Large scale inference under sparse and weak alternatives: non-asymptotic phase diagram for CsCsHM statistics

OSKAR STATTIN

Master in Applied and Computational Mathematics
Date: June 27, 2017
Supervisor: Tatjana Pavlenko
Examiner: Tatjana Pavlenko
Swedish title: Storskalig inferens för glesa och svaga alternativ: icke-asymptotiska fasdiagram för CsCsHM-testvariabler
School of Engineering Sciences

Abstract

Modern mätteknologi tillåter att generera och lagra gigantiska mängder data, varav en stor andel är redundanta och varav bara ett fåtal är användbara för ett givet problem. Områden där detta är vanligt är till exempel inom genomik, proteomik och astronomi, där stora multipla test ofta behöver utföras, med förväntan om endast några fåsignifikanta effekter. Ett antal nya testprocedurer har utvecklats för att testa dessa så-kallade svaga och glesa effekter i storskalig statistisk inferens. Den mest populära av dessa är troligen Higher Criticism, HC (se Donoho och Jin (2004)). En ny klass av goodness-of-fit-testvariabel döpt CsCsHM har nyligen blivit härledd (se Stepanova och Pavlenko (2017)) för samma typ av multipla testscenarion och har bevisat bättre asymptotiska egenskaper än den traditionella HC-metoden.

Den här rapporten utforskar det empiriska beteendet för båda testmetodikerna i närheten av detektionsgränsen, vilken är tröskeln för detektion av glesa och svaga effekter. Den här teoretiska, skarpa gränsen delar fasrymden, vilken är uppspänd av gleshets- och svaghetsparametrarna, i två delområden: det detektionsbara och det icke-detektionsbara området. Testsvariablernas metodik tillämpas även för variabelselektion för storskalig binär klassificering. Dessa tillämpas, förutom simuleringar, på riktig data. Resultaten pekar på att testvariablerna är jämförbara i prestation.

Abstract

High-throughput measurement technology allows to generate and store huge amounts of features, of which very few can be useful for any one single problem at hand. Examples include genomics, proteomics and astronomy, where massive multiple testing often needs to be performed, expecting a few significant effects and essentially a null background. A number of new test procedures have been developed for detecting these, so-called sparse and weak effects, in large scale statistical inference. The most widely used is Higher Criticism, HC (see e.g. Donoho and Jin (2004)). A new class of goodness-of-fit test statistics, called CsCsHM, has recently been derived (see Stepanova and Pavlenko (2017)) for the same type of multiple testing, it is shown to achieve better asymptotic properties than the traditional HC approach.

This report empirically investigates the behavior of both test procedures in the neighborhood of the detection boundary, i.e. the threshold for the detectability of sparse and weak effects. This theoretical boundary sharply separates the phase space, spanned by the sparsity and weakness parameters, into two subregions: the region of detectability and the region of undetectability. The statistics are also applied and compared for both methodologies for features selection in high dimensional binary classification problems. Besides the study of the methods and simulations, applications of both methods on realistic data are carried out. It is found that the statistics are comparable in performance accuracy.

Contents

Chapter 1

Introduction

As storage grows cheaper and data exponentially bigger, fast computational methods are needed to keep up. When acquiring data it is, with the availability of large storage and modern high throughput measurement technology, easy to measure a large number of variables without necessarily expecting a signal in any but a few of them. This phenomenon is sometimes called *data glut* and gives rise to a new kind of problem of massive multiple testing for significances against a null background. For some data types, such as microarrays of DNA or proteins, which are naturally huge in dimension, sometimes having as many as $\sim 10^6$ entries, this is an essential part of the analysis.

When testing multiple hypotheses, i.e that we have a non-zero signal in any of the variables, we need to take into consideration the possibility that one or several of the variables are significant as by chance, and not by merit. John Tukey suggested a testing method on a *meta level* or a *higher level*, which he called Higher Criticism (HC) to account for this phenomenon. Its purpose was to test the overall body of tests for a fixed significance. This idea was the basis for the HC type statistic proposed by Donoho and Jin in 2004 [1].

HC has since evolved and today there exists many different adaptations for various applications, two of these applications are signal detection [1] [2] and feature selection for classification [3] [4] [5]. See the symposium by Donoho and Jin for a comprehensive overview [6]. The HC framework has proved especially useful for very large data, because of its cheap computational properties, light overhead and statistical interpretability. It is a non-parametric method so it requires little to no tuning.

However there is a known problem with the asymptotics of HC, where the test statistic goes to infinity in probability as the sample size goes to infinity under H_0. To amend this problem Stepanova and Pavlenko used results from studies of empirical weighted processes to suggest a new statistic which has better asymptotical properties, see [2]. This new statistic is called after the mathematicians Csörgos, Csörgos, Horváth and Mason, abbreviated CsCsHM.

1.1 Problem statement

In this report we aim to answer how CsCsHM behaves for finite sample size in comparison to HC. This is done in a way inspired by Blomberg [7], with the the primary tool being heat maps of an empirical error measure over the phase space spanned by the sparsity and weakness parameters.

Two scenarios are considered: signal detection and variable selection for binary classification. The second is in some way an extension of the former, for which CsCsHM is adapted. Besides simulations, applications of both HC and CsCsHM on realistic data are performed for variable selection for classification.

1.2 Outline

This report consists of three main parts, the chapter 2: Background where the theoretical justifications for the work is presented alongside some orienting results on related topics. After this starting point, in the chapter 3: Numerical study the simulations and experiments performed are presented with methods and results. Finally the results are discussed and some conclusions are drawn in chapter 4: Discussion.

Chapter 2

Background

In this section an introduction to the theoretical results and definitions underlying the experiments are presented. For a more thorough background, see the cited sources, especially the works by Donoho and Jin.

2.1 Higher criticism in signal detection

Signal detection is the task of finding whether there exists a signal among the noise in a dataset. This can be formalized as a hypothesis testing scenario where the null hypothesis H_0 is that there is no signal, so the distribution consists of only noise. The alternative hypothesis H_1 is that a small fraction of the signals come from another separate distribution.

2.1.1 A general framework for signal detection

In the introduction it was mentioned that HC is used in scenarios where we have weak and sparse signals. This is formalized as the rare and weak (RW) framework which is based on two hypotheses.

The first hypothesis is of effect sparsity. It states that only a few of the observed signals are expected to differ from a global null hypothesis that everything is noise, meaning that the true signals are rare. The second hypothesis is of effect weakness, and states that the signals are hard to detect since they are weak.

Together these two hypotheses give us the challenging situation of having only a few informative signals among many non-informative, and these signals are hard to distinguish from each other.

This framework can be incorporated into a mixture model of two distributions F and G, where the former is the non-informative, and the latter is the informative distribution. This can be written as

$$X \sim (1-\epsilon)F + \epsilon G, \tag{2.1}$$

meaning that a vector $X = (X_1, \ldots, X_n)$ has each component independently and identically distributed drawn from one of the two distributions. The number of components of the feature vector, n, is large. The epsilon denotes the fraction of the components which are informative $\epsilon \in (0, 1)$, and is according to the sparsity hypothesis small.

For our purposes we will look at a mixture of the same type of distribution, where G has a slightly shifted mean but the same variance. Using the notation from 2.1 and integrating the mixture model into the hypothesis testing of signal detection, it can be written as

$$H_0 : X_i \sim F, \tag{2.2}$$
$$H_1^{(n)} : X_i \sim (1-\epsilon)F + \epsilon G. \tag{2.3}$$

By specifying the distributions of F and G as normal distributions, we arrive at the n individual tests,

$$H_{0,i} : X_i \sim \mathcal{N}(0,\ 1), \tag{2.4}$$
$$H_{1,i} : X_i \sim \mathcal{N}(\mu_0,\ 1), \quad \mu_0 > 0. \tag{2.5}$$

Here the null hypothesis H_0 is that there are no signals among all the signals, so the intersection of $H_{0,1} \cap H_{0,2} \cap \cdots \cap H_{0,n}$. The alternative is that there are signals, so that some $H_{1,i}$ is true. We write this as

$$H_0 : X_i \sim \mathcal{N}(0,\ 1), \tag{2.6}$$
$$H_1^{(n)} : X_i \sim (1-\epsilon)\mathcal{N}(0,\ 1) + \epsilon\mathcal{N}(\mu_0,\ 1), \tag{2.7}$$

where μ_0 is a small shift in mean and ϵ is as above. Further we assume that the signals are independently and identically distributed, and share a common amplitude and variance. In this framework, which we call $RW(\epsilon, \mu_0)$ we can test the null hypothesis, that the observations only consist of noise.

The phase space of β and r

A natural way to investigate the results of signal detection is to look at the borderland where the signals are so weak or sparse that it becomes nearly impossible to separate signal from noise. However the parametrization of $RW(\epsilon, \mu_0)$ makes for awkward values in the area close to where detection is impossible. Introducing the sparsity parameter β and strength parameter r in the following way allows for a more handy phase space on the domain $(0,1) \times (0,1)$.

The sparsity parameter β is defined by

$$\epsilon = n^{-\beta}. \tag{2.8}$$

This parameter leads to a natural partition of a dense region where $0 < \beta \leq 0.5$ and a sparse region where $0.5 < \beta < 1$. In these two regions the strength parameter is chosen differently, since different growth rates are interesting for the two paradigms. The strength parameter r is defined by

$$\mu_0 = \begin{cases} n^{-r} & 0 < \beta \leq 0.5, \quad 0 < r < 0.5, \\ \sqrt{2r\log(n)} & 0.5 < \beta < 1, \quad 0 < r < 1. \end{cases} \tag{2.9}$$

This gives us a region of undefined quadratical space in the dense region $0 < \beta < 0.5$ where $0.5 < r < 1$, where the signals are too weak with this parametrization to be interesting. See Figure 2.1 below for a visualization of the different areas.

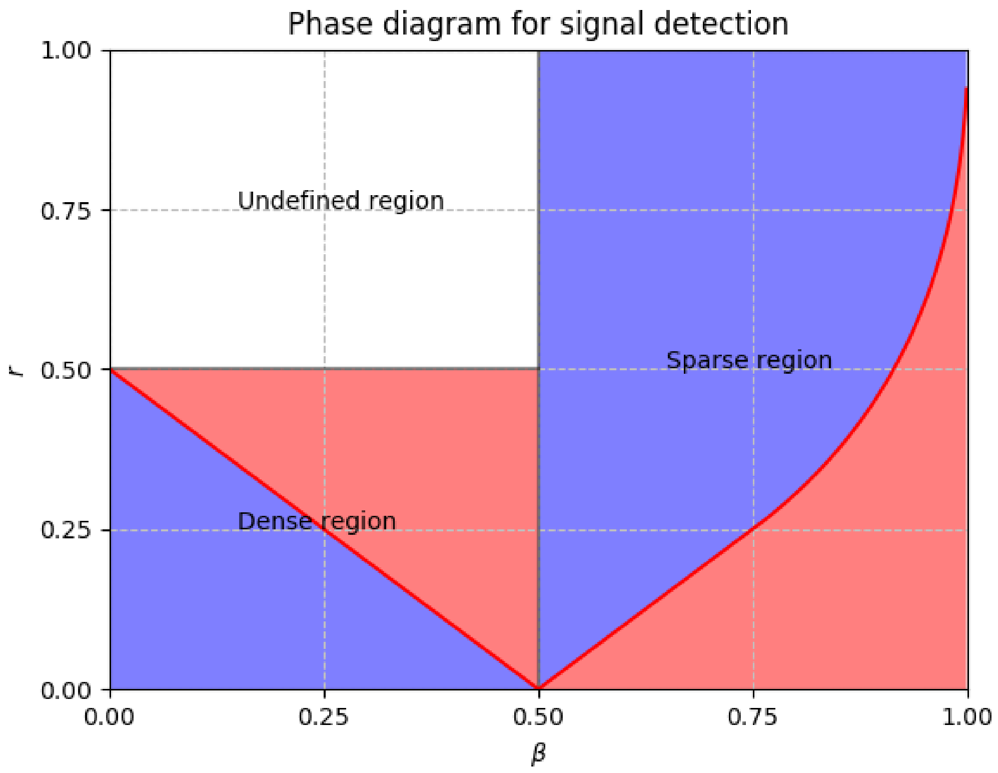

Figure 2.1: The phase diagram for signal detection with the detection boundary for the dense and sparse areas. Blue denotes the area of detectability, and red the area of non-detectability.

The detection boundary

In the phase space spanned by these two parameters a theoretical limit called the *detection boundary*, deduced by Ingester [8], separates the space into two regions: a region of undetectability and a region of detectability. Detection is impossible below and possible above this threshold, why the areas are sometimes called the regions of failure and success respectively.

For the normal mixture case the detection boundary has been proven to be

$$\rho^*(\beta) = \begin{cases} 1-\beta & 0 < \beta \leq 0.5, \\ \beta - 0.5 & 0.5 < \beta < 0.75, \\ 1-(1-\sqrt{\beta})^2 & 0.75 < \beta < 1. \end{cases} \tag{2.10}$$

Detection is thus possible wherever this boundary is exceeded by r, i.e $r > \rho^*(\beta)$, see [1].

2.1.2 Higher criticism

Donoho and Jin suggested in 2004 a procedure called "Higher Criticism", or HC, for testing in the RW-scenario of normal distributions [1]. Inspired by and named after John Tukey's ideas of a higher-order testing, the motivation for the proposed statistic was to compare the expected number of significant tests to the actual number. Instead of looking at individual tests' significance, the significance of the overall body of tests is used, making it a *second-order* or *higher* type of testing.

In Tukey's proposed method, one decides upon a significance level, say $\alpha = 0.05$, and the statistic is then formed as

$$HC_{n,0.05} = \sqrt{n}\frac{(\text{Fraction significant at 5\%} - 0.05)}{\sqrt{0.05 \times 0.95}}. \tag{2.11}$$

Donoho and Jin generalized this to include a wider range of significances, over all levels with a selected upper bound α_0, letting their HC-statistic become

$$HC_{n,\alpha} = \sqrt{n}\frac{(\text{Fraction significant at } \alpha - \alpha)}{\sqrt{\alpha \times (1-\alpha)}}. \tag{2.12}$$

If the body of tests is significant, then this statistic is expected to be large for some α, and if it is not, then it is expected to be small for all α:s. Thus the maximum value over all ranges of α was chosen as the test statistic for the significance of the body of tests:

$$HC_n^* = \max_{0<\alpha<\alpha_0} \{HC_{n,\alpha}\}. \tag{2.13}$$

This test statistic is then compared to a critical value, and if this value is exceeded the null hypothesis is rejected and a signal is considered detected.

Under the assumptions detailed in the previous section 2.1.1, a testing procedure for signal detection in the case of a mixture of Gaussians was suggested as follows. For every datapoint x_i in a vector X a p-value π_i is calculated as

$$P(\mathcal{N}(0,1) \geq x_i) = 1 - \Phi(x_i) = \pi_i, \tag{2.14}$$

where Φ is the cumulative density function of the standardised normal distribution. These values are then sorted in ascending order, $\pi_{(1)}, \ldots, \pi_{(p)}$, and are used to define an objective function and a test variable as

$$HC_{n,i} = \sqrt{n}\frac{i/n - \pi_{(i)}}{\sqrt{\pi_{(i)}(1-\pi_{(i)})}}, \quad i = 1, \ldots, n, \tag{2.15}$$

$$HC_n^* = \max_{0<i<\alpha_0 n} \{HC_{n,i}\}. \tag{2.16}$$

The null hypothesis can then be rejected if the test variable HC_n^* exceeds a critical value. This raw form of HC-statistic is referred to as "Orthodox higher criticism" by Donoho and Jin, and several amendments to its heavy tails have been proposed, see below. In Figure 2.2 the results of this procedure has been visualized for simulated data.

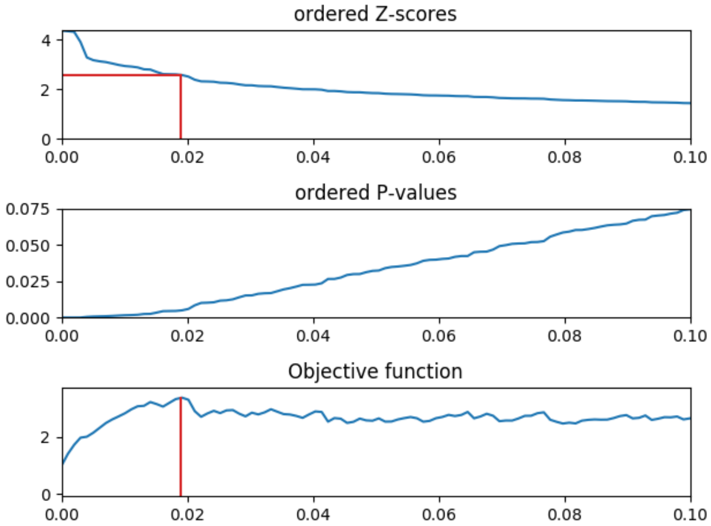

Figure 2.2: The ordered z-scores, p-values and corresponding HC objective function for simulated data. The red line indicates the test variable, and the z-score giving rise to it.

Two attractive properties of this method is firstly that it does not require any information about the parameters ϵ or μ_0 in contrast to other approaches such as likelihood-ratio tests. This is dubbed *optimal adaptivity*. Secondly, the testing is performed at a moderate cost, $\mathcal{O}(n \log n)$.

Properties of higher criticism

To better understand the objective function we can see it as a comparison of the expected fraction of observed significances under the null hypothesis to the actual fraction of observed significances. The nominator thus captures the difference between this expected behaviour of the p-values and their actual behaviour, and is similar to the Kolmogorov-Smirnov (KS) statistic.

A KS test variable K_n over a continuous variable $x \in \mathbb{R}$ is formed as

$$K_n := \sqrt{n} \sup_x |\mathbb{F}_n(x) - F(x)|, \tag{2.17}$$

where $F(x)$ is the theoretical cumulative density function in x, and $\mathbb{F}_n$ is the empirical distribution function (EDF) defined as

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i < x\}. \tag{2.18}$$

Closely related is the goodness-of-fit measure suggested by Anderson and Darling [9] which utilizes a similar structure, but squared instead of absolute valued and with a normalizing function $\psi(x)$,

$$AD_{n,\psi} = n \int_{-\infty}^{\infty} (\mathbb{F}_n(x) - F(x))^2 \psi(F(x)) f(x) dx. \tag{2.19}$$

Here the $\psi(x)$ is introduced as a type of normalizing function, and $f(x)$ is the probability density function in x. Under the null hypothesis we have that $n\mathbb{F}_n \sim \text{Bin}(n, F(x))$, so the variance is known to be $Var(\mathbb{F}_n(x)) = \frac{1}{n}F(x)(1 - F(x))$. With this as normalizing function, and using an L_∞ norm we arrive at the HC statistic with $\alpha_0 = 1$. For details see [10]. HC is thus rooted in similar statistics and not an isolated idea.

Donoho and Jin suggests two different normalizations for HC of this form,

$$HC^{2004} = \max_{\{1<i<\alpha_0 n\}} \sqrt{n}\frac{i/n - \pi_{(i)}}{\sqrt{\pi_{(i)}(1-\pi_{(i)})}}, \tag{2.20}$$

$$HC^{2008} = \max_{\{1<i<\alpha_0 n\}} \sqrt{n}\frac{i/n - \pi_{(i)}}{\sqrt{\frac{i}{n}(1-\frac{i}{n})}}. \tag{2.21}$$

For an approach from the goodness-of-fit angle, see [11] where several different normalizations are considered and analyzed.

Under the global null the expected distribution of the p-values is uniform, so we expect to see $\mathbb{U}_n(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{U_i < u\}$, where $U_i \sim U(0,1)$ being iid random variables, as the EDF. For a sequence of iid random variables $X_1, X_2, \ldots X_n$ with a continuous cumulative distribution function F on $\mathbb{R}$ the EDF can then be approximated as $\mathbb{F}_n(i) = i/n$ for the i:th p-value, according to the assumption that the p-values are uniformly distributed.

It has been noticed that convergence of test statistics on this form $HC(u) = \sqrt{n}(\mathbb{U}_n(u) - u)/\sqrt{u(1-u)}$ depend on their behaviour close to one and zero. Therefore a practical modification is to truncate the lower interval over which the u:s are taken, for example $(1/n, \alpha_0)$. This gives rise to a new test variable

$$HC^{+} = \max_{1/n<u<\alpha_0} HC_n(u), \tag{2.22}$$

where the right hand side can be on the form of any of HC^{2008} or HC^{2004}. However this alleviates the problem for finite samples sizes only, as it has been shown that

$$\lim_{n\to\infty} P(a_n \sup_{0<u<\alpha_0} \sqrt{n}\frac{(\mathbb{U}_n(u) - u)}{\sqrt{u(1-u)}} - b_n \leq x) = e^{-\frac{1}{2}e^{-x}}, \tag{2.23}$$

where $a_n = \sqrt{2\log\log n}$ and $b_n = 2\log\log n + 1/2\log\log\log n - 1/2\log(4\pi)$. The right hand side is the Gumbel distribution. Similar results can be shown for other truncated intervals, see [2]. This means that no matter the chosen interval, the extreme value distribution is reached and the statistic will tend to infinity for large enough n under H_0.

This has motivated research to find other statistics either by a different normalization, like the CsCsHM-statistic [2], or by ordered statistics such as the exact Berk-Jones statistic [10].

Determining the critical value and choice of test statistic

To use HC as a test on significance level α, we need a critical value $h(n, \alpha)$ that satisfies

$$P(HC_n^* > h(n, \alpha)|H_0) = \alpha. \tag{2.24}$$

As usual in hypothesis testing, the test statistic is calculated and if it exceeds this critical value, then the null is rejected in favour of the alternative hypothesis. To find this threshold, Theorems 1.1 and 1.2 from Donoho and Jin [1] are used. The former theorem states that under the null hypothesis H_0 the following holds,

$$\frac{HC_n^*}{\sqrt{2 \log \log n}} \xrightarrow{p} 1, \quad n \to \infty. \tag{2.25}$$

The second theorem states that for a sequence of problems indexed by n where $\alpha_n \to 0$ so slowly that $h(n, \alpha_n) = \sqrt{2 \log \log n}(1 + o(1))$, then a HC-test that rejects the null H_0 when $HC_n^* > h(n, \alpha_n)$, has full power when every alternative $H_1^{(n)}$ is defined so that $r > \rho^*(\beta)$. This means that the probability of rejecting the null hypothesis goes to one under every $H_1^{(n)}$ in the region of detectability as n goes to infinity,

$$P(\text{Reject } H_0|H_1^{(n)}) \to 1, \quad n \to \infty. \tag{2.26}$$

The HC test procedure is consistent against all alternatives and has full power in the region of detectability, if the threshold is taken for a fixed α as

$$h(n, \alpha) = \sqrt{2 \log \log n}(1 + o(1)), \tag{2.27}$$

where the $o(1)$ is a buffer since the results are asymptotical. Numerical simulations show that this estimation can be of varying satisfaction. However this means that HC has the property of optimal adaptivity.

The error rate of the HC procedure in the area of detectability is not necessary zero for finite size samples. It has been shown that there is no perfectly sharp boundary where it suddenly becomes impossible to detect signals, but rather a blurry area of transition between success and failure, see [7].

2.1.3 CsCsHM as an alternative testing procedure

A solution to the problem of the HC-statistic converging to an extreme value distribution was proposed by Stepanova and Pavlenko in [2]. By using a normalization with roots in results for weighted empirical processes, the test variable converges to a brownian bridge in distribution as the sample size goes to infinity. The new statistic is named after the mathematicians Csörgő, Csörgő, Horváth and Mason, or CsCsHM for short.

The results are based on what are called *Erdős-Feller-Kolmogorov-Petrovski* (EFKP) upper class functions of Brownian bridges, $\{B(u), 0 \leq u \leq 1\}$, of which an important one is

$$q(u) = \sqrt{u(1-u)\log\log(1/(u(1-u)))}. \tag{2.28}$$

This example comes from Khinchine's local law of the iterated logarithm which states that

$$\limsup_{u\to 0} \frac{W(u)}{\sqrt{u\log\log(1/u)}} \overset{a.s.}{=} \sqrt{2}, \tag{2.29}$$

where $W(u)$ is a standard Wiener process starting at zero. Relating this to the Brownian bridge, $\{B(u), 0 \leq u \leq 1\} \overset{D}{=} \{W(u) - uW(u), 0 \leq u \leq 1\}$, we have that

$$\limsup_{u\to 0} \frac{|B(u)|}{\sqrt{u(1-u)\log\log(1/(u(1-u))}} \overset{a.s.}{=} \sqrt{2}. \tag{2.30}$$

Using these observations as a starting point in the testing scenario $H_0 : F = F_0$ versus the alternative $H_1 : F > F_0$, the following test statistic is suggested

$$T_n^+(q) = \sup_{0<F_0(t)<1} \sqrt{n}\frac{\mathbb{F}_n(t) - F_0(t)}{q(F_0(t))}. \tag{2.31}$$

As detailed in Proposition 3.1 in [2], it is shown that this statistic with slightly different demands on $F_0(t)$ will converge in distribution to a normalized Brownian bridge. If q is an $EFKP$ upper-class function of a Brownian bridge, then under H_0 for any numbers $0 \leq a < b \leq 1$ as $n \to \infty$,

$$\sup_{a<F_0(t)<b} \sqrt{n}\frac{(\mathbb{F}_n(t) - F_0(t))}{q(F_0(t))} \xrightarrow{D} \sup_{a<u<b} \frac{B(u)}{q(u)}. \tag{2.32}$$

This gives us a theoretical justification of truncating the interval over which the statistic is formed, $(0, \alpha_0)$, and the interpretation of this interval as the body of significances which are becomes clearer.

For this setting, a test of asymptotic level α that rejects H_0 when

$$T_n^+(q) \geq t_\alpha^+(q), \tag{2.33}$$

where $t_\alpha^+(q)$ is chosen such that $P(\sup_{0<u<1} B(u)/q(u) \geq t_\alpha^+(q)) = \alpha$. Then for every alternative H_n^i in the area of detectability, i.e $r > \rho(\beta)$, the test based on $T_n^+(q)$ has full power, meaning that

$$P(T_n^+(q) \geq t_\alpha^+(q)|H_n^i) \to 1, \quad n \to \infty. \tag{2.34}$$

This means that asymptotically, the CsCsHM testing procedure will be able to perfectly separate cases where there are signals and where there are not, as long as the strength of the signal exceeds the detection boundary. This means CsCsHM also has the property of optimal adaptivity.

To choose the specific threshold, there are tabulations of the distribution of the random variable $\sup_{0<u<1} B(u)/q(u)$.

Two applied variants of this proposed statistic, with the same testing procedure for signal detection as the one for HC described previously, are

$$CsCsHM_1(\pi_{(i)}) = \frac{\sqrt{n}(i/n - \pi_{(i)})}{\sqrt{\pi_i(1-\pi_{(i)})\log(\log(\frac{1}{\pi_{(i)}(1-\pi_{(i)})}))}} \quad 1 \leq i \leq n, \tag{2.35}$$

$$CsCsHM_2(\pi_{(i)}) = \frac{\sqrt{n}(i/n - \pi_{(i)})}{\sqrt{\pi_i(1-\pi_{(i)})}\log(\log(\frac{1}{\pi_{(i)}(1-\pi_{(i)})}))}, \quad 1 \leq i \leq n, \tag{2.36}$$

where the difference lies in the square root in the denominator. When finding the maximum of these two a restriction of size of the $\pi_{(i)}$:s, similar to the one for HC, can be enforced. For a similar depiction of the z-scores and p-values as previously shown for HC, see Figure 2.3 below.

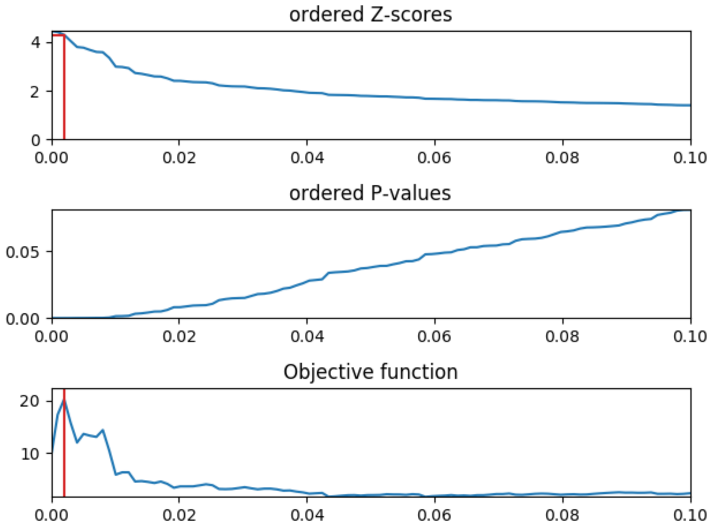

Figure 2.3: The ordered z-scores, p-values and corresponding CsCsHM$_1$ objective function for simulated data. The red line indicates the test variable, and the z-score giving rise to it.

2.2 Variable selection with higher criticism

Two natural extensions of signal detection are to be able to identify variables containing a signal, and to recover these signals. The focus is still on the interesting rare and weak framework with the dimensionality being much larger than the number of samples, $p >> n$. Notice that we revert to the notation of dimension as p and samples as n.

In binary classification we have n samples from a population. Each one of these consists of p dimensions, where each individual sample X_i comes from one of two classes, $Y_i \in \{c_1, c_2\}$. To represent this a matrix of size $(n \times p)$, called $\mathbf{X} = (X_1, \dots, X_n)^T$ is used, where every row $X_i = (x_{i1}, \dots, x_{ip})$ represents one sample. We will assume that each row of $\mathbf{X}$ is independently and identically distributed. The correlation matrix will be assumed to be identity.

The goal of classification is to learn a predictor from training data, that determines the classes of test data as well as possible.

According to the hypotheses of sparsity and weakness only a small number of the p variables will be informative, and among these the contrast mean (the difference in mean between the classes) will be small. Instead of using all the variables when performing classification, which could potentially be very computationally costly, variable selection opts to find the ones that seem to impact the choice of class of the samples the most. To select the variables, the HC framework can be employed.

2.2.1 Phase diagram for classification

While the scenario is similar for the two cases of variable selection and signal detection, the detection boundary will be slightly different for classification. This is because the number of samples will influence the boundary in the phase diagram spanned by (β, r). The effect is depicted in Figure 2.4. The phase diagram will include several regions where success is possible, probable or outright impossible.

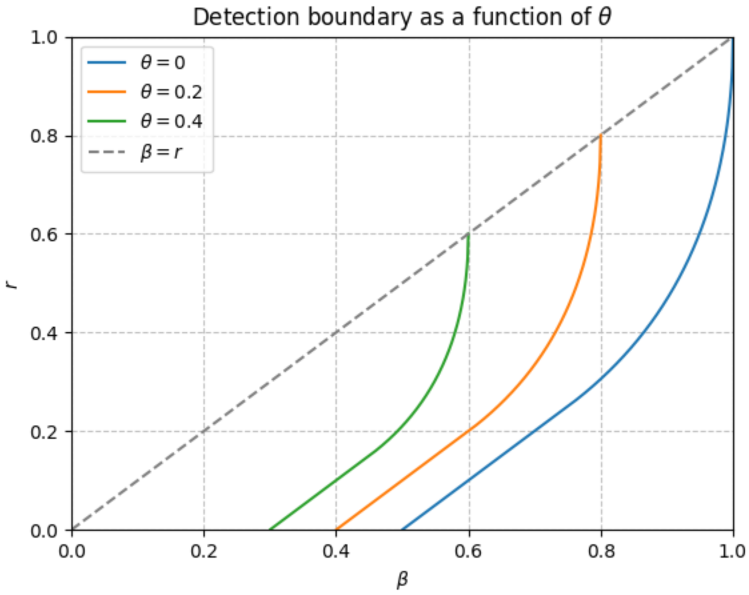

Figure 2.4: The effect of θ on the detection boundary for classification.

The question is whether a trained classifier can successfully classify data given a set of parameters in the phase space. As mentioned the amount of samples n comes into play, so a new parameter θ linking dimensionality and sample size is defined as $n = p^\theta$. We assume balanced data between classes $n_1 = n_2 = p^\theta/2$.

HC thresholding achieves the optimal detection boundary for classification, which can be written for $0 < \beta < 1 - \theta$ as

$$\rho_C^*(\beta) = (1-\theta)\rho^*(\frac{\beta}{1-\theta}), \tag{2.37}$$

as proved by Donoho and Jin [3]. Below this boundary all classifiers' misclassification rate tends to $1/2$, which is equivalent to guessing the labels of the test data. For different growth rates of θ, see [5] for a detailed exposition.

See Figure 2.5 below for a visualization of the different phase diagram regions, as proved by Ji and Jin in [12]. These regions come from a measure of how many of the variables the procedure manages to correctly recover. The lines in this figure in addition to the previous detection boundary are $r = \beta$ and the topmost is $r = (1 + \sqrt{1-\beta})^2$ which demarks the start of the "Exactly recoverable" region. It was reported by Jin and Ke in [13] that Orthodox HC achieves these optimal phase diagrams with hard thresholding.

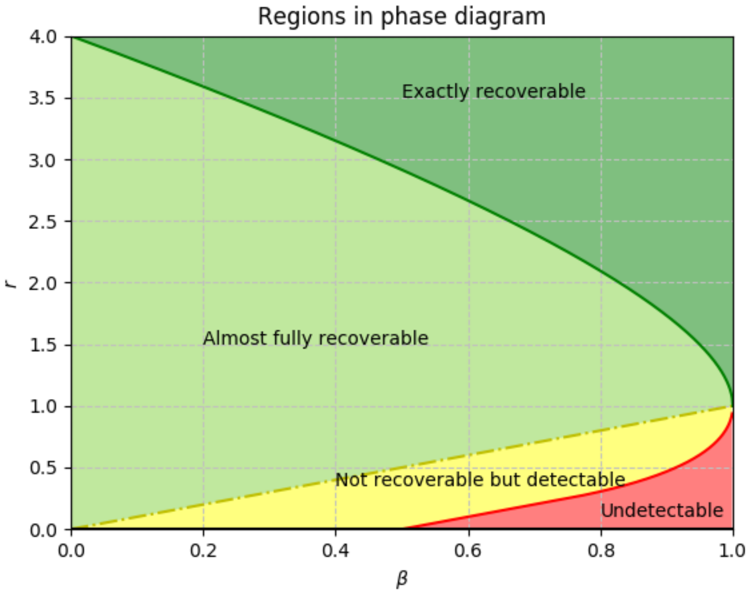

Figure 2.5: Depiction of the regions of varying difficulty for signal recovery.

2.2.2 Higher criticism thresholding

For HC thresholding (HCT), the labels will be defined as $Y \in \{-1, 1\}$, and the data modelled as coming from the following distribution,

$$X_i \sim \mathcal{N}(\mu_i \cdot Y_i, I_n), \tag{2.38}$$

where the covariance matrix I_n is identity, and μ_i is the contrast mean between the two classes. We will also assume that the prior probability for each class is equal.

A type of feature score similar to a t-statistic is formed, defined as

$$Z_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i \cdot x_{ij}), \tag{2.39}$$

where the size of the score will be big in absolute terms if the contrast mean is big, and small otherwise. We obtain two-sided p-values $\pi_i = P(|\mathcal{N}(0,1)| \geq |Z_i|) = 2\Phi(|Z_i|)$ for all the feature scores $\{Z_j\}_{j=1}^{p}$. Similarily to before, we sort these in ascending order, $\pi_{(1)}, \ldots, \pi_{(p)}$, and form the HC objective function for the transformed scores:

$$HC(i, \pi_{(i)}) = \sqrt{p} \frac{i/p - \pi_{(i)}}{\sqrt{i/p(1 - i/p)}}, \quad 1 \leq i \leq p. \tag{2.40}$$

This is the statistic we previously called HC^{2008} in Eq. 2.21 which was proposed by Donoho and Jin in 2008 in [4]. This objective function can, as discussed, be exchanged for another HC-type statistic with some different normalization.

The maximizing index of the objective function is found,

$$\hat{i}^{HC} = \arg\max_{1 \leq i \leq p} HC(i, \pi_{(i)}), \quad \pi_{(i)} \in (\epsilon_s, 0.1) \tag{2.41}$$

where we have introduced a restriction on the size of the p-values. The value of the $Z_{(j)}$ (where we have an ordering in the same way as for the $S_{(i)}$ values) corresponding to the index $\hat{i}^{HC}$ is then taken as the threshold value,

$$\hat{t}_p^{HC} = |Z|_{(\hat{i}^{HC})}. \tag{2.42}$$

Once the threshold is decided, it is used in a threshold function that chooses which variables to consider. A function of this kind is the hard threshold function, defined as $\eta_t^{hard}(z_i) = \text{sgn}(Z_i)\mathbb{1}_{\{|Z_i|>t\}}$.

One easy way to predict the classes when the interesting variables have been selected, is by using a discriminant rule. Then the class of a given sample is given by the sign of $L(X_i)$, given by

$$L(X_i) = \sum_{j=1}^{p} w_t(j) x_{ij}, \tag{2.43}$$

where $w_t(j) = \eta_t^{hard}(z_j)$. This is a case of linear discriminant analysis, LDA, see Hastie, Tibshirani, Friedman [14] and references therein. There are other ways to treat the remaining selected variable but this classifier will serve our purpose.

2.2.3 Variable selection using CsCsHM

Variable selection using the CsCsHM statistic is similar to the HCT procedure. Some slight differences in notation are used: we now call the class labels $Y_i \in \{0, 1\}$, and our decision rule will be formulated differently. In this section we will call the objective function T and the test statistic T^+ for CsCsHM to declutter the notation.

Feature scores $\{Z_i\}_{i=1}^{p}$ are formed as for HCT, but instead of a two-sided p-value, a one-sided transformation is used,

$$S_i = 1 - \Phi(Z_i), \tag{2.44}$$

where $\Phi(t)$ is as before the CDF of a standardised normal distribution. These are then sorted in ascending order as $S_{(1)}, \ldots, S_{(p)}$ and the objective function is taken to be either (2.35) or (2.36), so that we have $\{T(S_{(j)})\}_{j=1}^{p}$.

The maximizing argument of the objective function is then found,

$$S^* = \arg\max_{S_{(i)}} T(S_{(i)}), \quad 1 \leq i \leq p, \tag{2.45}$$

where the test statistic for CsCsHM would be $T^+ = T(S^*)$.

The selection of variables is then decided by the threshold function w_i that is zero or one for every variable. Taking the threshold as $\hat{\tau} = |Z^*|$, where Z^* is the z-score that is used to create the transformed S^*, every variable that exceeds this value is chosen as

$$w_i = \mathbb{1}\{|Z_i| > \hat{\tau}\}, \quad 1 \leq i \leq p. \tag{2.46}$$

This procedure of selecting variables with CsCsHM can be summed up by:

- create z-scores $\{Z_i\}_{i=1}^p$,
- transform these into p-values $S_i = 1 - \Phi(Z_i)$ for $1 \leq i \leq p$ and sort them into ascending order $S_{(1)}, \ldots, S_{(p)}$
- form the objective function $T(S_{(i)}), 1 \leq i \leq p$,
- find the maximizing argument $S^* = \arg\max T(S_{(i)}), 1 \leq i \leq p$ and take its corresponding z-score's absolute value as the threshold $\hat{\tau} = |Z^*|$,
- select variables that exceed the threshold in absolute value, $w_i = \mathbb{1}\{|Z_i| > \hat{\tau}\}$.

The predicted class label of a test data point X_0, with calculated selected variables as in w_i, is given by

$$\Psi(X_0) = \mathbb{1}\{\sum_{i=1:w_i>0}^{p} (x_{0i} - \frac{1}{2}(\tilde{\mu}_{1i} + \tilde{\mu}_{2i}))(\tilde{\mu}_{1i} - \tilde{\mu}_{2i}) \leq 0)\}, \tag{2.47}$$

where $\tilde{\mu}_{ki}$ is the estimated mean of class k for the i:th variable.

2.3 Extensions of higher criticism

There are a wide range of applications for both signal detection and binary classification in the rare and weak framework. Many are found in the fields of genomics and proteomics, where the huge amount of features can limit techniques with high computational costs.

As mentioned previously, the idea of HC is very flexible and has been applied successfully to many different areas. Some extensions that I neglect to address in this report are the big body of work on correlated signals, including innovated HC, and the work on non-gaussian mixtures.

Higher criticism for heterogenous and heteroscedastic mixtures

The HC framework described in this chapter is grounded on the underlying assumption that the informative signals have the same variance, $\sigma = 1$. This could be the case, but it is also likely that the signals will add some type of variance to the table. A formal investigation by Cai and Jin [15] derives the optimal detection boundary for this model.

The model is defined by

$$H_0 : X_i \sim N(0,1), \tag{2.48}$$

$$H_1^{(n)} : X_i \sim (1-\epsilon)N(0,1) + \epsilon N(A,\sigma^2), \quad 1 \leq i \leq n, \tag{2.49}$$

where A and σ are unknown, and all components are assumed iid. With the same type of parametrization as for the homogenous case (A is the same as μ_0), the detection boundary was derived for the sparse case to be,

$$\rho^*(\beta|\sigma) = \begin{cases} (2-\sigma^2)(\beta - 1/2), & 1/2 < \beta \leq 1 - \sigma^2/4, \\ (1-\sigma\sqrt{1-\beta})^2, & 1-\sigma^2/4 < \beta \leq 1, \end{cases} \quad 0 < \sigma < \sqrt{2}, \tag{2.50}$$

and for slightly larger σ,

$$\rho^*(\beta|\sigma) = \begin{cases} 0, & 1/2 < \beta \leq 1 - 1/\sigma^2, \\ (1-\sigma\sqrt{1-\beta})^2, & 1-1/\sigma^2 < \beta \leq 1, \end{cases} \quad \sigma \geq \sqrt{2}. \tag{2.51}$$

This means that the higher the variance of the signals are, the lower the detection boundary is "pushed" into the bottom right of the phase diagram.

Cai and Jin also show that the HC-type testing has full power above the detection boundary and keeps the property of optimal adaptivity for the critical value $\tau = \sqrt{2(1+\delta)\log\log n}$ when the heteroscedastic term is introduced. They suggest using an empirical threshold controlling for the Type I errors because of the slow asymptotics of the double $\log$ term.

Higher criticism for χ^2-mixtures

HC has also been applied to mixtures of χ^2-distributions. One possible such application suggested by Donoho and Jin [1] is covert operations, where the degrees of freedom would be $d = 2$, corresponding to the two parts of the communication.

The χ^2-distribution is also useful when looking at gene-expression data for example, where the genes are expected to be dependent in some type of "blocks", but that these in turn are independent of each other, see Pavlenko et al [16].

This gives us the hypotheses

$$H_0 : X \sim \chi_d^2(0), \tag{2.52}$$

$$H_1 : X \sim (1-\epsilon)\chi_d^2(0) + \epsilon\chi_d^2(\omega_0), \tag{2.53}$$

where ω_0 is the shift and d are the degrees of freedom. Here we assume that each block is independent of each other, have the same size p_0 and informative variables share a common amplitude ω_0^2.

Donho and Jin derived that the detection boundary is the same as in the normal case, see [1], with a region of undetectability underneath the detection boundary and a region of detectability above. HC keeps its property of optimal adaptivity for this scenario.

GWAS and genetic applications

In a genome-wide association study, GWAS, a big set of gene expressions (or single nucleotide polymorphisms, SNP) for different individuals are studied. The hypothesis is that the expression levels are associated with a trait that is often tied to a disease. This way, signal detection can be used to tie genetic factors to diseases, or detect known effects. Classification could possibly be used in preemptive diagnosis of diseases for which the genetic expression is known. The data is collected in microarrays, and these typically have a very large p, possibly in the millions. For a more detailed account of signal detection and variable selection in a genetic setting, see the article by Wu et al [17] and references therein.

Two microarray datasets that have been successfully used for differentiating patients with cancer and without cancer for colon cancer, prostate cancer and leukemia have been described by Alon et al [18] and Golub et al [19] respectively. Both Donoho and Jin [4] as well as Dettling [20] have applied various techniques for classification on these data sets.

Chapter 3

Numerical study

In this section the numerical studies performed are presented. The implementations were made using the `numpy`-library for `python`. For the phase diagram simulations the `mpi4py`-directive was used to parallelize the computations. These simulations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Tegner PDC.

The results of the simulations of the phase diagrams for signal detection are presented in Section 3.1.2, while the results for classification of cancer data as well as phase diagrams are found in Section 3.2.2.

3.1 Signal detection

For signal detection we want to investigate how the CsCsHM-statistic performs in comparison to HC.

3.1.1 Methods

The most interesting area for signal detection is where the task is difficult, but possible, therefore making the phase diagram close to the detection boundary an ideal place for numerical studies. To investigate the empirical finite size behaviour of the procedures for different test statistics, simulations were conducted in the region spanned by (β, r).

Defining an error measure

For the signal detection situation, we need to define an error measure in order to evaluate the empirical properties of the methods at certain points (β, r) in the phase space. We define this error in the same manner as Blomberg did in [7],

$$\widehat{\text{Err}} = \frac{\#H_0\text{falsely rejected} + \#H_1\text{falsely rejected}}{\#\text{simulations}}. \tag{3.1}$$

This error is calculated by simulating equally many cases where one of H_0 or H_1 is true, and noting how many times the correct hypothesis is rejected.

Simulations of H_0 and H_1

For pseudocode of how $\widehat{\text{Err}}$ is calculated for a (β, r) see Algorithm 1 below. Basically m simulations are performed, half of which the null is true and half of which the alternative is true. The test statistic (either HC or CsCsHM$_i$) is calculated for these data and then the error is the sum of all Type I (false positive) and Type II (false negative) errors. Since the threshold is taken from asymptotic theory we allow ourselves a tuning parameter $\delta \in (0, 1)$ which slightly shifts the threshold to minimize the error.

```
Result: $\widehat{\text{Err}}$ for one pair of $(\beta, r)$
for $i = 1 \to m/2$ do
    Simulate data where $H_0$ is true.
    Save $T^i_{H_0}$ = test statistic
end
for $i = 1 \to m/2$ do
    Simulate data where $H_1$ is true.
    Save $T^i_{H_1}$ = test statistic
end
$\widehat{\text{Err}} = \sum_{i=1}^{m/2} \mathbb{1}(T^i_{H_0} > \tau + \delta) + \mathbb{1}(T^i_{H_1} \leq \tau + \delta)$
```

Algorithm 1: Algorithm for heat map simulations. τ is a critical value and δ is chosen such that it minimizes the error.

The calculation of the statistics is done using the built-in algebra of the `numpy` library. Since the iterations in this algorithm are completely independent of each other, it is easily parallelized. Using the `mpi4py` directive, the code is executed on parallel processes and then the result for each map-coordinate (β, r) was averaged from the results from all runs. We then expect the error for each coordinate to converge to some specific value depending on the location in the phase space.

3.1.2 Results

The behaviour of the finite size samples are characterized by the heat maps created using Algorithm 1 described above, making them a good starting point.

In Figure 3.1 the dense areas of the phase space are visualized for the HC and CsCsHM statistic respectively. We see that there they are quite similar with comparable errors. HC seems to be more certain once we move away from the detection boundary, but the CsCsHM performs a tiny sliver better close to the boundary.

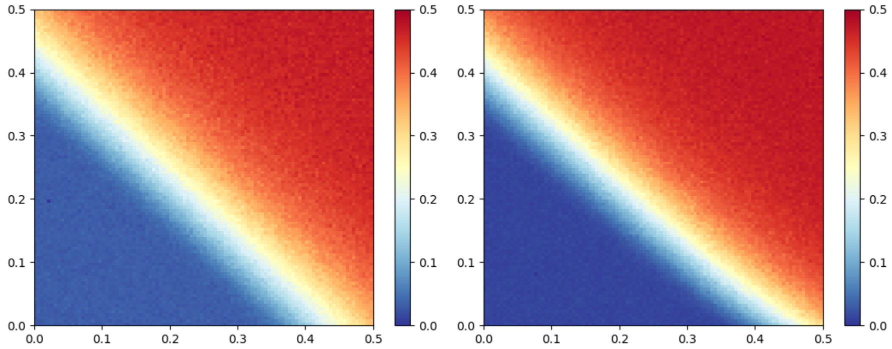

Figure 3.1: Empirical heat map of $\widehat{\text{Err}}$ in dense region of the phase space for CsCsHM_1 to the left and HC to the right. The grid size is (100×100) and $n = 10^4$. As usual the x-axis is spanned by β and the y-axis by r.

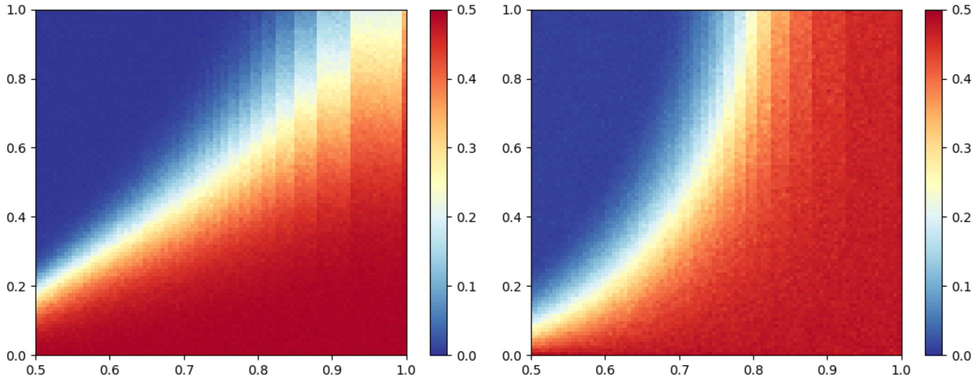

Figure 3.2: Empirical heat map of $\widehat{\mathrm{Err}}$ in sparse region of the phase space for CsCsHM$_1$ to the left and HC to the right. The grid size is (100×100) and $n = 10^4$. As usual the x-axis is spanned by β and the y-axis by r. Here $\alpha_0 = 0.1$ for CsCsHM and $\alpha_0 = 0.5$ for HC.

Moving on to the sparse heat maps, Figure 3.2, we see that the two statistics are comparable here as well. Both seem to perform well once it has some distance to the detection boundary. The HC statistic seems to be better at weaker signals around $\beta \in (0.5, 0.7)$ but CsCsHM is better for larger β:s. This phenomenon remains unexplained. What was noticed when experimenting was that the choice of α_0 affected the error rates, especially for CsCsHM, with higher α_0:s giving much worse results. HC was more resistant to this behaviour.

To explore the error we look at the test statistics' values under H_0 and H_1. In Figure 3.4 we see the behaviour of the HC test variables, and in Figure 3.3 for CsCsHM. For the latter we have removed the thin but long upper tail of the alternative hypothesis to make the visualization more clear. Under H_0 it seems that CsCsHM has a heavy upper tail, which makes it difficult to separate it from the alternative's left tail, which we can see in the bottom picture of the two figures. HC has lighter left tail and is therefore more separable for when H_1 or H_0 are true.

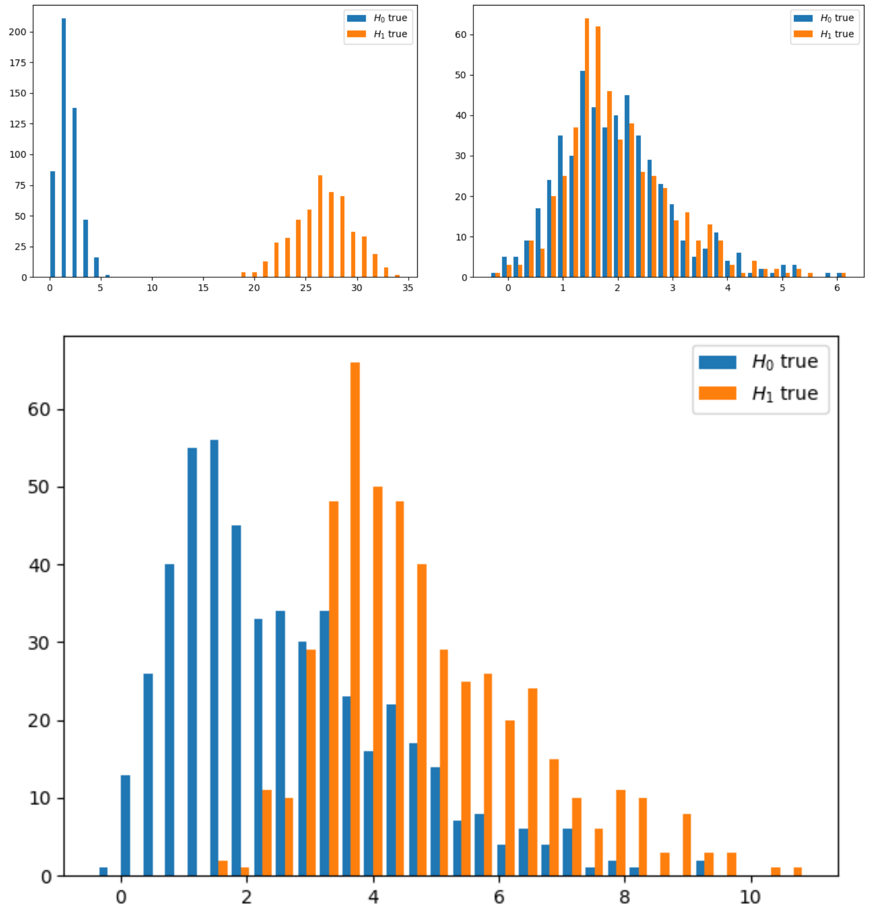

Figure 3.3: Histograms of the CsCsHM$_1$ test statistic for simulated data. The orange bins correspond to when the alternative H_1 is true and the blue bins when H_0 is true. The parameters in the phase space are: top left $(\beta = 0.55, r = 0.9)$, top right $(\beta = 0.9, r = 0.3)$ and bottom $(\beta = 0.55, r = 0.3)$.

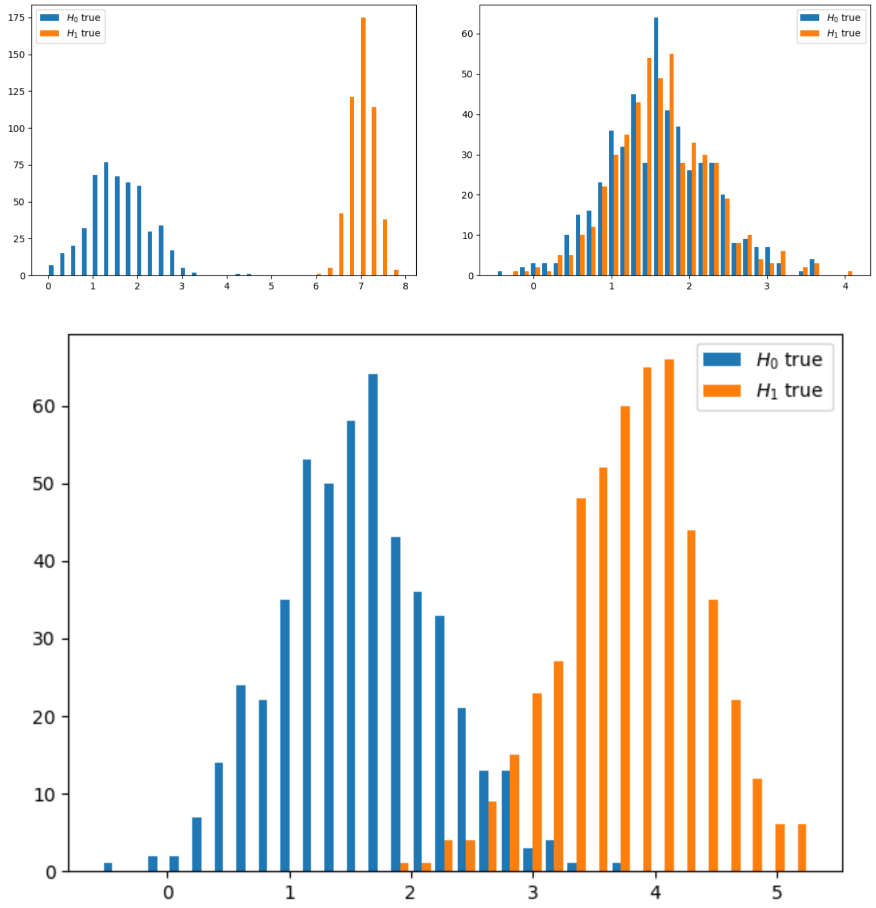

Figure 3.4: Histograms of the HC test statistic for simulated data. The orange bins correspond to when the alternative H_1 is true and the blue bins when H_0 is true. The parameters in the phase space are: top left $(\beta = 0.55, r = 0.9)$, top right $(\beta = 0.9, r = 0.3)$ and bottom $(\beta = 0.55, r = 0.3)$.

3.2 Classification

For classification we want to investigate CsCsHM thresholding throughout the phase space. In addition we want to investigate the performance on some cancer data sets in comparison to HCT.

3.2.1 Methods

To study the classification we will need an error measure to be able to look at the finite sample size behaviour of the CsCsHM thresholding procedure. To investigate the cancer data sets we need a procedure that allows us to use the framework without violating our assumptions too much.

Error measure for the phase diagram

For the classification situation we want to measure the capability of the classifier to correctly separate the two classes by selecting the interesting variables. A natural way to measure this possibility is to see how well it performs on a test data set. The error is thus taken as the misclassification rate on a set of data not used in training,

$$\widehat{\mathrm{Err}} = \frac{\#\text{falsely classified points}}{\#\text{total points}}. \tag{3.2}$$

With this error, heat maps spanning over the phase space were plotted with the intention to show how the boundary changes close to the theoretical detection boundary. To do this we generate training and test data of equal sizes with balanced distribution over classes, train our classifiers on the training data and finally we apply them on the test data. The color in the heat map is scaled after the error.

Procedure for the cancer data sets

In order to assess the classification procedures, they were tested on some real data. The data sets chosen have been explored with various techniques, where there are reported error rates for both simple and complex classification methods.

Two microarray datasets that have been successfully used for differentiating patients with cancer and without cancer have been described by Alon et al [18] and Golub et al [19]. The first is for colon cancer and the second for leukemia. Both Donoho and Jin [4] as well as Dettling [20] have applied various techniques for classification on these data sets. The colon cancer data set consists of $n = 62$ samples distributed over the classes as $n_1 = 40$ and $n_2 = 22$ with dimensionality $p = 2000$, and the leukemia data set consists of $n = 73$ samples distributed as $n_1 = 48$ and $n_2 = 25$ with dimensionality $p = 7129$.

It is of course difficult to decide the sparsity β or signal strength r for the data sets, but the setting is certainly $p >> n$ for these data sets. and the hypothesis is that only a few of the genes' expressions correlate with the disease.

In order to compare the CsCsHM to the HCT approach, both were applied to the data sets. The testing procedure was kept simple, forming z-scores, using the HC or CsCsHM procedure to find a threshold value to select interesting variables from these, and then finally applying this classifier on a test set not used in training.

Calling the set of all data as the union of the set of data in each class, $C = C_1 \cup C_2$, and denoting the size of a set A as $|A|$, the z-scores are formed as

$$z_j^* = \frac{1}{\sqrt{1/|C_1| + 1/|C_2|}} \frac{\bar{x}_{j1} - \bar{x}_{j2}}{s_j}, \quad 1 \le j \le p, \tag{3.3}$$

where the standard deviation s_j is estimated by

$$s_j^2 = \frac{1}{|C| - 2}\Big(\sum_{i \in C_1} (x_{ij} - \bar{x}_{j1})^2 + \sum_{i \in C_2} (x_{ij} - \bar{x}_{j2})^2 \Big)$$

and the class mean for class k as $\bar{x}_{ji} = \frac{1}{|C_k|} \sum_{j \in C_k} x_{ij}$. Since we do not know the underlying distribution of the gene expressions, we look at the difference of the class means which we can expect to be approximately normally distributed.

These scores are then normalized,

$$Z_j = \frac{z_j^* - \bar{z}^*}{std(z^*)}, \quad 1 \le j \le p, \tag{3.4}$$

where $std(z^*)$ is the standard deviation of the z-scores, and $\bar{z}^*$ is their mean.

Using either the HC or CsCsHM procedures, the threshold value is found from these scores. This value is used to choose variables, which are in turn used to classify each sample in a test set. The misclassification rate was then calculated. To calculate the mean misclassification rate, a 10-fold cross-validation split of the data was used: the classifier was trained on $9/10$th of the data and tested on the remaining $1/10$th. The testing set was rotated so that all parts was used for training and testing. This was done for three procedures, first HCT with the HC^+ statistic as defined in equation 2.22 with the normalization as in HC^{2008}

2.21, then for the alternative CsCsHM thresholding procedure with the two statistics CsCsHM$_1$ and CsCsHM$_2$ as found in 2.35 and 2.36.

3.2.2 Results

Heat maps

The heat map, see Figure 3.5, shows that the CsCsHM$_1$ classifier has a decent behaviour in the area just above the detection boundary. As has been noticed before for signal detection, there is no sharp boundary in the case of classification either. The classifier has a "grey area" for finite samples where classification is possible, but with a slightly elevated error rate of roughly 0.25. This area is quite big when $\beta > 0.5$. We also notice that the plot is quite non-smooth, despite the high number of samples, meaning that the results have a high variance.

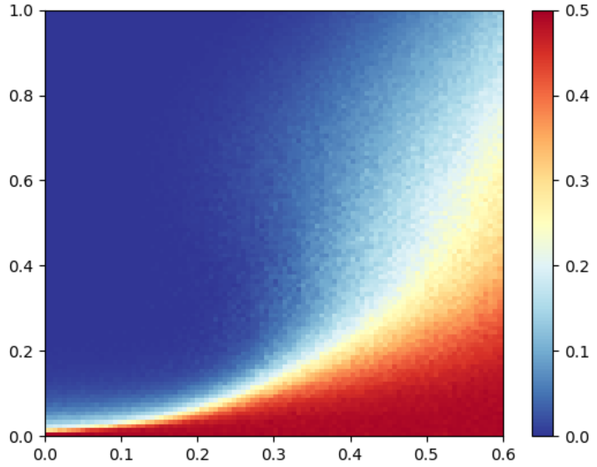

Figure 3.5: Empirical behaviour of finite sample size of CsCsHM$_1$ thresholding near the detection boundary for classification. Here $\theta = 0.4$, $\alpha_0 = 0.1$, the grid size is (100×100) and $n = 10^3$.

Cancer data classification

In Table 3.1 and Table 3.2 the results for the classification of the cancer data sets are presented. As we can see, the statistics behave differently. Although the error is comparable between all three classification procedures, the number of selected variables varies widely. The CsCsHM$_2$ statistic consistently chooses a lot more variables, while both HC and CsCsHM$_1$ choose fewer depending on the data set.

When looking at the objective function for all the feature scores (in a way similar to what is done in [6]) for the cancer data, we notice that they have peculiar s-shapes, see Figure 3.6. For both data sets we see that the choice of α_0 will heavily impact the test statistic. This is not a desirable behaviour since we want the method to be non-parametric.

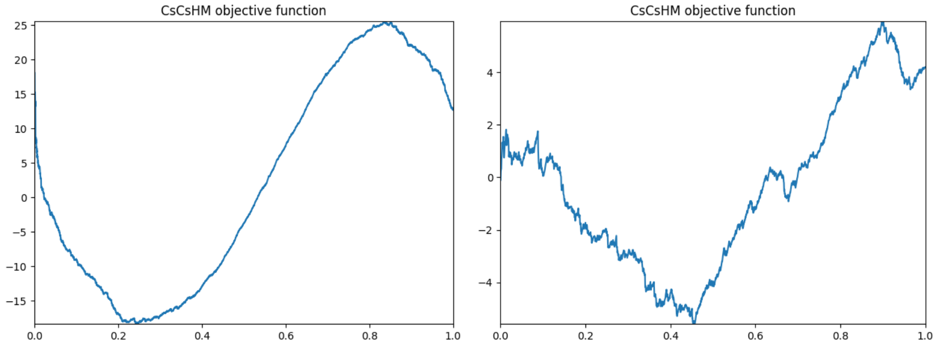

Figure 3.6: CsCsHM$_1$ objective function for every feature score for the two cancer data sets, leukemia to the left, and colon cancer to the right. The x-axis is taken as the fraction of the features (to simplify the interpretability of the effect of α_0).

Table 3.1: Mean misclassification rates and mean number of variables selected with standard deviation for the leukemia data set, with a size of (73×7129), for different procedures.

Method	Misclassification rate	N. variables selected
HC^+	0.024 (±0.00771)	151.4 (±4.35)
$CsCsHM_1$	0.026 (±0.00284)	68.4 (±3.09)
$CsCsHM_2$	0.0246 (±0.00437)	1275.3 (±661.37)

Table 3.2: Mean misclassification rates and mean number of variables selected with standard deviation for the colon cancer data set, with a size of (62×2000), for different procedures.

Method	Misclassification rate	N. variables selected
HC^+	$0.1058 \quad (\pm 0.01024)$	$35.8 \quad (\pm 4.78)$
$CsCsHM_1$	$0.1075 \quad (\pm 0.01176)$	$87.6 \quad (\pm 12.30)$
$CsCsHM_2$	$0.1075 \quad (\pm 0.01115)$	$137.1 \quad (\pm 22.83)$

Investigating the effect of α_0

In an attempt to try to understand how many variables are chosen relates to the position in the phase space, simulations for different points were made. In Figure 3.7 we let r vary over ten points for a fixed β and observe the behaviour of the error as well as the number of variables chosen. In Figure 3.8 we let β vary instead for a fixed r and observe the results. What can be noted is that when truncating the upper part of the objective functions with a small α_0, CsCsHM chooses fewer variables. The error behaves roughly the same.

Looking at Figure 3.8 we can see that for HC the number of variables selected as β is small seems to be the same as the number of informative signals. This number decreases as the sparsity increases, only to increase again when the sparsity is so high that the signals are very few. The same behaviour in reverse can be observed in Figure 3.7, where we start with very weak signals so the signals cannot be differentiated from the noise. For both images we notice that CsCsHM performs consistently worse than HC in mean error.

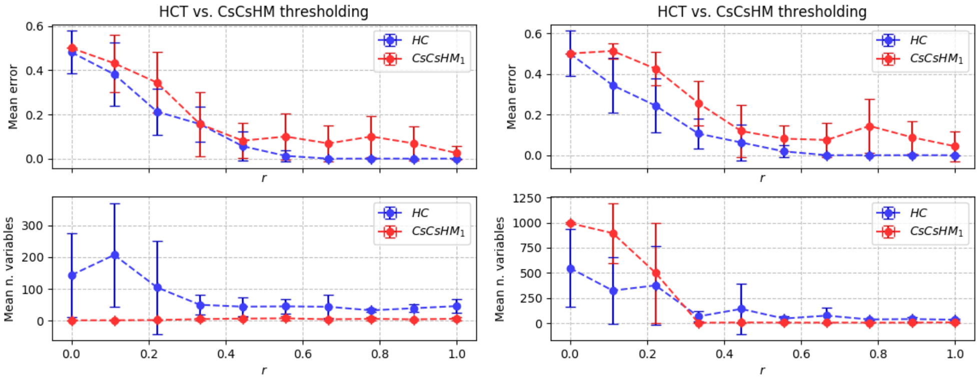

Figure 3.7: Mean error and mean number of selected variables for CsCsHM$_1$ thresholding and HCT for a fixed $\beta = 0.5$ on $p = 10^3$ dimensions and varying r for $\theta = 0.3$. To the left: $\alpha_0^{CsCsHM} = 0.1$ and $\alpha_0^{HC} = 0.5$ to the right $\alpha_0^{CsCsHM} = \alpha_0^{HC} = 1$

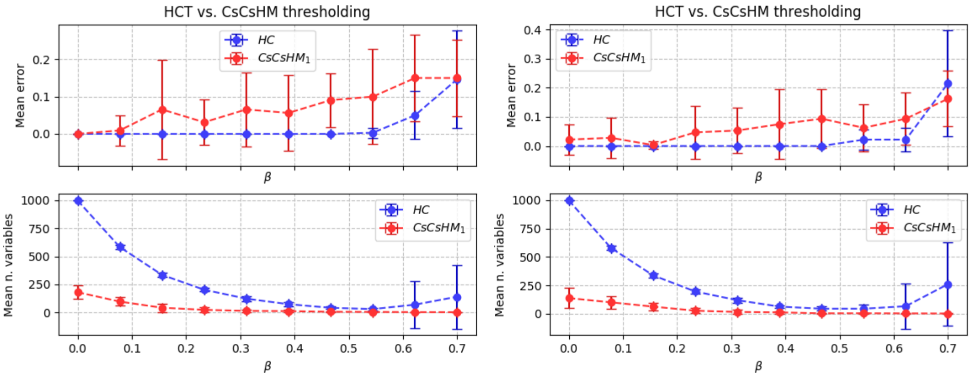

Figure 3.8: Mean error and mean number of selected variables for CsCsHM$_1$ thresholding and HCT for a fixed $r = 0.8$ on $p = 10^3$ dimensions and varying β for $\theta = 0.3$. To the left: $\alpha_0^{CsCsHM} = 0.1$ and $\alpha_0^{HC} = 0.5$ to the right $\alpha_0^{CsCsHM} = \alpha_0^{HC} = 1$.

Chapter 4

Discussion

In this section we attempt to explain the results and put them in a proper context. Some suggestions for future work are also proposed.

Remarks and conclusions

The behaviour of the different statistics for finite sample sizes is key to how useful they are in practice. These initial explorations done in this report are a good starting point but it remains to do a more thorough investigation where the connection of the statistics behaviour and the parameter α_0 is more precisely mapped out.

Heat maps of the empirical error show that the statistics are comparable when it comes to signal detection. The impact of α_0 was noted as a big factor for the performance of the CsCsHM type statistics, which puts them at a disadvantage for practical use. Moreover it seems that for classification the HC type statistic manages to select the correct signals to a higher extent, giving the classifier a lower error. However, CsCsHM has a similar error but slightly elevated, and chooses fewer variables. To get a lightweight model with as few variables as possible, one could argue that the easiest way is either using CsCsHM, or using HC and choosing a subset of the selected variables.

Classification on real data sets is a very practical hands-on way of comparing the classifiers, but the exact results are maybe not that interesting. Since the data sets are special cases chosen because they have been used before, there is nothing new brought to the table and the results could be circumstantial. This part did contribute to the understanding the number of variables that are chosen by the statistics.

To sum up, the conclusion is that despite CsCsHM:s better asymptotical properties, their performance are still in practice equal to slightly worse than the HC type statistic for sample sizes up to 10^4.

Suggestions for future work

Future numerical studies could systematically investigate the CsCsHM statistics behaviour for different α_0:s on simulated data. The impact of the sample size should also be more carefully investigated. For classification the number of correctly chosen variables from simulated data could be taken as a more accurate and interesting measure of performance than error on a test set. This would also lower the computational cost of the simulations.

In this report we have only considered the normal mixture case with homoscedasticity, and this is only a special case. There are a lot of interesting models including more challenging situations such as non-gaussianity and heteroscedasticity. Since the CsCsHM statistic was just recently proposed, there are not a lot of theoretical results in these areas, but empirical investigations of finite size sample data for HC and other statistics such as Berk-Jones would be interesting.

Bibliography

[1] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures, 2004. ISSN 00905364.

[2] Natalia Stepanova and Tatjana Pavlenko. Goodness-of-fit tests based on sup-functionals of weighted empirical processes. *arXiv preprint arXiv:1406.0526*, to appear 2017.

[3] David Donoho and Jiashun Jin. Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4449–4470, 2009.

[4] David Donoho and Jiashun Jin. Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences of the United States of America*, 105(39):14790–5, 2008. ISSN 1091-6490. doi: 10.1073/pnas.0807471105. URL `http://www.scopus.com/inward/record.url?eid=2-s2.0-54449086895{&}partnerID=tZOtx3y1`.

[5] Jiashun Jin. Impossibility of successful classification when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 106(22):8859–8864, 2009.

[6] David Donoho and Jiashun Jin. Higher Criticism for Large-Scale Inference: especially for Rare and Weak effects. *Statistical Science*, 30(1):1–25, 2015. ISSN 0883-4237. doi: 10.1214/14-STS506. URL `http://arxiv.org/abs/1410.4743`.

[7] Niclas Blomberg. Higher criticism testing for signal detection in rare and weak models, 2012.

[8] Yu I Ingster. Minimax detection of a signal for i (n)-balls. *Mathematical Methods of Statistics*, 7(4):401–428, 1998.

[9] Theodore W Anderson and Donald A Darling. A test of goodness of fit. *Journal of the American statistical association*, 49(268):765–769, 1954.

[10] Amit Moscovich, Boaz Nadler, Clifford Spiegelman, et al. On the exact berk-jones statistics and their p-value calculation. *Electronic Journal of Statistics*, 10(2):2329–2354, 2016.

[11] Leah Jager and Jon A Wellner. Goodness-of-fit tests via phi-divergences. *The Annals of Statistics*, pages 2018–2053, 2007.

[12] Pengsheng Ji and Jiashun Jin. UPS delivers optimal phase diagram in high-dimensional variable selection. *Annals of Statistics*, 40(1):73–103, 2012. ISSN 00905364. doi: 10.1214/11-AOS947.

[13] Jiashun Jin and Tracy Ke. Rare and Weak effects in Large-Scale Inference: methods and phase diagrams. *arXiv preprint*, page 31, 2014. ISSN 10170405. doi: 10.5705/ss.2014.138. URL `http://arxiv.org/abs/1410.4578`.

[14] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning (2nd edition)*. Springer-Verlag, 2009.

[15] T. Tony Cai, X. Jessie Jeng, and Jiashun Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 73(5):629–662, 2011. ISSN 13697412. doi: 10.1111/j.1467-9868.2011.00778.x.

[16] Tatjana Pavlenko, Anders Björkström, and Annika Tillander. Covariance structure approximation via gLasso in high-dimensional supervised classification. *Journal of Applied Statistics*, 39(8): 1643–1666, 2012. ISSN 0266-4763. doi: 10.1080/02664763.2012.663346. URL `http://www.tandfonline.com/doi/abs/10.1080/02664763.2012.663346`.

[17] Zheyang Wu, Yiming Sun, Shiquan He, Judy Cho, Hongyu Zhao, and Jiashun Jin. Detection boundary and Higher Criticism approach for rare and weak genetic effects. *Annals of Applied Statistics*, 8(2):824–851, 2014. ISSN 19417330. doi: 10.1214/14-AOAS724.

[18] U Alon, N Barkai, D a Notterman, K Gish, S Ybarra, D Mack, and a J Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 96(12):6745–6750, 1999. ISSN 00278424. doi: 10.1073/pnas.96.12.6745.

[19] T R Golub, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield, and E S Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999. ISSN 0036-8075. doi: 10.1126/science.286.5439.531.

[20] M Dettling. BagBoosting for tumor classification with gene expression data. *Bioinformatics (Oxford, England)*, 20(18):3583–3593, 2004. ISSN 1367-4803.