University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2003 *Paper* 138

Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates

Sandrine Dudoit* Mark J. van der Laan†
Katherine S. Pollard‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley, sandrine@stat.berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

‡Center for Biomolecular Science & Engineering, University of California, Santa Cruz, kpollard@gladstone.ucsf.edu

http://biostats.bepress.com/ucbbiostat/paper138

Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates

Sandrine Dudoit, Mark J. van der Laan, and Katherine S. Pollard

Abstract

The present article proposes general single-step multiple testing procedures for controlling Type I error rates defined as arbitrary parameters of the distribution of the number of Type I errors, such as the generalized family-wise error rate. A key feature of our approach is the test statistics null distribution (rather than data generating null distribution) used to derive cut-offs (i.e., rejection regions) for these test statistics and the resulting adjusted p-values. For general null hypotheses, corresponding to submodels for the data generating distribution, we identify an asymptotic domination condition for a null distribution under which single-step common-quantile and common-cut-off procedures asymptotically control the Type I error rate, for arbitrary data generating distributions, without the need for conditions such as subset pivotality. Inspired by this general characterization of a null distribution, we then propose as an explicit null distribution the asymptotic distribution of the vector of null-value shifted and scaled test statistics. In the special case of family-wise error rate (FWER) control, our method yields the single-step minP and maxT procedures based on minima of unadjusted p-values and maxima of test statistics, respectively, with the important distinction in the choice of null distribution. Single-step procedures based on consistent estimators of the null distribution are shown to also provide asymptotic control of the Type I error rate. A general bootstrap algorithm is supplied to conveniently obtain consistent estimators of the null distribution. The special cases of t- and F-statistics are discussed in detail. The companion articles focus on step-down multiple testing procedures for control of the FWER (van der Laan et al., 2003a) and on augmentations of FWER-controlling methods to control error rates such as the generalized family-wise error rate and the proportion of false positives among the rejected

hypotheses (van der Laan et al., 2003b). The proposed bootstrap multiple testing procedures are evaluated by a simulation study and applied to gene expression microarray data in the fourth article of the series (Pollard et al., 2004).

Contents

1 Introduction **3**
- 1.1 Motivation . . . 3
- 1.2 Outline . . . 5

2 Multiple hypothesis testing framework **6**
- 2.1 Basic set-up . . . 6
  - 2.1.1 Model . . . 7
  - 2.1.2 Parameters . . . 7
  - 2.1.3 Null hypotheses . . . 8
  - 2.1.4 Test statistics . . . 9
- 2.2 Multiple testing procedures . . . 10
- 2.3 Type I error rates . . . 10
- 2.4 Adjusted p-values . . . 13
- 2.5 Stepwise multiple testing procedures . . . 14
- 2.6 Type I error rate control and choice of null distribution . . . 16
  - 2.6.1 Type I error rate control . . . 16
  - 2.6.2 Sketch of proposed approach to Type I error rate control 16
  - 2.6.3 Choice of null distribution and null domination conditions . . . 18
  - 2.6.4 Contrast with other approaches . . . 19

3 Single-step procedures for control of general Type I error rates **21**
- 3.1 General procedures and asymptotic control results . . . 22
  - 3.1.1 Single-step common-quantile procedure . . . 23
  - 3.1.2 Single-step common-cut-off procedure . . . 25
  - 3.1.3 Common-quantile vs. common-cut-off procedures . . . 25
- 3.2 Explicit proposal for the test statistics null distribution . . . 26
- 3.3 Adjusted p-values . . . 28
  - 3.3.1 General Type I error rates . . . 29
  - 3.3.2 Application to control of the PCER . . . 30
  - 3.3.3 Application to control of the gFWER . . . 31

4 Bootstrap-based single-step procedures for control of general Type I error rates **34**

4.1 Asymptotic control for consistent estimator of the null distribution 36
4.2 Bootstrap estimation of the null distribution 44

5 Examples **47**
5.1 t-statistics for single-parameter hypotheses 47
5.1.1 Null distribution 48
5.1.2 Estimation of the null distribution 49
5.1.3 Example: Tests of means 50
5.1.4 Example: Tests of correlations 51
5.1.5 Example: Tests of regression parameters 52
5.2 F-statistics for multiple-parameter hypotheses 54
5.2.1 Null distribution 54
5.2.2 Estimation of the null distribution 59

6 Strong control, weak control, and subset pivotality **59**
6.1 Strong and weak control of a Type I error rate 60
6.2 Subset pivotality 62

1 Introduction

1.1 Motivation

DNA microarrays and other high-throughput biological assays have motivated us to investigate multiple testing methods in large multivariate settings, though our results apply to multiple testing in general. Current statistical inference problems in genomic data analysis are characterized by: (i) high-dimensional multivariate distributions, with typically unknown and intricate correlation patterns among variables; (ii) large parameter spaces; (iii) a number of variables (hypotheses) that is much larger than the sample size; and (iv) some non-negligible proportion of false null hypotheses, i.e., true positives. Multiple hypothesis testing methods are concerned with the simultaneous test of $m > 1$ null hypotheses, while controlling a suitably defined Type I error rate (i.e., false positive rate), such as the family-wise error rate or the false discovery rate. General references on multiple testing include Hochberg and Tamhane (1987), Shaffer (1995), and Westfall and Young (1993). A number of recent articles have addressed the question of multiple testing as it relates to the identification of differentially expressed genes in DNA microarray experiments (Dudoit et al., 2002; Efron et al., 2001; Golub et al., 1999; Manduchi et al., 2000; Pollard and van der Laan, 2003; Reiner et al., 2003; Tusher et al., 2001; Westfall et al., 2001; Xiao et al., 2002); a review of multiple testing methods in the context of microarray data analysis is given in Dudoit et al. (2003).

In multiple testing, decisions to reject the null hypotheses are based on cut-off rules for test statistics (or their associated p-values), so that a given Type I error rate is controlled at a specified level α. In practice, however, the joint distribution of the test statistics is typically unknown and replaced by an assumed *null distribution* in order to derive these cut-offs. Current approaches use a data generating distribution that satisfies the complete null hypothesis that *all* null hypotheses are true. Procedures based on such a null distribution typically rely on the subset pivotality condition stated in Westfall and Young (1993), p. 42–43, to ensure that control under a data generating distribution satisfying the complete null hypothesis does indeed give the desired control under the *true* data generating distribution. However, the subset pivotality condition is violated in important testing problems, since a data generating distribution satisfying the complete null hypothesis might result in a joint distribution for the vector of test statistics that is

different from their actual distribution. In fact, in many problems, there does not even exist a data generating null distribution that correctly specifies the joint distribution of the test statistics corresponding to the true null hypotheses (e.g., tests concerning correlations in Section 5.1).

Pollard and van der Laan (2003) formally define a statistical framework for testing multiple single-parameter null hypotheses of the form $H_{0j} = \mathrm{I}(\mu(j) \leq \mu_0(j))$, for one-sided tests, and $H_{0j} = \mathrm{I}(\mu(j) = \mu_0(j))$, for two-sided tests, where $\mu = (\mu(j) : j = 1, \ldots, m)$ is an m-vector of parameters and $\mu_0 = (\mu_0(j) : j = 1, \ldots, m)$ are the hypothesized null-values. They propose as null distribution the asymptotic distribution of the mean-zero centered test statistics and prove that, with this choice of null distribution, single-step multiple testing procedures based on common-cut-off rules for the test statistics or the corresponding marginal p-values (common-quantile procedures) provide asymptotic control of any Type I error rate that is a function of the distribution of the number of false positives. This general approach does not rely on subset pivotality. Pollard and van der Laan (2003) propose a bootstrap algorithm for estimating the null distribution and prove the important practical result that multiple testing procedures based on consistent estimators of the null distribution (e.g., from non-parametric or model-based bootstrap) asymptotically control the Type I error rate. These authors also generalize the equivalence of hypothesis testing and confidence regions to the multivariate setting, by demonstrating that their single-step multiple testing procedures, with asymptotic control of a particular Type I error rate at level α, are equivalent with constructing an asymptotic $(1-\alpha)$–confidence region for the parameter of interest (e.g., bootstrap-based) and rejecting the hypotheses for which the null-values are not included in the confidence region.

This manuscript and its companion (van der Laan et al., 2003a) are concerned with the choice of null distribution for single-step and step-down multiple testing procedures that provide asymptotic control of Type I error rates defined as arbitrary parameters of the distribution of the number of Type I errors. Examples of such error rates include the generalized family-wise error rate (gFWER), i.e., the probability of at least $(k+1)$ Type I errors, for some user-supplied integer $k \geq 0$, and the family-wise error rate (FWER), which is the gFWER in the special case $k = 0$. We build on the earlier work of Pollard and van der Laan (2003) as follows: (i) general collections of null hypotheses, corresponding to submodels for the data generating distribution, are considered; (ii) step-down procedures are provided for asymptotic control of the FWER; (iii) adjusted p-values are derived for each of the multiple

testing procedures. A general characterization and explicit construction are proposed for a test statistics null distribution that provides asymptotic control of the Type I error rate under the true data generating distribution, without the need for conditions such as subset pivotality. This null distribution is used to obtain cut-offs for the test statistics (or their corresponding unadjusted p-values), and also to derive the resulting adjusted p-values for single-step and step-down procedures.

1.2 Outline

The present article focuses on the choice of null distribution for *single-step* procedures controlling error rates defined as arbitrary parameters of the distribution of the number of Type I errors, such as the generalized family-wise error rate, gFWER. In the next section, we describe a general statistical framework for multiple hypothesis testing. In particular, Section 2.6 outlines the main features of our approach to Type I error control and the choice of a null distribution. Section 3 proposes single-step *common-quantile* (Procedure 1) and *common-cut-off* (Procedure 2) multiple testing procedures that provide asymptotic control of the Type I error rate. A key feature of our approach is the test statistics null distribution (rather than data generating null distribution) used to derive cut-offs (i.e., rejection regions) for these test statistics and the resulting adjusted p-values. For general null hypotheses, corresponding to submodels for the data generating distribution, we identify an asymptotic domination condition for a null distribution under which single-step common-quantile and common-cut-off procedures asymptotically control the Type I error rate, for arbitrary data generating distributions, without the need for conditions such as subset pivotality (Theorem 1). Inspired by this general characterization of a null distribution, we then propose as an explicit null distribution the asymptotic distribution of the vector of null-value shifted and scaled test statistics (Theorem 2). In the special case of family-wise error rate control, our approach yields the *single-step minP* and *single-step maxT* procedures based on minima of unadjusted p-values and maxima of test statistics, respectively (Section 3.3.3). In Section 4, procedures based on a consistent estimator of the null distribution are shown to also provide asymptotic control of the Type I error rate (Theorems 3 and 4, Corollary 1). Resampling procedures are supplied to conveniently obtain consistent estimators of the null distribution (bootstrap Procedures 3–5). Section 5 focuses on two particular examples of testing problems covered

by our framework: the test of single-parameter null hypotheses (e.g., tests of means, correlations, regression parameters) using t-statistics and the test of multiple-parameter hypotheses using F-statistics. Section 6 revisits the notions of strong and weak control of a Type I error rate, and the related condition of subset pivotality. Differences between our approach to Type I error control and earlier approaches are highlighted.

The companion article (van der Laan et al., 2003a) considers *step-down* approaches for controlling the family-wise error rate and provides procedures based on maxima of test statistics (*step-down maxT*) and minima of unadjusted p-values (*step-down minP*). In the third article of the series, van der Laan et al. (2003b) propose simple augmentations of FWER-controlling procedures which control the generalized family-wise error rate and the proportion of false positives among the rejected hypotheses, under general data generating distributions, with arbitrary dependence structures among variables. The proposed methods are evaluated by a simulation study and applied to gene expression microarray data in the fourth article of the series (Pollard et al., 2004). Software implementing the bootstrap single-step and step-down multiple testing procedures will be available in the R package `multtest`, released as part of the Bioconductor Project (`www.bioconductor.org`).

2 Multiple hypothesis testing framework

2.1 Basic set-up

For the remainder of the article, we adopt the following definitions for inverses of cumulative distribution functions (c.d.f.) and survivor functions. Let F denote a (non-decreasing and right-continuous) c.d.f. and let $\bar{F}$ denote the corresponding (non-increasing and right-continuous) survivor function, defined as $\bar{F} \equiv 1 - F$. For $\alpha \in [0, 1]$, define inverses as

$$F^{-1}(\alpha) \equiv \inf\{x : F(x) \geq \alpha\} \qquad \text{and} \qquad \bar{F}^{-1}(\alpha) \equiv \inf\{x : \bar{F}(x) \leq \alpha\}. \quad (1)$$

With these definitions, $\bar{F}^{-1}(\alpha) = F^{-1}(1 - \alpha)$.

Note that we follow the convention that lower case letters denote realizations of random variables, e.g., x is a realization of the random variable X.

2.1.1 Model

Let $X_1, \ldots, X_n$ be n independent and identically distributed (i.i.d.) random d-vectors, $X = (X(j) : j = 1, \ldots, d) \sim P \in \mathcal{M}$, where the *data generating distribution* P is known to be an element of a particular *statistical model* $\mathcal{M}$ (possibly non-parametric). For example, in cancer microarray studies, $(X_i(1), \ldots, X_i(g))$ may denote a g-vector of gene expression measures and $(X_i(g+1), \ldots, X_i(d))$ a $(d-g)$-vector of biological and clinical outcomes for patient i, $i = 1, \ldots, n$. In microarray data analysis and other current areas of application of multiple testing methods, the dimension d, of the data vector X, is usually much larger than the sample size n, i.e., one can have thousands of expression measures for less than one hundred patients.

2.1.2 Parameters

We consider general *parameters* defined as functions of the unknown data generating distribution P: $\mu = (\mu(j) : j = 1, \ldots, m)$, where $\mu(j) = \mu_j(P) \in \mathbb{R}$ and typically $m \geq d$. Parameters of interest include means, differences in means, correlations, and can refer to linear models, generalized linear models, survival models (e.g., Cox proportional hazards model), time-series models, dose-response models, etc. For instance, in microarray data analysis, one may be concerned with testing problems regarding the following parameters.

- *Location parameters.* E.g. means and medians for measuring differential expression for g genes.
  $\mu(j) \equiv E(X(j))$= mean expression level of gene j, $j = 1, \ldots, g$, in a particular population, $X \sim P$.
  $\mu(j) \equiv \mu_2(j) - \mu_1(j) = E(X_2(j)) - E(X_1(j))$ = difference in mean expression level for gene j, $j = 1, \ldots, g$, in Populations 1 and 2, $X_1 \sim P_1$, $X_2 \sim P_2$.

- *Scale parameters.* E.g. covariances and correlations.
  $\gamma(j, j') = Cor[X(j), X(j')]$ = pairwise correlation for the expression measures of genes j and j', $j \neq j' = 1, \ldots, g$, $X \sim P$.

- *Regression parameters.* E.g. slopes, main effects, and interactions, for measuring association of expression level $X(j)$ of gene j, $j = 1, \ldots, g$, with outcomes/covariates $(X(j) : j = g+1, \ldots, d)$, $X \sim P$.
  $\mu(j)$ = regression parameter for univariate Cox proportional hazards model for survival time $T = X(g+1)$ given the expression measure

$X(j)$ of gene j.
$\mu(j)$ = interaction effect for two drugs on expression level of gene j.
$\mu(j)$ = linear combination $a'\beta(j)$, e.g., contrast in ANOVA.

2.1.3 Null hypotheses

General submodel null hypotheses. In order to cover a broad class of testing problems, we define m null hypotheses in terms of a collection of *submodels*, $\mathcal{M}_j \subseteq \mathcal{M}$, $j = 1, \ldots, m$, for the data generating distribution P. The m *null hypotheses* are defined as $H_{0j} \equiv \mathrm{I}(P \in \mathcal{M}_j)$ and the corresponding *alternative hypotheses* as $H_{1j} \equiv \mathrm{I}(P \notin \mathcal{M}_j)$. Thus, H_{0j} is true, i.e., $H_{0j} = 1$, if $P \in \mathcal{M}_j$ and false otherwise.

Let $S_0 = S_0(P) \equiv \{j : H_{0j} \text{ is true}\} = \{j : P \in \mathcal{M}_j\}$ be the set of $m_0 = |S_0|$ true null hypotheses, where we note that S_0 depends on the true data generating distribution P. Let $S_0^c = S_0^c(P) \equiv \{j : H_{0j} \text{ is false}\} = \{j : P \notin \mathcal{M}_j\}$ be the set of $m_1 = m - m_0$ false null hypotheses, i.e., true positives. The goal of a multiple testing procedure is to accurately estimate the set S_0, and thus its complement S_0^c, while controlling probabilistically the number of false positives at a user-supplied level α.

Single-parameter null hypotheses. A familiar special case, considered in Section 5, is that where each null hypothesis refers to a single parameter, $\mu(j) = \mu_j(P) \in \mathbb{R}$, $j = 1, \ldots, m$. The parameters $\mu(j)$ could be expected values or pairwise correlations for components of a random d-vector $X \sim P$. One distinguishes between two types of testing problems for single parameters.

$$\begin{aligned} \text{One-sided tests} \quad & H_{0j} = \mathrm{I}\big(\mu(j) \le \mu_0(j)\big) \\ \text{vs.} \quad & H_{1j} = \mathrm{I}\big(\mu(j) > \mu_0(j)\big), \quad j = 1, \ldots, m. \end{aligned}$$

$$\begin{aligned} \text{Two-sided tests} \quad & H_{0j} = \mathrm{I}\big(\mu(j) = \mu_0(j)\big) \\ \text{vs.} \quad & H_{1j} = \mathrm{I}\big(\mu(j) \ne \mu_0(j)\big), \quad j = 1, \ldots, m. \end{aligned}$$

The hypothesized null-values, $\mu_0(j)$, are frequently zero (e.g., no difference in mean expression levels between two populations of patients).

2.1.4 Test statistics

The decisions to reject or not the null hypotheses are based on an m-vector of *test statistics*, $T_n = (T_n(j) : j = 1, \ldots, m)$, that are functions of the data, $X_1, \ldots, X_n$. Denote the (finite sample) joint distribution of the test statistics T_n by $Q_n = Q_n(P)$. It is assumed that large values of $T_n(j)$ provide evidence against the null hypothesis H_{0j}. For two-sided tests, one can take absolute values of the test statistics.

As in Pollard and van der Laan (2003), for single-parameter null hypotheses, $H_{0j} = \mathrm{I}(\mu(j) \leq \mu_0(j))$, $j = 1, \ldots, m$, we consider two main types of test statistics, *difference statistics*, $D_n(j)$, and *t-statistics* (i.e., standardized differences), $T_n(j)$,

$$\begin{aligned} D_n(j) &\equiv \big(\text{Estimator} - \text{Null-value}\big) = \sqrt{n}(\mu_n(j) - \mu_0(j)) \\ T_n(j) &\equiv \frac{\text{Estimator} - \text{Null-value}}{\text{Standard Error}} = \sqrt{n}\frac{\mu_n(j) - \mu_0(j)}{\sigma_n(j)}. \end{aligned} \tag{2}$$

Here, $\mu_n = (\mu_n(j) : j = 1, \ldots, m)$ denotes an m-vector of *estimators* for the parameter m-vector $\mu = (\mu(j) : j = 1, \ldots, m)$ and $\sigma_n/\sqrt{n} = (\sigma_n(j)/\sqrt{n} : j = 1, \ldots, m)$ denote the corresponding *estimated standard errors*. We consider *asymptotically linear estimators* μ_n of the parameter μ, with m-dimensional vector *influence curve* (IC), $IC(X \mid P) = (IC_j(X \mid P) : j = 1, \ldots, m)$, such that

$$\mu_n(j) - \mu(j) = \frac{1}{n}\sum_{i=1}^{n} IC_j(X_i \mid P) + o_P(1/\sqrt{n}), \tag{3}$$

where $E[IC(X \mid P)] = 0$ and $\Sigma(P)$ and $\rho(P)$, denote, respectively, the covariance and correlation matrices of the vector IC. In addition, $\sigma_n^2(j)$ are assumed to be consistent estimators of the IC variances, $\sigma^2(j) = E[IC_j^2(X \mid P)]$.

The influence curve of a given estimator can be derived as the mean-zero centered functional derivative of the estimator (as a function of the empirical distribution P_n for the entire sample of size n), applied to the empirical distribution based on a sample of size one (Gill, 1989; Gill et al., 1995). As illustrated in Section 5.1, this general representation for the test statistics includes standard one-sample and two-sample t-statistics, but also test statistics for correlations and regression parameters in linear and non-linear models. F-statistics for multiple-parameter null hypotheses are discussed in Section 5.2.

2.2 Multiple testing procedures

A *multiple testing procedure* (MTP) produces a set S_n of rejected hypotheses, that *estimates* S_0^c, the set of false null hypotheses,

$$S_n = S(T_n, Q_0, \alpha) \equiv \{j : H_{0j} \text{ is rejected}\} \subseteq \{1, \ldots, m\}. \tag{4}$$

As indicated by the long notation $S(T_n, Q_0, \alpha)$, the set S_n (or $\hat{S}_0^c$) depends on: (i) the data, $X_1, \ldots, X_n$, through the test statistics T_n; (ii) a null distribution, Q_0, for the test statistics, used to compute cut-offs for each $T_n(j)$ (and the resulting adjusted p-values); and (iii) the nominal level α of the MTP, i.e., the desired upper bound for a suitably defined Type I error rate. Multiple testing procedures such as those proposed in this and the companion articles, can be represented as

$$S_n = S(T_n, Q_0, \alpha) = \{j : T_n(j) > c_j\},$$

where $c_j = c_j(T_n, Q_0, \alpha)$, $j = 1, \ldots, m$, are possibly random *cut-offs*, or *critical values*, computed under the null distribution Q_0 for the test statistics.

2.3 Type I error rates

Type I and Type II errors. In any testing situation, two types of errors can be committed: a *false positive*, or *Type I error*, is committed by rejecting a true null hypothesis, and a *false negative*, or *Type II error*, is committed when the test procedure fails to reject a false null hypothesis. The situation can be summarized by Table 1 below, where the number of Type I errors is $V_n \equiv |S_n \cap S_0|$ and the number of Type II errors is $U_n \equiv |S_n^c \cap S_0^c|$. Note that both U_n and V_n depend on the unknown data generating distribution P through $S_0 = S_0(P)$. The numbers $m_0 = |S_0|$ and $m_1 = m - m_0$ of true and false null hypotheses are *unknown parameters*, the number of rejected hypotheses $R_n \equiv |S_n|$ is an *observable random variable*, and $m_1 - U_n$, U_n, $m_0 - V_n$, and V_n are *unobservable random variables* (depending on P, through $S_0(P)$).

Type I error rates. Ideally, one would like to simultaneously minimize the number V_n of Type I errors and the number U_n of Type II errors. A standard approach in the univariate setting is to prespecify an acceptable level α for the Type I error rate and seek tests which minimize the Type II error rate,

Table 1: Type I and Type II errors in multiple hypothesis testing.

		Null hypotheses not rejected	Null hypotheses rejected	
Null hypotheses	true	$\lvert S_n^c \cap S_0 \rvert$	$V_n = \lvert S_n \cap S_0 \rvert$ (Type I)	$m_0 = \lvert S_0 \rvert$
	false	$U_n = \lvert S_n^c \cap S_0^c \rvert$ (Type II)	$\lvert S_n \cap S_0^c \rvert$	$m_1 = \lvert S_0^c \rvert$
		$\lvert S_n^c \rvert$	$R_n = \lvert S_n \rvert$	m

i.e., maximize *power*, within the class of tests with Type I error rate at most α. In the multiple hypothesis case, a variety of generalizations are possible for the definition of Type I error rate (and of power). Here, we consider error rates that are defined as functions of the distribution of the number of Type I errors, that is, can be represented as parameters $\theta(F_{V_n})$, where F_{V_n} is the discrete cumulative distribution function (c.d.f.) on $\{0, \ldots, m\}$ for the number of Type I errors, V_n. Such a general representation covers the following commonly-used Type I error rates.

- *Per-comparison error rate* (PCER), or expected proportion of Type I errors among the m tests,

$$PCER \equiv E(V_n)/m = \int v dF_{V_n}(v)/m.$$

- *Per-family error rate* (PFER), or expected number of Type I errors,

$$PFER \equiv E(V_n) = \int v dF_{V_n}(v).$$

- *Median-based per-family error rate (mPFER)*, or median number of Type I errors,

$$mPFER \equiv \text{Median}(F_{V_n}) = F_{V_n}^{-1}(1/2).$$

- *Family-wise error rate* (FWER), or probability of at least one Type I error,
$$FWER \equiv Pr(V_n \geq 1) = 1 - F_{V_n}(0).$$

- *Generalized family-wise error rate* (gFWER), or probability of at least $(k+1)$ Type I errors, $k = 0, \ldots, m_0 - 1$,
$$gFWER(k) \equiv Pr(V_n \geq k+1) = 1 - F_{V_n}(k).$$

  When $k = 0$, the gFWER is the usual family-wise error rate, FWER.

For convenience, we work with normalized Type I error rates, so that $\theta(F_{V_n}) \in [0, 1]$. Note that the *false discovery rate* (FDR) of Benjamini and Hochberg (1995) cannot be represented as a parameter $\theta(F_{V_n})$, because it is defined also in terms of the distribution of R_n, the total number of rejected hypotheses (including the true positives, $m_1 - U_n = R_n - V_n$). The FDR is the expected proportion of Type I errors among the rejected hypotheses, i.e., $FDR = E(V_n/R_n)$, with the convention that $V_n/R_n = 0$ if $R_n = 0$. van der Laan et al. (2003b) provide simple augmentations of FWER-controlling procedures that control the *proportion of false positives among the rejected hypotheses*, V_n/R_n, at a user-supplied proportion $q \in (0, 1)$, under general data generating distributions P, with arbitrary dependence structures among variables. That is, for a level $\alpha \in (0, 1)$, the augmented procedures satisfy

$$PFP(q) \equiv Pr(V_n/R_n > q) \leq \alpha.$$

Assumptions for the mapping θ that defines the Type I error rates. We make the following assumptions for the mapping $\theta : F \to \theta(F)$, defining a Type I error rate as a parameter corresponding to a cumulative distribution function F on $\{0, \ldots, m\}$.
Monotonicity. Given two c.d.f.'s F_1 and F_2 on $\{0, \ldots, m\}$,

$$F_1 \geq F_2 \qquad \Longrightarrow \qquad \theta(F_1) \leq \theta(F_2). \tag{AMI}$$

Uniform continuity. Given two c.d.f.'s F_1 and F_2 on $\{0, \ldots, m\}$, define the distance measure d by $d(F_1, F_2) = \max_{x \in \{0,\ldots,m\}} \mid F_1(x) - F_2(x) \mid$. For two sequences of c.d.f.'s, $\{F_n\}$ and $\{G_n\}$,

$$\text{if } d(F_n, G_n) \to 0, \text{ as } n \to \infty, \text{ then } \theta(F_n) - \theta(G_n) \to 0. \tag{ACI}$$

2.4 Adjusted p-values

Unadjusted p-values. Consider the test of individual null hypotheses H_{0j} at single test level α, i.e., such that the chance of a Type I error for each H_{0j} is at most α (note that in this $m = 1$ case, FWER, PCER, and PFER coincide). Given a test statistic $T_n(j)$, with marginal null distribution Q_{0j}, the null hypothesis H_{0j} is rejected at single test level α, if $T_n(j) > c_j(Q_{0j}, \alpha)$, where the cut-off $c_j(\alpha) = c_j(Q_{0j}, \alpha)$ is defined in terms of the marginal survivor function, $\bar{Q}_{0j}(z) = 1 - Q_{0j}(z) = Pr_{Q_{0j}}(T_n(j) > z)$, as

$$c_j(Q_{0j}, \alpha) \equiv \bar{Q}_{0j}^{-1}(\alpha) = \inf\{z : \bar{Q}_{0j}(z) \leq \alpha\}.$$

For the test of single null hypothesis H_{0j}, the *unadjusted p-value* (a.k.a. *marginal* or *raw* p-value), $P_{0n}(j) = P(T_n(j), Q_{0j})$, is based only on the test statistic $T_n(j)$ for that hypothesis and is defined as

$$\begin{aligned} P_{0n}(j) &\equiv \inf\{\alpha \in [0,1] : \text{Reject } H_{0j} \text{ at single test level } \alpha, \text{ given } T_n(j)\} \\ &= \inf\{\alpha \in [0,1] : c_j(Q_{0j}, \alpha) < T_n(j)\}, \qquad j = 1, \ldots, m. \end{aligned} \tag{5}$$

That is, $P_{0n}(j)$ is the nominal level of the *single* hypothesis testing procedure at which H_{0j} would just be rejected, given $T_n(j)$. For continuous marginal null distributions Q_{0j}, the unadjusted p-values are given by $P_{0n}(j) = c_j^{-1}(T_n(j)) = \bar{Q}_{0j}(T_n(j))$, where c_j^{-1} is the inverse of the monotone decreasing function $\alpha \to c_j(\alpha) = c_j(Q_{0j}, \alpha)$.

Adjusted p-values. The definition of p-value can be extended to multiple testing problems as follows. Given any multiple testing procedure

$$S_n = S(T_n, Q_0, \alpha) = \{j : T_n(j) > c_j(T_n, Q_0, \alpha)\},$$

based on cut-offs $c_j(\alpha) = c_j(T_n, Q_0, \alpha)$, the *adjusted p-value*, $\tilde{P}_{0n}(j) = \tilde{P}(j, T_n, Q_0)$, for null hypothesis H_{0j}, is defined as

$$\begin{aligned} \tilde{P}_{0n}(j) &\equiv \inf\{\alpha \in [0,1] : \text{Reject } H_{0j} \text{ at MTP level } \alpha, \text{ given } T_n\} \\ &= \inf\{\alpha \in [0,1] : j \in S(T_n, Q_0, \alpha)\} \\ &= \inf\{\alpha \in [0,1] : c_j(T_n, Q_0, \alpha) < T_n(j)\}, \qquad j = 1, \ldots, m. \end{aligned} \tag{6}$$

That is, $\tilde{P}_{0n}(j)$ is the nominal level of the *entire* MTP (e.g., gFWER or FDR) at which H_{0j} would just be rejected, given T_n. For continuous null distributions Q_0, $\tilde{P}_{0n}(j) = c_j^{-1}(T_n(j))$, where c_j^{-1} is the inverse of the monotone decreasing function $\alpha \to c_j(\alpha) = c_j(T_n, Q_0, \alpha)$.

The particular mapping c_j, defining the cut-offs $c_j(T_n, Q_0, \alpha)$, will depend on the choice of MTP (e.g., single-step vs. stepwise, common cut-offs vs. common-quantile cut-offs). For instance, the adjusted p-values for the classical Bonferroni procedure for FWER control are $\tilde{P}_{0n}(j) = \min(mP_{0n}(j), 1)$. Adjusted p-values for general single-step common-quantile and common-cut-off Procedures 1 and 2 are derived in Section 3.3. Dudoit et al. (2003) provide adjusted p-values for commonly-used FWER and FDR controlling procedures.

We now have two representations for an MTP, in terms of *cut-offs*, or *critical values*, $c_j = c_j(T_n, Q_0, \alpha)$, for the test statistics $T_n(j)$,

$$S(T_n, Q_0, \alpha) = \{j : T_n(j) > c_j\},$$

and in terms of *adjusted p-values*, $\tilde{P}_{0n}(j) = \tilde{P}(j, T_n, Q_0)$,

$$S(T_n, Q_0, \alpha) = \{j : \tilde{P}_{0n}(j) \leq \alpha\}.$$

That is, hypothesis H_{0j} is rejected at nominal Type I error rate α if $\tilde{P}_{0n}(j) \leq \alpha$. As in the single hypothesis case, an advantage of reporting adjusted p-values, as opposed to only rejection or not of the hypotheses, is that the level of the test does not need to be determined in advance, that is, results of the multiple testing procedure are provided for all α. Adjusted p-values are convenient and flexible summaries of the strength of the evidence against each null hypothesis, in terms of the Type I error rate for the entire MTP. Plots of sorted adjusted p-values allow scientists to examine various false positive rates (e.g., gFWER, FDR, or PCER) associated with different sets of rejected hypotheses. They do not require researchers to preselect a particular definition of Type I error rate or α-level, but rather provide them with tools to decide on an appropriate combination of number of rejections and tolerable false positive rate for a particular experiment and available resources.

2.5 Stepwise multiple testing procedures

One usually distinguishes among two main classes of multiple testing procedures, single-step and stepwise procedures, depending on whether the cut-off vector $c = (c_j : j = 1, \ldots, m)$ for the test statistics T_n is constant or random (given Q_0), i.e., is independent or not of these test statistics.

In *single-step* procedures, each hypothesis H_{0j} is evaluated using a critical value $c_j = c_j(Q_0, \alpha)$ that is independent of the results of the tests of other

hypotheses and is not a function of the data $X_1, \ldots, X_n$ (unless these data are used to estimate the null distribution Q_0, as in Section 4).

Improvement in power, while preserving (asymptotic) Type I error rate control, may be achieved by *stepwise procedures*, in which rejection of a particular hypothesis depends on the outcome of the tests of other hypotheses. That is, the cut-offs $c_j = c_j(T_n, Q_0, \alpha)$ are allowed to depend on the data, $X_1, \ldots, X_n$, via the test statistics T_n. In *step-down* procedures, the hypotheses corresponding to the *most significant* test statistics (i.e., largest absolute test statistics or smallest unadjusted p-values) are considered successively, with further tests depending on the outcome of earlier ones. As soon as one fails to reject a null hypothesis, no further hypotheses are rejected. In contrast, for *step-up* procedures, the hypotheses corresponding to the *least significant* test statistics are considered successively, again with further tests depending on the outcome of earlier ones. As soon as one hypothesis is rejected, all remaining more significant hypotheses are rejected.

Step-down and step-up analogues of the classical Bonferroni procedure are the Holm (1979) and Hochberg (1988) procedures, respectively. In these procedures, based solely on the marginal distributions of the test statistics (i.e., on the unadjusted p-values only), the unadjusted p-value for the hypothesis with the kth most significant test statistic is multiplied by $(m - k + 1) \leq m$ rather than m. Let $O_n(j)$ denote the indices for the ordered unadjusted p-values $P_{0n}(j)$, so that $P_{0n}(O_n(1)) \leq \ldots \leq P_{0n}(O_n(m))$. The adjusted p-values for hypothesis $H_{0,O_n(j)}$ are

$$\begin{aligned}
\tilde{P}_{0n}(O_n(j)) &= \min\big(mP_{0n}(O_n(j)), 1\big) \qquad \text{[Bonferroni]} \qquad (7)\\
\tilde{P}_{0n}(O_n(j)) &= \max_{k=1,\ldots,j} \Big\{\min\big((m-k+1)\, P_{0n}(O_n(k), 1\big)\Big\} \qquad \text{[Holm]}\\
\tilde{P}_{0n}(O_n(j)) &= \min_{k=j,\ldots,m} \Big\{\min\big((m-k+1)\, P_{0n}(O_n(k), 1\big)\Big\} \qquad \text{[Hochberg].}
\end{aligned}$$

The present article focusses on single-step procedures, while the companion article considers step-down procedures (van der Laan et al., 2003a). Commonly-used single-step and stepwise MTPs for control of the FWER and FDR are reviewed in Dudoit et al. (2003).

2.6 Type I error rate control and choice of null distribution

2.6.1 Type I error rate control

A multiple testing procedure $S_n = S(T_n, Q_0, \alpha)$ is said to provide *finite sample control* of the Type I error rate $\theta(F_{V_n})$ at *level* $\alpha \in (0,1)$, if $\theta(F_{V_n}) \leq \alpha$, where $V_n = |S_n \cap S_0|$ denotes the number of Type I errors. Similarly, the MTP provides *asymptotic control* of the Type I error rate at level α, if $\limsup_n \theta(F_{V_n}) \leq \alpha$.

Note that the random variable V_n, for the number of Type I errors, is defined by the *true distribution* $Q_n = Q_n(P)$ for the test statistics T_n, i.e., by their distribution under the true underlying data generating distribution P. In practice, however, the distribution $Q_n(P)$ is *unknown and estimated by a null distribution* Q_0, in order to derive cut-offs for each test statistic $T_n(j)$ (and the resulting adjusted p-values). The choice of a suitable null distribution Q_0 is crucial, in order to ensure that (finite sample or asymptotic) control of the Type I error rate under this assumed distribution does indeed provide the required control under the true distribution $Q_n(P)$.

2.6.2 Sketch of proposed approach to Type I error rate control

The following discussion highlights important considerations in choosing a null distribution Q_0 and motivates our general approach to the problem of Type I error rate control. Recall that the distribution, F_{V_n}, for the number of Type I errors, $V_n = |S_n \cap S_0| = |S(T_n, Q_0, \alpha) \cap S_0(P)|$, depends on the following: the true distribution $Q_n = Q_n(P)$ of the test statistics T_n, the null distribution Q_0 used to derive the m-vector of cut-offs $c_j(T_n, Q_0, \alpha)$ for these test statistics, and the set $S_0(P)$ of true null hypotheses. Type I error control is therefore a statement about the true unknown distribution P, via $Q_n(P)$ and $S_0(P)$.

When needed, we may use the following long notation for the number of rejected hypotheses and Type I errors, respectively, when $T_n \sim Q$,

$$\begin{aligned} R(Q_0 \mid Q) = R(S(T_n, Q_0, \alpha) \mid Q) &\equiv |S(T_n, Q_0, \alpha)|, \\ V(Q_0 \mid Q) = V(S(T_n, Q_0, \alpha) \mid Q) &\equiv |S(T_n, Q_0, \alpha) \cap S_0(P)|. \end{aligned} \tag{8}$$

This notation acknowledges that the distribution of the above quantities is defined in terms of a null distribution Q_0 (for deriving cut-offs) and a

distribution Q for the test statistics T_n (here, the subset S_0 is kept fixed at the truth $S_0(P)$ and the nominal level α of the test is also held fixed). For a given MTP, $S(T_n, Q_0, \alpha)$, also adopt the following short-hand notation for the number of rejected hypotheses and Type I errors

$$\begin{aligned} R_n &\equiv R(S(T_n, Q_0, \alpha) \mid Q_n), & R_0 &\equiv R(S(T_n, Q_0, \alpha) \mid Q_0), \\ V_n &\equiv V(S(T_n, Q_0, \alpha) \mid Q_n), & V_0 &\equiv V(S(T_n, Q_0, \alpha) \mid Q_0), \end{aligned} \tag{9}$$

where $Q_n = Q_n(P)$ refers to the actual (finite sample) joint distribution of the random m-vector of test statistics T_n, under the true data generating distribution P, and Q_0 refers to an assumed null distribution for these test statistics. Control of Type I error rates of the form $\theta(F_{V_n})$ can be achieved by the following three-step approach, which provides some intuition behind single-step Procedures 1 and 2, and the general characterization (Theorem 1) and explicit construction (Theorem 2) of a test statistics null distribution Q_0.

Three-step road map to Type I error rate control.

1. **Null domination conditions for Type I error rate.** For proper control of the Type I error rate $\theta(F_{V_n})$, for $T_n \sim Q_n(P)$, select a null distribution Q_0 such that

$$\begin{aligned} \theta(F_{V_n}) &\leq \theta(F_{V_0}) \qquad \text{[finite sample control]} \\ \limsup_{n\to\infty} \theta(F_{V_n}) &\leq \theta(F_{V_0}) \qquad \text{[asymptotic control]}. \end{aligned} \tag{10}$$

2. Note that the number of Type I errors is never greater than the total number of rejected hypotheses, i.e., $V_0 \leq R_0$, so that $F_{V_0} \geq F_{R_0}$, and hence, by monotonicity Assumption AMI,

$$\theta(F_{V_0}) \leq \theta(F_{R_0}).$$

3. Control the parameter $\theta(F_{R_0})$, corresponding to the *observed* number of rejected hypotheses R_0, under the null distribution Q_0, i.e., assuming $T_n \sim Q_0$,

$$\theta(F_{R_0}) \leq \alpha.$$

Combining Steps 1–3 provides the desired control of the Type I error rate

$\theta(F_{V_n})$ at level $\alpha \in (0,1)$, that is,

$$\begin{aligned} \theta(F_{V_n}) \le \theta(F_{V_0}) \le \theta(F_{R_0}) &\le \alpha \qquad \text{[finite sample control]} \\ \limsup_{n\to\infty} \theta(F_{V_n}) \le \theta(F_{V_0}) \le \theta(F_{R_0}) &\le \alpha \qquad \text{[asymptotic control]}. \end{aligned}$$

Note that such an approach is conservative in two ways: from controlling $\theta(F_{R_0}) \ge \theta(F_{V_0})$ and from the null domination in Step 1. The latter step is usually the most involved and requires a judicious choice for the null distribution Q_0. This article focuses on procedures that provide *asymptotic control* of the Type I error rate (i.e., such that $\limsup_n \theta(F_{V_n}) \le \alpha$) and provides a general characterization (Theorem 1) and an explicit construction (Theorem 2) for a null distribution Q_0 that satisfies the asymptotic null domination condition in Step 1.

2.6.3 Choice of null distribution and null domination conditions

The above θ-specific null domination conditions in Step 1 of the road map hold under the following general null domination conditions for the distribution F_{V_n}, of the number of Type I errors V_n, and for the joint distribution Q_{n,S_0}, of the S_0-specific subvector $(T_n(j) : j \in S_0)$ of test statistics. Thus, under the latter two conditions, the road map provides the required (finite sample or asymptotic) control of *any* Type I error rate of the form $\theta(F_{V_n})$.

Null domination conditions for number of Type I errors. For each $x \in \{0, \ldots, m\}$

$$\begin{aligned} F_{V_n}(x) &\ge F_{V_0}(x) \qquad \text{[finite sample control]} \\ \liminf_{n\to\infty} F_{V_n}(x) &\ge F_{V_0}(x) \qquad \text{[asymptotic control]}, \end{aligned} \tag{11}$$

that is, the number of Type I errors, V_0, under the null distribution Q_0, is stochastically greater than the number of Type I errors, V_n, under the true distribution $Q_n = Q_n(P)$ for the test statistics T_n. In particular, (11) holds under the following null domination property for the joint distribution of the test statistics $(T_n(j) : j \in S_0)$.

Null domination conditions for test statistics $(T_n(j) : j \in S_0)$. The distribution of the S_0-specific subvector $(T_n(j) : j \in S_0)$ for $T_n \sim Q_0$ equals

or dominates the corresponding distribution for $T_n \sim Q_n = Q_n(P)$,

$$\begin{aligned} Q_{n,S_0} &\geq Q_{0,S_0} \qquad \text{[finite sample control]} \\ \liminf_{n\to\infty} Q_{n,S_0} &\geq Q_{0,S_0} \qquad \text{[asymptotic control]}. \end{aligned} \tag{12}$$

That is, in the finite sample case,

$$Pr_{Q_n}(T_n(j) \leq t_j, j \in S_0) \geq Pr_{Q_0}(T_n(j) \leq t_j, j \in S_0), \qquad \forall t_j \in I\!R, \quad j \in S_0.$$

For finite sample control, the null domination condition in Step 1 then follows by monotonicity Assumption AMI. For asymptotic control, one relies also on uniform continuity Assumption ACI. Note that null domination is a statement about the distribution of the test statistics $(T_n(j) : j \in S_0)$ corresponding only to the *true* null hypotheses. More specific (i.e., less stringent) forms of null domination can be derived for given definitions of the Type I error rate $\theta(F_{V_n})$ (cf. FWER control in van der Laan et al. (2003a)).

One of the main contributions of this and the companion article (van der Laan et al., 2003a) is the general characterization (Theorem 1) and explicit construction (Theorem 2) of a proper null distribution Q_0 for the test statistics T_n. Procedures based on such a distribution provide asymptotic control of arbitrary Type I error rates $\theta(F_{V_n})$, for testing null hypotheses $H_{0j} = \mathrm{I}(P \in \mathcal{M}_j)$, corresponding to submodels $\mathcal{M}_j \subseteq \mathcal{M}$ for general data generating distributions P (i.e., distributions P with general dependence structures among variables). Our proposed test statistics null distribution Q_0 can be used in testing problems which cannot be handled by traditional approaches based on a data generating null distribution P_0 (e.g., tests of parameters of survival models, tests of pairwise correlations). The construction of the null distribution Q_0 in Theorem 2 is inspired by null domination condition (12), for the test statistics T_n: Q_0 is defined as the limit distribution of a sequence of random variables that are stochastically greater than the test statistics for the true null hypotheses. The resulting null distribution therefore satisfies asymptotic null domination condition (11), for the number of Type I errors, and also θ-specific asymptotic null domination condition (10) in Step 1 of the road map, for any Type I error rate mapping $\theta(\cdot)$.

2.6.4 Contrast with other approaches

As detailed in Section 6, the following two main points distinguish our approach and that of Pollard and van der Laan (2003) from existing ap-

proaches to Type I error rate control, such as those discussed in Hochberg and Tamhane (1987) and Westfall and Young (1993).

Firstly, we are only concerned with control of the Type I error rate under the *true* data generating distribution P, i.e., under the joint distribution $Q_n = Q_n(P)$ of the test statistics T_n implied by P. The notions of weak and strong control are therefore irrelevant to our approach. In particular, our notion of null domination differs from that of subset pivotality (see Westfall and Young (1993), p. 42–43) in the following senses: (i) null domination is concerned only with the true data generating distribution P, i.e., it considers only the subset $S_0(P)$ of true null hypotheses and not all possible 2^m subsets $\Lambda_0 \subseteq \{1, \ldots, m\}$ of null hypotheses, and (ii) null domination does not require equality of the joint distributions Q_0 and $Q_n(P)$ for the S_0-specific test statistics, but the weaker domination.

Secondly, we propose a null distribution for the test statistics ($T_n \sim Q_0$) rather than a data generating null distribution ($X \sim P_0$). A common choice of data generating null distribution P_0 in multiple testing procedures is one that satisfies the *complete null hypothesis*, $H_0^C \equiv \prod_{j=1}^m H_{0j} = \prod_{j=1}^m \mathrm{I}(P \in \mathcal{M}_j)$, that all m null hypotheses are true, i.e., $P_0 \in \cap_{j=1}^m \mathcal{M}_j$. The data generating null distribution P_0 then implies a null distribution $Q_n(P_0)$ for the test statistics. As discussed in Pollard and van der Laan (2003), procedures based on $Q_n(P_0)$ do not necessarily provide proper (asymptotic) control under the true distribution P, as the assumed null distribution $Q_n(P_0)$ and the true distribution $Q_n(P)$ for the test statistics T_n may have different limits and, as a result, violate the required null domination condition for the Type I error rates (equation (10), p. 17, in Step 1 of the road map). For instance, for test statistics with Gaussian asymptotic distribution (Section 5.1), the correlation matrices for the subvector of test statistics $(T_n(j) : j \in S_0)$ may be different under the true distribution P and the assumed complete null distribution P_0: $\rho_{S_0}(P_0) \neq \rho_{S_0}(P)$. In fact, in many testing problems, there does not even exist a data generating null distribution $P_0 \in \cap_{j=1}^m \mathcal{M}_j$ that correctly specifies a joint distribution for the test statistics such that the required null domination condition is satisfied.

Thus, unlike current procedures which can only be applied to a limited set of multiple testing problems, our proposed test statistics null distribution leads to single-step and step-down procedures that provide the desired asymptotic Type I error rate control in general testing problems.

3 Single-step procedures for control of general Type I error rates

In this section, we propose *single-step common-quantile* and *single-step common-cut-off* procedures for controlling general Type I error rates that are defined as parameters $\theta(F_{V_n})$ of the distribution of the number of Type I errors (Procedures 1 and 2, respectively). The methods are based on a null distribution Q_0 for the test statistics T_n, such as the one proposed in Theorem 2, and are a generalization of the procedures discussed in Pollard and van der Laan (2003) for single-parameter null hypotheses. As shown in Section 3.3.3, our general approach includes as special cases, for control of the FWER, the *single-step minP* and *maxT* procedures based on minima of unadjusted p-values and maxima of test statistics, respectively (Dudoit et al., 2003; Westfall and Young, 1993). Before discussing the details of the procedures and presenting proofs of Type I error rate control, we first outline our proposed methods and main results.

Given a null distribution Q_0 and nominal level α, single-step common-quantile Procedure 1 rejects null hypothesis H_{0j}, $j = 1, \cdots, m$, provided the test statistic $T_n(j)$ is greater than the $\delta_0(\alpha)$–quantile, $d_j(Q_0, \delta_0(\alpha))$, of the marginal distribution Q_{0j}. For control of the Type I error rate $\theta(F_{V_n})$ at level α, $\delta_0(\alpha)$ is chosen so that the corresponding parameter $\theta(F_{R_0})$, for the distribution of the *observed* number of rejections R_0 under Q_0, is bounded by α. In the simpler common-cut-off Procedure 2, H_{0j} is rejected if $T_n(j)$ is greater than a common cut-off $e(Q_0, \alpha)$ chosen so that $\theta(F_{R_0}) \leq \alpha$.

Theorem 1 proves that Procedures 1 and 2 provide asymptotic control of the Type I error rate $\theta(F_{V_n})$, under the following asymptotic null domination condition concerning the joint distribution of the test statistics $(T_n(j) : j \in S_0)$ for the *true* null hypotheses (Assumption AQ0): In the limit, the number of Type I errors, V_n, under the true distribution $Q_n = Q_n(P)$ for the test statistics T_n, is stochastically smaller than the corresponding number of Type I errors, V_0, under the assumed null distribution Q_0, i.e., $\liminf_n F_{V_n}(x) \geq F_{V_0}(x)$, $\forall x \in \{0, \ldots, m\}$. From uniform continuity Assumption ACI, one can show that $\limsup_n \theta(F_{V_n}) \leq \theta(F_{V_0})$. Asymptotic control of the Type I error rate then follows as sketched in the three-step road map on p. 17. As illustrated in Section 3.3.3, the procedure can be applied to control any error rate that is a function of the distribution for the number of Type I errors, V_n, including the usual FWER, PCER, and also the gFWER. The key issue

is the choice of a suitable null distribution Q_0.

Theorem 2 provides an explicit proposal for a null distribution that satisfies the null domination condition of Theorem 1. This null distribution Q_0 is defined as the limit distribution of the null-value shifted and scaled test statistics: $Z_n(j) \equiv \nu_{0n}(j)(T_n(j) + \lambda_0(j) - E[T_n(j)])$. In this construction, the null-values $\lambda_0(j)$ are such that the limit distribution of Z_n is stochastically larger than that of T_n for the true null hypotheses ($j \in S_0$); hence, key Assumption AQ0 is satisfied. The $\nu_{0n}(j)$ are chosen to prevent a degenerate limit for the false null hypotheses ($j \notin S_0$).

3.1 General procedures and asymptotic control results

According to single-step Procedures 1 and 2, hypothesis H_{0j} is rejected at MTP level α if $T_n(j) > c_j$, where $c_j = c_j(Q_0, \alpha)$, $j = 1, \ldots, m$, are defined as either common quantiles for the margins Q_{0j} of a null distribution Q_0 (Procedure 1) or common cut-offs (Procedure 2). For an m-vector of cut-offs, $c = (c_j : j = 1, \ldots, m)$, and for $T_n \sim Q$, denote the number of rejected hypotheses and Type I errors by

$$R(c \mid Q) \equiv \sum_{j=1}^{m} \mathrm{I}(T_n(j) > c_j) \quad \text{and} \quad V(c \mid Q) \equiv \sum_{j \in S_0} \mathrm{I}(T_n(j) > c_j), \tag{13}$$

respectively. For a given cut-off vector c, and as in Section 2.6, adopt the following short-hand notation

$$\begin{aligned} R_n &\equiv R(c \mid Q_n), & R_0 &\equiv R(c \mid Q_0), \\ V_n &\equiv V(c \mid Q_n), & V_0 &\equiv V(c \mid Q_0), \end{aligned} \tag{14}$$

where $Q_n = Q_n(P)$ refers to the actual (finite sample) joint distribution of the test statistics T_n, under the true data generating distribution P, and Q_0 refers to an assumed null distribution for these test statistics.

3.1.1 Single-step common-quantile procedure

Procedure 1. Single-step common-quantile procedure for control of general Type I error rates $\theta(F_{V_n})$.
Given an m-variate null distribution Q_0 and $\delta \in [0,1]$, define an m-vector, $d(Q_0,\delta) = (d_j(Q_0,\delta) : j = 1,\ldots,m)$, of δ–quantiles,

$$d_j(Q_0,\delta) \equiv Q_{0j}^{-1}(\delta) = \inf\{z : Q_{0j}(z) \geq \delta\}, \qquad j = 1,\ldots,m, \tag{15}$$

where Q_{0j} denote the marginal cumulative distribution functions corresponding to Q_0. For a test of level $\alpha \in (0,1)$, choose δ as

$$\delta_0(\alpha) \equiv \inf\{\delta : \theta(F_{R(d(Q_0,\delta)|Q_0)}) \leq \alpha\}, \tag{16}$$

where we recall that $R(d(Q_0,\delta)|Q_0)$ denotes the number of rejected hypotheses for common-quantile cut-offs $d(Q_0,\delta)$, under the null distribution Q_0 for the test statistics T_n. The *single-step common-quantile* multiple testing procedure for controlling the Type I error rate $\theta(F_{V_n})$ at level α is defined in terms of the common-quantile cut-offs, $c(Q_0,\alpha) \equiv d(Q_0,\delta_0(\alpha))$, by the following rule.

Reject H_{0j} if $T_n(j) > d_j(Q_0,\delta_0(\alpha))$, $j = 1,\ldots,m$,

that is,

$$S(T_n,Q_0,\alpha) \equiv \big\{j : T_n(j) > d_j(Q_0,\delta_0(\alpha))\big\}.$$

Here, F_{V_n} denotes the c.d.f. for the number of Type I errors, $V_n \equiv V(d(Q_0,\delta_0(\alpha)) \mid Q_n)$, under the true distribution $Q_n = Q_n(P)$ for the test statistics T_n.

Theorem 1 [Asymptotic control of Type I error rate for single-step common-quantile Procedure 1] *Assume that there exists a random m-vector $Z \sim Q_0 = Q_0(P)$, so that, for all $c = (c_j : j = 1,\ldots,m) \in \mathbb{R}^m$ and $x \in \{0,\ldots,m\}$, the joint distribution $Q_n = Q_n(P)$ of the test statistics T_n satisfies the following asymptotic null domination property with respect to Q_0*

$$\liminf_{n\to\infty} Pr_{Q_n}\left(\sum_{j\in S_0} \mathrm{I}(T_n(j) > c_j) \leq x\right) \geq Pr_{Q_0}\left(\sum_{j\in S_0} \mathrm{I}(Z(j) > c_j) \leq x\right). \tag{AQ0}$$

In other words, the number of Type I errors, V_n, under the true distribution $Q_n = Q_n(P)$ for the test statistics T_n, is stochastically smaller in the

limit than the corresponding number of Type I errors, V_0, under the null distribution Q_0: $\liminf_n F_{V_n}(x) \geq F_{V_0}(x)\ \forall x$. *In addition, suppose that the mapping $\theta(\cdot)$ defining the Type I error rate is such that Assumptions AMI and ACI hold. Then, single-step Procedure 1, with common-quantile cut-offs $c(Q_0, \alpha) = d(Q_0, \delta_0(\alpha))$, provides asymptotic control of the Type I error rate $\theta(F_{V_n})$ at level α, that is,*

$$\limsup_{n\to\infty} \theta(F_{V_n}) \leq \alpha,$$

where V_n denotes the number of Type I errors for $T_n \sim Q_n(P)$

$$V_n \equiv V(c(Q_0, \alpha) \mid Q_n) = \sum_{j \in S_0} \mathrm{I}(T_n(j) > c_j(Q_0, \alpha)).$$

Note that the main null domination Assumption AQ0 (also in equation (11), above) only concerns the joint distribution of the subvector $(T_n(j) : j \in S_0)$ of test statistics for the *true* null hypotheses. An explicit construction for a null distribution Q_0 that satisfies Assumption AQ0 is proposed in Theorem 2, below. As discussed in Section 4, estimation of Q_0 with the non-parametric or model-based bootstrap results in an estimator Q_{0n} and corresponding estimated cut-off vector $c(Q_{0n}, \alpha)$, such that $\limsup_n \theta(F_{V(c(Q_{0n},\alpha)|Q_n)}) \leq \alpha$.

Proof of Theorem 1. Since $V_0 \leq R_0$, Q_0-a.s., then $F_{V_0}(x) \geq F_{R_0}(x)\ \forall x$. Hence, by monotonicity Assumption AMI and by definition of the cut-offs $c(Q_0, \alpha) = d(Q_0, \delta_0(\alpha))$, so that $\theta(F_{R_0}) \leq \alpha$, we have

$$\theta(F_{V_0}) \leq \theta(F_{R_0}) \leq \alpha. \tag{17}$$

Rewrite F_{V_n} as

$$F_{V_n} = F_{V_0} + (F_{V_n} - F_{V_0}) \geq F_{V_0} + \min(0, F_{V_n} - F_{V_0})$$

and again apply monotonicity Assumption AMI to show that

$$\theta(F_{V_n}) \leq \theta(F_{V_0} + \min(0, F_{V_n} - F_{V_0})).$$

Now, by the main null domination Assumption AQ0, $\liminf_n F_{V_n}(x) \geq F_{V_0}(x)$ $\forall x$, so that

$$\lim_{n\to\infty} (F_{V_0}(x) + \min(0, F_{V_n}(x) - F_{V_0}(x))) = F_{V_0}(x) \qquad \forall x,$$

and by uniform continuity Assumption ACI,

$$\limsup_{n\to\infty} \theta(F_{V_n}) \leq \lim_{n\to\infty} \theta(F_{V_0} + \min(0, F_{V_n} - F_{V_0})) = \theta(F_{V_0}). \tag{18}$$

Finally, combining (17) and (18), we get the desired asymptotic control of the Type I error rate

$$\limsup_{n\to\infty} \theta(F_{V_n}) \leq \alpha.$$

□

3.1.2 Single-step common-cut-off procedure

Note that one can readily define a *common-cut-off* analogue of Procedure 1 as follows. Asymptotic control of the Type I error rate $\theta(F_{V_n})$ by Procedure 2 also follows from the proof of Theorem 1.

Procedure 2. Single-step common-cut-off procedure for control of general Type I error rates $\theta(F_{V_n})$.
Given an m-variate null distribution Q_0 and for a test of level $\alpha \in (0,1)$, define a common cut-off $e(Q_0, \alpha)$, such that

$$e(Q_0, \alpha) \equiv \inf\{c : \theta(F_{R((c,\ldots,c)|Q_0)}) \leq \alpha\}, \tag{19}$$

where we recall that $R((c,\ldots,c)|Q_0)$ denotes the number of rejected hypotheses for common cut-off c, under the null distribution Q_0 for the test statistics T_n. The *single-step common-cut-off* multiple testing procedure for controlling the Type I error rate $\theta(F_{V_n})$ at level α is defined in terms of the common cut-offs, $c(Q_0, \alpha) = (e(Q_0, \alpha), \ldots, e(Q_0, \alpha))$, by the following rule.

$$\text{Reject } H_{0j} \text{ if } T_n(j) > e(Q_0, \alpha), \quad j = 1, \ldots, m,$$

that is,

$$S(T_n, Q_0, \alpha) \equiv \big\{j : T_n(j) > e(Q_0, \alpha)\big\}.$$

Here, F_{V_n} denotes the c.d.f. for the number of Type I errors, $V_n \equiv V((e(Q_0, \alpha), \ldots, e(Q_0, \alpha)) \mid Q_n)$, under the true distribution $Q_n = Q_n(P)$ for the test statistics T_n.

3.1.3 Common-quantile vs. common-cut-off procedures

As discussed in Section 3.3.3, for control of the FWER, single-step common-quantile Procedure 1 and common-cut-off Procedure 2 reduce to the single-step minP and maxT procedures, based on minima of unadjusted p-values

and maxima of test statistics, respectively (Dudoit et al., 2003; Westfall and Young, 1993). Procedures based on common cut-offs and common quantiles are equivalent when the test statistics $T_n(j)$, $j = 1, \ldots, m$, are identically distributed under Q_0, i.e., when the marginal distributions Q_{0j} do not depend on j: in this case, the significance rankings based on test statistics $T_n(j)$ and p-values $P_{0n}(j) = \bar{Q}_{0j}(T_n(j))$ coincide. In general, however, the two types of procedures produce different results, and considerations of balance, power, and computational feasibility should dictate the choice between the two approaches. In the case of non-identically distributed test statistics $T_n(j)$, not all tests are weighted equally in common-cut-off procedures and this can lead to unbalanced adjustments (Beran, 1988; Westfall and Young, 1993; Westfall, 2003). When the null distribution Q_0 is estimated by resampling (e.g., bootstrap, permutation), quantile-based procedures (Procedure 1, minP procedure for FWER control) tend to be sensitive to the number of resampling steps and to the discreteness of the estimated null distribution. This can result in more conservative procedures than those based directly on the test statistics (Procedure 2, maxT procedure for FWER control). Also, quantile-based procedures require more computation, because the unadjusted p-values $P_{0n}(j)$ must be estimated before one can consider their joint distribution. The reader is referred to Dudoit et al. (2003), Ge et al. (2003), and Pollard and van der Laan (2003) for further discussion of the relative merits of common-quantile vs. common-cut-off procedures.

3.2 Explicit proposal for the test statistics null distribution

One can make the following explicit proposal for the null distribution Q_0 used to derive cut-offs in Procedures 1 and 2.

Theorem 2 [General construction for null distribution Q_0] *Suppose there exists known m-vectors $\lambda_0 \in \mathbb{R}^m$ and $\tau_0 \in \mathbb{R}^{+m}$ of null-values, so that*

$$\begin{aligned} \limsup_{n\to\infty} E[T_n(j)] &\leq \lambda_0(j) \qquad \textit{and} \\ \limsup_{n\to\infty} Var[T_n(j)] &\leq \tau_0(j), \qquad \textit{for } j \in S_0. \end{aligned} \tag{20}$$

Let

$$\nu_{0n}(j) \equiv \sqrt{\min\left(1, \frac{\tau_0(j)}{Var[T_n(j)]}\right)} \tag{21}$$

and define an m-vector Z_n by

$$Z_n(j) \equiv \nu_{0n}(j)\Big(T_n(j) + \lambda_0(j) - E[T_n(j)]\Big), \qquad j = 1, \ldots, m. \tag{22}$$

Suppose that

$$Z_n \stackrel{\mathcal{L}}{\Rightarrow} Z \sim Q_0(P). \tag{23}$$

Then, for this choice of null distribution $Q_0 = Q_0(P)$, and for all $c = (c_j : j = 1, \ldots, m) \in I\!R^m$ and $x \in \{0, \ldots, m\}$,

$$\liminf_{n\to\infty} Pr_{Q_n}\left(\sum_{j\in S_0} \mathrm{I}(T_n(j) > c_j) \leq x\right) \geq Pr_{Q_0}\left(\sum_{j\in S_0} \mathrm{I}(Z(j) > c_j) \leq x\right),$$

so that asymptotic null domination Assumption AQ0 in Theorem 1 holds.

The asymptotic distribution $Q_0 = Q_0(P)$, of the null-value shifted and scaled test statistics Z_n, generalizes the null distribution of Pollard and van der Laan (2003) for single-parameter null hypotheses. In the later case, the null distribution Q_0 turns out to be a Gaussian distribution with mean vector zero (Section 5.1). The purpose of the null-values $\lambda_0(j)$ is to generate test statistics $(Z_n(j) : j \in S_0)$ that are stochastically larger in the limit than the original test statistics $(T_n(j) : j \in S_0)$, thereby resulting in a distribution Q_0 that satisfies key Assumption AQ0 of asymptotic null domination for the number of Type I errors. In contrast, the use of the scaling factors $\tau_0(j)$ in the construction of Z_n is not needed for proving control of the Type I error rate. The purpose of $\tau_0(j)$ is to avoid having a degenerate asymptotic null distribution and infinite cut-offs for the true positives ($j \notin S_0$), an important property for power considerations. This scaling is needed in particular for F-statistics which have asymptotically infinite means and variances under non-local alternative hypotheses (Section 5.2). Note that the null-values $\lambda_0(j)$ and $\tau_0(j)$ only depend on the marginal distributions of the test statistics $T_n(j)$ under the true null hypotheses and are generally known from univariate testing. For example, as discussed in Section 5, $\lambda_0(j) \equiv 0$ and $\tau_0(j) \equiv 1$ for t-statistics, and $\lambda_0(j) \equiv 1$ and $\tau_0(j) \equiv 2/(K-1)$ for F-statistics comparing K population means under the assumption of constant variances in the different populations. The null-values $\lambda_0(j)$ and $\tau_0(j)$ can possibly depend on the unknown data generating distribution P, as is the case for F-statistics for unequal population variances. In such a situation, one can replace the parameters $\lambda_0(j)$ and $\tau_0(j)$ by consistent estimators thereof.

In practice, one can estimate the null distribution Q_0 and the resulting single-step cut-offs using bootstrap Procedures 3–5, as discussed in detail in Section 4. For B bootstrap samples, one has an $m \times B$ matrix of test statistics, $\mathbf{T} = \big(T_n^b(j)\big)$, with rows corresponding to the m hypotheses and columns to the B bootstrap samples. The expected values, $E[T_n(j)]$, and variances, $Var[T_n(j)]$, are estimated by simply taking row means and variances of the matrix $\mathbf{T}$. The matrix of test statistics $\mathbf{T}$ can then be row-shifted and scaled using the supplied null-values $\lambda_0(j)$ and $\tau_0(j)$, to produce an $m \times B$ matrix $\mathbf{Z} = \big(Z_n^b(j)\big)$. The null distribution Q_0 is estimated by the empirical distribution of the columns of matrix $\mathbf{Z}$.

Proof of Theorem 2. Define an intermediate random vector $(\tilde{Z}_n(j) : j \in S_0)$, for the true null hypotheses, by

$$\tilde{Z}_n(j) \equiv T_n(j) + \max(0, \lambda_0(j) - E[T_n(j)]), \qquad j \in S_0.$$

Then, $T_n(j) \leq \tilde{Z}_n(j)$. In addition, since $\limsup_n E[T_n(j)] \leq \lambda_0(j)$ and $\limsup_n Var[T_n(j)] \leq \tau_0(j)$ for $j \in S_0$ (and thus $\lim_n \nu_{0n}(j) = 1$), it follows that $(\tilde{Z}_n(j) : j \in S_0)$ and $(Z_n(j) : j \in S_0)$ have the same limit distribution

$$(\tilde{Z}_n(j) : j \in S_0) \stackrel{\mathcal{L}}{\Rightarrow} (Z(j) : j \in S_0) \sim Q_{0,S_0}.$$

Thus,

$$\begin{aligned} \liminf_{n\to\infty} Pr\left(\sum_{j\in S_0} \mathrm{I}(T_n(j) > c_j) \leq x\right) &\geq \liminf_{n\to\infty} Pr\left(\sum_{j\in S_0} \mathrm{I}(\tilde{Z}_n(j) > c_j) \leq x\right) \\ &= Pr\left(\sum_{j\in S_0} \mathrm{I}(Z(j) > c_j) \leq x\right). \end{aligned}$$

□

3.3 Adjusted p-values

Rather than simply reporting rejection or not of a subset of null hypotheses at a prespecified level α, one can report *adjusted p-values* for single-step Procedures 1 and 2, computed under the assumed null distribution Q_0 for the test statistics T_n. While the definition of adjusted p-value in equation (6) of Section 2.4 holds for general null distributions, in this section, we consider

for simplicity a null distribution Q_0 with continuous and strictly monotone marginal c.d.f.'s, Q_{0j}, and survivor functions, $\bar{Q}_{0j} = 1 - Q_{0j}$, $j = 1, \ldots, m$. We first provide adjusted p-values for general Type I error rates $\theta(F_{V_n})$. Adjusted p-values for commonly-used MTPs can be shown to correspond to particular choices for the Type I error rate mapping $\theta(\cdot)$. Explicit formulae for adjusted p-values for PCER- and gFWER-controlling MTPs are given in Sections 3.3.2 and 3.3.3, respectively.

3.3.1 General Type I error rates

Result 1 [Adjusted p-values for common-quantile Procedure 1] *The adjusted p-values for single-step common-quantile Procedure 1, based on a null distribution Q_0 with continuous and strictly monotone marginal distributions, are given by*

$$\begin{aligned} & \tilde{P}_{0n}(j) &&= \theta(F_{R(d(Q_0, 1-P_{0n}(j))|Q_0)}) \qquad (24) \\ \textit{where} \quad & P_{0n}(j) &&= \bar{Q}_{0j}(T_n(j)) = 1 - Q_{0j}(T_n(j)), \qquad j = 1, \ldots, m. \end{aligned}$$

Here, $P_{0n}(j)$ is the unadjusted p-value for hypothesis H_{0j} under Q_0 and the common-quantile cut-offs are

$$d_h(Q_0, 1 - P_{0n}(j)) = Q_{0h}^{-1}(1 - P_{0n}(j)) = \bar{Q}_{0h}^{-1}(P_{0n}(j)) = \bar{Q}_{0h}^{-1}(\bar{Q}_{0j}(T_n(j))). \qquad (25)$$

In particular, for $h = j$, $d_j(Q_0, 1 - P_{0n}(j)) = T_n(j)$. Procedure 1 for controlling the Type I error rate $\theta(F_{V_n})$ at level α can then be stated equivalently as

$$S(T_n, Q_0, \alpha) = \big\{ j : \tilde{P}_{0n}(j) \leq \alpha \big\}.$$

Proof of Result 1. The common-quantile cut-offs in Procedure 1 can be represented as

$$d_j(Q_0, \delta_0(\alpha)) = Q_{0j}^{-1}(\delta_0(\alpha)) = Q_{0j}^{-1}(\theta_0^{-1}(\alpha)),$$

where θ_0^{-1} is the inverse of the monotone decreasing function $\delta \to \theta_0(\delta) \equiv \theta(F_{R(d(Q_0,\delta)|Q_0)})$, that is, $\theta_0^{-1}(\alpha) = \inf\{\delta : \theta_0(\delta) \leq \alpha\}$. Then,

$$\begin{aligned}
\tilde{P}_{0n}(j) &= \inf\{\alpha \in [0,1] : d_j(Q_0, \delta_0(\alpha)) < T_n(j)\} \\
&= \inf\{\alpha \in [0,1] : Q_{0j}^{-1}(\theta_0^{-1}(\alpha)) < T_n(j)\} \\
&= \inf\{\alpha \in [0,1] : \theta_0^{-1}(\alpha) \leq Q_{0j}(T_n(j))\} \\
&= \inf\{\alpha \in [0,1] : \alpha \geq \theta_0(Q_{0j}(T_n(j)))\} \\
&= \theta_0(Q_{0j}(T_n(j))) \\
&= \theta_0(1 - P_{0n}(j)) \\
&= \theta(F_{R(d(Q_0,1-P_{0n}(j))|Q_0)}).
\end{aligned}$$

□

Result 2 [Adjusted p-values for common-cut-off Procedure 2] *The adjusted p-values for single-step common-cut-off Procedure 2, based on a null distribution Q_0 with continuous and strictly monotone marginal distributions, are given by*

$$\tilde{P}_{0n}(j) = \theta(F_{R((T_n(j),\ldots,T_n(j))|Q_0)}), \qquad j = 1, \ldots, m. \tag{26}$$

The proof of this result is similar to that for common-quantile adjusted p-values in Result 1 and is therefore omitted.

3.3.2 Application to control of the PCER

Result 3 [Adjusted p-values for common-quantile Procedure 1 for PCER control] *For control of the PCER, the adjusted p-values for single-step common-quantile Procedure 1, based on a null distribution Q_0 with continuous and strictly monotone marginal distributions, reduce to the unadjusted p-values*

$$\tilde{P}_{0n}(j) = P_{0n}(j) = \bar{Q}_{0j}(T_n(j)), \qquad j = 1, \ldots, m. \tag{27}$$

Proof of Result 3. Let $Z \sim Q_0$ and consider the Type I error rate mapping $\theta(F) = \int x dF(x)/m$. Then,

$$\begin{aligned}
\tilde{P}_{0n}(j) &= \theta(F_{R(d(Q_0, 1-P_{0n}(j))|Q_0)}) \\
&= \frac{1}{m}\sum_{h=1}^{m} Pr_{Q_0}\big(Z(h) > \bar{Q}_{0h}^{-1}(P_{0n}(j))\big) \\
&= \frac{1}{m}\sum_{h=1}^{m} \bar{Q}_{0h}\big(\bar{Q}_{0h}^{-1}(P_{0n}(j))\big) \\
&= P_{0n}(j).
\end{aligned}$$

□

Result 4 [Adjusted p-values for common-cut-off Procedure 2 for PCER control] *For control of the PCER, the adjusted p-values for single-step common-cut-off Procedure 2, based on a null distribution Q_0 with continuous and strictly monotone marginal distributions, are given by*

$$\tilde{P}_{0n}(j) = \frac{1}{m}\sum_{h=1}^{m} \bar{Q}_{0h}\big(T_n(j)\big), \qquad j = 1, \ldots, m. \tag{28}$$

Proof of Result 4. Let $Z \sim Q_0$ and consider the Type I error rate mapping $\theta(F) = \int x dF(x)/m$. Then,

$$\begin{aligned}
\tilde{P}_{0n}(j) &= \theta(F_{R((T_n(j),\ldots,T_n(j))|Q_0)}) \\
&= \frac{1}{m}\sum_{h=1}^{m} Pr_{Q_0}\big(Z(h) > T_n(j)\big) \\
&= \frac{1}{m}\sum_{h=1}^{m} \bar{Q}_{0h}\big(T_n(j)\big).
\end{aligned}$$

□

3.3.3 Application to control of the gFWER

Result 5 [Adjusted p-values for common-quantile Procedure 1 for gFWER control] *For control of the gFWER, the adjusted p-values for*

single-step common-quantile Procedure 1, based on a null distribution Q_0 with continuous and strictly monotone marginal distributions, are given by

$$\tilde{P}_{0n}(j) = Pr_{Q_0}\left(P_0^\circ(k+1) \leq P_{0n}(j)\right), \qquad j = 1, \ldots, m. \tag{29}$$

Here, $P_0(j) \equiv \bar{Q}_{0j}(Z(j))$ denote unadjusted p-values under the test statistics null distribution Q_0, i.e., for $Z(j) \sim Q_{0j}$, and $P_0^\circ(j)$ denote the corresponding ordered unadjusted p-values, so that $P_0^\circ(1) \leq \ldots \leq P_0^\circ(m)$.

Proof of Result 5. Let $Z \sim Q_0$ and consider the Type I error rate mapping $\theta(F) = 1 - F(k)$. Then,

$$\begin{aligned}
\tilde{P}_{0n}(j) &= \theta(F_{R(d(Q_0, 1-P_{0n}(j))|Q_0)}) \\
&= Pr_{Q_0}\left(\sum_{h=1}^{m} \mathrm{I}\Big(Z(h) > d_h(Q_0, 1 - P_{0n}(j))\Big) > k\right) \\
&= Pr_{Q_0}\left(\sum_{h=1}^{m} \mathrm{I}\Big(\bar{Q}_{0h}(Z(h)) \leq \bar{Q}_{0h}(\bar{Q}_{0h}^{-1}(P_{0n}(j)))\Big) > k\right) \\
&= Pr_{Q_0}\left(\sum_{h=1}^{m} \mathrm{I}\Big(P_0(h) \leq P_{0n}(j)\Big) > k\right) \\
&= Pr_{Q_0}\left(P_0^\circ(k+1) \leq P_{0n}(j)\right).
\end{aligned}$$

□

Procedure 1 for control of the gFWER is thus based on the distribution, under Q_0, of the $(k+1)$*st ordered unadjusted p-value*, $P_0^\circ(k+1)$, and $(1-\delta_0(\alpha))$ is chosen as the α–quantile of the distribution of $P_0^\circ(k+1)$. In the special case of FWER control ($k = 0$), the procedure is based on the distribution of the *minimum* of the m unadjusted p-values, i.e., on $P_0^\circ(1) = \min_{j\in\{1,\ldots,m\}} P_0(j)$. Thus, for FWER control, the single-step minP procedure discussed in Dudoit et al. (2003) and Westfall and Young (1993) corresponds to Procedure 1 with $(1-\delta_0(\alpha))$ chosen as the α–quantile of the distribution of the minimum marginal p-value $P_0^\circ(1)$.

Consider the special case where the random vector $Z \sim Q_0$ has *independent* components $Z(j)$, with continuous marginal distributions Q_{0j}, $j = 1, \ldots m$. Then, $P_0(j) = \bar{Q}_{0j}(Z(j))$ are independent $U(0,1)$ random

variables and the $(k+1)$st ordered unadjusted p-value $P_0^\circ(k+1)$ has a $Beta(k+1, m-k)$ distribution. Thus, in this independence situation, the single-step procedure for control of the gFWER is very simple and is based only on the marginal null distributions, Q_{0j}. Adjusted p-values are

$$\tilde{P}_{0n}(j) = \frac{\Gamma(m+1)}{\Gamma(k+1)\Gamma(m-k)} \int_0^{P_{0n}(j)} z^k (1-z)^{m-k-1} dz, \tag{30}$$

where $\Gamma(i) = (i-1)!$ for a positive integer i. In particular, for FWER control $(k=0)$, the adjusted p-values for Procedure 1 reduce to the *single-step Šidák adjusted p-values* (Dudoit et al., 2003)

$$\tilde{P}_{0n}(j) = 1 - (1 - P_{0n}(j))^m . \tag{31}$$

Result 6 [Adjusted p-values for common-cut-off Procedure 2 for gFWER control] *For control of the gFWER, the adjusted p-values for single-step common-cut-off Procedure 2, based on a null distribution Q_0 with continuous and strictly monotone marginal distributions, are given by*

$$\tilde{P}_{0n}(j) = Pr_{Q_0}\left(Z^\circ(k+1) > T_n(j)\right), \qquad j = 1, \ldots, m, \tag{32}$$

where $Z^\circ(j)$ denotes the jth ordered component of $Z = (Z(j) : j = 1, \ldots, m) \sim Q_0$, so that $Z^\circ(1) \geq \ldots \geq Z^\circ(m)$.

Proof of Result 6. Again, let $Z \sim Q_0$ and consider the Type I error rate mapping $\theta(F) = 1 - F(k)$. Then,

$$\begin{aligned} \tilde{P}_{0n}(j) &= \theta(F_{R((T_n(j),\ldots,T_n(j))|Q_0)}) \\ &= Pr_{Q_0}\left(\sum_{h=1}^{m} \mathrm{I}\Big(Z(h) > T_n(j)\Big) > k\right) \\ &= Pr_{Q_0}\left(Z^\circ(k+1) > T_n(j)\right). \end{aligned}$$

□

Thus, for control of the gFWER at level α, the common cut-off in Procedure 2 is chosen as the $(1-\alpha)$–quantile of the distribution of the $(k+1)st$ *ordered component* $Z^\circ(k+1)$ of $Z \sim Q_0$. In the special case of FWER control $(k=0)$, the procedure is based on the *maximum* of the m variables $Z(j)$, i.e., on $Z^\circ(1) = \max_{j\in\{1,\ldots,m\}} Z(j)$. Thus, for FWER control, the single-step

maxT procedure discussed in Dudoit et al. (2003) and Westfall and Young (1993) corresponds to Procedure 2 with common cut-off chosen as the $(1-\alpha)$–quantile of the distribution of the maximum $Z^{\circ}(1)$.

Control of gFWER via control of FWER. Finally, we note that one can control the gFWER using simple modifications of Procedures 1 and 2 for control of the FWER. The main idea is to follow the FWER-controlling procedure exactly until the first non-rejection and then reject the null hypotheses specified by this procedure as well as the k hypotheses corresponding to the next k most significant test statistics (i.e., the k hypotheses with the next k smallest adjusted p-values). Such an approach to gFWER control is appealing, because it only requires working with the FWER and it guarantees at least k rejected hypotheses. More details and formal results are provided in van der Laan et al. (2003b).

4 Bootstrap-based single-step procedures for control of general Type I error rates

In practice, since the data generating distribution P is unknown, then so is the null distribution $Q_0 = Q_0(P)$ defined in Theorem 2. Estimation of Q_0 is then needed, especially to deal with the unknown dependence structure among the test statistics. Estimators Q_{0n} of the null distribution can be obtained according to the following three main approaches.

- *Bootstrap null distribution.* As detailed in Section 4.2, non-parametric or model-based bootstrap procedures provide a very general approach for obtaining consistent estimators of the null distribution $Q_0(P)$ proposed in Theorem 2.

- *Test statistics specific null distribution.* For the test of single-parameter null hypotheses with t-statistics, the null distribution $Q_0 = Q_0(P)$ is the m-variate Gaussian distribution $N(0, \rho(P))$, where $\rho(P)$ is the correlation matrix of the vector influence curve, $IC(X \mid P) = (IC_j(X \mid P) : j = 1, \ldots, m)$, for an asymptotically linear estimator μ_n of the parameter vector μ (see Sections 2.1.4 and 5.1 for details). In this case, one can estimate Q_0 by $Q_{0n} = N(0, \rho_n)$, using a consistent estimator ρ_n of the correlation matrix $\rho(P)$, such as the correlation matrix

corresponding to the $m \times m$ estimated IC covariance matrix

$$\Sigma_n \equiv \frac{1}{n} \sum_{i=1}^{n} IC_n(X_i) IC_n^T(X_i),$$

where $IC_n(X) = (IC_{jn}(X) : j = 1, \ldots, m)$ is an estimator of the m-vector influence curve $IC(X \mid P)$. For simple parameters such as means, the influence curves can be derived straightforwardly. For example, for estimation of the mean vector $\mu = E[X]$, for a random m-vector $X \sim P$, the influence curves are $IC_j(X \mid P) = X(j) - \mu(j)$, and the corresponding estimated influence curves are $IC_{jn}(X) = X(j) - \mu_n(j)$, where $\mu_n(j) = \bar{X}_n(j)$ is the empirical mean for the jth component of the m-vector X, $j = 1, \ldots, m$. Then, ρ_n is simply the sample correlation matrix. Influence curves for estimators of correlations and regression parameters are given in Section 5.1. In cases where the influence curves are not readily available, $\rho(P)$ can be estimated with the bootstrap (Section 5.1.2). For the test of multiple-parameter null hypotheses using F-statistics (Section 5.2), a null distribution Q_0 can be defined as a simple quadratic function of K independent Gaussian m-vectors, $Y_k \sim N(0, \Sigma_k)$, $k = 1, \ldots, K$. An estimator Q_{0n} of the null distribution Q_0 can be obtained by estimating each population covariance matrix Σ_k by the corresponding sample covariance matrix or using the bootstrap. An advantage of the test statistics specific estimation approach is that it yields a continuous null distribution and hence does not suffer from the discreteness of the bootstrap null distribution mentioned above.

- *Data generating null distribution.* In certain testing problems, one may define a test statistics null distribution, $Q_n(P_0)$, in terms of a data generating distribution P_0 that satisfies the complete null hypothesis $H_0^C = \prod_{j=1}^{m} H_{0j}$. Such a null distribution may be estimated by $Q_{0n} = Q_n(P_{0n})$, where, for example, P_{0n} is a permutation- or bootstrap-based estimator of P_0. However, as discussed in Pollard and van der Laan (2003), this approach can fail in important testing problems, as the true distribution $Q_n(P)$ and the assumed null distribution $Q_n(P_{0n})$ may have different limits (or correlation matrices in the case of Gaussian asymptotic distributions) and, as a result, violate the required asymptotic null domination condition on the Type I error rates (equation (10), p. 17).

In this section, we consider analogues of Procedures 1 and 2, based on a consistent estimator Q_{0n} of a null distribution Q_0, such as the distribution $Q_0 = Q_0(P)$ defined in Theorem 2. In such multiple testing procedures, the estimator Q_{0n} is used in place of Q_0, to estimate the cut-offs for the test statistics and the resulting adjusted p-values.

4.1 Asymptotic control for consistent estimator of the null distribution

Theorem 3 [Consistency of single-step common-quantile cut-offs in Procedure 1] *Let Q_0 be a specified m-variate null distribution and let Q_{0n} converge weakly to Q_0. Assume that Q_0 is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^m$, with uniformly bounded density, and that each marginal distribution Q_{0j} has continuous Lebesgue density f_{0j} with interval support, that is, $\{z : f_{0j}(z) > 0\} = (a_j, b_j)$, where a_j and b_j are allowed to equal $-\infty$ and ∞, respectively. For an arbitrary m-variate distribution Q and constant $\delta \in [0,1]$, define δ–quantiles for the marginal distributions Q_j by*

$$d_j(Q,\delta) \equiv Q_j^{-1}(\delta) = \inf\{z : Q_j(z) \geq \delta\}, \qquad j = 1, \ldots, m,$$

and let $d(Q,\delta) = (d_j(Q,\delta) : j = 1, \ldots, m)$ denote the corresponding quantile m-vector. Define a non-increasing function

$$\delta \to G_Q(\delta) \equiv \theta\left(F_{R(d(Q,\delta)|Q)}\right),$$

where

$$R(d(Q,\delta) \mid Q) \equiv \sum_{j=1}^{m} \mathrm{I}(Z(j) > d_j(Q,\delta))$$

is the number of rejected hypotheses for $Z \sim Q$. For a fixed level $\alpha \in (0,1)$, define

$$\delta(Q) \equiv G_Q^{-1}(\alpha) = \inf\left\{\delta : \theta\left(F_{R(d(Q,\delta)|Q)}\right) \leq \alpha\right\}.$$

In particular, for the null distribution Q_0, assume that $\delta(Q_0) \in (0,1)$ and that the function $G_{Q_0}(\delta)$ is continuous and has a positive derivative at $\delta(Q_0) = G_{Q_0}^{-1}(\alpha)$. Then, one has the following consistency results for the common quantiles.

$$\begin{aligned} \text{As } n \to \infty, \qquad \delta(Q_{0n}) - \delta(Q_0) &\to 0 \qquad \text{and} \\ d_j(Q_{0n}, \delta(Q_{0n})) - d_j(Q_0, \delta(Q_0)) &\to 0, \qquad \forall\, j = 1, \ldots, m. \end{aligned}$$

Proof of Theorem 3. In what follows, we adopt the short-hand notation $G_0(\delta) \equiv G_{Q_0}(\delta)$, $G_{0n}(\delta) \equiv G_{Q_{0n}}(\delta)$, $\delta_0 \equiv \delta(Q_0)$, and $\delta_{0n} \equiv \delta(Q_{0n})$. We begin by establishing the following five facts.

Fact 1. *For each $j = 1, \ldots, m$, Q_{0j}^{-1} is uniformly continuous on any interval $[a, b] \subset (0, 1)$.*
That is, if $x_n - y_n \to 0$ for sequences $\{x_n\}$ and $\{y_n\} \in [a, b]$, then $Q_{0j}^{-1}(x_n) - Q_{0j}^{-1}(y_n) \to 0$. This fact follows from the assumption that each marginal distribution Q_{0j} has continuous Lebesgue density f_{0j} with interval support, that is, $\{z : f_{0j}(z) > 0\} = (a_j, b_j)$, where a_j and b_j are allowed to equal $-\infty$ and ∞, respectively.

Fact 2. *For each $j = 1, \ldots, m$, as $n \to \infty$, $Q_{0n,j} - Q_{0j}$ converges uniformly to zero over the support (a_j, b_j) of Q_{0j}.*
By the weak convergence of Q_{0n} to Q_0, we have that $Q_{0n,j}$ converges pointwise to Q_{0j} at each continuity point of Q_{0j}. Since pointwise convergence of monotone functions to a continuous monotone function implies uniform convergence, it follows that $Q_{0n,j} - Q_{0j}$ converges uniformly to zero.

Fact 3. *For each $j = 1, \ldots, m$, $Q_{0n,j}^{-1} - Q_{0j}^{-1}$ converges uniformly to zero over any interval contained in $(0, 1)$, that is, $\forall\, \epsilon > 0$, $\sup_{\delta \in [\epsilon, 1-\epsilon]} \mid Q_{0n,j}^{-1}(\delta) - Q_{0j}^{-1}(\delta) \mid \to 0$, as $n \to \infty$.*
This statement follows from the facts that: (i) for an m-variate distribution Q, the quantile mapping $Q_j \to Q_j^{-1}(\delta)$, for each margin Q_j, is continuous w.r.t. to the supremum norm convergence at a Q_{0j} at which $f_{0j}(Q_{0j}^{-1}(\delta)) > 0$; (ii) pointwise convergence of monotone functions $Q_{0n,j}^{-1}$ to a continuous monotone function Q_{0j}^{-1} at each point δ implies uniform convergence; and (iii) Fact 2.

Fact 4. *Consider c.d.f.'s F_n and F such that $F_n - F$ converges uniformly to zero and F^{-1} is uniformly continuous on an interval $[a, b] \subset (0, 1)$. For a sequence $\{x_n\} \in \mathbb{R}$, suppose there is an integer $N > 0$, such that $x_n \in [a, b]$ and $FF_n^{-1}(x_n) \in [a, b]$, $\forall\, n > N$. Then $F_n^{-1}(x_n) - F^{-1}(x) = F^{-1}(x_n) - F^{-1}(x) + o(1)$.*

This fourth fact follows from the expansion

$$\begin{aligned} F_n^{-1}(x_n) - F^{-1}(x) &= \{F_n^{-1}(x_n) - F^{-1}(x_n)\} + \{F^{-1}(x_n) - F^{-1}(x)\} \\ &= \{F^{-1}FF_n^{-1}(x_n) - F^{-1}F_nF_n^{-1}(x_n)\} + \{F^{-1}(x_n) - F^{-1}(x)\}. \end{aligned}$$

The first term converges to zero, by uniform convergence of $F_n - F$ to zero, uniform continuity of F^{-1} on $[a, b]$, and the fact that x_n and $FF_n^{-1}(x_n) \in [a, b]$, $\forall\, n > N$.

In order to apply Fact 4 with $F = Q_{0j}$, $F_n = Q_{0n,j}$, $x = \delta_0$, and $x_n = \delta_{0n}$, we need to prove the following Fact 5.

Fact 5. *Suppose that $\delta_{0n} \to \delta_0$. Then, there is an interval $[a, b] \subset (0, 1)$ and integer $N_0 > 0$, such that, for all $n > N_0$, $\delta_{0n} \in [a, b]$ and $Q_{0j}(Q_{0n,j}^{-1}(\delta_{0n})) \in [a, b]$.*
The first statement follows from convergence of δ_{0n} to δ_0 and by the assumption that $\delta_0 \in (0, 1)$. Thus, there exist $\epsilon > 0$ and an integer $N(\epsilon) > 0$, such that $\epsilon < \delta_{0n} < 1 - \epsilon$, $\forall\, n > N(\epsilon)$. By monotonicity of $Q_{0n,j}^{-1}$, we have $Q_{0n,j}^{-1}(\epsilon) \le Q_{0n,j}^{-1}(\delta_{0n}) \le Q_{0n,j}^{-1}(1 - \epsilon)$, $\forall\, n > N(\epsilon)$. Next, by weak convergence of $Q_{0n,j}$ to Q_{0j}, $\forall\, \epsilon' > 0$, $\exists\, N(\epsilon')$, such that

$$Q_{0j}^{-1}(\epsilon) - \epsilon' \le Q_{0n,j}^{-1}(\delta_{0n}) \le Q_{0j}^{-1}(1 - \epsilon) + \epsilon', \qquad \forall\, n > \max(N(\epsilon), N(\epsilon')).$$

Hence, by monotonicity of Q_{0j}

$$Q_{0j}(Q_{0j}^{-1}(\epsilon) - \epsilon') \le Q_{0j}(Q_{0n,j}^{-1}(\delta_{0n})) \le Q_{0j}(Q_{0j}^{-1}(1-\epsilon) + \epsilon'),\ \forall\, n > \max(N(\epsilon), N(\epsilon')).$$

Finally, by continuity of Q_{0j} and letting $\epsilon' \downarrow 0$, there is an interval $[a, b] \subset (0, 1)$ and an integer $N_0 > 0$, such that, as required, $Q_{0j}(Q_{0n,j}^{-1}(\delta_{0n})) \in [a, b]$, for $n > N_0$.

Convergence of $(\delta_{0n} - \delta_0)$ to zero. In order to apply Facts 1 – 5 to establish consistency of the single-step common-quantile cut-offs, the main task is to prove that $(\delta_{0n} - \delta_0)$ converges to zero. Consider the functions $G_0(\delta) = G_{Q_0}(\delta)$ and $G_{0n}(\delta) = G_{Q_{0n}}(\delta)$, and note that $\delta_{0n} - \delta_0 = G_{0n}^{-1}(\alpha) - G_0^{-1}(\alpha)$. By monotonicity Assumption AMI on the mapping $\theta(\cdot)$, then G_{0n} and G_0 are monotone decreasing functions in δ. Also, for any m-variate distribution Q, $R(d(Q, \delta) \mid Q) \le m$, thus G_{0n} and G_0 are uniformly bounded above by $\theta(\Delta_{\{m\}})$, where $\Delta_{\{m\}}$ denotes the c.d.f. with unit mass at the singleton $\{m\}$.

Under the assumption that G_0 has a positive derivative at $\delta_0 = G_0^{-1}(\alpha)$, it follows that $G_{0n}^{-1}(\alpha) - G_0^{-1}(\alpha)$ converges to zero provided $G_{0n} - G_0$ converges uniformly to zero. Since pointwise convergence of monotone functions to a continuous monotone function implies uniform convergence, it suffices to show that $(G_{0n} - G_0)(\delta)$ converges to zero at any given $\delta \in (0,1)$. By uniform continuity Assumption ACI on the mapping $\theta(\cdot)$, the latter holds if the number of rejections, $R(d(Q_{0n}, \delta) \mid Q_{0n})$, under Q_{0n}, converges in distribution to the corresponding quantity, $R(d(Q_0, \delta) \mid Q_0)$, under Q_0. We have

$$\begin{aligned} R(d(Q_{0n}, \delta) \mid Q_{0n}) &= R(d(Q_{0n}, \delta) \mid Q_{0n}) - R(d(Q_0, \delta) \mid Q_{0n}) \\ &\quad + R(d(Q_0, \delta) \mid Q_{0n}). \end{aligned}$$

Let $Z_n \sim Q_{0n}$ and $Z \sim Q_0$, and denote the third term with $f(Z_n) \equiv \sum_{j=1}^m \mathrm{I}(Z_n(j) > d_j(Q_0, \delta))$. Since f is continuous Q_0-a.s., by the Continuous Mapping Theorem, we have that $f(Z_n)$ converges in distribution to $f(Z) = \sum_{j=1}^m \mathrm{I}(Z(j) > d_j(Q_0, \delta)) = R(d(Q_0, \delta) \mid Q_0)$. Thus, it remains to prove that the difference $R(d(Q_{0n}, \delta) \mid Q_{0n}) - R(d(Q_0, \delta) \mid Q_{0n})$ converges to zero a.s., that is, $\lim_n Pr_{Q_{0n}}(R(d(Q_{0n}, \delta) \mid Q_{0n}) \neq R(d(Q_0, \delta) \mid Q_{0n})) = 0$. For a fixed $\delta \in (0,1)$, use the following short-hand notation for the common quantiles, $c_0(j) = d_j(Q_0, \delta) = Q_{0j}^{-1}(\delta)$ and $c_{0n}(j) = d_j(Q_{0n}, \delta) = Q_{0n,j}^{-1}(\delta)$, and note that, from Fact 3, $c_{0n}(j) - c_0(j) \to 0$, as $n \to \infty$, $\forall\, j = 1, \ldots, m$. Define subsets $A_n \subseteq I\!R^m$ by

$$A_n \equiv \left\{ z \in I\!R^m : \left| \sum_{j=1}^m \mathrm{I}(z(j) > c_{0n}(j)) - \mathrm{I}(z(j) > c_0(j)) \right| > 0 \right\}.$$

Then, $R(d(Q_{0n}, \delta) \mid Q_{0n}) \neq R(d(Q_0, \delta) \mid Q_{0n})$ if and only if $Z_n \in A_n$. By absolute continuity of Q_0 w.r.t. the Lebesgue measure on $I\!R^m$, with uniformly bounded density, we have that $Q_0(A(\epsilon)) \to 0$ as $\epsilon \to 0$, for subsets $A(\epsilon) \subseteq I\!R^m$ defined as in Lemma 1, below. Thus, it follows from this Lemma that $\lim_n Q_{0n}(A_n) = 0$, that is, as required,

$$\begin{aligned} \lim_{n\to\infty} Q_{0n}(A_n) &= \lim_{n\to\infty} Pr_{Q_{0n}}(Z_n \in A_n) \\ &= \lim_{n\to\infty} Pr_{Q_{0n}}(R(d(Q_{0n}, \delta) \mid Q_{0n}) \neq R(d(Q_0, \delta) \mid Q_{0n})) = 0. \end{aligned}$$

Convergence of $(d_j(Q_{0n}, \delta_{0n}) - d_j(Q_0, \delta_0))$ to zero. We are now in a position to apply Facts 1 – 5 to establish consistency of the single-step common-quantile cut-offs, i.e., convergence of $d_j(Q_{0n}, \delta_{0n}) - d_j(Q_0, \delta_0)$ to zero. The

assumptions of Fact 4, with $F = Q_{0j}$, $F_n = Q_{0n,j}$, $x = \delta_0$, and $x_n = \delta_{0n}$, are satisfied from Facts 2, 1, and 5, respectively. Thus,

$$\begin{aligned} d_j(Q_{0n}, \delta_{0n}) - d_j(Q_0, \delta_0) &= Q_{0n,j}^{-1}(\delta_{0n}) - Q_{0j}^{-1}(\delta_0) \\ &= Q_{0j}^{-1}(\delta_{0n}) - Q_{0j}^{-1}(\delta_0) + o(1). \end{aligned}$$

By continuity of Q_{0j}^{-1} (Fact 1) and convergence of $\delta_{0n} - \delta_0$ to zero, it follows that $Q_{0j}^{-1}(\delta_{0n}) - Q_{0j}^{-1}(\delta_0)$ converges to zero.

□

Lemma 1 *Consider a sequence of m-vectors* $\{c_{0n}\} \in I\!R^m$*, with limit* $c_0 \in I\!R^m$*, i.e.,* $c_{0n}(j) \to c_0(j)$*, as* $n \to \infty$*,* $\forall\, j = 1, \ldots, m$*. Define subsets* $A_n \subseteq I\!R^m$ *by*

$$A_n \equiv \left\{ z \in I\!R^m : \left| \sum_{j=1}^{m} \mathrm{I}(z(j) > c_{0n}(j)) - \mathrm{I}(z(j) > c_0(j)) \right| > 0 \right\}$$

and for $\epsilon > 0$ *define*

$$A(\epsilon) \equiv \left\{ z \in I\!R^m : \sup_{\{c : ||c - c_0|| < \epsilon\}} \left| \sum_{j=1}^{m} \mathrm{I}(z(j) > c(j)) - \mathrm{I}(z(j) > c_0(j)) \right| > 0 \right\}.$$

Let Q_0 *be a specified m-variate distribution and let* Q_{0n} *converge weakly to* Q_0*. Further assume that there exists a sequence* $\{\epsilon_k\} \downarrow 0$*, such that* $A(\epsilon_k)$ *are continuity sets of* Q_0 *(i.e., the boundary sets* $\partial A(\epsilon_k)$ *have mass zero under* Q_0*,* $Q_0(\partial A(\epsilon_k)) = 0$*) and such that* $\lim_k Q_0(A(\epsilon_k)) = 0$*. Then,* $\lim_n Q_{0n}(A_n) = 0$*.*

Proof of Lemma 1. By the definition of weak convergence (van der Vaart and Wellner, 1996), for each continuity set $A(\epsilon_k) \subseteq I\!R^m$, $Q_{0n}(A(\epsilon_k)) - Q_0(A(\epsilon_k)) \to 0$, as $n \to \infty$. Thus, for all k and all $\epsilon > 0$, there exists an integer $N(k, \epsilon) > 0$, such that $Q_{0n}(A(\epsilon_k)) \leq Q_0(A(\epsilon_k)) + \epsilon$, for all $n > N(k, \epsilon)$. Next, by convergence of c_{0n} to c_0, $\forall\, k$, there exists an integer $N(k) > 0$, such that, for all $n > N(k)$, $A_n \subseteq A(\epsilon_k)$ and hence $Q_{0n}(A_n) \leq Q_{0n}(A(\epsilon_k))$. Thus, for all k and all $\epsilon > 0$,

$$Q_{0n}(A_n) \leq Q_{0n}(A(\epsilon_k)) \leq Q_0(A(\epsilon_k)) + \epsilon, \qquad \forall\, n > \max(N(k), N(k, \epsilon)).$$

But $\lim_k Q_0(A(\epsilon_k)) = 0$, hence, as required, $\lim_n Q_{0n}(A_n) = 0$.

□

An (simpler) analogue of Theorem 3 can be obtained for consistency of the single-step common cut-offs in Procedure 2.

Theorem 4 [Consistency of single-step common cut-offs in Procedure 2] *Let Q_0 be a specified m-variate null distribution and let Q_{0n} converge weakly to Q_0. For an arbitrary m-variate distribution Q, define a non-increasing function*

$$c \to G_Q(c) \equiv \theta\left(F_{R((c,\ldots,c)|Q)}\right),$$

where $R((c,\ldots,c) \mid Q) \equiv \sum_{j=1}^{m} \mathrm{I}(Z(j) > c)$ is the number of rejected hypotheses for $Z \sim Q$. For a fixed level $\alpha \in (0,1)$, define common cut-offs

$$e(Q) \equiv G_Q^{-1}(\alpha) = \inf\left\{c : \theta\left(F_{R((c,\ldots,c)|Q)}\right) \leq \alpha\right\}.$$

In particular, for the null distribution Q_0, assume that the function $G_{Q_0}(c)$ is continuous and has a positive derivative at $e(Q_0) = G_{Q_0}^{-1}(\alpha)$. Then, one has the following consistency results for the common cut-offs

$$e(Q_{0n}) - e(Q_0) \to 0, \qquad \text{as } n \to \infty.$$

Proof of Theorem 4. First note that $e_{0n} - e_0 = G_{0n}^{-1}(\alpha) - G_0^{-1}(\alpha)$, where we use the short-hand notation $G_0(\delta) = G_{Q_0}(\delta)$ and $G_{0n}(\delta) = G_{Q_{0n}}(\delta)$, and $e_0 = e(Q_0)$ and $e_{0n} = e(Q_{0n})$. Convergence of $e_{0n} - e_0$ to zero, follows from the first part of the proof that $\delta_{0n} - \delta_0 \to 0$ in Theorem 3. One simply needs to show that $(G_{0n} - G_0)(c)$ converges to zero at any given c. This follows by noting that: (i) from the Continuous Mapping Theorem, the number of rejections, $R((c,\ldots,c) \mid Q_{0n})$, under Q_{0n}, converges in distribution to the corresponding quantity, $R((c,\ldots,c) \mid Q_0)$, under Q_0, and (ii) by uniform continuity Assumption ACI on the mapping $\theta(\cdot)$ and fact (i), $(G_{0n} - G_0)(c) = \theta(F_{R((c,\ldots,c)|Q_{0n})}) - \theta(F_{R((c,\ldots,c)|Q_0)})$ converges to zero.

□

Having established consistency of the cut-offs for common-quantile and common-cut-off Procedures 1 and 2, based on a consistent estimator Q_{0n} of the null distribution Q_0, the following corollary proves consistency of the resulting Type I error rates.

Corollary 1 [Consistency of Type I error rate for single-step common-quantile and common-cut-off Procedures 1 and 2] *Let Q_0 be a specified m-variate null distribution and let Q_{0n} converge weakly to Q_0. Denote the number of Type I errors for Procedure 1 (or Procedure 2) based on the null distribution Q_0 and its approximation Q_{0n} by*

$$V(c_{0n} \mid Q_n) = \sum_{j \in S_0} \mathrm{I}\big(T_n(j) > c_{0n}(j)\big)$$

$$\text{and} \qquad V(c_0 \mid Q_n) = \sum_{j \in S_0} \mathrm{I}\big(T_n(j) > c_0(j)\big),$$

respectively, where $T_n \sim Q_n = Q_n(P)$. For Procedure 1, the common-quantile cut-offs $c_0(j) = c_j(Q_0, \alpha) = d_j(Q_0, \delta_0)$ and $c_{0n}(j) = c_j(Q_{0n}, \alpha) = d_j(Q_{0n}, \delta_{0n})$ are defined as in Theorem 3. For Procedure 2, the common cut-offs $c_0(j) = c_j(Q_0, \alpha) = e(Q_0)$ and $c_{0n}(j) = c_j(Q_{0n}, \alpha) = e(Q_{0n})$ are defined as in Theorem 4. Assume that the conditions of Theorem 3 (or Theorem 4) hold, so that $c_{0n}(j) - c_0(j) \to 0$, as $n \to \infty$, for each $j = 1, \ldots, m$. Further assume that the joint distribution $Q_n(P)$ of the test statistics T_n is such that

$$Q_n(A_{S_0}(\epsilon)) = Pr_{Q_n}(T_n \in A_{S_0}(\epsilon)) \to 0, \qquad \text{as } \epsilon \downarrow 0, \tag{33}$$

where subsets $A_{S_0}(\epsilon) \subseteq I\!R^m$ are defined as

$$A_{S_0}(\epsilon) \equiv \left\{ z \in I\!R^m : \sup_{\{c : ||c - c_0|| < \epsilon\}} \left| \sum_{j \in S_0} \mathrm{I}(z(j) > c(j)) - \mathrm{I}(z(j) > c_0(j)) \right| > 0 \right\}.$$

Then,

$$\lim_{n \to \infty} Pr_{Q_n}(V(c_{0n} \mid Q_n) \neq V(c_0 \mid Q_n)) = 0,$$

so that asymptotic control of the Type I error rate $\theta(F_{V(c_0|Q_n)})$, as in Theorem 1, for Procedure 1 (or Procedure 2) based on cut-offs $c(Q_0, \alpha)$, implies asymptotic control of the corresponding Type I error rate $\theta(F_{V(c_{0n}|Q_n)})$, for Procedure 1 (or Procedure 2) based on estimated cut-offs $c(Q_{0n}, \alpha)$. That is, $\limsup_n \theta(F_{V(c_{0n}|Q_n)}) \leq \alpha$ follows from $\limsup_n \theta(F_{V(c_0|Q_n)}) \leq \alpha$.

Proof of Corollary 1. Define subsets $A_{S_0,n} \subseteq I\!R^m$ by

$$A_{S_0,n} \equiv \left\{ z \in I\!R^m : \left| \sum_{j \in S_0} \mathrm{I}(z(j) > c_{0n}(j)) - \mathrm{I}(z(j) > c_0(j)) \right| > 0 \right\}$$

and note that

$$Pr_{Q_n}(V(c_{0n} \mid Q_n) \neq V(c_0 \mid Q_n)) = Pr_{Q_n}(T_n \in A_{S_0,n}) = Q_n(A_{S_0,n}).$$

From Theorem 3 (Theorem 4), the cut-offs for Procedure 1 (Procedure 2) are such that $c_{0n}(j) - c_0(j) \to 0$, as $n \to \infty$, $\forall\, j = 1, \ldots, m$. Thus, for each $\epsilon > 0$, there is an integer $N(\epsilon) > 0$, such that $A_{S_0,n} \subseteq A_{S_0}(\epsilon)$ and hence $Q_n(A_{S_0,n}) \leq Q_n(A_{S_0}(\epsilon))$, for each $n > N(\epsilon)$. By assumption, $\lim_{\epsilon \downarrow 0} Q_n(A_{S_0}(\epsilon)) = 0$, thus $Pr_{Q_n}(V(c_{0n} \mid Q_n) \neq V(c_0 \mid Q_n)) = Q_n(A_{S_0,n}) \to 0$, as $n \to \infty$, i.e., $V(c_{0n} \mid Q_n) - V(c_0 \mid Q_n)$ converges to zero. Hence, by uniform continuity Assumption ACI for the mapping $\theta(\cdot)$, $\theta(F_{V(c_{0n}|Q_n)}) - \theta(F_{V(c_0|Q_n)}) \to 0$. Thus, asymptotic control of the Type I error rate $\theta(F_{V(c_{0n}|Q_n)})$ follows from asymptotic control of the Type I error rate $\theta(F_{V(c_0|Q_n)})$, as established in Theorem 1.

□

The next lemma can be applied to prove asymptotic Type I error control for Procedures 1 and 2, based on a consistent estimator Q_{0n} of the null distribution $Q_0 = Q_0(P)$ defined in Theorem 2, without requiring assumption (33) of Corollary 1. In this special case, one can define $\tilde{Q}_n$ as the distribution of the intermediate random vectors $\tilde{Z}_n$ used in the proof of Theorem 2. Under the assumptions of Theorems 3 and 4, the estimated single-step cut-offs $c_{0n} = c(Q_{0n}, \alpha)$ converge to $c_0 = c(Q_0, \alpha)$. The lemma then shows that $\limsup_n \theta(F_{V(c_{0n}|Q_n)}) \leq \theta(F_{V(c_0|Q_0)})$.

Lemma 2 *Consider a sequence of m-vectors $\{c_{0n}\} \in I\!R^m$, with limit $c_0 \in I\!R^m$, i.e., $c_{0n}(j) \to c_0(j)$, as $n \to \infty$, $\forall\, j = 1, \ldots, m$. Let Q_0 be an m-variate distribution that satisfies the continuity set assumption of Lemma 1, and suppose that $\tilde{Q}_n$ is an m-variate distribution that converges weakly to Q_0 and that dominates a third m-variate distribution Q_n on the set S_0 of true null hypotheses, i.e., such that $Q_{n,S_0} \geq \tilde{Q}_{n,S_0}$. Then,*

$$\limsup_{n \to \infty} \theta(F_{V(c_{0n}|Q_n)}) \leq \theta(F_{V(c_0|Q_0)}).$$

Proof of Lemma 2. Define subsets $A_{S_0,n} \subseteq I\!R^m$ by

$$A_{S_0,n} \equiv \left\{ z \in I\!R^m : \left| \sum_{j \in S_0} \mathrm{I}(z(j) > c_{0n}(j)) - \mathrm{I}(z(j) > c_0(j)) \right| > 0 \right\}.$$

By the domination of Q_n by $\tilde{Q}_n$ and monotonicity Assumption AMI for the Type I error rate mapping $\theta(\cdot)$, it follows that, for all n, $\theta(F_{V(c_{0n}|Q_n)}) \leq \theta(F_{V(c_{0n}|\tilde{Q}_n)})$. From Lemma 1, $Pr_{\tilde{Q}_n}(V(c_{0n} \mid \tilde{Q}_n) \neq V(c_0 \mid \tilde{Q}_n)) = \tilde{Q}_n(A_{S_0,n}) \to 0$. Thus, by uniform continuity Assumption ACI for the mapping $\theta(\cdot)$, we have $\theta(F_{V(c_{0n}|\tilde{Q}_n)}) - \theta(F_{V(c_0|\tilde{Q}_n)}) \to 0$. Also, by weak convergence of $\tilde{Q}_n$ to Q_0, $\theta(F_{V(c_0|\tilde{Q}_n)}) - \theta(F_{V(c_0|Q_0)}) \to 0$. Hence,

$$\limsup_{n\to\infty} \theta(F_{V(c_{0n}|Q_n)}) \leq \limsup_{n\to\infty} \theta(F_{V(c_{0n}|\tilde{Q}_n)}) = \theta(F_{V(c_0|Q_0)}).$$

□

Note that, for an estimator Q_{0n} of the null distribution Q_0, the consistency results in Theorems 3 and 4 and Corollary 1 are conditional on the empirical distribution for an infinite sequence $X^\infty = (X_1, X_2, \ldots) \sim P^\infty$. That is, the results apply for every $X^\infty \sim P^\infty$ for which $Q_{0n} \overset{\mathcal{L}}{\Rightarrow} Q_0$. Consequently, if $Q_{0n} \overset{\mathcal{L}}{\Rightarrow} Q_0$, P^∞-a.s., then the above consistency results hold P^∞-a.s. Under regularity conditions, bootstrap estimators Q_{0n} of the null distribution Q_0 are consistent, in the sense that Q_{0n} converges weakly to Q_0, conditional on the empirical distribution P_n (van der Vaart and Wellner, 1996). Thus, under such regularity conditions, the consistency results in Theorems 3 and 4 and Corollary 1 hold P^∞-a.s. for bootstrap-based analogues of Procedures 1 and 2.

4.2 Bootstrap estimation of the null distribution

The null distribution $Q_0 = Q_0(P)$ of Theorem 2 can be estimated with the non-parametric or model-based bootstrap. Let $P_n^\star$ denote an estimator of the true data generating distribution P. For the *non-parametric bootstrap*, $P_n^\star$ is simply the empirical distribution P_n, that is, samples of size n are drawn at random with replacement from the observed $X_1, \ldots, X_n$. For the *model-based bootstrap*, $P_n^\star$ is based on a model $\mathcal{M}$ for the data generating distribution P, such as the family of m-variate Gaussian distributions.

A bootstrap sample consists of n i.i.d. realizations, $X_1^\#, \ldots, X_n^\#$, of a random variable $X^\# \sim P_n^\star$. Denote the m-vector of test statistics computed from such a bootstrap sample by $T_n^\# = (T_n^\#(j) : j = 1, \ldots, m)$. The null distribution Q_0 proposed in Theorem 2 can be estimated by the distribution

of the null-value shifted and scaled bootstrap statistics

$$Z_n^{\#}(j) \equiv \sqrt{\min\left(1, \frac{\tau_0(j)}{Var_{P_n^\star}[{T_n}^{\#}(j)]}\right)}\Big(T_n^{\#}(j) + \lambda_0(j) - E_{P_n^\star}[{T_n}^{\#}(j)]\Big). \quad (34)$$

In practice, one can only approximate the distribution of $Z_n^{\#} = (Z_n^{\#}(j) : j = 1, \ldots, m)$ by an empirical distribution over B bootstrap samples drawn from $P_n^\star$, as described next in Procedure 3.

Procedure 3. Bootstrap estimation of null distribution Q_0.

1. Generate B bootstrap samples, $(X_1^b, \ldots, X_n^b)$, $b = 1, \ldots, B$. For the bth sample, the X_i^b, $i = 1, \ldots, n$, are n i.i.d. realizations of a random variable $X^{\#} \sim P_n^\star$.

2. For each bootstrap sample, compute an m-vector of test statistics, $T_n^b = (T_n^b(j) : j = 1, \ldots, m)$, which can be arranged in an $m \times B$ matrix, $\mathbf{T} = \big(T_n^b(j)\big)$, with rows corresponding to the m hypotheses and columns to the B bootstrap samples.

3. Compute row means and variances of the matrix $\mathbf{T}$, to yield estimates of $E[T_n(j)]$ and $Var[T_n(j)]$, $j = 1, \ldots, m$.

4. Obtain an $m \times B$ matrix, $\mathbf{Z} = \big(Z_n^b(j)\big)$, of null-value shifted and scaled bootstrap statistics $Z_n^b(j)$, as in Theorem 2, by row-shifting and scaling the matrix $\mathbf{T}$ using the bootstrap estimates of $E[T_n(j)]$ and $Var[T_n(j)]$ and the user-supplied null-values $\lambda_0(j)$ and $\tau_0(j)$.

5. The bootstrap estimate Q_{0n} of the null distribution Q_0 from Theorem 2 is the empirical distribution of the columns Z_n^b of matrix $\mathbf{Z}$.

Procedure 4. Bootstrap estimation of common quantiles for Procedure 1 for gFWER control.

0. Apply Procedure 3 to generate an $m \times B$ matrix, $\mathbf{Z} = \left(Z_n^b(j)\right)$, of null-value shifted and scaled bootstrap statistics $Z_n^b(j)$. The bootstrap estimate Q_{0n} of the null distribution Q_0 from Theorem 2 is the empirical distribution of the columns Z_n^b of matrix $\mathbf{Z}$.

1. For Procedure 1, the bootstrap common-quantile cut-offs are simply the row quantiles of the matrix $\mathbf{Z}$. That is, $d_j(Q_{0n}, \delta)$ is the δ–quantile of the B-vector $(Z_n^b(j) : b = 1, \ldots, B)$ of bootstrap statistics for H_{0j}

$$d_j(Q_{0n}, \delta) \equiv \inf \left\{ z : \frac{1}{B} \sum_{b=1}^{B} \mathrm{I}(Z_n^b(j) \leq z) \geq \delta \right\}, \quad j = 1, \ldots, m.$$

2. For a test with nominal level $\alpha \in (0, 1)$, δ is chosen as

$$\delta_{0n}(\alpha) \equiv \inf\{\delta : \theta(F_{R(d(Q_{0n},\delta)|Q_{0n})}) \leq \alpha\}.$$

   That is, $\delta_{0n}(\alpha)$ corresponds to the smallest cut-offs $d(Q_{0n}, \delta)$ such that the value of the mapping $\theta(\cdot)$, applied to the distribution of the number of rejections $R(d(Q_{0n}, \delta) \mid Q_{0n})$, under the bootstrap distribution Q_{0n}, is at most α.

   In the case of gFWER control, and for a (limit) null distribution Q_0 with continuous and strictly monotone marginal distributions, $(1 - \delta_{0n}(\alpha))$ is the α–quantile of the bootstrap estimate of the distribution of the $(k+1)$st ordered unadjusted p-value (Section 3.3.3). Specifically, $\delta_{0n}(\alpha)$ is obtained as follows.

   (a) Compute an $m \times B$ matrix, $\mathbf{P} = \left(P_n^b(j)\right)$, of bootstrap unadjusted p-values, by row-ranking the matrix $\mathbf{Z}$, i.e., by replacing each $Z_n^b(j)$ by its rank over the B bootstrap samples, where 1 corresponds to the largest value of $Z_n^b(j)$ and B the smallest.

   (b) For each column of the matrix $\mathbf{P}$, compute the $(k+1)$st smallest p-value, $P_n^{\circ\, b}(k+1)$. For FWER control ($k = 0$), simply compute column minima.

   (c) The estimate $(1 - \delta_{0n}(\alpha))$ is the α–quantile of the B-vector $(P_n^{\circ\, b}(k+1) : b = 1, \ldots, B)$.

Procedure 5. Bootstrap estimation of common cut-offs for Procedure 2 for gFWER control.

0. Apply Procedure 3 to generate an $m \times B$ matrix, $\mathbf{Z} = \left(Z_n^b(j)\right)$, of null-value shifted and scaled bootstrap statistics $Z_n^b(j)$. The bootstrap estimate Q_{0n} of the null distribution Q_0 from Theorem 2 is the empirical distribution of the columns Z_n^b of matrix $\mathbf{Z}$.

1. In the case of gFWER control, and for a (limit) null distribution Q_0 with continuous and strictly monotone marginal distributions, the bootstrap common cut-off $c_j(Q_{0n}, \alpha) = e(Q_{0n}, \alpha)$ is equal to the $(1-\alpha)$–quantile of the bootstrap estimate of the distribution of the $(k+1)$st ordered component of $Z \sim Q_0$ (Section 3.3.3). Specifically, $e(Q_{0n}, \alpha)$ is obtained as follows.

   (a) For each column of the matrix $\mathbf{Z}$, compute the $(k+1)$st largest statistic $Z_n^{\circ\, b}(k+1)$. For FWER control ($k = 0$), simply compute column maxima.

   (b) The estimated common cut-off $e(Q_{0n}, \alpha)$ is the $(1-\alpha)$–quantile of the B-vector $(Z_n^{\circ\, b}(k+1) : b = 1, \ldots, B)$.

5 Examples

5.1 t-statistics for single-parameter hypotheses

In this section, we consider testing m one-sided single-parameter null hypotheses of the form $H_{0j} = \mathrm{I}(\mu(j) \leq \mu_0(j))$, against alternative hypotheses $H_{1j} = \mathrm{I}(\mu(j) > \mu_0(j))$, where $\mu(j) = \mu_j(P)$ is a real-valued parameter, $j = 1, \ldots, m$. As in Section 2.1.4, consider *t-statistics*

$$T_n(j) \equiv \sqrt{n} \frac{\mu_n(j) - \mu_0(j)}{\sigma_n(j)}, \tag{35}$$

where $\mu_n = (\mu_n(j) : j = 1, \ldots, m)$ is an *asymptotically linear estimator* of the parameter m-vector $\mu = (\mu(j) : j = 1, \ldots, m)$, with m-dimensional vector

influence curve, $IC(X \mid P) = (IC_j(X \mid P) : j = 1, \ldots, m)$, such that

$$\mu_n(j) - \mu(j) = \frac{1}{n} \sum_{i=1}^{n} IC_j(X_i \mid P) + o_P(1/\sqrt{n}), \tag{36}$$

and $\sigma_n^2(j)$ is a consistent estimator of $\sigma^2(j) \equiv E[IC_j^2(X \mid P)]$, $j = 1, \ldots, m$. Large values of $T_n(j)$ provide evidence against $H_{0j} = \mathrm{I}(\mu(j) \leq \mu_0(j))$. Two-sided tests of $H_{0j} = \mathrm{I}(\mu(j) = \mu_0(j))$ against alternatives $H_{1j} = \mathrm{I}(\mu(j) \neq \mu_0(j))$ can be handled similarly, by taking absolute values of the test statistics $T_n(j)$. Next, we propose a null distribution Q_0^*, that provides asymptotic control of Type I error rates of the form $\theta(F_{V_n})$, when used in Procedures 1 and 2.

5.1.1 Null distribution

Theorem 5 *The test statistics $T_n = (T_n(j) : j = 1, \ldots, m)$ satisfy asymptotic null domination Assumption AQ0 of Theorem 1, where the null distribution $Q_0^* = Q_0^*(P)$ is the m-variate Gaussian distribution $N(0, \rho(P))$ and $\rho(P)$ is the correlation matrix of the vector influence curve $IC(X \mid P)$. Thus, by Theorem 1, single-step Procedures 1 and 2, based on T_n and the null distribution Q_0^*, provide asymptotic control of general Type I error rates $\theta(F_{V_n})$, for the test of single-parameter null hypotheses of the form $H_{0j} = \mathrm{I}(\mu(j) \leq \mu_0(j))$, against alternative hypotheses $H_{1j} = \mathrm{I}(\mu(j) > \mu_0(j))$, $j = 1, \ldots, m$.*

Proof of Theorem 5. We verify Assumption AQ0 of Theorem 1 for the test statistics T_n of equation (35) and the null distribution $N(0, \rho(P))$. Firstly, note that the $T_n(j)$ can be rewritten as

$$\begin{aligned} T_n(j) &= \sqrt{n}\frac{\mu_n(j) - \mu(j)}{\sigma_n(j)} + \sqrt{n}\frac{\sigma(j)}{\sigma_n(j)}\frac{\mu(j) - \mu_0(j)}{\sigma(j)} \\ &= Z_n^*(j) + \frac{\sigma(j)}{\sigma_n(j)} d_n(j), \qquad j = 1, \ldots, m, \end{aligned} \tag{37}$$

in terms of deterministic shifts, $d_n(j) \equiv \sqrt{n}\frac{\mu(j) - \mu_0(j)}{\sigma(j)}$, and standardized statistics, $Z_n^*(j) \equiv \sqrt{n}\frac{\mu_n(j) - \mu(j)}{\sigma_n(j)}$. By condition (36), the Central Limit Theorem, and Slutsky's Theorem, we have that

$$Z_n^* \stackrel{\mathcal{L}}{\Rightarrow} Z^* \sim Q_0^*(P) \equiv N(0, \rho(P)),$$

where $\rho(P)$ is the correlation matrix of the vector influence curve $IC(X \mid P)$. For $j \in S_0$, $d_n(j) \leq 0$, so that $T_n(j) \leq Z_n^*(j)$. Thus, for all $c = (c_j : j = 1, \ldots, m) \in I\!R^m$ and $x \in \{0, \ldots, m\}$,

$$\begin{aligned} \liminf_{n\to\infty} Pr\left(\sum_{j\in S_0} \mathrm{I}(T_n(j) > c_j) \leq x\right) \\ \geq \liminf_{n\to\infty} Pr\left(\sum_{j\in S_0} \mathrm{I}(Z_n^*(j) > c_j) \leq x\right) \\ = Pr\left(\sum_{j\in S_0} \mathrm{I}(Z^*(j) > c_j) \leq x\right), \end{aligned}$$

and the main Assumption AQ0 is satisfied by using Q_0^*. □

Comparison to general proposal from Theorem 2. The above theorem involves a null distribution Q_0^* that was derived specifically in terms of the t-statistics in equations (35) and (36). It turns out that, under mild regularity conditions, this null distribution Q_0^* corresponds to the general proposal Q_0 in Theorem 2, with null-values $\lambda_0(j) \equiv 0$ and $\tau_0(j) \equiv 1$. To see this, consider the simple known variance case, $\sigma_n(j) \equiv \sigma(j)$. Then, $E[T_n(j)] = d_n(j)$, $Var[T_n(j)] = 1$, and $Cov[T_n] = Cor[T_n] = \rho(P)$. Hence, $T_n(j) = Z_n^*(j) + E[T_n(j)]$. The null distribution Q_0^* of Theorem 5 is the asymptotic distribution of the m-vector Z_n^*, that is, the $N(0, \rho(P))$ distribution. For null-values $\lambda_0(j) \equiv 0$ and $\tau_0(j) \equiv 1$, and in the known variance case, the m-vector Z_n defining the general null distribution Q_0 in Theorem 2 reduces to Z_n^*. Hence, $Q_0(P) = Q_0^*(P) = N(0, \rho(P))$.

5.1.2 Estimation of the null distribution

One can exploit the specific form of the t-statistics defined by equations (35) and (36) to derive a consistent estimator of the null distribution $Q_0^* = N(0, \rho(P))$ as follows.

When the m-vector influence curve $IC(X \mid P) = (IC_j(X \mid P) : j = 1, \ldots, m)$ for the estimator μ_n is available (e.g., tests of means and correlations below), one can estimate Q_0^* by the m-variate Gaussian distribution $Q_{0n}^* = N(0, \rho_n)$, where ρ_n is the correlation matrix corresponding to the

$m \times m$ estimated IC covariance matrix,

$$\Sigma_n \equiv \frac{1}{n}\sum_{i=1}^{n} IC_n(X_i)IC_n^T(X_i), \tag{38}$$

and $IC_n(X) = (IC_{jn}(X) : j = 1, \ldots, m)$ is an estimator of $IC(X \mid P)$.

In cases where the influence curve is not readily available, $\rho(P)$ can be estimated with the bootstrap as follows. For each bootstrap sample, $X_1^\#, \ldots, X_n^\# \sim P_n^\star$, compute the estimator $\mu_n^\#$. A bootstrap estimator of the correlation matrix $\rho(P)$ is given by the covariance matrix $\rho_n = Cov_{P_n^\star}[{Z_n}^\#]$ of the standardized statistics $Z_n^\#$

$$Z_n^\#(j) \equiv \sqrt{n}\frac{(\mu_n^\#(j) - E_{P_n^\star}[{\mu_n}^\#(j)])}{\sqrt{Var_{P_n^\star}[{\mu_n}^\#(j)]}}, \qquad j = 1, \ldots, m. \tag{39}$$

The estimated null distribution is then given by $Q_{0n}^* = N(0, \rho_n)$. A bootstrap estimator of the null distribution Q_0^* is also provided by the joint distribution of the standardized statistics $Z_n^\#$. Note that, when an estimator of the IC is available, using the bootstrap to estimate $\rho(P)$ does not necessarily pay off over direct estimation based on the sample. When the correlation matrix is sparse, shrinkage estimation methods may be beneficial.

Alternately, a consistent estimator of the null distribution Q_0 can be obtained using bootstrap Procedure 3 of Section 4, which follows the general construction of Theorem 2, with null-values $\lambda_0(j) \equiv 0$ and $\tau_0(j) \equiv 1$. As mentioned at the beginning of Section 4, above, one of the main advantages of a parametric estimator $Q_{0n}^* = N(0, \rho_n)$, is that it is continuous and hence does not suffer from the discreteness of non-parametric bootstrap estimators. Similar issues arise for the F-statistics discussed in Section 5.2.

5.1.3 Example: Tests of means

A familiar testing problem, that falls within the single-parameter hypothesis testing framework, is that where $X_1, \ldots, X_n$ are n i.i.d. random d-vectors, $X \sim P$, and the parameter of interest is the mean vector $\mu = \mu(P) = E[X] = (\mu(j) : j = 1, \ldots, d)$, where $\mu(j) = \mu_j(P) \equiv E[X(j)]$. The $m = d$ null hypotheses, $H_{0j} = \mathrm{I}(\mu(j) \leq \mu_0(j))$, then refer to individual components of the mean vector μ and the test statistics $T_n(j)$ are the usual one-sample t-statistics, where $\mu_n(j) = \bar{X}_n(j) = \frac{1}{n}\sum_i X_i(j)$ and $\sigma_n^2(j) = \frac{1}{n}\sum_i (X_i(j) - \bar{X}_n(j))^2$ are empirical means and variances for the

d components, respectively. In this simple case, the components of the m-vector influence curve are $IC_j(X \mid P) = X(j) - \mu(j)$ and can be estimated by $IC_{jn}(X) = X(j) - \bar{X}_n(j)$. Thus, a consistent estimator of Q_0^* is the m-variate Gaussian distribution, $N(0, \rho_n)$, where ρ_n is the $m \times m$ sample correlation matrix.

5.1.4 Example: Tests of correlations

Another common testing problem covered by this framework is that where the parameter of interest is the $d \times d$ correlation matrix for a random d-vector $X \sim P$: $\Gamma = \Gamma(P) = (\gamma_{jk}(P) : j,\ k = 1, \ldots, d)$, where $\gamma(j,k) = \gamma_{jk}(P) \equiv Cor[X(j), X(k)]$. Suppose we are interested in testing the $m = d(d-1)/2$ null hypotheses that the d components of X are uncorrelated, that is, null hypotheses $H_{0,jk} = \mathrm{I}(\gamma(j,k) = 0)$, $j = 1, \ldots, d$, $k = j+1, \ldots, d$. Common test statistics for this problem are $T_n(j,k) \equiv \sqrt{n}\gamma_n(j,k)$, where $\gamma_n(j,k)$ are the sample correlations. As discussed in Westfall and Young (1993), Example 2.2, p. 43, subset pivotality fails for this testing problem. To see this, consider the simple case where $d = 3$ and assume that $H_{0,12}$ and $H_{0,13}$ are true, so that $\gamma(1,2) = \gamma(1,3) = 0$. Then, the joint distribution of $(T_n(1,2), T_n(1,3))$ is asymptotically normal with mean vector zero, variance one, and correlation $\gamma(2,3)$, and thus depends on the truth or falsity of the third hypothesis $H_{0,23}$. In other words, the covariance matrix of the vector influence curve for the sample correlations is not the same under the true P as it is under a null distribution P_0 for which $\gamma(j,k) \equiv 0$, $\forall j \neq k$. Tests of correlations thus provide an example where standard procedures based on subset pivotality fail, while procedures based on the t-statistics specific null distribution from Theorem 5 or the general construction from Theorem 2 achieve the desired Type I error control.

The influence curves for the sample correlations $\gamma_n(j,k)$ can be obtained by applying the Delta-method to the function $f(\eta(j,k))$ defined as follows

$$\gamma(j,k) = f(\eta(j,k)) \equiv \frac{\mu(j,k) - \mu(j)\mu(k)}{\sqrt{\mu(j,j) - \mu^2(j)}\sqrt{\mu(k,k) - \mu^2(k)}},$$

where $\eta(j,k) = \eta_{jk}(P)$ is the 5×1 parameter vector
$\eta(j,k) \equiv [\mu(j), \mu(k), \mu(j,j), \mu(k,k), \mu(j,k)]^T$, $\mu(j) = \mu_j(P) \equiv E[X(j)]$, and $\mu(j,k) = \mu_{jk}(P) \equiv E[X(j)X(k)]$, $j,\ k = 1, \ldots, d$. Let $f'(\eta)$ denote the 1×5 gradient vector of $f(\eta)$. Then,

$$\gamma_n(j,k) - \gamma(j,k) = f'(\eta(j,k))[\eta_n(j,k) - \eta(j,k)] + o_P(1/\sqrt{n}),$$

where $\eta_n(j,k) = [\mu_n(j), \mu_n(k), \mu_n(j,j), \mu_n(k,k), \mu_n(j,k)]^T$ is a 5×1 estimator of $\eta(j,k)$ based on the sample moments. Hence, the influence curve for the estimator $\gamma_n(j,k)$ is

$$\begin{aligned} IC_{jk}(X \mid P) &= f'(\eta(j,k))(\eta_1(j,k) - \eta(j,k)) \\ &= f'(\eta(j,k))\big[X(j), X(k), X^2(j), X^2(k), X(j)X(k)\big]^T \\ &\quad -f'(\eta(j,k))\big[\mu(j), \mu(k), \mu(j,j), \mu(k,k), \mu(j,k)\big]^T. \end{aligned}$$

5.1.5 Example: Tests of regression parameters

Suppose $X_1, \ldots, X_n$ are n i.i.d. random $(m+1)$-vectors, with $X = (X(1), \ldots, X(m), Y) \sim P$, and consider the following model for the conditional expected value of the outcome Y given individual explanatory variables $X(j)$

$$E[Y \mid X(j)] = g(X(j) \mid \beta(j)), \qquad j = 1, \ldots, m, \tag{40}$$

where $\beta(j) = (\beta_0(j), \beta_1(j))$ are the regression coefficients for variable $X(j)$. The parameter vector of interest is the m-vector of slope parameters, $\beta_1 = (\beta_1(j) : j = 1, \ldots, m)$, and we wish to test the m null hypotheses $H_{0j} = \mathrm{I}(\beta_1(j) = 0)$, $j = 1, \ldots, m$. One can estimate the regression parameters $\beta(j)$ for each variable $X(j)$ using the method of least squares, that is, by seeking $\beta(j)$ that minimizes the sum of squared residuals, $\sum_i \big(Y_i - g(X_i(j) \mid \beta(j))\big)^2$. The *least squares estimator*, $\beta_n(j) = (\beta_{0n}(j), \beta_{1n}(j))$, is obtained by solving the following equation for β

$$\begin{aligned} 0 &= \frac{\partial}{\partial \beta} \sum_{i=1}^{n} \Big(Y_i - g(X_i(j) \mid \beta)\Big)^2, \\ \text{that is,} \qquad 0 &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial}{\partial \beta} g(X_i(j) \mid \beta)\right) \Big(Y_i - g(X_i(j) \mid \beta)\Big). \end{aligned}$$

Under regularity conditions, one can show that

$$\beta_n(j) - \beta(j) = \frac{1}{n} \sum_{i=1}^{n} c^{-1}(\beta(j)) \left(\frac{\partial}{\partial \beta} g(X_i(j) \mid \beta)\right)\Big|_{\beta=\beta(j)} \Big(Y_i - g(X_i(j) \mid \beta(j))\Big) + o_P(1/\sqrt{n}),$$

where

$$c(\beta) \equiv E \begin{bmatrix} \left(\frac{\partial}{\partial \beta_0} g(X(j) \mid \beta)\right)^2 & \left(\frac{\partial}{\partial \beta_0} g(X(j) \mid \beta)\right)\left(\frac{\partial}{\partial \beta_1} g(X(j) \mid \beta)\right) \\ \left(\frac{\partial}{\partial \beta_0} g(X(j) \mid \beta)\right)\left(\frac{\partial}{\partial \beta_1} g(X(j) \mid \beta)\right) & \left(\frac{\partial}{\partial \beta_1} g(X(j) \mid \beta)\right)^2 \end{bmatrix}.$$

Let $IC_j(X \mid P) = IC_j(X \mid \beta(j)) = \big(IC_{j0}(X \mid \beta(j)), IC_{j1}(X \mid \beta(j))\big)$ denote the two-dimensional vector influence curve for the least squares estimator $\beta_n(j)$ of the regression parameters corresponding to variable $X(j)$. From the above expansion

$$IC_j(X \mid \beta(j)) = c^{-1}(\beta(j)) \left(\frac{\partial}{\partial\beta} g(X(j) \mid \beta)\right)\Big|_{\beta=\beta(j)} \Big(Y - g(X(j) \mid \beta(j))\Big). \tag{41}$$

The m-dimensional vector influence curve for the least squares estimators, $\beta_{1n} = (\beta_{1n}(j) : j = 1, \ldots, m)$, of the m slope parameters is

$$IC(X \mid P) = (IC_{j1}(X \mid \beta(j)) : j = 1, \ldots, m).$$

The covariance matrix of the vector influence curve $IC(X \mid P)$ is

$$\Sigma(P) = E\left[IC(X \mid P) IC^T(X \mid P)\right]$$

and can be estimated as discussed in Section 5.1.2, using the sample covariance matrix based on an estimator $IC_n(X)$ of the vector influence curve.

E.g. Linear regression. A common model for a continuous outcome Y is the *linear model*

$$E[Y \mid X(j)] = g(X(j) \mid \beta(j)) = \beta_0(j) + \beta_1(j)X(j), \qquad j = 1, \ldots, m.$$

In this case, the influence curves are given by

$$IC_j(X \mid \beta(j)) = \frac{1}{Var[X(j)]} \begin{bmatrix} E[X^2(j)] & -E[X(j)] \\ -E[X(j)] & 1 \end{bmatrix} \begin{bmatrix} 1 \\ X(j) \end{bmatrix} \Big(Y - \beta_0(j) - \beta_1(j)X(j)\Big).$$

E.g. Logistic regression. For a binary outcome $Y \in \{0, 1\}$, the *logistic model* is

$$Pr(Y = 1 \mid X(j)) = g(X(j) \mid \beta(j)) = \frac{\exp(\beta_0(j) + \beta_1(j)X(j))}{1 + \exp(\beta_0(j) + \beta_1(j)X(j))}, \qquad j = 1, \ldots, m.$$

Here,

$$\left(\frac{\partial}{\partial\beta} g(X(j) \mid \beta)\right)\Big|_{\beta=\beta(j)} = \frac{\exp(\beta_0(j) + \beta_1(j)X(j))}{\Big(1 + \exp(\beta_0(j) + \beta_1(j)X(j))\Big)^2} \begin{bmatrix} 1 \\ X(j) \end{bmatrix},$$

and the influence curves can be derived by substituting for $\frac{\partial}{\partial\beta} g(X(j) \mid \beta)$ in equation (41), above.

5.2 F-statistics for multiple-parameter hypotheses

Consider random m-vectors $X_k \sim P_k$, from K different populations with data generating distributions P_k, $k = 1, \ldots, K$. Denote the mean vector and covariance matrix in population k by $\mu_k \equiv E[X_k]$ and $\Sigma_k \equiv Cov[X_k]$, respectively. We are interested in testing the m null hypotheses, $H_{0j} = \mathrm{I}(\mu_1(j) = \mu_2(j) = \ldots = \mu_K(j))$, that the jth component of the mean vectors is constant across populations, $j = 1, \ldots, m$. Suppose we observe i.i.d. samples, $X_{k,1}, \ldots, X_{k,n_k}$, of size n_k from population k, $k = 1, \ldots, K$. Let $n \equiv \sum_k n_k$ denote the total sample size and $\delta_{k,n} \equiv n_k/n$ the proportion of observations from population k in the sample, where it is assumed that $\forall k$, $\delta_{k,n} \to \delta_k > 0$ as $n \to \infty$.

As test statistics one can use the well-known *F-statistics*

$$T_n(j) \equiv \frac{1/(K-1)\sum_{k=1}^{K} n_k(\bar{X}_{k,n_k}(j) - \bar{X}_n(j))^2}{1/(n-K)\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_{k,i}(j) - \bar{X}_{k,n_k}(j))^2}, \qquad j = 1, \ldots, m, \tag{42}$$

where $\bar{X}_{k,n_k}$ denotes the sample mean vector for population k and $\bar{X}_n = \sum_k \delta_{k,n}\bar{X}_{k,n_k}$ denotes the overall mean vector. Next, we propose an F-specific null distribution Q_0^* that provides asymptotic control of Type I error rates of the form $\theta(F_{V_n})$ when used in Procedures 1 and 2.

5.2.1 Null distribution

Theorem 6 *The F-statistics $T_n = (T_n(j) : j = 1, \ldots, m)$ satisfy asymptotic null domination Assumption AQ0 in Theorem 1, where the null distribution $Q_0^* = Q_0^*(P)$ is the joint distribution of a random m-vector $Z^* = (Z^*(j) : j = 1, \ldots, m)$, of quadratic forms $Z^*(j)$, defined as*

$$Z^*(j) \equiv \frac{1}{(K-1)\sum_{k=1}^{K}\delta_k\sigma_k^2(j)}\left(\sum_{k=1}^{K}(1-\delta_k)Y_k^2(j) - \sum_{\substack{k=1\\k\neq l}}^{K}\sum_{l=1}^{K}\sqrt{\delta_k\delta_l}Y_k(j)Y_l(j)\right), \tag{43}$$

in terms of K independent Gaussian m-vectors $Y_k = (Y_k(j) : j = 1, \ldots, m) \sim N(0, \Sigma_k)$. In matrix form,

$$Z^*(j) = Y^T(j)A(j)Y(j), \qquad j = 1, \ldots, m, \tag{44}$$

where the $Y(j)$ are K-vectors, $Y(j) = (Y_k(j) : k = 1, \ldots, K)$, and $A(j)$ is a symmetric $K \times K$ matrix with entries

$$A_{kl}(j) \equiv \frac{1}{(K-1)\sum_{k=1}^{K} \delta_k \sigma_k^2(j)} \begin{cases} (1-\delta_k), & \text{if } k = l, \\ -\sqrt{\delta_k \delta_l}, & \text{if } k \neq l. \end{cases} \tag{45}$$

Thus, by Theorem 1, single-step Procedures 1 and 2, based on F-statistics T_n and the null distribution Q_0^*, provide asymptotic control of general Type I error rates $\theta(F_{V_n})$, for the test of multiple-parameter null hypotheses $H_{0j} = \mathrm{I}(\mu_1(j) = \mu_2(j) = \ldots = \mu_K(j))$, $j = 1, \ldots, m$.

In addition, the quadratic forms $Z^*(j)$ have means and variances given by

$$\begin{aligned} E[Z^*(j)] &= \frac{1}{(K-1)\sum_{k=1}^{K} \delta_k \sigma_k^2(j)} \sum_{k=1}^{K} (1-\delta_k)\sigma_k^2(j), \qquad (46) \\ Var[Z^*(j)] &= \frac{2}{(K-1)^2 (\sum_{k=1}^{K} \delta_k \sigma_k^2(j))^2} \Big(\big(\sum_{k=1}^{K} (1-2\delta_k)\sigma_k^4(j) \big) + \big(\sum_{k=1}^{K} \delta_k \sigma_k^2(j) \big)^2 \Big). \end{aligned}$$

In the special case of constant population variances, $\sigma_k^2(j) \equiv \sigma^2(j)$, then $E[Z^*(j)] = 1$, $Var[Z^*(j)] = 2/(K-1)$, and the quadratic forms $(K-1)Z^*(j)$ have marginal $\chi^2(K-1)$ distributions.

Proof of Theorem 6. Firstly, note that the denominators of the F-statistics can be written as

$$D_n(j) = \frac{n}{n-K} \sum_k \delta_{k,n} \sigma_{k,n_k}^2(j), \qquad j = 1, \ldots, m, \tag{47}$$

where $\sigma_{k,n_k}^2(j)$ are consistent estimators of the population variances $\sigma_k^2(j)$, i.e., of the diagonal elements of covariance matrices Σ_k, $k = 1, \ldots, K$. Thus, as $n \to \infty$,

$$D_n(j) \xrightarrow{P} D(j) = \sum_k \delta_k \sigma_k^2(j), \qquad j = 1, \ldots, m. \tag{48}$$

The numerators of the F-statistics $T_n(j)$ can be rewritten as *quadratic forms*

$$\begin{aligned} N_n(j) &= \frac{1}{K-1}\sum_k \left((1-\delta_{k,n})Y_{k,n_k}(j) - \sum_{l\neq k}\sqrt{\delta_{k,n}\delta_{l,n}}Y_{l,n_l}(j)\right)^2 \\ &= \frac{1}{K-1}\left(\sum_k (1-\delta_{k,n})Y^2_{k,n_k}(j) - \sum_{k\neq l}\sum_l \sqrt{\delta_{k,n}\delta_{l,n}}Y_{k,n_k}(j)Y_{l,n_l}(j)\right), \end{aligned}$$

where $Y_{k,n_k} = (Y_{k,n_k}(j) : j = 1,\ldots,m)$ are K independent m-vectors defined by $Y_{k,n_k}(j) \equiv \sqrt{n_k}(\bar{X}_{k,n_k}(j) - \bar{\mu}(j))$ and $\bar{\mu}(j) = \sum_k \delta_k\mu_k(j)$, $k = 1,\ldots,K$. Thus, the F-statistics $T_n = (T_n(j) : j = 1,\ldots,m)$ can be approximated by a random m-vector $Z^*_n = (Z^*_n(j) : j = 1,\ldots m)$ of quadratic forms

$$\begin{aligned} T_n(j) &\approx \frac{N_n(j)}{D(j)} \\ &\approx \frac{1}{(K-1)\sum_k \delta_k\sigma^2_k(j)}\left(\sum_k (1-\delta_k)Y^2_{k,n_k}(j) - \sum_{k\neq l}\sum \sqrt{\delta_k\delta_l}Y_{k,n_k}(j)Y_{l,n_l}(j)\right) \\ &\equiv Z^*_n(j), \qquad j = 1,\ldots,m, \end{aligned} \tag{49}$$

that is, by a simple quadratic function $f(Y_{1,n_1},\ldots,Y_{K,n_K}) = (f_j(Y_{1,n_1},\ldots,Y_{K,n_K}) : j = 1,\ldots,m)$ of the m-vectors Y_{k,n_k}, $k = 1,\ldots,K$. By the Central Limit Theorem, $(Y_{k,n_k}(j) : j \in S_0) \stackrel{\mathcal{L}}{\Rightarrow} (Y_k(j) : j \in S_0)$, where the $Y_k = (Y_k(j) : j = 1,\ldots,m)$ are independent m-vectors with $Y_k \sim N(0,\Sigma_k)$, $k = 1,\ldots,K$. For $j \notin S_0$, $Y_{k,n_k}(j) = \sqrt{n_k}(\bar{X}_{k,n_k}(j) - \mu_k(j)) + \sqrt{n_k}(\mu_k(j) - \bar{\mu}(j))$ converges to either $+\infty$ or $-\infty$ for some k. Applying the Continuous Mapping Theorem to the function $(f_j(Y_{1,n_1},\ldots,Y_{K,n_K}) : j \in S_0)$ proves that $(T_n(j) : j \in S_0)$ converges in distribution to $(Z^*(j) : j \in S_0)$, where $Z^* = f(Y_1,\ldots,Y_K)$ and the Y_k are independent m-vectors with $Y_k \sim N(0,\Sigma_k)$, $k = 1,\ldots,K$. That is, the limit distribution of $(T_n(j) : j \in S_0)$ is directly implied by the multivariate Gaussian distributions $N(0,\Sigma_k)$, where Σ_k denotes the $m \times m$ covariance matrix of the data generating distribution P_k for the kth population, $k = 1,\ldots,K$.

Therefore, the F-statistics T_n satisfy Assumption AQ0 of Theorem 1, where the null distribution $Q^*_0 = Q^*_0(P)$ is the joint distribution of the random m-vector $Z^* = (Z^*(j) : j = 1,\ldots,m)$ of quadratic forms $Z^*(j) = f_j(Y_1,\ldots,Y_K)$, defined in terms of K independent Gaussian m-vectors, $Y_k \sim N(0,\Sigma_k)$, $k = 1,\ldots,K$.

The moments of $Z^*(j)$ are obtained from standard results on quadratic forms (Theorem 1, p. 55, and Corollary 1.3, p. 57, in Searle (1971)). In the special case of constant variances across populations, $Diag(\Sigma_k) \equiv Diag(\Sigma)$, the matrices $(K-1)A(j)Cov[Y(j)]$ are idempotent, and hence, the quadratic forms $(K-1)Z^*(j)$ have marginal $\chi^2(K-1)$ distributions (Theorem 2, p. 57, in Searle (1971)).

□

The above theorem involves a null distribution Q_0^* that was derived specifically in terms of the F-statistics T_n in equation (42). In this case, $(T_n(j) : j \in S_0) \stackrel{\mathcal{L}}{\Rightarrow} Q^*_{0,S_0}$, but for $j \notin S_0$, $T_n(j) \to \infty$. Key Assumption AQ0 in Theorem 1 is nonetheless satisfied, as it only concerns test statistics corresponding to the *true* null hypotheses (i.e., $j \in S_0$). Convergence to Q_0^* is not needed for the false null hypotheses. Note that the distribution Q_0^* is entirely determined by the covariance matrices Σ_k and proportions δ_k (via the matrices $A(j)$ and the $N(0, \Sigma_k)$ distribution for the Y_k), thus the main task is to estimate these quantities from the sample. Properties of the marginal distributions Q^*_{0j} follow from standard univariate results on quadratic forms. The main contribution of the theorem is that it provides a null distribution Q_0^* resulting in multiple testing procedures that take into account the *joint* distribution of the test statistics, i.e., the correlation structure of the null distribution Q_0^* is implied by the correlation structure of the data generating distributions P_k, via Σ_k in the definition of the quadratic forms.

Gaussian distributions with constant population variances. In the special case when the X_k have Gaussian distributions with common covariance matrix Σ, i.e., $X_k \sim N(\mu_k, \Sigma)$, the test statistics have marginal F-distributions (Section 2.4, Searle (1971)): $T_n(j) \sim F(\nu_1, \nu_2, \lambda_n(j))$, with degrees of freedom $\nu_1 = K-1$ and $\nu_2 = n-K$, and non-centrality parameters

$$\lambda_n(j) = \frac{1}{\sigma^2(j)} \sum_k n_k(\mu_k(j) - \bar{\mu}(j))^2, \qquad \bar{\mu}(j) = \sum_k \delta_k \mu_k(j). \tag{50}$$

For the true null hypotheses, i.e., for $j \in S_0$, $\lambda_n(j) \equiv 0$. For the false null hypotheses (corresponding to non-local alternatives), i.e., for $j \notin S_0$, the non-centrality parameters $\lambda_n(j) \to \infty$ as $n \to \infty$. In addition, as $n \to \infty$,

$\nu_2 \to \infty$. The means and variances of the F-statistics are given by

$$E[T_n(j)] = \frac{(\nu_1 + \lambda_n(j))\nu_2}{\nu_1(\nu_2 - 2)} \to \begin{cases} 1, & \text{if } j \in S_0 \\ \infty, & \text{if } j \notin S_0 \end{cases} \tag{51}$$

and

$$\begin{aligned} Var[T_n(j)] &= \frac{2\nu_2^2(\nu_1^2 + (2\lambda_n(j) + \nu_2 - 2)\nu_1 + \lambda_n(j)(\lambda_n(j) + 2\nu_2 - 4))}{\nu_1^2(\nu_2 - 4)(\nu_2 - 2)^2} \\ &\to \begin{cases} 2/(K-1), & \text{if } j \in S_0 \\ \infty, & \text{if } j \notin S_0. \end{cases} \end{aligned} \tag{52}$$

Also, $(K-1)T_n(j) \stackrel{\mathcal{L}}{\Rightarrow} \chi^2(K-1, \lambda_n(j))$.

Comparison to general proposal from Theorem 2. Instead of the F-specific distribution Q_0^* from Theorem 6, one could use the general construction in Theorem 2, whereby the null distribution Q_0 is defined as the limit distribution of

$$Z_n(j) \equiv \sqrt{\min\left(1, \frac{\tau_0(j)}{Var[T_n(j)]}\right)}\Big(T_n(j) + \lambda_0(j) - E[T_n(j)]\Big), \qquad j = 1, \ldots, m.$$

Here, the null-values $\lambda_0(j)$ and $\tau_0(j)$ are based on the means and variances of the asymptotic distribution of the test statistics T_n for the true null hypotheses, i.e., on $E[Z^*(j)]$ and $Var[Z^*(j)]$, given in equation (46) of Theorem 6. In the special case of constant variances, $\sigma_k^2(j) \equiv \sigma^2(j)$, the null-values are independent of the unknown data generating distributions P_k: $\lambda_0(j) \equiv 1$ and $\tau_0(j) \equiv 2/(K-1)$. Otherwise, one needs to estimate δ_k and $\sigma_k^2(j)$ from the sample in order to use equation (46). Note that scaling the test statistics T_n is important in the construction of Z_n, as the F-statistics for non-local alternative hypotheses converge to infinity. Without the scaling, one could have asymptotically infinite cut-offs and hence no power against the alternatives.

The F-specific null distribution Q_0^* and the general null distribution Q_0 from Theorem 2 are the same for the true null hypotheses ($j \in S_0$), but may differ for the false null hypotheses. Thus, in choosing between Q_0^* and Q_0, the main issue is power at local alternatives (for non-local alternatives, the asymptotic power is one for both distributions).

5.2.2 Estimation of the null distribution

A consistent estimator of the general null distribution Q_0 of Theorem 2 can be obtained using bootstrap Procedure 3 of Section 4, with null-values $\lambda_0(j)$ and $\tau_0(j)$ based on equation (46) of Theorem 6. Unless the variances $\sigma_k^2(j)$ are constant across populations, this involves estimating $Diag(\Sigma_k)$ and δ_k by their empirical counterparts.

Alternately, one can exploit specific properties of F-statistics to derive a consistent estimator of the null distribution Q_0^* of Theorem 6. In this approach, the samples $X_{k,1}, \ldots, X_{k,n_k}$, $k = 1, \ldots, K$, are used to derive estimators Σ_{k,n_k} and $\delta_{k,n} = n_k/n$, of the population covariance matrices Σ_k and proportions δ_k, respectively. The null distribution Q_0^* is then estimated simply by the distribution of the m-vector of quadratic forms, defined in terms of independent m-vectors $Y_k \sim N(0, \Sigma_{k,n_k})$, using the sample analogue of equation (43). Note that unlike the general bootstrap estimator from Procedure 3, for the null distribution of Theorem 2, this F-specific estimator has the advantage of being continuous.

A third, F-specific bootstrap-based approach involves resampling the centered observations, $X_{k,i} - \bar{X}_{k,n_k}$, and defining the null distribution as the bootstrap distribution of the resulting F-statistics. In this method, the null distribution of the test statistics is based on a data generating null distribution. The last two approaches provide consistent estimators of the same null distribution Q_0^* described in Theorem 6.

6 Strong control, weak control, and subset pivotality

The multiple testing procedures proposed in this article are concerned with controlling a specified Type I error rate under only one distribution, namely, the *true* underlying data generating distribution P, i.e., the joint distribution $Q_n = Q_n(P)$ of the test statistics T_n implied by P. With such an approach, the notions of strong and weak control of a Type I error rate therefore become irrelevant. In this section, we attempt nonetheless to formalize these concepts and the associated property of subset pivotality, and discuss how they relate to the approach introduced in Section 2.6.

6.1 Strong and weak control of a Type I error rate

As discussed in Hochberg and Tamhane (1987), p. 3, and Westfall and Young (1993), p. 9–10, the multiple testing literature commonly distinguishes between weak and strong control of a Type I error rate. *Weak control* refers to control of the Type I error rate under a data generating distribution P_0 that satisfies the *complete null hypothesis*, $H_0^C = \prod_{j=1}^m H_{0j} = \prod_{j=1}^m \mathrm{I}(P \in \mathcal{M}_j)$, that all m null hypotheses are true, i.e., under a distribution $P_0 \in \cap_{j=1}^m \mathcal{M}_j$. In contrast, *strong control*, as defined in Westfall and Young (1993), considers *all possible subsets* $\Lambda_0 \subseteq \{1, \ldots, m\}$ of null hypotheses and refers to control of the Type I error rate under distributions satisfying *any* of these 2^m subsets of null hypotheses. In particular, strong control implies weak control for $\Lambda_0 = \{1, \ldots, m\}$. As detailed below, the definitions of weak and strong control implicitly assume a mapping, $\Lambda_0 \rightarrow P_{\Lambda_0}$, from subsets Λ_0 of null hypotheses to data generating distributions, $P_{\Lambda_0} \in \cap_{j \in \Lambda_0} \mathcal{M}_j$, satisfying these null hypotheses. While strong control does consider the subset $S_0 = S_0(P)$ of true null hypotheses corresponding to the true data generating distribution P, control under P is not guaranteed, unless the mapping $\Lambda_0 \rightarrow P_{\Lambda_0}$ results in $P_{S_0} = P$.

We note that in much of the multiple testing literature, Type I error rates are defined loosely as probabilities *given subsets of null hypotheses*, rather than as probabilities *under distributions satisfying subsets of null hypotheses*. For example, for control of the family-wise error rate, Westfall and Young (1993), p. 9, define FWEP as the family-wise error rate "... computed under the *partial null hypothesis* (meaning that some subcollection of nulls, say $H_{j_1}, \ldots, H_{j_t}$, is true)" and provide the following definition in equation (1.2)

$$FWEP = Pr(\text{Reject at least one } H_i,\ i = j_1, \ldots, j_t \mid H_{j_1}, \ldots, H_{j_t} \text{ are true}).$$

As discussed in Pollard and van der Laan (2003), such a quantity is not well-defined, because Type I error rates are *parameters of a distribution* for the number of Type I errors (and possibly the total number of rejected hypotheses, as for the FDR) and can only be defined meaningfully with respect to such a distribution. A more precise definition would be that FWEP is the Type I error rate *under a distribution* P_{Λ_0}, defined to satisfy a certain subset $\Lambda_0 = \{j_1, \ldots, j_t\}$ of null hypotheses, i.e., for a data generating distribution $P_{\Lambda_0} \in \cap_{j \in \Lambda_0} \mathcal{M}_j$. This immediately raises the issue of how to map from a subset Λ_0 of null hypotheses to a well-defined data generating distribution $P_{\Lambda_0} \in \cap_{j \in \Lambda_0} \mathcal{M}_j$. Except in very simple situations (e.g., for null hypothe-

ses concerning the mean vector of a multivariate Gaussian data generating distribution), each subset Λ_0 of null hypotheses corresponds to a *family* of possible distributions. One approach is to define the distribution P_{Λ_0} as a *projection* of the true underlying data generating distribution P onto the submodel $\cap_{j\in\Lambda_0}\mathcal{M}_j$, selecting, for example, the distribution with smallest Kullback-Leibler divergence with the true P. That is,

$$P_{\Lambda_0} = \Pi(P \mid \cap_{j\in\Lambda_0}\mathcal{M}_j) = \text{argmax}_{P'\in\cap_{j\in\Lambda_0}\mathcal{M}_j} \int \log\left(\frac{dP'(x)}{d\mu(x)}\right) dP(x),$$

for a dominating measure μ. Another possibility is to select the distribution P_{Λ_0} on the conservative boundary of the submodel $\cap_{j\in\Lambda_0}\mathcal{M}_j$. The reader is referred to Pollard and van der Laan (2003) for a discussion of multivariate null distributions and proposals for specifying such joint distributions based on projections of the data generating distribution P onto submodels satisfying the null hypotheses. However, as discussed by these authors, in many testing problems of interest, one simply cannot identify a data generating null distribution $P_0 \in \cap_{j=1}^m \mathcal{M}_j$ that provides asymptotic (or finite sample) control of the Type I error rate under the true distribution P. In such cases, the *assumed* null distribution $Q_{n,S_0}(P_0)$ and the *true* distribution $Q_{n,S_0}(P)$ for the S_0-specific subvector $(T_n(j) : j \in S_0)$ of test statistics have different limits, that violate the null domination condition for the Type I error rate, i.e., that result in $\limsup_n \theta(F_{V_n}) \geq \theta(F_{V_0}) = \alpha$. Instead, as in the present article for the test of single-parameter null hypotheses using t-statistics (Section 5.1), Pollard and van der Laan (2003) recommend using a test statistics null distribution Q_0, such as the Kullback-Leibler projection of $Q_n = Q_n(P)$ onto the space of multivariate Gaussian distributions with mean vector zero. The latter corresponds with the limit distribution $Q_0^*(P) = N(0, \rho(P))$ in Theorem 5.

As in Section 2.3, consider now a multiple testing procedure based on test statistics T_n and null distribution Q_0. For a level α test, denote the set of rejected hypotheses by $S_n = S(T_n, Q_0, \alpha) \subseteq \{1, \ldots, m\}$. The number of Type I errors under a data generating distribution $P_{\Lambda_0} \in \cap_{j\in\Lambda_0}\mathcal{M}_j$ is given by

$$V_n(\Lambda_0) \equiv V(Q_0 \mid Q_n(P_{\Lambda_0})) = \mid S(T_n, Q_0, \alpha) \cap \Lambda_0 \mid,$$

where $Q_n(P_{\Lambda_0})$ denotes the joint distribution of the test statistics T_n implied

by the data generating distribution P_{Λ_0}. For single-step Procedures 1 and 2,

$$V_n(\Lambda_0) = V(Q_0 \mid Q_n(P_{\Lambda_0})) = \sum_{j \in \Lambda_0} \mathrm{I}(T_n(j) > c_j(Q_0, \alpha)),$$

where $c(Q_0, \alpha) = (c_j(Q_0, \alpha) : j = 1, \ldots, m)$ are cut-offs for the test statistics derived under the null distribution Q_0. Strong control of the Type I error rate at level α would require that

$$\begin{aligned} \max_{\Lambda_0 \subseteq \{1,\ldots,m\}} \theta(F_{V_n(\Lambda_0)}) &\leq \alpha \qquad \text{[finite sample strong control]} \quad (53) \\ \limsup_{n\to\infty} \max_{\Lambda_0 \subseteq \{1,\ldots,m\}} \theta(F_{V_n(\Lambda_0)}) &\leq \alpha \qquad \text{[asymptotic strong control]}, \end{aligned}$$

where, as defined above, P_{Λ_0} is a data generating distribution satisfying the subset of null hypotheses Λ_0, that is, $P_{\Lambda_0} \in \cap_{j \in \Lambda_0} \mathcal{M}_j$. Thus strong control involves considering 2^m distributions, each of them corresponding to a subset Λ_0 of null hypotheses. Note also that this definition of strong control is completely dependent on the definition of the mapping $\Lambda_0 \to P_{\Lambda_0}$. Weak control corresponds to $\Lambda_0 = \{1, \ldots, m\}$ and $P_0 = P_{\{1,\ldots,m\}}$. Control under the true underlying distribution P does not necessarily follow from strong control, unless the mapping $\Lambda_0 \to P_{\Lambda_0}$ results in $P_{S_0} = P$ for $\Lambda_0 = S_0$. In other words, control under the true P could fail under strong control, when an improper mapping for P_{S_0} is used. In contrast, as discussed in Section 2.6, the methodology proposed in this series of articles is only concerned with control under the true P, i.e., for asymptotic control we focus on procedures that satisfy

$$\limsup_{n\to\infty} \theta(F_{V_n}) \leq \alpha,$$

where $V_n = V_n(S_0) = V(Q_0 \mid Q_n(P))$.

6.2 Subset pivotality

In practice, it is not feasible to consider all 2^m possible subsets of true null hypotheses and current single-step or stepwise multiple testing procedures are typically based on cut-offs derived under a data generating distribution P_0 that satisfies the complete null hypothesis, $H_0^C = \prod_{j=1}^m H_{0j}$. Strong control, and in particular control under the truth, is then claimed to follow from weak control under conditions such as subset pivotality. As stated in Condition 2.1, p. 42, in Westfall and Young (1993), "The distribution of $\mathbf{P}$ has the *subset*

pivotality property if the joint distribution of the subvector $\{P_i : i \in K\}$ is identical under the restrictions $\cap_{i \in K} H_{0i}$ and H_0^C, for all subsets $K = \{i_1, \ldots, i_j\}$ of true null hypotheses." In our notation, K is a subset $\Lambda_0 \subseteq \{1, \ldots, m\}$ and $\mathbf{P}$ refers to the vector $(P_{0n}(j) : j = 1, \ldots, m)$ of unadjusted p-values (Section 3.3).

As for the definitions of weak and strong control, the definition of subset pivotality implicitly assumes a mapping, $\Lambda_0 \rightarrow P_{\Lambda_0}$, from subsets Λ_0 of null hypotheses to data generating distributions, $P_{\Lambda_0} \in \cap_{j \in \Lambda_0} \mathcal{M}_j$, satisfying these null hypotheses. The (finite sample) subset pivotality condition for test statistics can be restated as follows, in terms of distributions P_{Λ_0}, for subsets Λ_0 of null hypotheses,

$$Q_{n,\Lambda_0}(P_0) = Q_{n,\Lambda_0}(P_{\Lambda_0}), \qquad \forall \Lambda_0 \subseteq \{1, \ldots, m\}. \tag{54}$$

Note that the subset pivotality condition considers all 2^m possible subsets Λ_0 of $\{1, \ldots, m\}$, and not simply the subset $\Lambda_0 = S_0(P)$ corresponding to the true underlying data generating distribution P. In this sense, when $P_{S_0} = P$, the condition is stronger than needed, since it is only of interest to control Type I error rates under the true P, that is, the only relevant condition is $Q_{n,S_0}(P_0) = Q_{n,S_0}(P)$ for $\Lambda_0 = S_0$. In general, however, subset pivotality may not guarantee control under the true P, if an improper mapping $\Lambda_0 \rightarrow P_{\Lambda_0}$ is used, so that $P_{S_0} \neq P$. Finally, as discussed in Section 2.6, subset pivotality (equation (54)) differs from our finite sample null domination condition (equation (12)) which: (i) only considers the subset $\Lambda_0 = S_0(P)$; (ii) does not require the test statistics null distribution Q_0 to be defined in terms of a data generating null distribution P_0, as $Q_n(P_0)$; and (iii) does not require equality of distributions, but the weaker domination: $Q_{0,S_0} \leq Q_{n,S_0}(P)$.

Software

Software implementing the bootstrap single-step and step-down multiple testing procedures will be available in the R package `multtest`, released as part of the Bioconductor Project (`www.bioconductor.org`).

References

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stistical Society*,

57:289–300, 1995.

R. Beran. Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, 83:679–686, 1988.

S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1):111–139, 2002.

B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

Y. Ge, S. Dudoit, and T. P. Speed. Resampling-based multiple testing for microarray data analysis. *TEST*, 12(1), 2003.

R. D. Gill. Non- and semi-parametric maximum likelihood estimators and the von Mises method. I. *Scand. J. Statist.*, 16(2):97–128, 1989. ISSN 0303-6898. With a discussion by J. A. Wellner and J. Præstgaard and a reply by the author.

R. D. Gill, M. J. van der Laan, and J. A. Wellner. Inefficient estimators of the bivariate survival function for three models. *Ann. Inst. H. Poincaré Probab. Statist.*, 31(3):545–597, 1995. ISSN 0246-0203.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M.L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–802, 1988.

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

S. Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6:65–70, 1979.

E. Manduchi, G. R. Grant, S. E. McKenzie, G. C. Overton, S. Surrey, and C. J. Stoeckert. Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics*, 16: 685–698, 2000.

K. S. Pollard, M. D. Birkner, S. Dudoit, and M. J. van der Laan. Multiple testing. Part IV. Assessment of multiple testing procedures: Simulation studies and applications to genomic data analysis. Technical report, Division of Biostatistics, UC Berkeley, 2004. (In preparation).

K. S. Pollard and M. J. van der Laan. Resampling-based multiple testing: Asymptotic control of Type I error and applications to gene expression data. Technical Report 121, Division of Biostatistics, UC Berkeley, 2003.

A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19 (3):368–375, 2003.

S. R. Searle. *Linear Models*. John Wiley & Sons, 1971.

J. P. Shaffer. Multiple hypothesis testing. *Annu. Rev. Psychol.*, 46:561–584, 1995.

V. Goss Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proc. Natl. Acad. Sci.*, 98:5116–5121, 2001.

M. J. van der Laan, S. Dudoit, and K. S. Pollard. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. Technical Report 139, Division of Biostatistics, UC Berkeley, 2003a. URL `www.bepress.com/ucbbiostat/paper139/`.

M. J. van der Laan, S. Dudoit, and K. S. Pollard. Multiple testing. Part III. Procedures for control of the generalized family-wise error rate and proportion of false positives. Technical Report 140, Division of Biostatistics, UC Berkeley, 2003b. URL `www.bepress.com/ucbbiostat/paper140/`.

A. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

P. H. Westfall. Resampling-based multiple testing for microarray data analysis. *TEST*, 12(1), 2003. (Discussion).

P. H. Westfall and S. S. Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons, 1993.

P. H. Westfall, D. V. Zaykin, and S. S. Young. Multiple tests for genetic effects in association studies. In S. Looney, editor, *Biostatistical Methods*, volume 184 of *Methods in Molecular Biology*, pages 143–168. Humana Press, Toloway, NJ, 2001.

Y. Xiao, M. R. Segal, D. Rabert, A. H. Ahn, P. Anand, L. Sangameswaran, D. Hu, and C. A. Hunt. Assessment of differential gene expression in human peripheral nerve injury. *BioMed Central Genomics*, 3(1):28, 2002.