False discovery control for multiple tests of association under general dependence

Nicolai Meinshausen
Seminar für Statistik
ETH Zürich

December 2, 2004

Abstract

We propose a confidence envelope for false discovery control when testing multiple hypotheses of association simultaneously. The method is valid under arbitrary and unknown dependence between the test statistics and allows for an exploratory approach when choosing suitable rejection regions while still retaining strong control over the proportion of false discoveries.

Key Words: False Discovery Proportion, False Discovery Rate, multiple testing
Short Title: False Discovery Control

1 Introduction

Testing of multiple hypotheses has attracted considerable interest recently due to the availability of ever larger data sets. Many fields have adopted the practice of screening over many candidate hypotheses to filter out a few interesting "discoveries". To control the probability of erroneously rejecting true null hypotheses, traditional techniques rely mostly on control of the family-wise error rate (Sidak, 1967; Simes, 1986; Holm, 1979; Westfall & Young, 1993). It is well known that control of the family-wise error rate is very conservative if the number of tests is large. The false discovery rate (FDR) was introduced by Benjamini & Hochberg

(1995) to alleviate this problem by controlling the expected proportion of false discoveries among all rejections.

We follow here the notation by Genovese & Wasserman (2002) and call the *false discovery proportion* FDP the ratio of the number V of true null hypotheses among all rejections R,

$$\mathrm{FDP} = \begin{cases} V/R & R \neq 0 \\ 0 & R = 0 \end{cases}.$$

If rejecting all hypotheses with p-value in the region $[0, t]$, we write $\mathrm{FDP} = \mathrm{FDP}(t)$ to indicate the dependence of the proportion of false discoveries on the chosen rejection region. The notion of $\mathrm{FDP}(t), t \in [0, 1]$, as a stochastic process is discussed in Genovese & Wasserman (2004). For details of the notation see Genovese & Wasserman (2002). The false discovery rate is defined in Benjamini & Hochberg (1995) as the expectation of this quantity,

$$\mathrm{FDR} = E(\mathrm{FDP}).$$

It was argued in Genovese & Wasserman (2002) and Genovese & Wasserman (2004) that instead of controlling the expectation of FDP, it is of considerable more interest to control the random variable FDP itself, as we are really concerned about the number of false rejections in the given experiment (and not in an average of FDP for hypothetical replications of the experiment). See as well Korn et al. (2004) or the original technical report Korn et al. (2001). In Genovese & Wasserman (2004), *confidence envelopes* for FDP are developed. Confidence envelopes $\overline{\mathrm{FDP}}(t)$ are random functions such that at any chosen level $\alpha > 0$,

$$P(\mathrm{FDP}(t) \leq \overline{\mathrm{FDP}}(t) \text{ for all } t \in [0, 1]) \;\geq\; 1 - \alpha. \tag{1}$$

A confidence envelope allows to choose the rejection region adaptively while still retaining control over the number of false discoveries. A confidence envelope for the proportion of false discoveries is equivalent to a confidence envelope for the number of false rejections.

Even though test statistics are strongly dependent in most applications, e.g. microarray gene expression data analysis, most work on FDR uses the assumption of independence. The confidence envelopes in Genovese & Wasserman (2004) are no exception. Some success in broadening the applicability of FDR-controlling

procedures to the case of dependent test statistics has been made in Storey & Tibshirani (2001); Benjamini & Yekutieli (2001); Storey et al. (2004).

In this paper, we consider multiple tests of association as they occur e.g. in microarray gene expression experiments. The structure of the data allows for permutation-based testing. In this setting, a method to control the FDP is proposed in Korn et al. (2004) which takes the dependence between test statistics into account. It does not, however, provide a confidence envelope for FDP. A simultaneous bound in the sense of (1) is useful as it allows the user to choose the appropriate rejection region in an exploratory fashion while still retaining strong control over the proportion of false discoveries. We show that confidence envelopes like (1) can be constructed for multiple tests of association under arbitrary and unknown dependence between the test statistics.

2 A simultaneous upper bound for the proportion of false discoveries

Let $X = (X_1, X_2, \ldots, X_m)$ be a m-dimensional random vector and Y a real-valued response variable. We are interested in finding components of X which are associated with the response variable Y. In microarray gene expression experiments, the components of X might e.g. correspond to expression values of m genes and Y to the some clinical variable. We note that the following approach is valid as well for designed experiments, where a non-random class variable y takes the place of Y.

Let $\mathcal{N} \subseteq \{1, \ldots, m\}$ be a set of components of X, which are jointly independent of the response variable,

$$Y \perp \{X_k, k \in \mathcal{N}\}. \tag{2}$$

The set $\mathcal{N}$ is potentially not uniquely defined, e.g. there might be several blocks of components of X, which are jointly independent of Y. In this case, let $\mathcal{N}$ be any set of interest which fulfills (2) . The null hypothesis $H_{0,k}$ for each component $k = 1, \ldots, m$ is that k belongs to the set $\mathcal{N}$. The alternative $H_{1,k}$ is true if $k \notin \mathcal{N}$.

Given a sample of size n of independent realizations of (Y, X), one can conduct tests of association individually for every component X_k, $k = 1, \ldots, m$. The

outcome of this multiple test procedure is a m-dimensional vector P of p-values. Denote the p-values under the original sample of (X, Y) by $P_1, P_2, \ldots$ and let $P_{(1)}, P_{(2)}, \ldots$ in the following be the ordered version such that $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$. We assume for the following that rejection of some hypothesis entails that all hypotheses with identical or more significant test results are rejected as well. P-values are allowed to be tied but we require that hypotheses that belong to tied p-values are either all rejected or neither of them is rejected. The total number of rejections $R(t)$ for a rejection region $[0, t]$ is

$$R(t) = \#\{k \in \{1, \ldots, m\} : P_k \leq t\},$$

while the number $V(t)$ of false rejections and $S(t)$ of correct rejections, with $R(t) = V(t) + S(t)$, are given by

$$\begin{aligned} V(t) &= \#\{k \in \mathcal{N} : P_k \leq t\}, \\ S(t) &= \#\{k \in \{1, \ldots, m\} \setminus \mathcal{N} : P_k \leq t\}. \end{aligned}$$

The false discovery proportion $\mathrm{FDP}(t)$ for a rejection region $[0, t]$ is then

$$\mathrm{FDP}(t) = \frac{V(t)}{\max\{R(t), 1\}}$$

We aim to bound $\mathrm{FDP}(t)$ from above by a random function $\overline{\mathrm{FDP}}(t)$ such that (1) is fulfilled. The bound relies essentially on a lower simultaneous bound $\underline{S}(t)$ for the number of true discoveries $S(t)$. Given a bound $\underline{S}(t)$ with

$$P(S(t) \geq \underline{S}(t) \text{ for all } t \in [0, 1]) \;\geq\; 1 - \alpha,$$

a simultaneous bound for the number of false discoveries is given by $\overline{V}(t) = R(t) - \underline{S}(t)$. A bound for the false discovery proportion which fulfills (1) is then $\overline{\mathrm{FDP}}(t) = \overline{V}(t)/\max\{R(t), 1\}$. All three bounds $\underline{S}(t), \overline{V}(t)$ and $\overline{\mathrm{FDP}}(t)$ for the total number of true discoveries, false discoveries and the proportion of false discoveries are completely equivalent. Given one, the other two can be calculated immediately.

2.1 Algorithm

The confidence envelope $\underline{S}(t)$ for $S(t)$ is obtained in three steps. Note that the algorithm works as well with raw test statistics without explicit computation of

p-values. Basing the approach on test statistics rather than on p-values is in particular useful if computation of p-values involves a computationally intensive step, as e.g. for the exact computation of p-values under the Wilcoxon-test. For simplicity of notation we chose to work with p-values exclusively.

Step 1: Compute p-values for all hypotheses under permuted samples of the response variable.
Let $\Pi = \{\pi_1, \pi_2, \ldots, \pi_w\}$ be a set of w random permutations of $\{1, 2, \ldots, n\}$, where every permutation has probability $1/n!$. For every permutation $\pi \in \Pi$, consider the permuted version of Y. (For computational efficiency we might focus on the set of permutations that lead to distinguishable outcomes if applied to Y, drawing each permutation independently from this set with appropriately reweighted probabilities.) For this permuted version and the original values of X, p-values for all m hypotheses are calculated. For each permutation, a m-dimensional vector $P^\pi = (P_1^\pi, P_2^\pi, \ldots, P_m^\pi)$ of p-values is obtained. The result for all permutations can be arranged in a $w \times m$ matrix of p-values,

$$\begin{pmatrix} P_1^{\pi_1} & P_2^{\pi_1} & \ldots & P_m^{\pi_1} \\ P_1^{\pi_2} & P_2^{\pi_2} & \ldots & P_m^{\pi_2} \\ \ldots & & & \end{pmatrix} \tag{3}$$

Step 2: Ordering of the p-values.
The first column of the previous matrix contains the p-values of the first hypothesis under random permutations of the class labels, while the second column contains the p-values of the second hypothesis under the same permutations. For each column, these p-values are now permuted randomly. Consecutively, for each row, the p-values are put into increasing order, so that the first row contains the ordered p-values of all m hypotheses under a random permutation of the class labels (where the permutations can be different for the various hypotheses). Next, the values within each column are put into increasing order and the result is written, for each column $1 \leq k \leq m$ as an w-dimensional vector $(Q_k^1, Q_k^2, \ldots, Q_k^w)$ with

$$Q_k^1 \leq Q_k^2 \leq \ldots \leq Q_k^w \qquad \forall 1 \leq k \leq m,$$

where w denotes again the cardinality of the set of permutations. The result of the second step is again an $w \times m$ matrix, given by

$$\begin{pmatrix} Q_1^1 & Q_2^1 & \dots & Q_m^1 \\ Q_1^2 & Q_2^2 & \dots & Q_m^2 \\ \dots & & & \end{pmatrix} \tag{4}$$

We denote the row vectors of this matrix by $Q^\ell = (Q_1^\ell, Q_2^\ell, \dots, Q_m^\ell)$ for $\ell = 1, \dots, w$. By construction, it holds that the smallest p-value from each permutation is larger than or equal to Q_1^1, while the second smallest p-value from each permutation is larger than or equal to Q_2^1 and so on. The first row Q^1 gives thus a natural bound for the empirical distributions of p-values under permutation of the response variable. The bound is, however, rather conservative. A less conservative bound is found in Step 3 by stepping through the rows $Q^1, Q^2, \dots, Q^w$ and selecting -as a bound for the empirical distribution of p-values under permutations of the response variable- the largest row which fulfills a suitable criterion.

Step 3: Compare with permutation matrix from Step 1.
For each row Q^ℓ, $\ell = 1, \dots, w$ of matrix (4), calculate the quantity $\beta(\ell)$ as the proportion of permutations for which the ordered p-values are bounded from below by Q^ℓ,

$$\beta(\ell) = \frac{|\{\pi \in \Pi : Q^\ell \leq P^\pi\}|}{w}, \tag{5}$$

where the inequality $Q^\ell \leq P^\pi$ is fulfilled if it is fulfilled for all m components, that is $Q^\ell \leq P^\pi \Leftrightarrow Q_k^\ell \leq P_{(k)}^\pi$, $\forall k = 1, \dots, m$, and $(P_{(1)}^\pi, P_{(2)}^\pi, \dots, P_{(m)}^\pi)$ is the ordered version of $(P_1^\pi, P_2^\pi, \dots, P_m^\pi)$. Find the largest row index $\ell(\alpha)$ such that $\beta(\ell)$ is still larger than $1 - \alpha$,

$$\ell(\alpha) = \max\{\ell \in \{1, \dots, w\} : \beta(\ell) \geq 1 - \alpha\}. \tag{6}$$

Note that it holds in general that $1 \leq \ell(\alpha) \leq (1 - \alpha)w$. Define for all $t \in [0, 1]$,

$$B(t) := \#\{k \in \{1, \dots, m\} : Q_k^{\ell(\alpha)} \leq t\}. \tag{7}$$

This function corresponds to the quantile bounding functions in Meinshausen & Bühlmann (2005). The lower bound for the number of true discoveries is defined as

$$\underline{S}(t) := \max_{0 \leq \tau \leq t} R(\tau) - B(\tau). \tag{8}$$

Upper bounds for the number of false discoveries and the false discovery proportion are then given by $\overline{V}(t) := R(t) - \underline{S}(t)$ and $\overline{\mathrm{FDP}}(t) := \overline{V}(t)/\max\{R(t), 1\}$ respectively.

2.2 Simultaneous probabilistic bound

We verify that $\overline{\mathrm{FDP}}(t) = \overline{V}(t)/\max\{R(t), 1\}$ is indeed a simultaneous upper probabilistic bound for $\mathrm{FDP}(t) = V(t)/\max\{R(t), 1\}$ in the sense of (1). For this, it is sufficient to show that $\overline{V}(t)$ is an upper bound for the number of false discoveries.

Theorem 1 *Let $\underline{S}(t)$ be defined as in (8) for some $\alpha > 0$ and $\overline{V}(t) = R(t) - \underline{S}(t)$. Under arbitrary distribution of (Y, X),*

$$\begin{aligned} P(S(t) \geq \underline{S}(t) \text{ for all } t \in [0,1]) &\geq 1-\alpha, \\ P(V(t) \leq \overline{V}(t) \text{ for all } t \in [0,1]) &\geq 1-\alpha, \end{aligned}$$

Proof of Theorem 1. Let $\tilde{P}$ be the vector with components

$$\tilde{P}_k = \begin{cases} P_k & k \in \mathcal{N} \\ 1 & k \in \mathcal{N}^c \end{cases} \qquad \forall\, k = 1 \ldots, m.$$

The ordered vector is denoted by $(\tilde{P}_{(1)}, \tilde{P}_{(2)}, \ldots, \tilde{P}_{(m)})$ so that $\tilde{P}_{(1)} \leq \tilde{P}_{(2)} \leq \ldots \leq \tilde{P}_{(m)}$. The number of false rejections, for a rejection region $[0, t]$, is given by $V(t) = \#\{k \in \mathcal{N} : P_k \leq t\}$. Hence, for $t < 1$,

$$V(t) = \#\{k \in \{1, \ldots, m\} : \tilde{P}_{(k)} \leq t\} \tag{9}$$

If it holds that

$$P(\tilde{P}_{(k)} \geq Q_k^{\ell(\alpha)} \text{ for all } k \in \{1, \ldots, m\}) \;\geq\; 1 - \alpha, \tag{10}$$

then, by comparing (9) with (7), it follows that

$$P(V(t) \leq B(t) \text{ for all } t) \;\geq\; 1 - \alpha. \tag{11}$$

Thus $B(t)$ is an upper simultaneous bound for $V(t)$ if (10) holds true. Continuing with this assumption for the moment, the number of true discoveries $S(t) = \#\{k \in$

$\{1, \ldots, m\} \setminus \mathcal{N} : P_k \leq t\}$ can be written as the total number of rejections less the number of false discoveries, $S(t) = R(t) - V(t)$. Note that any realization of $S(t)$ is monotonously increasing in t. Hence

$$S(t) = \max_{0 \leq \tau \leq t} S(\tau) = \max_{0 \leq \tau \leq t} R(\tau) - V(\tau).$$

If (10) is indeed true, then, using (11), it holds with probability at least $1 - \alpha$ that

$$S(t) \geq \max_{0 \leq \tau \leq t} R(\tau) - B(\tau) \quad \text{for all } t \in [0, 1],$$

and the result follows.

It hence remains to show (10). Consider the statistic T that includes the sample of $\{X_k, k \in \mathcal{N}\}$ and the ordered sample of Y. Conditional on T, $\tilde{P}$ and $\tilde{P}^\pi$ have the same multivariate distribution for a random permutation $\pi \in \Pi$, where $\tilde{P}^\pi$ is the vector of p-values $\tilde{P}$ after a random permutation π of the observed response variable Y. Let $\tilde{\ell}(\alpha)$ be defined as $\ell(\alpha)$ in (6), yet using p-values $\tilde{P}$ instead of P in the definition of β in (5). It holds that $\tilde{\ell}(\alpha) \geq \ell(\alpha)$ and hence $Q_k^{\tilde{\ell}(\alpha)} \geq Q_k^{\ell(\alpha)}$ for all k. Thus, for any T,

$$P(\tilde{P}_{(k)} \geq Q_k^{\ell(\alpha)} \text{ for all } k \in \{1, \ldots, m\} | T) \geq P(\tilde{P}^\pi_{(k)} \geq Q_k^{\tilde{\ell}(\alpha)} \text{ for all } k \in \{1, \ldots, m\} | T). \tag{12}$$

By definition of $\tilde{\ell}(\alpha)$, $P(\tilde{P}^\pi_{(k)} \geq Q_k^{\tilde{\ell}(\alpha)} \text{ for all } k \in \{1, \ldots, m\}) \geq 1 - \alpha$. Integrating over T and using (12), equation (10) follows. This completes the proof.

3 Properties

The algorithm discussed above is well suited for implementation. However, it does not provide good intuition about the rationale behind it. We discuss a few properties of the proposed bounds and discuss relation to previous work. Hopefully this sheds a little light on how and why the algorithm works.

3.1 Simultaneous bound

The main property of the estimate, and the motivation for the work, is that the bound (for either the total number of false discoveries or its proportion) is valid simultaneously for all possible rejection regions under arbitrary dependence

between the test statistics. The latter condition requires a special structure of the data. The algorithm is e.g. not going to work if one is just given a list of possibly dependent p-values. This is what distinguishes our estimate from the one proposed in Genovese & Wasserman (2004). Somewhat related to our method, a bound for the number of false rejections is proposed in Korn et al. (2004). Their false discovery control works like this: one specifies an absolute number u of false rejections one is willing to make. Based on this choice, the algorithm returns a number of rejections. The true number of false rejections is then larger than u only with small probability α. The same works as well when specifying proportions instead of absolute numbers of acceptable false discoveries. However, the estimate lacks the property of simultaneous control and is not a confidence envelope in the sense of property (1). Choosing the number of rejections adaptively invalidates in particular the conservative property of the bound in Korn et al.. As with the confidence envelope proposed in this paper, the estimate of Korn et al. utilizes permutation of the response variables. It will be seen in the numerical examples that the bounds are rather close numerically. This is no coincidence. Indeed, the number of false discoveries $V(t)$ is conservatively estimated in Korn et al. essentially by

$$\tilde{B}(t) = \#\{l \in \{1, \ldots, m\} : Q_k^{(1-\alpha)w} \leq t\}. \tag{13}$$

This is almost the bound $B(t)$ in (7), except for the fact that $(1-\alpha)w$ has been replaced by $k(\alpha) \leq (1-\alpha)w$. Note that $\tilde{B}(t)$ offers point-wise control in the sense that, for any chosen $t \in [0,1]$,

$$P(V(t) \leq \tilde{B}(t)) \ \geq \ 1-\alpha.$$

Instead of providing point-wise control at level α like $\tilde{B}(t)$, the bound $B(t)$ offers point-wise control at some level smaller than α. Thus it holds in general that $\tilde{B}(t) \leq B(t)$. The level at which $B(t)$ offers point-wise control is chosen such that $B(t)$ is a simultaneous bound at level α,

$$P(V(t) \leq B(t) \text{ for all } t \in [0,1]) \ \geq \ 1-\alpha.$$

However, we do not bound the number of false discoveries directly by $B(t)$. By inspection of the algorithm, it can be seen that the number of false discoveries is instead bounded by

$$\overline{V}(t) = R(t) - \max_{0 \leq \tau \leq t} (R(\tau) - B(\tau)),$$

which is smaller than or equal to $B(t)$. It still produces a valid simultaneous upper bound for $V(t)$ due to the simultaneous nature of the bound $B(t)$ and monotonicity of $S(t)$. Using $\overline{V}(t)$ is hence less conservative than $B(t)$. Indeed, it can be less conservative than the point-wise bound $\tilde{B}(t)$. Consider the limit $t \to 1$. The total number of rejections $R(t)$ converges in this setting to m. Both $\tilde{B}(t)$ and $B(t)$ converge as well to m irrespective of the level α. The upper bound $\overline{V}(t)$, in contrast, converges in general to some smaller value, giving an upper bound for the number of true null hypotheses, as explained below.

In conclusion, building a confidence envelope is only slightly more conservative than using point-wise control (sometimes even less conservative), while offering the possibility of choosing the rejection region in an exploratory fashion and still retaining strong control over the proportion of false discoveries.

3.2 Monotonicity

Note that the number of true discoveries $S(t)$ is monotonously increasing with t. When rejecting more hypotheses, the total number of correct rejections cannot decrease. By definition (8) of $\underline{S}(t)$,

$$\underline{S}(t) = \max_{0 \leq \tau \leq t} R(\tau) - B(\tau)$$

it is clear that the proposed lower bound is as well monotonously increasing with t. Interpretation would be somewhat difficult for a non-monotone bound: consider the statement that there are at least 5 true discoveries among the first 10 hypotheses, but only 4 true discoveries among the first 20? Clearly, given the first statement, we expect to find at least 5 true discoveries among the first 20 rejections. Monotonicity is hence a desirable feature of any lower bound for the number of true discoveries. Note that even when only requiring point-wise control, the method in Korn et al. (2004) cannot produce a monotonic lower bound for the number of true discoveries.

3.3 Total number of false null hypotheses

In some applications it might be of interest to just estimate the total number of false null hypotheses m_1 (Liang et al., 2002; Turkheimer et al., 2001). By

definition of S, it holds that $m_1 = S(1)$. The function $\underline{S}(t)$ leads thus directly to a lower probabilistic bound for m_1, by setting $\hat{m}_1 = \underline{S}(1)$. It follows from the above properties that indeed

$$P(m_1 \geq \hat{m}_1) \;\geq\; 1 - \alpha. \tag{14}$$

Note that no parameters have to be tuned to obtain the estimate $\hat{m}_1$ and it gives a natural estimate of the total amount of true alternative hypotheses. Property (14) is valid under arbitrary dependence between the test statistics. This has already been established in Meinshausen & Bühlmann (2005). The estimate $\hat{m}_1$ is in particular equivalent to the estimate in Meinshausen & Bühlmann (2005) under the quantile bounding function. In Meinshausen & Rice (2006), the asymptotic properties of a similar estimate are discussed for the case of independent test statistics. These papers are, however, just concerned with estimating the total number of false null hypotheses. It was only established later that similar techniques can be used to give confidence envelopes for the proportion of false discoveries, as shown in this paper.

4 Numerical Examples

4.1 Simulation study

We study the properties of the proposed bound for a few numerical examples. To reflect the dimensions of microarray gene expression datasets considered later, the number of hypotheses is chosen as $m = 1000$. The response variable Y has a Bernoulli distribution with $p = 0.5$. The predictor variable X is normally distributed $X \sim \mathcal{N}(\mu, \Sigma)$ where the covariance matrix Σ is given by $\Sigma_{ii} = 1$ and $\Sigma_{ij} = \rho$ for $i \neq j$ and some $0 \leq \rho \leq 1$. The mean is given by $\mu_i = 0$ if $Y_i = 0$ or $i \leq 600$ and $\mu_i = 1$ otherwise. Hence there are 600 true null hypotheses. Test statistics are calculated under a two-sided Wilcoxon test. 500 randomly chosen permutations of the response variable are used for the construction of the lower bound $\underline{S}$. The level is always chosen in the following as $\alpha = 0.05$. The lower bound $\underline{S}(t)$ is plotted as a function of the true number $S(t)$ of true discoveries in Figure 1. The more observations are made, the closer the lower bound $\underline{S}$ is to S. The bound is particularly tight if only a few hypotheses are rejected. For

the dependent test statistics the variance of false discoveries is larger, at least for fixed rejection regions. The bound has to accommodate this and hence tends to be more conservative.

[Figure 1 about here]

Next, we examine the probability of

$$\exists t \in [0,1] : \overline{\mathrm{FDP}}(t) < \mathrm{FDP}(t), \tag{15}$$

which is the same event as $\exists t \in [0,1] : \underline{S}(t) > S(t)$, e.g. there is at least one $t \in [0,1]$ for which the lower bound for the number of true discoveries is larger than the actual number of true discoveries. The probability of these events is limited by construction by the level α. The setup is exactly the same as above. However, we do not only use $m_1 = 400$ but test the procedure as well with a considerably smaller amount of false null hypotheses, $m_1 = 10$. The results are shown in Table 1 as an average over 500 simulations.

[Table 1 about here]

It can be seen that the bound is in general rather tight. The true level at which one is controlling seems to be in general above 3% when using $\alpha = 5\%$. The bound is unduly conservative only for independent test statistics and many false null hypotheses ($m_1 = 400, \rho = 0$). For $\rho = 0.4$ and $m_1 = 10$, the proportion of false discoveries is underestimated with the proposed bound in 5.4% of all simulations. This is above the 5% level but does not represent a significant deviation.

To illustrate the usefulness of a simultaneous bound, we consider again the above setup with $n = 60$, $\rho = 0$, $m = 1000$, and $m_1 = 10$ false null hypotheses. For a given threshold of b permitted false rejections (where b is chosen as either 5, 10, or 50), as many rejections as possible are made, once with the proposed method and once with the method of Korn et al. (2004), both at nominal level 5%. For the simultaneous bound and 500 simulations, the average number of rejected hypotheses is 8.36, 13.78, 54.02 for $b = 5, 10$, and 50 respectively. Of theses rejections, 7.19, 8.70, and 9.78 are on average correctly rejected false null hypotheses (there are $m_1 = 10$ false null hypotheses in total). The probability of rejecting more than b hypotheses for any $b \in \{5, 10, 50\}$ is 0.2% in the simulation study and

hence bounded by the nominal level 5%, as expected from the simultaneous nature of the bound. For the method of Korn et al., the average number of rejections is 10.46, 16.01, and 57.05, of which 8.04, 8.93, and 9.79 are on average correctly rejected null hypotheses. The average number of correctly rejected hypotheses is thus slightly larger than for the simultaneous bound and the probability of rejecting more than b hypotheses is bounded by 5% for each b separately. However, the probability that the number of false rejected null hypotheses is larger than b for some $b \in \{5, 10, 50\}$ cannot be controlled with this method and is indeed 12.6% in the simulation study, larger than the nominal level 5%.

4.2 Microarray gene expression data

We illustrate the method for three microarray gene expression data sets. In a microarray gene expression experiment the expression levels of thousands of genes (about 5000) are measured simultaneously. The number of observations is on the other hand typically small, in the region of about 50. Specifically, the datasets of Singh et al. (2002) about prostate cancer (102 observations, 6033 genes), Golub et al. (1999) about leukemia (72 observations, 3571 genes), and Alon et al. (1999) about colon cancer (62 observations, 2000 genes) are considered. The three datasets contain a binary response variable $Y \in \{0, 1\}$ which is typically the subtype of a specific cancer under consideration. We note that one might rather tend to consider these studies as designed experiments and think of a binary non-random class variable y in this context. The method is unaffected by this choice, however. Gene $k = 1, \ldots, m$ is called differentially expressed if the expression values X_k of gene k and the response variable Y are not independent. Here we consider the more specific case of over-expression, that is we test for each gene with the one-sided Wilcoxon-test if its mean expression level is higher for class 1 than for class 0.

[Figure 2 about here]

We compute the lower bound $\underline{S}(t)$ for the number of true discoveries as in the previous section. The result is shown in Figure 2. The lower bound bound has in all three experiments a distinctive shape, first growing almost as fast as $R(t)$ and then levelling off slowly. While the optimal number of rejections will in general

depend on the gain from making a true discovery versus the loss of making a false discovery, it does not seem worthwhile to reject hypotheses that fall into the "flat" region of the bound $\underline{S}(t)$.

Finally, we compare the bound $\overline{\mathrm{FDP}}(t)$ to the bound implied by the method in Korn et al. (2004). The results are shown in Figure 3 for the three datasets. The proposed bound $\overline{\mathrm{FDP}}(t)$ tends to be more conservative for smaller rejection regions. This is intuitively clear as the proposed bound holds simultaneously for all possible number of rejections. For larger rejection regions, the situation is reverse and the proposed bound is tighter than the point-wise bound. For reasons explained above, the proposed bound does not converge to 1 as $t \to 1$. Instead, it gives for $t = 1$ an upper bound for the total proportion of true null hypotheses. For all three datasets, the bound implies e.g. roughly that there are at least 10% of genes for which expression levels are higher for the specified class. This relative proportion corresponds of course to the highest level that $\underline{S}(t)$ reaches, compare Figure 2. The important feature of the proposed method is, in our view, that the bound for the number of true discoveries holds simultaneously for all rejection regions and allows thus for an explorative approach when choosing a rejection region without giving up strict control over the proportion of false discoveries.

[Figure 3 about here]

5 Acknowledgements

I would like to thank Andreas Buja, John Rice, Hans-Rudolf Künsch and my supervisor Peter Bühlmann for interesting discussions and helpful comments on an earlier version of the paper. The comments by an anonymous referee and an associate editor are also gratefully acknowledged.

References

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

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Ser.* B **57**, 289–300.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–1188.

Genovese, C. & Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. Roy. Statist. Ser.* B **64**, 499–517.

Genovese, C. & Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32**, 1035–1061.

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

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

Korn, E., Troendle, J., McShane, L. & Simon, R. (2001). Controlling the number of false discoveries: Application to high-dimensional genomic data. Tech. Rep. 3, Biometric Research Branch, National Cancer Institute, Bethesda.

Korn, E., Troendle, J., McShane, L. & Simon, R. (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *J. Statist. Plann. Inference* **124**, 379–398.

Liang, C.-L., Rice, J., de Pater, I., Alcock, C., Axelrod, T., Wang, A. & Marshall, S. (2002). Statistical methods for detecting stellar occultations by kuiper belt objects: the taiwanese-american occultation survey. *Statist. Sci.* **19**, 265–274.

Meinshausen, N. & Bühlmann, P. (2005). Lower bounds for the number of true null hypotheses for multiple testing of associations under general dependence structures. *Biometrika* **92**, to appear.

Meinshausen, N. & Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* **32**, to appear.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.* **84**, 608–610.

Simes, R. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 608–610.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T. & Sellers, W. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203–209.

Storey, J., Taylor, J. & Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J. Roy. Statist. Ser.* B **66**, 187–205.

Storey, J. & Tibshirani, R. (2001). Estimating false discovery rates under dependence, with application to DNA microarrays. Tech. rep., Department of Statistics, University of California, Berkeley.

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

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

Nicolai Meinshausen
Seminar für Statistik, Leonhardstrasse 27
ETH Zentrum, 8092 Zürich, Switzerland
nicolai@stat.math.ethz.ch

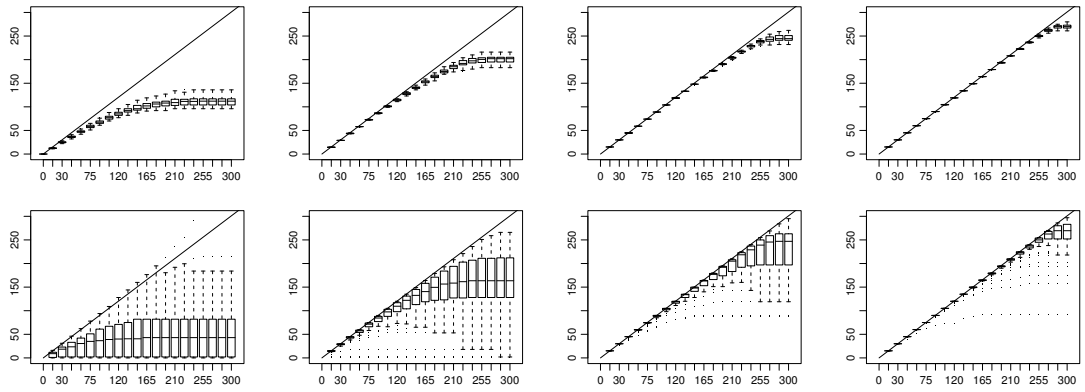

Figure 1: Boxplots of the lower bound $\underline{S}(t)$ as a function of the true value $S(t)$ for 100 simulations and $t \in [0, 1]$. The continuous line corresponds to $\underline{S}(t) = S(t)$. The number n of observations in each sample is increasing from $n = 20$ (left column) to $n = 80$ (right column) in steps of 20. The dependence between the test statistics is $\rho = 0$ (upper row) and $\rho = 0.4$ (lower row) respectively.

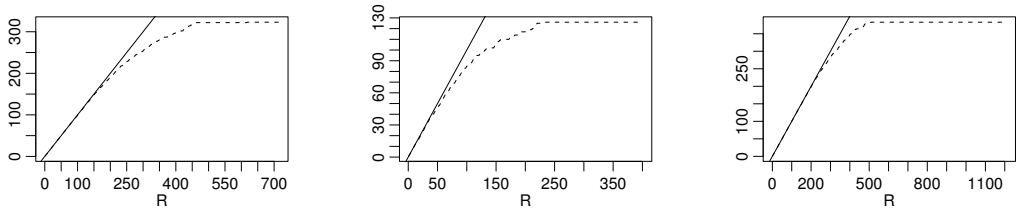

Figure 2: The lower bound $\underline{S}(t)$ as a function of the number $R(t)$ of rejections for (from left to right) leukemia, colon, and prostate cancer data, broken line. The unbroken line corresponds to the total number of rejections $R(t)$.

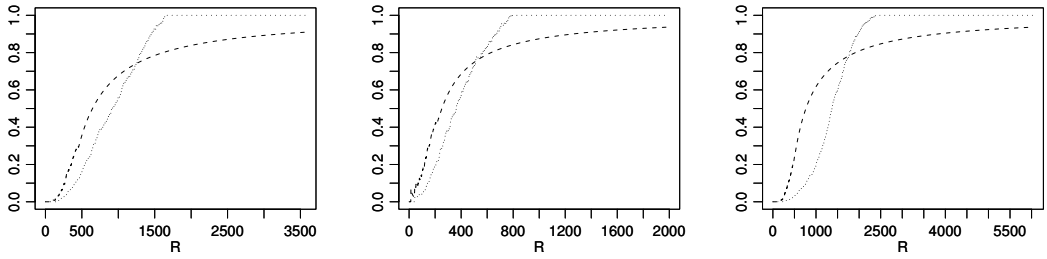

Figure 3: For the same data as in Figure 2, the simultaneous bound $\overline{\mathrm{FDP}}(t)$ as a function of the number of rejections $R(t)$, broken line, and the pointwise bound, dotted line.

Table 1: The probability (in %) of underestimating the true proportion of false discoveries (using $\alpha = 5\%$) for some $t \in [0, 1]$.

	$m_1 = 400$			$m_1 = 10$		
	$n = 20$	$n = 60$	$n = 100$	$n = 20$	$n = 60$	$n = 100$
$\rho = 0$	1.6	1.4	0.8	3.0	4.2	4.8
$\rho = 0.2$	3.2	3.8	3.0	3.8	3.6	4.6
$\rho = 0.4$	5.0	4.8	4.4	4.4	5.4	4.0