Statistica Sinica **26** (2016), 1-34
doi:http://dx.doi.org/10.5705/ss.2014.138

Invited Review Article

RARE AND WEAK EFFECTS IN LARGE-SCALE INFERENCE: METHODS AND PHASE DIAGRAMS

Jiashun Jin and Zheng Tracy Ke

Carnegie Mellon University and University of Chicago

Abstract: Often when we deal with 'Big Data', the true effects we are interested in are *Rare and Weak* (RW). Researchers measure a large number of features, hoping to find perhaps only a small fraction of them to be relevant to the research in question; the effect sizes of the relevant features are individually small so the true effects are not strong enough to stand out for themselves.

Higher Criticism (HC) and Graphlet Screening (GS) are two classes of methods that are specifically designed for the Rare/Weak settings. HC was introduced to determine whether there are any relevant effects in all the measured features. More recently, HC was applied to classification, where it provides a method for selecting useful predictive features for trained classification rules. GS was introduced as a graph-guided multivariate screening procedure, and was used for variable selection.

We develop a theoretical framework where we use an *Asymptotic Rare and Weak* (ARW) model simultaneously controlling the size and prevalence of useful/significant features among the useless/null bulk. At the heart of the ARW model is the so-called *phase diagram*, which is a way to visualize clearly the class of ARW settings where the relevant effects are so rare or weak that desired goals (signal detection, variable selection, etc.) are simply impossible to achieve. We show that HC and GS have important advantages over better known procedures and achieve the optimal phase diagrams in a variety of ARW settings.

HC and GS are flexible ideas that adapt easily to many interesting situations. We review the basics of these ideas and some of the recent extensions, discuss their connections to existing literature, and suggest some new applications of these ideas.

Key words and phrases: Classification, control of FDR, feature ranking, feature selection, graphlet screening, hamming distance, higher criticism, large-scale inference, rare and weak effects, phase diagram, sparse precision matrix, sparse signal detection, variable selection.

1. Introduction

We are often said to be entering the era of 'Big Data'. High-throughput devices measure thousands or even millions of different features per single subject on a daily basis. Such an activity is the driving force of many areas of science and technology, including a new branch of statistical practice which Efron (2011) calls *Large-Scale Inference* (LSI).

In many high-throughput data sets, the relevant effects are *Rare and Weak* (RW). The researchers expect that only a small fraction of these measured features are relevant for the research in question, and the effect sizes of the relevant features are individually small. The researchers do not know in advance which features are relevant and which are not, so they choose to measure all features within a certain range systematically and automatically, hoping to identify a small fraction of relevant ones in the future.

Examples include, but are not limited to, Genome-Wide Association Study (GWAS) and deep sequencing study, where we are in the so-called "large-p small-n" paradigm, with p being the number of SNPs and n the number of subjects. As technology on data acquisition advances, we are able to measure increasingly more features per subject. However, the number of relevant features do not grow proportionally, so the relevant effects are sparse; in addition, since n is usually not as large as we wish, the effect sizes of the relevant features (in the summary statistics, e.g., two-sample t-tests) are individually small.

Effect *rarity* is a useful hypothesis proposed as early as 1980's by Box and Mayer (1986). Later, this hypothesis was found to be valid in many applications (e.g., wavelet image processing Donoho and Johnstone (1994), cosmology and astronomy Jin et al. (2005), genetics and genomics Tibshirani et al. (2002)) and had inspired a long line of researches, where the common theme was to exploit sparsity (e.g., Donoho and Johnstone (1994, 1995)).

However, these works have been largely focused on the regime where the effects are rare but are individually strong (Rare/Strong), with limited attention to the more challenging Rare/Weak regime; the latter contains many new phenomena which we have not seen, to discover which, we need new methods and new theoretical frameworks (the notions of Rare/Strong and Rare/Weak effects are made precise in Section 2.3).

It is instructive to consider a *two-group study*. For a specific disease under consideration, suppose we have two groups of subjects, a treatment group and a control group, each subject being measured on the same set of features. In many such studies, signals are Rare and Weak (we call a feature a signal if it is relevant to the disease and a noise otherwise). In this paper, we investigate two interconnected LSI problems in the Rare/Weak regime.

- *Sparse signal detection*. We are interested in deciding whether there is any difference between two groups. In the Rare/Weak settings, the inter-group difference for any single relevant feature is not significant enough, so we have to combine the strengths of these features.
- *Sparse signal recovery*. We are interested in separating relevant features from the overwhelmingly more irrelevant ones (i.e., variable selection).

1.1. Higher criticism (HC) and graphlet screening (GS)

HC and GS are recent methods for detecting and recovering sparse signals, respectively, specifically designed for the Rare/Weak settings.

The term of "Higher Criticism" was coined by Tukey (1976, 1989) with the following story, in the context of multiple testing. A young scientist administers 250 uncorrelated tests, out of which 11 are signifiant at the level of 5%, and he is very excited about the findings. However, before he makes a big deal about it, a senior researcher tells him that, even in the purely null case, one would expect 12.5 significances, and finding only 11 significances is in fact disappointing. Tukey proposed HC as "a second-level significance testing"

$$HC_{250,0.05} = \frac{\sqrt{250}[(\text{Fraction significant at level } 0.05) - 0.05]}{\sqrt{0.05 \times 0.95}}.$$

If the young researcher really "had discovered something", this score should be large (say, ≥ 2). However, in Tukey's example, $HC_{250,0.05} = -0.43$, suggesting that the overall body of the evidence is consistent with the null of "no evidence".

Here, the problem of interest is related to the problem of multiple testing Benjamini and Hochberg (1995), but is different. Given p different (component) null hypotheses $H_1, \ldots, H_p$, the goal of multiple testing is to decide which null are true and which are not. Here, the goal is to test the (joint) null hypothesis that all (component) null are true against the alternative that a small fraction of them is untrue.

In modern multiple testing settings, p (number of tests) is usually large and the component tests may be rare and individually weak (not strong enough to stand out by itself). In such settings, it is desirable to generalize Tukey's HC beyond a single significance level of $\alpha = 0.05$. In light of this, Donoho and Jin (2004) proposed the following Higher Criticism statistic:

$$HC_p^* = \max_{0 \leq \alpha \leq \alpha_0} HC_{p,\alpha}, \quad \text{where } HC_{p,\alpha} = \frac{\sqrt{p}[(\text{Fraction significant at level } \alpha) - \alpha]}{\sqrt{\alpha(1-\alpha)}},$$

where $\alpha_0 \in (0,1)$ is a tuning parameter. Under the joint null, $HC_{p,\alpha}$ and $HC_{p,1-\alpha}$ have the same distribution, so it is unnecessary to consider the case $\alpha > 1/2$. For this reason, we usually set $\alpha_0 = 1/2$. The Higher Criticism statistic HC_p^* can be computed efficiently as follows.

- For each $1 \leq i \leq p$, obtain the individual P-value π_i.
- Sort the P-values in the ascending order: $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$.
- The Higher Criticism statistic HC_p^* can be equivalently written as

$$HC_p^* = \max_{\{1 \leq i \leq \alpha_0 p\}} HC_{p,i}, \text{ where } HC_{p,i} \equiv \frac{\sqrt{p}[i/p - \pi_{(i)}]}{\sqrt{\pi_{(i)}(1-\pi_{(i)})}}. \tag{1.1}$$

Later, we will discuss several variants of HC. To distinguish one from the other, we call that in (1.1) the *Orthodox HC (OHC)*. OHC is known to be heavy-tailed. To alleviate the problem, we recommend the following modified version Donoho and Jin (2004):

$$HC_p^+ = \max_{\{1\le i\le \alpha_0 p:\ \pi_{(i)}>1/p\}} HC_{p,i}. \tag{1.2}$$

Graphlet Screening is a recent idea for variable selection Jin, Zhang, and Zhang (2014). Consider a linear regression of n samples and p variables:

$$W = X\beta + z, \qquad X = X_{n,p} = [x_1, x_2, \ldots, x_p], \qquad z \sim N(0, I_n). \tag{1.3}$$

We assume X is normalized so that the Gram matrix $G = X'X$ has unit diagonals, and that G is sparse in that each row of G has relatively few large entries. We are interested in the challenging Rare/Weak regime where only a small fraction of the entries of β is nonzero and each nonzero is individually small.

Univariate Screening (US) (also called marginal regression Genovese et al. (2012) or Sure Screening Fan and Lv (2008)) is a popular and computationally efficient variable selection approach: we project W to each column of the matrix X, one at a time, and select the variable where the coefficients (W, x_j) are large in magnitude. The main challenge US faces is "signal cancellation" Roeder and Wasserman (2009): due to correlations among the columns of X, $|E[(W, x_j)]|$ could be small even when $|\beta_j|$ is large.

Our proposal is *graph-guided multivariate screening* or *Graphlet Screeing (GS)* for short, an approach to overcoming the challenge of "signal cancelation" with only a modest increase in computational cost. The idea of GS can be illustrated as follows. Suppose that in order to alleviate the effects "signal cancellation", we decide to use bivariate screening instead of US. We may use Exhaustive Bivariate Screening (EBS): we project W to all possible pairs of columns of X, one at a time, and recruit both variables in the pair if the norm of projected coefficients is large. Unfortunately, this approach is both ineffecient and computationally infeasible: it includes too many, $O(p^2)$ number of, pairs for screening so it needs signals stronger than necessary for successful screening.

The key idea of GS is to recognize that it is only necessary to screen a pair of columns *together* when two columns are highly correlated (otherwise, screening two columns separately is adequate). The sparsity of $G = X'X$ dictates that, out of all $O(p^2)$ pairs, only a small fraction of them have two strongly correlated columns. As a result, GS is computationally much less expensive than EBS.

More generally, fixing a relatively small integer $m_0 > 1$, GS can be viewed as a graph-guided m_0-variate screening. Fix a threshold $\delta > 0$. Let $\mathcal{G}^{*,\delta} = (V, E)$ be the graph where $V = \{1, \ldots, p\}$ and there is an edge between i and j if and only if $|G(i, j)| \ge \delta$. Let $d_p^* = d_p^*(\delta, G)$ be the maximal degree of $\mathcal{G}^{*,\delta}$, and let $\mathcal{A}^{*,\delta}(m_0) = \mathcal{A}^{*,\delta}(m_0, G) = \{\text{all connected subgraphs of } \mathcal{G}^{*,\delta} \text{ with size} \le m_0\}$.

- For each $1 \le m \le m_0$ and $\mathcal{I} \in \mathcal{A}^{*,\delta}(m_0)$, obtain the projection of W from R^n to the space $\{x_j : j \in \mathcal{I}\}$, denoted by $P^{\mathcal{I}}W$.
- For a threshold t that may depend on $(m, \mathcal{I}, X)$, retain all nodes in $\mathcal{I}$ if and only if $\|P^{\mathcal{I}}W\|$ exceeds the threshold (once a node is retained, it stays there until we finish screening for all $\mathcal{I} \in \mathcal{A}^{*,\delta}(m_0)$).

The target of GS is to retain almost all true signals while removing as much noise as possible. To filter out the false positives, we may need an additional step which we call *Graphlet Cleaning*; see details in Section 4.2.

GS is able to overcome the challenge of "signal cancellation" for it exploits the graphical structures in the design variables. As for the computational cost, since G is sparse, the maximum degree of $\mathcal{G}^{*,\delta}$—d_p^*—is relatively small given an appropriate choice of δ. It can be shown that $|\mathcal{A}^{*,\delta}(m_0)| \le Cp(ed_p^*)^{m_0}$ Jin, Zhang, and Zhang (2014), so the computational cost of GS is higher than that of US only by a factor of $C(ed_p^*)^{m_0}$. Of course, in the simplest case where G is diagonal, "signal cancellation" does not pose a challenge, and GS reduces to US.

1.2. Asymptotic rare and weak model and phase diagrams

We evaluate HC and GS by developing a new theoretical framework using the *Asymptotic Rare and Weak* (ARW) model, simultaneously controlling the signal prevalence and signal strengths. We visualize the ARW model with phase diagrams. The phase space refers to the two-dimensional space with axes simultaneously quantifying the signal prevalence and signal strengths. The phase space partitions into several subregions, where the desired inference (signal detection, variable selection, etc.) has distinctly different results; because of the partition of the phase space, we call it the phase diagram.

Phase diagram can be viewed as a new criterion for optimality which is particularly appropriate for the ARW model. Given a problem, different methods may have different phase diagrams, characterizing the subregions where they succeed and where they fail. When a method partitions the phase space in exactly the same way as the optimal methods, we say it achieves the optimal phase diagram. We show that in a wide variety of settings, HC and GS achieve the optimal phase diagrams for signal detection and signal recovery, respectively. In many of such settings, other methods (such as FDR-controlling methods Benjamini and Hochberg (1995), L^0/L^1-penalization methods Donoho and Stark (1998); Tibshirani (1996)) do not achieve the optimal phase diagrams.

HC and GS are flexible ideas and can be applied to many interesting settings. They are not tied to the ARW model and their advantages over existing methods remain in much broader settings.

1.3. Roadmap and highlights

In Section 2, we discuss the ARW model and phase diagrams for the problems of sparse signal detection and sparse signal recovery, respectively. In Section 3, we discuss the achievability of the phase diagrams in the presence of uncorrelated noise. We show that HC and the well-known method of Hard Thresholding achieve the optimal phase diagram for signal detection and signal recovery, respectively. We also review the recent developments on HC. In Section 4, we discuss the achievability for the more challenging case of correlated noise. For signal detection, we develop HC into the more sophisticated Innovated HC, and for signal recovery, we use GS. In Section 5, we suggest some new applications of HC and GS, supported by some preliminary numerical studies. In Section 6, we extend HC as a feature selection method in the context of classification, and address the classification phase diagram as well as the optimality of HC.

Our study exposes several noteworthy ideas highlighted below (other noteworthy ideas include but are not limited to Lemma 1 and Remarks 4, 12).

- Innovated Transformation (IT) is the key to many methods (e.g., Innovated HC and GS; see also Jin (2012), Fan, Jin, and Yao (2013)) that exploit graphic structures. Among all possible transformations, IT yields the largest (post-transformation) Signal-to-Noise Ratios (SNR) simultaneously at all (pre-transformation) signal sites, and so it is preferred; see Section 4.1.
- Contrary to what we might have expected, the optimal phase diagrams do not critically depend on the local graphic structures, and remain the same across a wide range of cases. Therefore, to have procedures that achieve the optimal phase diagram, we must exploit local graphic structures.
- The L^0/L^1-penalization methods are well-known approaches to variable selection, but they do not adequately exploit local graphic structures as GS does. The primary focus of these methods is usually to fully recover Rare/Strong signals, and in the more challenging Rare/Weak settings, they do not achieve the optimal phase diagram, even in very simple settings and even when the tuning parameters are ideally set; see Sections 4.2−4.4.
- In classification, a prevailing idea is to select a few important predictive features so that the (feature) FDR Benjamini and Hochberg (1995) is small. Recent studies reveal something very different: in some Rare/Weak settings, we must select features in a way so that the FDR is very high, so that we are able to include almost all useful features for classification. See Section 6.

2. The ARW Model and Phase Diagrams

In this section, we discuss phase diagrams and the watershed phenomena associated with the problems of signal detection and signal recovery. Discussions

on the achievability of the phase diagrams are long, so they are deferred to Sections 3−4. See Section 2.4 for our plan for the discussions on the achievability.

The ARW model was first proposed by Donoho and Jin (2004) for sparse signal detection. More recently, it was extended to more complicated forms Hall and Jin (2010) and to different settings including classification Donoho and Jin (2009); Ingster, Pouet, and Tsybakov (2009); Ji and Jin (2011); Jin (2009); Jin, Ke, and Wang (2015); Jin and Wang (2014); Jin, Zhang, and Zhang (2014).

In this paper, we use a version of the ARW model that is simple for presentation, yet contains all important ingredients associated with the major insights exposed in the above references. Consider a Stein's p-normal means model

$$Y = \beta + z, \qquad z \sim N(0, \Sigma), \tag{2.1}$$

where $\Sigma = \Sigma_{p,p}$ is the covariance matrix. Denote the precision matrix by

$$\Omega = \Omega_{p,p} = \Sigma_{p,p}^{-1}.$$

For simplicity, we usually drop subscripts 'p, p'. We assume Ω is sparse in the (strict) sense that each row of Ω has relatively few nonzeros. Such an assumption is only for simplicity in presentation; see Hall and Jin (2010); Jin, Zhang, and Zhang (2014) for discussions on more general Ω. Model (2.1) may arise from many applications, including the following.

- *Two-group study.* In the two-group study in Section 1, $\{Y_j, 1 \le j \le p\}$ can be viewed as the two-sample t-statistic associated with the j-th feature. In many such studies (e.g., Genetic Regulatory Network (GRN)), the precision matrix is sparse Peng et al. (2010).
- *Linear models with random designs.* Given $W \sim N(X\tilde{\beta}, I_n)$, where the rows of X are iid samples from the p-variate normal $N(0, \Omega)$ for a sparse (in strict sense) matrix Ω. Such settings can be found in Compressive Sensing Donoho (2006); Donoho, Maleki, and Montanari (2009) or Computer Security Fienberg and Jin (2012); Jin, Zhang, and Zhang (2014), where $\Omega = I_p$. Letting $\widetilde{W} = (1/\sqrt{n})X'W$ and $\beta = \sqrt{n}\tilde{\beta}$, then approximately $\widetilde{W} \sim N(\Omega\beta, \Omega)$, which is equivalent to model (2.1); the connection is solidified in Jin, Zhang, and Zhang (2014).

We assume Ω is known, as our primary goal is to investigate how the graphic structures of Ω affect the constructions of the optimal methods and optimal phase diagrams. Such an assumption is valid in many applications. For example, in the linear model, Ω plays the role of the Gram matrix which is known to us. When Ω is unknown but is sparse, it can be estimated by many recent algorithms, such as the glasso Friedman, Hastie, and Tibshirani (2007). The gained insight

here is readily extendable to the case where Ω is unknown but can be estimated reasonably well, as only large entries of Ω have a major influence on the problems of interest.

We choose a somewhat unconventional normalization such that

$$\Omega(i,i) = 1, \qquad 1 \le i \le p. \tag{2.2}$$

Let d_p^* be the maximum number of nonzeros in the rows of Ω:

$$d_p^* = d_p^*(\Omega) = \max_{1\le i\le p}\big\{\#\{1 \le j \le p : \Omega(i,j) \neq 0\}\big\}.$$

We assume $d_p^*(\Omega)$ *grows slowly enough*:

$$d_p^*(\Omega)p^{-\delta} \to 0, \qquad \text{for any fixed } \delta > 0. \tag{2.3}$$

Fixing $\epsilon \in (0,1)$ and $\tau > 0$, we model the vector β by

$$\beta_i \overset{iid}{\sim} (1-\epsilon)\nu_0 + \epsilon\nu_\tau, \qquad 1 \le i \le p, \tag{2.4}$$

where ν_a is the point mass at a. We are primarily interested in the case where ϵ is small and τ is small or moderately large, so that the signals (i.e., nonzero entries of β) are Rare/Weak. In our asymptotic framework, we let p be the driving asymptotic parameter, and tie (ϵ,τ) to p through fixed parameters ϑ and r:

$$\epsilon = \epsilon_p = p^{-\vartheta}, \qquad \tau = \tau_p = \sqrt{2r\log(p)}, \qquad 0 < \vartheta < 1, \;\; r > 0. \tag{2.5}$$

As p grows, the signals become increasingly sparser. To counter this effect, we have to let the signal strength parameter τ grow to ∞ slowly, so that the problems of detection and signal recovery are non-trivial. Let $S_p(\beta)$ and $s_p(\beta)$ be the support of β and the number of signals, respectively:

$$S_p(\beta) = \{1 \le i \le p : \beta_i \neq 0\}, \qquad s_p(\beta) = |S_p(\beta)|.$$

It is seen that with overwhelming probability,

$$s_p(\beta) \sim p\epsilon_p = p^{1-\vartheta}. \tag{2.6}$$

Definition 1. We call (2.1)–(2.5) the Asymptotic Rare/Weak model $ARW(\vartheta, r, \Omega)$.

See Genovese et al. (2012); Hall and Jin (2008, 2010); Ji and Jin (2011); Jin, Zhang, and Zhang (2014); Ke, Jin, and Fan (2014) for the ARW model in more general forms. The ARW model is subtle even when $\Omega = I_p$; see Donoho and Jin

(2004) for example. See Section 2.3 for remarks on the concepts of Rare/Weak and Rare/Strong.

2.1. Detecting rare and weak signals

We formulate the sparse signal detection problem as a hypothesis testing problem, where we test a *joint null hypothesis* that all β_i's are 0:

$$H_0^{(p)}: \qquad \beta = 0, \tag{2.7}$$

against a specific complement of the joint null:

$$H_1^{(p)}: \qquad \beta \text{ satisfies a Rare and Weak model (2.4)}-\text{(2.5)}. \tag{2.8}$$

In the simplest case where $\Omega = I_p$, we have that $Y_i \overset{iid}{\sim} N(0,1)$ under $H_0^{(p)}$ and that $Y_i \overset{iid}{\sim} (1-\epsilon_p)N(0,1) + \epsilon_p N(\tau_p, 1)$ under $H_1^{(p)}$, so the testing problem (2.7)−(2.8) is also the problem of detecting Gaussian mixtures Donoho and Jin (2004).

We have a watershed phenomenon: in the two-dimensional phase space $\{(\vartheta, r) : 0 < \vartheta < 1, r > 0\}$, there is a curve that partitions the whole phase space into two regions where the testing problem (2.7)−(2.8) has distinctly different results. One may think that this curve depends on the off-diagonals of Ω in a complicated way. Somewhat surprisingly, this is not the case, and the off-diagonals of Ω do not have a major influence on the partition.

In detail, define the *standard phase function for detection*

$$\rho^*(\vartheta) = \begin{cases} 0, & 0 < \vartheta \le \frac{1}{2}, \\ \vartheta - \frac{1}{2}, & \frac{1}{2} < \vartheta \le \frac{3}{4}, \\ (1 - \sqrt{1-\vartheta})^2, & \frac{3}{4} < \vartheta < 1. \end{cases} \tag{2.9}$$

Theorem 1. *Fixing $\vartheta \in (0,1)$ and $r > 0$, consider a sequence of $ARW(\vartheta, r, \Omega)$ indexed by p. As $p \to \infty$, if $r < \rho^*(\vartheta)$, then for any sequence of tests that test $H_1^{(p)}$ against $H_0^{(p)}$, the sum of Type I and Type II errors tends to 1; if $r > \rho^*(\vartheta)$, then there is a test for which the sum of Type I and Type II errors tends to 0.*

In (2.9), $\rho^*(\vartheta) = 0$ when $\vartheta < 1/2$. This does not mean that $H_0^{(p)}$ and $H_1^{(p)}$ are asymptotically separable for any (ϵ_p, τ_p); it only means that they are asymptotically separable even when $\tau_p \ll \sqrt{\log(p)}$; see Donoho and Jin (2004).

When $\Omega = I_p$, Theorem 1 was proved by Donoho and Jin (2004) (see also Ingster (1997, 1999)). When $\Omega \neq I_p$, the second claim follows from Theorem 4, and the proof of the first claim is similar to that in (Fan, Jin, and Yao, 2013, Thm. 1.1) so we skip it. The proof requires subtle analysis of the Hellinger distance as well as the following lemma.

Lemma 1 (Chromatic Number). *Fix a $p \times p$ matrix Ω. If each row of Ω has no more than K nonzeros, then we can color indices $1, \ldots, p$ in no more than K different colors so that for any pair of indices i, j with the same color, $\Omega(i,j) = 0$.*

2.2. Recovering rare and weak signals

Consider again the ARW model $Y = \beta + z$, where $z \sim N(0, \Sigma)$ and (β, Ω) satisfying (2.2)−(2.5). The main interest here is to separate the nonzero entries of β from the zero ones (i.e., signal recovery or variable selection). For any estimator $\hat{\beta}$, we measure the errors by the Hamming distance $h_p(\hat{\beta}, \beta) = \sum_{i=1}^p P\{\mathrm{sgn}(\hat{\beta}_i) \neq \mathrm{sgn}(\beta_i)\}$, which is the sum of the expected number of signals that have been classified as noise and the expected number of noise that have been classified as signals. Here, $\mathrm{sgn}(u) = -1, 0, 1$ according to $u < 0$, $u = 0$, or $u > 0$. The minimax Hamming distance is then

$$\mathrm{Hamm}_p^*(\vartheta, r; \Omega) = \inf_{\hat{\beta}} \{h_p(\hat{\beta}, \beta)\}. \tag{2.10}$$

Definition 2. $L_p > 0$ denotes a generic multi-$\log(p)$ term that may change from occurrence to occurrence and satisfies that $L_p p^{\delta} \to \infty$ and $L_p p^{-\delta} \to 0$ for any $\delta > 0$, as $p \to \infty$.

The following theorem was proved by Ji and Jin (2011) (see also Jin, Zhang, and Zhang (2014); Ke, Jin, and Fan (2014)).

Theorem 2 (Lower bound). *Fixing $\vartheta \in (0,1)$ and $r > 0$, consider a sequence of $ARW(\vartheta, r, \Omega)$ indexed by p. As $p \to \infty$,*

$$\mathrm{Hamm}_p^*(\vartheta, r; \Omega) \begin{cases} \geq L_p \cdot p^{1-(\vartheta+r)^2/(4r)}, & r > \vartheta, \\ \sim p^{1-\vartheta}, & 0 < r < \vartheta. \end{cases} \tag{2.11}$$

For many sequences of Ω (especially $\Omega = I_p$), the lower bound is tight. We address this in Section 3.2 ($\Omega = I_p$) and in Section 4 ($\Omega \neq I_p$).

In principle, $\mathrm{Hamm}_p^*(\vartheta, r; \Omega)$ may depend on Ω in a complicated way. Still, for many sequences of $\Omega = \Omega_{p,p}$ (with careful calibrations), there is a constant $c = c(\vartheta, r; \Omega)$ depending on (ϑ, r) and the calibrations we choose for Ω such that

$$\mathrm{Hamm}_p^*(\vartheta, r; \Omega) = L_p p^{1-c(\vartheta, r; \Omega)}.$$

Examples include (a) $\Omega = I_p$, (b) Ω is the diagonal block-wise matrix as in Section 4.4, (c) a long-memory time series model and a change point model discussed in Ke, Jin, and Fan (2014) (see Remark 9). In all these cases, $c(\vartheta, r; \Omega) = 1$ gives the curve $r = \rho_{exact}^*(\vartheta; \Omega)$:

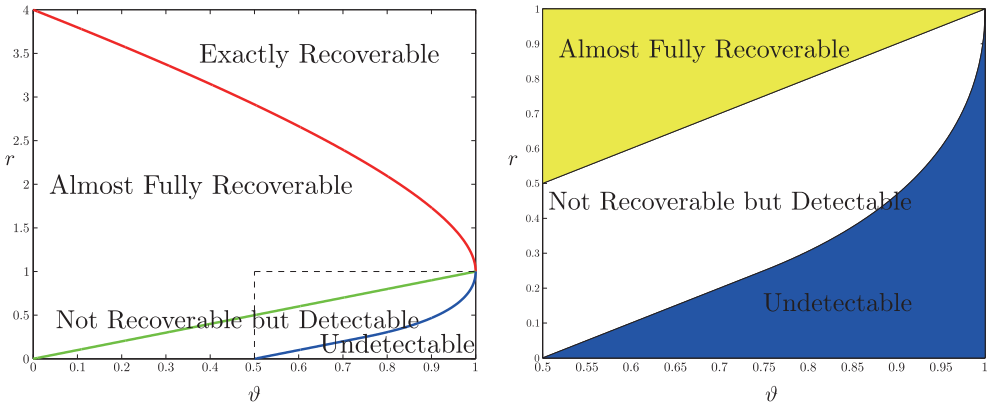

Figure 1. Phase diagrams ($\Omega = I_p$). Left: curves in red, green, and blue are $r = \rho^*_{exact}(\vartheta; I_p)$, $r = \vartheta$, and $r = \rho^*(\vartheta)$, correspondingly. Right: enlargement of the region bounded by the dashed lines in the left panel.

- Fixing (ϑ, r) in the interior of the region $\{0 < \vartheta < 1,\ r > \rho^*_{exact}(\vartheta; \Omega)\}$, it is possible to exactly recover the support of β with high probabilities.
- Fixing (ϑ, r) in the interior of the region $\{0 < \vartheta < 1,\ 0 < r < \rho^*_{exact}(\vartheta; \Omega)\}$, it is impossible to exactly recover the support of β with high probabilities.

2.3. Phase diagrams

The preceding results give rise to the phase diagrams: the three curves $r = \rho^*(\vartheta)$, $r = \vartheta$, and $r = \rho^*_{exact}(\vartheta; \Omega)$ partition the phase space $\{(\vartheta, r) : 0 < \vartheta < 1, r > 0\}$ into four different subregions, where the inference is distinctly different.

- *Region of Undetectable*: $\{(\vartheta, r) : 0 < \vartheta < 1, r < \rho^*(\vartheta)\}$. The signals are so Rare/Weak that it is impossible to detect their existence: $H_1^{(p)}$ and $H_0^{(p)}$ are nearly inseparable and any test is asymptotically powerless.
- *Region of Not Recoverable but Detectable*: $\{(\vartheta, r) : 0 < \vartheta < 1, \rho^*(\vartheta) < r < \vartheta\}$. It is possible to have an asymptotically full power test, but impossible to separate the signals from the noise: the Hamming errors of any estimator is comparable to the total number of signals as $\mathrm{Hamm}^*_p(\vartheta, r; \Omega) \gtrsim p^{1-\vartheta}$.
- *Region of Almost Fully Recoverable*: $\{(\vartheta, r) : 0 < \vartheta < 1, \vartheta < r < \rho^*_{exact}(\vartheta; \Omega)\}$. It is possible to recover almost all signals but not all of them; the Hamming distance is much smaller than $p^{1-\vartheta}$, but is also much larger than 1.
- *Region of Exactly Recoverable*: $\{(\vartheta, r) : 0 < \vartheta < 1, r > \rho^*_{exact}(\vartheta; \Omega)\}$. The signals are sufficiently strong so that $\mathrm{Hamm}^*_p(\vartheta, r; \Omega) = o(1)$, and it is possible to have exact recovery with overwhelming probabilities.

Only the last two regions may depend on Ω. See Figure 1 for the case $\Omega = I_p$ (see Section 4 for more general cases). We call the union of the last three subregions the *Region of Detectable*:

$$\{(\vartheta, r) : 0 < \vartheta < 1, r > \rho^*(\vartheta)\}. \tag{2.12}$$

Phase diagram is a flexible notion that has been extended to many different settings, including large-scale multiple testing Arias-Castro, Candes, and Plan (2011); Hall and Jin (2010); Jager and Wellner (2004), variable selection Ji and Jin (2011); Jin, Zhang, and Zhang (2014); Ke, Jin, and Fan (2014), classification Donoho and Jin (2009); Fan, Jin, and Yao (2013); Ingster, Pouet, and Tsybakov (2009); Jin (2009), spectral clustering Jin and Wang (2014); Jin, Ke, and Wang (2015).

Remark 1. In the four regions aforementioned, we may call signals corresponding to the first three regions Rare/Weak, and signals corresponding to the last one Rare/Strong. However, in more general settings, Rare/Weak and Rare/Strong are relative concepts that are scientifically meaningful but are not always easy to define mathematically.

2.4. Achievability of the phase diagrams

In the preceding section, we have only said that the optimal phase diagrams are achievable, without referring to any specific method. It is of primary interest to develop methods—preferably easy-to-implement and not tied to the ARW model—to achieve the optimal phase diagrams.

- We say a testing procedure achieves the *optimal phase diagram for detection* if for any (ϑ, r) in the interior of Region of Detectable, the power of the procedure tends to 1 as $p \to \infty$.
- We say a variable selection procedure $\hat{\beta}$ achieves the *optimal phase diagram for recovery* if $h_p(\hat{\beta}, \beta) \leq L_p \cdot \mathrm{Hamm}_p^*(\vartheta, r; \Omega)$ for sufficiently large p, where L_p is the generic multi-$\log(p)$ term as in Definition 2.

In Section 3, we address achievability for the case $\Omega = I_p$, and show that Orthodox Higher Criticism (OHC) and Hard Thresholding achieve the optimal phase diagrams for detection and recovery, respectively. In Section 4, we address the acheivabilty for more genera Ω, and show that Innovated HC and GS achieve the optimal phase diagrams for detection and recovery, respectively. Combining these with Theorems 1 and 2 gives the phase diagrams in Section 2.3.

3. Detecting and Recovering Signals in White Noise

We revisit the problems of signal detection and signal recovery, and show that when $\Omega = I_p$, OHC and Hard Thresholding achieve the optimal phase diagrams

for detection and recovery, respectively. We also review the recent applications and extensions of the HC idea. Discussion on the general Ω is in Section 4.

3.1. Optimal signal detection by higher criticism (white noise)

To apply Orthodox HC (OHC) to the testing problem (2.7)−(2.8), we compute HC_p^* following the three steps in (1.1) where the P-values π_i are given by $P(|N(0,1)| \geq |X_i|)$, $1 \leq i \leq p$. Fix $0 < \alpha < 1$. To use the HC_p^* for a level-α test, we must find the critical value $h(p,\alpha)$ defined by $P_{H_0^{(p)}}\{HC_p^* > h(p,\alpha)\} = \alpha$. Asymptotically, it is known that for any fixed $\alpha \in (0,1)$,

$$h(p,\alpha) = \sqrt{2\log\log(p)}(1+o(1)). \tag{3.1}$$

We say a sequence α_p *tends to* 0 *slowly enough* if $h(p,\alpha_p) \sim \sqrt{2\log\log(p)}$. Consider the HC-test where we reject $H_0^{(p)}$ if and only if

$$HC_p^* \geq h(p,\alpha_p). \tag{3.2}$$

The following theorem was proved by Donoho and Jin (2004), where $\rho^*(\vartheta)$ is the standard phase function defined in (2.9).

Theorem 3. *Fix (ϑ, r) in the phase space such that $r > \rho^*(\vartheta)$. Suppose as $p \to \infty$, the level α_p of the HC-test tends to 0 slowly enough, then the power of the HC-test tends to 1.*

Combining this with Theorem 1 (not requiring $\Omega = I_p$), for any fixed (ϑ, r) in Region of Detectable (see (2.12)), OHC yields an asymptotically full power test when $\Omega = I_p$. Therefore, OHC achieves the optimal phase diagram for detection. Theorem 3 continues to hold if we replace HC_p^* in (3.2) by HC^+ defined in (1.2) and replace $h(p,\alpha)$ by its counterpart $h^+(p,\alpha)$, defined through $P_{H_0^{(p)}}(HC_p^+ \geq h^+(p,\alpha)) = \alpha$.

Remark 2. For fixed $0 < \alpha < 1$, $h(p,\alpha) \sim h^+(p,\alpha) \sim \sqrt{2\log\log(p)}$. However, this approximation is asymptotic, and may not be accurate for finite p. In the literature it is known that Shorack and Wellner (1986, p.600), as $p \to \infty$, $b_p HC_p^* - c_p$ and $b_p HC_p^+ - c_p$ both converge weakly to the standard Gumbel distribution, where $b_p = \sqrt{2\log\log(p)}$ and $c_p = 2\log\log(p) + (1/2)[\log\log\log(p) - \log(4\pi)]$. As a result, for any fixed $\alpha \in (0,1)$ and large p, $h(p,\alpha) \approx h^+(p,\alpha) \approx b_p^{-1}\big[c_p - \log\log(1/(1-\alpha))\big]$. For $h^+(p,\alpha)$, this approximation is reasonably accurate, especially for large p and small α. See Table 1 of Donoho and Jin (2015).

Remark 3. We only need P-values to implement the OHC-test with no knowledge of the parameters (ϵ_p, τ_p), so the test is not tied to the specific model in

(2.7)−(2.8). In the idealized case where (ϵ_p, τ_p) are known, the optimal test is the Likelihood Ratio Test (LRT). There is an interesting phase transition associated with the limiting distribution of LRT. Write the log-likelihood ratio associated with (2.7)−(2.8) as $LR_p(\epsilon_p, \tau_p) = LR_p(\epsilon_p, \tau_p; Y) = \sum_{i=1}^p \log((1-\epsilon_p) + \epsilon_p e^{\tau_p Y_i - \tau_p^2/2})$. With the calibrations in (2.5), LR_p can have non-degenerate limits only when (ϑ, r) fall *exactly* onto the phase boundary $r = \rho^*(\vartheta)$; still, this alone is inadequate, and we must modify the calibrations slightly. In detail, let

$$r = \rho^*(\vartheta), \quad \tau_p = \sqrt{2r\log(p)}, \quad \epsilon_p = \begin{cases} p^{-\vartheta}, & \text{if } \frac{1}{2} < \vartheta \le \frac{3}{4}, \\ \tau_p^2\sqrt{r}\, p^{-\vartheta}, & \text{if } \frac{3}{4} < \vartheta < 1. \end{cases} \tag{3.3}$$

As $p \to \infty$, if (3.3) holds, then LR_p has weak limits as follows (see Jin (2003)):

$$LR_p \longrightarrow \begin{cases} N(\mp\frac{1}{2}, 1), & \text{if } \frac{1}{2} < \vartheta < \frac{3}{4}, \\ N(\mp\frac{1}{4}, \frac{1}{2}), & \text{if } \vartheta = \frac{3}{4}, \\ \nu_{\mp}^{(\vartheta)}, & \text{if } \frac{3}{4} < \vartheta < 1, \end{cases} \quad \text{under } H_0^{(p)} \text{ and } H_1^{(p)}, \text{ respectively.}$$

Here, $\nu_{\mp}^{(\vartheta)}$ are the distributions with the characteristic functions $\psi_{\mp}^{(\vartheta)}$ given by $\psi_-^{(\vartheta)}(t) = (1/\sqrt{2\pi})\int_{-\infty}^{\infty}[e^{\sqrt{-1}t\log(1+e^z)} - 1 - \sqrt{-1}te^z]e^{-(z/\vartheta)(1+\sqrt{1-\vartheta})}dz$ and $\psi_+^{(\vartheta)}(t)$ $= (1/\sqrt{2\pi})\int_{-\infty}^{\infty}[e^{\sqrt{-1}t\log(1+e^z)} - 1]e^{-(z/\vartheta)(1-\vartheta+\sqrt{1-\vartheta})}dz$, respectively.

3.2. Optimal signal recovery by Hard Thresholding (white noise)

We now discuss signal recovery for the case of $\Omega = I_p$. In this simple setting, $Y_i \overset{iid}{\sim} (1-\epsilon_p)N(0,1) + \epsilon_p N(\tau_p, 1)$, and a conventional approach to estimating β is to use Hard Thresholding (HT). Fixing a threshold $t > 0$, we estimate β by $\hat{\beta}_{t,i}^{HT} = Y_i \cdot 1\{|Y_i| \ge t\}$. A convenient choice of t is $t_q(p) = \sqrt{2q\log(p)}$, where $0 < q < 1$ is a fixed parameter. Ideally, when (ϑ, r) are given, we choose q as

$$q^{ideal} = \begin{cases} \frac{(\vartheta+r)^2}{4r}, & r > \vartheta, \\ \vartheta, & 0 < r < \vartheta. \end{cases} \tag{3.4}$$

With the ideal threshold $t_p^{ideal} = \sqrt{2q^{ideal}\log(p)}$, it follows from the Mills' ratio Wasserman (2006) that the Hamming distance between $\hat{\beta}_{t_p^{ideal}}^{HT}$ and β satisfies

$$h_p(\hat{\beta}_{t_p^{ideal}}^{HT}, \beta) \begin{cases} = L_p p^{1-(\vartheta+r)^2/(4r)}, & \text{if } r > \vartheta, \\ \sim p^{1-\vartheta}, & \text{if } 0 < r < \vartheta. \end{cases}$$

Combining this with Theorem 2 (which is for more general Ω), we conclude that when $\Omega = I_p$, HT achieves the minimax Hamming distance, up to a multi-$\log(p)$ factor; so it achieves the optimal phase diagram for recovery as in Section 2.3−2.4.

Remark 4. The ideal choice of q in (3.4) depends on (ϑ, r) and it is hard to set them in a data-driven fashion. A convenient choice of q is $q = 1$, corresponding to the universal threshold $t_p^* = \sqrt{2\log(p)}$ Wasserman (2006). When $r > \rho^*_{exact}(\vartheta; I_p)$, this threshold leads to $\mathrm{Hamm}_p(\hat{\beta}^{HT}_{t_p^*}, \beta) \leq C(\log(p))^{-1/2}$ (and so exact recovery is achieved).

Remark 5. When a method yields exact recovery with overwhelming probability, we say that it has the *Oracle Property*, a well-known notion in the variable selection literature Fan and Li (2001). In such a framework, we are using $P(\mathrm{sgn}(\hat{\beta}) \neq \mathrm{sgn}(\beta))$ as the measure of loss. Seemingly, such a measure is only appropriate for Rare/Strong signals. When signals are Rare and Weak, exact recovery is usually impossible, and the Hamming distance is a more appropriate measure of loss.

3.3. Applications

HC (and its variants) has found applications in GWAS and DNA Copy Number Variation (CNV), where the genetic effects are believed to be rare and weak. Parkhomenko et al. (2009) used HC to detect modest genetic effects in a genome-wide study of rheumatoid arthritis. Sabatti et al. (2008) used HC to quantify the strength of the overall genetic signals for each of the nine traits of interest. De la Cruz et at. (2010) used HC to test whether there are associated markers in a given set of markers, with applications to Crohn's disease. Jeng, Cai, and Li (2010, 2013) proposed a variant of HC called *Proportion Adaptive Segment Selection (PASS)*, which can be viewed as a two-way screening process (across different SNPs and across different subjects), simultaneously dealing with the rare genetic effects and rare genomic variation. See also He and Wu (2011); Martin et al. (2009); Mukherjee, Pillai, and Lin (2013); Roeder and Wasserman (2009); Wu et al. (2012).

HC has also found applications in modern experiments in Cosmology and Astronomy—another important source of rare and weak signals. Jin et al. (2005) and Cayon, Jin, and Treaster (2004) (see also Cayon et al. (2006)) applied HC to standardized wavelet coefficients of Wilkinson Microwave Anisotropy Probe (WMAP), addressing the problem nonGaussianity detection in the Cosmic Microwave Background (CMB). Compared to the widely used kurtosis-based non-Gaussianity detector, HC showed superior power and sensitivity, and pointed to the *cold spot* centered at galactic coordinate (longitude, latitude) $= (207.8°, -56.3°)$ (see Vielva (2010) for more discussions). Pires et al. (2009) applied many nonGaussianity detectors to gravitational weak lensing data and showed that HC is competitive, being more specifically focused on excess of observations in the tails of the distribution. Bennett et al. (2012) applied the HC

ideas to the problem of Gravitational Wave detection for a monochromatic periodic source in a binary system. They use a modified form of HC which offers a noticeable increase in the detection power, and yet is robust.

HC has also been applied to disease surveillance for early detection of disease outbreak and local anomaly detection in a graph McFowland, Speakman, and Neill (2013); Saligrama and Zhao (2012), where it is found to have competitive powers.

3.4. Connections and extensions

In model (2.7)−(2.8), the noise entries are iid from a distribution F that is known as $N(0,1)$. Delaigle, Hall, and Jin (2011) address the more realistic setting where F is unknown and is probably nonGaussian. They consider a two group model (a control group and a case group) and compute P-values for each individual feature using bootstrapped Student's t-scores (see also Greenshtein and Park (2012); Liu and Shao (2013)).

The testing problem (2.7)−(2.8) is a special case of the problem of testing $H_0^{(p)}$ of $X_i \overset{iid}{\sim} F$ versus $H_1^{(p)}$ of $X_i \overset{iid}{\sim} (1-\epsilon)F+\epsilon G$, where $\epsilon \in (0,1)$ is small, F and G are distinct distributions, and (ϵ, F, G) may depend on p. Cai, Jeng, and Jin (2011) considers the case where $F = N(0,1)$ and $G = N(\tau, \sigma^2)$; see also Bogdan et al. (2011). Park and Ghosh (2010) gave a nice review on large-scale multiple testing with a detailed discussion on HC. Cai and Wu (2014) consider the extension where $F = N(0,1)$ and G is a Gaussian location mixture, and Arias-Castro and Wang (2013) investigate the case where F is *unknown* but symmetric. In a closely related setting, Laurent, Marteau, and Maugis-Rabusseau (2013) consider the problem of testing whether the samples X_i are *iid* from a single normal, or a mixture of two normals with different (but unknown) means. Addario-Berry et al. (2010) and Arias-Castro, Candes, and Durand (2011) consider a setting similar to (2.7)−(2.8) but where the signals are structured, forming clusters in (unknown) geometric shapes; the work is closely related to that in (Hall and Jin, 2010, Section 6).

Gayraud and Ingster (2011) show that HC statistic continues to be successful in detecting very sparse mixtures in a functional setting. Haupt, Castro, and Nowak (2008); Haupt et al. (2010) consider the setting where adaptive sample scheme is available so that we can do inference and collect data in an alternating order.

HC can also be viewed as a measure for goodness-of-fit. Jager and Wellner (2007) introduced a new family of goodness-of-fit tests based on the ϕ-divergence, including HC as a special case. Wellner and Koltchinskii (2003) further investigated the Berk-Jones statistic, which is closely related to HC, and derive its limiting distribution. In Jager and Wellner (2004), they further investigated the

limiting distribution of a class of weighted Kolmogorov statistics, including HC as a especial case. The pontogram of Kendall (1980) is an instance of HC, applied to a special set of P-values.

3.5. Comparison with FDR-controlling methods

Benjamini and Hochberg's (1995) FDR-controlling procedure (BH's procedure) is a recent innovation in multiple testing. Consider p (component) null hypotheses $H_1, \ldots, H_p$, where for each H_i, we have obtained a P-value π_i, $1 \le i \le p$. Let $\pi_{(1)} < \cdots \pi_{(p)}$ be the sorted P-values, and let $R_i = \pi_{(i)}/(i/p)$. For any preselected FDR control parameter $0 < q < 1$, letting i_q^{FDR} be the largest index such that $R_i \le q$, BH's procedure rejects all i_q^{FDR} (component) nulls that have the smallest P-values. The procedure is shown to control the FDR at level q if the P-values are independent.

While both HC and BH's procedure are multiple testing ideas in the sparse signal settings (we say π_i contains a signal if H_i is false and a noise otherwise), the scientific goals are very different. The goal of BH's procedure is to tell which (component) nulls are true and which are false, and the goal of HC is to decide whether all component null are true or a small fraction of them is untrue. The difference in scientific goals dictates that the success of BH's procedure needs stronger signals than that of HC. In many Rare/Weak settings, while BH's procedure still controls the FDR, it yields very few discoveries. In this case, a more reasonable goal is to test whether any signals exist without demanding that we properly identify them all; this is what HC is specifically designed for.

While both methods use P-values for inference, they normalize P-values in very different ways. HC normalizes them by $HC_{p,i} = \sqrt{p}[i/p - \pi_{(i)}]/\sqrt{\pi_{(i)}(1-\pi_{(i)}}$, and BH's procedure uses the normalization of $R_i = \pi_{(i)}/(i/p)$. If our goal is global testing, the former is a better choice. For example, if for some i we have $p\pi_{(i)} \gg 1$, then it could happen that $R_i \sim 1$ but $HC_{p,i} \gg 1$ (so $HC_{p,i}$ provides strong evidence against the joint null, but R_i fails to do so).

HC is also intimately connected to the problem of constructing confidence bands for the *False Discovery Proportion* (FDP). See Cai, Jin, and Low (2007); De Una-Alvarez (2012); Ge and Li (2012). Motivated by a study of Kuiper Belt Objects (KBO) (e.g., Meinshausen and Rice (2006)), Cai, Jin, and Low (2007) develop HC into an estimator for the proportion of non-null effects, a problem that has attracted substantial attention in the area of large-scale multiple testing in the past decade. The literature along this line connects to Benjamini and Hochberg (1995) on controlling FDR, as well as to Efron (2004) on controlling the local FDR in gene microarray studies.

4. Detecting and Recovering Signals in Colored Noise

In this section, we extend the discussions in Section 3 to the case of $\Omega \neq I_p$. We propose Innovated Higher Criticism (IHC) for signal detection and Graphlet Screening (GS) for signal recovery. IHC and GS can be viewed as Ω-aware Higher Criticism and Ω-aware Hard Thresholding, respectively.

4.1. Innovated higher criticism and its optimality in signal detection

We revisit the testing problem (2.7)−(2.8), where

$$Y = \beta + z, \qquad z \sim N(0, \Sigma). \tag{4.1}$$

HC is a method of combining P-values which is shown to be successful when $\Sigma = I_p$. We are interested in adapting HC for a general sparse precision matrix Ω, and there are three perceivable ways of combining the P-values.

In the first one, we obtain individual P-values marginally in a brute-force fashion: $\pi_i = P(|N(0,1)| \geq |Y_i|/(\Sigma(i,i))^{1/2})$. We call the HC applied to these P-values the *Brute-force HC (BHC)*. BHC neglects the correlation structure, so we expect there is room for improvement.

For an alternative, denoting the unique square root of Ω by $\Omega^{1/2}$, it is tempting to use the *Whitened Transformation* $Y \mapsto \Omega^{1/2}Y \sim N(\Omega^{1/2}\beta, I_p)$, so that the noise is whitened. We then obtain individual P-values by $\pi_i = P(|N(0,1)| \geq |(\Omega^{1/2}Y)_i|)$, $1 \leq i \leq p$. We call the resultant HC the *Whitened HC*.

Our proposal is *Innovated HC (IHC)*. Underlying IHC is the idea to find a transformation $Y \mapsto MY = M\beta + Mz$ ($M = M_{p,p}$, may depend on Ω) for

- *Preserving sparsity*, i.e., making most entries of the vector $M\beta$ zero. This is important since the strength of HC lies in detecting very sparse signals.
- *Simultaneously maximizing SNR*, i.e., maximizing the Signal-to-Noise Ratio (SNR) for all i at which $\beta_i \neq 0$, where SNR is defined by $(M\beta)_i/\sqrt{(M\Sigma M')(i,i)}$ (noting that $MY \sim N(M\beta, M\Sigma M')$).

The best choice turns out to be $M = \Omega$, corresponding to the so-called Innovated Transformation (IT): $Y \mapsto \Omega Y \sim N(\Omega\beta, \Omega)$. This is related to the notion of *innovation* in time series literature and so the name of Innovated Transformation. Section 1.2 of Fan, Jin, and Yao (2013) discusses why $M = \Omega$ is the best choice.

Now, first, IT preserves the sparsity of β, due to the sparsity of Ω given in (2.3). Second, for most i at which $\beta_i \neq 0$, among the three choices of M, $M = I_p$ (corresponding to model (4.1)), $M = \Omega^{1/2}$, and $M = \Omega$, the SNR are $(\Sigma(i,i))^{-1/2}\beta_i$, $((\Omega^{1/2})(i,i))\beta_i$, and β_i, respectively, where in the last term, we have used the assumption of $\Omega(i,i) = 1$ (see Hall and Jin (2010) for the insight underlying these results and proofs, where the key is to combine the sparsity of

Ω and the ARW model of β). Since $\Sigma(i,i)^{-1/2} \leq (\Omega^{1/2})(i,i) \leq 1$ is always true, IT has the largest SNR, simultaneously at all i such that $\beta_i \neq 0$.

It is particularly interesting that, while WT yields uncorrelated noise, it does not yield the largest possible SNR, so WT is not the best choice. For the current setting, where signals are Rare and Weak and Ω is sparse, the advantage of larger SNR out-weights the disadvantage of sparse correlations among the noise, so we prefer IHC to WHC. Similarly, we prefer WHC to BHC.

Example 1. Suppose Ω is block-wise diagonal and satisfies $\Omega(i,j) = 1\{i = j\} + h_0 \cdot 1\{|i-j| = 1,\ \max\{i,j\} \text{ is even}\}$, $-1 < h_0 < 1$, $1 \leq i,j \leq p$. For all $1 \leq i \leq p$, $(\Sigma(i,i))^{-1/2} = \sqrt{1-h_0^2}$ and $(\Omega^{1/2})(i,i) = (1/2)[\sqrt{1+h_0} + \sqrt{1-h_0}]$, and so $(\Sigma(i,i))^{-1/2} \leq (\Omega^{1/2})(i,i) \leq 1$; IT yields larger SNR than that of WT, and WT yields larger SNR than that of model (4.1).

Remark 6. At the heart of IHC is entry-wise thresholding applied to the vector ΩY. This is equivalent to Univariate Screening (US) Fan and Lv (2008); Genovese et al. (2012). In detail, we can rewrite model (2.1) as a regression model $W \sim N(X\beta, I_p)$, with $W = \Omega^{1/2}Y$ and $X = \Omega^{1/2}$. US thresholds the vector $X'W$ entry-wise; here $X'W = \Omega Y$.

Similarly, IHC consists of three steps (the last two are the same as in OHC).

- Obtain two-sided P-values by $\pi_i = P(|N(0,1)| \geq |(\Omega Y)_i|)$, $1 \leq i \leq p$.
- Sort P-values: $\pi_{(1)} < \pi_{(2)} < \cdots < \pi_{(p)}$.
- Innovated Higher Criticism statistic is then $IHC_p^* = \max_{\{1 \leq i \leq p/2\}} IHC_{p,i}$, where $IHC_{p,i} = \sqrt{p}[(i/p) - \pi_{(i)}]/\sqrt{\pi_{(i)}(1-\pi_{(i)})}$.

Consider the IHC test where we reject $H_0^{(p)}$ if and only if $IHC_p^* \geq d_p^*(\Omega)h(p,\alpha)$, where $d_p^*(\Omega)$ satisfies (2.3) and $h(p,\alpha)$ satisfies (3.1). Then, $P_{H_0^{(p)}}(\text{reject } H_0^{(p)}) \leq \alpha$. The next result extends Theorem 3 from $\Omega = I_p$ to more general Ω.

Theorem 4. *Fix (ϑ, r) in the phase space. As $p \to \infty$, if $r > \rho^*(\vartheta)$ and $\alpha = \alpha_p$ tends to 0 slowly enough, then the power of the IHC-test tends to 1. If $r < \rho^*(\vartheta)$, then for any test, the sum of Type I and Type II testing errors tends to 1.*

Combining this with Theorem 1, for any (ϑ, r) in the Region of Detectable, IHC provides an asymptotically full power test, so it achieves the optimal phase diagram for detection given in Section 2.3.

The proof of Theorem 4 has two new ingredients, additional to that of Theorem 3. The first one is Lemma 1 in Section 2.1. The second one is the similarity between β and $\Omega\beta$: The most interesting region for signal detection is

$1/2 < \vartheta < 1$ Donoho and Jin (2004). For ϑ in this range, $\epsilon_p \ll 1/\sqrt{p}$; so β has about $p\epsilon_p$ nonzeros all equal to τ_p, $\Omega\beta$ has $\lesssim d_p^*(\Omega) \cdot p\epsilon_p$ nonzeros, where about $p\epsilon_p$ of them equal τ_p and all others do not exceed τ_p in magnitude. Since $d_p^*(\Omega)$ does not exceed a multi-log(p) term, we do not expect any difference between the detection boundary of Ω and that of I_p; both are $r = \rho^*(\vartheta)$.

4.2. Graphlet screening and its optimality in variable selection

For signal recovery, we rewrite model (2.1) as the linear regression model

$$W \sim N(X\beta, I_p), \qquad W \equiv \Omega^{1/2}Y, \qquad X \equiv \Omega^{1/2}. \tag{4.2}$$

In Section 1.1, we have mentioned that (a) Univariate Screening (US) is a popular approach to variable selection but faces the challenge of "signal cancellation", and (b) Exhaustive Multivariate Screening (EMS) may help alleviate the effects of "signal cancellation" but is both inefficient and computationally infeasible (it includes too many subsets for screening and needs signals stronger than necessary for successful screening).

Our proposal is to use Graphlet Screening (GS). We recognize that in EMS many subsets of variables can be safely skipped for screening. The key innovation of GS is to use a sparse graph to guide the screening. With the same notations as those in Section 1.1, we let $\mathcal{G} = (V, E)$ be the graph where $V = \{1, \ldots, p\}$ and there is an edge between nodes i and j if and only if $\Omega(i, j) \neq 0$. Let $\mathcal{A}(m_0) = \{\text{all connected subgraphs of } \mathcal{G} \text{ with size} \leq m_0\}$. GS only applies χ^2-screening to those subsets in $\mathcal{A}(m_0)$. When $\Omega = I_p$, GS reduces to Hard Thresholding, so it can be viewed as an Ω-aware Hard Thresholding.

By a well-known result in graph theory (see Jin, Zhang, and Zhang (2014) and references therein),

$$|\mathcal{A}(m_0)| \leq Cp(ed_p^*(\mathcal{G}))^{m_0}, \tag{4.3}$$

where $d_p^*(\mathcal{G})$ is the maximum degree of $\mathcal{G}$. By the definition and (2.3), $d_p^*(\mathcal{G}) = d_p^*(\Omega)$, and does not exceed a multi-log(p) term. As a result, GS has a much smaller computational cost than that of EMS (in fact, it is only larger than that of US by a multi-log(p) factor for fixed m_0), and also requires much weaker signals than EMS does for successful screening.

Remark 7. GS is a flexible idea and can be adapted to many different settings, where the implementation may vary from occurrence to occurrence. It is a screening method and it has been applied to variable selection Jin, Zhang, and Zhang (2014); Ke, Jin, and Fan (2014), which includes model (4.2) as a special case. It can also be viewed as a way to evaluate the combined significance of (a small number of) features, so it can be used for feature ranking; see Section 5 for more discussion.

We now describe how to apply GS to model (4.2) for signal recovery. List all elements in $\mathcal{A}(m_0)$ in the order of sizes, with ties broken lexicographically, $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_N$, $N \equiv |\mathcal{A}(m_0)|$. Our proposal is a two-step procedure, containing a Screen step and Clean step. Fix positive tuning parameters (u, v, q). In the Screen step, initialize with $\mathcal{S}_0 = \emptyset$. For $i = 1, \ldots, N$, letting $\mathcal{S}_{i-1}$ be the set of all retained indices up to stage $i-1$, we update $\mathcal{S}_{i-1}$ by

$$\mathcal{S}_i = \begin{cases} \mathcal{S}_{i-1} \cup \mathcal{I}_i, \text{ if } \|P^{\mathcal{I}_i} W\|^2 - \|P^{\mathcal{I}_i \cap \mathcal{S}_{i-1}} W\|^2 \geq 2q \log(p), \\ \mathcal{S}_{i-1}, \qquad \text{otherwise} \end{cases}$$

where for any $\mathcal{I} \subset \{1, \ldots, p\}$, $P^{\mathcal{I}}$ is the projection matrix from R^n to $\{x_j : j \in \mathcal{I}\}$. The set of all retained nodes in the Screen step is then $\mathcal{S}_N$.

In the Clean step, we set $\hat{\beta}_j^{gs} = 0$ for $j \notin \mathcal{S}_N$. For $j \in \mathcal{S}_N$, we let $\mathcal{G}_{\mathcal{S}_N}$ be the subgraph of $\mathcal{G}$ formed by restricting all nodes to $\mathcal{S}_N$. There is a natural decomposition of $\mathcal{G}_{\mathcal{S}_N}$ into components (maximum connected subgraphs): $\mathcal{G}_{\mathcal{S}_N} = \mathcal{G}_{\mathcal{S}_N,1} \cup \mathcal{G}_{\mathcal{S}_N,2} \cup \ldots \cup \mathcal{G}_{\mathcal{S}_N,L}$. We estimate $\{\beta_j : j \in \mathcal{G}_{\mathcal{S}_N,\ell}\}$, $1 \leq \ell \leq L$, by minimizing $\|P^{\mathcal{G}_{\mathcal{S}_N,\ell}}(W - \sum_{j \in \mathcal{G}_{\mathcal{S}_N,\ell}} \beta_j x_j)\|^2 + u^2 \sum_{j \in \mathcal{G}_{\mathcal{S}_N,\ell}} |\beta_j|_0$, subject to the constraint that either $\beta_j = 0$ or $|\beta_j| \geq v$. Putting these together gives the final estimate, denoted by $\hat{\beta} = \hat{\beta}^{gs}(m_0, u, v, q)$.

Theorem 5. *Fix (m_0, ϑ, r) such that $1 < r/\vartheta < 3 + 2\sqrt{2} \approx 5.828$ and $m_0 \geq (r - \vartheta)^2/(4\vartheta r)$. Suppose (2.2)–(2.3) hold, $\|\Omega^{-1}\| \leq C$, and $\max_{1 \leq i \leq p} \sum_{j=1}^p |\Omega(i,j)|^\gamma \leq C$, for some constants $\gamma \in (0,1)$ and $C > 0$. Suppose $|\Omega(i,j)| \leq 4\sqrt{2} - 5 \approx 0.6569$ for all $1 \leq i, j \leq p$, $i \neq j$. If we set the tuning parameters (u, v, q) in GS by $u = \sqrt{2\vartheta \log(p)}$, $v = \sqrt{2r \log(p)}$ and q an appropriately small constant, then as $p \to \infty$, $h_p(\hat{\beta}^{gs}, \beta) \leq L_p \mathrm{Hamm}_p^*(\vartheta, r; \Omega) = L_p p^{1-(\vartheta+r)^2/(4r)} + o(1)$.*

Here $\mathrm{Hamm}_p^*(\vartheta, r; \Omega)$ is the minimax Hamming distance as in (2.10). For all Ω considered in Theorem 5, GS achieves the optimal phase diagram for recovery (see Section 2.3). Theorem 5 is a special case of the results in Jin, Zhang, and Zhang (2014); see Section 2.6 there for the proof. The conditions on r/ϑ and on the off-diagonals of Ω are not necessary for the optimality of GS. In fact, $h_p(\hat{\beta}^{gs}, \beta) \leq L_p \mathrm{Hamm}_p^*(\vartheta, r; \Omega) + o(1)$ holds in much broader settings, but $\mathrm{Hamm}_p^*(\vartheta, r; \Omega)$ may not have such a simple expression. See more discussions in Jin, Zhang, and Zhang (2014) about the asymptotic minimaxity of GS in more general settings.

Remark 8. The tuning parameter m_0 is usually chosen subject to computational capacity. The choice of q is relatively flexible, as long as it falls into certain ranges. The choice of u is harder, but the best u is a function of ϵ_p; in some settings (e.g., Cai, Jin, and Low (2007)), we can estimate ϵ_p consistently, and we know

how to choose the best u. For these reasons, we essentially have only one tuning parameter v, which is connected to the tuning parameter in the subset selection and that of the lasso; see Jin, Zhang, and Zhang (2014).

4.3. Phase diagrams (colored noise)

The optimal phase diagram for general Ω consists of four subregions separated by three curves $r = \rho^*(\vartheta)$, $r = \vartheta$ and $r = \rho^*_{exact}(\vartheta;\Omega)$; $\rho^*_{exact}(\vartheta;\Omega)$ may depend on the off-diagonals of Ω in a complicated way, but we always have $\rho^*_{exact}(\vartheta;\Omega) \geq \rho^*_{exact}(\vartheta;I_p)$, since $\Omega = I_p$ is the easiest case for exact recovery under our normalization $\Omega(i,i) = 1$.

For Ω satisfying conditions of Theorem 5 and (ϑ, r) such that $1 < r/\vartheta < 3 + 2\sqrt{2}$, the minimax Hamming distance for Ω has the same convergence rate as that for the case of $\Omega = I_p$. Note that in the phase space, the curve $r = \rho^*_{exact}(\vartheta;I_p)$ and the line $r/\vartheta = 3 + 2\sqrt{2}$ intersect at the point $(\vartheta, r) = (1/2, (3+2\sqrt{2})/2)$. Therefore, $\rho^*_{exact}(\vartheta;\Omega) = \rho^*_{exact}(\vartheta;I_p)$, for all $1/2 < \vartheta < 1$. Consequently, the right half of the curve $r = \rho^*_{exact}(\vartheta;\Omega)$ coincides with the right half of the curve $r = \rho^*_{exact}(\vartheta;I_p)$; see Jin, Zhang, and Zhang (2014) for discussion on general Ω.

By Theorems 1−2 and Theorems 4−5, the optimal phase diagram for detection is achieved by IHC, and the optimal phase diagram for recovery is achieved by GS for a wide range of Ω, including but are not limited to those satisfying the conditions of Theorem 5. See Jin, Zhang, and Zhang (2014) for details.

4.4. An example, and comparisons with L^0/L^1-penalization methods

In general, it is hard to derive an explicit form for $r = \rho^*_{exact}(\vartheta;\Omega)$ for the whole range of ϑ. Still, examples for some $\Omega \neq I_p$ would shed light on how this curve depends on the off-diagonal entries of Ω.

We revisit Example 1 in Section 4.1, where Ω is block-wise diagonal, and each diagonal block is the 2×2 matrix with 1 on the diagonals and h_0 on the off-diagonals. It was shown in Jin, Zhang, and Zhang (2014) that $\mathrm{Hamm}^*_p(\vartheta, r, \Omega) = L_p p^{1-c(\vartheta,r;h_0)}$, where

$$c(\vartheta, r; h_0) = \min\Big\{ \frac{(\vartheta+r)^2}{4r},\ \vartheta + \frac{(1-|h_0|)}{2}r,\ 2\vartheta + \frac{\{[(1-h_0^2)r - \vartheta]_+\}^2}{4(1-h_0^2)r} \Big\}. \quad (4.4)$$

The curve $\rho^*_{exact}(\vartheta;\Omega)$ is then the solution of $c(\vartheta, r; h_0) = 1$, which depends on h_0. The top left panel of Figure 2 displays the phase diagram for $h_0 = 0.5$.

Somewhat surprisingly, even for very simple Ω such as the block-wise diagonal example above and even when the tuning parameters are ideally set, subset selection (L^0-penalization) and the lasso have non-optimal phase diagrams; in particular, their Region of Exactly Recoverable is smaller than that of GS. Figure 2 shows phase diagrams associated with GS, L^0/L^1-penalization methods for

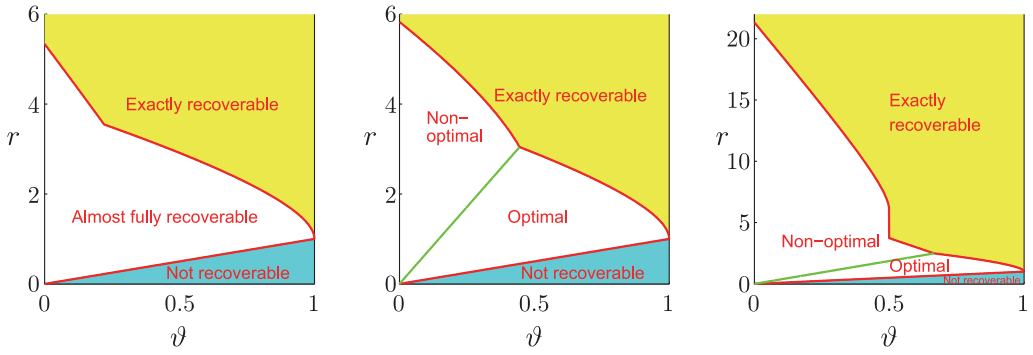

Figure 2. Phase diagrams (block-wise diagonal example, $h_0 = 0.5$). From left to right: GS, L^0, and L^1-penalization method. Note that the first two subregions described in Section 2.3 are combined into Region of Not Recoverable, for convenience.

the block-wise diagonal example, where the tuning parameters are set ideally to minimize the Hamming distance; see Ji and Jin (2011) and Jin, Zhang, and Zhang (2014). Given the non-optimality of L^0/L^1-penalization methods in such a simple Ω, we do not expect that for more general Ω they could be optimal.

We must note that the optimality of L^0/L^1-penalization methods in the literature are largely limited to settings, different from here, where they usually have Rare/Strong signals (i.e., somewhere above the curve $r = \rho^*_{exact}(\vartheta, \Omega)$) in Figure 1), and either L^2-loss or $P(\mathrm{sgn}(\hat{\beta}) \neq \mathrm{sgn}(\beta))$ is frequently used as the measure of loss. However, L^2-loss is more appropriate for prediction setting, not for variable selection, and $P(\mathrm{sgn}(\hat{\beta}) \neq \mathrm{sgn}(\beta))$ is more appropriate for Rare/Strong signals, not for Rare/Weak signals where it is merely impossible to fully recover the support of β. Since L^0-penalization method is the target of many penalization methods, including the lasso, SCAD Fan and Li (2001), MC+ Zhang (2010), we should not expect these penalization methods to be optimal as well.

Remark 9. Ke, Jin, and Fan (2014) studied a more complicate setting than that in Theorem 5 or that in Jin, Zhang, and Zhang (2014), where the Gram matrix is not sparse but is sparsifiable. They derived the phase diagrams for a case that Ω is the correlation matrix of a long-memory time series and for a change-point model. The change-point model is a special case of model (4.2) where X is an upper triangular matrix of 1's (therefore, $X\beta$ is piece-wise constant). For the change-point model, the phase space partitions into only 2 regions, separated by the curve $r = \rho^{*,cp}_{exact}(\vartheta)$, where $\rho^{*,cp}_{exact}(\vartheta) = \max\{4(1-\vartheta), (4-10\vartheta) + 2\sqrt{[(2-5\vartheta)^2 - \vartheta^2]_+}\}$.

4.5. Connections to the literature

Model (4.1) is closely related to the linear model $W \sim N(X\beta, I_n)$, where the Innovated Transformation reduces to that of $W \mapsto X'W$. Arias-Castro, Candes, and Durand (2011) applied HC to $X'W$ for signal detection, which is similar to IHC. Ingster, Tsybakov, and Verzelen (2010) considered a case that $W \sim N(X\beta, \sigma^2 I_n)$ where σ is unknown and $X_i \overset{iid}{\sim} N(0, (1/n)I_p)$. They proposed a modified IT, $W \mapsto \|W\|^{-1}X'W$, to adapt to the unknown σ. Mukherjee, Pillai, and Lin (2013) considered the binary-response logistic regression. They proposed HC-like statistics for signal detection and exposed interesting dependence of the detection boundary on the design matrix.

Another related setting is with data as iid samples $Y_1, \cdots, Y_n$ of $N(\beta, \Sigma)$. This reduces to model (4.1), noting that $(1/\sqrt{n})\sum_{i=1}^n Y_i \sim N(\beta, \Sigma)$ is the vector of sufficient statistics of β. When the data are nonGaussian, Zhong, Chen, and Xu (2013) proposed an "L_γ-thresholding test" which takes BHC as a special case of $\gamma = 0$.

GS, as a method to improve US, is different from the Iterative Sure Independence Screening (ISIS) Fan and Lv (2008); Fan, Samworth, and Wu (2009). ISIS first applies US to select a small set of variables M_1. In the second step, for each $j \notin M_1$, it runs a least-square algorithm on the model $M_1 \cup \{j\}$ and records the coefficient of j. These coefficients are then used to rank variables and expand M_1 to a set M_2. This procedure runs iteratively. ISIS alleviates 'signal cancellation' between variables in M_1 and those in $\{1, \cdots, p\}\backslash M_1$, but unlike GS, it does not deal with 'signal cancellation' among variables in $\{1, \cdots, p\}\backslash M_1$.

GS is closely related to LARS Efron et al. (2004) and the forward-backward greedy algorithm Zhang (2011) in utilizing local graphic structure of variables. The Screen step of GS is a step-wise forward algorithm and the Clean step is a backward algorithm.

5. Stylized Applications

HC and GS are flexible ideas that can be adapted to a broad set of problems and settings. In this section, we outline some potential applications.

5.1. Higher criticism for estimating the bandwidth of a matrix

The HC idea, although still in its early stage of development, is seeing increasing interest both in practice and in theory. In Section 3.3−3.4, we have reviewed applications and extensions of HC in many different settings. In this section, we illustrate a new application of HC.

Consider samples $X_i \in R^p$ from a Gaussian distribution: $X_i \overset{iid}{\sim} N(0, \Sigma)$, $1 \le i \le n$. The Gaussian assumption is not critical and is only for simplicity. In

many applications, with the Linkage Disequilibrium (LD) matrix being an iconic example, Σ is unknown but is banded; denote the bandwidth by $b = b(\Sigma)$ so that b is the smallest integer such that $\Sigma(i,j) = 0$ for all i, j with $|i-j| \geq b+1$.

We adapt HC to estimate $b(\Sigma)$. HC can also be adapted to test whether $b(\Sigma) \leq k_0$ or $b(\Sigma) > k_0$ for a given small integer k_0; the discussion is similar so we omit it to save space. Let the empirical covariance matrix be $S_n = (1/n)\sum_{i=1}^n X_i X_i'$. For $1 \leq k \leq p-1$, let $\xi^{(k)}$ and $\hat{\xi}^{(k)}$ be the $(p-k)\times 1$ vectors formed by the k-th (upper) off-diagonal of Σ and S_n, respectively: $\xi^{(k)} = (\Sigma(1,1+k), \ldots, \Sigma(p-k,p))'$, $\hat{\xi}^{(k)} = (S_n(1,1+k), \ldots, S_n(p-k,p))'$. We consider a Rare/Weak setting where each $\xi^{(k)}$ has a small fraction of nonzeros, and each nonzero is relatively small. For any i, j such that $\Sigma(i,j) = 0$, we have that approximately, $\sqrt{n}S_n(i,j) \sim N(0,1)$.

We propose the following HC estimator for $b(\Sigma)$. Fix an integer b_0 (a relatively small but conservative upper bound for $b(\Sigma)$) and a level $\alpha \in (0,1)$.

- For $k = 1, \ldots, b_0$, apply HC_p^+ in (1.2) to $\hat{\xi}^{(k)}$, where the P-value associated with the i-th entry of $\hat{\xi}^{(k)}$ is given by $P(|N(0,1)| \geq \sqrt{n}\hat{\xi}_i^{(k)})$, $1 \leq i \leq p-k$. Denote the resultant HC scores be $HC^{(1)}, \ldots, HC^{(b_0)}$, correspondingly.
- Estimate $b(\Sigma)$ by $\hat{b}^{HC} = \max\{1 \leq k \leq b_0 : HC^{(k)} \geq h^+(p, \alpha/b_0)\}$.

Here $h^+(p,\alpha)$ is as in Remark 2 which can be computed by simulations.

We conducted a small-scale simulation, where $(p, n, b(\Sigma), b_0, \alpha) = (5000, 200, 2, 10, 0.05)$. For $k = 1, 2$, and fixed (ϵ, τ), we generated the entries of $\xi^{(k)}$ randomly from $(1-\epsilon)\nu_0 + \epsilon\nu_\tau$. We then applied the above procedure and repeatd the whole simulation processes independently for 200 times, and recorded the error rates (the fraction of simulations where $\hat{b}^{HC} \neq b(\Sigma)$). We investigated six combinations of (ϵ, τ): $(0.01, 0.175)$, $(0.01, 0.2)$, $(0.01, 0.225)$, $(0.005, 0.225)$, $(0.005, 0.25)$, and $(0.01, 0.275)$; the corresponding error rates of $\hat{b}^{HC}$ were 6.5%, 0.5%, 0%, 8.5%, 3%, and 2%.

Remark 10. The choice of $h^+(p, \alpha/b_0)$ is from Bonferroni correction. It is acceptable for relatively small b_0. For large b_0, we may need to adjust the threshold, say, with the FDR-controlling method in Benjamini and Hochberg (1995).

5.2. Ranking features by graphlet screening

Consider the linear regression model of Section 1.1:

$$W = X\beta + z, \qquad X = X_{n,p} = [x_1, x_2, \ldots, x_p], \qquad z \sim N(0, I_n), \tag{5.1}$$

with assumptions on Ω and β in Section 2. We are interested in ranking the features so to have a competitive Receiver Operating Curve (ROC).

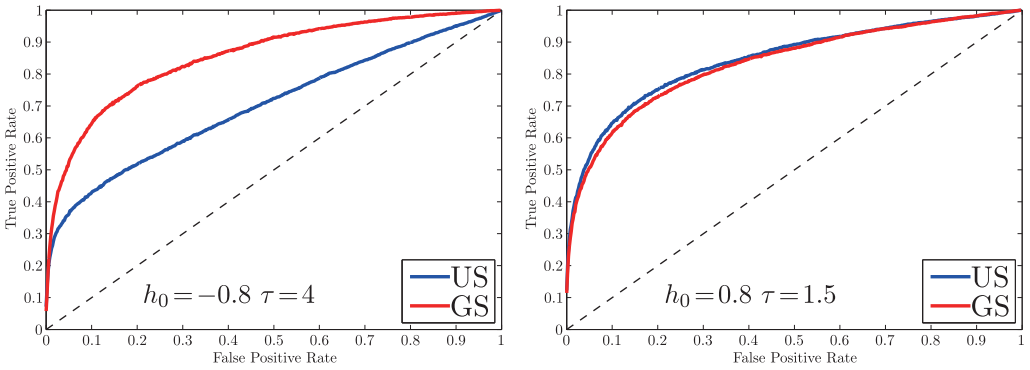

Figure 3. ROC curves associated with features ranking by US (blue) and GS (red). Signal cancellation is severe on the left and GS offers a significant improvement. It is not severe on the right so GS is comparable to US; see Section 5.2 for details.

We propose to rank the features by GS. Fix a threshold $\delta > 0$. Similar to that in Section 1.1, let $\mathcal{G}^{*,\delta} = (V, E)$ be the graph where $V = \{1, \ldots, p\}$ and there is an edge between i and j if and only if $|G(i,j)| \geq \delta$; since G is approximately sparse, $\mathcal{G}^{*,\delta}$ is sparse in that the maximum degree is small, given an appropriate choice of δ. Fixing $m_0 > 1$, let $\mathcal{A}^{*,\delta}(m_0) = \mathcal{A}^{*,\delta}(m_0, G) = \{\text{all connected subgraphs of } \mathcal{G}^{*,\delta} \text{ with size} \leq m_0\}$. Similarly to (4.3), $|\mathcal{A}^{*,\delta}(m_0)| \leq Cp(ed_p^*)^{m_0}$, where $d_p^* = d_p^*(\delta, G)$ is the maximal degree of $\mathcal{G}^{*,\delta}$. Our procedure consists of the following steps.

- For each $\mathcal{I} \in \mathcal{A}^{*,\delta}(m_0)$, compute a P-value as $\pi^{(\mathcal{I})} = P(\chi^2_{|\mathcal{I}|}(0) > \|P^{\mathcal{I}}Y\|^2)$.
- Let $\pi_j^{gs} = \min_{\mathcal{I} \in \mathcal{A}^{*,\delta}(m_0)}\{\pi^{(\mathcal{I})}\}$, for $1 \leq j \leq p$.
- Rank the significance of feature j according to π_j^{gs}.

Here $P^{\mathcal{I}}$ is the projection from R^n to $\{x_j : j \in \mathcal{I}\}$. The procedure is related to the hierarchical variable selection procedures Meinshausen (2008) but differs in significant ways.

We conducted a small-scale numerical study, where $(n, p, \epsilon) = (500,\ 1{,}000,\ 0.05)$. Let Σ be a $p \times p$ blockwise diagonal matrix with size-2 blocks, each block with diagonals 1 and off-diagonals h_0. Given (h_0, τ), we first generated $(\beta_{2j-1}, \beta_{2j}) \overset{iid}{\sim} (1-\epsilon)\nu_{(0,0)} + (\epsilon/2)\nu_{(\tau,\tau)} + (\epsilon/2)\nu_{(\tau,0)}$, for $j = 1, \cdots, p/2$, where ν_a is a point mass at a for any $a \in R^2$. Next, we generated $X_i \overset{iid}{\sim} N(\beta, (1/n)\Sigma)$, for $i = 1, \cdots, n$. We applied both US and GS (taking $m_0 = 2$) to rank features. Figure 3 displays the corresponding ROC curves, obtained from averaging 200 independent repetitions. We investigated the cases $(h_0, \tau) = (-0.8, 4), (0.8, 1.5)$. In the first case, signal cancellation is severe and GS significantly outperforms US; in the second case, GS has a similar performance as US.

Feature ranking is of interest in many high dimensional problems including, but are not limited to, (a) large-scale multiple testing, where it is of interest to develop methods that control the FDR while maximizing the power of multiple tests, (b) cancer classification where it is desirable to select a small fraction of features for the trained classification decision Donoho and Jin (2008, 2009); Jin (2009), and spectral clustering where it is desirable to perform a dimension reduction before we apply Principle Component Analysis (PCA) Jin and Wang (2014). As GS provides a better strategy in feature ranking than US, it is potentially useful in attacking all of these problems.

6. Feature Selection by Higher Criticism for Classification

Among many uses of Higher Criticism, one of particular interest is setting thresholds for feature selection in classification. Consider a two-class classification setting where $(Y^{(i)}, \ell_i)$, $1 \le i \le n$, are measurements from two different classes. Here, $Y^{(i)} \in R^p$ are the feature vectors and $\ell_i \in \{-1, 1\}$ are the class labels. We assume the classes are equally likely, so that after a standardizing transformation, $Y^{(i)} \sim N(\ell_i \cdot \mu, \Sigma)$, with $\mu \in R^p$ the contrast mean vector and Σ the $p \times p$ covariance matrix; such an assumption is only for simplicity in presentations. Given a fresh feature vector Y, the primary interest is to predict the associated class label $\ell \in \{-1, 1\}$.

For simplicity, we assume Σ is known and $\Omega = \Sigma^{-1}$ is sparse. The case Σ is unknown (but Ω is sparse) is discussed in Fan, Jin, and Yao (2013). Fisher's linear discriminant analysis (LDA) is a classical approach to classification. Let $w = (w_1, w_2, \ldots, w_p)'$ be a $p \times 1$ feature weight vector. For a fresh feature vector $Y = (Y_1, \cdots, Y_p)'$, Fisher's LDA takes the form $L(Y) = \sum_{i=1}^p w_i Y_i$, and classifies $\ell = \pm 1$ according to $L(Y) \gtrless 0$. When (Σ, μ) is known, it is known that the optimal weight vector satisfies $w \propto \Omega\mu$.

To adapt Fisher's LDA to the current setting, the key is to estimate μ. We are primarily interested in the Rare/Weak setting where only a small fraction of the entries of μ is nonzero and the nonzero entries are individually small. Define the feature z-vector $Z = (1/\sqrt{n}) \sum_{i=1}^n (\ell_i \cdot Y^{(i)}) \sim N(\sqrt{n}\mu, \Sigma)$. A standard approach to estimating μ is by some sort of thresholding scheme. For any $t > 0$, denote by $\eta_t(z)$ the clipping thresholding function $\eta_t(z) = \mathrm{sgn}(z)1\{|z| \ge t\}$ Donoho and Jin (2008); Fan, Jin, and Yao (2013). Our proposal is to use Innovated Thresholding which thresholds ΩZ coordinate-wise:

$$\hat{\mu}_{t,i}^{IT} = \eta_t((\Omega Z)_i), \qquad 1 \le i \le p. \tag{6.1}$$

One could also use Brute-force Thresholding which thresholds Z coordinate-wise, or Whitened Thresholding which thresholds $\Omega^{1/2} Z$ coordinate-wise. However,

these schemes are inferior to Innovated Thresholding, for Innovated Transformation yields largest Signal-to-Noise Ratio (Section 4.1). In (6.1), we use the clipping thresholding rule. One could also use hard-thresholding or soft-thresholding, but the difference is usually not significant; see Donoho and Jin (2008); Fan, Jin, and Yao (2013).

We now modify Fisher's LDA. Letting $L_t^{IT}(Y;\Omega) = (\hat{\mu}_t^{IT})'\Omega Y$, where $\hat{\mu}_t^{IT} = (\hat{\mu}_{t,1}^{IT}, \hat{\mu}_{t,2}^{IT}, \ldots, \hat{\mu}_{t,p}^{IT})'$, we classify ℓ as ± 1 according to $L_t^{IT}(Y;\Omega) \gtrless 0$. This connects to the modified HC in Zhong, Chen, and Xu (2013), but the focus there is on signal detection.

An important issue is how to set the threshold t. We propose *Higher Criticism Threshold (HCT)*, a variant of OHC.

1. Calculate (two-sided) P-values $\pi_i = P\{|N(0,1)| \geq |(\Omega Z)_i|\}$, $1 \leq i \leq p$.
2. Sort the P-values into ascending order: $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$.
3. Define the *Higher Criticism feature scores* by

$$HC(i;\pi_{(i)}) = \sqrt{p}\frac{i/p - \pi_{(i)}}{\sqrt{(i/p)(1-i/p)}}, \qquad 1 \leq i \leq p. \tag{6.2}$$

For a tuning parameter $\alpha_0 \in (0, 1/2]$, let $\hat{i}^{HC} = \mathrm{argmax}_{\{1 \leq i \leq \alpha_0 \cdot p\}}\{HC(i;\pi_{(i)})\}$. The HCT is then $\hat{t}_p^{HC} = \hat{t}_p^{HC}(Z_1, Z_2, \ldots, Z_p; \alpha_0) = |Z|_{\hat{i}^{HC}}$.

In practice, we usually set $\alpha_0 = 0.10$; HCT is relatively insensitive to different choices of α_0. In (6.2), the denominator of the HC objective function is different from that of OHC we used for testing problems (2.7)−(2.8), although in a similar spirit. See Donoho and Jin (2009) for explanations.

Once the threshold is decided, the associated Fisher's LDA is

$$L_{HC}^{IT}(Y;\Omega) = (\hat{\mu}_{HC}^{IT})'\Omega Y, \quad \text{where } \hat{\mu}_{HC}^{IT} = \hat{\mu}_t^{IT}\big|_{t=\hat{t}_p^{HC}}. \tag{6.3}$$

The HCT trained classification rule classifies $\ell = \pm 1$ according to $L_{HC}^{IT}(Y) \gtrless 0$.

Remark 11. The classification problem is closely connected to the testing problem (2.7)−(2.8) in Sections 3−4. For illustration, assume $\Omega = I_p$ and $\sqrt{n}\mu_j \overset{iid}{\sim} (1-\epsilon)\nu_0 + \epsilon\nu_\tau$. Given a test feature $Y \sim N(\ell \cdot \mu, I_p)$, the classification problem can be viewed as the problem of testing $H_0^{(p)}$ of $Y \sim N(-\mu, I_p)$ against $H_1^{(p)}$ of $Y \sim N(\mu, I_p)$. Although this is very similar to that of (2.7)−(2.8), there is a major difference. In (2.7)−(2.8), we have no other information than the prior distribution on μ, so all features are equally likely to be useful. In the classification problem, however, the training z-vector $Z \sim N(\sqrt{n}\mu, I_p)$ contains additional information about μ; for feature i, $1 \leq i \leq p$, the posterior probability that it is a useful feature is given by $P(\mu_i \neq 0|Z) = \epsilon e^{\tau Z_i - \tau^2/2}/[(1-\epsilon) + \epsilon e^{\tau Z_i - \tau^2/2}]$, which ≈ 1 if Z_i is large and positive and ≈ 0 if Z_i is large and negative. Seemingly,

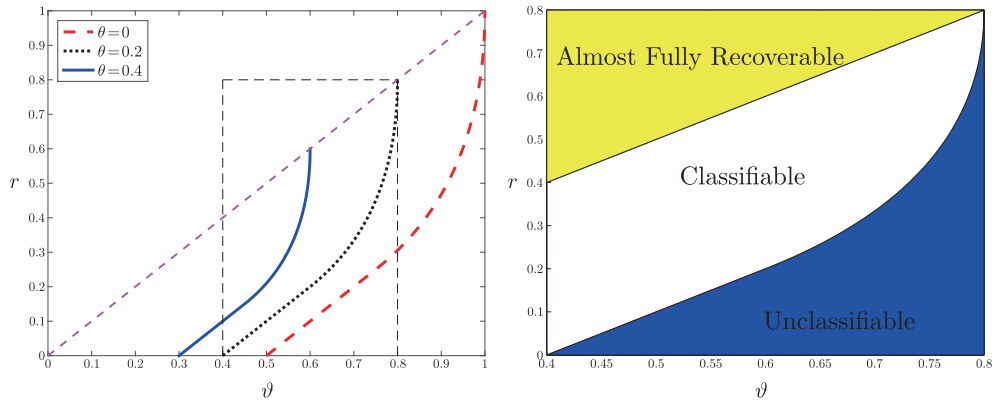

Figure 4. Left: $r = \rho^*_\theta(\vartheta)$ for $\theta = 0, 0.2, 0.4$. Right: enlargement the region bounded by grey dashed lines in the left panel; in the yellow region, it is not only possible to have successful classifications, but is also possible to separate useful features from useless ones.

the posterior distribution contains much more information on inference than the prior distribution does. This also suggests (one-sided) clipping hard thresholding, similar to that suggested by Fisher's LDA.

6.1. Phase diagram for classification

We first introduce the ARW model for classification. We model the contrast mean vector μ by $\sqrt{n}\mu_i \overset{iid}{\sim} (1-\epsilon)\nu_0 + \epsilon\nu_\tau$, $1 \le i \le p$. Fix (ϑ, r, θ) such that $r > 0$, $0 < \theta < 1$ and $0 < \vartheta < 1-\theta$. Similarly, we let $\epsilon = \epsilon_p = p^{-\vartheta}$, $\tau = \tau_p = \sqrt{2r\log(p)}$, and $n = n_p = p^\theta$. It was noted in Jin (2009) and Fan, Jin, and Yao (2013) that for any fixed $\theta \in (0,1)$, the most interesting range for ϑ is $0 < \vartheta < (1-\theta)$. When $\vartheta > (1-\theta)$, for successful classification, we need $\tau_p \gg \sqrt{\log(p)}$, but this corresponds to the Rare/Strong regime, which is relatively easy, for we can separate the nonzero entries of μ from zero ones by simple thresholding. For $\rho^*(\cdot)$ as in (2.9), let $\rho^*_\theta(\vartheta) = (1-\theta)\rho^*(\vartheta/(1-\theta))$, $0 < \vartheta < (1-\theta)$. The following theorem was proved in Fan, Jin, and Yao (2013).

Theorem 6. *Fix $(\vartheta, \theta, r) \in (0,1)^3$ such that $0 < \vartheta < (1-\theta)$. Suppose that $\Omega = \Sigma^{-1}$ satisfies (2.2)–(2.3) and that the spectral norm of Σ is bounded by a constant $C > 0$. If $r > \rho^*_\theta(\vartheta)$, then the classification error of the trained HCT classification rule in (6.3) tends to 0 as $p \to \infty$. If $0 < r < \rho^*_\theta(\vartheta)$, then the classification error of any trained classification rule is no less than $1/2 + o(1)$, where $o(1) \to 0$ as $p \to \infty$.*

There is a similar phase diagram associated with the classification problem.

- *Region of Classifiable*: $\{(\vartheta, r) : 0 < \vartheta < (1-\theta), r > \rho^*_\theta(\vartheta)\}$. In this region, the HC threshold $\hat{t}_p^{HC}$ satisfies $\hat{t}_p^{HC}/t_p^{ideal} \to 1$ in probability, where t_p^{ideal} is the ideal threshold that one would choose if the underlying parameters (ϑ, r, Ω) are known. Also, the classification error of HCT-trained classification rule tends to 0 as $p \to \infty$.
- *Region of Unclassifiable*: $\{(\vartheta, r) : 0 < \vartheta < (1-\theta),\ r < \rho^*_\theta(\vartheta)\}$. In this region, the classification error of any trained classification rule can not be substantially smaller than 1/2.

See more discussion in Donoho and Jin (2008, 2009); Jin (2009). Ingster, Pouet, and Tsybakov (2009) independently derived the classification boundary in a broader setting, but they did not discuss HC. In Figure 4, we plot the phase diagrams for $\theta = 0, 0.2, 0.4$.

The advantage of HC is its optimality in the ARW model. Note that HCT is a data-driven non-parametric statistic, the use of which does not require the knowledge of the ARW parameters. HC is not tied to the idealized model we discussed here, and can be useful for more general settings. See Donoho and Jin (2008) for applications of HC to cancer classification with microarray data sets.

Our proposal of threshold choice by HC is very different from Benjamini-Hochberg's FDR-controlling method (or Efron's local FDR approach), where the philosophy is to control the feature FDR, the expected fraction of falsely selected features out of all selected features, by a small number (e.g., 5%). However, this is not necessarily the right strategy when signals are Rare/Weak. Donoho and Jin (2009) identified a sub-region of Region of Classifiable where to obtain optimal classification behavior, we must set the feature selection threshold very low so that we include most of the useful features; but when we do this, we must include many useless features and the feature FDR is approximately 1.

References

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

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

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

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

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

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

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

Box, G. and Meyer, D. (1986). An analysis for unreplicated fractional factorials. *Technometrics* **28**, 11-18.

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

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

Cai, T. and Wu, Y. (2014). Optimal detection of sparse mixtures against a given null distribution. *IEEE Trans. Inform. Theory* **60**, 2217-2232.

Cayon, L. and Banday, A., Jaffe, T., Eriksen, H., Hansen, H., Gorski, K. and Jin, J. (2006). No Higher Criticism of the Bianchi-corrected Wilkinson Microwave Anisotropy Probe data. *Mon. Not. Roy. Astron. Soc.* **369**, 598-602.

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

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

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

De Una-Alvarez, J. (2012). The Beta-Binomial SGoF method for multiple dependent tests. *Statistical Applications in Genetics and Molecular Biology.* **11**, 1544-6115.

Donoho, D. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52**, 1289-1306.

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

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

Donoho, D. and Jin, J. (2009). Feature selection by higher criticism thresholding: optimal phase diagram. *Phil. Tran. Roy. Soc. A* **367**, 4449-4470.

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

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

Donoho, D. and Johnstone, I. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 1200-1224.

Donoho, D., Maleki, A. and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.* **106**, 18914-18919.

Donoho, D. and Stark, P. (1989). Uncertainty principles and signal recovery. *SIAM J. Appl. Math.* **49**, 906-931.

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

Efron, B. (2011). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* IMS Monographs, Cambridge Press.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407-840.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. Ser. B.* **70**, 849-911.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.* **10**, 2013-2038.

Fan, Y., Jin, J. and Yao, Z. (2013). Optimal classification by Higher Criticism in sparse Gaussian graphic model. *Ann. Statist.*, **41**, 2263-2702.

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

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

Gayraud, G. and Ingster, Y. (2011). Detection of sparse variable functions. *Electron. J. Statist.* **6**, 1409-1448.

Ge, Y. and Li, X. (2012). Control of the false discovery proportion for independently tested null hypotheses. *J. Probab. and Statist.* Article ID 320425, 19 pages.

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

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

Hall, P. and Jin, J. (2008). Properties of higher criticism under strong dependence. *Ann. Statist.* **36**, 381-402.

Hall, P. and Jin, J. (2010). Innovated Higher Criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38**, 1686-1732.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Jin, J. (2012). Comment on "Estimating false discovery proportion under arbitrary covariance dependence". *J. Amer. Statist. Assoc.* **107**, 1042-1045.

Jin, J., Ke, Z. and Wang, W. (2015). Phase transitions for high-dimensional clustering and related problems. arXiv.1502.06952.

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

Jin, J. and Wang, W. (2014). Important Features PCA for high dimensional clustering. arXiv.1407.5241.

Jin, J., Zhang, C.-H. and Zhang, Q. (2014). Optimality of Graphlet Screening in high dimensional variable selection. *J. Mach. Learn. Res.* **15**, 2723-2772.

Ke, Z., Jin, J. and Fan, J. (2014). Covariate assisted screening and estimation. *Ann. Statist.* **42**, 2202-2242.

Kendall, D. (1980). Discussion of "simulating the Ley hunter" by Simon Broadbent *J. Royal Stat. Soc. Ser. A.* **2**, 109-140.

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

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

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

McFowland, E., Speakman, S. and Neill, D. (2013). Fast generalized subset scan for anomalous pattern detection. *J. Mach. Learn. Res.* **14**, 1533-1561.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika* **95**, 265-278.

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

Mukherjee, R., Pillai, N. and Lin, X. (2013). Hypothesis testing for sparse binary regression. *Ann. Statist.* **43**, 352-381.

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

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

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2010). Partial correlation estimation by joint sparse regression model. *J. Amer. Statist. Assoc.* **104**, 735-746.

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

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

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

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

Shorack, G. and Wellner, J. (1986) *Empirical Processes with Applications to Statistics*. Vol **59**, SIAM.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B.* **58**, 267-288.

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

Tukey, J. (1976). T13 n: The Higher Criticism. *Course notes, Stat 411*. Princeton University.

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

Vielva, P. (2010). A comprehensive overview of the cold spot. *Adv. Astron.* Article ID 592094, 20 pages.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer, NY.

Wellner, J. and Koltchinskii, V. (2003). A note on the asymptotic distribution of Berk-Jones type statistics under the null hypothesis. *High Dimensional Probability III*. Birkhauser Basel, Germany.

Wu, Z., Sun, Y., He, S., Choy, J., Zhao, H. and Jin, J. (2012). Detection boundary and Higher Criticism approach for sparse and weak genetic effects. *Ann. Appl. Statist.* **8**, 824-851.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894-942.

Zhang, T. (2011). Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Trans. Inform. Theory* **57**, 4689-4708.

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

Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

E-mail: jiashun@stat.cmu.edu

Department of Statistics, University of Chicago, Chicago, IL 60637, USA.

E-mail: zke@galton.uchicago.edu

(Received April 2014; accepted June 2015)