Estimating the Number and Effect Sizes of Non-null Hypotheses

Jennifer Brennan [1] **Ramya Korlakai Vinayak** [1] **Kevin Jamieson** [1]

Abstract

We study the problem of estimating the distribution of effect sizes (the mean of the test statistic under the alternate hypothesis) in a multiple testing setting. Knowing this distribution allows us to calculate the power (type II error) of any experimental design. We show that it is possible to estimate this distribution using an inexpensive pilot experiment, which takes significantly fewer samples than would be required by an experiment that identified the discoveries. Our estimator can be used to guarantee the number of discoveries that will be made using a given experimental design in a future experiment. We prove that this simple and computationally efficient estimator enjoys a number of favorable theoretical properties, and demonstrate its effectiveness on data from a gene knockout experiment on influenza inhibition in *Drosophila*.

1. Introduction

Designing scientific experiments is something of a chicken and egg problem. In order to design an experiment with a specified power (type II error), we need to know the effect size (the mean of the test statistic under the alternate hypothesis). The effect size determines the required accuracy of each measurement, which increases with the number of *experimental replicates* (samples). Unfortunately, this effect size is typically unknown, and estimating the effect size for a single hypothesis test is as sample intensive as performing the original experiment. In the case of single hypothesis testing, this presents a fundamental barrier to efficient experimental design. By contrast, in the setting of multiple hypothesis testing, we show that it is possible to estimate the distribution of effect sizes present in the data using an inexpensive pilot experiment, which takes significantly fewer samples than would be required for the full experiment. For example, suppose a scientist would like to test 10,000 genes using an experimental measurement that is distributed $\mathcal{N}(\mu_i, \frac{1}{t})$ when the effect size is μ_i and t replicates are performed. Without knowledge of the likely effect sizes, it is unclear how to choose an experimental design. An experiment with too many replicates per hypothesis is wasteful; one with too few will lack the statistical power to identify alternate hypotheses. In this paper, we seek to facilitate experimental design in the multiple testing setting by answering the question *"How many hypotheses have an effect size of at least γ?"* using significantly fewer samples than would be needed to identify all discoveries with that effect size. These estimates suggest a trade-off between the cost of an experiment (as measured by the number of experimental replicates required to achieve a certain power) and the number and effect sizes of the discoveries that will be made. Figure 1 illustrates the application of our estimator to an inexpensive pilot study, allowing a scientist to evaluate possible experimental designs. The application to experimental design motivates an important property of our estimator: it must produce a *conservative* estimate of the number of hypotheses above a given effect size. If the scientist designs a costly experiment based on the results of this estimator, it is important to ensure that this experiment will generate at a minimum the estimated number of discoveries.

As a baseline, one approach to this estimation problem is to use a *plug-in estimator*, which estimates the entire distribution of effect sizes and then "plugs in" this estimate as if it were the true distribution. The plug-in estimator could start with the maximum likelihood estimate (MLE) of the distribution of effect sizes given the observed test statistics. The estimate for the fraction of hypotheses above some effect size γ would simply be the fraction of this distribution that exceeded γ. Unfortunately, such a plug-in estimator based on the MLE may vastly overestimate this fraction, as two distributions can have similar likelihoods but very different amounts of mass above some threshold.

In this work, we design an estimator for the fraction of hypotheses with effect sizes above a given threshold, for all thresholds simultaneously. Our estimator operates in the spirit of the Kolmogorov-Smirnov test, first creating an ℓ_∞ ball around the empirical CDF to define plausible

[1] Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA. Correspondence to: Jennifer Brennan <jrb@cs.washington.edu>.

Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

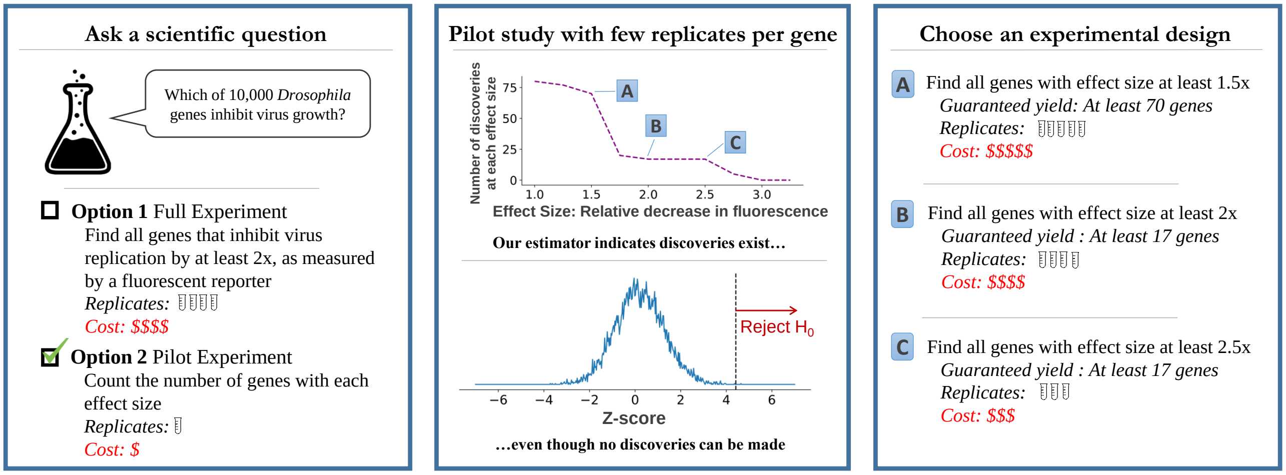

Ask a scientific question
Which of 10,000 *Drosophila* genes inhibit virus growth?
☐ Option 1 Full Experiment
Find all genes that inhibit virus replication by at least 2x, as measured by a fluorescent reporter
Replicates:
Cost: $$$$
☑ Option 2 Pilot Experiment
Count the number of genes with each effect size
Replicates:
Cost: $
Pilot study with few replicates per gene
Number of discoveries at each effect size
75
50
25
0
1.0
1.5
2.0
2.5
3.0
A
B
C
Effect Size: Relative decrease in fluorescence
Our estimator indicates discoveries exist…
Reject H0
−6
−4
−2
0
2
4
6
Z-score
…even though no discoveries can be made
Choose an experimental design
A Find all genes with effect size at least 1.5x
Guaranteed yield: At least 70 genes
Replicates:
Cost: $$$$$
B Find all genes with effect size at least 2x
Guaranteed yield : At least 17 genes
Replicates:
Cost: $$$$
C Find all genes with effect size at least 2.5x
Guaranteed yield : At least 17 genes
Replicates:
Cost: $$$

Figure 1. When applied to the results of a pilot experiment, our estimator can estimate the cost and number of discoveries guaranteed by different experimental designs. In this example, the original experiment design (Option 1) is expensive, with no guarantee on the number of discoveries that will be made. Our method suggests two alternatives to the original experimental design (B); the same guarantee on discoveries could be made at lower cost (C), or additional discoveries could be made at higher cost (A).

distributions, and then finding the element of the ball with the smallest amount of probability mass above γ. With high probability, this amount of mass does not exceed the true fraction of hypotheses with mean at least γ. We prove that this simple and computationally efficient estimator enjoys a number of favorable theoretical properties, including finite-sample upper and lower bounds on the value of the estimate.

1.1. Problem Statement

Let ν_* be a distribution on $\mathbb{R}$, and for $i = 1, 2, \ldots n$ let

$$\mu_i \sim \nu_*$$

be an unobserved latent variable drawn iid from ν_*. For each μ_i drawn from ν_*, we observe the test statistic

$$X_i \sim f_{\mu_i},$$

where f_μ is a known distribution parameterized by the effect size μ. For example, suppose the test statistics were Z-scores, which are distributed according to the standard normal distribution under the null hypothesis and shifted by the effect size under the alternate. Then, $f_\mu = \mathcal{N}(\mu, 1)$. While our estimator is well defined for any parametric f (e.g., any single-parameter exponential family), we focus on Gaussian test statistics for exposition. In the setting of Figure 1, ν_* represents the distribution of effect sizes and X_i are the observations.

Our goal is to estimate the probability that the effect size of an observation is greater than γ,

$$\zeta_{\nu_*}(\gamma) := \mathbb{P}_{\nu_*}(\mu > \gamma), \tag{1}$$

simultaneously for all $\gamma \in \mathbb{R}$.

The problem of counting the non-null hypotheses is most interesting when γ is small. For example, consider the case when the test statistics X_i are Z-scores. Under the hypothesis that all effect sizes are zero, the expected maximum Z-score is $\mathbb{E}[\max_i X_i] \approx \sqrt{\log n}$. Therefore, if we want to avoid any false discoveries, we cannot reject any hypotheses with test statistic less than $\Theta(\sqrt{\log n})$. If the effect sizes are at least this large, then we will be able to identify the alternate hypotheses through a standard Bonferroni correction (Dunn, 1961). In this regime, counting is no more difficult than identification. However, if the effect sizes are much smaller than this threshold (say, if all $\mu_i \ll 1$), identification could be impossible. Our estimator, by contrast, detects the existence of discoveries even in this low signal-to-noise regime.

1.2. Contributions

Our contributions are as follows:

- Given a parameterization f_μ, we propose an estimator that provides a conservative estimate of the fraction of effect sizes above a given threshold, simultaneously for all thresholds (Section 2).
- We provide finite-sample bounds on the error of our estimator (Theorem 2.1).
- In the low signal-to-noise regime and the setting of Gaussian mixtures, we compare our estimator's sample complexity to a known lower bound for hypothesis testing (detecting the presence of the alternate hypothesis), and we give a novel lower bound for the sample complexity

of estimation (estimating the fraction of means from the alternate hypothesis). We show that our method matches finite-sample rates for these problems, even though it is designed for more general distributions than the ones in these lower bounds (Section 2).

- We describe how to use this estimator to design pilot studies for scientific experimentation (Section 3). When testing n hypotheses in the low signal-to-noise regime, our technique detects treatments with positive effect sizes using a factor of n fewer replicates than it would take to identify them. Additionally, the results of the pilot experiment can be used to upper bound the cost of identifying the discoveries at each effect size.

1.3. Related Work

The problem of estimating the number of null hypotheses has been studied extensively in the statistics literature. Our goal in this work is to provide a conservative estimate of the number of hypotheses with effect size above some threshold (Eqn (1)). There are several lines of work related to this goal.

Simple Null Hypotheses A different but related problem is to estimate the number of non-null hypotheses, regardless of their effect sizes, i.e., $\mathbb{P}_{\nu_*}(\mu \neq 0)$. In this setting - also known as the simple null hypothesis - it is possible to compute p-values that are uniformly distributed under the null. For example, when observations are drawn $X_i \sim \mathcal{N}(\mu_i, 1)$, the p-value is $p_i = 1 - \Phi(X_i)$, where Φ is the standard normal CDF.

The graphical estimator of Schweder & Spjøtvoll (1982) was the first technique to estimate the number of nulls, using the principle that p-values are distributed uniformly under the null hypothesis and skewed toward zero under the alternate. Their technique estimates the density of the p-value distribution at 1. This same idea was improved in the context of estimating the number of nulls for adaptive control of the false discovery rate (FDR) (Benjamini & Hochberg, 2000; Storey, 2002). These later works provide finite-sample guarantees on overestimating the number of nulls in order to make non-asymptotic guarantees on FDR control. However, none of these results provide lower bounds on the estimated number of non-nulls. Motivated by adaptive FDR control, techniques for counting the number of non-null hypotheses have been extended to incorporate prior knowledge about the dependence structure of the hypotheses or the likelihood that each test will result in a discovery. See Li & Barber (2019) for a review of this area.

Bounds on the False Discovery Proportion - the high-probability analogue of FDR - can also be employed to report a guarantee on the number of significant effects. The simultaneous FDP estimator of Katsevich and Ramdas (2018) provides such bounds simultaneously for all sets in a path. A guarantee on the FDP of a set corresponds to a lower bound on the number of discoveries; maximizing over the guarantees provided by each set in the path gives an improved lower bound. With an assumption on the form of the test statistic under the alternate hypothesis, this algorithm can be modified to bound the number of discoveries above an arbitrary threshold. We compare to this baseline method in our experimental results.

Another technique for the simple null setting, again motivated by the uniform distribution of p-values under the null, is to test the extent to which the distribution of p-values deviates from the uniform distribution. Several estimators have taken this approach (Genovese et al., 2004; Meinshausen & Rice, 2006; Patra & Sen, 2016; Jin, 2008). Most similar to our work are the techniques that build one-sided confidence intervals around the empirical CDF of p-values (Genovese et al., 2004; Meinshausen & Rice, 2006), which provide finite-sample error bounds and a conservative estimator. Finally, there are estimators specific to the Gaussian setting, which estimate the zero-mean component in a mixture of Gaussians (Cai et al., 2007; Carpentier et al., 2019).

Extensions to one-sided null hypotheses ($H_0: \mu \leq 0$) further assume that p-values are subuniformly distributed when $\mu < 0$ (Meinshausen & Bühlmann, 2005; Li & Barber, 2019) or assume a gap between 0 and the smallest alternate effect size (Lee & Valiant, 2019). These works estimate the quantity $\mathbb{P}_{\nu_*}(\mu > 0)$. This problem is a special case of ours, because subuniformity holds only for the threshold of $\gamma = 0$.

Composite Null Hypotheses We seek to estimate the number of hypotheses with an effect size above some threshold. Here, p-values are neither subuniform nor necessarily well defined, so much of the previous work is not applicable. The Fourier transform technique (Jin, 2008) can be extended to address composite null hypotheses (Chen, 2019). However, this extension only provides asymptotic results, which are insufficient since we seek a conservative estimator.

Adapting the Generalized Likelihood Ratio Test Jiang & Zhang (2016) develop asymptotic power statements for the generalized likelihood ratio test for Gaussian observations. We discuss in Section 5 how this work could be used to create an estimator for our problem, and highlight the limitations that make this approach impractical.

Plug-in Estimation As discussed in Section 1, another approach to this problem is plug-in estimation, where an estimate $\widetilde{\nu}$ of the distribution ν_* is used to form an estimator $\widehat{\zeta}_n(\gamma) = \mathbb{P}_{\widetilde{\nu}}(\mu > \gamma)$. When f_μ is Gaussian, the task is to learn a mixture of Gaussians. In this setting, much effort has been devoted to recovering the mixture parameters (Pearson, 1894; Belkin & Sinha, 2010; Kalai et al., 2010; Hardt

& Price, 2015) or learning a mixture that is close to the original distribution in some metric, such as total variation (TV) distance (Moitra & Valiant, 2010; Daskalakis & Kamath, 2014). Outside of the Gaussian setting, recent works have provided guarantees for learning mixtures of binomial distributions in terms of the Wasserstein-1 distance (Tian et al., 2017; Vinayak et al., 2019). These types of theoretical guarantees do not lend themselves easily to guarantees on our problem, since two distributions can be close in TV or Wasserstein distance but have very different amounts of mass above some threshold γ.

Empirical Bayes Methods Our estimator takes advantage of multiple hypothesis testing by using the empirical distribution of the X_i to learn something about the latent distribution ν_*. The same idea can be seen in empirical Bayes methods, where the empirical distribution of X_i is used as the prior over X. Several papers have taken an empirical Bayes approach to multiple testing, but none address our exact problem. Efron (2007) uses an empirical Bayes method to estimate the distribution of X under the alternate hypothesis, which is distinct from our goal of estimating ν_* (note we cannot simply deconvolve Efron's estimate to get ν_*, as it is not guaranteed to have any parametric form). Stephens (2017) uses empirical Bayes methods and a strong unimodality assumption on ν_* to produce estimates and confidence intervals for each μ_i. While these confidence intervals could theoretically be used to estimate (1), the fact that Stephens' method produces a confidence interval for individual μ_i suggests that they will be too loose to compete with our method. Indeed, we see this looseness in the experimental results, where our estimator outperforms Stephens' in our regime of interest. Furthermore, this method only works for Gaussian and t-distributed observations.

2. Estimating Effect Sizes

Recall our goal, to estimate $\zeta_{\nu_*}(\gamma)$ from Eqn (1). Let $\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq t\}$ be the empirical CDF of the test statistics X_i and

$$F_\nu(t) = \mathbb{P}_{\mu\sim\nu,\, X\sim f_\mu}(X \leq t)$$

be the true CDF of test statistics under latent distribution ν. For any $\gamma \in \mathbb{R}$, our estimator is given by

$$\widehat{\zeta}_n(\gamma) = \min_{\nu : ||\widehat{F}_n - F_\nu||_\infty \leq \tau_{\alpha,n}} \int_\gamma^\infty \nu(x)dx \tag{2}$$

where the estimator is conservative with probability at least $1-\alpha$, and

$$\tau_{\alpha,n} = \sqrt{\frac{\log(2/\alpha)}{2n}}.$$

The intuition for this estimator is as follows. To conservatively estimate the amount of mass ζ above threshold γ, we look for the distribution with the smallest amount of mass above γ that could have plausibly generated the observations X_i. Our measurement of plausibility is based on high probability bounds on the deviation between the empirical CDF and its expectation. If $\widehat{F}_n$ was in fact drawn from F_ν, then with high probability the ℓ_∞ distance between $\widehat{F}_n$ and F_ν will not exceed $\tau_{\alpha,n}$. By restricting our search space to the ℓ_∞ ball around $\widehat{F}_n$ (seen in the constrained optimization from Eqn (2)), we do not overestimate the true amount of mass above γ, with high probability. Moreover, using different values of γ traces a curve for $\zeta_{\nu_*}(\gamma)$ (see the middle panel of Figure 1). We note that this estimator can be implemented as an efficient convex program. We simply discretize x over some range, and the estimator becomes a convex program in the vector $\mathbf{x}$. It can then be solved using off-the-shelf software (see Appendix C for details).

2.1. Main Results

Our estimator underestimates the true mass $\zeta_{\nu_*}(\gamma)$ for all γ simultaneously with high probability. Furthermore, we provide a finite sample bound on how much we underestimate $\zeta_{\nu_*}(\gamma)$ at every γ.

Theorem 2.1. *For $i = 1, \ldots, n$, let $\mu_i \sim \nu_*$ and $X_i \sim f_{\mu_i}$ where each draw is iid. Let our simultaneous estimator be given by (2). Then, the probability of overestimating the fraction of hypotheses with effect size above any threshold γ is bounded by α:*

$$\mathbb{P}\left(\exists\gamma : \widehat{\zeta}_n(\gamma) > \zeta_{\nu_*}(\gamma)\right) \leq \alpha.$$

Furthermore, with probability at least $1-\delta$, for all $\gamma \in \mathbb{R}$ and $\varepsilon \in (0, \zeta_{\nu_*}(\gamma)]$ we have $\zeta_{\nu_*}(\gamma) - \hat{\zeta}_n(\gamma) \leq \varepsilon$ whenever

$$n \geq \frac{\log\left(\frac{4}{\alpha\delta}\right)}{\left(\min_{\nu:\mathbb{P}_\nu((\gamma,\infty))\leq\zeta_{\nu_*}(\gamma)-\varepsilon} ||F_\nu - F_{\nu_*}||_\infty\right)^2}. \tag{3}$$

Remark 2.1 (Pointwise consistency)**.** *Our estimator is pointwise consistent. For any threshold γ and any $\varepsilon > 0$, there is some n large enough that the error in our estimate satisfies $\zeta_{\nu_*}(\gamma) - \hat{\zeta}(\gamma) < \varepsilon$. This follows from the fact that, for any $\varepsilon > 0$, the denominator of 3 is strictly positive.*

Our estimator is guaranteed not to overestimate $\zeta_{\nu_*}(\gamma)$, which is critical in the use of pilot studies to guide experimental design. The key quantity in this sample complexity result is the minimum ℓ_∞ distance between the true CDF F_{ν_*} and the set of CDFs corresponding to mixing distributions with less than ζ mass above γ. We call this set of mixing distributions S,

$$S(\zeta, \gamma) := \{\nu : \mathbb{P}_\nu((\gamma, \infty)) \leq \zeta\}. \tag{4}$$

Specifically, consider $S(\zeta_{\nu_*}(\gamma) - \varepsilon, \gamma)$, which appears in Eqn (3). If $\varepsilon = 0$, then we have $\nu_* \in S(\zeta_{\nu_*}, \gamma)$, so the minimum ℓ_∞ distance to F_{ν_*}, $\min_{\nu \in S(\zeta_{\nu_*}(\gamma),\gamma)} ||F_\nu - F_{\nu_*}||_\infty$, would be zero, implying that no finite sample can guarantee $\varepsilon = 0$. This reflects the fact that $\widehat{\zeta}_n(\gamma)$ is an underestimate at every γ; therefore, in order for ε to be zero, we must have estimated $\zeta_{\nu_*}(\gamma)$ exactly. As ε increases, $S(\zeta_{\nu_*}(\gamma) - \varepsilon, \gamma)$ shrinks, and the distance to F_{ν_*} increases, decreasing the required number of samples n.

To interpret the sample complexity in Theorem 2.1, we consider a simple model where test statistics are drawn from a mixture of two Gaussians. In this setting, which we denote $X_i \sim P(\zeta_*, \gamma_*)$, we have

$$\begin{aligned} \mu_i &\sim (1-\zeta_*)\delta_0 + \zeta_*\delta_{\gamma_*} \\ X_i &\sim \mathcal{N}(\mu_i, \sigma^2), \end{aligned} \tag{5}$$

where δ_x is the Dirac delta function at x. There are two natural questions we might ask: How many samples a re necessary to determine the existence of the mixture component at $\gamma_* > 0$, and how many samples are required to estimate the weight of this component? We call these the *testing* and *estimation* problems respectively. In the following sections, we address our algorithm's sample complexity for these problems, and compare to lower bounds. For ease of exposition, let $\alpha = \delta$, although the results hold for the more general case.

2.2. Global Null Testing

In the global null testing problem, we observe X_i according to (5), and we want to determine whether $\zeta_* > 0$ (i.e., testing $H_0 : P_{\nu_*}(\mu > 0) = 0$ vs $H_1 : P_{\nu_*}(\mu > 0) > 0$). Our test declares H_1 if $\widehat{\zeta}_n(0) > 0$, and H_0 if $\widehat{\zeta}_n(0) = 0$. Clearly this test erroneously declares H_1 with probability at most δ (it has type I error at most δ), since $\widehat{\zeta}_n(0) \le \zeta_*$ with probability at least $1-\delta$ (recall that we set $\alpha = \delta$ in Theorem 2.1). The next corollary bounds the sample complexity that guarantees a probability of detection of at least $1-\delta$ (i.e., that bounds the type II error by δ).

Corollary 2.1.1. *Let $\{X_i\}_{i=1}^n$ be drawn according to (5). Consider the simultaneous estimator $\widehat{\zeta}_n$ defined by (2). Then, with probability at least $1-\delta$, we have $\widehat{\zeta}_n(0) \le \zeta_*$ and $\widehat{\zeta}_n(0) > 0$ whenever*

$$n \ge \frac{2\log\left(\frac{2}{\delta}\right)}{\zeta_*^2\left(\Phi_\sigma\left(\frac{1}{2}\gamma_*\right) - \Phi_\sigma\left(-\frac{1}{2}\gamma_*\right)\right)^2},$$

where Φ_σ is the CDF of the distribution $\mathcal{N}(0,\sigma^2)$. Furthermore, if $\gamma_* < \sigma$, then the above can be simplified to

$$n \ge \frac{16\sigma^2\log\left(\frac{2}{\delta}\right)}{\zeta_*^2\gamma_*^2}.$$

Proof Sketch. To obtain this sample complexity result, we must lower bound the distance term in the denominator of Eqn (3). Recalling our definition of S in Eqn (4), we lower bound the associated minimax quantity by its value at a specific point, $t = \frac{1}{2}\gamma_*$,

$$\begin{aligned} \min_{\nu \in S(0,0)} \|F_\nu - F_{\nu_*}\|_\infty &= \min_{\nu \in S(0,0)} \sup_{t\in\mathbb{R}} |F_\nu(t) - F_{\nu_*}(t)| \\ &\ge \min_{\nu \in S(0,0)} F_\nu(\tfrac{1}{2}\gamma_*) - F_{\nu_*}(\tfrac{1}{2}\gamma_*) \end{aligned}$$

The constraint $\nu \in S(0,0)$ allows us to lower bound the first quantity by $\Phi(\frac{1}{2}\gamma_*)$, and we compute the second quantity exactly, giving the first conclusion of the corollary. The second conclusion follows from a quadratic approximation to the normal density. □

We compare Corollary 2.1.1 to the finite-sample lower bound arising from the "most biased coin problem" (Chandrasekaran & Karp, 2014; Jamieson et al., 2016). In this problem, the algorithm draws N observations X_i as per (5), where N is potentially a random variable, according to either $H_0 : X_i \sim \mathcal{N}(0,\sigma^2)$ or $H_1 : X_i \sim P(\zeta_*, \gamma_*)$. When γ_* and ζ_* are known and $\gamma_* \le \sigma$, Theorem 2 of Jamieson et al. (2016) states that any (potentially randomized) procedure that decides between these hypotheses with probability of error at most δ requires at least

$$\mathbb{E}[N] \ge \max\left\{\frac{1-\delta}{\zeta_*}, \frac{\sigma^2\log(1/\delta)}{2\zeta_*^2\gamma_*^2}\right\}$$

samples. To facilitate comparison with the sample complexity of our estimator, we show in Lemma A.3 that the small-γ_* sample complexity from Corollary 2.1.1 matches the stated lower bound up to constants both when δ is fixed and as $\delta \to 0$.

2.3. The Estimation Problem

In the estimation problem, we observe X_i according to (5), and we estimate ζ_* using our estimator $\widehat{\zeta}(0)$. Since $\widehat{\zeta}(0) \le \zeta_*$ with high probability, it remains to understand the magnitude of this underapproximation – the dependence of ε from Theorem 2.1 on the number of samples n. The following corollary describes the number of samples needed to guarantee an error bound $\varepsilon \le \frac{1}{2}\zeta_*$ with high probability.

Corollary 2.1.2. *Let $\{X_i\}_{i=1}^n$ be drawn according to (5). Let $\zeta_* > 0$ and $\gamma_* \in (0, \sigma]$. Then, with probability at least $1-\delta$, our estimate $\widehat{\zeta}_n$ from (2) satisfies $\widehat{\zeta}_n(0) \in (\frac{1}{2}\zeta_*, \zeta_*]$ as long as*

$$n \gtrsim \frac{\sigma^4\log\left(\frac{2}{\delta}\right)}{\zeta_*^2\gamma_*^4}.$$

Proof Sketch. Again, we obtain this result by lower bounding the denominator of Eqn (3). We note that this quantity is the optimal value of a convex optimization problem, and show that the associated optimal point is

$$\nu_{OPT} = (1 - \tfrac{1}{2}\zeta_*)\delta_0 + \tfrac{1}{2}\zeta_*\delta_{2\gamma_*}.$$

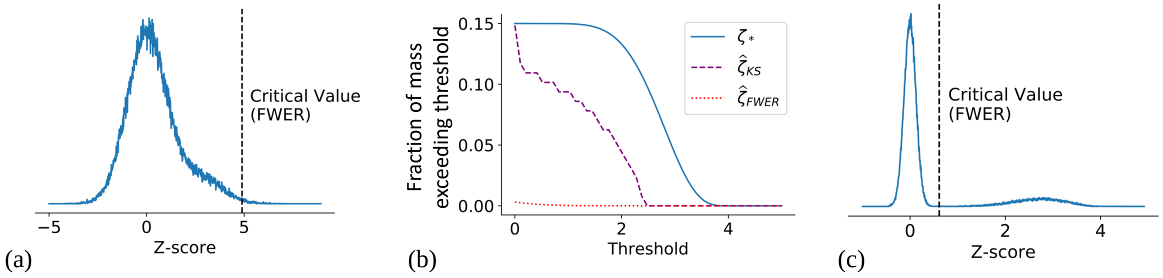


Figure 2. Our estimator applied to pilot experiments. (a) After observing $X_i \sim \mathcal{N}(\mu_i, 1)$ for $i = 1, \ldots, n$ with $n = 10^4$, only 0.3% of null hypotheses are rejected via a Bonferroni corrected test (indicated by the FWER critical value). However, the Z-scores appear skewed positive, suggesting additional discoveries exist. (b) Our estimator $\widehat{\zeta}_{KS}$ indicates that there are many discoveries to be made; for example, at least 9% of treatments have effect size at least 1, and at least 4% have effect size at least 2. Note that our estimator also counts more discoveries at each threshold than are identified by Bonferroni correction ($\widehat{\zeta}_{FWER}$), without exceeding the true value ζ_*. (c) The experimenter designs an experiment to identify the effects greater than 2, and allocates $\gamma^{-2}\log(n)\log(1/\widehat{\zeta}(\gamma)) = 8$ replicates per hypothesis. Now, 14% of the null hypotheses can be rejected.

Finally, we apply several Taylor series approximations to bound the optimal value as a polynomial in ζ_* and γ_*. □

We present a novel lower bound for the estimation problem which matches our result up to constants.

Lemma 2.2. *Consider data $\{X_i\}_{i=1}^n$ generated under the model $X_i \sim P(\zeta, \gamma)$, parameterized by $\zeta \in (0, \frac{1}{2})$ and $\gamma \in (0, \sigma)$ according to our canonical two-spike model (5). Fix a parameterization (ζ_*, γ_*). For any $\varepsilon \in (0, \frac{2}{3}\zeta_*)$, define the set A_ε of nearby parameterizations as*

$$A_\varepsilon = \left\{(\zeta, \gamma) \; : \; |\zeta_* - \zeta| \le 4\varepsilon, \; \tfrac{1}{3}\gamma_* \le \gamma \le 3\gamma_*\right\}.$$

Suppose $\widehat{\zeta}_n(X)$ is an estimator of ζ satisfying $\mathbb{P}(|\widehat{\zeta}_n(X) - \zeta| \ge \varepsilon) < \frac{1}{4}$ for any $(\zeta, \gamma) \in A_\varepsilon$. Then the estimator requires at least $n \gtrsim \frac{\sigma^4}{\varepsilon^2\gamma_*^4}$ samples on the instance (ζ_*, γ_*).

Instantiating Lemma 2.2 with $\varepsilon = \frac{1}{2}\zeta_*$, we see that the sample complexity in Corollary 2.1.2 matches the lower bound.

Finally, we remark that estimating the fraction of observations with mean γ_* to constant multiplicative error requires a factor of $\frac{\sigma^2}{\gamma_*^2}$ more samples than testing whether any observations have mean γ_*.

2.4. Proof of Theorem 2.1

We begin with the first part of the theorem, which bounds the probability of overestimating ζ_*. Let A be the event that $\widehat{F}_n$ stays within its Dvoretzky-Kiefer-Wolfowitz (DKW) confidence interval, i.e.,

$$A := \left\{||F_{\nu_*} - \widehat{F}_n||_\infty \le \tau_{\alpha,n}\right\}.$$

By the DKW inequality (Massart, 1990), we have $P(A^c) \le \alpha$. If we assume that event A holds, then

$$\begin{aligned}
\widehat{\zeta}_n(\gamma) &= \max\left\{\zeta \ge 0 : \min_{\nu \in S(\zeta,\gamma)} ||F_\nu - \widehat{F}_n||_\infty > \tau_{\alpha,n}\right\} \\
&\overset{(a)}{\le} \min\left\{\zeta \ge 0 : \min_{\nu \in S(\zeta,\gamma)} ||F_\nu - \widehat{F}_n||_\infty \le \tau_{\alpha,n}\right\} \\
&\overset{(b)}{\le} \zeta^*(\gamma)
\end{aligned}$$

where (a) holds because $S(\zeta, \gamma) \subseteq S(\zeta', \gamma)$ for all $\zeta \le \zeta'$, and (b) is true because, by event A, $\zeta_*(\gamma)$ is a member of the set. We note that on event A, this argument holds for all γ. We conclude that, with probability at least $1 - \alpha$, we have $\widehat{\zeta}_n(\gamma) \le \zeta^*(\gamma)$ for all γ.

To prove the power statement, let B be as follows

$$B := \left\{||F_{\nu_*} - \widehat{F}_n||_\infty \le \tau_{\delta,n}\right\}.$$

Suppose that B holds. Then,

$$\begin{aligned}
&\widehat{\zeta}_n(\gamma) = \max\left\{\zeta \ge 0 : \min_{\nu \in S(\zeta,\gamma)} ||F_\nu - \widehat{F}_n||_\infty \ge \tau_{\alpha,n}\right\} \\
&\overset{(a)}{\ge} \max\left\{\zeta \ge 0 : \min_{\nu \in S(\zeta,\gamma)} ||F_\nu - F_{\nu_*}||_\infty \right. \\
&\qquad\qquad\qquad\qquad \left. - ||\widehat{F}_n - F_{\nu_*}||_\infty \ge \tau_{\alpha,n}\right\} \\
&\overset{(b)}{\ge} \max\left\{\zeta \ge 0 : \min_{\nu \in S(\zeta,\gamma)} ||F_\nu - F_{\nu_*}||_\infty \ge \tau_{\alpha,n} + \tau_{\delta,n}\right\} \\
&\ge \max\left\{\zeta \ge 0 : \min_{\nu \in S(\zeta,\gamma)} ||F_\nu - F_{\nu_*}||_\infty \ge \sqrt{\frac{\log(4/\alpha\delta)}{n}}\right\},
\end{aligned}$$

where (a) is the triangle inequality and (b) applies event B. We can rewrite the assumption on n in the theorem statement as

$$\min_{\nu \in S(\zeta_* - \varepsilon, \gamma)} ||F_\nu - F_{\nu_*}||_\infty \ge \sqrt{\frac{\log(4/\alpha\delta)}{n}}$$

which gives us $\widehat{\zeta}_n(\gamma) \geq \zeta_* - \varepsilon$, simultaneously for all γ, whenever B holds. Since B holds with probability at least $1 - \delta$, this completes the proof. □

3. Applications to Pilot Experiments

In this section, we consider the task of analyzing a pilot experiment to count the number of discoveries when the effect sizes are small, say $\mu < 1$ for the case of Gaussian $\mathcal{N}(\mu, 1)$ observations. Figure 2 illustrates this process. First, the scientist allocates a small number of replicates to a large number of hypotheses in order to obtain many noisy estimates of effect sizes (Fig. 2(a)). The scientist then uses our estimator to obtain a guarantee on the number of discoveries to be made at each effect size (Fig. 2(b)). Finally, the scientist calculates the cost of identifying the discoveries that have been detected using a choice of fixed and sequential experimental designs. When the full experiment is run, it results in at least as many discoveries as our estimator has guaranteed (Fig. 2(c)).

The following proposition describes our estimator's performance on pilot data in the low signal-to-noise regime. In particular, if the pilot study design allocates its replicates equally across all n hypotheses, our estimator detects the alternate hypotheses using a factor of n fewer total replicates than it would take to identify these discoveries.

Proposition 3.1. *Consider a pilot experiment for n hypotheses, where an initial budget of $B = mt$ will be used to uniformly allocate t replicates to each of $m \leq n$ randomly chosen hypotheses. Suppose the true distribution of effect sizes is $\nu_* = (1 - \zeta_*)\delta_0 + \zeta_*\delta_{\gamma_*}$ and $f_\mu = \mathcal{N}(\mu, \frac{1}{t})$, as when computing Z-scores from t replicates. Then,*

$$\mathbb{P}(\widehat{\zeta}_n(0) > 0) \geq 1 - \delta$$

i.e., we detect the presence of positive effects with high probability, as long as

$$\gamma_* \geq 4\sqrt{\frac{\log(\frac{2}{\delta})}{\zeta_*^2 B}} \qquad \textit{and} \qquad m \geq \frac{4\log(\frac{2}{\delta})}{\zeta_*^2}.$$

Remark 3.1. *Consider the setting where the budget is constrained, say $B \lesssim n$, and let ζ_* be constant (so that the proportion of discoveries does not vanish with n). Proposition 3.1 suggests maximizing m: either taking $m = n$ if $B \geq n$ or taking $t = 1$ if $B < n$. With this budget allocation, our estimator detects the existence of alternate effects with just $B \approx \log(\frac{1}{\delta})\gamma_*^{-2}$ total replicates. Note that distinguishing observations from $\mathcal{N}(0, 1)$ and $\mathcal{N}(\gamma_*, 1)$ with probability $1 - \delta$ requires $\log(\frac{1}{\delta})\gamma_*^{-2}$ samples, so identifying all of the discoveries requires at least $n\log(\frac{1}{\delta})\gamma_*^{-2}$ replicates total. We conclude that, in this instance, any identification procedure requires n times more total replicates than our estimator requires for detection.*

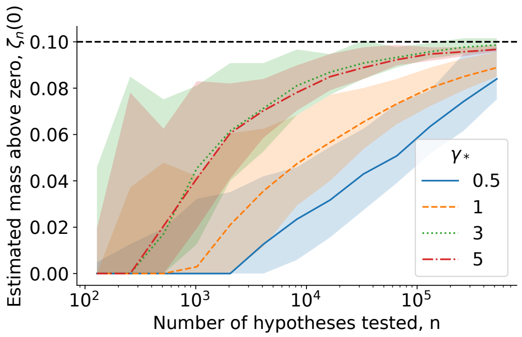


Figure 3. Median and 90% bootstrapped confidence intervals for $\widehat{\zeta}_n(0)$, where $\nu_* = (1 - \zeta_*)\delta_0 + \zeta_*\delta_{\gamma_*}$, for various γ_*. As n increases, for a constant $\zeta_* = 0.1$, our estimator $\widehat{\zeta}_n(0)$ converges to ζ_* without overestimating. As expected, the estimates are lower (have more error) when the alternate effect size γ_* is small.

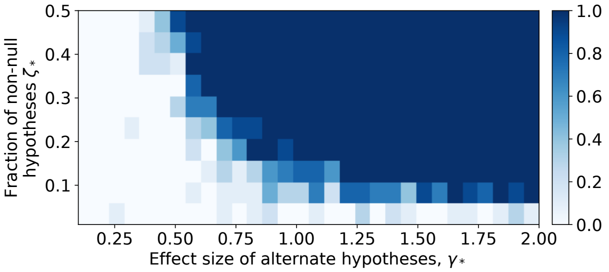


Figure 4. Empirical $\mathbb{P}(\widehat{\zeta}_n(0) > \frac{1}{2}\zeta_*)$, for various parameterizations (γ_*, ζ_*) of the two-spike Gaussian model (5). For a fixed value of $n = 10^4$, the probability of detecting at least half of the discoveries increases with both γ_* and ζ_*. Empirical probabilities were computed over ten trials.

Our estimator can also be used to choose between a fixed experimental design (in which each hypothesis is tested with a fixed number of replicates) and a sequential design (in which the next replicate is allocated after observing all previously drawn X_i). A sequential design, as in Jamieson & Jain (2018), can be more difficult to implement, but could result in significant savings if the alternate effect sizes span a large range. By providing information about the variety of effect sizes, our estimator quantifies the advantage of using a sequential experimental design.

4. Experiments

Details of our implementation can be found in Appendix C. A Python implementation is available at `https://github.com/jenniferbrennan/CountingDiscoveries/`.

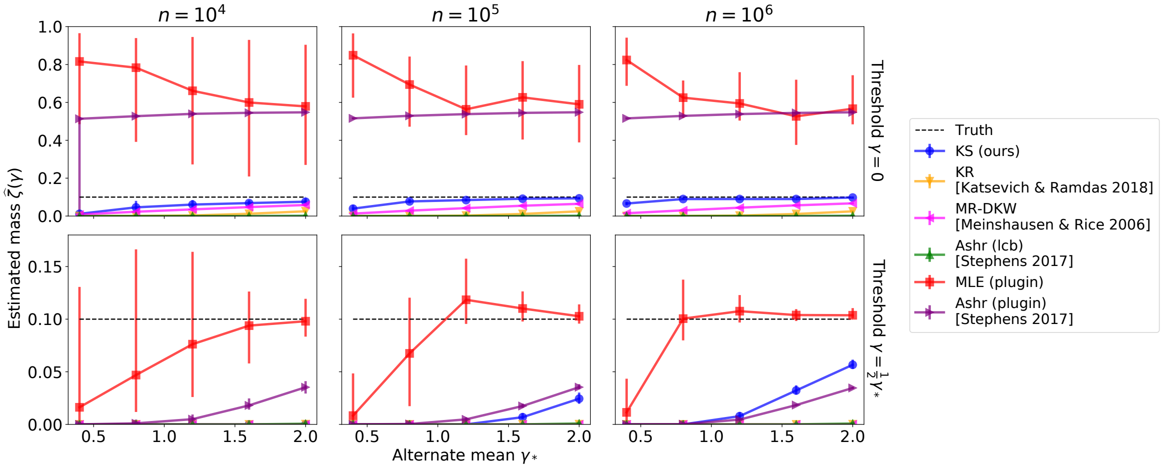


Figure 5. Our estimator outperforms the baselines in the mixture of two Gaussians setting, Eqn (5), returning $\widehat{\zeta}_n(\gamma)$ close to the truth without overestimating, for a wide variety of n and γ.

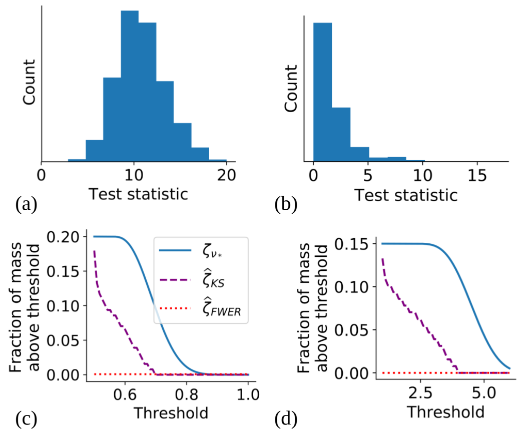
4.1. Experimental Results on Simulated Data

We evaluate our estimator on both real and simulated data. We begin with the mixture of two Gaussians described by Eqn (5). Figure 3 shows the rate of convergence of our estimator for different values of γ_*, the alternate effect size. Note that the estimate never exceeds the true value ζ_*, and that it improves as n increases. The variance of our estimator, shown with bootstrapped 90% confidence intervals, can be large for small n but decreases as n increases.

For a fixed value of n, we are interested in the probability that our estimator detects at least half of the discoveries, $\mathbb{P}_{\nu_*}(\widehat{\zeta}_n(0) \geq \frac{1}{2}\zeta_*)$, as a function of the fraction of discoveries ζ_* and their effect size γ_*. Our estimator exhibits a sharp transition between detecting fewer than half and detecting more than half, as shown in Figure 4.

We compare our estimator to several baselines found in the literature. Figure 5 shows the performance of various estimators in the mixture of two Gaussians setting. Each of the six panels shows how the estimators perform as the alternate mean γ_* varies, for different numbers of hypotheses n and tested thresholds γ. We see that the two plugin methods, the MLE and the plugin `ashr` estimate (Stephens, 2017), both fail to satisfy our constraint $\widehat{\zeta}(\gamma) \leq \zeta_*(\gamma)$. Among the four remaining estimators, ours comes closest to the truth. Notably, our estimator continues to improve as n increases, whereas the baselines do not. We conclude that our estimator outperforms existing methods in the regime of large n and $\gamma_* \approx \sigma$. Appendix D includes a plot comparing only the four estimators that satisfy our constraint.

We also demonstrate that our method works on distributions other than Gaussians by applying it to synthetic Poisson and binomial data (Figure 6). Details of the experiments can be found in Appendix C; we note that our test detected the presence of the alternate hypotheses even when no alternates were identifiable via Bonferroni-corrected multiple testing.



(a)

(b)

(c)

(d)

Figure 6. Performance of our estimator on binomial (left) and Poisson (right) data. Top panels show the observed histogram of X_i. Bottom panels show the true fraction of effects above each threshold (ζ_{ν_*}), as well as estimates using our method ($\widehat{\zeta}_{KS}$) and identification via Bonferroni-corrected multiple testing ($\widehat{\zeta}_{FWER}$). Our estimator gets closer to the truth, without overestimating.

4.2. Experimental Results on Real Data

We evaluated our estimator on Z-scores from an experiment to identify which genes contribute to influenza replication in *Drosophila*, described by Hao et al. (2008). The data, available in our supplementary material, consisted of Z-scores from two replicates for each of 13,071 genes. Figure 7(a) shows the empirical distribution of the 13,071 averaged Z-scores, which are the observations X_i. The theoretically motivated distribution $X_i \sim \mathcal{N}(\mu, \frac{1}{2})$ is a poor fit to this data, perhaps due to undocumented pre-processing steps

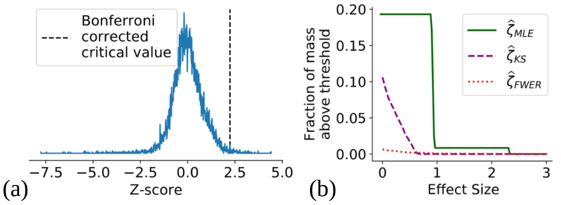


Figure 7. Two Z-scores were averaged for each of 13,071 *Drosophila* genes. Even though (a) indicates that very few discoveries could be made, (b) shows that the MLE suggests, and our estimator confirms, many discoveries exist. We note that the MLE provides no guarantee of a conservative estimate, and may drastically overestimate the true fraction at any point.

not annotated in the dataset, so we began by estimating the variance of these observations. We found that $\sigma^2 = \frac{1}{4}$ provided a good fit to the data; we used this value for the rest of our computations. Testing for significant effect sizes using Bonferroni correction at the 0.05 level (critical value shown in Figure 7(a)) resulted in 83 discoveries, representing 0.6% of genes. Given the low number of replicates performed in this experiment, we might suspect that there are more discoveries with smaller effect sizes.

Figure 7(b) shows the results of the plug-in MLE estimator $\widehat{\zeta}_{MLE}$, our estimator ($\widehat{\zeta}_{KS}$), and identification with Bonferroni correction ($\widehat{\zeta}_{FWER}$). The fitted MLE suggests that there are around 2600 discoveries to be made (20% of genes), with most effect sizes around 1. As discussed previously, the MLE can overestimate the true number of discoveries and their effect sizes. Our conservative estimator guarantees that there are at least 1400 genes (11% of all genes) with positive effects, including at least 190 genes (1.5%) with effect size of at least 0.5. Our estimator generally detected more discoveries than $\widehat{\zeta}_{FWER}$, excluding the influence of the 23 genes (0.2%) with $X_i > 3$. These observations fall into the *sparse regime* (Donoho & Jin, 2004), where our estimator has less power. These results could facilitate the design of an experiment to identify genes with effect sizes exceeding some threshold, or upper bound the cost of a sequential experiment to identify the top 200 genes.

5. Discussion and Future Work

We have presented an algorithm that estimates the fraction of a mixing distribution that lies above some threshold, subject to the constraint that the estimate does not exceed the true fraction with high probability. Our algorithm can be generalized to the following template:

1. Choose some distance metric on CDFs.
2. Find the set $\mathcal{A}$ of "plausible" F_ν given observation $\widehat{F}_n$, which are the CDFs such that $d(F_\nu, \widehat{F}_n) < \tau_{\alpha,n}$.
3. Choose $\tau_{\alpha,n}$ such that $\mathbb{P}(d(\widehat{F}_n, F_{\nu_*}) > \tau_{\alpha,n}) \leq \alpha$.

Returning the minimum amount of mass above the threshold, over the set of plausible distributions $\mathcal{A}$, guarantees with high probability that we do not overestimate the true mass. Our algorithm instantiates this template with the ℓ_∞ norm as the distance metric.

Another natural choice of metric is the likelihood of $\widehat{F}_n$ given F_ν. In order to use this in our template, we need finite sample bounds on the likelihood of $\widehat{F}_n$ given F_{ν_*}. Asymptotic versions of these results are worked out by Jiang & Zhang (2016) for the case of Gaussian X_i, and it would be easy to extend these to finite sample bounds. Extensions of Jiang & Zhang's results should show that the resulting estimator is optimal throughout the so-called *sparse* and *dense* regimes (Donoho & Jin, 2004). Unfortunately, their value of $\tau_{\alpha,n}$ depends on unknown constants, and therefore it would require extensive simulations with thousands of repetitions for each (ζ, γ) pair to obtain a reliable estimate of the critical values for pilot study analysis. In addition, using the likelihood-based approach for a new distribution f_μ requires an entirely new proof of the high-probability bound on the likelihood ratio.

We believe it would be possible to modify our estimator in order to improve its performance in the sparse regime, where effects are large but rare. Our estimator uses the DKW inequality (Massart, 1990) to measure the plausibility of a latent distribution ν, but the DKW inequality is not tight where the empirical CDF has low variance. Such points occur in the sparse regime, for example at $F_{\nu_*}(\frac{1}{2}\gamma_*)$. Applying a bound that uses variance information, such as an empirical Bernstein DKW (as surveyed in, e.g., (Howard & Ramdas, 2019)), could address this lack of power.

Any algorithm for this problem necessarily makes some assumptions about the data generating process, otherwise all observations could come from the alternate, with $X \sim P_1$ having density $\widehat{F}_n$. As discussed in Section 1.3, previous works have used various assumptions, such as unimodality of ν_* or "purity" of p-values around zero. Our key assumption is the parametric form of X_i, under both the null and the alternate. In practice, in order to decrease our reliance on this assumption, we could learn some parameter of the test statistic distribution from the observations themselves. This was our approach with the *Drosophila* data, when we fit the variance σ^2 of the Z-score. This approach is also taken by Efron (2007). Even more ambitiously, we could learn the conditional distribution $f(X|\mu)$ by fitting it jointly with the means μ, and then use this conditional distribution to generate F_ν from a candidate distribution ν.

Acknowledgements

We thank Lalit Jain and Weihao Kong for their insightful comments, and we thank Swati Padmanabhan for valuable discussions about convex optimization. Jennifer Brennan is supported by an NSF Graduate Research Fellowship. This work was supported in part by NSF IIS-1907907.

References

Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

Belkin, M. and Sinha, K. Polynomial learning of distribution families. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 103–112. IEEE, 2010.

Benjamini, Y. and Hochberg, Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics*, 25(1):60–83, 2000.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Cai, T. T., Jin, J., Low, M. G., et al. Estimation and confidence sets for sparse normal mixtures. *The Annals of Statistics*, 35(6):2421–2449, 2007.

Carpentier, A., Verzelen, N., et al. Adaptive estimation of the sparsity in the Gaussian vector model. *The Annals of Statistics*, 47(1):93–126, 2019.

Chandrasekaran, K. and Karp, R. Finding a most biased coin with fewest flips. In *Conference on Learning Theory*, pp. 394–407, 2014.

Chen, X. Uniformly consistently estimating the proportion of false null hypotheses via Lebesgue–Stieltjes integral equations. *Journal of Multivariate Analysis*, 2019.

Daskalakis, C. and Kamath, G. Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians. In *Conference on Learning Theory*, pp. 1183–1213, 2014.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Donoho, D. and Jin, J. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 06 2004. doi: 10.1214/009053604000000265.

Dunn, O. J. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

Efron, B. et al. Size, power and false discovery rates. *The Annals of Statistics*, 35(4):1351–1377, 2007.

Genovese, C., Wasserman, L., et al. A stochastic process approach to false discovery control. *The Annals of Statistics*, 32(3):1035–1061, 2004.

Hao, L., Sakurai, A., Watanabe, T., Sorensen, E., Nidom, C. A., Newton, M. A., Ahlquist, P., and Kawaoka, Y. Drosophila RNAi screen identifies host genes important for influenza virus replication. *Nature*, 454(7206):890, 2008.

Hardt, M. and Price, E. Tight bounds for learning a mixture of two Gaussians. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 753–760. ACM, 2015.

Howard, S. R. and Ramdas, A. Sequential estimation of quantiles with applications to a/b-testing and best-arm identification. *arXiv preprint arXiv:1906.09712*, 2019.

Jamieson, K. G. and Jain, L. A bandit approach to sequential experimental design with false discovery control. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 3660–3670. Curran Associates, Inc., 2018.

Jamieson, K. G., Haas, D., and Recht, B. The power of adaptivity in identifying statistical alternatives. In *Advances in Neural Information Processing Systems*, pp. 775–783, 2016.

Jiang, W. and Zhang, C.-H. Generalized likelihood ratio test for normal mixtures. *Statistica Sinica*, 26:955–978, 07 2016. doi: 10.5705/ss.202015.0086.

Jin, J. Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):461–493, 2008.

Kalai, A. T., Moitra, A., and Valiant, G. Efficiently learning mixtures of two Gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 553–562. ACM, 2010.

Katsevich, E. and Ramdas, A. Simultaneous high-probability bounds on the false discovery proportion in structured, regression, and online settings. *arXiv preprint arXiv:1803.06790*, 2018.

Lee, J. C. and Valiant, P. Uncertainty about uncertainty: Near-optimal adaptive algorithms for estimating binary mixtures of unknown coins. *arXiv preprint arXiv:1904.09228*, 2019.

Li, A. and Barber, R. F. Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):45–74, 2019.

Massart, P. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990.

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

Meinshausen, N. and Rice, J. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics*, 34(1): 373–393, 2006.

Moitra, A. and Valiant, G. Settling the polynomial learnability of mixtures of Gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 93–102. IEEE, 2010.

Patra, R. K. and Sen, B. Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):869–893, 2016.

Pearson, K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

Schweder, T. and Spjøtvoll, E. Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69(3):493–502, 1982.

Stephens, M. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017.

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

Tian, K., Kong, W., and Valiant, G. Learning populations of parameters. In *Advances in Neural Information Processing Systems*, pp. 5778–5787, 2017.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer-Verlag New York, 1 edition, 2009. ISBN 978-0-387-79051-0.

Vinayak, R. K., Kong, W., Valiant, G., and Kakade, S. Maximum likelihood estimation for learning populations of parameters. In *International Conference on Machine Learning*, pp. 6448–6457, 2019.