Familywise Error Rate Control via Knockoffs

Abstract

We present a novel method for controlling the k-familywise error rate (k-FWER) in the linear regression setting using the knockoffs framework first introduced by Barber and Candès. Our procedure, which we also refer to as knockoffs, can be applied with any design matrix with at least as many observations as variables, and does not require knowing the noise variance. Unlike other multiple testing procedures which act directly on p-values, knockoffs is specifically tailored to linear regression and implicitly accounts for the statistical relationships between hypothesis tests of different coefficients. We prove that knockoffs controls the k-FWER exactly in finite samples and show in simulations that it provides superior power to alternative procedures over a range of linear regression problems. We also discuss extensions to controlling other Type I error rates such as the false exceedance rate, and use it to identify candidates for mutations conferring drug-resistance in HIV.

Keywords: k-familywise error rate; knockoffs; multiple testing; linear regression; Lasso; negative binomial distribution.

1 Introduction

Multiple testing has received increasing attention with the advent of fields like genetics, technology, and astronomy which produce very high-dimensional datasets. The increasing number of hypotheses being simultaneously tested has motivated extensive research into procedures that maintain control of the familywise errors that abound when each hypothesis is only tested individually. For instance, the canonical criterion of the familywise error rate (FWER) controls the probability of falsely rejecting any of the true null hypotheses. A number of more modern landmark works have introduced new Type I error rates that allow for higher power by relaxing the FWER, including the false discovery rate (FDR) (Benjamini and Hochberg, 1995), the k-FWER (Hommel and Hoffmann, 1988; Lehmann and Romano, 2005), and the false discovery exceedance (FDX) (Genovese and Wasserman, 2004; van der Laan et al., 2004). Each one has a different interpretation, but all control an error rate defined over all hypotheses being tested, so that conclusions can be drawn by considering rejected hypotheses together.

Among multiple testing problems, some of most important deal with finding relationships between variables. Such investigations are often posed as a linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z},$$

where $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p] \in \mathbb{R}^{n \times p}$ is a design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a signal vector of interest, and $\boldsymbol{z} \in \mathbb{R}^n$ is the error term. The hypotheses of interest are which variables β_j, after controlling for all other variables, contribute to the model, or have nonzero coefficients. With the ability to encode correlations between variables, linear models capture far more real-life examples than sequence models. Examples abound particularly in genetics, where one searches for relationships between parts of the genome, often in the form of single nucleotide polymorphisms or expression levels, and continuous variables such as health

factors or drug response. Unfortunately, due to the dependence among the variables in the linear model, their respective tests do not in general exhibit any of the simple dependence structures, such as independence or positive dependence, that are required for many of the most powerful existing procedures.

In this work we focus on controlling the k-FWER, the probability of making at least k false discoveries, in the context of linear models. Our method uses the framework of knockoffs introduced by Barber and Candès (2015). The idea of knockoffs is to carefully construct artificial variables that serve as controls for the original variables. Barber and Candès show that these controls are easy to construct and can be used to automatically account for variable dependence to provide finite-sample FDR control for general design matrices without knowledge of the noise variance. Controlling the FDR can be highly desirable in a high-power setting, but results can be hard to interpret when few discoveries are made, as the realized false discovery proportion may be highly variable. The k-FWER, which in the case of $k = 1$ reduces to the standard FWER, always has a clear interpretation by explicitly bounding the probability of k or more false discoveries, making it a useful criterion in all settings, as evidenced by its wide acceptance in the scientific community. The k-FWER also provides a fundamental building block to other Type I error rates, such as the FDX and Per Family Error Rate (PFER), as we will discuss in Section 4. We leverage the attractive features of the knockoffs framework to construct a novel procedure for controlling the k-FWER that implicitly accounts for the exact dependence structure in linear regression problems. In particular, we prove finite-sample k-FWER control for general design matrices without any knowledge of the noise variance, and show in simulations that the power can be substantially greater than state-of-the-art alternatives.

Much previous work has studied controlling the k-FWER under varying assumptions on the statistical dependence among the hypothesis test statistics or p-values. The bulk

of such work has dealt with procedures that act directly on the p-values. When there are more observations than variables and the noise is i. i. d. Gaussian, ordinary least squares regression generates dependent t-statistics for all variables, allowing those procedures that can account for the dependence structure to be applied to the associated p-values. Unfortunately, the joint distribution of such p-values does not generally satisfy popular dependence assumptions such as positive regression dependence on subset (Benjamini and Yekutieli, 2001) or multivariate total positivity (Karlin and Rinott, 1980). Furthermore, many of the procedures that can account for general dependence structures do so nonparametrically through resampling. However, resampling procedures tend to require extra assumptions such as subset-pivotality (Westfall and Young, 1989) which do not hold in general in the regression setting, or only provide exact control asymptotically (Romano and Wolf, 2007). We mention here some work on controlling the k-FWER in finite samples and refer the reader to Guo et al. (2014) for a more thorough review. The most popular methods for FWER control are the Bonferroni (Dunn, 1961) and Holm's (Holm, 1979) procedures, neither of which require assumptions on the dependence among p-values. Under independence, the Bonferroni procedure can be improved using the Šidák correction (Šidák, 1967), or one can employ Hochberg's step-up procedure (Hochberg, 1988). In Lehmann and Romano (2005), step-down procedures generalizing Bonferroni and Holm's procedures are presented, while Romano and Wolf (2007) introduce a generic step-down procedure, all for controlling the k-FWER. Romano and Shaikh (2006) also present step-up procedures for controlling the k-FWER under arbitrary unknown dependence.

To avoid confusion, we point out that the recent work of Lockhart et al. (2014) provides p-values for coefficients in a linear model, however they deal with a different notion of a null hypothesis than used here. In their framework, the null hypotheses are defined sequentially with respect to a growing model, wherein each time the model size is increased by one,

the null hypothesis is that the new variable is uncorrelated with the response, conditional on only the variables already included in the model. In contrast, in our setting the null hypotheses are defined globally as simply whether elements of $\boldsymbol{\beta}$ (the full-model coefficient vector) are zero or not. For instance, if $\beta_1 \neq 0$ and $\beta_2 = 0$ but is correlated with (only) β_1, we would consider selecting β_2 to be a false discovery, while in the sequential setting, β_2 would be a true discovery as long as it is selected before β_1.

The remainder of the paper is structured as follows. Section 2 introduces notation and gives a short introduction to the knockoffs framework. Section 3 describes the knockoffs procedure for control of the k-FWER and proves this control along with tail bounds. Section 4 provides a brief discussion of how the procedure can be used to control the PFER and FDX. Section 5 compares our procedure to state-of-the-art alternatives from the literature, both in terms of practical considerations and power, in a series of simulations. Section 6 demonstrates an implementation on a real dataset from genetics, and Section 7 concludes with discussion and directions for future research.

2 Preliminaries for knockoffs

In this section, we introduce the knockoffs machinery of Barber and Candès (2015) at a minimal level to be sufficient for our exposition of k-FWER control. This material is largely borrowed from the reference Barber and Candès (2015). In referring to the knockoffs framework, we always assume that the number of observations n is at least the number of variables p, the design matrix $\boldsymbol{X}$ has full column rank so that the Gram matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, and the noise term $\boldsymbol{z}$ has independent Gaussian entries. We would like to briefly emphasize here that $n \geq p$ is necessary for the full-model multiple hypothesis testing problem to even be well-defined. For any linear regression problem, the "true" coefficient

vector is only statistically well-defined modulo addition with any vector in the null space of the design matrix. If $p > n$, then the design matrix has a nontrivial null space, thus allowing zeros and nonzeros in the coefficient vector to arise and disappear, changing the fundamental values of the null hypotheses, without changing the data-generating process at all. Except for this non-degeneracy assumption, the knockoffs machinery works for general designs $\boldsymbol{X}$ and does not even require knowledge of noise variance σ^2.

To start with, again, consider the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z},$$

where the noise vector $\boldsymbol{z}$ has independent $\mathcal{N}(0, \sigma^2)$ entries, and each column of $\boldsymbol{X}$ has been normalized to have unit ℓ_2-norm, that is, $\|\boldsymbol{X}_j\| = 1$ for all $1 \le j \le p$. The first step of this method is to construct the knockoff design, denoted as $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{n \times p}$, that obeys

$$\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}} = \boldsymbol{X}^\top \boldsymbol{X}, \quad \boldsymbol{X}^\top \widetilde{\boldsymbol{X}} = \boldsymbol{X}^\top \boldsymbol{X} - \mathrm{Diag}(\boldsymbol{s}), \tag{1}$$

where $\boldsymbol{s} \in \mathbb{R}^p$ has nonnegative entries and the superscript $\top$ denotes matrix transpose hereafter. There are multiple ways to construct this knockoff design; see Barber and Candès (2015, Section 2.1). The first equality forces $\widetilde{\boldsymbol{X}}$ to have the same correlation structure among its columns as $\boldsymbol{X}$. In the ideal case of $n \ge 2p$, it can be guaranteed that the $2p$ column vectors of $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$ are jointly linearly independent. By the second equality, for every $1 \le j \le p$, the original variable $\boldsymbol{X}_j$ and the knockoff counterpart $\widetilde{\boldsymbol{X}}_j$ have the same correlation with all the other $2p - 2$ variables, namely, $\boldsymbol{X}_i, \widetilde{\boldsymbol{X}}_i$ for $i \ne j$. At a high level, we can view the knockoff design as a control group as compared to the original design $\boldsymbol{X}$, which is treated as the case group.

Denote by $\boldsymbol{X}_{\mathrm{KO}} = [\boldsymbol{X}, \widetilde{\boldsymbol{X}}] \in \mathbb{R}^{n \times 2p}$ the concatenation of the original design and the knockoff design. With $\boldsymbol{X}_{\mathrm{KO}}$ in hand, the next step is to generate statistics for each variable.

One way to do so, suggested in Barber and Candès (2015), is by fitting the entire Lasso regularization path on the augmented design,

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{b}\in\mathbb{R}^{2p}}{\operatorname{argmin}} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}_{\text{KO}}\boldsymbol{b}\|^2 + \lambda\|\boldsymbol{b}\|_1, \tag{2}$$

and letting Z_j be the first λ such that $\widehat{\beta}_j$ is nonzero. Formally,

$$Z_j = \sup\{\lambda : \widehat{\beta}_j(\lambda) \neq 0\}.$$

Defining $\widetilde{Z}_j$ analogously for each knockoff variable $\widetilde{\boldsymbol{X}}_j$, the knockoff statistics (using slightly different notation than in the original paper) are

$$W_j = \max\{Z_j, \widetilde{Z}_j\}, \quad \chi_j = \operatorname{sgn}(Z_j - \widetilde{Z}_j),$$

where $\operatorname{sgn}(x) = -1, 0, 1$ if $x < 0, x = 0, x > 0$, respectively. As pointed out in the reference paper, many alternative statistics, including some based on least-squares, least angle regression (Efron et al., 2004), and sorted-ℓ_1-penalized estimation (Bogdan et al., 2015), can be used instead as long as they obey the sufficiency and antisymmetry properties defined therein. The following result, due to Barber and Candès (2015), characterizes the joint distribution of the null χ_j. We say j is a true null when $\beta_j = 0$ and a false null otherwise.

Lemma 1 (Barber and Candès (2015))**.** *Conditional on all W_j and all false null χ_j, all true null χ_j are jointly independent and uniformly distributed on $\{-1, 1\}$.*

This simple lemma is very helpful in proving k-FWER control. Its proof follows from the symmetry between $\boldsymbol{X}_j$ and $\widetilde{\boldsymbol{X}}_j$ if $\beta_j = 0$, which is provided by the construction (1). The lemma shows that χ_j can be interpreted as a one-bit p-value, in the sense that it has equal chance to take 1 or -1 if $\beta_j = 0$. In fact when $\beta_j = 0$, the knockoff symmetry

characterized in (1) introduces exchangeability between $\boldsymbol{X}_j$ and its knockoff counterpart $\widetilde{\boldsymbol{X}}_j$ in the Lasso path (2). Hence, $\boldsymbol{X}_j$ and $\widetilde{\boldsymbol{X}}_j$ are equally likely to enter the Lasso path first. Conversely, if $\beta_j \neq 0$, then $\boldsymbol{X}_j$ is likely to enter before $\widetilde{\boldsymbol{X}}_j$ so that $\chi_j = 1$. Thus a large W_j and a positive χ_j provide evidence against the jth null hypothesis $H_{0,j} : \beta_j = 0$.

3 k-familywise error rate control

Inspired by the interpretation of the statistics W_j and χ_j, it is reasonable to reject hypotheses with positive signs χ_j and large W_j. Parameterized by a positive integer v, the knockoffs procedure for controlling the k-FWER is as follows.

Step 1. Denote by $W_{\rho(1)} \geq W_{\rho(2)} \cdots \geq W_{\rho(p)}$ the order statistics of $\boldsymbol{W}$, where $\rho(1), \ldots, \rho(p)$ is a permutation of $1, \ldots, p$.

Step 2. Let $j^\star$ be the index of the vth -1 in the sequence $\chi_{\rho(1)}, \ldots, \chi_{\rho(p)}$. If fewer than v negatives appear, set $j^\star = p$.

Step 3. Reject all the null hypotheses $H_{0,\rho(j)}$ whenever $j \leq j^\star$ and $\chi_j = +1$.

More compactly, define the threshold

$$T_v = \sup\Big\{t > 0 : \#\{j : W_j \geq t, \chi_j = -1\} = v\Big\},$$

with the usual convention that $\sup \emptyset = -\infty$. The multiplicity of W_j is not accounted for since all W_j are unique with probability 1. Then, the knockoffs procedure rejects all $H_{0,j}$ with $W_j \geq T_v$ and $\chi_j = +1$.

Before characterizing the distribution of false discoveries made by the knockoffs procedure, we define some notation. Let $\mathcal{N}_0 = \{1 \leq j \leq p : \beta_j = 0\}$ be the set of true null hypotheses and $\mathrm{NB}(m, q)$ denote a negative binomial random variable, which counts the

number of successes before the mth failure in a sequence of independent Bernoulli trials with success probability q.

Lemma 2. *For any integer $v \geq 1$, the false discovery number*

$$V = \#\{j \in \mathcal{N}_0 : W_j \geq T_v \text{ and } \chi_j = +1\}$$

is stochastically dominated by $\mathrm{NB}(v, 1/2)$.

Proof of Lemma 2. First, we prove this lemma in the case where $\mathcal{N}_0 = \{1, \ldots, p\}$, that is, $\beta_j = 0$ for all j. Conditional on all W_j, Lemma 1 concludes that $\chi_{\rho(1)}, \ldots, \chi_{\rho(p)}$ are independent and each takes $+1$ and -1, respectively, with probability $1/2$. Note that the permutation ρ is deterministic conditional on the W_j. Recognizing that V is the number of positive χ_j before the vth negative or the pth trial happens, whichever comes first, we see that V is an early stopped negative binomial random variable. In the general case, false null χ_j will insert -1's into the process on the nulls, causing it to stop no later than when $\mathcal{N}_0 = \{1, \ldots, p\}$. Therefore, V is always stochastically dominated by $\mathrm{NB}(v, 1/2)$. □

The stochastic upper bound in Lemma 2 is tight in the following sense. The distribution of V can be made arbitrarily close to $\mathrm{NB}(v, 1/2)$ under the global null by taking $p \gg v$, as in this case at least v negative χ_j will appear in the sequence with high probability. Next we present the main result, which is immediate from Lemma 2 and the negative binomial cumulative distribution function.

Theorem 1. *For any integer $k \geq 1$ and significance $0 < \alpha < 1$, let v to be the largest integer satisfying*

$$\sum_{i=k}^{\infty} 2^{-i-v} \binom{i+v-1}{i} \leq \alpha. \tag{3}$$

Then the knockoffs procedure with parameter v controls the k-FWER at level α, that is, $\mathbb{P}(V \geq k) \leq \alpha$.

As a concrete example, taking $v = 4$ would provide 10-FWER control at level 0.05. As one may observe from (3), the integer v as a function of the level α cannot be continuous. Consequently, $\mathbb{P}(V \geq k)$ is in general lower than the target level α. In particular, for $\alpha \leq 1/2^k$ no positive integer v satisfies (3), so the naive procedure must reject nothing. This matter can be easily resolved by randomization of v, as we will show in Remark 1.

To better understand the knockoffs procedure, we may want to know how many false rejections are made when the k-FWER is not controlled. To this end, the following result bounds the tail probability of V, or the probability of making many more rejections than expected.

Corollary 1. *For arbitrary $a > 0$, the error rate of the knockoffs procedure with parameter v obeys*

$$\mathbb{P}(V \geq (1+a)v) \leq \theta(a)^v,$$

where $\theta(a) = \frac{(a+2)^{a+2}}{2^{a+2}(a+1)^{a+1}} < 1$.

Proof of Corollary 1. By Lemma 2, it suffices to prove the inequality when V is distributed as $\mathrm{NB}(v, 1/2)$. For any positive number $\eta < \log 2$, from the Markov inequality we get

$$\mathbb{P}(V \geq k) \leq \frac{\mathbb{E}(\mathrm{e}^{\eta V})}{\mathrm{e}^{(1+a)\eta v}} = \frac{1}{(2-\mathrm{e}^{\eta})^v \mathrm{e}^{(1+a)\eta v}}.$$

The desired bound follows from taking $\eta = \log(2+2a) - \log(2+a)$. □

Remark 1 (Power Improvement). As mentioned earlier, the knockoffs procedure suffers from a discretization problem, especially for small k, but this can be remedied by randomization as follows. For any desired level $\alpha \in (0, 1)$, there must exist an integer $v \geq 0$ such that

$$\mathbb{P}_v(V \geq k) \leq \alpha \leq \mathbb{P}_{v+1}(V \geq k),$$

where the subscript v or $v+1$ emphasizes the parameter of the knockoffs procedure. We can devise a mixture procedure that obeys exactly $\mathbb{P}(V \geq k) = \alpha$ by putting weights ω and $1-\omega$, respectively, on the knockoffs procedures with parameters v and $v+1$, where

$$\omega = \frac{\mathbb{P}_{v+1}(V \geq k) - \alpha}{\mathbb{P}_{v+1}(V \geq k) - \mathbb{P}_{v}(V \geq k)}.$$

Furthermore, as with any procedure controlling the k-FWER, power can always be improved without affecting the k-FWER by always making at least $k-1$ rejections. In the case of knockoffs, if we were going to make fewer than $k-1$ rejections, we can simply continue rejecting the indices with the largest W_j and positive χ_j until there are $k-1$. The benefit of this modification depends on the ordering of the hypotheses induced by W_j.

4 Controlling other error rates

This paper has been about controlling the k-FWER, but the procedure introduced can be used to control other Type I error rates as well, namely the PFER and the FDX.

Originally proposed by John Tukey in an unpublished work in 1953, the PFER is defined as $\mathbb{E}(V)$, or in words, the expected number of false discoveries. The control of this error rate under general p-value dependence has not received as much attention in the literature as other error rates, although both Gordon et al. (2007) and Meng et al. (2014) have discussed using the Bonferroni procedure for this purpose. Lemma 2 shows that the knockoffs procedure for controlling the k-FWER also controls the PFER at level v, as $\mathbb{E}(V) \leq \mathbb{E}\,\mathrm{NB}(v, 1/2) = \frac{1/2}{1-1/2} v = v$.

The FDX, also known as the γ-false discovery proportion, tail probability for the proportion of false positives, or false discovery excessive probability, is the probability that the FDP exceeds a specified bound γ. It can be viewed as a more stringent form of the

FDR, and has received much attention recently; see, for example Guo et al. (2014). A number of authors have noticed its intimate connection with the k-FWER, and many of the most successful FDX-controlling procedures in the literature can be posed as meta-procedures applied to a family of k-FWER-controlling procedures (van der Laan et al., 2004; Genovese and Wasserman, 2004; Romano and Wolf, 2007). We briefly review three such meta-procedures, any one of which could be combined with the knockoffs procedure introduced here, and defer further investigation to future work.

In van der Laan et al. (2004), the authors introduced a simple and intuitive procedure which augments any FWER-controlling procedure to control the FDX. This procedure was generalized to any k-FWER-controlling procedure in Genovese and Wasserman (2006). Once the k-FWER-controlling procedure makes R rejections, then if $(k-1)/R > \gamma$, the augmentation procedure makes no rejections, but if $(k-1)/R \leq \gamma$, r more rejections can be made, where r satisfies $(k-1+r)/(R+r) \leq \gamma$. This augmentation procedure controls the FDX exactly when the underlying k-FWER-controlling procedure also provides exact control.

Genovese and Wasserman (2004) proposed a test-inversion procedure for FDX control, similar to the closure principle of Marcus et al. (1976) for FWER control, which was then investigated further in Genovese and Wasserman (2006). The inversion procedure runs global null hypothesis tests on every subset of hypotheses, and then finds the largest subset S whose maximal intersection with any subset for which the global null was not rejected is at most $\gamma|S|$. Note that any k-FWER-controlling procedure is also trivially a test of the global null hypothesis, rejecting whenever k or more rejections are made. Rejecting S from the inversion procedure controls the FDX exactly, and although in general it takes exponential time, for some global tests it can be run in polynomial time (Genovese and Wasserman, 2004).

Given a procedure that can control the k-FWER for any $k \geq 1$, Romano and Wolf (2007) propose a heuristic that aims to control the FDX. In short, given a prescribed level γ and significance α, both between 0 and 1, this heuristic uses a k-FWER-controlling procedure to make rejections for increasing k until just before the number of rejections goes above $k/\gamma - 1$. Explicitly, let R_k be the number of rejections made by a procedure controlling the k-FWER. Then the Romano–Wolf heuristic defines $\hat{k}$ as the smallest k such that $R_k < k/\gamma - 1$ and makes rejections as if controlling the $\hat{k}$-FWER. Although not rigorous due to its adaptivity in $\hat{k}$, under some dependence assumptions, the Romano–Wolf heuristic is shown to enjoy finite sample or asymptotic FDX control for step-down procedures (Guo and Romano, 2007; Delattre and Roquain, 2013).

5 Comparison with other procedures

As mentioned in the introduction, the structure and dependence between coefficients in linear regression preclude the use of many existing procedures. The state-of-the-art procedures that can be found in existing literature and provide exact finite-sample control of the k-FWER in linear regression are:

(a) the generic step-down procedure of Romano and Wolf (2007) applied to the least-squares p-values

(b) the step-up procedure of Romano and Shaikh (2006) applied to the least-squares p-values

(c) the adaptation of Holm's procedure to k-FWER applied to the least-squares p-values (Lehmann and Romano, 2005)

(d) for 1-FWER, the Lasso pathwise testing procedure of Lee et al. (2013)

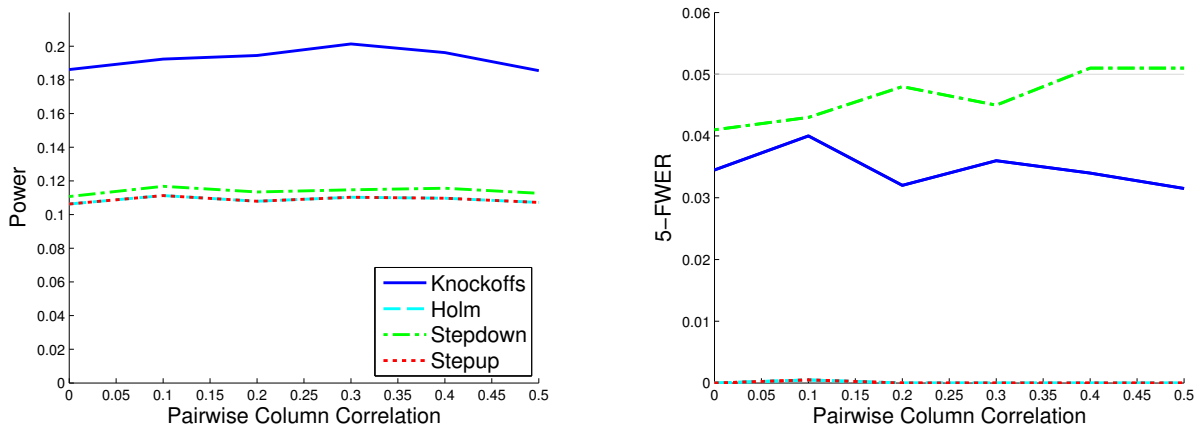
(e) also for the 1-FWER, the closure of any global hypothesis testing procedure, such as the χ^2 test, that can be applied to p-values with any known dependence, applied to the least-squares p-values

Procedure (d) requires the user to know σ^2 exactly, and both (d) and (e) take computation that is exponential in the dimension, p, making them infeasible to use for problems of even moderate size. As a result, we only compare our procedure to (a), (b), and (c). It should be noted that the problem dimensions we considered in simulations were still limited by procedure (b), whose computation time is $O(p^{k-1})$, since each threshold is computed as a maximum over subsets of size $k-1$ from a superset of size up to p. There are also works that obtain asymptotic control of the FWER under some assumptions on the distribution of the design matrix (see, for example, Chernozhukov et al. (2013); Javanmard and Montanari (2014)). As knockoffs applies under no assumptions on the design matrix and the error rates are controlled exactly, we do not compare to such works here.

In each of the following simulations, we performed many independent experiments to gauge how the performance of knockoffs, both in absolute terms and relative to previous methods, depends on correlation in the columns of $\boldsymbol{X}$, the sparsity of $\boldsymbol{\beta}$, and the signal to noise ratio. In each experiment, $\boldsymbol{X}$ is generated by normalizing the columns of a multivariate Gaussian matrix with independent and identically distributed rows, and $\boldsymbol{\beta}$ is generated by setting a pre-specified number of entries to zero, and setting the rest to the same nonzero magnitude, which is also prespecified. The following experiments are all performed in the sparse setting, as that is what the canonical statistics $\boldsymbol{W}$ that use the Lasso are best-suited for. However, nothing about the knockoffs framework to control any Type I error rate is particularly tied to sparsity, and it is of continuing interest to find different statistics $\boldsymbol{W}$ that achieve high power in all manner of settings. In all the following simulations, $n = 1000$, $p = 450$, $\sigma^2 = 25$, we control the 5-FWER at the 5% level, and we apply the modifications

in Remark 1. The step-up procedure is implemented using the critical values suggested in Romano and Shaikh (2006), namely their Equation (13). For a sake of reproducibility, the code to generate these figures is available at `http://wjsu.web.stanford.edu/code.html`.

Our first experiment took $\boldsymbol{\beta}$ to have 10 nonzero elements, all with magnitude 10, and varied the pairwise correlation between the columns of $\boldsymbol{X}$ from 0 to 0.5. Figure 1 shows

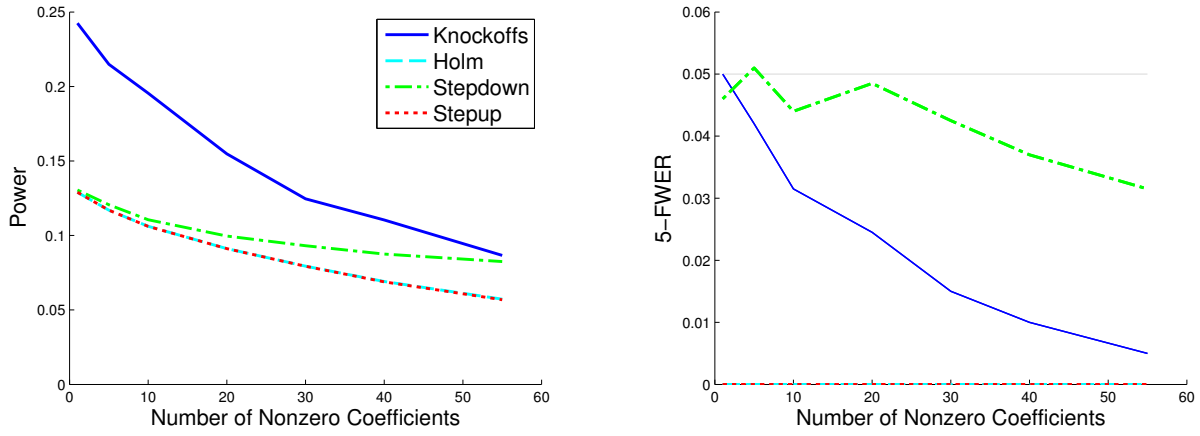


(a) (b)

Figure 1: Comparison of Holm's procedure, generic step-down procedure, step-up procedure, and knockoffs for controlling the 5-FWER at the 5% significance level. As functions of the column correlation of the design matrix, the procedures' powers are shown in (a), while the 5-FWER is given in (b), with the grey line denoting the nominal level of 5%. The curves for Holm and step-up lie on top of one another. Each point is an average over 2000 simulations.

the power (number of true discoveries divided by $\|\boldsymbol{\beta}\|_0$) of the knockoff procedure nearly doubling that of all alternative procedures. The power and 5-FWER of all four procedures is largely unaffected by the correlation in the columns of $\boldsymbol{X}$.

Our second experiment generated columns for $\boldsymbol{X}$ independently, and varied the sparsity of $\boldsymbol{\beta}$, with each nonzero coefficient having magnitude 10. Figure 2 shows the power of the

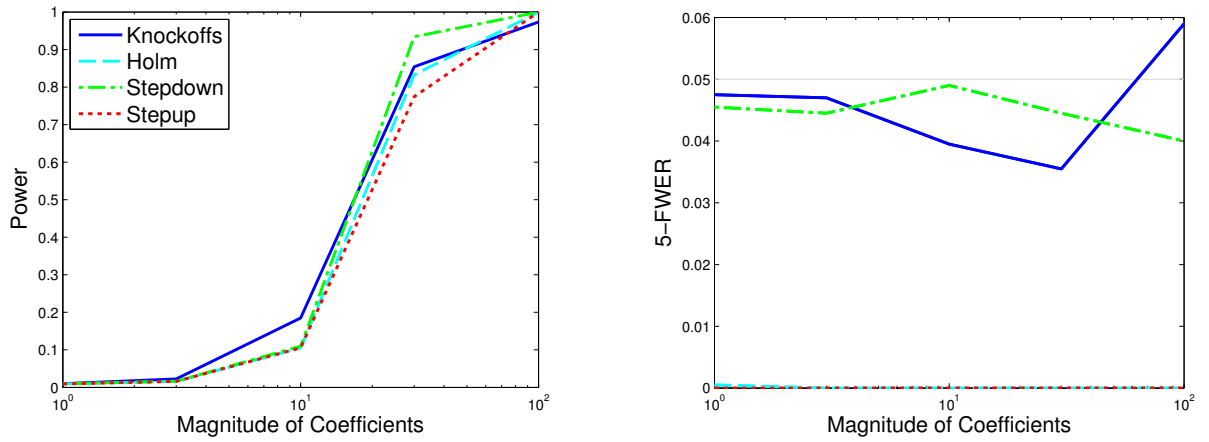


(a) (b)

Figure 2: Comparison of Holm's procedure, generic step-down procedure, step-up procedure, and knockoffs for controlling the 5-FWER at the 5% significance level. As functions of the number of nonzero coefficients, the procedures' powers are shown in (a), while the 5-FWER is given in (b), with the grey line denoting the nominal level of 5%. The curves for Holm and step-up lie on top of one another. Each point is an average over 2000 simulations.

knockoff procedure approximately doubling that of all alternative procedures in the sparsest regime and gradually losing its advantage as the sparsity approaches 10%. The 5-FWER of the knockoffs and step-down decrease as the coefficient vector becomes less sparse, with that of knockoffs becoming conservative especially quickly.

Our third experiment generated independent columns for $\boldsymbol{X}$, used $\boldsymbol{\beta}$ with 10 nonzero entries, and varied the magnitude of the nonzero entries on a logarithmic scale.



(a) (b)

Figure 3: Comparison of Holm's procedure, generic step-down procedure, step-up procedure and knockoffs for controlling the 5-FWER at the 5% significance level. As functions of the magnitude of the nonzero coefficients, the procedures' powers are shown in (a), while the 5-FWER is given in (b), with the grey line denoting the nominal level of 5%. Each point is an average over 2000 simulations.

Figure 3 shows the power of the knockoff procedure above all alternative procedures in the low- to middle-power regimes, while it actually has slightly less power in the very high-power regime, corresponding to a signal-to-noise ratio $\|\boldsymbol{\beta}\|^2/\sigma^2 > 350$. This reversal can be explained by the fact that with non-orthogonal columns and a not-extremely-sparse $\boldsymbol{\beta}$, the Lasso will not perfectly select all signal variables before the non-signal variables, even when the signal-to-noise ratio is extremely high (Su et al., 2015). As such, the Lasso-based $\boldsymbol{W}$ statistic used in knockoffs never achieves a power of 1; this phenomenon could be remedied by using one of the least-squares-based $\boldsymbol{W}$ mentioned in Barber and Candès (2015). The 5-FWER of all four procedures is again largely unaffected by the coefficient

magnitude.

6 Real data experiment

In this section, we apply our method to a data set on HIV drug resistance. Specifically, the data set, described and analyzed in Rhee et al. (2006) and also used in the original knockoffs paper Barber and Candès (2015), contains genotype information from samples of HIV Type 1, along with drug resistance measurements for 16 drugs across three classes. The three classes are protease inhibitors (PI), nucleoside reverse transcriptase inhibitors (NRTI), and nonnucleoside reverse transcriptase inhibitors (NNRTI), each of which has its own set of samples. Drug resistance was measured as the log-fold-increase of resistance as compared to a control, and the genetic information comes as single nucleotide polymorphisms (SNPs), so that the design matrix is binary with each column representing the presence or absence of a minor allele at a given locus.

In order to analyze the data, some cleaning was required. In particular, some samples do not have resistance measurements for some of the drugs, so these samples were removed on a drug-by-drug basis. Also, some SNPs have so few mutations that either their effect would be too hard to detect, or their inclusion actually causes rank-deficiency in the design matrix. As such, for each drug we only included polymorphisms with at least five mutations present in the culled sample; this was the minimum required to ensure all design matrices were full-rank.

We compare our knockoffs procedure to the step-down, step-up, and Holm procedures, as well as to the original knockoffs procedure for controlling FDR at level q. As k-FWER is often used as an exploratory analysis, and to make analysis comparable with knockoffs for FDR control, we set $\alpha = 0.5$ (FDR controls a mean, and with $\alpha = 0.5$, k-FWER controls a

Table 1: Multiple testing procedures applied to HIV drug resistance data sets

Drug	Type	Samples	SNPs	FDR ko	k-FWER ko	Step-down	Step-up	Holm
APV	PI	767	164	19/29	**10/10**	14/18	**14/15**	14/17
ATV	PI	328	104	22/28	**18/19**	18/20	**14/14**	17/19
IDV	PI	825	165	25/42	15/17	17/21	17/20	17/20
LPV	PI	515	141	17/18	**13/14**	**17/18**	**13/13**	**14/14**
NFV	PI	842	166	26/40	20/22	17/21	16/18	17/21
RTV	PI	793	163	20/26	**18/18**	17/23	15/17	15/20
SQV	PI	824	164	20/31	19/29	16/21	15/18	15/19
X3TC	NRTI	629	216	4/6	5/7	6/9	**5/6**	6/8
ABC	NRTI	623	216	16/35	16/31	8/11	8/11	8/11
AZT	NRTI	626	216	15/21	13/17	13/21	10/14	11/18
D4T	NRTI	625	216	15/26	13/21	**11/12**	**10/11**	**10/11**
DDI	NRTI	628	216	2/2	**5/5**	8/13	7/9	8/12
TDF	NRTI	351	148	6/6	**8/8**	9/11	**7/8**	**9/10**
DLV	NNRTI	730	231	10/25	10/16	11/25	11/20	11/22
EFV	NNRTI	732	236	11/21	11/19	10/17	10/16	10/16
NVP	NNRTI	744	236	10/23	8/13	7/15	7/12	7/13
Average Number of True Discoveries				14.9	12.6	12.4	11.2	11.8
			2-FWER	0.81	0.63	0.88	0.63	0.81

Summary: For each procedure, we report the number of true positives and the number of total discoveries, separated by a slash. Among the k-FWER-controlling procedures, entries are bold when fewer than $k = 2$ false discoveries are made. At the end of the table we report summary statistics for each procedure. ko stands for knockoffs.

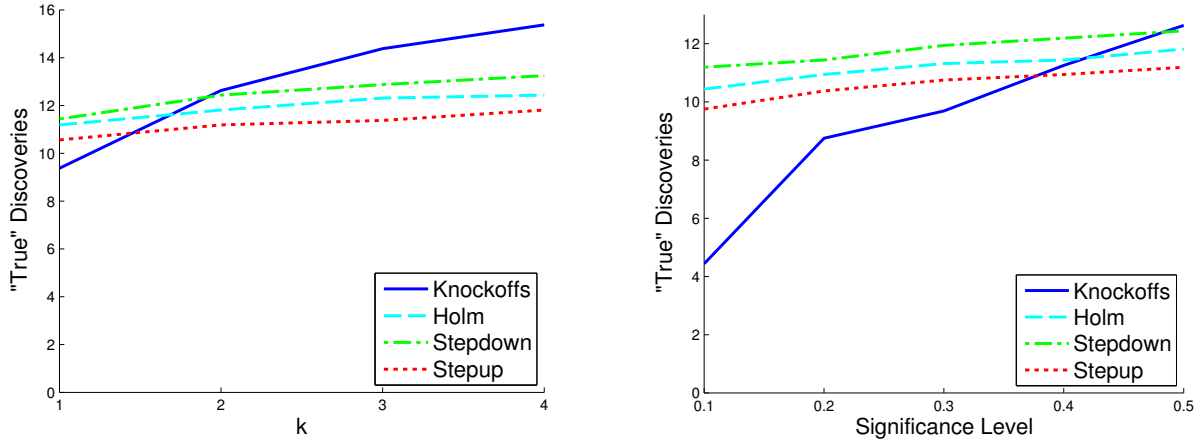
median). We set $k = 2$ and $q = 0.2$, and ran all five procedures on all 16 drugs, the results of which are summarized in Table 1.

Although the ground truth is unknown in this case, there exists an approximate ground truth from treatment-selected mutation (TSM) panels (Rhee et al., 2005). These panels list mutations that were found to be statistically significantly more frequent in virus samples from individuals treated with a drug in that class than samples from individuals who had not. Thus in our experiment evaluation, we consider a SNP discovery for a given drug to be true if it has a mutation listed in the TSM panel for that drug's class.

The table shows the number of total discoveries and false discoveries made by each method on each data set. As suspected, FDR-controlling knockoffs was more powerful than any of the k-FWER-controlling procedures, but is harder to interpret as it never makes a very large number of discoveries, and thus the FDP may be quite different from q. The remaining procedures have varying levels of 2-FWER, but recall that the error rates reported are likely to be overestimates, as there may be important SNPs that the TSM panels missed. Still, we see that on this data set, the step-down and Holm procedures commit more 2-familywise errors than knockoffs, while the step-up procedure has over 10% less power than knockoffs.

To better-understand the trade-offs with the choice of k and α, we also compared the average number of "true" discoveries (discoveries confirmed by TSM panels) as a function of k and α for all four k-FWER-controlling procedures. Figure 4 shows the relative performance of knockoffs improving as the error control (k and/or α) relaxes. To understand why this is, note that the least-squares-p-value-based procedures all have fairly constant power as k and α vary, because there is a group of variables with very strong signals whose p-values are so small that they are always rejected, almost regardless of k and α. But as discussed with regard to Figure 3, the Lasso path does not always perfectly order the very

strong signals first, so knockoffs using the Lasso $\boldsymbol{W}$ statistic loses some power to reject very strong signals. However, as k and/or α increase, knockoffs easily finds the strong signals and more while the other procedures become increasingly conservative. As evidenced in Figure 4, knockoffs quickly overtakes the available alternatives when k and α are not chosen too stringently, so that a large number of rejections are made. Because of this, we expect the advantages of knockoffs to be especially pronounced in large exploratory analyses.



(a) (b)

Figure 4: Average number of TSM-panel-confirmed ("true") discoveries of k-FWER-controlling procedures (a) as a function of k for $\alpha = 0.5$ and (b) as a function of α for $k = 2$.

7 Discussion

This work leaves a number of important avenues open for future research. First, we mentioned in Section 3 a number of methods that translate k-FWER-controlling procedures

into procedures for controlling the FDX. Investigating the best such method could yield a powerful method for controlling another important Type I error rate. Second, Barber and Candès (2015) mention in passing the possibility of multiple knockoffs, i.e., constructing $m \geq 1$ sets of knockoffs and replacing the one-bit p-values corresponding to the χ_j's with $m+1$-discretized p-values. In the setting of FDR control, one can search over many one-bit p-values and need only consider what fraction, on average, may be false discoveries. However to control the k-FWER, one must keep track of every false discovery, and we may expect the extra resolution of multiple knockoffs to provide more power to distinguish true discoveries from false ones. Lastly, we feel the knockoffs framework is still a largely untapped resource for generating multiple testing procedures. The investigation of alternative W_j statistics for ordering variables, and the extension to other regression settings such as logistic regression and higher-dimensional problems ($p > n$) are all important open subjects.

We have presented a novel method for controlling the k-FWER in the context of linear regression. Knockoffs requires no knowledge of the noise variance and implicitly takes into account the exact dependence structure of the problem, allowing it to provide considerable power improvements over state-of-the-art alternatives in a range of settings. This, along with its intuitive justification and ease of computation, makes knockoffs a useful practical tool for multiple hypothesis testing.

References

Barber, R. F. and E. J. Candès (2015, 10). Controlling the false discovery rate via knockoffs. *Ann. Statist. 43*(5), 2055–2085.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical

and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), pp. 289–300.

Benjamini, Y. and D. Yekutieli (2001, 08). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist. 29*(4), 1165–1188.

Bogdan, M., E. van den Berg, C. Sabatti, W. Su, and E. J. Candès (2015). SLOPE—adaptive variable selection via convex optimization. *The Annals of Applied Statistics 9*(3), 1103.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics 41*(6), 2786–2819.

Delattre, S. and E. Roquain (2013). New procedures controlling the false discovery proportion via Romano-Wolf's heuristic. *arXiv preprint arXiv:1311.4030*.

Dunn, O. J. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association 56*(293), 52–64.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407–499.

Genovese, C. and L. Wasserman (2004). A stochastic process approach to false discovery control. *The Annals of Statistics*, 1035–1061.

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

Gordon, A., G. Glazko, X. Qiu, and A. Yakovlev (2007, 06). Control of the mean number of false discoveries, bonferroni and stability of multiple testing. *Ann. Appl. Stat. 1*(1), 179–190.

Guo, W., L. He, and S. K. Sarkar (2014, 06). Further results on controlling the false discovery proportion. *Ann. Statist. 42*(3), 1070–1101.

Guo, W. and J. P. Romano (2007). A generalized Sidak-Holm procedure and control of generalized error rates under independence. *Statistical Applications in Genetics and Molecular Biology 6*(1).

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika 75*(4), 800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*(2), pp. 65–70.

Hommel, G. and T. Hoffmann (1988). Controlled uncertainty. In P. Bauer, G. Hommel, and E. Sonnemann (Eds.), *Multiple Hypothesenprfung / Multiple Hypotheses Testing*, Volume 70 of *Medizinische Informatik und Statistik*, pp. 154–161. Springer Berlin Heidelberg.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research 15*(1), 2869–2909.

Karlin, S. and Y. Rinott (1980). Classes of orderings of measures and related correlation inequalities. i. multivariate totally positive distributions. *Journal of Multivariate Analysis 10*(4), 467 – 498.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2013). Exact post-selection inference with the Lasso. *arXiv preprint arXiv:1311.6238*.

Lehmann, E. L. and J. P. Romano (2005, 06). Generalizations of the familywise error rate. *Ann. Statist. 33*(3), 1138–1154.

Lockhart, R., J. E. Taylor, R. J. Tibshirani, and R. Tibshirani (2014, 04). A significance test for the Lasso. *Ann. Statist. 42*(2), 413–468.

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

Meng, X., J. Wang, and X. Wu (2014). Multiple comparisons controlling expected number of false discoveries. *Communications in Statistics - Theory and Methods 43*(13), 2830–2843.

Rhee, S.-Y., W. J. Fessel, A. R. Zolopa, L. Hurley, T. Liu, J. Taylor, D. P. Nguyen, S. Slome, D. Klein, M. Horberg, J. Flamm, S. Follansbee, J. M. Schapiro, and R. W. Shafer (2005). HIV-1 protease and reverse-transcriptase mutations: Correlations with antiretroviral therapy in subtype B isolates and implications for drug-resistance surveillance. *Journal of Infectious Diseases 192*(3), 456–465.

Rhee, S.-Y., J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, and R. W. Shafer (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences 103*(46), 17355–17360.

Romano, J. P. and A. M. Shaikh (2006, 08). Stepup procedures for control of generalizations of the familywise error rate. *Ann. Statist. 34*(4), 1850–1873.

Romano, J. P. and M. Wolf (2007, 08). Control of generalized error rates in multiple testing. *The Annals of Statistics 35*(4), 1378–1408.

Su, W., M. Bogdan, and E. J. Candès (2015). False discoveries occur early on the Lasso path. *arXiv preprint arXiv:1511.01957*.

Su, W. and E. J. Candès (2015). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *arXiv preprint arXiv:1503.08393*.

van der Laan, M. J., S. Dudoit, and K. S. Pollard (2004). 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).

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association 62*(318), 626–633.

Westfall, P. H. and S. S. Young (1989). p value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association 84*(407), 780–786.