IMS Collections
Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman
Vol. 8 (2012) 57–63

DOI: 10.1214/11-IMSCOLL804

Multiple rank tests for pairwise comparisons*

Arthur Cohen and Harold Sackrowitz

Rutgers, The State University of New Jersey

Abstract: Consider a balanced one way layout without a normality assumption. That is, assume each population has an unknown translation parameter and has a continuous distribution. We wish to test the C_2^k pairwise differences in translation parameters assuming there are k populations.We propose a multiple testing method based on ranks that is analogous to the method developed for testing pairwise contrasts among means by Cohen, Sackrowitz and Chen (2010) [3] (CSC). This latter method has an intuitive and important practical property not shared by most multiple testing methods. Namely, for each individual pairwise hypothesis relevant acceptance sections are intervals. Furthermore, the CSC method is shown to do well in terms of power compared to competitive methods. In the nonparametric setting the analogous procedure has the desirable interval property and also stacks up well in a comparative simulation study.

Contents

1. Introduction

Nonparametric multiple testing is discussed in [4]. In [2], Campbell and Skillings study multiple rank tests for pairwise comparisons in a balanced one way layout without a normality assumption. They consider single step and stepwise procedures, noting that the latter have better power. Among the stepwise procedures are those that rerank at different steps as well as those that do not rerank. One is also offered a choice of joint ranking or separate ranking by pairs. A simulation study of power is partial to an ad hoc procedure.

In this paper we propose a nonparametric rank test analoque of a method developed in [3] for testing pairwise comparisons in a one way layout assuming normality. The multiple testing method is called PADD+. In a normal model PADD+ yields admissible tests for individual hypotheses and has an important monotonicity property for individual tests. Namely, for certain fixed variables, the acceptance sections for testing an individual hypothesis are intervals (and not a collection of disjoint

*Research supported by NSF grant 0894547 and NSA grant H-98230-10-1-0211
AMS 2000 subject classifications: Primary 62H15, 62G10
Keywords and phrases: ad hoc procedure, interval property, joint ranks, separate ranks, screening step, stepwise procedure

sets). Furthermore simulations indicate that the method leads to tests that are more powerful than usual conventional step-up or step-down methods. The latter are also shown to be inadmissible.

For the nonparametric model we show that the analogue of PADD+ has a desirable and intuitive monotonicity property that the ad hoc procedure does not have. Namely, convex acceptance sections. In terms of power, studied by simulation, this analoque of PADD+ is comparable to the performance of the ad hoc procedure.

In the next section we formally state the model. We then describe the ad hoc procedure and the rank-based version of PADD+, called RPADD+. An example illustrates the two procedures. In Section 3 we show that the ad hoc procedure does not have an important monotonicity property while RPADD+ has the property. Section 4 offers a simulation study comparing the power of the two different procedures.

2. Model and procedures

We describe the model in [2]. Let

$$Y_{ij} = \theta_i + \varepsilon_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, n,$$

where the θ_i's are unknown and the ε_{ij}'s are independent, indentically distributed, continuous random variables. Hypotheses of interest are

$$H_{ij} : \theta_i = \theta_j \quad vs. \quad K_{ij} : \theta_i \neq \theta_j$$

for every $i \neq j$ and $i, j = 1, \ldots, k$. Two different sets of rank statistics are defined as follows:

Joint Rank Range: For any p treatments ($p \leq k$), relabeled from 1 to p, observations are jointly ranked and the i^{th} treatment sum of ranks, R_i is calculated for $i = 1, \ldots, p$.

Separate Rank: Separately rank all the observations from only the i^{th} and j^{th} treatments. Let R_{ij} denote the sum of the ranks for treatment i in this separate ranking with treatment j.

In [2] Campbell and Skillings recommend the following ad hoc procedure labeled NAH.

Step 0: Order the treatments from smallest to largest according to the rank sums in the joint ranking. That is, let $R_1, \ldots, R_k$ be the rank sums and let $R_{(1)} \geq R_{(2)} \geq \ldots \geq R_{(k)}$ be the ordered rank sums.

Step 1: Conclude that treatments (1) and (k) differ if $R_{(1)} - R_{(k)}$ exceeds $r_{a_k,k}$, the upper a_k cutoff of the rank range for k treatments. Then continue to Step 2. If treatments (1) and (k) are not declared different then stop and report no differences.

Step 2: Declare treatments (1) and $(k-1)$ different if in the joint rankings of treatments $(1), \ldots, (k-1)$ the difference in the rank sums of treatment (1) and treatment $(k-1)$ exceeds $r_{a_{k-1},k-1}$. Also declare treatments (2) and (k) different if in the joint rankings of treatments (2),...,(k) the difference in the rank sums of treatment (2) and treatment (k) exceeds $r_{a_{k-1},k-1}$. Then continue to Step 3. If neither pair differ, stop.

Step $(k-p+1)$: For $p = k-2, \ldots, 2$ continue to test treatment subsets of the form $(i), (i+1), \ldots, (i+p-1)$. For each subset S the observations are reranked

and the difference in rank sums for treatments $i+p-1$ and i is compared to $r_{a_p,p}$. The procedure terminates by early stopping or with reported differences after step $(k-1)$.

Typically critical values are chosen so that Type I error rates are estimated to be less than or equal to α.

Next we describe the analogue of PADD+ called RPADD+. We begin by mimicking the definitions in [3]. Let $S = \{1, \ldots, k\}$. For any subset of integers $A \subset S$ let N(A) = the number of points in A. Let $R_A = \sum_{i \in A} R_i/N(A)$. Next define, for each sample point $\underline{y} = (y_{11}, \ldots, y_{1n}, \ldots, y_{k1}, \ldots, y_{kn})$ and for all $A \subset B \subseteq S$ with $A \neq \phi \neq B \setminus A$,

$$D_{\underline{y}}(A; B) = (R_A - R_{B \setminus A})/\sigma_{A,B} \tag{2.1}$$

where

$$\sigma^2_{A,B} = \omega(1/N(A) + 1/N(B \setminus A))/12 \tag{2.2}$$

and

$$\omega = n(kn)(kn+1).$$

Note (2.2) is an approximation derived using [4], p. 245. Let

$$D^*_{\underline{y}}(B) = \max_{A \subset B} D_{\underline{y}}(A; B). \tag{2.3}$$

Further let $V_{\underline{y}}(B)$ denote the A set for which the maximum is attained. At the first stage and first step of RPADD+ all non-empty 2 set partitions of S are considered. Since the sum of the ranks $R_{(i)}$ are ordered it is not necessary to look at every 2 set partition to find the maximum. $D_{\underline{y}}(A; S)$ is computed for all non-empty $A \subset S$. Letting $C_k(S)$ denote a constant at stage 1, step 1, and letting $D_1 = D^*_{\underline{y}}(S)$, if $D_1 \leq C_k(S)$, stop and accept all null hypotheses. If $D_1 > C_k(S)$, then partition S into $V_{\underline{y}}(S)$ and $S \setminus V_{\underline{y}(S)}$ and continue to step 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 m, S has been partitioned into $B_1, \ldots, B_{m+1}$ and we continue. Let $C\{B_1, \ldots, B_{m+1}\}$ be a constant determined by the partition $\{B_1, \ldots, B_{m+1}\}$. Compute $D_{m+1} = \max_{1 \leq j \leq m+1} D^*_{\underline{y}}(B_j)$. Next break B_j into $V_{\underline{y}}(B_j)$ and $B_j \setminus V_{\underline{y}}(B_j)$. Continue to stage $m+1$.

Thus we see that as we enter stage m the partition consists of m sets. Denote these by $B_{m,1}(\underline{x}), \ldots, B_{m,m}(\underline{x})$. If $D_m \leq C\{B_{m,1}(\underline{x}), \ldots, B_{m,m}(\underline{x})\}$, stop and then $\{B_{m,1}(\underline{x}), \ldots, B_{m,m}(\underline{x})\}$ is the final partition. If $D_m > C\{B_{m,1}(\underline{x}), \ldots, B_{m,m}(\underline{x})\}$ we continue and the partition will become finer. If $\{B_{m,1}(\underline{x}), \ldots, B_{m,m}(\underline{x})\}$ 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.

There is considerable flexibility in the choice of critical values $C\{B_{m,1}, \ldots, B_{m,m}\}$. One way to choose them is to simply allow them to depend on the stage m. 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.

Note that for an arbitrary set of indices, say $B = \{i_1, \ldots, i_q\}$, with $R_{(i_1)} \geq R_{(i_2)} \geq \cdots \geq R_{(i_q)}$ the relevant (q-1) statistics from (2.1) are

$$T_{iq}(\underline{R}) = \{\sum_{j=1}^{i} R_{(i_j)}/i - \sum_{j=i+1}^{q} R_{(i_j)}/(q-i)\}/\omega\{1/i + 1/(q-i)\}, \tag{2.4}$$

for $i = 1, 2, \ldots, q-1$. RPADD+ involves a final screening stage. Two constants $C_L \leq C_U$ are specified. Then $H_{ii'}$ will be rejected if and only if the indices i and i' lie in different sets of the final partition, $B_{m,1}(\underline{x}), \ldots, B_{m,m}(\underline{x})$ and $|R_i - R_{i'}| > C_L$ or i and i' lie in the same set of the final partition and $|R_i - R_{i'}| > C_U$.

We conclude this section with an example illustrating NAH and RPADD+.

Example 2.1. The data is taken from [5], page 525. Four brands of golf balls were tested. Distances of drives using Iron Byron, the USGA's robotic golfer, were observed for 10 balls of each brand. The rank sums were as follows: Brands C, B, A and D yielded 352, 250, 118 and 100 respectively. With $k = 4$, $n = 10$, NAH first considers $R_{(1)} - R_{(4)}$ and rejects if this difference exceeds $r_{a_4,4}$. From (2.2) we have that $\sqrt{\omega} = 36.97$. Choosing $a_4 = .05$ and using tables of the studentized range with degrees of freedom set equal to ∞ we find the studentized range critical value is 3.63. Since $352 - 100 > (3.63)(36.97)$ we reject $H_{(1)(4)}$. Next we consider $H_{(1)(3)}$ and $H_{(2)(4)}$. For $H_{(1)(3)}$ and $H_{(2)(4)}$ the critical value is 3.31 and since 352 - 118 and 250 - 100 both exceed (3.31)(36.97) we also reject $H_{(1)(3)}$ and $H_{(2)(4)}$. Finally consider $H_{(1)(2)}, H_{(2)(3)}$ and $H_{(3)(4)}$ with critical value 2.77. Only $H_{(2)(3)}$ is rejected. Thus NAH leads to rejecting $H_{(1)(4)}, H_{(2)(4)}, H_{(1)(3)}$ and $H_{(2)(3)}$.

Next we apply RPADD+ to the same data set. We choose critical values $C_1 = 2.56, C_2 = 2.19, C_3 = 1.84, C_L = 2.47$ based on the simulation study of Section 4. At step 1 we consider the maximum of three statistics in (2.4), namely (352 - 156)/42.69 = 4.59, (240 - 100)/42.69 = 3.28 and (301 - 109)/36.97 = 5.19. This leads to rejection of $H_{(1)(4)}, H_{(2)(4)}, H_{(1)(3)}$ and $H_{(2)(3)}$. At step 2 we examine $H_{(1)(2)}$ and $H_{(3)(4)}$. Here we calculate $(352-250)/(36.97\sqrt{2}) = 1.95 < 2.19$ and $(118-100)/(36.97\sqrt{2}) = .34 < 2.19$. Hence $H_{(1)(2)}$ and $H_{(3)(4)}$ are accepted. Finally at the screening stage we examine $(352 - 118)/(36.97\sqrt{2}) = 4.48 > 2.47$ as is $(352 - 100)/(36.97\sqrt{2}) = 4.82$, $(250-118)/(36.97\sqrt{2}) = 2.525$ and $(250-100)/(36.97\sqrt{2}) = 2.87$. Hence the screen stage does not switch any reject to an accept.

3. Interval property

A desirable property for a testing procedure to have is the interval property for each individual test. Without loss of generality we focus on $H_{12} : \theta_1 - \theta_2 = 0$ versus $K_{12} : \theta_1 - \theta_2 \neq 0$.

Definition 3.1. Let $\underline{R} = (R_1, \ldots, R_k)'$ be the vector of rank sums at stage 1, step 1 of the process. Let $\underline{R}^* = (R_1 + \Delta_1, R_2 - \Delta_1, R_3, \ldots, R_k)'$ for $\Delta_1 > 0$ and let $\underline{R}^{**} = (R_1 + \Delta_2, R_2 - \Delta_2, R_3, \ldots, R_k)'$ for $\Delta_2 > \Delta_1$. Assume that the individual test ϕ_{12} for H_{12} accepts at $\underline{R}$ and rejects at $\underline{R}^*$. Then ϕ_{12} has the interval property if and only if ϕ_{12} rejects at $\underline{R}^{**}$.

We now give an example demonstrating that NAH does not have the interval property.

Example 3.1. Let $k = 3$, $n = 6$, $r_{a_3,3} = 17$ and $r_{a_2,2} = 12$. That is, each sample point, $\underline{y}$ is an 18×1 vector consisting of 6 observations from each of 3 populations.

TABLE 1
NAH rankings at steps 1 and 2 for the first sample point.

	step 1			**step 2**	
population	**1**	**2**	**3**	**1**	**2**
	4	3	1	2	1
	5	6	2	3	4
individual	11	10	7	6	5
ranks	12	13	8	7	8
	15	14	9	10	9
	16	17	18	11	12
Total	63	63	45	39	39

TABLE 2
NAH rankings at steps 1 and 2 for the second sample point.

	step 1			**step 2**	
population	**1**	**2**	**3**	**1**	**2**
	4	3	1	2	1
	5	6	2	3	4
individual	12	10	7	7	5
ranks	15	11	8	10	6
	16	13	9	11	8
	17	14	18	12	9
Total	69	57	45	45	33

TABLE 3
NAH rankings at steps 1 and 2 for the third sample point.

	step 1			**step 2**	
population	**1**	**2**	**3**	**1**	**2**
	4	1	2	3	1
	6	3	5	4	2
individual	12	10	7	7	5
ranks	15	11	8	9	6
	16	13	9	10	8
	17	18	14	11	12
Total	70	56	45	44	34

We will exhibit, for each of three sample points, the 18 individual ranks used at step 1 as well as the 12 (reranked) ranks that NAH would use at step 2. The three sets of ranks appear in Tables 1–3.

Thus, when the first sample point $\underline{R}$ is observed, NAH will reject H_{13} at step 1 but H_{12} will be accepted based on step 2. At the second sample point $\underline{R}^*$, NAH will reject H_{13} at step 1 and will reject H_{12} at step 2. At the third sample point $\underline{R}^{**}$, NAH will reject H_{13} at step 1 but will accept H_{12} at step 2 since $44-34=10<12$.

The next lemma shows that RPADD, without the screen stage, has the interval property.

Lemma 3.1. *The individual test $\phi_{12}(\underline{R})$ induced by RPADD has the interval property.*

Proof. Define the $k\times 1$ vector $\underline{g}=(1,-1,0,\ldots,0)$. Note from (2.4) that $T_{1q}(\underline{R}+\Delta\underline{g}),\Delta>0$, is an increasing function of Δ while $T_{iq}(\underline{R}+\Delta\underline{g})=T_{iq}(\underline{R})$ for all Δ and $i=2,\ldots,q-1$. Assume $\Psi_{12}(\underline{R})$ accepts H_{12} but, for some $\Delta_1>0$, $\Psi_{12}(\underline{R}^*)=\Psi_{12}(\underline{R}+\Delta_1\underline{g})$ rejects H_{12}. This means that at $\underline{R}$ the procedure stopped before there was any value of q for which T_{1q} was the maximum statistic that exceeded the relevant critical value. Furthermore, there exists a q^* such that T_{1q^*} was the maximum statistic that did exceed the appropriate critical value. Since T_{1q} is an increasing function of Δ, it follows that $T_{1q}(\underline{R}^{**})=T_{1q}(\underline{R}+\Delta_2\underline{g})$ exceeds its

appropriate critical value for some $q \leq q^*$ which implies $\Psi_{12}(\underline{R}^{**})$ also rejects. □

We conclude this section with

Theorem 3.1. *RPADD+ has the interval property for testing* H_{12}.

Proof. Note that $(R_1 - R_2) < (R_1^* - R_2^*) < (R_1^{**} - R_2^{**})$. We can now follow the same steps as in the proof of Theorem 4.2 of [3] with one exception. Here we need to use Lemma 3.1 above for Case 3 of that reference to complete the proof. □

4. Simulation results

In this section we present the results of a simulation study comparing RPADD+ with the study for NAH presented in [2]. In that paper they reported the results of a large simulation study for a variety of k, n values and using the exponential, uniform, normal and double exponential distributions. They presented the results from the uniform as they were said to comparable for all distributions. We did simulations using the RPADD+ procedure for the uniform distribution using the choices of k, n and the parameter points of [2]. We report, in Tables 4 and 5, results only for $k = 4$, $n = 10$ and $k = 6$, $n = 10$ as they are representative of the overall power behavior.

The RPADD+ critical values were fine tuned using simulation to control Type I errors at rates comparable to those in [2]. Our initial PADD stage critical values were suggested by [1] for another model but work well here. For $k = 4$ these critical values were multiplied by 1.025 while for $k = 6$ the multiplier was 1.108. The screening cutoffs were again determined through simulation. Upper screening was not used.

For $k = 4$, $n = 10$ the constants were $C_1 = 2.56, C_2 = 2.19, C_3 = 1.84, C_L = 2.47$. For $k = 6$, $n = 10$ they were $C_1 = 2.94, C_2 = 2.58, C_3 = 2.31, C_4 = 2.05, C_5 = 1.73, C_L = 3.24$. All simulations are based on 5000 iterations.

TABLE 4
Size and power comparison of NAH and RPADD+ for k = 4, n = 10.

			Treatment Pair	size or power	
				Procedure	
k	n	means	(ij, j)	NAH	RPADD+
4	10	(0, 0, 0, 0)	FWER	.044	.049
			(1,2)	.011	.022
		(0, 1, 1, 2)	FWER	.020	.053
			(1,2)	.346	.438
			(1,4)	.928	.908
		(0, 2/3, 4/3, 2)	(1,2)	.160	.190
			(1,3)	.610	.717
			(1,4)	.930	.941
		(0, 0, 0, 2)	FWER	.046	.046
			(1,2)	.021	.021
			(1,4)	.941	.971
		(0, 0, 2, 2)	FWER	.051	.006
			(1,2)	.028	.003
			(1,4)	.944	.984

TABLE 5
Size and power comparison of NAH and RPADD+ for k = 6, n = 10.

			Treatment Pair	size or power	
				Procedure	
k	n	means	(i, j)	NAH	RPADD+
6	10	(0, 0, 0, 0, 0, 0)	FWER	.037	.049
			(1,2)	.003	.008
		(0, .4, .8, 1.2, 1.6, 0)	(1,2)	.031	.074
			(1,3)	.124	.125
			(1,4)	.341	.412
			(1,5)	.642	.750
			(1,6)	.885	.932
		(0, 0, 0, 0, 0, 2)	FWER	.048	.052
			(1,2)	.007	.007
			(1,4)	.884	.899
		(0, 0, 1, 1, 2, 2)	FWER	.034	.013
			(1,2)	.013	.002
			(1,3)	.204	.216
			(1,6)	.877	.937
		(0, 0, 0, 2, 2, 2)	FWER	.032	.006
			(1,2)	.008	.001
			(1,6)	.891	.912

FWER in the table is the strong familywise error rate = the probability of at least one Type I error. This depends on the entire parameter point as well as the number of null hypotheses that are true. In [2] it is labeled EERI.

Since both NAH and RPADD+ are translation invariant we take the populations to be uniform with width one and mean θ_i. That is, population i is uniform on $(\theta_i - 0.5, \theta_i + 0.5)$ In the tables the means are in expressed in standard units so each must be multiplied by .289.

References

[1] BENJAMINI, Y. and GAVRILOV, Y. (2009). A simple forward selection procedure based on false discovery rate control. *Annals of Applied Statistics*, **3**, 179–198.

[2] CAMPBELL, G. and SKILLINGS, J. H. (1985). Nonparametric stepwise multiple comparison procedures. *Journal of the American Statistical Association*, **80**, 998–1003.

[3] COHEN, A., SACKROWITZ, H. and CHEN, C. (2010). Multiple testing of pairwise comparisons. Submitted.

[4] HOCHBERG, Y. and TAMHANE, A. C. (1987). *Multiple Comparison Procedures*, Wiley, New York.

[5] MCCLAVE, J. T. and SINCICH, T. (2006). *Statistics, tenth edition*, Pearson Prentice Hall, New Jersey.