A regression framework for the proportion of true null hypotheses

Simina M. Boca, Jeffrey T. Leek

December 30, 2015

Abstract

The false discovery rate is one of the most commonly used error rates for measuring and controlling rates of false discoveries when performing multiple tests. Adaptive false discovery rates rely on an estimate of the proportion of null hypotheses among all the hypotheses being tested. This proportion is typically estimated once for each collection of hypotheses. Here we propose a regression framework to estimate the proportion of null hypotheses conditional on observed covariates. We provide both finite sample and asymptotic conditions under which this covariate-adjusted estimate is conservative - leading to appropriately conservative false discovery rate estimates. We demonstrate the viability of our approach through simulation and an application to estimating different prior probabilities for a range of sample sizes and allele frequencies in a GWAS meta-analysis.

1 Introduction

Multiple testing is a ubiquitous issue in modern scientific studies. Microarrays (Brown, 1995), next-generation sequencing (Shendure and Ji, 2008), and high-throughput metabolomics make it possible to simultaneously test the relationship between hundreds or thousands of biomarkers and an exposure or outcome of interest. These problems have a common structure consisting of a collection of variables, or features, for which measurements are obtained on multiple samples, with a hypothesis test being performed for each feature.

When performing thousands of hypothesis tests, the most widely used framework for controlling for multiple testing is the false discovery rate. For a fixed unknown parameter μ, and testing a single null hypothesis $H_0 : \mu = \mu_0$ versus some alternative hypothesis, for example, $H_1 : \mu = \mu_1$, the null hypothesis may either truly hold or not for each feature. Additionally, the test may lead to H_0 either being rejected or not being rejected. Thus, when performing m hypothesis tests for m different unknown parameters, Table 1 shows the total number of outcomes of each type, using the notation from Benjamini and Hochberg (1995). We note that U, T, V, and S, and as a result, also $R = V + S$, are random variables, while m_0, the number of null hypotheses, is fixed and unknown.

Table 1: Outcomes of testing multiple hypotheses.

	Fail to reject null	Reject null	Total
Null true	U	V	m_0
Null false	T	S	$m - m_0$
	$m - R$	R	m

The false discovery rate (FDR), introduced in Benjamini and Hochberg (1995), is the expected fraction of false discoveries among all discoveries. The false discovery rate depends on the overall fraction of null hypotheses, namely $\pi_0 = \frac{m_0}{m}$. This proportion can also be interpreted as the *a priori* probability that a null hypothesis is true, π_0.

When estimating the FDR, incorporating an estimate of π_0 can result in a more powerful procedure compared to the original Benjamini and Hochberg (1995) procedure; moreover, as m increases, the estimate of π_0 improves, which means that the power of the multiple-testing approach does not necessarily decrease when more hypotheses are considered (Storey, 2002).

Most modern adaptive false discovery rate procedures rely on an estimate of π_0 using the data of all tests being performed. But additional information, in the form of meta-data, may be available to

aid the decision about whether to reject the null hypothesis for a particular feature. One example is in set-level inference, including gene-set analysis, where each set has a different fraction of false discoveries. Another is in meta-analyses where each feature may have a different sample size and hence different power. Adjusting for covariates independent of the data conditional on the truth of the null hypothesis has also been shown to improve power in RNA-seq, eQTL, and proteomics studies (Ignatiadis et al., 2015).

In this paper, we build on the work of Benjamini and Hochberg (1995), Efron et al. (2001), and Storey (2002) and the more recent work of Scott et al. (2014), which frames the concept of *FDR regression* and extends the concepts of FDR and π_0 to incorporate covariates, represented by additional meta-data. Our focus will be on estimating the covariate-specific π_0. We will also show how this can be seen as an extension of our work in Boca et al. (2013) on set-level inference, where we developed an approach which focused on estimating the fraction of non-null variables in a set, introducing the idea of "atoms," non-overlapping sets based on the original annotations, and the concept of the "atomic FDR." We provide a more direct approach to estimating the covariate-specific π_0 and provide a number of theoretical frequentist properties for our estimator.

The remainder of the paper is organized as follows. In Section 2 we provide an overview of an example application application to GWAS meta-analysis. In Section 3, we review the definitions of FDR and π_0 and extend π_0 to consider conditioning on a specific covariate. In Section 4, we discuss estimation and inference procedures for the covariate-specific π_0 in the FDR regression framework. In Section 5, we consider special cases within the FDR regression framework, including how the no covariates case and the case where the features are partitioned return us to the "standard" estimation procedures. In Section 6, we explore some theoretical properties of the estimator, including showing that, under certain conditions, it is a conservative estimator of the covariate-level π_0, its variance has an upper bound which can be calculated from the given data, and it is an asymptotically conservative estimator of the covariate-level π_0. In Section 7 and Section 8, we consider simulations and an analysis of GWAS data. Finally, Section 9 provides our statement of reproducibility and Section 10 provides the discussion.

2 Example application: adjusting for sample size and allele frequency in GWAS meta-analysis

As we have described, there are a variety of situations where meta-data could be valuable for improving estimation of the prior probability a hypothesis is true or false. Here we consider an example from the meta-analysis of data from genome-wide association studies (GWAS) for body mass index (Locke et al., 2015).

In a GWAS, data are collected for a large number of genomic loci called single nucleotide polymorphisms (SNPs) Hirschhorn and Daly (2005). Each person has a copy of the DNA at each SNP inherited from their mother and from their father. At each locus there are usually one of two types of DNA, called alleles, that can be inherited, denoted A and a. In general, A refers to the variant that is more common in the population being studied and a to the variant that is less common. Each person has a genotype for that SNP of the form AA, Aa, or aa. The number of copies of a, commonly called the minor allele - is assumed to follow a binomial distribution.

In a GWAS, each individual has the alleles for hundreds of thousands of SNPs measured along with some outcomes of interest like body mass index (BMI). Then each SNP is tested for association with the outcome in a regression model and p-values are calculated for the association. GWAS studies have grown to sample sizes of tens of thousands of individuals. But the largest studies consist of meta-analyses combining multiple studies Neale et al. (2010); Hirschhorn and Daly (2005). In these studies, the sample size may not be the same for each SNP, for example if different individuals are measured with different technologies which measure different SNPs. As a result, the sample size could be considered as a meta-data covariate.

A second covariate of interest could be the frequency of the minor allele a in the population. The power to detect associations increases with increasing minor allele frequency. This is related to the idea that logistic regression is more powerful for outcomes that occur frequently.

Here we consider data from the Genetic Investigation of ANthropometric Traits (GIANT) consortium, specifically the genome-wide association study for BMI (Locke et al., 2015). The GIANT

consortium performed a meta-analysis of 329,224 individuals measuring 2,555,510 SNPs and tested each for association with BMI. Here we will consider using a regression model to estimate a prior probability for association for each SNP conditional on the SNP-specific sample size and allele frequency.

3 Covariate-specific π_0

We will now review the main concepts behind the FDR and the *a priori* probability that a null hypothesis is true, and consider the extension to the covariate-specific FDR, and the covariate-specific *a priori* probability. A natural mathematical definition of the FDR would be:

$$FDR = E\left[\frac{V}{R}\right].$$

However, R is a random variable that can be equal to 0, so the definition that is generally used is:

$$FDR = E\left[\frac{V}{R}\middle| R > 0\right] Pr(R > 0), \tag{1}$$

namely the expected fraction of false discoveries among all discoveries multiplied by the probability of making at least one rejection.

We index the m null hypotheses being considered by $1 \leq i \leq m$: $H_{01}, H_{02}, \ldots, H_{0m}$. For each i, the corresponding null hypothesis H_{0i} can be considered as being about a binary parameter θ_i, such that:

$$\theta_i = 1(H_{0i} \text{ true}).$$

Thus, assuming that θ_i are identically distributed, the *a priori* probability that a feature is null is:

$$\pi_0 = Pr(\theta_i = 1). \tag{2}$$

We now extend the definition of π_0 to consider conditioning on a covariate $\mathbf{X}_i$, where $\mathbf{X}_i$ is a column vector of length c, possibly with $c = 1$:

Definition 1

$$\pi_0(\mathbf{x}_i) = Pr(\theta_i = 1|\mathbf{X}_i = \mathbf{x}_i).$$

4 Estimation and inference for covariate-specific π_0 in the FDR regression framework

We will now discuss the estimation and inference procedures for $\pi_0(\mathbf{x}_i)$ in a FDR regression framework. We assume that a hypothesis test is performed for each i, summarized by a p-value P_i. At a given threshold $0 < \lambda < 1$, we consider the random variables Y_i:

$$Y_i = 1(P_i > \lambda). \tag{3}$$

Thus, Y_i is a dichotomous random variable that is 1 when the null hypothesis H_{0i} is not rejected at an α-level of λ and 0 when it is rejected. Thus, $m - R = \sum_{i=1}^{m} Y_i$ for a fixed, given λ. The null p-values will come from a Uniform(0,1) distribution, while the p-values for the features from the alternative

$$G(\lambda) = Pr(P_i \leq \lambda|\theta_i = 0). \tag{4}$$

The major assumption we make moving forward is that *conditional on the null, the p-values do not depend on the covariates.* In Theorem 2, we prove the major result we will use to derive the estimator for $\pi_0(\mathbf{x}_i)$.

Theorem 2 *Suppose that m hypotheses tests are performed and that conditional on the null, the p-values do not depend on the covariates. Then:*

$$E[Y_i|\mathbf{X}_i = \mathbf{x}_i] \quad = \quad (1-\lambda)\pi_0(\mathbf{x}_i) + \{1 - G(\lambda)\}\{1 - \pi_0(\mathbf{x}_i)\}.$$

Proof.

$$\begin{aligned} E[Y_i|\mathbf{X}_i = \mathbf{x}_i] &= Pr(P_i > \lambda|\mathbf{X}_i = \mathbf{x}_i) \\ &= Pr(P_i > \lambda|\theta_i = 1, \mathbf{X}_i = \mathbf{x}_i)P(\theta_i = 1|\mathbf{X}_i = \mathbf{x}_i) \\ &+ Pr(P_i > \lambda|\theta_i = 0, \mathbf{X}_i = \mathbf{x}_i)P(\theta_i = 0|\mathbf{X}_i = \mathbf{x}_i). \end{aligned}$$

Then, using the assumption that conditional on the null, the p-values do not depend on the covariates:

$$\begin{aligned} E[Y_i|\mathbf{X}_i = \mathbf{x}_i] &= Pr(P_i > \lambda|\theta_i = 1)P(\theta_i = 1|\mathbf{X}_i = \mathbf{x}_i) + Pr(P_i > \lambda|\theta_i = 0)P(\theta_i = 0|\mathbf{X}_i = \mathbf{x}_i) \\ &= (1-\lambda)\pi_0(\mathbf{x}_i) + \{1 - G(\lambda)\}\{1 - \pi_0(\mathbf{x}_i)\}. \end{aligned}$$

In Corollary 3, we show the corresponding result for the no-covariate case. This result is easy to prove directly, but we consider it as a corrolary to Theorem 2 to show that there are no identifiability problems with the extension to covariates.

Corollary 3 *Suppose that m hypotheses tests are performed and that conditional on the null, the p-values do not depend on the covariates. Then:*

$$E[Y_i] = (1-\lambda)\pi_0 + \{1 - G(\lambda)\}\{1 - \pi_0\}.$$

Proof. Applying the law of iterated expectations:

$$E[Y_i] = E[E[Y_i|\mathbf{X}_i]] = (1-\lambda)E[\pi_0(\mathbf{X}_i)] + \{1 - G(\lambda)\}\{1 - E[\pi_0(\mathbf{X}_i)]\}.$$

We complete the proof by using:

$$\begin{aligned} \pi_0 &= Pr(\theta_i = 1) = \int Pr(\theta_i = 1, \mathbf{X}_i = \mathbf{x})d\nu(\mathbf{x}) \\ &= \int Pr(\theta_i = 1|\mathbf{X}_i)dF_{\mathbf{X}_i} = E[Pr(\theta_i = 1|\mathbf{X}_i)] = E[\pi_0(\mathbf{X}_i)], \end{aligned}$$

where ν is typically either the Lebesgue measure over a subset $\mathbb{R}$ or the counting measure over a subset of $\mathbb{Q}$, and $F_{\mathbf{X}_i}$ is the cumulative distribution function for $\mathbf{X}_i$. Here we are implicitly assuming some distribution for $\mathbf{X}_i$ as well. Everywhere else we are conditioning on $\mathbf{X}$.

We first review the procedure which applies Corollary 3 to lead to the estimator of π_0 for the no-covariate case, which is also used in Storey (2002), then develop a procedure based on Theorem 2 to obtain an estimator of $\pi_0(\mathbf{x})$. Both of them are based on assuming reasonably powered tests and a large enough λ, so that:

$$G(\lambda) \approx 1.$$

Corollary 3 then leads to:

$$\pi_0 \approx \frac{E[Y_i]}{1-\lambda},$$

resulting in:

$$\pi_0 \approx \frac{\frac{\sum_{i=1}^m E[Y_i]}{m}}{1-\lambda}.$$

Using a method-of-moments approach, we consider the estimator:

$$\hat{\pi}_0 = \frac{\frac{\sum_{i=1}^m Y_i}{m}}{1-\lambda} = \frac{m-R}{(1-\lambda)m}, \tag{5}$$

which is used in Storey (2002). Applying the same steps with Theorem 2, we get:

$$\pi_0(\mathbf{x}_i) \approx \frac{E[Y_i|\mathbf{X}_i = \mathbf{x}_i]}{1-\lambda}.$$

We can use a regression framework to estimate $E[Y_i|\mathbf{X}_i = \mathbf{x}_i]$, then estimate $\pi_0(\mathbf{x})$ by:

$$\hat{\pi}_0(\mathbf{x}_i) = \frac{\hat{E}[Y_i|\mathbf{X}_i = \mathbf{x}_i]}{1-\lambda}.$$

We now denote by $\mathbf{Y}$ the random vector of length m with the ith element Y_i and by $\mathbf{X}$ the matrix of dimension $m \times (c+1)$, which has the ith row consisting of $(1\ \mathbf{X}_i^T)$. Moving forward, we will denote by $\mathbf{x}$ the observed values of the random matrix $\mathbf{X}$.

We consider estimators of the form:

$$\begin{aligned} \hat{E}[\mathbf{Y}|\mathbf{X}=\mathbf{x}] &= \mathbf{S}\mathbf{Y}, \\ \hat{E}[Y_i|\mathbf{X}_i=\mathbf{x}_i] &= \mathbf{S}_i^T\mathbf{Y}, \end{aligned} \tag{6}$$

where $\mathbf{S} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-}\mathbf{Z}^T$ for some $m \times p$ matrix $\mathbf{Z}$ with $p < m$ and $rank(\mathbf{Z}) = d \leq p$ and $\mathbf{S}_i^T$ is the ith row of $\mathbf{S}$; in particular, we can have $\mathbf{Z} = \mathbf{X}$ for linear regression or have $\mathbf{Z}$ also include polynomial or spline terms. If $d = p$, then $\mathbf{Z}^T\mathbf{Z}$ is invertible; if $d < p$, one can use any pseudoinverse of $\mathbf{Z}^T\mathbf{Z}$, since the projection matrix is unique. If we assume that the p-values are independent, we can also use bootstrap samples of them to obtain a confidence interval for $\hat{\pi}(\mathbf{x}_i)$. The details for the entire estimation and inference procedure are in Algorithm 1.

4.1 Algorithm 1: Estimation and inference for $\hat{\pi}_0(\mathbf{x}_i)$

a) Obtain the p-values $P_1, P_2, \ldots, P_m$, for the m hypothesis tests.

b) For a given threshold λ, obtain $Y_i = 1(P_i > \lambda)$ for $1 \leq i \leq m$.

c) Choose a design matrix $\mathbf{Z}$, estimate $E[Y_i|\mathbf{X}_i = \mathbf{x}_i]$ by:

$$\hat{E}[Y_i|\mathbf{X}_i=\mathbf{x}_i] = \mathbf{S}_i^T\mathbf{Y},$$

where $\mathbf{S} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-}\mathbf{Z}^T$, and $\pi_0(\mathbf{x}_i)$ by:

$$\hat{\pi}_0(\mathbf{x}_i) = \frac{\hat{E}[Y_i|\mathbf{X}_i=\mathbf{x}_i]}{1-\lambda} = \frac{\mathbf{S}_i^T\mathbf{Y}}{1-\lambda}. \tag{7}$$

d) Take B bootstrap samples of $P_1, P_2, \ldots, P_m$ and calculate the bootstrap estimates $\hat{\pi}^b(\mathbf{x}_i)$ for $1 \leq b \leq B$ using the procedure described above.

e) Form a $1-\alpha$ upper confidence interval for $\hat{\pi}(\mathbf{x}_i)$ by taking the $1-\alpha$ quantile of the $\hat{\pi}^b(\mathbf{x}_i)$ as the upper confidence bound, the lower confidence bound being 0.

5 Special cases for covariate-specific π_0

5.1 No covariates

If we do not consider any covariates, the usual estimator $\hat{\pi}_0$ from Equation 5 can be deduced from applying Algorithm 1 by fitting a linear regression with just an intercept.

5.2 Partioning the features

Now assume that the set of features is partitioned into S sets, namely that a collection of sets $\mathcal{S} = \{A_s : 1 \leq s \leq S\}$ is considered such that all sets are non-empty, pairwise disjoint, and have the set of all the features as their union. Note that the index s does not need to indicate any kind of ordering of the sets. For example, such partioning could be induced by considering all possible atoms resulting from gene-set annotations, or could consist of brain regions of interest in a functional imaging analysis, when considering only the genes or voxels that are annotated (Boca et al., 2013). We can consider this in the covariate framework we developed by taking $\mathbf{x}_i$ to be a vector of length $S-1$, which consists of 0s at all positions with the exception of a value of 1 at the index corresponding to the single set $A_s \in \mathcal{S}$ such that $i \in A_s$, for $1 \leq s \leq S-1$. Set A_S representing the "baseline set," so that $\mathbf{x}_i$ is a vector of length $S-1$ consisting of just 0s if $i \in A_S$. In notation commonly used in linear algebra:

$$\mathbf{x}_i = \begin{cases} \mathbf{e}_s \text{ for } i \in A_s \text{ , given that } 1 \leq s \leq S-1, \\ \mathbf{0} \text{ for } i \in A_S. \end{cases} \tag{8}$$

Taking into account the partition, a natural way of estimating $\pi_0(\mathbf{x}_i)$ is to just apply the estimator $\hat{\pi}_0$ from Equation 5 to each of the S sets:

$$\begin{aligned}\hat{\pi}_0(\mathbf{e}_s) &= \frac{\frac{\sum_{i\in A_s} Y_i}{|A_s|}}{1-\lambda} \text{ for } 1 \le s \le S-1, \\ \hat{\pi}_0(\mathbf{0}) &= \frac{\frac{\sum_{i\in A_S} Y_i}{|A_S|}}{1-\lambda}.\end{aligned}$$

A related idea has been proposed for partitioning hypotheses into sets to improve power Efron (2008). These results can also be obtained by estimating $\hat{\pi}_0(\mathbf{x}_i)$ via Algorithm 1 by fitting a linear regression with an intercept and the covariates $\mathbf{x}_i$.

6 Theoretical results

We now proceed to explore some theoretical properties of the estimator $\hat{\pi}_0(\mathbf{x}_i)$. In what follows, $\mathbf{1}$ is the $m \times 1$ vector consisting of just 1s. We will also use the notation:

$$\pi_0(\mathbf{x}) = \begin{bmatrix} \pi_0(\mathbf{x}_1) \\ . \\ . \\ . \\ \pi_0(\mathbf{x}_m) \end{bmatrix}.$$

Lemma 4 below gives the bias of $\hat{\pi}_0(\mathbf{x}_i)$. Note that $\frac{1-G(\lambda)}{1-\lambda}\{1-\pi_0(\mathbf{x}_i)\} \ge 0$, since $\lambda \le 1, G(\lambda) \le 1$, and $\pi_0(\mathbf{x}_i) \le 1$, $E[\hat{\pi}_0(\mathbf{x}_i)] \ge \pi_0(\mathbf{x}_i)$. The second term could, however, be negative, and depends on the level of non-linearity present in $\pi_0(\mathbf{x}_i)$ and misspecification of the model as encapsulated in the design matrix $\mathbf{Z}$.

Lemma 4 *The bias of $\hat{\pi}_0(\mathbf{x}_i)$ is:*

$$E[\hat{\pi}_0(\mathbf{x}_i)] - \pi_0(\mathbf{x}_i) = \frac{1-G(\lambda)}{1-\lambda}\{1-\pi_0(\mathbf{x}_i)\} + \frac{G(\lambda)-\lambda}{1-\lambda}\{S_i^T\pi_0(\mathbf{x}) - \pi_0(\mathbf{x}_i)\} \tag{9}$$

Proof By Eq. 7:

$$E[\hat{\pi}_0(\mathbf{x}_i)|] - \pi_0(\mathbf{x}_i) = \frac{S_i^T E(\mathbf{Y}|\mathbf{X})}{1-\lambda} - \pi_0(\mathbf{x}_i)$$

Using the result of Theorem 2:

$$\begin{aligned}E[\hat{\pi}_0(\mathbf{x}_i)] - \pi_0(\mathbf{x}_i) &= \frac{S_i^T[(1-\lambda)\pi_0(\mathbf{x}) + \{1-G(\lambda)\}\{\mathbf{1}-\pi_0(\mathbf{x})\}]}{1-\lambda} - \pi_0(\mathbf{x}_i) \\ &= \frac{S_i^T[\{G(\lambda)-\lambda\}\pi_0(\mathbf{x}) + \mathbf{1}\{1-G(\lambda)\}}{1-\lambda} - \pi_0(\mathbf{x}_i) \\ &= S_i^T\mathbf{1}\frac{G(\lambda)-\lambda}{1-\lambda} + \frac{G(\lambda)-\lambda}{1-\lambda}S_i^T\pi_0(\mathbf{x}) - \pi_0(\mathbf{x}_i)\end{aligned}$$

Given that $\mathbf{S} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^-\mathbf{Z}^T$ and that the first column of $\mathbf{Z}$ is $\mathbf{1}, \mathbf{S}_i^T\mathbf{1} = 1$. This is a known result used in linear regression. It can be obtained using the fact that $\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^-\mathbf{Z}^T = \mathbf{Z}_1(\mathbf{Z}_1^T\mathbf{Z}_1)^-\mathbf{Z}_1^T$, where $\mathbf{Z}_1$ is a matrix consisting of d linearly independent columns of $\mathbf{Z}$, including the first column, then applying the formula for the inverse of a block matrix. Thus:

$$\begin{aligned}E[\hat{\pi}_0(\mathbf{x}_i)] - \pi_0(\mathbf{x}_i) &= \frac{G(\lambda)-\lambda}{1-\lambda} + \frac{G(\lambda)-\lambda}{1-\lambda}S_i^T\pi_0(\mathbf{x}) - \pi_0(\mathbf{x}_i) \\ &= \frac{1-G(\lambda)}{1-\lambda}\{1-\pi_0(\mathbf{x}_i)\} + \frac{G(\lambda)-\lambda}{1-\lambda}\{S_i^T\pi_0(\mathbf{x}) - \pi_0(\mathbf{x}_i)\}\end{aligned}$$

Theorem 5 shows that, if the model is correctly specified, i.e. $\pi_0(\mathbf{x}) = \mathbf{Z}\beta$ for some vector β of length $c+1$, then $\hat{\pi}_0(\mathbf{x}_i)$ is a conservative estimate of $\pi_0(\mathbf{x}_i)$.

Theorem 5 *If $\pi_0(\mathbf{x}) = \mathbf{Z}\beta$ for some vector β of length $c+1$, then $\hat{\pi}_0(\mathbf{x}_i)$ is a conservative estimate of $\pi_0(\mathbf{x}_i)$, i.e.:*

$$E[\hat{\pi}_0(\mathbf{x}_i)] \geq \pi_0(\mathbf{x}_i)$$

Proof In this case, using the fact that $\mathbf{S}$ is a projection matrix onto the space spanned by the columns of $\mathbf{Z}$ and therefore $\mathbf{SZ} = \mathbf{Z}$:

$$\mathbf{S}_i^T \pi_0(\mathbf{x}) - \pi_0(\mathbf{x}_i) = \mathbf{S}_i^T \mathbf{Z}\beta - \mathbf{Z}_i\beta = \mathbf{Z}_i\beta - \mathbf{Z}_i\beta = 0,$$

so:

$$E[\hat{\pi}_0(\mathbf{x}_i)] - \pi_0(\mathbf{x}_i) = \frac{1 - G(\lambda)}{1 - \lambda}\{1 - \pi_0(\mathbf{x}_i)\}$$

Remark 6 *If the same π_0 is shared by all the features, i.e. it does not change based on any covariates, then $\hat{\pi}_0$ is a conservative estimate of π_0. This result is also described elsewhere, for example in Storey (2002). We note here that it can also be obtained as a direct consequence of Theorem 5. Theorem 5 also applies to the case where the covariates concern the partitioning of the features, as in Section 5.2.*

Lemma 7 gives a bound on $Var[\hat{\pi}_0(\mathbf{x}_i)]$ in terms of $\mathbf{S}$ and λ. We note that this bound can always be calculated from the given data.

Lemma 7 *Assuming that Y_i are independent conditional on $\mathbf{X}$ and that all the features are independent:*

$$Var[\hat{\pi}_0(\mathbf{x}_i)] \leq \frac{S_{ii}}{4(1-\lambda)^2},$$

where S_{ii} are the diagonal elements of $\mathbf{S}$.

Proof. By Eq. 7:

$$Var[\hat{\pi}_0(\mathbf{x}_i)] = Var\left[\frac{\mathbf{S}_i^T Y}{1-\lambda}|\mathbf{X} = \mathbf{x}\right] = \frac{\mathbf{S}_i^T Var(\mathbf{Y}|\mathbf{X} = \mathbf{x})\mathbf{S}_i}{(1-\lambda)^2}.$$

By independence of Y_i conditional on $\mathbf{X}$ and independence of the features:

$$Var[\hat{\pi}_0(\mathbf{x}_i)] = \sum_{j=1}^{m} Var(Y_j|\mathbf{X}_j = \mathbf{x}_j)\frac{S_{ij}^2}{(1-\lambda)^2}.$$

Since $Y_j|\mathbf{X}_j = \mathbf{x}_j$ is a Bernoulli random variable, its variance is $P[Y_j = 1|\mathbf{X}_j = \mathbf{x}_j]\{1 - P[Y_j = 1|\mathbf{X}_j = \mathbf{x}_j]\}$, which has $\frac{1}{4}$ as its maximal value, attained at $P[Y_j|\mathbf{X}_j = \mathbf{x}_j] = \frac{1}{2}$. This leads to:

$$Var[\hat{\pi}_0(\mathbf{x}_i)] \leq \frac{\sum_{j=1}^{m} S_{ij}^2}{4(1-\lambda)^2} = \frac{S_{ii}}{4(1-\lambda)^2},$$

the last equality being a direct consequence of $\mathbf{S}$ being a symmetric idempotent matrix.

Theorem 8 shows that, if $S_{ii} \to 0$ as $m \to \infty$ holds alongside the assumptions of Lemma 7, then $\hat{\pi}_0(\mathbf{x}_i)$ is a consistent estimator of $E[\hat{\pi}_0(\mathbf{x}_i)]$.

Theorem 8 *If Y_i are independent conditional on $\mathbf{X}$, all the features are independent, and $S_{ii} \to 0$ as $m \to \infty$,*

$$\hat{\pi}_0(\mathbf{x}_i) \to_P E[\hat{\pi}_0(\mathbf{x}_i)].$$

Proof. By Chebyshev's inequality, for all $\epsilon > 0$:

$$Pr(|\hat{\pi}_0(\mathbf{x}_i) - E[\hat{\pi}_0(\mathbf{x}_i)]| \geq \epsilon) \leq \frac{Var[\hat{\pi}_0(\mathbf{x}_i)]}{\epsilon^2}.$$

Then, by using the stated assumptions and Lemma 7, we get that

$$\lim_{m\to\infty} Pr(|\hat{\pi}_0(\mathbf{x}_i) - E[\hat{\pi}_0(\mathbf{x}_i)]| \geq \epsilon) \to 0.$$

Is it likely or even possible that $S_{ii} \to 0$ as $m \to \infty$? In general this will be the case, unless there are some $\mathbf{x}_i$ which have very high leverage on the regression line by being far from the overall mean of the $\mathbf{x}_i$ vectors. The reason for this is that $\mathbf{S}$ being idempotent implies that $Tr(\mathbf{S}) = rank(\mathbf{S})$, and given that $\mathbf{S} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-}\mathbf{Z}^T$, $Tr(\mathbf{S}) = rank(\mathbf{Z}) = d$, which means that the mean value of S_{ii} is $\frac{d}{m}$. The diagonal elements of $\mathbf{S}$ are also the leverages for the individual data points, with a "rule of thumb" of $S_{ii} > 2\frac{d}{m}$ often being used to identify high leverage points (Hoaglin and Welsch, 1978). It can also be shown that $\frac{1}{m} \leq S_{ii} \leq 1$: We first note that $0 \leq S_{ii} \leq 1$, by once again using the fact that $\mathbf{S}$ is idempotent:

$$\begin{aligned} S_{ii} &= \sum_{j=1}^{m} S_{ij}^2 \\ => S_{ii}(1 - S_{ii}) &= \sum_{j \neq i} S_{ij}^2 \geq 0. \\ => & \quad 0 \leq S_{ii} \leq 1. \end{aligned}$$

We get the improved lower bound by using the fact that $\mathbf{S}_i^T\mathbf{1} = 1$ and Cauchy's inequality:

$$1 = (\sum_{j=1}^{m} S_{ij})^2 \leq m \sum_{j=1}^{m} S_{ij}^2 = mS_{ii}.$$

Furter using the fact that the mean value of S_{ii} is d/m and the inequalities between the arithmetic mean and the minimum and maximum values, we obtain:

$$\frac{1}{m} \leq \min_i S_{ii} \leq \frac{d}{m} \leq \max_i S_{ii} \leq 1.$$

Hoaglin and Welsch (1978) discuss the case where $S_{ii} = 1$, which occurs when the model is fully saturated, predicting the outcome exactly.

Thus, by Theorems 5 and 8, under reasonable conditions, $\hat{\pi}_0(\mathbf{x}_i)$ is a conservative and an asymptotically conservative estimator of $\pi_0(\mathbf{x}_i)$.

We note that our approach to estimating $\pi_0(\mathbf{x}_i)$ does not place any restrictions on its range. Thus, in practice, the values will be truncated to be between 0 and 1. In the following theorem, we show that implementing this truncation decreases the mean squared error of the estimator. The approach is similar to that taken in Theorem 2 of Storey (2002).

Theorem 9 *Let*

$$\hat{\pi}_0^C(\mathbf{x}_i) = \begin{cases} 0 & \hat{\pi}_0(\mathbf{x}_i) < 0 \\ \hat{\pi}_0(\mathbf{x}_i) & 0 \leq \hat{\pi}_0(\mathbf{x}_i) \leq 1 \\ 1 & 1 < \hat{\pi}_0(\mathbf{x}_i) \end{cases}$$

Then:

$$E[(\hat{\pi}_0(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2] \geq E[(\hat{\pi}_0^C(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2].$$

Proof. We prove this result by showing that:

$$E[(\hat{\pi}_0(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) > 1] > E[(\hat{\pi}_0(\mathbf{x}_i)^C - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) > 1] \tag{10}$$

and:

$$E[(\hat{\pi}_0(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) < 0] > E[(\hat{\pi}_0^C(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) < 0]. \tag{11}$$

Then, we can combine them as follows:

$$\begin{aligned} & E[(\hat{\pi}_0(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2] - E[(\hat{\pi}_0^C(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2] = \\ = \; & E[(\hat{\pi}_0(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) > 1] - E[(\hat{\pi}_0(\mathbf{x}_i)^C - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) > 1] P(\hat{\pi}_0(\mathbf{x}_i) > 1) \\ + \; & E[(\hat{\pi}_0(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) < 0] - E[(\hat{\pi}_0^C(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) < 0] P(\hat{\pi}_0(\mathbf{x}_i) < 0) \\ \geq \; & 0. \end{aligned}$$

In 10:

$$\begin{aligned} & E[(\hat{\pi}_0(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) > 1] - E[(\hat{\pi}_0^C(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) > 1] = \\ = \; & E[(\hat{\pi}_0(\mathbf{x}_i) - 1)(\hat{\pi}_0(\mathbf{x}_i) + 1 - 2\pi_0(\mathbf{x}_i)) | \hat{\pi}_0(\mathbf{x}_i) > 1] > 0, \end{aligned}$$

because in this region $\hat{\pi}_0(\mathbf{x}_i) + 1 > 2 \geq 2\pi_0(\mathbf{x}_i)$.
In 11:

$$\begin{aligned} & E[(\hat{\pi}_0(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) < 0] - E[(\hat{\pi}_0^C(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) < 0] = \\ = \; & E[(a - \hat{\pi}_0(\mathbf{x}_i))(2\pi_0(\mathbf{x}_i) - \hat{\pi}_0(\mathbf{x}_i) - 0) | \hat{\pi}_0(\mathbf{x}_i) < 0] > 0, \end{aligned}$$

because in this region $2\pi_0(\mathbf{x}_i) \geq 0 > \hat{\pi}_0(\mathbf{x}_i)$.

7 Simulations

We first describe simulations which give a better idea of the usefulness of Lemma 4 and Theorem 5. We implemented a variety of scenarios, with different values of $\pi_0(\mathbf{x}_i)$ and $\mathbf{Z}$, representing different levels of linearity and model misspecification. In each case, there are $m = 1,000$ features and λ was taken to be 0.8. For the scenarios where $\mathbf{x}_i$ is a scalar, its values were taken to be evenly spaced, while for the scenarios where it is a vector, the values the first component were taken to be evenly spaced, while the second component was a step function, with the first $m/2$ values being equal to 1 and the remaining $m/2$ values being equal to 0. We then randomly generated whether each feature was from the null or alternative distributions, so that the null hypothesis was true for the features for which a success was drawn from the Bernoulli distribution with probability $\pi_0(\mathbf{x}_i)$.

For the null features, p-values were randomly sampled from a $U(0,1)$ distribution, while for the alternative features, they were sampled from a $\beta(a,b)$ distribution, with $a = 1, b = 50$. Sampling the true positive p-values from a Beta distribution is justified in light of recent statistical research (Allison et al., 2002; Pounds and Morris, 2003; Allison et al., 2006; Leek and Storey, 2011). Plots of $\pi_0(\mathbf{x}_i)$ versus $\mathbf{x}_i$ and $\hat{E}[\hat{\pi}_0(\mathbf{x}_i)]$ and $\pi_0(\mathbf{x}_i)$ versus the index i are in Figure 1. Note that, as expected, the closer we get to having a correctly specified model with a linear estimator, the better the estimation is.

Next, we use the same set of simulations to estimate the variance of $\hat{\pi}_0(\mathbf{x}_i)$ and compare it to the bound from Lemma 7. Plots of $\widehat{Var}[\hat{\pi}_0(\mathbf{x}_i)]$ and its upper bound versus the index i are presented in Figure 2.

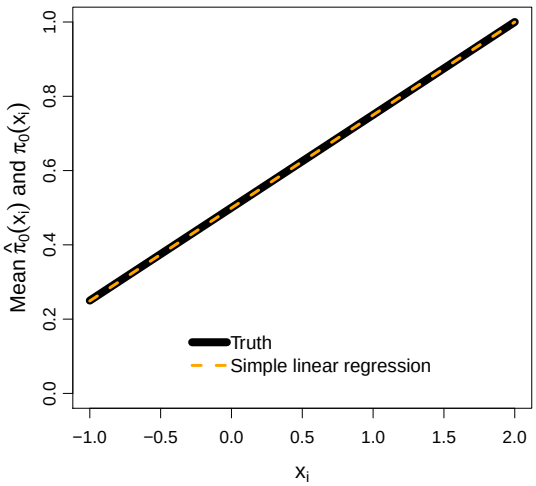
Finally, we used the same scenarios, but varied the number of features in order to see whether Theorem 8, which says that $\hat{\pi}_0(\mathbf{x}_i)$ is a consistent estimator of $E[\hat{\pi}_0(\mathbf{x}_i)]$, holds. The number of features was taken to be either $m = 10, 100, 1,000$ or $10,000$ and the components of $\mathbf{x}_i$ were set as before. For each value of m considered, we calculated $\max_{i=1}^m S_{ii}$. The results, shown in Table 2, indeed justify the assumptions of Theorem 8. In general, $\mathbf{Z}^T\mathbf{Z}$ can be written as a matrix of the sample means of pairs of the p variables (i.e. $\overline{Z_i Z_j}$ for the variables i^{th} and j^{th} variables) multiplied by m, therefore all the terms in $\mathbf{S} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ include combinations of the individual variables Z_{ij} and the sample means of combinations, with number of terms depending on p, which is fixed, multiplied by $1/m$, if $\mathbf{Z}^T\mathbf{Z}$ is invertible. Thus, as long as all the means are bounded as $m \to \infty$, as they would be in the case of equally spaced values, then $S_{ii} \to 0$ as $m \to \infty$, fulfilling the conditions for Theorem 8.

Table 2: Values of $max_{i=1}^m S_{ii}$ for the simulation scenarios in Figures 1 and 2, using the panel labels in Figure 2, for different values for the number of features, m. Scenarios a) and b) fit the same model, therefore have the same upper bounds.

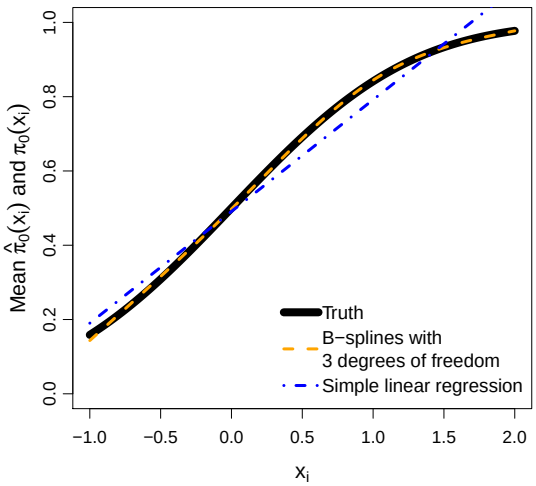
Scenario	m=10	m=100	m=1,000	m=10,000
a), b)	2.159	0.246	0.025	0.002
c)	4.697	0.726	0.076	0.008
d)	4.798	0.755	0.079	0.008
e)	6.250	3.540	0.491	0.051

Figure 1: Different simulation scenarios. The true function $\pi_0(\mathbf{x}_i)$ is plotted in a thick black line, while the empirical means of $\hat{\pi}_0(\mathbf{x}_i)$, assuming different modelling approaches are shown in the dashed orange line and the dashed-and-dotted blue line. In c), different terms are used in the regression for x_{i1}, while the true values are used for x_{2i}.

(a) $\pi_0(x_i) = x_i/4 + 1/2$



(b) $\pi_0(x_i) = \phi(x_i)$



(c) $\pi_0(\mathbf{x}_i) = \sin(x_{i1})/4 + x_{i2}/4 + 1/2$, where $x_{i2} = 1(i \leq m/2)$.

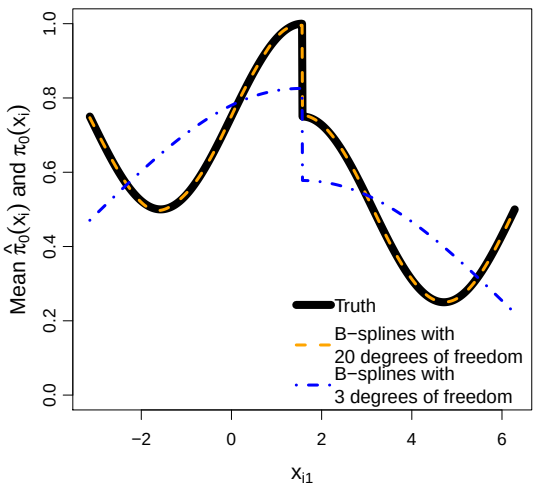
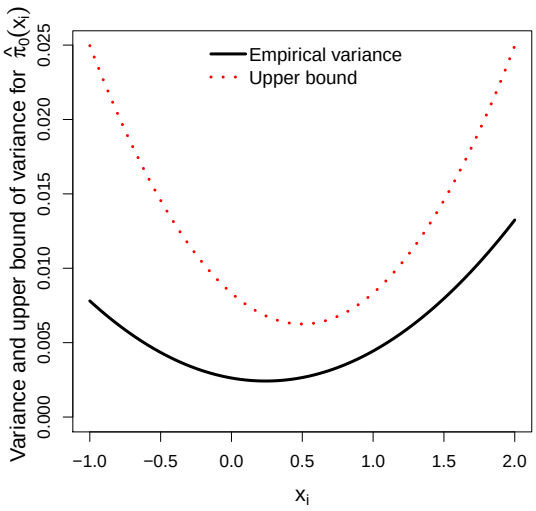
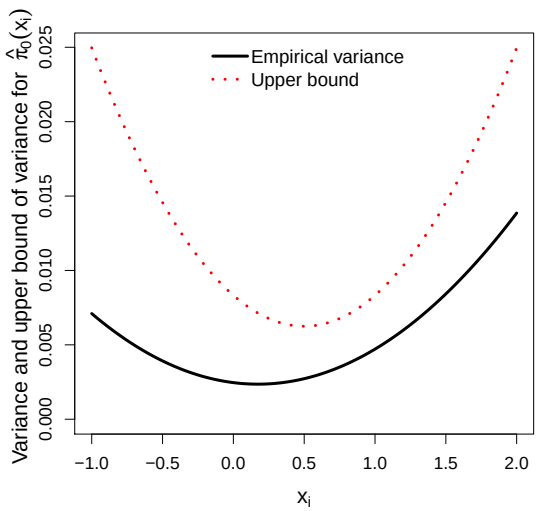


Figure 2: The same simulation scenarios as in Figure 1. The empirical variance of $\hat{\pi}_0(\mathbf{x}_i)$ is in a continuous black line, while the upper bound obtained in Lemma 7 is plotted in a dotted red. Scenarios a) and b) fit the same model on the same values of x_i, resulting in identical plots.

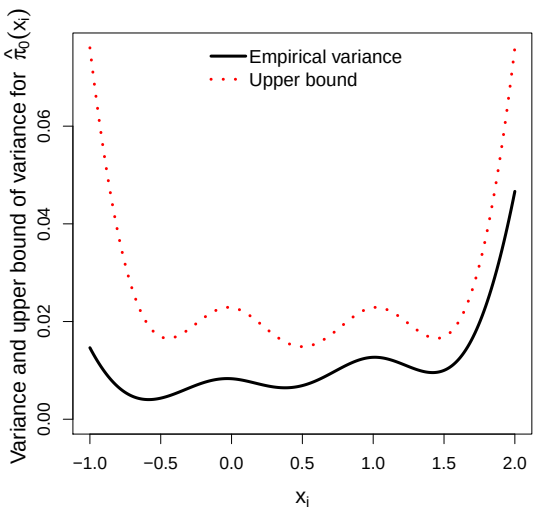
(a) $\pi_0(x_i) = x_i/4 + 1/2$



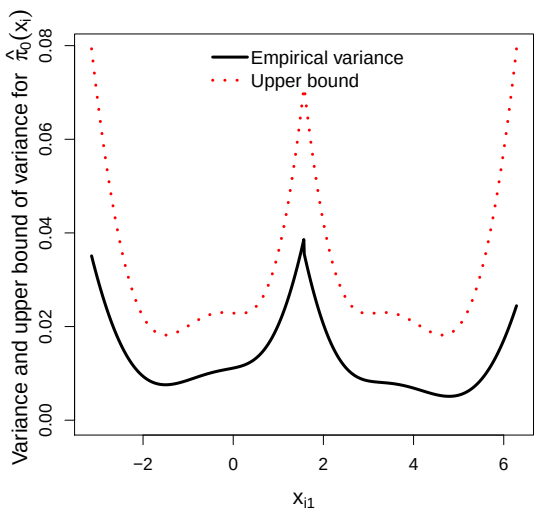
(b) $\pi_0(x_i) = \phi(x_i)$:
Fit linear regression.



(c) $\pi_0(x_i) = \phi(x_i)$:
Fit B-splines with 3 degrees of freedom.



(d) $\pi_0(\mathbf{x}_i) = \sin(x_{i1})/4 + x_{i2}/4 + 1/2$,
where $x_{i2} = 1(i \leq m/2)$:
Fit B-splines with 3 degrees of freedom for x_{i1}, true values for x_{i2}.



(e) $\pi_0(\mathbf{x}_i) = \sin(x_{i1})/4 + x_{i2}/4 + 1/2$,
where $x_{i2} = 1(i \leq m/2)$:
Fit B-slines with 20 degrees of freedom for x_{i1}, true values for x_{i2}.

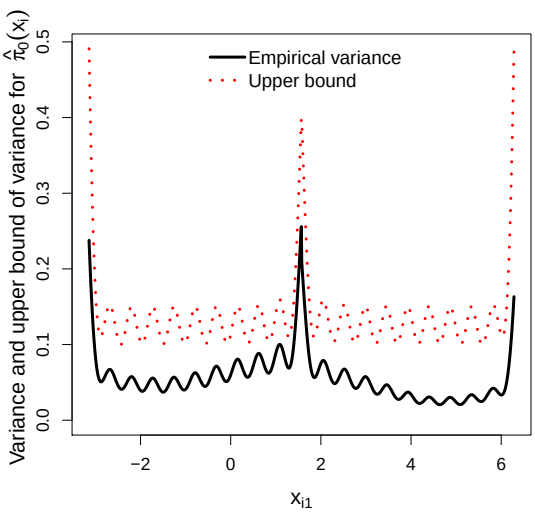
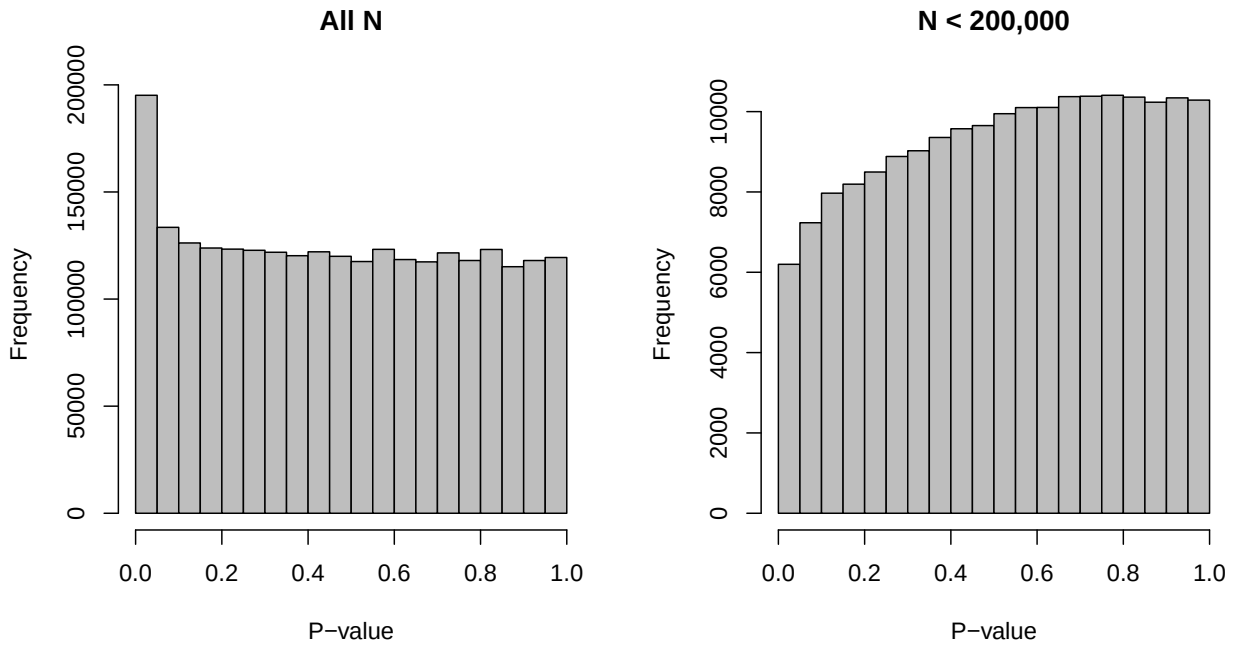


Figure 3: Histograms of p-values for the SNP-BMI tests of association from the GIANT consortium. Panel a) shows the distribution for all sample sizes N (2,500,573 SNPs), while panel b) shows the subset N <200,000 (187,114 SNPs).



8 Data analysis

Here we consider data from the Genetic Investigation of ANthropometric Traits (GIANT) consortium, specifically the genome-wide association study for body mass index (BMI) (Locke et al., 2015). From a total of 2,555,510 SNPs, we removed the SNPs which did not have minor allele frequencies (MAFs) listed for the HapMap CEU population, leading to 2,500,573 SNPs. For each of these SNPs, we considered the p-values from the test of association with BMI and the meta-data covariates consisting of the number of individuals (N) considered for each SNP and the minor allele frequencies (MAFs) in the HapMap CEU population, since it is well-known that both sample size and MAF have an impact on p-values, with larger sample sizes and MAFs leading to more significant results.

The model we considered uses natural cubic splines with 5 degrees of freedom to model N and 3 discrete categories for the MAFs. Figure 3 shows the dependence of p-values on sample sizes within this dataset. Figure 4 shows the estimates of $\pi_0(\mathbf{x}_i)$ plotted against the sample size N, stratified by the CEU MAFs for a random subset of 50,000 SNPs.

Our results are consistent with intuition - larger sample sizes and larger MAFs lead to a smaller fraction of SNPs estimated to be null. Applying this estimator to the false discovery rate calculation will mean increased power to detect associations for SNPs with large sample sizes and large MAFs, with potentially reduced power for SNPs with the opposite characteristics.

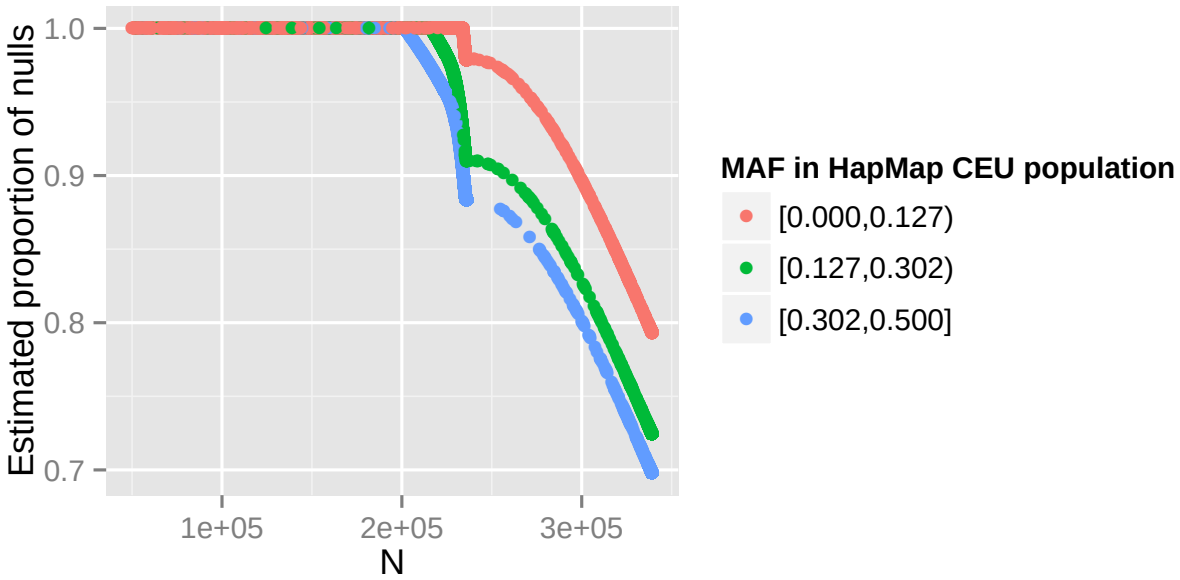
9 Reproducibility

All analyses and simulations in this paper are fully reproducible and the code is available on Github at: `https://github.com/SiminaB/Fdr-regression`

10 Discussion

Here we have introduced a regression framework for the proportion of true null hypotheses in a multiple testing framework. We have provided conditions for conservative and consistent estimation of this proportion conditional on covariates. Using simulations we have shown that while the regression

Figure 4: Plot of the estimates of $\pi_0(\mathbf{x}_i)$ against the sample size N, stratified by the MAF categories for a random subset of 50,000 SNPs.



estimates may be incorrect under model misspecification the upper bounds on the variance of the estimator hold even for inaccurate models.

Applying our estimator to GWAS data from the GIANT consortium demonstrated that, as expected, the estimate of the fraction of null hypotheses decreases with both sample size and minor allele frequency. It is a well known and problematic phenomenon that p-values for all features decrease as the sample size increases. This is because the null is rarely precisely true for any given feature. One interesting consequence of our estimates is that we can calibrate what fraction of p-values appear to be drawn from the non-null distribution as a function of sample size, potentially allowing us to quantify the effect of the"large sample size means small p-values" problem directly.

A range of other applications for our methodology are also possible by modifying our regression framework, including estimating false discovery rates for gene sets (Boca et al., 2013), estimating science-wise false discovery rates (Jager and Leek, 2013), or improving power in high-throughput biological studies (Ignatiadis et al., 2015).

References

Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7(1):55–65.

Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., Prolla, T. A., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39(1):1–20.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Boca, S. M., Corrada Bravo, H., Caffo, B., Leek, J. T., and Parmigiani, G. (2013). A decision-theory approach to interpretable set analysis for high-dimensional data. *Biometrics*. doi: 10.1111/biom.12060.

Brown, O. P. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470.

Efron, B. (2008). Simultaneous inference: When should hypothesis testing problems be combined? *The annals of applied statistics*, pages 197–223.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108.

Hoaglin, D. C. and Welsch, R. E. (1978). The hat matrix in regression and anova. *The American Statistician*, 32(1):17–22.

Ignatiadis, N., Klaus, B., Zaugg, J., and Huber, W. (2015). Data-driven hypothesis weighting increases detection power in big data analytics. *bioRxiv*, page 034330.

Jager, L. R. and Leek, J. T. (2013). Empirical estimates suggest most published medical research is true. *Biostatistics*. doi: 10.1093/biostatistics/kxt007.

Leek, J. T. and Storey, J. D. (2011). The joint null criterion for multiple hypothesis tests. *Statistical Applications in Genetics and Molecular Biology*, 10(1).

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206.

Neale, B. M., Medland, S. E., Ripke, S., Asherson, P., Franke, B., Lesch, K.-P., Faraone, S. V., Nguyen, T. T., Schäfer, H., Holmans, P., et al. (2010). Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 49(9):884–897.

Pounds, S. and Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242.

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P., and Kass, R. E. (2014). False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, (just-accepted):00–00.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145.

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