

Bounding the k -family-wise error rate using resampling methods

Tijl De Bie^{1,2} and John Shawe-Taylor³

¹ OKP Research Group, K.U.Leuven, Belgium

² Department of Engineering Mathematics, University of Bristol, UK

³ Department of Computer Science, University College London, UK

Abstract. The multiple hypothesis testing (MHT) problem has long been tackled by controlling the family-wise error rate (FWER), which is the probability that any of the hypotheses tested is unjustly rejected. The best known method to achieve FWER control is the Bonferroni correction, but more powerful techniques such as step-up and step-down methods exist. A particular challenge to be dealt with in MHT problems is the unknown dependency structure between the tests. The above-mentioned approaches make worst-case assumptions in this regard, which makes them extremely conservative in practical situations, where positive dependencies between the tests are abundant. In this paper we consider randomisation strategies to overcome this problem, and provide a rigorous statistical analysis of the finite-sample behaviour. Furthermore, we extend our results to an approach to control the k -FWER as introduced by [7]. Another result is a uniform bound on the k -FWER, uniform over a specified set of values of k , which additionally allows to control the false discovery proportion (FDP, see e.g. [7]). Our methods are essentially assumption free, and by effectively taking into account dependencies between the tests their strong power is ensured in all situations.

1 Introduction

In order to detect interesting phenomena in data, it is common practice to test whether the data is unlikely, given certain assumptions on the probability distribution of the data. The set of these assumptions is usually referred to as the null hypothesis. The null hypothesis should represent the default belief about the data distribution. If the data is found unlikely given the null hypothesis, one may decide to reject the null hypothesis, and more interesting alternatives may be considered. One speaks of a true rejection if the test rejects the null hypothesis while in reality it does not hold. If the null hypothesis is rejected even though it actually holds, it is called a false rejection.

For concreteness, in this paper we will consider gene expression data. In such a data set, the expression level (a real value) of a gene is measured in a number of tissue samples, each of which belongs to one of a limited number of tissue types (for example, ‘cancerous’ and ‘non-cancerous’), to which we will refer as

the label. The null hypothesis is that gene expression is not related to the tissue type. Then, we can compute a measure of relatedness (such as the correlation with the labels, or the Wilcoxon rank sum test statistic), and assess whether its value is likely given the hypothesis of independence between the gene expression and the label.

For a single gene, this hypothesis testing task is very well understood. But matters become more complicated if several such hypothesis tests are carried out simultaneously, i.e. for all genes in the genome simultaneously. In such cases we are interested to limit the number of false rejections while uncovering (rejecting) as many genes that are truly related to the label as possible. However, the probability to see at least one false rejection grows larger with an increasing number of tests. Therefore, one often attempts to control the so-called family-wise error rate (FWER, see e.g. [10]), which is the probability that *at least one* hypothesis test unjustly rejects the null hypothesis. Not surprisingly, though, for a large number of hypothesis tests, such an approach may be overconservative.

More useful results can be obtained if one is prepared to tolerate a certain number of false rejections. In this context, [7] have defined the k -FWER, which is the probability that *at least k* hypothesis tests falsely reject the null hypothesis. An upper bound on the k -FWER is then sufficient to guarantee the presence of at most $k - 1$ false rejections. A related quantity, the False Discovery Proportion (FDP), which is the number of false rejections divided by the total number of rejections, is of great interest in practice. The FDP is an empirical variant of the False Discovery Rate (FDR), which is essentially defined as the expectation of the FDP, the introduction of which by [2] represented an important step forward in the study of multiple hypothesis tests.

In this paper we propose methods to bound the performance measures FWER, k -FWER and FDP based on resampling techniques, by making use of null hypotheses that are invariant to a group of transformations of the data. For example, the gene expression data being specified as unrelated to the label under the null hypothesis implies that, by permuting the labels, the probability of the data remains the same. Such approaches automatically take dependencies between the expression of different genes into account, which make them less conservative than other approaches.

2 Hypothesis Testing and multiple hypothesis testing

The aim of hypothesis testing is to draw statistical conclusions about the state of the world based on the evidence provided by a finite sample. We will refer to the particular aspects of the state that are considered by means of pattern functions. We use the notation

$$\pi : X \mapsto \mathbb{R}$$

to denote a pattern function from the data space \mathcal{X} to the reals. We expect that there will be some level of output or pattern strength $\pi(X)$ that will indicate the presence of a ‘significant’ pattern.

Hypothesis testing relies on defining a null hypothesis, being the assumption that the data X is drawn according to a distribution \mathcal{D} from a specified set of distributions Ω_0 , i.e. the assumption that $X \sim \mathcal{D}$ with $\mathcal{D} \in \Omega_0$. Typically, the null hypothesis encodes the negation of what we consider interesting. Clearly one should aim to keep the null hypothesis as natural as possible, since any conclusions will be predicated on the appropriateness of Ω_0 .

If we perform a hypothesis test with null hypothesis Ω_0 and pattern function π at significance level p , we look for a pattern strength σ such that

$$\forall \mathcal{D} \in \Omega_0 : P_{X \sim \mathcal{D}}(\pi(X) \geq \sigma) \leq p.$$

If now our actual observation X satisfies $\pi(X) \geq \sigma$, then we conclude that the pattern function π is significant at the level p in the sense that the probability of such a pattern strength being observed under the null hypothesis is at most p . Then one usually decides to disbelieve or to ‘reject’ the null hypothesis. Observing data for which $\pi(X) \geq \sigma$ while the null hypothesis actually holds results in a *type I error* (or a false rejection), and it is controlled by the value of p . Then the test is falsely rejected. Vice versa, observing data X for which $\pi(X) < \sigma$ while the null hypothesis is false results in a *type II error*, the probability of which is hard to evaluate or to control in general.

Consider now a set of hypothesis tests, associated to a class of pattern functions

$$\Pi = \{\pi_\alpha : \alpha \in A\}$$

for some index set A . In our example, X is the gene expression data of all genes in a number of tissues, together with the labels of these tissues. Then $\pi_\alpha(X)$ is a measure of how strongly the expression of gene α relates to the labels, as measured by a correlation, or by the Wilcoxon rank sum test statistic, or by any other measure deemed appropriate. In this case it is common practice to control the FWER, which is the probability of at least one rejection while the null hypothesis holds. Formally, if individual hypothesis tests reject the null hypothesis for pattern strengths $\pi_\alpha(X) \geq \sigma_\alpha$, the FWER is given by $P_{X \sim \mathcal{D}}(\exists \alpha \in A : \pi_\alpha(X) \geq \sigma_\alpha)$. So to upper bound the FWER under Ω_0 at significance level p , we must choose pattern strengths σ_α such that

$$\forall \mathcal{D} \in \Omega_0 : P_{X \sim \mathcal{D}}(\exists \alpha \in A : \pi_\alpha(X) \geq \sigma_\alpha) \leq p.$$

A more useful null hypothesis to consider is one containing all distributions where the values of all pattern functions $\pi_\alpha(X)$ are jointly distributed as they would be under a distribution from Ω_0 , except for a subset $\{\pi_\alpha : \alpha \in \bar{A} \subseteq A\}$ of them. For fixed \bar{A} denote this distribution by $\Omega_0(\bar{A})$. Let us furthermore define

$$\bar{\Omega}_0 = \bigcup_{\bar{A} \subseteq A} \Omega_0(\bar{A}).$$

In the gene expression example, $\bar{\Omega}_0$ would correspond to all but a subset of the genes to be independent of the label.

With each α , we can then associate an individual null hypothesis Ω_0^α , which holds if $\pi_\alpha(X)$ is distributed as it would be under Ω_0 . Formally:

$$\Omega_0^\alpha = \bigcup_{\bar{A} \subseteq A: \alpha \in A \setminus \bar{A}} \Omega_0(\bar{A}).$$

Hence, under $\Omega_0(\bar{A})$, we know that $|\bar{A}|$ of the *individual* hypotheses Ω_0^α are actually false, to be precise those with $\alpha \in \bar{A}$. Therefore, under $\Omega_0(\bar{A})$, a rejection of the individual null hypothesis Ω_0^α with $\alpha \in \bar{A}$ is called a true rejection.

Since the FWER bounds the number of *false* rejections, it can be upper bounded under $\bar{\Omega}_0$ at significance level p , by choosing a pattern strength such that

$$\forall \bar{A} \subseteq A, \forall \mathcal{D} \in \Omega_0(\bar{A}) : P_{X \sim \mathcal{D}}(\exists \alpha \in A \setminus \bar{A} : \pi_\alpha(X) \geq \sigma_\alpha) \leq p.$$

If the FWER under $\bar{\Omega}_0$ is upper bounded by p in this way, and for a number of tests it holds that $\pi_\alpha(X) \geq \sigma$ such that Ω_0^α is rejected, we must conclude that all these rejections are individually significant at the level p (since under any $\mathcal{D} \in \bar{\Omega}_0$, with probability at least $1 - p$ over the data, we do not expect any false rejection at all).

When controlling the FWER, a main task consists in determining the smallest possible value for σ for which the above inequality holds. If the index set A is finite or countably infinite there is a standard approach to this problem, known as the Bonferroni correction (see e.g. [10] and references therein). We recapitulate it in the following Lemma.

Lemma 1 (Bonferroni Correction). *Suppose $A \subseteq \mathbb{N}$ is a countable index set for a class of pattern functions $\{\pi_\alpha : \alpha \in A\}$. Let $q_\alpha \in (0, 1)$ such that*

$$\sum_{\alpha \in A} q_\alpha \leq 1.$$

If we choose σ_α such that $\forall \mathcal{D} \in \Omega_0$:

$$P_{X \sim \mathcal{D}}(\pi_\alpha(X) \geq \sigma_\alpha) \leq pq_\alpha,$$

then $\forall \bar{A} \subseteq A, \forall \mathcal{D} \in \Omega_0(\bar{A})$:

$$P_{X \sim \mathcal{D}}(\exists \alpha \in A \setminus \bar{A} : \pi_\alpha(X) \geq \sigma) \leq p.$$

Proof. The result follows from a simple application of the so-called union bound. In particular if $B_\alpha(X)$ is the event that pattern π_α is significant, then

$$\begin{aligned} P_{X \sim \mathcal{D}}(\exists \alpha \in A \setminus \bar{A} : \pi_\alpha(X) \geq \sigma_\alpha) &= P_{X \sim \mathcal{D}} \left(\bigcup_{\alpha \in A \setminus \bar{A}} B_\alpha(X) \right) \\ &\leq \sum_{\alpha \in A \setminus \bar{A}} P_{X \sim \mathcal{D}}(B_\alpha(X)) \leq p \sum_{\alpha \in A \setminus \bar{A}} q_\alpha \leq p. \end{aligned}$$

■

Remark 1. Suppose we observe a set $\tilde{A} \subseteq A$ for which $\pi_\alpha(X) \geq \sigma_\alpha$, for $\alpha \in \tilde{A}$. Since the assertion

$$\nexists \alpha \in A \setminus \tilde{A} : \pi_\alpha(X) \geq \sigma_\alpha$$

holds with probability $1 - p$, we can conclude with this probability that $\tilde{A} \subseteq \bar{A}$, implying that each α for which $\pi_\alpha(X) \geq \sigma_\alpha$ represents a test for which Ω_0^α fails.

This result, while exact, has limited practical use, because it fails to take into account the correlations between different pattern functions and is therefore necessarily conservative in its estimation. This is particularly apparent if we consider an uncountable set A , in which case the Bonferroni correction is infinitely conservative (it would not reject a single hypothesis).

3 Randomisation testing of a single hypothesis

We will now consider randomisation testing as a technique for testing a single hypothesis. Thereafter we will discuss ways to control the FWER and the k-FWER based on such randomisation testing strategies.

Suppose that for a group of data transformations $g \in G$ from the data space \mathcal{X} onto itself, all distributions $\mathcal{D} \in \Omega_0$ satisfy that $P_{X \sim \mathcal{D}}(X) = P_{X \sim \mathcal{D}}(g(X))$. Then, without knowing which distribution $\mathcal{D} \in \Omega_0$ the data is sampled from, we can generate a number of samples $g(X)$ for all $g \in G$ with the same probability. For example with the gene expression data, one may consider the null hypothesis Ω_0 containing all distributions for which all X with label vector randomly permuted are equally likely.

It is well known that for null hypotheses that are invariant with respect to a certain group of data transformations, one can use a randomisation test to test the hypothesis. Such a randomisation test works as follows. Compute the pattern strength for all transformations of the data $\{g(X) : g \in G\}$. Then, rank these pattern strengths $\{\pi(g(X)) : g \in G\}$ in decreasing order. A test which rejects the null hypothesis if $\pi(X) < \sigma$ is then significant at a level equal to the rank that σ would have in this ordering divided by $|G|$. If we assume for simplicity that there are no ties between the $\pi(g(X))$, we can summarise this result in the following theorem (see e.g. [8]):

Theorem 1. *Given a pattern function π , a null hypothesis Ω_0 invariant under the group of transformations $g \in G$, and a fixed $p \in (0, 1)$. Given the set $\{g(X) : g \in G\}$. Choose $\sigma(X)$ such that*

$$p = \frac{\#\{g \in G : \pi(g(X)) \geq \sigma(X)\}}{|G|}.$$

Then, $\forall \mathcal{D} \in \Omega_0$:

$$P_{X \sim \mathcal{D}}(\pi(X) \geq \sigma(X)) = p.$$

Proof. From the choice of $\sigma(X)$, it follows that

$$\begin{aligned} p|G| &= \#\{g \in G : \pi(g(X)) \geq \sigma(X)\} \\ &= \sum_{g \in G} I(\pi(g(X)) \geq \sigma(X)) \end{aligned}$$

where $I(\cdot)$ is the indicator function. Subsequently taking the expectation with respect to any data distribution $\mathcal{D} \in \Omega_0$ yields:

$$\begin{aligned} p|G| &= \sum_{g \in G} E_{X \sim \mathcal{D}} \{I(\pi(g(X)) \geq \sigma(X))\} \\ &= |G| P_{X \sim \mathcal{D}}(\pi(X) \geq \sigma(X)) \end{aligned}$$

where we used the fact that all $\mathcal{D} \in \Omega_0$ are invariant with respect to the transformations in G . Hence the result follows. \blacksquare

In practice, such as in the example of the gene expression data for any reasonable number of microarrays, $|G|$ is too large to compute the complete set $\{\pi(g(X)) : g \in G\}$. Therefore, one often makes use of a random sample of G of size m , further denoted by G_m (here, we assume that G_m is sampled with repeats—slightly tighter bounds may be obtained if no repeats are allowed). We will show that in this case a slightly more conservative strategy is appropriate. The result follows from a sample tail bound for the Binomial distribution as the lemma below shows. We first introduce the necessary notation following Langford [6].

Definition 1 (Binomial Tail Distribution and Tail Inversion). *The probability*

$$\text{Bin}(m, h, p) = \sum_{j=0}^h \binom{m}{j} p^j (1-p)^{m-j}$$

equals the probability that m random coin flips with bias p produce h or fewer heads. The probability

$$\overline{\text{Bin}}(m, h, \delta) = \max\{q : \text{Bin}(m, h, q) \geq \delta\}$$

equals the largest true bias such that the probability of observing h or fewer heads is at least δ .

It is easy to see that both $\text{Bin}(m, h, p)$ and $\overline{\text{Bin}}(m, h, \delta)$ are monotonically increasing with h . We will make use of the following Lemma, which is a reformulation of Theorem 3.3 in [6].

Lemma 2. *Consider a random coin with a bias p , the probability of a head. Denote by $\hat{p}m$ the number of heads in a sample S of m flips. Then, with probability $1 - \delta$ over the coin flips S ,*

$$p \leq \overline{\text{Bin}}(m, \hat{p}m, \delta).$$

Proof. Define $\hat{p}^* = \sup\{\hat{p} : \overline{\text{Bin}}(m, \hat{p}m, \delta) < p\}$. Then by the definition of $\overline{\text{Bin}}$ we have

$$\text{Bin}(m, \hat{p}^*m, p) \leq \delta.$$

It follows that if p is the true bias and $\hat{p}m$ is the number of heads in the random sample S of size m then

$$P_S(\overline{\text{Bin}}(m, \hat{p}m, \delta) < p) = P_S(\hat{p}m \leq \hat{p}^*m) = \text{Bin}(m, \hat{p}^*m, p) \leq \delta,$$

from which the result follows. \blacksquare

We can now prove the first result in this paper, which states how the type I error of a single hypothesis test can be controlled based on a sample G_m of data transformations.

Theorem 2. *Given a pattern function π , a null hypothesis Ω_0 invariant under the group of transformations $g \in G$, a fixed $\delta \in (0, 1)$ and a fixed $\hat{p} \in (0, 1)$. Given a uniformly sampled multiset G_m from G of size m . Choose $\sigma(X)$ such that*

$$\hat{p} = \frac{\#\{g \in G_m : \pi(g(X)) \geq \sigma(X)\}}{m}.$$

Then with probability at least $1 - \delta$ over the m -sample G_m , $\forall \mathcal{D} \in \Omega_0$:

$$P_{X \sim \mathcal{D}}(\pi(X) \geq \sigma(X)) \leq \overline{\text{Bin}}(m, \hat{p}m, \delta).$$

Proof. From Lemma 2 and from the choice of $\sigma(X)$, it follows that with probability at least $1 - \delta$ over the m -sample G_m :

$$\frac{\#\{g \in G : \pi(g(X)) \geq \sigma(X)\}}{|G|} \leq \overline{\text{Bin}}(m, \hat{p}m, \delta).$$

Hence,

$$\begin{aligned} \overline{\text{Bin}}(m, \hat{p}m, \delta)|G| &\geq \#\{g \in G : \pi(g(X)) \geq \sigma(X)\} \\ &\geq \sum_{g \in G} I(\pi(g(X)) \geq \sigma(X)) \end{aligned}$$

Taking expectations with respect to any $\mathcal{D} \in \Omega_0$, and exploiting the invariance of Ω_0 with respect to transformations from G , yields

$$\overline{\text{Bin}}(m, \hat{p}m, \delta)|G| \geq |G|P_{X \sim \mathcal{D}}(\pi(X) \geq \sigma(X)),$$

from which the result follows. \blacksquare

Remark 2. This approach is usually referred to as the Bootstrap method.

It is instructive to keep in mind that for $\delta < 0.5$ and hence for all practical purposes, $\hat{p} \leq \overline{\text{Bin}}(m, \hat{p}m, \delta) \leq \hat{p} + \sqrt{\frac{\ln(\frac{1}{\delta})}{2m}}$, see [6]. This means that for increasing m and $\delta < 0.5$, the value of $\overline{\text{Bin}}(m, \hat{p}m, \delta)$ converges from above to \hat{p} , the proportion of $g \in G_m$ in which the null hypothesis is rejected. Hence, unsurprisingly, in practice it is advantageous to take the sample size m as large as possible in order to obtain a FWER-controlled test as tight and powerful as possible.

4 Controlling the FWER

In a similar way, the FWER under $\bar{\Omega}_0$ can be controlled. Recall that under $\bar{\Omega}_0$, there is an unspecified subset $\bar{A} \subseteq A$ such that the distributions of the pattern functions π_α with $\alpha \in \bar{A}$ may deviate from their distribution under any $\mathcal{D} \in \Omega_0$. The null hypothesis Ω_0^α for each α , however, is that $\alpha \in A \setminus \bar{A}$. In particular for the transformation invariant Ω_0 , the null hypothesis Ω_0^α says that gene α is among those genes whose expression levels are jointly independent of the label. For these genes α , it holds that $P_{X \sim \mathcal{D}}(X) = P_{X \sim \mathcal{D}}(g(X))$ as under Ω_0 . However, for the genes $\alpha \in \bar{A}$ this is not necessarily the case.

Theorem 3. *Given a pattern class Π , the set $\bar{\Omega}_0$ of distributions as defined above, a fixed $\delta \in (0, 1)$ and a fixed $\hat{p} \in (0, 1)$. Given a uniformly sampled multiset G_m from G of size m . Choose $\sigma_\alpha(X)$ such that*

$$\hat{p} = \frac{\#\{g \in G_m : \exists \alpha \in A : \pi_\alpha(g(X)) \geq \sigma_\alpha(X)\}}{m}.$$

Then with probability at least $1 - \delta$ over m -samples G_m , $\forall \bar{A} \subseteq A$, $\forall \mathcal{D} \in \Omega_0(\bar{A})$,

$$P_{X \sim \mathcal{D}}(\exists \alpha \in A \setminus \bar{A} : \pi_\alpha(X) \geq \sigma_\alpha(X)) \leq \overline{\text{Bin}}(m, \hat{p}m, \delta).$$

Proof. Let us choose $\sigma_\alpha(X)$ such that

$$\hat{p} = \frac{\#\{g \in G_m : \exists \alpha \in A : \pi_\alpha(g(X)) \geq \sigma_\alpha(X)\}}{m}.$$

Clearly, for any \bar{A} ,

$$\#\{g \in G : \exists \alpha \in A : \pi_\alpha(g(X))\} \geq \#\{g \in G : \exists \alpha \in A \setminus \bar{A} : \pi_\alpha(g(X))\}$$

Then, from Lemma 2, it follows that with probability at least $1 - \delta$ over m -samples G_m

$$\frac{\#\{g \in G : \exists \alpha \in A \setminus \bar{A} : \pi_\alpha(g(X)) \geq \sigma(X)\}}{|G|} \leq \overline{\text{Bin}}(m, \hat{p}m, \delta)$$

Hence,

$$\begin{aligned} \overline{\text{Bin}}(m, \hat{p}m, \delta)|G| &\geq \#\{g \in G : \exists \alpha \in A \setminus \bar{A} : \pi_\alpha(g(X)) \geq \sigma(X)\} \\ &\geq \sum_{g \in G} I(\exists \alpha \in A \setminus \bar{A} : \pi_\alpha(g(X)) \geq \sigma(X)) \end{aligned}$$

Taking expectations with respect to any $\mathcal{D} \in \bar{\Omega}_0$ yields

$$\overline{\text{Bin}}(m, \hat{p}m, \delta)|G| \geq |G|P_{X \sim \mathcal{D}}(\exists \alpha \in A \setminus \bar{A} : \pi_\alpha(X) \geq \sigma(X)),$$

from which the result follows. ■

In practice, this suggests the following testing procedure. Choose a value for \hat{p} , a confidence parameter δ , generate a random sample G_m of data transformations, and choose $\sigma(X)$ as in the Theorem. Then, with a probability of at least $1 - \delta$ over the m -sample G_m , the FWER under $\bar{\Omega}_0$ (i.e. the probability to falsely reject at least one hypothesis) is upper bounded by $\overline{\text{Bin}}(m, \hat{p}m, \delta)$. Hence, with a probability of at least $1 - \delta$ over all m -samples G_m , all individual null hypotheses Ω_0^α can be rejected at a confidence of at least $1 - \overline{\text{Bin}}(m, \hat{p}m, \delta)$.

5 Controlling the k -FWER

In the preceding section we have focused on the FWER, the probability that any of the individual hypotheses Ω_0^α is falsely rejected. A practically interesting generalisation of Theorem 3 which is proved in a very similar way, bounds the probability to observe at least k false rejections while $\bar{\Omega}_0$ holds.

Theorem 4. *Given a pattern class Π , the set $\bar{\Omega}_0$ of distributions as defined above, a fixed $\delta \in (0, 1)$, a fixed $\hat{p} \in (0, 1)$, and a fixed $k \in \{1, 2, \dots, |A|\}$. Given a uniformly sampled multiset G_m from G of size m . Choose $\sigma_\alpha(X)$ such that*

$$\hat{p} = \frac{\#\{g \in G_m : \exists A_k \subset A \text{ with } |A_k| = k \text{ and } \forall \alpha \in A_k : \pi_\alpha(g(X)) \geq \sigma_\alpha(X)\}}{m}.$$

Then, with probability at least $1 - \delta$ over m -samples G_m , $\forall \bar{A} \subseteq A, \forall \mathcal{D} \in \Omega_0(\bar{A})$,

$$\begin{aligned} P_{X \sim \mathcal{D}} (\exists A_k \subset A \setminus \bar{A} \text{ with } |A_k| = k \text{ and } \forall \alpha \in A_k : \pi_\alpha(X) \geq \sigma(X)) \\ \leq \overline{\text{Bin}}(m, \hat{p}m, \delta). \end{aligned}$$

Hence, let us consider the multiple hypothesis test which rejects the individual null hypothesis Ω_0^α if

$$\pi_\alpha(X) \geq \sigma(X).$$

Then, with probability at least $1 - \delta$ over m -samples G_m , this multiple hypothesis test has a k -FWER of at most $\overline{\text{Bin}}(m, \hat{p}m, \delta)$. More informally speaking, if it is unlikely to observe k or more false rejections (as quantified by the above Theorem), at most $k - 1$ of the rejections may be regarded as false rejections. This means that we have no choice but considering at least all but $k - 1$ of the rejections as actual positives, though of course we will not be able to distinguish the actual positives from the false rejections.

6 A uniform bound on the k -FWER

We know now how to upper bound the probability to see at least k false rejections given a required k -FWER of p . However, in practical cases, it is more convenient to decide how many false rejections we want to tolerate only after the total number of rejections in the given data is known. Since Theorem 4 only holds for specific k , deciding which k to control the k -FWER needs to be done a priori, before the data has been considered.

Therefore, it is of interest to control the k -FWER uniformly, over a set of k -values $\mathcal{K} \subseteq \{1, 2, \dots, |A|\}$ one considers a priori of potential interest. The following theorem implies a methodology to obtain such a uniform control.

Theorem 5. *Consider a pattern class Π , a null hypothesis $\bar{\Omega}_0$, a fixed $\delta \in (0, 1)$, a fixed $\hat{p} \in (0, 1)$, and a fixed $\mathcal{K} \subseteq \{1, 2, \dots, |A|\}$. Given a uniformly sampled multiset G_m from G of size m . Choose functions $\sigma_\alpha : \mathcal{X} \times \mathcal{K} \rightarrow \mathbb{R}$ non-increasing in k satisfying:*

$$\hat{p} = \frac{\#\{g \in G_m : \exists k \in \mathcal{K}, A_k \subset A \text{ with } (|A_k| = k \text{ and } \forall \alpha \in A_k : \pi_\alpha(g(X)) \geq \sigma_\alpha(X, k))\}}{m}.$$

Then, with probability at least $1 - \delta$ over m -samples G_m , $\forall \bar{A} \subseteq A$, $\forall \mathcal{D} \in \Omega_0(\bar{A})$,

$$P_{X \sim \mathcal{D}} \left(\exists k \in \mathcal{K}, A_k \subset A \setminus \bar{A} \text{ with } (|A_k| = k \text{ and } \forall \alpha \in A_k : \pi_\alpha(X) \geq \sigma(X, k)) \right) \leq \overline{\text{Bin}}(m, \hat{p}m, \delta). \quad (1)$$

This means that with probability at least $1 - \delta$ over the m -sample G_m and simultaneously for all $k \in \mathcal{K}$, the multiple hypothesis test which rejects the individual null hypotheses for $\alpha \in A$ with

$$\pi_\alpha(X) \geq \sigma_\alpha(X, k)$$

has a k -FWER of at most $\overline{\text{Bin}}(m, \hat{p}m, \delta)$.

It remains to be decided which threshold function $\sigma_\alpha(X, \cdot)$ to use. We will discuss a convenient choice for a practical multiple hypothesis test in the empirical results section, where we make the simplification of choosing σ independently of α .

7 The false discovery proportion

For brevity, in the remainder we will assume that G_m is such that Eq. (1) holds, which holds with probability at least $1 - \delta$ over G_m . Then, given the data X and a threshold function $\sigma(X, k)$, we can compute the number of rejections as a function of k :

$$\#\text{rej}(k) = \#\{\alpha \in A : \pi_\alpha(X) \geq \sigma_\alpha(X, k)\}.$$

Now, Theorem 5 provides a uniform upper bound on the k -FWER for $k \in \mathcal{K}$ with \mathcal{K} a given set; hence, it holds for any $k \in \mathcal{K}$ in particular, even if it is selected in a data-dependent way. Thanks to this, and based on the number of rejections $\#\text{rej}(k)$, we can make a number of statements.

At most $k - 1$ of the total number of rejections $\#\text{rej}(k)$ may be considered false rejections, since it is unlikely that for k or more different $\alpha \in A$, the inequality $\pi_\alpha(X) \geq \sigma(k)$ holds under the null hypothesis $\bar{\Omega}_0$ (for any \bar{A}). Hence, the number of true rejections must be at least the $\#\text{rej}(k)$ minus $k - 1$. Note that the number of true rejections can only increase for decreasing $\sigma(X, k)$, i.e. for increasing k . Hence, the number true rejections for a given k can be lower bounded as

$$\#\text{trej}(k) \geq \max_{l \in \mathcal{K}: l \leq k} (\#\{\alpha \in A : \pi_\alpha(X) \geq \sigma(X, l)\} - l + 1) = \underline{\#\text{trej}(k)}.$$

Note that this lower bound $\underline{\#\text{trej}(k)}$ is increasing with k , such that for $k = \max\{k : k \in \mathcal{K}\}$ it gives a lower bound for the total number of true rejections:

$$\#\text{trej} \geq \underline{\#\text{trej}(\max\{k : k \in \mathcal{K}\})} = \max_{l \in \mathcal{K}} (\#\{\alpha \in A : \pi_\alpha(X) \geq \sigma(l)\} - l + 1).$$

Based on the lower bound on $\#\text{trej}(k)$ we can obtain a tighter upper bound on the number of false rejections as

$$\#\text{frej}(k) \leq \#\text{rej}(k) - \underline{\#\text{trej}(k)} = \overline{\#\text{frej}(k)}.$$

We can now bound the false discovery proportion (FDP), namely the expected number of false rejections divided by the total number of rejections. We denote it by FDP:

$$\text{FDP}(k) \leq \frac{\overline{\#\text{frej}(k)}}{\#\text{rej}(k)} = \overline{\text{FDP}(k)}.$$

8 Empirical study

We test our approach on 2 microarray data sets (the Alon data set [1] and the Golub data set [4]), both consisting of a number of microarray experiments on each of two tissue types, the details of which are summarized in Table 1. We are interested in finding out which genes are related to the tissue type, and we use the Wilcoxon rank sum test evaluated on each of the genes as the pattern functions π_α , where α is the gene index. The null hypothesis Ω_0 is that each gene is unrelated to the outcome, such that any permutation of the label vector is equally likely for the given data. The null hypothesis $\bar{\Omega}_0$ acknowledges that a subset of the genes may not be independent of the label vector. Then, witnesses of the invalidity of the null hypothesis are of interest to distinguish between both tissue types. We generated a set of $m = 1000$ random permutations g and associated data sets $g(X)$ by taking the given data set and randomly permuting the labels.

A practical issue is the choice of the threshold or threshold function $\sigma_\alpha(X, \cdot)$, which we choose here to be the same for all α , i.e. $\sigma_\alpha(X, \cdot) = \sigma(X, \cdot)$. Recall that $\sigma(X, \cdot)$ is a (non-increasing) function $\sigma(X, \cdot) : \mathcal{K} \rightarrow \mathbb{R}$. We will choose the function $\sigma(X, \cdot)$ from a specified set of functions $\{\sigma_\tau(X, \cdot)\}$. This set of

Table 1. Summary of the two microarray data sets used.

Source	# genes	# arrays tissue 1	# arrays tissue 2
[1]	2000	40	22
[4]	7129	47	25

functions is in principle free to choose. A convenient choice is by sorting the sets $\{\pi_\alpha(g(X)) : \alpha \in A\}$ for all $g \in G_m$ in decreasing order. Then we define the function value $\sigma_r(X, k)$ with $k \in \mathcal{K}$ as the r th largest value over all g among all k th numbers in the decreasingly sorted $\{\pi_\alpha(g(X)) : \alpha \in A\}$. Hence, r can vary from 1 to m , such that m functions σ_r can be defined.

In the uniform bound, the set of values \mathcal{K} over which the k -FWER should be bounded should be motivated by the maximal number of false positives one is willing to tolerate. However, since the experiments in this paper are artificial (even though based on real data), we consider several such intervals to illustrate the behaviour of the method.

The results are shown in Figures 1 with $\delta = 0.05$. We have chosen \hat{p} such that $\overline{\text{Bin}}(m, \hat{p}m, \delta)$ is equal to 0.05 for each experiment. The plots show the pointwise bound (which holds for each value of k individually), and the uniform bound for three different choices of \mathcal{K} . Clearly, for large values of k , the uniform bound benefits from the fact that it leads to a non-decreasing lower bound on the number of true positives, which makes its lower bound on the number of true positives tighter in this range. However, for small values of k the bound is generally somewhat looser. Obviously, the larger the set of values \mathcal{K} , the more conservative the test for each individual value of k .

Hence, the uniform bound allows one to focus on a zone of interest, while retaining the possibility to choose the desired working point later in a data-dependent way. A better prior view on the region of interest will strengthen the statistical conclusions that can be drawn, and hence the power of the test.

9 Discussion

In the past decade, a vast amount of work has been done on multiple testing, in terms of new definitions that match practice more adequately, in terms of dealing with more general or more specialised situations, and in terms of theoretical improvements of existing ideas. A good recent overview is given in [11]. Here we provide a brief overview of the work that is related or relevant to the current paper.

The FDR has been introduced in [2]. Only later, e.g. in [3], it has been argued that the FDP is more relevant in practice than the FDR. The k -FWER has been put forward as an interesting means for multiple testing by e.g. [7], who also discuss a step-down procedure to control the FDP.

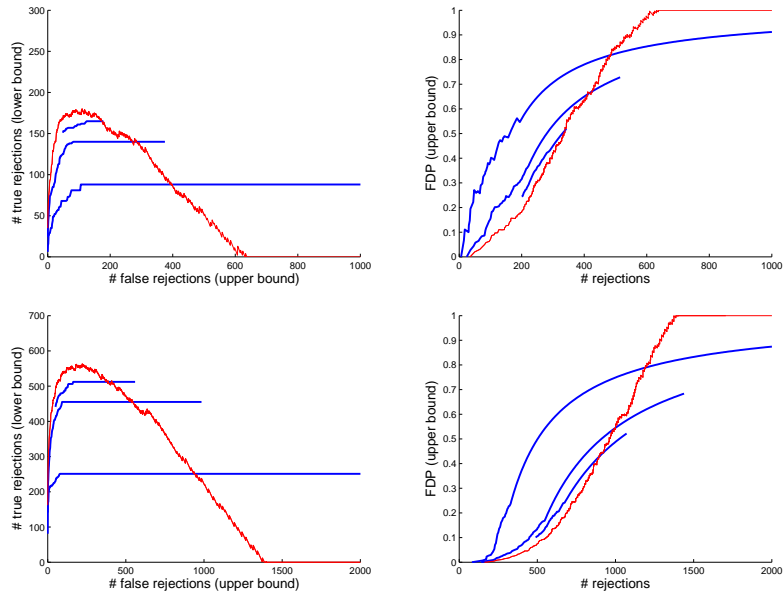


Fig. 1. The performance of our approaches for the Alon data [1] (top) and on the Golub data [4] (bottom) are analysed. On the left hand side the lower bound $\overline{\#trej}(k)$ is shown as a function of $\overline{\#frej}(k)$ for 3 different sets \mathcal{K} : these are the 3 bold curves that increase monotonically. The non-monotonic curve is computed by using the pointwise bound on the k -FWER, and plots the number of positives minus k versus k , for the thresholds required to guarantee the (pointwise) k -FWER. On the right hand side, the $\overline{FDP}(k)$ is plotted as a function of the number of positives for the same 3 choices of \mathcal{K} (in a bold line). Again, additionally the pointwise FDP estimate is compared, by plotting k divided by the number of positives for the thresholds obtained to guarantee the required (pointwise) k -FWER (obviously upper bounded by 1) as a function of the number of positives.

Most work assumes independence of the pattern functions (the test statistics) of the individual tests, or they operate in the worst-case scenario failing to take into account positive dependencies. However, some approaches do allow for positively dependent tests (such as [13], with some additional assumptions), or more generally in [5, 13].

Recent research has considered uniform bounds for the k -FWER, FDR, or the FDP, such as in [3] (assuming independence) or in [12] (valid only asymptotically, also for dependent tests) and notably [9]. The latter is also based on permutation testing, and it is similar to our work in many respects. Therefore we would like to highlight the two main important differences.

First of all, in order to arrive at a significance of a test, the method described in [9] does not directly consider the probability that one of the hypothesis tests fails over X assuming the null hypothesis. Rather they consider the probability *jointly over X and G_m* that one of the tests fails for one of the possible values of k . In contrast, our method separates the randomness in choosing G_m from the randomness in X , such that the statements made are actually about X itself, and not jointly about X and the (in a sense irrelevant) G_m . A result of this difference is that our approach will usually result in stronger statements concerning X if m is larger, and less strong if m is smaller. This means that for small sample size m our approach will be more conservative (and hence less susceptible to atypical S) than the method of [9], which we believe may be appropriate.

This feature is especially relevant if the same G_m is used in several multiple tests: when using [9] an ‘atypical’ G_m may then have unforeseen effects for each of these multiple tests, making them invalid. In contrast, in our approach this risk is explicitly bounded by δ . Note that this situation is irrelevant in the example given in this paper, though it is relevant in other applications. For example, in sequence analysis for bioinformatics, it is of interest to test for the overrepresentation of a sequence pattern in a large number of gene sequences. The null hypothesis is usually a small-order Markov assumption, such that it is possible to generate a randomisation based on G_m under this null hypothesis. However, usually there is not just one such sequence pattern to analyse, but a large number of them. Now, sequence analysis algorithms usually exploit similarities between these sequence patterns to speed up searches by searching for all of them simultaneously. Therefore, for computational reasons, one may be obliged to generate just one random sample S instead of a different one for each of the sequence patterns considered.

Another difference with [9], which is probably even more important in practice, is the fact that our method allows user-defined choices of \mathcal{K} . This simple adjustment of the method makes it much more suitable for practical applications, as a uniform bound over *all* values of k tends to be too conservative, in particular in the more interesting range of small k -values (see Figure 1).

10 Conclusions

We have described a methodology to bound the k -FWER in a multiple hypothesis test, uniformly over a set of values for k . This allows one to select k in a data-dependent way, as guided by this k -FWER or a uniform upper bound on the FDP which is derived from it. Being based on randomisation testing, our approach automatically and effectively takes dependencies between hypotheses into account.

We should point out that there is space for further increase in power of the proposed methods, by means of step-up or step-down procedures. This would be the case in particular for large sets \bar{A} of positives. We leave this further work.

Acknowledgements We are grateful to Nello Cristianini, who brought this problem to our attention, and for other insightful comments and suggestions. TDB acknowledges support from the CoE EF/05/007 SymBioSys, and from GOA/2005/04, both from the Research Council K.U.Leuven. We also acknowledge the support of the EU Network of Excellence, PASCAL under grant ref IST-2002-506778.

References

1. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
2. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Ser. B*, 57:289–300, 1995.
3. C. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *Ann. Statist.*, 32:1035–1061, 2004.
4. D. K. Golub, T. R. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E.S. Lander. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, 96(12):6745–6750, 1999.
5. E. Korn, J. Troendle, L. McShane, and R. Simon. Controlling the number of false discoveries: Application to high-dimensional genomic data. *J. Statist. Plann. Inference*, 124:379–398, 2004.
6. J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.
7. E. L. Lehmann and J. P. Romano. Generalizations of the familywise error rate. *The Annals of Statistics*, 33:1138–1154, 2005.
8. E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer, third edition, 2005.
9. N. Meinshausen. False discovery control for multiple tests of association under general dependence. *Scandinavian Journal of Statistics*, 33:227, 2006.

10. T. Nichols and S. Hayasaka. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, 12:419–446, 2003.
11. A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19:368–375, 2003.
12. J. Storey, J. Taylor, and D. Siegmund. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J. Roy. Statist. Ser. B*, 66:187–205, 2004.
13. D. Yekutieli and Y. Benjamini. Resampling-based false discovery rate controlling test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82:171–196, 1999.