False discovery proportion estimation by permutations: confidence for significance analysis of microarrays

Hemerik, J.; Goeman, J.J.

False discovery proportion estimation by permutations: confidence for SAM

Jesse Hemerik*[1] and Jelle J. Goeman[2]

[1,2] Leiden University Medical Center, The Netherlands

November 29, 2017

Summary

SAM ("Significance Analysis of Microarrays") is a highly popular permutation-based multiple testing method that estimates the false discovery proportion (FDP), the fraction of false positives among all rejected hypotheses. Perhaps surprisingly, until now this method had no known properties. This paper extends SAM by providing $(1-\alpha)$-confidence upper bounds for the FDP, so that exact confidence statements can be made. As a special case, an estimate of the FDP is obtained that underestimates the FDP with probability at most 0.5. Moreover, using a closed testing procedure, this paper decreases the upper bounds and estimates in such a way that the confidence level is maintained. We base our methods on a general result on exact testing with random permutations.

Keywords: Confidence, False Discovery Proportion, Multiple Testing, Permutation

*Address for correspondance: Jesse Hemerik, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Postzone S5-P, Postbus 9600, 2300 RC Leiden, The Netherlands.
E-mail: j.b.a.hemerik@lumc.nl

1 Introduction

When multiple hypotheses are tested, interest is often in estimating the false discovery proportion (FDP), the number of false positives divided by the total number of rejections. When there are no rejections, the FDP is defined to be zero. When there is unknown dependence in the data, a challenge is to find methods that are powerful but also require few assumptions on the dependence structure (van der Laan et al., 2004; Meinshausen, 2006; Genovese and Wasserman, 2006). A highly popular method for estimation of the FDP is Significance Analysis of Microarrays (SAM) (Tusher et al., 2001). SAM is a very general method that is not at all limited to microarray data. It requires no parametric assumpions and almost no assumptions on the dependence structure in the data. Instead, it adapts to the dependence stucture by using permutations. Consequently Tusher et al. (2001) have been cited more than 10,000 times.

The SAM procedure in Tusher et al. is based on two ideas. The first is estimation of the FDP based on permutations of the data. The second idea is a specific choice of the test statistic, involving a small fudge factor. In this paper, focus is on the first idea. We do not consider a specific type of test statistics, but allow any test statistics to be used.

The rationale of SAM is the following. SAM rejects all hypotheses with test statistics lying in the user-defined rejection region. The number of false positives is estimated by considering permuted versions of the data. The median of the numbers of rejections for the permuted versions of the data is the estimate of the number of false positives. This value divided by the number of rejections is an estimate of the FDP. No properties of the estimate have been proven (although Dudoit et al. (2002, 2003) note that SAM estimates the Per-Family Error Rate). Until now SAM has been a very sensible, but quite heuristic method.

Based on Storey (2002; 2004) an adaptation of SAM was suggested to decrease the estimate. It is based on the idea that when there are relatively many false hypotheses, the original SAM method tends to overestimate the number of false positives. The reason is that the many false hypotheses cannot lead to false positives and SAM did not take this into account. Hence it was suggested to multiply the basic SAM estimate by an estimate $\widehat{\pi}_0$ of the fraction of true hypotheses π_0. Usually this leads to a lower estimate of the FDP. This multiplication by the plug-in $\widehat{\pi}_0$ has been implemented in the *samr* *R*-package, the main software for SAM (Chu et al., 2001). Like the original SAM method, this newer procedure has no known properties.

The first aim of this paper is to construct a $(1-\alpha)$-confidence upper

bound for the FDP for $\alpha \in [0, 1)$. That is, we extend SAM by providing a confidence interval around the estimate of the FDP. Thus, for small α, we obtain a high-confidence upper bound for the FDP. For $\alpha = 0.5$, we obtain an estimate of the FDP, which underestimates the FDP with probability at most 0.5. This estimate coincides with the estimate of the *samr* package without multiplication by $\widehat{\pi}_0$. Our $(1-\alpha)$-upper bound is the $(1-\alpha)$-quantile of the permutation distribution of the number of rejections. It was inspired by a work by Meinshausen (2006), who provides upper bounds for the FDP that are uniformly valid over multiple rejection regions.

The further contributions of this paper are procedures that decrease the basic $(1 - \alpha)$-bound, in such a way that the exact properties are maintained. In particular, for $\alpha = 0.5$ the estimate is improved. These uniform improvements do not require additional assumptions. As with the plug-in method based on $\widehat{\pi}_0$, the gain is largest when there are relatively many false hypotheses. The improvements are derived using a result by Goeman and Solari (2011), who provide uniform FDP bounds by using a closed testing procedure. Our derivation reveals surprising connections between SAM and closed testing.

All our methods are based on a general result on exact testing with randomly sampled permutations, which extends the work of Phipson and Smyth (2010). This result also allows proving properties of other existing methods based on random permutations. All methods in this paper have been implemented in the *R* package *confSAM*.

This paper is built up as follows. In Section 2.1 the basic $(1 - \alpha)$-bound for the FDP is discussed. A closed-testing based improvement of this method is presented in Section 3, including a fast approximation of this improvement. In Section 4 a conservative shortcut is constructed for the method from Section 3. The proposed methods are applied to simulated data in in Section 5. Section 6 contains an analysis of real data.

2 Basic upper bound

In this section the basic $(1 - \alpha)$-bound for the FDP is discussed.

2.1 Setting and notation

Througout the paper, we consider the following setting. Let X be data, taking values in a sample space $\mathcal{X}$. Consider hypotheses $H_1, ..., H_m$ and test statistics $T_i : \mathcal{X} \to \mathbb{R}$, $1 \leq i \leq m$. For each $1 \leq i \leq m$, let $D_i \subseteq \mathbb{R}$ be a

rejection region associated with hypothesis H_i and test statistic T_i. That is, H_i is rejected if and only if $T_i(X) \in D_i$, so that

$$\mathcal{R} = \{1 \leq i \leq m : T_i(X) \in D_i\}$$

is the set of indices of rejected hypotheses. We simply call $\mathcal{R}$ the set of rejected hypotheses. We write $\mathcal{R}^c = \{1, ..., m\} \setminus \mathcal{R}$. Let

$$\mathcal{N} = \{1 \leq i \leq m : H_i \text{ is true}\}$$

be the set of true hypotheses. Let

$$N = \#\mathcal{N}, \quad R = \#\mathcal{R}$$

and

$$V = \#\mathcal{N} \cap \mathcal{R},$$

the number of false positives. Since sets and numbers such as $\mathcal{R}$, R and V depend on the data, we denote them as functions on $\mathcal{X}$. For example, for $x \in \mathcal{X}$, $\mathcal{R}(x)$ denotes the set of rejected hypotheses for data x. The set $\mathcal{N}$ does not depend on the data, since the hypotheses are fixed. Thus $V(x) = \#(\mathcal{R}(x) \cap \mathcal{N})$. When no argument is written, this means that the argument is X. For example, R is short for $R(X)$. The false discovery proportion is

$$FDP = \frac{V}{R}$$

if $R > 0$ and 0 otherwise.

All methods in this paper are based on permutations or other transformations of the data. Let G be a set of transformations $g : \mathcal{X} \to \mathcal{X}$, such that G is a group under composition of maps. Throughout the paper we use the word 'group' in the strict algebraic sense, rather than loosely in the meaning of 'set' as is often done in the statistical literature. We write $g(x)$ as gx. In practice G is often a group of permutation maps. Sometimes other groups of transformations will be used, such as rotations (Langsrud, 2005; Solari et al., 2014) and multiplication of part of the data by -1 (Pesarin and Salmaso (2010), pp. 54 and 168).

The following assumption, made throughout this paper, underlies many permutation-based multiple testing methods, e.g. Westfall and Young's max$T$ method (1993), Meinshausen and Bühlmann (2005) and Meinshausen (2006).

Assumption 1. The joint distribution of the test statistics $T_i(gX)$ with $i \in \mathcal{N}$, $g \in G$, is invariant under all transformations in G of X.

In applications, an argument needs to be given for this distributional assumption. As a mathematical example where the assumption is satisfied, consider a basic randomized trial where H_i implies that the distribution of the expression level of gene i is the same for cases and controls. Typically each T_i only depends on the expression levels measured for gene i. Then Assumption 1 is satisfied if the multivariate distribution of the expression levels corresponding to $\mathcal{N}$ is the same for cases and controls.

It is allowed to define $\mathcal{N}$ simply as the largest set of hypotheses for which Assumption 1 is satisfied, as in Meinshausen and Bühlmann (2005). This is a less usual definition of $\mathcal{N}$, but Assumption 1 is then guaranteed to hold.

Throughout the paper, random transformations from G are used. The vector of random transformations is defined as follows.

Definition 2. Let G' be the vector $(id, g_2, ..., g_w)$, where id is the identity in G and $g_2, ..., g_w$ are random elements from G. Write $g_1 = id$. The random transformations can be drawn either with or without replacement: the statements in this paper hold for both cases. In the latter case, $w \leq \#G$. If we draw $g_2, ...g_w$ without replacement, then we take them to be uniformly distributed on $G \setminus \{id\}$, otherwise uniform on G.

For $\mathcal{I} \subseteq \{1, ..., m\}$ and $x \in \mathcal{X}$, write

$$R_{\mathcal{I}}(x) = \#\mathcal{R}(x) \cap \mathcal{I}.$$

Let

$$R_{\mathcal{I}}^{(1)} \leq ... \leq R_{\mathcal{I}}^{(w)}$$

be the sorted values $R_{\mathcal{I}}(g_j X)$, $1 \leq j \leq w$. We have $R_{\{1,...,m\}} = R$, so write $R^{(j)} := R_{\{1,...,m\}}^{(j)}$, $1 \leq j \leq w$.

Throughout the paper, $\alpha \in [0, 1)$ and $k = \lceil (1 - \alpha)w \rceil$, the smallest integer at least as large as $(1 - \alpha)w$. The minimum of two numbers a and b is denoted by $a \wedge b$.

2.2 Upper bound and median unbiased estimate

Here the upper bound and estimate for the FDP are constructed. We first prove the permutation principle that underlies our methods. It is known that the permutation test is exact when the set of transformations (e.g. permutations) has a group structure (Hoeffding, 1952). For example, the set of all possible permutation maps is a group. In recent decades, permutation methods have become popular. Often random permutations are used,

to limit the computation time. Usually a p-value based on random permutations is seen as an estimate of the true permutation p-value. However, it is also possible to compute an exact p-value based on random permutations, if they are suitably sampled from a group. Phipson and Smyth (2010) provide formulas for exact p-values based on random permutations under some assumptions. Their results imply that to obtain a valid test, the original observation should be included with the random permutations. However, they ignore the role of the group structure, which is fundamental to permutation methods. Moreover, it has not been clear how results on testing with random permutations generalise to other permutation methods (such as SAM and Meinshausen, 2006). We now state a general result on testing with random transformations. This result can be used to prove properties of various permutation-based multiple testing methods. We will illustrate this in Theorem 4, where we apply Theorem 3 in the SAM context.

Theorem 3. *Let $S : \mathcal{X} \to \mathbb{R}$ be a test statistic. Let $S^{(1)}(X, G') \leq ... \leq S^{(w)}(X, G')$ be the ordered test statistics $S(g_j X)$, $1 \leq j \leq w$.*

Consider a null hypothesis H_0 which implies that the joint distribution of the test statistics $S(gX)$, $g \in G$, is invariant under all transformations in G of X. Then under H_0, $\mathbb{P}(S(X, G') > S^{(k)}(X, G')) \leq \alpha$.

Proof. From the group structure of G, it follows that for all $1 \leq j \leq w$, $G' g_j^{-1}$ and G' have the same distribution, if we disregard the order of the elements. Let j have the uniform distribution on $\{1, ..., w\}$ and write $h = g_j$. Under H_0,

$$\begin{aligned}
&\mathbb{P}\big\{S(X) > S^{(k)}(X, G')\big\} = \\
&\mathbb{P}\big\{S(X) > S^{(k)}(X, G'h^{-1})\big\} = \\
&\mathbb{P}\big\{S(hX) > S^{(k)}(hX, G'h^{-1})\big\}.
\end{aligned}$$

Since $(G'h^{-1})(hX) = G'(h^{-1}hX)$, the above equals

$$\begin{aligned}
&\mathbb{P}\big\{S(hX) > S^{(k)}(h^{-1}hX, G')\big\} = \\
&\mathbb{P}\big\{S(hX) > S^{(k)}(X, G')\big\}
\end{aligned}$$

Since $h = g_j$ with j uniform, this equals

$$\mathbb{E}\Big[w^{-1}\#\big\{1 \leq j \leq w : S^{(j)}(X, G') > S^{(k)}(X, G')\big\}\Big] \leq \alpha,$$

as was to be shown. □

The value $R^{(k)}$ is the $(1-\alpha)$-quantile of the numbers of rejections for the permuted versions of the data. The following theorem states that this simple quantile is a $(1-\alpha)$-upper bound for the number of false positives V.

Theorem 4. *The number $\overline{V} := R^{(k)} \wedge R$ is a $(1-\alpha)$-upper bound for V, i.e.*

$$\mathbb{P}\big(V \leq \overline{V}\big) \geq 1-\alpha.$$

Proof. Let

$$V^{(1)} \leq \ldots \leq V^{(w)}$$

be the sorted values $V(g_j X) = \#\big(\mathcal{R}(g_j X) \cap \mathcal{N}\big)$, $1 \leq j \leq w$. With Theorem 3 it follows that

$$\mathbb{P}\big(V > V^{(k)}\big) \leq \alpha.$$

Since $V^{(k)} \leq R^{(k)}$,

$$\mathbb{P}\big(V \leq R^{(k)}\big) \geq 1-\alpha$$

and the result follows. □

Note that $V \leq \overline{V}$ holds if and only if $V/R \leq \overline{V}/R$, provided $R > 0$. Thus $\overline{V}/R$, which is interpreted as 0 when $R = 0$, is a $(1-\alpha)$-upper bound for the FDP. Note that taking $\alpha = 0.5$ in the above theorem provides an estimate $\overline{V}$ of V with the property that $\mathbb{P}\big(V \leq \overline{V}\big) \geq 0.5$. We will call such an estimate *median unbiased*, in line with the existing notion of a median unbiased estimator of a parameter.

By assumption, the dependence structure of the test statistics $T_i(X)$, $i \in \mathcal{N}$, is maintained by their permutation distribution. The quantile $\overline{V}$ is based on the permutation distribution of the number of rejections R, which is based on the permutation distribution of the test statistics. Hence $\overline{V}$ is adapted to the joint distribution of the test statistics $T_i(X)$, $i \in \mathcal{N}$. Therefore the method does not need to take into account a worst-case scenario for their dependence structure. Thus, by using permutations, relatively tight bounds for the FDP tend to be obtained.

2.3 Choice of rejection regions

The rejection regions D_i can be freely chosen, provided that they are not based on the data. When they are based on the data, this may introduce some selection bias, especially when the regions are cherry-picked in such a way that the number of rejections is large compared the estimate of V.

When the rejection regions D_i do not depend on the data, it is not always possible to choose sensible rejection regions when little is known about the distribution of the test statistics. We can use permutation p-values as test statistics however, such that we can always choose a sensible rejection region, for instance $(0, 0.01]$. This leads to about $0.01N$ false positives on average. In the setting of SAM, such p-values based on permutations (or other transformations) can nearly always be calculated. These p-values can be based on random permutations, as in Theorem 3. Moreover, it is allowed to base all m p-values on a single collection of random permutations (independent from $g_1, ..., g_w$), since the test statistics are essentially assumption-free. Note that permutation p-values are never smaller than one divided by the number of permutations. Thus, when the rejection region is $(0, c)$, more than c^{-1} permutations should be used.

When the cutoff c is very small, the number of random permutations needed is very large, which may make using permutation p-values time-consuming. A possible practical solution is the following. Often the tail of the permutation distribution of each T_i can be modeled by e.g. a generalized Pareto distribution (Knijnenburg et al., 2009). In that case, draw a small number of random permutations, compute the corresponding values of the test statistic T_i and fit such a distribution to these values. Then use $D_i = (q_i, \infty)$ as the rejection region for T_i, where q_i is the $(1 - c)$-quantile of the distribution determined for T_i. Note that this means that the rejection regions are data-dependent. However, since they depend on permutations of the data and not on cherry-picking the regions that give the strongest results, the selection bias tends to be very limited or absent. Since this paper focuses on proving exact properties however, we will keep the assumption that the rejection regions D_i are fully independent.

3 Closed testing for improved bounds

Especially when there are many false hypotheses, the basic bound $\overline{V}$ does not exhaust α. The reason is that $\overline{V}$ then tends to be substantially larger than the bound $V^{(k)}$, as can be seen from their definitions. The bound $V^{(k)}$ cannot be computed in practice but has been shown to be a $(1 - \alpha)$-upper bound in the proof of Theorem 4. The closed testing principle (Marcus et al., 1976) is a powerful method for familywise error rate control. Goeman and Solari (2011) show how closed testing can be used to obtain upper bounds for the FDP. By relating SAM to closed testing, we will be able to derive a potentially smaller upper bound for V than the basic bound $\overline{V}$. The

improved bound is still valid with probability $1-\alpha$.

3.1 Closed testing

We recall the closed testing principle and how it can be used to obtain uniform upper bounds for the FDP. For each nonempty $\mathcal{I} \subseteq \{1,...,m\}$, denote by $H_{\mathcal{I}}$ the intersection hypothesis $\bigcap_{i\in\mathcal{I}} H_i$, the hypotheses that all hypotheses H_i, $i \in I$, are true. Suppose that for each nonempty $\mathcal{I} \subseteq \{1,...,m\}$ an α-level test for $H_{\mathcal{I}}$ is defined. These $2^m - 1$ tests are called local tests. The closed testing procedure rejects all $H_{\mathcal{I}}$ for which all $H_{\mathcal{J}}$ with $\mathcal{I} \subseteq \mathcal{J} \subseteq \{1,...,m\}$ are rejected by their local tests. By Marcus et al. (1976), the probability that the closed testing procedure rejects at least one true intersection hypothesis is at most α. Thus the procedure strongly controls the familywise error rate at level α.

The FDP upper bounds are derived as follows. Let

$$\mathcal{C} = \{\mathcal{I} \subseteq \{1,...,m\} : \quad H_{\mathcal{I}} \text{ is rejected by the closed testing procedure}\}.$$

For each $\mathcal{K} \subseteq \{1,...,m\}$ define

$$\overline{V}_{\mathrm{ct}}(\mathcal{K}) = \max\{\#\mathcal{I} :\ \mathcal{I} \subseteq \mathcal{K},\ \mathcal{I} \notin \mathcal{C}\},$$

where the maximum is defined to be zero if the set is empty. By Goeman and Solari (2011) the following holds.

Theorem 5. *Uniformly over all $\mathcal{K} \subseteq \{1,...,m\}$, $\overline{V}_{ct}(\mathcal{K})$ is a $(1-\alpha)$-upper bound for $\#\mathcal{K} \cap \mathcal{N}$, i.e.*

$$P\left[\bigcap_{\mathcal{K}\subseteq\{1,...,m\}} \{\#\mathcal{K} \cap \mathcal{N} \leq \overline{V}_{ct}(\mathcal{K})\}\right] \geq 1-\alpha.$$

The proof is in Goeman and Solari (2011). Note that $\#\mathcal{K} \cap \mathcal{N}$ is the number of false positives if $\mathcal{K}$ is the rejected set. Thus the theorem provides bounds for the numbers of false positives that are uniform over all possible rejected sets. Thus, if the rejected set is chosen based on the data, then the corresponding upper bound is still valid with probability at least $1-\alpha$.

A closed testing procedure depends on its local tests. For different local tests, different closed testing procedures are obtained. The more power the closed testing procedure has, the lower the resulting FDP bound tends to be. One particular closed testing procedure leads directly to the basic bound $\overline{V}$. The reader can check that this is the closed testing procedure based on the local tests that reject $H_{\mathcal{I}}$ if and only if $R_{\mathcal{I}} > R^{(k)}$. In Section 3.2 a more powerful closed testing procedure is considered, which allows improvement of the basic bound $\overline{V}$.

3.2 Improved FDP bounds

To obtain an improvement of the bound $\overline{V}$ for the number of false positives, we use a more sophisticated closed testing procedure. For each nonempty $\mathcal{I} \subseteq \{1, ..., m\}$ consider the local test that rejects $H_{\mathcal{I}}$ if and only if $R_{\mathcal{I}} > R_{\mathcal{I}}^{(k)}$. (See the notation defined in Section 2.1.) This test has level at most α by theorem 3. Throughout the rest of this paper, let $\overline{V}_{\text{ct}}(\mathcal{K})$ refer to the closed testing procedure based on these local tests. Note that $\overline{V}_{\text{ct}}(\mathcal{R})$ provides an upper bound for $V = \#\mathcal{R} \cap \mathcal{N}$, the number of true hypotheses in $\mathcal{R}$. We write $\overline{V}_{\text{ct}} := \overline{V}_{\text{ct}}(\mathcal{R})$.

The bound $\overline{V}_{\text{ct}}(\mathcal{R})$ is ideal in the sense that no smaller $(1-\alpha)$-bound for V is given in this paper or elsewhere in the literature, under our assumptions. In practice however, it is often computationally infeasible to calculate this bound without the use of shortcuts. Indeed, when naively computing $\overline{V}_{\text{ct}}$, a huge number of local tests needs to performed unless m is small. This section is devoted to deriving an exact shortcut for calculating $\overline{V}_{\text{ct}}$. In Section 4 a conservative shortcut will be derived, i.e. an efficient method for finding an upper bound for $\overline{V}_{\text{ct}}$.

The following lemma offers a shortcut for determining whether $H_{\mathcal{I}}$ is rejected by our closed testing procedure.

Lemma 6. *For $\mathcal{I} \subseteq \{1, ..., m\}$, $\mathcal{I} \in \mathcal{C}$ if and only if $R_{\mathcal{I}} > R_{\mathcal{I} \cup \mathcal{R}^c}^{(k)}$.*

Proof. To prove the first implication, note that

$$\begin{aligned}
&\mathcal{I} \in \mathcal{C} \Leftrightarrow \\
&\text{For all } \mathcal{I} \subseteq \mathcal{J} \subseteq \{1, ..., m\},\ R_{\mathcal{J}} > R_{\mathcal{J}}^{(k)} \Rightarrow \\
&R_{\mathcal{I} \cup \mathcal{R}^c} > R_{\mathcal{I} \cup \mathcal{R}^c}^{(k)} \Leftrightarrow \\
&R_{\mathcal{I}} > R_{\mathcal{I} \cup \mathcal{R}^c}^{(k)}.
\end{aligned}$$

For the reverse implication, suppose

$$\#\mathcal{I} \cap \mathcal{R} > R_{\mathcal{I} \cup \mathcal{R}^c}^{(k)} \tag{1}$$

and let $\mathcal{I} \subseteq \mathcal{J} \subseteq \{1, ..., m\}$. Then

$$R_{\mathcal{J}} = \#\mathcal{I} \cap \mathcal{R} + \#(\mathcal{J} \setminus \mathcal{I}) \cap \mathcal{R} \tag{2}$$

and, because obviously $\#\mathcal{A} \geq R_{\mathcal{A} \cup \mathcal{B}}^{(k)} - R_{\mathcal{B}}^{(k)}$ for $\mathcal{A}, \mathcal{B} \subseteq \{1, ..., m\}$,

$$\#(\mathcal{J} \setminus \mathcal{I}) \cap \mathcal{R} \geq R_{((\mathcal{J} \setminus \mathcal{I}) \cap \mathcal{R}) \cup (\mathcal{I} \cup \mathcal{R}^c)}^{(k)} - R_{\mathcal{I} \cup \mathcal{R}^c}^{(k)} \geq R_{\mathcal{J}}^{(k)} - R_{\mathcal{I} \cup \mathcal{R}^c}^{(k)}. \tag{3}$$

Combining (1), (2) and (3) yields

$$R_{\mathcal{J}} > R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c} + R^{(k)}_{\mathcal{J}} - R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c} = R^{(k)}_{\mathcal{J}}.$$

Thus all $H_{\mathcal{J}}$ with $\mathcal{I} \subseteq \mathcal{J} \subseteq \{1, ..., m\}$ are rejected by their local tests, so that $\mathcal{I} \in \mathcal{C}$. □

Due to this shortcut, te determine whether $\mathcal{I} \in \mathcal{C}$ it is not necessary to perform all local tests for the hypotheses $H_{\mathcal{J}}$ with $\mathcal{I} \subseteq \mathcal{J}$. Instead, it suffices to check if $R_{\mathcal{I}} > R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c}$.

Using this fact and additional observations, the following exact shortcut is obtained for determining $\overline{V}_{\mathrm{ct}}$.

Proposition 7. *The bound* $\overline{V}_{ct}$ *equals*

$$R \wedge \left[\min \left\{ 1 \leq M \leq R : \text{ for all } \mathcal{I} \subseteq \mathcal{R} \text{ with } \#\mathcal{I} = M,\ M > R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c} \right\} - 1 \right]. \tag{4}$$

Proof. By Lemma 6, $\overline{V}_{\mathrm{ct}}(\mathcal{R}) =$

$$\max\{\#\mathcal{I} : \mathcal{I} \subseteq \mathcal{R}, R_{\mathcal{I}} \leq R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c}\} =$$
$$\max\{\#\mathcal{I} : \mathcal{I} \subseteq \mathcal{R}, \#\mathcal{I} \leq R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c}\} =$$
$$R \wedge \left[\min \left\{ 1 \leq M \leq R : \text{ for all } \mathcal{I} \subseteq \mathcal{R} \text{ with } \#\mathcal{I} \geq M,\ \#\mathcal{I} > R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c} \right\} - 1 \right].$$

For any $\mathcal{I} \subseteq \mathcal{R}$, if $\#\mathcal{I} > R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c}$ then for all $\mathcal{I} \subseteq \mathcal{J} \subseteq \mathcal{R}$, $\#\mathcal{J} > R^{(k)}_{\mathcal{J}\cup\mathcal{R}^c}$. Hence the above equals (4). □

Using Proposition 7, $\overline{V}_{\mathrm{ct}}$ can be calculated much faster than by naive computation based on the definition of $\overline{V}_{\mathrm{ct}}$. When R or $\overline{V}_{\mathrm{ct}}$ is large however, calculating $\overline{V}_{\mathrm{ct}}$ is often still infeasible; the computation time is roughly proportional to $\binom{R}{\overline{V}_{\mathrm{ct}}+1}$ if $\overline{V}_{\mathrm{ct}} < R$. Hence in Section 3.3 a method for approximating $\overline{V}_{\mathrm{ct}}$ is defined. Moreover, in Section 4 a conservative shortcut is derived, which calculates an upper bound to $\overline{V}_{\mathrm{ct}}$ in relatively little time. The performance of these methods is illustrated with simulations in Sections 5.5 and 5.6.

3.3 Approximation method

We propose a method for approximating the bound $\overline{V}_{\mathrm{ct}}$, for cases where computing $\overline{V}_{\mathrm{ct}}$ is infeasible. Proposition 7 states that

$$\overline{V}_{\mathrm{ct}} = R \wedge \left[\min \left\{ 1 \leq M \leq R : \ M > \mu(M) \right\} - 1 \right], \tag{5}$$

where

$$\mu(M) := \max\{ R^{(k)}_{\mathcal{I} \cup \mathcal{R}^c} : \quad \mathcal{I} \subseteq \mathcal{R} \text{ and } \#\mathcal{I} = M \}.$$

Determining $\mu(M)$ can be computationally infeasible, since the number of subsets $\mathcal{I} \subseteq \mathcal{R}$ with $\#\mathcal{I} = M$ can be huge. Hence we propose to draw a big number of random sets $\mathcal{I} \subseteq \mathcal{R}$ with $\#\mathcal{I} = M$ and calculate the maximum for this collection of subsets. That is, we consider an approximation

$$\mu^*(M) = \max\{ R^{(k)}_{\mathcal{I} \cup \mathcal{R}^c} : \quad \mathcal{I} \in \mathcal{S} \},$$

where $\mathcal{S}$ is some large random subcollection of $\{ \mathcal{I} \subseteq \mathcal{R} : \ \#\mathcal{I} = M \}$. This leads to the estimate

$$\overline{V}^*_{\mathrm{ct}} = R \wedge \left[\min \left\{ 1 \leq M \leq R : \ M > \mu^*(M) \right\} - 1 \right]. \tag{6}$$

Note that $\mu^*(M) \leq \mu(M)$ and consequently $\overline{V}^*_{\mathrm{ct}} \leq \overline{V}_{\mathrm{ct}}$. Hence $\overline{V}^*_{\mathrm{ct}}$ is not guaranteed to be a $(1-\alpha)$-confidence upper bound. However, in many cases, including the simulation scenarios considered in Section 5.6 , $\overline{V}^*_{\mathrm{ct}}$ is still a $(1-\alpha)$-confidence bound for V. The reason is that $\overline{V}^*_{\mathrm{ct}}$ converges to $\overline{V}_{\mathrm{ct}}$ for $\#\mathcal{S} \to \infty$. For details see Section 5.6.

4 Conservative shortcut

Here we will construct a conservative shortcut for the closed testing-based method that provides $\overline{V}_{\mathrm{ct}}$. The shortcut is much more computationally efficient and can often be used when there are thousands of rejections. The upper bound $\overline{V}_{\mathrm{sc}}$ that the shortcut provides is always less than or equal to the basic bound $\overline{V}$. On the other hand, it only improves $\overline{V}$ in specific settings, some of which are considered in Section 5.5. The bound $\overline{V}_{\mathrm{sc}}$ is often larger than $\overline{V}_{\mathrm{ct}}$. It is never smaller than $\overline{V}_{\mathrm{ct}}$, which guarantees its validity as a $(1-\alpha)$-confidence bound. Thus the following ordering holds:

$$\overline{V} \geq \overline{V}_{\mathrm{sc}} \geq \overline{V}_{\mathrm{ct}} \geq \overline{V}^*_{\mathrm{ct}}.$$

The bound $\overline{V}_{\text{ct}}$ depends on $\mu(M)$, the computation of which can be computationally infeasible. Lemma 8 provides an upper bound $U(M)$ for this maximum which can often be computed in a limited amount of time even when there are thousands of rejections. This will lead to the conservative shortcut for the full closed testing-based method.

Note that the upper bounds $\overline{V}$ and $\overline{V}_{\text{ct}}$ are functions of $\mathcal{R}(g_1X), ..., \mathcal{R}(g_wX)$, the rejected sets for the transformed versions of the data. Reindex the transformations such that $R(g_1X) \leq ... \leq R(g_wX)$. Hence $R(g_jX) = R^{(j)}$ for all $1 \leq j \leq w$. The collection of rejected sets can be represented by a binary $m \times w$-matrix $\mathbf{M}$, where

$$\mathbf{M}_{(i,j)} = \begin{cases} 1 & \text{if } i \in \mathcal{R}(g_jX), \\ 0 & \text{otherwise.} \end{cases}$$

Since the upper bound $\overline{V}_{\text{ct}}$ can be viewed as a function of this matrix, the problem of finding shortcuts for the closed testing-based method is essentially a combinatorial one.

To have an intuitive understanding of Lemma 8 below, it is useful to view the quantities considered in it as functions of the matrix $\mathbf{M}$. For each $1 \leq j \leq w$, we define

$$S_j = R_{\mathcal{R}}(g_jX).$$

Note that this is the sum of the elements of $\mathbf{M}$ that are both in the j-th column and in the rows corresponding to the rejected set $\mathcal{R} = \mathcal{R}(X)$. Next, define for each $i \in \mathcal{R}$

$$\Sigma_i = \sum_{1 \leq j \leq w} R_{\{i\}}(g_jX).$$

This is simply the sum of the elements of the i-th row of $\mathbf{M}$. Let $\Sigma_{(1)} \leq ... \leq \Sigma_{(R)}$ be the sorted values Σ_i and for $1 \leq M \leq R$ let

$$\Sigma = \Sigma(M) := \Sigma_{(1)} + ... + \Sigma_{(R-M)}.$$

We now state the main result on which the conservative shortcut is based.

Lemma 8. *For each $s \in \mathbb{N}$ define*

$$N_s = \#\{1 \leq j < k : \ R^{(j)} < R^{(k)} - s\},$$
$$M_s = k - 1 - N_s.$$

For each $s \in \mathbb{N}$ and $N_s < j \leq w$, let

$$K_j^s = \max\left\{0, S_j - (R^{(j)} - R^{(k)} + s)\right\}$$

and let $K^s_{(1)} \geq ... \geq K^s_{(w-N_s)}$ *be these values sorted from large to small.*

For $1 \leq M \leq R$ *let*

$$U(M) = R^{(k)} - 1 - \max A,$$

where

$$A = \left\{ s \in \mathbb{N} : \Sigma(M) > \sum_{1 \leq j \leq N_s} S_j + \sum_{N_s < j \leq w} \min\{S_j, R^{(j)} - R^{(k)} + s\} + \sum_{1 \leq j \leq M_s} K^s_{(j)} \right\}.$$

Then

$$U(M) \geq \mu(M).$$

Proof. Let $1 \leq M \leq R$. Write

$$\mathcal{B} = \left\{ \mathcal{R}^{\mathsf{c}} \subseteq \mathcal{I} \subseteq \{1, ..., m\} : \ \#\mathcal{I} = \#\mathcal{R}^{\mathsf{c}} + M \right\}.$$

Note that

$$\mu(M) = \max\{R^{(k)}_{\mathcal{I}} : \mathcal{I} \in \mathcal{B}\}.$$

The proof consists of three parts. In part 1 we show that $0 \in A$ implies $R^{(k)} > \mu(M)$. In part 2 we note that for all $s \in \mathbb{N}$, $s \in A$ implies $R^{(k)} - s > \mu(M)$. We then conclude that by definition $U(M) \geq \mu(M)$.

–Part 1.

Suppose $0 \in A$. Consider any $\mathcal{I} \in \mathcal{B}$. Write $\mathcal{I}^{\mathsf{c}} = \{1, ..., m\} \setminus \mathcal{I}$. Note that $\#\mathcal{I}^{\mathsf{c}} = R - M$. Hence, by choice of Σ,

$$\sum_{1 \leq j \leq w} R_{\mathcal{I}^{\mathsf{c}}}(g_j X) \geq \Sigma. \tag{7}$$

Since $0 \in A$,

$$\sum_{1 \leq j \leq w} R_{\mathcal{I}^{\mathsf{c}}}(g_j X) > \sum_{1 \leq j \leq N_0} S_j + \sum_{N_0 < j \leq w} \min\{S_j, R^{(j)} - R^{(k)}\} + \sum_{1 \leq j \leq M_0} K^0_{(j)}. \tag{8}$$

First note that

$$\sum_{1 \leq j \leq N_0} R_{\mathcal{I}^{\mathsf{c}}}(g_j X) \leq \sum_{1 \leq j \leq N_0} S_j,$$

since $R_{\mathcal{I}^{\mathsf{c}}}(g_j X) \leq S_j$ for all j. Hence with (8) it follows that

$$\sum_{N_0 < j \leq w} R_{\mathcal{I}^{\mathsf{c}}}(g_j X) > \sum_{N_0 < j \leq w} \min\{S_j, R^{(j)} - R^{(k)}\} + \sum_{1 \leq j \leq M_0} K^0_{(j)}. \tag{9}$$

Suppose that $R_{\mathcal{I}}^{(k)} = R^{(k)}$. This implies that

$$\#\big\{N_0 < j \le w : R_{\mathcal{I}}(g_j X) < R^{(k)}\big\} \le M_0, \tag{10}$$

and equivalently

$$\#\big\{N_0 < j \le w : R_{\mathcal{I}}(g_j X) \ge R^{(k)}\big\} > (w - N_0) - M_0. \tag{11}$$

For the indices j in the set at (10),

$$R_{\mathcal{I}^c}(g_j X) \le S_j.$$

Moreover, for the indices j in the set at (11),

$$R_{\mathcal{I}^c}(g_j X) = R^{(j)} - R_{\mathcal{I}}(g_j X) \le \min\{S_j, R^{(j)} - R^{(k)}\}.$$

These observations imply that $\sum_{N_0<j\le w} R_{\mathcal{I}^c}(g_j X)$ is at most the right side of (9), which contradicts (9). Hence $R^{(k)} \neq R_{\mathcal{I}}^{(k)}$, i.e. $R^{(k)} > R_{\mathcal{I}}^{(k)}$. Since this holds for all $\mathcal{I} \in \mathcal{B}$, by definition $R^{(k)} > \mu(M)$.
–Part 2.
Consider any $\mathcal{I} \in B$. Let $s \in \mathbb{N}$. In part 1 we supposed that $0 \in A$; we now more generally suppose that $s \in A$. It follows like in part 1 that $R^{(k)} - s > R_{\mathcal{I}}^{(k)}$ and consequently $R^{(k)} - s > \mu(M)$.
–Part 3.
Since this holds for all $s \in A$, we have $R^{(k)} - \max A > \mu(M)$, i.e.

$$R^{(k)} - 1 - \max A = U(M) \ge \mu(M).$$

□

By (5), Lemma 8 and the fact that $\overline{V}_{\mathrm{ct}} \le \overline{V}$,

$$\overline{V}_{\mathrm{sc}} := \overline{V} \wedge \left[\min\Big\{1 \le M \le R : \ M > U(M)\Big\} - 1\right]$$

is an upper bound for $\overline{V}_{\mathrm{ct}}$. Recall that the calculation of $\overline{V}_{\mathrm{sc}}$ is usually feasible when there are many thousands of rejections, but this shortcut only provides an improvement over $\overline{V}$ in some situations with many false hypotheses, as is illustrated with simulations in Section 5.5.

5 Simulations

We investigate the performance of the discussed methods on simple simulated data. In sections 5.2 and 5.3 variants of the basic SAM method are investigated as upper bounds (for $\alpha = 0.05$) and as estimates (for $\alpha = 0.5$). Some of the variants considered are based on plug-in estimates of the fraction of true hypotheses π_0 as in the *samr* package. In section 5.4 the closed testing-based bound $\overline{V}_{\text{ct}}$ is compared to the basic bound $\overline{V}$. The performance of the shortcut is illustrated in Section 5.5. All simulations were performed with *R*.

5.1 Simulated data and tests used

Here we describe the simulated data and tests used for all simulations. The simulated data matrix was the $2n \times m$-matrix

$$\mathbf{X} = \mathbf{X}' + \mathbf{Z}.$$

It can be seen as representing m gene expression levels of $2n$ persons. Here $\mathbf{X}'$ is a $2n \times m$-matrix of independent normally distributed variables with variance 1. For some $0 \leq F \leq m$, in the first F columns of $\mathbf{X}$ the first n entries had mean 1 and all other entries had mean 0. The matrix $\mathbf{Z}$ was used to make the entries in each row of $\mathbf{X}$ correlated. It is defined by $\mathbf{Z}_{ji} := s_i Z_j$, where $s_i = 1 - 2\mathbb{1}_{\{i>m/2\}}$ and each Z_j is independent and normally distributed with mean 0 and standard deviation σ_Z. For $1 \leq j \leq 2n$ and $1 \leq i < i' \leq m$ we have $Cov(\mathbf{X}_{ji}, \mathbf{X}_{ji'}) = E(\pm Z_j^2) = \pm {\sigma_Z}^2$, hence the correlation is

$$\rho(\mathbf{X}_{ji}, \mathbf{X}_{ji'}) = \frac{\pm {\sigma_Z}^2}{1 + {\sigma_Z}^2}.$$

For each $1 \leq i \leq m$, let H_i be the null hypothesis that $\mathbf{X}_{1,i}..., \mathbf{X}_{2n,i}$ are independent and standard normally distributed. Thus the first F hypotheses were false, such that the fraction of true hypotheses was $\pi_0 = (m - F)/m$.

Under H_i, the test statistic

$$T_i := \sum_{j=1}^{n} \mathbf{X}_{j,i} - \sum_{j=n+1}^{2n} \mathbf{X}_{j,i}$$

is normally distributed with variance $2n \cdot (1 + {\sigma_Z}^2)$, so that we can efficiently calculate the corresponding two-sided p-value, i.e. the probability under H_i of a larger value of $|T_i|$ than observed. The test statistics used in the

simulations were these p-values. For each hypothesis we used the same rejection region D of the form $(0, c)$, where $c \in (0, 1)$ is some cutoff.

As the group of transformations we used the $(2n)!$ maps that shuffle the rows of $\mathbf{X}$, leaving each individual row intact. These can e.g. be seen as permutations of cases and controls. In all the simulations except those of Section 5.5 we used $w = 100$, i.e. each time we drew 99 random permutations and added the identity. For larger w similar results are obtained (see also Marriott, 1979). The values of m, π_0, the cutoff c, α and $|\rho|$ are specified per case below.

5.2 Performance of variants of SAM as bounds

For $m = 1000$, $\alpha = 0.05$ and rejection region $D = (0, 0.01)$, we investigate the performance of variants of SAM as $(1-\alpha)100\%$-confidence upper bounds of the FDP. Some of the variants of SAM discussed here are based on an estimate $\widehat{\pi}_0$ of $\pi_0 = N/m$. Like the *samr* package, we calculated $\widehat{\pi}_0$ as

$$\frac{\#\{1 \leq i \leq m : P_i > q_i\}}{0.5 \cdot m},$$

where q_i is the 0.5-quantile of the permutation distribution of P_i. We write $\overline{FDP} := R^{(k)}/R$, where $\overline{FDP} = 0$ for $R = 0$. Note that $\overline{FDP}$ is potentially larger than 1. We also write $\widehat{\pi}_0' = \widehat{\pi}_0 \wedge 1$ and $\overline{FDP}' = \overline{FDP} \wedge 1$.

Table 1 shows 95%-confidence intervals for the probabilities that the bounds were smaller than the real FDP, for different values of π_0 and $|\rho|$. The value $|\rho| = 0.5$ corresponds to $\sigma_Z = 1$. From Table 1 it can be seen that $\overline{FDP}'$ is the only bound with the desired property $\mathbb{P}(\text{upper bound} < FDP) \leq \alpha$. For the other bounds, this probability is much larger than α for many settings (see also Korn et al., 2007), especially under dependence. This is related to the known fact that the estimate $\widehat{\pi}_0$ often has low accuracy under dependence (Qiu et al., 2005; Qiu and Yakovlev, 2006; Kim and van de Wiel, 2008; Schwartzman and Lin, 2011). For $\alpha = 0.1$ we got similar results.

The tables in Sections 5.2 and 5.3 are based on 5000 simulations per setting, which took about half an hour per setting on a good PC.

5.3 Performance of variants of SAM as estimators

We performed the same simulations as in Section 5.2, but with $\alpha = 0.5$. For $\alpha = 0.5$ we write $\widehat{FDP} := \overline{FDP}$ and we let $\widehat{FDP}' = \widehat{FDP} \wedge 1$. Table

π_0	$\lvert\rho\rvert$	$\overline{FDP}'$	$\widehat{\pi}_0 \cdot \overline{FDP}$	$\widehat{\pi}_0' \cdot \overline{FDP}$	$\widehat{\pi}_0' \cdot \overline{FDP}'$
1	0	$.038 \pm .006$	$\mathbf{.063} \pm .007$	$\mathbf{.063} \pm .007$	$\mathbf{.505} \pm .014$
1	0.5	$.047 \pm .006$	$\mathbf{.114} \pm .010$	$\mathbf{.114} \pm .010$	$\mathbf{.381} \pm .013$
0.95	0	$.012 \pm .004$	$.028 \pm .005$	$.028 \pm .005$	$.028 \pm .005$
0.95	0.5	$.043 \pm .006$	$\mathbf{.106} \pm .009$	$\mathbf{.106} \pm .009$	$\mathbf{.238} \pm .012$
0.8	0	$.000 \pm .001$	$.002 \pm .002$	$.002 \pm .002$	$.002 \pm .002$
0.8	0.5	$.029 \pm .005$	$\mathbf{.088} \pm .009$	$\mathbf{.088} \pm .009$	$\mathbf{.181} \pm .011$
0.5	0	$.000 \pm .001$	$.000 \pm .001$	$.000 \pm .001$	$.000 \pm .001$
0.5	0.5	$.016 \pm .004$	$.055 \pm .007$	$.055 \pm .007$	$\mathbf{.116} \pm .009$

Table 1: 95%-confidence intervals for $\mathbb{P}(\text{upper bound} < FDP)$, for $\alpha = 0.05$. Probabilities larger than 0.05 are shown in boldface.

2 shows for the different estimates the probability of underestimating the FDP, for different values of π_0 and of the correlation.

The simulations confirm that, as we have proven, $\mathbb{P}(\widehat{FDP}' \leq FDP) \leq 0.5 = \alpha$, i.e. it is a median unbiased estimator. Note also that for the estimate $\widehat{\pi}_0' \cdot \widehat{FDP}'$, this does not hold in all situations. For many of the simulated settings however, all estimates were median unbiased.

π_0	$\lvert\rho\rvert$	$\widehat{FDP}'$	$\widehat{\pi}_0 \cdot \widehat{FDP}$	$\widehat{\pi}_0' \cdot \widehat{FDP}$	$\widehat{\pi}_0' \cdot \widehat{FDP}'$
1	0	0.44	0.51	0.51	0.69
1	0.5	0.45	0.47	0.47	0.47
0.95	0	0.32	0.43	0.43	0.43
0.95	0.5	0.44	0.46	0.47	0.47
0.8	0	0.08	0.29	0.29	0.29
0.8	0.5	0.37	0.45	0.45	0.45
0.5	0	0.00	0.16	0.16	0.16
0.5	0.5	0.28	0.40	0.40	0.40

Table 2: Estimates of $\mathbb{P}(\text{estimate} < FDP)$ for $\alpha = 0.5$. Each estimate is based on 5000 simulations, such that it differs less than 0.015 from the real value with probability at least 95%.

In Table 3 95%-confidence intervals are shown (assuming normality) for the expected absolute errors of the different estimators. Note that for large π_0, $\widehat{FDP}'$ was a more accurate estimator than the estimators that use $\widehat{\pi}_0$

or $\widehat{\pi}_0'$. For small π_0 and no correlation, the other estimates were more accurate. When there was correlation, the estimates based on $\widehat{\pi}_0$ were often less accurate than $\widehat{FDP}'$. The reason for this may be that $\widehat{\pi}_0$ and $\widehat{\pi}_0'$ were less accurate estimators of π_0 under dependence than under independence. The low accuracy of $\widehat{\pi}_0$ under dependence is a known issue (Qiu et al., 2005; Qiu and Yakovlev, 2006; Kim and van de Wiel, 2008; Schwartzman and Lin, 2011).

π_0	$\lvert\rho\rvert$	$\widehat{FDP}'$	$\widehat{\pi}_0 \cdot \widehat{FDP}$	$\widehat{\pi}_0' \cdot \widehat{FDP}$	$\widehat{\pi}_0' \cdot \widehat{FDP}'$
1	0	**.091** $\pm$ .004	.305 $\pm$ .012	.298 $\pm$ .012	.104 $\pm$ .004
1	0.5	**.332** $\pm$ .011	.543 $\pm$ .018	.469 $\pm$ .014	.348 $\pm$ .011
0.95	0	.099 $\pm$ .002	**.096** $\pm$ .002	**.096** $\pm$ .002	**.096** $\pm$ .002
0.95	0.5	**.387** $\pm$ .009	.520 $\pm$ .017	.456 $\pm$ .014	.399 $\pm$ .009
0.8	0	.050 $\pm$ .001	**.034** $\pm$ .001	**.034** $\pm$ .001	**.034** $\pm$ .001
0.8	0.5	**.236** $\pm$ .008	.268 $\pm$ .010	.256 $\pm$ .009	.244 $\pm$ .008
0.5	0	.044 $\pm$ .000	**.015** $\pm$ .000	**.015** $\pm$ .000	**.015** $\pm$ .000
0.5	0.5	.145 $\pm$ .006	.144 $\pm$ .007	.143 $\pm$ .007	**.142** $\pm$ .007

Table 3: 95%-confidence intervals for $\mathbb{E}|\text{estimate} - FDP|$ for $\alpha = 0.5$. In each row, the smallest average error is shown in boldface.

Apart from recording the absolute errors we also recorded the relative differences

$$\left|\frac{\text{estimate}}{FDP} - 1\right|.$$

For this error measure we got similar results. In particular, $\widehat{FDP}'$ was the best estimator of the FDP for large π_0.

The closed testing-based estimate $\overline{V}_{\text{ct}}/R$ (for $\alpha = 0.5$) and its approximation $\overline{V}^*_{\text{ct}}/R$ are often more accurate than $\widehat{FDP}'$ (results not shown). For $|\rho| = 0$ and $\pi_0 \leq 0.8$ however, the estimates based on $\hat{\pi}_0$ still performed best.

5.4 Performance of the closed testing-based bound

Here we illustrate that the bound $\overline{V}_{\text{ct}}$ based on the full closed testing procedure often improves the basic bound $\overline{V}$. We computed $\overline{V}_{\text{ct}}$ using the

shortcut in Proposition 7. Recall that calculating $\overline{V}_{\mathrm{ct}}$ is often computationally infeasible when R or $\overline{V}_{\mathrm{ct}}$ is large, hence we took $m = 100$. Further, we took $\alpha = 0.1$ and $D = (0, 0.01)$ as the rejection region. We calculated 95%-confidence intervals (assuming normality) for the expected values of the $(1-\alpha)$-upper bounds. The results are shown in Table 4. Recall that $\overline{V}/R$ and $\overline{V}_{\mathrm{ct}}/R$ are the $(1-\alpha)$-confidence FDP bounds corresponding to $\overline{V}$ and $\overline{V}_{\mathrm{ct}}$.

π_0	$\lvert\rho\rvert$	R	$\overline{V}/R$	$\overline{V}_{\mathrm{CT}}/R$
0.9	0	8.8 ± 0.1	0.35 ± 0.01	0.33 ± 0.01
0.9	0.2	7.6 ± 0.2	0.47 ± 0.01	0.46 ± 0.01
0.7	0	23.9 ± 0.5	0.18 ± 0.01	0.12 ± 0.00
0.7	0.2	20.2 ± 1.1	0.23 ± 0.02	0.18 ± 0.02
0.5	0	39.1 ± 0.5	0.15 ± 0.01	0.08 ± 0.00
0.5	0.2	40.0 ± 1.8	0.17 ± 0.01	0.11 ± 0.01

Table 4: 95%-confidence intervals for $\mathbb{E}$(upper bound). The column below "R" shows the average number of rejections. The value $|\rho| = 0.2$ corresponds to $\sigma_Z = 0.5$.

The table shows that if π_0 is near 1, the basic bound $\overline{V}$ and the closed testing-based bound $\overline{V}_{\mathrm{ct}}$ are close, but when there are many false hypotheses, closed testing provides a substantial improvement. The same holds for $\alpha = 0.5$.

The simulations were computationally intensive, especially for $\pi_0 = 0.5$ when there were many rejections. For this value of π_0, 100 simulations took about 40 hours on a good PC.

5.5 Performance of the conservative shortcut

We illustrate that in some settings for $\alpha = 0.5$ the estimate $\overline{V}_{sc}$ obtained with the conservative shortcut defined in Section 4 is lower than the basic estimate $\overline{V}$. In these simulations $m = 2000$. We also took $w = 2000$ and $D = (0, 0.1)$, because the shortcut usually does not improve $\overline{V}$ if the cutoff and w are small. For different values of π_0 and the correlation, we calculated confidence intervals (assuming normality) for the expected absolute difference from the real FDP, for $\widehat{FDP}'$ and $\widehat{FDP}_{sc} := \overline{V}_{sc}/R$. The results are shown in Table 5. The computation time was a few minutes per 100

simulations.

As expected, the shortcut only improved $\widehat{FDP}'$ when π_0 was far from 1. The shortcut provides less small bounds than the full closed testing procedure, but is computationally feasible for larger datasets. For such datasets, it is the best computationally feasible bound that has been proven to be a $(1-\alpha)$-confidence bound.

π_0	$\lvert\rho\rvert$	$\widehat{FDP}'$	$\widehat{FDP}_{sc}$
0.8	0	0.117 ± 0.006	0.117 ± 0.006
0.8	0.2	0.145 ± 0.010	0.145 ± 0.010
0.5	0	0.157 ± 0.002	0.150 ± 0.002
0.5	0.2	0.140 ± 0.006	0.140 ± 0.006
0.1	0	0.177 ± 0.001	0.105 ± 0.001
0.1	0.2	0.171 ± 0.002	0.157 ± 0.003

Table 5: 95%-confidence intervals for $\mathbb{E}|\text{estimator} - FDP|$. The value $|\rho| = 0.2$ corresponds to $\sigma_Z = 0.5$.

5.6 Performance of the approximation method

We now investigate the approximation method (Section 3.3), which provides smaller bounds than the conservative shortcut. Its validity as a $(1-\alpha)$-confidence bound has not been proven (for finite $\#\mathcal{S}$), hence we use simulations to investigate its validity.

Firstly we show that in the simulation settings where computation of $\overline{V}_{\text{ct}}$ was feasible, the estimate $\overline{V}^*_{\text{ct}}$ is on average close $\overline{V}_{\text{ct}}$. In the settings of Section 5.4 ($\alpha = 0.1$), we constructed $\mathcal{S}$ as a collection of 1000 independent, uniformly distributed random subsets from $\{\mathcal{I} \subseteq \mathcal{R} : \ \#\mathcal{I} = M\}$ (duplicates were allowed). In table 6 it can be seen that $\overline{V}^*_{\text{ct}}$ was on average close to $\overline{V}_{\text{ct}}$. This means that they were usually equal and sometimes $\overline{V}^*_{\text{ct}}$ was equal to $\overline{V}_{\text{ct}} - 1$ or $\overline{V}_{\text{ct}} - 2$. Taking $\#\mathcal{S}$ smaller (larger) than 1000 naturally resulted in a larger (smaller) average estimation error (result not shown).

The estimate $\overline{V}^*_{\text{ct}}$ of $\overline{V}_{\text{ct}}$ is good, but not perfect. This is irrelevant for our purposes however, as long as $\overline{V}^*_{\text{ct}}$ has the desired property of being a $(1-\alpha)$-confidence bound. In the last column of Table 7 it can be seen that this is the case for the simulation setting of Sections 5.2 and 5.3. (Note that we took $m = 1000$ again, since the approximation method is feasible for

π_0	$\lvert\rho\rvert$	$\overline{V}_{\mathrm{ct}}/R$	$\lvert\overline{V}_{\mathrm{ct}}/R - \overline{V}^*_{\mathrm{ct}}/R\rvert$
0.9	0	0.33 ± 0.01	0.000
0.9	0.2	0.46 ± 0.01	0.000
0.7	0	0.12 ± 0.00	0.005
0.7	0.2	0.18 ± 0.02	0.006
0.5	0	0.08 ± 0.00	0.005
0.5	0.2	0.11 ± 0.01	0.006

Table 6: The last column shows the average absolute difference between the two upper bounds of the FDP. The second-last column shows confidence intervals for the expected values of $\overline{V}_{\mathrm{ct}}/R$. The value $|\rho| = 0.2$ corresponds to $\sigma_Z = 0.5$.

large m.) Here $\#\mathcal{S}$ was again taken to be 1000.

It was also interesting to compare $\overline{V}^*_{\mathrm{ct}}$ to the bound $V^{(k)}$. The bound $V^{(k)}$, which is unknown in practice, was shown to be a $(1-\alpha)$-confidence bound in the proof of Theorem 4. Table 7 shows that the probability that $\overline{V}^*_{\mathrm{ct}} < V^{(k)}$ was very small in the simulation settings of Sections 5.2 and 5.3, with $\overline{V}^*_{\mathrm{ct}} > V^{(k)}$ being much more likely. Since $V^{(k)}$ is a $(1-\alpha)$-upper bound, it is then not surprising that $\overline{V}^*_{\mathrm{ct}}$ is also a $(1-\alpha)$-upper bound in these settings.

For other values of α (and for p-values based on a t-statistic), we similarly found that $\overline{V}^*_{\mathrm{ct}}$ was a $(1-\alpha)$-confidence bound. Based on these findings, it seems reasonable to use $\overline{V}^*_{\mathrm{ct}}$ as a $(1-\alpha)$-confidence upper bound in practice, given that the test statistics are p-values as was the case in our simulation settings. We recommend taking $\#\mathcal{S}$ as large as possible in practice (try $\#\mathcal{S} \geq 10^4$).

6 Application to data

We illustrate the performance of the $(1-\alpha)$-upper bound $\overline{V}/R$ on real data. We analyse the freely available dataset that was used for the original SAM paper by Tusher et al. (2001). The dataset contains gene expression levels of about 7000 genes measured with a microarray. For each gene there are eight observations, of which four from unirradiated cells and four from irradiated cells. In each of these two groups there are two observations from one cell line and two observations from another cell line (making four observations

π_0	$\lvert\rho\rvert$	$\mathbb{E}R$	$\mathbb{P}(\overline{V}^*_{\mathrm{ct}} < V^{(k)})$	$\mathbb{P}(\overline{V}^*_{\mathrm{ct}} < V)$
1	0	10.0	0.000	0.039
1	0.5	10.0	0.018	0.050
0.95	0	27.9	0.000	0.014
0.95	0.5	18.2	0.013	0.048
0.8	0	81.6	0.000	0.002
0.8	0.5	40.0	0.007	0.035
0.5	0	188.3	0.000	0.000
0.5	0.5	87.8	0.002	0.024

Table 7: The same simulation setting as in Sections 5.2 and 5.3 was taken. The second last column shows the estimated probability that $\overline{V}^*_{\mathrm{ct}}$ was smaller than the bound $V^{(k)}$. The last column shows the error rate. The estimates are based on 5000 simulations per setting (taking up to 5 hours).

per cell line). More details are in Tusher et al. (2001).

We performed the same analysis as Tusher et al., with the addition that we calculated $(1-\alpha)$-confidence upper bounds for the FDP. Not all 8! permutations were used but only the permutation maps that permuted within the two cell lines. There are $4!4! = 576$ such permutation maps. Note that this set of permutation maps has a group structure. This group consists of 36 classes of 16 equivalent permutations that always give the same test statistic. Using one permutation from each class leads to the same analysis as with 576 permutations, so we only use 36 distinct permutations. The same permutations are used in Tusher et al. (2001).

For gene i, H_i is defined as the hypothesis that the distribution of the expression level of gene i is the same for all cells. Note that Assumption 1 is satisfied if the joint distribution of the gene expression levels corresponding to $\mathcal{N}$ is the same for cases and controls. As a biological argument for this exhangeability, note that it seems unlikely that the treatment would affect the joint distribution of the gene expressions corresponding to $\mathcal{N}$, while leaving the marginal distributions unchanged.

In Tusher et al. and here, the user chooses a threshold $\Delta \geq 0$. Based on Δ and the data, the rejection region D is calculated. This region is of the form $(-\infty, c_1) \cup (c_2, \infty)$, with $c_1,\ c_2 \in \mathbb{R}$. Details on how the cut-offs c_1 and c_2 are based on Δ and the data are in Tusher et al.. The larger Δ is, the fewer hypotheses are rejected and the smaller the FDP tends to be.

The dependence of the cut-offs on the data might lead to bias. The bias is minor or absent however, as long as Δ is not cherry-picked after looking at the data. In the analysis here and in Tusher et al. no plug-in estimate of π_0 was used.

Considering the same values of the threshold Δ as Tusher et al. and some larger values, we calculated the corresponding estimates of the FDP as well as the basic $(1-\alpha)$-confidence upper bound for the FDP. The results are shown in Table 8. Here $\overline{FDP}_\gamma$ stands for $\overline{V}/R$ for $1-\alpha=\gamma$, so that e.g. $\overline{FDP}_{0.9}$ is a 90%-confidence upper bound for the FDP. $\widehat{\overline{FDP}}_{\text{mean}}$ stands for $\widehat{V}_{\text{mean}}/R$ where $\widehat{V}_{\text{mean}}$ is the mean of the values $R(gX)$, $g \in H$, where H is the set of 36 permutations . This is the estimate that is reported in Table 1 in Tusher et al. (2001). Keep in mind that the bounds are not uniform over Δ or α.

Δ	R	$\widehat{\overline{FDP}}_{\text{mean}}$	$\overline{FDP}_{0.5}$	$\overline{FDP}_{0.9}$	$\overline{FDP}_{0.95}$
0.3	571	0.56	0.45	0.97	1
0.4	282	0.46	0.34	0.99	1
0.6	162	0.35	0.25	0.98	1
0.9	80	0.24	0.13	0.88	0.98
1.2	46	0.18	0.09	0.67	0.98
1.8	26	0.14	0.08	0.46	0.85
2.5	12	0.12	0.08	0.42	0.75
3	10	0.12	0.10	0.30	0.70
3.5	3	0.06	0	0.33	0.33

Table 8: For different values of the threshold Δ, estimators and bounds for the FDP are shown. R is the number of rejected hypotheses. The value $\overline{FDP}_{0.5}$ is a median unbiased estimator of the FDP and $\overline{FDP}_{0.95}$ is a 95%-confidence upper bound for the FDP.

Some of our results are slightly different from those in Table 1 in Tusher et al. (2001), which may be due to a minor difference in the code or the data used. Note that for every Δ the estimate $\overline{FDP}_{0.5}$ based on the median is smaller the estimate $\widehat{\overline{FDP}}_{\text{mean}}$ based on the mean. This is because the permutation distribution of R tended to be skewed to the right. Note that for $\alpha = 0.05$ and smaller values of Δ, we obtain trivial 95%-confidence bounds. For example, for $\Delta = 0.6$ we do not have 95% confidence that at

least one of the 162 rejected hypotheses is false. For larger values of Δ the cut-offs are stricter and we do get useful 95%-confidence bounds.

Note that since there are only 36 permutations, the 95%-confidence bound for V is the second largest value among $R(gX)$, $g \in H$. Thus it is in fact a $(35/36)100\% \approx 97.2\%$ confidence bound. For $\Delta = 3.5$ there are 3 rejections and we know with 97.2% confidence that at least two of these are true findings. We also know with 50% confidence that all three rejections are true findings. For $\Delta = 3$ there are 10 rejections and we know with 90% confidence (and indeed $(33/36)100\% \approx 91.7\%$ confidence) that at least 7 of these are true findings, although we cannot generally pinpoint which of the rejected hypotheses are false.

Calculating $\overline{V}_{\mathrm{ct}}$ was only feasible for $\Delta \geq 2.5$ and sometimes offered an improvement over $\overline{V}$. For example, for $\Delta = 3$ and $\alpha = 0.05$, the bound was 0.6 instead of 0.7. Usually the basic bound was not improved for $\Delta \geq 2.5$, due to the relatively small number of rejections for such Δ and the discreteness of the already small bound $\overline{V}_{\mathrm{ct}}$.

For $\Delta < 2.5$, when computing $\overline{V}_{\mathrm{ct}}$ was not feasible, we performed the approximation method (with $\#\mathcal{S} = 10^4$). The results are shown in Table 9. The improvements are relatively small in this situation, since there is no proof that π_0 is far away from 1 for these data.

In many practical situations FDP bounds (and the FDP itself) tend to decrease with R, but this is only a tendency. Examples of exceptions can be seen in both Table 8 and Table 9. Hence a user might find that decreasing Δ post hoc would both increase R and decrease the bound, which would be very tempting. This could lead to selection bias however; Δ should be chosen before looking at the data.

Δ	R	$\overline{FDP}^*_{0.5}$	$\overline{FDP}^*_{0.9}$	$\overline{FDP}^*_{0.95}$
0.3	571	**0.43**	**0.81**	1
0.4	282	0.34	**0.82**	1
0.6	162	0.25	**0.88**	1
0.9	80	0.13	0.88	**0.94**
1.2	46	0.09	0.67	**0.96**
1.8	26	0.08	0.46	**0.81**

Table 9: For different values of the threshold Δ, estimators and bounds for the FDP, derived with the approximation method, are shown. Here $\overline{FDP}^*_\gamma$ stands for $\overline{V}^*_{\mathrm{ct}}/R$ for $1 - \alpha = \gamma$. Where it improved the basic bound (see Table 8), the result is shown in boldface.

Conclusion

SAM is a widely applied method, since it requires few assumptions on the dependence structure of the data and nevertheless adapts to this structure. Until now SAM had no known properties. In this paper the assumptions underlying SAM have been made explicit. Moreover it has been shown how SAM can be extended to provide a $(1-\alpha)$-confidence upper bound for the FDP. For $\alpha = 0.5$ a median unbiased estimate of the FDP is obtained. The *samr* *R*-package multiplies this estimate by an estimate of the fraction of true hypotheses π_0 to obtain a lower estimate of the FDP. We have shown using simulations that this often still results in a median unbiased estimate of the FDP, although in many cases the estimate becomes less accurate. For $\alpha = 0.05$ and $\alpha = 0.1$, multiplying the $(1-\alpha)$-confidence bound by the estimate of π_0 often does not result in a $(1-\alpha)$-confidence bound.

We have shown that by using a closed testing procedure the basic bound can be decreased, in such a way that the confidence level is maintained. The improvement over the basic bound can be appreciable, as simulations illustrate. The improved bound only depends on rejected sets for permuted versions of the data. Once these are known, the computation time is not influenced by the complexity of the test statistics. Hence the choice of test statistics typically does not determine the computational feasibility of the method.

When there are many rejected hypotheses, the closed testing-based method is often computationally infeasible. Therefore we have included a fast approximation of this method, which still provides confidence for our simulation settings. We have also constructed a conservative shortcut, which provides larger bounds but has proven validity. This shortcut only improves the basic bound in specific settings. Both these fast alternatives to the closed testing-based method are feasible when there are many thousands of rejections.

Our methods provide an FDP bound for the prespecified rejection region. The region cannot generally be picked after looking at the data, since the bounds are not uniform over multiple rejection regions. There exists a limited amount of literature on uniformly valid FDP bounds (Meinshausen, 2006; Goeman and Solari, 2011). An example is the method by Meinshausen (2006), which is closely related to SAM. There are opportunities to improve some of these methods, similarly to the way SAM has been improved here.

This may be the subject of future research.

Theorem 3 provides a general permutation principle which can be used to prove properties of methods based on random permutations (SAM; Meinshausen, 2006; Westfall and Young, 1993). This result is related to Phipson and Smyth (2010) but is more generally useful. We have used it to prove the validity of the methods in this paper. It may be used to prove properties of other permutation-based procedures in the future.

Bibliography

Chu, G., Li, J., Narasimhan, B., Tibshirani, R., and Tusher, V. Significance analysis of microarrays users guide and technical document. 2001.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. Multiple hypothesis testing in microarray experiments. Technical report. Available at http://www.bepress.com/ucbbiostat/paper110/. 2002.

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

Genovese, C. R. and Wasserman, L. Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476): 1408–1417, 2006.

Goeman, J. J. and Solari, A. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011.

Hoeffding, W. The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 23:169–192, 1952.

Kim, K. I. and van de Wiel, M. A. Effects of dependence in high-dimensional multiple testing problems. *BMC bioinformatics*, 9(1):114, 2008.

Knijnenburg, T. A., Wessels, L. F., Reinders, M. J., and Shmulevich, I. Fewer permutations, more accurate p-values. *Bioinformatics*, 25(12):i161–i168, 2009.

Korn, E. L., Li, M.-C., McShane, L. M., and Simon, R. An investigation of two multivariate permutation methods for controlling the false discovery proportion. *Statistics in medicine*, 26(24):4428, 2007.

Langsrud, Ø. Rotation tests. *Statistics and computing*, 15(1):53–60, 2005.

Marcus, R., Eric, P., and Gabriel, K. R. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

Marriott, F. Barnard's Monte Carlo tests: How many simulations? *Applied Statistics*, pages 75–77, 1979.

Meinshausen, N. False discovery control for multiple tests of association under general dependence. *Scandinavian Journal of Statistics*, 33(2):227–237, 2006.

Meinshausen, N. and Bühlmann, P. Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika*, 92(4):893–907, 2005.

Pesarin, F. and Salmaso, L. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010.

Phipson, B. and Smyth, G. K. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1):39, 2010.

Qiu, X. and Yakovlev, A. Some comments on instability of false discovery rate estimation. *Journal of bioinformatics and computational biology*, 4 (05):1057–1068, 2006.

Qiu, X., Klebanov, L., and Yakovlev, A. Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

Schwartzman, A. and Lin, X. The effect of correlation in false discovery rate estimation. *Biometrika*, 98(1):199–214, 2011.

Solari, A., Finos, L., and Goeman, J. J. Rotation-based multiple testing in the multivariate linear model. *Biometrics*, 70(4):954–961, 2014.

Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

Storey, J. D., Taylor, J. E., and Siegmund, D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

Tusher, V. G., Tibshirani, R., and Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

Westfall, P. H. and Young, S. S. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993.