Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects

David Donoho and Jiashun Jin

Stanford University and Carnegie Mellon University

Abstract

In modern high-throughput data analysis, researchers perform a large number of statistical tests, expecting to find perhaps a small fraction of significant effects against a predominantly null background. Higher Criticism (HC) was introduced to determine whether there are *any* non-zero effects; more recently, it was applied to feature selection, where it provides a method for selecting useful predictive features from a large body of potentially useful features, among which only a rare few will prove truly useful.

In this article, we review the basics of HC in both the testing and feature selection settings. HC is a flexible idea, which adapts easily to new situations; we point out how it adapts to clique detection and bivariate outlier detection. HC, although still early in its development, is seeing increasing interest from practitioners; we illustrate this with worked examples. HC is computationally effective, which gives it a nice leverage in the increasingly more relevant "Big Data" settings we see today.

We also review the underlying theoretical "ideology" behind HC. The *Rare/Weak* (RW) model is a theoretical framework simultaneously controlling the size and prevalence of useful/significant items among the useless/null bulk. The RW model shows that HC has important advantages over better known procedures such as False Discovery Rate (FDR) control and Family-wise Error control (FwER), in particular, certain optimality properties. We discuss the rare/weak *phase diagram*, a way to visualize clearly the class of RW settings where the true signals are so rare or so weak that detection and feature selection are simply impossible, and a way to understand the known optimality properties of HC.

Dedications. To the memory of John W. Tukey 1915–2000 and of Yuri I. Ingster 1946–2012, two pioneers in mathematical statistics.

Key words. Classification; control of FDR; feature selection; Higher Criticism; large covariance matrix; Large-Scale Inference; Rare and Weak effects; phase diagram; sparse signal detection.

AMS 2010 subject classification. 62G10, 62H30, 62G32.

1 Introduction

A data deluge is now flooding scientific and technical work [3]. In field after field, high-throughput devices gather many measurements per individual; depending on the field these could be gene expression levels, or spectrum levels, or peak detectors or wavelet transform coefficients; there could be thousands or even millions of different feature measurements per single subject.

High-throughput measurement technology automatically measures systematically generated features and contrasts; these features are not custom-designed for any one project. Only a small proportion of the measured features are expected to be relevant for the research in question; but researchers don't know in advance which those will be; they instead measure every contrast fitting within their systematic scheme, intending later to identify a small fraction of relevant ones post-facto.

This flood of high throughput measurements is driving a new branch of statistical practice: what Efron [44] calls *Large-Scale Inference* (LSI). For this paper, two specific LSI problems are of interest:

- *Massive multiple testing for sparse intergroup differences.* Here we have two groups, a treatment and a control, and for each measured variable, we test whether the two groups are different on that measurement, obtaining, say a P-value per feature. Of course many individual features are unrelated to the specific intervention being studied, and those would be expected to show no significant differences – but we don't know which these are. We expect that even when there are true inter-group differences, only a small fraction of measured features will be affected – but, again, don't know which features they are. We must therefore use the whole collection of P-values to correctly decide if there's any difference between the two groups.

- *Sparse feature selection.* A large number of features are available for training a linear classifier, but we expect that most of those features are in fact useless for separating the underlying classes. We must decide which features to use in designing a class prediction rule.

Higher Criticism (HC) and its elaborations can be useful in both of these LSI settings; under a particular asymptotic model discussed below - the *Asymptotic Rare Weak* (ARW) model - HC offers theoretical optimality in selecting features. In this paper we will review the basic notions of HC, some variations and settings where it applies. HC is a flexible idea, and can be adapted to a range of new problem areas; we briefly discuss three simple examples.

1.1 HC basics

John Tukey [113, 114, 115] coined the term "Higher Criticism" [1] and motivated it by the following story. A young scientist administers a total of 250 independent tests, out of which 11 are significant at the level of 5%. The youngster is excited about the findings and plans to trumpet them until a senior researcher tells him that, even in the purely null case, one would expect to have 12.5 significances. In that sense, finding only 11 significances is actually disappointing. Tukey proposes a kind of *second-level significance testing*, based on the statistic

$$HC_{N,0.05} = \sqrt{N}(\text{Fraction significant at } 5\% - 0.05)/\sqrt{.05 \times 0.95},$$

where $N = 250$ is the total number tests. Obviously this score has an interpretation similar to Z- and t- statistics; so Tukey suggests that a value of 2 or larger indicates *significance of the overall body of tests*. In Tukey's example,

$$HC_{N,0.05} = -0.43.$$

If the young researcher really 'had something' this score should be strongly positive, for example 2 or more; but here the score is negative, implying that the overall body of the evidence is consistent with the null hypothesis of no difference. Donoho and Jin [40] saw that in the modern context of large N and rare/weak signals, it was advantageous to generalize beyond the single significance level $\alpha = 0.05$. They maximized over *all* levels α between 0 and some preselected upper bound $\alpha_0 \in (0, 1)$. So generalize Tukey's proposal and set

$$HC_{N,\alpha} = \sqrt{N}(\text{Fraction significant at } \alpha - \alpha)/\sqrt{\alpha \times (1-\alpha)}.$$

If the overall body of tests is significant, then we expect $HC_{N,\alpha}$ to be large for *some* α. Otherwise, we expect $HC_{N,\alpha}$ to be small over all α in a wide range. In other words, the significance of the overall body of test is captured in the following Higher Criticism statistic

$$HC^*_N = \max_{\{0 \le \alpha \le \alpha_0\}} HC_{N,\alpha}, \tag{1.1}$$

where $\alpha_0 \in (0, 1)$ is a tuning parameter we often set at $\alpha_0 = 1/2$.

Higher Criticism (HC) can be computed efficiently as follows. Consider a total of N uncorrelated tests.

[1] In mid-twentieth century humanities studies, the term Higher Criticism became popular to label a certain school of Biblical scholarship.

- For the i-th one, get the corresponding individual P-value π_i, producing in all a body of P-values $\pi_1, \pi_2, \ldots, \pi_N$.
- Sort the P-values in the ascending order:

$$\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(N)}.$$

- The Higher Criticism statistic in (1.1) can be equivalently written as follows

$$HC_N^* = \max_{\{1 \le i \le \alpha_0 N\}} HC_{N,i}, \qquad HC_{N,i} \equiv \sqrt{N} \frac{(i/N) - \pi_{(i)}}{\sqrt{\pi_{(i)}(1-\pi_{(i)})}}. \tag{1.2}$$

In words, we are looking at a test for equality of a binomial proportion $\pi_{(i)}$ to an expected value i/N, maximizing this statistic across a range of i. We think that the evidence against being purely null is located somewhere in this range, but we can't say in advance where that might be. The computational cost of HC is $O(N \log(N))$ and is moderate.

Figure 1 illustrates the definition of Higher Criticism. Consider an example where the (one-sided) P-values π_i are produced by Z-values z_i through $\pi_i = 1 - \Phi(z_i)$, $1 \le i \le N$, where Φ denotes the CDF of $N(0,1)$. The first panel shows the sorted Z-values in the descending order, the second panel shows the sorted P-values, and the last panel shows $HC_{N,i}$. In this example, $HC_N^* = 7.1$, reached by $HC_{N,i}$ at $i = .0085 \times N$.

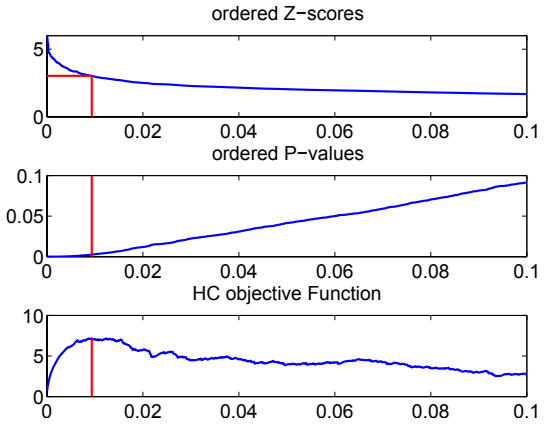

Figure 1: Illustration of HC. The component score maximizing the HC objective is located at the red line. Bottom panel: the HC objective function $HC_{N,i}$ versus i/N. Middle panel: the underlying P-vales $\pi_{(i)}$ versus i/N. Top Panel, the corresponding ordered Z-scores $z_{(i)}$ versus i/N.

Remark. Asymptotic theory shows that the component scores $HC_{N,i}$ can be poorly behaved for i very small (e.g., 1 or 2). We often recommend the following modified version

$$HC_N^+ = \max_{\{1 \le i \le \alpha_0 N:\ \pi_{(i)} > 1/N\}} HC_{N,i}.$$

Remark. As the last remark illustrates, small variations on the above prescription will sometimes be useful; e.g., the modification of the underlying Z-like scores; in (2.8)-(2.9) below. Moreover, there are several other statistics such as Berk-Jones and Average Likelihood Ratio offer cognates or substitutes; see Section 2.7 below. The real point of HC is less the specific definition (1.2) and more a viewpoint about the nature of evidence against the null hypothesis; namely, that although the evidence may be cumulatively substantial, it is diffuse, individually very weak and affecting a relatively small fraction of the individual P-values or Z-scores in our study.

So HC can be viewed as a family of methods for which the above definitions give a convenient entry point. To make utterly clear when needed we label definition (1.2) the Orthodox Higher Criticism (OHC).

1.2 The Rare/Weak effects viewpoint, and phase diagram

Effect *sparsity* was proposed as a useful hypothesis already in the 1980's by Box and Meyer [17]; it proposes that relatively few of the observational units or factorial levels can be expected to show any difference from a global null hypothesis of no effect, and that *a priori* we have no opinion about which units or levels those might be.

The Effect *weakness* hypothesis assumes that individual effects are not individually strong enough to be detectable, once traditional multiple comparisons ideas are taken into account. [2]

The Rare/Weak viewpoint combines *both* hypotheses in analysis of large-scale experiments; it is intended to be a flexible concept and to vary from one setting to another.

The next section operationalizes these ideas in a specific model, where N independent Z-scores follow a mixture with a fraction $(1-\epsilon)$ which are truly null effects and so distributed $N(0,1)$, while the remaining ϵ fraction have a common effect size τ and are distributed $N(\tau,1)$. In this situation, the Rare/Weak viewpoint studies the regime where ϵ is small, the locations of the nonzero effects are scattered irregularly through the scores and the effect size τ is, at moderate N, only 2 or 3 standard deviations.

For large N one can develop a precise theory; see Section 6 below. There we develop the *Asymptotic Rare/Weak* (ARW) model, a framework that assigns parameters to the *rare* and *weak* attributes of a non-null situation; a key phenomenon is that in the two-dimensional parameter space there are three separate regions (or *phases*) where an inference goal is *relatively easy*, *non-trivial but possible*, and *impossible* to achieve, correspondingly. The ARW phase diagram offers revealing comparisons between HC and other seemingly similar methods, such as FDR control.

2 HC for detecting sparse and weak effects

In [40] Higher Criticism was originally proposed for detecting sparse Gaussian mixtures. Suppose we have N test statistics X_i, $1 \le i \le N$, (reflecting many individual genes, voxels, etc.). Suppose that these tests are standardized so that each individual test, under its corresponding null hypothesis, would have mean 0 and standard deviation 1. We are interested in testing whether *all* test statistics are distributed $N(0,1)$, versus the alternative that a small fraction is distributed as normal with an elevated mean τ. In effect, we want an *overall* test of a complete null hypothesis:

$$H_0^{(N)}: \qquad X_i \overset{iid}{\sim} N(0,1), \qquad 1 \le i \le N, \tag{2.3}$$

against an alternative in its complement:

$$H_1^{(N)}: \qquad X_i \overset{iid}{\sim} (1-\epsilon)N(0,1) + \epsilon N(\tau,1), \qquad 1 \le i \le N. \tag{2.4}$$

To use HC for such a case, we calculate the (one-sided) P-values by

$$\pi_i = 1 - \Phi(X_i), \qquad 1 \le i \le N.$$

We then apply the basic definition of HC to the collection of P-values. [3]

In Figure 2, we show the simulated HC values of $H_0^{(N)}$ and $H_1^{(N)}$ based on 100 independent repetitions, where the parameters are set as $(N,\epsilon,\tau) = (10^6, 10^{-3}, 2)$. It is seen that the simulated HC values under $H_1^{(N)}$ are well separated from those under $H_0^{(N)}$.

2.1 Critical value for using HC as a level-α test

Fix $0 < \alpha < 1$. To use HC as a level-α test, we must find a critical value $h(N,\alpha)$ so that

$$P_{H_0^{(N)}}\{HC_N^* > h(N,\alpha)\} \le \alpha.$$

[2] E.g., Bonferroni-based family-wise error rate control.

[3] If we thought that under the alternative the mean might be either positive or negative, we would of course use two-sided P-values

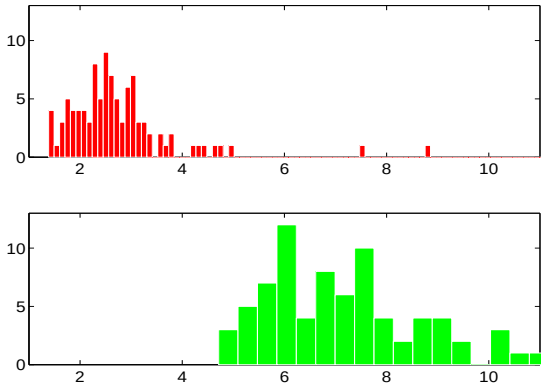

Figure 2: Simulated Higher Criticism values. Top panel: simulation under null hypothesis $H_0^{(N)}$. Bottom panel: simulation under alternative hypothesis $H_1^{(N)}$.

HC_N^* can be connected with the maximum of a standardized empirical process; see Donoho and Jin [40]. Using this connection, it follows from [109, Page 600] that as $N \to \infty$, $b_N HC_N^* - c_N$ and $b_N HC_N^+ - c_N$ converge weakly to the same limit—the standard Gumbel distribution, where $b_N = \sqrt{2\log\log(N)}$ and $c_N = 2\log\log(N) + (1/2)[\log\log\log(N) - \log(4\pi)]$. As a result, for any fixed $\alpha \in (0,1)$ and very large N (the subscript G stands for Gumbel below),

$$h(N,\alpha) \approx h_G(N,\alpha) \approx \sqrt{2\log\log(N)}, \qquad \text{where} \quad h_G(N,\alpha) = b_N^{-1}\big[c_N - \log\log(\tfrac{1}{1-\alpha})\big]. \quad (2.5)$$

When N is moderately large and α is moderately small, the approximations may not be accurate enough, and it is hard to derive an accurate closed-form approximation (even in much simpler cases; see [31] for non-asymptotic bound on extreme values of normal samples). In such cases, it is preferable to determine $h(N,\alpha)$ by simulations. Table 1 displays $h_G(N,\alpha)$ and $h(N,\alpha)$ (where $\alpha_0 = 1/2$ as in (1.1)) computed from 10^5 independent simulations. One sees that: (a) $h_G(N,\alpha)$ approximate the percentiles of HC_N^* poorly, but approximate those of HC_N^+ reasonably well, especially when N get larger and α get smaller; (b) the tail of HC_N^* is fat but that of HC_N^+ is relatively thin; (c) the percentiles of HC_N^+ and HC_N^* increase with N only very slowly; therefore, the values of $h(N,\alpha)$ for a few selected N represent those of a wide range of N. Very recently, Li and Siegmund [90] proposes a new approximation to $h(N,\alpha)$ which is more accurate when N is moderately large.

level	statistic	N			
		10^3	5×10^3	2.5×10^4	1.25×10^5
$\alpha = .05$	HC_N^+	3.17 (3.00)	3.22 (3.08)	3.26 (3.14)	3.30 (3.19)
	HC_N^*	4.77 (3.00)	4.73 (3.08)	4.74 (3.14)	4.75 (3.19)
$\alpha = .01$	HC_N^+	3.95 (3.83)	3.97 (3.87)	3.96 (3.90)	3.99 (3.93)
	HC_N^*	10.08 (3.83)	9.88 (3.87)	10.20 (3.90)	9.92 (3.93)
$\alpha = .005$	HC_N^+	4.29 (4.18)	4.28 (4.20)	4.26 (4.22)	4.28 (4.24)
	HC_N^*	13.78 (4.18)	14.39 (4.20)	14.34 (4.22)	13.95 (4.24)
$\alpha = .001$	HC_N^+	5.03 (5.00)	5.02 (4.98)	4.98 (4.97)	4.98 (4.97)
	HC_N^*	30.27 (5.00)	30.36 (5.02)	31.95 (4.97)	31.49 (4.97)

Table 1: Simulated values $h(N,\alpha)$ based on 10^5 repetitions. Numbers in brackets are $h_G(N,\alpha)$.

2.2 Two gene microarray data sets

In this paper (Sections 2.3.1 and 3.1), we use two standard gene microarray data sets to help illustrate the use of HC: the lung cancer data analyzed by Gordon *et al.* [56], and the leukemia

data analyzed by Golub *et al.* [54] (for the latter, we use the cleaned version published by Dettling [35], which contains measurements for 3571 genes). Both data sets are available at `www.stat.cmu.edu/~jiashun/Research/software/`. See Table 2, where the partition of samples into the training set and the test set is the same as in [56] and [54], respectively.

Data Name	# training samples	# test samples	# genes
Leukemia	27 (ALL), 11 (AML)	20 (ALL), 14 (AML)	3571
Lung Cancer	16 (MPM), 16 (ADCA)	15 (MPM), 134 (ADCA)	12533

Table 2: MPM: malignant pleural mesothelioma. ADCA: adenocarcinoma. ALL: acute lymphoblastic leukemia. AML: acute myelogenous leukemia.

2.3 Detecting rare and weak effects in genomics and genetics

When the genomics revolution began 10-15 years ago, many scientists were hopeful that the common disease common variant hypothesis would apply. Under this hypothesis, there would be, for each common disease, a specific gene that is clearly responsible. Such hopes were dashed over the coming years, and today, much research starts from the hypothesis that numerous genes are differentially expressed in affected patients, but with individually small effect sizes [30]. HC, with its emphasis on detecting rare and weak effects, seems well-suited to this new environment.

2.3.1 Two worked examples

Let's apply HC to the two gene microarray datasets. For each in turn, let x_{ij} be the expression level for the i-th sample and the j-th gene, $1 \le i \le n$, $1 \le j \le p$. Let C and D be the set of indices of samples from the training set and the test set, respectively. For notational consistency with later sections, we only use the data in the training set, but using the whole data gives similar results.

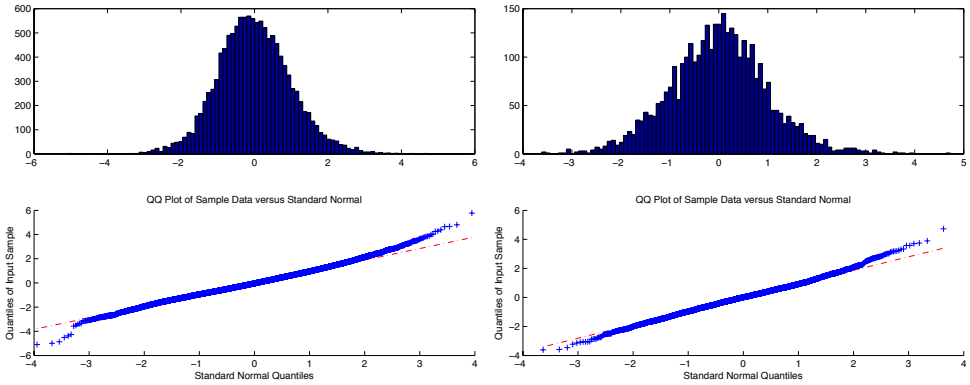

Figure 3: Left Column: histogram (top) and qq-plot (bottom) of $Z = (Z_1, Z_2, \ldots, Z_p)'$ for the lung cancer data set. Right: Corresponding plots for the leukemia data set.

Write $C = C_1 \cup C_2$, where C_1 and C_2 are the sets of indices of the training samples from Class 1 and 2, respectively. Fix $1 \le j \le p$. Let $\bar{x}_{jk} = \frac{1}{|C_k|}\sum_{j \in C_k} x_{ij}$ denote the average expression value of gene j for all samples in Class k, $k = 1, 2$, and $s_j^2 = \frac{1}{(|C|-2)}[\sum_{i \in C_1}(x_{ij} - \bar{x}_{j1})^2 + \sum_{i \in C_2}(x_{ij} - \bar{x}_{j2})^2]$ the pooled variance. Define the t-like statistic

$$z_j^* = \frac{1}{\sqrt{1/|C_1| + 1/|C_2|}} \frac{\bar{x}_{j1} - \bar{x}_{j2}}{s_j}, \qquad 1 \le j \le p.$$

In the null case, if the data $\{x_{ij}\}_{1 \le i \le n, 1 \le j \le p}$ are independent and identically distributed across different genes, and for each gene j, $\{x_{ij}\}_{i=1}^n$ are normal samples with the same variance

in each class, then z_j^* has the Student's t-distribution with $df = |C| - 2$ when gene j is not differentially expressed. Using this we may calculate individual P-values for each gene and apply HC. However, as pointed out in Efron [43], a problem one frequently encounters in analyzing gene microarray data is the so-called *discrepancy between the empirical null and theoretical null*, meaning that there is a gap between the aforementioned t-distribution and the empirical null distribution associated with $\{z_j^*\}_{j=1}^p$. This gap might be caused by unsuspected between-gene variance components or other factors. We follow Efron's suggestion and standardize z_j^*:

$$Z_j = \frac{z_j^* - \bar{z}^*}{sd(z^*)}, \qquad 1 \le j \le p, \tag{2.6}$$

where $\bar{z}^*$ and $sd(z^*)$ represent the empirical mean and standard deviation associated with $\{z_j^*\}_{j=1}^p$, respectively. We call $Z_1, Z_2, \ldots, Z_p$ the standardized Z-scores, and believe that in the purely null case they are approximately normally distributed. In Figure 3, we show the histogram (top) and the qqplot (bottom) associated with the standardized Z-scores. The figure suggests that for both data sets, the standardization in (2.6) is effective.

We apply HC to $\{Z_j\}_{j=1}^p$ for both the leukemia and the lung cancer data, where the individual (two-sided) P-values are obtained assuming $Z_j \sim N(0,1)$ if the j-th gene is not differentially expressed. The resulting HC scores are 6.1057 and 13.3025 in two cases. The P-values associated with the scores (computed by numerical simulations) are $\approx 5 \times 10^{-5}$ and $< 10^{-5}$, respectively. They suggest the definite presence of signals, sparsely scattered in the Z-vector. And, indeed, the qqplots exhibit a visible 'curving away' from the identity lines.

An alternative approach to computing these two P-values uses random shuffles. Denote the data matrix by $X = (x_{ij})_{1 \le i \le n, 1 \le j \le p}$. First, we randomly shuffle the rows of X, independently between different columns. For the permuted data, we follow the steps above and calculate the standardized Z-vector. Due to our shuffling, the signals wash out and the Z-vector can be viewed as containing no effect signals. Next, apply HC to the Z-vector, obtaining an HC statistic specific to that shuffle. Repeat the whole process for 1000 independent shuffles. As a result, we have 1001 HC scores: one, based on the original data matrix; all others, based on shuffles. The shuffles are used to calculate the P-value of the original HC score. For the (training) leukemia data and the (training) lung cancer data, the resultant P-values are approximately 0.01 and < 0.001, respectively. Both data sets have standardized Z-vectors exhibiting very subtle departures from the null hypothesis, but still permit reliable rejection of the null hypothesis.

2.3.2 Applications to Genome-wide association study

By now the literature of Genome-wide association studies (GWAS) has several publications applying HC or its relatives. Parkhomenko *et al.* [101] used HC to detect modest genetic effects in a genome-wide study of rheumatoid arthritis. It was further suggested in Martin *et al.* [92] that the implementation of HC in GWAS may provide evidence for the presence of additional remaining SNPs modestly associated with the trait.

Sabatti *et al.* [106] used HC in a GWAS for metabolic traits. HC enabled them to quantify the strength of the overall genetic signal for each of the nine traits (Triglycerides, HDL, ...) they were interested in, where to deal with the possible dependence caused by Linkage Disequilibrium (LD) between SNPs, they computed individual P-values by permutations. See also De la Cruz *et al.* [32] where the authors considered the problem of testing whether there are associated markers in a given region or a given set of markers, with applications to analysis of a SNP data set on Crohn's disease.

Wu *et al.* [121] adapted HC for detecting rare and weak genetic signals using the information of LD. He and Wu [63] used HC and innovated HC for signal detection for large-scale exonic single-nucleotide polymorphism data, and suggested modifications of HC in such settings.

Motivated by GWAS, Mukherjee *et al.* [96] considered the signal detection problem using logistic regression coefficients rather than 2-sample Z-scores, and discovered an interesting relationship between the sample size and the detectability when both response variable and design variables are discrete. See Section 2.10 for more discussions on signal detection problem associated with regression models. To address applications in GWAS, Roeder and Wasserman [103] made an interesting connection between HC and weighted hypothesis testing.

2.3.3 Applications to DNA Copy Number Variation

Computational biology continues to innovate and GWAS is no longer the only game in town. DNA Copy Number Variation (CNV) data grew rapidly in importance after the GWAS era began, and today provide an important window on genetic structural variation. Jeng et al. [71, 72] applied HC-style thinking to CNV data and proposed a new method called *Proportion Adaptive Segment Selection (PASS)*. PASS can be viewed as a two-way screening procedure for genomic data, which targets both the signal sparsity across different features (SNPs) and the sparsity across different subjects—so-called rare variation in genomics.

2.4 Applications to Cosmology and Astronomy

HC has been applied in several modern experiments in Astronomy and Cosmology, where typically the experiment produces data which can be interpreted as images (of a kind), and where there is a well-defined null hypothesis, whose overthrow would be considered a shattering event.

Studies of the Cosmic Microwave Background (CMB) offer several examples. CMB is a relic of radiation emitted when the Universe was about 370,000 years old. This radiation exhibits characteristic of an almost perfect blackbody at a temperature of 2.726 Kelvin. In the simplest inflation models, CMB temperature fluctuations should in each pixel behave as a realization of a zero-mean Gaussian random variable. The resulting Gaussian field (on the sphere) is completely determined by its power spectrum. In recent decades, a large number of studies have been devoted to the subject of detecting non-Gaussian signatures (hot spots, cold spots, excess kurtosis, ...) in the CMB.

Jin *et al* [79], and Cayon *et al* [27] (see also [29, 26]), applied HC to standardized wavelet coefficients of CMB data from the Wilkinson Microwave Anisotropy Probe (WMAP). HC would be sensitive to small collection of such coefficients departing from the standard null, without requiring that individual coefficients depart in a pronounced way. Compared to the kurtosis-based non-Gaussianity detector (widely used in cosmology when the departure from Gaussianity is in the not-very-extreme tails), HC showed superior power and sensitivity, and pointed in particular to the *cold spot* centered at galactic coordinate (longitude, latitude) = $(207.8°, -56.3°)$. In [116], Vielva reviews the cold spot detection problem, and shows that HC rejects Gaussianity, confirming earlier detections by other methods.

Gravitational weak lensing calculations measure the distortion of background galaxies supposedly caused by intervening large-scale structure. Pires *et al* [102] applied many nonGaussianity detectors to weak lensing data, including the empirical Skewness, the empirical Kurtosis, and HC, and showed that HC is competitive, while of course being more specifically focused on excess of observations in the tails of the distribution.

Most recently, Bennett *et al.* [11] applied the HC ideas to the problem of Gravitational Wave detection. They use HC as a second-pass method operating on F-statistic and C-statistics for a monochromatic periodic source in a binary system; such statistics contain a large number of relatively weak signals spread irregularly across many frequency bands. They use a modified form of HC, which is both sensitive and robust, and offer a noticeable increase in the detection power (e.g., a 30% increase in detectability for phase-wandering source over multiple time intervals).

2.5 Applications to disease surveillance and local anomaly detection

In disease surveillance, we have aggregated count data c_i representing cases of a certain disease (e.g. influenza) at the i-th spatial region (e.g. zip code), $1 \le i \le N$. When disease breaks out, the counts will have elevated values in one or a few small geographical regions. Neill and Lingwall [99, 98] use HC for disease surveillance and spatio-temporal cluster detection: they suppose we have historical counts for each spatial location measured over time $t = 1, 2, \ldots, T$. The P-value of c_i is calculated by $(d_i + 1)/(T + 1)$, where d_i is the number of historical counts larger than c_i at the i-th location.

Disease outbreak detection is a special case of *local anomaly* detection as studied in Saligrama and Zhao [107]. Suppose we have a graph $G = (V, E)$ with usual graph metric, where a random variable is associated with each node. A simple scenario of *local anomaly* they consider assumes

that for all nodes outside the anomaly, the associated random variables have the same density f_0, and for nodes inside the anomaly, the associated density is different from f_0. Saligrama and Zhao [107] investigate several models and statistics for local anomaly detection; HC is found to be competitive in this setting.

2.6 Estimating the proportion of non-null effects

As presented so far, HC offers a test statistic. In the setting of Section 2, we sample X_i from the two-component mixture

$$X_i \overset{iid}{\sim} (1-\epsilon)N(0,1) + \epsilon N(\tau, 1), \qquad 1 \le i \le N. \tag{2.7}$$

The detection problem which HC addresses, involves testing $H_0^{(N)} : \epsilon = 0$ versus $H_1^{(N)} : \epsilon > 0$. Alternatively, one could estimate the mixing proportion ϵ. Motivated by a study of Kuiper Belt Objects (KBO) (e.g., [95]), Cai *et al.* [23] (see also [95]) develop HC into an estimator for ϵ, focusing on the regime where $\epsilon > 0$ is very small. [4]

In the growing literature of large-scale multiple testing, the problem of estimating the proportion of non-null effects has attracted considerable attention in the past decade, though sometimes with different goals. For example, rather than knowing the true proportion of nonzero effects, one might only want to estimate the largest proportion within which the false discovery rate can be controlled. The literature along this line connects to work of Benjamini and Hochberg [10] on controlling False Discovery Rate (FDR), and Efron [43] on controlling the local FDR in gene microarray studies. See [22, 78, 76] and references therein.

2.7 Statistics with HC-like constructions

HC can be viewed as a measure of the goodness-of-fit between two distributions, namely between the distribution F_N of the empirical P-values and the model uniform distribution F_0. In this viewpoint, HC is effectively computing the distance measure

$$\mu_1(F_N, F_0) = \sqrt{N} \cdot \max_{i=1}^{N} \frac{|F_N(i/N) - F_0(i/N)|}{\sqrt{F_0(i/N)(1-F_0(i/N))}} \tag{2.8}$$

(or, more properly, a restricted form, where $i = 1$ and $i > \alpha_0 N$ are omitted in the maximum) or the reverse

$$\mu_2(F_N, F_0) = \sqrt{N} \cdot \max_{i=1}^{N} \frac{|i/N - F_0(\pi_{(i)})|}{\sqrt{F_0(\pi_{(i)})(1-F_0(\pi_{(i)}))}}. \tag{2.9}$$

We call these the *theoretically standardized* and *empirically standardized* goodness of fit, respectively.

To understand HC, then, one might consider how it differs from other measures of the discrepancy between two distributions. HC includes the element of standardization, which for many readers will suggest comparison to the Anderson-Darling statistic [4]:

$$A(F_N, F_0) = N \cdot \int \frac{|F_N(x) - F_0(x)|^2}{F_0(x)(1-F_0(x))} dx.$$

HC, however, involves maximization rather than integration, which makes it a kind of weighted Kolmogorov-Smirnov statistic. Jager and Wellner [69] investigated the limiting distribution of a class of weighted Kolmogorov statistics, including HC as a special case.

Another perspective is to view the P-values underlying HC as obtained from the normal approximation to a one-sample test for a known binomial proportion, and to consider instead the exact test based on likelihood ratios, or asymptotic tests based instead on KL divergence

[4] Among the many competing methods, we mention just [8] . Let $\pi_i = 1 - \Phi(X_i)$ in Model (2.7), then π_i are iid samples from the density $f_{\epsilon,\tau}(x) = (1-\epsilon) + \epsilon g_{\epsilon,\tau}(x)$, where $g_{\epsilon,\tau}(x)$ is monotone decreasing in $0 < x < 1$ and is unbounded at 0. Balabdaoui *et al* [8] studies the behavior of the maximum likelihood estimator of $f_{\epsilon,\tau}(x)$ at 0, and uses it to derive an alternative estimator for ϵ.

between the binomial with parameter π_i and the binomial with parameter i/N. This perspective reveals a similarity of HC to the *Berk-Jones* (BJ) statistic [12]. The similarity was carefully studied in [40, Section 1.6]; see details therein. Using the divergence $D(p_0, p_1) = p_0 \log(p_0/p_1) + (1-p_0)\log((1-p_0)/(1-p_1))$, the Berk-Jones statistic can be written as

$$BJ = \max_{i=1}^{N} N \cdot D(\pi_i, i/N).$$

Wellner and Koltchinskii [118] derive the limiting distribution of the Berk-Jones statistic, finding that it shares many theoretical properties in common with HC.

In [70], Jager and Wellner introduced a new family of goodness-of-fit tests based on the ϕ-divergence, including HC as a special case, and show all such tests achieve the optimal detection boundary in [40] (see the discussion below in Section 6).

Reintroducing the element of integration found in the Anderson-Darling statistic, Walther [117] proposed an *average likelihood ratio* (ALR) approach. If $LR_{i,N}$ denotes the usual likelihood ratio for a one-sided test of the binomial proportion, ALR takes the form

$$ALR = \sum_{i=1}^{\alpha_0 N} w_{i,N} LR_{i,N}; \qquad LR_{i,N} \equiv \exp(N \max\{D(\pi_i, i/N), 0\}),$$

with weights $w_{i,N} = (2i\log(N/3))^{-1}$. Walther shows that ALR compares favorably with HC and BJ for finite sample performance, while having similar asymptotic properties under the ARW model discussed below. See [40, 115] for more discussions on the relative merits of HC, BJ, and ALR.

Additionally, as a measure of goodness-of-fit, HC is closely related to other goodness-of-fit tests, motivated, however, by the goal of optimal detection of presence of mixture components representing rare/weak signals. We remark that the pontogram of Kendall and Kendall [85] is an instance of HC, applied to a special set of P-values.

Gontscharuk *et al* [55] introduced the notion of *local levels* for goodness-of-fit tests, and studied the asymptotic behavior when applying the framework to one version of HC; for HC, the local level associated with $HC_{N,i}$ and a critical value χ roughly translates to $P[HC_{N,i} \geq \chi]$.

2.8 Connection to FDR-controlling methods

HC is connected to Benjamini and Hochberg's (BH) False Discovery Rate (FDR) control method in large-scale multiple testing [10]. Given N uncorrelated tests where $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(N)}$ are the sorted P-values, introduce the ratios

$$r_k = \pi_{(k)}/(k/N), \qquad 1 \leq k \leq N.$$

Given a prescribed level $0 < q < 1$ (e.g., $q = 5\%$), and let $k = k_q^{FDR}$ be the largest index such that $r_k \leq q$. BH's procedure rejects all tests whose P-values are among the k_q^{FDR} smallest, and accepts all others. The procedure controls the FDR in that the expected fraction of false discoveries is no greater than q.

The contrast between FDR control method and HC can be captured in a few simple slogans. We think of the BH procedure as targeting *rare but strong* signals, with the main goal to select the few strong signals embedded in a long list of null signals, without making too many false selections. HC targets the more delicate regime where the signals are *rare and weak*. In the rare/weak settings, the signals and the noise may be almost indistinguishable; and while the BH procedure still controls the FDR, it yields very few discoveries. In this case, a more reasonable goal is to test whether any signals exist without demanding that we properly identify them all; this is what HC is specifically designed for. See also Benjamini [9].

HC is also intimately connected to the problem of constructing confidence bands for the *False Discovery Proportion* (FDP). See Cai et al. [23], Ge and Li [51], and de Una-Alvarez [37].

2.9 Innovated HC for detecting sparse mixtures in colored noise

So far, the underlying P-values were always assumed independent. Dai *et al.* [30], points out the importance of the correlated case for genetics and genomics; we suppose many other application areas have similar concerns. Hall and Jin [58] showed that directly using HC in such cases could be unsatisfactory, especially under strong correlations. Hall and Jin [59] pointed out that correlations (when known or accurately estimated) need not be a nuisance or curse, but could sometimes be a blessing if used properly. They proposed *Innovated Higher Criticism*, which applies HC in a transformed coordinate system; in analogy to time series theory, Hall and Jin called this the innovations domain. Innovated HC was shown to be successful when the correlation matrix associated with the noise entries has polynomial off-diagonal decay.

2.10 Signal detection problem associated with regression models

Suppose we observe an $n \times 1$ vector Y which satisfies a a linear regression model

$$Y = X\beta + z, \qquad z \sim N(0, I_n),$$

where X is the $n \times N$ design matrix, β is the $N \times 1$ vector of regression coefficients, and z is the noise vector. The problem of interest is now to test whether all regression coefficients β_i are 0 or a small fraction of them is nonzero. The setting considered in [58, 59] is a special case, where the number of variables N is the same as the sample size n.

Arias-Castro *et al.* [6] and Ingster, Tsybakov, and Verzelen [68] considered the more general case where N is much larger than n. The main message is that, under some conditions, what has been previously established for the Gaussian sequence model extends to high-dimensional linear regression. Motivated by GWAS, Mukherjee *et al.* [96] considered a similar problem with binary response logistic regression. They exposed interesting new phenomena governing the detectability of non-null β when both response variable and design variables are discrete.

Meinshausen [93] considers the problem of variable selection associated with a linear model. Adapting HC to the case of correlated noise with unknown variance, he uses the resultant method for hierarchical testing of variable importance setting. The method is shown to be able to significantly improve testing powers. Charbonnier [28] generalizes HC from one-sample testing problem to two-sample testing problem. It considers two linear models and tries to test if the regression coefficient vectors are the same. Also related is Suleiman and Ferrari [110], where the authors use constrained likelihood ratios for detecting sparse signals in highly noisy 3D data.

2.11 Signal detection when noise distribution is unknown/nonGaussian

In model (2.3)-(2.4), the noise entries are iid samples from $N(0, 1)$. In many applications, the noise distribution is unknown and is probably nonGaussian. To use HC for such settings, we need an approach to computing P-values π_i, $1 \le i \le N$.

Delaigle and Hall [33] and Delaigle *et al.* [34] address this problem in the settings where the data are arranged in a 2-D array $\{X(i, j)\}$, $1 \le i \le n$, $1 \le j \le N$. In this array, different columns are independent, and entries in the j-th column are iid samples from a distribution F_j which is unknown and presumably nonGaussian. We need to associate a P-value with each column. The authors consider the situation where the statistician plans to summarize each column using the Student's t-statistic obtaining scores $t_1, t_2, \ldots, t_N$. They propose to generate bootstrapped t-scores $t_1^*, t_2^*, \ldots, t_N^*$ to compute the P-values. They further investigate how the relative magnitude of n and N affects closeness between the marginal distribution of the derived P-values and that of the standard Uniform distribution, provided that the mean of F_j is 0. A similar setting is considered by Greenshtein and Park [57] and by Liu and Shao [91]. The first paper proposes a modified Anderson-Darling statistic and investigates robustness for different range of (n, N). The second paper proposes a test based on extreme values of Hotelling's T^2, and studies the case where the sparse signals appear in groups and the underlying distributions are not necessarily normal.

2.12 Detecting sparse mixtures more generally

More generally, the problem of detecting sparse mixtures considers hypotheses

$$H_0^{(N)}: \quad X_i \overset{iid}{\sim} F, \qquad vs. \qquad H_1^{(N)}: \quad X_i \overset{iid}{\sim} (1-\epsilon)F + \epsilon G,$$

where $\epsilon \in (0,1)$ is small and F and G are two distributions that are presumably different; (ϵ, F, G) may depend on N. In Donoho and Jin [40], $F = N(0,1)$ and $G = N(\tau, 1)$ for some $\tau > 0$.

Cai *et al.* [21] extends the study and considers the case where $F = N(0,1)$ and $G = N(\tau, \sigma^2)$, so the mixture in the alternative hypothesis is not only heterogeneous but also heteroscedastic, and σ models the heteroscedasticity. They found that σ has a surprising phase-change effect over the detection problem. The heteroscedastic model is also considered in Bogdan *et al.* [15] and Bogdan *et al.* [16] from a Bayesian perspective. Park and Ghosh [100] gave a nice review on recent topics on multiple testing where HC is discussed in detail.

Cai and Wu [25] extend the study to the more general case where $F = N(0,1)$ and G is a Gaussian location mixture with a general mixing distribution, and study the detection boundary as well as the detectability of HC.

Arias-Castro and Wang [7] investigate the case where F is *unknown* but symmetric, and develop distribution free tests to tackle several interesting problems, including that of testing of symmetry.

In addition, Gayraud and Ingster [50] consider the problem of detecting sparse mixtures in the functional setting, and shows that the HC statistic continues to be successful in the very sparse case; Laurent *et al.* [89] consider the problem of testing whether the samples X_i come from a single normal, or a mixture of two normals with different means (both means are unknown).

In a closely related setting, Addario-Berry *et al.* [1] and Arias-Castro *et al.* [5] consider structured signals, forming clusters in geometric shapes that are unknown to us. The setting is closely related to that considered in [59, Section 6]. Haupt et al [61, 62] consider a more complicated setting where adaptive sample scheme is available, where we can do inference and collect data in an alternating order.

3 Higher Criticism for feature selection

Higher Criticism has applications far beyond the testing of a global null hypothesis.

Consider a classification problem where we have training samples (X_i, Y_i), $1 \le i \le n$, from two different classes. We denote X_i by the feature vectors and $Y_i = \pm 1$ the class labels. For simplicity, we assume two classes are equally likely and the feature vectors $X_i \in R^p$ are Gaussian distributed with identical covariances, so that, after a standardizing transformation, the feature vector $X_i \sim N(Y_i \cdot \mu, I_p)$, with vector μ being the contrast mean and I_p the $p \times p$ identity matrix. Given a fresh feature vector X, the goal is to predict the associated class label $Y \in \{-1, 1\}$.

We are primarily interested in the case where $p \gg n$ and where the contrast mean vector μ is unknown but has nonzero coordinates that are both rare and weak. That is, only a small fraction of coordinates of μ is nonzero, and each nonzero coordinate is individually small and contributes weakly to the classification decision.

In the classical $p < n$ setting, consider traditional Fisher linear discriminant analysis (LDA). Letting $w = (w(j), 1 \le j \le p)$ denote a sequence of feature weights, Fisher's LDA takes the form

$$L(X) = \sum_{j=1}^{p} w(j) X(j).$$

It is well-known that the optimal weight vector $w \propto \mu$, but unfortunately μ is unknown to us and in the $p > n$ case can be hard to estimate; especially when the nonzero coordinates of μ are rare and weak; in that case, the empirical estimate $\bar{X}$ is noisy in every coordinate, and only a few coordinates 'stick out' from the noise background.

Feature selection (i.e., selecting a small fraction of the available features for classification) is a standard approach to attack the challenges above. Define a vector of feature scores

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i \cdot X_i); \tag{3.10}$$

this contains the evidence in favor of each feature's significance. We will select a subgroup of features for our classifier, using hard thresholding of the feature scores. For a threshold value $t > 0$ still to be determined, define the hard threshold function

$$w_t(z) = \mathrm{sgn}(z) \cdot 1_{\{|z|>t\}},$$

which selects the features having sufficiently large evidence, and preserves the sign of such feature scores. The post-feature-selection Fisher's LDA rule is then

$$L_t(X) = \sum_{j=1}^{p} w_t(j) X(j),$$

and we simply classify Y as ± 1 according to $L_t(X) \gtrless 0$. This is related to the modified HC in [123], but there the focus is on signal detection instead of feature selection.

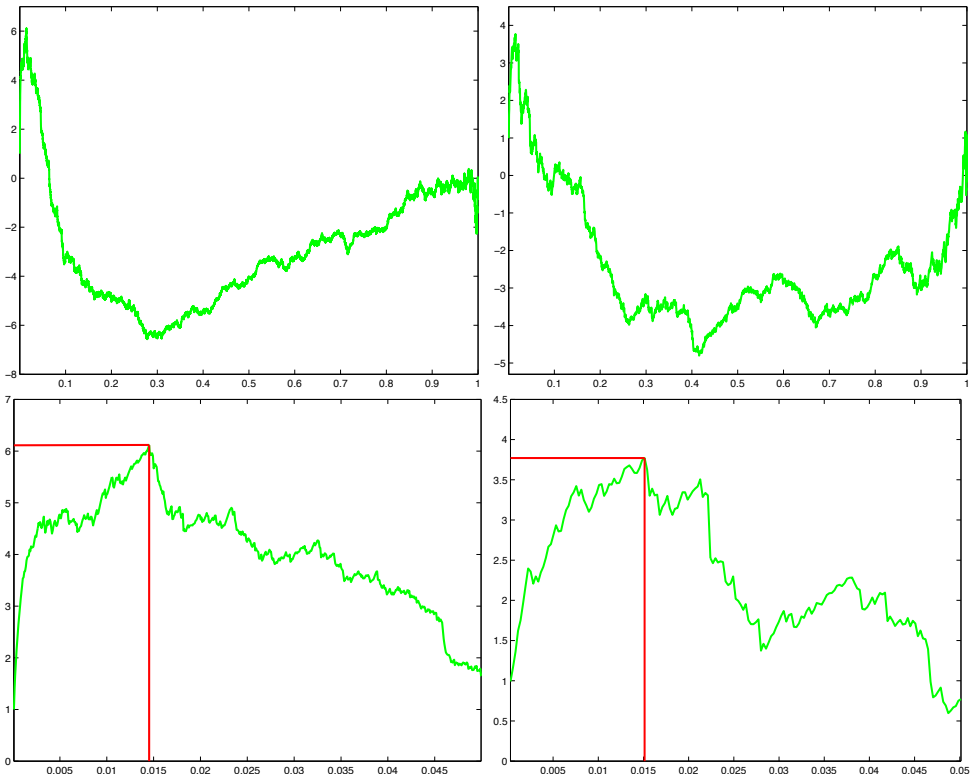

Figure 4: Top row: plot of feature scores $HC_{N,i}$ versus i/N for lung cancer data (left) and leukemia data (right). Bottom row: enlargements of plots in the top row.

How should we set the threshold t? Consider *HC feature selection*, where a simple variant of HC is used to set the threshold. To apply HC to feature selection, we fix $\alpha_0 \in (0, 1/2]$ and follow three steps (to be consistent with OHC described in Section 1.1, we switch back from p to N; note that $N = p$ in this section):

- Calculate a (two-sided) P-value $\pi_j = P\{|N(0,1)| \geq |Z(j)|\}$ for each $1 \leq j \leq N$.
- Sort the P-values into ascending order: $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(N)}$.

- Define the *Higher Criticism feature scores* by

$$HC(i;\pi_{(i)}) = \sqrt{N}\frac{i/N - \pi_{(i)}}{\sqrt{(i/N)(1-i/N)}}, \qquad 1 \le i \le N. \tag{3.11}$$

  Obtain the maximizing index of $HC(i;\pi_{(i)})$:

$$\hat{i}^{HC} = \text{argmax}_{\{1\le i\le \alpha_0\cdot N\}}\{HC(i,\pi_{(i)})\}.$$

  The *Higher Criticism threshold (HCT)* for feature selection is then by

$$\hat{t}_N^{HC} = \hat{t}_N^{HC}(Z_1, Z_2, \ldots, Z_N; \alpha_0, n) = |Z|_{\hat{i}^{HC}}.$$

In modern high-throughput settings where *a priori* relatively few features are likely to be useful, we set $\alpha_0 = 0.10$. [5] [6] See [42] for explanation.

Once the threshold is decided, LDA with HC feature selection is

$$L_{HC}(X) = \sum_{j=1}^{p} w_{HC}(j)X(j), \qquad \text{where} \qquad w_{HC}(j) = \text{sgn}(Z(j))1\{|Z(j)| \ge \hat{t}_p^{HC}\},$$

and the HCT trained classification rule will classify $Y = \pm 1$ according to $L_{HC}(X) \gtrless 0$.

The classifier above is a computationally inexpensive approach, especially when compared to resampling-based methods (such as cross-validations, boosting, etc.). This gives HC a lot of computational advantage in the now very relevant "Big Data" settings.

3.1 Applications to gene microarray data

We now apply the HCT classification rule to the two microarray data sets discussed earlier. Again Z_j is the standardized Z-score associated with the j-th gene, using all samples in the training set C.

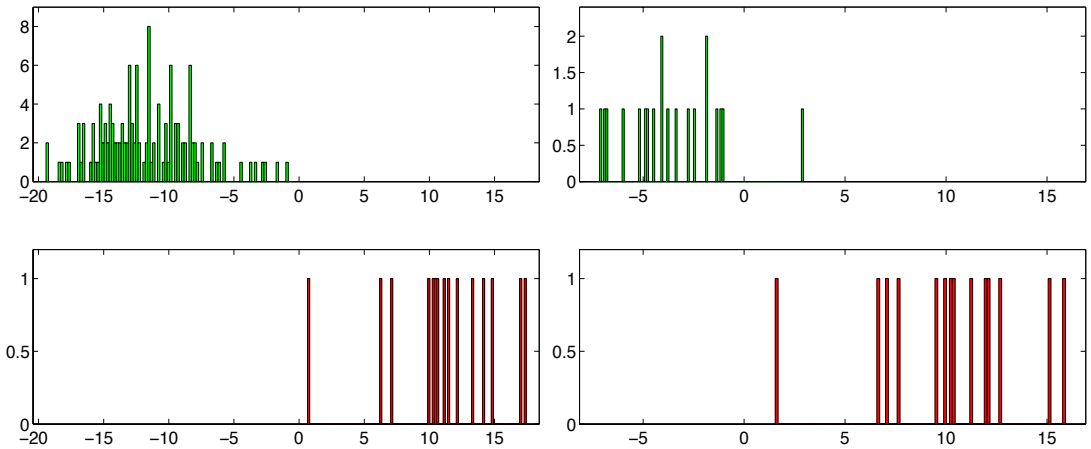

Figure 5: Top row: histogram of the test scores corresponding to Class 1. Bottom row: Class 2. Left column: lung cancer data. Right column: leukemia data.

First, we apply HC to $Z = (Z_1, Z_2, \ldots, Z_p)'$ to obtain the HC threshold; this will also determine how many features we keep for classification. The scores $HC_{N,i}$ are displayed in Figure 4. For the lung cancer data, the maximizing index is $\hat{i}^{HC} = 182$, at which the HC score is

[5] In practice, HCT is relatively insensitive to different choices of α_0.

[6] Note the denominator of the HC objective function is different from from the denominator used earlier, in testing, although the spirit is similar. The difference is analogous to the one between the two goodness of fit tests (2.8) and (2.9).

6.112, and we retain all 182 genes with the largest Z-scores (in absolute value) for classification (equivalently, a gene is retained if and only if the Z-score exceeds $\hat{t}_p^{HC} = 2.65$). For the leukemia data, $\hat{i}^{HC} = 54$, with HC score 3.771, and the threshold $\hat{t}_p^{HC} = 2.68$.

Next, for each sample X_i in the test set D, we calculate the HCT-based LDA score. Recall that for any $i \in C$, the data associated with the i-th sample is $X_i = \{x_{ij}\}_{j=1}^p$. The HCT-based LDA score $lda_i = lda(X_i)$ is given by:

$$lda_i = \sum_{j=1}^{p} w_{HC}(j)\Big(\frac{x_{ij} - \bar{x}_j}{s_j}\Big),$$

where we recall $w_{HC}(j) = \mathrm{sgn}(Z_j)1\{|Z_j| \geq \hat{t}_p^{HC}\}$, and $(\bar{x}_j, s_j, w_{HC}(j))$ only depend on the training samples in C. The scores $\{lda_i\}_{i=1}^n$ are displayed in Figure 5, where we normalized each score by a common factor of $1/\sqrt{\hat{i}^{HC}}$ for clarity. The scores corresponding to Class 1 are displayed in the top row in green (ADCA for lung cancer data, and ALL for leukemia), and the scores for Class 2 are displayed in the bottom row in red; the left column displays lung cancer data and the right column displays leukemia data). For lung cancer data, LDA-HCT correctly classifies each sample. For leukemia data, LDA-HCT correctly classifies each sample, with one exception: sample 29 in the test set (number 67 in the whole data set).

We employed these two data sets because they gave such a clear illustration. In our previous paper [41], we considered each data in Dettling's well-known compendium, which includes the colon cancer data and the prostate data. The results were largely as good as or better than other classifiers, many involving much fancier-sounding underlying principles.

3.2 Threshold choice by HC: applications to biomarker selection

Wehrens and Fannceschi [119] used HC thresholding for biomarker selection to analyze metabolomics data from spiked apples. They considered P-values that are calculated from Principal Component scores and reported a marked improvement in biomarker selection, compared to the standard selection obtained by existing practices. The paper concludes that HC thresholds can differ considerably from current practice, so it is no longer possible to blindly apply the selection thresholds used historically; the data-specific cutoff values provided by HC open the way to objective comparisons between biomarker selection methods, not biased by arbitrary habitual threshold choices.

3.3 Comparison to other classification approaches

LDA-HCT classifier is closely related to other threshold-based LDA feature selection rules: PAM by Tibshirani *et al.* [112] and FAIR by Fan and Fan [46]. HCT picks the threshold based on feature Z-scores by Higher Criticism, while the other methods set this threshold differently. For the same datasets we discussed earlier, the error rates for PAM and FAIR were reported in [46, 112]; as it turns out, LDA-HCT has smaller error rates.

Comparisons with some of the more popular 'high-tech' classifiers (including Boosting [36], SVM [19], and Random Forests [18]) were reported in [41]. More complex methods usually need careful tuning to perform well, but HCT-LDA is very simple, both conceptually and computationally. When used on the ensemble of standard data sets published in Dettling, HCT-LDA happens to be minimax-regret optimal: it suffers the least performance loss, relative to the best method, across the ensemble.

Hall *et al.* [60] apply HC for classification in a different manner. They view HC as a goodness-of-fit diagnostic. Their method first uses the training vectors to obtain the empirical distributions of each class, and then uses HC to tell which of these distributions best fits each test vector. They classify each test vector using the best-fitting class distribution. While this rule is sensible, it turns out that in a formal asymptotic analysis using the rare/weak model, it is outperformed substantially by HCT-LDA.

3.4 Connection to Feature Selection by controlling feature-FDR

False Discovery Rate control methods offer a popular approach for feature selection. Fix $0 < q < 1$. $FDRT_q$ selects features in a way so that

$$\text{feature-FDR} \equiv E\left[\frac{\#\{\text{Falsely selected features}\}}{\#\{\text{All selected features}\}}\right] \leq q.$$

In the simple setting considered in Section 3, this can be achieved by applying Benjamini-Hochberg's FDR controlling method to all feature P-values. The approach appeals to the common belief that, in order to have optimal classification behavior, we should select features in a way so that the feature-FDR stays small.

However, such beliefs have theoretical support only when signals are rare/strong. In principle, the optimal q associated with the optimal classification behavior should depend on the underlying distribution of the signals (e.g., sparsity and signal strength); and when signals are rare/weak, the optimal FDR level turns out to be much larger than 5%, and in some cases is close to 1. In [42], we studied the optimal level in an asymptotic rare/weak setting, and derived the leading asymptotics of the optimal FDR. In Section 6.2 below we give more detail.

In several papers [2, 86, 87], Strimmer and collaborators compared the approach of feature selection by HCT with both that of control of the FDR and that of control of the False non-Discovery Rate (FNDR), analytically and also with synthetic data and several real data sets on cancer gene microarray. In their papers, they also compared the EBayes approach of Efron [45], which presets an error rate threshold (say, 2.5%), and targets a threshold where the prediction error falls below the desired error rate. Their numerical studies confirm the points explained above: HCT adapts well to different sparsity level and signal strengths, while the methods of controlling FNDR, and EBayes do not perform as well (in misclassification sense); and HC typically selects more false features than other approaches. The goal of the HC feature selection, as we will see, is to optimize the classification error, not to control the FDR. In fact, [86, Table 2] found that HCT had the best classification performance for the cancer microarray data sets they investigated.

3.5 Feature selection by HCT when features are correlated

Above, we assumed the feature vector $X_i \sim N(Y_i \cdot \mu, I_p)$ for $Y_i = \pm 1$. A natural generalization is to assume $X_i \sim N(Y_i \cdot \mu, \Sigma)$, where $\Sigma = \Sigma_{p,p}$ is a unknown covariance matrix. Two problems arise: how to estimate the precision matrix $\Omega = \Sigma^{-1}$ and how to incorporate the estimated precision matrix into the HCT classifier. In the latter, the key is to extend the idea of threshold choice by HCT to the setting where not only the features are correlated, but the covariance matrix is unknown and must be estimated.

The authors of [64] address the first problem by proposing *Partial Correlation Screening (PCS)* as a new row-wise approach to estimating the precision matrix. PCS starts by computing the $p \times p$ empirical scatter matrix $S = (1/n)\sum_{i=1}^{n} X_i X_i'$. Assume the rows of Ω are relatively sparse. To estimate a row of Ω, the algorithm only needs to access relatively few rows of S. For this reason, the method is able cope with much larger p (say, $p = 10^4$) than existing approaches (e.g. Bickel and Levina [13], glasso [49], Neighborhood method [94], and CLIME [24]). [47] addresses the second problem by combining the ideas in Donoho and Jin [41] on threshold choices by HCT with those in Hall and Jin [58] on Innovated HC. This combination injects an estimate of Ω into the HCT classification method; it is asymptotically optimal if Ω is sufficiently sparse and we have a reasonably good estimate of Ω (e.g., [24]).

4 Testing problems about a large covariance matrix

In this section and the next we briefly develop stylized applications of HC to settings which may seem initially far outside the original scope of the idea. In each case, HC requires merely the ability to compute a collection of P-values for a collection of statistics under an intersection null hypothesis. This allows us to easily obtain HC-tests in diverse settings.

Consider a data matrix $X = X_{n,p}$, where the rows of X are iid samples from $N(0, \Sigma)$. We are interested in testing $\Sigma = I_p$ versus the hypothesis that Σ contains a sub-structure. First, we consider the case where the substructure is a small-size clique. In Section 4.1, we approach the testing problem by applying HC to the whole body of pairwise empirical correlations, and to the maximum row-wise correlation (for each variable, this is the maximum of each variable's correlations with all other variables). Second, in Section 4.2, we consider the case where the matrix $\Sigma = I + H$ follows the so-called *spiked covariance model* [83], a low-rank perturbation of the identity matrix. We apply HC to the eigenvalues of the empirical covariance matrix.

4.1 Detecting a possible clique in the covariance matrix

In this section, the global null hypothesis is $\Sigma = I_p$ while the alternative is that Σ contains a small clique. Formally, Σ can be written as $\Sigma = \Gamma \Sigma_0 \Gamma'$, where Γ is a permutation matrix, and for an integer $1 \le k < p$ and $a \in [0, 1)$,

$$\Sigma_0(i,j) = \begin{cases} 1\{i=j\} + a\{i \neq j\}, & \max\{i,j\} \le k, \\ 1\{i=j\}, & \max\{i,j\} > k. \end{cases} \tag{4.12}$$

The parameter a can take negative values as long as Σ_0 remains positive definite.

We suggest two different approaches for detecting the cliques using HC. In each of the two approaches, the key is to obtain P-values.

In the first approach, we obtain individual P-values from pairwise correlations. In detail, write the data matrix $X = X_{n,p}$ as

$$X = [x_1, x_2, \ldots, x_p].$$

The pairwise correlation between the i-th and j-th variable is

$$\rho_{ij} = \frac{(x_i, x_j)}{\|x_i\|\|x_j\|}.$$

Recall that $t_k(0)$ denotes the central Student's t-distribution with $df = k$. The following lemma summarizes some basic properties of ρ_{ij} [105].

Lemma 4.1 *Suppose $\Sigma = I_p$. If $i \neq j$, then for any $\rho \in (-1, 1)$, $P(\rho_{ij} \ge \rho) = P(t_{n-1}(0) \ge \sqrt{n-1}\rho/\sqrt{1-\rho^2})$. Also. If $(i,j) \neq (k,\ell)$, then ρ_{ij} and $\rho_{k\ell}$ are independent.*

This says that the collection of random variables

$$\{\rho_{ij} : 1 \le i \le j \le p\}$$

are pairwise independent (but not jointly independent). It can be further shown that the correlation matrix between different ρ_{ij} is very sparse, so a simple but reasonable approach is to apply OHC to $\{\rho_{ij} : 1 \le i \le j \le p\}$ directly; numerically, the correlation between different ρ_{ij} won't significantly affect the performance of OHC. On the other hand, since the correlation matrix between ρ_{ij} can be calculated explicitly, a slightly more complicated method is to incorporate the correlation structures into HC, following the idea of Innovated HC [59].

In the second approach, we obtain P-values from the maximum correlation in each row:

$$\rho_i^* = \max_{j \neq i} \rho_{ij}, \qquad 1 \le i \le p.$$

Lemma 4.2 *Suppose $\Sigma = I_p$. For $1 \le i \le p$ and $\rho \in (-1, 1)$,*

$$P(\rho_i^* \le \rho) = \left[P\left(t_{n-1}(0) \le \frac{\sqrt{(n-1)}\rho}{\sqrt{1-\rho^2}} \right) \right]^{p-1} \equiv F_{p,n}(\rho).$$

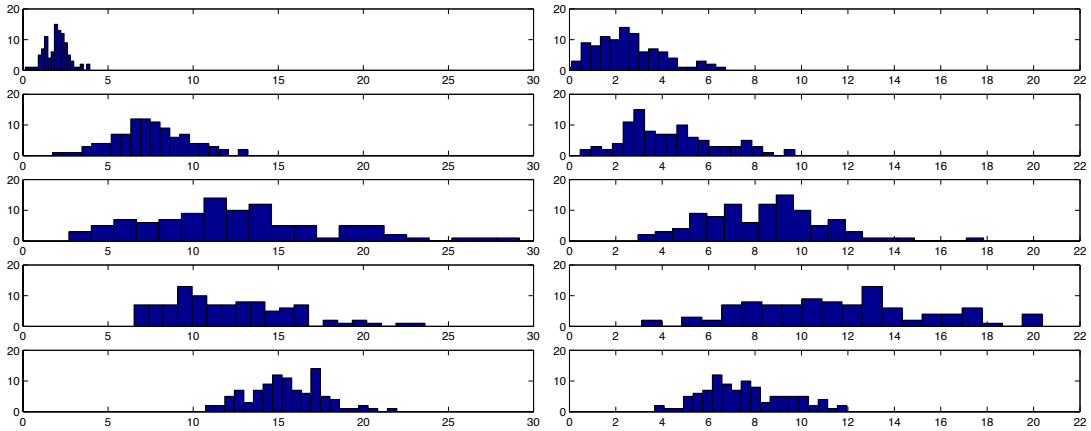

Figure 6: Simulated scores of OHC applied to pairwise correlations (left column) and maximum correlations (right column). The top panel represents the case of no cliques. The others represent cliques with various size and strength (introduced in the text). In comparison, OHC applied to maximum correlations has smaller power than OHC applied to pairwise correlations.

For a proof, see [105] for example. Let $\pi_i^* = F_{p,n}(\rho_i^*)$ so under the global null, $\pi_i^* \sim \text{Unif}(0,1)$. We simply use these P-values in the standard HC framework. [7]

We conducted a small-scale simulation as follows. Fix $(p,n) = (1000, 500)$. We consider 5 different combinations of $(k,a) = (1,0), (5,0.25), (15,0.2), (45,0.1), (135,0.05)$. For each combination, define Σ_0 as in (4.12). Note that for the first combination, $\Sigma = I_p$. Also, since the OHC is permutation invariant, we take $\Gamma = I_p$ for simplicity so that $\Sigma = \Sigma_0$. For each Σ, we generate n samples $X_1, X_2, \ldots, X_n$ from $N(0,\Sigma)$, and obtain ρ_{ij} for all $1 \le i < j \le p$.

In the first approach, we sort all $N = p(p-1)/2$ different P-values $\{\pi_{ij} : 1 \le i < j \le p\}$ in ascending order and write them as follows:

$$\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(N)}, \qquad N = p(p-1)/2.$$

We then apply the Orthodox Higher Criticism (OHC) and obtain the HC score

$$OHC_N^+ = \max_{i:1/N \le \pi_{(i)} \le 1/2} \sqrt{N}[(i/N) - \pi_{(i)}]/\sqrt{\pi_{(i)}(1-\pi_{(i)})}. \tag{4.13}$$

In the second approach, we sort all p different P-values π_j^*, $1 \le j \le p$, in the ascending order and denote them by

$$\pi_{(1)}^* < \pi_{(2)}^* < \ldots < \pi_{(p)}^*.$$

We then apply the HC by (4.13), but with $\pi_{(i)}$ replaced by $\pi_{(i)}^*$.

The histograms of OHC_N^+ based on 100 repetitions are displayed in Figure 6, which suggests that OHC yields satisfactory detection. For all four types of cliques, the OHC applied in the second approach has smaller power in separation than that in the first approach.

4.2 Detecting low rank perturbations of the identity matrix

Now we test whether $\Sigma = I_p$ or instead we have a low-rank perturbation $\Sigma = I + H$, where the rank r of H is relatively small compared to p. [8] Consider the spectral decomposition

$$\Sigma = Q\Lambda Q',$$

[7] π_i^* are equi-correlated: for any $1 \le i \ne j \ne p$, $\text{Cov}(\pi_i^*, \pi_j^*) = c_0(n,p)$ for a small constant $c_0(n,p)$ that does not depend on i or j and can be calculated numerically. It can be shown that $c_0(p,n) = O(1/p)$, and the equi-correlation does not have a major influence asymptotically. Numerical study confirms that correcting for the equi-correlation only has a negligible difference, so we only report results without the correction.

[8] The model $I + H$ is an instance of the so-called spiked covariance model [83]; there are of course hypothesis tests specifically developed for this setting using random matrix theory. We thought it would be interesting to derive what the HC viewpoint offers in this situation.

where Q is a $p \times p$ orthogonal matrix, and Λ is a diagonal matrix, with the first r entries $1 + h_i$, $h_i > 0$, $1 \le i \le r$, and other diagonal entries 1. We assume the eigenbasis Q is unknown to us. In a 'typical' eigenbasis, the coordinates of Q will be "uniformly small", so that even if some of the eigenvalue excesses h_i are nonzero 0, the matrix Σ can be close to the corresponding coordinates of I_p. Therefore, the pairwise covariances may be a very poor tool for diagnosing departure form the null.

Instead we work with the empirical spectral decomposition and apply HC to the sorted empirical eigenvalues. Denote the empirical covariance matrix by

$$S_n = (1/n)X'X,$$

and let

$$\lambda_1 > \lambda_2 > \ldots > \lambda_n$$

be the (nonzero) eigenvalues of S_n arranged in the descending order. The sorted eigenvalues play a role analogous to the sorted P-values in the earlier sections, since the perturbation of I_p by a low-rank matrix H will inflate a small fraction of the empirical eigenvalues, similar to the way the top few order statistics are inflated in the rare/weak model. We define our approximate Z-scores by standardizing each λ_i using its mean and standard deviation under the null. The resulting t-like statistics, which we call the *eigenHC*, is

$$eigenHC_{n,i} = \frac{(\lambda_i - E_0[\lambda_i])}{\mathrm{SD}_0(\lambda_i)}, \qquad 1 \le i \le p, \tag{4.14}$$

where $E_0[\lambda_i]$ and $SD_0(\lambda_i)$ are the mean and standard deviation of λ_i evaluated under the null hypothesis $\Sigma = I_p$, respectively. Note that $E_0[\lambda_i]$ and $SD_0(\lambda_i)$ can be conveniently evaluated by Monte-Carlo simulations. [9]

In Figure 7, we present a realization of $\{eigenHC_{n,i} : 1 \le i \le p\}$ in the case of $n = p = 1000$. The figure looks vaguely similar to realizations of a normalized uniform empirical process, which suggests that the normalization in (4.14) makes sense. We consider the test statistic

$$eigenHC_n^* = \max_{1 \le i \le \alpha_0 n} \{eigenHC_{n,i}\},$$

where α_0 is a tuning parameter we set here to 1/2.

We conducted a small-scale simulation experiment, with $(p, n) = (1000, 1000)$. For each of the 5 different combinations of $(r, h) = (0, 0), (5, 1), (15, 0.5), (45, 0.2), (135, 0.05)$, we let Λ be the $p \times p$ diagonal matrix with first r coordinates equal to $(1 + h)$ and remaining coordinates all 1. We then randomly generated a $p \times p$ orthogonal matrix Q (according to the uniform measure on orthogonal matrices), and set

$$\Sigma = Q\Lambda Q'.$$

Note that when $(r, h) = (0, 0)$, $\Sigma = I_p$. Next, for each Σ, we generated data $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(0, \Sigma)$, and appled $eigenHC_n^*$ to the synthetic data. Simulated results for 100 such synthetic datasets are reported in Figure 7, illustrating that HC can yield satisfactory results even for small r or h. [10]

Testing hypotheses about large covariance matrices has received much attention in recent years. For example, Arias-Castro *et al.* [5] tests that the underlying covariance matrix is the identity versus the alternative where there is a small subset of correlated components. The correlated components may have a certain combinatorial structure known to the statistician.

[9] Since these are the eigenvalues of a standard Wishart matrix, much existing analytic information is applicable. For example, under the null distribution, the top several eigenvalues are dependent and non Gaussian; Johnstone [83] showed that the distribution of the top eigenvalue is Tracy-Widom. Here we don't use such refined information, but only Monte-Carlo simulations.

[10] Our point here is not that HC should replace formal methods using random matrix theory, but instead that HC can be used in structured settings where theory is not yet available. A careful comparison to formal inference using random matrix theory – not possible here – would illustrate the benefits of theoretical analysis of a specific situation – as exemplified by random matrix theory, in this case – over the direct application of a general procedure like HC.

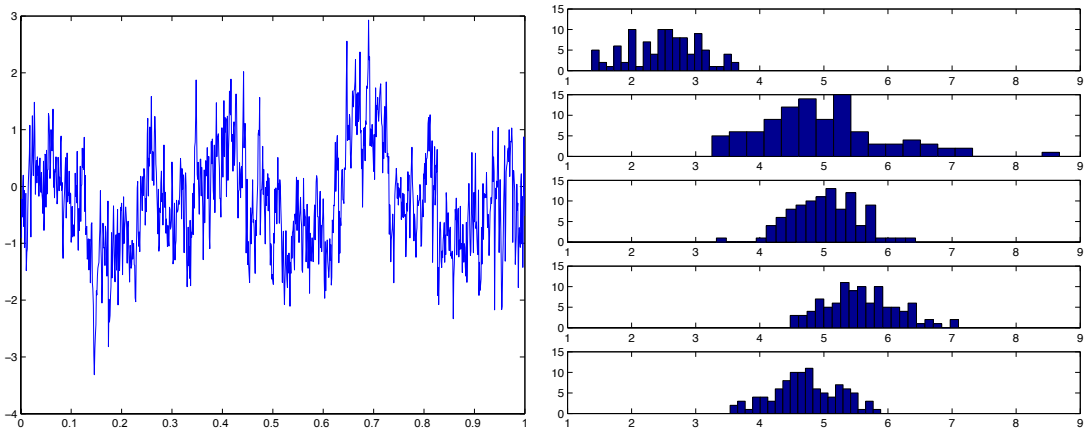

Figure 7: Left column: x-axis is i/n, y-axis is $eigenHC_{n,i}$. Right column: Simulated scores of $eigenHC_n^*$. The top panel represents the unperturbed case. The others represent contamination with different ranks (introduced in text).

Butucea and Ingster [20] consider testing the null model that the coordinates are iid $N(0,1)$, against a rare/weak model where a small fraction of them has significantly nonzero means. Muralidharan [97] is also related; it adapts HC to test column dependences in gene microarray data.

5 Sparse correlated pairs among many uncorrelated pairs

Suppose we observe independent samples (X_i, Y_i), $1 \le i \le n$, from a bivariate distribution with zero means and unit variances, which is generally unknown to us. Under the null hypothesis, the (X_i) are independent of the corresponding (Y_i) (and each other); but under the alternative, for *most* pairs (X_i, Y_i), independence holds, while for a small fraction (X_i, Y_i), the two coordinates may be correlated and each may have an elevated mean. In short, some small collection of the pairs is correlated, unlike the bulk of the data.

Since the underlying distribution of the pairs (X_i, Y_i) is unknown to us, we base our test statistics on ranks (r_i, s_i) of the data (X_i, Y_i). Our strategy is to compare the number of rank-pairs in the upper right corner to the number that would be expected under independence.

For $1 \le k \le n$, let

$$S_k = \#\{1 \le i \le n : \min\{r_i, s_i\} \ge k\} = \sum_{i=1}^{n} 1\{\min\{r_i, s_i\} \ge k\}.$$

Under the null, we have $P(\min\{r_i, s_i\} \ge k) = P(r_i \ge k)P(s_i \ge k) = (1 - k/n)^2$, so

$$E[S_k] = P(r_i \ge k)P(s_i \ge k) = n(1 - k/n)^2,$$

and

$$\mathrm{Var}(S_k) = n[(1 - k/n)^2(1 - (1 - k/n)^2)].$$

Therefore, the HC idea applies as follows. Define

$$pairHC_{n,k} = \sqrt{n}\frac{S_k/n - (1 - k/n)^2}{\sqrt{(1 - k/n)^2(1 - (1 - k/n)^2)}},$$

and

$$pairHC_n^* = \max_{(1-\alpha_0)n \le k \le n} pairHC_{n,k}.$$

Here, α_0 is a tuning parameter and is set to $1/2$ below.

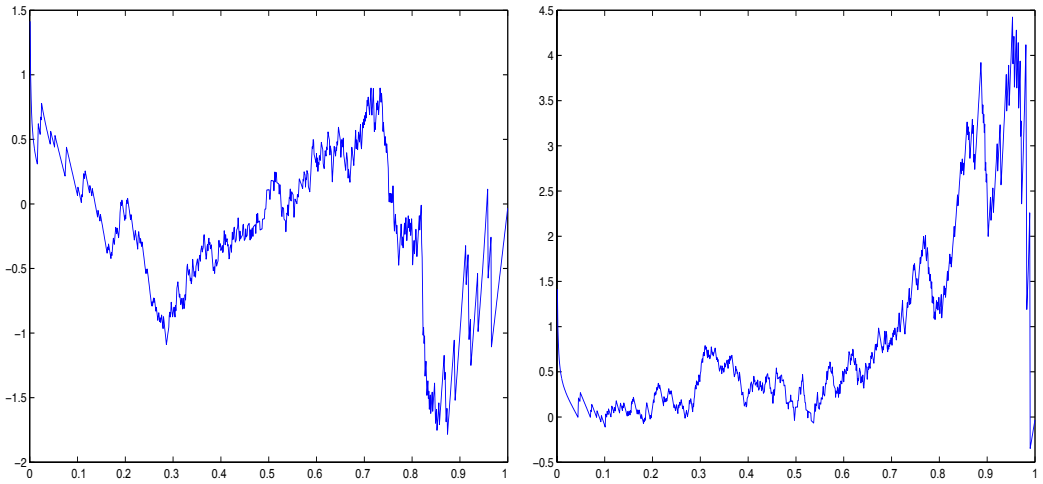

Figure 8: Plot of $pairHC_{n,k}$ versus k/n under the null (left) and under the alternative (right; $(\epsilon, \tau, \rho) = (0.05, 1, 0.25)$). x-axis is k/n ($n = 1000$).

To illustrate this procedure, suppose that (X_i, Y_i), $1 \le i \le n$, are iid samples from a mixture of two bivariate normals

$$(1-\epsilon)N(0, I_2) + \epsilon N(\tau \mathbf{1}_2, \Sigma), \qquad \mathbf{1}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where (ϵ, τ, ρ) are parameters. In Figure 8, we show a plot of $pairHC_{n,k}$ for $n = 1000$ and $k = 1, 2, \ldots, n$ under the null and under the alternative where $(\epsilon, \tau, \rho) = (0.05, 1, 0.25)$.

We conducted a small simulation experiment as follows.

- Fix $n = 1000$ and define 5 different settings where $(\epsilon, \tau, \rho) = (0, 0, 0)$, $(0.02, 0, 2.5)$, $(0.02, 0.50, 2)$, $(0.01, 0.50, 2.5)$, and $(0.01, 0.25, 3)$. Note that the first setting corresponds to the null case.
- Within each setting, conduct 100 Monte Carlo repetitions, each time generating a synthetic dataset with the given parameters and applying $pairHC_n^*$.

The results are reported in Figure 9, which suggests that HC yields good separation even when the signals are relatively rare and weak.

6 Asymptotic Rare/Weak model

In this section, we review the rare/weak signal model and discuss the advantages of HC in this setting.

Return to the problem (2.3)-(2.4) of detecting a sparse Gaussian mixture. We introduce an asymptotic framework which we call the *Asymptotic Rare/Weak* (ARW) model. We consider a sequence of problems, indexed by the number N of P-values (or Z-scores, or other base statistics); in the N-th problem, we again consider mixtures $(1-\epsilon)N(0,1) + \epsilon N(\tau, 1)$, but now we tie the behavior of (ϵ, τ) to N, in order to honor the spirit of the Rare/Weak situation. In detail, let $\vartheta \in (0, 1)$, and set

$$\epsilon = \epsilon_N = N^{-\vartheta},$$

so that, as $N \to \infty$, the non-null effects in $H_1^{(N)}$ become increasingly rare. To counter this effect, we let τ_N tend to ∞ slowly, so that the testing problem is (barely) solvable. In detail, fix $r > 0$, and set

$$\tau_N = \sqrt{2r \log(N)}.$$

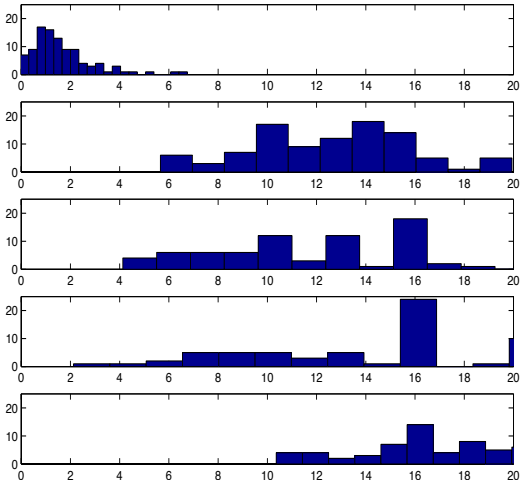

Figure 9: Simulated scores of $pairHC_n^*$. The top panel represents the null case. The others represent the alternative cases with different (ϵ, τ, ρ), introduced in the text ($n = 1000$).

With these assumptions the Rare/Weak setting corresponds to $\vartheta > 1/2$, rare enough that a shift in the overall mean is not detectable, and $r < 1$, weak enough that a shift in the maximum observation is not detectable.

The key phenomenon in this model is a *threshold for detectability*. As a measure of detectability, consider the best possible sum of Type I and Type II errors of the optimal test. Then, there will be a precise threshold separating values of (ϑ, r) where the presence of the mixture is detectable, from those where it is not detectable.

Let

$$\rho(\vartheta) = \begin{cases} \vartheta - 1/2, & 1/2 < \vartheta \le 3/4, \\ (1 - \sqrt{(1-\vartheta)})^2, & 3/4 < \vartheta < 1. \end{cases}$$

When $r > \rho(\vartheta)$, the hypotheses separate asymptotically: the best sum of Type I and Type II errors tends to 0 as N tends to ∞. On the other hand, when $r < \rho(\vartheta)$, the sum of Type I and Type II errors of any test can not get substantially smaller than 1. The result was first proved by Ingster [65, 66], and then independently by Jin [74, 75].

In other words, in the two-dimensional ϑ-r *phase space*, the curve $r = \rho(\vartheta)$ separates the bounded region $\{(\vartheta, r) : 1/2 < \vartheta < 1, 0 < r < 1\}$ into two separate subregions, the *detectable region* and the *undetectable region*. For (ϑ, r) in the interior of the detectable region, two hypotheses separate asymptotically and it is possible to separate them. For (ϑ, r) in the undetectable region, two hypotheses merge asymptotically, and it is impossible to separate them. Hence the phase diagram splits into two 'phases': See Figure 10 for illustration.

Fix (ϑ, r) in the detectable region. Suppose we reject $H_0^{(N)}$ if and only

$$HC_N^* \ge h(N, \alpha_N),$$

where α_N tends to 0 slowly enough so that $h(N, \alpha_N) = O(\sqrt{2\log\log(N)})$. Then when $H_1^{(N)}$ can be detected by the optimal test, HC also detects it, as $N \to \infty$.

Since HC can be applied without knowing the underlying parameter (ϑ, r), we say HC is optimally adaptive. HC thus has an advantage over the Neyman-Pearson likelihood ratio test (LRT), which requires precise information about the underlying ARW parameters. The HC approach can be applied much more generally; it is only for theoretical analysis that we focus on the narrow ARW model.

A similar phase diagram holds in classification problem considered in Section 3, provided that we calibrate the parameters appropriately. Consider a sequence of classification problems indexed by (n, p) where n is the number of observations and p the number of features available to the classifier. Suppose that two classes are equally likely so that $P(Y_i = 1) = P(Y_i = -1) = 1/2$

for all $1 \le i \le n$. For Z in (3.10), recall that $Z \sim N(\sqrt{n}\mu, I_p)$. We calibrate with

$$\sqrt{n}\mu(j) \overset{iid}{\sim} (1-\epsilon)\nu_0 + \epsilon\nu_\tau,$$

where ν_a denotes the point mass at a. Similarly, we use an ARW model, where we fix $(\vartheta, r, \theta) \in (0,1)^3$ and let

$$\epsilon = \epsilon_p = p^{-\vartheta}, \qquad \tau = \tau_p = \sqrt{2r\log(p)}, \qquad n = n_p = p^\theta.$$

Note that when $p \to \infty$, n_p grows with p, but is still much smaller than p. We call such growth *regular growth*. The results below hold for other types of growth of n; see Jin [77] for example.

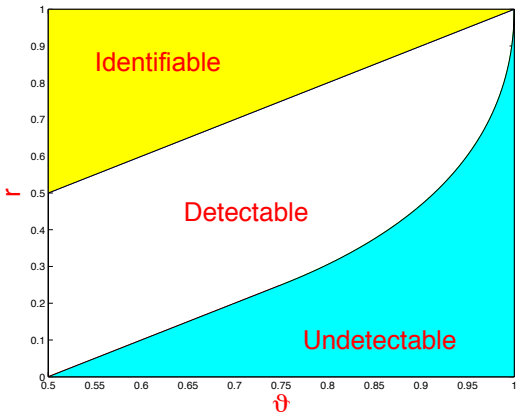

Figure 10: Phase diagram for the detection problem. The detection boundary separates the ϑ-r plane into the detectable region and the undetectable region. In the identifiable region, it is not only able to reliably tell the existence of nonzero coordinates, but also possible to identify individual nonzero coordinates.

It turns out that there is a similar phase diagram associated with the classification problem. Towards this end, define

$$\rho_\theta(\vartheta) = (1-\theta)\rho(\frac{\vartheta}{1-\theta}), \qquad 0 < \vartheta < (1-\theta).$$

Fix $\theta \in (0,1)$. In the two-dimensional phase space, the most interesting region for (ϑ, r) is the rectangular region $\{(\vartheta, r) : 0 < \vartheta, r < (1-\theta)\}$. The region partitions into two subregions.

- (*Region of Success*). If $r > \rho_\theta(\vartheta)$, then the HC threshold $\hat{t}_p^{HC}/t_p^{ideal} \to 1$ in probability; t_p^{ideal} is the ideal threshold that one would choose if the underlying parameters (ϑ, r) are known. Note that HCT is driven by data, without the knowledge of (ϑ, r). Also, the classification error of HCT-trained classification rule tends to 0 as $p \to \infty$.
- (*Region of Failure*). When $r < \rho_\theta(\vartheta)$, the classification error of any trained classification rule tends to $1/2$, as $p \to \infty$.

See Figure 11. The above includes the case where $n_p \to \infty$ but $n_p/p^a \to 0$ for any fixed $a > 0$ as the special case of $\theta = 0$. See more discussion in [41, 77, 42]. Ingster *et al.* [67] derived independently the classification boundary, in a broader setting than that in [41, 77, 42], but they didn't discuss HC.

The conceptual advantage of HC lies in its ability to perform optimally under the ARW framework—without needing to know the underlying ARW parameters: HC is a data-driven non-parametric statistic that is not tied to the idealized model we discussed here, and yet works well in this model.

The phase diagrams above are for settings where the test statistics or measured features X_i are independent normals with unit variances (the normal means may be different). In more complicated settings, how to derive the phase diagrams is an interesting but nontrivial problem.

Delaigle *et al.* [34] studies the problem of detecting sparse mixtures by extending model (2.3)-(2.4) to a setting where X_i are the Student's t-scores based on (possibly) nonGaussian data, where the marginal density of X_i is unknown but is approximately normal. Fan *et al.* [47] extends the classification problem considered in Section 3 to a setting where the measured features are correlated; the covariance matrix is unknown and must be estimated. In general, the approximation errors (either in the underlying marginal density or the estimated covariance matrix) have a negligible effect on the phase diagrams when the true effects are sufficiently sparse and a non-negligible yet subtle effect when the true effects are moderately sparse; see [34, 47].

6.1 Phase diagram in the non-asymptotic context

The phase diagrams depict an asymptotic situation; it is natural to ask how they behave for finite N. This has been studied in [42, Figure 4], Sun [111], Blomberg [14], and [104]. In principle, for finite N, we would not experience "perfectly sharp" phase transition as visualized in Figures 10 and 11. However, numeric studies reveal that for reasonably large N, the transition zone between the region where inference can be rather satisfactory and the region where inference is nearly impossible is comparably narrow, increasingly so as N increases. Sun [111] used such ideas to study a GWAS on Parkinson's disease, and argued that standard designs for GWAS are inefficient in many cases. Xie [122] and Wu [120] used the phase diagram as a framework for sample size and power calculations.

6.2 Phase diagram for FDR-controlling methods

Continuing the discussion in Section 3.4, we investigate the optimal FDR control parameter q in the rare/weak setting. Suppose we select features by applying Benjamini-Hochberg's FDR-controlling method to the ARW. The 'ideal' FDR control parameter $q^{ideal}(\vartheta, r, p)$ is the feature-FDR associated with t_p^{ideal} (i.e., we have a discovery if and only if the feature Z-score exceeds t_p^{ideal} in magnitude). In Donoho and Jin [42], it is shown that as $p \to \infty$,

$$q^{ideal}(\vartheta, r, p) = \begin{cases} o(1), & r > \vartheta, \\ \frac{\vartheta - r}{2r} + o(1), & \vartheta/3 < r < \vartheta, \\ 1 - o(1), & \rho_\theta(\vartheta) < r < \vartheta/3. \end{cases} \tag{6.15}$$

which gives an interesting 3-phase structure: see Region I, II, III in Figure 11. Somewhat surprisingly, the optimal FDR is very close to 1 in one of the three phases (i.e., Region III); in this phase, to obtain optimal classification behavior, we set the feature selection threshold very low so that we include most of the useful features; but when we do this, we necessarily include many useless features, which dominate in numbers among all selected features. Similar comments apply when replacing Benjamini-Hochberg's FDR control by the local FDR (Lfdr) approach of Efron [45]; see [42] for details.

6.3 Phase diagrams in other rare/weak settings

Phase diagrams offer a new criterion for measuring performance in multiple testing in the rare/weak effects model.

This framework is useful in many other settings. Consider a linear regression model $Y = X\beta + z$, $z \sim N(0, I_n)$, where $X = X_{n,p}$ and $p \geq n$. The signals in the coefficient vector β are rare and weak, and the goal is variable selection (different from that in Section 2.10). In a series of papers [52, 73, 82, 84], we use the Hamming error as the loss function for variable selection, and study phase diagrams in settings where the matrices X get increasingly "bad" so the problem get increasingly harder. These studies propose several new variable selection procedures including UPS [73], Graphlet Screening [82], and CASE [84]. The study is closely related to [39] on Compressed Sensing.

[80, 81] present ARW phase diagrams for sparse spectral clustering, and [48] presents phase diagrams for computer privacy and confidentiality.

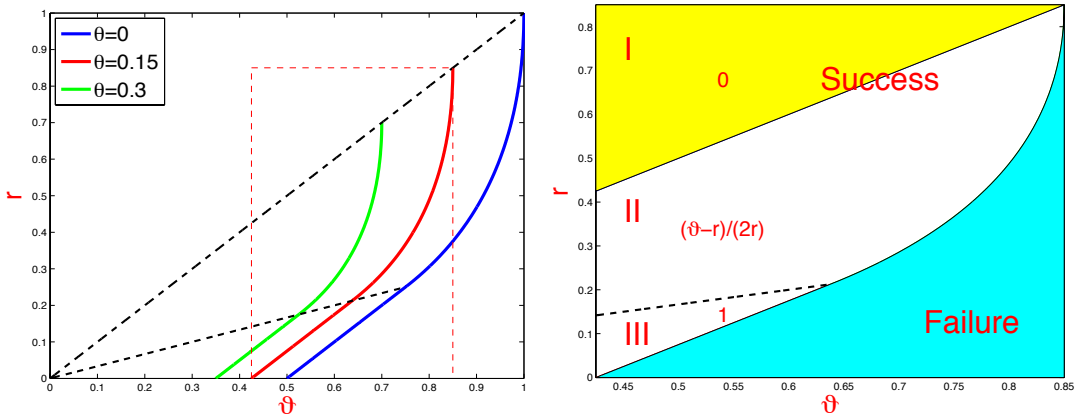

Figure 11: Left: curves $r = \rho_\theta(\vartheta)$ for $\theta = 0, .15, .3$. Two dashed black lines are $r = \vartheta$ and $r = \vartheta/3$, respectively. For $\theta = 0.15$, the most interesting region is represented by the rectangular box. Right: enlargement of the rectangular box. The curve $r = \rho_\theta(\vartheta)$ ($\theta = 0.15$) splits the box into two subregions: Failure (cyan) and Success (white and yellow). The two lines $r = \vartheta$ and $r = \vartheta/3$ further split Region of Success into three subregions, I, II, and III, where the leading terms of $q^{ideal}(\vartheta, r, p)$ in (6.15) are shown. In the yellow region, it is not only possible to have successful classifications, but is also possible to separate useful features from useless ones.

References

[1] Addario-Berry, L., Broutin, N., Devroye, L. and Lugosi, G. (2010). On combinatorial testing problems. *Ann. Statist.* **38** 3063–3092.

[2] Ahdesmaki, M. and Strimmer, K. (2010). Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *Ann. Appl. Stat.* **4**(1) 503–519.

[3] Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. *Wired Magazine* 16.07 June 23 2008.

[4] Anderson, T.W. and Darling, D.A. (1952). Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *Ann. Math. Statist.* **23** 193–212.

[5] Arias-Castro, E., Candes, E. and Durand, A. (2011). Detection of an anomalous cluster in a network *Ann. Statist.* **39**(1) 278–304.

[6] Arias-Castro, E., Candes, E. and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism *Ann. Statist.* **39**(5) 2533–2556.

[7] Arias-Castro, E. and Wang, M. (2013). Distribution free tests for sparse heterogeneous mixtures. *arXiv:1308.0346*.

[8] Balabdaoui, F., Jankowski, H., Pavlides, M., Seregin, A. and Wellner, J. (2011). On the Grenander estimator at zero. *Stat. Sinica* **21** 873–899.

[9] Benjamini, Y. (2010). Discovering the false discovery rate. *J. R. Statist. Soc. B.* **72** 405–416.

[10] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B* **57** 289–300.

[11] Bennett, M.F., Melatos, A., Delaigle, A. and Hall. P. (2012). Reanalysis of F-statistics gravitational-wave search with the higher criticism statistics. *Astrophys. J.* **766**(99) 1-10.

[12] Berk, R.H. and Jones, D.H. (1979). Goodness-of-fit test statistics that dominates the Kolmogorov statistic. *Z. Wahrsch. verw. Geb.* **47** 47–59.

[13] Bickel, P. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227.

[14] Blomberg, N. (2012). Higher criticism testing for signal detection in rare and weak models. *Master Thesis, KTH Royal Institute of Technology, Stockholm, Sweden.*

[15] Bogdan, M., Chakrabarti, A., Frommlet, F. and Ghosh, J. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *Ann. Statist.* **39**(3) 1551–1579.

[16] Bogdan, M., Ghosh, J. and Tokdar, T. (2008). A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* **1** 211–230.

[17] Box, M. and Meyer, D. (1986). An analysis for unreplicated fractional factorials. *Technometrics* **28**, 11–18.

[18] Breiman, L. (2001). Random forests. *Mach. Learn* **24** 5–32.

[19] Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Knowl. Discov. Data Min.* **2** 121–167.

[20] Butucea, C. and Ingster, Y.I. (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, **19**(5B), 2652–2688.

[21] Cai, T., Jeng, J. and Jin, J. (2011). Detecting sparse heterogeneous and heteroscedastic mixtures. *J. Roy. Statist. Soc. B.* **73** 629–662.

[22] Cai, T. and Jin, J. (2010). Optimal rate of convergence of estimating the null density and the proportion of non-null effects in large-scale multiple testing. *Ann. Statist.* **38**(1) 100–145.

[23] Cai, T., Jin, J. and Low, M. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* **35**(6) 2421–2449.

[24] Cai, T., Liu, W. and Luo, X. (2010). A constrained L^1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607.

[25] Cai, T. and Wu, Y. (2012). Optimal detection for sparse mixtures. *arXiv:1211.2265.*

[26] Cayon, L. and Banday, A.J. *et al.* (2006). No Higher Criticism of the Bianchi-corrected Wilkinson Microwave Anisotropy Probe data. *Mon. Not. Roy. Astron. Soc.* **369**(2) 598–602.

[27] Cayon, L., Jin, J. and Treaster, A. (2004). Higher Criticism statistic: detecting and identifying non-Gaussianity in the WMAP first year data. *Mon. Not. Roy. Astron. Soc.* **362** 826–832.

[28] Charbonnier, G. (2012). Inference of gene regulatory network from non independently and identically distributed transcriptomic data. *Ph.D Thesis*, Université d'Évry Val-d'Essonne, France.

[29] Cruz, M., Cayon, L., Martinez-Gonzalez, E., Vielva, P. and Jin, J. (2007). The non-Gaussian cold spot in the 3 year Wilkinson Microwave Anisotropy Probe data. *Astrophys. J.* **655**(1) 11–20.

[30] Dai, H., Charnigo, R., Srivastava, T., Talebizadeh, Z. and Qing. S. (2012). Integrating P-values for Genetic and Genomic Data Analysis. *J. Biom. Biostat.* 2012 3–7.

[31] Dasgupta, A., Lahiri, S. and Stoyanov, J. (2014). Sharp fixed n bounds and asymptotic expansions for the means and the median of a Gaussian sample maximum, and applications to Donoho-Jin model. *Statistical Methodology.* **20**, 40-62.

[32] De la Cruz, O., Wen, X., Ke, B., Song, M. and Nicolae, D. L. (2010). Gene, region and pathway level analyses in whole-genome studies. *Genet. Epidemiol.* **34** 222-231.

[33] Delaigle, A. and Hall, P. (2009). *Higher criticism in the context of unknown distribution, non-independence and classification*. In Perspectives in Math- ematical Sciences I: Probability and Statistics, 109–138 (eds N. Sastry, M. Delampady, B. Rajeev and T.S.S.R.K. Rao. World Scientific.

[34] Delaigle, A., Hall, J. and Jin, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student's t statistic. *J. Roy. Statist. Soc. B.* **73** 283–301.

[35] Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics* **20** 3583–3593.

[36] Dettling, M. and Bühlmann, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics* **19** 1061–1069.

[37] de Una-Alvarez, J. (2012). The Beta-Binomial SGoF method for multiple dependent tests. *Statistical applications in genetics and molecular biology.* **11**(3) 1544-6115.

[38] Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy. *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* 202–210 ACM Press.

[39] Donoho, D. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52**(4) 1289–1306.

[40] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962-994.

[41] Donoho, D. and Jin, J. (2008). Higher Criticism thresholding: optimal feature selection when useful features and rare and weak. *Proc. Natl. Acad. Sci.* **105**(39) 14790–14795.

[42] Donoho, D. and Jin, J. (2009). Feature selection by Higher Criticism thresholding: optimal phase diagram. *Phil. Tran. Roy. Soc. A* **367** 4449–4470.

[43] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104.

[44] Efron, B. (2011). *Large-Scale Inference: Empirical Bayes methods for estimation, testing, and prediction*. IMS Monographs, Cambridge Press.

[45] Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **99** 96–104.

[46] Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Ann. Statist.* **36**(6) 2605–2637.

[47] Fan, Y., Jin, J. and Yao, Z. (2013). Optimal feature selection by Higher Criticism in sparse Gaussian graphic model. *Ann. Statist.*, **41**(5), 2537–2571.

[48] Fienberg, S. and Jin, J. (2012). Privacy-preserving data sharing in high dimensional regression and classification settings. *J. Privacy and Confidentiality* **4**(1) Article 10.

[49] Friedman, J., Hastie, T. and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

[50] Gayraud, G. and Ingster, Y.I. (2011). Detection of sparse variable functions. *arXiv:1011.6369*.

[51] Ge, Y. and Li, X. (2012). Control of the False Discovery Proportion for independently tested null hypotheses. *J. Probab. and Statist.* 2012 Article ID 320425 19 pages.

[52] Genovese, C., Jin, J., Wasserman, L. and Yao, Z. (2012). A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.* **13** 2107–2143.

[53] Goldstein, D.B. (2009). Common genetic variation and human traits. *New England J. Med.* **360**, 1696–1698.

[54] Golub, T. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286** 531–536.

[55] Gontscharuk, V., Landwehr, S. and Finner, H. (2012). On the behavior of local levels of higher criticism tests. *Electron. J. Statist.* 1–27.

[56] Gordon, G.J. *et al.* (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res.* **62** 4963–4967.

[57] Greenshtein, E. and Park, J. (2012). Robust test for detecting a signal in a high dimensional sparse normal vector. *J. Statist. Planning and Inference* **142** 1445–1456.

[58] Hall, P. and Jin, J. (2008). Properties of higher criticism under strong dependence. *Ann. Statist.* **36**(1) 381–402.

[59] Hall, P. and Jin, J. (2009). Innovated Higher Criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38**(3) 1686–1732.

[60] Hall, P., Pittelkow, Y. and Ghosh, M. (2008). Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *J. Roy. Statist. Soc. B* **70** 158–173.

[61] Haupt, J., Castro, R. and Nowak, R. (2008). Adaptive discovery of sparse signals in noise. *Signals, Systems and Computers, 2008 42nd Asilomar Conference.* 1727–1731.

[62] Haupt, J., Castro, R. and Nowak, R. (2010). Improved bounds for sparse recovery from adaptive measurements. *Information Theory Proceedings (ISIT), 2010* 1565–1567.

[63] He, S. and Wu, Z. (2011). Gene-based Higher Criticism methods for large-scale exonic single-nucleotide polymorphism data. *BMC Proceedings 5* (Suppl 9):S65.

[64] Huang, S. and Jin, J. (2014). Partial correlation screening for estimating large precision matrix, with applications to classifications. *Manusrcipt.*

[65] Ingster, Y.I. (1997). Some problems of hypothesis testing leading to infinitely divisible distribution. *Math. Methods Statist.* **6** 47–69.

[66] Ingster, Y.I. (1999). Minimax detection of a signal for l_n^p-balls. *Math. Methods Statist.* **7** 401–428.

[67] Ingster, Y.I., Pouet, C. and Tsybakov, A. (2009). Classification of sparse high-dimensional vectors. *Phil. Trans. R. Soc. A* **367** 4427–4448.

[68] Ingster, Y.I., Tsybakov, A. and Verzelen, N. (2010). Detection boundary in sparse regression. *Electron. J. Statist.* **4** 1476–1526.

[69] Jager, L. and Wellner, J.A. (2004). On the "Poisson boundaries" of the family of weighted Kolmogorov statistics. *IMS Monograph* **45** 319–331.

[70] Jager, L. and Wellner, J.A. (2007). Goodness-of-fit tests via phi-divergences. *Ann. Statist.* **35**(5) 2018–2053.

[71] Jeng, J., Cai, T. and Li, H. (2010). Optimal sparse segment identification with application in copy number variation analysis. *J. Amer. Statist. Assoc.* **105**(491) 1156–1166.

[72] Jeng, J., Cai, T. and Li, H. (2013). Simultaneous discovery of rare and common segment variants. *Biometrika* **100**(1) 157–172.

[73] Ji, P. and Jin, J. (2011). UPS delivers optimal phase diagram in high dimensional variable selection. *Ann. Statist.* **40**(1) 73-103.

[74] Jin, J. (2003). Detecting and estimating sparse mixtures. *Ph.D. Thesis*, Department of Statistics, Stanford University.

[75] Jin, J. (2004). Detecting a target in very noisy data from multiple looks. *IMS Monograph* **45** 255–286.

[76] Jin, J. (2007). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *J. Roy. Statist. Soc.* **70**(3) 461–493.

[77] Jin, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proc. Natl. Acad. Sci.* **106**(22) 8859–9964.

[78] Jin, J. and Cai, T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506.

[79] Jin, J., Starck, J.-L., Donoho, D., Aghanim, N. and Forni, O. (2005). Cosmological non-gaussian signature detection: Comparing performance of different statistical tests. *EURASIP J. Appl. Signal Processing* **15** 2470–2485.

[80] Jin, J. and Wang, W. (2014) Important features Principle Component Analysis for high-dimensional clustering. *arXiv:1407.5241.*

[81] Jin, J. and Wang, W. (2014) Optimal feature selection for Important Features PCA in high dimensional clustering. *Manuscript.*

[82] Jin, J., Zhang, C.-H. and Zhang, Q. (2014). Optimality of Graphlet Screening in high dimensional variable selection. *J. Mach. Learn. Res.*, to appear.

[83] Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29**(2), 295–327.

[84] Ke, T., Jin, J. and Fan, J. (2014). Covariate assisted screening and estimation. *Ann. Statist.*, to appear.

[85] Kendall, D.G. and Kendall, W.S. (1980). Alignments in two-dimensional random sets of points. *Advances in Applied Probability*, **12**, 380-424.

[86] Klaus, B. and Strimmer, K. (2010). Thresholding for feature selection in genomic: Higher Criticism versus False Non-discovery Rate. *Proceedings of the 7th International Workshop on Computational Systems Biology, WCSB 2010 (June 16-18, 2010, Luxembourg, Luxembourg)* 59–62.

[87] Klaus, B. and Strimmer, K. (2013). Signal identification for rare and weak features: Higher Criticism and False Discovery Rate. *Biostat.*, **14**(1), 129-143.

[88] Krista, I., Strohmer, T. and Wertz, T. (2014). Localization of matrix factorizations. *Found.Comput.Math.*, to appear.

[89] Laurent, B., Marteau, C. and Maugis-Rabusseau, C. (2013). Non-asymptotic detection of two-component mixture with unknown means. *aiXiv:1304.6924.*

[90] Li, J. and Siegmund, D. (2014). Higher Criticism: p-values and criticism. *Manuscript.*

[91] Liu, W. and Shao. Q.M. (2013). A Cramér Rao moderate deviation theorem for Hotelling's T^2-statistic with applications to global tests. *Ann. Statist.* **41**(1) 296–322.

[92] Martin, L., Gao, G., Kang, G., Fang, Y. and Woo, J. (2009). Improving the signal-to-noise ratio in genome-wide association studies. *Genetic Epidemiology* **33** Suppl 1 29–32.

[93] Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika* **95**(2) 265–278.

[94] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs with the lasso. *Ann. Statist.* **34** 1436–1462.

[95] Meinshausen, N. and Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* **34**(1) 373–393.

[96] Mukherjee, R., Pillai, N. and Lin, X. (2013). Hypothesis testing for sparse binary regression. *arXiv:1308.0764*.

[97] Muralidharan, O. (2010). Detecting column dependence when rows are correlated and estimating the strength of the row correlation. *Electron. J. Statist.* **4** 1527–1546.

[98] Neill, D. (2006). Detection of spatial and spatio-temporal Clusters. *Ph.D Thesis*, School of Computer Science, Carnegie Mellon University.

[99] Neill, D. and Lingwall, J. (2007). A nonparametric scan statistic for multivariate disease surveillance. *Advances in Disease Surveillance* **4** 106-106.

[100] Park, J. and Ghosh, J. (2010). A guided random walk through some high dimensional problems. *Sankhya* **72-A** 81–100.

[101] Parkhomenko, E., Tritchler, D. and Lemire, M. *et al.* (2009). Using a higher criticism statistic to detect modest effects in a genome-wide study of rheumatoid arthritis. *BMC Proceedings* **3** (Suppl 7):S40.

[102] Pires, S., Starck, J.-L., Amara, A., Refregier, A. and Teyssier, R. (2009). Cosmological models discrimination with Weak Lensing. *Astron. Astrophys.* **505** 969–979.

[103] Roeder, K. and Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statist. Sci.* **24**(4) 398–413.

[104] Rohban, M.H., Ishwar, P., Orteny, P., Karl, W.C. and Saligrama, V. (2013). An impossibility result for high dimensional supervised learning. *IEEE Information Theory Workshop, 2013*, to appear.

[105] Ruben H (1960) Probability content of regions under spherical normal distribution, I. *Ann. Statist.* **31**(3), 598–618.

[106] Sabatti, C., Service, S. and Hartikainen, A. L. *et al.* (2008). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics* **41** 35–46.

[107] Saligrama, V. and Zhao, M. (2012). Local anomaly detection. *AISTATS 2012*.

[108] Sharpnack, J. and Singh, A. (2010). Identifying graph-structured activation patterns in networks. In *Neural Information Processing Systems (NIPS) 2010*.

[109] Shorack, G. and Wellner, J. (1986). *Empirical processes with applications to statistics*. John Wiley & Sons.

[110] Suleiman, R.F.R., Mary, D. and Ferrari, A. (2013). Minimax sparse detection based on one-class classifiers. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 5553-5557.

[111] Sun, L. (2011). On the efficiency of genome-wide scans: a multiple hypothesis testing perspective. *U.P.B. Sci. Bull., Series A* **73**(1) 19–26.

[112] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci.* **99** 6567–6572.

[113] Tukey, J.W. (1976). T13 N: The Higher Criticism. *Course notes, Stat 411, 1976. Princeton University.*

[114] Tukey, J.W. (1989). Higher Criticism for individual significances in several tables or parts of tables. *Internal working paper, Princeton University.*

[115] Tukey, J.W. (1994). *The problem of Multiple Comparisons in The Collected Works of John W. Tukey.* Vol. III, 1948–1983, Edited by Henry I. Braun. (Original manuscript 1953.) Chapman & Hall.

[116] Vielva, P. (2010). A comprehensive overview of the cold spot. *Advances in Astronomy.* 2010 Article ID 592094 20 pages.

[117] Walther, G. (2013). The average likelihood ratio for large-scale multiple testing and detecting sparse mixtures. *IMS Collections. From probability to statistics and back: High dimensional models and processes* **9** 317–326.

[118] Wellner, J.A. and Koltchinskii, V. (2003). A note on the asymptotic distribution of Berk-Jones type statistics under the null hypothesis. In *High Dimensional Probability III* (J. Hoffmann-Jorgensen, M. B. Marcus and J.A. Wellner, eds.) Birkhäuser, Basel.

[119] Wehrens, R. and Franceschi, P. (2012). Thresholding for biomarker selection in multivariate data using Higher Criticism. *Mol. BioSyst.* **8**(9) 2339–2346.

[120] Wu, M., Sanchez, B.N. and Song, P. (2013). Study design in high-dimensional classification analysis. *Manuscript.*

[121] Wu, Z., Sun, Y., He, S., Choy, J., Zhao, H. and Jin, J. (2012). Detection boundary and Higher Criticism approach for sparse and weak genetic effects. *Ann. Appl. Statist.*, to appear.

[122] Xie, J., Cai, T. and Li, H. (2011). Sample size and power analysis for sparse signal recovery in genome-wide association studies. *Biometrika* **98** 273-290.

[123] Zhong, P., Chen, S. and Xu, M. (2013). Test alternative to higher criticism for high dimensional means under sparsity and column-wise dependence. *Ann. Statist.* **41**(6) 2820-2851.