IMS Collections
Borrowing Strength: Theory Powering Applications –
A Festschrift for Lawrence D. Brown
Vol. 6 (2010) 144–157

DOI: 10.1214/10-IMSCOLL610

Multiple testing of pairwise comparisons

Arthur Cohen, Harold Sackrowitz and Chuanwen Chen

Rutgers, The State University of New Jersey

Abstract: Multiple testing for pairwise comparisons in a one way fixed and balanced analysis of variance model is studied. Normality, independence and homogeneity of variance is assumed. Two sided alternatives are considered. The usual stepwise procedures are shown to lack an intuitive and important interval property for acceptance sections of individual tests. This renders them inadmissible in terms of both types of errors. Alternative procedures that do have the interval property are suggested. The new procedures are compared to the standard procedures in terms of power and in terms of practicality. One of the new procedures follows logical restrictions which may be a desirable property in some instances.

Contents

1. Introduction

Multiple testing of pairwise comparisons in analysis of variance models is a long standing statistical practice. The classical Tukey procedure is the standard method which is offered in most textbooks. With the rise of multiple testing in the last two decades, given impetus by fields of application calling for many tests, new and less conservative approaches to tests of pairwise differences have arisen. The Tukey method is a single step method as opposed to stepwise methods suggested by [7, 13, 14, 15], and others. [11] (Section 9.3), have an excellent summary of this topic.

For multiple testing of means of a normal random vector stepwise procedures are popular methods. See, for example, [4]. Many of these procedures are based on P-values determined from marginal distributions of test statistics that are correlated. In the case of pairwise comparisons the test statistics used are correlated and have a known covariance matrix (except for a scalar multiple). This sort of situation occurs in many multiple testing settings. Some stepwise procedures do try to incorporate the dependence by using a closed testing framework (see [9]) or by using resampling

Research was supported by NSF Grant DMS-08-94547 and NSA Grant H-98230-06-0076.

Keywords and phrases: admissibility, logical restrictions, maximum residual down (MRD), step-down procedure, Tukey's method.

AMS 2000 subject classifications: Primary 62H15, 62C15.

based procedures (see [11]). Nevertheless many standard stepwise procedures would use the same critical values regardless of the correlation.

In constructing stepwise procedures it is common to begin with tests for the individual hypotheses that are known to have desirable properties. For example, [13] says to begin with unbiased tests. Next we realize that all stepwise procedures induce new tests on the individual testing problems. Furthermore, in a multivariate normal model with correlation, theoretical properties of these new induced tests depend heavily on the precise form of the covariance matrix. The nature of these induced tests is typically overlooked. However, when studied, we find that many of the stepwise procedures lack an important theoretical as well as intuitive property. Namely, certain acceptance sections (as indicated by the specific correlation structure) for the induced tests are not intervals when they absolutely should be. The basis of our approach is to develop procedures whose induced tests have acceptance regions satisfying the conditions mandated by the particular covariance structure.

We clarify these issues with the simplest special case of our model (described in detail in Section 2). Let X_i have a $N(\mu_i, 1)$ distribution for i = 1, 2, 3 where the X_i are independent. We wish to test the hypotheses $H_{12} : \mu_1 = \mu_2$ vs. $K_{12} : \mu_1 \neq \mu_2$, $H_{13} : \mu_1 = \mu_3$ vs. $K_{13} : \mu_1 \neq \mu_3$ and $H_{23} : \mu_2 = \mu_3$ vs. $K_{23} : \mu_2 \neq \mu_3$. If we were testing H_{ij} only we would use a test of the form: Reject H_{ij} if and only if $|X_i - X_j| \geq c$. This test has very good properties. It is the unique UMPU, is admissible and has an interval acceptance region. Most stepwise procedures for all three hypothesis testing problems would be based on these individual test statistics alone while using various choices of critical values. Thus the correlation does not enter into the functional form of the test statistics.

In the example above, the test of H_{ij} induced by most stepwise procedures will no longer enjoy any of the desirable properties of the original tests. In fact, it will likely have some undesirable properties.

For example, using the procedure of [7] or S1 of [14], the acceptance section for testing H_{12} will not even always be an interval for every fixed X_3. In fact, for every fixed X_3 there are values x and $0 < \Delta_1 < \Delta_2$ such that the induced test will exhibit the following behavior. When $X_1 = x = X_2$ is observed H_{12} will be accepted. Next we let X_1 and X_2 move apart. The induced test will lead to rejection of H_{12} when $X_1 = x + \Delta_1, X_2 = x - \Delta_1$ but it will again lead to acceptance of H_{12} when $X_1 = x + \Delta_2, X_2 = x - \Delta_2$. We will see more of this later in Figure 1. We will also see, by Lemma A.1, that this implies inadmissibility (defined below).

In 4 dimensions a more intuitive sense of this objectionable behavior can be obtained by looking at Example 2.1 of [13]. Here we focus on the decision for testing H_{34}. In that example there are four means to be compared pairwise. The observed values are $X_1 = 0.0,\ X_2 = 3.2,\ X_3 = 3.3$ and $X_4 = 6.55$. The purpose of that example was to show that the closed test will reject H_{34} but accept H_{13} even though $|X_1 - X_3| = 3.3 > 3.25 = |X_3 - X_4|$. This behavior was considered to be, practically, undesirable. It was pointed out that using the Royen procedure one would accept H_{34} for this data. What had not been noticed is the following. Suppose we keep X_1 and X_2 fixed while bringing X_3 and X_4 even closer together. For example, take $X_1 = 0.0,\ X_2 = 3.2,\ X_3 = 3.32$ and $X_4 = 6.53$. Surprisingly, for this data, the Royen procedure (available on SAS) will now reject H_{34}.

This type of behavior is also unappealing for practical purposes and cannot occur with the methods we present. Furthermore, it is often enough to render the test inadmissible. Here admissibility is with respect to the usual 0 - 1 testing loss function. That is, there is a loss of 1 for a mistake and 0 for a correct decision. Thus if a test of any one of the hypotheses is inadmissible there exists a test with smaller probability of Type I error and greater power. Furthermore, for the procedure as a whole using the better tests would result in fewer expected Type I errors and fewer expected Type II errors. We do not construct procedures that are theoretically better (this is not computationally feasible). We do offer procedures that are admissible and are, at least, competitive if not more favorable in terms of operating characteristics as demonstrated by simulation. See [2]. This is a serious shortcoming of many existing popular stepwise procedures.

Lastly is the issue of what is called logical restrictions. We notice that if H_{12} and H_{13} are true then H_{23} must also be true. Logical restrictions are discussed thoroughly in [11] and [14]. Such restrictions are important in the evaluation of multiple testing procedures (MTP). Recognizing this can result in better choices of critical values and greater power (see [14]). However, typical MTPs do not obey logical restrictions in their actions. That is, observed data sets can easily result in the acceptance of H_{12} and H_{13} but the rejection of H_{23}. When this is done the interpretation is that we are "convinced" that $\mu_2 \neq \mu_3$ but do not have enough evidence to claim either $\mu_1 \neq \mu_2$ or $\mu_1 \neq \mu_3$.

The desirability of obeying logical restrictions in the actions depends on the application at hand. For example, when the goal is some type of sorting or classification scheme and every population must be placed somewhere, saying that we are not sure may not be an option.

Suppose that the level of state funding for school districts depends on teaching success as measured by performance on standardized exams. Districts are compared by doing all pairwise tests on exam scores. If it is concluded that two districts differ in performance then they will be funded at different levels. Otherwise they are funded at the same level. We cannot say that A and B get equal funding and B and C get equal funding yet A and C get different funding. In this instance a multiple testing procedure of pairwise comparisons that follows logical restrictions in the actions is necessary.

Another common situation of this type is the assignment of grades. Instructors assign grades by comparing students on the semester's body of work. Students who are found to perform comparably receive the same grade otherwise they get different grades. Instructors cannot say that they are not sure.

The first method we introduce is labeled PADD (which reads partitioned average difference down) and does follow logical restrictions in its actions. It is based on sample averages determined by partitions of individual treatment means. The second method does not follow logical restrictions in actions and has the flexibility to accept or reject each and every hypothesis regardless of the decisions made for the other hypotheses. A method that does have such flexibility is the Tukey single step. The procedure we recommend is a combination of PADD followed by a screening stage using a modification of the Tukey method based on pairwise differences of sample means and our new step-down method. The combination procedure will be designated as PADD+. The + signifies the screening stage.

We apply PADD+, Royen and Shaffer's S1 to a 6-dimensional, example and discuss their performance based on a simulation study. The power functions of PADD+ and Royen were almost identical and somewhat preferable to S1. PADD+ has the advantage of admissibility, convex acceptance sections. It is also easy to compute and can easily handle high dimensions. Furthermore, in cases where logical restrictions in the actions must be adhered to one can simply omit the screening stage.

In response to shortcomings of typical stepwise procedures, [3] proposed a new method called Maximum Residual Down (MRD) for a multivariate normal model with a nonsingular known covariance matrix. This method is derived so that the focus is to do well for each individual test in terms of expected Type I error and expected Type II error. As such any optimality property it has for individual testing risk will hold for any risk that is an increasing function of these risks. This includes the classification risk function used by [10, 8] and [6]. The method is not only admissible for such risk functions but simulations indicate a superiority of the method in terms of numbers of errors, compared to popular stepwise methods in the models considered. In light of this we seek an extension of the MRD method to apply to the problem of multiple testing of pairwise comparisons. The MRD construction depended heavily on the non-singularity of the covariance matrix. In this situation we can consider a random vector which is multivariate normal but now the covariance matrix is singular. Nevertheless we initially derive a method based on the analogues of the residuals.

The theoretical optimality property of the new method is proven for a normal balanced one way analysis of variance model assuming the variance of all observations is known. For the more realistic case where the variance is unknown, we would simply replace the role of the variance by the mean square error (MSE).

In the next section we state the model and give some preliminaries. In Section 3 we outline the new method and give an example and discuss performance based on a simulation study. In Section 4 we state the results concerning admissibility. All proofs are given in the appendix.

2. Model and preliminaries

The model is that of a balanced fixed effects one way analysis of variance. That is $X_{ij}, i = 1 \ldots, I; j = 1, \ldots J$ are independent, normally distributed with mean μ_i and variance σ^2. For now assume σ^2 is known, let $J = 1$ without loss of generality and also suppress j in X_{ij}. Later σ^2 will be unknown and $J > 1$. For all $i \neq i'$ we wish to test the $q = C_2^I$ null hypotheses $H_{ii'} : \delta_{ii'}{=}\mu_i - \mu_{i'} = 0$ vs. $K_{ii'} : \delta_{ii'} \neq 0$.

We note that the C_2^I pairwise differences are subject to logical restrictions. Despite the existence of logical restrictions in the parameter space all procedures that have been proposed in the past do not follow logical restrictions in their actions. In many applications this seems reasonable.

For suppose we are comparing three populations A, B and C where sample means are ordered. One might want to say that A is different from C but cannot conclude that A differs from B or that B differs from C. However, in settings where the outcomes of acceptance and rejection require contradictory actions violations in logical restrictions is not an option. Recall the example given in the introduction.

We now define the weak and strong familywise error rate (FWER).

Definition 2.1. The weak FWER is the probability of rejecting at least one of the q hypotheses when all null hypotheses are true.

Definition 2.2. Suppose a subset of the q hypotheses are true and the remaining hypotheses are not true. Then the probability of rejecting at least one true hypothesis is called the strong FWER. It is understood that this error rate is a function of those parameters related to the hypotheses that were not true.

At this point we describe the Tukey single step procedure and then describe a class of step-down procedures. This latter class includes those procedures recommended by [7, 13], and [14]. The truncated procedure of [15] is an extension of these latter two for any collection of contrasts. The Tukey procedure based on the distribution of the range rejects $H_{ii'}$ if and only if

$$Y_{ii'} = |X_i - X_{i'}|/\sigma > C_T. \tag{2.1}$$

Typically C_T depends on α where $(1 - \alpha)$ is the probability of accepting all hypotheses when they are all true.

To describe a class of step-down procedures let $0 \leq \mathrm{C_1} \leq \cdots \leq \mathrm{C_q}$ be a sequence of critical values. Let $Y_{(1)} \leq Y_{(2)} \cdots \leq Y_{(q)}$ correspond to order statistics for $Y_{(i)}$.

(i) If $Y_{(q)} > C_q$ reject $H_{(q)}$. Otherwise stop and accept $H_{(\gamma)}, \gamma = 1, \ldots, q$.

(ii) If $H_{(q)}$ is rejected, reject $H_{(q-1)}$ if $Y_{(q-1)} > C_{q-1}$. Otherwise stop and accept $H_{(1)}, \ldots, H_{(q-1)}$.

(iii) In general, at stage m, if $Y_{(q-m+1)} > C_{q-m+1}$ reject $H_{(q-m+1)}$. Otherwise stop and accept $H_{(1)}, \ldots, H_{(q-m+1)}$. The critical values can be chosen in a variety of ways. Oftentimes C_q is chosen at stage 1 to control weak FWER (same as in the Tukey procedure) and the other $C's$ are chosen to control strong FWER.

We evaluate the collection of q tests by evaluating each individual test by its expected Type I error and expected Type II error. In terms of admissibility, multiple testing procedures that are inadmissible for these individual problems would remain so if the risk was any non-decreasing function of the collection of individual expected values.

3. PADD+

In this section we describe PADD and PADD+. PADD is an outgrowth of the MRD method developed in [3]. The motivation for the method is based on a theorem that yields a necessary and sufficient condition for admissibility of a test for a hypothesis concerned with a single parameter when there are other parameters in the model. In particular, when the $m \times 1$ random vector $\mathbf{Y}$ has a full rank exponential family density

$$f(\underline{y} \mid \underline{\theta}) = h(\underline{y})\beta(\underline{\theta}) \exp \sum_{i=1}^{m} t_i(\underline{y})\theta_i.$$

Suppose we wish to test $H : \theta_1 = 0$ vs. $K : \theta_1 \neq 0$. Then (see Lemma A.1 of the Appendix) a necessary and sufficient condition for a test $\phi_1(\underline{y})$ to be admissible is that for fixed $t_2, \ldots, t_m$, $\phi_1(\underline{y})$ can only be zero (i.e. accept) on an interval of t_1 values. We call the $t_i(\underline{y})$ residuals.

The situation in this paper, dealing with all pairwise differences of mean parameters, does not directly deal with a single full rank exponential family. However by taking linearly independent subsets of size I-1 from the variables that represent sample mean differences, we can generate residuals for each full rank subset. These residuals turn out to be average differences for all possible groupings of sample means. We will utilize these residuals in such a way that they satisfy the conditions required for admissibility. Some clarification of the above can be gained by studying the Appendix.

Both methods, PADD and PADD+, are based on what we have (see above) called residuals. Decision theory tells us how the residuals should be used to attain the desirable optimality properties. The situation in this paper is more complicated than the MRD paper because the density representing the data does not include all the pairwise differences to be tested. Nevertheless, by adaptively generating residuals when considering subsets of all the pairwise differences we come up with a complete collection of residuals. These residuals turn out to be average differences for all possible groupings of sample means. We proceed to formally define PADD and PADD+.

Let $S = [1, \ldots, I]$. For any subset of integers $A \subset S$ let N(A) = the number of points in A. Let $\overline{X}_A = \sum_{i \in A} X_i / N(A)$. Next define, for all $A \subset B \subseteq S$ with $A \neq \phi \neq B \setminus A$, for each sample point $\underline{x}$,

$$D_{\underline{x}}(A; B) = (\overline{X}_A - \overline{X}_{B \setminus A})) / \sigma(1/N(A) + 1/N(B \setminus A))^{1/2} \tag{3.1}$$

and

$$D^*_{\underline{x}}(B) = \max_{A \subset B} D_{\underline{x}}(A; B). \tag{3.2}$$

Thus $D^*_{\underline{x}}(B)$ is the largest possible standardized difference in subset means when the set of $[X_i : i \in B]$ is broken into two non-empty subsets whose union is $[X_i : i \in B]$. We further let $V_{\underline{x}}(B)$ denote the set for which the maximum is attained. That is,

$$D^*_{\underline{x}}(B) = D_{\underline{x}}(V_{\underline{x}}; B) \textit{ when } B \textit{ is split into } V_{\underline{x}}(B) \textit{ and } B \setminus V_{\underline{x}}(B).$$

At the first stage of PADD all non-empty 2 set partitions of S are considered. $D_{\underline{x}}(A; S)$ is computed for all non-empty $A \subset S$. Letting $C(S)$ denote a constant at stage 1 and letting $D_1 = D^*_{\underline{x}}(S)$, if $D_1 \leq C(S)$ stop and accept all null hypotheses. If $D_1 > C(S)$ then partition S into $V_{\underline{x}}(S)$ and $S \setminus V_{\underline{x}}(S)$ and continue to stage 2.

At each successive stage, until the procedure stops, one of the sets in the current partition will be split into two sets as follows: Suppose that after stage n, S has been partitioned into $B_1, B_2, \ldots, B_{n+1}$ and we continue. Let $C\{B_1, \ldots, B_{n+1}\}$ be a constant determined by the partition $\{B_1, \ldots, B_{n+1}\}$. Compute $D_{n+1} = \max_{1 \leq k \leq n+1} D^*_{\underline{x}}(B_k)$. If $D_{n+1} \leq C\{B_1, \ldots, B_{n+1}\}$ we stop. If $D_{n+1} > C\{B_1, \ldots, B_{n+1}\}$ find k* so that $D_{n+1} = D^*_{\underline{x}}(B_{k^*})$. Next break B_{k^*} into $V_{\underline{x}}(B_{k^*})$ and $B_{k^*} \setminus V_{\underline{x}}(B_{k^*})$. Continue to stage n + 1.

Thus we see that as we enter stage n the partition consists of n sets. Denote these by $B_{n,1}(\underline{x}), \ldots, B_{n,n}(\underline{x})$. If $D_n \leq C\{B_{n,1}(\underline{x}), \ldots, B_{n,n}(\underline{x})\}$, stop and then $\{B_{n,1}(\underline{x}), \ldots, B_{n,n}(\underline{x})\}$ is the final partition. If $D_n > C\{B_{n,1}(\underline{x}), \ldots, B_{n,n}(\underline{x})\}$ we continue and the partition will become finer. If $\{B_{n,1}(\underline{x}), \ldots, B_{n,n}(\underline{x})\}$ is the final partition then $H_{ii'}$ is accepted provided i and i′ are in the same set of the partition. Otherwise $H_{ii'}$ is rejected.

Note that computationally this step merely requires computation of (I-1) statistics that are expressed in terms of the order statistics. That is, suppose we have J observations in each column and let $\overline{X}_i$ be the sample mean for each column. The ordered sample means are $\overline{X}_{(1)} < \overline{X}_{(2)} < \cdots < \overline{X}_{(I)}$. The relevant (I - 1) statistics are as follows:

$$\sqrt{\mathrm{J}} \left\{ \sum_{\mathrm{j}=1}^{\mathrm{i}} \bar{X}_{(\mathrm{j})}/\mathrm{i} - \sum_{\mathrm{j}=\mathrm{i}+1}^{\mathrm{I}} \bar{X}_{(\mathrm{j})}/(\mathrm{I}-\mathrm{i}) \right\} / \sigma \sqrt{\{1/\mathrm{i} + 1/(\mathrm{I}-\mathrm{i})\}}.$$

Use of the order statistics also facilitates the computations at all stages. (See the example below).

There is considerable flexibility in the choice of critical values $C\{B_{n,1}, \ldots, B_{n,n}\}$. One way to choose them is to simply allow them to depend on the stage n. Another way to choose them is to let them depend on the number of indices in the largest set of the partition. Still another way is to let them depend on the total number of pairwise comparisons to be made by adding up the pairwise comparisons in each set of the partition.

Since at any given stage no single index can be in more than one subset it implies that PADD follows logical restrictions. We regard this as too restrictive in some applications in that it does not allow for the possibility of deciding on each and every hypothesis separately. For example if I = 3, the PADD does not allow the possibility of rejecting one hypothesis and accepting the two others. Intuitively there are sample points where one may want this option. Furthermore there are sample points in which the $Y_{ii'}$'s are small and yet PADD could reject $H_{ii'}$.

When more flexibility is needed we propose to supplement PADD with Tukey-like procedures. That is we specify two additional constants, $C_L \leq C_U$. Then $H_{ii'}$ will be rejected if and only if the indices i and i' lie in different sets of the final partition, $\{B_{n,1}(\underline{x}), \ldots, B_{n,n}(\underline{x})\}$ and $Y_{ii'} > C_L$ or i and i' lie in the same set of the final partition but $Y_{ii'} > C_U$. We call the combined procedure PADD+.

In the ANOVA model if σ^2 is unknown and $J > 1$, then the role of X_i would be replaced by $\overline{X}_i = \sum_{j=1}^{J} X_{ij}/J$, $i = 1, \ldots, I$, and $s^2 = MSE$ would replace σ^2. $I(J-1)s^2/\sigma^2$ has a chi-squared distribution with I(J-1) degrees of freedom.

We conclude this section with an example. The data appear in exercise 21, p. 385 of the multiple comparison section of [5]. The data are survival times of rats exposed to nitrogen dioxide. There are I = 6 groups with J = 14 rats in each group. With sample sizes of 14 and a list of variances given in the exercise, we feel that the assumptions for analysis are satisfied. The 6 means in increasing order are

$$\bar{X}_{(1)} = 166, \bar{X}_{(2)} = 184, \bar{X}_{(3)} = 202, \bar{X}_{(4)} = 212, \bar{X}_{(5)} = 266, \bar{X}_{(6)} = 303.$$

We calculate s = 41.01. We will select constants at any stage determined by the number of indices in the set of the partition with the largest number of indices. The actual constants are determined by simulation in deference to an FWER of $\alpha = .05$. For the screen stage the critical values are $C_L = 2.687$, $C_U = 3.491$ and for the PADD stage they are $C_i = F_t^{-1}(i\alpha/(2(7 - i(1 - \alpha/2))))$, for $i = 2, \ldots, 6$ where F_t is the cdf of the t-distribution with 78 degrees of freedom. These C_i are suggested by the work of [1] for a different problem with correlation but also work well here. Note C_6 is used at stage 1 since there are 6 indices in S.

At stage 1 we need to compute the following 5 statistics:

$$\sqrt{14}\left(\bar{X}_{(1)} - \left(\sum_{j=2}^{6} \bar{X}_{(j)}\right)/5\right)/s\left(1 + {}^{1}\!/\!_{5}\right)^{{}^{1}\!/\!_{2}} = -5.6138, \tag{3.3}$$

$$\sqrt{14}\left(\bar{X}_{(6)} - \left(\sum_{j=1}^{5} \bar{X}_{(j)}\right)/5\right)/s\left(1 + {}^{1}\!/\!_{5}\right)^{{}^{1}\!/\!_{2}} = 9.7503, \tag{3.4}$$

$$\sqrt{14}\left(\left(\bar{X}_{(1)} + \bar{X}_{(2)}\right)/2 - \sum_{j=3}^{6} \bar{X}_{(j)}/4\right)/s\left({}^{1}\!/\!_{2} + {}^{1}\!/\!_{4}\right)^{{}^{1}\!/\!_{2}} = -7.4547, \tag{3.5}$$

$$\sqrt{14}\left(\left(\bar{X}_{(6)} + \bar{X}_{(5)}\right)/2 - \sum_{j=1}^{4} \bar{X}_{(j)}/4\right)/s\left({}^{1}\!/\!_{2} + {}^{1}\!/\!_{4}\right)^{{}^{1}\!/\!_{2}} = 9.8499, \tag{3.6}$$

(3.7)

$$\sqrt{14}\left(\left(\bar{X}_{(1)} + \bar{X}_{(2)} + \bar{X}_{(3)}\right)/3 - \left(\bar{X}_{(4)} + \bar{X}_{(5)} + \bar{X}_{(6)}\right)/3\right)/s\left({}^{2}\!/\!_{3}\right)^{{}^{1}\!/\!_{2}} = -9.7442,$$

Using the data we find the maximum of the absolute value of the above 5 statistics occurs for (3.6) with the value of the statistic 9.8499. Since this exceeds $C_6 = 3.3532$ the partition at stage 2 consists of the two sets {1,2,3,4} and {5,6} so that we reject the following 8 hypotheses:

$$H_{(1)(6)}, H_{(2)(6)}, H_{(3)(6)}, H_{(4)(6)}, H_{(1)(5)}, H_{(2)(5)}, H_{(3)(5)}, \text{and} H_{(4)(5)}.$$

At this stage we consider the following 4 statistics:

$$\begin{aligned}
&\sqrt{14}\left(\bar{X}_{(5)} - \bar{X}_{(6)}\right)/s\sqrt{2} = -2.3868,\\
&\sqrt{14}\left(\bar{X}_{(1)} - \left(\bar{X}_{(2)} + \bar{X}_{(3)} + \bar{X}_{(4)}\right)/3\right)/s\left(1 + {}^1\!/\!_3\right)^{1/2} = -2.2124,\\
&\sqrt{14}\left(\bar{X}_{(4)} - \left(\bar{X}_{(1)} + \bar{X}_{(2)} + \bar{X}_{(3)}\right)/3\right)s\left(1 + {}^1\!/\!_3\right)^{1/2} = 2.6075,\\
&\sqrt{14}\left(\bar{X}_{(1)} + \bar{X}_{(2)} - \bar{X}_{(3)} - \bar{X}_{(4)}\right)/2s = -2.9196.
\end{aligned}$$

The largest of these in absolute value is $\sqrt{14}\left(\bar{X}_{(1)} + \bar{X}_{(2)} - \bar{X}_{(3)} - \bar{X}_{(4)}\right)/2s = 2.9196$ which exceeds $C_4 = 2.6914$. We are now left with the partition {1,2}, {3,4} and {5,6}. Thus at this stage, hypotheses $H_{(1)(3)}, H_{(1)(4)}, H_{(2)(3)}$ and $H_{(2)(4)}$ are rejected. The t-statistics for the remaining pairs in the partition are all less than $C_L = 2.687$. Thus these hypotheses are not rejected. Now all 12 hypotheses previously rejected are reconsidered. Their pairwise t-statistics are compared with $C_L = 2.687$. Hypotheses $H_{(1)(3)}, H_{(2)(3)}, H_{(2)(4)}$ give rise to t-statistics which are less than 2.687 and now are accepted. The final decisions for PADD+ are to reject 9 hypotheses.

For this data set [14] S1 procedure, [13] procedure and Tukey's single step procedure at FWER = .05 all reject the same 9 hypotheses.

The most natural procedure to compare with PADD+ is Royen's method. A simulation was done to study the power function behavior of the two methods for a general six population problem with 14 observations per population. This was done over a wide variety of more than 50 parameter points in the following way.

The simulation computed the average probability of the Type I errors over all hypotheses for which the null hypothesis is true at a given parameter point. For example, if the parameter point is (0, 2, 0, 2, 3, 4) the only true null hypotheses are H_{13} and H_{24}. Thus the probability of Type I error for this parameter point would be the average of the probabilities of rejecting H_{13} and H_{24}. Note that this probability depends on all six parameters (0, 2, 0, 2, 3, 4). Also, in the simulation, power was taken as the average of the probabilities of rejecting each of the null hypotheses that was not true.

There was, essentially, no difference in power between PADD+ and the Royen method for the more than 50 varied parameter points considered.

4. Properties of PADD and PADD+

In this section we will state that PADD and PADD+ do possess the intuitive and desirable interval property for acceptance sections. This latter property is a necessary and sufficient condition for admissibility in the variance known case. We also state that most stepwise procedures based on P-values determined from marginal distributions of test statistics lack this property and therefore are inadmissible. Proofs will be given in the Appendix.

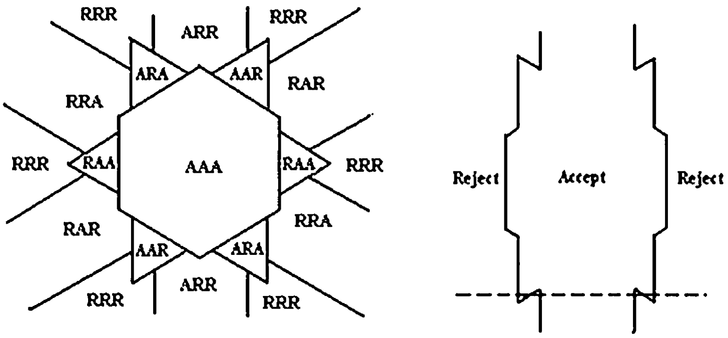

FIG 1. *Regions of action by the Step-down procedure and its induced individual test.*

Figures 1 and 2 offer cross sections of procedures for the case where $I = 3, \sigma^2$ is known and $X_1 + X_2 + X_3$ is fixed. Figure 1 is concerned with a step-down procedure such as given by [7]. In Figure 1 we see regions of actions (A = accept, R = reject) for all three hypothesis testing problems. Note that in all the figures the horizontal lines are the axes representing the variable, say $Y = X_1 - X_2$, which when normalized is the test statistic for testing the null hypothesis $H_{12} : \mu_{12} = 0$ vs. $K_{12} : \mu_{12} \neq 0$. Thus in Figure 1 we note a violation of the interval property for many fixed values of X_3 when $X_1 + X_2 + X_3$ is also fixed. That is, we see that for some values of X_3 the Holm procedure, as a function of Y, has the following pattern as Y goes from $-\infty$ to ∞: reject, accept, reject, accept, reject, accept, reject. This indeed is a disturbing property, which also holds for the procedures offered by [14], [13], and [15].

Figure 2 displays the regions corresponding to Figure 1 for PADD.

In the Appendix we will prove

Theorem 4.1. *For the analysis of variance model of Section 2 with σ^2 known the PADD procedure is admissible.*

Theorem 4.2. *For the analysis of variance model of Section 2 with σ^2 known the PADD+ procedure is admissible.*

We will also prove in Theorem A.4 in the Appendix that under very mild conditions on the critical values, the step-down procedures of [7, 13] and [14] are inadmissible.

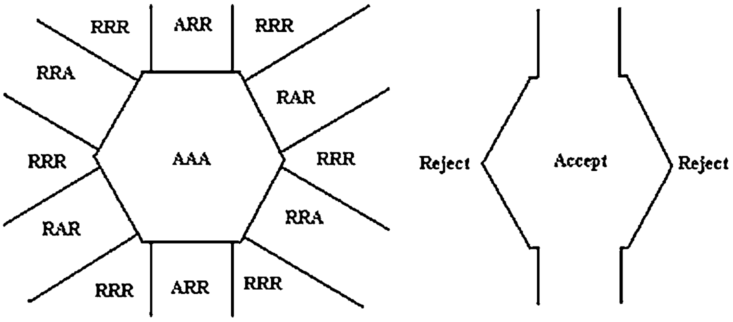

FIG 2. *Regions of action by the PADD procedure and its induced individual test.*

Should σ^2 be unknown we replace σ^2 with s^2 and find that the recommended PADD+ procedure has the following properties:

(i) It is translation and scale invariant.
(ii) For fixed s^2, it has the desirable interval acceptance sections.
(iii) The Tukey step corresponds to the usual Tukey procedure at different levels.
(iv) Simulations used to determine critical values are feasible for modest I.
(v) When $I(J-1)$ is large, the procedure is very close to the procedure described when σ^2 is known.
(vi) Should the practical nature of the problem dictate a procedure that follows logical restrictions, PADD suffices.

Another option is to replace σ^2 by $T = s^2 + J\sum_{i=1}^{I}\bar{X}_i^2$. Whereas this option has interval acceptance sections, is admissible, is scale invariant, it is not translation invariant and critical values should be chosen for each given value of T. This is not feasible for simulation and the determination of critical values.

Appendix: Proofs

In this appendix we prove the two theorems stated in Section 4 and an additional theorem concerned with inadmissibility. The symmetry of the problem in terms of hypotheses and the vector risk function (no hypothesis is treated any differently than any other) enables us to focus on any one particular hypothesis, say H_{12} : $\mu_1 = \mu_2$ vs. $K_{12} : \mu_1 \neq \mu_2$. Without loss of generality we will take $\sigma^2 = 1$.

To start let the $I \times 1$ vector $\underline{U} = \Gamma\underline{X}$ where

$$\Gamma = \begin{pmatrix} 1 \; -1 \; 0 \quad \ldots 0 \\ 0 \; 1 \quad -1 \ldots 0 \\ \vdots \\ 0 \ldots 0 \quad 1 \quad -1 \\ 1 \; 1 \quad \ldots \quad\quad 1 \end{pmatrix}.$$

Then $\underline{U} \sim N(\underline{\nu} = \Gamma\underline{\mu}, \Gamma\Gamma^{'})$. The density of $\underline{U}$ is

$$f_{\underline{U}}(\underline{u}|\underline{\nu}) = (2\pi)^{-I/2}|\Gamma\Gamma^{'}|^{-1/2} \exp -(1/2)(\underline{u}-\underline{\nu})^{'}(\Gamma\Gamma^{'})^{-1}(\underline{u}-\underline{\nu}). \tag{A.1}$$

In exponential family form (A.1) is

$$f_{\underline{U}}(\underline{u}|\underline{\nu}) = h(\underline{u})\beta(\underline{\nu}) \exp \underline{u}'(\Gamma\Gamma^{'})^{-1}\underline{\nu}. \tag{A.2}$$

If we let $\underline{W} = (\Gamma\Gamma^{'})^{-1}\underline{U}$

$$f_{\underline{W}}(\underline{w}|\underline{\nu}) = h^*(\underline{w})\beta(\underline{\nu}) \exp \sum_{i=1}^{I} w_i\nu_i. \tag{A.3}$$

The hypothesis of interest is $H_{12} : \nu_1 = 0$ vs. $K_{12} : \nu_1 \neq 0$. Now we give

Lemma A.1. *A necessary and sufficient condition for a test $\varphi(\underline{w})$ of H_{12} vs. K_{12} to be admissible, is that for almost every fixed $w_2, \ldots, w_I$, the acceptance sections of the test are convex in w_1.*

Proof. See Matthes and Truax [12]. □

Note $\underline{W} = (\Gamma\Gamma')^{-1}\underline{U} = (\Gamma\Gamma')^{-1}\Gamma\underline{X} = (\Gamma')^{-1}\underline{X}$. Hence to study a test function $\phi(\underline{w}) = \phi^*(\underline{x})$ as w_1 varies and $w_2, \dots, w_I$ remain fixed we can consider sample points ($\underline{\mathrm{x}}$+r$\underline{\mathrm{g}}$) where $\underline{g}$ is the first column of Γ' and r varies. This is true since when $\underline{x}$ is replaced by ($\underline{\mathrm{x}}$+r$\underline{\mathrm{g}}$) then $\underline{w}$ becomes

$$(\Gamma^{'})^{-1}(\underline{x} + r\underline{g}) = \underline{w} + (r, 0, \dots, 0)^{'} = (w_1 + r, w_2, \dots, w_I)^{'}.$$

We now let $\Psi_{12}(\underline{x})$ be the test of H_{12} determined by PADD. This test can be described as follows: Suppose that, entering the n^{th} stage the PADD partition is $B_{n,1}(\underline{x}), \dots, B_{n,n}(\underline{x})$. Suppose $\{1\}$ and $\{ 2\}$ were always in the same set of previous partitions but now both $\{ 1\}$ and $\{2\}$ belong to different sets in the partition. Then reject H_{12}. Otherwise accept H_{12} when the PADD procedure stops. Thus the test of H_{12} would stop and reject H_{12} the first time $\{ 1\}$ and $\{2\}$ are split (i.e. they are contained in different sets of the current partition).

To study $\Psi_{12}(\underline{x})$ we will need only focus on $D_{\underline{x}}(A, B)$ for sets B such that either $B \cap \{1, 2\} = \emptyset$ or $\{1, 2\} \subset B$. Otherwise H_{12} would have been rejected earlier. Next we state, for each B, some facts concerning $D_{\underline{x}}(A, B)$ and $D^*_{\underline{x}}(B)$ for points of the form $\underline{x} = (z_1 + r, z_2 - r, z_3, \dots, z_I)'$. These facts are immediate consequences of the definitions.

(F1) Suppose $B \cap \{1, 2\} = \phi$ That is, B contains neither $\{1\}$ nor $\{2\}$. Then $D_{\underline{x}}(A, B)$ and $D^*_{\underline{x}}(B)$ are constants as functions of r for all $A \subset B$.

(F2) Suppose $\{1, 2\} \subseteq A \subset B$. Then $D_{\underline{x}}(A, B)$ is constant as a function of r.

(F3) $x_i \geq x_j$ for every $i \varepsilon V_{\underline{x}}(B), j \varepsilon B \backslash V_{\underline{x}}(B)$ and any B.

(F4) This fact follows from (F3). Suppose $\{1, 2\} \subseteq B$ and consider maximizing $D_{\underline{x}}(A, B)$ for sets $A \subset B$ where A is such that either $\{1\}$ or $\{2\}$, but not both, is in A. The numerator of the maximized $D_{\underline{x}}(A, B)$ when maximized over such sets A will be of the form.

$$\text{(A.4)} \qquad \frac{\sum_{i \in A_1} z_i + \max(z_1 + r, z_2 - r)}{N(A_1) + 1} - \frac{\sum_{i \in A_2} z_i + \min(z_1 + r, z_2 - r)}{N(A_2) + 1},$$

where $A_1, A_2, \{1, 2\}$ is a partition of B. Note (A.4) > 0.

Another useful, but simple fact is

(F5) As functions of r both $\max(z_1{+}r, z_2{-}r)$ and $-\min(z_1{+}r, z_2{-}r)$ are strictly decreasing for $r < (z_2 - z_1)/2$ and strictly increasing for $r > (z_2 - z_1)/2$.

(F6) As a function of r, $|z_1 - z_2 + 2r|$ is strictly decreasing for $r < (z_2 - z_1)/2$ and strictly increasing for $r > (z_2 - z_1)/2$.

Next we establish two properties of the PADD procedure as a test of H_{12} versus K_{12}.

Lemma A.2. *Suppose* $\Psi_{12}(\underline{x})$ *accepts* H_{12} *and* $\Psi_{12}(\underline{x}^*)$ *rejects* H_{12} *where* $\underline{x}^* = \underline{x} + r\underline{g}$ *for some* $r > 0$ *and* $g = (1 - 1, 0, \dots, 0)'$. *Then* $|x_1 - x_2| \leq |x^*_1 - x^*_2|$.

Proof. Suppose that when $\underline{X} = \underline{x}$ is observed the PADD procedure stops at stage n and the final partition is $B_{n,1}(\underline{x}), \dots, B_{n,n}(\underline{x})$. Since ($\underline{x}$) is an accept point, $\{1, 2\} \subseteq B_{m,i}(\underline{x})$ for some $i = 1, \dots, m$ at each stage $m = 1, \dots, n$. That is, $\{1\}$ and $\{2\}$ were not split at any stage. Remember that each partition is the result of a maximization process. Since $\underline{X} = \underline{x}^*$ is a reject point, something must have changed in the maximization process as $\{1\}$ and $\{2\}$ have become split. □

By facts (F1) and (F2), $D_{\underline{x}}(A, B)$ and $D_{\underline{x}^*}(A, B)$ can differ only for sets for which $\{1, 2\} \subseteq B$ and A contains $\{1\}$ or $\{2\}$ but not both. Thus for some such A and B in one of the partitions, $D_{\underline{x}}(A, B)$ has increased and so $D^*_{\underline{x}^*}(B) > D^*_{\underline{x}}(B)$

and the set $V_{\underline{x}^*}(B)$ splits $\{1\}$ and $\{2\}$. By (F4) the measure $D_{\underline{x}}(A, B)$ for such sets can increase only when $\max(z_1 + r, z_2 - r)$ increases. The result now follows from (F5).

A similar proof yields

Lemma A.3. *Suppose $\Psi_{12}(\underline{x})$ rejects H_{12} and $\Psi_{12}(\underline{x}^*)$ accepts H_{12} where $\underline{x}^* = \underline{x} + r\underline{g}$ for some $r > 0$ and $g = (1, -1, 0, \ldots, 0)'$. Then $|x_1 - x_2| \geq |x_1^* - x_2^*|$.*

Next we define

Property M. A test is said to have Property M if there exists three points $x, \underline{x}^* = \underline{x} + r_1\underline{g}, \underline{x}^{**} = \underline{x} + r_2\underline{g}$ with $0 < r_1 < r_2$ such that $\underline{x}$ and $\underline{x}^{**}$ are accept points and $\underline{x}^*$ is a reject point.

Note Property M is necessary and sufficient for a test to be inadmissible by virtue of Lemma A.1.

We now state

Theorem A.1. *The PADD test $\Psi_{12}(\underline{x})$ is admissible as a test of H_{12} versus K_{12}.*

Proof. Note that $\Psi_{12}(\underline{x})$ is admissible if it does not have Property M. Suppose it did have Property M. Then from Lemma A.2 and Lemma A.3 we have $|x_1 - x_2| \leq |x_1^* - x_2^*|$ and $|x_1^* - x_2^*| \geq |x_1^{**} - x_2^{**}|$. Or equivalently, □

$$|x_1 - x_2| \leq |x_1 - x_2 + 2r_1| \geq |x_1 - x_2 + 2r_2|, \tag{A.5}$$

which is a contradiction of (F6).

Define the test function

$$\Psi_{12}^*(x) = \left\{ \begin{array}{ll} 0 & \text{if } |x_1 - x_2| < C_L \\ 1 & \text{if } |x_1 - x_2| > C_U \\ \Psi_{12}(x) & \text{otherwise.} \end{array} \right\} \tag{A.6}$$

This test function corresponds to PADD+.

Theorem A.2. *The PADD+ test $\Psi_{12}^*(\underline{x})$ is admissible as a test of H_{12} versus K_{12}.*

Proof. Suppose $\Psi_{12}^*(\underline{x})$ has Property M, i.e. $\Psi_{12}^*(\underline{x})$ is inadmissible. This can occur in 8 ways. They are listed in the Table 1. □

Table 1

Possible behaviors leading to Property M for Ψ_{12}^*

Sample points	$\boldsymbol{x}$	$\boldsymbol{x}^*$	$\boldsymbol{x}^{**}$	$\boldsymbol{x}$	$\boldsymbol{x}^*$	$\boldsymbol{x}^{**}$
		Case 1			Case 2	
Actions of Ψ_{12}	A	A	A	A	A	R
$\mid x_1 - x_2 \mid, \mid x_1^* - x_2^* \mid, \mid x_1^{**} - x_2^{**} \mid$	$< C_U,$	$> C_U,$	$< C_U$	$< C_U,$	$> C_U,$	$< C_L$
		Case 3			Case 4	
Actions of Ψ_{12}	A	R	A	A	R	R
$\mid x_1 - x_2 \mid, \mid x_1^* - x_2^* \mid, \mid x_1^{**} - x_2^{**} \mid$	$< C_U,$	$> C_L,$	$< C_U$	$< C_U,$	$> C_L,$	$< C_L$
		Case 5			Case 6	
Actions of Ψ_{12}	R	A	A	R	A	R
$\mid x_1 - x_2 \mid, \mid x_1^* - x_2^* \mid, \mid x_1^{**} - x_2^{**} \mid$	$< C_L,$	$> C_U,$	$< C_U$	$< C_L,$	$> C_U,$	$< C_L$
		Case 7			Case 8	
Actions of Ψ_{12}	R	R	A	R	R	R
$\mid x_1 - x_2 \mid, \mid x_1^* - x_2^* \mid, \mid x_1^{**} - x_2^{**} \mid$	$< C_L,$	$> C_L,$	$< C_U$	$< C_L,$	$> C_L,$	$< C_L$

We will see that each scenario would require that $|x_1 - x_2| \leq |x_1^* - x_2^*| \geq |x_1^{**} - x_2^{**}|$ but this would violate (F6). Case 1 is immediate as it requires $|x_1 - x_2| < C_U, |x_1^* - x_2^*| > C_U$ and $|x_1^{**} - x_2^{**}| < C_U$.

Cases 2, 5, 6, and 8 are similar to Case 1. Case 3 is impossible as Ψ_{12} is admissible by Theorem 4.1. Next we will consider Case 4 and Case 7.

In Case 4 we have $|x_1^* - x_2^*| > C_L > |x_1^{**} - x_2^{**}|$. Furthermore since $\underset{\sim}{x}$ is an accept point of Ψ_{12} and $\underset{\sim}{x}^*$ is a reject point of Ψ_{12} it follows from Lemma A.2 that $|x_1 - x_2| \leq |x_1^* - x_2^*|$. This gives the contradiction.

In Case 7 we have $|x_1 - x_2| < C_L < |x_1^* - x_2^*|$. Furthermore since $\underset{\sim}{x}^*$ is a reject point of Ψ_{12} and $\underset{\sim}{x}^{**}$ is an accept point of Ψ_{12} it follows from Lemma A.3 that $|x_1^* - x_2^*| \geq |x_1^{**} - x_2^{**}|$. This gives the contradiction.

The critical values for the Holm [7] procedure are all different. Some critical values for the Royen [13] and Shaffer [14] procedures can be the same. However if $I \geq 4$ there are at least 3 critical values that are different. In particular it is always true that $0 < C_1 < C_2 < C_3 < C_q$.

Theorem A.3. *For $I \geq 3$ the Holm procedure is inadmissible. For $I \geq 4$ and $(3C_2 + C_1)/2 > C_3$, the Shaffer and Royen procedures are inadmissible.*

Proof. We prove the theorem for Shaffer's procedure since the others will follow similarly. We exhibit three sample points $x, \underset{\sim}{x}^* = \underset{\sim}{x} + r_1\underset{\sim}{g}, \underset{\sim}{x}^{**} = \underset{\sim}{x} + r_2\underset{\sim}{g}$ with $0 < r_1 < r_2$ such that $\underset{\sim}{x}$ and $\underset{\sim}{x}^{**}$ are accept points for H_{12} while $\underset{\sim}{x}^*$ is a reject point. That is, we show that the individual test of H_{12} determined by Shaffer's method has Property M. Now choose the $I \times 1$ vector $\underset{\sim}{x}$ so that $x_1 = (3C_2 + C_1)/2, x_2 = C_2, x_3 = 0, x_4 = 2C_q + (3C_2 + C_1)/2$, and $x_i = 2x_{i-1}, i = 5, \ldots, I$. For this $\underset{\sim}{x}$ one can verify that all hypotheses are rejected except H_{12} and H_{23} . Now choose $r_1 = \varepsilon > 0$ and ε small, $\varepsilon < (C_2 - C_1)/4$. Recall $\underset{\sim}{g} = (1, -1, 0, \ldots, 0)$ so that $x_1^* = x_1 - \varepsilon$ and $x_1^* - x_2^* = (C_2 + C_1)/2 - 2\varepsilon > C_1$. Also $x_2^* = x_2 + \varepsilon$ so that $x_2^* - x_3^* = C_2 + \varepsilon$. Furthermore $x_i = x_i^*$ for $i \geq 3$. In light of this and the choices of $\underset{\sim}{x}$ and $\underset{\sim}{x}^*$ all hypotheses are rejected. Next choose $r_2 = (C_2 + C_1)/4$ so that $x_1^{**} - x_2^{**} = 0$. This ensures that the procedure accepts H_{12} at $\underset{\sim}{x}^{**}$. □

Remark A.5. The condition regarding the critical value would easily hold in virtually all practical situations.

Remark A.6. If σ^2 is unknown and s^2 replaces σ^2 in the Holm, Shaffer and Royen procedures they are still inadmissible under mild conditions. The proof requires only some modifications of the proof of Theorem A.4.

References

[1] Benjamini, Y. and Gavrilov, Y. (2009). A simple forward selection procedure based on false discovery rate control. *Ann. Appl. Statist.* **3** 179–198.

[2] Cohen, A. and Sackrowitz, H. B. (2008). Multiple testing to two-sided alternatives with dependent data. *Statist. Sinica* **18** 1593–1602.

[3] Cohen, A., Sackrowitz, H. B. and Xu, M. (2009). A new multiple testing method in the dependent case. *Ann. Statist.* **37** 1518–1544.

[4] Dudoit, S. and Van Der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics.* Springer.

[5] Devore, J. L. (2008). *Probability and Statistics for Engineering and the Sciences: Enhanced*, 7th ed. Duxbury Pr.

[6] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. Roy. Statist. Soc. Ser. B* **64** 499–518.

[7] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6** 65–70.

[8] Ishwaran, H. and Rao, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *J. Amer. Statist. Assoc.* **98** 438–455.

[9] Lehmacher, W., Wassmer, G. and Reitmeir, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* **47** 511–521.

[10] Lehmann, E. L. (1957). A theory of some multiple decision problems, I. *Ann. Math. Statist.* **28** 1–25.

[11] Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*, 2nd ed. Springer.

[12] Matthes, T. K. and Truax, D. R. (1967). Tests of composite hypotheses for the multivariate exponential family. *Ann. Math. Statist.* **38** 681–697.

[13] Royen, T. (1989). Generalized maximum range tests for pairwise comparisons of several populations. *Biometrical Journal* **31** 905–929.

[14] Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *J. Amer. Statist. Assoc.* **81** 826–831.

[15] Westfall, P. H. and Tobias, R. D. (2007). Multiple testing of general contrasts: Truncated closure and the extended Shaffer-Royen method. *J. Amer. Statist. Assoc.* **102** 487–494.