Statistica Sinica **35** (2025), 1155-1181
doi:https://doi.org/10.5705/ss.202023.0128

UNIFICATION OF RARE AND WEAK MULTIPLE TESTING MODELS USING MODERATE DEVIATIONS ANALYSIS AND LOG-CHISQUARED P-VALUES

Alon Kipnis

Reichman University

Abstract: Rare and Weak models for multiple hypothesis testing assume that only a small proportion of the tested hypotheses concern non-null effects and the individual effects are only moderately large, so they generally do not stand out individually, for example in a Bonferroni analysis. Such models have been studied in quite a few settings, for example in some cases studies focused on an underlying Gaussian means model for the hypotheses being tested; in some others, Poisson and Binomial. Such seemingly different models have the following common structure. Summarizing the evidence of individual tests by the negative logarithm of its P-value, the model is asymptotically equivalent to a situation in which most negative log P-values have a standard exponential distribution but a small fraction might have an alternative distribution which is approximately noncentral chisquared on one degree of freedom. This log-chisquared approximation is different from the log-normal approximation of Bahadur. The latter is unsuitable for analyzing Rare and Weak multiple-testing models.

We characterize the asymptotic performance of global tests combining asymptotic log-chisquared P-values in terms of the chisquared mixture parameters: the scaling parameter controlling heteroscedasticity, the non-centrality parameter, and the parameter controlling the rarity of individual non-null effects. In a phase space involving the last two parameters, we derive a region where all tests are asymptotically powerless. Outside of this region, the Berk-Jones and the Higher Criticism tests have maximal power. Inference techniques based on the minimal P-value, false-discovery rate controlling, and Fisher's combination test have sub-optimal asymptotic phase diagrams. Our analysis yields the asymptotic power of global testing in various new rare and weak models, including two-sample heteroscedastic normal mixtures and binomial experiments with perturbed probabilities of success.

Key words and phrases: Heterogeneous mixture, higher criticism, multiple testing, P-values, sparse mixture.

1. Introduction

1.1. Motivation

Consider a multiple hypothesis testing situation, each test involves a different feature of the data where different features are independent. We are interested

*Corresponding author. E-mail: alon.kipnis@runi.ac.il

in testing a global null hypothesis against the following alternative: the non-null effects are concentrated in a small, but unknown, subset of the hypotheses. In the most challenging situation, effects are not only rare but also weak in the sense that the non-null test statistics are unlikely to provide evidence after Bonferroni's correction. Rare and weak multiple hypothesis testing problems of this nature arise in a wide range of situations (Donoho and Jin, 2015). Specific examples include:

- *Sparse (rare) signal detection.* We are interested in intercepting a transmission that occupies few frequency bands out of potentially many, while the occupied bands are unknown to us (Tandra and Sahai, 2008; Bayer and Seljak, 2020). The features are periodogram ordinates associated with individual frequency bands. Evidence for the presence of a signal can be gathered by testing each ordinate against the same exponential distribution.

- *Classification.* Classifying images or other high-dimensional signals usually involves hundreds or more features. In a one-versus-all classification setup, we view the typical response of the features under each class as the null hypothesis. Testing against this null amounts to determining whether the tested signal is associated with that class or not. A situation of wide interest is when inter-class discrimination is due to a small proportion of features out of potentially many, and we do not know which ones they are likely to be (Donoho and Jin, 2009; Ingster, Pouet and Tsybakov, 2009; Jin, 2009).

- *Detecting rare changes between two high-dimensional distributions.* Testing whether two high-dimensional datasets are simply two different realizations of the same data-generating mechanism is a classical problem in statistics, computer science, and information theory (Acharya et al., 2012; Balakrishnan and Wasserman, 2018; Donoho and Kipnis, 2022). This scenario is formulated as a two-sample testing problem; the null hypothesis states that both samples were obtained from the same high-dimensional parent distribution. The alternative hypothesis states that differences between the mechanisms occur in a small and unknown subspace of the parameters.

Applications as above have motivated a significant body of work in rare and weak multiple testing settings throughout the past two decades, providing fruitful insights for signal detection, feature selection, and classification problems in high dimensions (Jin and Ke, 2016). Specific examples of rare and weak multiple testing settings include normal mixtures (Ingster and Suslina, 2012; Jin, 2003; Donoho and Jin, 2004; Abramovich et al., 2006), binomial mixtures (Mukherjee, Pillai and Lin, 2015), linear regression model under Gaussian noise (Arias-Castro, Candès and Plan, 2011; Ingster, Tsybakov and Verzelen, 2010), Poisson mixtures (Arias-Castro and Wang, 2015), heteroscedastic normal mixtures (Cai, Jeng and Jin, 2011), general mixtures (Cai and Wu, 2014; Arias-Castro and Wang, 2017),

mixture of unknown distributions (Delaigle and Hall, 2009; Delaigle, Hall and Jin, 2011; Arias-Castro and Wang, 2017), and several two-sample settings (Donoho and Kipnis, 2022; Galili, Kipnis and Yakhini, 2023).

1.2. Contributions

In this article, we study one rare and weak multiple testing setting that subsumes the vast majority of these previously studied ones. Our setting is not tied to a specific data-generating model. Instead, we model the behavior of a collection of P-values, each P-value summarizes the evidence of one test statistic against the global null. These P-values may be obtained either from one- or two-sample tests and may represent responses over a variety of models. More generally, the advantages of modeling the distribution of the P-values rather than the data are discussed in Lambert (1981), Lambert and Hall (1982), Sackrowitz and Samuel-Cahn (1999), and Boos and Stefanski (2011).

Recall that a deviation from the mean of a sequence $X_1, X_2, \ldots$, of standardized and identically and independently distributed random variables is said to be *moderate* if it is of the form $\sqrt{q \log(n)/n}$ for some $q > 0$ (Rubin and Sethuraman, 1965; Dembo and Zeitouni, 1998). For such deviations, Cramér's theorem implies

$$\Pr\left[\left|\frac{X_1 + \ldots + X_n}{\sqrt{n}}\right| > \sqrt{2q \log(n)}\right] \sim n^{-q}, \quad \text{uniformly in } q \geq a > 0, \qquad (1.1)$$

provided the moment-generating function exists. Our key insight says that the log-chisquared approximation (see (4.2) below) – and not the log-normal approximation of Bahadur (1960) and Lambert and Hall (1982) – is accurate for characterizing the asymptotic power of testing in rare and weak models involving departures on the moderate scale. Consequently, our setting unifies all previously studied rare and weak settings in which moderate deviations analysis applies under one setting we denote as the Rare Moderate Departures (RMD) model. This unification provides new characterizations for the power of some global testing procedures in those earlier studied settings and in several new settings as summarized in Table 1. Additionally, our analysis guides the calibration of parameters of new high-dimensional signal models to experience a phase transition between the rarity and strength of individual effects; see Galili, Kipnis and Yakhini (2023) for an example in the context of survival analysis.

1.3. The log-chisquared approximation

In order to introduce the log-chisquared perturbation model, suppose that the i-th test statistic yields the P-value p_i, for $i = 1, \ldots, n$. We further assume that $p_i \sim \text{Unif}(0, 1)$ under the global null, corresponding to the case where the model underlying the i-th test statistics has a continuous distribution (we relax this assumption later on). Consequently, $-2\log(p_i) \sim \text{Exp}(2)$, where Exp(2)

Table 1. Rare and weak multiple testing settings that are carried under our RMD formulation and their noncentral chisquared parameters. Models appearing in bold are new.

Departures model	Chisquared parameters		studied in
	ρ	σ	
Normal means (heteroscedastic)	r	s	(Cai, Jeng and Jin, 2011)
Two-sample normal means (homoscedastic)	$r/2$	1	(Donoho and Kipnis, 2022)
Two-sample normal means (heteroscedastic)	$r/2$	$\sqrt{(1+s^2)/2}$	
Poisson means	r	1	(Arias-Castro and Wang, 2015)
Two-sample Poisson means	$r/2$	1	(Donoho and Kipnis, 2022)
Binomial success probability	r	$s \cdot r$	

is the exponential distribution with mean 2 (or rate 1/2), also known as the chisquared distribution with two degrees of freedom χ^2_2. Our model proposes the following alternative: Roughly $n\epsilon$ of the P-values depart from their uniform distribution and instead obey

$$-2\log(p_i) \overset{D}{\approx} (\mu + \sigma Z)^2, \qquad Z \sim \mathcal{N}(0,1). \tag{1.2}$$

Here $\overset{D}{\approx}$ indicates a specific form of approximation in distribution that we formalize in Section 2 below. Leaving the details of this approximation aside for now, (1.2) says that $-2\log(p_i)$ is approximately distributed as a scaled noncentral chisquared random variable (RV) over one degree of freedom with noncentrality parameter μ, and scaling parameter σ. We focus on the case where the rarity parameter ϵ vanishes while the intensity parameter μ is only moderately large, making our global testing problem challenging; in some cases, impossible. As we shall see, in this regime the non-null effects are not only rare but are also weak in the sense that they generally do not stand out individually in a Bonferroni analysis.

1.4. Log-chisquared versus log-normal

The emergence of the log-chisquared approximation for P-values is somewhat surprising because this approximation is different from the log-normal approximation developed in Bahadur (1960) and Lambert and Hall (1982). In Section 4, we show that the log-chisquared distribution fits the distribution of the P-values under moderate departures significantly better than the log-normal distribution. Furthermore, the log-normal approximation does not indicate the correct asymptotic performance of tests under rare and weak multiple testing settings. To summarize this last point, we establish here that a rare

multiple hypothesis testing setting in which the departures are on the moderate scale corresponds to detecting a few noncentral chisquared signals against an exponential background, rather than detecting a few normal signals as one might have proposed in view of the log-normal approximation. Potential applications of this improved approximation, beyond the asymptotic power analysis of multiple hypothesis testing we discuss in this paper, include better estimates of the so-called "reproducibility probability" of experiments (Boos and Stefanski, 2011) and empirical Bayes method for identifying discoveries in large-scale inference (Efron et al., 2001; Pounds and Morris, 2003); we leave these topics as future work.

We note that the logarithmic scoring scale for P-values goes back to Fisher, who initially suggested it as a method of ranking success in card-guessing games (Fisher, 1924). For global testing, Fisher proposed the statistic (Fisher, 1992)

$$F_n := \sum_{i=1}^{n} -2\log(p_i), \tag{1.3}$$

which has a χ^2_{2n} distribution under the global null. A test based on F_n is known to be effective in the presence of small effects distributed across the bulk of cases, but not effective under relatively rare and somewhat stronger but individually still weak as our model proposes; see a formal statement about the inadequacy of F_n in our setting in Theorem 6 below. The logarithmic scale for P-values is now standard in genome-wide association studies (GWAS) (Balding, 2006; Pearson and Manolio, 2008; Price et al., 2010; Harold et al., 2009) and in other areas (Li, 2012; Quaino et al., 2014; Boos and Stefanski, 2011; Gibson, 2021). Our setting yields an explicit model for testing rare and weak effects in these applications: testing chisquared departures against an exponential background. A similar model arises in detecting the presence of rare and weak sinusoids in white noise based on the periodogram. For this setting, Fisher's periodogram test is based on the largest periodogram ordinate (Fisher, 1929) which is analogous to a Bonferroni analysis.

1.5. Paper organization

In Section 2 we define the RMD setting and analyze the asymptotic properties of tests. In Section 3 we explore several rare and weak signal detection problems that conform to the RMD model formulation. In Section 4 we compare our log-chisquared approximation for the distribution of P-values under the alternative hypothesis and the classical log-normal approximation. Additional discussions are provided in Section 5. All proofs are in the Supplementary Material.

2. Rare Moderate Departures Setting and Analysis

2.1. Multiple testing with rare effects

The description in the Introduction above depicts the following global hypothesis testing setting involving a sequence of P-values $p_1, \ldots, p_n$.

$$\begin{aligned} H_0 &: \ -2\log(p_i) \sim \mathrm{Exp}(2), \quad i = 1, \ldots, n, \\ H_1^{(n)} &: \ -2\log(p_i) \sim (1-\epsilon)\mathrm{Exp}(2) + \epsilon Q_i^{(n)}, \quad i = 1, \ldots, n, \end{aligned} \tag{2.1}$$

where $Q_i^{(n)}$ is a probability distribution specifying the non-null behavior of the i-th P-value.

We calibrate the rarity parameter ϵ to n according to

$$\epsilon = \epsilon_n := n^{-\beta}, \tag{2.2}$$

where $\beta \in (0, 1)$. This calibration proposes that for an overwhelming majority of the individual tests, the response under the alternative is indistinguishable from the null.

The expected proportion of the non-null effects (rarity) of most interest under RMD is $n^{-\beta}$ for $\beta \in (1/2, 1)$. Indeed, for less rare effects ($\beta < 1/2$), the signal may be viewed as "dense" in the sense that tests that are powerful against small but frequent departures can also be asymptotically powerful (Arias-Castro, Candès and Plan, 2011).

2.2. The log-chisquared approximation

We compare $Q_i^{(n)}$ to the non-central and scaled chisquared distribution as in the right-hand side of (1.2) with non-centrality parameter μ is calibrated to n as in:

$$\mu = \mu_n(\rho) := \sqrt{2\rho\log(n)}, \qquad \rho > 0, \tag{2.3}$$

and with a fixed scaling parameter σ. Specifically, define the moderately perturbed and scaled chisquared distribution

$$\chi^2(\rho, \sigma) \stackrel{D}{=} (\mu_n(\rho) + \sigma Z)^2, \qquad Z \sim \mathcal{N}(0, 1),$$

where $\stackrel{D}{=}$ indicates equality in distribution. For the sake of formalizing the approximation in (1.2), we introduce the function

$$\alpha(q; \rho, \sigma) := \left(\frac{\sqrt{q} - \sqrt{\rho}}{\sigma}\right)^2, \tag{2.4}$$

and note that

$$\lim_{n\to\infty} \frac{-\log \Pr\left[\chi^2(\rho,\sigma) \geq 2q\log(n)\right]}{\log(n)} = \alpha(q;r,\sigma), \qquad q > \rho.$$

A sequence of distributions $\{Q_i^{(n)}\}_{i=1}^n$ with $\Pr[Q_i^{(n)} \geq 2q\log(n)] > 0$ for all $i = 1,\ldots,n$ is said to be uniformly moderate chisquared if, for every $q > \rho$,

$$\lim_{n\to\infty} \max_{i=1,\ldots,n} \left| \frac{-\log \Pr\left[Q_i^{(n)} \geq 2q\log(n)\right]}{\log(n)} - \alpha(q;\rho,\sigma) \right| = 0. \tag{2.5}$$

Namely, we require that the moderate tail probability of $Q_i^{(n)}$ is identical to that of the non-central and scaled chisquared $\chi^2(\rho,\sigma)$. Henceforth, we refer to hypothesis testing problems of the form (2.1) in which $\{Q_i^{(n)}\}_{i=1}^n$ is uniformly moderate chisquared as Rare Moderate Departures (RMD) model with log-chisquared parameters (ρ,σ). A useful criterion for the validity of (2.5) is

$$Q_i^{(n)} \stackrel{D}{=} \left(\mu_n(\rho) + \sigma Z\right)^2 \left\{1 + o_p(1)\right\}, \quad \rho > 0, \quad n \to \infty, \tag{2.6}$$

where $o_p(1)$ indicates a sequence of RVs tending to zero in probability uniformly in i as $n \to \infty$.

2.3. Asymptotically uniform P-values

We now extend our setting (2.1) to situations in which $p_1,\ldots,p_n$ are not uniformly distributed under the null. We do so by considering, instead of (2.1),

$$\begin{aligned} H_0^{(n)} &: \ -2\log(p_i) \sim E_i^{(n)}, \quad i = 1,\ldots,n, \\ H_1^{(n)} &: \ -2\log(p_i) \sim (1-\epsilon)E_i^{(n)} + \epsilon Q_i^{(n)}, \quad i = 1,\ldots,n, \end{aligned} \tag{2.7}$$

where $Q_i^{(n)}$ satisfies (2.5) and where the probability distribution $E_i^{(n)}$ converges to Exp(2) in the sense that

$$\lim_{n\to\infty} \max_{i=1,\ldots,n} \left| \frac{-\log \Pr\left[E_i^{(n)} \geq 2q\log(n)\right]}{\log(n)} - q \right| = 0 \tag{2.8}$$

for every fixed $q > 0$. This extension of the RMD setting is particularly useful when the distribution of the P-values under the null is only super uniform as in some discrete models (Westfall and Wolfinger, 1997), or when we consider asymptotic P-values rather than exact P-values which is common in large-scale inference from multiple tests (Efron, 2012). Most of the properties of RMD models we derive in this paper hold under this extended setting. In this sense, condition (2.8) provides a range of deviation from the specification of the null distribution under which current and previous results concerning rare and moderately large effects hold.

2.4. Strong log-chisquared approximation

Stronger forms of the chisquared and exponential approximations (2.5) and (2.8) are needed to establish an information-theoretic limit of global testing under models (2.1) and (2.7). For the chisquared approximation, we require that

$$\lim_{n\to\infty} \max_{i=1,\dots,n} \left| \log \left[\frac{dQ_i^{(n)}}{d\chi^2(\rho,\sigma)}(2q\log(n)) \right] \right| \Big/ \log(n) = 0 \tag{2.9}$$

for any $q > \rho$. For the exponential approximation, we require that

$$\lim_{n\to\infty} \max_{i=1,\dots,n} \left| \log \left[\frac{dE_i^{(n)}}{d\mathrm{Exp}(2)}(2q\log(n)) \right] \right| \Big/ \log(n) = 0 \tag{2.10}$$

for any fixed $q > 0$. The type of equivalence between $Q_i^{(n)}$ and $\chi^2(\rho,\sigma)$ described in (2.9) is similar to the setting of Cai and Wu (2014). Henceforth, we refer to hypothesis testing problems of the form (2.1) under the condition (2.9) and (2.10) as the *strong* RMD. In the Supplementary Material, we show that (2.9) implies (2.5) and that (2.10) implies (2.8). Note that (2.9) holds whenever the distribution of each $Q_i^{(n)}$ has a density and satisfies (2.6).

2.5. Asymptotic power and phase transition

RMD models experience a *phase transition* phenomenon in the following sense. For some choice of the parameters r, β, and σ, the two hypotheses are completely indistinguishable. In another region, some tests can asymptotically distinguish $H_1^{(n)}$ from $H_0^{(n)}$ with probability tending to one. Formally, for a given sequence of statistics $\{T_n\}_{n=1}^{\infty}$, we say that $\{T_n\}_{n=1}^{\infty}$ is *asymptotically powerful* if there exists a sequence of thresholds $\{h_n\}_{n=1}^{\infty}$ such that

$$\Pr_{H_0^{(n)}}(T_n > h_n) + \Pr_{H_1^{(n)}}(T_n \le h_n) \to 0,$$

as n goes to infinity. In contrast, we say that $\{T_n\}_{n=1}^{\infty}$ is *asymptotically powerless* if

$$\Pr_{H_0^{(n)}}(T_n > h_n) + \Pr_{H_1^{(n)}}(T_n \le h_n) \to 1,$$

for any sequence $\{h_n\}_{n\in\mathbb{N}}$. The so-called phase transition curve is the boundary of the region in the parameter space (β, r) in which all tests are asymptotically powerless.

Our would-be phase transition curve is

$$\rho^*(\beta,\sigma) := \begin{cases} (2-\sigma^2)(\beta-1/2) & 1/2 < \beta < 1-\sigma^2/4, \quad 0<\sigma^2<2, \\ \left(1-\sigma\sqrt{1-\beta}\right)^2 & 1-\sigma^2/4 \le \beta < 1, \quad 0<\sigma^2<2, \\ 0 & 1/2 < \beta < 1-1/\sigma^2, \quad \sigma^2 \ge 2, \\ \left(1-\sigma\sqrt{1-\beta}\right)^2 & 1-1/\sigma^2 \le \beta < 1, \quad \sigma^2 \ge 2. \end{cases} \tag{2.11}$$

2.6. Information theoretic lower bound

One side of the phase transition follows from an information-theoretic lower bound. This bound requires the strong RMD formulation of (2.9) and (2.10).

Theorem 1. *Consider the hypothesis testing problem* (2.7). *For every* $i = 1,\dots,n$, *assume that* $Q_i^{(n)}$ *is absolutely continuous with respect to* $E_i^{(n)}$, *and let*

$$L_i^{(n)}(x) := \frac{dQ_i^{(n)}}{dE_i^{(n)}}(x) \tag{2.12}$$

be the likelihood ratio between the mixture components. Suppose that there exists $\gamma > 0$ *such that, for any* $q \in (r, 1+\gamma)$,

$$\lim_{n\to\infty} \max_{i=1,\dots,n} \frac{-\log\left(\mathbb{E}_{X\sim E_i^{(n)}}\left[L_i^{(n)}(X)\mathbf{1}_{\{X>2q\log(n)\}}\right]\right)}{\log(n)} \ge \alpha(q;\rho,\sigma), \tag{2.13a}$$

and

$$\lim_{n\to\infty} \max_{i=1,\dots,n} \frac{-\log\left(\mathbb{E}_{X\sim Q_i^{(n)}}\left[L_i^{(n)}(X)\mathbf{1}_{\{X\le 2q\log(n)\}}\right]\right)}{\log(n)} \ge \alpha^*(q;\rho,\sigma), \tag{2.13b}$$

where

$$\alpha^*(q;\rho,\sigma) := \min_{y\in[0,q]} \{2\alpha(y;\rho,\sigma) - y\}. \tag{2.13c}$$

If $\rho < \rho^*(\beta,\sigma)$, *all tests are asymptotically powerless.*

Theorem 1 implies the following:

Corollary 1. *Consider the hypothesis testing problem* (2.7) *under the strong RMD formulation of* (2.9) *and* (2.10). *If* $\rho < \rho^*(\beta,\sigma)$, *all tests are asymptotically powerless.*

Theorem 1 provides conditions for the impossibility of discriminating $H_0^{(n)}$ from $H_1^{(n)}$ in (2.7) that are more general than those provided in Cai and Wu (2014) and in other studies when specialized to our setting.

Figure 1 depicts $\rho^*(\beta,\sigma)$ for three choices of σ. The function $\rho^*(\beta,\sigma)$ was first derived in Cai, Jeng and Jin (2011) to describe the detection boundary of rare and weak normal means with heteroscedastic components. Theorems 1 extends this result from Cai, Jeng and Jin (2011) to general rare and weak multiple

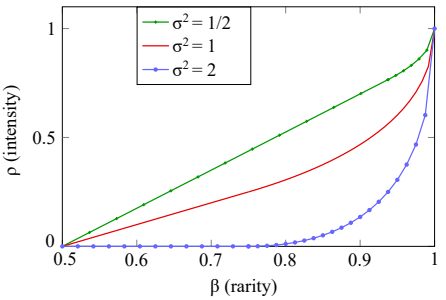

Figure 1. Phase Diagram. The phase transition curve $\rho^*(\beta, \sigma)$ of (2.11) defines the detection boundary in all Rare Moderate Departure models. For $\rho < \rho^*(\beta, \sigma)$, all tests are asymptotically powerless. For $\rho > \rho^*(\beta, \sigma)$, some tests, including Higher Criticism and Berk-Johns, are asymptotically powerful.

testing models obeying the RMD formulation. We discuss several such models in Section 3 below.

2.7. Optimal tests

To complete the phase transition characterization of RMD models initiated in Theorem 1, we consider two tests that are asymptotically powerful whenever $\rho > \rho^*(\beta, \sigma)$.

2.7.1. Higher criticism test

The Higher Criticism (HC) of the P-values $p_1, \ldots, p_n$ is defined as

$$\mathrm{HC}_n^* := \max_{1 \le i \le n\gamma_0} \sqrt{n} \frac{i/n - p_{(i)}}{\sqrt{p_{(i)}\left(1 - p_{(i)}\right)}},$$

where $p_{(i)}$ is the i-th order statistic of $p_1, \ldots, p_n$, and $0 < \gamma_0 < 1$ is a fixed parameter (Donoho and Jin, 2004). The HC test rejects $H_0^{(n)}$ for large values of HC_n^*.

In order to characterize the asymptotic power of HC under (2.7), we restrict the potential sub-uniformity of $p_1, \ldots, p_n$ under $H_0^{(n)}$ beyond what is permitted by (2.8) by requiring

$$\max_{i=1,\ldots,n} -\log \Pr\left[E_i^{(n)} \ge 2q\log(n)\right] \le q\log(n) - \log(1 + n^{(q-1)/2}) \tag{2.14}$$

for all n larger than some $n_0 \in \mathbb{N}$ and for all $q \in (0, 1]$. This restriction is not a concern when $p_1, \ldots, p_n$ are P-values since any super-uniform seqeunce of RVs $\{E_i^{(n)}\}$ satisfies (2.14).

Theorem 2. *Consider the hypothesis testing problem* (2.7) *under* (2.5), *and suppose that* $\{E_i^{(n)}\}$ *obey* (2.8) *and* (2.14). *Fix* $\gamma_0 \in (0, 1/2)$. *If* $\rho > \rho^*(\beta, \sigma)$, *then* HC_n^* *is asymptotically powerful.*

2.7.2. Berk-jones test

Define the P-values

$$\pi_i := \Pr\left(\mathrm{Beta}(i, n-i+1) < p_{(i)}\right), \quad i = 1, \ldots, n,$$

where $\mathrm{Beta}(a, b)$ is the Beta distribution with shape parameters $a, b > 0$. The Berk-Jones (BJ) test statistic is defined as (Berk and Jones, 1979; Moscovich, Nadler and Spiegelman, 2016)

$$M_n := \min\{M_n^-, M_n^+\}, \quad M_n^- := \min_i \pi_i, \quad M_n^+ := \min_i (1 - \pi_i).$$

Theorem 3. *Consider the hypothesis testing problems* (2.1) *under the RMD condition* (2.5). *If* $\rho > \rho^*(\beta, \sigma)$, *than* $1/M_n$ *is asymptotically powerful.*

2.8. Sub-optimal tests

2.8.1. Bonferroni and false-discovery rate controlling

Bonferroni and false-discovery rate (FDR) controlling methods are two popular approaches for inference in a multiple testing scenario (Efron, 2012). For testing against the family $H_0^{(n)}$, Bonferroni type inference uses the minimal P-value $p_{(1)}$ as the test statistics. The Benjamini-Hochberg (BH) FDR controlling procedure with parameter $q \in (0, 1)$ selects the smallest k^* P-values, where k^* is the largest integer k satisfying $p_{(k)} \le qk/n$ (Benjamini and Hochberg, 1995). A global test based on this procedure rejects $H_0^{(n)}$ at level α if at least one P-value is selected when $q = h(\alpha)$, for some critical value $h(\alpha) < 1$ designed to reject $H_0^{(n)}$ with probability at most α under $H_0^{(n)}$. Namely,

$$\text{Reject } H_0^{(n)} \text{ if and only if } \quad \min_{1 \le i \le n} \frac{p_{(i)}}{i/n} \le h(\alpha). \tag{2.15}$$

Note that since the BH procedure controls the family-wise error rate at level q under $H_0^{(n)}$, one can use $h(\alpha) = \alpha$, but our analysis is not restricted to this choice of $h(\alpha)$.

For a RMD model, both procedures turn out to be asymptotically powerful (respectively, powerless) within the exact same region. The phase transition curve distinguishing powerfulness from powerlessness is given by

$$\rho_{\mathsf{Bonf}}(\beta, \sigma) := \begin{cases} \left(1 - \sigma\sqrt{1-\beta}\right)^2, & 1/2 < \beta < 1, \quad \sigma^2 < 2, \\ \left(1 - \sigma\sqrt{1-\beta}\right)^2 & 1 - 1/\sigma^2 \le \beta < 1, \quad \sigma^2 > 2, \\ 0, & \beta < 1 - 1/\sigma^2, \quad \sigma^2 > 2. \end{cases} \tag{2.16}$$

Theorem 4. *Consider the hypothesis testing problem* (2.7) *under the RMD conditions* (2.5) *and* (2.8). $T_n^{\mathsf{Bonf}} = -\log(p_{(1)})$ *is asymptotically powerless whenever* $\rho < \rho_{\mathsf{Bonf}}(\beta,\sigma)$ *and asymptotically powerful whenever* $\rho > \rho_{\mathsf{Bonf}}(\beta,\sigma)$.

Theorem 5. *Consider the hypothesis testing problem* (2.7) *under the RMD conditions* (2.5) *and* (2.8). *A test based on* (2.15) *is asymptotically powerless whenever* $r < \rho_{\mathsf{Bonf}}(\beta,\sigma)$ *and asymptotically powerful whenever* $r > \rho_{\mathsf{Bonf}}(\beta,\sigma)$.

Theorems 4 and 5 imply that both Bonferroni and FDR type inference are asymptotically optimal for $\sigma < 2$ only when $\beta < 1/2$ or $(4-\sigma^2)/4 < \beta$. This situation is similar to the case of the Gaussian means model studied in Donoho and Jin (2004), implying that under small variances and moderate rarity the evidence for discriminating $H_0^{(n)}$ from $H_1^{(n)}$ are not amongst sets of the form $\{p_i, : p_i < qk/n,\, q \in (0,1),\, k = 1,\dots,n\}$. Asymptotically, in this case, optimal discrimination is achieved by considering P-values in the much wider range $\{p_i, : p_i < n^{-(1-\delta)}\}$ for some $\delta > 0$. This range is considered by HC and BJ, but not by FDR or Bonferroni.

2.8.2. Fisher's combination test

We conclude this section by noting that Fisher's combination test (1.3) is asymptotically powerless for all $\beta > 1/2$.

Theorem 6. *Consider the hypothesis testing problem* (2.7) *under* (2.5). F_n *of* (1.3) *is asymptotically powerless whenever* $\beta > 1/2$.

3. Examples of Rare Moderate Departures Models

We consider below various examples of rare and weak multiple testing settings that are carried under our RMD formulation. We summarize those in Table 1. In most cases, these settings were previously studied, however, without the RMD formulation and without deriving all properties stated in Theorems 1–6. We indicate these earlier studies at the end of each example.

3.1. Heteroscedastic normal mixture

Consider testing the presence of a rare location and variance departure in a Gaussian model as in

$$\begin{aligned} &H_0: \; X_i \sim \mathcal{N}(0,1), \qquad i = 1,\dots,n, \\ &H_1: \; X_i \sim (1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(\mu, s^2), \quad i = 1,\dots,n, \end{aligned} \tag{3.1}$$

with $s > 0$. The relation between the model (3.1) to (2.1) is via the test

$$p_i = \bar{\Phi}(X_i), \quad \bar{\Phi}(x) := \Pr(\mathcal{N}(0,1) > x), \quad i = 1,\dots,n. \tag{3.2}$$

Standard facts about Mills' ratio (Grimmett and Stirzaker, 2020) imply

$$-2\log\left(\bar{\Phi}(x)\right) \sim -2\log\left(\frac{\phi(x)}{|x|}\right) = x^2(1+o(1)), \tag{3.3}$$

as $x \to \infty$. Consequently, under H_1, the distribution of $-2\log(p_i)$ is of the form

$$(1-\epsilon)\mathrm{Exp}(2) + \epsilon\, Q_i(\mu, s),$$

where $Q_i(\mu, s)$ is a probability distribution obeying

$$Q_i(\mu, s) \stackrel{D}{=} (sZ+\mu)^2\{1+o_p(1)\}, \qquad Z \sim \mathcal{N}(0,1), \tag{3.4}$$

as $\mu \to \infty$. For $\mu = \mu_n(r) = \sqrt{2r\log(n)}$, the last evaluation implies that $Q_i(\mu, s)$ satisfies (2.5). Since each $Q_i(\mu, s)$ also has a density, the P-values of (3.2) correspond to the strong RMD model formulation with log-chisquared parameters $(\rho, \sigma) = (r, s)$.

Previous studies of the setting (3.1) were conducted by Cai, Jeng and Jin (2011), which derived the optimal phase transition curve $\rho^*(\beta, \sigma)$ and showed that it is attained by HC of the P-values (3.2). The homoscedastic case $s^2 = 1$ was initially studied by Ingster (1996), Jin (2003), and Donoho and Jin (2004).

3.2. Two-sample heteroscedastic normal mixture

A two-sample version of (3.1) takes the form:

$$\begin{aligned} &H_0: \ X_i, Y_i \sim \mathcal{N}(\nu_i, 1), \quad i = 1, \ldots, n, \\ &H_1: \ \begin{cases} X_i \sim \mathcal{N}(\nu_i, 1), \\ Y_i \sim (1-\epsilon)\mathcal{N}(\nu_i, 1) + \epsilon\mathcal{N}(\nu_i + \mu, s^2) \end{cases}, \quad i = 1, \ldots, n, \end{aligned} \tag{3.5}$$

where $\nu_1, \ldots, \nu_n$ is a sequence of *unknown* means. For this setting, consider the P-values

$$p_i := \bar{\Phi}\left(\frac{Y_i - X_i}{\sqrt{2}}\right). \tag{3.6}$$

Notice that, with $\tilde{Y}_i \sim \mathcal{N}(\nu_i + \mu, s^2)$ and $X_i \sim \mathcal{N}(\nu_i, 1)$, Mills' ratio (3.3) implies

$$-2\log\left\{\bar{\Phi}\left(\frac{\tilde{Y}_i - X_i}{\sqrt{2}}\right)\right\} \stackrel{D}{=} \left(\sqrt{\frac{1+s^2}{2}}Z + \frac{\mu}{\sqrt{2}}\right)^2\{1+o_p(1)\}, \quad Z \sim \mathcal{N}(0,1),$$

as $\mu \to \infty$. Therefore, under H_1, we have that the distribution of $-2\log(p_i)$ is of the form

$$(1-\epsilon)\mathrm{Exp}(2) + \epsilon Q_i(\mu, s), \tag{3.7}$$

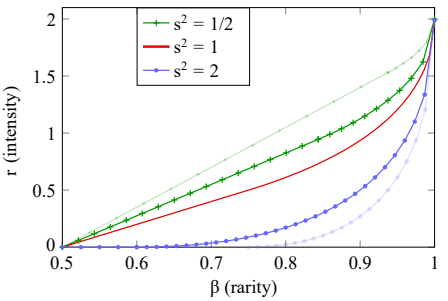

Figure 2. Two-Sample Phase Diagram. The phase transition curve $\rho^*_{\mathsf{two-sample}}(\beta, s)$ of (3.8) defines the detection boundary for an asymptotically log-chisquared perturbation model (2.1). For $r < \rho^*_{\mathsf{two-sample}}(\beta, s)$, all tests are powerless. For $r > \rho^*_{\mathsf{two-sample}}(\beta, s)$, the Higher Criticism and the Berk-Jones tests are asymptotically powerful. The faint lines correspond to $2\rho^*(\beta, s)$, where we have $\rho^*_{\mathsf{two-sample}}(\beta, 1) = 2\rho^*(\beta, 1)$.

where $Q_i(\mu, s)$ is a probability distribution obeying

$$Q_i(\mu, s) \stackrel{D}{=} \left(\sqrt{\frac{1+s^2}{2}}Z + \frac{\mu}{\sqrt{2}}\right)^2 \{1 + o_p(1)\}, \quad Z \sim \mathcal{N}(0,1),$$

as $\mu \to \infty$. It follows that with μ calibrated to n as in (2.3), $Q_i(\mu, s)$ satisfies (2.5) with mean parameter $\mu'_n(r) = \mu_n(r)/\sqrt{2} = \sqrt{r\log(n)}$ and scaling parameter $\sqrt{(1+s^2)/2}$, hence the P-values (3.6) corresponds to the strong RMD model formulation with log-chisquared parameters $\rho = r/2$ and $\sigma = \sqrt{(1+s^2)/2}$.

In order to derive a phase transition curve for this model, we start from (2.11), adjusting for the scaling factor 2 in the non-centrality parameter compared to (2.3), and substituting $\sqrt{(1+s^2)/2}$ for the standard deviation. We obtain:

$$\rho^*_{\mathsf{two-sample}}(\beta, s) := \begin{cases} (3-s^2)(\beta - 1/2) & 1/2 < \beta < (7-s^2)/8, \quad 0 < s^2 < 3, \\ 2[1 - \{(1+s^2)/2\}\sqrt{1-\beta}]^2 & (7-s^2)/8 \le \beta < 1, \quad 0 < s^2 < 3, \\ 0 & 1/2 < \beta < (s^2-1)/(s^2+1), \quad s^2 \ge 3, \\ 2[1 - \{(1+s^2)/2\}\sqrt{1-\beta}]^2 & (s^2-1)/(s^2+1) \le \beta < 1, \quad s^2 \ge 3. \end{cases} \tag{3.8}$$

Figure 2 depicts $\rho^*_{\mathsf{two-sample}}(\beta, s)$ for several values of s and compare it with $2\rho^*(\beta, s)$.

To the best of our knowledge, the curve $\rho^*_{\mathsf{two-sample}}(\beta, s)$ is new; the case $s = 1$ was considered in Donoho and Kipnis (2022).

3.3. Poisson means

Consider the hypothesis testing problem

$$\begin{aligned} H_0 &: \ X_i \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda_i), \quad i = 1, \ldots, n, \\ H_1 &: \ X_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon)\text{Pois}(\lambda_i) + \epsilon \text{Pois}(\lambda_i'), \quad i = 1, \ldots, n, \end{aligned} \tag{3.9}$$

where $\lambda_1, \ldots, \lambda_n$ is a sequence of *known* means and where each λ_i' is obtained by perturbing λ_i upwards. For this model, we have the P-values

$$p_i = \bar{\mathsf{P}}(X_i; \lambda_i), \quad i = 1, \ldots, n, \tag{3.10}$$

where $\bar{\mathsf{P}}(x; \lambda_i) := \Pr[\text{Pois}(\lambda_i) \geq x]$. We suppose that the Poisson rates increase with n such that

$$\frac{\min \lambda_i}{\log(n)} \to \infty, \tag{3.11}$$

and the perturbed means are given by

$$\lambda_i' = \lambda_i + \mu_n(r)\sqrt{\lambda_i}, \qquad i = 1, \ldots, n. \tag{3.12}$$

Noting that $\log(n)/\lambda_i' \to 0$ and $\lambda_i' - \lambda_i \to \infty$, the behavior of p_i under $H_1^{(n)}$ is obtained using a moderate deviation estimate of the RVs $\Upsilon_{\lambda_i'} \sim \text{Pois}(\lambda_i')$. This is provided by the following proposition.

Proposition 1. *Suppose that* λ_i *and* λ_i' *satisfy* (3.11) *and* (3.12). *Let* $X_i \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda_i')$ *and* $S_i = -2\log \bar{\mathsf{P}}(X_i; \lambda_i)$. *For every* $q > \rho \geq 0$,

$$\lim_{n\to\infty} \max_{1 \leq i \leq n} \left| \frac{-\log \Pr[S_i \geq 2q\log(n)]}{\log(n)} - \alpha(q; \rho, 1) \right|. \tag{3.13}$$

We conclude that under H_1, (3.10)–(3.12),

$$p_i \sim (1-\epsilon)E_i^{(n)} + \epsilon Q_i^{(n)}, \quad i = 1, \ldots, n$$

where $\{Q_i^{(n)}\}_{i=1}^n$ obey (2.5) and $\{E_i^{(n)}\}_{i=1}^n$ obey (2.8). Consequently, the model of (3.9) with P-values (3.10) is RMD with log-chisquared parameters $\rho = r$ and $\sigma = 1$.

Arias-Castro and Wang (2015) studied the Poisson Rates model (3.9). They derived the optimal phase transition $\rho^*(\beta, 1)$, the Bonferroni phase transition $\rho_{\mathsf{Bonf}}(\beta, 1)$, and showed that a version of HC is asymptotically powerful whenever $r > \rho^*(\beta, 1)$.

3.4. Two-sample poisson means

A two-sample version of (3.9) is given as:

$$H_0:\ X_i, Y_i \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda_i), \quad i = 1, \dots, n.$$
$$H_1:\ \begin{cases} X_i \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda_i) \\ Y_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon)\text{Pois}(\lambda_i) + \epsilon \text{Pois}(\lambda_i') \end{cases}, \quad i = 1, \dots, n. \tag{3.14}$$

Here $\lambda_1, \dots, \lambda_n$ is a sequence of *unknown* Poisson rates that satisfy (3.11), while $\lambda_1', \dots, \lambda_n'$ are defined as in (3.12). We summarize the significance of the pair (X_i, Y_i) associated with the i-th coordinate by the RV:

$$p_i := \bar{\Phi}\left(\sqrt{2Y_i} - \sqrt{2X_i}\right). \tag{3.15}$$

In order to analyze the behavior of $p_1, \dots, p_n$ under $H_0^{(n)}$ and $H_1^{(n)}$, note that the transformed Poisson RV $\sqrt{X}$, $X \sim \text{Pois}(\lambda)$, is variance stable:

$$2\sqrt{X} - 2\sqrt{\lambda} \to \mathcal{N}(0,1).$$

Under H_1, (3.11) and (3.12) imply

$$\sqrt{\lambda_i'}(1 + o(1)) = \sqrt{\lambda_i} + \frac{\mu_n(r)}{2}, \tag{3.16}$$

where $o(1) \to 0$ as $n \to \infty$ uniformly in i. Consequently, with $\Upsilon_{\lambda_i'} \sim \text{Pois}(\lambda_i')$,

$$\begin{aligned} \sqrt{2\Upsilon_{\lambda_i'}} - \sqrt{2X_i} &= \sqrt{2\Upsilon_{\lambda_i'}} - \sqrt{2\lambda_i'} - \left(\sqrt{2X_i} - \sqrt{2\lambda_i}\right) + \left(\sqrt{2\lambda_i'} - \sqrt{2\lambda_i}\right) \\ &\overset{D}{=} \left(Z + \frac{\mu_n(r)}{\sqrt{2}}\right)\{1 + o_p(1)\}, \quad Z \sim \mathcal{N}(0,1), \end{aligned}$$

as $n \to \infty$. By setting

$$\pi_i := \bar{\Phi}\left(\sqrt{2\Upsilon_{\lambda_i'}} - \sqrt{2X_i}\right),$$

combining Mill's ratio (3.3) and (3.16), we obtain

$$\begin{aligned} -2\log(\pi_i) &\overset{D}{=} \left(Z + \frac{\mu_n(r)}{\sqrt{2}}\right)^2 \{1 + o_p(1)\} \\ &= \left(Z + \sqrt{r\log(n)}\ \right)^2 \{1 + o_p(1)\}. \end{aligned} \tag{3.17}$$

The last evaluation suggests that (2.6) holds with log-chisquared parameters $(\rho, \sigma) = (r/2, 1)$. The full argument follows from the proposition below.

Proposition 2. *Suppose that* $\lambda_1, \dots, \lambda_n$ *satisfy*

$$\min_{i=1,\dots,n} \frac{\lambda_i}{\log(n)} \to \infty \quad \textit{as} \quad n \to \infty.$$

Set $\lambda_i'(r) := \lambda_i + \sqrt{r\lambda_i \log(n)}$ *for some* $r \geq 0$. *Assume that* $X \sim \text{Pois}(\lambda_i)$,

$Y \sim \text{Pois}(\lambda_i')$, *and let*

$$\pi_i := \bar{\Phi}\left(\sqrt{2Y_i} - \sqrt{2X_i}\right), \quad i = 1, \ldots, n.$$

Then

$$\lim_{n\to\infty} \max_{1\le i\le n} \left| \frac{-\log \Pr\left[-2\log(\pi_i) \ge 2q\log(n)\right]}{\log(n)} - \alpha\left(q; \frac{r}{2}, 1\right)\right| = 0.$$

Donoho and Kipnis (2022) studied a two-sided perturbation model similar to (3.14) and proposed to use P-values of an exact binomial test as in

$$p_i' := \Pr\left[\left|\text{Bin}\left(X_i + Y_i, \frac{1}{2}\right) - \frac{X_i + Y_i}{2}\right| \le \left|\frac{X_i - Y_i}{2}\right|\right], \tag{3.18}$$

which have several advantages over (3.15) in practice. Our RMD formulation shows that the asymptotic properties of the tests in Section 2 based on either collection of P-values under (3.5) are identical, e.g., the optimal phase transition is given by $\rho^*_{\mathsf{two-sample}}(\beta, 1)$.

3.5. Perturbed binomial experiments

Suppose that our data consists of n independent samples from a binomial distribution:

$$X_i \overset{\text{i.i.d.}}{\sim} \text{Bin}(m_i, q_i), \quad m_i \in \mathbb{N}, \quad i = 1, \ldots, n. \tag{3.19}$$

We are interested in testing the null hypothesis $H_0 : q_1 = \ldots = q_n = 1/2$ against an alternative in which we have $q_i > 1/2$ for a small fraction of the indices. It is natural to use P-values from the exact binomial test

$$p_i := p_{\mathsf{Bin}}(X_i) := \Pr\left[\text{Bin}\left(m_i, \frac{1}{2}\right) \ge X_i\right], \tag{3.20}$$

although other options are available, e.g. testing for overdispersion (Dean, 1992). The alternative hypothesis is specified as

$$H_1^{(n)} : X_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon)\text{Bin}\left(m_i, \frac{1}{2}\right) + \epsilon\text{Bin}\left(m_i, \frac{1}{2} + \delta\right), \quad i = 1, \ldots, n,$$

for small δ and ϵ.

For $X_i \sim \text{Bin}(m_i, q_i)$, the normal approximation

$$X_i \approx \mathcal{N}\left(m_i q_i, m_i q_i (1 - q_i)\right).$$

suggests that deviations on the moderate scale arise by the calibration

$$m_i = \frac{2\log(n)}{s}\{1 + o(1)\} \quad \text{and} \quad \delta = \sqrt{\frac{s \cdot r}{4}}, \tag{3.21}$$

where $s > 0$ and $r \geq 0$ are parameters satisfying $r \cdot s < 1$. The proposition below implies that under such calibration and with $\epsilon = n^{-\beta}$ for $\beta \in (1/2, 1)$, (3.19) with the P-values (3.20) corresponds to RMD model with $\rho = r$ and $\sigma^2 = 1 - sr$.

Proposition 3. *Consider $X \sim \mathrm{Bin}(m, 1/2 + \delta)$ with δ and m calibrated to n as in (3.21). Then for all $q \geq r$,*

$$\lim_{n\to\infty} \max_{1\leq i\leq n} \left| \frac{-\log \Pr\left[-2\log\left\{p_{\mathsf{Bin}}(X)\right\}\right] \geq 2q\log(n)}{\log(n)} - \alpha\left(q; r, \sqrt{1 - s\cdot r}\right)\right| = 0. \tag{3.22}$$

The optimal phase transition of the RMD model corresponding to (3.19) under (3.20) is defined by triplets (s, r, β) satisfying

$$r = \begin{cases} (1 + s\cdot r)\,(\beta - 1/2) & 1/2 < \beta < (3 + s\cdot r)/4, \\ \{1 - \sqrt{(1 - s\cdot r)(1-\beta)}\}^2 & (3 + s\cdot r)/4 \leq \beta < 1. \end{cases}$$

When $s < 1/(\beta - 1/2)$, the last relation defines the curve

$$\rho^*_{\mathsf{Bin}}(\beta, s) := \tag{3.23}$$
$$\begin{cases} \dfrac{\beta - 1/2}{1 - s(\beta - 1/2)} & 1/2 < \beta < \left(3 + \dfrac{1 - \sqrt{1-s}}{1 + \sqrt{1-s}}\right)/4, \quad s \leq 1, \\ \left(\dfrac{1 - \sqrt{(1-\beta)(1-s\beta)}}{1 + s(1-\beta)}\right)^2 & \left(3 + \dfrac{1 - \sqrt{1-s}}{1 + \sqrt{1-s}}\right)/4 \leq \beta < 1, \quad s \leq 1, \\ \dfrac{\beta - 1/2}{1 - s(\beta - 1/2)} & 1/2 < \beta < 1/2 + 1/s, \quad s \geq 1, \\ \infty & 1/2 + 1/s \leq \beta, \quad s \geq 1; \end{cases}$$

see an illustration in Figure 3. Consequently, for $r < \rho^*_{\mathsf{Bin}}(\beta, s)$ or $s > 1/(\beta - 1/2)$, all tests are asymptotically powerless while some tests are asymptotically powerful when $s \leq 1$ and $r > \rho^*_{\mathsf{Bin}}(\beta, s)$.

To the best of our knowledge, the curve $\rho^*_{\mathsf{Bin}}(\beta, s)$ is new. Mukherjee, Pillai and Lin (2015) studied the model (3.19) in the context of sparse binary regression under a coarser calibration that corresponds to the limit $s \to 0$ in (3.21). In this case, $\rho^*_{\mathsf{Bin}}(\beta, s)$ converges to $\rho^*(\beta, 1)$, in accordance with the results of Mukherjee, Pillai and Lin (2015).

4. Log-Chisquared Versus Log-Normal

4.1. The log-normal approximation

The log-normal approximation of a P-value under the alternative hypothesis is a tool developed by Bahadur to study the interplay among the test's size, power, and the "cost" of attaining new data which is most commonly associated with the

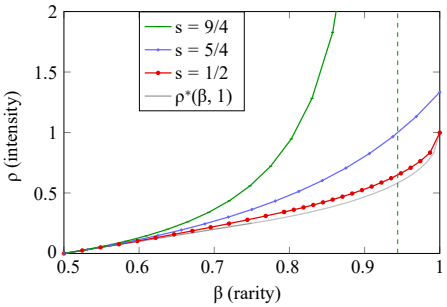

Figure 3. Phase transitions of multiple binomial experiments with perturbed success probability (3.19). The phase transition curve $\rho^*_{\mathsf{Bin}}(\beta, s)$ of (3.23) defines the detection boundary in the multiple binomials model of (3.19) under the calibration (3.21). The parameter s controls the number of experiments in individual binomial trials according to (3.21) (larger s means fewer trials). $\rho^*_{\mathsf{Bin}}(\beta, s = 9/4)$ asymptotes to the dashed line. The case $s \to 0$ corresponds to the homoscedastic case analyzed in Mukherjee, Pillai and Lin (2015).

sample size (Bahadur, 1960; Gleser, 1964; Lambert and Hall, 1982). Informally, suppose that the alternative hypothesis is characterized by a parameter θ and that a_n is a sequence tending to infinity with n describing the cost of sampling from the population of interest. Bahadur's log-normal approximation says that, under some conditions including asymptotic normality of the test statistic, a P-value π under the alternative $H_1 = H_1(\theta, a_n)$ obeys

$$\frac{\log(\pi) + a_n c(\theta)}{\sqrt{a_n}} \xrightarrow{D} \mathcal{N}\left(0, \tau^2(\theta)\right), \tag{4.1}$$

as $n \to \infty$. In the terminology of Lambert and Hall (1982), $c(\theta)$ is Bahadur's half-slope describing the asymptotic behavior of the test's size, i.e., the rate at which π goes to zero. The test's power is determined both by $\tau(\theta)$ and $c(\theta)$. It is convenient to write (4.1) as

$$-2\log(\pi) \overset{D}{\approx} \mathcal{N}\left(a_n c(\theta), a_n \tau^2(\theta)\right). \tag{4.2}$$

In the sections below, we compare our log-chisquared approximation to the log-normal approximation of (4.2) for P-values under the alternative hypothesis.

4.2. Formal comparison

It is well-recognized that (4.2) is a *large deviation* estimate of the test statistic in the sense that $c(\theta)$ is a transformation of the statistic's rate function whenever this statistic satisfies a large deviation principle (Sievers, 1969; Gleser, 1984; Singh, 1980). In contrast, in all RMD models of Section 3 the alternative hypothesis corresponds to a *moderate deviation* of each test statistic from its null

(Rubin and Sethuraman, 1965). Consequently, the log-normal approximation of (4.2) cannot correctly indicate the asymptotic power of tests under the RMD formulation.

We formally show this last point in the homoscedastic RMD normal mixture model:

$$\begin{aligned} H_0 &: X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad i = 1, \dots, n, \\ H_1 &: X_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(\mu,1), \quad i = 1, \dots, n, \end{aligned} \tag{4.3}$$

with the P-values $p_i = \bar{\Phi}(X_i)$. This is the model (3.1) with $s = 1$. Under H_1,

$$-2\log p_i \sim (1-\epsilon)\text{Exp}(2) + \epsilon Q_i, \tag{4.4}$$

where the probability distribution Q_i is the subject of our approximation. When μ and ϵ are calibrated to n as in (2.3) and (2.2), respectively, we propose in this paper the log-chisquared approximation

$$Q_i = Q_i^{(n)} \overset{D}{\approx} (\mu_n(r) + Z)^2 . \tag{4.5}$$

On the other hand, we have

$$(\mu + Z)^2 = (\mu^2 + 2\mu Z)\{1 + o_p(1)\}, \quad \mu \to \infty,$$

implying the log-normal approximation:

$$Q_i^{(n)} \overset{D}{\approx} \mathcal{N}(\mu_n^2(r), 4\mu_n^2(r)) = \mathcal{N}\left(2r\log(n), 8r\log(n)\right). \tag{4.6}$$

In particular, $\theta = r$, $a_n = \log(n)$, $c(\theta) = 2r$, and $\tau^2(\theta) = 8r$ in the notation of (4.2). We now compare approximations (4.5) and (4.6). Observe that the success of HC and the BJ tests follows from the behavior of

$$\Pr(\pi_i < n^{-q}), \quad -2\log(\pi_i) \sim Q_i^{(n)}$$

for $r < q < 1$ as $n \to \infty$; see the proofs of Theorems 2 and 3 in the Supplementary Materials and an extended analysis in Donoho and Kipnis (2024). With $Q_i^{(n)}$ as in (4.5),

$$\begin{aligned} \Pr(\pi_i < n^{-q}) &= \Pr(-2\log(\pi_i) > 2q\log(n)) \\ &\sim \Pr\left(\{\mu_n(r) + Z\}^2 \geq 2q\log(n)\right) \\ &\sim \Pr\left(Z \geq \sqrt{\log(n)}(\sqrt{2q} - \sqrt{2r})\right). \end{aligned} \tag{4.7}$$

A standard evaluation of the behavior of HC under $H_1^{(n)}$ uses (4.7) to show that it is asymptotically powerful for $r > \rho(\beta, 1)$ (Donoho and Jin, 2004). On the

other hand, with $Q_i^{(n)}$ as in (4.6),

$$\begin{aligned}\Pr(\pi_i < n^{-q}) &= \Pr(-2\log(\pi_i) > 2q\log(n)) \\ &\sim \Pr\left(\mu_n(r)^2 + 2\mu_n(r)Z \geq 2q\log(n)\right) \\ &= \Pr\left(Z \geq \sqrt{\log(n)}\frac{q-r}{\sqrt{2r}}\right). \end{aligned} \tag{4.8}$$

Since

$$\frac{q-r}{\sqrt{2r}} \geq \sqrt{2q} - \sqrt{2r}, \quad q \geq r > 0,$$

using the log-normal approximation in a formal exercise of would-be power analysis of HC by replacing (4.7) with (4.8), incorrectly predicts that HC is powerless for some $r > \rho(\beta, 1)$. Specifically, the log-normal approximation incorrectly predicts the phase transition curve:

$$\rho^\dagger(\beta, 1) = \begin{cases} (2/3)\,(2\beta - 1) & 7/8 < \beta, \\ 3 - 2\beta - 2\sqrt{(2-\beta)(1-\beta)} & 1/2 < \beta \leq 7/8, \end{cases}$$

which satisfies $\rho^\dagger(\beta, 1) > \rho(\beta, 1)$ for $\beta \in (1/2, 1)$.

4.3. Empirical comparison

We now provide an empirical comparison between the log-normal and the log-chisquared approximation under moderate departures using Monte Carlo simulations involving fixed effects (i.e., not rare). In each simulation, we sample data $x_1, \ldots, x_n$ independently from $\mathcal{N}(\mu_n(r), 1)$ and consider $\pi_i = \bar{\Phi}(x_i)$ as a P-value under $H_0 : X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $i = 1, \ldots, n$. Of course, in this model, we can characterize the distribution of π_i analytically, as well as the deviation of this distribution from the log-chisquared and the log-normal distributions, respectively. The purpose of the simulation is to illustrate the better fit of the log-chisquared approximation. Figure 4 illustrates the results of one simulation with $n = 1000$ (top panels) and one simulation with $n = 100000$ (bottom panels), while the departure intensity parameter $r = 1$ is fixed in both cases. The panels on the left show the histogram of $\{-2\log(\pi_i)\}_{i=1}^n$ with the density of the normal distribution $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ and the density of the noncentral chisquared distribution $\chi_2^2(\hat{\lambda})$, where $\hat{\mu}$ and $\hat{\sigma}^2$ are the standard mean and variance estimates and $\hat{\lambda} = \hat{\mu} - 2$ is the non-centrality estimate. The middle and right panels illustrate QQ-plots of the empirical distribution of $\{-2\log(\pi_i)\}_{i=1}^n$ against $\chi_2^2(\hat{\lambda})$ and $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$, respectively, showing the better fit of the empirical distribution of $\{-2\log(\pi_i)\}_{i=1}^n$ to $\chi_2^2(\hat{\lambda})$.

Figure 5 illustrates the results of 1,000 Monte Carlo simulations with many configurations of n and r. For each configuration, we conducted an Anderson-Darling (AD) goodness-of-fit test of $\{-2\log(p_i)\}_{i=1}^n$ against $\chi_2^2(\hat{\lambda})$ and $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$.

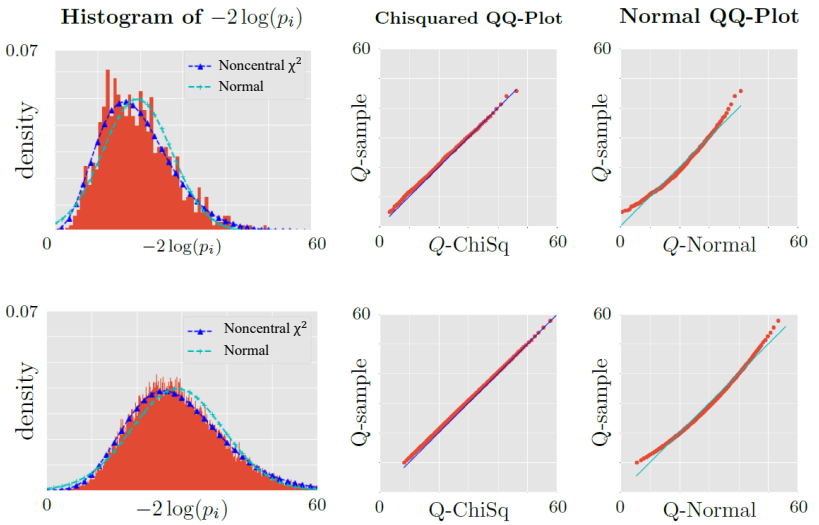

Figure 4. Comparing log-normal and log-chisquared approximations to moderately perturbed P-values $\pi_1, \ldots, \pi_n$. Here $\pi_i \sim \bar{\Phi}(X_i)$, $X_i = \mathcal{N}(\sqrt{2r\log(n)}, 1)$, with $n = 10^3$ (top) and $n = 10^5$ (bottom). Left: histogram of $\{-2\log(\pi_i)\}_{i=1}^n$. Middle: QQ-plots of the empirical distribution of $\{-2\log(\pi_i)\}_{i=1}^n$ against the noncentral chisquared distribution. Right: QQ-plots of the empirical distribution of $\{-2\log(\pi_i)\}_{i=1}^n$ against the normal distribution.

The test against the normal (respectively, chisquared) rejects when the AD statistic exceeds its simulated 0.95-th quantile under the null obtained by sampling 10,000 times from the normal (chisquared) distribution. It follows from Figure 5 that the log-chisquared distribution fits the distribution of the P-values under fixed moderate effects much better than the log-normal distribution. The lack of fit of the log-chisquared distribution is only visible when the sample size is very large or when the signal is very weak. The analysis in Section 2 implies that this lack of fit is insignificant when the effects are also rare in the sense that the log-chisquared approximation provides the information-theoretic limit of signal detection under RMD.

5. Additional Discussion

5.1. Heteroscedasticity in RMD models

The phase transition described by $\rho(\beta, \sigma)$ can be seen as the result of two phenomena: (i) location shift controlled by ρ, and (ii) heteroscedasticity controlled by σ^2. Roughly speaking, increasing the effect of either (i) or (ii) eases detection and reduces the phase transition curve, as seen in Figure 1. We refer to Cai, Jeng and Jin (2011) for a more comprehensive discussion on the effect of (ii) on the phase transition curves. With obvious changes, this discussion is also relevant to the curves $\rho_{\mathsf{two-sample}}(\beta; \sigma)$ of (3.8) and $\rho^*_{\mathsf{Bin}}(\beta, s)$ of (3.23) with $s > 0$.

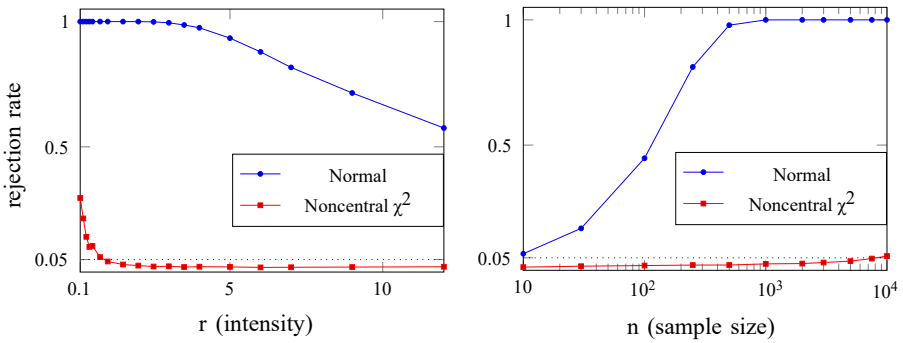

Figure 5. Comparing the fit of the empirical distribution of moderately perturbed P-values to the normal and noncentral chisquared distributions. Both panels show the rejection rate of the Anderson-Darling (AD) Goodness-of-fit test at a significance level of 0.05 (smaller rejection rate indicates a better fit). Left: rejection rate versus perturbation intensity parameter r; $n = 1000$ is fixed. Right: rejection rate versus sample size n; $r = 2$ is fixed.

Comparing the effect of heteroscedasticity in one- versus two-sample setting, we see that

$$\rho_{\mathsf{two-sample}}(\beta, 1) = 2\rho(\beta, 1),$$

an observation first made in Donoho and Kipnis (2022). Interestingly, as shown in Figure 2, this relation between $\rho_{\mathsf{two-sample}}(\beta, s)$ and $\rho(\beta, s)$ does not hold when $s \neq 1$. Specifically, detection in the two-sample homeostatic setting ($s = 1$) asymptotically requires twice the effect size. On the other hand, compared to the one-sample case, more than twice the effect size is needed for overdispersed mixtures ($s > 1$) and less for underdispersed ones ($s < 1$).

5.2. Other generalizations of rare and weak models

Cai and Wu (2014) considered general rare and weak signal detection models characterized by the asymptotic behavior of the likelihood ratio between the mixture components on the moderate deviation scale which is similar to (2.9). For rare and weak departures from the exponential distribution, the information-theoretic lower bound of Theorem 1 generalizes (Cai and Wu, 2014, Thm. 3) by allowing for non-identically distributed coordinates and by providing conditions that involve integrated versions of the likelihood ratio.

Another generalization of rare and weak signal detection models is provided by Arias-Castro and Wang (2017), which considered a symmetric null distribution and proposed non-parametric HC- and Bonferroni-type tests that possess interesting optimality properties. Our RMD formulation applies to the setting of Arias-Castro and Wang (2017) when the non-symmetric behavior of an individual test statistic under the alternative hypothesis is on the moderate deviation scale.

For the HC test, Donoho and Kipnis (2024) considered rare mixtures of P-values with non-null component $Q_i^{(n)}$ obeying

$$\max_i \Pr_{p_i \sim Q_i^{(n)}} [p_i > 2q\log(n)] = \max_i \mathbb{E}_{X\sim \mathrm{Exp}(2)}[L_i(X)\mathbf{1}_{\{X>2q\log(n)\}}] \\ = n^{-\alpha'(q;\rho)+o(1)},$$

for some continuous, non-negative bivariate function $\alpha'(q;\rho)$ that is increasing in q and decreasing in ρ. They showed that HC of such P-values is powerless in the region

$$\Xi_{\mathrm{HC}} \equiv \left\{(\rho,\beta) \,:\, \max_{q\in[0,1]} \left\{\frac{1+q}{2} - \alpha'(q;\rho)\right\} < \beta\right\}.$$

The region Ξ_{HC} coincide with $\{(\rho,\beta) \,:\, \rho < \rho(\beta,\sigma)\}$ in the RMD setting for which we have $\alpha'(q;\rho) = \max\{\alpha(q;\rho,\sigma),0\}$.

Examples of rare and weak multiple testing settings with non-moderate departures include the sparse positive dependence model of Arias-Castro, Huang and Verzelen (2020), rare mixtures involving distributions of polynomial tails with a location shift of order $\mu_n(r)$ as the alternative studied in Arias-Castro and Ying (2019), and the two-sample Poisson means in the low-counts case of Donoho and Kipnis (2022).

Supplementary Material

The Supplementary Material contains proofs of all Theorems and Propositions.

Acknowledgments

The author would like to thank David Donoho for discussions and comments on an early version of this manuscript and two anonymous reviewers who provided valuable suggestions that have improved the manuscript. Parts of this article were presented at the 2022 IEEE International Symposium on Information Theory (ISIT) (Kipnis, 2022). Parts of the work were done while the author was with the Department of Statistics at Stanford University.

References

Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics* **34**, 584–653.

Acharya, J., Das, H., Jafarpour, A., Orlitsky, A., Pan, S. and Suresh, A. (2012). Competitive classification and closeness testing. In *JMLR: Workshop and Conference Proceedings* **23**, 22.1–22.18.

Arias-Castro, E., Candès, E. J. and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics* **39**, 2533–2556.

Arias-Castro, E., Huang, R. and Verzelen, N. (2020). Detection of sparse positive dependence. *Electronic Journal of Statistics* **14**, 702–730.

Arias-Castro, E. and Wang, M. (2015). The sparse Poisson means model. *Electronic Journal of Statistics* **9**, 2170–2201.

Arias-Castro, E. and Wang, M. (2017). Distribution-free tests for sparse heterogeneous mixtures. *Test* **26**, 71–94.

Arias-Castro, E. and Ying, A. (2019). Detection of sparse mixtures: Higher criticism and scan statistic. *Electronic Journal of Statistics* **13**, 208–230.

Bahadur, R. R. (1960). Stochastic comparison of tests. *The Annals of Mathematical Statistics* **31**, 276–295.

Balakrishnan, S. and Wasserman, L. (2018). Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics* **12**, 727–749.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7**, 781–791.

Bayer, A. E. and Seljak, U. (2020). The look-elsewhere effect from a unified Bayesian and frequentist perspective. *Journal of Cosmology and Astroparticle Physics* **2020**, 009.

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)* **57**, 289–300.

Berk, R. H. and Jones, D. H. (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **47**, 47–59.

Boos, D. D. and Stefanski, L. A. (2011). P-value precision and reproducibility. *The American Statistician* **65**, 213–221.

Cai, T. T., Jeng, X. J. and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **73**, 629–662.

Cai, T. T. and Wu, Y. (2014). Optimal detection of sparse mixtures against a given null distribution. *IEEE Transactions on Information Theory* **60**, 2217–2232.

Dean, C. B. (1992). Testing for overdispersion in poisson and binomial regression models. *Journal of the American Statistical Association* **87**, 451–457.

Delaigle, A. and Hall, P. (2009). Higher criticism in the context of unknown distribution, non-independence, and classification. In *Perspectives in Mathematical Sciences I: Probability and Statistics*, 109–138. World Scientific.

Delaigle, A., Hall, P. and Jin, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on student's t-statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **73**, 283–301.

Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*. Springer-Verlag, New York.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32**, 962–994.

Donoho, D. and Jin, J. (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**, 4449–4470.

Donoho, D. and Jin, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science* **30**, 1–25.

Donoho, D. L. and Kipnis, A. (2022). Higher criticism to compare two large frequency tables, with sensitivity to possible rare and weak differences. *The Annals of Statistics* **50**, 1447–1472.

Donoho, D. L. and Kipnis, A. (2024). The impossibility region for detecting sparse mixtures using the higher criticism. *The Annals of Applied Probability* **34**, 4921–4939.

Efron, B. (2012). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

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**, 1151–1160.

Fisher, R. A. (1924). A method of scoring coincidences in tests with playing cards. In *Proceedings of the Society for Psychical Research* **34**, 181–185.

Fisher, R. A. (1929). Tests of significance in harmonic analysis. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **125**, 54–59.

Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in Statistics*, 66–70. Springer.

Galili, B., Kipnis, A. and Yakhini, Z. (2023). Detecting rare and weak deviations of non-proportional hazard in survival analysis. *arXiv:2310.00554*.

Gibson, E. W. (2021). The role of p-values in judging the strength of evidence and realistic replication expectations. *Statistics in Biopharmaceutical Research* **13**, 6–18.

Gleser, L. J. (1964). On a measure of test efficiency proposed by RR bahadur. *The Annals of Mathematical Statistics* **35**, 1537–1544.

Gleser, L. J. (1984). Large deviation indices and Bahadur exact slopes. *Statistics & Decisions* **1**, 193–204.

Grimmett, G. and Stirzaker, D. (2020). *Probability and Random Processes*. 2nd Edition. Oxford University Press, Oxford.

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M. L. et al. (2009). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nature genetics* **41**, 1088–1093.

Ingster, J. (1996). On some problems of hypothesis testing leading to infinitely divisible distributions. *Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS) Preprint*, No. 215. DOI: 10.20347/WIAS.PREPRINT.215.

Ingster, Y. and Suslina, I. A. (2012). *Nonparametric Goodness-of-Fit Testing under Gaussian Models*. Springer Science & Business Media.

Ingster, Y. I., Pouet, C. and Tsybakov, A. B. (2009). Classification of sparse high-dimensional vectors. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**, 4427–4448.

Ingster, Y. I., Tsybakov, A. B. and Verzelen, N. (2010). Detection boundary in sparse regression. *Electronic Journal of Statistics* **4**, 1476–1526.

Jin, J. (2003). *Detecting and Estimating Sparse Mixtures*. Ph.D. thesis. Stanford University.

Jin, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proceedings of the National Academy of Sciences* **106**, 8859–8864.

Jin, J. and Ke, Z. T. (2016). Rare and weak effects in large-scale inference: Methods and phase diagrams. *Statistica Sinica* **26**, 1–34.

Kipnis, A. (2022). Rare and weak detection models under moderate deviations analysis and log-chisquared p-values. In *2022 IEEE International Symposium on Information Theory (ISIT)*, 2052–2057.

Lambert, D. (1981). Influence functions for testing. *Journal of the American Statistical Association* **76**, 649–657.

Lambert, D. and Hall, W. (1982). Asymptotic lognormality of p-values. *The Annals of Statistics* **10**, 44–64.

Li, W. (2012). Volcano plots in analyzing differential expressions with mrna microarrays. *Journal of Bioinformatics and Computational Biology* **10**, 1231003.

Moscovich, A., Nadler, B. and Spiegelman, C. (2016). On the exact berk-jones statistics and their p-value calculation. *Electronic Journal of Statistics* **10**, 2329–2354.

Mukherjee, R., Pillai, N. S. and Lin, X. (2015). Hypothesis testing for high-dimensional sparse binary regression. *Annals of Statistics* **43**, 352–381.

Pearson, T. A. and Manolio, T. A. (2008). How to interpret a genome-wide association study. *Jama* **299**, 1335–1344.

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**, 1236–1242.

Price, A. L., Zaitlen, N. A., Reich, D. and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459–463.

Quaino, P., Juarez, F., Santos, E. and Schmickler, W. (2014). Volcano plots in hydrogen electrocatalysis – uses and abuses. *Beilstein Journal of Nanotechnology* **5**, 846–854.

Rubin, H. and Sethuraman, J. (1965). Probabilities of moderate deviations. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* **27**, 325–346.

Sackrowitz, H. and Samuel-Cahn, E. (1999). P-values as random variables—expected p-values. *The American Statistician* **53**, 326–331.

Sievers, G. L. (1969). On the probability of large deviations and exact slopes. *The Annals of Mathematical Statistics* **40**, 1908–1921.

Singh, K. (1980). Large deviations and exact Bahadur's slope of L-statistics. Technical report. Department of Statistics, Stanford University.

Tandra, R. and Sahai, A. (2008). SNR walls for signal detection. *IEEE Journal of Selected Topics in Signal Processing* **2**, 4–17.

Westfall, P. H. and Wolfinger, R. D. (1997). Multiple tests with discrete distributions. *The American Statistician* **51**, 3–8.

(Received April 2023; accepted July 2023)