JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

KATHERINE S. POLLARD
MERRILL D. BIRKNER
MARK J. VAN DER LAAN
SANDRINE DUDOIT

Test statistics null distributions in multiple testing : simulation studies and applications to genomics

Journal de la société française de statistique, tome 146, nº 1-2 (2005), p. 77-115

<http://www.numdam.org/item?id=JSFS_2005__146_1-2_77_0>

TEST STATISTICS NULL DISTRIBUTIONS IN MULTIPLE TESTING: SIMULATION STUDIES AND APPLICATIONS TO GENOMICS

Katherine S. POLLARD *, Merrill D. BIRKNER **,

Mark J. VAN DER LAAN**, Sandrine DUDOIT** (1)

RÉSUMÉ

Les tests d'hypothèses multiples sont fréquemment utilisés dans le domaine de la recherche biomédicale et génomique, en l'occurrence, pour l'identification de gènes différentiellement exprimés et co-exprimés à partir des données issues de puces à ADN. Nous avons développé des procédures de tests multiples avec rééchantillonnage, à pas simple mais aussi pas à pas, pour contrôler une vaste classe de taux d'erreurs de première espèce, définis par des probabilités de queues de distributions et des espérances de fonctions arbitraires du nombre de faux positifs et du nombre total d'hypothèses nulles rejetées. Parmi les contributions fondamentales de notre méthodologie, notons la caractérisation générale et la construction explicite d'une *distribution nulle pour les statistiques de test* (plutôt qu'une distribution génératrice de données nulle). Cette distribution garantit le contrôle du taux d'erreurs de première espèce pour des problèmes de tests multiples pour des lois génératrices de données présentant une structure de dépendance quelconque, des hypothèses nulles définies de manière générale en terme de sous-modèles, et des statistiques de test arbitraires.

Cet article présente des études par simulation pour la comparaison de distributions nulles des statistiques de test, sous deux scénarios particulièrement pertinents à l'analyse de données biomédicales et génomiques : les tests sur les coefficients de régression pour des modèles linéaires dans le cas où les covariables et les erreurs peuvent être dépendantes et les tests sur les coefficients de corrélation. Les études par simulation démontrent que le choix d'une distribution nulle peut considérablement influer les taux d'erreurs de première espèce d'une procédure donnée de tests multiples. Les procédures fondées sur notre distribution nulle bootstrap non-paramétrique pour les statistiques de test contrôlent le taux d'erreurs de première espèce au niveau nominal, alors que des procédures comparables, fondées sur des distributions bootstrap paramétriques nulles pour les données, peuvent être très anti-conservatrices ou conservatrices. L'analyse de données sur l'expression de microARN dans des tissus cancéreux et non-cancéreux (Lu *et al.*, 2005), par tests

* Department of Biomolecular Engineering, University of California, Santa Cruz

** Division of Biostatistics, University of California, Berkeley

(1) Auteur correspondant. Division of Biostatistics, University of California, Berkeley 140 Earl Warren Hall, #7360, Berkeley, CA 94720-7360, sandrine@stat.berkeley.edu

pour coefficients de régression logistique et coefficients de corrélation, illustre la flexibilité et la puissance de notre méthodologie.

Mots clés : Biopuces, bootstrap, classification, co-expression, corrélation, cut-off, distribution nulle, erreur de première espèce, expression différentielle, facteur de confusion, génomique, maxT, microARN, non paramétrique, permutation, puissance, P-valeurs ajustées, rééchantillonnage, région de rejet, régression linéaire, régression logistique, simulation, statistique de test, taux global d'erreur, tests multiples.

ABSTRACT

Multiple hypothesis testing problems arise frequently in biomedical and genomic research, for instance, when identifying differentially expressed and co-expressed genes in microarray experiments. We have developed generally applicable resampling-based single-step and stepwise multiple testing procedures (MTP) for controlling a broad class of Type I error rates, defined as tail probabilities and expected values for arbitrary functions of the numbers of false positives and rejected null hypotheses. A key feature of the methodology is the general characterization and explicit construction of a *test statistics null distribution* (rather than data generating null distribution), which provides Type I error control in testing problems involving general data generating distributions (with arbitrary dependence structures among variables), null hypotheses defined in terms of submodels, and test statistics.

This article presents simulation studies comparing test statistics null distributions in two testing scenarios of great relevance to biomedical and genomic data analysis: tests for regression coefficients in linear models where covariates and error terms are allowed to be dependent and tests for correlation coefficients. The simulation studies demonstrate that the choice of null distribution can have a substantial impact on the Type I error properties of a given multiple testing procedure. Procedures based on our proposed non-parametric bootstrap test statistics null distribution typically control the Type I error rate "on target" at the nominal level, while comparable procedures, based on parameter-specific bootstrap data generating null distributions, can be severely anti-conservative or conservative. The analysis of microRNA expression data from cancerous and non-cancerous tissues (Lu *et al.*, 2005), using tests for logistic regression coefficients and correlation coefficients, illustrates the flexibility and power of our proposed methodology.

Keywords: Adjusted *p*-value, bootstrap, cluster analysis, co-expression, confounding variable, correlation, cut-off, differential expression, family-wise error rate, genomics, linear regression, logistic regression, maxT, microarray, microRNA, multiple hypothesis testing, non-parametric, null distribution, permutation, power, rejection region, resampling, simulation study, test statistic, Type I error rate.

1. Introduction

1.1. Motivation

The genomic age has brought growing interest in multiple testing. As new high-throughput biotechnologies, such as, DNA microarrays, mass spectrometry, and capillary sequencing, facilitate the collection of high-dimensional biological datasets, researchers are becoming increasingly reliant on statistical methods for assessing the significance of biological findings over families of thousands or even millions of hypothesis tests. The identification of differentially expressed genes and co-expressed genes from genome-wide mRNA expression data are classic examples. Other applications include: tests of association between gene expression measures and Gene Ontology (GO) annotation (Dudoit and van der Laan (2005); `www.geneontology.org`); the identification of transcription factor binding sites in ChIP-Chip experiments, where chromatin immunoprecipitation (ChIP) of transcription factor-bound DNA is followed by microarray hybridization (Chip) of the IP-enriched DNA (Keles *et al.*, 2004); tests of association between phenotypes and amino acid mutations, e.g., viral replication capacity and HIV-1 sequence variation (Birkner *et al.*, 2005b,c ; van der Laan *et al.*, 2005); the genetic mapping of complex traits using single nucleotide polymorphisms (SNP) (Birkner *et al.* (2005a); `www.obelinks.org`). These testing problems are particularly challenging, as they involve inference for high-dimensional multivariate distributions, with complex and unknown dependence structures among variables. Therefore, existing methods, based solely on the marginal distributions of the test statistics and/or simplifying assumptions about their joint distribution, are generally not appropriate.

Motivated by the aforementioned biomedical and genomic applications and the limitations of existing multiple testing methods, we have developed and implemented (in R and SAS) resampling-based single-step and stepwise *multiple testing procedures* (MTP) for controlling a broad class of Type I error rates, in testing problems involving general data generating distributions (with arbitrary dependence structures among variables), null hypotheses (defined in terms of submodels for the data generating distribution), and test statistics (e.g., t-statistics, F-statistics) (Birkner *et al.*, 2005b; Dudoit and van der Laan, 2005; Dudoit *et al.*, 2004a,b; Pollard and van der Laan, 2004; Pollard *et al.*, 2005; van der Laan *et al.*, 2004a,b, 2005). In particular, procedures that take into account the *joint* distribution of the test statistics are provided to control Type I error rates defined as tail probabilities and expected values for arbitrary functions $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n. The following quantities are derived to summarize the results of a MTP: rejection regions (i.e., cut-offs) for the test statistics, confidence regions for the parameters of interest, and adjusted p-values.

As demonstrated in the early article of Pollard and van der Laan (2004), a key feature of our proposed MTPs is the *test statistics null distribution* (rather than data generating null distribution) used to obtain rejection regions, confidence regions, and adjusted p-values. Whether testing single or multiple hypotheses, one needs the (joint) distribution of the test statistics in order to derive a procedure that probabilistically controls Type I errors. In practice, however, the true distribution of the test statistics is unknown and replaced by a null distribution. The choice of a suitable null distribution is crucial, in order to ensure that (finite sample or asymptotic) control of the Type I error rate under the *assumed* null distribution does indeed provide the required control under the *true* distribution. This issue is particularly relevant for large-scale testing problems such as those described above in biomedical and genomic research.

Common approaches use a data generating distribution, such as a permutation distribution, that satisfies the *complete null hypothesis* that all null hypotheses are true. Procedures based on such a *data generating null distribution* typically rely on the *subset pivotality* condition stated in Westfall and Young (1993), p. 42–43, to ensure that control under the data generating null distribution does indeed give the desired control under the true data generating distribution. However, the subset pivotality condition is violated in many important testing problems, since a data generating null distribution may result in a joint distribution for the test statistics that has a different dependence structure than their true distribution. In fact, in most 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.

Indeed, subset pivotality fails for two types of testing problems that are highly relevant in biomedical and genomic data analysis: tests concerning correlation coefficients and tests concerning regression coefficients. Tests of correlation arise, for example, when seeking to discover sets of co-expressed genes based on microarray expression measures. Tests concerning regression coefficients in linear and non-linear models (e.g., logistic model, Cox proportional hazards survival model) are commonly used, particularly in medical applications, to identify genes or genomic regions associated with a possibly censored outcome (e.g., survival, tumor class, response to treatment). While subset pivotality holds for some regression models, such as the simple linear model with independent covariates and error terms, it fails for many models used in practice (e.g., linear regression model of Section 3.1 and logistic regression model of Section 4).

1.2. Outline

The present article, inspired by the early work of Pollard and van der Laan (2004), concerns the choice of a test statistics null distribution in multiple testing. Specifically, it investigates the Type I error and power properties of multiple testing procedures based on our general bootstrap test statistics null distribution (Dudoit and van der Laan, 2005; Dudoit *et al.*, 2004b; Pollard and van der Laan, 2004) and various parameter-specific bootstrap data generating null distributions (Westfall and Young, 1993). For the purpose of comparing null distributions, we focus on control of the family-wise error rate (FWER), using the single-step maxT procedure, a common-cut-off procedure exploiting the joint distribution of the test statistics. Note, however, that each null distribution could be employed with any other MTP, including our stepwise joint augmentation and empirical Bayes procedures, for controlling generalized tail probability (gTP) error rates, $gTP(q, g) = Pr(g(V_n, R_n) > q)$, for an arbitrary function $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n (Dudoit and van der Laan, 2005; Dudoit *et al.*, 2004a,b; Pollard and van der Laan, 2004; van der Laan *et al.*, 2004a,b, 2005).

Section 2 provides an overview of our general framework for multiple hypothesis testing and our approach to Type I error control and the choice of a test statistics null distribution. Section 3 describes simulation studies comparing test statistics null distributions in two testing scenarios. The first simulation study considers tests for regression coefficients in linear models with dependent covariates and error terms and compares our general non-parametric bootstrap test statistics null distribution (Procedure 2) to a bootstrap null distribution which involves resampling residuals (Westfall and Young (1993), Section 3.4.1, p. 106–109). The second simulation study considers tests for correlation coefficients and compares our general non-parametric bootstrap test statistics null distribution (Procedure 2) to a bootstrap null distribution which involves resampling individual variables independently (Westfall and Young (1993), Section 6.3, p. 194). The simulation studies demonstrate that the choice of null distribution can have a substantial impact on the Type I error properties of a given multiple testing procedure, such as the single-step maxT MTP. Section 4 applies the single-step maxT procedure, based on the general non-parametric bootstrap test statistics null distribution of Procedure 2, to a dataset of microRNA (miRNA) expression measures from cancerous and non-cancerous tissues (Lu *et al.*, 2005). The first testing problem concerns parameters in a (non-linear) logistic regression model relating cancer status to miRNA expression measures and a tissue type confounding variable, while the second concerns pairwise correlation coefficients for miRNA expression measures. Our methods identify 90 (58 % of the 155 studied) single miRNAs significantly associated with cancer status, as well as hundreds of pairs of miRNAs with significantly correlated profiles across tissue samples. Finally, Section 5 closes with conclusions and a discussion of ongoing efforts.

2. Methods

2.1. Multiple hypothesis testing framework

The present section introduces a general statistical framework for multiple hypothesis testing and discusses in turn the main ingredients of a multiple testing problem. The reader is referred to Dudoit and van der Laan (2005) and Dudoit *et al.* (2004b) for details.

2.1.1 Data generating distribution and parameters

Consider a *random sample*, $\mathcal{X}_n \equiv \{X_i : i = 1, \ldots, n\}$, of n independent and identically distributed (i.i.d.) random variables from a *data generating distribution* P: $X_i \overset{i.i.d.}{\sim} P$, $i = 1, \ldots, n$. Suppose that the data generating distribution P is an element of a particular *statistical model* $\mathcal{M}$, i.e., a set of possibly non-parametric distributions, $P \in \mathcal{M}$. Let P_n denote the corresponding *empirical distribution*, which places probability $1/n$ on each realization of X.

Define *parameters* as arbitrary functions of the data generating distribution P: $\Psi(P) = \psi = (\psi(m) : m = 1, \ldots, M)$, where $\psi(m) = \Psi(P)(m) \in \mathbb{R}$.

2.1.2 Null and alternative hypotheses

General submodel hypotheses. In order to cover a broad class of testing problems, define M pairs of null and alternative hypotheses in terms of a collection of *submodels*, $\mathcal{M}(m) \subseteq \mathcal{M}$, $m = 1, \ldots, M$, for the data generating distribution P. The M *null hypotheses* are defined as $H_0(m) \equiv \mathrm{I}(P \in \mathcal{M}(m))$ and the corresponding *alternative hypotheses* as $H_1(m) \equiv \mathrm{I}(P \notin \mathcal{M}(m))$. Here, $\mathrm{I}(\cdot)$ is the indicator function, equaling 1 if the condition in parentheses is true and 0 otherwise. Thus, $H_0(m)$ is true, i.e., $H_0(m) = 1$, if the data generating distribution P belongs to submodel $\mathcal{M}(m)$; $H_0(m)$ is false otherwise, i.e., $H_0(m) = 0$.

This general submodel representation covers tests for means, quantiles, correlation coefficients, and regression coefficients in linear and non-linear models (e.g., logistic, survival, time-series, and dose-response models).

Parametric hypotheses. In many testing problems, the submodels concern *parameters*, i.e., each null hypothesis may refer to a single parameter, $\psi(m) = \Psi(P)(m) \in \mathbb{R}$. One distinguishes between two types of testing problems for parametric hypotheses, one-sided and two-sided tests.

$$\begin{array}{rll} \textit{One-sided tests} & H_0(m) = \mathrm{I}\big(\psi(m) \leqslant \psi_0(m)\big) & \\ \text{vs.} & H_1(m) = \mathrm{I}\big(\psi(m) > \psi_0(m)\big), & m = 1, \ldots, M. \\ \textit{Two-sided tests} & H_0(m) = \mathrm{I}\big(\psi(m) = \psi_0(m)\big) & \\ \text{vs.} & H_1(m) = \mathrm{I}\big(\psi(m) \neq \psi_0(m)\big), & m = 1, \ldots, M. \end{array}$$

The hypothesized *null values*, $\psi_0(m)$, are frequently zero. For instance, in microarray data analysis, one may be interested in testing the null hypotheses $H_0(m)$ of no differences in mean gene expression measures between two populations of patients or of no pairwise correlations in gene expression measures.

Sets of true and false null hypotheses. Let $\mathcal{H}_0 = \mathcal{H}_0(P) \equiv \{m : H_0(m) = 1\} = \{m : P \in \mathcal{M}(m)\}$ be the set of $h_0 \equiv |\mathcal{H}_0|$ *true null hypotheses*, where we note that $\mathcal{H}_0$ depends on the data generating distribution P. Let $\mathcal{H}_1 = \mathcal{H}_1(P) \equiv \mathcal{H}_0^c(P) = \{m : H_1(m) = 1\} = \{m : P \notin \mathcal{M}(m)\}$ be the set of $h_1 \equiv |\mathcal{H}_1| = M - h_0$ *false null hypotheses*, i.e., *true positives*. The goal of a multiple testing procedure is to accurately estimate the set $\mathcal{H}_0$, and thus its complement $\mathcal{H}_1$, while probabilistically controlling false positives.

Complete null hypothesis. The *complete null hypothesis*, $H_0^C \equiv \prod_{m=1}^M H_0(m) = \prod_{m=1}^M \mathrm{I}(P \in \mathcal{M}(m)) = \mathrm{I}(P \in \cap_{m=1}^M \mathcal{M}(m))$, is true if and only if all M individual null hypotheses $H_0(m)$ are true, i.e., if and only if the data generating distribution P belongs to the intersection $\cap_{m=1}^M \mathcal{M}(m)$ of the M submodels.

2.1.3 Multiple testing procedures

Test statistics. A *testing procedure* is a *data-driven rule* for deciding which null hypotheses should be rejected, i.e., which $H_0(m)$ should be declared false (zero), so that $P \notin \mathcal{M}(m)$. The decisions to reject or not the null hypotheses are based on an M–vector of *test statistics*, $T_n = (T_n(m) : m = 1, \ldots, M)$, that are functions $T_n(m) = T(m; X_1, \ldots, X_n)$ of the data, $X_1, \ldots, X_n$. Denote the typically unknown (finite sample) *joint distribution* of the test statistics T_n by $Q_n = Q_n(P)$.

For the test of single-parameter null hypotheses, $H_0(m) = \mathrm{I}(\psi(m) \leqslant \psi_0(m))$ or $H_0(m) = \mathrm{I}(\psi(m) = \psi_0(m))$, $m = 1, \ldots, M$, consider two main types of test statistics, *difference statistics*,

$$T_n(m) \equiv \text{Estimator} - \text{Null value} = \sqrt{n}(\psi_n(m) - \psi_0(m)), \tag{1}$$

and *t-statistics* (i.e., standardized differences),

$$T_n(m) \equiv \frac{\text{Estimator} - \text{Null value}}{\text{Standard error}} = \sqrt{n}\frac{\psi_n(m) - \psi_0(m)}{\sigma_n(m)}. \tag{2}$$

Here, $\hat{\Psi}(P_n) = \psi_n = (\psi_n(m) : m = 1, \ldots, M)$ denotes an *estimator* for the parameter $\Psi(P) = \psi = (\psi(m) : m = 1, \ldots, M)$ and $(\sigma_n(m)/\sqrt{n} : m = 1, \ldots, M)$ denote the estimated *standard errors* for components $\psi_n(m)$ of ψ_n. Test statistics for other types of null hypotheses include χ^2-statistics, F-statistics, and likelihood ratio statistics.

Multiple testing procedure. A *multiple testing procedure* (MTP) provides *rejection regions*, $\mathcal{C}_n(m)$, i.e., sets of values for each test statistic $T_n(m)$ that lead to the decision to reject the corresponding null hypothesis $H_0(m)$ and declare that $P \notin \mathcal{M}(m)$, $m = 1, \ldots, M$. In other words, a MTP produces a random (i.e., data-dependent) subset $\mathcal{R}_n$ of rejected hypotheses that estimates $\mathcal{H}_1$, the set of true positives,

$$\mathcal{R}_n = \mathcal{R}(T_n, Q_{0n}, \alpha) \equiv \{m : T_n(m) \in \mathcal{C}_n(m)\} = \{m : H_0(m) \text{ is rejected}\}, \tag{3}$$

where $\mathcal{C}_n(m) = \mathcal{C}(m; T_n, Q_{0n}, \alpha)$, $m = 1, \ldots, M$, denote possibly random rejection regions. The long notation $\mathcal{R}(T_n, Q_{0n}, \alpha)$ and $\mathcal{C}(m; T_n, Q_{0n}, \alpha)$ emphasizes that the MTP depends on:

1. the *data*, $\mathcal{X}_n = \{X_i : i = 1, \ldots, n\}$, through the M–vector of *test statistics*, $T_n = (T_n(m) : m = 1, \ldots, M)$;
2. an M–variate (estimated) test statistics *null distribution*, Q_{0n}, for deriving rejection regions, confidence regions, and adjusted p-values;
3. the *nominal level* α of the MTP, i.e., the desired upper bound for a suitably defined Type I error rate.

Rejection regions. Rejection regions are typically defined in terms of intervals, such as, $\mathcal{C}_n(m) = (u_n(m), +\infty)$, $\mathcal{C}_n(m) = (-\infty, l_n(m))$, or $\mathcal{C}_n(m) =$

$(-\infty, l_n(m)) \cup (u_n(m), +\infty)$, where $l_n(m) = l(m; T_n, Q_{0n}, \alpha)$ and $u_n(m) = u(m; T_n, Q_{0n}, \alpha)$ are to-be-determined lower and upper *critical values*, or *cut-offs*, computed under the null distribution Q_{0n} for the test statistics T_n. Rejection regions of the form $\mathcal{C}_n(m) = (-\infty, l_n(m)) \cup (u_n(m), +\infty)$ allow the use of asymmetric cut-offs for two-sided tests. Unless specified otherwise, assume that large values of the test statistic $T_n(m)$ provide evidence against the corresponding null hypothesis $H_0(m)$, that is, consider rejection regions of the form $\mathcal{C}_n(m) = (c_n(m), +\infty)$, based on cut-offs $c_n(m) = c(m; T_n, Q_{0n}, \alpha)$. For two-sided tests of single-parameter null hypotheses using difference or t-statistics (Equations (1) and (2)), one could take absolute values of the test statistics.

2.1.4 *Type I error rates and power*

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 ($\mathcal{R}_n \cap \mathcal{H}_0$), and a *false negative*, or *Type II error*, is committed when the test procedure fails to reject a false null hypothesis ($\mathcal{R}_n^c \cap \mathcal{H}_1$). The situation can be summarized by Table 1, below, where the number of rejected hypotheses is $R_n \equiv |\mathcal{R}_n| = \sum_{m=1}^{M} \mathrm{I}(T_n(m) \in \mathcal{C}_n(m))$, the number of Type I errors is $V_n \equiv |\mathcal{R}_n \cap \mathcal{H}_0| = \sum_{m \in \mathcal{H}_0} \mathrm{I}(T_n(m) \in \mathcal{C}_n(m))$, and the number of Type II errors is $U_n \equiv |\mathcal{R}_n^c \cap \mathcal{H}_1| = \sum_{m \in \mathcal{H}_1} \mathrm{I}(T_n(m) \notin \mathcal{C}_n(m))$. Note that both U_n and V_n depend on the unknown data generating distribution P through the unknown set of true null hypotheses $\mathcal{H}_0 = \mathcal{H}_0(P)$. Therefore, the numbers $h_0 = |\mathcal{H}_0|$ and $h_1 = |\mathcal{H}_1| = M - h_0$ of true and false null hypotheses are *unknown parameters*, the number of rejected hypotheses R_n is an *observable random variable*, and the entries in the body of the table, U_n, $h_1 - U_n$, V_n, and $h_0 - V_n$, are *unobservable random variables*.

		Null hypotheses: not rejected	Null hypotheses: rejected	
Null hypotheses	true	$\lvert\mathcal{R}_n^c \cap \mathcal{H}_0\rvert$	$V_n = \lvert\mathcal{R}_n \cap \mathcal{H}_0\rvert$ (Type I)	$h_0 = \lvert\mathcal{H}_0\rvert$
	false	$U_n = \lvert\mathcal{R}_n^c \cap \mathcal{H}_1\rvert$ (Type II)	$\lvert\mathcal{R}_n \cap \mathcal{H}_1\rvert$	$h_1 = \lvert\mathcal{H}_1\rvert$
		$M - R_n$	$R_n = \lvert\mathcal{R}_n\rvert$	M

TABLE 1. — *Type I and Type II errors in multiple hypothesis testing.* This table summarizes the different types of decisions and errors in multiple hypothesis testing. The number of rejected hypotheses is $R_n \equiv |\mathcal{R}_n| = \sum_{m=1}^{M} \mathrm{I}(T_n(m) \in \mathcal{C}_n(m))$, the number of Type I errors is $V_n \equiv |\mathcal{R}_n \cap \mathcal{H}_0| = \sum_{m \in \mathcal{H}_0} \mathrm{I}(T_n(m) \in \mathcal{C}_n(m))$, and the number of Type II errors is $U_n \equiv |\mathcal{R}_n^c \cap \mathcal{H}_1| = \sum_{m \in \mathcal{H}_1} \mathrm{I}(T_n(m) \notin \mathcal{C}_n(m))$.

Ideally, one would like to simultaneously minimize both the number of Type I errors and the number of Type II errors. Unfortunately, this is not feasible and one seeks a *trade-off* between the two types of errors. A standard approach is to specify an acceptable level α for a suitably defined Type I error rate and derive testing procedures that aim to minimize a Type II error rate, i.e., maximize power, within the class of tests with Type I error rate at most α.

Type I error rate. When testing multiple hypotheses, there are many possible definitions for the Type I error rate of a test procedure. Accordingly, we adopt a general definition for Type I error rates, as *parameters*, $\theta_n = \theta(F_{V_n,R_n})$, of the joint distribution F_{V_n,R_n} of the numbers of Type I errors $V_n = |\mathcal{R}_n \cap \mathcal{H}_0|$ and rejected hypotheses R_n.

This article focuses on the *family-wise error rate* (FWER), that is, the probability of at least one Type I error,

$$FWER \equiv Pr(V_n > 0) = 1 - F_{V_n}(0). \tag{4}$$

The FWER is controlled, in particular, by the classical Bonferroni procedure (Section 2.3).

Power. As with Type I error rates, *power* can be defined generally as a *parameter*, $\vartheta_n = \vartheta(F_{U_n,R_n})$, of the joint distribution F_{U_n,R_n} of the numbers of Type II errors $U_n = |\mathcal{R}_n^c \cap \mathcal{H}_1|$ and rejected hypotheses R_n.

The present article assesses multiple testing procedures in terms of their *average power*, or expected proportion of rejected true positives,

$$AvgPwr \equiv \frac{1}{h_1} E[h_1 - U_n] = 1 - \frac{1}{h_1} \int u dF_{U_n}(u). \tag{5}$$

A variety of other Type I and II error rates are discussed in Dudoit and van der Laan (2005).

2.1.5 Adjusted p-values

The notion of p-value extends directly to multiple testing problems as follows. Consider any multiple testing procedure $\mathcal{R}_n(\alpha) = \mathcal{R}(T_n, Q_{0n}, \alpha)$, with rejection regions $\mathcal{C}_n(m;\alpha) = \mathcal{C}(m; T_n, Q_{0n}, \alpha)$. Then, one can define an M–vector of *adjusted p-values*, $\widetilde{P}_{0n} = (\widetilde{P}_{0n}(m) : m = 1, \ldots, M)$, as

$$\begin{aligned} \widetilde{P}_{0n}(m) &\equiv \inf\{\alpha \in [0,1] : \text{Reject } H_0(m) \text{ at nominal MTP level } \alpha\} \qquad (6) \\ &= \inf\{\alpha \in [0,1] : m \in \mathcal{R}_n(\alpha)\} \\ &= \inf\{\alpha \in [0,1] : T_n(m) \in \mathcal{C}_n(m;\alpha)\}, \qquad m = 1, \ldots, M. \end{aligned}$$

That is, the adjusted p-value $\widetilde{P}_{0n}(m)$, for null hypothesis $H_0(m)$, is the *smallest nominal Type I error level* of the multiple hypothesis testing procedure (e.g., FWER or any other Type I error rate) at which one would reject $H_0(m)$, given T_n. Note that the *unadjusted p-value* $P_{0n}(m)$, for the individual test of null hypothesis $H_0(m)$, corresponds to the special case $M = 1$.

As in single hypothesis tests, the smaller the adjusted p-value, the stronger the evidence against the corresponding null hypothesis. Thus, one rejects $H_0(m)$ for small adjusted p-values $\widetilde{P}_{0n}(m)$. This leads to two equivalent representations for a MTP, in terms of rejection regions for the test statistics and in terms of adjusted p-values,

$$\mathcal{R}_n(\alpha) = \{m : T_n(m) \in \mathcal{C}_n(m;\alpha)\} = \{m : \widetilde{P}_{0n}(m) \leqslant \alpha\}. \tag{7}$$

2.2. Type I error control and choice of a null distribution

2.2.1 General test statistics null distribution

One of the main tasks in specifying a multiple testing procedure is to derive rejection regions for the test statistics such that the *Type I error rate is controlled* at a desired level α, i.e., such that

$$\begin{aligned} \theta(F_{V_n,R_n}) &\leqslant \alpha \qquad \text{[finite sample control]} \\ \limsup_{n\to\infty} \theta(F_{V_n,R_n}) &\leqslant \alpha \qquad \text{[asymptotic control]}. \end{aligned} \tag{8}$$

Note that the Type I error parameter $\theta(F_{V_n,R_n})$ is defined under the *true distribution* $Q_n = Q_n(P)$ of the test statistics T_n, which is a function of the true underlying data generating distribution P. In practice, however, the distribution $Q_n(P)$ is *unknown and replaced by a null distribution* Q_0 (or estimate thereof, Q_{0n}). 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 null distribution does indeed provide the required control under the true distribution $Q_n(P)$. For proper control, the null distribution Q_0 must be such that the Type I error rate under this null distribution dominates the Type I error rate under the true distribution $Q_n(P)$. That is, the following *null domination* condition must be satisfied,

$$\begin{aligned} \theta(F_{V_n,R_n}) &\leqslant \theta(F_{V_0,R_0}) \qquad \text{[finite sample control]} \\ \limsup_{n\to\infty} \theta(F_{V_n,R_n}) &\leqslant \theta(F_{V_0,R_0}) \qquad \text{[asymptotic control]}, \end{aligned} \tag{9}$$

where V_0 and R_0 denote, respectively, the numbers of Type I errors and rejected hypotheses under Q_0, i.e., for $T_n \sim Q_0$.

For error rates $\theta(F_{V_n})$, defined as arbitrary parameters of the distribution of the number of Type I errors V_n, we propose as null distribution $Q_0 = Q_0(P)$, the *asymptotic distribution of the M-vector Z_n of null value shifted and scaled test statistics* (Dudoit and van der Laan, 2005; Dudoit *et al.*, 2004b; Pollard and van der Laan, 2004; van der Laan *et al.*, 2004a),

$$Z_n(m) \equiv \sqrt{\min\left(1, \frac{\tau_0(m)}{Var[T_n(m)]}\right)}\Big(T_n(m) - E[T_n(m)]\Big) + \lambda_0(m),\ m = 1, \dots, M. \tag{10}$$

Single-step and stepwise procedures based on such a null distribution do indeed provide the desired asymptotic control of the Type I error rate $\theta(F_{V_n})$, for general data generating distributions (with arbitrary dependence structures among variables), null hypotheses (defined in terms of submodels for the data generating distribution), and test statistics (e.g., t-statistics, F-statistics).

The construction of the null distribution Q_0 is inspired by null domination Condition (9). As detailed in Theorem 2, p. 32, in Dudoit *et al.* (2004b), the null values $\lambda_0(m)$ and $\tau_0(m)$ are chosen such that $\limsup_n E[T_n(m)] \leqslant \lambda_0(m)$ and $\limsup_n Var[T_n(m)] \leqslant \tau_0(m)$, for $m \in \mathcal{H}_0$. By shifting the test statistics $T_n(m)$ using the location parameters $\lambda_0(m)$, one obtains a sequence of random variables $Z_n(m)$ that are asymptotically stochastically greater than the test statistics $T_n(m)$ for the true null hypotheses $\mathcal{H}_0$. Thus, the number of Type I errors V_0, under the null distribution Q_0, is asymptotically stochastically greater than the number of Type I errors V_n, under the true distribution $Q_n = Q_n(P)$. The resulting null distribution Q_0 therefore satisfies asymptotic null domination Condition (9), under general monotonicity and continuity assumptions for the Type I error rate mapping θ (Assumptions AMI and ACI, p. 12, Dudoit *et al.* (2004b)). In contrast, the scaling parameters $\tau_0(m)$ are not needed for Type I error control. The purpose of $\tau_0(m)$ is to avoid a degenerate null distribution and infinite cut-offs for the true positives ($m \in \mathcal{H}_1$), an important property for power considerations. Note that the null values $\lambda_0(m)$ and $\tau_0(m)$ only depend on the *marginal* distributions of the test statistics $T_n(m)$ for the true null hypotheses $\mathcal{H}_0$ and are generally known from univariate testing. For the test of single-parameter null hypotheses using t-statistics, the null values are $\lambda_0(m) = 0$ and $\tau_0(m) = 1$. For testing the equality of K population means using F-statistics, the null values are $\lambda_0(m) = 1$ and $\tau_0(m) = 2/(K-1)$, under the assumption of equal variances in the different populations.

For a broad class of testing problems, such as the test of single-parameter null hypotheses using t-statistics (Equation (2)), the null distribution Q_0 is an M-variate Gaussian distribution with mean vector zero and covariance matrix $\Sigma^\star(P)$: $Q_0 = Q_0(P) \equiv \mathrm{N}(0, \Sigma^\star(P))$. For tests where the parameter of interest is the M-dimensional mean vector $\Psi(P) = \psi = E[X]$, the estimator ψ_n is simply the M-vector of empirical means and $\Sigma^\star(P)$ is the correlation matrix $Cor[X]$ of $X \sim P$. More generally, for an asymptotically linear estimator ψ_n, $\Sigma^\star(P)$ is the correlation matrix of the vector influence curve (IC). This situation covers standard one-sample and two-sample t-statistics for testing mean parameters, but also test statistics for correlation coefficients (Equation (24)) and regression coefficients in linear and non-linear models (Equations (18) and (28)).

In practice, however, since the data generating distribution P is unknown, then so is the proposed null distribution $Q_0 = Q_0(P)$. Resampling procedures, such as the bootstrap procedures of Section 2.4, may be used to conveniently obtain consistent estimators Q_{0n} of the null distribution Q_0 and of the resulting test statistic cut-offs and adjusted p-values (Dudoit and van der

Laan, 2005; Dudoit *et al.*, 2004b; Pollard and van der Laan, 2004; van der Laan *et al.*, 2004a).

2.2.2 *Contrast with other approaches*

As detailed in Dudoit and van der Laan (2005), Dudoit *et al.* (2004b), and Pollard and van der Laan (2004), the following two main points distinguish our approach from existing approaches to Type I error control and the choice of a null distribution (e.g., in Hochberg and Tamhane (1987) and Westfall and Young (1993)).

Type I error control under the true data generating distribution. 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 $Q_n = Q_n(P)$ for the test statistics T_n implied by P. The concepts of *weak control* and *strong control* are therefore irrelevant in our context.

In particular, the notion of *null domination*, introduced in Equation (9) and discussed in detail in Dudoit and van der Laan (2005) and Dudoit *et al.* (2004b), differs from that of *subset pivotality* (Westfall and Young (1993), p. 42–43) in the following senses: (i) null domination is only concerned with the *true* data generating distribution P, i.e., it only considers the subset $\mathcal{H}_0(P)$ of true null hypotheses and not all possible 2^M subsets $\mathcal{J}_0 \subseteq \{1,\ldots,M\}$ of null hypotheses, and (ii) null domination does not require equality of the joint distributions $Q_{0,\mathcal{H}_0}$ and $Q_{n,\mathcal{H}_0}(P)$ for the $\mathcal{H}_0$–specific test statistics, but the weaker domination of $Q_{n,\mathcal{H}_0}(P)$ by $Q_{0,\mathcal{H}_0}$.

Null distribution for the test statistics. 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 is one that satisfies the complete null hypothesis, $H_0^C = \mathrm{I}(P \in \cap_{m=1}^M \mathcal{M}(m))$, that all M null hypotheses are true, i.e., $P_0 \in \cap_{m=1}^M \mathcal{M}(m)$. 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 (2004), procedures based on $Q_n(P_0)$ do not necessarily provide proper (asymptotic) Type I error control under the true distribution P. Indeed, the assumed null distribution $Q_n(P_0)$ and the true distribution $Q_n(P)$ for the test statistics T_n may converge to distributions with different dependence structures and, as a result, may violate null domination Condition (9) for the Type I error rate. For instance, for test statistics with Gaussian asymptotic distributions, the $\mathcal{H}_0$–specific correlation matrix under the true distribution P may be different from the corresponding correlation matrix under the assumed complete null distribution P_0, that is, one may have $\Sigma^\star_{\mathcal{H}_0}(P) \neq \Sigma^\star_{\mathcal{H}_0}(P_0)$. In the two-sample testing problem, for the commonly-used permutation null distribution P_0, Pollard and van der Laan (2004) show that $\Sigma^\star_{\mathcal{H}_0}(P) = \Sigma^\star_{\mathcal{H}_0}(P_0)$ only if (i) the two populations have the same covariance matrices or (ii) the two sample sizes are equal.

Consequently, approaches based on permutation or other data generating null distributions P_0 (e.g., Korn *et al.* (2004), Troendle (1995, 1996), and Westfall and Young (1993)) are only valid under certain assumptions for the true data generating distribution P. In fact, in most testing problems, there does not exist a data generating null distribution $P_0 \in \cap_{m=1}^{M} \mathcal{M}(m)$ that correctly specifies a joint distribution for the test statistics, i.e., such that the required null domination Condition (9) for the Type I error rate is satisfied.

Thus, unlike current procedures which can only be applied to a limited set of multiple testing problems, the test statistics null distribution Q_0 of Equation (10) leads to single-step and stepwise procedures that provide the desired (asymptotic) Type I error rate control for general data generating distributions, null hypotheses, and test statistics. The 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 and the associated assumption of subset pivotality. Such problems include tests for correlation coefficients and regression coefficients in models where covariates and error terms are allowed to be dependent (Sections 3 and 4).

2.3. Multiple testing procedures

The classical *single-step Bonferroni procedure* is perhaps the most widely-used procedure for controlling the family-wise error rate. For a test at nominal FWER level $\alpha \in [0,1]$, the procedure rejects any hypothesis $H_0(m)$ with unadjusted p-value $P_{0n}(m)$ less than or equal to the common single-step cut-off α/M. The corresponding adjusted p-values are given by,

$$\widetilde{P}_{0n}(m) = \min\big(M\, P_{0n}(m), 1\big), \qquad m = 1, \ldots, M. \tag{11}$$

While simple, this *marginal* procedure can be very conservative for even moderate numbers M of hypotheses. As illustrated in Dudoit *et al.* (2003, 2004a) and van der Laan *et al.* (2005), substantial gains in power can be achieved by taking into account the *joint* distribution of the test statistics, as in the following procedure.

PROCEDURE 1. — **[FWER-controlling single-step maxT procedure]** *The single-step maxT procedure is a joint common-cut-off procedure based on the distribution of the maximum test statistic,* $\max_m Z(m)$, *for an* M*-vector* $Z = (Z(m) : m = 1, \ldots, M) \sim Q_0$ *with the test statistics null distribution* Q_0. *For controlling the FWER at nominal level* $\alpha \in [0,1]$, *the common cut-off* $c(Q_0, \alpha)$, *for the test statistics* $T_n = (T_n(m) : m = 1, \ldots, M)$, *is the* $(1-\alpha)$*-quantile of the distribution of* $\max_m Z(m)$ *under* Q_0,

$$c(Q_0, \alpha) \equiv \inf\left\{c \in \mathbb{R} : Pr_{Q_0}\left(\max_{m \in \{1,\ldots,M\}} Z(m) \leqslant c\right) \geqslant (1-\alpha)\right\}. \tag{12}$$

The corresponding adjusted p-values are given by

$$\widetilde{P}_{0n}(m) = Pr_{Q_0}\left(\max_{m \in \{1,\ldots,M\}} Z(m) \geqslant T_n(m)\right), \qquad m = 1, \ldots, M. \tag{13}$$

For a test at nominal FWER level α, one has two equivalent representations of the set $\mathcal{R}_n(\alpha)$ of rejected hypotheses, in terms of cut-offs for the test statistics and in terms of adjusted p-values,

$$\mathcal{R}_n(\alpha) = \{m : T_n(m) > c(Q_0, \alpha)\} = \{m : \widetilde{P}_{0n}(m) \leqslant \alpha\}.$$

The reader is referred to our earlier articles and book in preparation, for a variety of other joint multiple testing procedures, controlling a broad class of Type I error rates defined as tail probabilities and expected values for arbitrary functions $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n (Dudoit and van der Laan, 2005; Dudoit *et al.*, 2004a,b; Pollard and van der Laan, 2004; van der Laan *et al.*, 2004a,b, 2005)

2.4. Bootstrap-based multiple testing procedures

The test statistics null distribution $Q_0 = Q_0(P)$ defined in Equation (10) depends on the true data generating distribution P and is therefore typically unknown. It can be estimated with the (non-parametric or model-based) bootstrap as detailed in Procedure 2, below. Bootstrap-based test statistic cut-offs and adjusted p-values for FWER-controlling single-step maxT Procedure 1 may then be obtained as in Procedure 3.

PROCEDURE 2. — **[Bootstrap estimation of the test statistics null distribution Q_0]** *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 data, $\mathcal{X}_n = \{X_i : i = 1, \ldots, n\}$. For the model-based bootstrap, $P_n^\star$ belongs to a model $\mathcal{M}$ for the data generating distribution P, such as a family of multivariate Gaussian distributions. One then proceeds as follows to generate the bootstrap test statistics null distribution.*

1. *Obtain the bth bootstrap dataset, $\mathcal{X}_n^b \equiv \{X_i^b : i = 1, \ldots, n\}$, $b = 1, \ldots, B$, by generating n i.i.d. random variables X_i^b with distribution $P_n^\star$.*
2. *For each bootstrap dataset $\mathcal{X}_n^b$, compute an M–vector of test statistics, $T_n^B(\cdot, b) = (T_n^B(m, b) : m = 1, \ldots, M)$, which can be arranged in an $M \times B$ matrix, $\mathbf{T}_n^B \equiv \big(T_n^B(m, b)\big)$, with rows corresponding to the M null hypotheses and columns to the B bootstrap samples.*
3. *For each null hypothesis $H_0(m)$, compute the empirical means $E[T_n^B(m, \cdot)] \equiv \sum_b T_n^B(m, b)/B$ and variances $Var[T_n^B(m, \cdot)] \equiv \sum_b (T_n^B(m, b) - E[T_n^B(m, \cdot)])^2/B$ of the B bootstrap test statistics $T_n^B(m, b)$ (i.e., row means and variances of the matrix $\mathbf{T}_n^B$), to yield estimates of $E[T_n(m)]$ and $Var[T_n(m)]$, respectively, $m = 1, \ldots, M$.*
4. *Obtain an $M \times B$ matrix, $\mathbf{Z}_n^B \equiv \big(Z_n^B(m, b)\big)$, of null value shifted and scaled bootstrap statistics $Z_n^B(m, b)$, as in Equation (10), by row-shifting and scaling the matrix $\mathbf{T}_n^B$ using the bootstrap estimates of $E[T_n(m)]$ and*

$Var[T_n(m)]$ and the user-supplied null values $\lambda_0(m)$ and $\tau_0(m)$. That is,

$$Z_n^B(m,b) \equiv \sqrt{\min\left(1, \frac{\tau_0(m)}{Var[T_n^B(m,\cdot)]}\right)}\left(T_n^B(m,b) - E[T_n^B(m,\cdot)]\right) + \lambda_0(m). \tag{14}$$

5. *The bootstrap estimate Q_{0n} of the null distribution Q_0 from Equation (10) is the empirical distribution of the B columns $Z_n^B(\cdot,b)$ of matrix $\mathbf{Z}_n^B$.*

As detailed in Sections 3.1.2 and 3.2.2, below, general bootstrap Procedure 2 differs in a number of key aspects from commonly-used bootstrap procedures. The latter procedures typically derive a test statistics null distribution $Q_n(P_{0n})$ by first creating a data generating distribution P_{0n} that satisfies the complete null hypothesis, $H_0^C = \mathrm{I}(P \in \cap_{m=1}^M \mathcal{M}(m))$, that all M null hypotheses are true. For example, for tests concerning regression coefficients in Section 3.1.2, procedure `Bootstrap e` resamples residuals to generate bootstrap samples for which the outcome Y is independent of each covariate $X(j)$. "Raw" test statistics T_n are then computed, rather than null value shifted and scaled test statistics Z_n.

PROCEDURE 3. — **[Bootstrap estimation of common cut-offs and adjusted p-values for single-step maxT Procedure 1]**

0. *Apply Procedure 2 to generate an $M \times B$ matrix, $\mathbf{Z}_n^B = \left(Z_n^B(m,b)\right)$, of null value shifted and scaled bootstrap statistics $Z_n^B(m,b)$.*
1. *Compute the maximum statistic, $\max_m Z_n^B(m,b)$, $b = 1,\ldots,B$, for each bootstrap dataset $\mathcal{X}_n^b$, i.e., each column of the matrix $\mathbf{Z}_n^B$.*
2. *For controlling the FWER at nominal level $\alpha \in [0,1]$, the bootstrap single-step maxT common cut-off $c(Q_{0n},\alpha)$ is the $(1-\alpha)$–quantile of the empirical distribution of the B maxima $\{\max_m Z_n^B(m,b) : b = 1,\ldots,B\}$.*
3. *The bootstrap single-step maxT adjusted p-value for null hypothesis $H_0(m)$ is the proportion of maxima $\{\max_m Z_n^B(m,b) : b = 1,\ldots,B\}$ exceeding the corresponding observed test statistic $T_n(m)$,*

$$\widetilde{P}_{0n}(m) \equiv \frac{1}{B}\sum_{b=1}^{B} \mathrm{I}\left(\max_{m\in\{1,\ldots,M\}} Z_n^B(m,b) \geqslant T_n(m)\right), \quad m = 1,\ldots,M. \tag{15}$$

Note that Procedure 3 can be applied, as in Section 3, below, to any matrix $\mathbf{Z}_n^B$ of resampled statistics (e.g., from other bootstrap or permutation procedures).

3. Simulation studies

This section presents two separate simulation studies comparing our general *non-parametric bootstrap test statistics null distribution* (Procedure 2; Dudoit and van der Laan (2005) ; Dudoit *et al.* (2004b) ; Pollard and van der Laan (2004) ; van der Laan *et al.* (2004a)) to *parameter-specific bootstrap* *data generating null distributions* proposed in Westfall and Young (1993). Specifically, the first simulation study considers tests for *regression coefficients* in linear models where the error term is allowed to depend on the covariates and compares the bootstrap null distribution of Procedure 2 to a bootstrap null distribution which involves *resampling residuals* (Westfall and Young (1993), Section 3.4.1, p. 106–109). The second simulation study considers tests for *correlation coefficients* and compares the bootstrap null distribution of Procedure 2 to a bootstrap null distribution which involves *resampling individual variables independently* (Westfall and Young (1993), Section 6.3, p. 194). For both testing problems and each null distribution, Procedure 3, i.e., the resampling version of single-step maxT Procedure 1, is applied to control the family-wise error rate.

As detailed in Sections 3.1.4 and 3.2.4, the simulation results demonstrate that the choice of null distribution can have a substantial impact on the Type I error properties of a given multiple testing procedure, such as the single-step maxT MTP. The general non-parametric bootstrap test statistics null distribution of Procedure 2 typically controls the Type I error rate "on target" at the nominal level α. In contrast, bootstrapping residuals for tests for regression coefficients can lead to severely anti-conservative procedures, while the independent bootstrap for tests for correlation coefficients can lead to conservative procedures.

3.1. Simulation Study 1: Tests for linear regression coefficients in models with dependent covariates and error terms

The first simulation study concerns tests for regression coefficients in linear models where the error term is allowed to depend on the covariates. This represents an important and practical testing scenario, since in many biomedical and genomic applications, error terms and covariates cannot be assumed to be independent and may have a complex and unknown joint distribution (e.g., logistic regression model relating cancer status to miRNA expression measures in Section 4).

3.1.1 Simulation model

Data generating distribution. Consider a data structure $(X,Y) \sim P$, where X is an M–dimensional covariate row vector and Y a univariate outcome. Assume that the pair (X,Y) has an $(M+1)$–dimensional Gaussian distribution P, that satisfies

$$E[X] = 0, \qquad Cov[X] = \sigma_{xx}, \tag{16}$$
$$E[Y|X] = X\psi, \qquad Var[Y|X] = \sigma_{y|X} = s(X),$$

where ψ is an M–dimensional column vector of regression parameters, σ_{xx} an $M \times M$ covariance matrix, and $s(X)$ a scalar function of the covariates X. That is, one can express the outcome Y in terms of the familiar *linear regression model*

$$Y = X\psi + \epsilon, \qquad \text{where} \qquad \epsilon|X \sim \mathrm{N}(0, s(X)), \tag{17}$$

so that,

$$Y|X \sim \mathrm{N}(X\psi, s(X)).$$

Suppose one has a random sample, $\mathcal{XY}_n \equiv \{(X_i, Y_i) : i = 1, \ldots, n\}$, of n independent and identically distributed pairs $(X_i, Y_i) \sim P$, from the above specified Gaussian data generating distribution P. Let $\mathbf{X}_n$ and $\mathbf{Y}_n$ denote, respectively, the $n \times M$ design matrix and the $n \times 1$ outcome vector.

Null and alternative hypotheses. The hypotheses of interest concern the M components of the regression parameter vector ψ. Specifically, consider two-sided tests of the M null hypotheses $H_0(m) = \mathrm{I}(\psi(m) = \psi_0(m))$ vs. the alternative hypotheses $H_1(m) = \mathrm{I}(\psi(m) \neq \psi_0(m))$, $m = 1, \ldots, M$. For simplicity, and without loss of generality, set the null values $\psi_0(m)$ equal to zero, i.e., $\psi_0(m) \equiv 0$.

3.1.2 Multiple testing procedures

Test statistics. The M null hypotheses are tested based on standard t-statistics for *ordinary least squares* (OLS) regression,

$$T_n(m) \equiv \frac{\psi_n(m)}{\sigma_n(m)}, \qquad m = 1, \ldots, M, \tag{18}$$

where $\psi_n = (\psi_n(m) : m = 1, \ldots, M)$ is an M–vector of least squares estimators for the regression parameters, with estimated $M \times M$ covariance matrix σ_n,

$$\psi_n \equiv (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \mathbf{Y}_n, \tag{19}$$

$$\sigma_n \equiv \frac{(\mathbf{Y}_n - \mathbf{X}_n\psi_n)^\top (\mathbf{Y}_n - \mathbf{X}_n\psi_n)}{n - M} (\mathbf{X}_n^\top \mathbf{X}_n)^{-1}.$$

Define an n–vector $\mathbf{e}_n$ of residuals by

$$\mathbf{e}_n \equiv \mathbf{Y}_n - \mathbf{X}_n\psi_n = (e_i \equiv Y_i - X_i\psi_n : i = 1, \ldots, n). \tag{20}$$

The simulation study compares the Type I error and power properties of FWER-controlling single-step maxT Procedure 1, based on the following two different bootstrap test statistics null distributions ($B = 10{,}000$ bootstrap samples).

`Bootstrap XY` **null distribution — Bootstrapping covariate/outcome pairs (X, Y).** The general non-parametric bootstrap test statistics null distribution of Procedure 2 involves *resampling covariate/outcome pairs* (X_i, Y_i) and computing *null value shifted and scaled test statistics* for each bootstrap sample. Specifically, one proceeds as follows for the bth bootstrap sample, $b = 1, \ldots, B$.

1. Sample n covariate/outcome pairs (X_i^b, Y_i^b) at random, with replacement from the set of n observations $\mathcal{XY}_n = \{(X_i, Y_i) : i = 1, \ldots, n\}$. Let $\mathcal{XY}_n^b \equiv \{(X_i^b, Y_i^b) : i = 1, \ldots, n\}$ denote the resulting bootstrap dataset.
2. Compute an M–vector $T_n^B(\cdot, b) = (T_n^B(m, b) : m = 1, \ldots, M)$ of bootstrap test statistics as in Equation (18), based on the bootstrap dataset $\mathcal{XY}_n^b$.
3. Compute an M–vector $Z_n^B(\cdot, b) = (Z_n^B(m, b) : m = 1, \ldots, M)$ of bootstrap null value shifted and scaled test statistics,

$$Z_n^B(m,b) \equiv \sqrt{\min\left(1, \frac{1}{Var[T_n^B(m,\cdot)]}\right)}\left(T_n^B(m,b) - E[T_n^B(m,\cdot)]\right),$$

where $\lambda_0(m) = 0$, $\tau_0(m) = 1$, and $E[T_n^B(m, \cdot)] \equiv \sum_b T_n^B(m, b)/B$ and $Var[T_n^B(m, \cdot)] \equiv \sum_b (T_n^B(m, b) - E[T_n^B(m, \cdot)])^2/B$ denote, respectively, the empirical mean and variance of the B bootstrap test statistics $T_n^B(m, b)$ for null hypothesis $H_0(m)$, $m = 1, \ldots, M$ (i.e., row means and variances of the matrix $\mathbf{T}_n^B$, as in Procedure 2).

The test statistics null distribution is the empirical distribution Q_{0n} of the $B = 10,000$ M–vectors $\{Z_n^B(\cdot, b) : b = 1, \ldots, B\}$, i.e., of the columns of matrix $\mathbf{Z}_n^B$.

`Bootstrap e` **null distribution — Bootstrapping residuals** e**.** In contrast, the parameter-specific bootstrap test statistics null distribution proposed in Section 3.4.1, p. 106–109, of Westfall and Young (1993), involves *resampling residuals* e_i and computing *raw test statistics* (without shifting and scaling) for each bootstrap sample. Specifically, one proceeds as follows for the bth bootstrap sample, $b = 1, \ldots, B$.

1. Sample n residuals at random, with replacement from the set of n observed residuals $\{e_i : i = 1, \ldots, n\}$ defined in Equation (20). Let $\mathbf{e}_n^b = (e_i^b : i = 1, \ldots, n)$ denote the resulting n–vector of bootstrap residuals.
2. Generate n bootstrap covariate/outcome pairs, by randomly pairing each of the n observed covariate vectors X_i with a bootstrap residual e_i^b, that is, by defining a bootstrap outcome n–vector $\mathbf{Y}_n^b \equiv \mathbf{e}_n^b$ as the vector of bootstrap residuals. Let $\mathcal{XY}_n^b \equiv \{(X_i, Y_i^b) : i = 1, \ldots, n\}$ denote the resulting bootstrap dataset.
3. Compute an M–vector $T_n^B(\cdot, b) = (T_n^B(m, b) : m = 1, \ldots, M)$ of bootstrap test statistics as in Equation (18), based on the bootstrap dataset $\mathcal{XY}_n^b$.

The test statistics null distribution is the empirical distribution Q_{0n} of the $B = 10,000$ M–vectors $\{T_n^B(\cdot, b) : b = 1, \ldots, B\}$, i.e., of the columns of matrix $\mathbf{T}_n^B$.

Thus, bootstrap procedures `Bootstrap XY` and `Bootstrap e` differ in two key aspects: (i) the (re)sampling units, `Bootstrap XY` resamples covariate/outcome pairs (X_i, Y_i), while `Bootstrap e` resamples residuals e_i; (ii) the bootstrap test statistics, `Bootstrap XY` relies on null value shifted and scaled test statistics Z_n, while `Bootstrap e` relies on "raw" test statistics T_n. In other words, procedure `Bootstrap e` derives the test statistics null distribution by first creating a data generating null distribution in (i), that corresponds to the complete null hypothesis that the outcome Y is independent of each covariate $X(j)$. Note that bootstrapping covariate/outcome pairs (X_i, Y_i) preserves the correlation structure of the data, while bootstrapping residuals and randomly pairing residuals and covariates destroys this correlation structure.

Single-step maxT procedure. Adjusted p-values for single-step maxT Procedure 1 may be obtained by applying Procedure 3 with bootstrap null distributions `Bootstrap XY` and `Bootstrap e`. Specifically, adjusted p-values for `Bootstrap XY` and `Bootstrap e` are computed, respectively, from the empirical distributions of the B maxima of shifted and scaled test statistics $\{\max_m Z_n^B(m,b) : b = 1,\ldots,B\}$ and raw test statistics $\{\max_m T_n^B(m,b) : b = 1,\ldots,B\}$. For a test at nominal FWER level α, one rejects null hypotheses with adjusted p-values less than or equal to α.

3.1.3 Simulation study design

Simulation parameters. The following model parameters are varied in the simulation study.

- *Sample size, n.* $n = 25$, 100.
- *Number of hypotheses, M.* $M = 10$, 20.
- *Covariance matrix of the covariates, σ_{xx}.* The covariance matrix σ_{xx} of the covariates X has unit diagonal elements and off-diagonal elements set to a common value ς, i.e., $\sigma_{xx}(m,m) = 1$, for $m = 1,\ldots,M$, and $\sigma_{xx}(m,m') = \varsigma$, for $m \neq m' = 1,\ldots,M$. The following values are considered for the common covariance: $\varsigma = 0.10$, 0.50, 0.80.
- *Conditional variance of outcome Y given covariates X, $s(X)$.* $Var[Y|X] = \sigma_{y|X} = s(X) = \sum_{m \notin \mathcal{H}_0} X(m)$.
- *Proportion of true null hypotheses, $\frac{h_0}{M}$.* $\frac{h_0}{M} = 0.50$, 0.75.
- *Alternative regression parameters, $(\psi(m) : m \notin \mathcal{H}_0)$.* For each simulation model, regression parameters $(\psi(m) : m \notin \mathcal{H}_0)$, for the true positives, are generated as $|\mathcal{H}_0^c| = M - h_0$ independent uniform random variables over the interval $[0, \frac{\mu}{\sqrt{n}}]$. That is, $\psi(m) \stackrel{i.i.d.}{\sim} U(0, \frac{\mu}{\sqrt{n}})$, $m \notin \mathcal{H}_0$. The following values are considered for the shift parameter: $\mu = 0.10$, 0.25.

Estimating Type I error rate and power. For each simulation model (i.e., each combination of parameter values n, M, ς, $s(X)$, h_0/M, and μ), generate $A = 500$ random samples, $\mathcal{XY}_n^a \equiv \{(X_i^a, Y_i^a) : i = 1,\ldots,n\}$, of covariate/outcome pairs $(X,Y) \sim P$. For each such simulated dataset, compute adjusted p-values $\widetilde{P}_{0n}^a(m)$ for single-step maxT Procedure 3, based on each of the two bootstrap null distributions (`Bootstrap XY` and `Bootstrap e`). For a given nominal Type I error level α, compute the numbers of rejected hypotheses $R_n^a(\alpha)$, Type I errors $V_n^a(\alpha)$, and Type II errors $U_n^a(\alpha)$,

$$R_n^a(\alpha) \equiv \sum_{m=1}^{M} \mathrm{I}(\widetilde{P}_{0n}^a(m) \leqslant \alpha), \tag{21}$$

$$V_n^a(\alpha) \equiv \sum_{m \in \mathcal{H}_0} \mathrm{I}(\widetilde{P}_{0n}^a(m) \leqslant \alpha),$$

$$U_n^a(\alpha) \equiv \sum_{m \notin \mathcal{H}_0} \mathrm{I}(\widetilde{P}_{0n}^a(m) > \alpha).$$

The *actual Type I error rate* is estimated as follows and then compared to the *nominal Type I error level* α,

$$FWER(\alpha) \equiv \frac{1}{A} \sum_{a=1}^{A} \mathrm{I}(V_n^a(\alpha) > 0). \tag{22}$$

The *average power* of a given MTP is estimated by

$$AvgPwr(\alpha) \equiv 1 - \frac{1}{h_1} \frac{1}{A} \sum_{a=1}^{A} U_n^a(\alpha). \tag{23}$$

The simulation error for the actual Type I error rate and power is of the order $1/\sqrt{A} = 1/\sqrt{500} \approx 0.045$.

Graphical summaries. Simulation results are displayed using the following two main types of graphical summaries.

- **Type I error control comparison.** For a given data generating model, plot, for each MTP, the *difference between the nominal and actual Type I error rates* vs. the *nominal Type I error rate*, i.e., plot

$$(\alpha - FWER(\alpha)) \qquad \text{vs.} \qquad \alpha,$$

  for $\alpha \in \{0, 0.01, 0.02, \ldots, 0.50\}$, i.e., values of α in `seq(from = 0, to = 0.50, by=0.01)`. Positive (negative) differences correspond to (anti-) conservative MTPs; the higher the curve, the more conservative the procedure.

- **Power comparison.** For a given data generating model, *receiver operator characteristic* (ROC) curves may be used to compare different MTPs in terms of power. ROC curves are obtained by plotting, for each MTP, *power* vs. *actual Type I error rate*, i.e., $AvgPwr(\alpha)$ vs. $FWER(\alpha)$, for a range of nominal Type I error levels α. However, due to possibly large variations in power between simulation models, consider instead the following modified display, which facilitates comparisons across models. For a given model, plot the *difference in power* between two procedures vs. the *actual Type I error rate*, i.e., plot

$$(AvgPwr^{Boot\ XY}(\alpha^{Boot\ XY}(a)) - AvgPwr^{Boot\ e}(\alpha^{Boot\ e}(a))) \qquad \text{vs.} \qquad a,$$

where $\alpha^j(\cdot)$ is defined such that $FWER^j(\alpha^j(a)) = a$, $j \in \{Boot\ XY, Boot\ e\}$, for $a \in \{FWER^{Boot\ XY}(\alpha) : \alpha \in \{0, 0.01, 0.02, \ldots, 0.50\}\} \cap \{FWER^{Boot\ e}(\alpha) : \alpha \in \{0, 0.01, 0.02, \ldots, 0.50\}\}$.

3.1.4 Simulation results

Our comparison of the test statistics null distributions `Bootstrap XY` and `Bootstrap e` focusses primarily on Type I error control. All figures are displayed at the end of the paper.

Figure 1 displays differences between nominal and actual Type I error rates for four simulation models, where one parameter is varied as the others remain constant. In general, procedures based on the residual bootstrap null distribution `Bootstrap e` are *anti-conservative* over the entire range of the nominal level α, while procedures based on the general non-parametric bootstrap null distribution `Bootstrap XY` control the Type I error rate close to the target nominal level α. In some testing scenarios, the actual Type I error rate for `Bootstrap e` exceeds the nominal Type I error level by as much as 0.20. The following trends are observed.

- *Covariance matrix of the covariates, σ_{xx}.* (Figure 1, Panels (b) vs. (a)) As the correlation ς between covariates increases, the actual Type I error rate for `Bootstrap XY` gets closer to the nominal level α. In contrast, procedure `Bootstrap e` becomes more anti-conservative as the correlation ς increases.
- *Sample size, n.* (Figure 1, Panels (c) vs. (a)) As the number of observations n increases, the actual Type I error rate for `Bootstrap XY` gets closer to the nominal level α.
- *Alternative regression parameters, $(\psi(m) : m \notin \mathcal{H}_0)$.* (Figure 1, Panels (d) vs. (a)) As the magnitude of the parameter μ, defining the regression coefficients $(\psi(m) : m \notin \mathcal{H}_0)$ for the true positives, increases, the actual Type I error rate for `Bootstrap XY` gets closer to the nominal level α. In contrast, procedure `Bootstrap e` becomes more anti-conservative as the shift μ increases.
- *Proportion of true null hypotheses, $\frac{h_0}{M}$.* No clear trends are noticeable for the proportion of true null hypotheses (data not shown).

For most simulation models, the differences in power are within simulation error (i.e., less than $1/\sqrt{A} = 1/\sqrt{500} \approx 0.045$), for the two versions of bootstrap-based single-step maxT Procedure 3 (Figure 2). The main noticeable trends are, as expected, that power increases with sample size n and effect size μ.

3.2. Simulation Study 2: Tests for correlation coefficients

The second simulation study concerns tests for correlation coefficients, a testing scenario of great interest in genomic applications. Indeed, as illustrated in Section 4, below, a common question in microarray and other high-throughput gene expression assays, is the identification of co-expressed genes, i.e., pairs of genes with correlated expression profiles.

3.2.1 Simulation model

Data generating distribution. Consider a J–dimensional Gaussian random row vector $X \sim P = \mathrm{N}(0, \sigma)$, with mean vector zero and covariance matrix $\sigma = (\sigma(j,j') : j, j' = 1, \ldots, J)$ equal to the corresponding correlation matrix $\rho = (\rho(j,j') : j, j' = 1, \ldots, J)$.

Suppose one has a random sample, $\mathcal{X}_n \equiv \{X_i : i = 1, \ldots, n\}$, of n i.i.d. random variables $X_i \sim P$, from the above specified Gaussian data generating distribution P.

Null and alternative hypotheses. The hypotheses of interest concern the $M \equiv \binom{J}{2} = J(J-1)/2$ distinct entries, $\psi = (\psi(m) : m = 1, \ldots, M)$, of the $J \times J$ correlation matrix ρ. One may recode pairs of row and column indices $\{(j,j') : j = 1, \ldots, (J-1),\ j' = j+1, \ldots, J\}$, for the upper triangle of ρ, into a single index $m = 1, \ldots, M$, defined by $m \equiv (j-1)(2J-j)/2 + (j'-j)$.

Consider two-sided tests of the $M = J(J-1)/2$ null hypotheses $H_0(m) = \mathrm{I}(\psi(m) = \psi_0(m))$ vs. the alternative hypotheses $H_1(m) = \mathrm{I}(\psi(m) \neq \psi_0(m))$, $m = 1, \ldots, M$. For simplicity, and without loss of generality, set the null values $\psi_0(m)$ equal to zero, i.e., test the null hypotheses of no pairwise correlations.

3.2.2 Multiple testing procedures

Test statistics. The M null hypotheses are tested based on the following t-statistics,

$$T_n(m) \equiv \sqrt{n-2}\,\frac{\psi_n(m)}{\sqrt{1-\psi_n^2(m)}}, \qquad m = 1, \ldots, M, \tag{24}$$

where $\psi_n = (\psi_n(m) : m = 1, \ldots, M)$ is the M–vector of distinct empirical correlation coefficients. Specifically, the empirical correlation coefficient for the pair of random variables $(X(j), X(j'))$, corresponding to the mth null hypothesis, is defined as

$$\psi_n(m) = \rho_n(j,j') \equiv \frac{\sigma_n(j,j')}{\sqrt{\sigma_n(j,j)\sigma_n(j',j')}}, \tag{25}$$

based on empirical means $\bar{X}_n(j)$ and covariances $\sigma_n(j,j')$,

$$\bar{X}_n(j) \equiv \frac{1}{n}\sum_{i=1}^{n} X_i(j), \qquad \sigma_n(j,j') \equiv \frac{1}{n}\sum_{i=1}^{n}(X_i(j) - \bar{X}_n(j))(X_i(j') - \bar{X}_n(j')).$$

For Gaussian data generating distributions, the t-statistics in Equation (24) have marginal t-distributions with $(n-2)$ degrees of freedom, under the null hypotheses that the corresponding correlation coefficients are zero, i.e., $\psi(m) = 0$. One could also use unstandardized test statistics,

$$T_n(m) \equiv \sqrt{n}\psi_n(m), \qquad m = 1, \ldots, M. \tag{26}$$

The simulation study compares the Type I error and power properties of FWER-controlling single-step maxT Procedure 1, based on the following two different bootstrap test statistics null distributions ($B = 10,000$ bootstrap samples).

`Bootstrap X` **null distribution — Bootstrapping entire J–vectors X.** The general non-parametric bootstrap test statistics null distribution of Procedure 2 involves *resampling entire J–vectors X_i* and computing *null value shifted and scaled test statistics* for each bootstrap sample. Specifically, one proceeds as follows for the bth bootstrap sample, $b = 1, \ldots, B$.

1. Sample n J–vectors X_i^b at random, with replacement from the set of n observations $\mathcal{X}_n = \{X_i : i = 1, \ldots, n\}$. Let $\mathcal{X}_n^b \equiv \{X_i^b : i = 1, \ldots, n\}$ denote the resulting bootstrap dataset.
2. Compute an M–vector $T_n^B(\cdot, b) = (T_n^B(m, b) : m = 1, \ldots, M)$ of bootstrap test statistics as in Equation (24), based on the bootstrap dataset $\mathcal{X}_n^b$.
3. Compute an M–vector $Z_n^B(\cdot, b) = (Z_n^B(m, b) : m = 1, \ldots, M)$ of bootstrap null value shifted and scaled test statistics,

$$Z_n^B(m, b) \equiv \sqrt{\min\left(1, \frac{1}{Var[T_n^B(m, \cdot)]}\right)} \Big(T_n^B(m, b) - E[T_n^B(m, \cdot)]\Big),$$

where $\lambda_0(m) = 0$, $\tau_0(m) = 1$, and $E[T_n^B(m, \cdot)] \equiv \sum_b T_n^B(m, b)/B$ and $Var[T_n^B(m, \cdot)] \equiv \sum_b (T_n^B(m, b) - E[T_n^B(m, \cdot)])^2/B$ denote, respectively, the empirical mean and variance of the B bootstrap test statistics $T_n^B(m, b)$ for null hypothesis $H_0(m)$, $m = 1, \ldots, M$ (i.e., row means and variances of the matrix $\mathbf{T}_n^B$, as in Procedure 2).

The test statistics null distribution is the empirical distribution Q_{0n} of the $B = 10,000$ M–vectors $\{Z_n^B(\cdot, b) : b = 1, \ldots, B\}$, i.e., of the columns of matrix $\mathbf{Z}_n^B$.

`Bootstrap X(j)` **null distribution — Bootstrapping independent entries $X(j)$ of the J–vectors X.** In contrast, the parameter-specific bootstrap test statistics null distribution proposed in Section 6.3, p. 194, of Westfall and Young (1993), involves *resampling each component $X_i(j)$ of the J–vectors X_i independently* and computing *raw test statistics* (without shifting and scaling) for each bootstrap sample. Specifically, one proceeds as follows for the bth bootstrap sample, $b = 1, \ldots, B$.

1. For each variable $X(j)$, $j = 1, \ldots, J$, sample n j–specific entries $X_i^b(j)$, $i = 1, \ldots, n$, at random, with replacement from the set of n j–specific observations $\{X_i(j) : i = 1, \ldots, n\}$. The ith bootstrap J–vector $X_i^b = (X_i^b(j) : j = 1, \ldots, J)$, $i = 1, \ldots, n$, is obtained by combining J such independently sampled variables. Let $\mathcal{X}_n^b \equiv \{X_i^b : i = 1, \ldots, n\}$ denote the resulting bootstrap dataset.
2. Compute an M–vector $T_n^B(\cdot, b) = (T_n^B(m, b) : m = 1, \ldots, M)$ of bootstrap test statistics as in Equation (24), based on the bootstrap dataset $\mathcal{X}_n^b$.

The test statistics null distribution is the empirical distribution Q_{0n} of the $B = 10,000$ M–vectors $\{T_n^B(\cdot, b) : b = 1, \ldots, B\}$, i.e., of the columns of matrix $\mathbf{T}_n^B$.

As in the regression example of Section 3.1, bootstrap procedures `Bootstrap X` and `Bootstrap X(j)` differ in two key aspects: (i) the (re)sampling units, `Bootstrap X` resamples entire J–vectors X_i, while `Bootstrap X(j)` resamples independent components $X_i(j)$; (ii) the bootstrap test statistics, `Bootstrap X` relies on null value shifted and scaled test statistics Z_n, while `Bootstrap X(j)` relies on "raw" test statistics T_n. In other words, procedure `Bootstrap X(j)` derives the test statistics null distribution by first creating a data generating null distribution in (i), that corresponds to the complete null hypothesis that the J variables $X(j)$, $j = 1, \ldots, J$, are independent.

Single-step maxT procedure. Adjusted p-values for single-step maxT Procedure 1 may be obtained by applying Procedure 3 with bootstrap null distributions `Bootstrap X` and `Bootstrap X(j)`. Specifically, adjusted p-values for `Bootstrap X` and `Bootstrap X(j)` are computed, respectively, from the empirical distributions of the B maxima of shifted and scaled test statistics $\{\max_m Z_n^B(m, b) : b = 1, \ldots, B\}$ and raw test statistics $\{\max_m T_n^B(m, b) : b = 1, \ldots, B\}$. For a test at nominal FWER level α, one rejects null hypotheses with adjusted p-values less than or equal to α.

3.2.3 Simulation study design

Simulation parameters. The following model parameters are used in the simulation study.

- *Sample size, n.* $n = 25$.
- *Number of hypotheses, M.* $M = 45$.
- *Proportion of true null hypotheses,* $\frac{h_0}{M}$. $\frac{h_0}{M} = \frac{25}{45} \approx 0.56$.
- *Correlation matrix, ρ.* The correlation matrix $\rho = (\rho(j, j') : j, j' = 1, \ldots, J)$ (here, equal to the covariance matrix σ) has the following block diagonal form,

$$\rho = \begin{bmatrix} \varrho_{J/2\times J/2} & O_{J/2\times J/2} \\ O_{J/2\times J/2} & \varrho_{J/2\times J/2} \end{bmatrix},$$

  where $O_{J/2\times J/2}$ denotes a $J/2 \times J/2$ matrix of zeros and $\varrho_{J/2\times J/2}$ a $J/2 \times J/2$ matrix with unit diagonal elements and off-diagonal elements set to a common value ϱ, i.e., $\varrho_{J/2\times J/2}(j, j) = 1$, for $j = 1, \ldots, J/2$, and $\varrho_{J/2\times J/2}(j, j') = \varrho$, for $j \neq j' = 1, \ldots, J/2$. The following values are considered for the common block correlation coefficient: $\varrho = 0.30, 0.50, 0.60$.

Note that the only parameter that is varied in the simulation study is the correlation matrix ρ, that is, the parameter of interest in the multiple testing problem.

Estimating Type I error rate and power. As in Section 3.1.3, above, for Simulation Study 1.

Graphical summaries. As in Section 3.1.3, above, for Simulation Study 1.

3.2.4 Simulation results

Our comparison of the test statistics null distributions `Bootstrap X` and `Bootstrap X(j)` focusses primarily on Type I error rate control.

Figure 3 displays differences between nominal and actual Type I error rates for three simulation models, where the common block correlation coefficient ϱ is varied as the other parameters remain constant. In general, procedures based on the independent covariates bootstrap null distribution `Bootstrap X(j)` are *conservative* over the entire range of the nominal level α, while procedures based on the general non-parametric bootstrap null distribution `Bootstrap X` control the Type I error rate close to the target nominal level α. The most extreme differences are observed for large nominal Type I error levels α. In some testing scenarios, the nominal level for `Bootstrap X(j)` exceeds the actual Type I error rate by as much as 0.25. As the correlation parameter ϱ increases, procedure `Bootstrap X(j)` becomes more conservative.

As in the regression simulation study of Section 3.1, we find that, for most simulation models, the differences in power are within simulation error (i.e., less than $1/\sqrt{A} = 1/\sqrt{500} \approx 0.045$), for the two versions of bootstrap-based single-step maxT Procedure 3 (Figure 4). The main noticeable trends are, as expected, that power increases with sample size n and effect size ϱ.

Similar trends are observed for standardized (Equation (24)) and unstandardized (Equation (26)) correlation test statistics (data not shown for unstandardized statistics).

4. microRNA data analysis

In addition to playing the important role of passing genetic messages from DNA to the protein-making machinery of the cell, *ribonucleic acids* (RNA) serve many other cellular functions. A new class of small, non-coding RNAs, known as *microRNAs* (miRNA), are currently the subject of intense study due to their provocative roles as gene regulators (`miRBase`, `microrna.sanger.ac.uk`; Wienholds and Plasterk (2005)). By binding to *messenger RNA* (mRNA), miRNAs regulate gene expression post-transcriptionally and affect the abundance of a wide range of proteins, in diverse biological processes. By now, hundreds of miRNAs have been identified, in various multicellular organisms, including the fruitfly *Drosophila melanogaster* (*D. melanogaster*) and humans, and many are evolutionary conserved. Although the biological functions of miRNAs are still largely unknown, miRNAs are predicted to regulate up to 30 % of all protein-coding genes. Each mammalian miRNA is believed to regulate approximately 200 genes and many genes have several target sites for one or several different miRNAs. The large number of miRNA genes, their diverse expression patterns, and the abundance of miRNA targets, suggest the involvement of miRNAs in a variety of diseases, including cancers and viruses. More than half of the known human miRNA genes are located in genomic regions related to cancers, such as, fragile sites, minimal regions of loss of heterozigosity, minimal regions of amplification, and common breakpoint regions.

In their recent study, monitoring miRNA levels in cells derived from cancerous and non-cancerous tissues, Lu *et al.* (2005) made an astonishing discovery: predictors based on abundance of the several hundred known mammalian miRNAs are better able to distinguish developmental lineage, differentiation state, and cancer state, than the best corresponding predictors based on genome-wide mRNA expression profiles from the same cells. miRNA expression profiling may therefore be a valuable tool for the classification of poorly differentiated tumors.

The analysis in Lu *et al.* (2005) includes a comparison of miRNA expression measures between cancerous and non-cancerous tissues, using the FWER-controlling *marginal* Bonferroni procedure, with modified two-sample t-statistics (adjusted p-values for the Bonferroni procedure are given in Equation (11)). For a test at nominal FWER level $\alpha = 0.05$, the authors found that 59 % of the miRNAs were significantly less abundant in cancerous compared to non-cancerous tissues. Only a few miRNAs were over-expressed in cancerous tissues and none significantly so. Furthermore, miRNA measures were observed to vary greatly among the 19 different tissue types represented in the dataset (e.g., stomach, colon, pancreas); tissue type is therefore regarded as a *confounding variable.*

Motivated by these findings, we have undertaken further analyses of this publicly available miRNA dataset. We first consider the identification of *differentially expressed* miRNAs between cancerous and non-cancerous tissues. Our approach is based on tests for regression coefficients in (non-linear) logistic models relating cancer status to miRNA expression measures, while adjusting for the confounding tissue type variable. The second analysis concerns the identification of *co-expressed* miRNAs, i.e., pairs of miRNAs with correlated expression profiles across tissue samples. Both testing problems are addressed using FWER-controlling *joint* single-step maxT Procedure 1, based on the general non-parametric bootstrap test statistics null distribution of Procedure 2 (adjusted p-values given in Equation (13)). This method identifies 90 (58 % of the 155 studied) significantly differentially expressed miRNAs, as well as hundreds of significantly co-expressed pairs of miRNAs.

4.1. miRNA dataset of Lu *et al.* (2005)

Lu *et al.* (2005) measured expression levels for 217 known human miRNAs, by a bead-based flow cytometric profiling method, in cells from 46 cancerous and 140 non-cancerous tissues ($n = 186$ target samples in total). The pre-processed, $\log_2$-transformed data are available from the authors' website (`www.broad.mit.edu/cancer/pub/miGCM`: miRNA expression measures in file `miGCM_218.gct`; probe sequence information in file `supplementary_table_1.xls`; target sample information, such as cancer status and tissue type, in file `supplementary_table_2.xls`). The analyses in Sections 4.2 and 4.3, below, exclude cell lines and any miRNA with expression measures below a detection threshold of $\log_2 32 = 5$ in more than half of the $n = 186$ target samples.

The data for each of the $n = 186$ target samples consist of a binary *outcome/phenotype* Y_i for cancer status (1 for cancerous vs. 0 for non-cancerous tissues), a J–dimensional *covariate/genotype* vector $X_i = (X_i(j) : j = 1, \ldots, J)$ of real-valued expression measures for each of $J = 155$ miRNAs, and a 19–dimensional tissue type indicator vector W_i, $i = 1, \ldots, n$.

4.2. Differentially expressed miRNAs between cancerous and non-cancerous tissues: Tests for logistic regression coefficients

The first analysis of the Lu *et al.* (2005) dataset concerns the identification of *differentially expressed* miRNAs between cancerous and non-cancerous tissues. In order to adjust for the confounding tissue type variable, we apply bootstrap-based single-step maxT Procedure 3 to test hypotheses concerning regression coefficients in logistic models relating cancer status to miRNA expression measures and tissue type. Note that another approach could be based on standard two-sample t-statistics. For such simple tests, data generating null distributions, such as the permutation distribution, lead to proper Type I error control under the conditions that (i) the two populations have the same covariance matrices or (ii) the two sample sizes are equal (Pollard and van der Laan, 2004). The multiple testing methodology proposed in Section 2, however, allows one to use more general and flexible models, such as the logistic regression model of Equation (27), below, which facilitates adjustment for covariates and also provides a simple predictor of cancer status.

4.2.1 Multiple testing procedures

Our approach for identifying differentially expressed miRNAs involves fitting, for each miRNA, a *logistic regression model* that includes expression measure $X(j)$ and tissue type W as covariates. Specifically, the logistic regression model for the jth miRNA is

$$\text{logit}(E[Y|X(j), W]) \equiv \alpha(j) + \beta(j)X(j) + \gamma(j)W, \qquad j = 1, \ldots, J, \tag{27}$$

where $\text{logit}(z) \equiv \log(z/(1-z))$ is the *logit function*, $\alpha(j)$ a baseline effect parameter, $\beta(j)$ a main effect parameter for the expression measure $X(j)$ of the jth miRNA, and $\gamma(j)$ an miRNA-specific 19–dimensional parameter vector adjusting for tissue type W.

The parameter of interest in the logistic model of Equation (27) is $\beta(j)$, the scalar coefficient for the expression measure $X(j)$ of the jth miRNA, $j = 1, \ldots, J$. Thus, for each miRNA, one considers the two-sided test of the null hypothesis $H_0(j) = \mathrm{I}(\beta(j) = 0)$, of no association of the expression measure $X(j)$ with cancer status Y, vs. the alternative hypothesis $H_1(j) = \mathrm{I}(\beta(j) \neq 0)$. Two-sided tests are used to identify both over- and under-expressed miRNAs in cancerous tissues.

The J null hypotheses are tested based on the following t-statistics,

$$T_n(j) \equiv \frac{\beta_n(j) - \beta_0(j)}{\sigma_n(j)}, \qquad j = 1, \ldots, J, \tag{28}$$

where the null values $\beta_0(j)$ are zero and $\beta_n(j)$ are logistic regression parameter estimates with estimated standard errors $\sigma_n(j)$ (as implemented in the R function `glm`, with the call `glm (Y ~ X(j) + W, family = "binomial")`, using the binomial family and iteratively reweighted least squares (IWLS)).

In order to simultaneously test the $J = 155$ null hypotheses of no association of miRNA measures with cancer status, we apply FWER-controlling single-step maxT Procedure 1, with the general non-parametric bootstrap test statistics null distribution of Procedure 2. That is, test statistic cut-offs and adjusted p-values are computed as in Procedure 3 ($B = 5,000$ bootstrap samples). Note that fitting the logistic regression model of Equation (27) allows the identification of differentially expressed miRNAs, while adjusting for the confounding variable tissue type.

4.2.2 Results

For the logistic regression analysis, bootstrap-based single-step maxT Procedure 3 yields 90 miRNAs (58% of the 155 studied) and 53 miRNAs (34% of the 155 studied) with adjusted p-values less than nominal FWER levels of $\alpha = 0.05$ and $\alpha = 0.01$, respectively, thus indicating that some miRNAs are very significantly differentially expressed between cancerous and non-cancerous tissues (Figure 5, Panel (a); Supplementary Table 2, website companion `www.stat.berkeley.edu/ sandrine/Docs/Papers/SFdS05/SFdS.html`). All of the 90 miRNAs that are significantly differentially expressed at level $\alpha = 0.05$ have negative test statistics ($T_n(j) < -3.8$), suggesting under-expression in cancerous compared to non-cancerous tissues. These findings are in agreement with the original publication of Lu *et al.* (2005), the main distinctions being that our analysis takes into account the *joint distribution* of the test statistics and allows *adjusting* for the tissue type *confounding variable*, using the logistic regression model of Equation (27) for the comparison of cancerous and non-cancerous tissues.

Five of the highly significant miRNAs listed in Supplementary Table 2 are located in minimal deleted regions, minimal amplified regions, and breakpoint regions involved in human cancers (Calin *et al.*, 2004). Specifically, `miR-23b` and `let-7d` have been associated with urothelial cancer; `miR-100` with breast, lung, ovarian, and cervical cancers; `miR-22` with hepatocellular cancer; `miR99a` with lung cancer.

4.3. Co-expressed miRNAs: Tests for correlation coefficients

A biological question of great interest in gene expression experiments is the identification of *co-expressed* genes, here, miRNAs with correlated expression measures across tissue samples. While some tests of association between expression measures and a binary outcome, such as cancer status, could be performed with standard multiple testing methods (e.g., MTPs based on a permutation data generating null distribution), correlation tests are a problem for which our bootstrap-based MTPs truly allow one to perform previously unavailable analyses.

4.3.1 Multiple testing procedures

Consider the $M \equiv J(J-1)/2 = 155 \times 154/2 = 11,935$ distinct Pearson correlation coefficients between pairs of miRNA expression profiles,

$$\rho(j,j') \equiv Cor[X(j), X(j')], \qquad j = 1, \ldots, J-1,\ j' = j+1, \ldots, J.$$

It is of interest to identify all pairs of miRNAs with significantly correlated expression profiles across the $n = 186$ target samples. Thus, for each distinct pair (j,j') of miRNAs, one considers the two-sided test of the null hypothesis $H_0(j,j') = \mathrm{I}(\rho(j,j') = 0)$, of no association in expression measures, vs. the alternative hypothesis $H_1(j,j') = \mathrm{I}(\rho(j,j') \neq 0)$.

The M null hypotheses are tested based on the following test statistics,

$$T_n(j,j') \equiv \sqrt{n}\rho_n(j,j'), \qquad j = 1, \ldots, J-1,\ j' = j+1, \ldots, J, \tag{29}$$

where $\rho_n(j,j')$ are empirical correlation coefficients, as defined previously in Equation (25).

In order to simultaneously test the $M = 11,935$ null hypotheses of no association in expression measures between pairs of miRNAs, we again apply FWER-controlling single-step maxT Procedure 1, with the general non-parametric bootstrap test statistics null distribution of Procedure 2. That is, test statistic cut-offs and adjusted p-values are computed as in Procedure 3 ($B = 5,000$ bootstrap samples).

4.3.2 Results

Interestingly, bootstrap-based single-step maxT Procedure 3 yields 8,916 miRNA pairs (or nearly 75 % of all $M = 11,935$ pairs) with adjusted p-values less than a nominal FWER level $\alpha = 0.05$ and 7,479 with p-values approximately equal to zero (Figure 5, Panel (b); Supplementary Table 3, website companion). Correlation coefficients found to be significantly different from zero at nominal level $\alpha = 0.05$ range from 0.26 to 0.99, with median value 0.55. The most significant are given in Supplementary Table 3. Several of the identified pairs of miRNAs are composed of miRNAs in the same family, e.g., `hsa-miR-10a` and `hsa-miR-10b`. Only 8 % of all pairwise correlation coefficients are negative and none are significantly so.

The two most significantly correlated miRNAs are a pair of paralogs, `miR-17-5p` (chromosome 17) and `miR-106a` (chromosome X), which belong to miRNA clusters believed to be up-regulated by the proto-oncogene c-MYC (O'Donnell *et al.*, 2005). `miR-19a`, `miR-19b`, and `miR-20` are also members of these paralogous miRNA clusters. Several other co-expressed miRNAs are linked to cancer. In particular, `miR-107` has been shown to increase cell growth in lung carcinomas (Cheng *et al.*, 2005). `miR-143` and `miR-145`, located within 1.7kb on human chromosome 5, are expressed at lower levels in cancerous and pre-cancerous tissue compared to normal colon tissue (Michael *et al.*, 2003).

The fact that a majority of pairwise correlation coefficients are significantly different from zero, even after adjusting for multiple tests (nominal FWER level 0.05), suggests a great deal of structure in miRNA expression. In order to focus on highly correlated pairs of miRNAs, the correlation tests could be repeated with a null value larger than zero, e.g., with null hypotheses $H_0(j,j') = \mathrm{I}(\rho(j,j') \leqslant 0.5)$, $j = 1,\ldots,J-1$, $j' = j+1,\ldots,J$.

4.3.3 Cluster analysis

The above multiple testing analysis clearly suggests the existence of clusters of miRNAs with highly correlated expression measures. We therefore decided to perform hierarchical clustering of the miRNAs, in order to identify general expression patterns and groups of co-expressed miRNAs. We use the *hierarchical ordered partitioning and collapsing hybrid* (HOPACH) algorithm, with Pearson correlation distance (van der Laan and Pollard, 2003). HOPACH is implemented in the Bioconductor R package `hopach` (Pollard and van der Laan, 2005). Supplementary Figure 6 (website companion) displays a pseudocolor image of the 155×155 miRNA correlation matrix, with rows and columns ordered according to the final level of the HOPACH tree. Similarly expressed miRNAs appear near each other and are visualized as blocks in the pseudocolor image. It will be of great interest to investigate the biological and medical implications of the identified clusters of co-expressed miRNAs.

5. Conclusions

This investigation of multiple testing procedures has focused on the choice of a test statistics null distribution, in testing problems for which subset pivotality may not hold. Subset pivotality is a condition under which data generating distributions satisfying the complete null hypothesis produce valid test statistics null distributions (Westfall and Young (1993), p. 42–43). Commonly-used procedures, based on permutation or bootstrap data generating null distributions, rely on the subset pivotality condition to justify Type I error control under the true distribution. However, subset pivotality is violated in many important testing problems, since a data generating null distribution may result in a joint distribution for the test statistics that has a different dependence structure than their true distribution. In fact, in most situations, 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.

Indeed, subset pivotality fails for two types of testing problems that are highly relevant in biomedical and genomic data analysis: tests concerning correlation coefficients and tests concerning regression coefficients. Correlation tests abound in molecular biology, where similarities between measurable properties of large numbers of genes and genome sequences are of great interest. Non-linear regression models are also frequently used to assess genotype/phenotype associations, while adjusting for potential confounding variables. Procedures based on a data generating null distribution, such as a permutation distribution, do not provide a correct test statistics null distribution in these settings.

Motivated by limitations of existing approaches, Pollard and van der Laan (2004), and subsequently Dudoit *et al.* (2004b), propose a general characterization and explicit construction of a test statistics null distribution that controls Type I errors, without requirements such as subset pivotality, in testing problems involving general data generating distributions (i.e., arbitrary dependence structures among variables). Resampling procedures, such as the bootstrap procedures of Section 2.4, are provided to conveniently obtain consistent estimators of the null distribution and of the resulting test statistic cut-offs and adjusted p-values. Pollard and van der Laan (2004) compare MTPs based on the proposed bootstrap test statistics null distribution and several other null distributions in the two-sample testing problem. The former null distribution performs competitively, whenever the sample sizes are large enough to avoid ties in the resampled distribution and poorly estimated variances in the denominators of t-statistics.

The goal of the present article was to evaluate the practical performance of different test statistics null distributions in testing problems where subset pivotality fails. Specifically, the simulation studies of Section 3 compare our general non-parametric bootstrap test statistics null distribution (Procedure 2) to parameter-specific bootstrap data generating null distributions, in the following two settings: tests for regression coefficients in linear models where covariates and error terms are allowed to be dependent and tests for correlation coefficients. The general non-parametric bootstrap distribution (Procedure 2, `Bootstrap XY` and `Bootstrap X`) differs from corresponding parameter-specific bootstrap distributions (`Bootstrap e` and `Bootstrap X(j)`) in two key aspects: (i) the (re)sampling units, `Bootstrap XY` and `Bootstrap X` resample "raw" observations, while `Bootstrap e` and `Bootstrap X(j)` resample, respectively, residuals e_i and independent components $X_i(j)$; (ii) the bootstrap test statistics, `Bootstrap XY` and `Bootstrap X` rely on null value shifted and scaled test statistics Z_n, while `Bootstrap e` and `Bootstrap X(j)` rely on "raw" test statistics T_n. In other words, procedures `Bootstrap e` and `Bootstrap X(j)` derive the test statistics null distribution by first creating a data generating distribution that satisfies the complete null hypothesis.

The simulation studies, involving a range of data generating models, demonstrate that the choice of null distribution can have a substantial impact on the Type I error properties of a given multiple testing procedure. The single-step maxT procedure, based on the general non-parametric bootstrap null distribution of Procedure 2, does indeed control the family-wise error rate at or slightly below the target nominal level. Interestingly, comparable MTPs based on parameter-specific bootstrap data generating null distributions, are anti-conservative for tests for regression coefficients (`Bootstrap e`) and conservative for tests for correlation coefficients (`Bootstrap X(j)`). Power is similar for the different null distributions in both testing problems.

Section 4 illustrates the flexibility and power of our proposed methodology, by applying the single-step maxT procedure, with general non-parametric bootstrap test statistics null distribution (Procedures 1, 2, and 3), to a novel genomic dataset from a study of microRNA expression in cancerous and non-cancerous tissues (Lu *et al.*, 2005). Tests for regression coefficients, in a logistic model adjusting for the confounding tissue type variable, identify 90 miRNAs as being significantly *differentially expressed* between cancerous and non-cancerous tissues (nominal FWER level 0.05). This corroborates the original article's discovery that miRNA expression profiling has great potential for cancer diagnosis. Stepwise, augmentation, and empirical Bayes procedures could be used for more powerful analyses and control of a broader class of Type I error rates (Dudoit and van der Laan, 2005; Dudoit *et al.*, 2004a ; van der Laan *et al.*, 2004a,b, 2005).

We also investigated several questions not addressed in the original publication of Lu *et al.* (2005). Firstly, we performed multiple testing to identify pairs of *co-expressed* miRNAs. The fact that a majority of pairwise correlation coefficients are significantly different from zero, even after adjusting for multiple tests (nominal FWER level 0.05), suggests a great deal of structure in miRNA expression. This prompted us to perform hierarchical clustering of the miRNA profiles. The HOPACH algorithm yielded sensible ordering of the miRNAs, with groups of similarly expressed miRNAs visualized as blocks in the pseudo-color image of the correlation matrix (Supplementary Figure 6). Further investigation of the identified clusters of co-expressed miRNAs could reveal biologically and medically interesting connections between miRNAs.

We note that the large number of significant findings in both testing problems is likely due to a reasonably large sample size ($n = 186$) relative to the number of tests ($M = 155$ regression coefficients and $M = 11,935$ correlation coefficients), as compared to similar studies of mRNA expression. This analysis of a rich dataset, using novel and rigorous statistical methods, highlights the possibility for meaningful biological and medical discovery from high-throughput gene expression studies.

Software

Our proposed resampling-based multiple testing procedures are implemented in the R package `multtest`, released as part of the Bioconductor Project, an open-source software project for the analysis of biomedical and genomic data (Pollard *et al.* (2005); `multtest` package, Version 1.7.3, Bioconductor Release 1.7, `www.bioconductor.org`). Birkner *et al.* (2005b) illustrate the implementation in SAS (Version 9) of the bootstrap-based single-step maxT procedure and augmentation procedures for controlling the generalized family-wise error rate (gFWER) and tail probabilities for the proportion of false positives (TPPFP) among the rejected hypotheses.

The hierarchical ordered partitioning and collapsing hybrid (HOPACH) algorithm is implemented in the Bioconductor R package `hopach` (Pollard and van der Laan (2005); `hopach` package, Version 1.2.1, Bioconductor Release 1.7).

Website companion

The website companion to this article provides additional tables, figures, and references:
`www.stat.berkeley.edu/~sandrine/Docs/Papers/SFdS05/SFdS.html`.

Acknowledgment

We wish to thank the Editors for inviting us to contribute to this special issue of the *Journal de la Société Française de Statistique* and for their constructive comments on an earlier version of this manuscript.

References

BIRKNER M.D., COURTINE M., DUDOIT S., VAN DER LAAN M.J.,CLÉMENT K., ZUCKER J.D. (2005a). Statistical methods for detecting genotype/phenotype associations in the ObeLinks Project, Technical report, Division of Biostatistics, University of California, Berkeley, (In preparation).

BIRKNER M.D., POLLARD K.S., VAN DER LAAN M.J., DUDOIT S. (2005b). Multiple testing procedures and applications to genomics, Technical Report 168, Division of Biostatistics, University of California, Berkeley, URL `www.bepress.com/ucbbiostat/paper168`.

BIRKNER M.D., SINISI S.E., VAN DER LAAN M.J. (2005c). Multiple testing and data adaptive regression: An application to HIV-1 sequence data, *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 8, URL `www.bepress.com/sagmb/vol4/iss1/art8`.

CALIN G.A., SEVIGNANI C., DUMITRU C.D., HYSLOP T., NOCH E., YENDAMURI S., SHIMIZU M., RATTAN S., BULLRICH F., NEGRINI M., CROCE C.M. (2004). Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers, *Proc. Natl. Acad. Sci.*, 101(9), p. 2999–3004.

CHENG A.M., BYROM M.W., SHELTON J., FORD L.P. (2005). Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis, *Nucleic Acids Research*, 33(4), p. 1290–1297.

DUDOIT S., VAN DER LAAN M.J. (2005). *Multiple Testing Procedures and Applications to Genomics*, Springer, (In preparation).

DUDOIT S., SHAFFER J.P., BOLDRICK J.C. (2003). Multiple hypothesis testing in microarray experiments, *Statistical Science*, 18(1), p. 71–103.

DUDOIT S., VAN DER LAAN M.J., BIRKNER M.D. (2004a). Multiple testing procedures for controlling tail probability error rates, Technical Report 166, Division of Biostatistics, University of California, Berkeley, URL `www.bepress.com/ucbbiostat/paper166`.

DUDOIT S., VAN DER LAAN M.J., POLLARD K.S. (2004b). Multiple testing. Part I. Single-step procedures for control of general Type I error rates, *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 13, URL `www.bepress.com/sagmb/vol3/iss1/art13`.

HOCHBERG Y., TAMHANE A.C. (1987). *Multiple Comparison Procedures*, Wiley–Interscience.

KELEŞ S., VAN DER LAAN M.J., DUDOIT S., CAWLEY S.E. (2004). Multiple testing methods for ChIP-Chip high density oligonucleotide array data, Technical

Report 147, Division of Biostatistics, University of California, Berkeley, URL www.bepress.com/ucbbiostat/paper147.

KORN E.L., TROENDLE J.F., MCSHANE L.M., SIMON R. (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data, *Journal of Statistical Planning and Inference*, 124(2), p. 379–398.

LU J., GETZ G., MISKA E.A., ALVAREZ-SAAVEDRA E., LAMB J., PECK D., SWEET-CORDERO A., EBERT B.L., MAK R.H., FERRANDO A.A., DOWNING J.R., JACKS T., HORVITZ H.R., GOLUB T.R. (2005). MicroRNA expression profiles classify human cancers, *Nature*, 435(9), p. 834–838, URL www.broad.mit.edu/cancer/pub/miGCM.

MICHAEL M.Z., O'CONNOR S.M., VAN HOLST PELLEKAAN N.G.,YOUNG G.P., JAMES R.J. (2003). Reduced accumulation of specific microRNAs in colorectal neoplasia, *Molecular Cancer Research*, 1(12), p. 882–891.

O'DONNELL K.A., WENTZEL E.A., ZELLER K.I., DANG C.V., MENDELL J.T. (2005). c-Myc-regulated microRNAs modulate E2F1 expression, *Nature*, 435(7043), p. 839–843.

POLLARD K.S.,VAN DER LAAN M.J. (2004). Choice of a null distribution in resampling-based multiple testing, *Journal of Statistical Planning and Inference*, 125(1–2), p. 85–100.

POLLARD K.S., VAN DER LAAN M.J. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter 13: Cluster Analysis of Genomic Data, p. 209–228, Statistics for Biology and Health, Springer-Verlag, New York, URL www.bepress.com/ucbbiostat/paper167.

POLLARD K.S., DUDOIT S., VAN DER LAAN M.J. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter 15: Multiple Testing Procedures: The multtest Package and Applications to Genomics, p. 249–271, Statistics for Biology and Health, Springer-Verlag, New York, URL www.bepress.com/ucbbiostat/paper164.

TROENDLE J.F. (1995). A stepwise resampling method of multiple hypothesis testing, *Journal of the American Statistical Association*, 90(429), p. 370–378.

TROENDLE J.F. (1996). A permutational step-up method of testing multiple outcomes, *Biometrics*, 52(3), p. 846–859.

VAN DER LAAN M.J., POLLARD K.S. (2003). A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap, *Journal of Statistical Planning and Inference*, 117(2), p. 275–303.

VAN DER LAAN M.J., DUDOIT S., POLLARD K.S. (2004a). Multiple testing. Part II. Step-down procedures for control of the family-wise error rate, *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 14, URL www.bepress.com/sagmb/vol3/iss1/art14.

VAN DER LAAN M.J., DUDOIT S., POLLARD K.S. (2004b). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives, *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 15, URL www.bepress.com/sagmb/vol3/iss1/art15.

VAN DER LAAN M.J., BIRKNER M.D., HUBBARD A.E. (2005). Empirical Bayes and resampling based multiple testing procedure controlling tail probability of the proportion of false positives, *Statistical Application in Genetics and Molecular Biology*, 41(1): Article 29, URL www.bepress.com/sagmb/vol4/iss1/art29.

WESTFALL P.H., YOUNG S.S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*, John Wiley & Sons.

WIENHOLDS E., PLASTERK R.H.A. (2005). MicroRNA function in animal development, *Federation of European Biochemical Societies Letters*, 579(26), p. 5911–5922.

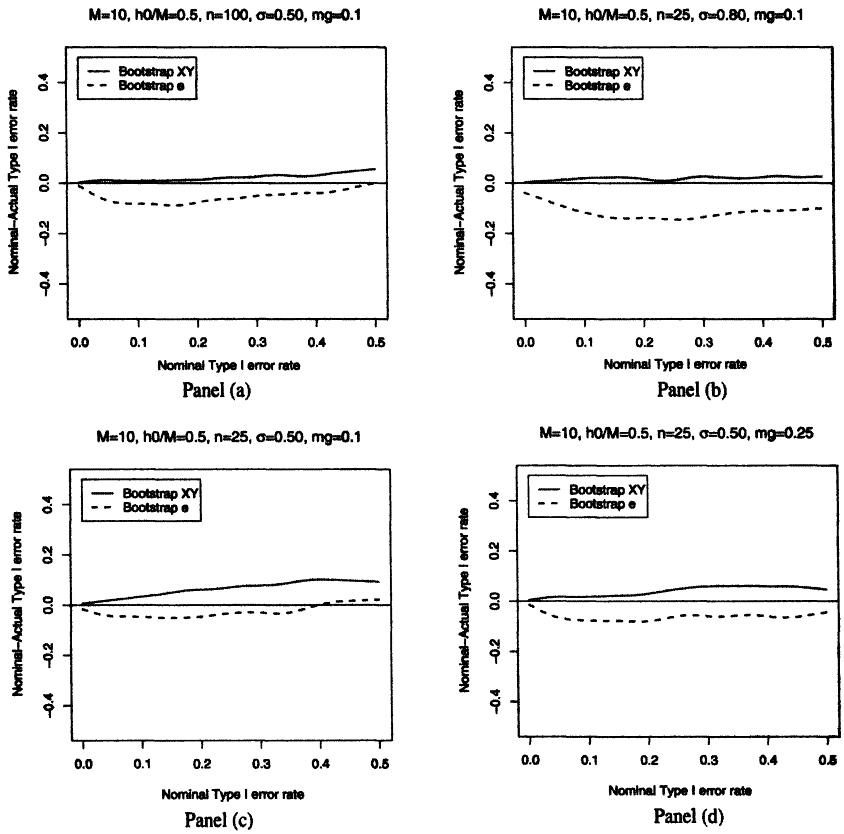

FIG 1. — *Simulation Study 1: Tests for linear regression coefficients, Type I error control comparison.* Plots of differences between nominal and actual Type I error rates vs. nominal Type I error rate, for single-step maxT procedure based on general non-parametric bootstrap null distribution `Bootstrap XY` and residual bootstrap null distribution `Bootstrap e`. The null hypotheses are tested using the t-statistics of Equation (18). Panel (a): Model with sample size $n = 25$; $M = 10$ null hypotheses; common covariance for the covariates $\varsigma = 0.50$; proportion $h_0/M = 0.50$ of true null hypotheses; shift parameter for alternative regression coefficients $\mu = 0.10$. Panel (b): $n = 25$; $M = 10$; $\varsigma = 0.80$; $h_0/M = 0.50$; $\mu = 0.10$. Panel (c): $n = 100$; $M = 10$; $\varsigma = 0.50$; $h_0/M = 0.50$; $\mu = 0.10$. Panel (d): $n = 25$; $M = 10$; $\varsigma = 0.50$; $h_0/M = 0.50$; $\mu = 0.25$.

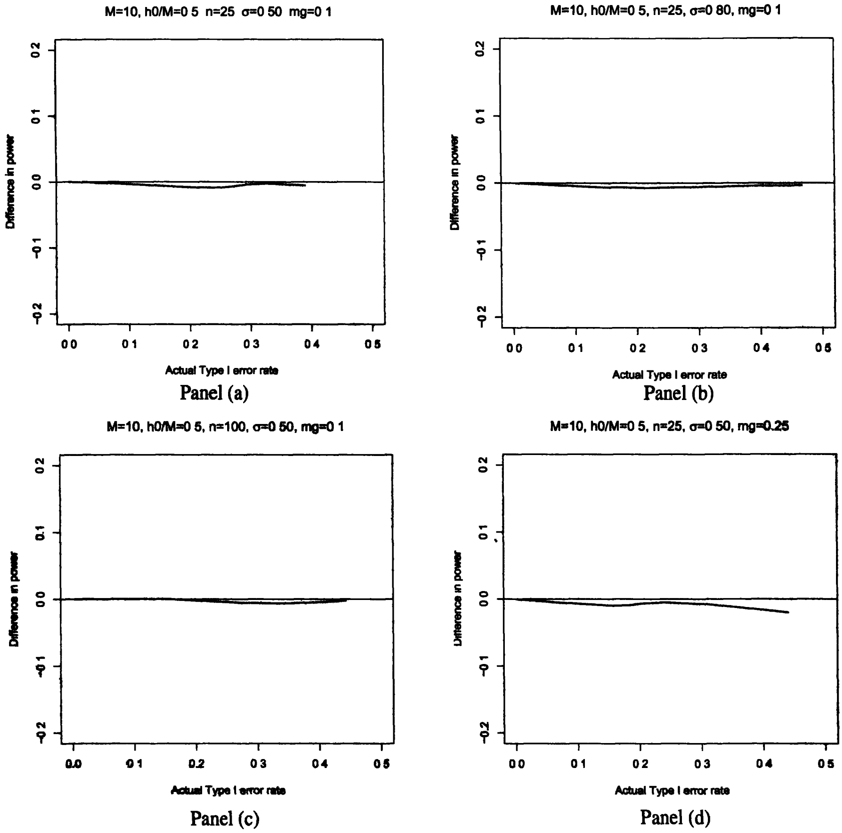

FIG 2. — *Simulation Study 1: Tests for linear regression coefficients, power comparison.* Plots of difference in power vs. actual Type I error rate, for single-step maxT procedure based on general non-parametric bootstrap null distribution **Bootstrap XY** and residual bootstrap null distribution **Bootstrap e**. The null hypotheses are tested using the t-statistics of Equation (18). Positive differences indicate greater power for **Bootstrap XY**. Panel (a): Model with sample size $n = 25$; $M = 10$ null hypotheses; common covariance for the covariates $\varsigma = 0.50$; proportion $h_0/M = 0.50$ of true null hypotheses; shift parameter for alternative regression coefficients $\mu = 0.10$. Panel (b): $n = 25$; $M = 10$; $\varsigma = 0.80$; $h_0/M = 0.50$; $\mu = 0.10$. Panel (c): $n = 100$; $M = 10$; $\varsigma = 0.50$; $h_0/M = 0.50$; $\mu = 0.10$. Panel (d): $n = 25$; $M = 10$; $\varsigma = 0.50$; $h_0/M = 0.50$; $\mu = 0.25$.

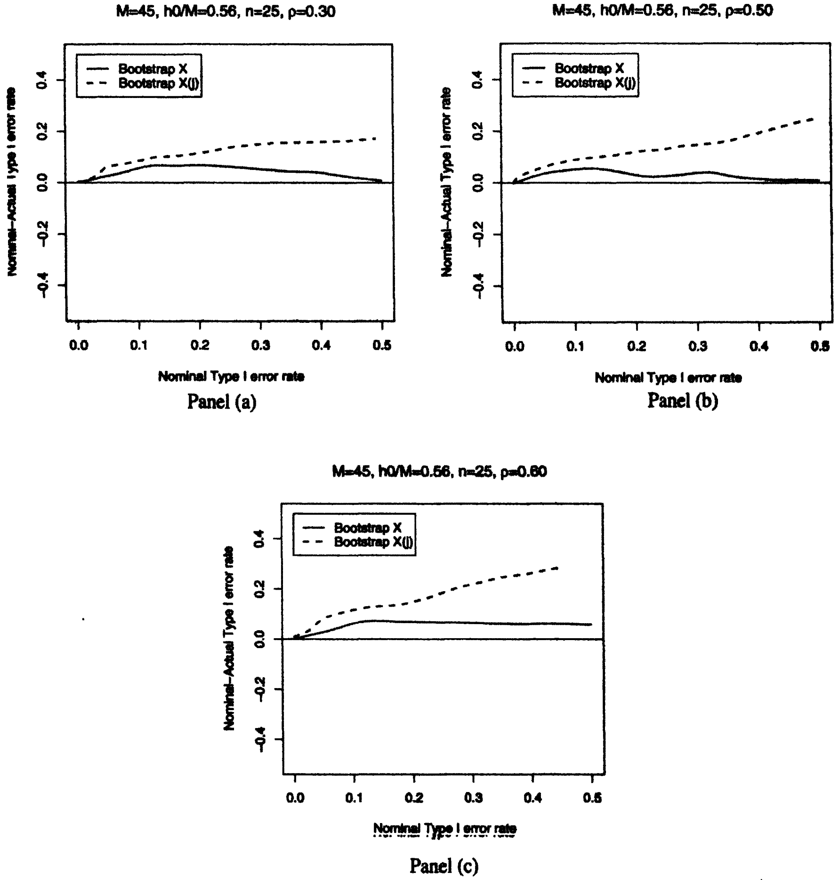

FIG 3. — *Simulation Study 2: Tests for correlation coefficients, Type I error control comparison.* Plots of differences between nominal and actual Type I error rates vs. nominal Type I error rate, for single-step maxT procedure based on general non-parametric bootstrap null distribution `Bootstrap X` and independent covariates bootstrap null distribution `Bootstrap X(j)`. The null hypotheses are tested using the t-statistics of Equation (24). Model with sample size $n = 25$; $M = 45$ null hypotheses; proportion $h_0/M = 25/45$ of true null hypotheses. Panel (a): common correlation coefficient for the two blocks $\varrho = 0.30$. Panel (b): $\varrho = 0.50$. Panel (c): $\varrho = 0.60$.

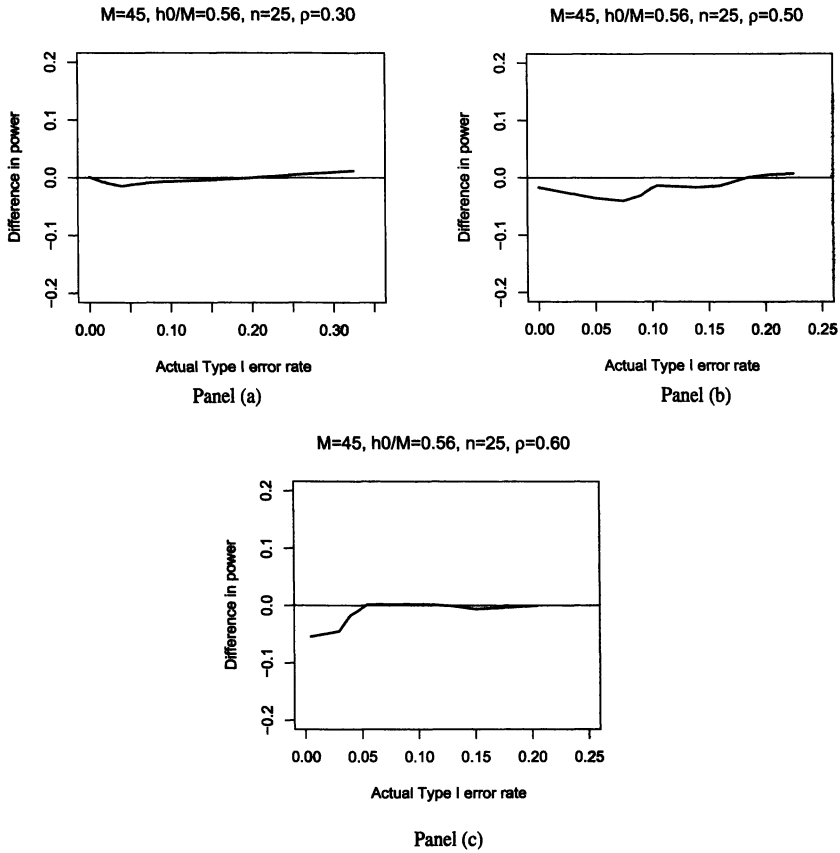

FIG 4. — *Simulation Study 2: Tests for correlation coefficients, power comparison.* Plots of difference in power vs. actual Type I error rate, for single-step maxT procedure based on general non-parametric bootstrap null distribution `Bootstrap X` and independent covariates bootstrap null distribution `Bootstrap X(j)`. The null hypotheses are tested using the t-statistics of Equation (24). Positive differences indicate greater power for `Bootstrap X`. Model with sample size $n = 25$; $M = 45$ null hypotheses; proportion $h_0/M = 25/45$ of true null hypotheses. Panel (a): common correlation coefficient for the two blocks $\varrho = 0.30$. Panel (b): $\varrho = 0.50$. Panel (c): $\varrho = 0.60$.

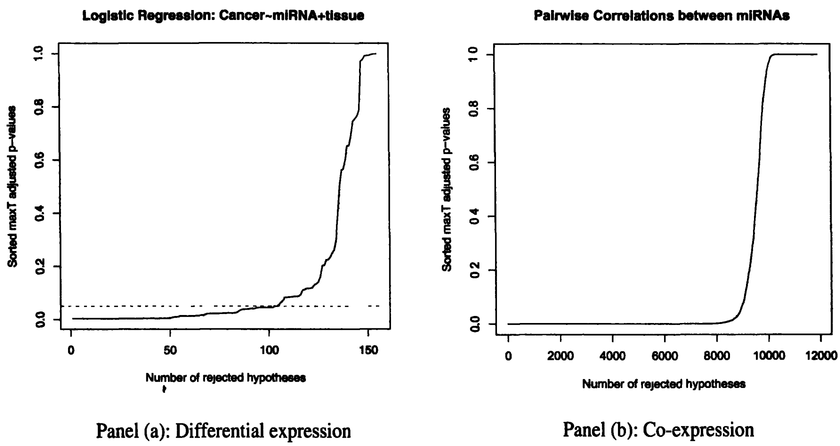

FIG 5. — *miRNA data analysis: Adjusted p-values.* Plots of sorted adjusted *p*-values for bootstrap-based single-step maxT Procedure 3. Panel (a): Identification of differentially expressed miRNAs, based on tests for logistic regression coefficients. Panel (b): Identification of pairs of co-expressed miRNAs, based on tests for correlation coefficients.