Graphical Procedures for Multiple Comparisons Under General Dependence

Christopher J. Bennett\textsuperscript{a} and Brennan S. Thompson\textsuperscript{b}

March 7, 2014

Abstract

It has been more than half a century since Tukey first introduced graphical displays that relate non-overlap of confidence intervals to statistically significant differences between parameter estimates. In this paper, we show how Tukey’s graphical overlap procedure can be modified to accommodate general forms of dependence within and across samples. We also develop a procedure that can be used to more effectively resolve rankings within the tails of the distributions of parameter values, thereby generalizing existing methods for “multiple comparisons with the best.” We show that these new procedures retain the simplicity of Tukey’s original procedure while maintaining asymptotic control of the familywise error rate under very general conditions. Simple examples are used throughout to illustrate the procedures.

Keywords: bootstrap; familywise error rate; overlap

\textsuperscript{†} The authors thank Professor Liu, an associate editor, and two anonymous referees for comments and suggestions that have helped to improve the manuscript.

\textsuperscript{a}(Corresponding author) Bates White, LLC, 1300 Eye Street N.W., Washington D.C. 20002, U.S.A. E-mail: chris.bennett@bateswhite.com

\textsuperscript{b} Department of Economics, Ryerson University, 350 Victoria Street, Toronto, Ontario M5B 2K3, CANADA E-mail: brennan@ryerson.ca
1 Introduction

The focus of this paper is on graphical procedures that may be used to rank a finite collection of (unknown) population parameters. The principle difficulty in developing formal statistical procedures for this problem—graphical or otherwise—is that it requires one to make decisions concerning each and every pairwise comparison.

Tukey (1953) showed how to construct a set of simultaneous confidence intervals for pairwise differences in means. Moreover, Tukey demonstrated that tests based on inverting these simultaneous confidence intervals maintain control over the probability of finding one or more spurious differences (i.e., the familywise error rate) when the samples are balanced and independently drawn from normal populations with a common variance. Tukey’s original procedure has since been extended through the use of resampling-based methods and iterative stepwise refinements so as to improve power and allow for more general sampling schemes; see, e.g., Westfall and Young (1993) and Romano and Wolf (2005a,b).

Complementing this test procedure, Tukey (1953) also developed a graphical procedure that involves constructing uncertainty intervals around each of the parameter estimates so that statistically significant differences can be detected on the basis of whether or not the relevant uncertainty intervals overlap. Tukey’s graphical representation is considered particularly appealing because it offers users more than a 0-1 (“Yes-No”) decision regarding differences between parameters. Indeed, the graphical representation allows users to determine both statistical and practical significance of pairwise differences, while also providing them with a measure of uncertainty concerning the locations of the individual parameters.

In this paper, we use resampling-based methods, together with a slight twist on the construction of the uncertainty intervals, to extend Tukey’s (1953) graphi-
cal procedure beyond the case of independent balanced samples drawn from normal populations with a common variance. The key step in our construction lies in parameterizing the lengths of the uncertainty intervals in such a way that the probability of non-overlap can be directly estimated. The resulting formulation (i) enables us to maintain the simplicity of Tukey’s original procedure and control the familywise error rate, at least asymptotically, without imposing any stringent distributional assumptions; and (ii) lends itself to the construction of a $p$-value scale, which enables users of the graphical display to also visually determine (at least approximately) the strength of evidence against equality of a given pair of parameters.

Beyond generalizing Tukey’s overlap procedure to accommodate general forms of dependence, we also investigate more powerful alternatives to his basic procedure for use in situations where the ordering of two (or more) parameters is likely to remain unresolved. The first alternative is an iterative stepwise method which, analogously to Holm’s (1979) improvement over the Bonferroni correction, is capable of resolving at least as many pairwise comparisons as the basic procedure without sacrificing (asymptotic) control over the familywise error rate. The second, and arguably more novel, alternative is designed specifically for situations in which an investigator is willing to abandon pursuit of a global ranking in return for the ability to resolve more comparisons in the tail of the distribution of parameter values. We show that this “tail ranking” procedure, which builds on Hsu’s (1981,1984) procedures for “multiple comparisons with the best”, gains a sizable power advantage over the basic and stepwise ranking procedures in the (user-defined) tail region of interest, albeit at the cost of being silent on the ranking of parameters outside the tail.

The outline for the paper is as follows. First, we begin in Section 2 with a formal description of the inference problem. Next, in Section 3, we outline the resampling scheme that is used to construct the uncertainty intervals, and discuss the corre-
sponding decision rule for the procedure. The large sample properties of the overlap procedure are also examined here. Global refinements of the overlap procedure are then taken up in Section 4, while our tail ranking procedure is developed in Section 5. Simulation results are presented in Section 6, and an empirical illustration, in which we attempt to rank a large collection of hedge funds by the mean of their risk-adjusted returns, is presented in Section 7. Section 8 concludes.

\section{Problem Formulation}

We seek to rank a collection of \( k \geq 2 \) parameters. The parameters of interest in this ranking, \( \theta_i = \theta_i(P), i \in K = \{1, \ldots, k\} \), are unknown but are presumed to be consistently estimable from observable data generated from some (unknown) probability mechanism \( P \). That is, for each \( i \in K \), we have available to us a \( \sqrt{n} \)-consistent estimator \( \hat{\theta}_{n,i} \) of \( \theta_i \), where \( n \) denotes the sample size (here we focus on the case of balanced samples; the treatment of unbalanced samples, which requires only minor adjustments, is discussed and illustrated in Appendix C).

In our discussion of this problem, it is helpful to partition the collection of parameter pairs into two sets, namely the set of parameter pairs for which equality holds in the population, and the set of parameter pairs which are strictly ordered in the population. Accordingly, we define the sets

\[ D_0(P) = \{(i, j) \in K \times K : i < j, \theta_i(P) = \theta_j(P)\}, \quad (2.1) \]

and

\[ D_1(P) = \{(i, j) \in K \times K : \theta_i(P) < \theta_j(P)\}. \quad (2.2) \]

The set \( D_1(P) \) may be empty, in which case all of the parameters are equal and no
parameter pairs are strictly ordered; otherwise, \( D_1(P) \) is non-empty and some or all of the parameter pairs are strictly ordered. Because it is the elements of \( D_1(P) \) that are of primary interest, the task before us essentially amounts to inferring from the available data which ordered pairs \((i, j) \in K \times K\) belong to \( D_1(P) \). In doing so, we seek to guarantee control over the familywise error rate, at least asymptotically.

Formally, let \( \delta_n(i, j) = 1 \) if we decide that \((i, j) \in D_1(P)\), and \( \delta_n(i, j) = 0 \) otherwise. A Type I error occurs whenever \((i, j) \in D_0(P)\) and \( \delta_n(i, j) = 1 \) (i.e., when \( \theta_i = \theta_j \) and we incorrectly decide that \( \theta_i < \theta_j \)). Consequently, asymptotic control of the familywise error rate at the nominal level \( \alpha \) will require that the decision rule \( \delta_n \) satisfy

\[
\text{Prob} \left[ \sum_{(i,j) \in D_0(P)} \delta_n(i, j) \geq 1 \right] \leq \alpha, \tag{2.3}
\]

in the limit as \( n \to \infty \).

In the case of comparing \( k \geq 2 \) population means on the basis of independent sample averages \( \hat{\theta}_{n,i}, \ i \in K \), with \( \hat{\theta}_{n,i} \sim N(\theta_i, \sigma^2/n) \), Tukey (1953) proposed a graphical procedure that involves plotting each parameter estimate \( \hat{\theta}_{n,i} \) together with its corresponding uncertainty interval

\[
C^\text{Tukey}_{n,i}(\alpha) = \left[ \hat{\theta}_{n,i} \pm Q_{k,\nu}^{\alpha} \frac{S_\nu}{2\sqrt{n}} \right], \tag{2.4}
\]

where \( S_\nu \) is the pooled estimator of \( \sigma \) based on \( \nu = k(n-1) \) degrees of freedom (d.f.), and \( Q_{k,\nu}^\alpha \) denotes the \((1 - \alpha)100^{th}\) percentile of the Studentized range distribution with parameter \( k \) and d.f. \( \nu \) (see, e.g., Hochberg and Tamhane, 1987, p. 31). In this
context, Tukey showed that

$$
\delta_{n}^{\text{Tukey}}(i, j) = \begin{cases} 
1, & \text{if } \hat{\theta}_{n,i} < \hat{\theta}_{n,j} \text{ and } C_{n,i}^{\text{Tukey}}(\alpha) \cap C_{n,j}^{\text{Tukey}}(\alpha) = \emptyset \\
0, & \text{otherwise,}
\end{cases}
$$

(2.5)
satisfies (2.3) for all $n$. In other words, Tukey’s graphical decision rule that ranks $\theta_i$ below $\theta_j$ whenever $\hat{\theta}_{n,i}$ is less than $\hat{\theta}_{n,j}$ and the uncertainty intervals $C_{n,i}^{\text{Tukey}}(\alpha)$ and $C_{n,j}^{\text{Tukey}}(\alpha)$ do not overlap achieves control of the familywise rate at the nominal level $\alpha$, even in finite samples.

Tukey’s (1953) graphical overlap procedure is particularly attractive because significant differences and their directions can be detected visually. Unfortunately, the procedure has proven difficult to extend satisfactorily beyond the case of independent balanced samples drawn from normal populations with a common variance; see, e.g., Gabriel (1978), Andrews et al. (1980), Hochberg et al. (1982), and Hsu and Peruggia (1994). In the next section, we revisit Tukey’s (1953) graphical overlap procedure in an attempt to extend it in a natural way to allow for both heterogeneous samples and arbitrary dependence across samples.

## 3 The Overlap Procedure

In an attempt to remain faithful to Tukey’s (1953) original construction, we consider a graphical procedure based on presenting each parameter estimate $\hat{\theta}_{n,i}$, together with its corresponding uncertainty interval,

$$
C_{n,i}(\gamma) = \left[ \hat{\theta}_{n,i} - \frac{\hat{\sigma}_{n,i}}{\sqrt{n}} \gamma, \hat{\theta}_{n,i} + \frac{\hat{\sigma}_{n,i}}{\sqrt{n}} \gamma \right],
$$

(3.1)

$$
\equiv [L_{n,i}(\gamma), U_{n,i}(\gamma)]
$$

5
whose length is determined by the parameter $\gamma$, together with the sample size $n$ and
the standard error $\hat{\sigma}_{n,i}$ of the estimator $\sqrt{n}\hat{\theta}_{n,i}$. Like Tukey, we infer that $\theta_i < \theta_j$ if
$\hat{\theta}_{n,i} < \hat{\theta}_{n,j}$ and the uncertainty intervals for $\theta_i$ and $\theta_j$ are non-overlapping. That is,
the overlap procedure records
\[
\delta_n(i, j) = \begin{cases} 
1, & \text{if } \hat{\theta}_{n,i} < \hat{\theta}_{n,j} \text{ and } C_{n,i}(\gamma) \cap C_{n,j}(\gamma) = \emptyset \\
0, & \text{otherwise.} 
\end{cases} 
\] (3.2)

The departure from Tukey (1953), and others, lies in modifying the way in which
the half-lengths of the individual intervals are determined. Rather than construct si-
multaneous confidence sets for the pairwise differences and then apportion the margin
of error to the individual parameter estimates, we propose a direct approach for esti-
mating the half-lengths via the parameter $\gamma$. In particular, in order to maintain the
desired control over the familywise error rate, our strategy for choosing $\gamma$ is based on
directly estimating the probability of observing a pair of non-overlapping uncertainty
intervals as a function of $\gamma$ when all of the parameters are equal.

Towards formally describing this procedure, first note that, as a function of $\gamma$, the
probability of declaring at least one significant difference when all $k$ parameters are
equal is simply the probability that the lower bound of one uncertainty interval lies
above the upper bound of another uncertainty interval:

\[
Q_n(\gamma; P) := \text{Prob}_P\left[ \sqrt{n} \max_{i,j \in K} \{L_{n,i}(\gamma) - U_{n,j}(\gamma)\} > 0 \right]. 
\] (3.3)

The “ideal” choice of $\gamma$ when all $k$ parameters are equal is thus

\[
\gamma_n(\alpha) = \inf \{ \gamma : Q_n(\gamma; P) \leq \alpha \}. 
\] (3.4)
Note that $\gamma_n(\alpha)$ is ideal in the sense that, if obtained, would guarantee finite sample control of the familywise error rate (Equation 2.3).

Unfortunately, because $P$ is unknown, we cannot compute $Q_n(\gamma; P)$, and $\gamma_n(\alpha)$ is thus infeasible. We therefore turn our attention to estimating $\gamma_n(\alpha)$ so as to achieve at least asymptotic control of the familywise error rate.

Constructing a feasible counterpart to $\gamma_n(\alpha)$ of course requires that we first formulate an empirical estimator of $Q_n(\gamma; P)$. Towards this end, let $\hat{P}_n$ be an estimate of $P$, and let $\hat{\theta}^*_{n,i}$ and $\hat{\sigma}^*_{n,i}$ denote the corresponding versions of $\hat{\theta}_{n,i}$ and $\hat{\sigma}_{n,i}$ that are obtained under sampling from $\hat{P}_n$. For example, $\hat{P}_n$ is often chosen to be the empirical distribution when the data are i.i.d., whereas block bootstrap methods are commonly employed when the data exhibit some form of serial dependence; see Lahiri (2003). The idea is to estimate $Q_n(\gamma; P)$ using its bootstrap analogue

$$Q_n(\gamma; \hat{P}_n) = \text{Prob}_{\hat{P}_n} \left[ \sqrt{n} \max_{i,j \in K} \{ L_{n,i}^*(\gamma) - U_{n,j}^*(\gamma) \} > 0 \right], \quad (3.5)$$

where

$$L_{n,i}^*(\gamma) = \left( \hat{\theta}^*_{n,i} - \hat{\theta}_{n,i} \right) - \frac{\gamma}{\sqrt{n}} \hat{\sigma}^*_{n,i}, \quad (3.6)$$

and

$$U_{n,i}^*(\gamma) = \left( \hat{\theta}^*_{n,i} - \hat{\theta}_{n,i} \right) + \frac{\gamma}{\sqrt{n}} \hat{\sigma}^*_{n,i}, \quad (3.7)$$

which, in turn, leads naturally to an estimator of $\gamma$ as

$$\hat{\gamma}_n(\alpha) = \inf \left\{ \gamma : Q_n(\gamma; \hat{P}_n) \leq \alpha \right\}. \quad (3.8)$$

Plugging $\hat{\gamma}_n(\alpha)$ into $C_{n,i}(\gamma)$ gives rise to a simple graphical device for visualizing statistically and practically significant differences.
The basic overlap plot is illustrated in Figure 1 using a toy example in which we attempt to rank the $k = 5$ busiest U.S. commercial airlines (American, Delta, Southwest, United, and US) by the mean of their daily average arrival delay times during 2007 ($n = 365$).\footnote{These data are obtained from the Research and Innovative Technology Administration, which coordinates the U.S. Department of Transportation research program (the entire dataset is freely available at http://www.transtats.bts.gov/). Other analyses of these data include Wicklin (2011).} The plot includes point estimates, which range from 5.24 minutes (Southwest) to 13.90 minutes (American), along with uncertainty intervals constructed using the overlap procedure with $\alpha = 0.05$ and $B = 9,999$ i.i.d. bootstrap replications (here, we obtain $\gamma_n^* = 1.32$). The ordering of point estimates together with the non-overlap of some of the uncertainty intervals allows us to rank Southwest the lowest, Delta second lowest, and American highest in terms of average arrival delay times. However, given that the uncertainty intervals for United and US (the airlines with the second and third highest point estimates, respectively) overlap, we are unable to reconcile the ranking of these airlines. As we show later in Section
4. a refinement of this ranking can be obtained without sacrificing control over the familywise error rate. First, however, we establish the large sample properties of this basic overlap procedure.

### 3.1 Large Sample Properties

In order to examine the large sample properties of the overlap procedure that result from the use of $\gamma_n^*(\alpha)$ in (3.1), we adopt the following high-level assumptions.

**Assumption 3.1.**

i. For every $\mathbf{x} = (x_1, \ldots, x_k)$, the $k$-variate distribution function

$$
\mathcal{L}_n(\mathbf{x}) = \text{Prob}_P \left[ \sqrt{n} \left( \hat{\theta}_{n,i} - \theta_i \right) \leq x_i, \ 1 \leq i \leq k \right]
$$

converges to a *continuous* and *strictly increasing* limiting distribution function $\mathcal{L}_\infty(\mathbf{x})$ as $n$ tends to infinity.

ii. For every $\mathbf{x} = (x_1, \ldots, x_k)$, the $k$-variate distribution function

$$
\mathcal{L}_n^*(\mathbf{x}) = \text{Prob}_{P_\mathbf{\hat{x}}} \left[ \sqrt{n} \left( \hat{\theta}_{n,i}^* - \hat{\theta}_{n,i} \right) \leq x_i, \ 1 \leq i \leq k \right]
$$

converges in probability to $\mathcal{L}_\infty(\mathbf{x})$ as $n$ tends to infinity.

iii. For every $i \in K$, $\hat{\sigma}_{n,i}$ and $\hat{\sigma}_{n,i}^*$ converge to a (common) positive constant $\sigma_i$ in probability.

The conditions set out in Assumption 3.1 mirror Assumptions 3.1 and 4.1 of Romano and Wolf (2005b). As remarked by Romano and Wolf (2005b), Part (i) of Assumption 3.1 is very weak and holds, for example, in the case of a limiting
multivariate normal distribution with nonsingular covariance matrix. Part (ii) of Assumption 3.1 requires that the bootstrap distribution consistently estimate the limit distribution $L_\infty$, and Part (iii) requires consistency of the bootstrap variance estimators. Romano and Wolf (2005b, pp. 1249-1254) provide a detailed discussion of some fairly flexible scenarios under which these conditions are satisfied. Sufficient conditions for Assumption 3.1 are also discussed at length for various scenarios in both the i.i.d. and dependent cases in Shao and Tu (1995) and Lahiri (2003), respectively.

We are now in a position to state our main result.

**Theorem 3.2.** Suppose the conditions in Assumption 3.1 are satisfied. Then, the decision rule $\delta_n$ as defined in (3.2) with $\gamma = \gamma^n_\alpha(\alpha)$ satisfies

1. **Error Rate Control (Type I)**

$$\lim_{n \to \infty} \text{Prob}_P \left[ \sum_{(i,j) \in D_0(P)} \delta_n(i,j) \geq 1 \right] \leq \alpha$$  

    \hspace{1cm} (3.9)

2. **Consistency**

$$\lim_{n \to \infty} \text{Prob}_P \left[ \prod_{(i,j) \in D_1(P)} \delta_n(i,j) = 1 \right] = 1$$  

    \hspace{1cm} (3.10)

3. **Control of Directional Errors**

$$\lim_{n \to \infty} \text{Prob}_P \left[ \sum_{(i,j) \in D_1(P)} \delta_n(j,i) \geq 1 \right] = 0$$  

    \hspace{1cm} (3.11)

The first property, which relates to control of the familywise error rate, states that the probability of inferring a strict ordering where non exists (i.e., where a pair of parameters are in fact equal) is guaranteed to be bounded from above by the prespecified nominal level $\alpha$, at least asymptotically. We show in the proof of Theorem
3.2 (see Appendix A.2) that (3.9) in fact holds with equality when all \( k \) parameters are equal.

The second and third properties relate to inferring the correct ordering of unequal parameters. Specifically, the second property states that the procedure will identify all of the strict orderings with probability tending to one with the sample size, whereas the third property states that the probability of getting the order of any strictly ordered pair incorrect tends to zero with the sample size.

### 3.2 Confidence Intervals for Pairwise Differences

In addition to allowing significant differences to be determined visually, the overlap plot also provides us with confidence intervals for the differences between any given pair of parameters under consideration. Specifically, an asymptotically valid \( 1 - \alpha \) level confidence interval for \( \theta_i - \theta_j \) is easily obtained as the difference in point estimates plus or minus the average widths of the two uncertainty intervals, i.e.,

\[
C_n(i,j)(\gamma_n^*(\alpha)) = \left[ (\hat{\theta}_{n,i} - \hat{\theta}_{n,j}) \pm \gamma_n^*(\alpha) \left( \frac{\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}}{\sqrt{n}} \right) \right].
\] (3.12)

For example, referring back to Figure 1, we see that an asymptotically valid 95% confidence interval for the difference in mean daily average arrival delays between Delta and Southwest (the second lowest and lowest ranked airlines, respectively) is given by \([1.64 \pm 1.25]\).

We show in Appendix B that \( C_n(i,j)(\gamma_n^*(\alpha)) \) is indeed guaranteed to have asymptotic coverage probability for \( \theta_i - \theta_j \) of at least \( 1 - \alpha \) whenever the overlap procedure maintains asymptotic control over the familywise error rate at the nominal level \( \alpha \).
3.3 Approximate \( p \)-values for Significant Pairwise Differences

Beyond a binary Yes/No decision, one can also infer the strength of evidence from the distance between non-overlapping uncertainty intervals. Concretely, suppose that \( \hat{\theta}_{n,i} < \hat{\theta}_{n,j} \) and that \( L_{n,j}(\alpha) \), the lower endpoint of the uncertainty interval for \( \theta_j \), is above \( U_{n,i}(\alpha) \), the upper endpoint of the uncertainty interval for \( \theta_i \). While equality of the parameters would certainly be rejected at the nominal level \( \alpha \), we can also quantify the strength of evidence against equality in the form of an approximate \( p \)-value by examining how

\[
\Delta_{n,(i,j)}(\alpha) = L_{n,j}(\alpha) - U_{n,i}(\alpha)
\]

measures on a \( p \)-value scale, which is a graphical device that may be appended to the overlap plot showing the association between differences in uncertainty interval endpoints and \( p \)-values. Indeed, we may use the fact that the probability of observing non-overlap lengths greater than or equal to \( \Delta > 0 \) can be estimated from

\[
Prob_{\hat{P}_n}\left\{ \max_{i,j \in K} \left[ L_{n,i}(\gamma_n^*) - U_{n,j}(\gamma_n^*) \right] > \Delta \right\}
\]

(3.13)

to produce a \( p \)-value scale.

Figure 2 displays such a \( p \)-value scale for the airline example considered above. Using this scale, we see that the distance between the uncertainty intervals for American (the highest ranked airline) and Southwest (the lowest ranked airline), which is \( 13.05 - 5.79 = 7.26 \), corresponds to a (multiplicity-adjusted) \( p \)-value of slightly less than 0.01.

4 Global Refinement

In this section, we consider situations in which a ranking remains incomplete after an application of the overlap procedure, and investigate the potential for an iterative
Figure 2: $p$-value scale for comparison of mean daily average arrival delays

stepwise procedure to refine the ranking without sacrificing control over the desired error rate.

In order to facilitate our discussion of a possible refinement strategy, we define

$$D_n(\gamma) = \bigcup_{i=1}^k \{i\} \times A_{n,i}(\gamma)$$

(4.1)

where

$$A_{n,i}(\gamma) = \{j \in K : C_{n,i}(\gamma) \cap C_{n,j}(\gamma) \neq \emptyset\}.$$  

(4.2)

The set $D_n(\gamma)$ is the collection of unresolved pairwise comparisons for a given value of the threshold parameter $\gamma$. For example, when an application of the overlap procedure with $\gamma = \gamma_n^*$ fails to rank any two parameters, we have $D_n(\gamma_n^*) = K \times K$. At the other extreme, when an application of the overlap procedure produces a complete ranking of the parameters under consideration, we have $D_n(\gamma_n^*) = \{(i, i) : i \in K\}$.

The basic idea behind the refinement strategy involves updating the set $D_n(\gamma)$ generated by an application of the overlap procedure and using the updated set to re-estimate the probability of committing a Type I error as a function of $\gamma$. For example, since the basic overlap procedure gives rise to $D_{n,1} = D_n(\gamma_n^*)$, we may re-estimate $\gamma$
in the second step using

\[
\gamma^*_D(n, 1)(\alpha) = \inf \left\{ \gamma : \text{Prob}_{\hat{P}_n} \left[ \sqrt{n} \max_{(i,j) \in D_{n,1}} \max \{ \Delta^*_n(i,j)(\gamma), \Delta^*_n(j,i)(\gamma) \} > 0 \right] \leq \alpha \right\}.
\]

(4.3)

where \( \Delta^*_n(i,j)(\gamma) = L^*_n(i)(\gamma) - U^*_n(j)(\gamma) \). In this way, the nominal level \( \alpha \) threshold \( \gamma^*_D(n, 1)(\alpha) \) is estimated with each \( i \in K \) compared only with members \( j \in K \) for which the ranking remains unresolved in the basic overlap plot.

Using \( \gamma^*_D(n, 1) \) to update the overlap plot gives rise to (weakly) narrower uncertainty intervals \( C_{n,i}(\gamma^*_D(n, 1)) \), \( i \in K \), and a possible refinement of the ranking. Iterating this procedure (at most \( k-1 \) times), as described formally in Algorithm 4.1 below, then gives rise to the refined decision rule, denoted by \( \delta^{\text{global}}_n \).

Algorithm 4.1.

1. Compute \( \gamma^*_n \) and set \( \delta^{\text{global}}_n(i,j) = \delta_n(i,j) \).

2. If \( D_{1,n} := D_n(\gamma^*_n) \subset K \times K \), compute \( \gamma^*_D(n, 1) \) and set

\[
\delta^{\text{global}}_n(i,j) = \begin{cases} 
1 & \text{if } C_{n,i}(\gamma^*_D(n, 1)) \cap C_{n,j}(\gamma^*_D(n, 1)) = \emptyset \text{ and } \hat{\theta}_{n,i} > \hat{\theta}_{n,j} \\
0 & \text{otherwise}
\end{cases}
\]

(4.4)

Else, stop.

3. If \( D_{2,n} := D_n(\gamma^*_D(n, 1)) \subset D_{1,n} \), compute \( \gamma^*_D(n, 2) \) and set

\[
\delta^{\text{global}}_n(i,j) = \begin{cases} 
1 & \text{if } C_{n,i}(\gamma^*_D(n, 2)) \cap C_{n,j}(\gamma^*_D(n, 2)) = \emptyset \text{ and } \hat{\theta}_{n,i} > \hat{\theta}_{n,j} \\
0 & \text{otherwise}
\end{cases}
\]

(4.5)
Else, stop.

To illustrate the use of the global refinement procedure outlined above, we revisit the airline example first introduced in Section 3. As the uncertainty intervals for US and United overlap in the basic plot illustrated in Figure 1, we apply Algorithm 4.1 in the hope of refining the ranking. In doing so, we obtain $\gamma_{2,n}^* = 1.02$ (recall that $\gamma_{1,n}^* = 1.32$), which implies substantially shorter intervals. Figure 3 shows the updated plot with the shorter intervals, from which we see that the ordering of US and United is now resolved.

Theorem 4.2 below shows that this refinement is obtained without sacrificing (asymptotic) control over the familywise error rate.

**Theorem 4.2.** Suppose that the conditions in Assumption 3.1 are satisfied. Then, in addition to satisfying conditions (1)-(3) of Theorem 3.2, the decision rule $\delta_{n}^{\text{global}}$
defined in Algorithm 4.1 satisfies

$$
\lim_{n \to \infty} \text{Prob}_P \left[ \sum_{(i,j) \in D_0(P)} \delta_n^{\text{global}}(i,j) \geq 1 \right] = \alpha,
$$

whenever $D_0(P)$ is non-empty. Additionally, $\delta_n^{\text{global}}$ satisfies

$$
\delta_n^{\text{global}}(i,j) \geq \delta_n(i,j) \text{ for every } (i,j) \in K \times K.
$$

Equation (5.4) in Theorem 4.2 shows that the limiting familywise error rate of the global refinement procedure is exactly equal to $\alpha$ when at least one pair of parameters under consideration is equal. Thus, the global refinement procedure is asymptotically non-conservative. In contrast, the limiting familywise error rate of the basic overlap procedure (i.e., without refinement) is exactly equal to $\alpha$ only if all of the parameters under consideration are equal, and is therefore conservative whenever $D_0(P)$ is a strict subset of $K \times K$.

Equation (4.7) together with equation (5.4) demonstrate that the global refinement procedure orders at least as many pairs as the basic overlap procedure while still maintaining control over the familywise error rate at the nominal level $\alpha$. These improvements are analogous to the well-documented improvements that are obtained through the use of iterative stepdown procedures in multiple testing; see, e.g., Holm (1979), Romano and Wolf (2005a), and Romano and Wolf (2005b, p. 1248).

5 Tail Rankings

We now turn our attention to situations in which the collection of parameters $K$ may be (substantially) larger than the ranking one ultimately seeks to produce. Concretely,
suppose that \( r < k = |K| \) and that interest centers on identifying and ranking only those parameters within the top \( r \) so that, for example, when \( r = 1 \) one cares only about identifying the top parameter, when \( r = 2 \) one cares only about identifying and ranking the top two parameters, and so on.\(^2\)

Because the basic overlap procedure and its global refinement are conservative when \( r < k \), we propose an alternative procedure for use in these situations. This new procedure is designed to increase power by eliminating irrelevant pairwise comparisons (i.e. the comparisons of parameter pairs outside the tail). More concretely, the procedure involves testing only whether the top \( j \) parameter estimates are statistically distinguishable from the bottom \( k - j + 1 \) for every \( 1 \leq j \leq r \).\(^3\)

By looking only at \( r \) consecutive pairs of minimums and maximums we can reduce the width of the uncertainty intervals relative to the basic (or even the globally refined) overlap procedure, thereby enabling us to resolve more comparisons in the tail region of interest.

To formalize this idea, let \( [1], [2], \ldots, [k] \), be the random indices such that \( \hat{\theta}_{n,[1]} > \hat{\theta}_{n,[2]} > \cdots > \hat{\theta}_{n,[k]} \). For example, if \( k = 3 \) and \( \hat{\theta}_{n,2} > \hat{\theta}_{n,3} > \hat{\theta}_{n,1} \), then \( [1] = 2, [2] = 3 \), and \( [3] = 1 \). We then define \( \text{smax}_{i \in K} L_{n,i}(\gamma) \) (\( \text{smax}_{i \in K} U_{n,i}(\gamma) \)) to be the lower (upper) endpoint of the uncertainty interval associated with the \( s \)th largest point estimate in the collection \( K \), so that \( \text{smax}_{i \in K} L_{n,i}(\gamma) = \hat{\theta}_{n,[s]} + \gamma \text{se}(\hat{\theta}_{n,[s]}) \) and \( \text{smax}_{i \in K} U_{n,i}(\gamma) = \hat{\theta}_{n,[s]} - \gamma \text{se}(\hat{\theta}_{n,[s]}) \), where \( \text{se}(\hat{\theta}_{n,i}) = \hat{\sigma}_{n,i}^{2}/\sqrt{n} \).

Interest then centers on the quantity

\[
e_{n,\ell}(\gamma) \equiv \min_{1 \leq s \leq \ell} \text{smax}_{i \in K} L_{n,i}(\gamma) - \max_{\ell < s \leq k} \text{smax}_{i \in K} U_{n,i}(\gamma),
\]

\(^{2}\)We focus of the upper tail for sake of exposition. Rankings of similarly defined subsets of the \( k \) parameters can, of course, be treated analogously.

\(^{3}\)Here we exploit the fact that \( a_i > a_{i+1} \) for \( i = 1, \ldots, r - 1 \) and \( a_r > a_j \) for every \( j > r \) if, and only if, \( \min\{a_1, \ldots, a_j\} > \max\{a_{j+1}, \ldots, a_k\} \) for every \( j = 1, \ldots, r \).
which is positive only if the lower end points associated with the top \( \ell \) parameter estimates are above the upper bounds of the remaining \( k - \ell + 1 \) intervals. We use \( e_{n,\ell}(\gamma) \) as the basis for distinguishing the top \( \ell \) parameters from the bottom \( k - \ell + 1 \) parameters.

Indeed, when it happens that \( e_{n,\ell}(\gamma) > 0 \), we declare the top \( \ell \) parameter estimates to be statistically distinguishable from the bottom \( k - \ell + 1 \) parameter estimates.

Since an error is committed when \( e_{n,\ell}(\gamma) > 0 \) and yet \( \theta_{[i]} \geq \theta_{[j]} \) for some \( i > \ell \) and some \( j \leq \ell \) (i.e. when a parameter that belongs to the top \( \ell \) is declared to be outside the top \( \ell \)), it is desirable that \( \gamma \) be chosen so as to maintain probabilistic control over the commission of one or more such errors, at least asymptotically. More formally, this amounts to selecting the maximal \( \gamma \) which ensures that

\[
Q_{n,r}(\gamma; P) := \text{Prob} \left[ \sum_{\ell=1}^{r} \max \left\{ \sqrt{n} e_{n,\ell}(n) d_{n,\ell}, 0 \right\} > 0 \right] < \alpha
\]  

(5.2)

at least asymptotically, where \( d_{n,\ell} \) is an indicator that is equal to one if \( \theta_{[i]} \geq \theta_{[j]} \) for some \( i > \ell \) and some \( j \leq \ell \) and zero otherwise.

As before, the problem as stated is infeasible because \( P \) is unknown and we therefore estimate the \( 1-\alpha \) quantile of the distribution \( Q_{n,r}(\gamma; P) \), denoted here by \( \gamma_{n,r}(\alpha) \), under the assumption that \( \theta_1 = \theta_2 = \cdots = \theta_k \), in which case \( d_{n,\ell} = 1 \) for every \( n > 1 \) and every \( 1 \leq \ell \leq r \), and, hence, the bootstrap analogue of (5.2) reduces to

\[
Q_{n,r}(\gamma; \hat{P}_n) := \text{Prob}_{\hat{P}_n} \left[ \sum_{\ell=1}^{r} \max \left\{ \sqrt{n} e_{n,\ell}^*(\gamma), 0 \right\} > 0 \right],
\]

(5.3)

where

\[
e_{n,\ell}^*(\gamma) \equiv \min \max_{1 \leq s \leq \ell} L_{n,i}^*(\gamma) - \max \max_{\ell < s \leq k} U_{n,i}^*(\gamma).
\]

Since \( Q_{n,r}(\gamma; \hat{P}_n) \) is known (or, at least, can be simulated), we use its \( 1-\alpha \) quantile, denoted here by \( \gamma_{n,r}^*(\alpha) \), to form the the \( r \) “tail” uncertainty intervals \( C_{n,[i]}(\gamma_{n,r}^*(\alpha)) \),
Along with the \( r \) uncertainty intervals, we also plot the maximum upper bound among all uncertainty intervals with center points outside the top \( r \). It is only by including this upper bound that we are able to distinguish between parameters inside and outside the top \( r \).

Figure 4: Tail ranking of airlines.

Figure 4 shows the tail overlap plot in the context of our earlier toy example involving the ranking of airlines by mean daily average delay time (here, we obtain \( \gamma^*_n,r = 0.77 \)). For illustrative purposes, we set \( r = 4 \) so that Southwest (the airline with the lowest point estimate) is the only airline not (explicitly) depicted in the plot. The dotted horizontal line, which in general marks the location of the maximum upper bound among uncertainty intervals with estimated center points outside the top \( r \), allows us to infer whether the top \( r \) parameter estimates can be distinguished from the “rest”, which in this case with \( r = 4 \) is just Southwest (note that this horizontal line need not, in general, correspond to the upper bound of the uncertainty interval with the \((r + 1)^{th}\) largest center point).
From Figure 4 we see that the lower bound of the uncertainty interval for American is above all other upper bounds. From this, we may infer that American has the highest parameter value. Similarly, since the lower bounds associated with American and United are above all other upper bounds, we may infer that United and American have the two highest parameter values. It thus follows that American can be ranked highest, and United ranked second highest. Continuing in this fashion, we are able to rank US third highest and Delta fourth highest.

Theorem 5.1 below shows that the practice of inferring rankings from the tail overlap procedure in this fashion is statistically sound in the sense that we are still able to maintain probabilistic control over the familywise error rate, at least asymptotically.

**Theorem 5.1.** Suppose that the conditions in Assumption 3.1 are satisfied. Then,

1. **Error Rate Control:**

\[
\lim_{n \to \infty} \text{Prob} \left[ \sum_{\ell=1}^{r} \max \left\{ \sqrt{n} e_{n,\ell}(\gamma_{n,r}^*(\alpha)) d_{n,\ell}, 0 \right\} > 0 \right] \leq \alpha, \tag{5.4}
\]

with equality when \( \theta_1 = \theta_2 = \cdots = \theta_k \).

2. **Consistency:** For all \( i, j \in K \),

\[
\lim_{n \to \infty} \text{Prob} \left[ L_{n,i}(\gamma_{n,r}^*(\alpha)) - U_{n,j}(\gamma_{n,r}^*(\alpha)) > 0 \mid \theta_i > \theta_j \right] = 1. \tag{5.5}
\]

### 6 Finite Sample Performance

In this section, we examine the finite-sample performance of both the basic overlap procedure and the tail ranking procedure by way of several Monte Carlo experiments.
6.1 Basic Overlap Procedure

As a point of comparison for the basic overlap procedure, we consider max-$T$-type procedures for all pairwise differences. These procedures base their decision rules on individual statistics of the form

$$T_{n,(i,j)} = \frac{\sqrt{n} |\hat{\theta}_{n,i} - \hat{\theta}_{n,j}|}{s_{n,(i,j)}}.$$  

For $s_{n,(i,j)}$, we consider the non-studentized case (max-$T$) with $s_{n,(i,j)} = 1$, and the studentized case (st.max-$T$) with

$$s_{n,(i,j)} = \sqrt{\hat{\sigma}^2_{n,i} + \hat{\sigma}^2_{n,j} - 2\hat{s}_{n,(i,j)}}, \tag{6.1}$$

where $\hat{\sigma}^2_{n,i}$ and $\hat{\sigma}^2_{n,j}$ are, as above, the estimated variances of $\sqrt{n}\hat{\theta}_{n,i}$ and $\sqrt{n}\hat{\theta}_{n,j}$, and $\hat{s}_{n,(i,j)}$ is the estimated covariance between $\sqrt{n}\hat{\theta}_{n,i}$ and $\sqrt{n}\hat{\theta}_{n,j}$. Additionally, in every trial we follow Romano and Wolf (2005b) and estimate the nominal level $\alpha$ critical value for each procedure using the $1 - \alpha$ quantile of $\max_{i \neq j} T^*_n,(i,j)$, where $T^*_n,(i,j)$ is the appropriately recentered bootstrap counterpart of the respective $T_{n,(i,j)}$. See Romano and Wolf (2005b) for further discussion regarding these procedures.

We consider three different designs in our Monte Carlo simulations. Taken together, these designs are intended to showcase the robustness of the overlap procedure. The first two designs involve generating $n$ i.i.d. random vectors from a $k$-dimensional multivariate normal distribution with mean vector $(\theta_1, \ldots, \theta_k)'$, and covariance matrix given by a Toeplitz matrix whose first row is $(\sigma^2_1, \rho, \ldots, \rho)$. In Design 1, we set $\sigma^2_i = 1$ for every $1 \leq i \leq k$, whereas in Design 2, we set $\sigma^2_1 = 5$ and $\sigma^2_i = 1$ for $i = 2, \ldots, k$. That is, in Design 1, the marginal variances are identical, whereas in Design 2, there is a single variance that differs substantially from the others.
In Design 3, we use a Gaussian copula, with common correlation coefficient \( \rho \), to generate \( n \) i.i.d. random vectors from a \( k \)-dimensional multivariate exponential distribution with rate parameter vector \((1/\theta_1, \ldots, 1/\theta_k)'\). Thus, the mean vector in Design 3 is \((\theta_1, \ldots, \theta_k)'\), while the marginal variances are \((\theta_1^2, \ldots, \theta_k^2)'\). We include this last design to showcase how the overlap procedure performs with non-identical skewed marginal distributions.

In order to examine control of the familywise error rate, we set \( \theta_i = 0 \) for \( 1 \leq i \leq k \) in Designs 1 and 2, and \( \theta_i = 1 \) for \( 1 \leq i \leq k \) in Design 3. We consider sample sizes \( n \in \{50, 100, 500, 1000\} \) and correlation parameters \( \rho \in \{0, 0.5, 0.9\} \). In all of
our simulations, we set the nominal level to be $\alpha = 0.05$. The simulation results corresponding to $k = 10$ (45 pairwise comparisons) and $k = 50$ (1,225 pairwise comparisons) are reported in Table 1. The results for $k = 10$ are based on 100,000 Monte Carlo replications with $B = 1,999$ bootstrap samples, whereas the results for $k = 50$ are based on 10,000 Monte Carlo replications with $B = 499$ bootstrap samples.

From Table 1, it is clear that control of the familywise error rate is, on the whole, satisfactory for all three designs for both $k = 10$ and $k = 50$. That is, the empirical familywise error rates are generally below the nominal level $\alpha$, with only the following exceptions: (i) the non-studentized max-$T$ procedure when the marginal variances differ substantially (Design 2) and the sample size is small; and (ii) both max-$T$-type procedures when $k = 50$ and the correlation level is low. There is also some evidence that the overlap procedure is relatively conservative when the sample size is small and the degree of correlation is high. We might anticipate, therefore, that the overlap procedure may exhibit lower average power relative to the max-$T$-type procedures in this latter case (see below).

We now turn our attention to the power of the overlap and max-$T$-type procedures. Specifically, we consider average power, which is the proportion of false null hypotheses that are rejected (Romano and Wolf, 2005b, pp. 1242-1243). Towards this end, we set $\theta_i = 0.1i$, $1 \leq i \leq k$, in Designs 1 and 2, and $\theta_i = 1 + 0.1i$, $1 \leq i \leq k$, in Design 3. The average power recorded from the simulations for $k = 10$ and $k = 50$ are reported in Table 2.

The results in Table 2 demonstrate that, when the degree of correlation increases, the average power of each of the procedures is uniformly higher (for any $n$). Also, when the marginal variances are identical (Design 1), all of the procedures offer very similar performance. However, when the variances are non-identical (strikingly so in Design
Table 2: Proportion of Strict Orderings Inferred at the 5% Nominal Level

<table>
<thead>
<tr>
<th>k</th>
<th>ρ</th>
<th>n</th>
<th>Design 1: N(0.1i, 1) marginals</th>
<th>Design 2: N(0.1i, 1 + 4(i = 1)) marginals</th>
<th>Design 3: Exp(1 + 0.1i) marginals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>max-T</td>
<td>st.max-T</td>
<td>Overlap</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>50</td>
<td>0.188</td>
<td>0.163</td>
<td>0.173</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>50</td>
<td>0.357</td>
<td>0.345</td>
<td>0.349</td>
</tr>
<tr>
<td>500</td>
<td>0</td>
<td>50</td>
<td>0.713</td>
<td>0.712</td>
<td>0.712</td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>50</td>
<td>0.818</td>
<td>0.818</td>
<td>0.818</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>50</td>
<td>0.354</td>
<td>0.325</td>
<td>0.330</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>50</td>
<td>0.526</td>
<td>0.515</td>
<td>0.515</td>
</tr>
<tr>
<td>500</td>
<td>0</td>
<td>50</td>
<td>0.818</td>
<td>0.817</td>
<td>0.817</td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>50</td>
<td>0.900</td>
<td>0.899</td>
<td>0.899</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
<td>50</td>
<td>0.710</td>
<td>0.690</td>
<td>0.690</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>50</td>
<td>0.817</td>
<td>0.811</td>
<td>0.810</td>
</tr>
<tr>
<td>500</td>
<td>0</td>
<td>50</td>
<td>0.963</td>
<td>0.993</td>
<td>0.993</td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>50</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

2, but also in Design 3), the max-T procedure generally performs significantly worse than either the studentized version (st.max-T) or the overlap procedure. Furthermore, it is only in larger samples and when the degree of correlation is rather high that we tend to see any practical difference in average power between the st.max-T and overlap procedures. In particular, when the degree of correlation is high and the sample size is rather large, there is evidence to suggest that the st.max-T procedure delivers higher average power than the overlap procedure.\(^4\)

\(^4\)Note that the overlap procedure is equivalent to a max-T-type procedure with \(s_{n,(i,j)} = \hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}\). This observation suggests that, at least in terms of 0-1 significance testing, the performance of the proposed overlap procedure should be on par with the st.max-T procedure whenever the estimators \(\theta_{n,i}, i \in K\), are uncorrelated, and on par with the max-T (and st.max-T) procedure whenever the estimators \(\theta_{n,i}, i \in K\), are correlated.
6.2 Tail Ranking Procedure

To examine the finite-sample performance of the tail ranking procedure outlined in Section 5, we use the same three basic designs as above with all $k$ means equal (i.e., we set $\theta_i = 0$ for $1 \leq i \leq k$ in Designs 1 and 2, and $\theta_i = 1$ for $1 \leq i \leq k$ in Design 3). This allows us to simultaneously examine both the control of the familywise error rate, as well as the potential gain in power over the basic overlap procedure, which we gauge using the average value of the ratio $\gamma_{n,r}^*/\gamma_n^*$.

Fixing $r = k/10$, we again consider both $k = 10$ and $k = 50$. Note that the number of comparisons made in the tail ranking procedure is $\sum_{i=1}^{r} (k - i) = r[k - (r + 1)/2]$, since we compare $\theta_{[i]}$ to $\theta_{[j]}$ for every $1 \leq i \leq r$ and every $i < j \leq k$. Thus, with $k = 10$ and $r = 1$, there are 9 comparisons (as opposed to 45 when we considered all pairwise comparisons), and with $k = 50$ and $r = 5$, there are 235 comparisons (as opposed to 1,225 when we considered all pairwise comparisons). Since all of the parameters are equal, the familywise error rate is the probability that at least one of these comparisons results in us inferring a strict ordering, i.e., the probability of inferring that $\theta_{[i]} > \theta_{[j]}$ for at least one pair $(i,j)$ with $1 \leq i \leq r$ and $i < j \leq k$. Our results are reported in Table 3.\footnote{As before, the results for $k = 10$ are based on 100,000 Monte Carlo replications with $B = 1,999$ bootstrap samples, whereas the results for $k = 50$ are based on 10,000 Monte Carlo replications with $B = 499$ bootstrap samples.}

It is clear from Table 3 that control of the familywise error rate (denoted FWER in the table) is satisfactory for all three designs for both $k = 10$ and $k = 50$. For $k = 10$, the empirical familywise error rates are always very close to the nominal level; indeed, the absolute discrepancy in rejection probabilities is never greater than 0.003 whenever the estimators are uncorrelated and have equal variances. For $k = 50$, the empirical familywise error rates are also quite close to the nominal level $\alpha$ for Designs 1 and 2, but tend to be slightly
Table 3: TITLE

<table>
<thead>
<tr>
<th></th>
<th>Design 1: N(0,1) marginals</th>
<th>Design 2: N(0,1 + 4F(i = 1)) marginals</th>
<th>Design 3: Exp(1) marginals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \gamma^<em>_n, \gamma^</em>_n ) FWER</td>
<td>( \gamma^<em>_n, \gamma^</em>_n ) FWER</td>
<td>( \gamma^<em>_n, \gamma^</em>_n ) FWER</td>
</tr>
<tr>
<td>10</td>
<td>0.312 0.053</td>
<td>0.356 0.051</td>
<td>0.295 0.046</td>
</tr>
<tr>
<td></td>
<td>0.320 0.051</td>
<td>0.365 0.051</td>
<td>0.308 0.047</td>
</tr>
<tr>
<td></td>
<td>0.326 0.050</td>
<td>0.372 0.050</td>
<td>0.320 0.050</td>
</tr>
<tr>
<td></td>
<td>0.327 0.050</td>
<td>0.373 0.050</td>
<td>0.322 0.050</td>
</tr>
<tr>
<td>0.5</td>
<td>0.312 0.049</td>
<td>0.436 0.053</td>
<td>0.288 0.046</td>
</tr>
<tr>
<td></td>
<td>0.320 0.050</td>
<td>0.447 0.051</td>
<td>0.303 0.048</td>
</tr>
<tr>
<td></td>
<td>0.326 0.051</td>
<td>0.456 0.050</td>
<td>0.318 0.049</td>
</tr>
<tr>
<td></td>
<td>0.327 0.050</td>
<td>0.457 0.052</td>
<td>0.321 0.050</td>
</tr>
<tr>
<td>0.9</td>
<td>0.313 0.046</td>
<td>0.632 0.052</td>
<td>0.291 0.043</td>
</tr>
<tr>
<td></td>
<td>0.320 0.048</td>
<td>0.639 0.052</td>
<td>0.305 0.047</td>
</tr>
<tr>
<td></td>
<td>0.326 0.050</td>
<td>0.644 0.051</td>
<td>0.320 0.050</td>
</tr>
<tr>
<td></td>
<td>0.327 0.050</td>
<td>0.645 0.051</td>
<td>0.323 0.049</td>
</tr>
<tr>
<td>50</td>
<td>0.183 0.050</td>
<td>0.210 0.050</td>
<td>0.170 0.043</td>
</tr>
<tr>
<td></td>
<td>0.193 0.047</td>
<td>0.221 0.052</td>
<td>0.183 0.043</td>
</tr>
<tr>
<td></td>
<td>0.200 0.047</td>
<td>0.229 0.048</td>
<td>0.195 0.047</td>
</tr>
<tr>
<td></td>
<td>0.201 0.048</td>
<td>0.230 0.048</td>
<td>0.197 0.045</td>
</tr>
<tr>
<td>0.5</td>
<td>0.185 0.048</td>
<td>0.280 0.051</td>
<td>0.164 0.040</td>
</tr>
<tr>
<td></td>
<td>0.193 0.048</td>
<td>0.293 0.051</td>
<td>0.178 0.045</td>
</tr>
<tr>
<td></td>
<td>0.200 0.048</td>
<td>0.304 0.051</td>
<td>0.193 0.047</td>
</tr>
<tr>
<td></td>
<td>0.201 0.046</td>
<td>0.305 0.049</td>
<td>0.196 0.044</td>
</tr>
<tr>
<td>0.9</td>
<td>0.187 0.043</td>
<td>0.546 0.053</td>
<td>0.169 0.039</td>
</tr>
<tr>
<td></td>
<td>0.194 0.047</td>
<td>0.555 0.050</td>
<td>0.181 0.039</td>
</tr>
<tr>
<td></td>
<td>0.200 0.045</td>
<td>0.561 0.050</td>
<td>0.195 0.043</td>
</tr>
<tr>
<td></td>
<td>0.201 0.049</td>
<td>0.561 0.048</td>
<td>0.198 0.041</td>
</tr>
</tbody>
</table>

lower in Design 3 (where the marginal distributions are skewed), particularly when the sample size is small and/or the degree of correlation is high.

Table 3 also demonstrates that the tail ranking procedure offers a large potential power gain over the basic overlap procedure. For \( k = 10 \) \((k = 50)\), the ratio \( \gamma^*_n, \gamma^*_n \) is close to 0.3 (0.2) for all sample sizes and correlation levels for Designs 1 and 3, but increases with the degree of correlation in Case 2, to around 0.6 when \( \rho = 0.9 \).
7 Empirical Illustration

In this section, we present a larger-scale example in which we seek to rank a collection of hedge funds by the mean of their risk-adjusted returns to investors. The hedge funds entering into the ranking are those included in the Centre for International Securities and Derivatives Market (CISDM) database with a complete monthly return history for the period January 1994 – June 2013 (n = 234). This leaves us with k = 156 funds.

Since the parameters of interest, namely the means of the risk-adjusted returns, can be expressed as $\theta_i = \theta_i(m_i) = m_{i,1}/\sqrt{m_{i,2} - m_{i,1}^2}$ for $1 \leq i \leq k$, where $m_i = (m_{i,1}, m_{i,2})$ is the vector of first and second uncentered moments of the $i^{th}$ return series, it follows under mild conditions that the plug-in estimator $\hat{\theta}_n$ of the $k \times 1$ vector $\theta$, when scaled and recentered, converges in distribution to a mean-zero Gaussian random vector whose covariance matrix has diagonal elements

$$\sigma_i^2 = \nabla \theta_i(m_i)' \Psi_i \nabla \theta_i(m_i),$$

with $\Psi_i := \lim_{n \to \infty} \text{Var}(\sqrt{n}\hat{m}_{n,i})$ and $\nabla \theta(m_i)' = (m_{i,2}(m_{i,2} - m_{i,1}^2)^{-3/2}, -m_{i,1}[2(m_{i,2} - m_{i,1}^2)]^{-3/2})$. Thus, a consistent standard error for $\sqrt{n}\hat{\theta}_{n,i}$ is

$$\hat{\sigma}_{n,i} = \sqrt{\nabla \theta_i(\hat{m}_{n,i})' \hat{\Psi}_{n,i} \nabla \theta_i(\hat{m}_{n,i})},$$

where $\hat{\Psi}_{n,i}$ is a heteroskedasticity and autocorrelation (HAC) consistent estimate of $\Psi_i$. In obtaining $\hat{\Psi}_{n,i}$, we utilize the prewhitened quadratic spectral kernel and the corresponding automatic choice of bandwidth of Andrews and Monahan (1992); see Ledoit and Wolf (2008) for further details.

Moreover, because the individual return series exhibit serial correlation, we use the
circular block bootstrap with $B = 499$ bootstrap replications and a fixed blocksize $b = 5$ when estimating the parameters $\gamma_n^*$ and $\gamma_{n,r}^*$.\footnote{Existing data-driven methods for choosing the blocksize may also be adapted for use in the present context. While we do not investigate this issue in the present paper, the reader may wish to consult Romano and Wolf (2005b), Romano et al. (2008), and references therein for a discussion of the choice of block sizes in related contexts.} For $\hat{\sigma}_{n,i}^*$, the bootstrap counterpart of $\hat{\sigma}_{n,i}$, we use the “natural” standard error that exploits the block dependence structure; see Götze and Künsch (1996).

Figure 5: Hedge funds: Basic overlap

Figure 5 depicts the basic overlap plot at the nominal level $\alpha = 0.10$ (here, we obtain $\gamma_n^* = 3.86$). Even at this higher nominal level, the uncertainty intervals associated with the funds are large enough to severely handicap the procedure’s ability to
resolve most pairwise comparison of funds. For example, we are unable to distinguish
the performance of the “best” fund from 51 other funds with lower point estimates
because their uncertainty intervals overlap with the “top” fund’s uncertainty interval.

We next consider an application of the global refinement procedure outlined in
Section 4. While we are able to obtain a very slight reduction in the length of the
uncertainty intervals (γ_{2,n}^* = 3.84, while γ_{1,n}^* = 3.86), the equivalence classes for 146 of
the 156 funds (including that of the “best” fund) are unchanged. As the differences
between the uncertainty intervals depicted in Figure 5 and the refined uncertainty
intervals are not easily detected visually, we do not plot these refined uncertainty
intervals.

Finally, since neither the basic overlap procedure nor its global refinement are able
to resolve many of the pairwise comparisons between funds, and because interest in
such settings often centers on identifying the top performing funds, we apply the tail
ranking procedure with r = 5. Keeping α = 0.1 and using the same circular block
bootstrap as above with b = 5, we obtain γ_{n,r}^* = 0.69.

The corresponding overlap plot, shown in Figure 6, reveals that the uncertainty
interval of the “best” fund (Equity Income Partners LP) lies strictly above all other
intervals and, as such, can be declared the best performing fund with respect to mean
risk-adjusted returns. The funds with the second and third highest point estimates
(Millennium USA LP and Millennium International Ltd, respectively) can be ranked
above all the funds with lower point estimates (i.e., all funds except the “best” fund,
Equity Income Partners LP). Accordingly, we can say that both of these two funds
are in contention for the second and third highest positions.

Moreover, as the dotted horizontal line indicates the location of the maximum
upper bound among uncertainty intervals with estimated center points outside the top
r = 5, we conclude that the funds with the third and fourth highest point estimates
Figure 6: Hedge funds: Tail ranking with $r = 5$

(Bryn Mawr Capital LP and TIG Arbitrage Associates LP, respectively) can not be distinguished from at least some of the funds outside the top $r = 5$. These two funds, and any of those funds whose uncertainty intervals overlap with their uncertainty intervals, can simply be considered candidates for the fourth and fifth highest positions.

8 Concluding Remarks

We have shown that Tukey’s (1953) overlap procedure can be generalized in a natural way to accommodate general forms of dependence within and across samples. Additionally, we have developed two new complementary procedures: the first being an iterative stepwise method, and the second being an extension of Hsu’s (1981, 1984) procedures for “multiple comparisons with the best.” These new procedures maintain asymptotic control of the classical familywise error rate, and exhibit good performance.
in finite samples.

An obvious extension of these procedures would be to provide control of some generalized notions of the familywise error rate (see, e.g., Benjamini and Hochberg, 1995, and Ramsey et al., 2011), thereby building on Tukey’s own suggestion that one relax the familywise error rate when the number of comparisons becomes large (see, e.g., Basford and Tukey, 1999). This we leave to future work.
References


A Technical Appendix

Section A.1 establishes a number of useful results, including the consistency of the bootstrap estimator $Q_n(\cdot; \hat{P}_n)$ of $Q_n(\cdot; P)$ (Lemma A.2) and the consistency of the bootstrap estimator $\gamma^*_n(\alpha)$ of the quantile $\gamma_n(\alpha)$ (Lemma A.3). Section A.2 contains the proofs of the main theorems from the paper.

A.1 Large sample properties of $Q_n$

Define the map $Q: \mathbb{R}_+ \rightarrow [0, 1]$ by

$$Q(\gamma; P) := \text{Prob}_P[g(X, \sigma, \gamma) > 0],$$

where $X$ is a $k \times 1$ random vector with joint distribution $\mathcal{L}_\infty$, $\sigma$ is a $k \times 1$ vector (possibly random), and $g: \mathbb{R}^k \times \mathbb{R}^{k+1}_+ \rightarrow \mathbb{R}$ is defined by

$$g(x, y, \gamma) = \max_{j \in K} \{x_j - \gamma y_j\} - \min_{j \in K} \{x_j + \gamma y_j\}. \quad (A.2)$$

Because the function $Q(\gamma; P)$ plays such a prominent role in our analysis, we begin by establishing several of its key properties.

**Lemma A.1.** $Q(\gamma; P)$ is a continuous and weakly decreasing function satisfying

$$Q(0; P) = 1 \text{ and } \lim_{\gamma \to \infty} Q(\gamma; P) = 0.$$

**Proof.** That $Q(\gamma; P)$ is weakly decreasing in $\gamma$ follows immediately from the fact that $g$ is strictly decreasing in $\gamma$. In order to prove continuity, we first write

$$\lim_{\delta \to 0} Q(\gamma + \delta; P) = \lim_{\delta \to 0} \text{Prob}_P[g(X, \sigma, \gamma + \delta) \geq 0]$$

$$= \lim_{\delta \to 0} \text{Prob}_P[g(X, \sigma, \gamma) + (g(X, \sigma, \gamma + \delta) - g(X, \sigma, \gamma)) \geq 0] \quad (A.3)$$

From the second line, we see that in order to establish continuity of $Q(\gamma; P)$ in $\gamma$, it suffices to show that, for all $\epsilon > 0$,

$$\lim_{\delta \to 0} \text{Prob}_P[|g(X, \sigma, \gamma + \delta) - g(X, \sigma, \gamma)| > \epsilon] = 0. \quad (A.4)$$

By continuity of $g$, however, we obtain

$$\lim_{\delta \to \infty} |g(X(\omega), \sigma, \gamma + \delta) - g(X(\omega), \sigma, \gamma)| = 0 \text{ for all } \omega,$$

35
which is sufficient for (A.4). Hence, \( \lim_{\delta \to 0} Q(\gamma + \delta; P) = Q(\gamma; P) \) is established.

Towards showing that \( Q(0, P) = 1 \), first note that

\[
Q(0, P) = \Pr_P \left[ \max_{j \in K} X_j - \min_{j \in K} X_j > 0 \right] = \Pr_P \left[ \max_{i, j \in K} |X_i - X_j| > 0 \right] \quad \text{(A.5)}
\]

Because the joint distribution \( L_\infty \) of the random vector \( X \) is continuous and strictly increasing in its arguments (cf. Assumption 3.1), we have \( \Pr_P[X_i = X_j] = 0 \) for all \( i \neq j \) and, hence,

\[
\Pr_P \left[ \max_{i, j \in K} |X_i - X_j| > 0 \right] = 1.
\]

Last, we wish to show that \( \lim_{\gamma \to \infty} Q(\gamma; P) = 0 \). To obtain the desired result, we exploit the inequality

\[
\max_{j \in K} \{X_j - \gamma \sigma_j\} - \min_{j \in K} \{X_j + \gamma \sigma_j\} \leq \max_{j, \ell \in K} |X_\ell - X_j| - \gamma \min_{j \in K} \sigma_j \quad \text{(A.6)}
\]

to obtain

\[
\Pr_P[g(X, \sigma, \gamma) \geq 0] \leq \Pr_P \left[ \max_{i, j \in K} |X_i - X_j| - \gamma \min_{j \in K} \sigma_j > 0 \right]. \quad \text{(A.7)}
\]

Because (A.7) holds for all \( \gamma \in \mathbb{R}_+ \) and \( \max_{i \neq j} |X_i - X_j| \) is bounded in probability, the desired result is obtained upon taking limits as \( \gamma \to \infty \).

Having introduced and examined the limiting function \( Q(\gamma; P) \), we now set out to prove the consistency of

\[
Q_n(\gamma; \hat{P}_n) := \Pr_{\hat{P}_n} \left[ g(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n), \hat{\sigma}_n, \gamma) > 0 \right]
\]

as an estimator of

\[
Q_n(\gamma; P) := \Pr_P \left[ g(\sqrt{n}(\hat{\theta}_n - \theta), \hat{\sigma}_n, \gamma) > 0 \right],
\]

and to show that the difference between the quantiles \( \gamma^*_n(\alpha) \) and \( \gamma_n(\alpha) \) tends to zero in probability as the sample size tends to infinity.

**Lemma A.2.** Suppose that the conditions of Assumption 3.1 are satisfied. Then, for all \( \epsilon, \delta > 0 \), there exists \( N \) such that for all \( n \geq N \)

\[
\Pr_P \left[ \sup_{\gamma \in \mathbb{R}_+} |Q_n(\gamma; \hat{P}_n) - Q_n(\gamma; P)| > \epsilon \right] < \delta \quad \text{(A.8)}
\]
Proof. Let $\epsilon, \delta > 0$ be given. It suffices to show that, for $n$ sufficiently large,

$$\sup_{\gamma \in \mathbb{R}_+} |Q_n(\gamma; P) - Q(\gamma; P)| < \epsilon/2$$

(A.9)

and

$$\text{Prob}_P \left[ \sup_{\gamma \in \mathbb{R}_+} |Q_n(\gamma; \hat{P}_n) - Q(\gamma; P)| > \frac{\epsilon}{2} \right] < \delta. \quad \text{(A.10)}$$

We first establish (A.9). To this end, let $X_{n,j}$ be a random variable satisfying $X_{n,j} \sim L_{n,j}$ where $L_{n,j}$ is the $j$th marginal of the joint distribution $L_{n}$ as defined in Assumption 3.1(i). After recalling that Assumption 3.1 provides us with $X_n \xrightarrow{d} X (X \sim L_\infty)$ and $\hat{\sigma}_{n,j} \xrightarrow{p} \sigma_j$ for every $j \in K$, we obtain

$g(X_n, \hat{\sigma}_{n,j}, \gamma) \xrightarrow{d} g(X, \sigma, \gamma)$

as an immediate consequence of the continuous mapping theorem. Hence, for every $\gamma \in \mathbb{R}_+$ there exists $N_\gamma$ such that for all $n \geq N_\gamma$

$$|Q_n(\gamma; P) - Q(\gamma; P)| < \epsilon/4.$$

(A.11)

That the convergence is uniform over $\gamma \in \mathbb{R}_+$ is then an immediate consequence of Polyá’s theorem (Athreya and Lahiri, 2006, p. 242).

Towards establishing (A.10), we note that arguments analogous to those used above, together with Assumption 3.1 parts (ii) and (iii), imply that for all $\gamma \in \mathbb{R}_+$ there exist $N'_\gamma$ sufficiently large such that, for all $n \geq N'_\gamma$,

$$\text{Prob}_P \left[ |Q_n(\gamma; \hat{P}_n) - Q(\gamma; P)| > \epsilon/4 \right] < \delta,$$

which in turn gives

$$\text{Prob}_P \left[ \sup_{\gamma \in \mathbb{R}_+} |Q_n(\gamma; \hat{P}_n) - Q(\gamma; P)| > \epsilon/2 \right] < \delta \quad \text{(A.12)}$$

for all $n$ sufficiently large by Polyá’s theorem.

Lemma A.3. Suppose that the conditions of Assumption 3.1 are satisfied. Then, for all $\epsilon, \delta > 0$, there exists $N$ such that for all $n \geq N$

$$\text{Prob}_P \left[ |\gamma_n(\alpha) - \gamma^*_n(\alpha)| > \delta \right] < \epsilon$$

Proof. The proof is adapted from Beran (1984). Let $\epsilon, \delta > 0$ be given and define the
generalized inverse

\[ Q^{-1}(\alpha; P) = \inf\{ \gamma : Q(\gamma; P) \leq \alpha \}. \]

Lemma A.2, which shows that \( Q_n(\gamma; P) \) converges uniformly to \( Q(\gamma; P) \) as \( n \) tends to \( \infty \), implies that

\[ Q_n(Q^{-1}(\alpha; P) + \delta/2; P) < \alpha < Q_n(Q^{-1}(\alpha; P) - \delta/2; P) \]  \hspace{1cm} (A.13)

for sufficiently large \( n \). An immediate consequence of (A.13) together with \( Q_n \) being decreasing in \( \gamma \) is that

\[ Q^{-1}(\alpha; P) - \delta/2 < \gamma_n(\alpha) < Q^{-1}(\alpha; P) + \delta/2 \]  \hspace{1cm} (A.14)

for \( n \) sufficiently large, where we have used the fact that \( \gamma_n(\alpha) = Q_n^{-1}(\alpha; P) \).

Similarly, from Lemma A.2, we have that \( Q_n(\gamma, ˆP_n) \) converges uniformly (in probability) to \( Q(\gamma; P) \). Consequently, there exists \( n \) sufficiently large such that

\[ \Pr_p \left[ Q_n(Q^{-1}(\alpha; P) + \delta/2; ˆP_n) < \alpha < Q_n(Q^{-1}(\alpha; P) - \delta/2; ˆP_n) \right] \geq 1 - \epsilon \]  \hspace{1cm} (A.15)

and, hence,

\[ \Pr_p \left[ Q^{-1}(\alpha; P) - \delta/2 < \gamma^*_n(\alpha) < Q^{-1}(\alpha; P) + \delta/2 \right] \geq 1 - \epsilon \]  \hspace{1cm} (A.16)

for \( n \) sufficiently large. Upon combining (A.14) and (A.16) we obtain

\[ \Pr_p \left[ |\gamma^*_n(\alpha) - \gamma_n(\alpha)| > \delta \right] < \epsilon \]

for \( n \) sufficiently large, which—given that \( \epsilon \) and \( \delta \) are arbitrary—establishes the desired convergence. \( \square \)

### A.2 Proofs of the Main Results

**Proof of Theorem 3.2 (Part 1: FWE Control):** Let \( \alpha \in (0,1) \) be given. We are required to show that

\[ \lim_{n \to \infty} \Pr_p \left[ \sum_{(i,j) \in D_0(P)} \delta_n(i,j) \geq 1 \right] \leq \alpha, \]  \hspace{1cm} (A.17)

where \( D_0(P) \) is the set pairs \( (i,j) \in K \times K \) \( i \neq j \) for which \( \theta_i = \theta_j \). We proceed in two steps. First, we establish that (A.25) holds with equality when \( D_0 = K \times K \). We then show
that the inequality in (A.25) remains satisfied when $D_0 \subset K \times K$.

Thus, suppose $D_0(P) = K \times K$ so that all of the parameters are equal under $P$. From the definition of the indicator function $\delta_n$, we then have

$$
Prob_P \left[ \sum_{(i,j) \in D_0(P)} \delta_n(i,j) \geq 1 \right] = Prob_P \left[ g(X_n, \hat{\sigma}, \gamma_n^*(\alpha)) > 0 \right] \tag{A.18}
$$

where $X_n \sim \mathcal{L}_n$, and $\hat{\sigma}$ is a consistent estimator of the vector $\sigma$ (cf. Assumption 3.1). Exploiting the fact that $g$ is monotone decreasing in $\gamma$ permits us to write

$$
Prob_P \left[ \sum_{(i,j) \in D_0(P)} \delta_n(i,j) \geq 1 \right] \leq Prob_P \left[ g(X_n, \hat{\sigma}, \gamma(\alpha) - \epsilon) > 0, |\gamma(\alpha) - \gamma_n^*(\alpha)| < \epsilon \right]
+ Prob_P \left[ g(X_n, \hat{\sigma}, \gamma_n^*(\alpha)) > 0, |\gamma(\alpha) - \gamma_n^*(\alpha)| \geq \epsilon \right] \tag{A.19}
$$

and

$$
Prob_P \left[ \sum_{(i,j) \in D_0(P)} \delta_n(i,j) \geq 1 \right] \geq Prob_P \left[ g(X_n, \hat{\sigma}, \gamma(\alpha) + \epsilon) > 0, |\gamma(\alpha) - \gamma_n^*(\alpha)| < \epsilon \right]
+ Prob_P \left[ g(X_n, \hat{\sigma}, \gamma_n^*(\alpha)) > 0, |\gamma(\alpha) - \gamma_n^*(\alpha)| \geq \epsilon \right] \tag{A.20}
$$

where $\gamma(\alpha)$ is defined implicitly by the equation $Q(\gamma(\alpha); P) = \alpha$. Consequently, in light of Lemmas A.2 and A.3, we obtain

$$
Q(\gamma(\alpha) + \epsilon; P) \leq \lim_{n \to \infty} Prob_P \left[ \sum_{(i,j) \in D_0(P)} \delta_n(i,j) \geq 1 \right] \leq Q(\gamma(\alpha) - \epsilon; P) \tag{A.21}
$$

upon taking limits of (A.19) and (A.20) as $n \to \infty$. The proof is thus completed upon taking the limit as $\epsilon \to 0$ and recalling that $Q(\cdot; P)$ is continuous with $Q(\gamma(\alpha); P) = \alpha$.

Having established that (A.17) is satisfied with equality when $D_0(P) = K \times K$, it remains to be shown that (A.17) is satisfied when $D_0(P) \subset K \times K$. However, when $D_0(P) \subset K \times K$,
it is the case that

\[ \text{Prob}_P \left[ \sum_{(i,j) \in D_0(P)} \delta_n(i,j) \geq 1 \right] = \text{Prob}_P \left[ \max_{(i,j) \in D_0(P)} \left\{ \max \{ L_{n,i}(\gamma^*_n(\alpha)) - U_{n,j}(\gamma^*_n(\alpha)), L_{n,j}(\gamma^*_n(\alpha)) - U_{n,i}(\gamma^*_n(\alpha)) \} \right\} > 0 \right] \leq \text{Prob}_P \left[ g(X_n, \hat{\sigma}, \gamma^*_n(\alpha)) > 0 \right] \]

The desired result is then established because

\[ \lim_{n \to \infty} \text{Prob}_P \left[ g(X_n, \hat{\sigma}, \gamma^*_n(\alpha)) > 0 \right] = \alpha \]

as was shown above.

\[ \square \]

**Proof of Theorem 3.2 (Part 2: Consistency):** First, we note that

\[ \text{Prob}_P \left[ \prod_{(i,j) \in D_1(P)} \delta_n = 1 \right] = 1 - \text{Prob}_P \left[ \sum_{(i,j) \in D_1(P)} \delta_n(i,j) = 0 \right] \geq 1 - \sum_{(i,j) \in D_1(P)} \text{Prob}_P [\hat{\theta}_{n,i} > \hat{\theta}_{n,j}] - \sum_{(i,j) \in D_1(P)} \text{Prob}_P [C_{n,i} \cap C_{n,j} \neq \emptyset] \]

(A.22)

Establishing the desired result thus requires showing that the last two terms in the second line converge to zero in probability. As for the first of these two terms, we have

\[ \sum_{(i,j) \in D_1(P)} \text{Prob}_P [\hat{\theta}_{n,i} > \hat{\theta}_{n,j}] = \sum_{(i,j) \in D_1(P)} \text{Prob}_P [- \Delta_n(i,j) \leq \sqrt{n}(\theta_j - \theta_i)] \]

with \( \Delta_n(i,j) = \sqrt{n}[(\hat{\theta}_{n,i} - \hat{\theta}_{n,j}) - (\theta_i - \theta_j)] \). Now, \( \Delta_n(i,j) = O_P(1) \) for each \( (i,j) \in D_1(P) \), whereas \( \sqrt{n}(\theta_j - \theta_i) \to \infty \) as \( n \to \infty \). Consequently,

\[ \lim_{n \to \infty} \sum_{(i,j) \in D_1(P)} \text{Prob}_P [- \Delta_n(i,j) \leq \sqrt{n}(\theta_j - \theta_i)] = 0 \]

as desired. It remains to be shown that

\[ \sum_{(i,j) \in D_1(P)} \text{Prob}_P [C_{n,i} \cap C_{n,j} \neq \emptyset] \to 0 \]
as \( n \) tends to \( \infty \). To establish the latter convergence result it suffices to show that each component in the sum tends to zero. Consequently, let \((i, j)\) be an element of \( D_1(P) \) and note that

\[
\text{Prob}_P[C_{n,i} \cap C_{n,j} \neq \emptyset] \leq \text{Prob}_P[\sqrt{n} |\hat{\theta}_{n,i} - \hat{\theta}_{n,j}| < \gamma_n^*(\alpha)(\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j})]
\]  

(A.23)

Upon recalling that \( \gamma_n^*(\alpha) = O_P(1) \) and \( \hat{\sigma}_{n,j} + \hat{\sigma}_{n,j} = O_P(1) \), the desired convergence is established because

\[
\lim_{n \to \infty} \text{Prob}_P[\sqrt{n} |\hat{\theta}_{n,i} - \hat{\theta}_{n,j}| > M] = 1
\]

for all \( M > 0 \).

Proof of Theorem 3.2 (Part 3: Directional Errors): Using the fact that \( 0 \leq \delta_n(j, i) + \delta_n(i, j) \leq 1 \) for every \( i, j \in K \), we have

\[
\text{Prob}_P \left[ \sum_{(i,j) \in D_1(P)} \delta_n(j, i) \geq 1 \right] \leq \text{Prob}_P \left[ \sum_{(i,j) \in D_1(P)} [1 - \delta_n(i, j)] \geq 1 \right]
\]

\[
\leq \text{Prob}_P \left[ \prod_{(i,j) \in D_1(P)} \delta_n(i, j) = 0 \right]
\]

(A.24)

\[
\leq 1 - \text{Prob}_P \left[ \prod_{(i,j) \in D_1(P)} \delta_n(i, j) = 1 \right]
\]

The proof is then complete upon taking limits because \( \text{Prob}_P \left[ \prod_{(i,j) \in D_1(P)} \delta_n(i, j) = 1 \right] \to 1 \) as \( n \to \infty \) is obtained from the previously established consistency of the ranking procedure.

\[\Box\]

A.3 Stepwise Refinement

We now prove Theorem 4.2. The proof of Part 1 is adapted from the proof of Theorem 3.1 of Romano and Wolf (2005b).

Proof of Theorem 4.2 (Part 1: FWE Control): Let \( \alpha \in (0, 1) \) be given. We are required to show that

\[
\lim_{n \to \infty} \text{Prob}_P \left[ \sum_{(i,j) \in D_0(P)} \delta_n^{\text{global}}(i, j) \geq 1 \right] = \alpha,
\]

(A.25)

when \( D_0(P) \) is non-empty subset of \( K \times K \). From Part 2 of Theorem 3.2, we have that, with probability tending to 1, all unequal parameters will be separated in the first step.
of the iterative overlap procedure. Since \( \gamma^*_D(\alpha) \leq \gamma_n(\alpha) \), where \( \gamma^*_D(\alpha) \) is defined analogously to (4.3), the equality in (A.25) then follows from the proof that

\[
\lim_{n \to \infty} \text{Prob}_n \left[ \sqrt{n} \max_{(i,j) \in \mathcal{D}_n} \max \left\{ \Delta_{n,i,j}(\gamma^*_D(\alpha)), \Delta_{n,j,i}(\gamma^*_D(\alpha)) \right\} > 0 \right] = \alpha,
\]

which is similar to the proof of Part 1 of Theorem 3.2 and, hence, omitted.

\[\Box\]

**Proof of Theorem 4.2 (Part 2: Consistency):** Because \( \delta_n(i,j) \geq \delta_n(i,j) \) for all \( n \) and every \( i,j \in K \) (cf. Part 4 below), the desired result follows at once from part (ii) of Theorem 3.2.

\[\Box\]

**Proof of Theorem 4.2 (Part 3: Directional Errors):** Similarly, because \( \delta_n(i,j) \geq \delta_n(i,j) \) for all \( n \) and every \( i,j \in K \) (cf. Part 4 below), the desired result follows at once from part (iii) of Theorem 3.2.

\[\Box\]

**Proof of Theorem 4.2 (Part 4: Weak Refinement):** Let \( n \) and \( i,j \in K \) with \( i \neq j \) be given. From inspection of Algorithm 4.3, it must be the case that either \( \delta_n(i,j) = \delta_n(i,j) \) or \( \delta_n(i,j) = 1 \) and \( \delta_n(i,j) = 0 \). Either way, \( \delta_n(i,j) \) satisfies

\[
\delta_n(i,j) \geq \delta_n(i,j),
\]

and, hence, \( \delta_n(i,j) \) yields a weak refinement of the rankings provided by \( \delta_n(i,j) \).

\[\Box\]

### A.4 Tail Ranking

We begin by introducing some additional notation and collecting some useful results. First, we define the map \( Q_r : \mathbb{R}^+ \to [0,1] \) by

\[
Q_r(\gamma; P) := \text{Prob}_P \left[ \sum_{\ell=1}^{r} \max \{ g_{\ell}(X,\sigma,\gamma),0 \} > 0 \right],
\]

where \( X \) is a \( k \times 1 \) random vector with joint distribution \( \mathcal{L}_\infty \), \( \sigma \) is a \( k \times 1 \) vector (possibly random), and \( g_{\ell} : \mathbb{R}^k \times \mathbb{R}_+^{k+1} \to \mathbb{R} \) is defined by

\[
g_{\ell}(x,y,\gamma) = \min_{1 \leq s \leq \ell} \max_{j \in K} \{ x_j - \gamma y_j \} - \max_{\ell \leq s \leq k} \min_{j \in K} \{ x_j + \gamma y_j \}.
\]
Then, with only slight modifications to the arguments set out in Section A.1, we can show that $Q_r(\gamma; P)$ is continuous and weakly decreasing function satisfying

$$Q_r(0; P) = 1 \text{ and } \lim_{\gamma \to \infty} Q_r(\gamma; P) = 0.$$ 

It is then also straightforward in light of the arguments presented in Section A.1 to prove the consistency of

$$Q_{n,r}(\gamma; \hat{P}_n) := \text{Prob}_{\hat{P}_n}\left[ \sum_{\ell=1}^{r} \max\{ g_\ell(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n), \hat{\sigma}_n^*, \gamma), 0 \} > 0 \right]$$

as an estimator of

$$Q_{n,r}(\gamma; P) := \text{Prob}_P\left[ \sum_{\ell=1}^{r} \max\{ g_\ell(\sqrt{n}(\hat{\theta}_n - \theta), \hat{\sigma}_n, \gamma), 0 \} > 0 \right],$$

and to show that the difference between the quantiles $\gamma_{n,r}^*(\alpha)$ and $\gamma_{n,r}(\alpha)$ tends to zero in probability as the sample size tends to infinity. This brings us to the following useful lemma, whose proof is similar to the proof of Part 1 of Theorem 3.2 and, hence, omitted.

**Lemma A.4.** Suppose that the conditions in Assumption 3.1 are satisfied. Then,

$$\lim_{n \to \infty} Q_{n,r}(\gamma_{n,r}^*(\alpha); P) = \alpha$$

**Proof of Theorem 5.1:** Suppose first that $\theta_1 = \theta_2 = \cdots = \theta_k$. We seek to show that, in this case,

$$\lim_{n \to \infty} \text{Prob}_P\left[ \sum_{\ell=1}^{r} \max\{ \sqrt{n} e_{n,\ell}(\gamma_{n,r}^*(\alpha))d_{n,\ell}, 0 \} > 0 \right] = \alpha \quad (A.29)$$

Since $d_{n,\ell} = 1$ for $n > 1$ and every $1 \leq \ell \leq r$ when the parameters are all equal, we may write

$$\text{Prob}_P\left[ \sum_{\ell=1}^{r} \max\{ \sqrt{n} e_{n,\ell}(\gamma_{n,r}^*(\alpha))d_{n,\ell}, 0 \} > 0 \right] = \text{Prob}_P\left[ \sum_{\ell=1}^{r} \max\{ \sqrt{n} e_{n,\ell}(\gamma_{n,r}^*(\alpha)), 0 \} > 0 \right] = Q_{n,r}(\gamma_{n,r}^*(\alpha; r); P) \quad (A.30)$$

in this case. That equation (A.29) holds then follows upon taking limits and appealing to Lemma A.4. That the left-hand side of (A.29) remains (asymptotically) bounded from
above by $\alpha$ outside of the case $\theta_1 = \cdots = \theta_k$ then follows from the fact that, for every $n > 1$,

$$
Prob_P \left[ \sum_{\ell=1}^{r} \max \{ \sqrt{n} e_{n,\ell}(\gamma_{n,r}^*(\alpha))d_{n,\ell}, 0 \} > 0 \right]
$$

$$
= Prob_P \left[ \sum_{\ell=1}^{r} \max \{ g_{\ell}(\sqrt{n}\hat{\theta}_n, \hat{\sigma}_n, \gamma_{n,r}^*(\alpha))d_{n,\ell}, 0 \} > 0 \right]
$$

$$
\leq Prob_P \left[ \sum_{\ell=1}^{r} \max \{ g_{\ell}(\sqrt{n}(\hat{\theta}_n - \theta), \hat{\sigma}_n, \gamma_{n,r}^*(\alpha))d_{n,\ell}, 0 \} > 0 \right]
$$

(A.31)

$$
\leq Prob_P \left[ \sum_{\ell=1}^{r} \max \{ g_{\ell}(\sqrt{n}(\hat{\theta}_n - \theta), \hat{\sigma}_n, \gamma_{n,r}^*(\alpha))d_{n,\ell}, 0 \} > 0 \right]
$$

$$
= Q_{n,r}(\gamma_{n}^*(\alpha; r); P)
$$

This completes the proof of the first part of Theorem 5.1. The second part of the theorem, namely consistency, is obvious upon writing

$$
Prob_P \left[ L_{n,i}(\gamma_{n}^*(\alpha)) - U_{n,j}(\gamma_{n}^*(\alpha)) > 0 \mid \theta_i > \theta_j \right]
$$

$$
= Prob_P \left[ \sqrt{n}(\hat{\theta}_{n,i} - \hat{\theta}_{n,j}) + \gamma_{n}^*(\alpha)(\hat{\sigma}_{n,j} - \hat{\sigma}_{n,i}) > 0 \mid \theta_i > \theta_j \right]
$$

$$
= Prob_P \left[ \sqrt{n}(\hat{\theta}_{n,i} - \theta_i) - \sqrt{n}(\hat{\theta}_{n,j} - \theta_j) + \gamma_{n}^*(\alpha)(\hat{\sigma}_{n,j} - \hat{\sigma}_{n,i}) > \sqrt{n}(\theta_j - \theta_i) \mid \theta_i > \theta_j \right]
$$

(A.32)

Indeed, since $\sqrt{n}, (\theta_j - \theta_i)$ diverges towards $-\infty$ and

$$
\sqrt{n}(\hat{\theta}_{n,i} - \theta_i) - \sqrt{n}(\hat{\theta}_{n,j} - \theta_j) + \gamma_{n}^*(\alpha)(\hat{\sigma}_{n,j} - \hat{\sigma}_{n,i}) = O_P(1),
$$

it follows that

$$
\lim_{n \to \infty} Prob_P \left[ L_{n,i}(\gamma_{n}^*(\alpha)) - U_{n,j}(\gamma_{n}^*(\alpha)) > 0 \mid \theta_i > \theta_j \right] = 1.
$$

\[\square\]
B Validity of the CI’s Constructed from the Overlap Procedure

This section establishes the validity of the confidence intervals derived from the overlap plot. Towards this end, define

\[ M_{i,j} = \max \left\{ \sqrt{n} (\hat{\theta}_{i,n} - \theta_i - \gamma \hat{\sigma}_{n,i}, \sqrt{n} (\hat{\theta}_{j,n} - \theta_j - \gamma \hat{\sigma}_{n,j}) \right\} \]

and

\[ m_{i,j} = \min \left\{ \sqrt{n} (\hat{\theta}_{i,n} - \theta_i + \gamma \hat{\sigma}_{n,i}, \sqrt{n} (\hat{\theta}_{j,n} - \theta_j + \gamma \hat{\sigma}_{n,j}) \right\}. \]

Then, using

\[ M_{i,j} - m_{i,j} = \sqrt{n} |(\hat{\theta}_{i,n} - \hat{\theta}_{j,n}) - (\theta_i - \theta_j)| - \gamma (\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}), \tag{B.1} \]

we obtain the chain of inequalities

\[ Q_n(\gamma, P) = \text{Prob} \left[ \max_{j \in K} \{ \sqrt{n} (\hat{\theta}_{n,j} - \theta_j - \gamma \hat{\sigma}_{n,j}) \} - \min_{j \in K} \{ \sqrt{n} (\hat{\theta}_{n,j} - \theta_j + \gamma \hat{\sigma}_{n,j}) \} \geq 0 \right] \]

\[ \geq \text{Prob} [M_{i,j} - m_{i,j} \geq 0] \]

\[ = \text{Prob} \left[ (\hat{\theta}_{i,n} - \hat{\theta}_{j,n}) - (\theta_i - \theta_j) \geq \gamma \left( \frac{\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}}{\sqrt{n}} \right) \right] \]

\[ = 1 - \text{Prob} \left[ (\hat{\theta}_{i,n} - \hat{\theta}_{j,n}) - \gamma \left( \frac{\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}}{\sqrt{n}} \right) \leq \theta_i - \theta_j \leq (\hat{\theta}_{i,n} - \hat{\theta}_{j,n}) + \gamma \left( \frac{\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}}{\sqrt{n}} \right) \right] \]

\[ \tag{B.2} \]

The inequality in B.2 shows that

\[ \lim_{n \to \infty} \text{Prob} \left[ (\hat{\theta}_{i,n} - \hat{\theta}_{j,n}) - \gamma \left( \frac{\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}}{\sqrt{n}} \right) \leq \theta_i - \theta_j \leq (\hat{\theta}_{i,n} - \hat{\theta}_{j,n}) + \gamma \left( \frac{\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}}{\sqrt{n}} \right) \right] \geq 1 - \alpha, \]

provided that \( \lim_{n \to \infty} Q_n(\gamma, P) = \alpha \). In other words, B.2 shows that the confidence interval

\[ C_n(i,j) = \left[ (\hat{\theta}_{i,n} - \hat{\theta}_{j,n}) - \gamma \left( \frac{\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}}{\sqrt{n}} \right), (\hat{\theta}_{i,n} - \hat{\theta}_{j,n}) + \gamma \left( \frac{\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}}{\sqrt{n}} \right) \right] \]

is guaranteed to have asymptotic coverage probability at least \( 1 - \alpha \) whenever the overlap procedure maintains asymptotic control over the FWE at the nominal level \( \alpha \).
C  Mutually Independent Unbalanced Samples

This section outlines how the resampling-based overlap procedure can be suitably modified to treat mutually independent unbalanced samples.

To begin, we introduce the following notation. First, for $1 \leq i \leq k$, let $\hat{\theta}_{n,i}$ denote an estimator of $\theta_i$ based on $n_i$ observations from the $i$th sample. Similarly, let $\hat{\sigma}_{n,i}$ be an estimator of the variance of $\sqrt{n_i} \hat{\theta}_{n,i}$ that is based on the same $n_i$ observations. Lastly, we define

$$\lambda_{in} \equiv \lambda_i(n_1, \ldots, n_k) = \frac{\prod_{j \neq i} n_j}{\sum_j \left( \prod_{l \neq j} n_l \right)},$$

so that when $k = 2$, for example, $\lambda_{1n} = n_2/(n_1 + n_2)$ and $\lambda_{2n} = n_1/(n_1 + n_2)$.

In light of these definitions, natural analogues of $Q_n(\gamma; P)$ and $Q_n(\gamma; \hat{P}_n)$ (cf. equations 3.3 and 3.5) are

$$Q_n(\gamma, P) \equiv \text{Prob}_{\hat{P}} \left[ \max_{i \in K} \left\{ \lambda_{in}^{1/2} \left[ \sqrt{n_i} (\hat{\theta}_{n,i} - \theta_0) - \gamma \hat{\sigma}_{n,i} \right] \right\} - \min_{i \in K} \left\{ \lambda_{in}^{1/2} \left[ \sqrt{n_i} (\hat{\theta}_{n,i} - \theta_0) + \gamma \hat{\sigma}_{n,i} \right] \right\} \geq 0 \right],$$

and

$$Q_n(\gamma, \hat{P}_n) \equiv \text{Prob}_{\hat{P}_n} \left[ \max_{i \in K} \left\{ \lambda_{in}^{1/2} \left[ \sqrt{n_i} (\hat{\theta}_{n,i}^* - \hat{\theta}_{n,i}) - \gamma \hat{\sigma}_{n,i}^* \right] \right\} - \min_{i \in K} \left\{ \lambda_{in}^{1/2} \left[ \sqrt{n_i} (\hat{\theta}_{n,i}^* - \hat{\theta}_{n,i}) + \gamma \hat{\sigma}_{n,i}^* \right] \right\} \geq 0 \right],$$

with the corresponding feasible estimator of $\gamma$ defined accordingly as

$$\gamma_n^*(\alpha) = \inf \left\{ \gamma : Q_n(\gamma; \hat{P}_n) \leq \alpha \right\}. \quad (C.1)$$

The following theorem shows that the overlap procedure when modified for unbalanced samples retains the asymptotic properties as in the balanced case.

**Theorem C.1.** Suppose that the conditions in Assumption 3.1 are satisfied and that the sampling process satisfies

$$\epsilon < \lambda_i(n_1, \ldots, n_k) < 1 - \epsilon$$

for some (small) $\epsilon > 0$ and every $1 \leq i \leq k$. Then, the decision rule $\delta_n$ as defined in (3.2) with $\gamma = \gamma_n^*(\alpha)$ satisfies conditions (1)-(3) of Theorem 3.2.

The proof of Theorem C.1 is analogous to the proof of Theorem 3.2 and is therefore omitted.