University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2006 *Paper* 171

A Method to Increase the Power of Multiple Testing Procedures Through Sample Splitting

Daniel Rubin* Sandrine Dudoit†
Mark J. van der Laan‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley, daniel.rubin@fda.hhs.gov

†Division of Biostatistics, School of Public Health, University of California, Berkeley, sandrine@stat.berkeley.edu

‡Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

A Method to Increase the Power of Multiple Testing Procedures Through Sample Splitting

Daniel Rubin, Sandrine Dudoit, and Mark J. van der Laan

Abstract

Consider the standard multiple testing problem where many hypotheses are to be tested, each hypothesis is associated with a test statistic, and large test statistics provide evidence against the null hypotheses. One proposal to provide probabilistic control of Type-I errors is the use of procedures ensuring that the expected number of false positives does not exceed a user-supplied threshold. Among such multiple testing procedures, we derive the "most powerful" method, meaning the test statistic cutoffs that maximize the expected number of true positives. Unfortunately, these optimal cutoffs depend on the true unknown data generating distribution, so could never be used in a practical setting. We instead consider splitting the sample so that the optimal cutoffs are estimated from a portion of the data, and then testing on the remaining data using these estimated cutoffs. When the null distributions for all test statistics are the same, the obvious way to control the expected number of false positives would be to use a common cutoff for all tests. In this work, we consider the common cutoff method as a benchmark multiple testing procedure. We show that in certain circumstances the use of estimated optimal cutoffs via sample splitting can dramatically outperform this benchmark method, resulting in increased true discoveries, while retaining Type-I error control. This paper is an updated version of the work presented in Rubin et al. (2005), later expanded upon by Wasserman and Roeder (2006).

1 Introduction

The ingredients for a general type of multiple testing problem are as follows.

- Data: Suppose we observe a random sample $\mathcal{X} = \{X_i\}_{i=1}^n$ of n i.i.d. random variables $X_i \sim P$, where P is an unknown data generating distribution, and $P \in \mathcal{P}$ for $\mathcal{P}$ a statistical model.

- Null Hypotheses: We wish to use the observed data $\mathcal{X}$ to test M null hypotheses $\{H_{0,m} : P \in \mathcal{P}_m \subseteq \mathcal{P}\}_{m=1}^M$ concerning P, where $\mathcal{P}_1, ..., \mathcal{P}_M$ are submodels of $\mathcal{P}$. Let $\mathcal{H}_0$ denote the indices of the true null hypotheses, so that $m \in \mathcal{H}_0$ if and only if hypothesis m is true, meaning $P \in \mathcal{P}_m$, and note that $\mathcal{H}_0$ of course is not a random variable and depends on the unknown data generating distribution P.

- Test Statistics: Test statistics $T_m = T_m(\mathcal{X})$ are functions of the observed data associated with each of the M hypotheses, constructed so that large values of a statistic provides evidence against the corresponding null hypothesis.

- Cutoffs: A multiple testing procedure is based on a vector of cutoffs $(c_1, ..., c_M)$ such that a null hypothesis is rejected when the test statistic exceeds the cutoff value, or $\{T_m > c_m\}$. In this work, we will consider the data, null hypotheses, and test statistics to be given, and attempt to determine an appropriate cutoff vector.

A false positive is the type of error said to occur if a test rejects when the null hypothesis is actually true. Let

$$FP = \sum_{m \in \mathcal{H}_0} I(T_m > c_m) \tag{1}$$

denote the total number of false positives, and note that this is a random variable whose distribution depends on the unknown data generating distribution P.

Much of the multiple testing literature has concerned the control of Type-I error measures, meaning the development of methods meant to stochastically limit the number of false positives. Two commonly used error measures are the family-wise error (FWE) and the false discovery rate (FDR), defined as

$$FWE(P) = P(FP \geq 1) \tag{2}$$

and

$$FDR(P) = E_P[\frac{FP}{\max(R,1)}] \tag{3}$$

where

$$R = \sum_{m=1}^{M} I(T_m > c_m) \tag{4}$$

is the total number of rejections. A multiple testing procedure would be said to control the FWE or FDR at a user-supplied level α if $FWE(P) \leq \alpha$ or $FDR(P) \leq \alpha$ for all P belonging to the model $\mathcal{P}$.

In this paper we focus on tests controlling a different Type-I error measure, the expected number of false positives (EFP), given by

$$EFP(P) = E_P[FP] = \sum_{m \in \mathcal{H}_0} P(T_m > c_m). \tag{5}$$

Specifically, we consider sets of tests that can control the expected number of false positives at user defined levels α, $0 \leq \alpha \leq M$, so that

$$EFP(P) \leq \alpha \text{ for all } P \in \mathcal{P}. \tag{6}$$

There are several advantages to this error measure. Unlike the family-wise error and false discovery rate, we note from (5) that $EFP(P)$ only depends on the marginal distributions of the M test statistics. We will see that the EFP can be controlled without assuming independence of the test statistics, as is often done with methods built to control the FDR. Neither will it be necessary to estimate the unknown and possibly complex test statistic dependence structure, as is done in certain resampling schemes for multiple testing, discussed in Dudoit et al. (2004) and Pollard and van der Laan (2004). Further, the EFP is more flexible than the family-wise error because with a large number of tests M, procedures controlling the FWE at a level $\alpha < 1$ will have virtually no power to detect alternatives. The EFP can provide less stringent control of the number of false positives when set to higher levels α, and can consequently lead to a large number of rejections even when there are many hypotheses being tested simultaneously. It is also immediate from Markov's inequality that by controlling the EFP at level α, we can control the tail behavior of the distribution of false positives, because for any $t > 0$,

$$P(FP \geq t) \leq E_P[FP]/t \leq \alpha/t. \tag{7}$$

The special case of $t = 1$ shows that control of the EFP at level α also controls the family-wise error at level α.

In order to bound the expected number of false positives, we will assume that the test statistics obey a natural requirement, called the *null domination condition*. Suppose the existence of a probability distribution $P_0 \in \mathcal{P}$, such that if a null hypothesis is true under the data generating distribution P, the associated test statistic is stochastically larger under P_0 than under P. That is, let $S_m(\cdot) = P(T_m > \cdot)$ denote the survivor function of test statistic T_m under P, and $S_m^0(\cdot) = P_0(T_m > \cdot)$ the corresponding survivor function under P_0. The null domination condition can be stated formally as

$$S_m(\cdot) \leq S_m^0(\cdot) \text{ for all } m \in \mathcal{H}_0. \tag{8}$$

Consider a vector of test statistics cutoffs $(c_1, ..., c_M)$ such that the expected number of rejections R under P_0 is controlled at level α, so that

$$E_{P_0}[R] = \sum_{m=1}^{M} P_0(T_m > c_m) \leq \alpha. \tag{9}$$

The null domination condition (8) then implies

$$EFP(P) = \sum_{m \in \mathcal{H}_0} P(T_m > c_m) \leq \sum_{m \in \mathcal{H}_0} P_0(T_m > c_m) \leq \sum_{m=1}^{M} P_0(T_m > c_m) \leq \alpha,$$

so that the cutoffs control $EFP(P)$ for any $P \in \mathcal{P}$ at the desired level as in (6). Consequently, we will let $\mathcal{C}$ denote the set of cutoff vectors satisfying (9), and restrict our study to multiple testing procedures whose cutoffs belong to this set. We note that $\mathcal{C}$ could theoretically depend on the choice of P_0 yielding the null domination condition.

The obvious way to choose cutoffs in $\mathcal{C}$ would be to use a common quantile for all test statistics. That is, we could let $c_m = (S_m^0)^{-1}(\alpha/M)$ and observe that $E_{P_0}[R] = \alpha$. For a choice of P_0 such that $S_m^0(\cdot)$ does not depend on m, as could occur if all test statistics were based on pivotal statistics and had the same null distribution, this reduces to using a common cutoff $c_m = (S^0)^{-1}(\alpha/M)$ for all M test statistics. In this case, we call the procedure the *common cutoff method*. This method is a very natural way to simultaneously test hypotheses while controlling the expected number of false positives, and it is intuitively appealing to use a common cutoff for all tests with the same null distribution, if nothing is known about differences among alternatives between tests.

In this work, we consider the common cutoff method as a benchmark procedure, and ask "When can we do better?"

To answer this question, it is first necessary for us to define what we mean by "better." Unlike univariate testing, there are many ways in which we could define the power of a multiple testing procedure. Two extreme examples would be to define the power as the probability of rejecting all true positives, or the probability of rejecting at least one true positive. In this paper, we follow the approach of Storey (2005), and Wasserman and Roeder (2006) and define the power of a set of cutoffs in $\mathcal{C}$ as the expected number of true positives (TP) under the data generating distribution P, or

$$E_P[TP] = \sum_{m \notin \mathcal{H}_0} P(T_m > c_m) = \sum_{m \notin \mathcal{H}_0} S_m(c_m). \tag{10}$$

When given data, null hypotheses, and test statistics, the only remaining ingredient necessary to perform multiple tests is the vector of cutoffs. With the expected number of false and true positives as measures of Type-I error and power, it is natural to inquire about the most powerful cutoff vector that controls the EFP at level α. If it were known, we could use it to test at a desired EFP level and optimize the expected number of valid discoveries (true positives). Because we have restricted to cutoff vectors in $\mathcal{C}$ in order to control the EFP, we formally define the *optimal cutoffs* as the vector

$$c(P) = (c_1(P), ..., c_M(P)) = \max_{c \in \mathcal{C}}^{-1} E_P[TP] = \max_{c \in \mathcal{C}}^{-1} \sum_{m \notin \mathcal{H}_0} S_m(c_m). \tag{11}$$

In section 2 we derive a simple analytical characterization of the optimal cutoffs $c(P)$ as a function of the marginal test statistic survivor functions $S_1(\cdot), ..., S_M(\cdot)$. We then specialize to the case where all M tests are one-sided z-tests with a standard Normal null distribution, and a shifted Normal alternative distribution. As one might expect, the optimal cutoffs we derive depend heavily on the true unknown data generating distribution P, and thus could only be used by an oracle, rather than a practicing statistician or data analyst. We attempt to overcome this difficulty in section 3, by considering the case where the data $\mathcal{X}$ is built from a sample of independent and identically distributed random variables $\{X_1, ..., X_n\}$. We discuss the procedure of using a small fraction of the data to estimate the optimal cutoffs, and the remaining data to perform the testing under these estimated cutoffs. In section 5 we report simulation results comparing the power of various sample splitting techniques with the common cutoff method and the oracle power using the optimal cutoffs.

Estimated shift alternatives can sometimes be pooled across tests to gain strength, and the resulting sample splitting procedure can then dramatically outperform the common cutoff technique, while retaining the user-specified EFP control. We consider the practical utility of our sample splitting approach, and conclude that the method could potentially be used to increase power in a variety of real-world applications.

2 Optimal Cutoffs

Below we present the main theorem of the paper, which identifies the cutoffs solving the optimization problem in (11).

Theorem 1. *In the setting of section 1, fix a level* $0 < \alpha < M - |\mathcal{H}_0|$ *for control of the expected number of false positives. Let* $S_m^0(\cdot) = P_0(T_m > \cdot)$ *be the survival function of test statistic* T_m *under the null distribution* P_0 *and* $S_m(t) = P(T_m > t)$ *be the survival function of* T_m *under the true data generating distribution* P*. Assume* $S_m(t) \geq S_m^0(t)$*,* $m \notin \mathcal{H}_0$*, and the null domination condition* $S_m(t) \leq S_m^0(t)$*,* $m \in \mathcal{H}_0$ *as in (8). Let*

$$\begin{aligned} c_m(\lambda) &= \infty \text{ if } m \in \mathcal{H}_0 \text{ (meaning test } m \text{ never rejects)} \\ \text{and} \quad c_m(\lambda) &= \max_x{}^{-1} S_m(x) - \lambda S_m^0(x) \text{ if } m \notin \mathcal{H}_0. \end{aligned}$$

For each $m \notin \mathcal{H}_0$*, if* $x = c_m(\lambda)$ *is finite and* S_m^0 *and* S_m *are twice differentiable with densities* $f_m(t) \equiv -\frac{d}{dt} S_m(t)$ *and* $f_m^0(t) \equiv -\frac{d}{dt} S_m^0(t)$*, respectively, then:*

$$\begin{aligned} -f_m(x) + \lambda f_m^0(x) &= 0 \\ -f_m'(x) + \lambda f_m^{0\prime}(x) &< 0, \end{aligned} \tag{12}$$

where $f_m', f_m^{0\prime}$ *are the derivatives of* f_m, f_m^0*.*

If $R(c)$ *and* $TP(c)$ *denote the numbers of rejections and true positives incurred by the set of tests using cutoff vector* $c = (c_1, ..., c_M)$*, and* $\lambda > 0$ *solves* $E_{P_0}[R(c(\lambda))] - \alpha = 0$*, then* $E_P[TP(c(P))] = E_P[TP(c(\lambda))]$*, for* $c(P)$ *the optimal cutoff vector defined in (11). Thus, if* $c(P)$ *is unique, then* $c(P) = c(\lambda)$*.*

Proof. The proof of this theorem requires a generalization of the Lagrange multiplier method to handle situations in which the constrained maximization problem

is solved by points on the edge of the parameter space (i.e. ∞ or $-\infty$) so that the derivative cannot be set equal to zero.

Define the function $g : \mathbb{R}^{M+1} \to \mathbb{R}$ by

$$\begin{aligned} g(c,\lambda) &= E_P[TP(c)] - \lambda(E_{P_0}[R(c)] - \alpha) \\ &= \sum_{m \notin \mathcal{H}_0} S_m(c_m) - \lambda \left(\sum_{m=1}^{M} S_m^0(c_m) - \alpha \right). \end{aligned}$$

By the fact that $g(\cdot,\cdot)$ is an additive function in functions of c_m it follows that $c(\lambda) = \max_c^{-1} g(c,\lambda)$, where $c_m(\lambda) = \max_x^{-1}\{S_m(x) - \lambda S_m^0(x)\}$ for $m \notin \mathcal{H}_0$ and $c_m(\lambda) = \max_x^{-1}(-\lambda S_m^0(x))$ for $m \in \mathcal{H}_0$.

Since λ solves $\sum_{m=1}^{M} S_m^0(c_m(\lambda)) - \alpha = 0$, this implies

$$E_P[TP(c(\lambda))] = g(c(\lambda),\lambda) \geq g(c(P),\lambda) = E_P[TP(c(P))].$$

By definition of $c(P)$, we also have $E_P[TP(c(P))] \geq E_P[TP(c(\lambda))]$. This proves $E_P[TP(c(P))] = E_P[TP(c(\lambda))]$.

If $\lambda < 0$, then $c_m(\lambda) = -\infty$ for $m \notin \mathcal{H}_0$ which means that $\sum_{m=1}^{M} S_m^0(c_m(\lambda)) = M - |\mathcal{H}_0|$. Therefore, if $\alpha < M - |\mathcal{H}_0|$, then we can exclude $c(\lambda)$, $\lambda < 0$, as possible solutions. If $\lambda > 0$, then $c_m(\lambda) = \infty$ for $m \in \mathcal{H}_0$, as stated in the theorem. Finally, (12) is just applying that a finite maximum of a twice differentiable function satisfies that the derivative at the maximum equals zero and the second derivative at the maximum is negative. $\square$

Because Theorem 1 is stated in a fairly abstract manner, it may enhance understanding to contemplate its application to the following multiple testing problem. Consider real-valued parameters $\mu_m = \mu_m(P)$ estimated by $\hat{\mu}_m = \hat{\mu}_m(\mathcal{X})$ for $1 \leq m \leq M$, and the issue of testing the null hypotheses $\mu_m = \mu_{m,0}$ against the alternatives that $\mu_m > \mu_{m,0}$. When the data $\mathcal{X}$ are composed of n i.i.d. measurements, asymptotic normality results are available from many areas of statistics showing that for commonly used parameters μ_m and estimators $\hat{\mu}_m$,

$$\sqrt{n}(\hat{\mu}_m - \mu_m) \sim N(0,\sigma^2). \tag{13}$$

Such results are well known for sample moments and correlations, as well as estimated coefficients in regression models, or maximum likelihood estimators under regularity

conditions. For an asymptotic variance estimator $\hat{\sigma}^2$, we can write the test statistic as,

$$T_m \equiv \sqrt{n}(\hat{\mu}_m - \mu_{m,0})/\hat{\sigma} = \sqrt{n}(\hat{\mu}_m - \mu_m)/\hat{\sigma} + \sqrt{n}(\mu_m - \mu_{m,0}))/\hat{\sigma}. \tag{14}$$

When the parameter μ_m is allowed to vary with sample size n, such that $(\mu_m - \mu_{m,0}) \sim hn^{-1/2}$ and $\hat{\sigma}$ is consistent for the asymptotic variance σ, we can obtain a nondegenerate limiting distribution $T_m \sim N(d_m, 1)$ under the local alternative, for $d_m = h/\sigma$. Hence, it is natural to form one-sided tests for μ_m by using T_m given in (14) as a test statistic.

Under the asymptotic approximation $T_m \sim N(d_m, 1)$, the testing problem becomes a Gaussian shift problem. We can consider testing the one sided hypothesis $H_{0,m}$: $d_m \leq 0 \leftrightarrow \mu_m \leq \mu_{m,0}$ against the alternative that $d_m > 0 \leftrightarrow \mu_m > \mu_{m,0}$. Below we present a corollary to Theorem 1, giving the optimal cutoffs in closed form when testing multiple shifts.

Theorem 1 teaches us that the optimal cutoffs $c(P)$ for one-sided testing in the Gaussian shift problem can be solely expressed in terms of the solution of one maximization problem:

$$\phi(d, \lambda) \equiv \max_x^{-1} S^0(x - d) - \lambda S^0(x), \tag{15}$$

where $S^0(\cdot)$ denotes the $N(0, 1)$ survivor function. Specifically, we have

$$\begin{aligned} c_m(\lambda) &= \infty \text{ if } d_m \leq 0 \\ c_m(\lambda) &= \phi(d_m, \lambda) \text{ if } d_m > 0 \end{aligned}$$

and λ is obtained by solving

$$0 = \sum_{m=1}^{M} S^0(c_m(\lambda)) - \alpha.$$

The univariate maximization problem (15) is handled by 1) setting the derivative equal to zero, 2) if a solution exists, then we check if it is a maximum (i.e. second derivative is negative) and 3) if no solution exists, then the derivative is either always positive or always negative.

Corollary 1. *Consider the setting of Theorem* 1 *with* $S_m(x) = S^0(x - d_m)$ *and* $f_m(x) = f^0(x - d_m)$*, where* $f^0(x) = 1/\sqrt{2\pi}\exp(-x^2/2)$ *and* S^0 *are the respective density and survival functions of a standard Normal distribution,* $m = 1, \ldots, M$.

Define

$$g(d, \lambda) = \begin{cases} \frac{\log(\lambda) + 0.5d^2}{d} & \text{if } d > 0 \\ \infty & \text{if } d \leq 0. \end{cases}$$

We have that $E_P[TP(c(P))] = E_P[TP(c(\lambda))]$*, where 1)* $c_m(\lambda) = g(d_m, \lambda)$ *and 2)* $\lambda > 0$ *is the unique solution of* $\sum_{m=1}^{M} S^0(c_m(\lambda)) - \alpha = 0$*.*

Proof. We apply Theorem 1. Thus $c_m = \infty$ if $d_m \leq 0$. We need to find $c_m(\lambda) = \max_x^{-1}\{S^0(x-d_m)-\lambda S^0(x)\}$ for $m \notin \mathcal{H}_0$. Setting the derivative of $\{S^0(x-d_m)-\lambda S^0(x)\}$ equal to zero yields that $c_m(\lambda) = g(d_m, \lambda)$ for $m \notin \mathcal{H}_0$. In addition, at this solution we have that the second derivatives $\frac{d^2}{dc_m^2}\{S^0(c_m - d_m) - \lambda S^0(c_m)\} = -d_m f^0(c_m - d_m) < 0$ for $m \notin \mathcal{H}_0$ are strictly negative if and only if $d_m > 0$. Thus $c_m(\lambda)$ is indeed the wished unique maximum. Now, the application of Theorem 1 yields the proof of the first result about $c(\lambda)$. □

One intuitive implication of this corollary is that when all alternatives are thought to be some common value, the optimal cutoffs are in fact equal to the common cutoffs. However, when it is believed that there are differences among alternatives across tests, an essential complication limiting the applicability of Theorem 1 to practical problems is that the shifts d_m would be unknown before performing the testing, and in fact knowledge of these shifts would eliminate the need of testing in the first place. We attempt to partially rectify this difficulty in the following section.

Note that Theorem 1 can also be applied to form optimal cutoffs when using as test statistics $|T_m|$, which would correspond to the two-sided test of $H_{0,m} : d_m = 0 \leftrightarrow \mu_m = \mu_{m,0}$ against the alternative that $d_m \neq 0 \leftrightarrow \mu_m \neq \mu_{m,0}$. However, we have omitted these results to focus on the one-sided problem for illustrative purposes.

3 Sample Splitting Approach

While the optimal cutoffs of Corollary 1 could not be used without special knowledge, improvements in power relative to the common cutoff procedure are possible in some circumstances when the observed data $\mathcal{X}$ consists of a series of i.i.d. measurements $X_i \sim P$. Our basic approach will be to split the observations into two parts, use one part for performing the tests, and the other part for estimating the optimal cutoffs associated with these tests. To ease exposition, we will focus for the remainder of this section on the problem of one-sided testing for means. However, it is straightforward to apply our sample splitting approaches to the general testing situations described in (13) and (14), and two-sided tests present no essential difficulties.

For $\mathcal{X} = \{X_1, ..., X_n\}$ a sample of n i.i.d. realizations, suppose that X_i is the random vector $X_i = (X_{i,1}, ..., X_{i,M})$, where we can think of $X_{i,m}$ as representing a measurement of covariate m on subject i. Suppose all M "covariates" have been standardized to have unit variance, and let μ_m denote $E[X_{i,m}]$. We consider the problem of testing the null hypotheses $H_{0,m} : \mu_m \leq 0$ against the alternatives that $\mu_m > 0$. For instance, one could imagine that $X_{i,m} = Y_{i,m} - Z_{i,m}$ for Y and Z representing measurements on matched treatment and control subjects, and an interest in performing multiple paired t-tests to search for the set of covariates having elevated means in the treatment group. We we will let $\bar{X}_m$ denote the sample mean for covariate m, and use as test statistics $T_m \equiv \sqrt{n}\bar{X}_m$. The asymptotic approximation that $T_m \sim N(d_m, 1)$ for $d_m = \sqrt{n}\mu_m$ is justified by the Central Limit Theorem, so that the multiple testing problem can be reduced to the Gaussian shift problem as discussed in the previous section. The shifts d_m, and consequently the optimal cutoffs of Corollary 1, thus depend on the true unknown covariate means μ_m.

3.1 Failure of the Naive Plug-in Method

A simple idea for approximating the optimal cutoff procedure would be to use the usual test statistics $T_m = \sqrt{n}\bar{X}_m$, but estimate the shifts $d_m = \sqrt{n}\mu_m$ with the plug-in estimators $\hat{d}_m = \sqrt{n}\bar{X}_m = T_m$, and then use the optimal cutoffs as defined in Corollary 1 under these estimated shifts. Unfortunately, algebraic manipulation yields that hypothesis m is then rejected when

$$\begin{aligned}
\{T_m > \log(\lambda)\hat{d}_m^{-1} + \frac{1}{2}\hat{d}_m, \hat{d}_m > 0\} &= \{T_m > \log(\lambda)T_m^{-1} + \frac{1}{2}T_m, T_m > 0\} \\
&= \{T_m^2 > 2\log(\lambda), T_m > 0\} \\
&= \{T_m > \sqrt{\max(2\log(\lambda), 0)}\}.
\end{aligned}$$

Therefore, each of the M tests would have a common rejection region, and choosing λ to control the expected number of false positives reduces this plug-in technique exactly to the common cutoff procedure.

3.2 Sample Splitting

Rather than using all n random variables in the sample to estimate the shift alternatives, we can implement a nontrivial procedure by using only a "held-out" set of

observations. Consider splitting the n observations $\mathcal{X} = \{X_1, ..., X_n\}$ into two samples of sizes n_1 and n_2, and letting $\bar{X}_{m,1}$ and $\bar{X}_{m,2}$ denote the sample means for covariate m built from the two respective groups. The test statistics built from the second sample are then $T_{m,2} \equiv \sqrt{n_2}\bar{X}_{m,2}$, with limiting distribution $T_{m,2} \sim N(d_m, 1)$ for $d_m = \sqrt{n_2}\mu_m$.

One sample splitting technique that attempts to gain power relative to the common cutoff method is to estimate these shifts with $\hat{d}_m = \sqrt{n_2}\bar{X}_{m,1}$. The estimated shifts remain unbiased for the true shift alternatives, but now become independent of the test statistics $T_{m,2}$. The idea is then to estimate the optimal cutoffs (for tests $T_{m,2}$) as in Corollary 1, by using the $\hat{d}_m$ as surrogates for the shifts d_m. The procedure reduces to rejecting hypothesis m when for an appropriate choice of λ,

$$\{T_{m,2} > \log(\lambda)\hat{d}_m^{-1} + \frac{1}{2}\hat{d}_m,\ \hat{d}_m > 0\}. \tag{16}$$

Note that the cutoff for $T_{m,2}$ in (16) is monotone in λ, and hence increasing the tuning parameter λ corresponds to decreasing the allowed Type-I error, measured by the expected number of false positives. For $S^0(\cdot)$ the standard Normal survivor function, choosing λ to solve

$$\sum_{m=1}^{M} I(\hat{d}_m > 0) S^0(\log(\lambda)\hat{d}_m^{-1} + \frac{1}{2}\hat{d}_m) = \alpha \tag{17}$$

controls the EFP of the overall procedure at level α. To see this, let $\lambda^\star$ be the value solving (17), and recall that the sample splitting ensures $\hat{d}_m \perp T_{m,2}$. Observe that when $m \in \mathcal{H}_0 \leftrightarrow d_m \leq 0$ and $T_{m,2} \sim N(d_m, 1)$, we have the null domination condition that $P(T_{m,2} > t) \leq S^0(t)$ for all t. Letting P_0 denote the distribution on the observed data such that $T_{m,2} \sim N(0, 1)$ for all m, it immediately follows from conditioning on the $\hat{d}_m$ values that,

$$
\begin{aligned}
E_P[FP] &\leq E_{P_0}[FP] \\
&\leq E_{P_0}[R] \\
&= \sum_{m=1}^{M} E_{P_0}[I(\hat{d}_m > 0) I(T_{m,2} > \log(\lambda^\star)\hat{d}_m^{-1} + \frac{1}{2}\hat{d}_m)] \\
&= E_{P_0}[\sum_{m=1}^{M} I(\hat{d}_m > 0) P_0(T_{m,2} > \log(\lambda^\star)\hat{d}_m^{-1} + \frac{1}{2}\hat{d}_m | \hat{d}_m)] \\
&= E_{P_0}[\sum_{m=1}^{M} I(\hat{d}_m > 0) S^0(\log(\lambda^\star)\hat{d}_m^{-1} + \frac{1}{2}\hat{d}_m)] \\
&\leq E_{P_0}[\alpha] = \alpha. \qquad (18)
\end{aligned}
$$

Wasserman and Roeder (2006) have suggested splitting the dataset, again estimating the shifts with $\hat{d}_m = \sqrt{n_2}\bar{X}_{m,1}$ and estimating the optimal cutoffs as in (16), but instead using the full data statistic $T_m = \sqrt{n}\bar{X}_m$ for testing. Because of the decomposition $T_m = a\hat{d}_m + bT_{m,2}$ for $a = n_1/\sqrt{nn_2}$ and $b = \sqrt{n_2/n}$, their procedure rejects hypothesis m on the event

$$
\begin{aligned}
& \{T_m > \log(\lambda)\hat{d}_m^{-1} + \frac{1}{2}\hat{d}_m,\ \hat{d}_m > 0\} \\
= \ & \{a\hat{d}_m + bT_{m,2} > \log(\lambda)\hat{d}_m^{-1} + \frac{1}{2}\hat{d}_m,\ \hat{d}_m > 0\} \\
= \ & \{T_{m,2} > \frac{\log(\lambda)\hat{d}_m^{-1} + \frac{1}{2}\hat{d}_m - a\hat{d}_m}{b},\ \hat{d}_m > 0\} \qquad (19)
\end{aligned}
$$

for an appropriate choice of λ. It can be verified as in (18) that the procedure controls the expected number of false positives at level α when choosing λ to solve

$$
\sum_{m=1}^{M} I(\hat{d}_m > 0) S^0(\log(\lambda)\hat{d}_m^{-1} b^{-1} + \frac{1}{2}\hat{d}_m b^{-1} - a\hat{d}_m b^{-1}) = \alpha. \qquad (20)
$$

Observe that the cutoff $\frac{\log(\lambda)\hat{d}_m^{-1} + \frac{1}{2}\hat{d}_m - a\hat{d}_m}{b}$ is a function of $\hat{d}_m$. Thus, as with using the rejection region of (16), this procedure can also be interpreted as an attempt to use $T_{m,2}$ for testing after constructing cutoffs based upon $\{\hat{d}_m\}_{m=1}^M$. Wasserman and Roeder have concluded that using the full data test statistic T_m instead of $T_{m,2}$ can increase power, although our simulation results in section 4 do not necessarily replicate this result. For future denotation in this work, we will refer to the procedures using the rejection regions of (16) and (19) as the EOC1 and EOC2 procedures (estimated optimal cutoffs methods 1 and 2).

3.3 Prior Information and Pooled Alternatives

Another approach to the testing problem, which is not always as impractical as it first appears, is to guess the shifts $d_m = \sqrt{n}\mu_m$ from prior information, and then estimate the optimal cutoffs of the form $c(\lambda) = \log(\lambda)d_m^{-1} + \frac{1}{2}d_m$. Here λ can be chosen to control the Type-I constraint as in (18), and it is important to stress that poor guessed shifts do not compromise the expected number of false positives. Preliminary results of ours suggest that using guessed optimal cutoffs can yield a more powerful procedure than using common cutoffs, even if the cutoffs are not guessed accurately, so long as the ordering is roughly correct. In addition, Wasserman and Roeder (2006) have reported theoretical robustness results for procedures equivalent to guessing optimal cutoffs.

More generally, prior information can be combined with sample splitting to increase the power of testing procedures. Suppose that withheld data is used to estimate shifts $\{\hat{d}_m\}_{m=1}^M$ as in section 3.2, and consider the following scenarios where one could pool shift estimates across different tests.

1. The tests are ordered so that it is thought $m \to \mu_m$ is a smooth function of the test index. This could easily occur in practice if data was collected across time or at different spatial locations, test results were desired at every time or location, and one guessed that the μ_m varied smoothly in time or space. In this case, the estimated shift alternatives could possibly be made more accurate by smoothing the $\hat{d}_m$ as a function of the index m.

2. It could be guessed that the μ_m obeyed a certain approximate ordering. In this case, one could use isotonic regression on the $\{m, \hat{d}_m\}_{m=1}^M$ pairs in an attempt to improve accuracy.

3. One could imagine that the μ_m only realized a small number of values across the M covariates, if it were thought the covariates were clustered into a small number of unknown groups, and measurements in each group were highly correlated. It might then be possible to apply clustering techniques to the covariates, or even the $\{\hat{d}_m\}_{m=1}^M$ values themselves, and pool estimated shifts within each cluster.

After enhancing shift estimates through pooling, we would then implement the estimated optimal cutoff procedures of (16) and (19) exactly as before. As with simply

guessing shifts, pooling withheld-data estimated shifts based on unfounded assumptions will not mar the EFP Type-I error rate.

4 Simulations

We compared the power of the sample splitting methods of (16) and (19) introduced in the previous section with that of the common cutoff method and unattainable optimal cutoff method of section 2. Each simulated dataset $\mathcal{X}$ consisted of $n = 100$ i.i.d. observations, where each observation was a vector of measurements for $M = 1000$ simulated covariates. We performed 1000 tests of the null hypothesis that covariate m had mean $\mu_m \leq 0$ against the alternative $\mu_m > 0$, as discussed previously. All testing was performed to control the expected number of false positives at level $\alpha = 0.05$, which by (7) also controlled the family-wise error at this standard level. The first 9950 hypotheses were true nulls, with the covariates having mean zero. The final 50 hypotheses were false nulls, with the covariate mean being set to a common value μ that we varied as a simulation parameter. We considered μ between 0.2 and 0.6, in increments of 0.05. These choices covered the range of values over which the probability of a true positive rejection with common cutoffs varied from roughly zero to roughly one. Each entry of the $n \times M = 100 \times 1000$ data matrix $\mathcal{X}$ was simulated from an independent Gaussian distribution with variance one.

For each choice of shift μ we performed 1000 simulations. In each simulation, we implemented the common cutoff method and optimal cutoff procedure (which depended on knowledge of the alternatives). In addition, we implemented both the EOC1 and EOC2 sample splitting procedures of section 3.2, holding out a proportion p of the $n = 100$ samples to estimate the alternatives. We examined values of $p \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

From the estimated shifts $(\hat{d}_1, ..., \hat{d}_M)$ formed in the EOC1 and EOC2 sample splitting procedures, we also implemented a *smoothed EOC1 cutoff* method. For this technique, we assumed that the 1000 hypotheses were ordered in such a manner that $m \to d_m$ was a smooth function of m. Recall that the ordering was such that the final 50 covariates had mean μ while the others had mean zero, so this was in fact a step function. Of course, such an assumption would be unwarranted in most multiple testing applications, but we considered this procedure because it demonstrated the gains one

could incur by pooling shift estimates across tests if special knowledge was available. We then used the **smooth.spline()** function in R to estimate the alternative shifts by the fitted values of the smoothed "data" $\{m, \hat{d}_m\}_{m=1}^{1000}$, and used these estimated alternative shifts to estimate the optimal cutoffs.

Hence, for each of the nine choices of shift μ (and six choices of hold-out proportion p for sample splitting techniques), we performed 1000 simulations computing the multiple testing results for five procedures (common cutoffs, oracle cutoffs, EOC1 sample splitting, EOC2 sample splitting, smoothed cutoff sample splitting). In this paper we have been defining the power of a set of tests through $E_P[TP]$, the expected number of true positives. When reporting simulation results we used the scaled version $\frac{E_P[TP]}{\text{Number of true positives}} = \frac{E_P[TP]}{50}$ for simplicity, which corresponded in our example to the probability that a true positive was rejected. The power of the common cutoff procedure was calculated analytically. For the remaining four multiple testing procedures, we estimated this scaled power for each choice of μ and p by computing the proportion of true positives rejected across all 1000 simulations.

The results are displayed in figures $1-6$. We can clearly see that the optimal cutoff method provides the most power, and any other result would immediately raise a red flag. For these simulations, the optimal cutoff power substantially exceeded that of the common cutoffs, giving credence to the hope that a large gain over the benchmark standard (common cutoffs) could be possible if the optimal cutoffs could be guessed or estimated accurately. It is apparent that both the EOC1 and EOC2 sample splitting procedures are less powerful than using common cutoffs for all choices of simulation parameters. It seems that both of these sample splitting procedures degrade as the hold-out proportion p grows, with the degradation being more rapid for the EOC2 cutoffs. The EOC1 and EOC2 power track each other closely for small values of p, with the EOC2 procedure being slightly more powerful for $p \leq 0.2$. Wasserman and Roeder (2006) have reported similar simulations in which the EOC2 cutoffs outperform the EOC1 cutoffs for $p = 1/2$, although both methods are outperformed by the common cutoff procedure, whose power can be computed analytically and compared with their results. Most interesting to us are the results for the smoothed cutoffs, because this sample splitting technique is actually more powerful than the benchmark common cutoffs for small p. For $p \leq 0.1$, the power of this procedure was not significantly less than that the optimal cutoff procedure.

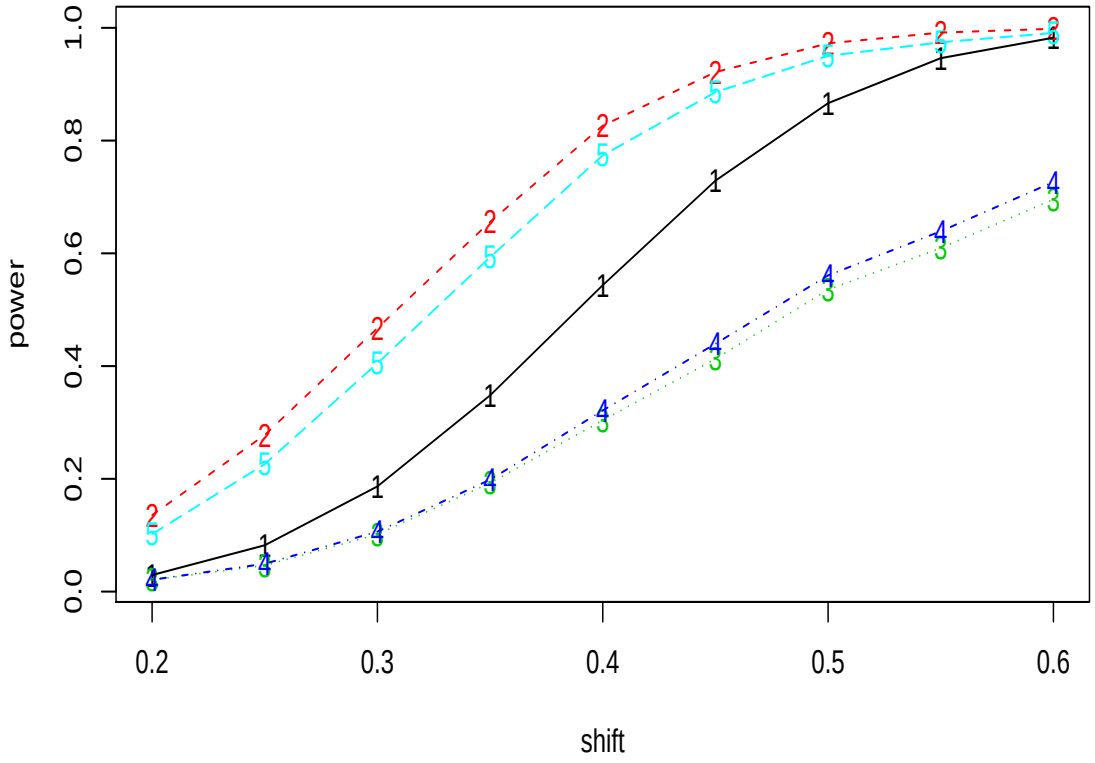

Figure 1: The power of various multiple testing methods is shown. The shift μ is reported on the x-axis. The y-axis represents $E_P[TP]/50$, or the probability that a true positive is rejected, which was common across all true positives in the simulations. Each line represents the power of a multiple testing method as a function of the shift. Lines labeled $1, 2, 3, 4, 5$ denote the powers of the common cutoff method, optimal cutoff method (only available to an oracle), EOC1 sample splitting cutoffs, EOC2 sample splitting cutoffs, and the smoothed EOC1 sample splitting cutoffs. Here $p = 0.05$ is the proportion of the data split to estimate the optimal cutoffs for the sample splitting procedures. The other five figures are based on different choices of p.

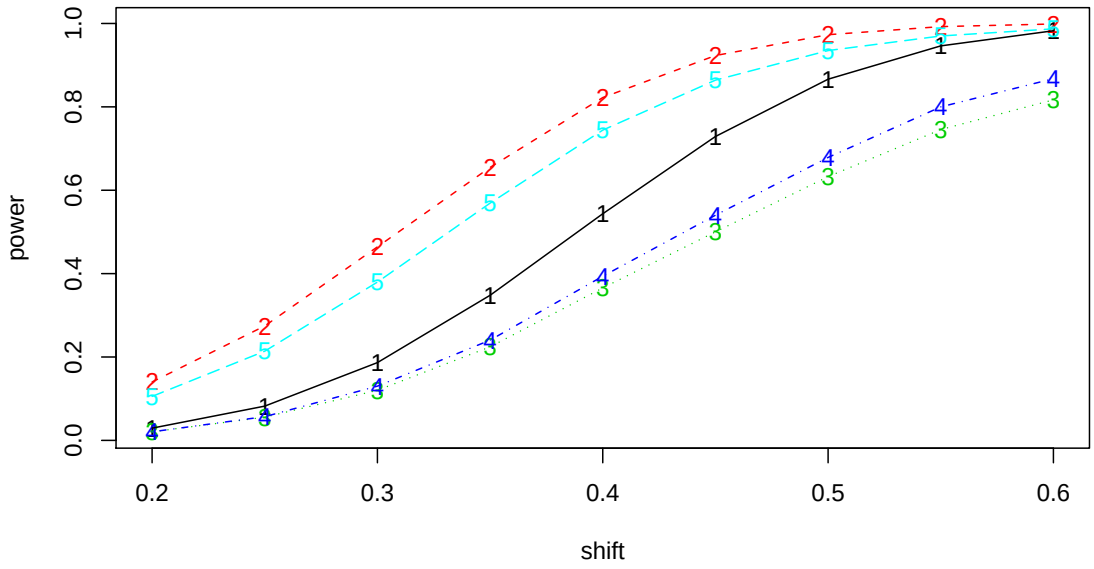

Figure 2: $p = 0.1$. 1 = common cutoff, 2 = oracle cutoffs, 3 = EOC1 sample splitting cutoffs, 4 = EOC2 sample splitting cutoffs, 5 = smoothed EOC1 sample splitting cutoffs

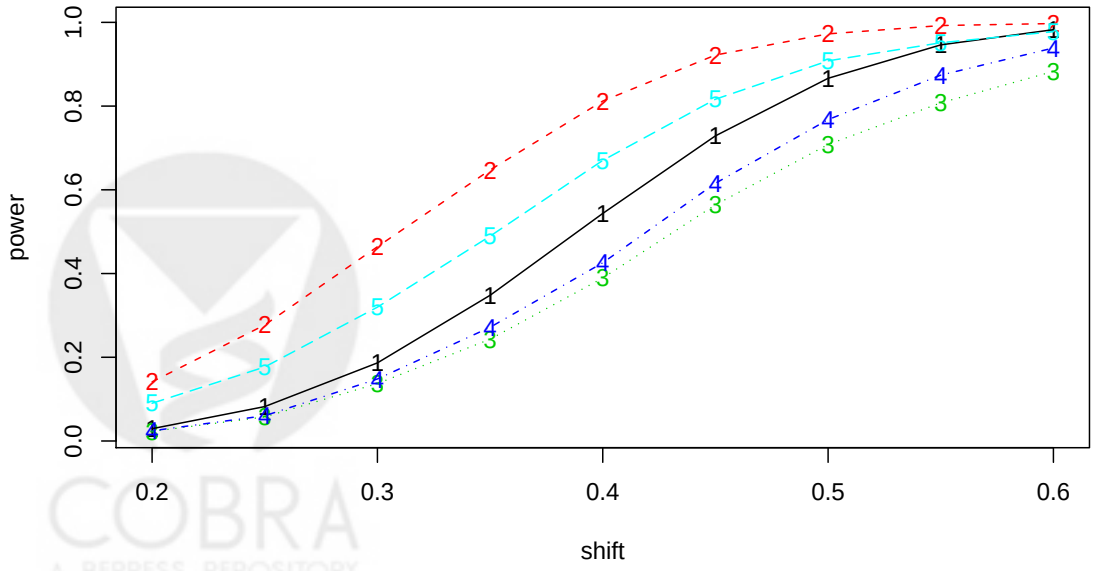

Figure 3: $p = 0.2$. 1 = common cutoff, 2 = oracle cutoffs, 3 = EOC1 sample splitting cutoffs, 4 = EOC2 sample splitting cutoffs, 5 = smoothed EOC1 sample splitting cutoffs

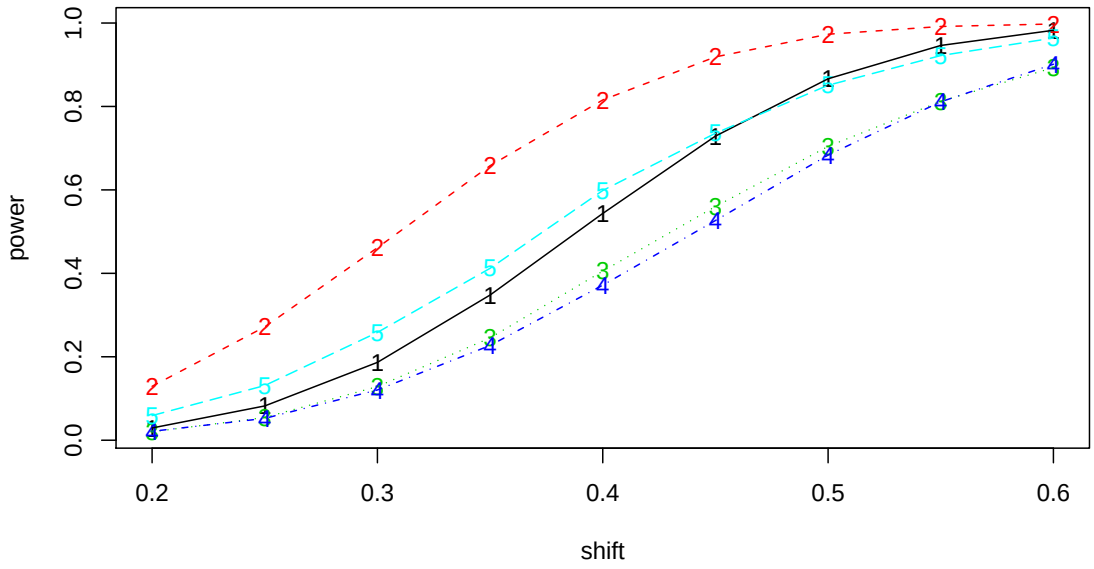

Figure 4: $p = 0.3$. 1 = common cutoff, 2 = oracle cutoffs, 3 = EOC1 sample splitting cutoffs, 4 = EOC2 sample splitting cutoffs, 5 = smoothed EOC1 sample splitting cutoffs

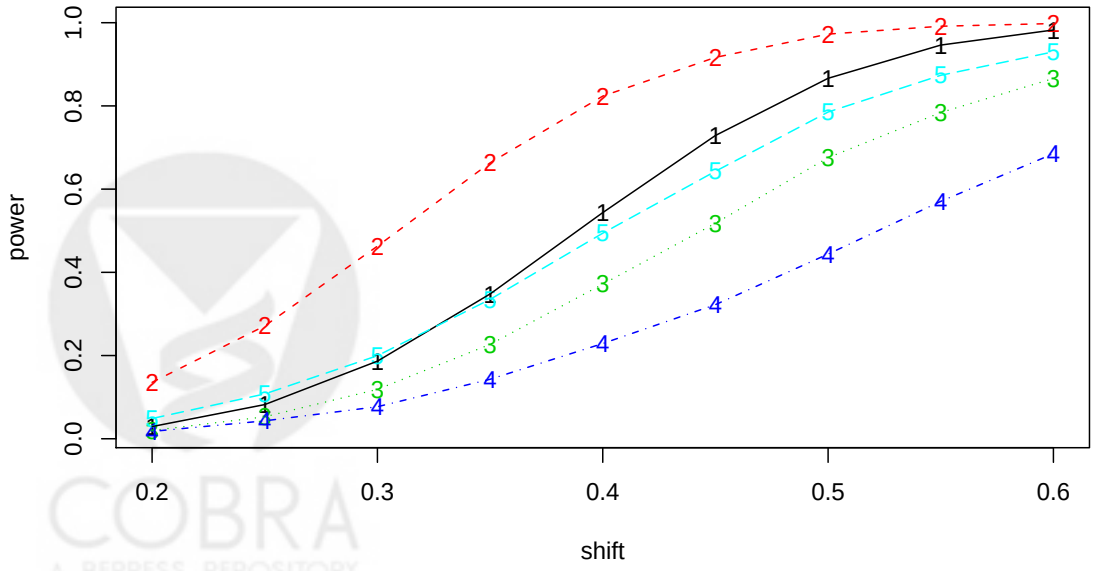

Figure 5: $p = 0.4$. 1 = common cutoff, 2 = oracle cutoffs, 3 = EOC1 sample splitting cutoffs, 4 = EOC2 sample splitting cutoffs, 5 = smoothed EOC1 sample splitting cutoffs

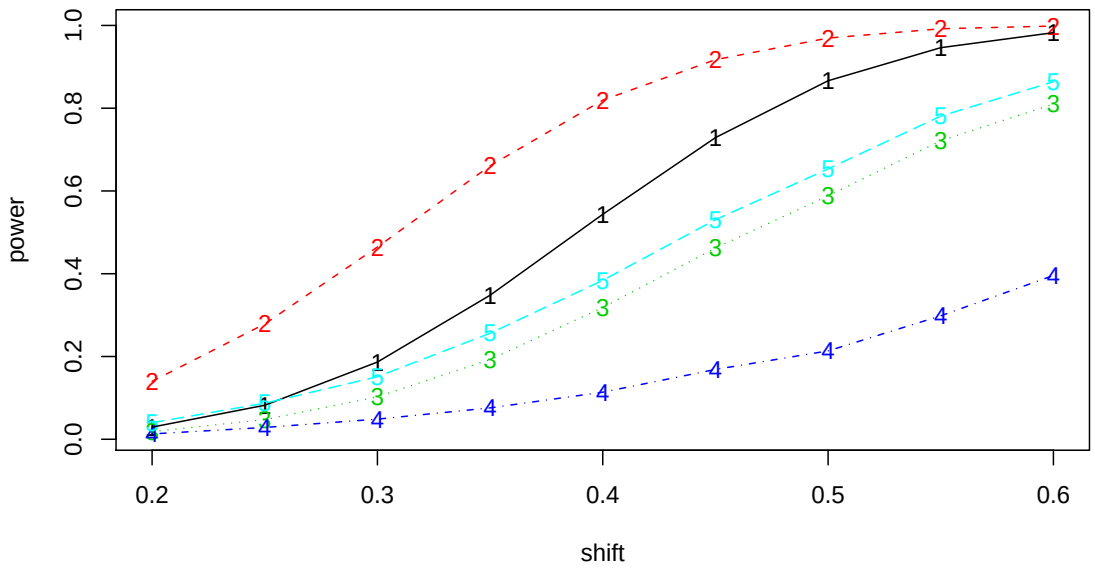

Figure 6: $p = 0.5$. 1 = common cutoff, 2 = oracle cutoffs, 3 = EOC1 sample splitting cutoffs, 4 = EOC2 sample splitting cutoffs, 5 = smoothed EOC1 sample splitting cutoffs

5 Discussion

The optimal cutoffs introduced in section 2 could only be used by an oracle, and provide more expected true positives than could any practical method controlling the expected number of false positives below a nominal level. Ideally, one might hope to use an adaptive procedure that could estimate the optimal cutoffs, and reasonably approximate their optimal power for a wide range of $P \in \mathcal{P}$. One would posit that such a method would be more powerful than the nonadaptive common cutoff against broad classes of alternatives.

Discouragingly, we do not believe in the existence of such an adaptive procedure that would be more generally applicable and powerful than the common cutoff technique, and could be used as a default multiple testing procedure. The simulation results of section 4 demonstrate that straightforward sample splitting approximations to the optimal cutoff method in fact can be much less powerful than the common cutoff procedure, even in extremely simple testing scenarios. The general problem with any sample splitting technique is that withholding a moderately sized proportion p of the

data for optimal cutoff estimation can drastically reduce the power of tests built from the remaining data. Knowledge or approximate knowledge of the optimal cutoffs for these tests built from $n(1-p)$ samples does not compensate for the power reduction due to the sample size decrease.

An essential rejoinder to this pessimistic outlook is that sample splitting procedures can considerably improve upon the common cutoff technique when alternatives are accurately estimated using only a small proportion of withheld data, such as when $p = 0.1$. For this to be possible, it may be necessary for estimated alternatives to somehow be pooled across hypotheses. Such an approach clearly would not be possible in all multiple testing situations, but as mentioned in section 3, could be applicable if alternative estimates are available from previous experiments or a meta-analysis, the hypotheses can be indexed so that the alternatives are thought to be ordered, the shift alternatives are thought to be an unknown smooth function of the hypothesis index, or if it is thought that the alternatives are clustered into groups. Our intuition is that guessing or accurately pooling estimated alternatives can be done in a variety of scientific contexts. We view the contribution of this work as the development of a set of tools that allow one to apply prior beliefs to a small amount of withheld data, in a fashion that can potentially yield an increased number of valid discoveries.

References

Dudoit, S., van der Laan, M., and Pollard, K.S. (2004). Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates. Statistical Applications in Genetics and Molecular Biology, **3**, No. 1, Article 13.

Pollard, K.S. and van der Laan, M. (2004). Choice of a null distribution in resampling-based multiple testing. Journal of Statistical Planning and Inference, **125**, 85-100.

Rubin, D., van der Laan, M., and Dudoit S. (2005). Multiple testing procedures which are optimal at a simple alternative. Technical report 171, Division of Biostatistics, School of Public Health, University of California, Berkeley.

Storey J.D. (2005). The optimal discovery procedure: A new approach to simultaneous significance testing. UW Biostatistics Working Paper Series, Working Paper 259.

Wasserman, L. and Roeder, K. (2006). Weighted Hypothesis Testing. Personal communication.