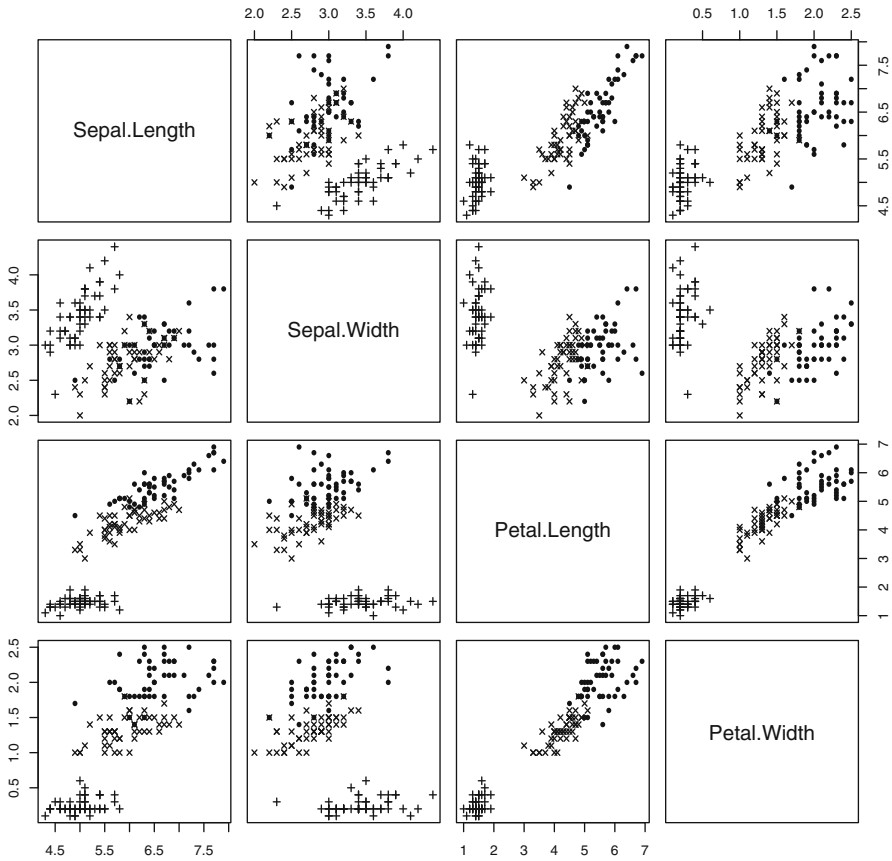
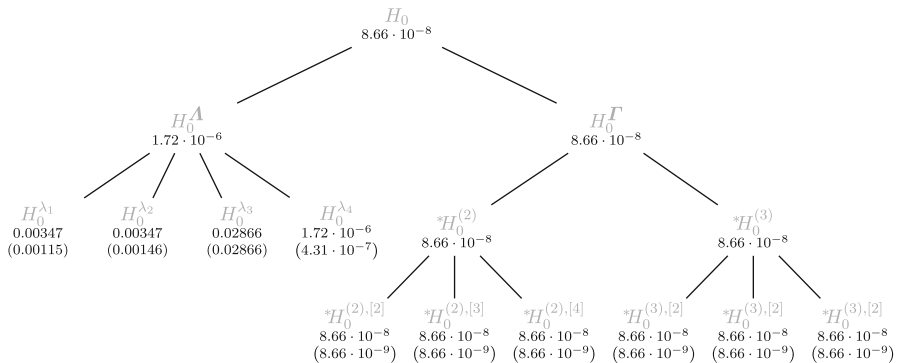**Fig. 4** Matrix of scatterplots of Fisher's iris data

Fig. 5 Adjusted and unadjusted *p*-values (in *round brackets*) of the augmentation UI procedure, applied on the iris data. The rejected hypotheses at the nominal level 0.05 are displayed in *gray*

Table 2 Scheme of the various phases leading to $^*p_{(l),[j]}$, $l = 2, 3$, $j = 2, 3, 4$

Phase	p-value	$l = 2$			$l = 3$		
		$j = 2$	$j = 3$	$j = 4$	$j = 2$	$j = 3$	$j = 4$
1	p_{CPC}	4.33×10^{-9}	4.33×10^{-9}	4.33×10^{-9}	4.33×10^{-9}	4.33×10^{-9}	4.33×10^{-9}
	$p_{(l),[j]}$	1.00000	0.14698	1.00000	1.00000	0.65809	0.78466
2	$p_{\text{CPC}}^{\text{Holm}}$	8.66×10^{-9}	8.66×10^{-9}	8.66×10^{-9}	8.66×10^{-9}	8.66×10^{-9}	8.66×10^{-9}
	$p_{(l),[j]}^{\text{Holm}}$	1.00000	0.14698	1.00000	1.00000	0.65809	0.78466
3	$^*p_{(l),[j]}$	8.66×10^{-9}	8.66×10^{-9}	8.66×10^{-9}	8.66×10^{-9}	8.66×10^{-9}	8.66×10^{-9}

same number of "minimal" p-values. Table 2 describes the phases of the Holm procedure, leading to the 6 values of $^*p_{(l),[j]}$, $l = 2, 3$, $j = 2, 3, 4$. The minimal p-values referred to $H_0^{\boldsymbol{\Lambda}}$ are instead obtained by simultaneous Bartlett's tests of univariate homoscedasticity among the 3 groups, as described at the end of Sect. 3.1.

Once the p-values for all the minimal hypotheses have been computed, their adjusted version in (17), for the first step of the augmentation procedure, has been obtained following the phases sketched in Table 3. Here, according to the last row, the adjusted (minimal) p-values are lower than the adopted nominal size; thus, at the end of the first step, all the minimal hypotheses have been rejected. As $r_{\text{FWER}} = 10$, the augmentation procedure has no influence, so that $p_{(i)}^{\text{Aug}} = p_{(i)}^{\text{Holm}}$, $i = 1, \ldots, 10$. This is the reason why the adjusted p-values in Fig. 5 agree with those obtained by Holm's procedure.

Finally, if we compare the p-value obtained by the Box test (which is lower than 10^{-30}) with the significance of the UI procedure (8.66×10^{-8}), we see that the latter is more conservative. The inferential approach used in this example allows us to conclude that heteroscedasticity is fully verified, as the scatters in the three species differ in orientation and in size and/or shape.

Example 2: Blue crabs

The crab data set of Campbell and Mahon (1974) on the genus *Leptograpsus*, which has been analyzed further in Ripley (1996), is very popular in multivariate literature.

Here, following the setting of Peel and McLachlan (2000) (see also McLachlan and Peel 2000, pp. 90–92), the 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) on the rear width (RW) and the length along the midline (CL) of the carapace. Mardia's test suggests that it is reasonable to assume that the two group-conditional distributions are bivariate normal. On the contrary Box's M-test, with F-approximation, rejects homoscedasticity at any reasonable significance level, because the p-value is 2.76×10^{-14}.

In order to obtain more information about the nature of the observed heteroscedasticity, the following considerations could be useful. In Fig. 6 the scatter plot of RW

Table 3 Phases leading to $p_{(i)}^{\text{Holm}}$, $i = 1, \ldots, m$, for the $m = 10$ minimal hypotheses

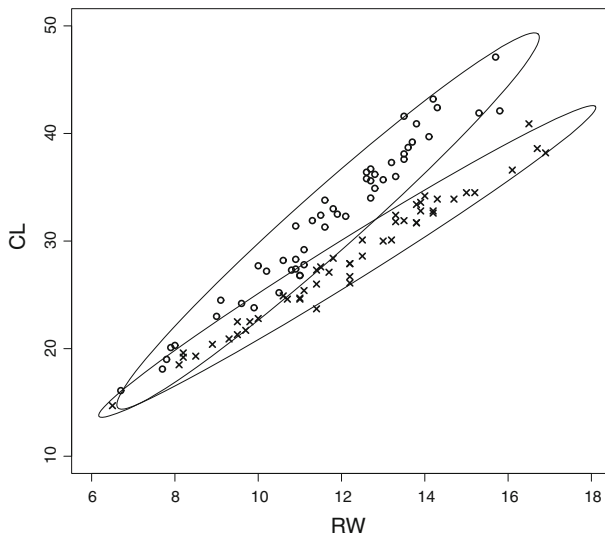| Phases	$*H_0^{(l),[j]}$						$H_0^{\lambda_4}$	$H_0^{\lambda_1}$	$H_0^{\lambda_2}$	$H_0^{\lambda_3}$
i	1	2	3	4	5	6	7	8	9	10
$(m-i+1)$	10	9	8	7	6	5	4	3	2	1
$P_{(i)}$	8.66×10^{-9}	8.66×10^{-9}	8.66×10^{-9}	8.66×10^{-9}	8.66×10^{-9}	8.66×10^{-9}	4.31×10^{-7}	0.00115	0.00146	0.02866
$(m-i+1)P_{(i)}$	8.66×10^{-8}	7.79×10^{-8}	6.92×10^{-8}	6.06×10^{-8}	5.19×10^{-8}	4.33×10^{-8}	1.72×10^{-6}	0.00347	0.00293	0.02866
$p_{(i)}^{\text{Holm}}$	8.66×10^{-8}	8.66×10^{-8}	8.66×10^{-8}	8.66×10^{-8}	8.66×10^{-8}	8.66×10^{-8}	1.72×10^{-6}	0.00347	0.00347	0.02866

Fig. 6 Scatter plot of variables RW and CL for $n_1 = 50$ males and $n_2 = 50$ females *blue* crabs (*circle* denotes male and *multiple symbol* female). The ellipses of equal (95%) concentration are also superimposed

versus CL, in both groups, is shown. Once we obtain the sample mean vectors and the unbiased covariance matrices

$$
\bar{x}_1 = \begin{pmatrix} 11.678 \\ 31.862 \end{pmatrix} \quad S_1 = \begin{pmatrix} 4.296 & 14.460 \\ 14.460 & 51.071 \end{pmatrix}
$$
$$
\bar{x}_2 = \begin{pmatrix} 12.138 \\ 28.102 \end{pmatrix} \quad S_2 = \begin{pmatrix} 5.947 & 14.253 \\ 14.253 & 35.042 \end{pmatrix}
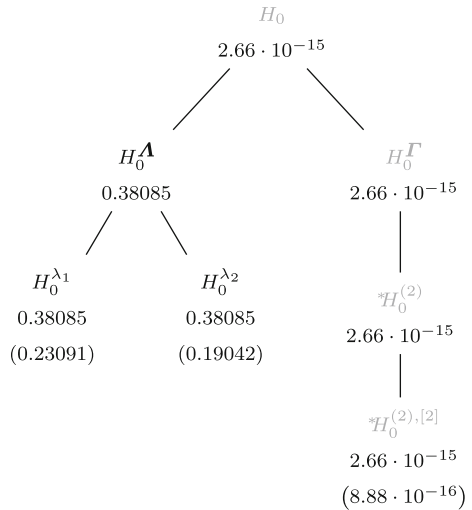\tag{20}
$$

we can superimpose on the scatter plot in Fig. 6 the ellipses of equal (95%) concentration arising from a bivariate normal distribution with mean vector $\bar{x}_h$ and covariance matrix S_h, $h = 1, 2$. The principal component analysis of S_h, $h = 1, 2$, returns

$$
L_1 = \begin{pmatrix} 55.180 & 0 \\ 0 & 0.187 \end{pmatrix} \quad G_1 = \begin{pmatrix} 0.273 & -0.962 \\ 0.962 & 0.273 \end{pmatrix}
$$
$$
L_2 = \begin{pmatrix} 40.861 & 0 \\ 0 & 0.129 \end{pmatrix} \quad G_2 = \begin{pmatrix} 0.378 & -0.926 \\ 0.926 & 0.378 \end{pmatrix}.
\tag{21}
$$

Although the Box M-test points out heteroscedasticity, the scatter plot in Fig. 6 shows strong similarity between size and shape of the two ellipses. This remark is confirmed by applying the augmentation UI procedure graphically represented in Fig. 7.

Here, focusing attention on the hierarchy of the sub-family having $H_0^{\boldsymbol{\Lambda}}$ as overall null hypothesis, we obtained an unadjusted p-value of 0.23091 for the comparison between the variances associated with the first principal component in the two groups (first diagonal element in L_h, $h = 1, 2$), and an unadjusted p-value of 0.19042 for the

Fig. 7 Adjusted and unadjusted
p-values (in *round brackets*) of
the augmentation UI procedure,
applied on the crab data. The
rejected hypotheses at the
nominal level 0.05 are displayed
in *gray*

$$H_0$$
$$2.66 \cdot 10^{-15}$$

$$H_0^{\Lambda}$$
$$0.38085$$

$$H_0^{\Gamma}$$
$$2.66 \cdot 10^{-15}$$

$$H_0^{\lambda_1}$$
$$0.38085$$
$$(0.23091)$$

$$H_0^{\lambda_2}$$
$$0.38085$$
$$(0.19042)$$

$${}^{*}H_0^{(2)}$$
$$2.66 \cdot 10^{-15}$$

$${}^{*}H_0^{(2),[2]}$$
$$2.66 \cdot 10^{-15}$$
$$\left(8.88 \cdot 10^{-16}\right)$$

comparison between the variances associated with the second principal component in the two groups (second diagonal element in L_h, $h = 1, 2$). Note that, being $k = 2$, $p = 2$ simultaneous F-tests have been performed. Naturally, if at the chosen nominal level the null hypothesis of homometroscedasticity is not rejected before adjustment, it cannot be rejected after adjustment.

As regards the orientation of the two ellipses in Fig. 6, they show a slight difference in the directions of their main axes, reflected also in the values of the elements along the diagonal in the sample eigenvalue matrices G_1 and G_2 given in (21). This conjecture is confirmed in the hierarchy of the sub-family having H_0^{Γ} as overall null hypothesis: due to a practically null unadjusted p-value associated with Flury's test for H_0^{CPC} ($p_{CPC} = 4.44 \times 10^{-16}$), the null hypothesis of homotroposcedasticity (equal orientation) is rejected at any reasonable significance level with an adjusted p-value of 2.66×10^{-15}. Naturally, for coherence, also the homoscedasticity hypothesis H_0 is rejected with a lower significance with respect to the Box test.

Finally, regarding the augmentation procedure, it is interesting to note that in the first step, 1 out of the 3 minimal hypotheses is rejected ($r_{FWER} = 1$). The combination between $c = 0.01$ and $r_{FWER} = 1$ is too low to add further rejections, according to (18), in the second step. In other words, $a(c, \alpha) = 0$. Thus, also in this example, $p_{(i)}^{Aug} = p_{(i)}^{Holm}$, $i = 1, \ldots, 3$.

Example 3: Genuine and forged bank notes

From Flury and Riedwyl (1983), and based on (Flury 1988, pp. 51–56), the following $p = 2$ variables, measured (in millimeters) on former Swiss 1000-franc bank notes, will be considered:

$$LEFT = \text{width of bank note, measured on left side,}$$
$$RIGHT = \text{width of bank note, measured on right side.}$$

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 suggests that it is reasonable to assume bivariate normality for the group of forged bills, but it leaves some doubt on the other one. Since Flury (1988) tacitly accepts the hypothesis of bivariate normality in the group of genuine bills, we will retain it to hold from now on. On the contrary Box's M-test rejects homoscedasticity providing a p-value of 0.003 with F-approximation.

In order to comprehend more deeply the nature of the presumed heteroscedasticy, we begin by analyzing the two groups and obtaining their sample mean vectors and unbiased covariance matrices

$$\bar{x}_1 = \begin{pmatrix} 129.943 \\ 129.722 \end{pmatrix} \quad S_1 = \begin{pmatrix} 0.133 & 0.086 \\ 0.086 & 0.126 \end{pmatrix}$$
$$\bar{x}_2 = \begin{pmatrix} 130.300 \\ 130.193 \end{pmatrix} \quad S_2 = \begin{pmatrix} 0.065 & 0.047 \\ 0.047 & 0.089 \end{pmatrix}. \tag{22}$$

Figure 8 shows the scatter plots of the variables LEFT and RIGHT in both groups, as well as the ellipses of equal (95%) concentration. Note that, unlike the example on the blue crabs discussed above, the scatter plots for genuine and forged bank notes are separately displayed in order to avoid a large overlapping between sample points. Although Box's M-test points out heteroscedasticity, a strong similarity between the orientation of the two ellipses in Fig. 8a and b appears, while some difference in their size emerges. As a first step, in order to confirm or discard this conjecture, the principal component analysis of S_h, $h = 1, 2$,

$$L_1 = \begin{pmatrix} 0.215 & 0 \\ 0 & 0.043 \end{pmatrix} \quad G_1 = \begin{pmatrix} 0.721 & -0.693 \\ 0.693 & 0.721 \end{pmatrix}$$
$$L_2 = \begin{pmatrix} 0.125 & 0 \\ 0 & 0.029 \end{pmatrix} \quad G_2 = \begin{pmatrix} 0.613 & -0.790 \\ 0.790 & 0.613 \end{pmatrix} \tag{23}$$

could be taken into account. Based on (23), but also according to the graphical representation, the two sample matrices of eigenvalues L_1 and L_2 seem to differ while, on the other hand, the sample eigenvector matrices G_1 and G_2 are very similar. Applying the homometroscedasticity test, it provides raw (non-multiplicity corrected) p-values of 0.00751 and 0.04059, respectively, for the equality of variances associated with the first and the second principal components in the two groups (that is the first and second diagonal element in L_h, $h = 1, 2$). After adjustment, these p-values become 0.02253 and 0.08119, respectively. Thus, at the chosen nominal level, $H_0^{\lambda_1}$ is rejected and, for coherence, also H_0^{Λ} and H_0 are rejected. These results are all displayed in Fig. 9.

As regards the conjecture on the equality between orientations, being the component hypothesis $^*H_0^{(2),[2]}$ of H_0^{Γ} not rejected before adjustment, for consonance, homotroposcedasticiy is not rejected also after adjustment. Consequently, also in this
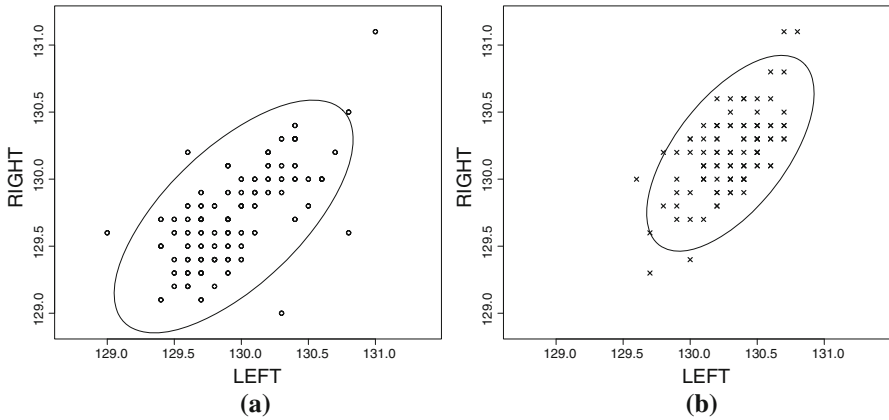
Fig. 8 Scatter plots, and related ellipses of equal (95%) concentration, of variables *LEFT* and *RIGHT* in two groups of Swiss bank notes. Coinciding points are marked by *single symbol* only. **a** Genuine, **b** forged

Fig. 9 Adjusted and unadjusted *p*-values (in *round brackets*) of the augmentation UI procedure, applied on the genuine and forged bank notes data. The rejected hypotheses at the nominal level 0.05 are displayed in *gray*
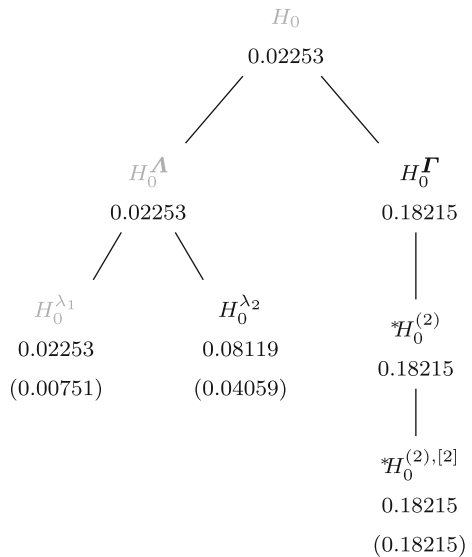
case, homoscedasticity is rejected for coherence. Finally, regarding the augmentation procedure, the same considerations stated at the end of Example 2 apply.

6 A simulation study

In the following, we present the results of a wide simulation study developed in order to show the performance of the proposed augmentation UI procedure in different situations of gradual departure from homoscedasticity. As there are a lot of factors that may be varied (the number of groups k, the dimension p of the observed variables, the overall size of the sample n, the nominal level of the test, the size, shape, and

orientation of the group scatters), some of them have been necessarily considered fixed.

In more detail, two different scenarios of simulation have been considered. In the first, data have been generated from $k = 2$ groups drawn from two bivariate normals ($p = 2$) with mean vectors $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Small to moderately large sizes, $n = 60$, $n = 150$ and $n = 300$, have been considered with $n_1 = n_2$ in all the three cases. Here, $\boldsymbol{\Sigma}_1$ has been randomly generated and its spectral decomposition

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Gamma}_1\boldsymbol{\Lambda}_1\boldsymbol{\Gamma}_1' = \begin{pmatrix} \gamma_{11}^{(1)} & \gamma_{12}^{(1)} \\ \gamma_{21}^{(1)} & \gamma_{22}^{(1)} \end{pmatrix}\begin{pmatrix} \lambda_1^{(1)} & 0 \\ 0 & \lambda_2^{(1)} \end{pmatrix}\begin{pmatrix} \gamma_{11}^{(1)} & \gamma_{12}^{(1)} \\ \gamma_{21}^{(1)} & \gamma_{22}^{(1)} \end{pmatrix}',$$

with $\lambda_1^{(1)} \geq \lambda_2^{(1)}$, has been computed. In order to simulate different ways of gradual departure from homoscedasticity, $\boldsymbol{\Sigma}_2$ has been defined by the following relation with $\boldsymbol{\Sigma}_1$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}\boldsymbol{\Gamma}_1\begin{pmatrix} d\lambda_1^{(1)} & 0 \\ 0 & \frac{\lambda_2^{(1)}}{d} \end{pmatrix}\left[\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}\boldsymbol{\Gamma}_1\right]',$$

where

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

is the rotation matrix of angle θ, and $d \geq 1$ is a sort of "deformation" constant: if $d = 1$, the concentration ellipsoids related to $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are homometroscedastic; the flattening of the concentration ellipsoids related to $\boldsymbol{\Sigma}_2$ increases in line with the increase of d. Several values of θ and d have been considered: 8 values for θ (0, $\pi/36$, $\pi/18$, $\pi/12$, $\pi/6$, $\pi/4$, $\pi/3$, $\pi/2$ in radiants, i.e., 0°, 5°, 10°, 15°, 30°, 45°, 60°, 90° in degrees), and 5 values for d (1, 1.25, 1.5, 2 and 4). All combinations of these two factors have been taken into consideration in the simulation.

The null hypotheses H_0, $H_0^{\boldsymbol{\Lambda}}$, and $H_0^{\boldsymbol{\Gamma}}$—that we retain the principal vertices of the hierarchy displayed in Fig. 3—have been assessed using the proposed augmentation UI procedure. Regarding H_0, Box's M-test with F-approximation (labeled by **Box** in Table 4) has been also taken into account for useful comparisons.

Table 4 displays the empirical power (rejection rates), with a nominal level of 0.05 and $c = 0.01$, calculated simulating 10000 samples for each of the 40 possible settings and for each of the 3 values of n. The numerical results have been arranged in a sort of 8×5 block matrix ($8 =$ number of θ values $\times 5 =$ number of d values), with blocks of dimension 4×3 ($4 = 3 + 1 =$ number of considered null hypotheses plus Box's M-test for $H_0 \times 3 =$ number of n values). In more detail, homotroposcedasticity holds in the first row of this block matrix, while homometroscedasticity holds in the first column. Consequently, the block of position (1, 1) is the only case of homoscedasticity; a gradual departure from this situation is obtained moving in the South-East direction in the matrix.

Table 4 Empirical power for H_0^Λ, H_0^Γ, and H_0, arising from the UI procedure, compared to Box's M-test. Samples are drawn with sizes $n = 60, 150, 300$, by two groups of bivariate normal, varying the shape (d) and the orientation (θ) of the concentration ellipsoids

θ	d	1			1.25			1.5			2			4		
	n	60	150	300	60	150	300	60	150	300	60	150	300	60	150	300
	Test															
0°	Box	0.0485	0.0501	0.0526	0.0908	0.1714	0.3350	0.2082	0.5095	0.8473	0.5650	0.9501	0.9996	0.9959	1.0000	1.0000
	H_0	0.0422	0.0427	0.0444	0.0824	0.1547	0.2813	0.1790	0.4416	0.7781	0.4801	0.9132	0.9989	0.9891	1.0000	1.0000
	H_0^Λ	0.0322	0.0343	0.0368	0.0708	0.1443	0.2745	0.1649	0.4342	0.7771	0.4705	0.9117	0.9989	0.9886	1.0000	1.0000
	H_0^Γ	0.0109	0.0088	0.0080	0.0134	0.0132	0.0107	0.0177	0.0163	0.0159	0.0207	0.0244	0.0248	0.0345	0.0323	0.0255
5°	Box	0.1400	0.2697	0.4153	0.2167	0.4357	0.6814	0.3568	0.7265	0.9466	0.6888	0.9800	1.0000	0.9968	1.0000	1.0000
	H_0	0.1190	0.2333	0.3917	0.1745	0.3646	0.5833	0.2701	0.6032	0.8823	0.5551	0.9444	0.9996	0.9896	1.0000	1.0000
	H_0^Λ	0.0367	0.0354	0.0407	0.0737	0.1537	0.2962	0.1688	0.4532	0.7997	0.4811	0.9227	0.9991	0.9871	1.0000	1.0000
	H_0^Γ	0.0854	0.2079	0.3718	0.1112	0.2628	0.4310	0.1276	0.3054	0.4953	0.1596	0.3781	0.5698	0.2256	0.4434	0.5959
10°	Box	0.3514	0.5771	0.7166	0.4446	0.7111	0.8686	0.5865	0.8739	0.9810	0.8261	0.9913	1.0000	0.9990	1.0000	1.0000
	H_0	0.3109	0.5392	0.6844	0.3902	0.6400	0.8026	0.4871	0.7856	0.9454	0.7059	0.9720	0.9998	0.9945	1.0000	1.0000
	H_0^Λ	0.0364	0.0386	0.0416	0.0772	0.1619	0.3111	0.1878	0.4680	0.8175	0.5134	0.9299	0.9990	0.9896	1.0000	1.0000
	H_0^Γ	0.2883	0.5240	0.6757	0.3426	0.5846	0.7325	0.3806	0.6267	0.7805	0.4314	0.6851	0.8233	0.5226	0.7292	0.8387
15°	Box	0.5254	0.7245	0.8430	0.6243	0.8360	0.9322	0.7338	0.9345	0.9911	0.8980	0.9967	1.0000	0.9993	1.0000	1.0000
	H_0	0.4853	0.7045	0.8205	0.5643	0.7823	0.8968	0.6450	0.8754	0.9752	0.8030	0.9874	1.0000	0.9964	1.0000	1.0000
	H_0^Λ	0.0413	0.0440	0.0434	0.0839	0.1723	0.3315	0.1884	0.4899	0.8267	0.5182	0.9338	0.9996	0.9902	1.0000	1.0000
	H_0^Γ	0.4676	0.6957	0.8142	0.5314	0.7486	0.8577	0.5673	0.7821	0.8871	0.6098	0.8177	0.9062	0.6853	0.8427	0.9142

—

Table 4 continued

θ	Test	d=1 n=60	150	300	d=1.25 n=60	150	300	d=1.5 n=60	150	300	d=2 n=60	150	300	d=4 n=60	150	300
30°	**Box**	0.7846	0.8988	0.9456	0.8527	0.9432	0.9840	0.9062	0.9834	0.9988	0.9723	0.9994	1.0000	0.9999	1.0000	1.0000
	H_0	0.7472	0.8802	0.9352	0.8099	0.9272	0.9714	0.8584	0.9603	0.9929	0.9267	0.9960	1.0000	0.9991	1.0000	1.0000
	H_0^Λ	0.0488	0.0472	0.0448	0.0885	0.1799	0.3381	0.2012	0.4972	0.8320	0.5350	0.9362	0.9995	0.9908	1.0000	1.0000
	H_0^Γ	0.7382	0.8768	0.933	0.7955	0.9157	0.9590	0.8247	0.9255	0.9664	0.8482	0.9401	0.9720	0.8774	0.9472	0.9731
45°	**Box**	0.8714	0.9477	0.9732	0.9215	0.9775	0.9928	0.9554	0.9946	0.9994	0.9887	0.9998	1.0000	0.9999	1.0000	1.0000
	H_0	0.8284	0.9234	0.9612	0.8967	0.9691	0.9866	0.9316	0.9836	0.9963	0.9651	0.9987	1.0000	0.9996	1.0000	1.0000
	H_0^Λ	0.0470	0.0480	0.0478	0.0940	0.1825	0.3377	0.2050	0.4981	0.8330	0.5411	0.9369	0.9995	0.9912	1.0000	1.0000
	H_0^Γ	0.8226	0.9211	0.9598	0.8896	0.9632	0.9812	0.9151	0.9686	0.9851	0.9281	0.9736	0.9870	0.9434	0.9776	0.9866
60°	**Box**	0.9068	0.9619	0.9814	0.9482	0.9865	0.9958	0.9708	0.9968	0.9997	0.9928	0.9999	1.0000	0.9999	1.0000	1.0000
	H_0	0.8530	0.9379	0.9679	0.9340	0.9845	0.9949	0.9669	0.9945	0.9986	0.9866	0.9995	0.9999	0.9999	1.0000	1.0000
	H_0^Λ	0.0484	0.0478	0.0475	0.0966	0.1845	0.3337	0.2126	0.5046	0.8297	0.5441	0.9388	0.9994	0.9928	1.0000	1.0000
	H_0^Γ	0.8482	0.9352	0.9670	0.9310	0.9820	0.9933	0.9597	0.9897	0.9940	0.9710	0.9904	0.9949	0.9749	0.9903	0.9953
90°	**Box**	0.9278	0.9714	0.9866	0.9615	0.9916	0.9983	0.9795	0.9980	0.9997	0.9960	1.0000	1.0000	1.0000	1.0000	1.0000
	H_0	0.8817	0.9522	0.9760	0.9542	0.9898	0.9984	0.9785	0.9973	0.9995	0.9945	0.9997	0.9999	1.0000	1.0000	1.0000
	H_0^Λ	0.0495	0.0517	0.0499	0.0953	0.1866	0.3388	0.2062	0.5061	0.8262	0.5484	0.9409	0.9996	0.9925	1.0000	1.0000
	H_0^Γ	0.8783	0.9504	0.9752	0.9520	0.9891	0.9979	0.9760	0.9954	0.9984	0.9892	0.9960	0.9985	0.9902	0.9963	0.9986

—

—

In the second scenario, data have been generated from $k = 2$ groups drawn from two trivariate normals ($p = 3$) with mean vectors $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. The same sizes $n = 60$, $n = 150$ and $n = 300$, have been considered with $n_1 = n_2 = n_3$ in all the three cases. As before, $\boldsymbol{\Sigma}_1$ has been randomly generated and its spectral decomposition has been obtained as follows

$$
\boldsymbol{\Sigma}_1 = \boldsymbol{\Gamma}_1 \boldsymbol{\Lambda}_1 \boldsymbol{\Gamma}_1' =
\begin{pmatrix}
\gamma_{11}^{(1)} & \gamma_{12}^{(1)} & \gamma_{13}^{(1)} \\
\gamma_{21}^{(1)} & \gamma_{22}^{(1)} & \gamma_{23}^{(1)} \\
\gamma_{31}^{(1)} & \gamma_{32}^{(1)} & \gamma_{33}^{(1)}
\end{pmatrix}
\begin{pmatrix}
\lambda_1^{(1)} & 0 & 0 \\
0 & \lambda_2^{(1)} & 0 \\
0 & 0 & \lambda_3^{(1)}
\end{pmatrix}
\begin{pmatrix}
\gamma_{11}^{(1)} & \gamma_{12}^{(1)} & \gamma_{13}^{(1)} \\
\gamma_{21}^{(1)} & \gamma_{22}^{(1)} & \gamma_{23}^{(1)} \\
\gamma_{31}^{(1)} & \gamma_{32}^{(1)} & \gamma_{33}^{(1)}
\end{pmatrix}',
$$

with $\lambda_1^{(1)} \geq \lambda_2^{(1)} \geq \lambda_3^{(1)}$. Further, $\boldsymbol{\Sigma}_2$ has been defined by the following relation with $\boldsymbol{\Sigma}_1$

$$
\boldsymbol{\Sigma}_2 =
\begin{pmatrix}
\cos\theta & 0 & \sin\theta \\
0 & 1 & 0 \\
-\sin\theta & 0 & \cos\theta
\end{pmatrix}
\boldsymbol{\Gamma}_1
\begin{pmatrix}
d\lambda_1^{(1)} & 0 & 0 \\
0 & \lambda_2^{(1)} & 0 \\
0 & 0 & \frac{\lambda_3^{(1)}}{d}
\end{pmatrix}
\left[
\begin{pmatrix}
\cos\theta & 0 & \sin\theta \\
0 & 1 & 0 \\
-\sin\theta & 0 & \cos\theta
\end{pmatrix}
\boldsymbol{\Gamma}_1
\right]',
$$

where

$$
\begin{pmatrix}
\cos\theta & 0 & \sin\theta \\
0 & 1 & 0 \\
-\sin\theta & 0 & \cos\theta
\end{pmatrix}
$$

is one of the three possible 3×3 rotation matrices of angle θ. For the rest, the same previous considerations, both on the simulation factors and on the structure of Table 5 with respect to Table 4, apply.

As expected, regarding the comparison between the two different ways of testing homoscedasticity, the Box M-test is generally better, in terms of size and power, than the test obtained by the union-intersection procedure, even if the differences in terms of estimated power are not so remarkable; this is true in both the considered scenarios. Note that the results on the Box M-test with χ^2-approximation have not been reported in Tables 4 and 5 simply because they coincide with those obtained by the F-approximation. Moreover, all the tests are consistent and their power increases in line with the degree of departure from their respective null hypotheses. Also, regarding the proposed UI procedure for H_0, it is worth recalling that strong control of the error rate for the overall null hypothesis H_0 guarantees control for all other implied (component) hypotheses of the hierarchy, due to the consonance property.

7 Concluding remarks and discussion

When considering data sets coming from two or more multivariate normal distributions, the classical definition of homoscedasticity can be decomposed into two facets: equality of the orientation of the components in character hyperspace as defined by the

Table 5 Empirical power for H_0^Λ, H_0^Γ, and H_0, arising from the UI procedure, compared to Box's M-test. Samples are drawn with sizes $n = 60, 150, 300$, by two groups of trivariate normal, varying the shape (d) and the orientation (θ) of the concentration ellipsoids

θ	Test	$d=1$			1.25			1.5			2			4		
	n	60	150	300	60	150	300	60	150	300	60	150	300	60	150	300
0°	**Box**	0.0511	0.0509	0.0513	0.0657	0.0953	0.1709	0.1148	0.2531	0.5267	0.2668	0.7048	0.9694	0.8776	1.0000	1.0000
	H_0	0.0489	0.0374	0.0383	0.0707	0.0861	0.1538	0.1179	0.2326	0.4822	0.2593	0.6634	0.9600	0.8532	0.9998	1.0000
	H_0^Λ	0.0273	0.0257	0.0303	0.0432	0.0717	0.1431	0.0868	0.2188	0.4759	0.2285	0.6565	0.9592	0.8466	0.9998	1.0000
	H_0^Γ	0.0229	0.0124	0.0087	0.0305	0.0168	0.0134	0.0366	0.0204	0.0161	0.0455	0.0298	0.0219	0.0730	0.0389	0.0249
5°	**Box**	0.1014	0.1864	0.3165	0.1286	0.2860	0.4950	0.1881	0.4573	0.7752	0.3726	0.8291	0.9902	0.9182	1.0000	1.0000
	H_0	0.0689	0.1170	0.2244	0.0947	0.1825	0.3493	0.1442	0.3277	0.6297	0.2943	0.7185	0.9747	0.8686	0.9997	1.0000
	H_0^Λ	0.0288	0.0299	0.0313	0.0429	0.0765	0.1519	0.0818	0.2245	0.5116	0.2309	0.6673	0.9656	0.8502	0.9997	1.0000
	H_0^Γ	0.0431	0.0917	0.2020	0.0565	0.1183	0.2450	0.0724	0.1450	0.2852	0.0946	0.1775	0.3385	0.1347	0.2253	0.3842
10°	**Box**	0.2399	0.4739	0.6695	0.2945	0.5827	0.7837	0.3771	0.7229	0.9155	0.5594	0.9282	0.9974	0.9571	1.0000	1.0000
	H_0	0.1704	0.3736	0.5803	0.2157	0.4637	0.6720	0.2780	0.5793	0.8222	0.4242	0.8313	0.9899	0.8973	1.0000	1.0000
	H_0^Λ	0.0301	0.0357	0.0409	0.0470	0.0896	0.1736	0.0839	0.2548	0.5408	0.2423	0.6887	0.9737	0.8536	1.0000	1.0000
	H_0^Γ	0.1469	0.3551	0.5677	0.1811	0.4203	0.6208	0.2171	0.4560	0.6579	0.2565	0.5061	0.7099	0.3244	0.5713	0.7484
15°	**Box**	0.3926	0.6757	0.8299	0.4674	0.7538	0.8968	0.5480	0.8529	0.9646	0.7088	0.9648	0.9989	0.9772	1.0000	1.0000
	H_0	0.3081	0.5881	0.7705	0.3776	0.6660	0.8309	0.4317	0.7473	0.9114	0.5639	0.9031	0.9941	0.9310	0.9998	1.0000
	H_0^Λ	0.0308	0.0372	0.0416	0.0514	0.0986	0.1792	0.0891	0.2669	0.5539	0.2552	0.7149	0.9716	0.8718	0.9998	1.0000
	H_0^Γ	0.2890	0.5765	0.7646	0.3478	0.6381	0.8026	0.3816	0.6787	0.8328	0.4361	0.7142	0.8621	0.5048	0.7670	0.8857

Table 5 continued

θ	Test	d = 1 n=60	150	300	d = 1.25 n=60	150	300	d = 1.5 n=60	150	300	d = 2 n=60	150	300	d = 4 n=60	150	300
30°	Box	0.7308	0.9128	0.9672	0.7824	0.9444	0.9858	0.8337	0.9708	0.9948	0.9128	0.9960	1.0000	0.9958	1.0000	1.0000
	H_0	0.6602	0.8733	0.9535	0.7133	0.9088	0.9696	0.7540	0.9363	0.9879	0.8282	0.9796	0.9996	0.9796	1.0000	1.0000
	H_0^Λ	0.0428	0.0442	0.0472	0.0589	0.1054	0.1959	0.1107	0.2759	0.5702	0.2780	0.7281	0.9774	0.8862	0.9997	1.0000
	H_0^Γ	0.6500	0.8699	0.9527	0.6999	0.9022	0.9657	0.7309	0.9173	0.9748	0.7727	0.9329	0.9802	0.8223	0.9481	0.9838
45°	Box	0.8684	0.9683	0.9899	0.9039	0.9808	0.9965	0.9276	0.9910	0.9989	0.9668	0.9988	1.0000	0.9982	1.0000	1.0000
	H_0	0.8003	0.9454	0.9825	0.8509	0.9686	0.9918	0.8809	0.9801	0.9964	0.9274	0.9938	0.9999	0.9910	1.0000	1.0000
	H_0^Λ	0.0373	0.0444	0.0484	0.0630	0.1078	0.1910	0.1079	0.2888	0.5633	0.2941	0.7309	0.9758	0.8904	0.9999	1.0000
	H_0^Γ	0.7935	0.9438	0.9824	0.8444	0.9656	0.9910	0.8715	0.9737	0.9940	0.9026	0.9821	0.9959	0.9333	0.9860	0.9963
60°	Box	0.9208	0.9819	0.9950	0.9453	0.9918	0.9989	0.9638	0.9968	0.9996	0.9840	0.9994	1.0000	0.9994	1.0000	1.0000
	H_0	0.8653	0.9693	0.9902	0.9109	0.9853	0.9961	0.9332	0.9916	0.9988	0.9632	0.9977	1.0000	0.9956	1.0000	1.0000
	H_0^Λ	0.0390	0.0445	0.0487	0.0642	0.1083	0.1932	0.1123	0.2950	0.5620	0.2986	0.7303	0.9745	0.8919	0.9998	1.0000
	H_0^Γ	0.8608	0.9686	0.9902	0.9079	0.9836	0.9958	0.9281	0.9890	0.9979	0.9538	0.9942	0.9985	0.9742	0.9958	0.9987
90°	Box	0.9576	0.9921	0.9984	0.9742	0.9966	0.9990	0.9811	0.9988	0.9998	0.9934	0.9998	1.0000	0.9998	1.0000	1.0000
	H_0	0.9189	0.9831	0.9951	0.9531	0.9936	0.9989	0.9675	0.9970	0.9999	0.9866	0.9994	0.9996	0.9985	1.0000	1.0000
	H_0^Λ	0.0385	0.0448	0.0443	0.0650	0.1061	0.1897	0.1083	0.2938	0.5574	0.2873	0.7315	0.9730	0.8849	1.0000	1.0000
	H_0^Γ	0.9167	0.9822	0.9950	0.9512	0.9931	0.9988	0.9659	0.9960	0.9998	0.9828	0.9984	0.9999	0.9930	0.9987	0.9998

eigenvectors (here defined as homotroposcedasticity), and equality of the spread (variance) of cases along the components as defined by the eigenvalues (here defined as homometroscedasticity). The spectral decomposition of the covariance matrices represents some geometrical properties related to the underlying group distributions that can be particularly useful in many fields, say, for example, for biometrical data. Thus, instead of limiting the attention on the usual test for the hypothesis of equality of the scatters, an augmentation UI procedure with control of the FDX has been introduced. When homoscedasticity is rejected by the new inferential method, one can still take advantage, for example to obtain a better parametrization, by the equality of shape and size or by the equality in orientation of the group scatters, whenever one of them holds true. We are firmly convinced that the appeal of this inferential procedure resides in the fact that a more detailed statistical test is useful to disclose richer information, as is acknowledged by the analysis of some real data sets and by a wide simulation study. Nevertheless, further effort can be devoted to assess its performance by simulations with different kinds of dependence among the underlying test statistics.

An interesting alternative is to apply the same augmentation UI procedure to each of the two hierarchical sub-families having homotroposcedasticity, $H_0^{\mathbf{\Gamma}}$, and homometroscedasticity, $H_0^{\mathbf{\Lambda}}$, as overall null hypotheses, providing two multiple testing procedures to assess them. Logically, FDX-control would be hence guaranteed on each hypothesis but not on the implying hypothesis H_0 of homoscedasticity. Note that, for the tests referred to homometroscedasticity, in which we have independence among test statistics, one could improve the performance in the first step of the augmentation procedure by substituting the Holm method with the Šidák-based procedure introduced by Holland and Copenhaver (1987).

Naturally, the multiple testing procedure proposed in the present work reflects one of the possible choices to cope with the multiplicity problem arising from the hierarchical family of hypotheses $\mathcal{H}$. However, any other alternative method has to take into account a threefold choice: the procedure (we have leaned towards the simple, coherent and consonant UI), the type of error rate (we have chosen the FDX-control because of its robustness to dependence), and the method used to control it (we have adopted augmentation for its simplicity and validity under arbitrary dependence). For example, as suggested by a referee, one could well prefer to employ closed multiple testing procedures, described in Marcus et al. (1976), that are the most powerful procedures in FWER-controlling (see Shaffer 1995, p. 568). Besides, one could lean towards a step-down procedure that begins by testing H_0 by means of Box's M-test, and then steps down through the hierarchy of implied minimal hypotheses. If H_0 is not rejected, then all of its implied hypotheses are retained without further tests; thus the minimal hypotheses are tested if, and only if, H_0 is rejected. Although this procedure is probably more powerful than our proposal (this comparison is in principle quite similar to that described in Hochberg and Tamhane 1987, pp. 2–5, between the protected least significant difference test and the Bonferroni procedure), it seems to present the drawback of giving only weak control of the error rate (under H_0 only; see Shaffer 1995, p. 566) and consonance may not hold.

Finally, an aspect requiring further study, as highlighted by a referee, is related to the following consideration. The individual relevance (i.e. power) of $H_0^{\mathbf{\Lambda}}$ strongly depends

on the dimension p and, moreover, it decreases when the number of groups k increases (we recall that the number of minimal hypotheses to test is $m = k(p-1) + 1$). A similar comment applies for H_0 and H_0^{Γ}. Hence, a source of possible improvement arises by looking for recent inferential methods that, taking advantage of the hierarchical structure of the hypotheses, allows for a better distribution of α on the various nodes of the tree (see, among others, Bretz et al. 2009; Burman et al. 2009; Goeman and Finos 2010). Many of these inferential methods can be seen as special cases of the sequential rejection principle of Goeman and Solari (2010), spurring us towards further developments in this direction.

References

Anderson E (1935) The irises of the Gaspe peninsula. Bull Am Ir Soc 59:2–5

Banfield JD, Raftery AE (1993) Model-based gaussian and non-gaussian clustering. Biometrics 49(3):803–821

Bartlett MS (1937) Properties of sufficiency and statistical tests. Proc R Stat Soc Lond Ser A Math Phys Sci 160(901):268–282

Benjamini Y (2010) Discovering the false discovery rate. J R Stat Soc Ser B (Methodol) 72(4):405–416

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol) 57(1):289–300

Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni dell'Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8(1):3–62

Bretz F, Maurer W, Brannath W, Posch M (2009) A graphical approach to sequentially rejective multiple test procedures. Stat Med 28(4):586–604

Burman CF, Sonesson C, Guilbaud O (2009) A recycling framework for the construction of Bonferroni-based multiple tests. Stat Med 28(5):739–761

Campbell NA, Mahon RJ (1974) A multivariate study of variation in two species of rock crab of genus Leptograpsus. Aust J Zool 22(3):417–425

Celeux G, Govaert G (1995) Gaussian parsimonious clustering models. Pattern Recognit 28(5):781–793

Dudoit S, van der Laan MJ (2008) Multiple testing procedures with applications to genomics. Springer, New York

Farcomeni A (2008) A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Stat Methods Med Res 17(4):347–388

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7(2):179–188

Flury BN (1984) Common principal components in k groups. J Am Stat Assoc 79(388):892–898

Flury BN (1988) Common principal components and related multivariate models. Wiley, New York

Flury BN, Constantine G (1985) The F-G diagonalization algorithm. Appl Stat 35:177–183

Flury BN, Gautschi W (1986) An algorithm for simultaneous orthogonal transformation of several positive definite matrices to nearly diagonal form. SIAM J Sci Stat Comput 7:169–184

Flury BN, Riedwyl H (1983) Angewandte multivariate statistik. Verlag Gustav Fischer, Jena

Gabriel KR (1969) Simultaneous test procedures–some theory of multiple comparisons. Ann Math Stat 40(1):224–250

Genovese CR, Wasserman L (2006) Exceedance control of the false discovery proportion. J Am Stat Assoc 101(476):1408–1417

Goeman J, Finos L (2010) The inheritance procedure: multiple testing of tree-structured hypotheses (unpublished preprint dowloadable from http://www.msbi.nl/dnn/Default.aspx?tabid=202)

Goeman J, Solari A (2010) The sequential rejection principle of familywise error control. Ann Stat (to appear)

Greselin F, Ingrassia S (2009) Weakly homoscedastic constraints for mixtures of t distributions. In: Fink A, Lausen B, Seidel W, Ultsch A (eds) Advances in data analysis, data handling and business intelligence. Springer, Berlin, pp 219–228

Greselin F, Ingrassia S (2010) Constrained monotone EM algorithms for mixtures of multivariate t distributions. Stat Comput 20(1):9–22

Hawkins DM (1981) A new test for multivariate normality and homoscedasticity. Technometrics 23(1): 105–110

Hochberg Y, Tamhane AC (1987) Multiple comparison procedures. Wiley, New York

Holland BS, Copenhaver MDP (1987) An improved sequentially rejective Bonferroni test procedure. Biometrics 43(2):417–423

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6(2):65–70

Jolicoeur P (1963) The degree of generality of robustness in Martes Americana. Growth 27:1–27

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

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

Mardia KV (1985) Mardia's test of multinormality. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, vol 5. Wiley, New York, pp 217–221

McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York

Murtagh F, Raftery A (1984) Fitting straight lines to point patterns. Pattern Recognit 17(5):479–483

Peel D, McLachlan GJ (2000) Robust mixture modelling using the t distribution. Stat Comput 10(4):339–348

Rencher AC (1998) Multivariate statistical inference and applications. Wiley, New York

Ripley B (1996) Pattern recognition and neural network. Cambridge University Press, Cambridge

Rosenthal R, Rubin DB (1983) Ensemble adjusted p-values. Psychol Bull 94(3):540–541

Shaffer JP (1995) Multiple hypothesis testing. Ann Rev Psychol 46(1):561–584

Sheskin DJ (2000) Handbook of parametric and nonparametric statistical procedures. Chapman & Hall, London

Van der Laan MJ, Duduoit S, Pollard KS (2004) Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. Stat Appl Genet Mol Biol 3(1):Article 15

Westfall PH, Young SS (1993) Resampling-based multiple testing: examples and methods for p-value adjustment. Wiley, New York

Wright SP (1992) Adjusted p-values for simultaneous inference. Biometrics 48(4):1005–1013