Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures

BY NICOLAI MEINSHAUSEN AND PETER BÜHLMANN
Seminar für Statistik, ETH Zürich, Leonhardstrasse 27,
8092 Zürich, Switzerland
meinshausen@stat.math.ethz.ch
buhlmann@stat.math.ethz.ch

SUMMARY

We propose probabilistic lower bounds for the number of false null hypotheses when testing multiple hypotheses of association simultaneously. The bounds are valid under general and unknown dependence structures between the test statistics. The power of the proposed estimator to detect the full proportion of false null hypotheses is discussed and compared to other estimators. The proposed estimator is shown to deliver a tight probabilistic lower bound for the number of false null hypotheses in a multiple testing situation even under strong dependence between test statistics.

Some key words: Family-wise error rate; Multiple testing; Number of false null hypotheses.

1 INTRODUCTION

When testing multiple hypotheses simultaneously, it is often of interest to select a subset of hypotheses which show a significant deviation from the null hypothesis. Adjusting for the multiplicity of the testing problem is commonly achieved by calculating a suitable error rate like the family-wise error rate, see for example Westfall & Young (1993) and Holm (1979), or the false discovery rate, as introduced by Benjamini & Hochberg (1995). Instead of selecting a subset of significant hypotheses, however, one might sometimes rather be interested in just testing a global null hypothesis; see Donoho & Jin (2004) for a recent development in this field and possible areas of application.

Here we consider an intermediate approach. The goal is to estimate the total number m_1 of false null hypotheses among all m tested hypotheses. For a chosen level α, we propose probabilistic lower bounds $\hat{m}_1$, for the total number m_1 of false null hypotheses, for which it holds under arbitrary and unknown dependence between the test statistics, that

$$\mathrm{pr}(\hat{m}_1 \leq m_1) \geq 1 - \alpha. \tag{1.1}$$

The estimator $\hat{m}_1$ can be used as a global test of significance, as the global null hypothesis $m_1 = 0$ can be rejected at level α if $\hat{m}_1 > 0$. On the other hand, estimates of m_1 are useful for tighter estimation of error rates. Storey (2002) showed for example that less conservative estimates of the false discovery rate are possible if an estimate of m_1 is available. Likewise, with an estimate of m_1 to hand, more powerful procedures are possible if the multiplicity adjustment is carried out using the per-comparison or the per-family error rate; see for example Shaffer (1995) and Dudoit et al. (2003) for an overview of the most common multiple hypotheses testing procedures.

In the context of gene expression microarray experiments, it is often of interest to test for differential expression; that is, to test the null hypothesis for each gene that its expression level follows the same distribution under various clinical classes (Golub et al., 1999; Alon et al., 1999). As well as being of interest in its own right, a lower bound on the number m_1 of differentially expressed genes is helpful for tighter estimation of common error rates.

A second application is provided by the Taiwanese-American occultation survey, one goal of which is to estimate the number of objects in the Kuiper Belt (Liang et al., 2002). This number is inferred from the rate of occultations of stars by Kuiper belt objects, which results in a very high-dimensional multiple testing problem. In this case, one is exclusively interested in estimating the number m_1 of false null hypotheses and not in identifying precisely which hypotheses show a significant deviation from the null hypothesis. As a third example, consider the detection and quantification of climate change. Frei & Schär (2001) examined the existence of a trend in the occurrence of extreme precipitation events in the alpine region. Precipitation events are recorded at a large number of stations. No recording station might show a significant effect when taking the multiplicity of the testing problem into account. With the proposed estimators it is nevertheless possible to give a probabilistic lower bound for the number of stations where an increase in extreme precipitation events is indeed occurring.

Allowing arbitrary dependence requires a special structure of the data. However, for multiple testing of associations the requirements are in general fulfilled. The gene-expression example and the detection of trends in extreme precipitation events are amenable to the analysis presented in this paper. In contrast, the astronomical example does not allow for permutation-based testing, which is central to our approach. Incidentally, the gene-expression and extreme-precipitation examples are also those applications in which the issue of dependence among test statistics is particularly pressing. Expression levels are sometimes heavily correlated among genes, and the occurrence of extreme precipitation events is likewise very much correlated among recording stations, especially if they are located in the same geographical region.

Starting with Schweder & Spjøtvoll (1982), estimators have been developed for m_1 that are conservative in the sense that

$$E(\hat{m}_1) \leq m_1. \tag{1.2}$$

The number of true null hypotheses is estimated in Schweder & Spjøtvoll (1982) by a linear fit of the empirical distribution of p-values; see also the

recent application to neuroimaging data in Turkheimer et al. (2001). Another idea in the paper of Schweder & Spjøtvoll (1982) that also appears in Storey (2002) is to estimate the number of true null hypotheses by the number of p-values greater than some threshold λ and then divide by $1 - \lambda$. Suggestions for an adaptive choice of λ are proposed in Storey (2002). For independent test statistics, an estimator with property (1.1) was proposed in Genovese & Wasserman (2004). The estimator proposed in this paper is to our knowledge the first to provide a lower bound for m_1 under general dependence structures between the test statistics.

2 METHODS

2.1 Setting and notation

Let $y \in \mathcal{Y}$ be a class variable with $\mathcal{Y} = \{1, \ldots, h\}$ for some $h \in \mathbb{N}$ or, more generally, a variable with $\mathcal{Y} = \mathbb{R}$. Let $(X_y)_{y \in \mathcal{Y}}$ be a family of m-dimensional random variables with components $X_y = \{X_{y,1}, \ldots, X_{y,m}\}$. In multiple testing of associations, one is interested in whether or not the distribution of the components of X_y are independent of $y \in \mathcal{Y}$.
Assume that there is some set $\mathcal{S} \subseteq \{1, \ldots, m\}$ such that the joint distribution of $\{X_{y,k}, k \in \mathcal{S}\}$ is identical for all values of the variable $y \in \mathcal{Y}$:

$$\{X_{y,k}; k \in \mathcal{S}\} = \{X_{y',k}; k \in \mathcal{S}\} \qquad \text{for all } y, y' \in \mathcal{Y}. \tag{2.3}$$

Let $\mathcal{N}$ be a subset of $\{1, \ldots, m\}$ such that (2.3) is fulfilled and such that there is no subset that fulfils (2.3) and has larger cardinality. The cardinality of $\mathcal{N}$ is denoted by m_0. The quantity m_0 can be interpreted as the number of true null hypotheses in the sense that it describes the number of components of X_y whose distribution is not dependent on the class variable $y \in \mathcal{Y}$. Note that the definition of the set $\mathcal{N}$ of true null hypotheses depends on the joint distribution of all components in this set. In particular, consider the case in which the marginal distributions of two components $X_{y,l}$ and $X_{y,k}$ are both independent of y, but their joint distribution is not. Then k and l are not both members of any set $\mathcal{S}$ that fulfils (2.3) and hence do not both count towards the number of true null hypotheses. The

number of false null hypotheses is defined as $m_1 = m - m_0$. Note that the setting is also applicable to cases where y is random.

2.2 *A simple example*

We begin with a simple example to clarify ideas and notation. The setting is similar to that of linear discriminant analysis. Let the class variable be binary with $\mathcal{Y} = \{0, 1\}$. Let $X_{y=0}$ and $X_{y=1}$ both follow Gaussian distributions with common but unknown covariance matrix Σ:

$$\begin{aligned} X_{y=0} &\sim \mathcal{N}_m(0, \Sigma) \\ X_{y=1} &\sim \mathcal{N}_m(\theta, \Sigma). \end{aligned}$$

The vector θ of means has components $\theta = (\theta_1, \ldots, \theta_m)$. The null hypothesis for each component $k = 1, \ldots, m$ is that the distribution is identical under either $y = 0$ or $y = 1$, which is equivalent to $\theta_k = 0$. The set $\mathcal{N}$ of true null hypotheses is thus given by $\mathcal{N} = \{k : \theta_k = 0\}$. In the context of gene expression microarray data, the class variable y might distinguish between cancerous and non-cancerous tissue, and the question arises of whether or not the expression levels for genes show a systematic upward or downward shift between these conditions.

2.3 *Confidence Interval*

It is assumed that an n-dimensional vector $(y_1, \ldots, y_n) \in \mathcal{Y}^n$ of class variables is available, along with corresponding observations of $X_{y_1}, \ldots, X_{y_n}$, which are assumed to be independent. We suppose that a suitable test is provided for independence of the marginal distributions of $X_{y,k}$, $k = 1, \ldots, m$, from the class variable y. The outcome of such a test, applied to every component $k = 1, \ldots, m$, is a set of p-values $P_1, \ldots, P_m$, where $P_k \sim \text{Un}[0, 1]$ if $k \in \mathcal{N}$. For example, for a two-sample problem with $y \in \{0, 1\}$, as in §2.2, a t-test or a Wilcoxon test is appropriate for testing for a shift in location between the two groups. In general, the test will be adapted to the problem at hand. The number of hypotheses with p-values in a given rejection region $[0, \gamma]$ is denoted by $R(\gamma)$:

$$R(\gamma) = \sum_{k \in \{1, \ldots, m\}} 1\{P_k \leq \gamma\}.$$

The number of false rejections, denoted by $V(\gamma)$, is the number of p-values P_k below γ, where k is a member of the set $\mathcal{N}$:

$$V(\gamma) = \sum_{k \in \mathcal{N}} 1\{P_k \leq \gamma\}.$$

We first introduce the key concept of a bounding function. Unless stated otherwise let Γ be the interval $[0, 1]$. A bounding function at level α is a random function $G_\alpha(\gamma)$ which is monotonically increasing with γ for every realisation such that

$$\text{pr}[\sup_{\gamma \in \Gamma}\{V(\gamma) - G_\alpha(\gamma)\} > 0] < \alpha. \tag{2.4}$$

We will show explicitly in §2.5 how a bounding function can be constructed. The proposed estimator of m_1 is given as the maximal difference between the realised number of rejections $R(\gamma)$ and a bounding function $G_\alpha(\gamma)$ at level α:

$$\hat{m}_1 = \sup_{\gamma \in \Gamma}\{R(\gamma) - G_\alpha(\gamma)\}. \tag{2.5}$$

The estimator of m_0 is simply $\hat{m}_0 = m - \hat{m}_1$. As mentioned above, $\Gamma = [0, 1]$ unless stated explicitly. Note that both $R(\gamma)$ and $G_\alpha(\gamma)$ are monotonically increasing with γ. Furthermore, the number $R(\gamma)$ of p-values less than or equal to γ is constant except for a set of at most m points of discontinuity, at which the supremum in (2.5) is attained. The supremum can hence be efficiently evaluated by maximising over the finite random set of realised p-values. We show that the estimator of m_0 indeed provides an probabilistic upper bound for the number of true null hypotheses.

Theorem 1. A one-sided (1-α) confidence interval for m_0 is given by $[0, \hat{m}_0]$. A one-sided (1-α) confidence interval for m_1 is given by $[\hat{m}_1, m]$. In particular,

$$\text{pr}(\hat{m}_1 \leq m_1) \geq 1 - \alpha.$$

A proof is given in the Appendix. Note that Theorem 1 allows for arbitrary dependence among the components of the m-dimensional X_y; we only require independence of the n observations $X_{y_1}, \ldots, X_{y_n}$, i.e. for the data sample.

The properties of the estimator are solely determined by the choice of the bounding function. In particular, the power to detect true non-null hypotheses is markedly different for different choices of the bounding functions. We are going to discuss in the sequel a general method for obtaining tight bounding functions.

2.4 Sufficient criterion for a bounding function

It is not possible to verify criterion (2.4). Criterion (2.4) requires knowledge of the distribution of V and hence of m_0, which is the very quantity one is trying to estimate. We shall show that the distribution of V can in some sense be bounded from above by the computable distribution of a random variable V^π, obtained by permutations of the class variables $(y_1, \ldots, y_n)$. Let Z be the sample with ordered values $(y_{(1)}, \ldots, y_{(n)})$ of the class variables $(y_1, \ldots, y_n)$:

$$Z = \{(y_{(i)}, X_{y_i})\}_{i=1,\ldots,n}.$$

Let π be a random permutation of $\{1, \ldots, n\}$ and define the action of a π on Z by the permutation of the class labels according to π, $\pi(Z) = \{(y_{\pi(i)}, X_{y_i})\}_{i=1,\ldots,n}$. Define the random variable P_k^π, $k = 1, \ldots, m$, as the p-value of the kth hypothesis under randomly permuted class labels, where each of the $n!$ permutations of the set $\{1, \ldots, n\}$ has equal probability:

$$P_k^\pi(Z) = P_k\{\pi(Z)\}.$$

The random variable $V^\pi(\gamma)$ is now defined as the number of components, k, for which P_k^π is smaller than γ:

$$V^\pi(\gamma) = \sum_{k \in \{1,\ldots,m\}} 1\{P_k^\pi \leq \gamma\}.$$

The distribution of V^π is determined by the unknown distribution of the test statistics. However, the distribution of V^π conditional on Z is computable if we use all $n!$ permutations of the class variables $(y_1, \ldots, y_n)$. The distribution of V^π thus yields, in a sense made precise below, a useful upper bound for the distribution of V.

Theorem 2. A random, $\sigma(Z)$-measurable and monotonically increasing function $G_\alpha(\gamma)$ is a bounding function according to (2.4) if, for any $Z = z$,

$$\mathrm{pr}[\sup_{\gamma\in\Gamma}\{V^\pi(\gamma) - G_\alpha(\gamma)\} > 0|Z = z] < \alpha. \tag{2.6}$$

Proof. Given the definition (2.4) of a bounding function, it is sufficient for a proof of Theorem 2 to show that a $\sigma(Z)$-measurable function G_α which fulfils (2.6) also satisfies, for any $Z = z$,

$$\mathrm{pr}[\sup_{\gamma\in\Gamma}\{V(\gamma) - G_\alpha(\gamma)\} > 0|Z = z] < \alpha.$$

The random variable $V(\gamma)$ is given by $V(\gamma) = \sum_{k\in\mathcal{N}} 1\{P_k \leq \gamma\}$. By definition (2.3) of the set of null hypotheses $\mathcal{N}$, the joint distribution of $\{P_k, k \in \mathcal{N}\}$, conditional on $Z = z$, is identical to the distribution of $\{P_k^\pi, k \in \mathcal{N}\}$, conditional on $Z = z$. Thus

$$\begin{aligned}&\mathrm{pr}[\sup_{\gamma\in\Gamma}\{V(\gamma) - G_\alpha(\gamma)\} > 0|Z = z] = \\ &\mathrm{pr}[\sup_{\gamma\in\Gamma}\{\sum_{k\in\mathcal{N}} 1\{P_k^\pi \leq \gamma\} - G_\alpha(\gamma)\} > 0|Z = z].\end{aligned}$$

The theorem follows since $\sum_{k\in\mathcal{N}} 1\{P_k^\pi \leq \gamma\} \leq \sum_{k\in\{1,\ldots,m\}} 1\{P_k^\pi \leq \gamma\} = V^\pi(\gamma)$, and if we integrate out over Z.

2.5 Quantile bounding functions and computation

We propose to use the quantile function of $V^\pi(\gamma)$ as a bounding function. Let $Q_z^\beta(\gamma)$ be the β-quantile of $V^\pi(\gamma)$, conditional on $Z = z$. This function can be computed by random permutations of the class variables. Let $\beta(\alpha)$ be the minimal value of $\beta \in [0, 1]$ such that (2.6) is fulfilled for $Q_z^\beta(\gamma)$. The quantile function $Q_z^{\beta(\alpha)}(\gamma)$ is then a valid bounding function. Note that any function G_α which fulfils (2.6) is bounded from below by the $(1 - \alpha)$-quantile of $V^\pi(\gamma)$; that is $G_\alpha(\gamma) \geq Q_z^{1-\alpha}(\gamma)$ for any bounding function G_α. It follows that $1 - \alpha \leq \beta(\alpha) \leq 1$.

Let Π be a set of random permutations of the class variable. For the finite set Π, the computation of the quantile functions can be limited to the set of quantiles $\beta \in \{1, 1 - 1/|\Pi|, 1 - 2/|\Pi|, \ldots, 1/|\Pi|\}$. For $\beta = 1$, criterion

(2.6) is surely fulfilled. The value $\beta(\alpha)$ is found by checking iteratively, starting with $\beta = 1$ and then for successively lower values of β, whether or not criterion (2.6) is fulfilled for the quantile function $Q_z^{\beta}(\gamma)$. Note that, if the criterion is not fulfilled for some β, then it cannot be fulfilled for any value lower than β. The value $\beta(\alpha)$ is the lowest value for which criterion (2.6) is fulfilled.

To check whether or not criterion (2.6) is fulfilled for the quantile function $Q_z^{\beta}(\gamma)$, calculate for every $\pi \in \Pi$ the p-values P_k^{π}, $k = 1, \ldots, m$, of all hypotheses. Check, for every permutation $\pi \in \Pi$, whether or not $V^{\pi}(\gamma) \leq Q_z^{\beta}(\gamma)$ for all $\gamma \in \{P_1^{\pi}, \ldots, P_m^{\pi}\}$. If this condition is fulfilled, set $c(\pi) = 0$. Otherwise, set $c(\pi) = 1$. Criterion (2.6) is fulfilled if and only if $\sum_{\pi} c(\pi) < \alpha|\Pi|$.

By (2.5), the estimator of m_1 is then given by

$$\hat{m}_1 = \sup_{\gamma \in \Gamma} \{R(\gamma) - Q_z^{\beta(\alpha)}(\gamma)\}.$$

As a result of the monotonicity of $Q_z^{\beta(\alpha)}(\gamma)$, the supremum is attained by some value of γ in the finite, random set of realised p-values $\{P_1, \ldots, P_m\}$. Evaluation of the supremum is hence achieved by maximising over a finite set of points. It holds by positivity of the bounding function that $0 \leq \hat{m}_1 \leq m$.

It might seem that the computational burden of this procedure is prohibitive if a permutation-based test is used for computation of the p-values, as the algorithm as laid out here involves in these cases a double permutation. It is therefore of interest to note that the algorithm also works when we use, instead of p-values, raw test statistics.

2.6 Connection to the family-wise error rate

Another possible choice of a bounding function is given by

$$G_{\alpha}(\gamma) = \begin{cases} 0 & \gamma \leq g(\alpha) \\ \infty & \gamma > g(\alpha) \end{cases},$$

where $g(\alpha)$ is the largest value in $[0, 1]$ such that (2.6) is fulfilled. By Bonferroni's inequality, $g(\alpha) \geq 1/m$. The estimate (2.5) for this bounding

function is given by

$$\hat{m}_1^{\mathrm{fw}} = R\{g(\alpha)\},$$

and is equal to the number of rejections when controlling the family-wise error rate at level α.

2.7 *Asymptotic power*

Here we compare the asymptotic powers of $\hat{m}_1^{\mathrm{fw}}$ and $\hat{m}_1$ to detect the correct proportion of false null hypotheses. The ability of the estimators to identify a large proportion of all false null hypotheses depends of course on the power of the individual tests. We settle here for the simple setting of a two-sample problem, where a one- or two-sided Wilcoxon test is used to test whether or not the distribution of a random variable $X_{y=0}$ is shifted compared to the distribution of another random variable $X_{y=1}$. The total number n of observations is given by $n = n_0 + n_1$, where n_0 is the number of independent observations of $X_{y=0}$ and n_1 is the number of independent observations of $X_{y=1}$.

We are particularly interested in how well the estimators can cope with a large number m of tests. Thus for the following analysis m is increasing with n, so that $m = m(n) \to \infty$ for $n \to \infty$. Both $X_{y=0}$ and $X_{y=1}$ are assumed to be infinite-dimensional. For n observations, the first $m(n)$ components are tested for association with the class variable. We need three reasonable assumptions.

Assumption 1. There exists some $c > 0$ such that, for all false null hypotheses $k \in \mathcal{N}^c$,

$$|\mathrm{pr}(X_{y=0,k} < X_{y=1,k}) - 1/2| > c.$$

Assumption 2. The dependence between test statistics is such that, for some $\tau \in (0, 1)$,

$$\sup_{\gamma \in \Gamma} \sum_{k,l=1}^{m} \left|\mathrm{cov}\big(1\{P_k \le \gamma\}, 1\{P_l \le \gamma\}\big)\right| = o(m^{1+\tau}) \qquad \text{for } m = m(n) \to \infty.$$

Assumption 3. The proportion of false null hypotheses converges to $\kappa \in (0, 1)$, while the proportion of observations from class $y = 1$ converges to

some $\nu \in (0,1)$:

$$\begin{aligned} m_1(n)/m(n) &\to \kappa \qquad \text{for } n \to \infty \\ n_1/n &\to \nu \qquad \text{for } n \to \infty. \end{aligned}$$

Assumption 1 could be relaxed by replacing c with a sequence c_n such that $c_n \to 0$ sufficiently slowly as $n \to \infty$. However, it suffices in its current form to illustrate the difference in power between the estimators. Assumption 2 is a weak condition regarding the strength of correlation between test statistics. For example it is fulfilled if test statistics are block-dependent and the size of the largest block is increasing at most as $o(m^\tau)$. For independent test statistics, the assumption is fulfilled for any $\tau > 0$. The second part of Assumption 3 seems reasonable. An interesting field for further research would be to study the behaviour of the estimators for $\kappa = 0$, where the proportion of false null hypotheses is vanishing for $n \to \infty$; see Meinshausen & Rice (2005) for the case of independent test statistics.

Theorem 3. Let Assumptions 1-3 be fulfilled and let $n^{-1} \log m(n) \to \infty$ *for* $n \to \infty$*. Then, for* $n \to \infty$*, in probability,*

$$\begin{aligned} \hat{m}_1^{\text{fw}}/m_1 &\to 0 \\ \hat{m}_1/m_1 &\to 1. \end{aligned}$$

From an asymptotic point of view, estimation of m_1 with $\hat{m}_1$ is thus more powerful than estimation with $\hat{m}_1^{\text{fw}}$. Note that the number of hypotheses increases very quickly in the result above as a function of the number of observations.

In general, the power of $\hat{m}_1^{\text{fw}}$ to detect the presence of false null hypotheses deteriorates with the number of tested hypotheses. The estimator $\hat{m}_1^{\text{fw}}$ is equal to the number of rejections that can be made under control of the family-wise error rate, as already mentioned above, and it is well known that the family-wise error rate is very conservative if the number of tested hypotheses is large. The result in Theorem 3 is thus perhaps not very surprising. However, Theorem 3 shows that, for the purpose of estimating m_1, more powerful estimators are available which do not suffer from vanishing power for an increasing number of tested hypotheses.

2.8 Composite null hypotheses

The method was primarily developed to test for identical distribution of the components of X_y for all $y \in \mathcal{Y}$. In practice, one might like to allow for more general composite null hypotheses, and here we show how the proposed method can be generalised. Suppose that the family X_y, $y \in \mathcal{Y}$, of random variables is parameterised by a vector $\theta = (\theta_1, \ldots, \theta_m) \in \Theta^m$. Consider first the case of point null hypotheses $\theta_k = 0$. The set of true null hypotheses therefore corresponds to the set $\mathcal{N} = \{k : \theta_k = 0\}$ and the number of true null hypotheses is given by $m_0 = \sum_{k=1}^{m} 1\{\theta_k = 0\}$.
Now suppose that the null hypothesis is given rather by $\theta_k \in \Theta_0$ for every component $k = 1, \ldots, m$, and some $\Theta_0 \subset \Theta$. In this case the number of true null hypotheses is given by $m_0 = \sum_{k=1}^{m} 1\{\theta_k \in \Theta_0\}$. The proposed method can be applied without further modifications to this problem under the perhaps crucial assumption that one can couple together the values $\theta_k \in \Theta_0$ and $\theta_k = 0$ in the following sense. Let $P_k(\theta_k)$ be the p-value of the kth hypothesis under parameter value θ_k. Suppose now that the parametrisation is so chosen that almost surely the p-values under any $\theta_k \in \Theta_0$ are at least as large as under $\theta_k = 0$:

$$\theta_k \in \Theta_0 \quad \Rightarrow \quad P_k(\theta_k) \geq P_k(0) \tag{2.7}$$

almost surely. Then the proposed estimators $\hat{m}_1$ have the desired property that $\mathrm{pr}(\hat{m}_1 \leq m_1) \geq 1 - \alpha$, where m_1 is now defined as $m_1 = \sum_{k=1}^{m} 1\{\theta_k \in \Theta_0\}$. This follows by an inspection of the proof of Theorem 2. Such a coupling can be achieved for a large number of potentially interesting composite null hypotheses. As an example, consider again the setting of §2.2. Let the null hypotheses be given not by $\theta_k = 0$ but instead by $\theta_k \in \Theta_0 = (-\infty, 0]$, so that m_1 measures only the number of hypotheses in which the shift in mean for class $y = 1$ compared to class $y = 0$ is positive. If we use a sensible test like the t-test or the Wilcoxon test, it is obvious that (2.7) is fulfilled in this case.

2.9 Estimation of error rates

There is by now a multitude of error rates for multiple hypothesis testing; see Shaffer (1995) or Dudoit et al. (2003) for an overview. The most

important ones are the family-wise error rate, the per-comparison error rate, which is defined as $E(V)/m$, the expected number of Type I errors V divided by the total number m of hypotheses. Furthermore there is the per-family error rate, $E(V)$. Finally there is the false discovery rate, which is defined as $E(Q)$, where Q is the proportion of falsely rejected hypotheses, that is $Q = V/R$ if $R > 0$ and $Q = 0$ if $R = 0$. Storey (2002) was the first to make use of an estimator of m_0 to give a less conservative estimator of the false discovery rate. Our proposed estimators of m_0 can also be used to give less conservative estimators of the per-comparison and per-family error rates. The value of the per-comparison and per-family error rate are given for a fixed rejection region $[0, \gamma]$ by

$$\begin{aligned} \text{PCER} &= m_0\gamma/m, \\ \text{PFER} &= m_0\gamma. \end{aligned}$$

The value of m_0 is unknown but bounded by m. The error rates can thus be trivially bounded from above by $\text{PCER} \leq \gamma$ and $\text{PFER} \leq m\gamma$. These bounds are rather conservative if there are many false null hypotheses. If we use for example the proposed estimator $\hat{m}_0$ of m_0, less conservative estimators are obtained. For the per-comparison error rate, the proposed estimator of the per-comparison error rate is

$$\hat{\text{PCER}} = \hat{m}_0\gamma/m.$$

This estimator is always smaller than the conservative upper bound: $\hat{\text{PCER}} \leq \gamma$. We are still on the safe side, however, as the estimator is, by Theorem 1, larger than the true value of the per-comparison error rate with high probability:

$$\text{pr}(\hat{\text{PCER}} \geq \text{PCER}) \geq 1-\alpha.$$

A similar result holds for the per-family error rate. In Storey (2002), it was shown that a useful estimator for the false discovery rate, when rejecting all hypotheses with p-value less than γ, is given by $m_0\gamma/R(\gamma)$. Let $\hat{m}_0$ be some estimator of m_0. A plug-in estimator for the false discovery rate is then $\hat{\text{FDR}} = \hat{m}_0\gamma/R$. In particular, the estimator of m_0 in Storey (2002) is

$$\hat{m}_0^{\text{st}} = \frac{m - R(\lambda)}{1-\lambda}. \tag{2.8}$$

This estimator has the property that $E(\hat{m}_0^{\lambda}) \geq m_0$ and $E(\hat{\text{FDR}}) \geq \text{FDR}$. Instead of using $\hat{m}_0^{\text{st}}$ as an estimator of m_0, it is possible to use different estimators, such as our $\hat{m}_0$. We compare both estimators in the sequel.

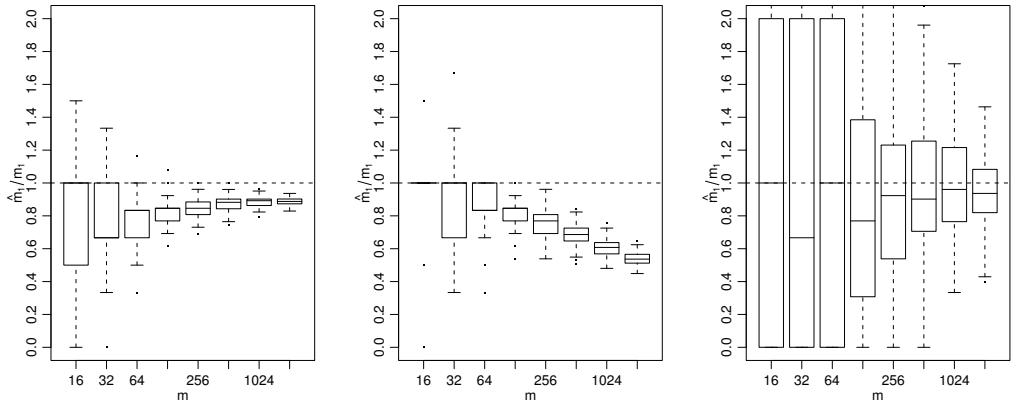
3 Numerical Examples

3.1 Simulated data

The set-up for the numerical comparison is the same as in the example of §2.2. The set $\mathcal{N}$ of true null hypotheses is generated by randomly drawing m_0 elements from the set $\{1, \ldots, m\}$. For $k \in \mathcal{N}$, $\theta_k = 0$, whereas for false null hypotheses with $k \in \mathcal{N}^c$, $\theta_k = 1$. The Wilcoxon test is used to test for a shift in mean between the distributions of $X_{y=0}$ and $X_{y=1}$ for all m components. The total number n of observations is assumed to be even and there are $n/2$ independent observations of $X_{y=0}$ and $n/2$ observations of $X_{y=1}$. The covariance matrix Σ is defined by $\Sigma = aK^{-1}$, where a is a scale factor, chosen so that the diagonal of Σ has unit entries and K is an $m \times m$ matrix with unit entries in the diagonal and $K_{ij} = \zeta/2$ if $|i - j| = 1$ or $\{i, j\} = \{1, m\}$, and $K_{ij} = 0$ otherwise. Independent test statistics are obtained if $\zeta = 0$. If $\zeta = 0.995$, this gives a covariance matrix with non-diagonal entries in the range of 0 to 0.9. About 90% of all correlations are below 0.01.

For $n = 60$ observations, the empirical distribution of $\hat{m}_1/m_1$, at level $\alpha = 0 \cdot 05$, is shown in Fig. 1 for 100 simulations and independent test statistics under an increasing number m of hypotheses. The number of false null hypotheses m_1 is kept at a constant proportion 0·1 of all hypotheses. It can be observed in Fig. 1(b) that the power of a method that controls the family-wise error rate, corresponding to $\hat{m}_1^{\text{fw}}$, vanishes for large m as expected from Theorem 3. The proposed estimator $\hat{m}_1$ shows qualitatively different behaviour. The power actually increases for increasing m, converging to a positive value close to 1. In Fig. 1(c), the smoother estimator of m_1, proposed in Storey & Tibshirani (2003) and denoted by $\hat{m}_1^{\text{st,sm}}$, is shown for comparison. The bias of this estimator is smaller, but the variance is substantially larger than for any of the proposed estimators.

Next, a more thorough simulation study is done for $m = 1000$ hypothe-



(a) $\hat{m}_1$ (b) $\hat{m}_1^{\mathrm{fw}}$ (c) $\hat{m}_1^{\mathrm{st,sm}}$

Figure 1: Box-plots for the ratio $\hat{m}_1/m_1$ as a function of the number m of tested hypotheses for independent test statistics, for (a) the proposed estimate $\hat{m}_1$, (b) the number of rejections $\hat{m}_1^{\mathrm{fw}}$ when controlling the family-wise error rate, and (c) the smoother estimate $\hat{m}_1^{\mathrm{st,sm}}$.

ses. The number of false null hypotheses is varied with $m_1 \in \{0, 100, 500\}$. The estimators $\hat{m}_1$ and $\hat{m}_1^{\mathrm{fw}}$ are compared in Table 1. Additionally, the estimator $\hat{m}_1^{\mathrm{st}}$ is shown, as proposed in Storey (2002); see equation (2.8). The parameter λ has to be chosen heuristically and the commonly-made choice $\lambda = 0{\cdot}5$ is used. A bootstrap method for obtaining an optimal choice of λ was proposed in Storey (2002). The resulting estimator is denoted by $\hat{m}_1^{\mathrm{st,b}}$. Finally, the smoother estimator $\hat{m}_1^{\mathrm{st,sm}}$ proposed in Storey & Tibshirani (2003) is shown. If there is no single false null hypothesis, $m_1 = 0$, the estimators $\hat{m}_1^{\mathrm{fw}}$ and $\hat{m}_1$ estimate m_1 correctly by 0 in at least $100(1-\alpha)$ percent of the simulations, as expected from property (1.1). In contrast, in this case the estimators $\hat{m}_1^{\mathrm{st}}$, $\hat{m}_1^{\mathrm{st,sm}}$ and $\hat{m}_1^{\mathrm{st,b}}$ produce large estimators of m_1, especially for dependent test statistics. Note that these last three estimators are thresholded at 0 and m respectively, and the conservative property that $E(\hat{m}_1) < m_1$ is thereby lost. Hence the average value of $\hat{m}_1^{\mathrm{st}}$ is often larger than m_1 in the simulations shown here.

The power of $\hat{m}_1^{\mathrm{fw}}$ to detect a sizeable proportion of all false null hypotheses are in general poor, as already expected from theoretical considerations above. Furthermore, the estimator $\hat{m}_1^{\mathrm{st,b}}$, with a bootstrap choice of

Table 1: *Simulation study. The average value (mean), standard deviation (sd), root mean square error (rsme) and the probability* $\mathrm{pr}(\hat{m}_1 > m_1)$ *of overestimation (pr) for different estimators of* m_1*, the number of false null hypotheses. Except for pr, values are rounded to the nearest integer.*

	$\zeta = 0$				$\zeta = 0.995$			
				$m_1 = 0$				
	mean	sd	rmse	pr	mean	sd	rmse	pr
$\hat{m}_1$	0	0	0	0·02	0	0	0	0·03
$\hat{m}_1^{\mathrm{fw}}$	0	0	0	0·00	0	0	0	0·00
$\hat{m}_1^{\mathrm{st}}$	10	17	20	0·37	96	136	166	0·48
$\hat{m}_1^{\mathrm{st,sm}}$	19	29	35	0·45	211	259	333	0·55
$\hat{m}_1^{\mathrm{st,b}}$	61	72	94	0·88	278	300	408	0·64
				$m_1 = 100$				
$\hat{m}_1$	86	4	14	0·00	72	10	30	0·00
$\hat{m}_1^{\mathrm{fw}}$	45	5	55	0·00	46	13	56	0·00
$\hat{m}_1^{\mathrm{st}}$	99	31	30	0·48	152	169	176	0·48
$\hat{m}_1^{\mathrm{st,sm}}$	91	60	60	0·42	250	291	326	0·52
$\hat{m}_1^{\mathrm{st,b}}$	163	75	98	0·89	357	292	388	0·67
				$m_1 = 500$				
$\hat{m}_1$	435	14	66	0·00	428	22	75	0·00
$\hat{m}_1^{\mathrm{fw}}$	224	11	276	0·00	229	51	276	0·00
$\hat{m}_1^{\mathrm{st}}$	495	22	23	0·44	510	109	109	0·54
$\hat{m}_1^{\mathrm{st,sm}}$	486	48	50	0·38	529	232	233	0·59
$\hat{m}_1^{\mathrm{st,b}}$	543	55	70	0·86	651	154	215	0·73

λ, seems unsuitable for dependent test statistics. The smoother estimator $\hat{m}_1^{\mathrm{st,sm}}$ likewise has a large bias and variance for dependent test statistics. The original estimator $\hat{m}_1^{\mathrm{st}}$ with fixed λ seems to be the most useful among $\hat{m}_1^{\mathrm{st}}$, $\hat{m}_1^{\mathrm{st,b}}$ and $\hat{m}_1^{\mathrm{st,sm}}$, at least for the data examined here.

This leaves $\hat{m}_1^{\mathrm{st}}$, with a appropriate predetermined choice of λ, and $\hat{m}_1$ as sensible estimators of m_1. In terms of root mean squared error, $\hat{m}_1^{\mathrm{st}}$ is best for independent test statistics and larger proportions m_1/m of false null hypotheses. For dependent test statistics, either $\hat{m}_1^{\mathrm{st}}$ or $\hat{m}_1$ has the lowest root mean squared error. The probability of overestimating m_1 is conservatively controlled with $\hat{m}_1$ at level α, as expected from Theorem 1. With $\hat{m}_1^{\mathrm{st}}$, the probability of overestimating m_1 is usually around $0{\cdot}5$. The high variance of $\hat{m}_1^{\mathrm{st}}$ under dependent test statistics suggests that there is a rather high probability of overestimating m_1 by a large amount.

3.2 Microarray data

With microarray studies it is possible to monitor the expression values of several thousand genes simultaneously. A common aim with microarray studies is to find differentially expressed genes, that is genes whose expression values show systematic variation among different groups. Given a class variable y like tumour type or clinical outcome, it can be tested for each gene k if the expression values $X_{y,k}$ are associated with y. We look at three microarray studies, in all of which the response variable is binary $y \in \mathcal{Y} = \{0, 1\}$. In the study on breast cancer from van't Veer et al. (2002), y corresponds to the clinical outcome; in the leukaemia study in Golub et al. (1999), the class variable y distinguishes between two different subtypes of leukaemia; and finally, in a colon cancer study in Alon et al. (1999), y indicates absence or presence of colon cancer. The number of genes involved is $m = 5408$ for the breast cancer study, $m = 3571$ for the leukaemia study and $m = 2000$ for the colon cancer study.

In Table 2, estimators of m_1 with the property that $\mathrm{pr}(\hat{m}_1 > m_1) < \alpha$ are compared. For the estimator $\hat{m}_1$, the approach laid out in §2.5 is used. The estimator $\hat{m}_1^{\mathrm{fw}}$ is equivalent to the number of rejections when controlling the family-wise error rate. We use the step-down method of Westfall & Young (1993) to control the family-wise error rate. Also shown is the

Table 2: *Estimators $\hat{m}_1$ of the number m_1 of differentially expressed genes, with* $\mathrm{pr}(\hat{m} > m_1) < \alpha$*, for three gene expression microarray datasets.*

	$\alpha = 0{\cdot}05$			$\alpha = 0{\cdot}01$		
	colon	leukaemia	breast	colon	leukaemia	breast
$\hat{m}_1^{\mathrm{fw}}$, Bonferroni	55	266	2	32	191	0
$\hat{m}_1^{\mathrm{fw}}$, Step-down	64	281	3	36	202	0
$\hat{m}_1$	286	957	355	245	811	126

number of rejections for control of the family-wise error rate, based on the Bonferroni correction.

With the estimator $\hat{m}_1$, a consistently higher proportion of false null hypotheses are detected than with control of the family-wise error rate. The gain of using the proposed estimator compared to control of the family-wise error rate depends on the number of tested hypotheses. Indeed, the least dramatic gain, which still represents roughly a factor of four, is for the colon cancer and leukaemia data with the lowest number of tested hypotheses. The gain is most pronounced for the breast-cancer data, where not a single rejection can be made when controlling the family-wise error rate at level $\alpha = 0{\cdot}01$, while the estimator $\hat{m}_1$ at the same level indicates that there are more than 100 true null hypotheses.

ACKNOWLEDGEMENT

The authors would like to thank J. Rice and A. Buja for helpful discussions and comments. The comments of the editor and two anonymous referees also helped to improve the quality of this manuscript.

APPENDIX

Proofs

Proof of Theorem 1. It suffices to show that $\mathrm{pr}(\hat{m}_1 > m_1) < \alpha$, where $\hat{m}_1 = \sup_{\gamma \in \Gamma}\{R(\gamma) - G_\alpha(\gamma)\}$. The number of rejections can be split into $R(\gamma) =$

$S(\gamma) + V(\gamma)$, where $S(\gamma)$ is the number of correct rejections. Let $\mathcal{N}^c$ be the complement of $\mathcal{N}$ in $\{1, \ldots, m\}$. Then $S(\gamma) = \sum_{k \in \mathcal{N}^c} 1\{P_k \leq \gamma\}$. Note that $\sup_{\gamma \in \Gamma}\{S(\gamma)\} = S(1) = m_1$. Thus

$$\begin{aligned}
\text{pr}(\hat{m}_1 > m_1) &= \text{pr}[\sup_{\gamma \in \Gamma}\{R(\gamma) - G_\alpha(\gamma)\} > m_1] \\
&= \text{pr}[\sup_{\gamma \in \Gamma}\{V(\gamma) + S(\gamma) - G_\alpha(\gamma)\} > m_1] \\
&\leq \text{pr}[\sup_{\gamma \in \Gamma}\{V(\gamma) - G_\alpha(\gamma)\} + S(1) > m_1] \\
&\leq \text{pr}[\sup_{\gamma \in \Gamma}\{V(\gamma) - G_\alpha(\gamma)\} > 0].
\end{aligned}$$

The function $G_\alpha(\gamma)$ is a bounding function at level α. The quantity

$$\text{pr}[\sup_{\gamma \in \Gamma}\{V(\gamma) - G_\alpha(\gamma)\} > 0]$$

is thus strictly smaller than α by definition of G_α, and the claim follows.

Lemma A1. Let $Q_z^\beta(\gamma)$ be the β-quantile of $V^\pi(\gamma)$, conditional on $Z = z$, under a rank-based test. Let Γ be the corresponding discrete set of p-values. It holds for any $\nu > 0$ and $z \in \mathcal{Z}$ under Assumption 2 that there exists a sequence $\delta_m \sim m^{-\frac{1}{2}+\frac{\tau}{2}}$ such that $\inf_{\beta \geq \nu} Q_z^\beta(\gamma)/m_0 \geq \gamma - \delta_m$. Furthermore, $Q_z^{1-\beta}(\gamma)/m \leq \gamma/\beta$ for all $\gamma \in (0, 1)$.

Proof. For the first claim, it is sufficient to show that $\text{pr}\{\gamma - V^\pi(\gamma)/m_0 > \delta_m | Z = z\} \to 0$ for $m \to \infty$ and all $\gamma \in \Gamma$. Replace $V^\pi(\gamma) = \sum_{k=1}^m 1\{P_k^\pi \leq \gamma\}$ by the smaller random variable $\sum_{k \in \mathcal{N}} 1\{P_k^\pi \leq \gamma\}$, where the sum stretches only over components k in the set $\mathcal{N}$ of true null hypotheses. As a rank-based test is used, it holds that the distribution of $\{P_k^\pi; k \in \mathcal{N}\}$, conditional on Z, is identical to the distribution of $\{P_k; k \in \mathcal{N}\}$. Hence it is sufficient to show that $\text{pr}(\gamma - \frac{1}{m_0}\sum_{k \in \mathcal{N}} 1\{P_k \leq \gamma\} > \delta_m) \to 0$ for $m \to \infty$. Note that $E(\frac{1}{m_0}\sum_{k \in \mathcal{N}} 1\{P_k \leq \gamma\}) = \gamma$. It follows by Assumption 2 and $\kappa < 1$ that $\text{var}(\frac{1}{m_0}\sum_{k \in \mathcal{N}} 1\{P_k \leq \gamma\}) = o(m^{-1+\tau})$. The first part of the claim follows thus by Chebychev's inequality.
For the second part it is sufficient to show that, for every $\gamma \in \Gamma$, $\text{pr}\{V^\pi(\gamma)/m > \gamma/\beta | Z = z\} < \beta$, where $V^\pi(\gamma) = \sum_{k=1}^m 1\{P_k^\pi \leq \gamma\}$. Let Π be the set of

all possible permutations of $\{1,\ldots,n\}$. Then the above is equivalent to showing that

$$\frac{1}{n!}\sum_{\pi\in\Pi} 1\{\sum_{k=1}^{m} 1\{P_k^\pi \leq \gamma\} > m\gamma/\beta\} < \beta. \tag{3.9}$$

Assume to the contrary that (3.9) is not fulfilled. This implies that, for at least $\beta n!$ of all permutations, $\sum_{k=1}^{m} 1\{P_k^\pi \leq \gamma\} > m\gamma/\beta$ and hence $\frac{1}{n!}\sum_{\pi\in\Pi}\sum_{k=1}^{m} 1\{P_k^\pi \leq \gamma\} > m\gamma$. However, as a rank-based test is used, it has to hold that $\text{pr}\{P_k^\pi \leq \gamma | Z = z\} = \frac{1}{n!}\sum_{\pi\in\Pi} 1\{P_k^\pi \leq \gamma\} \leq \gamma$, which leads to a contradiction. Hence (3.9) is fulfilled and the claim follows.

Proof of Theorem 3. The estimator is given by $\hat{m}_1^{\text{fw}} = R\{g(\alpha)\}$. According to (2.6), the value of $g(\alpha)$ is the minimal value of g such that, for a given $Z = z$, $\text{pr}\{V^\pi(1-g) > 0 | Z = z\} < \alpha$, which is equivalent to $Q_z^{1-\alpha}(1-g)\} < 0$. By Lemma A1, there exists some sequence $\delta_m \sim m^{-\frac{1}{2}+\frac{\tau}{2}}$ so that $Q_z^{1-\alpha}(\gamma)/m_0 \geq \gamma - \delta_m$. It follows that $m_0\{1-g(\alpha)-\delta_m\} = 0$. Let $\gamma_{\min}$ be the minimal p-value under a Wilcoxon test, $\gamma_{\min} = n_0!n_1!/n!$. If $\gamma_{\min} > 1-g(\alpha)$, it follows that $R\{1-g(\alpha)\} = 0$ and hence $\hat{m}_1^{\text{fw}} = 0$. Hence it suffices to show that $m_0(\gamma_{\min} - \delta_m) \to \infty$ for $n \to \infty$ as then $\gamma_{\min} > 1 - g(\alpha)$ eventually, implying that $R\{1-g(\alpha)\} \to 0$ for $n \to \infty$. By Stirling's formula, it holds that $-\log\gamma_{\min} = cn\{1+o(1)\}$ for some $c > 0$ and $n \to \infty$. On the other hand, for some $d > 0$, $-\log\delta_m = d\log m\{1+o(1)\}$. As $\log m(n)/n \to \infty$ for $n \to \infty$, it follows that $\delta_m/\gamma_{\min} \to 0$ for $n \to \infty$. It thus suffices to show that $m_0\gamma_{\min} \to \infty$, which is, since $\kappa < 1$, equivalent to showing that $m\gamma_{\min} \to \infty$ for $n \to \infty$. This follows again by $-\log\gamma_{\min} = O(n)$ and $\log m(n)/n \to \infty$ for $n \to \infty$.
For the proposed estimator $\hat{m}_1 = \max_{\gamma\in\Gamma}\{R(\gamma) - Q_z^{\beta(\alpha)}(\gamma)\}$, it is first shown that $\text{pr}(\hat{m}_1/m_1 > 1+\epsilon) \to 0$ for any $\epsilon > 0$. It clearly holds that $\beta(\alpha) \geq 1-\alpha$. By Assumption 2 and Lemma A1, there exists some sequence $\delta_m \sim m^{-\frac{1}{2}+\frac{\tau}{2}}$ such that, for all $\gamma \in \Gamma$, $Q_z^{\beta(\alpha)}(\gamma)/m_0 \geq \gamma - \delta_m$. Since $R(\gamma) \leq m_1 + \sum_{k\in\mathcal{N}} 1\{P_k \leq \gamma\}$, it holds that

$$\text{pr}(\hat{m}_1/m_1 > 1+\epsilon) \leq \text{pr}\{\sup_{\gamma\in\Gamma}(\sum_{k\in\mathcal{N}} 1\{P_k \leq \gamma\} - m_0(\gamma-\delta_m) > \epsilon m_1\}.$$

As $m_0\delta_m = o(m_1)$, the term $m_0\delta_m$ can without loss of generality be neglected. Note that $|\Gamma| \leq n^2$ for the Wilcoxon test. By Bonferroni's in-

equality, it thus remains to be shown that $\mathrm{pr}(\sum_{k\in\mathcal{N}} 1\{P_k \leq \gamma\} - m_0\gamma > \epsilon m_1) = o(n^{-2})$ for all $\gamma \in \Gamma$ and $n \to \infty$. It holds that $E(\sum_{k\in\mathcal{N}} 1\{P_k \leq \gamma\}) = m_0\gamma$. Furthermore, by Assumption 2, $\mathrm{var}(\sum_{k\in\mathcal{N}} 1\{P_k \leq \gamma\}) = o(m^{1+\tau})$. By Chebychev's inequality and since $\kappa \in (0,1)$, it follows that $\mathrm{pr}(\sum_{k\in\mathcal{N}} 1\{P_k \leq \gamma\} - m_0\gamma > \epsilon m_1) = O(m^{\tau-1})$. As $\log m(n)/n \to \infty$ for $n \to \infty$, it follows that $\mathrm{pr}(\sum_{k\in\mathcal{N}} 1\{P_k \leq \gamma\} - m_0\gamma > \epsilon m_1) = o(n^{-2})$, which proves the claim.

It remains to be shown that $\mathrm{pr}(\hat{m}_1/m_1 < 1-\epsilon) \to 0$ for any $\epsilon > 0$ and $n \to \infty$. By Lemma A1, $Q_z^{1-\beta}(\gamma)/m \leq \gamma/\beta$ for all $\gamma \in (0,1)$. As $\beta(\alpha) \leq \alpha/|\Gamma|$ and $|\Gamma| \leq n^2$, it follows that $Q_z^{\beta(\alpha)}(\gamma) \leq m\gamma n^2/\alpha$ for all $\gamma \in (0,1)$. Let

$$\gamma_n = \max\{\gamma \in \Gamma : \gamma \leq n^{-2}/\log n\}.$$

Then, from the above results and since $\kappa > 0$, $Q_z^{\beta(\alpha)}(\gamma_n)/m_1 = o(1)$ for $n \to \infty$. Since

$$\hat{m}_1 = \sup_{\gamma\in\Gamma}\{R(\gamma) - Q_z^{\beta(\alpha)}(\gamma)\} \geq R(\gamma_n) - Q_z^{\beta(\alpha)}(\gamma_n)$$

and $Q_z^{\beta(\alpha)}(\gamma_n)/m_1 = o(1)$ for $n \to \infty$, it remains to be shown that, for any $\epsilon > 0$, $\mathrm{pr}\{R(\gamma_n)/m_1 < 1-\epsilon\} \to 0$ for $n \to \infty$. By Assumption 2, $\mathrm{var}\{R(\gamma)/m_1\} = o(1)$. By Chebychev's inequality it hence suffices to show that, for any $\epsilon > 0$, $E\{R(\gamma_n)/m_1\} > 1-\epsilon$ for $m = m(n)$ large enough. The number of rejections $R(\gamma_n) = \sum_{k=1}^{m} 1\{P_k \leq \gamma_n\}$ is bounded from below by $\sum_{k\in\mathcal{N}^c} 1\{P_k \leq \gamma_n\}$ and it thus suffices to show under Assumption 1 that, for any false null hypothesis $k \in \mathcal{N}^c$, $\mathrm{pr}(P_k \leq n^{-2}/\log n) \to 1$ for $n \to \infty$. This follows from Lemma A2 below, which completes the proof.

Lemma A2. Let $X_{y=0}$ and $X_{y=1}$ be two independent random variables fulfilling Assumption 1. The number of independent observations of each variable is given by n_0 and n_1 respectively, and $n = n_0 + n_1$. Let P be the p-value of a false null hypothesis under a one- or two-sided Wilcoxon test. Under Assumption 3, it holds for any $\delta > 0$ that $\mathrm{pr}(P < n^{-\delta}) \to 1$ for $n \to \infty$.

Proof. It suffices to show the result for a one-sided Wilcoxon test, where the null hypothesis is $H_0 : \mathrm{pr}(X_{y=0} < X_{y=1}) = 1/2$ and the alternative

is given by H_A : $\mathrm{pr}(X_{y=0} < X_{y=1}) > 1/2$. By Assumption 1, all false null hypotheses satisfy $\mathrm{pr}(X_{y=0} < X_{y=1}) > 1/2 + c$ for some $c > 0$. Let $R_1, \ldots, R_n$ be the ranks of the combined observations of $X_{y=0}$ and $X_{y=1}$. The test statistic is given by $W = \sum_{i=1}^{n_1} R_i$, where the sum is understood to stretch only over observations where $y = 1$. Under the null hypothesis, $E(W) = n_1(n+1)/2$. Let $w_c = (1+c)n_1(n+1)/2$. Under Assumption 3, $n_1/n \to \nu \in (0,1)$ for $n \to \infty$. Hence it follows by Theorem 2.1 in Stone (1967) that, under the null hypothesis H_0, $\mathrm{pr}(W > w_c) = O\{\exp(-cn)\}$ for some constant $c > 0$. Thus, for any value of $\delta > 0$, $\mathrm{pr}(W > w_c) = o(n^{-\delta})$ for $n \to \infty$. It thus remains to be shown that, under the alternative, $\mathrm{pr}(W \leq w_c) \to 0$ for $n \to \infty$. Under the alternative hypothesis, $E(W) \geq (1+2c)n_1(n+1)/2$ and $\mathrm{var}(W) = O(n^3)$. From Chebychev's inequality, it indeed follows that, under the alternative, $\mathrm{pr}(W \leq w_c) \to 0$ for $n \to \infty$, which completes the proof.

REFERENCES

ALON, U., BARKAI, N., NOTTERMAN, D., GISH, K., YBARRA, S., MACK, D. & LEVINE, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Cell Biol.* **96**, 6745–50.

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

DONOHO, D. & JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32**, 962–95.

DUDOIT, S., SHAFFER, J. & BOLDRICK, J. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18**, 71–103.

FREI, C. & SCHÄR, C. (2001). Detection probabilities of trends in rare events: Theory and application to heavy precipitation in the Alpine region. *J. Climate* **14**, 1568–84.

GENOVESE, C. & WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **3**, 1035–61.

GOLUB, T., SLONIM, D., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J., COLLER, H., LOH, M., DOWNING, J., CALIGURI, M., BLOOMFIELD, C. & LANDER, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–7.

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.

LIANG, C.-L., RICE, J., DE PATER, I., ALCOCK, C., AXELROD, T., WANG, A. & MARSHALL, S. (2002). Statistical methods for detecting stellar occultations by Kuiper belt objects: the Taiwanese-American occultation survey. *Statist. Sci.* **19**, 265–74.

MEINSHAUSEN, N. & RICE, J. (2005). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.*

SCHWEDER, T. & SPJØTVOLL, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika* **69**, 493–502.

SHAFFER, J. (1995). Multiple hypothesis testing: A review. *Ann. Rev. Psychol.* **46**, 561–84.

STONE, M. (1967). Extreme tail probabilities for the null distribution of the two-sample Wilcoxon statistic. *Biometrika* **54**, 629–40.

STOREY, J. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc.* B **64**, 479–98.

STOREY, J. & TIBSHIRANI, R. (2003). Statistical significance for genome-wide studies. *Proc. Nat. Acad. Sci.* **100**, 9440–5.

TURKHEIMER, F., SMITH, C. & SCHMIDT, K. (2001). Estimation of the number of true null hypotheses in multivariate analysis of neuroimaging data. *NeuroImage* **13**, 920–30.

VAN'T VEER, L., DAL, H., VAN DER VIJVER, M., HE, Y., HART, A., MAO, M., PETERSE, H., VAN DER KOOY, K., MARTON, M., WITTEVEEN, A., SCHREIBER, G., KERKHOVEN, R., ROBERTS, C., LINSLEY, P., BERNARDS, R. & FRIEND, S. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **406**, 742–7.

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