Upper bounds for the number of true null hypotheses and novel estimates for error rates in multiple testing

Nicolai Meinshausen and Peter Bühlmann,
Seminar für Statistik, ETH Zürich, Switzerland

January 13, 2004

Abstract

When testing multiple hypotheses simultaneously, a quantity of interest is the number m_0 of true null hypotheses.

We present a general framework for finding upper probabilistic bounds for m_0, that is estimates $\widehat{m}_0$ with the property $P[\widehat{m}_0 \geq m_0] \geq 1-\alpha$ for any chosen level α. A conservative, one-sided $(1-\alpha)$ confidence interval for m_0 is then given by $[0, \widehat{m}_0]$. Moreover, $\widehat{m}_0$ can be used for novel estimates of type I errors in multiple testing such as the false discovery rate.

Control of the family-wise error rate emerges as a special case in our framework but suffers from vanishing power for a large number of tested hypotheses. We present a different estimate such that the ability to detect true non-null hypotheses increases with the number of tested hypotheses. A detailed algorithm is provided. The method is valid under general and unknown dependence between the test statistics.

We develop the method primarily for multiple testing of associations between random variables. The method is illustrated with simulation studies and applications to microarray data.

1 Introduction

Assume we have m parameters $\theta_i \in \mathbb{R}$, $i = 1, \ldots, m$ of interest in a multiple testing situation. Let $\Theta_0 \subseteq \mathbb{R}$ and denote by $H_{0,i}$ the null hypothesis for the i-th hypothesis,

$$H_{0,i} \quad : \quad \theta_i \in \Theta_0.$$

Denote by h_i the components of the vector $h \in \{0,1\}^m$, taking the value 0 if the i-th hypothesis is a true null hypothesis and the value 1 otherwise. The total number m_0 of true null hypotheses and the number $m_1 = m - m_0$ of true non-null hypotheses is then given by

$$\begin{aligned} m_0 &= \sum_{i=1}^{m} 1_{[h_i=0]}, \\ m_1 &= \sum_{i=1}^{m} 1_{[h_i=1]}. \end{aligned}$$

The usual goal in a multiple testing situation is to identify the hypotheses that are the most significant on an individual basis and adjust for the multiplicity of the testing problem by calculating a

suitable error rate like the family-wise error rate $FWER$, see e.g. Westfall and Young (1993) and Holm (1979), or the false discovery rate FDR as introduced by Benjamini and Hochberg (1995). Knowledge about m_0 can be useful for tighter estimation of these error rates. Storey (2002) showed e.g. that less conservative estimates of the false discovery rate are possible if an estimate of m_0 is available. Likewise, with an estimate of m_0 at hand, more powerful procedures are possible if the multiplicity adjustment is carried out using the per-comparison or the per-family error rate, see e.g. Shaffer (1995) and Dudoit et al. (2003) for an overview of the most common multiple hypotheses testing procedures.

The number m_0 of true null hypotheses is a quantity of interest in its own right, however.

In applications like microarray studies, the sample size is typically small while the number of tested hypotheses is very large. The power of multiple testing procedures is hence often low and it can happen that not a single hypothesis is significant if the multiplicity of the testing problem is properly taken care of.

In such a case, there are obviously two possible reasons for such a (non-)result. Either the test is not powerful enough and hence fails to reveal true non-null hypotheses or, as a second possible explanation, almost all hypotheses are true null hypotheses.

True non-null hypotheses -if existent- could be identified by a more powerful testing procedure. Increased power is in general achieved by collecting more data but collecting data is a costly process and the effort will be in vain if almost all hypotheses are true null.

The point which we are trying to make is that there is "information" in the data about the number of true non-null hypotheses even if there are no (or just very few) significant test-results. This information can be exploited with our proposed estimate and we are able to provide an upper probabilistic bound for the number of true null hypotheses.

The upper bound will clearly depend on the number of observations and better, smaller bounds will be achieved with more observations. For a small number of observations, the bounds are nevertheless much better than those implied by common multiple testing procedures like control of the family-wise error rate.

An upper bound for the number of true null hypotheses is clearly equivalent to a lower bound for the number of true non-null hypotheses. If this lower bound is substantially above the number of significant hypotheses, we know that with high probability, a low number of significant results in the multiple testing procedure is due to lack of power and not due to absence of true non-null hypotheses. Hence we know that collecting further data will lead to more "discoveries" and might be worth the effort. In the microarray study in the section on numerical results, we find e.g. a lower bound of more than 100 for the number of true non-null hypotheses, while not a single rejection can be made with the family-wise error rate.

Starting with Schweder and Spjøtvoll (1982), estimates have been developed for m_0 that are conservative in the sense that

$$E[\widehat{m}_0] \geq m_0. \tag{1.1}$$

The basic idea behind these estimates is the linearity of the cumulative distribution function for p-values of true null hypotheses (if only point null hypotheses are considered). The number of true null hypotheses is estimated in Schweder and Spjøtvoll (1982) by a linear fit of the empirical distribution of p-values, see as well the recent application to neuroimaging data in Turkheimer et al. (2001). Another idea in the paper of Schweder and Spjøtvoll (1982) that reappears in Storey (2002) is to estimate the number of true null hypotheses by the number of p-values greater than

some threshold λ and divide by $1-\lambda$. Suggestions for an adaptive choice of λ are proposed in Storey (2002). Note, however, that this estimate is only suitable for testing point null hypotheses. Additionally the estimate is not confined by the values of 0 and m respectively. Thresholding alleviates this problem, but the conservative property (1.1) might be lost.
Another recent idea to estimate m_0 has been put forward by Nettleton and Gene Hwang (2003), following a proposal in Mosig et al. (2001). The properties of the resulting estimate are analyzed in Nettleton and Gene Hwang (2003). It is clear, however, that the resulting estimate does not posses a property like (1.1) or (1.2).
We present a general framework for finding estimates of m_0 with the property

$$P[\widehat{m}_0 \geq m_0] \geq 1-\alpha. \tag{1.2}$$

This estimate can be viewed in the probabilistic sense as an upper bound for the number of true null hypotheses or, equivalently, as a lower bound for the number of true non-null hypotheses.
Several estimates are possible in our general framework, depending on the choice of a so-called bounding function. A special choice is proposed and the resulting estimate is shown to have positive power to detect true non-null hypotheses even in the limit of infinitely many tested hypotheses.
The following section 2 introduces the notation and covers the theory and properties of the resulting estimate. In section 3 we present numerical studies both with simulated and microarray data, demonstrating the power of the proposed method and illustrating the properties of a new estimate of the false discovery rate.

2 Theory

The methods are presented in the context of testing associations, but generalizations to different applications are easily possible. Multiple testing of associations arises in microarray data analysis, where a common goal is to identify genes that are differentially expressed with respect to a response variable Y like for example tumour type, see e.g. Golub et al. (1999).
Let $(\Omega, \mathcal{F}, P)$ be a probability space. The data consist in general of n independent copies

$$(\underline{X}_k, Y_k)_{k=1,\ldots,n}$$

of the random variable $(\underline{X}, Y) : \Omega \mapsto \mathcal{X} \times \mathcal{Y}$, where usually $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{Y} = \{0,1\}$ or $\mathcal{Y} = \mathbb{R}$. Each component of $\underline{X}$ is tested for association with the response variable Y.
Given any test for association between the i-th component of $\underline{X}$ and Y, let the p-value of this test be denoted by $P_i, i = 1, \ldots, m$. As our approach is permutation-based, rank-based tests which result in discrete p-values are a natural choice but tests with continuous p-values are applicable as well.
We assume that the case of independence between the i-th component of $\underline{X}$ and Y is included in the i-th null hypothesis, and it holds in this case for all values γ in the set of possible p-values that $P[P_i \leq \gamma] = \gamma$. We do not restrict ourselves to point null hypothesis of independence but require that p-values for other true null hypotheses are stochastically equal or greater than in the case of independence.

The number of hypotheses with p-values in a given rejection region $[0, \gamma]$ is denoted by $R(\gamma)$, the number of false rejections by $V(\gamma)$ and the number of correct rejections by $S(\gamma)$.

$$R(\gamma) = \sum_{i=1}^{m} 1_{[P_i \leq \gamma]}, \tag{2.1}$$

$$V(\gamma) = \sum_{i=1}^{m} 1_{[h_i=0]} 1_{[P_i \leq \gamma]}, \tag{2.2}$$

$$S(\gamma) = \sum_{i=1}^{m} 1_{[h_i=1]} 1_{[P_i \leq \gamma]}. \tag{2.3}$$

The total number $R(\gamma)$ is the sum of false and correct rejections, $R(\gamma) = S(\gamma) + V(\gamma)$ and we have $R(1) = m$, $V(1) = m_0$ and $S(1) = m_1$.

2.1 Confidence Interval

We present an estimate $\widehat{m}_1$ of m_1, which is a lower probabilistic bound for the number m_1 of true non-null hypotheses at any desired level α,

$$P[\widehat{m}_1 \leq m_1] \geq 1 - \alpha.$$

The estimate is applicable to arbitrary and unknown dependence between test statistics or p-values. The estimate is furthermore equivalent to an estimate $\widehat{m}_0 = m - \widehat{m}_1$ of the number m_0 of true null hypotheses with the property

$$P[\widehat{m}_0 \geq m_0] \geq 1 - \alpha.$$

We first introduce the key concept of a bounding function. Unless stated otherwise let Γ be the interval $[0, 1]$.

Definition 2.1 (bounding function) *A bounding function at level α is a random, $\mathcal{F}$-measurable function $G^\alpha(\gamma)$ which is, for every $\omega \in \Omega$, monotonously increasing with γ such that*

$$P\Big[\sup_{\gamma \in \Gamma} \big\{V(\gamma) - G^\alpha(\gamma)\big\} > 0\Big] < \alpha. \tag{2.4}$$

Further below, we will show explicitly how a bounding function can be constructed.
The proposed estimate 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.

Definition 2.2 *Let $G^\alpha(\gamma)$ be a bounding function at level α. The estimates $\widehat{m}_1$ and $\widehat{m}_0 = m - \widehat{m}_1$ are defined by*

$$\widehat{m}_1 = \sup_{\gamma \in \Gamma} \big\{R(\gamma) - G^\alpha(\gamma)\big\}. \tag{2.5}$$

As mentioned above, $\Gamma = [0, 1]$ unless stated explicitly.

Remark 2.1 *Note that both $R(\gamma)$ and $G^\alpha(\gamma)$ are monotonously increasing with γ. $R(\gamma)$ is furthermore constant except for a set of at most m points of discontinuity, at which the supremum in (2.5) is attained. Evaluation of the supremum can hence be restricted to the finite random set of realized p-values.*

We show that the estimate of m_0 indeed provides an upper probabilistic bound for the number of true null hypotheses.

THEOREM 2.1 (CONFIDENCE INTERVAL) *A one-sided (1-α) confidence interval for m_0 is given by $[0, \widehat{m}_0]$. A one-sided (1-α) confidence interval for m_1 is given by $[m_1, \widehat{m}_1]$. In particular,*

$$\begin{aligned} P[\widehat{m}_0 \geq m_0] &\geq 1-\alpha, \\ P[\widehat{m}_1 \leq m_1] &\geq 1-\alpha. \end{aligned}$$

A proof is given in the appendix.
The properties of the estimate are solely determined by a choice of the bounding function. The power to detect true non-null hypotheses in particular is markedly different for different choices of the bounding functions.
We are going to discuss in the following a general method to obtain tight bounding functions.

2.2 Sufficient Criterion for a bounding function

It is not possible to verify directly criterion (2.4) of whether a function is a bounding function or not. Criterion (2.4) requires knowledge of the distribution of V and hence of m_0, which is the very quantity we are trying to estimate. We show in the following that the distribution of V can in some sense be bounded from above by the computable distribution of another random variable V^0.
The computation of bounding functions and estimates $\widehat{m}_0$ and $\widehat{m}_1$ will be discussed in section 2.3.
It is maybe instructive to consider first the case of independent test statistics. Here $V(\gamma)$ is distributed Binomial(m_0,γ). The number m_0 of true null hypotheses is unknown but bounded from above by m. A stochastically larger random variable is hence for example given by V^0, if V^0 is distributed Binomial(m,γ).
We find now a bound for the distribution of V for the case of unknown and arbitrary dependence between the test statistics.
Let $Z \in \mathcal{Z} = (\mathcal{X} \times \mathcal{Y})^n$ be a sample of size n with ordered observations of Y,

$$Z = (\underline{X}_k, Y_{(k)})_{k=1,\ldots,n},$$

where $(Y_{(k)})_{k=1,\ldots,n}$ is the ordered sample of the response variable $(Y_k)_{k=1,\ldots,n}$. Define the action of a random permutation S on Z as the permutation of all Y-values:

$$S(Z) = (\underline{X}_k, Y_{S(k)})_{k=1,\ldots,n}$$

The p-value of the i-th hypothesis under a given sample of size n was denoted by $P_i : \Omega \to [0,1]$. Define the random variable $P_i^0 : \Omega \to [0,1]$, $l = 1, \ldots, m$ as the p-value of the i-th hypothesis under a randomly permuted Y-sample,

$$P_i^0\big(Z\big) = P_i\big(S(Z)\big).$$

The idea is now, that P_i^0 and P_i have the same distribution if the i-th component of $\underline{X}$ is independent of Y.

DEFINITION 2.3 *The random variable* $V^0(\gamma) : \Omega \to \{0, 1, \ldots, m\}$ *is defined as*

$$V^0(\gamma) = \sum_{i=1}^{m} 1_{[P_i^0 \leq \gamma]}.$$

The distribution of V^0 is determined by the unknown dependence between the test statistics. Under the assumption of independence between Y and $\underline{X}$, however, Z is a sufficient statistic. The distribution of V^0, conditional on Z, is in particular given in this case under all n! permutations of the observations of the response variable Y. The distribution of V^0 yields thus (in a sense made precise below) a useful upper bound for the distribution of V.

PROPOSITION 2.1 *A random,* $\sigma(Z)$*-measurable, and monotonously increasing function* $G^\alpha(\gamma)$ *is a bounding function according to (2.4) if*

$$P\Big[\sup_{\gamma\in\Gamma}\big\{V^0(\gamma) - G^\alpha(\gamma)\big\} > 0\Big|Z = z\Big] < \alpha. \tag{2.6}$$

In a given data set, it hence suffices to construct a bounding function G^α for the realized value of Z and evaluate the estimate with this bounding function.

2.3 ALGORITHM

An algorithm for the computation of the estimates $\widehat{m}_0$ and $\widehat{m}_1$ is given below for the case of unknown and arbitrary dependence between test statistics.
The estimates of m_0 or m_1 are given by (2.5),

$$\begin{aligned} \widehat{m}_1 &= \sup_{\gamma\in\Gamma}\big\{R(\gamma) - G^\alpha(\gamma)\big\}, \\ \widehat{m}_0 &= m - \widehat{m}_1. \end{aligned} \tag{2.7}$$

The set Γ was chosen as the interval $[0, 1]$. For ease of implementation, this set can for numerical computations be approximated by a substantial number of equally spaced points between 0 and 1. The properties of the estimate do not rely heavily on the precise choice of the set Γ in this case. If not mentioned otherwise, we will assume in the numerical results that Γ consists of 1000 equally spaced points between 0 and 1. The supremum is hence replaced by a maximum in the following.
We will discuss in the following the construction of a bounding function. To find a tight bounding function, we propose to select a parameterized family of functions and search for the "smallest" function in this family that fulfills condition (2.6).

DEFINITION 2.4 *Let* $\mathcal{G}$ *be a function family of real-valued, monotonously increasing functions* $g_\xi : [0, 1] \to \{0, 1, \ldots, m\}$*, indexed by parameter* $\xi \in [0, 1]$*, with the following properties*

(i) $g_\xi(1) = m$, $\forall\xi$

(ii) $\xi_1 \leq \xi_2 \Leftrightarrow g_{\xi_1}(\gamma) \leq g_{\xi_2}(\gamma)$, $\forall\gamma \in [0, 1]$.

Three possible function families will be presented further below.
Let $\xi_{\min} \in [0, 1]$ be

$$\xi_{\min} = \text{argmin}_{\xi\in[0,1]}\big\{\xi : P\big[\max_{\gamma\in\Gamma}\big\{V^0(\gamma) - g_\xi(\gamma)\big\} > 0\big] \leq \alpha\big\}.$$

According to Proposition 2.1, $g_{\xi_{\min}}$ is a bounding function at level α and the estimates of m_1 and m_0 can be evaluated with this bounding function.
The algorithm follows three steps:

Step 1: Fix a function family $\mathcal{G}$, according to Definition 2.4.

Step 2: Find the value of $\xi_{\min}$ as described below.

Step 3: Evaluate $\widehat{m}_1$ and $\widehat{m}_0$ according to (2.5) with the bounding function $g_{\xi_{\min}} \in \mathcal{G}$.

Suitable function families will be presented in section 2.4.
Regarding Step 2, we will show in the following how for a given $\xi \in [0,1]$ a decision can be made whether $\xi_{\min}$ is greater or smaller than ξ. With this information, the value of $\xi_{\min}$ can be found iteratively.
Let ξ be given. Consider all n! permutations of the observations of the response variable Y (or a random subset thereof). Calculate the p-values $P_i^0, i = 1, \ldots, m$ of all hypotheses under the original observed values of $(\underline{X}_k)_{k=1,\ldots,n}$ with randomly permuted values of $(Y_k)_{k=1,\ldots,n}$. Check, for every $\gamma_i \in \Gamma$, that

$$\sum_{i=1}^{m} 1_{[P_i^0 \leq \gamma_i]} \leq g_\xi(\gamma_i).$$

If this condition is fulfilled for every $\gamma_i \in \Gamma$, set $c(p) = 0$. Otherwise, set $c(p) = 1$. If, summing over all permutations,

$$\sum_p c(p) < \sum_p \alpha, \tag{2.8}$$

we have $\xi_{\min} < \xi$. Otherwise $\xi_{\min} \geq \xi$.
Finally, the evaluation of the estimates according to (2.5) in Step 3 is straightforward.

2.4 Function Families

The properties of the estimate of m_1 are determined by the choice of the function family $\mathcal{G}$. We present three families $\mathcal{G}^a$, $\mathcal{G}^b$ and $\mathcal{G}^c$ and discuss their relative strengths and weaknesses.
Let the members of function family $\mathcal{G}^a$ be defined for any value of $0 \leq u \leq m$.

$$g_\xi^a(\gamma) = \begin{cases} u & \gamma \leq 1 - \xi \\ m & \gamma > 1 - \xi \end{cases} \tag{2.9}$$

As will be seen below, this family leads to estimates of m_1 that correspond to control of the generalized family-wise error rate.
We introduce a second family $\mathcal{G}^b$, whose members are defined for any constant $0 < \lambda < 1$ as

$$g_\xi^b(\gamma) = \begin{cases} \xi & \gamma \leq \lambda \\ m & \gamma > \lambda \end{cases} \tag{2.10}$$

Function family $\mathcal{G}^b$ still requires a somewhat arbitrary choice of a parameter λ. No parameter is needed for family $\mathcal{G}^c$ with members

$$g_\xi^c(\gamma) = Q^\xi(z, \gamma), \tag{2.11}$$

where $Q^\xi(z,\gamma)$ is the ξ-quantile of $V^0(\gamma)$, conditional on $Z = z$.

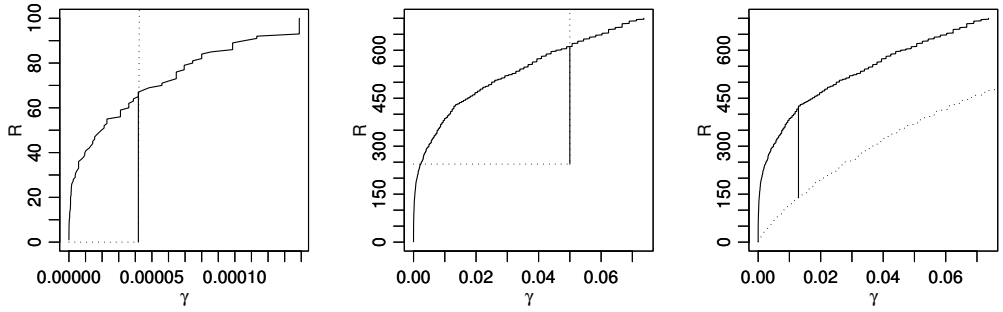

Figure 1: The cumulated amount $R(\gamma)$ of observed p-values for the colon cancer data (continuous line) and the bounding functions resulting from function family $\mathcal{G}^a$ (left, with $u = 0$), family $\mathcal{G}^b$ (middle, with $\lambda = 0.05$) and family $\mathcal{G}^c$ (right). The supremum of the difference between $R(\gamma)$ and the bounding function is plotted as a vertical line and the length of the vertical line is the estimate of m_1.

A choice of the function class leads, as discussed, to estimates of m_1 and m_0. We focus in the following on the properties of the estimates of m_1. The properties of the estimates of m_0 follow immediately. The estimates are denoted by $\widehat{m}_1^a$, $\widehat{m}_1^b$, and $\widehat{m}_1^c$ respectively.

We write in the following $[\cdot]_+$ as a shorthand notation for $\max\{0, \cdot\}$.

PROPOSITION 2.2 *The estimates $\widehat{m}_1^a$, $\widehat{m}_1^b$ and $\widehat{m}_1^c$ are given for $Z = z$ by*

$$\widehat{m}_1^a = \big[R(1 - \xi_{\min}) - u\big]_+, \tag{2.12}$$

$$\widehat{m}_1^b = \big[R(\lambda) - \xi_{\min}\big]_+, \tag{2.13}$$

$$\widehat{m}_1^c = \big[\sup_{\gamma \in \Gamma} \big\{R(\gamma) - Q^{\xi_{\min}}(z, \gamma)\big\}\big]_+. \tag{2.14}$$

For all estimates, $0 \leq \xi_{\min} \leq 1$. The value of $\xi_{\min}$ is bounded from above in the case of estimate $\widehat{m}_1^a$ by $\min\{\xi : P[V(\xi) > u] \leq \alpha\}$. For estimate $\widehat{m}_1^b$ the value of $\xi_{\min}$ is given by $Q^{1-\alpha}(z, \lambda)$. Finally, for estimate $\widehat{m}_1^c$ the value of $\xi_{\min}$ is bounded from below by $1 - \alpha$. If Γ is a finite set, then as well $\xi_{\min} \leq 1 - \alpha/|\Gamma|$. All three estimates take values between 0 and m,

$$0 \leq \widehat{m}_1^a, \widehat{m}_1^b, \widehat{m}_1^c \leq m. \tag{2.15}$$

It can be seen that the estimate $\widehat{m}_1^a$ is for $u = 0$ identical to the maximal possible number of rejections when controlling the family-wise error rate $P[V > 0]$ at level α. For a positive value of u, the estimate is identical to the number of rejections under control of the generalized family-wise error rate $P[V > u]$ at level α. The power of the resulting estimate is very poor for many tested hypotheses as will be seen later.

The second estimate $\widehat{m}_1^b$ is determined in contrast by the number of rejections $R(\lambda)$ at a fixed value λ less an appropriate quantity. The estimate is powerful for large numbers of tested hypotheses (as made rigorous below). On the downside, the estimate involves the somewhat arbitrary choice of the parameter λ.

No parameter has to be chosen in function family (c) and, compared to function family (b), we gain at least in the asymptotic sense as the best possible choice of λ is made automatically.

2.5 Asymptotic power

We look at the power of the different estimates to detect true non-null hypotheses in the limit of large numbers m of tested hypotheses. The power is here defined as the expected proportion of correctly identified true non-null hypotheses,

$$E\big[\frac{\widehat{m}_1}{m_1}\big]. \tag{2.16}$$

The power converges to 1 for all three estimates $\widehat{m}_1^a$, $\widehat{m}_1^b$, $\widehat{m}_1^c$ in the limit $n \to \infty$ of infinitely many observations. We examine in the following the more interesting limit of a fixed number of observations and increasingly many hypotheses, $m \to \infty$. In particular, let us for notational simplicity assume in the following that $\underline{X}$ is infinite-dimensional. We test the first m components of $\underline{X}$ for association with Y and examine the behaviour of the estimates of m_1 for $m \to \infty$.
In this section, we make the dependence of all functions on the value of m explicit by including it as a first argument, e.g. writing $R(m, \gamma)$ for the number of rejections among the first m hypotheses.
We make two assumptions on the test statistics. First,

(A1) There exists a function $F(\gamma)$ with $F(\gamma) \geq \gamma$, right continuous at $\gamma = 0$, such that pointwise in γ,

$$\frac{R(m,\gamma)}{m} \xrightarrow{a.s.} F(\gamma) \qquad \text{for } m \to \infty.$$

The function $F(\gamma)$ is equivalent to the cumulative distribution function of all p-values in the limit of infinitely many tested hypotheses. Assumption (A1) is similar (though less strict) to an assumption made in Storey et al. (2004). It is argued there that the assumption is fulfilled in most cases of practical interest.
For our result about the asymptotic power, the dependence between the test statistics has to be constrained. Consider the random variable $V^0(m, \gamma)$ as a function of the number m of included hypotheses,

$$V^0(m,\gamma) = \sum_{i=1}^{m} 1_{[P_l^0 \leq \gamma]}.$$

The value of $V^0(m, \gamma)$, conditional on Z, is equal to the number of hypotheses with a p-value below γ under a random permutation of the Y-values.
Note that the extreme case for the growth rate of this variance is quadratic in m,

$$\mathrm{Var}\big(\sum_{i=1}^{m} 1_{[P_l^0 \leq \gamma]} \big| Z\big) = O(m^2) \qquad \text{for } m \to \infty.$$

For example, this maximal growth rate is attained in the case where all test statistics have identical values. We exclude such extreme cases and require that

(A2) For any $\gamma \in \Gamma$ and $Z = z \in \mathcal{Z}$,

$$\mathrm{Var}\big(\sum_{i=1}^{m} 1_{[P_l^0 \leq \gamma]} \big| Z\big) = o(m^2) \qquad \text{for } m \to \infty.$$

Note that (A2) does not have to hold uniformly for all $Z \in \mathcal{Z}$.
As the computation of the proposed estimates of m_1 is permutation-based, it is natural to use a rank-based test (e.g. Wilcoxon test) for the necessary testing of association.
We examine the asymptotic power of the estimates $\widehat{m}_1^a$, $\widehat{m}_1^b$, and $\widehat{m}_1^c$ for rank-based tests. This implies that Γ in Definition 2.2 is a finite subset of $[0,1]$, namely the set of possible p-values, which is fixed for fixed sample size n, although the number of tests m is allowed to increase. Stronger assumptions would be necessary to treat the case of tests with continuous p-values.

THEOREM 2.2 *Assume that Γ in Definition 2.2 is a fixed, finite subset of $[0,1]$. The estimates $\widehat{m}_1^a$, $\widehat{m}_1^b$, and $\widehat{m}_1^c$, divided by m, converge a.s. under Assumptions (A1) and (A2) to*

$$\begin{aligned} \widehat{m}_1^a/m &\xrightarrow{a.s.} 0, \\ \widehat{m}_1^b/m &\xrightarrow{a.s.} F(\lambda) - \lambda, \\ \widehat{m}_1^c/m &\xrightarrow{a.s.} \max_{\gamma \in \Gamma} \{F(\gamma) - \gamma\}. \end{aligned}$$

The power vanishes hence for control of the family-wise error rate, estimate $\widehat{m}_1^a$. Positive asymptotic power is achieved with estimates $\widehat{m}_1^b$ and $\widehat{m}_1^c$ as long as $F(\gamma) > \gamma$, which requires by assumption (A1) that the proportion m_1/m of true non-null hypotheses is not vanishing for $m \to \infty$.
Family $\widehat{m}_1^c$ is seen to result asymptotically in the best power. For an optimal choice of the constant λ, the asymptotic power of both $\widehat{m}_1^b$ and $\widehat{m}_1^c$ is equivalent. It is not clear, however, how this optimal constant for estimate $\widehat{m}_1^b$ can be found in a given problem without introducing bias. Hence we usually prefer estimate $\widehat{m}_1^c$.

2.6 Estimation of Error Rates

Besides the mentioned advantage of knowledge about m_0 in a decision-making context, an estimate of m_0 is useful to give tighter estimates of error rates in multiple testing procedures. 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 (omitting the family-wise error rate *FWER*),

Per-comparison error rate *PCER*. The per-comparison error rate is defined as $E[V]/m$, the expected number of Type I errors divided by the total number of hypotheses.

Per-family error rate (*PFER*). The per-family error rate is defined simply as the expected number of Type I errors, $E[V]$.

False discovery rate (*FDR*). The false discovery rate is defined as $E[Q]$, where Q is the proportion of falsely rejected hypotheses

$$Q = \begin{cases} V/R & R > 0 \\ 0 & R = 0 \end{cases}.$$

Storey (2002) was the first to make use of an estimate of m_0 to give a less conservative estimate of the false discovery rate *FDR*. In section 3 we will show, however, that the proposed estimate of m_0 in Storey (2002) has a very large variance for dependent test statistics and that our proposed estimate avoids this problem.

The proposed estimates of m_0 can as well be used to give less conservative estimates 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

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

The value of m_0 is unknown but bounded by m. The value of the error rates can thus be bounded from above by $PCER \leq \gamma$ and $PFER \leq m\gamma$. These bounds are rather conservative if there are a lot of true non-null hypotheses. We can use our estimate $\widehat{m}_0^c$ of m_0, to produce less conservative estimates. It is sufficient to restrict ourselves to the case of $PCER$, as the case of $PFER$ follows by multiplying with m. The proposed estimate of $PCER$ is

$$\widehat{PCER} = \frac{\widehat{m}_0^c}{m}\,\gamma.$$

This estimate is always smaller than the conservative upper bound, $\widehat{PCER} \leq \gamma$. We are still on the safe side, however, as the estimate is, by Theorem 2.1, larger than the true value of $PCER$ with arbitrarily high probability $1-\alpha$,

$$P[\widehat{PCER} \geq PCER] \geq 1-\alpha.$$

Likewise for the per-family error rate $PFER$. The proposed estimates can hence be useful for estimating the mentioned error rates in a less conservative fashion than with the trivial bound $m_0 \leq m$.

In Storey (2002), an estimate $\widehat{FDR}^{\lambda} = \widehat{m}_0^{\lambda}\gamma/R$ of the false discovery rate of a rejection region $[0, \gamma]$ is proposed, where the estimate $\widehat{m}_0^{\lambda}$ is given by

$$\widehat{m}_0^{\lambda} = \frac{m - R(\lambda)}{1-\lambda}, \tag{2.17}$$

having the property $E[\widehat{m}_0^{\lambda}] \geq m_0$ and $E[\widehat{FDR}^{\lambda}] \geq FDR$. Note that this conservative property is only valid under assumption of a point null hypotheses, e.g. $H_0 : \theta_i = 0$ for all $i = 1, \ldots, m$ and is hence limited in its applicability. Furthermore, the variance of this estimate of m_0 is becoming very large for dependent test statistics as will be seen in the numerical examples in section 3. The true false discovery rate is then very frequently underestimated.

Our estimate $\widehat{m}_0^c$ converges to m as α tends to zero. It can then be shown that for α sufficiently small, $E[\widehat{m}_0^c] \geq m_0$ and, as for the estimate in Storey (2002), $E[\widehat{FDR}^c] \geq FDR$ for the new estimate

$$\widehat{FDR}^c = \widehat{m}_0^c\gamma/R.$$

This estimate is always less conservative than the Hochberg-type estimate $m\gamma/R$ while the risk of underestimating the true false discovery rate is only marginally higher. This is an improvement over the estimate in Storey (2002).

3 Numerical Results

We will demonstrate the power of the proposed estimates with simulated and microarray data. The estimates $\widehat{m}_1^a$, $\widehat{m}_1^b$ and $\widehat{m}_1^c$ are calculated as described above. Knowledge about the dependence between the test statistics is not used for the construction of the estimates of m_1 and m_0. We use the algorithm of section 2.3.

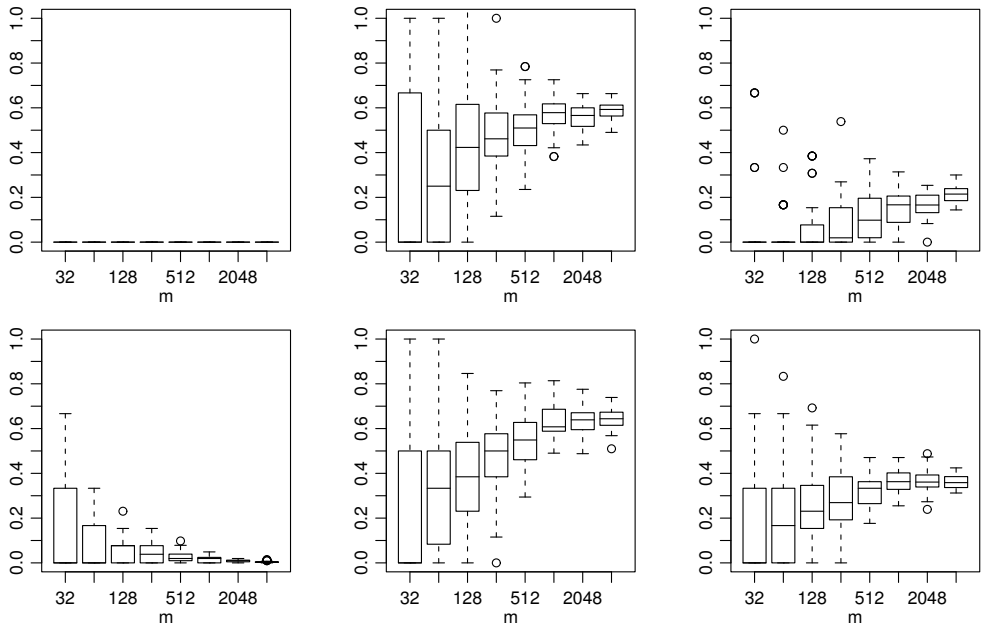

Figure 2: The ratio $\widehat{m}_1/m_1$ as a function of m, the number of tested hypotheses. Shown are $\widehat{m}_1^a$ (left, with $u = 0$), $\widehat{m}_1^b$ (middle, with $\lambda = 0.1$) and $\widehat{m}_1^c$ (right) for a sample size of 10 in the upper row and 20 in the lower row. The estimate $\widehat{m}_1^a$ corresponds to control of the *FWER* and it can be seen that the power is rather poor and vanishes for large m. The proposed estimates $\widehat{m}_1^b$ and $\widehat{m}_1^c$ show increasing power to detect true non-null hypotheses for many tested hypotheses.

3.1 Simulation Study

We test m components of a random variable X for a association with a binary response variable Y for varying values of m. The response variable has a Bernoulli distribution with $p = 0.5$. The random variable $\underline{X}$ is normally distributed,

$$X \sim \mathcal{N}(\mu, \Sigma)$$

with a covariance matrix of the form

$$\Sigma_{ij} = \begin{cases} 1 & i = j \\ \rho^2 & i \neq j \end{cases} . \tag{3.1}$$

and a mean vector μ with components $\mu_i = 1_{[Y=1]} 1_{[h_i=1]}$, where $h_i = 1_{[i>0.9m]}$ determines whether the hypothesis i is a true null hypotheses (if $h_i = 0$) or not (if $h_i = 1$).
For each hypothesis $i = 1, \ldots, m$ it is tested with the Wilcoxon test if the mean of the distribution of X_i is independent of Y.

For $n = 10$, $n = 20$ and $\rho = 0$ we show in Figure 2 the empirical distribution of $\widehat{m}_1/m_1$ (at level $\alpha = 0.05$) for 50 simulations as a function of the number m of tested hypotheses. The power of a

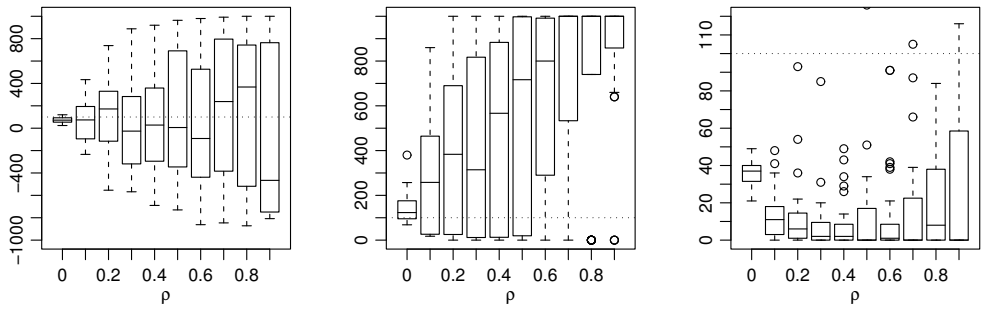

Figure 3: We show for $m = 1000$ and $m_1 = 100$ (dotted line) true non-null hypotheses the distribution of three estimates of m_1 for increasing degrees of dependence ρ between the test statistics. On the left and middle panel we show $\widehat{m}_1^\lambda = m - \widehat{m}_0^\lambda$, where $\widehat{m}_0^\lambda$ is the estimate of m_0 in Storey (2002), once with the common choice of $\lambda = 0.5$ (left) and once with the bootstrapped value of λ (middle). On the right, the proposed estimate $\widehat{m}_1^c$ is shown. Note the different scale of this plot.

FWER-controlling method (which corresponds to $\widehat{m}_1^a$ with $u = 0$ in our case) vanishes for large m as expected from Theorem 2.2. In fact, for $n = 10$ the estimate vanishes identically for all values of m. The proposed estimates $\widehat{m}_1^b$ and $\widehat{m}_1^c$ show qualitatively a different behaviour. First, their power is already quite large for $n = 10$. Second, the power actually increases for increasing m, converging to a positive value for very large values of m, as expected from Theorem 2.2.
In the next simulation we examine the effect of increasing dependence between the test statistics. In particular, we compare our estimate $\widehat{m}_0^c$ to the estimate $\widehat{m}_0^\lambda$ of m_0, as proposed in Schweder and Spjøtvoll (1982) and Storey (2002), equation (2.17).
We set for better comparison $\widehat{m}_1^\lambda = m - \widehat{m}_0^\lambda$. The parameter λ has to be chosen heuristically. A bootstrap method for an optimal choice of λ was proposed in Storey (2002).
The distribution of $\widehat{m}_1^\lambda$ (with the most common choice of $\lambda = 0.5$) and $\widehat{m}_1^\lambda$ (with the value of λ chosen by the bootstrap method) is shown in Figure 3 for $m = 1000$ hypotheses and $m_1 = 100$ true non-null hypotheses.
It can be seen that the estimate $\widehat{m}_1^\lambda$ is not suitable for strong dependence between the test statistics for either choice of λ. In fact, $\widehat{m}_1^\lambda$ is quite often negative or magnitudes larger than the true $m_1 = 100$ even for a moderate dependence like $\rho = 0.1$. Thresholding the estimate $\widehat{m}_1^\lambda$ at 0 (or $\widehat{m}_0^\lambda$ at m) resolves this problem, but the conservative property $E[\widehat{m}_1^\lambda \leq m_1$ will in general be lost.
No thresholding is necessary for the proposed estimates $\widehat{m}_1^a$, $\widehat{m}_1^b$ and $\widehat{m}_1^c$ as their range is limited naturally by $[0, m]$. Although the estimate $\widehat{m}_1^c$ (or equivalently $\widehat{m}_0^c$) is slightly negatively affected by increasing dependence ρ between the test statistics, the mean squared error is much better than for the estimate proposed in Storey (2002).
We did not show the corresponding results for the estimate $\widehat{m}_1^a$ (the number of rejection while controlling the family-wise error rate) as the estimate vanishes identically for our setting. The results for the estimate $\widehat{m}_1^b$ are similar to the results for estimate $\widehat{m}_1^c$ with the disadvantage that a value of λ has to be chosen rather arbitrarily.
Finally, we compare the three estimates $m\gamma/R$ (Benjamini-Hochberg type), $\widehat{m}_0^\lambda\gamma/R$ (as proposed in Storey (2002) with $\lambda = 0.5$), and the new estimate $\widehat{m}_0^c\gamma/R$ in terms of probability of under-

estimating the true false discovery rate and in terms of mean squared error. It is clear that the risk of underestimating the true false discovery rate is smallest for the most conservative estimate, setting $\widehat{m}_0 = m$, but only marginally higher for the less conservative estimate $\widehat{m}_0^c\gamma/R$. In contrast, the estimate $\widehat{m}_0^\lambda\gamma/R$ has a substantial risk of underestimating the true false discovery rate.

The probability $P[\widehat{FDR} < FDR]$ of underestimating the true false discovery rate.

	$m_0 = 900$				700				500			
$\widehat{FDR}$	$\rho = 0$	0.2	0.4	0.6	0	0.2	0.4	0.6	0	0.2	0.4	0.6
$\widehat{m}_0^c\gamma/R$	.12	.12	.08	.10	.00	.04	.06	.03	.00	.02	.00	.00
$m\gamma/R$	.01	.10	.08	.10	.00	.01	.02	.02	.00	.00	.00	.00
$\widehat{m}_0^\lambda\gamma/R$	.34	.30	.31	.30	.25	.36	.39	.36	.53	.58	.25	.36

Regarding the mean squared error, the estimate of Storey (2002) is best for independent test statistics. For reasonably strong dependent test statistics, however, the proposed estimate $\widehat{m}_0^c\gamma/R$ does not only have lower risk of underestimating the true false discovery rate but also a lower mean squared error.

The mean squared error $E[(\widehat{FDR} - FDR)^2]$, multiplied by 10^3.

	$m_0 = 900$				700				500			
$\widehat{FDR}$	$\rho = 0$	0.2	0.4	0.6	0	0.2	0.4	0.6	0	0.2	0.4	0.6
$\widehat{m}_0^c\gamma/R$	2.4	66	121	169	.83	10	21	42	.68	3.2	6.2	24
	(.56)	(2.3)	(2.8)	(3.2)	(.05)	(.65)	(1.2)	(1.8)	(.03)	(.3)	(.46)	(1.2)
$m\gamma/R$	3.1	71	134	173	3.3	10	24	57	3.5	5.5	9.0	26
	(.023)	(.18)	(.23)	(.024)	(.006)	(.05)	(.095)	(.15)	(.003)	(.028)	(.052)	(.12)
$\widehat{m}_0^\lambda\gamma/R$	.08	113	187	221	.06	10	26	55	.013	1.8	5.4	22
	(.026)	(.27)	(.34)	(.37)	(.006)	(.09)	(.15)	(.21)	(.003)	(.04)	(.069)	(.14)

Our limited simulation experience suggests that Storey's estimate $\widehat{m}_0^\lambda\gamma/R$ is preferable for independent test statistics, whereas the proposed estimate $\widehat{m}_0^c\gamma/R$ is best for reasonably strong dependence between the test statistics.

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, e.g. genes whose expression values shows a systematic variation among different groups. Given a response variable Y like tumour type or clinical outcome, it can be tested for each gene if the expression values $\underline{X}$ are associated with Y. We look specifically at three microarray studies. The response variable is binary in each case and predicts the clinical outcome of breast cancer, van't Veer et al. (2002), distinguishes between different subtypes of leukemia, Golub et al. (1999), or indicates absence and presence of colon cancer, Alon et al. (1999).

We compare the estimates $\widehat{m}_1^a$ (with $u = 0$), $\widehat{m}_1^b$ (with $\lambda = 0.05$ and $\lambda = 0.1$) and $\widehat{m}_1^c$. The estimate $\widehat{m}_1^a$ is equivalent to the number of rejections when controlling the $FWER$ at level 0.05 and 0.01. For the estimates $\widehat{m}_1^b$ and $\widehat{m}_1^c$, we use the approach as laid out in section 2.3. Instead of $\widehat{m}_1^a$, however, we use the more powerful step-down method of Westfall and Young (1993) to control the family-wise error rate, which is slightly less conservative than the permutation-based approach under the complete null hypothesis. Additionally the number of rejections for control of $FWER$, using the Bonferroni correction, is shown.

		colon $m = 2000$	leukemia $m = 3571$	breast $m = 5893$
$\alpha = 0.05$	$\widehat{m}_1^a$, Bonferroni	55	266	2
	$\widehat{m}_1^a$, Step-down	64	281	3
	$\widehat{m}_1^b$, $\lambda = 0.1$	363	1049	392
	$\widehat{m}_1^b$, $\lambda = 0.05$	367	1043	370
	$\widehat{m}_1^c$	286	957	355
$\alpha = 0.01$	$\widehat{m}_1^a$, Bonferroni	32	191	0
	$\widehat{m}_1^a$, Step-down	36	202	0
	$\widehat{m}_1^b$, $\lambda = 0.1$	256	866	97
	$\widehat{m}_1^b$, $\lambda = 0.05$	266	908	136
	$\widehat{m}_1^c$	245	811	126

With the estimates $\widehat{m}_1^b$ or $\widehat{m}_1^c$ consistently more true non-null hypotheses are detected than with control of the family-wise error rate, which is equivalent to the estimate $\widehat{m}_1^a$.
Note that the gain of using the proposed estimates compared to control of the *FWER* depends on the number of tested hypotheses. Indeed, the least dramatic gain (which is still roughly a factor four) is for the colon cancer and leukemia 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 *FWER* at level $\alpha = 0.01$ while the estimate $\widehat{m}_1^c$ at the same level indicates that there are more than 100 true null hypotheses.
From a pragmatic point of view, estimation of m_1 is probably most useful if the number of rejections for control of *FWER* is zero or close to zero. In the case of the leukemia data, more than 200 differentially expressed genes are found with control of *FWER*, already more than most biologists probably want to deal with. For the breast cancer data on the other hand only very few rejections can be made under control of *FWER*, while we get evidence for a substantial amount of true non-null hypotheses when using our proposed estimates.

4 Conclusion

We presented a general framework to obtain lower probabilistic bounds for the number m_1 of true non-null hypotheses or, equivalently, upper probabilistic bounds for the number of true null hypotheses.
The number m_0 of true null hypotheses is bounded with very high probability from above -for arbitrary and unknown dependence between the test statistics- by the proposed estimate of m_0.
The properties of a particular estimate of m_0 depend on the choice of the so-called bounding function. For a special choice of this function, control of the family-wise error rate is recovered. The power of *FWER* vanishes, however, in the limit of many tested hypotheses.
We are able to make a different choice of the bounding function such that the power to detect true non-null hypotheses remains positive even for infinitely many tested hypotheses. In fact, the power of the proposed estimate is increasing with the number of tested hypotheses.
We showed with theoretical considerations and numerical examples that the proposed estimate of m_0 is, even under strong dependence between the test statistics, very powerful for delivering a tight probabilistic upper bound for the number of true null hypotheses in a multiple testing situation.

Finally, our method can be used for novel estimates of error rates. In particular, we demonstrate its use for obtaining good estimates of the false discovery rate.

References

Alon, U., N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Cell Biology 96*, 6745–6750.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B 57*, 289–300.

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

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

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

Mosig, M., E. Lipkina, G. Khutoreskayaa, E. Tchourzynaa, M. Sollera, and A. Friedmanna (2001). A whole genome scan for quantitative trait loci affecting milk protein percentage in israeli-holstein cattle, by means of selective milk dna pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics 157*, 1683–1698.

Nettleton, D. and J. Gene Hwang (2003). Estimating the number of false null hypothesis when conducting many tests. Technical report, Iowa State University.

Schweder, T. and E. Spjøtvoll (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 493–502.

Shaffer, J. (1995). Multiple hypothesis testing: A review. *Annual Review of Psychology 46*, 561–584.

Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B 64*, 479–498.

Storey, J., J. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B 66*, 187–205.

Turkheimer, F., C. Smith, and K. Schmidt (2001). Estimation of the number of true null hypotheses in multivariate analysis of neuroimaging data. *NeuroImage 13*, 920–930.

van't Veer, L., H. Dal, M. van der Vijver, Y. He, A. Hart, M. Mao, H. Peterse, K. van der Kooy, M. Marton, A. Witteveen, G. Schreiber, R. Kerkhoven, C. Roberts, P. Linsley, R. Bernards, and S. Friend (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature 406*, 742–747.

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

5 Proofs

Proof of Theorem 2.1. It suffices to show that

$$P[\widehat{m}_1 > m_1] < \alpha, \tag{5.1}$$

where $\widehat{m}_1 = \sup_{\gamma\in\Gamma}\left\{R(\gamma) - G^\alpha(\gamma)\right\}$. The number of rejections can be split into

$$R(\gamma) = S(\gamma) + V(\gamma).$$

Note that $\sup_{\gamma\in\Gamma}\{S(\gamma)\} = S(1) = m_1$. Thus

$$\begin{aligned} P[\widehat{m}_1 > m_1] &= P[\sup_{\gamma\in\Gamma}\left\{R(\gamma) - G^\alpha(\gamma)\right\} > m_1] \\ &= P[\sup_{\gamma\in\Gamma}\left\{V(\gamma) + S(\gamma) - G^\alpha(\gamma)\right\} > m_1] \\ &\leq P[\sup_{\gamma\in\Gamma}\left\{V(\gamma) - G^\alpha(\gamma)\right\} + S(1) > m_1] \\ &\leq P[\sup_{\gamma\in\Gamma}\left\{V(\gamma) - G^\alpha(\gamma)\right\} > 0]. \end{aligned}$$

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

$$P[\sup_{\gamma\in\Gamma}\left\{V(\gamma) - G^\alpha(\gamma)\right\} > 0]$$

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

Proof of Proposition 2.1. The random variable $V(\gamma)$ is given by

$$V(\gamma) = \sum_{i=1}^{m} 1_{[h_i=0]} 1_{[P_i\leq\gamma]}.$$

If the i-th component of $\underline{X}$ and Y are independent, then $P[P_i \leq \gamma] = \gamma$ for every γ in the set of possible p-values. P-values under other true null hypotheses are stochastically greater than in the case of independence.
As P_i^0 and P_i follow the same distribution in the case of independence between the i-th components of $\underline{X}$ and Y, it follows that the distribution of $V(\gamma)$ is bounded from above by the distribution of

$$V^0(\gamma) = \sum_{i=1}^{m} 1_{[P_i^0\leq\gamma]}.$$

Similarly it follows that the distribution of $V(\gamma)$, conditional on $Z = z$, is bounded from above by the distribution of $V^0(\gamma)$, conditional on $Z = z$. Thus, for any given $Z = z$,

$$\begin{aligned} P\Big[\sup_{\gamma\in\Gamma}\left\{V(\gamma) - G^\alpha(\gamma)\right\} > 0\Big|Z = z\Big] &\leq P\Big[\sup_{\gamma\in\Gamma}\left\{V^0(\gamma) - G^\alpha(\gamma)\right\} > 0\Big|Z = z\Big] \\ &< \alpha. \end{aligned}$$

It follows

$$P\Big[\sup_{\gamma\in\Gamma}\left\{V(\gamma) - G^\alpha(\gamma)\right\} > 0\Big] < \alpha,$$

and the function $G^\alpha(\gamma)$ is hence a bounding function at level α.

Proof of Proposition 2.2. We prove the claims separately for each function family.
(Family $\mathcal{G}^a$) The functions g^a_ξ of family $\mathcal{G}^a$ are constant except for one point of discontinuity at $\gamma = 1 - \xi_{\min} \in \Gamma$. As $R(\gamma)$ is monotonously increasing in γ, the maximum in

$$\widehat{m}_1 = \sup_{\gamma\in\Gamma} \big\{R(\gamma) - g_{\xi_{\min}}(\gamma)\big\}. \tag{5.2}$$

is attained either at $\gamma = 1 - \xi_{\min}$ or at $\gamma = 1$. In the latter case, the difference $R(1) - g_{\xi_{\min}}(1)$ is zero. In the former case, for $\gamma = 1 - \xi_{\min}$, the difference is given by $R(\gamma) - g^a_{\xi_{\min}}(\gamma)$. Using $g^a_{\xi_{\min}}(\gamma) = u$ for $\gamma = 1 - \xi_{\min}$, the estimate $\widehat{m}^a_1$ is of the form

$$\widehat{m}^a_1 \quad = \quad \max\big\{0, R(1 - \xi_{\min}) - u\big\}.$$

As $0 \le u, R(\gamma)$ and $R(\gamma) \le m$, it follows that $0 \le \widehat{m}^a_1 \le 1$ as claimed.
(Family $\mathcal{G}^b$) The functions g^b_ξ of family $\mathcal{G}^b$ are again constant except for one point of discontinuity, which is now at $\gamma = \lambda$. At this value of γ, $g^b_\xi(\gamma) = \xi_{\min}$ and $\xi_{\min}$ is hence determined as the minimal value of ξ such that,

$$P\big[V^0(\lambda) - \xi_{\min} > 0\big|Z = z\big] \le \alpha.$$

Hence $\xi_{\min}$ is given by the $(1-\alpha)$-quantile $Q^{1-\alpha}(z,\lambda)$ of $V^0(\lambda)$. The estimate $\widehat{m}^b_1$ is thus of the form

$$\widehat{m}^b_1 \quad = \quad \max\big\{0, R(\lambda) - Q^{1-\alpha}(z,\lambda)\big\},$$

The estimate $\widehat{m}^b_1$ is bounded as $0 \le \widehat{m}^b_1 \le m$, as again $0 \le R(\lambda), Q^{1-\alpha}(z,\lambda)$ and $R(\lambda) \le m$.
(Family $\mathcal{G}^c$) The form

$$\widehat{m}^c_1 = \sup_{\gamma\in\Gamma} \big\{R(\gamma) - Q^{\xi_{\min}}(z,\gamma)\big\}$$

for the estimate $\widehat{m}^c_1$ follows directly from Definition 2.2. Note that it holds for any $\gamma \in \Gamma$ by definition of $Q^\xi(z,\gamma)$ as the ξ-quantile of $V^0(\gamma)$, conditional on $Z = z$,

$$P\big[V^0(\gamma) - Q^\xi(z,\gamma) > 0\big|Z = z\big] < 1 - \xi.$$

On the one hand, by definition of $Q^\xi(\gamma)$,

$$P\big[V^0(\gamma) - Q^{1-\alpha}(z,\gamma) > 0\big|Z = z\big] \ge \alpha,$$

and therefore

$$P\big[\sup_{\gamma\in\Gamma} \big\{V^0(\gamma) - Q^{1-\alpha}(z,\gamma)\big\} > 0\big|Z = z\big] \ge \alpha.$$

It follows that $\xi_{\min} \ge 1 - \alpha$. If Γ is a finite set, it follows on the other hand by Bonferronis inequality that

$$P\big[\sup_{\gamma\in\Gamma} \big\{V^0(\gamma) - Q^{1-\alpha/|\Gamma|}(z,\gamma)\big\} > 0\big|Z = z\big] < \alpha,$$

where $|\Gamma|$ is the cardinality of the set Γ. Hence $\xi_{\min} \le 1 - \alpha/|\Gamma|$.
Both $R(\gamma)$ and $Q^{\xi_{\min}}(z,\gamma)$ are positive and smaller than m for all $\gamma \in [0,1]$. Hence, $\widehat{m}^c_1 \le m$. Furthermore, $R(1) = Q^{\xi_{\min}}(z,1) = m$ and the estimate $\widehat{m}^c_1$ is hence always positive.

LEMMA 5.1 *For a rank-based test, let $Q^\beta(m,z,\gamma)$ be the β-quantile of $V^0(m,\gamma)$, conditional on $Z=z$. It holds for any $z\in\mathcal{Z}$ and $\gamma\in[0,1]$ under Assumption (A2) that*

$$\Big|\frac{Q^\beta(m,z,\gamma)}{m}-\gamma\Big|=o(1)\qquad \text{for } m\to\infty.$$

PROOF OF LEMMA 5.1. It was shown before, in the proof of Proposition 2.1, that $P[P_i^0\le\gamma]=\gamma$ for every γ in the set of possible p-values. As a rank-based test is used, we can extend the argument to conclude that $P[P_i^0\le\gamma|Z]=\gamma$ for any given $Z\in\mathcal{Z}$. Thus, for a given $Z=z$,

$$E\big[V^0(m,\gamma)\big|Z=z\big]=E\big[\sum_{i=1}^m 1_{[P_i^0\le\gamma]}\big|Z=z\big]=m\gamma.$$

By Chebychev's inequality it follows for any c,

$$P\Big[\big|V^0(m,\gamma)-m\gamma\big|>\sqrt{\frac{\mathrm{Var}\big(V^0(m,\gamma)\big)}{c}}\big|Z=z\Big]\le c.$$

Choosing $c(\beta)=\min\{1-\beta,\beta\}$, it follows

$$\big|Q^{1-\beta}(m,z,\gamma)-m\gamma\big|\le\sqrt{\frac{\mathrm{Var}\big(V^0(m,\gamma)\big|Z\big)}{c(\beta)}}.$$

Dividing by m, the claim follows by Assumption (A2).

PROOF OF THEOREM 2.2. We prove the claims separately for each function family.
(Family $\mathcal{G}^a$) For function family $\mathcal{G}^a$, the estimate is given by

$$\widehat{m}_1^a=\big[R(m,1-\xi_{\min})-u\big]_+.$$

Hence, by Assumption (A1),

$$\frac{\widehat{m}_1^a}{m}=\frac{[R(m,1-\xi_{\min})-u]_+}{m}\le\frac{R(m,1-\xi_{\min})}{m}\xrightarrow{a.s.}F(1-\xi_{\min})\qquad\text{for } m\to\infty.$$

Note that $F(0)=0$ and $F(\gamma)$ is right-continuous at $\gamma=0$. For a proof of the claim, it is hence sufficient to show that $\xi_{\min}\to1$ for $m\to\infty$.
The value of $\xi_{\min}$ is according to (2.6) the minimal value such that for a given $Z=z$,

$$P\big[V^0(m,1-\xi_{\min})>u\big]\le\alpha.$$

Hence $\xi_{\min}$ is the minimal value of $\xi\in[0,1]$ such that $Q^{1-\alpha}(m,z,1-\xi)=u$. Note that, for any $Z=z$, $Q^{1-\alpha}(m,z,1-\xi)$ is monotonously increasing for decreasing ξ from $Q^{1-\alpha}(m,z,1-\xi)=0$ for $\xi=1$ to $Q^{1-\alpha}(m,z,1-\xi)=m$ for $\xi=0$. Furthermore, it follows by Lemma 5.1 that the value of $Q^{1-\alpha}(m,z,1-\xi)$ is diverging for $m\to\infty$ for any value of Z and $0\le\xi<1$. Combining the last two observations, it follows that $\xi_{\min}\to1$ for $m\to\infty$, proving the claim for function family $\mathcal{G}^a$.
(Family $\mathcal{G}^b$) For function family $\mathcal{G}^b$, the estimate is given by

$$\widehat{m}_1^b=\big[R(m,\lambda)-Q^{1-\alpha}(m,z,\lambda)\big]_+,$$

where $Q^{1-\alpha}(m, z, \gamma)$ is the $1-\alpha$ quantile of the distribution of $V^0(m, \gamma)$, conditional on $Z = z$. By Lemma 5.1,

$$\frac{Q^{1-\alpha}(m, z, \lambda)}{m} = \lambda + o(1) \qquad \text{for } m \to \infty.$$

By Assumption (A1) furthermore

$$\frac{R(m, \lambda)}{m} \xrightarrow{a.s.} F(\lambda).$$

Under Assumption (A1), $F(\lambda) \geq \lambda$. Hence, using the last two equations,

$$\frac{\widehat{m}_1^b}{m} \xrightarrow{a.s.} F(\lambda) - \lambda \qquad \text{for } m \to \infty,$$

proving the claim for function family (b).
(Family $\mathcal{G}^c$) For function family $\mathcal{G}^c$,

$$\widehat{m}_1^c = \max_{\gamma \in \Gamma} \big\{ R(m, \gamma) - Q^{\xi_{\min}}(m, z, \gamma) \big\},$$

where $Q^{\xi_{\min}}(m, z, \gamma)$ is again the $\xi_{\min}$ quantile of the distribution of $V^0(m, \gamma)$, conditional on $Z = z$. Let $|\Gamma|$ be again the cardinality of the finite set of possible p-values. It was shown in Proposition 2.2 that $1 - \alpha \leq \xi_{\min} \leq 1 - \alpha/|\Gamma|$. Thus, by Assumption (A2) and Lemma 5.1, for any $\gamma \in \Gamma$,

$$\frac{Q^{\xi_{\min}}(m, z, \gamma)}{m} = \gamma + o(1) \qquad \text{for } m \to \infty.$$

By Assumption (A1), $R(m, \gamma)/m \xrightarrow{a.s.} F(\gamma)$. Hence, for every $\gamma \in \Gamma$,

$$\frac{R(m, \gamma)}{m} - \frac{Q^{\xi_{\min}}(m, z, \gamma)}{m} \xrightarrow{a.s.} F(\gamma) - \gamma \qquad \text{for } m \to \infty.$$

As Γ is a finite set,

$$\begin{aligned} \frac{\widehat{m}_1^c}{m} &= \max_{\gamma \in \Gamma} \Big\{ \frac{R(m, \gamma)}{m} - \frac{Q^{\xi_{\min}}(m, z, \gamma)}{m} \Big\} \\ &\xrightarrow{a.s.} \max_{\gamma \in \Gamma} \big\{ F(\gamma) - \gamma \big\} \qquad \text{for } m \to \infty, \end{aligned}$$

proving the claim for function family $\mathcal{G}^c$.