ROYAL INSTITUTE
OF TECHNOLOGY

Higher Criticism Testing for Signal Detection in Rare And Weak Models

Niclas Blomberg

Master Thesis at KTH Royal Institute of Technology
Stockholm, September 2012
Supervisor: Tatjana Pavlenko

Abstract

In several application fields today − genomics and proteomics are examples − we need models for selecting a small subset of useful features from high-dimensional data, where the useful features are both *rare* and *weak*, this being crucial for e.g. supervised classification of sparse high-dimensional data. A preceding step is to detect the presence of useful features, *signal detection*. This problem is related to testing a very large number of hypotheses, where the proportion of false null hypotheses is assumed to be very small. However, reliable signal detection will only be possible in certain areas of the two-dimensional sparsity-strength parameter space, the *phase space*.

In this report, we focus on two families of distributions, $\mathcal{N}$ and χ^2. In the former case, features are supposed to be independent and normally distributed. In the latter, in search for a more sophisticated model, we suppose that features depend in blocks, whose empirical separation strength asymptotically follows the non-central χ^2_ν-distribution.

Our search for informative features explores Tukey's *higher criticism (HC)*, which is a *second-level significance testing procedure*, for comparing the fraction of observed significances to the expected fraction under the global null.

Throughout the phase space we investgate the estimated error rate, $\widehat{\mathrm{Err}}$ = (#Falsely rejected H_0+ #Falsely rejected H_1)/#Simulations, where H_0: absence of informative signals, and H_1: presence of informative signals, in both the $\mathcal{N}$-case and the χ^2_ν-cases, for $\nu = 2, 10, 30$.

In particular, we find, using a feature vector of the approximately same size as in genomic applications, that the analytically derived *detection boundary* is too optimistic in the sense that close to it, signal detection is still failing, and we need to move far from the boundary into the *success region* to ensure reliable detection. We demonstrate that $\widehat{\mathrm{Err}}$ grows fast and irregularly as we approach the detection boundary from the success region.

In the χ^2_ν-case, $\nu > 2$, no analytical detection boundary has been derived, but we show that the empirical success region there is smaller than in the $\mathcal{N}$-case, especially as ν increases.

Contents

1 Introduction

Think of a microarray measurement of a human genome, and consider the proposal that an extremely small fraction of the genes are over (or under) expressed for cancer patients. Further, suppose that these over expressed genes are in fact just slightly over expressed, with a weak amplitude, μ_0. Then, apparently, it will be an intricate task to determine whether a microarray measurement contains over expressed "cancer genes", or not − a task that is the main consideration in this report.

This issue of high-dimensional measurements in *rare* and *weak* settings arises in many modern applications, not only in genomics, but in e.g. proteomics, cosmology, astronomy, and in robustness and covert communication problems (see e.g. Donoho & Jin (2009) or Meinhausen & Rice (2006), and references therein).

We face the problem of *signal detection*, a multiple testing problem where the proportion of false null hypotheses usually is small. We test the global null H_0: absence of informative signals, against the global alternative H_1: presence of informative signals.

In most studies on the subject features are assumed to be independent and normally distributed. However, we also encounter the problem assuming that features depend in blocks. In this case, the empirical blockwise separation strength is proven in Pavlenko *et al* (2012) to be χ^2-distributed. Hence, we study the $\mathcal{N}$- and the χ^2-cases.

We start by constructing a two-dimensional sparsity-strength parameter space, and call it the *phase space*. Using likelihood ratio tests (LRT) we can perfectly separate H_1 from H_0, but only in a subset of phase space, the *success region* (Donoho & Jin (2004)). The success region is defined through the *detection boundary*, a curve splitting the success and failure regions; however, in the χ^2_ν-cases, $\nu > 2$, no such detection boundary has been derived.

Our testing procedure, *higher criticism (HC)*, was initially proposed by Tukey in 1976. He adopted the term from the traditional method of higher criticism for studying ancient literature, where not only the literature but the circumstances it was written under is considered − a higher level study. The idea here is to form "Z-scores of P-values", performing *second-level significance testing*. In contrast to LRT and most other multiple testing procedures, higher criticism does not need information of the sparsity and strength parameters − it is *adaptive*. Also, e.g. Cai *et al* (2011) have proven that higher criticism perfectly separates H_1 from H_0 everywhere that LRT does, i.e. throughout the success region. Hence, we say that higher criticism has the property of *optimal adaptivity*.

The goal in this report is to empirically investigate the possibility of signal detection throughout the phase space, in both the $\mathcal{N}$- and χ^2_ν-cases, using higher criticism. Here, the qustions of interest are: Will the empirical results verify the analytical ones, in particular considering the detection boundary? What does the empirical success region and detection boundary look like in χ^2_ν-cases, $\nu > 2$, where no analytical detection boundary has been derived? Is the success region

appropriately described as a homogeneous region? For which dimensionality does higher criticism start to exhibit its optimal performance?

1.1 $\mathcal{N}$- or χ^2-distributions – two models

1.1.1 The $\mathcal{N}$-case

Consider the feature vector $X = (X_1, X_2, \ldots, X_p)$, where $p \sim 5 \cdot 10^3$ in genomic applications (compare Pawitan *et al* (2005)), but can be up to $\sim 10^{11}$ in astronomy (compare Meinhausen & Rice (2006)). In a first, most basic, model formulation, we use the following two assumptions:

1. Features are independent and normally distributed.
2. Informative signals have a common and weak amplitude, μ_0.

Now, suppose that $X_i, (1 \leq i \leq p)$, has the probability ϵ of being informative and $(1-\epsilon)$ of being uninformative. We model uninformative signals as $\mathcal{N}(0,1)$, and informative as $\mathcal{N}(\mu_0, 1)$, and then test H_0: absence of informative signals ($\epsilon = 0$), versus H_1: presence of informative signals ($\epsilon \in (0,1)$). The global null hypothesis is:

$$H_0^{(p)} : \quad X_i \overset{\text{IID}}{\sim} \mathcal{N}(0,1), \quad 1 \leq i \leq p, \tag{1}$$

and the global alternative is

$$H_1^{(p)} : \quad X_i \overset{\text{IID}}{\sim} (1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(\mu_0, 1), \quad 1 \leq i \leq p, \tag{2}$$

where ϵ, here, can be seen as the fraction of informative signals and μ_0 as their (common and weak) strength (Cai *et al* (2011)).

1.1.2 The χ^2-case

Considering the human genome it is obviously very naive to treat all genes as independent. Instead, we could expect them to depend in blocks; so that one block of genes regulate a certain function, yet another block regulate another function, and so on. The dependence within blocks are naturally much stronger than between, why we assume that blocks are independent (Pavlenko *et al* (2012)). With this it will be possible to filter out whole segments of genes that are uninformative.

As mentioned above, the block structure implies χ^2-distributed block strengths. Hence, we postulate the following three assumptions:

1. Features are blockwise independent and block separation strengths are χ^2-distributed.
2. All blocks are of the same size, p_0.
3. Informative blocks have a common amplitude, $\omega_0^2 = \mu_0^2$.

Now, by analogy with (1) and (2), where $X = (X_1, \ldots, X_b)$ is our set of b features (blocks), p_0 is the block size or degrees of freedom, $b = p/p_0$, and ϵ is the probability of a block X_i ($1 \leq i \leq b$) to be informative, we formulate the global null and alternative for the χ^2-case as:

$$H_0^{(b)}: \quad X_i \overset{\text{IID}}{\sim} \chi^2_{p_0}(0), \quad 1 \leq i \leq b, \tag{3}$$

and,

$$H_1^{(b)}: \quad X_i \overset{\text{IID}}{\sim} (1-\epsilon)\chi^2_{p_0}(0) + \epsilon\chi^2_{p_0}(\omega_0^2), \quad 1 \leq i \leq b, \tag{4}$$

where, again, ϵ can be seen as the fraction of informative blocks, and ω_0^2 is their (common and weak) strength.

The block size is usually $5 \leq p_0 \leq 15$, when working with feature vectors of size $p \sim 10^6$. If bigger, the number of blocks, b, would become too small to give reliable results.

1.2 The phase space

In the *phase space* we parameterize sparsity and strength. We have already introduced ϵ (for sparsity), and μ_0 and ω_0^2 (for strength), but, the phase space is more conveniently parameterized on $(0,1)^2$. Thus, we transform ϵ into a sparsity parameter, β, and μ_0 and ω_0^2 respectively into a strenght parameter, r.

1.2.1 The sparsity parameter, β

The *sparsity parameter* β is related to ϵ as follows.

In the $\mathcal{N}$-case:

$$\epsilon = \epsilon(\beta) = p^{-\beta}, \quad 0 < \beta < 1 \tag{5a}$$

In the χ^2-case:

$$\epsilon = \epsilon(\beta) = b^{-\beta}, \quad 0 < \beta < 1 \tag{5b}$$

1.2.2 The strength parameter, r

Cai *et al* (2011) argue that the detection problem behaves very differently in two regimes: the *dense regime*, $0 < \beta < 1/2$, and the *sparse regime*, $1/2 \leq \beta < 1$. In the sparse regime $\epsilon \ll 1/\sqrt{p}$, and the most interesting situation is when μ_0 grows with p at a rate of $\sqrt{\log p}$; with other growth rates, it is either too easy or impossible to separate the two hypotheses. In contrast, in the dense case where $\epsilon \gg 1/\sqrt{p}$, the most interesting situation is when μ_0 degenerates to 0 at an algebraic order, so that moment-based statistics could be successful. (Note, however, that moment-based statistics are still not preferred as β is in general unknown.) On this basis, Cai *et al* (2011) relate the strength parameter r to μ_0 and ω_0^2 as:

$$\mu_0 = \mu_0(r;\beta) = \begin{cases} p^{-r}, & 0 < \beta < 1/2;\ 0 < r < 1/2 \quad \text{(dense)} \\ \sqrt{2r\log p}, & 1/2 \leq \beta < 1;\ 0 < r < 1 \quad \text{(sparse)} \end{cases} \tag{6}$$

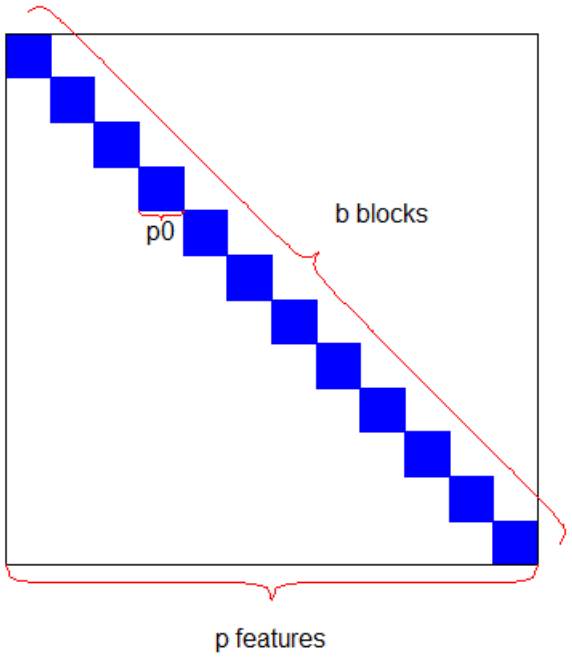


Figure 1: An illustrative scheme over block structure. We normally have $p < 10^4$, $5 \leq p_0 \leq 15$, and $b = p/p_0$ in genomic applications, thus a lot more blocks than viewed here.

$$\omega_0^2 = \omega_0^2(r;\beta) = \begin{cases} b^{-2r}, & 0 < \beta < 1/2;\ 0 < r < 1/2 \quad \text{(dense)} \\ 2r \log b, & 1/2 \leq \beta < 1;\ 0 < r < 1 \quad \text{(sparse)} \end{cases} \tag{7}$$

Note that the parameters are undefined for $0 < \beta < 1/2$ and $1/2 \leq r < 1$. Intuitively, in this area informative signals are too weak; μ_0 ranges from $p^{-1/2}$ to p^{-1}.

1.2.3 The detection boundary separates success and failure regions

If parameters are known, then in the success region it can be shown (see Cai *et al* (2011)) that the likelihood ratio test (LRT) obeys

$$P_{H_0}(\text{reject } H_0) + P_{H_1}(\text{reject } H_1) \to 0, \text{ as } p \text{ (or } b) \ \to \infty, \tag{8}$$

where left hand sum of (8) can be interpreted as the sum of Type I and II errors; hence we see that within the success region the LRT perfectly separates the alternative from the null.

Specifically, Cai *et al* (2011) report that using LRT the following *detection boundary*, $r = \rho^*_{\mathcal{N}}(\beta)$, can be derived (see also Figure 2):

$$\rho^*_{\mathcal{N}}(\beta) = \begin{cases} 1/2 - \beta, & 0 < \beta < 1/2 \\ \beta - 1/2, & 1/2 < \beta \leq 3/4 \\ (1 - \sqrt{1-\beta})^2, & 3/4 < \beta \leq 1 \end{cases} \tag{9a}$$

Further, Donoho & Jin (2004) claim that in the χ^2_2-case (note the subscript 2) we have the same detection boundary, in the sparse regime:

$$\rho^*_{\chi^2_2}(\beta) = \begin{cases} \beta - 1/2, & 1/2 < \beta \leq 3/4 \\ (1 - \sqrt{1-\beta})^2, & 3/4 < \beta \leq 1 \end{cases} \tag{9b}$$

The success region is now defined, in the dense regime, $0 < \beta < 1/2$, as the area below the boundary, $r < \rho^*_{\mathcal{N}}(\beta)$; and, in the sparse, $1/2 < \beta \leq 1$, as the area above, $r > \rho^*_{\mathcal{N},\chi^2_2}$. See Figure 2.

The failure region is the complement of the success region, and there the left hand sum in (8) approaches 1 for any test. It is worth noticing that the detection boundary in (9) is the same considering supervised classification, where the goal is to select a set of useful features that are most informative for class difference (see Jin (2009)).

Note that the left hand sum of (8) is of high importance in this report, and will be converted to its empirical version, $\widehat{\text{Err}}$, in Section 2.

1.3 Higher critcism

Considering the generally unrealistic requirement of known parameters (β, r) of the LRT, we want to find an adaptive method that works even without such *oracle* knowledge. Here *higher criticism (HC)* comes into the picture, a non-parametric procedure for signal detection (as well as feature selection, see Donoho & Jin (2009)), which, like the LRT, is successful in the entire success region − i.e. has the property of *optimal adaptivity*.

1.3.1 The testing procedure

Let us now describe the procedure of higher criticism, in the $\mathcal{N}$-case. It will work identically in the χ^2-case, see motivation below.

We have a set of p features, $X = (X_1, X_2, \ldots, X_p)$, from which we define the empirical culmulative distribution function and empirical survival function of X_i respectively:

$$F_p(t) = \frac{1}{p} \sum_{i=1}^{p} \mathbf{1}_{\{X_i < t\}},$$

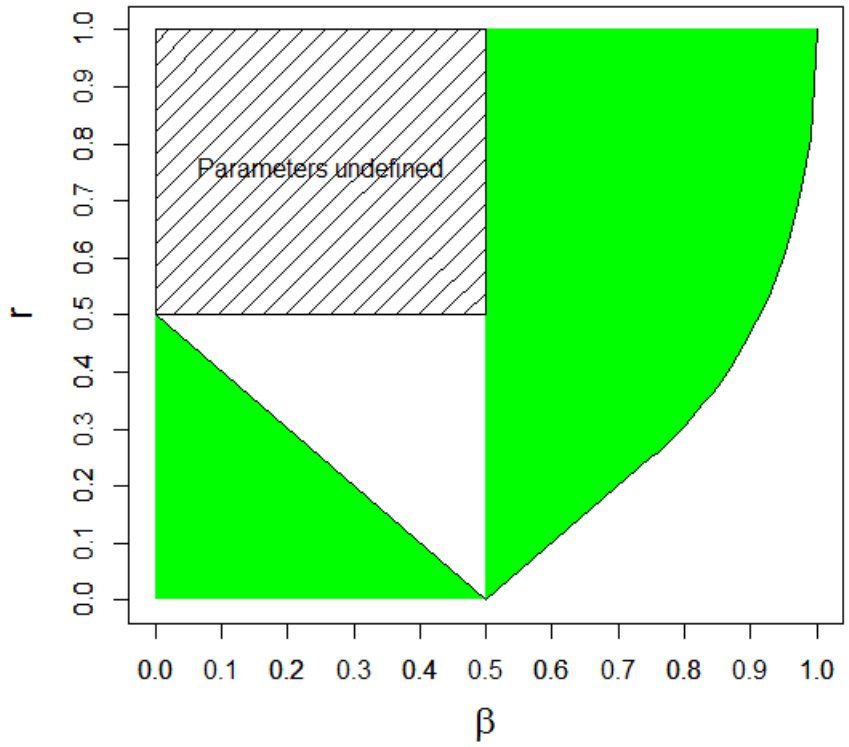


Figure 2: The detection boundary in (8) with the success region shaded (green). The undefined area is due to signals being too weak.

$$\bar{F}_p(t) = 1 - F_p(t).$$

Now, look at the standardized form of $\bar{F}_p(t) - \bar{\Phi}(t)$, to compare the fraction of obseved significances to the expected fraction under the global null:

$$\frac{\bar{F}_p(t) - \bar{\Phi}(t)}{\sqrt{\bar{\Phi}(t)(1 - \bar{\Phi}(t))}}\sqrt{p}.$$

From this we can define the *HC objective function*, $HC_p(i)$, through the following steps. Start by computing P-values of $x = (x_1, x_2, \ldots, x_p)$,

$$\pi_i = \bar{\Phi}(x_i) \equiv P(\mathcal{N}(0,1) \geq x_i), \quad 1 \leq i \leq p.$$

Second, sort the P-values in the ascending order $\pi_{(1)} \leq \pi_{(2)} \leq \ldots \leq \pi_{(p)}$. Consider the value t that satisfies $\bar{\Phi}(t) = \pi_{(i)}$. Since there are exactly i P-values less than or equal to $\pi_{(i)}$, exaclty i features are greater than or equal to t. Hence, for this particular t, $\bar{F}_p(t) = i/p$, which will hold for all t that satisfy $\bar{\Phi}(t) = \pi_{(i)}$. And, now, the standardized form of $\bar{F}_p(t) - \bar{\Phi}(t)$ becomes the *HC objective function*

$$HC_p(i) = \frac{i/p - \pi_{(i)}}{\sqrt{\pi_{(i)}(1 - \pi_{(i)})}}\sqrt{p}, \tag{10}$$

which is the "Z-score of the P-value". Note that there are some different versions of (10) that perform equally good (see Meinshausen & Rice (2006) for a general discussion on bounding functions, alternative to $\pi_{(i)}(1 - \pi_{(i)})$).

Next, we define the *HC test statistic*,

$$HC_p^* = \max_{1 \leq i \leq \alpha_0 p} HC_p(i)(t), \tag{11}$$

where it is sufficient only to look at the $\alpha_0 p$, $\alpha_0 \in (0, 1]$, first indices and still capture the peak of $HC_p(i)$ (we choose an appropriate α_o in Section 2).

The idea is to investigate whether we can detect deviations from the null, under which HC_p^* has certain properties. If the global null hypothesis is true, the distribution of its P-values is $\pi_i \overset{\text{IID}}{\sim} U(0, 1)$, and so asymptotically $HC_p(i) \in \mathcal{N}(0, 1)$. Thus, in (11) we look for the largest standardized discrepancy between the observed behaviour of $\pi_{(i)}$ and the expected under H_0, and reject H_0 when HC_p^* is large.

As seen in Shorack & Wellner (2009), results from empirical processes give that when $HC_p(i) \in \mathcal{N}(0, 1)$, $HC_p^* \approx \sqrt{2 \log \log p}$, which grows to ∞ very slowly. In contrast, under the alternative, $HC_p(i)$ has an elevated mean for some i, and HC_p^* could grow to ∞ algebraically fast. Therefore, an appropriate criteria for rejecting the null hypothesis is when

$$HC_p^* \geq \sqrt{2(1 + \delta) \log \log p}, \tag{12}$$

for some δ. In Section 2, we find a way to empirically choose δ in order to optimally separating H_0 and H_1 using higher criticism.

Cai *et al* (2011) show that the test criteria in (12) satisfies (8) throughout the success region, meaning that where the LRT can successfully separate H_1 from H_0, so can higher criticism. In Section 2 of this report, an empirical counterpart to (8) is formulated, $\widehat{\text{Err}}$, in equal connection to (12), being the key expression in the experiment.

Finally, note that also in the χ^2-case P-values are uniformly distributed, the *HC objective function* is asymptotically normal, and hence the *HC test statistic* $\approx \sqrt{2 \log \log b}$.

1.3.2 Higher criticism mimicking ideal behaviour

Let us now consider the ideal case when data comes from

$$F(t) \in (1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(\mu_0, 1),$$

with known parameters ϵ and μ_0, and with

$$\bar{F}(t) = 1 - F(t).$$

(Again, we look at only the $\mathcal{N}$-case, but can easily translate the expressions to fit the χ^2-case.)

Then, by analogy with (10) and (11) we can define the *ideal HC objective function* as

$$HC_{ideal}(t) = \frac{F(t) - \Phi(t)}{\sqrt{\Phi(t)(1-\Phi(t))}}\sqrt{p}, \tag{13}$$

and the *ideal HC test statistic* as

$$HC^*_{ideal} = \max_t HC_{ideal}(t). \tag{14}$$

Here we explore the connection between the empirical and ideal cases discussed in Cai *et al* (2011). For any fixed t,

$$E[HC_p(i)] = \begin{cases} 0, & \text{under } H_0, \\ HC_{ideal}(t), & \text{under } H_1, \end{cases} \quad p \to \infty. \tag{15}$$

Now, with this connection between $HC_p(i)$ and $HC_{ideal}(t)$ in (15), and because HC^*_p is a straight forward maximation of $HC_p(i)$, as HC^*_{ideal} is of $HC_{ideal}(t)$, we expect HC^*_p to approach HC^*_{ideal} in probability, as $p \to \infty$, why we say that HC mimicks ideal behaviour. This idea is presented in Figure 4.

1.3.3 Feature selection by thresholding with higher criticism

Higher criticism was initially proposed for multiple hypotheses testing as in the signal detection problem, but has recently also been used for feature selection (see Donoho & Jin (2009)). In feature selection the aim is not to detect but to identify the informative signals, thereby selecting useful features.

Now, assuming that data is standardized, we order the observed features in the decreasing rearrangement, $x_{(1)} \geq x_{(2)} \geq \ldots \geq x_{(p)}$, and define the *higher criticism threshold (HCT)* as

$$HCT^*_p = x_{(i^*)}, i^* = \arg\max_i HC_p(i), \tag{16}$$

or, analogously, for the *ideal HC threshold*,

$$HCT^*_{ideal} = F(t^*), \quad t^* = \arg\max_t HC_{ideal}(t). \tag{17}$$

The HCT equals the observed x_{i^*} where i^* maximizes the HC objective function. The higher criticism feature selector then picks features corresponding to $x_{(i)}$'s that are greater than or equal to the threshold.

Interestingly, the threshold is automatically set somewhat higher than μ_0 (or ω_0^2 in the χ^2-case), the strength of the informative signals. That way we miss some of the useful features, but this is a beneficial tradeoff for capturing less noise.

Continuing the argumentation at the end of Section 1.3.2, we here reason that because of (15) and the straigt forward argument maximizing in (16) and (17), we expect $HCT_p^* \approx HCT_{ideal}^*$. Figure 4 demonstrates this idea.

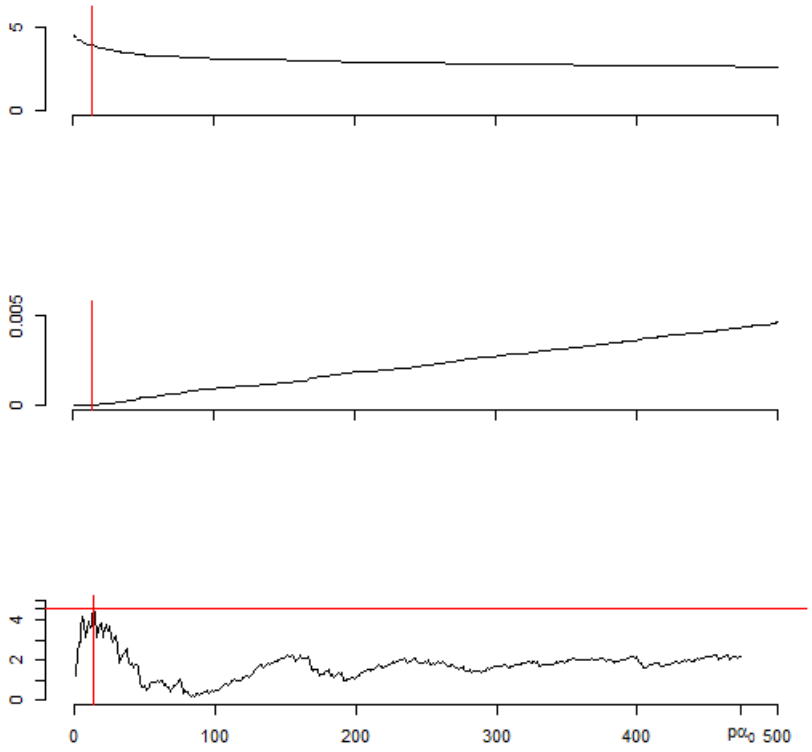


Figure 3: In all three graphs $(p, \beta, r) = (10^5, 0.7, 0.3)$. Up: Simulated signals $x_{(i)}$, ordered, with $1 \leq i \leq p\alpha_0$ (where the choice of α_0 is discussed in Section 2.3). Middle: Corresponding ordered $\pi_{(i)}$-values of data. Down: Corresponding $HC_p(i)$. We have $i^* = \arg\max_i HC_p(i)$ (vertical lines), and $HC_p^* = \max_{1<i<\alpha_0 p} HC_p(i)$ (horizontal line).

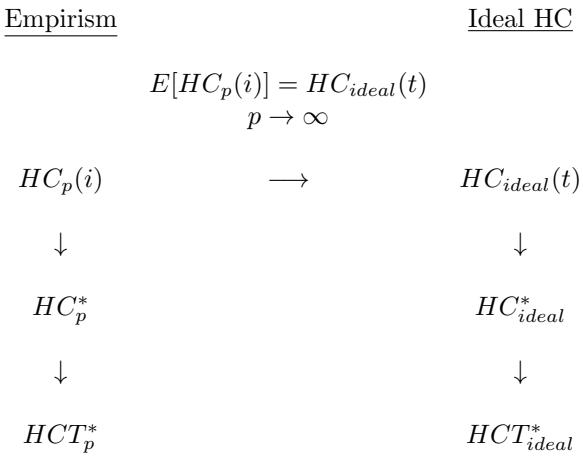


Figure 4: Illustrative chart motivating our choice of higher criticism. Because of the asymptotic connection between $HC_{ideal}(t)$ and $HC_p(i)$, we also expect HC_p^* to approach HC_{ideal}^*, and HCT_p^* to approach HCT_{ideal}^* in probability, as $p \to \infty$.

2 Method development

2.1 Designing the experiment

The aim to investigate signal detection throughout phase space leaves us several choices. Constrained by computational time, we have to carefully choose the amount of (β, r)-points to study, many points close to the detection boundary or a more all-covering study thoughout phase space; the number of and which χ^2_ν-cases to study; the value of p and b; how to optimize the choice of δ; and so on.

One possible way to go is by empirically classifying a (β, r)-point into success or failure, thereby determining the empirical detection boundary observing where we shift between success and failure points. However, labeling a point as either successful or failing gives varying results depending on what upper limit for success we choose for the sum of Type I and II errors, and, also, it poorly describes the error probability at a certain point. Instead, points are best described as more or less successful on a continous scale from 0 to 1, where we count the fraction of Type I and II errors.

We decide on investigating the whole phase space, to get an overall picture of the behaviour of the Type I and II errors. It will be an effortful task, but with a profitable result.

2.2 The estimated error rate, $\widehat{\text{Err}}$

For a certain point (β, r) in phase space we compute an empirical counterpart to the left hand side of (8), namely the estimated error rate ($\widehat{\text{Err}}$), which is an empirical probability:

$$\widehat{\text{Err}} = (\#\text{Falsely rejected } H_0 + \#\text{Falsely rejected } H_1)/\#\text{Simulations}, \quad (18)$$

where we simulate H_0 and H_1 equally many times. Clearly, a small $\widehat{\text{Err}}$ (close to 0) indicates success of detection, while a large (close to 1) indicates failure. An equvalent way of seeing (19) is as the empirical probability of Type I and II errors.

A falsely rejected H_0 occurs when we simulate under H_0 but unexpectedly (compare with (12)) $HC^*_p \geq \sqrt{2(1+\delta)\log\log p}$, (or with b instead of p). Conversely, H_1 is falsely rejected when under it, surprisingly, $HC^*_p < \sqrt{2(1+\delta)\log\log p}$.

2.3 Calibrating parameters

We start by specifying p and b. The choice of these parameters is essentially constrained by the computational time of the algorithms.

Comparing the $\mathcal{N}$-, χ^2_2-, χ^2_{10}-, and χ^2_{30}-cases, we could either choose a fixed p with a decreasing b as p_0 increases, or we could choose to fix $p = b$. The first alternative reflects reality, but, on the other hand, $p = b$ compares the cases under more similar conditions.

With the time constraints mentioned, we choose $p = b = 10^4$, which is small but still guarantees stable results.

Next, we choose α_0 in the HC objective function, (11). In several articles, see e.g. Donoho & Jin (2008) and (2009), α_0 is chosen as 0.1. However, that choice suits only the sparse case. Since we also deal with the dense case, and, further, need to optimize the computational time of our algorithms, we choose α_0 as a function of ϵ. After careful studies of the location of the peak of the HC objective function we come to choose $\alpha_0 = 15\epsilon$, however with the constraint $0.05 < \alpha_0 < 0.8$. This choice will essentially guarantee that the peak of the HC objective function will be captured.

The optimal value of δ is the one which minimizes $\widehat{\text{Err}}$ over the range $\delta = 0.2 \times [1, 2, \ldots, 10]$. However, there is another way of choosing the critical value in (12), which we could test in future experiments. Cai *et al* (2011) suggest to control the Type I error at a prescribed level α. We then simulate HC_p^*-scores under the null hypothesis N times, where $N\alpha \gg 1$ (e.g. $N\alpha = 50$). We let $t(\alpha)$ be the top α percentile of the simulated scores, and use $t(\alpha)$ as the critcal value. Cai *et al* argue that critical values determined this way are usually much more accurate than $\sqrt{2(1+\delta)\log\log p}$.

3 Simulations

All computations are done using R (version 2.15). We use a strong computer with a 10-core processor.

3.1 Developing algorithms

We investigate $\widehat{\text{Err}}$ at equidistant points throughout the phase space, with 0.01 distance between points in the $\mathcal{N}$-case and 0.03 in the χ^2_ν-cases. In the algorithms below we describe how we evaluate $\widehat{\text{Err}}$ for one such point.

The first step is to evaluate $\widehat{\text{Err}}$ on single time for one (β, r)-point. We choose #Simulations $= m_1 = 100$. The procedure is described in Algorithm 1 below.

Second, Algorithm 1 is repeated $m_2 = 100$ or 500 times. From these m_2 replications we need to take a specially designed average, because we observe that the standard mean is not optimal. In Figure 5 we show one out of many examples of a histogram of $m_2 = 50$ replications of Algorithm 1. In Algorithm 2 below we describe how the average is taken.

Now, we observe that in the upper left corner of the sparse half of phase space (roughly above the line $r = \beta + 0.2, 0.5 \le \beta \le 0.8$), we frequently get an unwelcome extreme left tail of the HC objective function. It is a natural behaviour since the signals there are very strong, making the denominator of (11) very small, for a small i. In Algorithm 3 below we show one plausible way to cut this tail off.

Algorithm 1. Find $\widehat{\text{Err}}$ for one (β, r)-point.

Input: β; r; distr $= \mathcal{N}$, χ^2_2, χ^2_{10}, or χ^2_{30}; $p = b = 10^4$; $m_1 = 100$.

Output: $\widehat{\text{Err}}$

%if distr $= \chi^2_{p_0}$ exchange p for b everywhere below

%in first for-loop simulate H_0 and H_1 m_1 times respectively

for $i = 1$ *to* m_1 **do**

  %simulate H_0:

  draw p features from central distribution

  calculate HC^*_p

  listH0 = listH0 + HC^*_p

  %simulate H_1:

  draw $(1-\epsilon)p$ features from central + ϵp from non-central distribution

  calculate HC^*_p

  listH1 = listH1 + HC^*_p

end

$\delta = 0.2 \times [1, \ldots, 10]$

for $i = 1$ *to length* (δ) **do**

  critval $= \sqrt{2(1 + \delta[i]) \log \log p}$

  TypeI = count elements in listH0 $\geq$ critval

  TypeII = count elements in listH1 $<$ critval

  Error = (TypeI + TypeII)

  listError = listError + Error

end

%minimize listError, i.e. choose optimal δ which gives lowest $\widehat{\text{Err}}$

$\widehat{\text{Err}} = \min(\text{listError})/m_1$

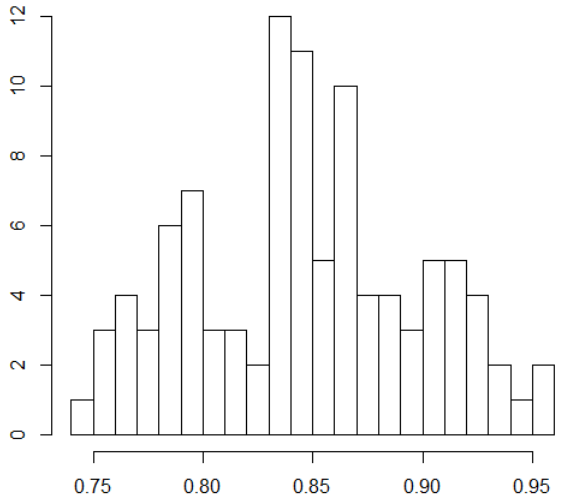


Figure 5: Histogram of $m_2 = 50$ replications of Algorithm 1. On x-axis: $\widehat{\text{Err}}$ (for $(\beta, r) = (0.2, 0.3)$). The pattern differs slightly for different (β, r). In Algorithm 2 we take a specially designed average of the m_2 replications.

Algorithm 2. Take a specially designed average of m_2 repetitions of $\widehat{\text{Err}}$.

Input: listErr, a list of m_2 $\widehat{\text{Err}}$'s from m_2 replications of Algorithm 1.
Output: (Average of) $\widehat{\text{Err}}$.
$\widehat{\text{Err}}$ = mean{
  mean(listErr),
  median(listErr),
  mean(three most frequent values in listErr)}

Algorithm 3. Remove unwelcome extreme left tail of $HC_p(i)$.

Input: $HC_p(i)$.
Output: $HC_p(i)$ with removed extreme left tail.
%exchange p for b in the χ^2-case
%identify tail in the if condition
if $HC_p(1) > 10$ *or*
$[HC_p(1) + HC_p(2) + HC_p(3)]/3 > HC_p(i), \forall\{4 \le i \le 50\}$ **then**
  i_{tail} = minimum i s.t. $HC_p(i) < \text{median}(HC_p(i),\ 1 \le i \le 100)$
end
return $HC_p(i), i_{tail} \le i \le p$

3.2 Interpretation of results

All graphs referred to in Section 3.2 are found in Appendix.

3.2.1 $\widehat{\text{Err}}$ throughout the phase space

Figure 6-9 show that the empirical success region is a lot smaller than we would anticipate from the detection boundary in (9). Further, it is smaller in the χ^2_ν-case than in the $\mathcal{N}$-case, and shrinks when ν grows. In the dense regime, χ^2-case, the empirical detection boundary appears to be "tilted", the slope of it is flatter than in the $\mathcal{N}$-case.

Considering the overall small size of the empirical success regions, we raise the question of what the behaviour of p (and b) looks like. Can it alone explain why higher criticism does not perform as we would expect?

Further, in Figure 7 we can observe some sort of shift at $\beta \approx 0.07$, and in Figure 9 some sort of shift at $\beta \approx 0.9$. The latter obsevation we can probably explain by high sensitivity to p (or b) under extremely sparse circumstances. With $\beta \geq 0.9$ and $p = 10^4$ we have $\epsilon p \leq 3$ informative signals. The former, however, needs further investigation (we do not entirely exclude a systematic error).

Let us also point out that we could obviously reduce the set of (β, r)-points at which computations are done, in future experiments. In Figure 6-9 we could exclude areas where $\widehat{\text{Err}}$ is constant, e.g. the upper-right triangle in the dense regime, and the lower-down triangle in the sparse.

3.2.2 $\widehat{\text{Err}}$ at the detection boundary

In Figure 10 we have estimated $\widehat{\text{Err}}$ wandering from left to right on the very detection boundary, in the χ^2-cases using the detection boundary in (9a).

We clearly see that our data follows some patterns that are magnified in the $\mathcal{N}$-case. The point sets cannot be fitted with polynomes. Nonetheless, Figure 10 demonstrate the inhomogeniety of the empirical success region, saying that it is safer to be in the middle of the β-range for dense and sparse cases respectively. In other words, we observe higher error probabilities in extremely sparse or weak situations, at the ends of the detection boundary. Note that, again, in the extremely sparse case, $\beta \geq 0.9$, we have (by completary experiments not presented in the graphs) indications of very high sensitivity to p (or b).

3.2.3 Exploring the asymptotic properties of ideal HC

Finally, we study the asymptotic behaviour of ideal HC, which need less computational times than the empirical HC.

We again use the criteria in (12), where we now choose $\delta = 0$, which logically will be optimal, in the sence that it generates the largest success region. We observe that if we elevate δ, the success region shrinks marginally. However, we want to observe the optimal success regions for different values of p, why $\delta = 0$ is a good choice here.

In Figure 11 we see the behaviour of $\hat{\rho_{\mathcal{N}}}^*(\beta)$ for $p = 10^4, 10^{10}, 10^{16}$. In the sparse case, clearly, between $p = 10^4$ and 10^{10} no radical changes are present, and it is not until $p \sim 10^{16}$ that the curve starts to fit the analytical detection boundary. In the dense case, however, we perform almost as good with $p = 10^{10}$ as with $p = 10^{16}$.

We also see in Figure 11 (a) that the line for $p = 10^4$ is very similar to the yellow color gradient ($\widehat{\text{Err}} = 0.5$) in Figure 6 (a). In the sparse case we do not have that similarity, which probably is an effect of the different parameter configurations in (6) and (7).

4 Discussion and scopes for future

To further study signal detection with higher criticism we would investigate the asymptotic behaviour more deeply, making it possible to explicitly express the empirical success region as a function of $(\beta, r; p, b, p_0, t_{\widehat{\text{Err}}})$, where $t_{\widehat{\text{Err}}}$ is the upper limit for success for the fraction of Type I and II errors.

Also, we would want to extend our experiment, studying the Type I and II errors separately. The two errors surely tend to zero at different rates, why data from H_0 will have a different empirical success region than data from H_1.

The empirical results from the simulations show that for p in the same order as for genomic applications (up to 10^4), and also for higher $p \sim 10^{11}$ as in some astronomic applications (Meinhausen & Rice (2006)), higher criticism is not successful near the detection boundary. Thus, in future work, we want to find ways to improve the performance close to the boundary. It would also be interesting to compare the performance of higher criticism (in both $\mathcal{N}$- and χ^2-cases) with other multiple comparing procedures, such as false discovery rate controlling, maximum- and range-comparing, and Bonferroni correction (see e.g. Donoho & Jin (2004)).

However, the main issue for future work is to develop theories for the χ^2_ν-case (and possibly for some other distributions as well, such as the general Gaussian or Subbotin distribution, see e.g. Donoho & Jin (2004), analog to those already developed for the $\mathcal{N}$-case. In these settings the detection boundary will be one focus.

From signal detection we may turn to the closely related problems of feature selection and supervised classification, where we select useful features for training of our classifier, thereby reducing complexity of the high-dimensional model. It is also worth noting that HC-based feature selection is much simpler than many other techniques, requiring no tuning parameter or cross-validation. We have clear indications (see Jin (2009)) on signal identification being even more limited than signal detection, why it would be also interesting to study the area in phase space where signals are *detectable* but not *identifiable*.

Finally, we want to run the procedures on real data with n observations, each with p features. Then, it can be very advantageous to first estimate β and r for at least some of the observations, to ensure that data belongs to the success region. There are techniques (however complicated) for estimating β and r in

the $\mathcal{N}$-case (see Meinhausen & Rice (2006)), but we will want to develop analog techniques for the χ^2-case.

A Appendix

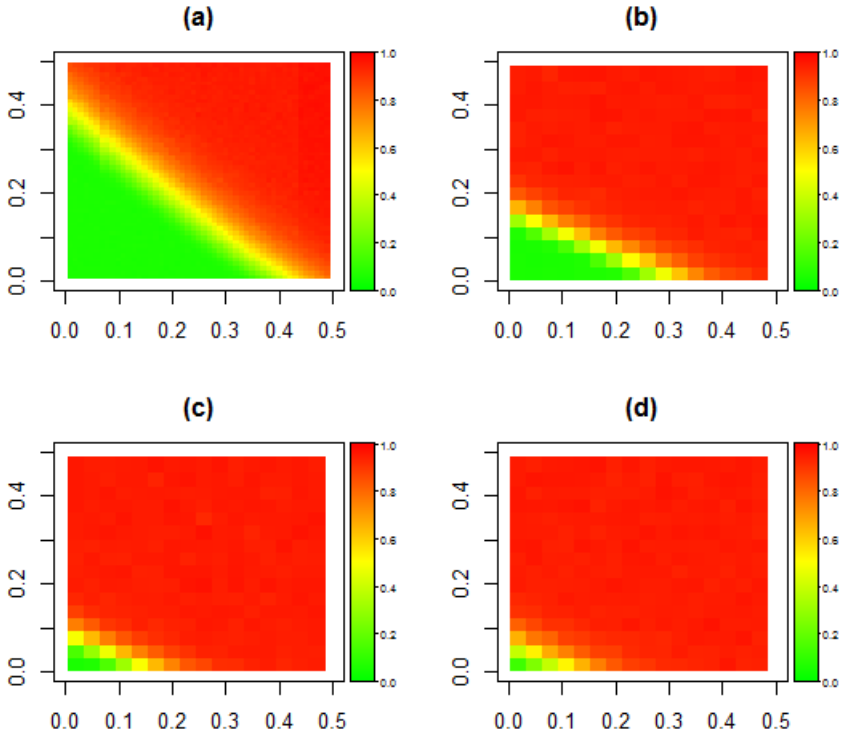
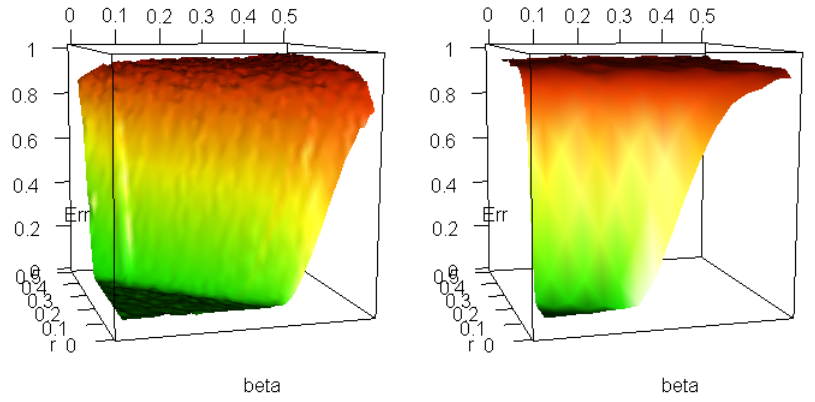
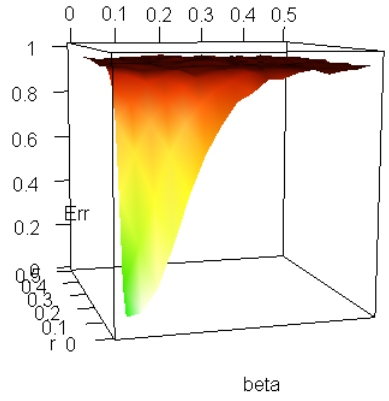


Figure 6: Color intensity map showing behaviour of $\widehat{\text{Err}}$ from 0 (green) to 1 (red), as a function of β and r, in the dense regime. (a) $\mathcal{N}$-case. (b) χ^2_2-case. (c) χ^2_{10}-case. (d) χ^2_{30}-case. In $\mathcal{N}$-case: $m_1 = 100$, $m_2 = 500$, 50^2 points in graph. In χ^2-cases: $m_1 = 100$, $m_2 = 100$, 17^2 points in graph.

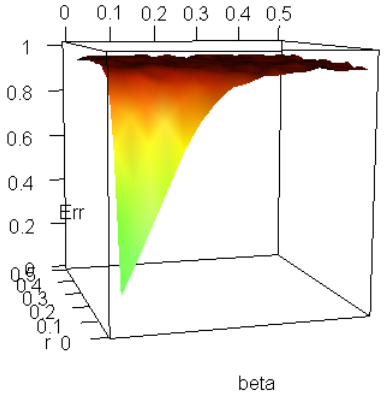


(a)

(b)



(c)



(d)

Figure 7: Surfaces corresponding to graphs in Figure 6. $\widehat{\text{Err}}$ as a function of β and r, in the dense regime.

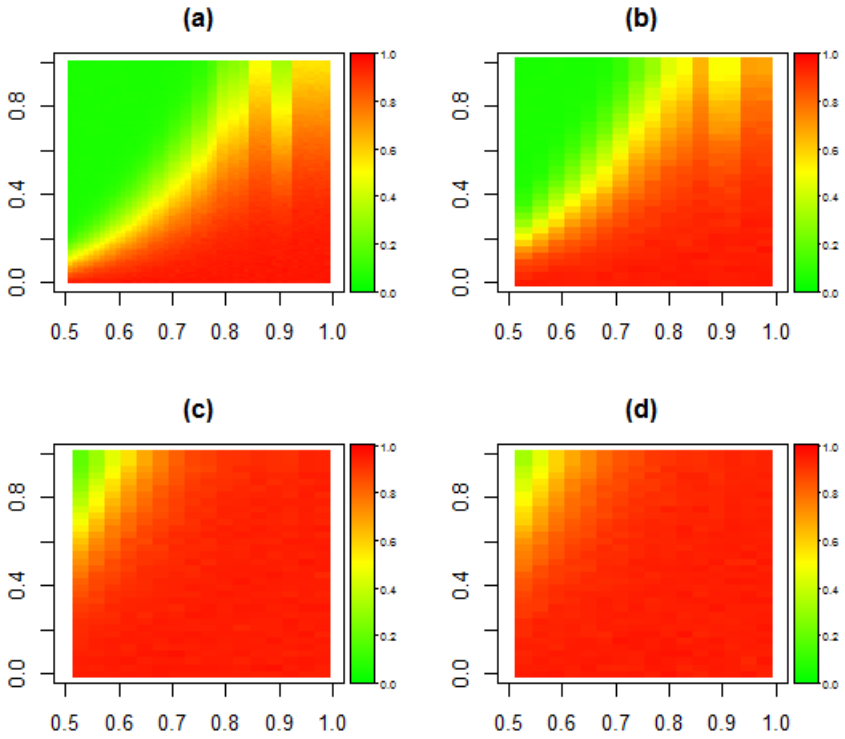


Figure 8: Color intensity map showing behaviour of $\widehat{\mathrm{Err}}$ from 0 (green) to 1 (red), as a function of β and r, in the sparse regime. (a) $\mathcal{N}$-case. (b) χ^2_2-case. (c) χ^2_{10}-case. (d) χ^2_{30}-case. In $\mathcal{N}$-case: $m_1 = 100$, $m_2 = 500$, 99×50 points in graph. In χ^2-cases: $m_1 = 100$, $m_2 = 100$, 33×17 points in graph.

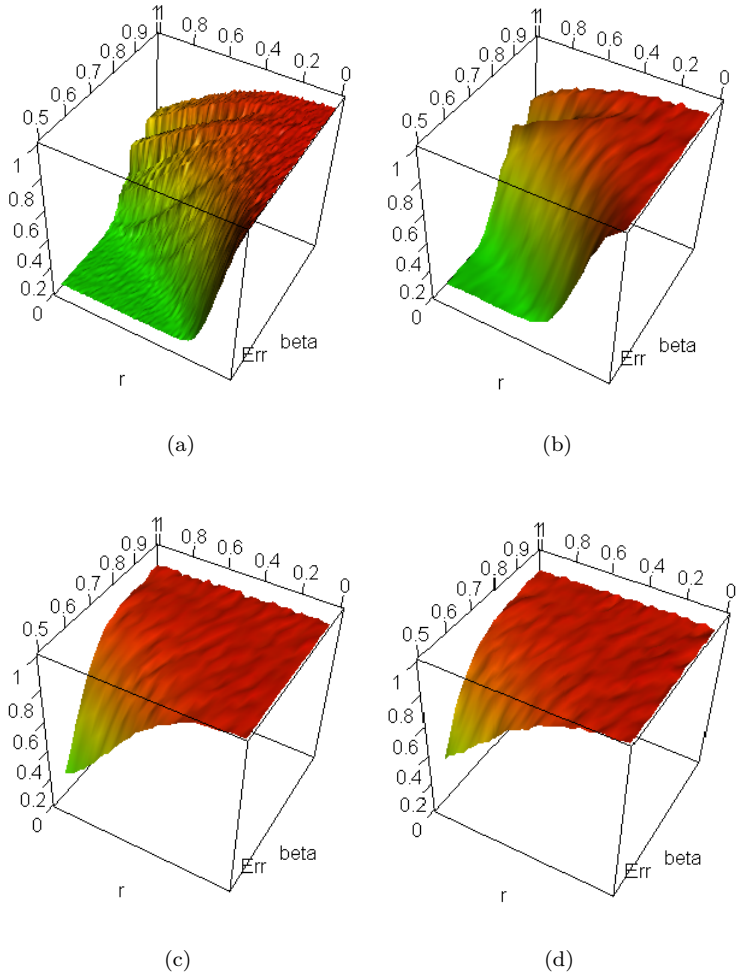
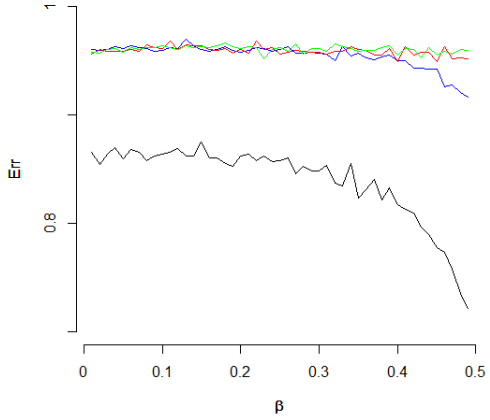
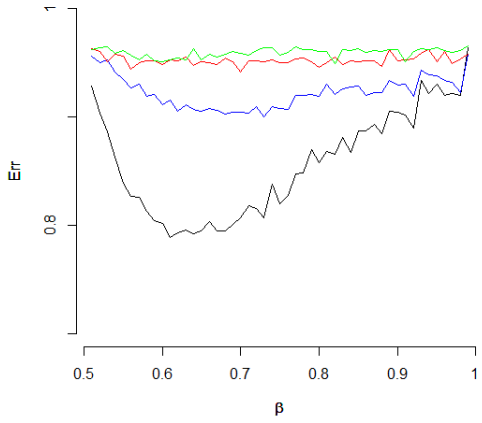


Figure 9: Surfaces corresponding to graphs in Figure 8. $\widehat{\text{Err}}$ as a function of β and r, in the sparse regime.

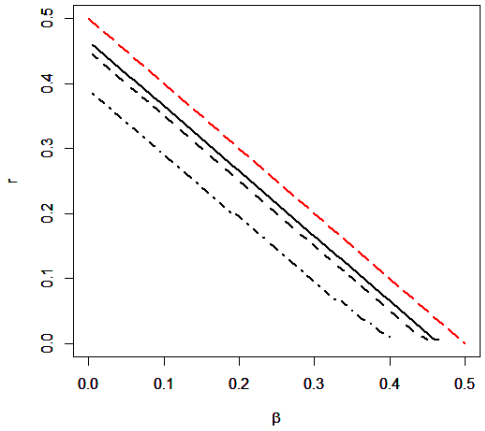


(a)

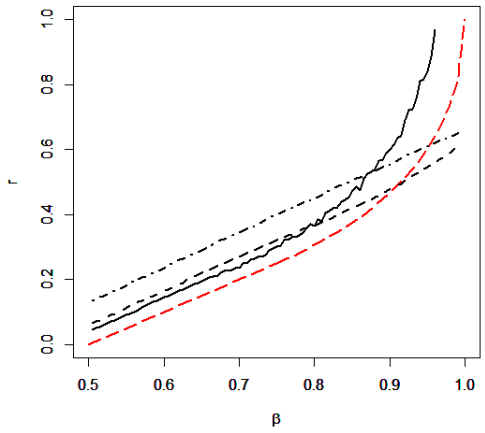


(b)

Figure 10: $\widehat{\mathrm{Err}}$ on the boundary (9), computed at equidistant points in (a) dense regime, $0 < \beta < 1/2$; (b) sparse regime $1/2 < \beta < 1$. From bottom: $\mathcal{N}$-case (black), χ^2_2-case (blue), χ^2_{10}-case (red), χ^2_{30}-case (green). $n_1 = 100$, $n_2 = 500$.



(a)



(b)

Figure 11: Detection boundary in (a) dense, (b) sparse $\mathcal{N}$-case using ideal HC. Red dashed line: analytical detection boundary; (-·-): $p = 10^4$; (- - -): $p = 10^{10}$; (—): $p = 10^{16}$.

References

[1] Cai T. T., Jeng X., Jin X. J. (2011): Optimal detection of heterogeneous and heteroscedastic mixtures. *Royal Statistical Society.*

[2] Donoho D., Jin J. (2004): Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics.*

[3] Donoho D., Jin J. (2008): Higher criticism thresholding: Optimal feature selection when useful features are rare and weak.. *PNAS.*

[4] Donoho D., Jin J. (2009): Feature selection by higher criticism thresholding achieves the optimal phase diagram. *The Royal Society.*

[5] Jin J. (2009): Impossibility of successful classification when useful features are rare and weak. *PNAS.*

[6] Meinhausen N., Rice J. (2006): Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics.*

[7] Pavlenko T., Björkström A., Tillander A. (2012): Covariance structure approximation via gLasso in high-dimensional supervised classification. *Journal of Applied Statistics.*

[8] Pawitan Y., Bjhle J., Amler L., Borg A.L., Egyhazi S., Hall P., Han X., Holmberg L., Huang F., Klaar S., Liu E.T., Miller L., Nordgren H., Ploner A., Sandelin K., Shaw P.M., Smeds J., Skoog L., Wedren S., Bergh J. (2005): Gene expression profiling spares early breast cancer patients from adjuvant therapy: Derived and validated in two population-based cohorts. *Breast Cancer.*