Graphical Procedures for Multiple Comparisons Under General Dependence

Christopher J. Bennett[a] and Brennan S. Thompson[b]

May, 2012

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 revisit Tukey's graphical procedure and show that it can be modified to allow for general forms of dependence across samples. We show that this modification of Tukey's overlap procedure maintains asymptotic control of the familywise error rate under general conditions, and we present simulation results that demonstrate good control of the familywise error rate in finite samples. Two empirical applications highlight the utility of this new procedure: first, the overlap procedure is applied to U.S. income data to generate a ranking of poverty by geographic region; and second, the overlap procedure is applied to airline transportation data to produce a ranking of mean daily average arrival delay times for major U.S. airlines.

Keywords: bootstrap; familywise error rate; overlap

[a](Corresponding author) Department of Economics, Vanderbilt University, VU Station B #351819, 2301 Vanderbilt Place, Nashville, TN 37235-1819, U.S.A. E-mail: **chris.bennett@vanderbilt.edu**

[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, he 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 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.

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 corresponding decision rule for the procedure. The large sample properties of the overlap procedure are also examined here. A refinement of the overlap procedure along the lines of Holms' (1979) improvement over the Bonferroni correction is then taken up in Section 4. Section 5 presents simulation results, followed by two empirical illustrations in Section 6. Lastly, Section 7 concludes.

2 Problem Formulation

We seek to rank a collection of $k > 2$ 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 θ_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 in Appendix C and also illustrated in Section 6).

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. Towards this end, we define the sets

$$D_0(P) = \{(i,j) \in K^2 : i < j, \theta_i(P) = \theta_j(P)\}, \tag{2.1}$$

and

$$D_1(P) = \{(i,j) \in K^2 : \theta_i(P) < \theta_j(P)\}. \tag{2.2}$$

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 $(i,j) \in K^2$ 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 α will require that our decision rule δ_n satisfy

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

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

$$C_{n,i}^{\mathrm{T}} = \left[\hat{\theta}_{n,i} \pm Q_{k,\nu}^{\alpha} \frac{S_\nu}{2\sqrt{n}}\right], \tag{2.4}$$

where S_ν is the pooled estimator of σ 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. ν; see, e.g., Hochberg and Tamhane (1987, p. 31). In this context, Tukey showed that

$$\delta_n^{\mathrm{T}}(i,j) = \begin{cases} 1, & \text{if } \hat{\theta}_{n,i} < \hat{\theta}_{n,j} \text{ and } C_{n,i}^{\mathrm{T}} \cap C_{n,j}^{\mathrm{T}} = \emptyset \\ 0, & \text{otherwise,} \end{cases} \tag{2.5}$$

satisfies (2.3) for all n. In other words, Tukey's graphical decision rule that ranks θ_i below θ_j whenever $\hat{\theta}_{n,i}$ plots below $\hat{\theta}_{n,j}$ and the uncertainty intervals $C_{n,i}^{\mathrm{T}}$ and $C_{n,j}^{\mathrm{T}}$ do not overlap achieves control of the familywise rate at the nominal level α, 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, for example, 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,

$$C_{n,i}(\gamma) = \left[\hat{\theta}_{n,i} \pm \gamma \frac{\hat{\sigma}_{n,i}}{\sqrt{n}}\right], \tag{3.1}$$

whose length is determined by the parameter γ, 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 $\hat{\theta}_i < \hat{\theta}_j$ and the uncertainty intervals for θ_i and θ_j are non-overlapping. That is, the overlap procedure records

$$\delta_n(i,j) = \begin{cases} 1, & \text{if } \hat{\theta}_i < \hat{\theta}_j \text{ and } C_{n,i}(\gamma) \cap C_{n,j}(\gamma) = \emptyset \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

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 simultaneous confidence sets for the pairwise differences and then apportion the margin of error to the individual parameter estimates, this new overlap procedure involves a direct approach for estimating the half-lengths via the parameter γ. In particular, in order to maintain the desired control over the familywise error rate, our strategy for choosing γ is based on directly estimating the probability of observing a pair of non-overlapping uncertainty intervals as a function of γ when all of the parameters are equal.

Towards formally describing this procedure, first note that, as a function of γ, the probability of declaring at least one significant difference when all k parameters are

equal is given by

$$Q_n(\gamma; P) := Prob_P\Big[\max_{i\in K}\{\sqrt{n}(\hat{\theta}_{n,i} - \theta_i) - \gamma\hat{\sigma}_{n,i}\} - \min_{i\in K}\{\sqrt{n}\,(\hat{\theta}_{n,i} - \theta_i) + \gamma\hat{\sigma}_{n,i}\} > 0\Big]. \quad (3.3)$$

Equation (3.3) shows that the "ideal" choice of γ when all k parameters are equal is

$$\gamma_n(\alpha) = \inf\Big\{\gamma : Q_n(\gamma; P) \leq \alpha\Big\}. \quad (3.4)$$

Note that $\gamma_n(\alpha)$ is ideal in the sense that, if obtained, we could 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 dependence; see Lahiri (2003). Our estimator of $Q_n(\gamma; P)$ is then given by

$$Q_n(\gamma; \hat{P}_n) = Prob_{\hat{P}_n}\Big[\max_{i\in K}\{\sqrt{n}\,(\hat{\theta}^*_{n,i} - \hat{\theta}_{n,i}) - \gamma\hat{\sigma}^*_{n,i}\} - \min_{i\in K}\{\sqrt{n}\,(\hat{\theta}^*_{n,i} - \hat{\theta}_{n,i}) + \gamma\hat{\sigma}^*_{n,i}\} > 0\Big], \quad (3.5)$$

which, in turn, gives rise to our estimator of γ,

$$\gamma^*_n(\alpha) = \inf\Big\{\gamma : Q_n(\gamma; \hat{P}_n) \leq \alpha\Big\}. \quad (3.6)$$

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

$$\mathcal{L}_n(\mathbf{x}) = Prob_P\big[\sqrt{n}\,(\hat{\theta}_{n,i} - \theta_i) \leq x_i,\ 1 \leq i \leq k\big]$$

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

$$\mathcal{L}^*_n(\mathbf{x}) = Prob_{\hat{P}_n}\big[\sqrt{n}\,(\hat{\theta}^*_{n,i} - \hat{\theta}_{n,i}) \leq x_i,\ 1 \leq i \leq k\big]$$

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 σ_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 $\mathcal{L}_\infty$, 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 δ_n as defined in (3.2) with $\gamma = \gamma_n^*(\alpha)$ satisfies*

1. *Error Rate Control (Type I)*

$$\lim_{n\to\infty} Prob_P\left[\sum_{(i,j)\in D_0(P)} \delta_n(i,j) \geq 1\right] \leq \alpha \tag{3.7}$$

2. *Consistency*

$$\lim_{n\to\infty} Prob_P\left[\prod_{(i,j)\in D_1(P)} \delta_n(i,j) = 1\right] = 1 \tag{3.8}$$

3. *Control of Directional Errors*

$$\lim_{n\to\infty} Prob_P\left[\sum_{(i,j)\in D_1(P)} \delta_n(j,i) \geq 1\right] = 0 \tag{3.9}$$

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 pre-specified nominal level α, at least asymptotically. We show in the proof of Theorem 3.2 (see Appendix A.2) that (3.7) 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.

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.,

$$C_{n,(i,j)} = \Big[(\hat{\theta}_{n,i} - \hat{\theta}_{n,j}) \pm \gamma\Big(\frac{\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}}{\sqrt{n}}\Big)\Big]. \tag{3.10}$$

We show in Appendix B that $C_{n,(i,j)}$ is indeed guaranteed to have asymptotic coverage probability for $\theta_i - \theta_j$ at least $1-\alpha$ whenever the overlap procedure maintains asymptotic control over the familywise error rate at the nominal level α. The simplicity of this construction in practical settings is also illustrated in Section 6.1.

4 Refinements of the Initial Ranking

Our focus in this section is on situations in which the ranking remains incomplete after an initial application of the proposed overlap procedure. In particular, we investigate the potential for an iterative stepwise procedure to generate refinements of an incomplete ranking while still maintaining control over the desired error rate.

In order to facilitate our discussion of a possible refinement strategy, we define the γ-equivalence class of an element $i \in K$ as the set

$$[i]_\gamma = \{j \in K : C_{n,i}(\gamma) \cap C_{n,j}(\gamma) \neq \emptyset\}, \tag{4.1}$$

and we denote by $\mathcal{K}(\gamma)$ the partition of K formed by the set of all γ-equivalence classes.

Setting $\gamma_{1,n} = \gamma_n^*(\alpha)$ gives rise to the the partition $\mathcal{K}(\gamma_{1,n})$ that is generated by an initial application of the overlap procedure. Of particular interest is the number of elements in $\mathcal{K}(\gamma_{1,n})$. For example, $\mathcal{K}(\gamma_{1,n})$ contains a single element, namely the

set K itself, when the overlap procedure fails to rank any two parameters. At the opposite extreme, in which the overlap procedure produces a complete ranking of the parameters under consideration, $\mathcal{K}(\gamma_{1,n})$ contains k singletons.

The basic idea behind the refinement strategy involves updating the set of equivalence classes generated by an application of the overlap procedure and using the updated equivalence class to re-estimate the probability of committing a Type I error as a function of γ. Concretely, at the second step, we compute this probability using

$$Q_n(\gamma; \hat{P}_n, \mathcal{K}(\gamma_{1,n})) = Prob_{\hat{P}_n}\left[\max_{A\in\mathcal{K}}\left\{\max_{i\in A}\{\sqrt{n}\,(\hat{\theta}^*_{n,i}-\hat{\theta}_{n,i})-\gamma\hat{\sigma}^*_{n,i}\}-\min_{i\in A}\{\sqrt{n}\,(\hat{\theta}^*_{n,i}-\hat{\theta}_{n,i})+\gamma\hat{\sigma}^*_{n,i}\}\right\} > 0\right] \tag{4.2}$$

Thus, $Q_n(\gamma, \hat{P}_n; \mathcal{K})$ is defined so that the maximum difference for each of the sets in $\mathcal{K}$ is recorded first, with the overall maximum subsequently computed as the largest of these individual maximums.

Notice that $Q_n(\gamma; \hat{P}_n)$ (c.f. Equation 3.5) is recovered from (4.2) after replacing $\mathcal{K}(\gamma_{1,n})$ with $\mathcal{K}_0 = \{K\}$. Moreover, notice that $|\mathcal{K}(\gamma_{1,n})| \geq |\mathcal{K}_0|$ and, hence,

$$Q_n(\gamma, \hat{P}_n; \mathcal{K}(\gamma_{1,n})) \geq Q_n(\gamma, \hat{P}_n; \mathcal{K}_0). \tag{4.3}$$

The importance of the inequality in (4.3) stems from the fact that

$$\gamma_{2,n}(\alpha; \mathcal{K}(\gamma_{1,n})) := \inf\left\{\gamma : Q_n(\gamma; \hat{P}_n, \mathcal{K}(\gamma_{1,n})) \leq \alpha\right\}, \tag{4.4}$$

can be no larger than $\gamma_{1,n}(\alpha, \mathcal{K}_0)$. Consequently, using $\gamma_{2,n}$ in a second application of the overlap procedure gives rise to narrower uncertainty intervals $C_{n,i}(\gamma_{2,n})$, $i \in K$, 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 δ_n^{global}.

Algorithm 4.1.

1. Set $\gamma_{1,n} = \gamma_n(\alpha; \mathcal{K}_0)$ and $\delta_n^{\text{global}}(i,j) = \delta_n(i,j)$.

2. If $|\mathcal{K}_0| < |\mathcal{K}_{1,n}| < k$, then compute $\gamma_{2,n} = \gamma_{2,n}(\alpha; \mathcal{K}_{1,n})$ and set

$$\delta_n^{\text{global}}(i,j) = \begin{cases} 1 & \text{if } C_{n,i}(\gamma_{2,n}) \cap C_{n,j}(\gamma_{2,n}) = \emptyset \text{ and } \hat{\theta}_{n,i} > \hat{\theta}_{n,j} \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

   Else, stop.

3. If $|\mathcal{K}_{1,n}| < |\mathcal{K}_{2,n}| < k$, then compute $\gamma_{3,n} = \gamma_{3,n}(\alpha, \mathcal{K}_{2,n})$ and set

$$\delta_n^{\text{global}}(i,j) = \begin{cases} 1 & \text{if } C_{n,i}(\gamma_{3,n}) \cap C_{n,j}(\gamma_{3,n}) = \emptyset \text{ and } \hat{\theta}_{n,i} > \hat{\theta}_{n,j} \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

   Else, stop.

   $\vdots$

The large sample properties of the modified procedure that is obtained through the application of this algorithm are outlined in Theorem 4.2 below:

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* δ_n^{global} *defined in Algorithm 4.1 satisfies*

$$\lim_{n\to\infty} Prob_P \left[\sum_{(i,j)\in D_0(P)} \delta_n^{\text{global}}(i,j) \geq 1 \right] = \alpha, \tag{4.7}$$

whenever $D_0(P)$ *in non-empty. Additionally,* δ_n^{global} *satisfies*

$$\delta_n^{\text{global}}(i,j) \geq \delta_n(i,j) \text{ for every } (i,j) \in K^2. \tag{4.8}$$

Equation (4.7) in Theorem 4.2 shows that the limiting familywise error rate of the global refinement procedure is exactly equal to α 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 α only if *all* of the parameters under consideration are equal, and is therefore conservative whenever $D_0(P)$ is a strict subset of K^2.

Equation (4.8) together with equation (4.7) 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 α. 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 Simulation Study

In this section, we examine the finite-sample performance of the overlap procedure via a simulation study focused on ranking $k = 10$ means. Specifically, we generate n i.i.d. random vectors from a 10-dimensional multivariate normal distribution with mean vector $(\theta_1, \ldots, \theta_{10})'$, and covariance matrix given by a Toeplitz matrix whose first row is $(\sigma_i^2, \rho, \rho, \rho, \rho, \rho, \rho, \rho, \rho, \rho)$. Temporarily fixing $\sigma_i^2 = 1$ for $i = 1, \ldots, 10$, we examine the following three cases for samples sizes of $n = 100$, 500, and 1,000, and

Table 1: Empirical FWEs and Proportion of Strict Orderings Inferred: Case 1

		FWE			Proportion of Strict Orderings Inferred		
		max-t			max-t		
ρ	n	Basic	Stud	Overlap	Basic	Stud	Overlap
-0.1	100	0.049	0.049	0.047	-	-	-
	500	0.051	0.051	0.050	-	-	-
	1000	0.049	0.050	0.049	-	-	-
0	100	0.049	0.050	0.047	-	-	-
	500	0.050	0.051	0.049	-	-	-
	1000	0.051	0.051	0.050	-	-	-
0.1	100	0.049	0.050	0.047	-	-	-
	500	0.050	0.051	0.049	-	-	-
	1000	0.050	0.051	0.050	-	-	-
0.5	100	0.049	0.050	0.043	-	-	-
	500	0.049	0.050	0.047	-	-	-
	1000	0.051	0.050	0.050	-	-	-

NOTE: $\theta_i = 0$ and $\sigma_i^2 = 1$, for $i = 1, \ldots, 10$. The nominal level is 5%.

Table 2: Empirical FWEs and Proportion of Strict Orderings Inferred: Case 2

		FWE			Proportion of Strict Orderings Inferred		
		max-t			max-t		
ρ	n	Basic	Stud	Overlap	Basic	Stud	Overlap
-0.1	100	0.013	0.013	0.012	0.170	0.160	0.164
	500	0.014	0.014	0.014	0.610	0.608	0.608
	1000	0.013	0.014	0.013	0.751	0.750	0.750
0	100	0.014	0.014	0.014	0.193	0.182	0.185
	500	0.014	0.014	0.014	0.631	0.629	0.630
	1000	0.015	0.015	0.014	0.767	0.766	0.766
0.1	100	0.014	0.014	0.013	0.221	0.208	0.211
	500	0.013	0.014	0.013	0.654	0.652	0.652
	1000	0.014	0.014	0.014	0.783	0.782	0.782
0.5	100	0.014	0.014	0.012	0.391	0.377	0.378
	500	0.014	0.014	0.014	0.766	0.765	0.765
	1000	0.014	0.013	0.013	0.872	0.871	0.871

NOTE: $\theta_i = 0$, for $i = 1, \ldots, 5$, $\theta_i = i/10$, for $i = 6, \ldots, 10$, and $\sigma_i^2 = 1$ for $i = 1, \ldots, 10$. The nominal level is 5%.

Table 3: Empirical FWEs and Proportion of Strict Orderings Inferred: Case 3

		FWE			Proportion of Strict Orderings Inferred		
		max-t			max-t		
ρ	n	Basic	Stud	Overlap	Basic	Stud	Overlap
-0.1	100	0.0005	0.0006	0.0005	0.333	0.321	0.326
	500	0.0000	0.0000	0.0000	0.696	0.695	0.695
	1000	0.0000	0.0000	0.0000	0.806	0.805	0.805
0	100	0.0006	0.0006	0.0005	0.357	0.345	0.349
	500	0.0000	0.0000	0.0000	0.713	0.712	0.712
	1000	0.0000	0.0000	0.0000	0.819	0.818	0.818
0.1	100	0.0004	0.0004	0.0004	0.384	0.372	0.375
	500	0.0000	0.0000	0.0000	0.731	0.730	0.730
	1000	0.0000	0.0000	0.0000	0.831	0.831	0.830
0.5	100	0.0001	0.0002	0.0001	0.526	0.515	0.515
	500	0.0000	0.0000	0.0000	0.818	0.817	0.817
	1000	0.0000	0.0000	0.0000	0.900	0.900	0.899

NOTE: $\theta_i = i/10$ and $\sigma_i^2 = 1$, for $i = 1, \ldots, 10$. The nominal level is 5%.

Table 4: Empirical FWEs and Proportion of Strict Orderings Inferred: Case 1a

		FWE			Proportion of Strict Orderings Inferred		
		max-t			max-t		
ρ	n	Basic	Stud	Overlap	Basic	Stud	Overlap
-0.1	100	0.053	0.049	0.047	-	-	-
	500	0.051	0.050	0.049	-	-	-
	1000	0.051	0.050	0.049	-	-	-
0	100	0.052	0.049	0.045	-	-	-
	500	0.051	0.052	0.051	-	-	-
	1000	0.051	0.050	0.049	-	-	-
0.1	100	0.052	0.049	0.046	-	-	-
	500	0.051	0.052	0.051	-	-	-
	1000	0.051	0.051	0.051	-	-	-
0.5	100	0.053	0.051	0.046	-	-	-
	500	0.050	0.051	0.048	-	-	-
	1000	0.052	0.050	0.050	-	-	-

NOTE: $\theta_i = i$, for $i = 1, \ldots, 10$, $\sigma_1^2 = 5$, and $\sigma_i^2 = 1$, for $i = 2, \ldots, 10$. The nominal level is 5%.

Table 5: Empirical FWEs and Proportion of Strict Orderings Inferred: Case 2a

		FWE			Proportion of Strict Orderings Inferred		
		max-t			max-t		
ρ	n	Basic	Stud	Overlap	Basic	Stud	Overlap
-0.1	100	0.035	0.014	0.015	0.059	0.140	0.142
	500	0.034	0.013	0.014	0.467	0.570	0.571
	1000	0.033	0.014	0.015	0.641	0.723	0.723
0	100	0.035	0.014	0.015	0.063	0.159	0.158
	500	0.034	0.013	0.015	0.479	0.590	0.591
	1000	0.034	0.013	0.016	0.651	0.738	0.739
0.1	100	0.037	0.014	0.016	0.068	0.181	0.177
	500	0.035	0.013	0.017	0.492	0.612	0.613
	1000	0.034	0.013	0.016	0.662	0.754	0.754
0.5	100	0.039	0.013	0.025	0.093	0.325	0.288
	500	0.037	0.013	0.028	0.547	0.718	0.706
	1000	0.037	0.013	0.029	0.711	0.838	0.821

NOTE: $\theta_i = 0$, for $i = 1, \ldots, 5$, $\theta_i = i/10$, for $i = 6, \ldots, 10$, $\sigma_1^2 = 5$, and $\sigma_i^2 = 1$, for $i = 2, \ldots, 10$. The nominal level is 5%.

Table 6: Empirical FWEs and Proportion of Strict Orderings Inferred: Case 3a

		FWE			Proportion of Strict Orderings Inferred		
		max-t			max-t		
ρ	n	Basic	Stud	Overlap	Basic	Stud	Overlap
-0.1	100	0.0024	0.0009	0.0009	0.189	0.262	0.268
	500	0.0004	0.0000	0.0000	0.586	0.665	0.666
	1000	0.0001	0.0000	0.0000	0.721	0.785	0.785
0	100	0.0025	0.0007	0.0008	0.199	0.283	0.289
	500	0.0004	0.0001	0.0001	0.595	0.681	0.682
	1000	0.0001	0.0000	0.0000	0.728	0.796	0.797
0.1	100	0.0023	0.0006	0.0007	0.209	0.306	0.312
	500	0.0004	0.0000	0.0001	0.605	0.698	0.699
	1000	0.0001	0.0000	0.0000	0.737	0.809	0.809
0.5	100	0.0030	0.0003	0.0017	0.260	0.435	0.425
	500	0.0004	0.0000	0.0002	0.648	0.781	0.771
	1000	0.0001	0.0000	0.0000	0.772	0.872	0.861

NOTE: $\theta_i = i/10$, for $i = 1, \ldots, 10$, $\sigma_1^2 = 5$, and $\sigma_i^2 = 1$, for $i = 2, \ldots, 10$. The nominal level is 5%.

correlation parameter values of $\rho = -0.1$, 0, 0.1, and 0.5:

Case 1: $\theta_i = 0$, for $i = 1, \ldots, 10$.

Case 2: $\theta_i = 0$, for $i = 1, \ldots, 5$, and $\theta_i = i/10$, for $i = 6, \ldots, 10$.

Case 3: $\theta_i = i/10$, for $i = 1, \ldots, 10$.

In each of these cases, there are $10(10-1)/2 = 45$ unique pairwise comparisons. In Case 1, no unique pairs of means are strictly ordered. In Case 2, 35 unique pairs of means are strictly ordered. In Case 3, all 45 unique pairs of means are strictly ordered.

In Tables 1-3, we report the empirical familywise error rates (FWEs) and the proportion of strict orderings identified (i.e., the proportion of parameter pairs $(i, j) \in D_1(P)$ that are correctly identified) for these three cases. The nominal level is $\alpha = 0.05$, and we use $B = 1,999$ bootstrap replications. The number of Monte Carlo replications is 100,000. As a point of comparison, we also report results for a single-step max-t (both basic and studentized) test of pairwise differences in means. Calculations of the critical values for these tests are obtained using bootstrap methods as described in Romano and Wolf (2005a, pp. 102-103).

From Table 1, it is clear that control of the familywise error rate is satisfactory for all three procedures. In terms of the proportion of strict orderings identified, Tables 2 and 3 show that (i) all of the procedures offer virtually identical performance; and (ii) when the correlation level increases, the proportion of strict orderings identified by each of the procedures is uniformly higher (for any n).

In order to examine the effect of non-identical marginal variances, we also consider the same three cases (referred to as Cases 1a, 2a, and 3a, respectively), now with $\sigma_1^2 = 5$ and $\sigma_i^2 = 1$ for $i = 2, \ldots, 10$. For these three cases, the empirical FWEs

and the proportion of strict orderings correctly identified are reported in Tables 4-6. It is again evident that control of the familywise error rate is satisfactory for all three procedures. In terms of the proportion of strict orderings identified, however, the basic (non-studentized) max-t procedure performs significantly worse than the studentized version and the overlap procedure, with the latter two tending to perform quite similarly.

6 Empirical Illustrations

In this section, we consider two distinct empirical illustrations of the overlap procedure: (i) ranking U.S. geographic regions by poverty incidence, and (ii) ranking major U.S. airlines by mean daily average arrival delay time. The first serves to illustrate the application of the overlap procedure to mutually independent unbalanced samples, whereas the latter serves to illustrate the application of the basic overlap procedure and its stepwise refinements to a setting with dependent samples. In both of these illustrations, we set $\alpha = 0.05$, and use $B = 9,999$ bootstrap replications in constructing the uncertainty intervals.

6.1 Poverty Rates for U.S. Regions

In this illustration, we present a ranking of the $k = 4$ U.S. census regions (Northeast, Midwest, West and South) based on the proportion of their population that is poor (using the U.S. Census Bureau poverty thresholds). Data for 2009 is obtained from the 2010 March Supplement of the Current Population Survey, conducted by the U.S. Census Bureau for the Bureau of Labour Statistics (this dataset is available at

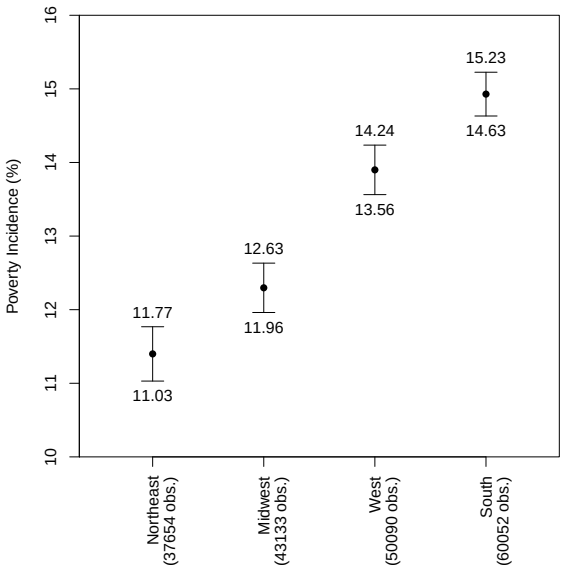


Figure 1: Poverty incidence in U.S. regions

http://www.census.gov/cps/).[1] There are a total of 190,929 individual respondents to the survey; the number of observations for each region is given in brackets beside its label in Figure 1.

In Figure 1, we plot point estimates of the proportion of poor individuals by region (represented by solid dots), which range from 11.4 percent (Northeast) to 14.9 percent (South), along with uncertainty intervals constructed using the overlap procedure (represented by solid "whiskers"; the values printed above and below these "whiskers" are the upper and lower endpoints of the intervals, respectively). As none of the uncertainty intervals overlap, we are able to obtain a complete ranking of the regions. Specifically, we find that the Northeast has the lowest level of poverty, the Midwest has second lowest level, the West has the third lowest level, and the South has the fourth lowest (i.e., highest) level.

Beyond conveying this ordinal ranking, the graphical representation of the overlap

[1]For the sake of this illustration, we ignore the complex design of the survey. We do, however, use the sampling weights provided and adjust our asymptotic variance estimates accordingly.

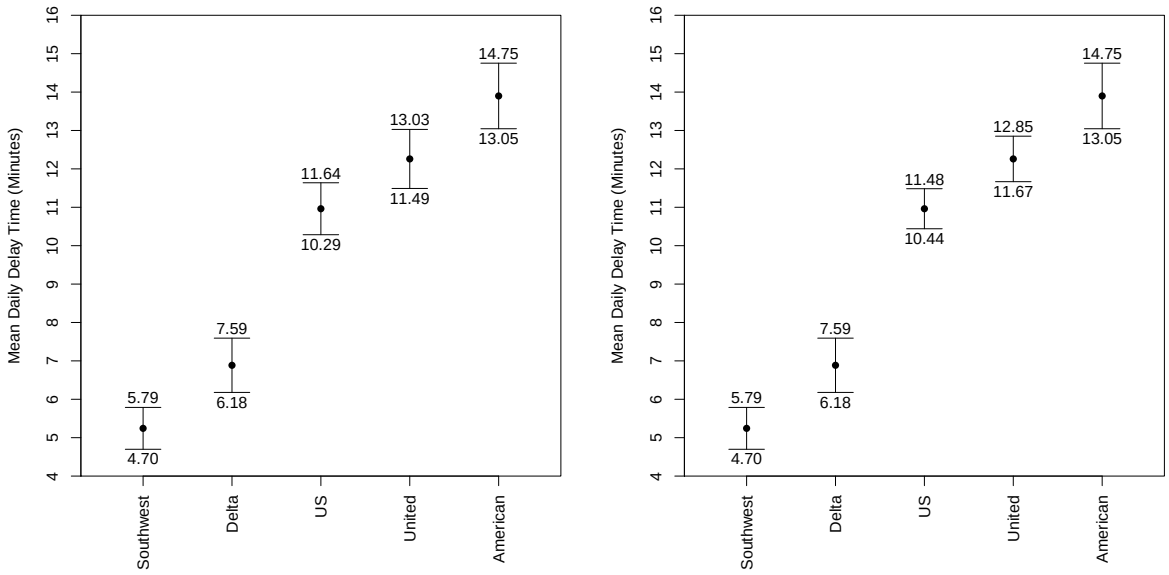
procedure also conveys the level of uncertainty for the location of the individual parameters. Moreover, we can use this information to construct asymptotically valid confidence intervals for the differences between parameters. Specifically, a confidence interval for the difference between two parameters is centered at the difference of the point estimates (which are the midpoints of the uncertainty intervals), with a margin of error equal to the average length of the two uncertainty intervals. For example, an asymptotically valid 95% confidence interval for the difference in poverty levels between the Midwest and the Northeast is given by

$$\left[12.3 - 11.4 \pm \frac{(12.63 - 11.96) + (11.77 - 11.03)}{2}\right] = \left[0.9 \pm 0.705\right].$$

6.2 Delay Times for U.S. Airlines

In this illustration, we present a ranking of the $k = 5$ busiest U.S. commercial airlines (American, Delta, Southwest, United, and US) by the mean of their daily average arrival delay times. Data for the entire year of 2007 ($n = 365$) are obtained from the Research and Innovative Technology Administration, which coordinates the U.S. Department of Transportation research program (this dataset is available at http://www.transtats.bts.gov/). This data has recently been analyzed by Wicklin (2011), among others. It is useful to note that the cross-sample correlations range from 0.43 to 0.66, with a median value of 0.55.

In Figure 2a, we provide point estimates of the mean daily average arrival delay time for the 5 airlines, which range from 5.24 minutes (Southwest) to 13.90 minutes (American), along with uncertainty intervals constructed using the overlap procedure (here, we obtain $\hat{\gamma}^*_{1,n} = 1.32$). As the intervals for the third and fourth best airlines (US and United) overlap, we apply Algorithm 4.1 to refine this ranking. In doing



(a) Initial Ranking (b) Refined Ranking

Figure 2: Mean daily average arrival delays

so, we obtain $\gamma^*_{2,n} = 1.02$, which implies substantially shorter intervals for these two airlines, that no longer overlap (see Figure 2b). Thus, application of the refinement procedure produces a complete ranking, with Southwest first best, Delta second best, US third best, United fourth best, and American fifth best (i.e., worst).

7 Concluding Remarks

We have shown that Tukey's (1953) overlap procedure extends in a natural way to settings in which there are general forms of dependence across samples. We have also demonstrated that these new procedures maintain asymptotic control of the classical familywise error rate, and exhibited their performance in simulations and two empirical illustrations. Future work will focus on extending our procedure to provide control of some generalized notions of the familywise error rate, thereby building on Tukey's own suggestion (see, e.g., Basford and Tukey, 1999) that one

relax the familywise error rate when the number of comparisons becomes large.

References

Andrews, H. P., Snee, R. D., and Sarner, M. H. (1980). Graphical display of means. *The American Statistician*, 34(4):195–199.

Athreya, K. and Lahiri, S. (2006). *Measure Theory And Probability Theory*. Springer, New York.

Basford, K. E. and Tukey, J. W. (1999). *Graphical Analysis of Multiresponse Data: Illustrated with a Plant Breeding Trial*. Chapman & Hall/CRC, Boca Raton.

Beran, R. (1984). Bootstrap methods in statistics. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 86(1):14–30.

Gabriel, K. R. (1978). A simple method of multiple comparisons of means. *Journal of the American Statistical Association*, 73(364):724–729.

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

Hochberg, Y., Weiss, G., and Hart, S. (1982). On graphical procedures for multiple comparisons. *Journal of the American Statistical Association*, 77(380):767–772.

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

Hsu, J. C. and Peruggia, M. (1994). Graphical representations of Tukey's multiple comparison method. *Journal of Computational and Graphical Statistics*, 3(2):143–161.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York.

Romano, J. P. and Wolf, M. (2005a). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Romano, J. P. and Wolf, M. (2005b). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.

Tukey, J. W. (1953). The problem of multiple comparisons. In *The Collected Works of John H. Tukey VIII. Multiple Comparisons: 1948-1983*, pages 1–300. Chapman and Hall, New York.

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

Wicklin, R. (2011). Visualizing airline delays and cancelations. *Journal of Computational and Graphical Statistics*, 20(4):284–286.

A Technical Appendix

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

A.1 Large sample properties of Q_n

For a given collection $\mathcal{K} = \{K_i\}_{i=1}^m$ of disjoint non-empty subsets of K, define the map $Q : \mathbf{R}_+ \to [0, 1]$ by

$$Q(\gamma; P, \mathcal{K}) := Prob_P[g(X, \sigma, \gamma; \mathcal{K}) > 0], \tag{A.1}$$

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

$$g(\mathbf{x}, \mathbf{y}, \gamma; \mathcal{K}) = \max_{1 \le i \le m} \left\{ \max_{j \in K_i} \{x_j - \gamma y_j\} - \min_{j \in K_i} \{x_j + \gamma y_j\} \right\}. \tag{A.2}$$

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

Lemma A.1. *If $m < k$, then $Q(\gamma; P, \mathcal{K})$ is a continuous and weakly decreasing function satisfying*

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

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

$$\begin{aligned} \lim_{\delta \to 0} Q(\gamma + \delta; P, \mathcal{K}) &= \lim_{\delta \to 0} Prob_P[g(X, \sigma, \gamma + \delta; \mathcal{K}) \ge 0] \\ &= \lim_{\delta \to 0} Prob_P[g(X, \sigma, \gamma; \mathcal{K}) + (g(X, \sigma, \gamma + \delta; \mathcal{K}) - g(X, \sigma, \gamma; \mathcal{K})) \ge 0] \end{aligned} \tag{A.3}$$

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

$$\lim_{\delta \to 0} Prob_P[|g(X, \sigma, \gamma + \delta, \mathcal{K}) - g(X, \sigma, \gamma, \mathcal{K})| > \epsilon] = 0. \tag{A.4}$$

By continuity of g, however, we obtain

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

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

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

$$\begin{aligned} Q(0,P) &= Prob_P\left[\max_{1\le i\le m}\left\{\max_{j\in K_i} X_j - \min_{j\in K_i} X_j\right\} > 0\right] \\ &= Prob_P\left[\max_{1\le i\le m}\left\{\max_{\ell,j\in K_i} |X_\ell - X_j|\right\} > 0\right] \end{aligned} \tag{A.5}$$

Because the joint distribution $\mathcal{L}_\infty$ of the random vector X is continuous and strictly increasing in its arguments (c.f. Assumption 3.1), we have $Prob_P[X_i = X_j] = 0$ for all $i \neq j$. Consequently,

$$Prob_P\left[\max_{1\le i\le m}\left\{\max_{\ell,j\in K_i} |X_\ell - X_j|\right\} > 0\right] = 1$$

since $m < k$ and, hence, at least one K_i in the partition contains two or more indices.

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

$$\max_{j\in K_i}\{X_j - \gamma\sigma_j\} - \min_{j\in K_i}\{X_j + \gamma\sigma_j\} \le \max_{j,\ell\in K_i} |X_\ell - X_j| - \gamma \min_{j\in K_i}\sigma_j \tag{A.6}$$

to obtain

$$Prob_P[g(X,\sigma,\gamma;\mathcal{K}) \ge 0] \le Prob\left[\max_{1\le i\le m}\left\{\max_{j,\ell\in K_i} |X_\ell - X_j| - \gamma \min_{j\in K_i}\sigma_j\right\} > 0\right]. \tag{A.7}$$

Because (A.7) holds for all $\gamma \in \mathbf{R}_+$ 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 $Q(\gamma;P,\mathcal{K})$, we now set out to prove the consistency of

$$Q_n(\gamma;\hat{P}_n,\mathcal{K}) := Prob_{\hat{P}_n}\left[g(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n), \hat{\sigma}_n^*, \gamma;\mathcal{K}) > 0\right]$$

as an estimator of

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

and to show that the difference between the quantiles $\gamma_n^*(\alpha;\mathcal{K})$ and $\gamma_n(\alpha,\mathcal{K})$ tends to zero

in probability as the sample size tends to infinity.

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

$$Prob_P\left[\sup_{\gamma \in \mathbf{R}_+} |Q_n(\gamma; \hat{P}_n, \mathcal{K}) - Q_n(\gamma; P, \mathcal{K})| > \epsilon\right] < \delta \tag{A.8}$$

Proof. Let $\epsilon, \delta > 0$ be given. It suffices to show that, for n sufficiently large,

$$\sup_{\gamma \in \mathbf{R}_+} |Q_n(\gamma, P, \mathcal{K}) - Q(\gamma, P, \mathcal{K})| < \epsilon/2 \tag{A.9}$$

and

$$Prob_P\left[\sup_{\gamma \in \mathbf{R}_+} |Q_n(\gamma; \hat{P}_n, \mathcal{K}) - Q(\gamma; P, \mathcal{K})| > \frac{\epsilon}{2}\right] < \delta. \tag{A.10}$$

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

$$g(X_n, \hat{\sigma}_n, \gamma; \mathcal{K}) \xrightarrow{d} g(X, \sigma, \gamma; \mathcal{K})$$

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

$$|Q_n(\gamma; P, \mathcal{K}) \to Q(\gamma; P, \mathcal{K})| < \epsilon/4. \tag{A.11}$$

That the convergence is uniform over $\gamma \in \mathbf{R}_+$ 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 $\gamma \in \mathbf{R}_+$ there exist N'_γ sufficiently large such that, for all $n \geq N'_\gamma$,

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

which in turn gives

$$Prob_P\left[\sup_{\gamma \in \mathbf{R}_+} |Q_n(\gamma; \hat{P}_n, \mathcal{K}) - Q(\gamma; P, \mathcal{K})| > \epsilon/2\right] < \delta \tag{A.12}$$

for all n sufficiently large by Polyá's theorem. □

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

$$Prob_P\left[|\gamma_n(\alpha;\mathcal{K}) - \gamma_n^*(\alpha;\mathcal{K})| > \delta\right] < \epsilon$$

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

$$Q^{-1}(\alpha; P, \mathcal{K}) = \inf\{\gamma : Q(\gamma; P, \mathcal{K}) \leq \alpha\}.$$

Lemma A.2, which shows that $Q_n(\gamma; P, \mathcal{K})$ converges uniformly to $Q(\gamma; P, \mathcal{K})$ as n tends to ∞, implies that

$$Q_n(Q^{-1}(\alpha; P, \mathcal{K}) + \delta/2; P, \mathcal{K}) < \alpha < Q_n(Q^{-1}(\alpha; P, \mathcal{K}) - \delta/2, ; P, \mathcal{K}) \tag{A.13}$$

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

$$Q^{-1}(\alpha; P, \mathcal{K}) - \delta/2 < \gamma_n(\alpha;\mathcal{K}) < Q^{-1}(\alpha; P, \mathcal{K}) + \delta/2 \tag{A.14}$$

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

Similarly, from Lemma A.2, we have that $Q_n(\gamma, \hat{P}_n; \mathcal{K})$ converges uniformly (in probability) to $Q(\gamma; P, \mathcal{K})$. Consequently, there exists n sufficiently large such that

$$Prob_P\left[Q_n(Q^{-1}(\alpha; P, \mathcal{K})+\delta/2; \hat{P}_n, \mathcal{K}) < \alpha < Q_n(Q^{-1}(\alpha; P, \mathcal{K})-\delta/2; \hat{P}_n, \mathcal{K})\right] \geq 1-\epsilon \tag{A.15}$$

and, hence,

$$Prob_P\left[Q^{-1}(\alpha; P, \mathcal{K}) - \delta/2 < \gamma_n^*(\alpha;\mathcal{K}) < Q^{-1}(\alpha; P, \mathcal{K}) + \delta/2\right] \geq 1 - \epsilon \tag{A.16}$$

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

$$Prob_P\left[|\gamma_n^*(\alpha;\mathcal{K}) - \gamma_n(\alpha;\mathcal{K})| > \delta\right] < \epsilon$$

for n sufficiently large, which—given that ϵ and δ are arbitrary—establishes the desired convergence. □

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

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

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.17) holds with equality when $D_0 = K \times K$. We then show that the inequality in (A.17) 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 δ_n and the fact that $Q_n(\gamma; P) := Q_n(\gamma; P, \{K\})$, we have

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

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

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

and

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

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

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

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 $Q(\cdot; P, \{K\})$ is continuous with $Q(\gamma(\alpha); P, \{K\}) = \alpha$.

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 must be the case that

$$\begin{aligned} &Prob_P\left[\sum_{(i,j)\in D_0(P)} \delta_n(i,j) \geq 1\right] \\ &\quad = Prob_P\left[\max_{A\in\mathcal{D}_0}\left\{\max_{j\in A}\{\sqrt{n}\,\hat{\theta}_{n,j} - \gamma_n^*(\alpha)\hat{\sigma}_{n,j}\} - \min_{j\in A}\{\sqrt{n}\,\hat{\theta}_{n,j} + \gamma_n^*(\alpha)\hat{\sigma}_{n,j}\}\right\} > 0\right] \\ &\quad = Prob_P\Big[g(X_n, \hat{\sigma}, \gamma_n^*(\alpha); \mathcal{D}_0) > 0\Big] \end{aligned}$$

where $\mathcal{D}_0$ partitions K into its equivalence classes. The desired result is then established because

$$Prob_P\Big[g(X_n, \hat{\sigma}, \gamma_n^*(\alpha); \mathcal{D}_0) > 0\Big] \leq Prob_P\Big[g(X_n, \hat{\sigma}, \gamma_n^*(\alpha); \{K\}) > 0\Big] \tag{A.22}$$

for all n, and

$$\lim_{n\to\infty} Prob_P\Big[g(X_n, \hat{\sigma}, \gamma_n^*(\alpha); \{K\}) > 0\Big] = \alpha$$

as was shown above.

□

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

$$\begin{aligned} Prob_P\left[\prod_{(i,j)\in D_1(P)} \delta_n = 1\right] &= 1 - Prob_P\left[\sum_{(i,j)\in D_1(P)} \delta_n(i,j) = 0\right] \\ &\geq 1 - \sum_{(i,j)\in D_1(P)} Prob_P\big[\hat{\theta}_{n,i} > \hat{\theta}_{n,j}\big] - \sum_{(i,j)\in D_1(P)} Prob_P\big[C_{n,i} \cap C_{n,j} \neq \emptyset\big] \end{aligned} \tag{A.23}$$

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

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

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,

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

as desired. It remains to be shown that

$$\sum_{(i,j)\in D_1(P)} Prob_P\big[C_{n,i} \cap C_{n,j} \neq \emptyset\big] \to 0$$

as n tends to ∞. 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

$$Prob_P[C_{n,i} \cap C_{n,j} \neq \emptyset] \leq Prob_P\big[\sqrt{n}\,|\hat{\theta}_{n,i} - \hat{\theta}_{n,j}| < \gamma^*_n(\alpha)\big(\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}\big)\big] \tag{A.24}$$

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

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

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

$$\begin{aligned} Prob_P\left[\sum_{(i,j)\in D_1(P)} \delta_n(j,i) \geq 1\right] &\leq Prob_P\left[\sum_{(i,j)\in D_1(P)} [1 - \delta_n(i,j)] \geq 1\right] \\ &\leq Prob_P\left[\prod_{(i,j)\in D_1(P)} \delta_n(i,j) = 0\right] \\ &\leq 1 - Prob_P\left[\prod_{(i,j)\in D_1(P)} \delta_n(i,j) = 1\right] \end{aligned} \tag{A.25}$$

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

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, and let $\mathcal{D}_0(P)$ denote the partition of K formed by the set of all equivalence classes

$$[i] = \{j \in K : \theta_i(P) \cap \theta_j(P)\}. \tag{A.26}$$

Note that $|\mathcal{D}_0(P)| = k$ when no two parameters are equal. Because FWE is trivially satisfied in this case, we may safely assume that $|\mathcal{D}_0(P)| < k$. (This assumption plays an important role in equation (A.27) below where a zero probability might otherwise arise if $|\mathcal{D}_0(P)| = k$.) Now, after recalling that the second part of Theorem 3.2 implies that, with probability tending to 1, all unequal parameters will be separated in the first step of the iterative overlap procedure, and using the fact that $\gamma_n^*(\alpha; \mathcal{D}_0(P)) \leq \gamma_n^*(\alpha; \mathcal{K})$, we have

$$\begin{aligned}
&\lim_{n\to\infty} Prob_P\left[\sum_{(i,j)\in D_0(P)} \delta_n^{\text{global}}(i,j) \geq 1\right] \\
&\quad = \lim_{n\to\infty} Prob_P\left[\max_{A\in\mathcal{D}_0}\left\{\max_{j\in A}\{\sqrt{n}\,\hat{\theta}_{n,j} - \gamma_n^*(\alpha;\mathcal{D}_0)\hat{\sigma}_{n,j}\} - \min_{j\in A}\{\sqrt{n}\,\hat{\theta}_{n,j} + \gamma_n^*(\alpha;\mathcal{D}_0)\hat{\sigma}_{n,j}\}\right\} > 0\right] \\
&\quad = \lim_{n\to\infty} Q_n(\gamma_n^*(\alpha;\mathcal{D}_0); P, \mathcal{D}_0) \\
&\quad = \alpha
\end{aligned}$$

with the last line an immediate consequence of Lemmas A.2 and A.3.

□

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

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

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 $\delta_n^{global}(i,j) = \delta_n(i,j)$ or $\delta_n^{global}(i,j) = 1$ and $\delta_n(i,j) = 0$. Either way, δ_n^{global} satisfies

$$\delta_n^{global}(i,j) \geq \delta_n(i,j), \tag{A.27}$$

and, hence, δ_n^{global} yields a weak refinement of the rankings provided by δ_n. □

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

$$M_{i,j} = \max\Big\{\sqrt{n}\,(\hat{\theta}_{i,n} - \theta_i) - \gamma\hat{\sigma}_{n,i}, \sqrt{n}\,(\hat{\theta}_{j,n} - \theta_j) - \gamma\hat{\sigma}_{n,j}\Big\},$$

and

$$m_{i,j} = \min\Big\{\sqrt{n}\,(\hat{\theta}_{i,n} - \theta_i) + \gamma\hat{\sigma}_{n,i}, \sqrt{n}\,(\hat{\theta}_{j,n} - \theta_j) + \gamma\hat{\sigma}_{n,j}\Big\}.$$

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

$$\begin{aligned}
Q_n(\gamma, P) &= Prob_P\Big[\max_{j\in K}\{\sqrt{n}(\hat{\theta}_{n,j} - \theta_j) - \gamma\hat{\sigma}_{j,n}\} - \min_{j\in K}\{\sqrt{n}\,(\hat{\theta}_{n,j} - \theta_j) + \gamma\hat{\sigma}_{n,j}\} \geq 0\Big] \\
&\geq Prob_P\big[M_{i,j} - m_{i,j} \geq 0\big] \\
&= Prob_P\left[|(\hat{\theta}_{i,n} - \hat{\theta}_{j,n}) - (\theta_i - \theta_j)| \geq \gamma\Big(\frac{\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}}{\sqrt{n}}\Big)\right] \\
&= 1 - Prob_P\left[(\hat{\theta}_{i,n} - \hat{\theta}_{j,n}) - \gamma\Big(\frac{\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}}{\sqrt{n}}\Big) \leq \theta_i - \theta_j \leq (\hat{\theta}_{i,n} - \hat{\theta}_{j,n}) + \gamma\Big(\frac{\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}}{\sqrt{n}}\Big)\right]
\end{aligned} \tag{B.2}$$

The inequality in B.2 shows that

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

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

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

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 α.

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 $\hat{\theta}_{n_i,i}$ denote an estimator of θ_i based on n_i observations from the i^{th} sample. Similarly, let $\hat{\sigma}_{n_i,i}$ be an estimator of the variance of $\sqrt{n_i}\hat{\theta}_{n_i,i}$ that is based on the same n_i observations. Lastly, we define

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

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

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

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

and

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

with the corresponding feasible estimator of γ defined accordingly as

$$\gamma^*_{\mathbf{n}}(\alpha) = \inf\left\{\gamma : Q_{\mathbf{n}}(\gamma; \hat{P}_{\mathbf{n}}) \leq \alpha\right\}. \tag{C.1}$$

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 δ_n as defined in (3.2) with $\gamma = \gamma^*_{\mathbf{n}}(\alpha)$ 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.