The Annals of Statistics
2013, Vol. 41, No. 5, 2537–2571
DOI: 10.1214/13-AOS1163
© Institute of Mathematical Statistics, 2013

OPTIMAL CLASSIFICATION IN SPARSE GAUSSIAN GRAPHIC MODEL

BY YINGYING FAN[1], JIASHUN JIN[2] AND ZHIGANG YAO[3]

University of Southern California, Carnegie Mellon University and Ecole Polytechnique Fédérale de Lausanne

Consider a two-class classification problem where the number of features is much larger than the sample size. The features are masked by Gaussian noise with mean zero and covariance matrix Σ, where the precision matrix $\Omega = \Sigma^{-1}$ is unknown but is presumably sparse. The useful features, also unknown, are sparse and each contributes weakly (i.e., rare and weak) to the classification decision.

By obtaining a reasonably good estimate of Ω, we formulate the setting as a linear regression model. We propose a two-stage classification method where we first select features by the method of *Innovated Thresholding* (IT), and then use the retained features and Fisher's LDA for classification. In this approach, a crucial problem is how to set the threshold of IT. We approach this problem by adapting the recent innovation of Higher Criticism Thresholding (HCT).

We find that when useful features are rare and weak, the limiting behavior of HCT is essentially just as good as the limiting behavior of ideal threshold, the threshold one would choose if the underlying distribution of the signals is known (if only). Somewhat surprisingly, when Ω is sufficiently sparse, its off-diagonal coordinates usually do not have a major influence over the classification decision.

Compared to recent work in the case where Ω is the identity matrix [*Proc. Natl. Acad. Sci. USA* **105** (2008) 14790–14795; *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** (2009) 4449–4470], the current setting is much more general, which needs a new approach and much more sophisticated analysis. One key component of the analysis is the intimate relationship between HCT and Fisher's separation. Another key component is the tight large-deviation bounds for empirical processes for data with unconventional correlation structures, where graph theory on vertex coloring plays an important role.

1. Introduction.

Consider a two-class classification problem, where we have n labeled training samples (X_i, Y_i), $1 \le i \le n$. Here, X_i are p-dimensional

Received January 2013; revised August 2013.

[1] Supported in part by NSF CAREER Award DMS-11-50318, Grant DMS-09-06784 and USC Marshall summer research funding.

[2] Supported in part by NSF Grant DMS-12-08315.

[3] Supported in part by NSF Award SES-1061387.

MSC2010 subject classifications. Primary 62G05; secondary 62G32.

Key words and phrases. Chromatic number, Fisher's LDA, Fisher's separation, phase diagram, precision matrix, rare and weak model, sparse graph.

feature vectors and $Y_i \in \{-1, 1\}$ are the corresponding class labels. For simplicity, we assume two classes are *equally likely*, and the data are centered so that

$$X_i \sim N(Y_i \cdot \mu, \Sigma_{p,p}), \tag{1.1}$$

where μ is the contrast mean vector between two classes, and $\Sigma_{p,p}$ is the $p \times p$ covariance matrix. Given a fresh feature vector

$$X \sim N(Y \cdot \mu, \Sigma_{p,p}), \tag{1.2}$$

the goal is to train (X_i, Y_i) to decide whether $Y = -1$ or $Y = 1$. We denote $\Sigma_{p,p}^{-1}$ by $\Omega_{p,p}$, and whenever there is no confusion, we drop the subscripts "p, p" (and also that of any estimator of it, say, $\hat{\Omega}_{p,p}$).

We are primarily interested in the so-called "$p \gg n$" regime. In many applications where $p \gg n$ (e.g., genomics), we observe the following aspects.

- *Signals are rare*. Due to large p, the useful features (i.e., the nonzero coordinates of μ) are rare. For example, for a given type of cancer or disease, there are usually only a small number of relevant features (i.e., genes or proteins). When we measure increasingly more features, we tend to include increasingly more *irrelevant* ones.
- *Signals are individually weak*. The training data can be summarized by the z-vector

  $$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i X_i \sim N(\sqrt{n}\mu, \Sigma). \tag{1.3}$$

  Due to the small n, signals are weak in the sense that, individually, the nonzero coordinates of $\sqrt{n}\mu$ are small or moderately large at most.
- *Precision matrix Ω is sparse*. Take Genetic Regulatory Network (GRN) for example. The feature vector $X = (X(1), \dots, X(p))'$ represents the expression levels of p different genes, and is approximately distributed as $N(\mu, \Sigma)$. For any $1 \leq i \leq p$, it is believed that for all except a few j, $1 \leq j \leq p$, the gene pair (i, j) are conditionally independent given all other genes. In other words, each row of Ω has only a few nonzero entries, so Ω is sparse [13].

In many applications, Ω is unknown and has to be estimated. In many other applications such as complicate disease or cancer, decades of biomedical studies have accumulated huge databases which are sometimes referred to as "data-for-data" [37]. Such databases can be used to accurately estimate Ω independently of the data at hand, so Ω can be assumed as known. In this paper, we investigate both the case where Ω is known and the case where Ω is unknown. In either case, we assume Ω has unit diagonals:

$$\Omega(i, i) = 1, \qquad 1 \leq i \leq p. \tag{1.4}$$

Such an assumption is only for simplicity, and we do not use such information for inference.

1.1. *Fisher's LDA and modern challenges.* Fisher's Linear Discriminant Analysis (LDA) [24] is a well-known method for classification, which utilizes a weighted average of the test features $L(X) = \sum_{j=1}^{p} w(j)X(j)$, and predicts $Y = \pm 1$ if $L(X) >< 0$. Here, $w = (w(1), \dots, w(p))'$ is a preselected weight vector. Fisher showed that the optimal weight vector satisfies

$$w \propto \Omega\mu. \tag{1.5}$$

In the classical setting where $n \gg p$, μ and Ω can be conveniently estimated and Fisher's LDA is approachable. Unfortunately, in the modern regime where $p \gg n$, Fisher's LDA faces immediate challenges.

- It is challenging to estimate Ω simply because there are $O(p^2)$ unknown parameters but we have only $O(np)$ different measurements.
- Even in the simplest case where $\Omega = I_p$, challenges remain, as the signals are rare and weak. See [16] for the delicacy of the problem.

The paper is largely focused on addressing the second challenge. It shows that successful classification can be achieved by simultaneously exploiting the sparsity of μ (aka. signal sparsity) and the sparsity of Ω (aka. graph sparsity). For the first challenge, encouraging progresses have been made recently (e.g., [9, 25]), and the problem is more or less settled. Still, the paper has a two-fold contribution along this line. First, we show that the performances of the methods in [9, 25] can be substantially improved if we add an additional re-fitting step; see details in Section 4. Second, we carefully analyze how the errors in estimating Ω may affect the classification results.

1.2. *Innovated thresholding.* We wish to adapt Fisher's LDA to the current setting. Recall that the optimal choice of weight vector is $w \propto \Omega\mu$. If we have a reasonably good estimate of Ω (see Section 1.8 for more discussion on estimating Ω), say, $\hat{\Omega}$, all we need is a good estimate of μ.

When μ is sparse, one usually estimates it with some type of thresholding [18]. Let Z be the training z-vector as in (1.3). For some threshold t to be determined, there are three obvious approaches to thresholding:

- *Brute-force Thresholding* (BT). We apply thresholding to Z directly using the so-called clipping rule [16]: $\hat{\mu}_t^Z(i) = \mathrm{sgn}(Z(i))1\{|Z(i)| \geq t\}$. Alternatively, one may use soft thresholding or hard thresholding. However, numeric studies (e.g., [16]) suggest that different thresholding schemes only have small differences in classification errors, provided that these schemes use the same threshold picked from the range of interest. For this reason, we only study the clipping rule; same below.
- *Whitened Thresholding* (WT). We first whiten the noise by the transformation $Z \mapsto \hat{\Omega}^{1/2}Z \approx N(\sqrt{n}\Omega^{1/2}\mu, I_p)$, and then apply the thresholding to the vector $\hat{\Omega}^{1/2}Z$ in a similar fashion.

- *Innovated Thresholding* (IT). We first take the transformation $Z \mapsto \hat{\Omega} Z$ and then apply the thresholding by

$$\hat{\mu}_t^{\hat{Z}}(i) = \mathrm{sgn}(\hat{Z}(i)) \cdot 1\{|\hat{Z}(i)| \geq t\}, \qquad \text{where } \hat{Z} \equiv \hat{\Omega} Z. \tag{1.6}$$

The transformation $Z \mapsto \hat{\Omega} Z$ is connected to the term of *Innovation* in the literature of time series [27], and so the name of Innovated Thresholding.

It turns out that, among the three approaches, IT is the best. To see the point, note that for any $p \times p$ nonsingular matrix M, one could always estimate μ by applying the thresholding to MZ entry-wise (in BT, WT, and IT, $M = I_p$, $\Omega^{1/2}$, and Ω approximately). The deal is, what is the best M?

Toward this end, write $M = [m_1, m_2, \ldots, m_p]'$. For any $1 \leq i \leq p$, it is seen that $(MZ)(i) \sim N(\sqrt{n} m_i' \mu, m_i' \Sigma m_i)$. Therefore, if we bet on $\mu(i) \neq 0$, we should choose m_i to optimize the Signal-to-Noise Ratio (SNR) of $(MZ)(i)$. By the Cauchy–Schwarz inequality, the optimal m_i satisfies that $m_i \propto \Omega\mu$. Writing $\Omega = [\omega_1, \omega_2, \ldots, \omega_p]$, it is seen that

$$\Omega\mu = \mu(i)\omega_i + \sum_{k \neq i} \mu(k)\omega_k \equiv (I) + (II). \tag{1.7}$$

When we bet on $\mu(i) \neq 0$, $(I) \propto \omega_i$ which is accessible to us. However, (II) is a very noisy vector and is inaccessible to us, estimating which is equally hard as estimating μ itself.

In summary, if we bet on $\mu(i) \neq 0$, then the "best" accessible choice is $m_i \propto \omega_i$. As this holds for all i and we do not know where the signals are, the optimal choice for M is $M = \Omega$. This says that IT is not only the best among the three choices above, but is also the best choice in more general situations.

The heuristics above are consolidated in Sections 1.7–1.9, where we show that IT based classifiers achieve the optimal phase diagram for classification, while BT or WT based classifiers do not, even in very simple settings.

REMARK. The advantage of IT over WT and BT can be illustrated with the following example, which is further discussed later in Section 1.9 where we compare the phase diagrams of IT, WT, and BT. Suppose Ω is a block diagonal matrix where for $h \in (-1, 1)$ and $1 \leq i, j \leq p$,

$$\begin{aligned} \Omega(i, j) = 1\{i = j\} + h \cdot 1\{j - i = 1, i \text{ is odd}\} \\ + h \cdot 1\{i - j = 1, i \text{ is even}\}. \end{aligned} \tag{1.8}$$

According to the block structure of Ω, we also partition the vector μ into $p/2$ blocks, and each block has two entries. For simplicity, we suppose each block of μ has either no signal, or a single signal with a strength $\tau/\sqrt{n} > 0$. BT, WT, and IT apply thresholding to Z, $\Omega^{1/2}Z$, and ΩZ, correspondingly, where $Z \sim N(\sqrt{n}\mu, \Sigma)$ is the training z-vector as above. In this simple example, the

SNR for Z, $\Omega^{1/2}Z$, and ΩZ are $\sqrt{(1-h^2)}\tau$, $[\sqrt{(1+h)}+\sqrt{(1-h)}]\tau/2$ and τ correspondingly, with the last one being the largest (for the mean vector of $\Omega^{1/2}Z$ or ΩZ, the nonzero coordinates have two different magnitudes; the SNR is computed based on the larger magnitude).

REMARK. In (1.7), the point that (II) is generally noninformative in designing the best m_i can be further elaborated as follows: since we do not know the locations of other nonzero coordinates of μ, it makes sense to model $\{\sqrt{n}\mu(j) : 1 \leq j \leq p, j \neq i\}$ as i.i.d. samples from

$$(1-\varepsilon_p)\nu_0 + \varepsilon_p H_p, \qquad \varepsilon_p > 0\text{: small,} \tag{1.9}$$

where ν_0 is the point mass at 0 and H_p is some distribution with no mass at 0. Under general "rare and weak" conditions for μ and sparsity condition for Ω, entries of $E[(II)]$ are uniformly small.

In the literature of variable selection, IT is also called *marginal regression* [26]. The connection is not surprising, as approximately, $\hat{\Omega}^{1/2}Z \approx \Omega^{1/2}Z \sim N(\sqrt{n}\Omega^{1/2}\mu, I_p)$ which is a regression model. Both methods apply thresholding to ΩZ entry-wise, but marginal regression uses the hard thresholding rule, and IT uses the clipping thresholding rule [16].

With that being said, challenges remain on how to set the threshold t of IT [see (1.6)]. If we set t too small or too large, the resultant estimator $\hat{\mu}_t^{\hat{Z}}$ has too many or too few nonzeros. Our proposal is to set the threshold in a data driven fashion by using the recent innovation of Higher Criticism Thresholding (HCT).

1.3. *Threshold choice by higher criticism.*

Higher Criticism (HC) is a notion mentioned in passing by Tukey [44]. In recent years, HC was found to be useful in sparse signal detection [15], large-scale multiple testing [2, 10, 45], goodness-of-fit [33], and was applied to nonGaussian detection in Cosmic Microwave Background [12] and genomics [29, 39]. HC as a method for threshold choice in feature selection was first introduced in [16] (see also [28]), but the study has been focused on the case where Ω is the identity matrix. The case we consider in the current paper is much more complicated, where how to use HC for threshold choice is a nontrivial problem.

Our proposal is as follows. Let $\hat{\Omega}$ be a reasonably good estimate of Ω and let Z be the training z-vector as in (1.3). As in (1.6), denote for short

$$\hat{Z} = \hat{Z}(Z, \hat{\Omega}, p, n) = \hat{\Omega}Z. \tag{1.10}$$

The proposed approach contains three simple steps.

- For each $1 \leq j \leq p$, obtain a p-value by $\pi_j = P(|N(0,1)| \geq |\hat{Z}(j)|)$.
- Sort all the p-values in the ascending order $\pi_{(1)} < \pi_{(2)} < \cdots < \pi_{(p)}$.

- Define the HC functional $\mathrm{HC}_{p,j} = \sqrt{p}[j/p - \pi_{(j)}]/\sqrt{(1-j/p)j/p}$, $1 \leq j \leq p$. Let $\hat{j}$ be the index at which $\mathrm{HC}_{p,j}$ takes the maximum. The Higher Criticism Threshold (HCT)—denoted by $|\hat{Z}_{(\hat{j})}|$—is defined as the $\hat{j}$th largest coordinate of $(|\hat{Z}(1)|, \ldots, |\hat{Z}(p)|)'$.

Moreover, for stability, we need the following refinement. Define

$$s_p^* = \sqrt{2\log(p)}, \qquad \tilde{s}_{p,n}^* = \sqrt{2\max\{0, \log(p/n^2)\}}. \tag{1.11}$$

It is well-understood (e.g., [15, 27]) that the threshold should not be larger than s_p^*. At the same time, the threshold should not be too small, especially when n is small. The HCT we use in this paper is

$$t_p^{\mathrm{HC}} = \begin{cases} |\hat{Z}_{(\hat{j})}|, & \text{if } \tilde{s}_{p,n}^* \leq |\hat{Z}_{(\hat{j})}| \leq s_p^*, \\ \tilde{s}_{p,n}^*, & \text{if } |\hat{Z}_{(\hat{j})}| < \tilde{s}_{p,n}^*, \\ s_p^*, & \text{if } |\hat{Z}_{(\hat{j})}| > s_p^*. \end{cases} \tag{1.12}$$

See Sections 1.5 and 3 for more detailed discussion.

1.4. *HCT trained classifier.* We are now ready for classification. Let $\hat{\Omega}$ be as above, and let $\hat{\mu}_{\mathrm{HC}}^{\hat{Z}} = \hat{\mu}^{\hat{Z}}(Z, \hat{\Omega}, p, n)$ be defined as

$$\hat{\mu}_{\mathrm{HC}}^{\hat{Z}}(j) = \mathrm{sgn}(\hat{Z}(j)) \cdot 1\{|\hat{Z}(j)| \geq t_p^{\mathrm{HC}}\}, \qquad 1 \leq j \leq p. \tag{1.13}$$

Compared to $\hat{\mu}_t^{\hat{Z}}$ in (1.6), the only difference is that we have replaced t by t_p^{HC}. Introduce the HCT classification statistic

$$L_{\mathrm{HC}}(X, \hat{\Omega}) = L_{\mathrm{HC}}(X, \hat{\Omega}; Z, p, n) = (\hat{\mu}_{\mathrm{HC}}^{\hat{Z}})'\hat{\Omega}X. \tag{1.14}$$

The HCT trained classifier (or HCT classifier for short) is then the decision rule that decides $Y = \pm 1$ according to $L_{\mathrm{HC}}(X, \hat{\Omega}) >< 0$.

The innovation of the procedure is two-fold: using IT for feature selection and using HCT for threshold choice in the more complicated case where Ω is unknown and is nonidentity. The work is connected to other works on HC [16, 27], but the procedure and the delicate theory it entails are new.

A question is whether IT has any advantages over exsiting variable selection methods (e.g., the Lasso [42], SCAD [22], Dantzig selector [11]). The answer is yes, for the following reasons. First, compared to these methods, IT is computationally much faster and much more approachable for delicate analysis. Second, our goal is classification, not variable selection. For classification, especially when features are rare and weak, the choice of different variable selection methods is secondary, while the choice of the tuning parameter is crucial. The threshold of IT can be conveniently set by HCT, but how to set the tuning parameter of the Lasso, SCAD, or Dantzig selector remains an open problem, at least in theory.

How does the HCT classifier behave? In Sections 1.5–1.6, we set up a theoretic framework and derive a lower bound for classification errors. In Sections 1.7–1.8, we investigate the HCT classifier for the cases where Ω is known and unknown separately, and show that the HCT classifier yields optimal phase diagram in classification.

1.5. *Asymptotic rare and weak model.* Motivated by the application examples aforementioned, we use a *Rare and Weak* signal model as follows. We model the scaled contrast mean vector $\sqrt{n}\mu$ as

$$\sqrt{n}\mu(j) \overset{\text{i.i.d.}}{\sim} (1-\varepsilon_p)\nu_0 + \varepsilon_p H_p, \qquad 1 \le j \le p, \tag{1.15}$$

where as in (1.9), ν_0 is the point mass at 0, H_p is some distribution with no mass at 0, and $\varepsilon_p \in (0,1)$ is small [note that (ε_p, H_p) depend on p but not on j]. We use p as the driving asymptotic parameter, and link parameters (n, ε_p, H_p) to p through some fixed parameters. In detail, fixing parameters $(\beta, \theta) \in (0,1)^2$, we model

$$\varepsilon_p = p^{-\beta}, \qquad n = n_p = p^{\theta}. \tag{1.16}$$

As p tends to ∞, the sample size n_p grows to ∞ but in a slower rate than that of p; the signals get increasingly sparser but the number of signals tends to ∞. The interesting range of parameters (β, θ, H_p) partitions into three regimes, according to the sparsity level.

- *Relatively Dense* (RD). In this regime, $0 < \beta < (1-\theta)/2$. The signals are relatively dense and successful classification is possible even when signals are very faint [e.g., H_p concentrates its mass around a term $\tau_p \ll \sqrt{2\log(p)}$]. In such cases, (a) successful feature selection is impossible as signals are too weak, and (b) feature selection is unnecessary for the signals are relatively dense.
- *Rare and Weak* (RW). In this regime, $(1-\theta)/2 < \beta < (1-\theta)$, and the signals are moderately sparse. For successful classification, we need moderately strong signals [i.e., nonzero coordinates of $\sqrt{n}\mu \asymp \sqrt{\log(p)}$]. In this case, feature selection is subtle but could be substantially helpful. In contrast, classification is impossible if signals are much weaker than $\sqrt{\log(p)}$, and consistent feature selection is possible (and so the problem of classification is much less challenging) if the signals are much stronger than $\sqrt{\log(p)}$.
- *Rare and Strong* (RS). In this regime, $\beta > (1-\theta)$, and the signals are very sparse. For successful classification, we need very strong signals [signal strength $\gg \sqrt{\log(p)}$]. In this case, feature selection is comparably easier to carry out (but substantially helpful) since the signals are strong enough to stand out for themselves.

While the statements hold broadly, the most transparent way to understand them is probably to consider the case where H_p is a point mass at τ_p (say): in the above three regimes, the minimum τ_p required for successful classification (up to some multi-$\log(p)$ factors in the first and last regimes) are $1/(\varepsilon_p\sqrt{(p/n_p)})$, $\sqrt{\log(p)}$, and $\sqrt{n_p/(p\varepsilon_p)}$ correspondingly; the proof is elementary so is omitted.

In summary, feature selection is impossible in the RD regime and is relatively easy in the RS regime. For these reasons, we are primarily interested in the RW regime where we assume

$$(1-\theta)/2 < \beta < (1-\theta). \tag{1.17}$$

The RD/RS regimes are further discussed in Section 1.10, where we address the connection between our work and [8, 21, 40]. For β in this range, the most interesting range for the signal strength is when H_p concentrates its mass at the scale of $\sqrt{\log(p)}$. In light of this, we fix $r > 0$ and calibrate the signal strength parameter τ_p by

$$\tau_p = \sqrt{2r\log(p)}. \tag{1.18}$$

Except in Section 1.6 where we address the lower bound arguments, we assume H_p is a point mass [compare (1.15)]:

$$H_p = \nu_{\tau_p}, \qquad \text{where as in (1.18), } \tau_p = \sqrt{2r\log(p)} \text{ and } 0 < r < 1. \tag{1.19}$$

We focus on the case $0 < r < 1$, as the case $r > 1$ corresponds to RS regime where the classification is comparably easier. This models a setting where the signal strengths are equal. The case where the signal strengths are unequal is discussed in Section 1.12.

Next, we model Ω. Motivated by the previous example on Genetic Regulatory Network, we assume each row of Ω has relatively few nonzeros. Such a matrix naturally induces a sparse graph $\mathcal{G} = \mathcal{G}(\Omega) = (V, E)$, where $V = \{1, 2, \dots, p\}$ and there is an edge between nodes i and j if and only if $\Omega(i, j) \neq 0$; see [5] for basic terminology in graph theory.

DEFINITION 1.1. Fix p and $1 \leq K < p$. We call a $p \times p$ positive definite matrix Ω K-sparse if each row of Ω has at most K nonzeros. For any graph $\mathcal{G}$, we call $\mathcal{G}$ K-sparse if the degree of each node $\leq K$.

When Ω is K-sparse, the induced graph $\mathcal{G}(\Omega)$ is $(K-1)$ sparse, since by convention, there is no edge between a node and itself.

The class of K-sparse graphs is much broader than the class of banded graphs (we call $\mathcal{G}$ a banded graph with bandwidth K if nodes i and j are not connected whenever $|i-j| > K$). In fact, even when $\mathcal{G}$ is K-sparse with $K = 2$, we cannot always shuffle the nodes of $\mathcal{G}$ and make it a banded graph with a small bandwidth.

Let $\mathcal{M}_p$ be the class of all $p \times p$ positive definite correlation matrices. Fixing $a \in (0, 1)$, $b > 0$, and a sequence of integers K_p, introduce

$$\mathcal{M}_p^*(a, K_p) = \{\Omega \in \mathcal{M}_p \text{ and is } K_p\text{-sparse, } |\Omega(i, j)| \leq a, i \neq j\} \tag{1.20}$$

and

$$\widetilde{\mathcal{M}}_p^*(a,b,K_p) = \{\Omega \in \mathcal{M}_p^*(a,K_p), \|\Omega^{-1}\| \le b\}, \tag{1.21}$$

where $\|\cdot\|$ is the spectral norm. In comparison, $\widetilde{\mathcal{M}}_p^*(a,b,K_p)$ is slightly smaller than $\mathcal{M}_p^*(a,K_p)$. The following short-hand notation is frequently used in this paper.

DEFINITION 1.2. We use L_p to denote a strictly positive generic multi-$\log(p)$ term that may vary from occurrence to occurrence but always satisfies that for any fixed $c > 0$, $\lim_{p\to\infty}\{L_p p^{-c}\} = 0$ and $\lim_{p\to\infty}\{L_p p^c\} = \infty$.

In this paper, we are primarily interested in the case where K_p is at most multi-logarithmically large unless stated otherwise:

$$\lim_{p\to\infty} K_p = \infty, \qquad K_p \le L_p; \tag{1.22}$$

the first requirement is only for convenience. In our classification setting, $X_i \sim N(Y_i\mu, \Sigma)$, $X \sim N(Y\mu, \Sigma)$, and $Y = \pm 1$ with equal probabilities. The following notation is frequently used in the paper.

DEFINITION 1.3. We say the classification problem (1.1)–(1.2) satisfies the Asymptotic Rare Weak model ARW$(\beta, r, \theta, \Omega)$ if (1.15)–(1.16), (1.19) and (1.22) hold.

REMARK. The normalization in ARW is different from that in conventional asymptotic settings. In the latter, we usually fix μ and let n increase, so the classification problem becomes increasingly easier as n increase. In ARW, to focus on the "most interesting parameter regime", we fix $\sqrt{n}\mu$ and let n increase. As a result, the SNR in the summarizing training Z-vector remain the same, but the SNR in the testing vector X decrease rapidly as n increase. Therefore, the classification problem becomes increasingly harder as n increase.

1.6. *Lower bound.* Introduce the *standard phase boundary* function

$$\rho(\beta) = \begin{cases} 0, & 0 < \beta \le 1/2, \\ \beta - 1/2, & 1/2 < \beta < 3/4, \\ (1 - \sqrt{1-\beta})^2, & 3/4 \le \beta < 1, \end{cases} \tag{1.23}$$

and let

$$\rho_\theta^*(\beta) = (1-\theta)\rho\bigl(\beta/(1-\theta)\bigr), \qquad (1-\theta)/2 < \beta < (1-\theta).$$

The function ρ has appeared before in determining phase boundaries in a seemingly unrelated problem of multiple hypothesis testing [15, 30, 31]. The following theorem is proved in the supplementary material [23].

THEOREM 1.1. *Fix* $(\beta, r, \theta) \in (0,1)^3$ *such that* $(1-\theta)/2 < \beta < (1-\theta)$ *and* $0 < r < \rho_\theta^*(\beta)$. *Suppose* (1.15)–(1.16), (1.18), *and* (1.22) *hold and that for sufficiently large* p, $\Omega \in \mathcal{M}_p^*(a, K_p)$ *and the support of* H_p *is contained in* $[-\tau_p, \tau_p]$. *Then as* $p \to \infty$, *for any sequence of trained classifiers*, *the misclassification error* $\gtrsim 1/2$.

Note that in Theorem 1.1, we do not require the signals to have the same strength. Also, recall that in our classification setting (1.1)–(1.2), two classes are assumed as equally likely; extension to the case where two classes are unequally likely is straightforward. Theorem 1.1 was discovered before in [16, 32], but the study has been focused on the case where $\Omega = I_p$ and H_p is the point mass at τ_p. The proof in the current case is much more difficult and needs a few tricks, where graph theory on vertex coloring plays a key role. The following lemma is adapted from [5], Section V.1.

LEMMA 1.1. *Fix* $K \geq 1$. *For any graph* $\mathcal{G} = (V, E)$ *that is* K*-sparse*, *the chromatic number of* $\mathcal{G}$ *is no greater than* $(K+1)$.

Recall that when Ω is K sparse, then the induced graph $\mathcal{G} = \mathcal{G}(\Omega)$ is $(K-1)$ sparse, and so the chromatic number of $\mathcal{G}(\Omega) \leq K$. As a result, we can color the nodes of $\mathcal{G}(\Omega)$ with no more than K different colors, where there is no edge between any pair of nodes with the same color.

Despite its seemingly simplicity, Lemma 1.1 has far-reaching implications. Lemma 1.1 is the corner stone for proving the lower bound and for analyzing the HCT classifier (where we need tight convergence rate of empirical processes for data with unconventional correlation structures).

1.7. *HCT achieves optimal phase diagram in classification* (Ω *is known*). One noteworthy aspect of HCT classifier is that it achieves the optimal phase diagram. In this section, we show this for the case where Ω is known. In this case, the HCT classifier $L_{\mathrm{HC}}(X, \hat{\Omega})$ reduces to $L_{\mathrm{HC}}(X, \Omega)$ (the term formed by replacing $\hat{\Omega}$ by Ω everywhere in the definition of former). Since we predict the label associated with X as ± 1 according to $L_{\mathrm{HC}}(X, \Omega) >< 0$, the predicted label is correct if and only if $Y \cdot L_{\mathrm{HC}}(X, \Omega) > 0$. The following theorem is proved in Section 3.3.

THEOREM 1.2. *Fix* $(\beta, r, \theta, a) \in (0,1)^4$ *such that* $(1-\theta)/2 < \beta < (1-\theta)$ *and* $r > \rho_\theta^*(\beta)$. *Consider a sequence of classification problems* $\mathrm{ARW}(\beta, r, \theta, \Omega)$ *with* $\Omega \in \widetilde{\mathcal{M}}_p^*(a, b, K_p)$ *for sufficiently large* p. *Then as* p *tends to* ∞, $P(Y \cdot L_{\mathrm{HC}}(X, \Omega) < 0) \to 0$. *When* $r < \beta$, *the condition on* Ω *can be relaxed to that of* $\Omega \in \mathcal{M}_p^*(a, K_p)$.

Call the two-dimensional space $\{(\beta, r): 0 < \beta < 1, 0 < r < 1\}$ the phase space. Theorems 1.1–1.2 say that the phase space partitions into two separate regions, *Region of Impossibility* and *Region of Possibility*, where the classification problem is distinctly different.

- *Region of Impossibility*. $\{(\beta, r) : (1-\theta)/2 < \beta < (1-\theta), 0 < r < \rho_\theta^*(\beta)\}$. Fix (β, r) in the interior of this region and consider a sequence of classification problems with $p^{1-\beta}$ signals where each signal $\leq \sqrt{2r\log(p)}$ in strength. Then for any sequence of "sparse" Ω, successful classification is impossible. This is the most difficult case where not much can be done for classification aside from random guessing.
- *Region of Possibility*. $\{(\beta, r) : (1-\theta)/2 < \beta < (1-\theta)\}, \rho_\theta^*(\beta) < r < 1\}$. Fix (β, r) in the interior of this region and suppose signals have equal strength of $\sqrt{2r\log(p)}$. HCT classifier $L_{\mathrm{HC}}(X, \Omega)$ yields successful classification (the results hold much more broadly where equal signal strength assumption can be largely relaxed).

We call the curve $r = \rho_\theta^*(\beta)$ the *separating boundary*. Somewhat surprisingly, the separating boundary does not depend on the off-diagonals of Ω. The partition of phase diagram was discovered by [16, 35], and independently by [32], but the focus was on the case where $\Omega = I_p$. See also [28]. The study in the current case is much more difficult. Similar phase diagrams are also found in sparse signal detection [15], variable selection [34], and spectral clustering [36].

Why HCT works? The key insight is that there is an intimate relationship between the HC functional and Fisher's separation; the latter plays a key role in determining the optimal classification behavior, but is, unfortunately, an *oracle* quantity which depends on unknown parameters. In Sections 2–3, we outline a series of theoretic results, explaining why the HCT classifier is the right approach and how it achieves the optimality.

1.8. *Optimality of HCT classification* (Ω *is unknown*). When Ω is unknown, we first estimate it with the training data.

DEFINITION 1.4. For any sequence of $\Omega_{p,p} \in \mathcal{M}_p^*(a, K_p)$, we say an estimator $\hat{\Omega}_{p,p}$ is acceptable if it is symmetric and independent of the test feature vector X, and that there is a constant $C > 0$ such that for sufficiently large p, $\hat{\Omega}_{p,p}$ is K_p'-sparse where $K_p' \leq L_p$, and $|\hat{\Omega}_{p,p}(i, j) - \Omega_{p,p}(i, j)| \leq CK_p^2\sqrt{\log(p)}/\sqrt{n_p}$ for all $1 \leq i, j \leq p$.

Usually, the $(L_p/\sqrt{n_p})$-rate cannot be improved, even when Ω is diagonal. For K_p-sparse Ω satisfying (1.22), acceptable estimators can be constructed based on the recent CLIME approach in [9]. If additionally Ω satisfies the mutual incoherence condition [38], Assumption 1, then the glasso [25] is also acceptable, provided the tuning parameters are properly set. If Ω is banded, then the Bickel and Levina Thresholding (BLT) method [4] is also acceptable, up to some modifications.

With that being said, the numeric performances of all these estimators can be improved with an additional step of *re-fitting*. See Section 4 for details.

Naturally, the estimation error of $\hat{\Omega}$ has some negative effects on the HCT classifier. Fortunately, for a large fraction of parameters (β, r) in Region of Possibility, such effects are negligible and HCT continues to yield successful classification. In detail, recalling that $n_p = p^\theta$, $\theta \in (0, 1)$, we suppose:

- *Condition* (a). $r > \max\{(1-2\theta)/4, \rho^*_\theta(\beta)\}$,
- *Condition* (b). When $0 < \theta \leq 1/3$ and $(1-\theta)/2 < \beta < (1-2\theta)$, $|r - \sqrt{1-2\theta}| \geq \sqrt{1-2\theta-\beta}$.

The following theorem is proved in Section 3.3.

THEOREM 1.3. *Fix* $(\beta, r, \theta, a) \in (0,1)^4$ *such that* $(1-\theta)/2 < \beta < (1-\theta)$, *and conditions* (a)–(b) *hold. Consider a sequence of classification problems* $\mathrm{ARW}(\beta, r, \theta, \Omega)$ *such that* $\Omega \in \mathcal{M}^*_p(a, K_p)$ *when* $r < \beta$ *and* $\Omega \in \widetilde{\mathcal{M}}^*_p(a, b, K_p)$ *when* $r \geq \beta$. *For the HCT classifier* $L_{\mathrm{HC}}(X, \hat{\Omega})$, *if* $\hat{\Omega}$ *is acceptable, then as p tends to* ∞, $P(Y \cdot L_{\mathrm{HC}}(X, \hat{\Omega}) < 0) \to 0$.

We remark that, first, when $0 < \theta \leq 1/4$ and $(1-\theta)/2 < \beta < 3(1-2\theta)/4$, condition (a) can be relaxed to that of $r > \max\{\beta/3, \rho^*_\theta(\beta)\}$. Second, when $\theta > 1/2$, conditions (a)–(b) automatically hold when $r > \rho^*_\theta(\beta)$. As a result, we have the following corollary, the proof of which is omitted.

COROLLARY 1.1. *When* $\theta > 1/2$, *Theorem* 1.3 *holds with conditions* (a)–(b) *replaced by that of* $r > \rho^*_\theta(\beta)$.

This says that as long as $n_p \gg \sqrt{p}$, the estimation errors of any acceptable estimator $\hat{\Omega}$ have negligible effects over the classification decision.

1.9. *Comparison with BT and WT.* In disguise, many methods are what we called "Brute-forth Thresholding" or "BT," including but not limited to [3, 19, 20, 43]. Since Ω is hard to estimate, Bickel and Levina [3], Fan and Fan [20], and Tibshirani et al. [43] neglect the off-diagonals in Σ for classification. In a seemingly different spirit, Efron [19] proposes a procedure where he first selects features by neglecting the off-diagonals in Σ and then estimates the correlation structures among selected features. However, under the Rare and Weak model, selected features tend to be uncorrelated. Therefore, at least for many cases, the approach fails to exploit the "local" graphic structure of the data and is "BT" in disguise. It is also noteworthy that [43] proposes to set the threshold of BT by cross validation, which is unstable, especially when n_p is small.

When we replace IT by either BT or WT in HCT classifier, the phase diagram associated with the resultant procedure is no longer optimal. While the claim holds very broadly, it can be conveniently illustrated with a simple example as follows.

Consider the same setting as in Theorem 1.2, except that Ω is the matrix defined in (1.8). That is, Ω is the diagonal block-wise matrix where each diagonal block is the 2×2 matrix with 1 on the diagonals and h on the off-diagonals, $h \in (-1, 1)$. In this simple case, by Theorem 1.2, HCT classifier gives successful classification when $r > \rho_\theta^*(\beta)$, and fails when $r < \rho_\theta^*(\beta)$. In comparison, if we use BT (which treats Σ as diagonal and does not incorporate correlations for classification), the separating function for success and failure becomes $r = \rho_\theta^*(\beta)/(1 - h^2)$, which is higher than $r = \rho_\theta^*(\beta)$ in the β-r plane (a similar claim holds for WT, but the separating function is $r = 2\rho_\theta^*(\beta)/[1 + \sqrt{1 - h^2}]$; note $2/[1 + \sqrt{1 - h^2}] > 1$ for all $h \neq 0$). Recall that when Ω is given, the *only* difference between the HCT classifier built over IT and the HCT classifier built over BT is that, for any threshold t, BT and IT estimate μ by

$$\hat{\mu}_t^Z(i) = \operatorname{sgn}(Z(i))1\{|Z(i)| \geq t\} \quad \text{and} \quad \hat{\mu}_t^{\tilde{Z}}(i) = \operatorname{sgn}(\tilde{Z}(i))1\{|\tilde{Z}(i)| \geq t\},$$

respectively, where $\tilde{Z} = \Omega Z$; see Section 1.2 for details. We have the following theorem, the proof of which is elementary so is omitted.

THEOREM 1.4. *Fix* $(\beta, \theta, r) \in (0, 1)^3$ *such that* $(1 - \theta)/2 < \beta < (1 - \theta)$. *Consider a sequence of classification problems* $\mathrm{ARW}(\beta, r, \theta, \Omega)$ *where* Ω *is the diagonal block-wise matrix defined in* (1.8). *Suppose we apply HCT classifier built over the Brute-force Thresholding* (*BT*) *as in Section* 1.2. *As* $p \to \infty$, *the classification error* $\to 0$ *if* $r > \rho_\theta^*(\beta)/(1 - h^2)$, *and the classification error* $\to 1/2$ *if* $r < \rho_\theta^*/(1 - h^2)$.

1.10. *Comparison with works focused on the RS regime.* The work is closely related to the recent approach by Shao et al. [40], the ROAD approach by Fan et al. [21], and the LPD approach by Cai and Liu [8]. While all approaches attempt to mimic Fisher's LDA, the difference lies in how we estimate the "ideal weight vector" w prescribed in (1.5). In our notation, Shao et al. [40] estimates w by $(\Sigma^*)^{-1}\hat{\mu}_t^Z$, where Σ^* is the regularized estimation of Σ as in Bickel and Levina [4] for an appropriate threshold, and $\hat{\mu}_t^Z$ is the estimation of μ by Brute-force Thresholding. ROAD estimates w by minimizing $(1/2)w'\hat{\Sigma}w + \lambda\|w\|_1 + (1/2)\gamma(w'Z - 1)^2$, and LPD estimates w by minimizing $\|w\|_1$ subject to the constraint of $\|\hat{\Sigma}\beta - Z\|_\infty \leq \lambda$, where λ and γ are tuning parameters.

In disguise, these works focused on the "Rare and Strong" regime according to our terminology. In fact, Shao et al. [40] assumes the minimum signal strength (smallest coordinate in magnitude of $\sqrt{n_p}\mu$) is of the order of $\sqrt{n_p}$, and the main results of Fan et al. [21] and Cai and Liu [8] (i.e., [21], Theorem 3, [8], Theorem 1) assume a sparsity constraint that can be roughly translated to $\beta > (1 - \theta/2)$ in our notation. Seemingly, this concerns the RS Regime we mentioned earlier.

Compared to these works, our work focuses on the most challenging regime where the signals are Rare and Weak, and we need much more sophisticated methods for feature selection and for threshold choices.

1.11. *Comparison with other popular classifiers.* HCT classifier also has advantages over other well-known classifiers such as the Support Vector Machine (SVM) [7], Random Forest [6], and Boosting [14]. These methods need tuning parameters and are internally very complicated, but they do not outperform HCT classifier even when we replace the IT by BT; see details in [16], where we have compared all these methods with three well-known gene microarray data sets in the context of cancer classification.

HCT is also closely related to PAM [43], but is different in some important aspects. First, HCT exploits the correlation structure while PAM does not. Second, while both methods perform feature selection, PAM sets the threshold by cross validations (CVT), and HCT sets the threshold by Higher Criticism. When n is small, CVT is usually unstable. In [16], we have shown that HCT outperforms CVT when analyzing three microarray data sets aforementioned. In Section 4, we further compare HCT with CVT with simulated data.

1.12. *Summary and possible extensions.* We propose HCT classifier for two-class classification, where the major methodological innovation is the use of IT for feature selection and the use of HC for threshold choice.

IT is based on an “optimal” linear transform that maximizes SNR in all signal locations, and has advantages over BT and WT. IT also has a three-fold advantage over the well-known variable selection methods such as the Lasso, SCAD, and Dantzig selector: (a) IT is computationally faster, (b) IT is more approachable in terms of delicate analysis, and (c) the tuning parameter of IT can be conveniently set, but how to set the tuning parameters of the other methods remains an open problem.

The idea of using HC for threshold choice goes back to [16], where the focus is on the case where Ω is the identity matrix (see also [28]). In this paper, with considerable efforts, we extend the ideas to the case where Ω is unknown but is presumably sparse, and show that HC achieves the optimal phase diagram in classification. The optimality of HC is not coincidental, and the underlying reason is the intimate relationship between the HC functional and Fisher’s separation. This is explained in Sections 2–3 with details.

In Theorems 1.2–1.3 and Sections 2–3, we assume the signals have the same signs and strengths. The first assumption is largely for simplicity and can be removed. The second assumption can be largely relaxed, and both Theorems 1.2–1.3 and the intimate relationship between HC and Fisher’s separation continue to hold to some extent if the signal strengths are unequal. One such example is where the signal distribution H_p, after scaled by a factor of $(\log(p))^{-1/2}$, has a continuous density over a closed interval contained in $(0, \infty)$ which does not depend on p.

In the paper, we require Ω to be K_p-sparse where $K_p \leq L_p$ (see Definition 1.2) and does not exceed a multi-$\log(p)$ term. This assumption is mainly used to control the chromatic number of the induced graph $\mathcal{G}(\Omega)$. Since the chromatic number of a graph could be much smaller than its maximum degree, the assumption on Ω

can be relaxed to that of the chromatic number of $g(\Omega)$ does not exceed a multi-$\log(p)$ term. Also, when Ω has many small nonzero coordinates, we can always regularize it first with a threshold $t > 0$: $\Omega^*(i, j) = \Omega(i, j)1\{|\Omega(i, j)| \geq t\}$, and the main results continue to hold if Ω^* is K-sparse and the difference between two matrices is "sufficiently small."

1.13. *Content.* The remaining part of the paper is organized as follows. In Section 2, we introduce two functionals: Fisher's separation and ideal HC, and show that the two functionals are intimately connected to each other. In Section 3, we derive a large-deviation bound on the empirical c.d.f., and then use it to characterize the stochastic fluctuation of the HC functional and that of Fisher's separation. Theorems 1.2–1.3 are proved at the end of this section. All other claims (theorems and lemmas) are proved in the supplementary material [23]. Section 4 contains numeric examples.

1.14. *Notation.* In this paper, $C > 0$ and $L_p > 0$ denote a generic constant and a generic multi-$\log(p)$ term respectively, which may vary from occurrence to occurrence. For two positive sequences $\{a_p\}_{p=1}^{\infty}$ and $\{b_p\}_{p=1}^{\infty}$, we say that $a_p \gtrsim b_p$ (or $a_p \lesssim b_p$) if there is a sequence $\{\Delta_p\}_{p=1}^{\infty}$ such that $\Delta_p \to 0$ and $a_p(1 + \Delta_p) \geq b_p$ [or $a_p(1 + \Delta_p) \leq b_p$]. We say that $a_p \sim b_p$ if $a_p \gtrsim b_p$ and $a_p \lesssim b_p$, and we say that $a_p \asymp b_p$ if there is a constant $c_0 > 1$ such that for sufficiently large p, $c_0^{-1} \leq a_p/b_p \leq c_0$.

The notation Ω and Σ are always associated with each other by $\Omega = \Sigma^{-1}$, and (X_i, Y_i) represents a training sample while (X, Y) represents a test sample. The summarizing z-vector for the training data set is denoted by Z, with $\tilde{Z} = \Omega Z$ and $\hat{Z} = \hat{\Omega} Z$, where $\hat{\Omega}$ is some estimate of Ω.

2. Ideal threshold and ideal HCT.

In Sections 2–3, we discuss the behavior of HCT classifier. We limit our discussion to the ARW$(\beta, r, \theta, \Omega)$ model, but the key ideas are valid beyond the ARW model and extensions are possible; see discussions in Section 1.12.

The key insight behind the HCT methodology is that in a broad context,

$$\text{HCT} \approx \text{ideal HCT} \approx \text{ideal threshold}.$$

The ideal HCT is the nonstochastic counterpart of HCT, and the ideal threshold is the threshold one would choose if the underlying signal structure were known.

In this section, we elaborate the intimate connection between the ideal HCT and the ideal threshold, and their connections to Fisher's separation. We also investigate the performance of "ideal classifier" where we assume Ω is known and the threshold is set ideally.

The connection between HCT and ideal HCT is addressed in Section 3, which is new even in the case of $\Omega = I_p$; compare [17]. Theorems 1.2–1.3 are also proved in Section 3.

2.1. *Fisher's separation and classification heuristics.* Fix a threshold $t > 0$ and let $\hat{\Omega}$ be an acceptable estimator of Ω. We are interested in the classifier that estimates $Y = \pm 1$ according to $L_t(X, \hat{\Omega}) >< 0$, whereas in (1.13)–(1.14),

$$L_t(X, \hat{\Omega}) = \big(\hat{\mu}_t^{\hat{Z}}\big)' \hat{\Omega} X \qquad \text{where } \hat{\mu}_t^{\hat{Z}}(j) = \operatorname{sgn}\big(\hat{Z}(j)\big) 1\big\{|\hat{Z}(j)| \geq t\big\}.$$

For any fixed $p \times 1$ vector Z and $p \times p$ positive definite matrix A, we introduce

$$M_p(t, Z, \mu, A) = M_p(t, Z, \mu, A; n_p) = \big(\hat{\mu}_t^{Z}\big)' A \mu$$

and

$$V_p(t, Z, A) = V_p(t, Z, A; \Omega) = \big(\hat{\mu}_t^{Z}\big)' A \Omega^{-1} A \hat{\mu}_t^{Z},$$

where loosely, "M" and "V" stand for the mean and variance, respectively. In our model, given $(\mu, \hat{Z}, \hat{\Omega})$, the test sample $X \sim N(Y \cdot \mu, \Omega^{-1})$; see (1.2) and note that $\hat{\Omega}$ is independent of X since it is acceptable. It follows that

$$L_t(X, \hat{\Omega}) \sim N\big(Y \cdot M_p(t, \hat{Z}, \mu, \hat{\Omega}), V_p(t, \hat{Z}, \hat{\Omega})\big),$$

and the misclassification error rate of $L_t(X, \hat{\Omega})$ is

$$P\big(Y \cdot L_t(X, \hat{\Omega}) < 0 | \mu, \hat{Z}, \hat{\Omega}\big) = \bar{\Phi}\bigg(\frac{M_p(t, \hat{Z}, \mu, \hat{\Omega})}{\sqrt{V_p(t, \hat{Z}, \hat{\Omega})}}\bigg), \tag{2.1}$$

where $\bar{\Phi} = 1 - \Phi$ denotes the survival function of $N(0, 1)$.

The right-hand side of (2.1) is closely related to the well-known Fisher's separation (Sep) [1], which measures the standardized interclass distance $\mathrm{Sep}(t, \hat{Z}, \mu, \hat{\Omega}) = \mathrm{Sep}(t, \hat{Z}, \mu, \hat{\Omega}; \Omega, p)$:

$$\mathrm{Sep}(t, \hat{Z}, \mu, \hat{\Omega}; \Omega, p) = \frac{E[L_t(X, \hat{\Omega}) | Y = 1] - E[L_t(X, \hat{\Omega}) | Y = -1]}{SD(L_t(X, \hat{\Omega}))}. \tag{2.2}$$

In fact, it is seen that $\mathrm{Sep}(t, \hat{Z}, \mu, \hat{\Omega}) = 2M_p(t, \hat{Z}, \mu, \hat{\Omega})/\sqrt{V_p(t, \hat{Z}, \hat{\Omega})}$, and (2.1) can be rewritten as

$$P\big(Y \cdot L_t(X, \hat{\Omega}) < 0 | \mu, \hat{Z}, \hat{\Omega}\big) = \bar{\Phi}\big(\tfrac{1}{2} \mathrm{Sep}(t, \hat{Z}, \mu, \hat{\Omega})\big).$$

By (1.15) and (1.19), the overall misclassification error rate is then

$$P\big(Y \cdot L_t(X, \hat{\Omega}) < 0\big) = E_{\varepsilon_p, \tau_p} E\big[\bar{\Phi}\big(\tfrac{1}{2} \mathrm{Sep}(t, \hat{Z}, \mu, \hat{\Omega})\big)\big], \tag{2.3}$$

where E is the expectation with respect to the law of $(\hat{Z}, \hat{\Omega} | \mu)$, and $E_{\varepsilon_p, \tau_p}$ is the expectation with respect to the law of μ; see (1.15) and (1.19).

We introduce two proxies for Fisher's separation. Throughout this paper,

$$\tilde{Z} = \Omega Z. \tag{2.4}$$

For the first proxy, recall that $\hat{Z} = \hat{\Omega} Z$ [e.g., (1.10)]. Heuristically, $\hat{\Omega} \approx \Omega$ and so $\hat{Z} \approx \tilde{Z}$. We expect that $\mathrm{Sep}(t, \hat{Z}, \mu, \hat{\Omega}) \approx \mathrm{Sep}(t, \tilde{Z}, \mu, \Omega)$; the latter is Fisher's separation for the idealized case where Ω is known and is defined as

$$\mathrm{Sep}(t, \tilde{Z}, \mu, \Omega) = 2M_p(t, \tilde{Z}, \mu, \Omega)/\sqrt{V_p(t, \tilde{Z}, \Omega)}. \tag{2.5}$$

For the second proxy, we note that when p is large, some regularity appears, and we expect that $M_p(t, \tilde{Z}, \mu, \Omega) \approx m_p(t, \varepsilon_p, \tau_p, \Omega)$ and $V_p(t, \tilde{Z}, \Omega) \approx v_p(t, \varepsilon_p, \tau_p, \Omega)$, where

$$\begin{aligned} m_p(t, \varepsilon_p, \tau_p, \Omega) &= E\bigl[M_p(t, \tilde{Z}, \mu, \Omega)\bigr], \\ v_p(t, \varepsilon_p, \tau_p, \Omega) &= E\bigl[V_p(t, \tilde{Z}, \Omega)\bigr]. \end{aligned} \tag{2.6}$$

In light of this, a second proxy separation is the *population* Sep:

$$\widetilde{\mathrm{Sep}}(t) = \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega) = 2m_p(t, \varepsilon_p, \tau_p, \Omega)/\sqrt{v_p(t, \varepsilon_p, \tau_p, \Omega)}.$$

In summary, we expect to see that

$$\mathrm{Sep}(t, \hat{Z}, \mu, \hat{\Omega}) \approx \mathrm{Sep}(t, \tilde{Z}, \mu, \Omega) \approx \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega),$$

and that

$$P\bigl(Y \cdot L_t(X, \hat{\Omega}) < 0\bigr) \approx \bar{\Phi}\bigl(\tfrac{1}{2}\widetilde{\mathrm{Sep}}(t)\bigr). \tag{2.7}$$

In Section 3, we solidify the above connections. But before we do that, we study the ideal threshold—the threshold that maximizes $\widetilde{\mathrm{Sep}}(t)$.

2.2. *Ideal threshold.* Ideally, one would choose t to minimize the classification error of $L_t(X, \hat{\Omega})$. In light of (2.7), this is almost equivalent to choosing t as the ideal threshold.

DEFINITION 2.1. The ideal threshold $T_{\mathrm{ideal}}(\varepsilon_p, \tau_p, \Omega)$ is the maximizing point of the second proxy: $T_{\mathrm{ideal}}(\varepsilon_p, \tau_p, \Omega) = \mathrm{argmax}_{\{0<t<\infty\}} \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega)$.

In general, $\widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega)$ and $T_{\mathrm{ideal}}(\varepsilon_p, \tau_p, \Omega)$ may depend on Ω in a complicated way. Fortunately, it turns out that for large p and all Ω in $\mathcal{M}_p^*(a, K_p)$ [see (1.20)], the leading terms of $\widetilde{\mathrm{Sep}}(t)$ and $T_{\mathrm{ideal}}(\varepsilon_p, \tau_p, \Omega)$ do not depend on the off-diagonals of Ω and have rather simple forms.

DEFINITION 2.2 (Folding). Denote $\Psi_\tau(t) = P(|N(\tau, 1)| \leq t)$. When $\tau = 0$, we drop the subscript and write $\Psi(t)$. Also, denote $\bar{\Psi}_\tau = 1 - \Psi_\tau(t)$ and $\bar{\Psi}(t) = 1 - \Psi(t)$.

In detail, let

$$\widetilde{W}_0(t) = \widetilde{W}_0(t, \varepsilon_p, \tau_p; \Psi) = \varepsilon_p \bar{\Psi}_{\tau_p}(t) / \sqrt{\bar{\Psi}(t) + \varepsilon_p \bar{\Psi}_{\tau_p}(t)}, \tag{2.8}$$

$$t_p^*(\beta, r) = \min\left\{2, \frac{r+\beta}{2r}\right\}\tau_p \tag{2.9}$$

and

$$\delta(\beta, r) = \begin{cases} \beta - r, & r \le \beta/3, \\ \dfrac{(\beta+r)^2}{8r}, & \beta/3 < r < \beta, \\ \beta/2, & \beta \le r < 1. \end{cases} \tag{2.10}$$

Elementary calculus shows that for large p,

$$\operatorname*{argmax}_{\{0 \le t < \infty\}} \{\widetilde{W}_0(t)\} \sim t_p^*(\beta, r), \qquad \sup_{\{0 \le t < \infty\}} \widetilde{W}_0(t) = L_p \cdot p^{-\delta(\beta, r)}. \tag{2.11}$$

It turns out that there is an intimate relationship between $\widetilde{\operatorname{Sep}}(t, \varepsilon_p, \tau_p, \Omega)$ and $\widetilde{W}_0(t, \varepsilon_p, \tau_p)$, where the latter does not depend on the off-diagonals of Ω. To see the point, we discuss the cases $r < \beta$ and $r \ge \beta$ separately.

In the first case, for a as in $\mathcal{M}_p^*(a, K_p)$, we let

$$\begin{aligned} c_0(\beta, r, a) &= \delta(\beta, a^2 r) - \delta(\beta, r), \\ \tilde{c}_0(\beta, r, a) &= \tilde{c}_1(\beta, r, a) - \delta(\beta, r), \end{aligned} \tag{2.12}$$

where $c_0(\beta, r, a) > 0$ for $r < \beta$; if $a < 1/3$, $\tilde{c}_1(\beta, r, a) = \beta$, and otherwise,

$$\tilde{c}_1(\beta, r, a) = \begin{cases} \dfrac{(3a-1)r}{3-a} + \beta, & r \le \dfrac{3-a}{1+5a}\beta, \\ \dfrac{3-a}{1+a}\dfrac{(\beta+r)^2}{8r}, & \dfrac{3-a}{1+5a}\beta < r < \beta. \end{cases}$$

The following lemma is proved in the supplementary material [23].

LEMMA 2.1. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $\rho_\theta^*(\beta) < r < \beta$ and $(1-\theta)/2 < \beta < (1-\theta)$. In the* ARW$(\beta, r, \theta, \Omega)$ *model, as $p \to \infty$,*

$$\begin{aligned} &\sup_{t>0} \sup_{\{\Omega \in \mathcal{M}_p^*(a, K_p)\}} \left| p^{(\theta-1)/2} \widetilde{\operatorname{Sep}}(t, \varepsilon_p, \tau_p, \Omega) - 2\tau_p \widetilde{W}_0(t, \varepsilon_p, \tau_p) \right| \\ &\quad \le L_p p^{-\max\{\beta - r/2, (3\beta + r)/4\}} \\ &\qquad + L_p \left[p^{-\min\{r, (\beta - r)/2, (1-a)(\beta - ar)\}} + p^{-c_0(\beta, r, a)} + p^{-\tilde{c}_0(\beta, r, a)} \right] \\ &\qquad \times \sup_{\{0 < t < \infty\}} \widetilde{W}_0(t, \varepsilon_p, \tau_p). \end{aligned}$$

Note that $\delta(\beta, r) < \max(\beta - r/2, (3\beta + r)/4)$. As a result, approximately, $\widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega) \propto \widetilde{W}_0(t, \varepsilon_p, \tau_p)$ for all $\Omega \in \mathcal{M}_p^*(a, K_p)$. Combining this with (2.11), we expect to have

$$T_{\text{ideal}}(\varepsilon_p, \tau_p, \Omega) \sim t_p^*(\beta, r),$$
$$\sup_{0<t<\infty} \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega) = L_p p^{(1-\theta)/2-\delta(\beta,r)}. \tag{2.13}$$

Next, consider the case $r \geq \beta$. The lemma below is proved in the supplementary material [23].

LEMMA 2.2. *Fix* $(\beta, r, \theta, a) \in (0, 1)^4$ *such that* $r \geq \beta$ *and* $(1 - \theta)/2 < \beta < (1 - \theta)$. *Let* $\Delta_1 = d_0 \log(\log(p))/\sqrt{\log p}$ *and* $\Delta_2 = 2\sqrt{\log(K_p \log p)}$, *where* $d_0 > 0$ *is some constant. In the* $\mathrm{ARW}(\beta, r, \theta, \Omega)$ *model with* $\Omega \in \widetilde{\mathcal{M}}_p^*(a, b, K_p)$, *as* $p \to \infty$,

(a) $\sup_{\{0<t<\sqrt{2\beta\log(p)}-\Delta_1\}} \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega) \lesssim \frac{5}{3}\tau_p K_p^{-1} p^{(1-\theta-\beta)/2}$,

(b) $\sup_{\{t\geq\tau_p+\Delta_2\}} \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega) \lesssim \frac{5}{3}\tau_p K_p^{-1} p^{(1-\theta-\beta)/2}$,

(c) $\sup_{\{\sqrt{2\beta\log p}-\Delta_1\leq t<\tau_p\}} \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega) \gtrsim 2\tau_p K_p^{-1} p^{(1-\theta-\beta)/2}$ *and* $\sup_{\{t>0\}} \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega) \leq L_p p^{(1-\theta-\beta)/2}$.

A direct result of Lemma 2.2 is that, for all $\Omega \in \widetilde{\mathcal{M}}_p^*(a, b, K_p)$ [see (1.20)],

$$\sqrt{2\beta\log(p)} \lesssim T_{\text{ideal}} \lesssim \sqrt{2r\log(p)},$$
$$\sup_{\{0<t<\infty\}} \{\widetilde{\mathrm{Sep}}(t)\} \asymp L_p p^{(1-\theta-\beta)/2}, \tag{2.14}$$

where $T_{\text{ideal}} = T_{\text{ideal}}(\varepsilon_p, \tau_p, \Omega)$ and $\widetilde{\mathrm{Sep}}(t) = \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega)$ for short. In this case, the function $\widetilde{\mathrm{Sep}}(t)$ sharply increases and decreases in the intervals $(0, \sqrt{2\beta\log(p)})$ and $(\sqrt{2r\log(p)}, \infty)$, respectively, but is relatively flat in the interval $(\sqrt{2\beta\log(p)}, \sqrt{2r\log(p)})$; in this interval, the function reaches the maximum but varies slowly at the magnitude of $O(L_p p^{(1-\theta-\beta)/2})$. In the current case, on one hand, it is not critical to pin down T_{ideal}, as $\widetilde{\mathrm{Sep}}(t) = L_p p^{(1-\theta-\beta)/2}$ for all t in the whole interval. On the other hand, it is hard to pin down T_{ideal} uniformly for all Ω under consideration, if possible at all.

2.3. *Ideal HCT.* Ideal HCT is a counterpart of HCT and a nonstochastic threshold that HCT tries to estimate. Introduce a functional which is defined over all survival functions associated with a positive random variable:

$$\mathrm{HC}(t, G) = \sqrt{p}[G(t) - \bar{\Psi}(t)]/\sqrt{G(t)(1 - G(t))}, \qquad t > 0.$$

We are primarily interested in thresholds that are neither too small or too large as far as HCT concerns; see (1.11). In light of this, we introduce the HCT functional

$$T_{\mathrm{HC}}(G) = \operatorname*{argmax}_{\{\bar{\Psi}^{-1}(1/2) < t < s_p^*\}} \mathrm{HC}(t, G),$$

where the term $\bar{\Psi}^{-1}(1/2)$ is chosen for convenience, and can be replaced by some other positive constants. Recall that $\tilde{Z} = \Omega Z$ and $\hat{Z} = \hat{\Omega} Z$ [e.g., (2.4) and (1.10)]. For any $t > 0$, let

$$\bar{F}_p(t) = \frac{1}{p} \sum_{j=1}^{p} 1\{|\hat{Z}(j)| \geq t\} \tag{2.15}$$

and

$$\begin{aligned} \widetilde{F}_p(t) &= \frac{1}{p} \sum_{j=1}^{p} 1\{|\tilde{Z}(j)| \geq t\}, \\ \widetilde{F}(t) &= \widetilde{F}(t, \varepsilon_p, \pi_p, \Omega) = E_{\varepsilon_p, \pi_p}[\widetilde{F}_p(t)]. \end{aligned} \tag{2.16}$$

Note that the only difference between $\widetilde{F}_p(t)$ and $\widetilde{F}(t)$ is the subscript p. Heuristically, for large p, we expect to have $\bar{F}_p(t) \approx \widetilde{F}_p(t) \approx \widetilde{F}(t)$. As a result, we expect that

$$T_{\mathrm{HC}}(\bar{F}_p) \approx T_{\mathrm{HC}}(\widetilde{F}_p) \approx T_{\mathrm{HC}}(\widetilde{F}),$$

where $T_{\mathrm{HC}}(\bar{F}_p)$ is the HCT where Ω is unknown and has to be estimated, $T_{\mathrm{HC}}(\widetilde{F}_p)$ is the HCT when Ω is known, and $T_{\mathrm{HC}}(\widetilde{F})$ is a nonstochastic counterpart of $T_{\mathrm{HC}}(\widetilde{F}_p)$. Note that in disguise, $T_{\mathrm{HC}}(\bar{F}_p)$ is the same as t_{HC}^*, the HCT defined in (1.12).

DEFINITION 2.3. We call $T_{\mathrm{HC}}(\widetilde{F})$ the ideal Higher Criticism Threshold (ideal HCT).

Similarly, the leading term of ideal HCT has a simple form that is easy to analyze. Fix $1 \leq j \leq p$. Let $D_j = \{k : 1 \leq k \leq p, \Omega(j,k) \neq 0\}$, and let

$$\begin{aligned} g_1(t) &= g_1(t; \Omega, \varepsilon_p, \tau_p) \\ &= \frac{1}{p} \sum_{j=1}^{p} P\bigl(|\tilde{Z}(j)| \geq t, \mu(k) \neq 0 \text{ for some } k \in D_j, k \neq j\bigr). \end{aligned}$$

The following is a counterpart of $\widetilde{W}_0(t)$ defined in (2.8) and can be well approximated by the latter:

$$W_0(t) = W_0(t, \varepsilon_p, \tau_p, \Omega) = \frac{\varepsilon_p \bar{\Psi}_{\tau_p}(t) + g_1(t)}{\sqrt{\bar{\Psi}(t) + \varepsilon_p \bar{\Psi}_{\tau_p}(t) + g_1(t)}}. \tag{2.17}$$

The following lemmas are proved in the supplementary material [23].

LEMMA 2.3. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $r > \rho_\theta^*(\beta)$ and $(1-\theta)/2 < \beta < (1-\theta)$. In the* $\mathrm{ARW}(\beta, r, \theta, \Omega)$ *model, as $p \to \infty$,*

$$\sup_{\{t > \bar{\Psi}^{-1}(1/2)\}} \sup_{\{\Omega \in \mathcal{M}_p^*(a, K_p)\}} \{|p^{-1/2}\mathrm{HC}(t, \tilde{F}) - W_0(t, \varepsilon_p, \tau_p, \Omega)|\} \le L_p p^{-\beta}.$$

LEMMA 2.4. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $r > \rho_\theta^*(\beta)$ and $(1-\theta)/2 < \beta < (1-\theta)$. In the* $\mathrm{ARW}(\beta, r, \theta, \Omega)$ *model, as $p \to \infty$, we have*

$$\begin{aligned}\sup_{\{t > 0\}} \sup_{\{\Omega \in \mathcal{M}_p^*(a, K_p)\}} &\bigl|W_0(t, \varepsilon_p, \tau_p, \Omega) - \widetilde{W}_0(t, \varepsilon_p, \tau_p)\bigr| \\ &\le L_p\Bigl[p^{-3\beta/2} + p^{-c_0(\beta, r, a)} \sup_{\{t>0\}} \widetilde{W}_0(t)\Bigr].\end{aligned}$$

If additionally $r \ge \beta$, then:

(a) $\sup_{\{0 \le t < \sqrt{2\beta \log(p)} - \Delta_1\}} W_0(t, \varepsilon_p, \tau_p, \Omega) \lesssim (\frac{1}{\sqrt{2}}) p^{-\beta/2}$,

(b) $\sup_{\{\tau_p \le t < \infty\}} W_0(t, \varepsilon_p, \tau_p, \Omega) \lesssim (\frac{1}{\sqrt{2}}) p^{-\beta/2}$,

(c) $\frac{3}{4} p^{-\beta/2} \lesssim \sup_{\{\sqrt{2\beta\log(p)} - \Delta_1 < t < \tau_p\}} W_0(t, \varepsilon_p, \tau_p, \Omega) \le L_p p^{-\beta/2}$,

where $\Delta_1 = d_0 \log\log(p)/\sqrt{\log(p)}$ is defined in Lemma 2.2.

Lemmas 2.3–2.4 say that, approximately, $\mathrm{HC}(t, \tilde{F}) \propto W_0(t)$, and that two functions $\widetilde{W}_0(t)$ and $W_0(t)$ are generally close.

2.4. *Relationship between two ideal thresholds and classification by the ideal classifier.* Together, Lemmas 2.1–2.4 consolidate the intimate relationship between the ideal threshold and the ideal HCT. To see the point, we discuss the cases $r < \beta$ and $r \ge \beta$ separately.

For the first case, write $T_{\mathrm{ideal}} = T_{\mathrm{ideal}}(\varepsilon_p, \tau_p, \Omega)$ and $\widetilde{\mathrm{Sep}}(t) = \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega)$ for short as before. The following theorem is proved in the supplementary material [23].

THEOREM 2.1. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $\rho_\theta^*(\beta) < r < \beta$ and $(1-\theta)/2 < \beta < (1-\theta)$. In the* $\mathrm{ARW}(\beta, r, \theta, \Omega)$ *model with $\Omega \in \mathcal{M}_p^*(a, K_p)$, as $p \to \infty$, there is a constant $c_1 = c_1(\beta, r, a) > 0$ such that $|T_{\mathrm{HC}}(\tilde{F}) - T_{\mathrm{ideal}}| \le L_p p^{-c_1(\beta, r, a)}$, and so $\widetilde{\mathrm{Sep}}(T_{\mathrm{HC}}(\tilde{F})) \sim \widetilde{\mathrm{Sep}}(T_{\mathrm{ideal}}) = L_p p^{(1-\theta)/2 - \delta(\beta, r)}$.*

Consider the second case. Lemma 2.4 says that $\sqrt{2\beta\log(p)} \lesssim T_{\mathrm{HC}}(\tilde{F}) \lesssim \sqrt{2r\log(p)}$. While it is hard to further elaborate how close two ideal thresholds are, in light of (2.14), HC classification with any t in this range is successful, so it is not critical to pin down the ideal HCT. The following theorem is proved in the supplementary material [23].

THEOREM 2.2. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $r \geq \beta$ and $(1-\theta)/2 < \beta < (1-\theta)$. In the* ARW$(\beta, r, \theta, a)$ *model where $\Omega \in \widetilde{\mathcal{M}}_p^*(a, b, K_p)$, as $p \to \infty$, we have that $2\tau_p K_p^{-1} p^{(1-\theta-\beta)/2} \lesssim \widetilde{\mathrm{Sep}}(T_{\mathrm{HC}}(\widetilde{F})) \leq \widetilde{\mathrm{Sep}}(T_{\mathrm{ideal}}(\varepsilon_p, \tau_p, \Omega)) = L_p p^{(1-\theta-\beta)/2}$.*

To conclude this section, we investigate the "ideal" classifier $L_t(X, \Omega)$, where Ω is known to us. Note that for each fixed t, the misclassification error of $L_t(X, \Omega)$ is $P(Y \cdot L_t(X, \Omega) < 0) = E_{\varepsilon_p, \pi_p} E[\bar{\Phi}(\frac{1}{2}\mathrm{Sep}(t, \tilde{Z}, \mu, \Omega)]$. The following theorem is proved in the supplementary material [23].

THEOREM 2.3. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $(1-\theta)/2 < \beta < (1-\theta)$ and $r > \rho_\theta^*(\beta)$. In the* ARW(β, r, θ, a) *model with $\Omega \in \widetilde{\mathcal{M}}_p^*(a, b, K_p)$, as $p \to \infty$,*

$$\min_t P\bigl(Y \cdot L_t(X, \Omega) < 0 | t\bigr) = \bar{\Phi}\bigl(\bigl(1 + o(1)\bigr) \cdot \tfrac{1}{2}\widetilde{\mathrm{Sep}}(T_{\mathrm{ideal}})\bigr).$$

When $r < \beta$, the condition $\Omega \in \widetilde{\mathcal{M}}_p^*(a, b, K_p)$ can be relaxed to that of $\Omega \in \mathcal{M}_p^*(a, K_p)$.

Combining Theorem 2.3 with Theorems 2.1–2.2,

$$\min_t P\bigl(Y \cdot L_t(X, \Omega) < 0 | t\bigr) = \bar{\Phi}\bigl(h(t) \cdot \widetilde{\mathrm{Sep}}(T_{\mathrm{HC}}(\widetilde{F}))\bigr),$$

where $h(t) = h(t; \beta, r, \theta, a, \Omega_p, p)$ satisfies $h(t) = 1/2 + o(1)$ when $r < \beta$ and $h(t) = L_p$ when $r \geq \beta$. Recall that in both cases, $\widetilde{\mathrm{Sep}}(T_{\mathrm{ideal}}) = L_p \widetilde{\mathrm{Sep}}(T_{\mathrm{HC}}(\widetilde{F})) = L_p p^{(1-\theta)/2 - \delta(\beta, r)}$, where the exponent $(1-\theta)/2 - \delta(\beta, r)$ is strictly positive by the assumption of $r > \rho_\theta^*(\beta)$. Therefore, if (β, r) fall in Region of Possibility and if we set t as either of the two ideal thresholds, then $L_t(X, \Omega)$ not only gives successful classification, but the classification error converges to 0 very fast.

3. Classification by HCT.

In the preceding section, we have been focusing on two ideal thresholds. In this section, we study the empirical quantities, and characterize the stochastic fluctuation of HCT and Sep defined in (2.2). We conclude the section by proving Theorems 1.2–1.3. The main results in this section are new, even in the idealized case where $\Omega = I_p$.

3.1. *Stochastic control on the HC functional.*

Recall that

$$\mathrm{HC}(t, \bar{F}_p) = \sqrt{p}\bigl[\bar{F}_p(t) - \bar{\Psi}(t)\bigr] / \sqrt{\bar{F}_p(t)\bigl(1 - \bar{F}_p(t)\bigr)}.$$

When $\bar{F}_p(t) = 0$, the above is not well defined, and we modify the definition slightly by replacing $\bar{F}_p(t)$ with $1/p$. The change does not affect the proof of the results. The stochastic fluctuation of HCT comes from that of $\bar{F}_p(t)$, which consists of two components: that of estimating Ω and that of the data. This is captured in the following triangle inequality [see (2.15)–(2.16)]:

$$\bigl|\bar{F}_p(t) - \widetilde{F}(t)\bigr| \leq \bigl|\widetilde{F}_p(t) - \widetilde{F}(t)\bigr| + \bigl|\bar{F}_p(t) - \widetilde{F}_p(t)\bigr|.$$

Consider $|\widetilde{F}_p(t) - \widetilde{F}(t)|$ first. The key is to study

$$\sqrt{p}(\widetilde{F}_p(t) - \widetilde{F}(t))/\sqrt{\widetilde{F}(t)(1 - \widetilde{F}(t))}.$$

When $\Omega = I_p$, this is the *standard uniform stochastic processes* [41] and much is known about its stochastic fluctuation. In the more general case where $\Omega \neq I_p$, it is usually hard to derive a tight bound on the tail probability of this process. Fortunately, when Ω is K_p-sparse, tight bounds are possible, and the key is graph theory on the chromatic number introduced in Lemma 1.1.

Recall that $s_p^* = \sqrt{2\log(p)}$ [e.g., (1.11)]. The following lemma is the direct result of Lemma 1.1 and the well-known Bennet's inequality [41], and is proved in the supplementary material [23].

LEMMA 3.1. *Fix* $(\beta, r, \theta, a) \in (0, 1)^4$ *and consider an* $\text{ARW}(\beta, r, \theta, \Omega)$ *model with* $\Omega \in \mathcal{M}_p^*(a, K_p)$. *As* $p \to \infty$, *there is a constant* $C > 0$ *such that with probability at least* $1 - o(p^{-1})$, *for all t satisfying* $\bar{\Psi}^{-1}(1/2) < t < s_p^*$,

$$\sqrt{p}|\widetilde{F}_p(t) - \widetilde{F}(t)|/\sqrt{\widetilde{F}(t)(1 - \widetilde{F}(t))} \le CK_p^3(\log(p))^{7/4}.$$

Next, consider $|\widetilde{F}_p(t) - \bar{F}_p(t)|$. Recall that $n_p = p^\theta$. By definition, if $\hat{\Omega}$ is an acceptable estimator of Ω, then there is a constant $C > 0$ such that with probability at least $1 - o(p^{-1})$,

$$\max_{\{1 \le i, j \le p\}} \{|\hat{\Omega}(i, j) - \Omega(i, j)|\} \le CK_p^2\sqrt{2\log(p)} \cdot p^{-\theta/2}. \tag{3.1}$$

As a result, we have the following lemma, whose proof is straightforward and thus omitted. Recall that $\hat{Z} = \hat{\Omega}Z$ and $\tilde{Z} = \Omega Z$ [e.g., (1.10) and (2.4)].

LEMMA 3.2. *For any acceptable estimator* $\hat{\Omega}$, $\max_{\{1 \le j \le p\}}\{|\hat{Z}(j) - \tilde{Z}(j)|\} \le CK_p^3 \log(p)p^{-\theta/2}$ *with probability at least* $1 - o(1/p)$.

Write for short $\eta_p = CK_p^3 \log(p)p^{-\theta/2}$. By Lemma 3.2, with probability at least $1 - o(1/p)$, for all $1 \le j \le p$, $|1\{|\hat{Z}(j)| \ge t\} - 1\{|\tilde{Z}(j)| \ge t\}| \le 1\{t - \eta_p \le |\tilde{Z}(j)| \le t + \eta_p\}$. As a result,

$$|\widetilde{F}_p(t) - \bar{F}_p(t)| \le \widetilde{F}_p(t - \eta_p) - \widetilde{F}_p(t + \eta_p),$$

where we note that heuristically,

$$\widetilde{F}_p(t - \eta_p) - \widetilde{F}_p(t + \eta_p) \approx \widetilde{F}(t - \eta_p) - \widetilde{F}(t + \eta_p) \approx 2\eta_p|\widetilde{F}'(t)|.$$

Combining these, with probability at least $1 - o(1/p)$, for any $t > \bar{\Psi}^{-1}(\frac{1}{2})$,

$$\frac{\sqrt{p}|\widetilde{F}_p(t) - \bar{F}_p(t)|}{\sqrt{\widetilde{F}(t)(1 - \widetilde{F}(t))}} \le 2\sqrt{2p}\eta_p|\widetilde{F}'(t)|/\sqrt{\widetilde{F}(t)} = 2\sqrt{2}p^{(1-\theta)/2}|\widetilde{F}'(t)|/\sqrt{\widetilde{F}(t)}.$$

Recall $s_p^* = \sqrt{2\log(p)}$. The above heuristic is captured in the following lemma, which is proved in the supplementary material [23].

LEMMA 3.3. *Fix $(\beta, r, \theta, a) \in (0,1)^4$. In the* ARW$(\beta, r, \theta, \Omega)$ *model with* $\Omega \in \mathcal{M}_p^*(a, K_p)$, *there exists a constant* $C > 0$ *such that with probability at least* $1 - o(1/p)$, *for all t such that* $\bar{\Psi}^{-1}(\frac{1}{2}) < t < s_p^*$,

$$\sqrt{p}|\bar{F}_p(t) - \tilde{F}_p(t)| \cdot [\tilde{F}(t)(1 - \tilde{F}(t))]^{-1/2} \leq L_p \max\{(p^{(1-\theta)}\tilde{F}(t))^{1/2}, 1\}.$$

Combining Lemmas 3.1 and 3.3, we have the following theorem, which is proved in the supplementary material [23].

THEOREM 3.1. *Fix $(\beta, r, \theta, a) \in (0,1)^4$. In the* ARW$(\beta, r, \theta, \Omega)$ *model with* $\Omega \in \mathcal{M}_p^*(a, K_p)$, *as $p \to \infty$, with probability at least* $1 - o(p^{-1})$,

$$|\mathrm{HC}(t, \bar{F}_p) - \mathrm{HC}(t, \tilde{F})| \leq L_p[(p^{1-\theta}\tilde{F}(t))^{1/2} + 1] \qquad \forall \bar{\Psi}^{-1}(\tfrac{1}{2}) < t < s_p^*.$$

By Theorem 3.1, in order for $|T_{\mathrm{HC}}(\bar{F}_p) - T_{\mathrm{HC}}(\tilde{F})|$ to be small, we must have that for all t in the vicinity of $T_{\mathrm{HC}}(\tilde{F})$,

$$|\mathrm{HC}(t, \bar{F}_p) - \mathrm{HC}(t, \tilde{F})| \ll \mathrm{HC}(t, \tilde{F}).$$

When $\theta > 1/2$, this holds for all (β, r) in Region of Possibility since it can be checked that $L_p[(p^{1-\theta}\tilde{F}(t))^{1/2} + 1] \ll \mathrm{HC}(t, \tilde{F})$. When $\theta \leq 1/2$, this might not hold for all (β, r) in this region, as the estimation error of $\hat{\Omega}$ is simply too large. This explains why we need to restrict HCT to be no less than $\tilde{s}_{p,n}^*$ as in (1.11). This also explains why we need conditions (a)–(b) in Theorem 1.3, but we do not need such conditions in Theorem 1.2 and Corollary 1.1.

In the ARW$(\beta, r, \theta, \Omega)$ model, $n_p = p^\theta$. Therefore,

$$\tilde{s}_{p,n}^* = s_p(\theta) \qquad \text{if we let } s_p(\theta) = \sqrt{2\max\{(1-2\theta), 0\}\log(p)};$$

see (1.11). Accordingly, the HCT defined in (1.12) can be rewritten as

$$t_p^{\mathrm{HC}} = \begin{cases} T_{\mathrm{HC}}(\bar{F}_p), & \text{if } s_p(\theta) \leq T_{\mathrm{HC}}(\bar{F}_p) \leq s_p^*, \\ s_p(\theta), & \text{if } T_{\mathrm{HC}}(\bar{F}_p) < s_p(\theta), \\ s_p^*, & \text{if } T_{\mathrm{HC}}(\bar{F}_p) > s_p^*. \end{cases}$$

It is worthy to note here that the ideal threshold always falls below s_p^*, which is defined as $\sqrt{2\log(p)}$; see Section 2.2 and especially (2.9). It is also worthy to note that when $\theta < 1/2$ and when Ω is unknown, the estimation error of Ω may have a major effect over the classification error, especially when the threshold is small. To alleviate such an effect, one possible approach is to set a number $s_p(\beta, r, \theta)$ (say), and never allow the threshold to be smaller than $s_p(\beta, r, \theta)$. Since

(β, r) are unknown to us [but $\theta \equiv \log(n_p)/\log(p)$ is known to us], so from a practical perspective, we must select $s_p(\beta, r, \theta)$ in a way so that it does not depend on (β, r). Our calculations show that $s_p(\theta) = \sqrt{2\max\{(1-2\theta), 0\}\log(p)}$ is one of such choices.

The main result in this section is as follows, which is proved in the supplementary material [23].

THEOREM 3.2. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $(1-\theta)/2 < \beta < 1-\theta$ and $r > \rho_\theta^*(\beta)$. In the* ARW$(\beta, r, \theta, \Omega)$ *model with $\Omega \in \mathcal{M}_p^*(a, K_p)$,*

(1) *If $\theta > \frac{1}{2}$, then as $p \to \infty$, there are positive constants $c_2 = c_2(\beta, r, a, \theta)$ and $d_0 = d_0(\beta, r, a, \theta)$ such that with probability at least $1 - o(1/p)$, $|t_p^{\mathrm{HC}} - T_{\mathrm{ideal}}(\varepsilon_p, \tau_p, \Omega)| \le L_p p^{-c_2}$ when $r < \beta$, and $t_p^{\mathrm{HC}} \in [\sqrt{2\beta\log p} - \Delta_1, \tau_p)$ when $r \ge \beta$, where $\Delta_1 = d_0 \log(\log(p))/\sqrt{\log(p)}$.*

(2) *If $0 < \theta \le \frac{1}{2}$ and (β, r, θ) satisfy the conditions in Theorem 1.3, then with probability at least $1 - o(1/p)$, $|t_p^{\mathrm{HC}} - T_{\mathrm{ideal}}(\varepsilon_p, \tau_p, \Omega)| \le L_p p^{-c_3}$ for some constant $c_3 = c_3(\beta, r, a) > 0$ when $r < \beta$, and $t_p^{\mathrm{HC}} \in [\sqrt{2\beta\log p} - \Delta_1, \tau_p)$ for $\Delta_1 = d_1 \log(\log(p))/\sqrt{\log p}$ when $r \ge \beta$, where $d_1 = d_1(\beta, r, a) > 0$ is a constant.*

3.2. *Stochastic fluctuation of Fisher's separation.* Similarly, the stochastic fluctuation of $\mathrm{Sep}(t, \hat{Z}, \mu, \hat{\Omega})$ contains two parts: that from $\tilde{Z} = \Omega Z$, and that from the estimation $\hat{\Omega}$. In detail,

$$\left|\mathrm{Sep}(t, \hat{Z}, \mu, \hat{\Omega}) - \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega)\right| \le 2 \cdot (I + II),$$

where $I = \frac{1}{2}|\mathrm{Sep}(t, \tilde{Z}, \mu, \Omega) - \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega)|$ and $II = \frac{1}{2}|\mathrm{Sep}(t, \hat{Z}, \mu, \hat{\Omega}) - \mathrm{Sep}(t, \tilde{Z}, \mu, \Omega)|$.

Consider I first. Recall that

$$\mathrm{Sep}(t, \tilde{Z}, \mu, \Omega) = 2M_p(t, \tilde{Z}, \mu, \Omega)/\sqrt{V_p(t, \tilde{Z}, \Omega)}.$$

Heuristically, $M_p(t, \tilde{Z}, \mu, \Omega) = m_p(t, \varepsilon_p, \tau_p, \Omega) + O_p(\sqrt{m_p(t, \varepsilon_p, \tau_p, \Omega)})$ and $V_p(t, \tilde{Z}, \mu, \Omega) = v_p(t, \varepsilon_p, \tau_p, \Omega) + O_p(\sqrt{v_p(t, \varepsilon_p, \tau_p, \Omega)})$; see (2.6). Combining these with the definitions, we expect that

$$\begin{aligned} &\mathrm{Sep}(t, \tilde{Z}, \mu, \Omega) \\ &\quad = \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega) \\ &\qquad \times \left[1 + O_p\left(\frac{1}{\sqrt{m_p(t, \varepsilon_p, \tau_p, \Omega)}} + \frac{1}{\sqrt{v_p(t, \varepsilon_p, \tau_p, \Omega)}}\right)\right], \end{aligned} \tag{3.2}$$

where in the square brackets, the second term is much smaller than 1. This is elaborated in the following lemma which is proved in the supplementary material [23]. In detail, let

$$q(t) = q(t; \beta, r, \theta, \Omega_p, p) = \begin{cases} p^{(1-\theta)/2 - \max\{4\beta - 2r, 3\beta + r\}/4}, & r < \beta, \\ 0, & r \geq \beta. \end{cases}$$

LEMMA 3.4. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $r \geq \rho_\theta^*(\beta)$ and $(1-\theta)/2 < \beta < (1-\theta)$. In the* ARW$(\beta, r, \theta, \Omega)$ *model with $\Omega \in \widetilde{\mathcal{M}}_p^*(a, b, K_p)$, as $p \to \infty$, with probability at least $1 - o(1/p)$,*

$$\sup_{\{t>0\}} |\mathrm{Sep}(t, \tilde{Z}, \mu, \Omega) - \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega)| \leq L_p[q(t) + p^{-\theta/2}].$$

When $r < \beta$, the condition on Ω can be relaxed to that of $\Omega \in \mathcal{M}_p^*(a, K_p)$.

Next, we consider *II*. The following lemma, which is proved in the supplementary material [23], characterizes the order of *II*.

LEMMA 3.5. *Under the same conditions as in Lemma 3.4, as $p \to \infty$, with probability at least $1 - o(1/p)$, for all t such that $s_p(\theta) < t < s_p^*$, $|\,\mathrm{Sep}(t, \hat{Z}, \mu, \hat{\Omega}) - \mathrm{Sep}(t, \tilde{Z}, \mu, \Omega)| \leq L_p[p^{-\theta}(p\widetilde{F}(t))^{1/2} + q(t) + p^{-\theta/2}]$. When $r < \beta$, the condition on Ω can be relaxed to that of $\Omega \in \mathcal{M}_p^*(a, K_p)$.*

Combining Lemmas 3.4–3.5 gives the following theorem, the proof of which is omitted (note that Theorem 3.3 is parallel to Theorem 3.1).

THEOREM 3.3. *Under the same conditions as in Lemma 3.4, as $p \to \infty$, with probability at least $1 - o(p^{-1})$, for all t such that $s_p(\theta) < t < s_p^*$,*

$$|\mathrm{Sep}(t, \hat{Z}, \mu, \hat{\Omega}) - \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega)| \leq L_p[p^{-\theta}(p\widetilde{F}(t))^{1/2} + q(t) + p^{-\theta/2}].$$

When $r < \beta$, the condition on Ω can be relaxed to that of $\Omega \in \mathcal{M}_p^*(a, K_p)$.

3.3. *Proof of Theorems 1.2–1.3.* We are now ready to prove Theorems 1.2–1.3, where Ω is assumed as known and unknown, respectively. The proofs are similar, so we only show Theorem 1.3. Consider $L_{\mathrm{HC}}(X, \hat{\Omega})$, where $\hat{\Omega}$ is an acceptable estimator. The misclassification error is

$$P(Y \cdot L_{\mathrm{HC}}(X, \hat{\Omega}) < 0) = E_{\varepsilon_p, \tau_p} E[\bar{\Phi}(\tfrac{1}{2}\,\mathrm{Sep}(t_p^{\mathrm{HC}}, \hat{Z}, \mu, \hat{\Omega}))]. \tag{3.3}$$

We now prove for the case of $r < \beta$ and $r \geq \beta$ separately.

In the first case, we note that $L_p[p^{-\theta}(p\widetilde{F}(t))^{1/2} + p^{-\theta/2}] \leq L_p p^{\min\{0, 1/2 - \theta\}}$ for $s_p(\theta) < t < s_p^*$. Write $T_{\mathrm{ideal}} = T_{\mathrm{ideal}}(\varepsilon_p, \tau_p, \Omega)$ and $\widetilde{\mathrm{Sep}}(t) = \widetilde{\mathrm{Sep}}(t, \varepsilon_p, \tau_p, \Omega)$ for short as before. By Theorem 3.3, with probability $1 - o(1/p)$,

$$\begin{aligned} &|\mathrm{Sep}(t_p^{\mathrm{HC}}, \hat{Z}, \mu, \hat{\Omega}) - \widetilde{\mathrm{Sep}}(t_p^{\mathrm{HC}})| \\ &\qquad \leq L_p[p^{\min\{0, 1/2 - \theta\}} + p^{(1-\theta)/2 - \max\{\beta - r/2, (3\beta + r)/4\}}]. \end{aligned} \tag{3.4}$$

At the same time, by Theorem 3.2, with probability $1 - o(1/p)$, $|t_p^{\mathrm{HC}} - T_{\mathrm{ideal}}|$ is algebraically small. Note that $\widetilde{\mathrm{Sep}}(t)$ is a nonstochastic function, and that in the vicinity of T_{ideal}, the second derivative of $\widetilde{\mathrm{Sep}}$ at t has the same magnitude as that of $\widetilde{\mathrm{Sep}}(t)$, up to a multi-$\log(p)$ term (the first derivative is 0 at $t = T_{\mathrm{ideal}}$). By Taylor's expansion and Lemma 2.1,

$$\widetilde{\mathrm{Sep}}\big(t_p^{\mathrm{HC}}\big) = \big(1 + o(1)\big)\widetilde{\mathrm{Sep}}(T_{\mathrm{ideal}}) = L_p p^{(1-\theta)/2-\delta(\beta,r)}, \tag{3.5}$$

where $\delta(\beta, r)$ is as in (2.10). By definitions, $\max\{4\beta - 2r, 3\beta + r\}/4 > \delta(\beta, r)$. Inserting (3.4)–(3.5) into (3.3) gives

$$P\big(Y \cdot L_{\mathrm{HC}}(X, \hat{\Omega}) < 0\big) = \big(1 + o(1/p)\big)\bar{\Phi}\big(L_p p^{(1-\theta)/2-\delta(\beta,r)}\big) + o(1/p), \tag{3.6}$$

and the claim follows since $(1 - \theta)/2 - \delta(\beta, r) > 0$.

In the second case, $\sqrt{2\beta \log p} \lesssim t_p^{\mathrm{HC}} \lesssim \sqrt{2r \log p}$ with probability at least $1 - o(1/p)$. Combining this with Theorem 3.3, with probability at least $1 - o(1/p)$,

$$\big|\mathrm{Sep}\big(t_p^{\mathrm{HC}}, \hat{Z}, \mu, \hat{\Omega}\big) - \widetilde{\mathrm{Sep}}\big(t_p^{\mathrm{HC}}\big)\big| \le L_p p^{\min\{0,1/2-\theta\}}. \tag{3.7}$$

At the same time, by similar argument as that of the proof of Theorem 2.2,

$$2\tau_p K_p^{-1} p^{(1-\theta-\beta)/2} \lesssim \widetilde{\mathrm{Sep}}\big(t_p^{\mathrm{HC}}\big) \le \widetilde{\mathrm{Sep}}(T_{\mathrm{ideal}}) = L_p p^{(1-\theta-\beta)/2}.$$

Combining this with (3.3) and (3.7) gives

$$P\big(Y \cdot L_{\mathrm{HC}}(X, \hat{\Omega}) < 0\big) = \big(1 + o(1/p)\big)\bar{\Phi}\big(\tfrac{1}{2} L_p p^{(1-\theta)/2-\delta(\beta,r)}\big) + o(1/p), \tag{3.8}$$

and the claim follows since $\frac{1-\theta}{2} - \delta(\beta, r) > 0$. This proves Theorem 1.3.

We conclude this section by a remark on the convergence rate. At the end of Section 2, we show that the "ideal" classifier $L_t(X, \Omega)$ has very fast convergence rate with t being either the ideal threshold or the ideal HCT. In comparison, the convergence rate of $L_{\mathrm{HC}}(X, \hat{\Omega})$ is unfortunately much slower (but is still algebraically fast). To explain this, we note that the rate of convergence of t_p^{HC} to $T_{\mathrm{HC}}(\widetilde{F})$ and the rate of convergence of $\hat{\Omega}$ to Ω are both algebraically fast; if these convergence rates can be improved, then the misclassification error rate of $L_{\mathrm{HC}}(X, \hat{\Omega})$ can be improved as well.

4. Simulations.

We have conducted a small-scale numerical study. The idea is to select a few sets of representative parameters for experiments, and compare the performance of HCT classifier (HCT) with three other methods: ordinary HCT (oHCT), pseudo HCT (pHCT), and CVT. All these methods are very similar to HCT, except that (a) in pHCT, we assume Ω is known to us, (b) in CVT, we set the threshold of IT by a 5-fold cross validation, and (c) in oHCT, we pretend Σ is diagonal, and estimate Ω accordingly. Note that CVT reduces to PAM [43] if we do not utilize the correlation structure; see more discussion in [16].

4.1. *Estimating* Ω. For some of the procedures, we need to estimate Ω. We use Bickel and Levina's Thresholding (BLT) procedure [4]. Alternatively, one could use the glasso [25] or the CLIME [9]. But since the main goal is to investigate the performance of HCT, we do not include glasso and CLIME in the study: if HCT performs well with Ω estimated by BLT, we expect it to perform even better if Ω is estimated more accurately.

At the same time, each of these methods can be improved numerically with an additional *re-fitting* stage. Take the BLT for example. For the training data $\{(X_i, Y_i)\}_{i=1}^n$, let $\bar{X} = \frac{1}{n}\sum_{i=1}^n Y_i X_i$, and let $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (Y_i X_i - \bar{X})'(Y_i X_i - \bar{X})$ be the empirical covariance matrix. BLT starts by obtaining an estimate of Σ using thresholding:

$$\Sigma^*(i,j) = \hat{\Sigma}(i,j)1\{|\hat{\Sigma}(i,j)| \geq \eta\}, \qquad 1 \leq i, j \leq p, \tag{4.1}$$

and then estimate Ω by $\hat{\Omega}^{**} = (\Sigma^*)^{-1}$. Here, $\eta > 0$ is a tuning parameter.

We propose the following refitting stage to improve the estimator. Fixing a tuning parameter $\zeta > 0$, we further improve $\hat{\Omega}^{**}$ via coordinate-wise thresholding and call the resultant estimator $\hat{\Omega}^*$:

$$\hat{\Omega}^*(i,j) = \hat{\Omega}^{**}(i,j)1\{|\hat{\Omega}^{**}(i,j)| \geq \zeta\}. \tag{4.2}$$

For each $1 \leq i \leq p$, let $S_i = \{1 \leq j \leq p : \hat{\Omega}^*(i,j) \neq 0\}$, and let A_i be the submatrix of $\hat{\Sigma}$ formed by restricting the rows/columns of $\hat{\Sigma}$ to S_i. Denote the final estimate of Ω by $\hat{\Omega} = [\omega_1, \omega_2, \ldots, \omega_p]$. We define ω_i as follows. Write $S_i = \{j_1, j_2, \ldots, j_k\}$, where $k = |S_i|$. Let e_i be the $p \times 1$ vector such that $e_i(j) = 1\{i = j\}$, $1 \leq j \leq p$, and let ξ_i be the $k \times 1$ vector formed by restricting the rows of e_i to S_i. Define $\eta_i = A_i^{-1}\xi_i$. We let $\omega_i(j_\ell) = \eta_i(\ell)$, $1 \leq \ell \leq k$, and let $\omega_i(j) = 0$ if $j \notin S_i$.

The resultant estimation of the refitting procedure is a symmetric matrix, which is also positive definite, provided that K_p is sufficiently small (say, $\sqrt{\log(p)}K_p \ll \sqrt{n}$) and that the smallest eigenvalue of Ω is bounded from below by a constant $C > 0$; recall that K_p is the maximum of the number of nonzeros in each row of Ω.

4.2. *Numerical experiments.* Fix $(p, n, \varepsilon_p, H_p, \Omega)$ and an integer m, each simulation experiment contains the following main steps.

1. Generate a $p \times 1$ vector μ according to $(\sqrt{n}\mu(j)) \overset{\text{i.i.d.}}{\sim} (1 - \varepsilon_p)\nu_0 + \varepsilon_p H_p$.
2. Generate training data (X_i, Y_i), $1 \leq i \leq n$, by letting $Y_i = 1$ for $i \leq n/2$ and $Y_i = -1$ for $i > n/2$, and $X_i \sim N(Y_i \cdot \mu, \Omega^{-1})$.
3. Generate m test vectors, each of which has the form of $X \sim N(Y \cdot \mu, \Omega^{-1})$, where $Y = \pm 1$ with equal probabilities.
4. Use the training data to build all four classifiers (HCT, oHCT, pHCT and CVT), apply them to the test set, and then record the test errors.

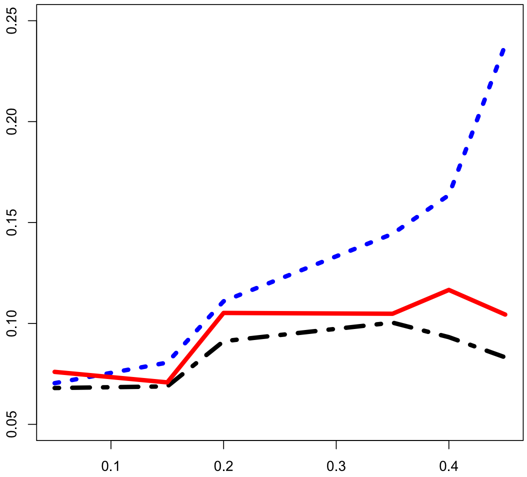

FIG. 1. *Comparison of classification errors by HCT* (*solid*), *oHCT* (*dashed*) *and pHCT* (*dash-dotted*). *The x-axis is a, and the y-axis is the classification error* (*Experiment* 1a).

When we need to estimate Ω, we use BLT with the aforementioned refitting stage. The study contains three different experiments, which we now discuss separately.

Experiment 1. In this experiment, we compare HCT with oHCT and pHCT. The experiment contains three sub-experiments 1a, 1b and 1c.

In Experiment 1a, we fix $(p, n, \varepsilon_p, \tau_p, m) = (3000, 2000, 0.1, 4, 500)$, and let H_p be the point mass at τ_p. Also, we choose Ω to be the tridiagonal matrix

$$\Omega(i, j) = 1\{i = j\} + a \cdot 1\big\{|i - j| = 1\big\}, \qquad 1 \le i, j \le p, \tag{4.3}$$

where a takes values from $\{0.05, 0.15, 0.2, 0.35, 0.4, 0.45\}$. The results are reported in Figure 1. The tuning parameter η in (4.1), which varies with the values of a, n and p, is calculated from trials of comparing $(\Sigma^*)^{-1}$ with the true Ω. The tuning parameter ζ in (4.2), which also varies with the values of a, n and p, is chosen so that there are only k nonzero coordinates in each row of $\hat{\Omega}^*$ after thresholding of $\hat{\Omega}^{**}$. We let $k = 2, 3$ if Ω is tridiagonal and $k = 4, 5$ if Ω is five-diagonal (see experiments below). In this experiment, η is set accordingly from $\{0.1, 0.1, 0.15, 0.15, 0.2, 0.25\}$ and ζ is from $\{0.05, 0.1, 0.1, 0.2, 0.25, 0.3\}$. The results suggest that HCT outperforms oHCT, but is slightly inferior to pHCT since we have to pay a price for estimating Ω. As a increases, the correlation structure becomes increasingly influential, so the advantage of HCT over oHCT becomes increasingly prominent (but differences between HCT and pHCT remain almost the same).

In Experiment 1b, for various $(p, n, \varepsilon_p, \tau_p)$, we choose $m = 500$ and let Ω be either of the following tridiagonal matrix or five-diagonal matrix. In the first case, Ω is a $p \times p$ tridiagonal matrix with 1 on the diagonal and a on the off-diagonal. In the second case, Ω is a $p \times p$ five-diagonal matrix with 1 on the diagonal,

TABLE 1

Classification errors by HCT, oHCT and pHCT. Ω is tridiagonal (left two columns) or five-diagonal matrix (right column). Rows 1–2: *Experiment* 1b. *Row* 3: *Experiment* 1c

	$\boldsymbol{n=1000, p=2000}$ $\boldsymbol{a=0.05, \varepsilon_p=0.1, \tau_p=4}$	$\boldsymbol{n=2000, p=3000}$ $\boldsymbol{a=0.45, \varepsilon_p=0.2, \tau_p=3}$	$\boldsymbol{n=2000, p=3000}$ $\boldsymbol{a_1=0.45, a_2=0.2}$ $\boldsymbol{\varepsilon_p=0.1, \tau_p=4}$
oHCT	0.0508	0.2818	0.1492
pHCT	0.0384	0.0698	0.1015
HCT	0.0523	0.0742	0.1053

	$\boldsymbol{n=500, p=1000}$ $\boldsymbol{a=0.05, \varepsilon_p=0.1, \tau_p=4}$	$\boldsymbol{n=2000, p=3000}$ $\boldsymbol{a=0.45, \varepsilon_p=0.05, \tau_p=5}$	$\boldsymbol{n=2000, p=3000}$ $\boldsymbol{a_1=0.35, a_2=0.2,}$ $\boldsymbol{\varepsilon_p=0.05, \tau_p=4}$
oHCT	0.0560	0.2629	0.2183
pHCT	0.0571	0.1398	0.1893
HCT	0.0572	0.1438	0.1959

	$\boldsymbol{n=1000, p=2000}$ $\boldsymbol{H_p=U(3.5,4.5)}$ $\boldsymbol{a=0.05, \varepsilon_p=0.1}$	$\boldsymbol{n=2000, p=3000}$ $\boldsymbol{H_p=U(2.5,3.5)}$ $\boldsymbol{a=0.45, \varepsilon_p=0.2}$	$\boldsymbol{n=2000, p=3000}$ $\boldsymbol{H_p=U(3.5,4.5)}$ $\boldsymbol{a_1=0.45, a_2=0.2, \varepsilon_p=0.1}$
oHCT	0.0444	0.2672	0.1648
pHCT	0.0522	0.0733	0.1159
HCT	0.0508	0.0843	0.0977

a_1 on the first off-diagonal, and a_2 on the second off-diagonal. Experiment 1c uses a very similar setting, except that we take H_p as the uniform distribution over $[\tau_p - 0.5, \tau_p + 0.5]$. We select ζ and η similarly as in Experiment 1a. The results based on 25 repetitions for Experiment 1b–1c are reported in Table 1, which suggest that HCT outperforms oHCT and that pHCT slightly outperforms HCT.

Experiment 2. In this experiment, we compare the pHCT with the CVT assuming Ω is known (the case Ω is unknown is discussed in Experiment 3). Experiment 2 contains two sub-experiments, 2a and 2b.

In Experiment 2a, we consider 6 different combinations of $(p, n, \varepsilon_p, \tau_p)$ with $m = 500$, and let Ω be the tridiagonal matrix as in (4.3) with $a = 0.2$. Averages of the selected thresholds and classification errors across different replications are reported in Table 2. The results over 25 repetitions suggest that the threshold choices by HC and cross validations are considerably different, with the former being more accurate and more stable. Note that HCT is also computationally much more efficient than the CVT.

TABLE 2
Comparison of thresholds (*Column* 2, 4, 6) *and classification errors* (*Column* 3, 5, 7) *by pHCT and CVT.* $(p, \tau_p) = (3000, 1.8)$, *and* $\varepsilon_p = 0.1$ (*top*) *and* 0.05 (*bottom*). *Left to right*: $n = 100, 50, 20$ (*Experiment* 2a)

	Threshold	**Error**	**Threshold**	**Error**	**Threshold**	**Error**
pHCT	1.9	0.05	2.16	0.002	1.99	0
CVT	2.5	0.08	1	0.018	1	0
pHCT	2.39	0.18	2.06	0.10	2.13	0.02
CVT	1.9	0.224	2.00	0.14	1.1	0.09

In Experiment 2b, we set $(p, \varepsilon_p, m) = (3000, 0.05, 500)$, $n \in \{20, 40\}$, and let Ω be the same as in Experiment 2a. We let τ_p range from 1 to 2.5 with an increment of 0.1. The classification errors over 25 repetitions by pHCT and CVT are in Figure 2, where a conclusion similar to that in Experiment 2a can be drawn.

Experiment 3. We compare the performance of HCT with CVT for the case where Ω is unknown and needs to be estimated. Note that for small n (say, less than 500) we might not have reasonable accuracy on estimating Ω using BLT. For small p, say 100–300, the CVT is computationally very slow and it is very likely that the refitting procedure for BLT would not have decent performance. We take $(p, n, \varepsilon_p) = (5000, 500, 0.1)$ and let Ω be the block diagonal matrix consisting 10 diagonal blocks, each is a big five-diagonal matrix $C = C_{500,500}(a_1, a_2)$, where $C(i, j) = 1\{i = j\} + a_1 \cdot 1\{|i - j| = 1\} + a_2 \cdot 1\{|i - j| = 2\}$, $1 \le i, j \le 500$, and $a_1 = 0.45$, $a_2 = 0.1$. We let τ_p range from 1 to 3 with an increment of 0.2. The tuning parameters ζ and η are set in the same way as in Experiment 1. The results

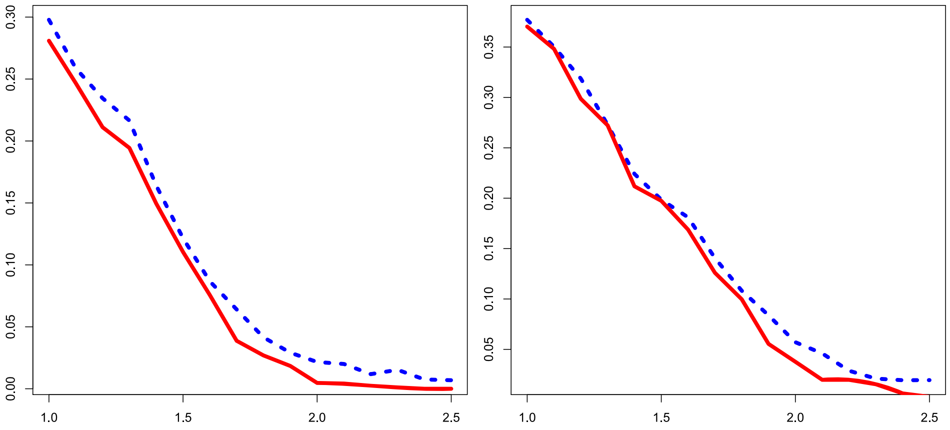

FIG. 2. *Classification errors of pHCT* (*solid*) *and CVT* (*dashed*) *for* $n = 20$ (*left*) *and* 40 (*right*) *and various* τ_p (*x-axis*) (*Experiment* 2b).

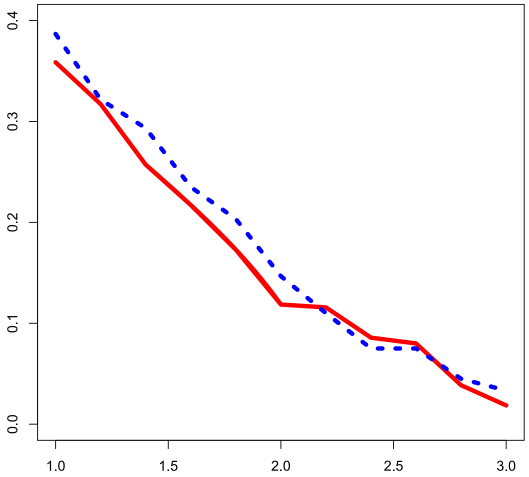

FIG. 3. *Classification errors by HCT (solid) and CVT (dashed) for various τ_p (x-axis) (Experiment 3).*

are reported in Figure 3. Due to high computational cost, we only conduct $m = 6$ repetitions, so the results are a bit noisy. Still, it is seen that HCT outperforms CVT.

In summary, for a reasonably large sample size n, HCT outperforms oHCT and is only slightly inferior to pHCT. The reason we need a relatively large n is mainly due to that we need to estimate Ω. The relative performance of pHCT, HCT, and oHCT is intuitive, since pHCT utilizes the true correlation structure among the features, HCT estimates the correlation structure, while oHCT ignores it. The comparisons of pHCT with CVT in Experiments 2a–2b suggest that if Ω is known, then HCT dominates CVT. Experiment 3 shows that when p is several times larger than n (e.g., 10 times larger), HCT has smaller classification errors than CVT does, and the precision matrix Ω can bez estimated reasonably well.

For larger p, the advantages of the HCT are even more prominent than those considered here. We skip the comparisons for larger p due to high computational cost, which mainly comes from the BLT procedure (we must run the algorithm many times to select a good tuning parameter η). In the future, if we could find a more efficient method for estimating Ω, then HCT will be both more effective and more convenient to use for large p.

SUPPLEMENTARY MATERIAL

Supplementary material for "Optimal classification in sparse Gaussian graphic model" (DOI: 10.1214/13-AOS1163SUPP; .pdf). We include all technical proofs omitted from the main text.

REFERENCES

[1] ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

[2] ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556. MR2906877

[3] BICKEL, P. J. and LEVINA, E. (2004). Some theory of Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010. MR2108040

[4] BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

[5] BOLLOBÁS, B. (1998). *Modern Graph Theory. Graduate Texts in Mathematics* **184**. Springer, New York. MR1633290

[6] BREIMAN, L. (2001). Random forests. *Mach Learning* **24** 5–32.

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

[8] CAI, T. and LIU, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.* **106** 1566–1577. MR2896857

[9] CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

[10] CAI, T. T., JIN, J. and LOW, M. G. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* **35** 2421–2449. MR2382653

[11] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[12] CAYON, L., JIN, J. and TREASTER, A. (2005). Higher criticism statistic: Detecting and identifying non-Gaussianity in the *WMAP* first-year data. *Mon. Not. R. Astron. Soc.* **362** 826–832.

[13] DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175.

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

[15] DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195

[16] DONOHO, D. and JIN, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **105** 14790–14795.

[17] DONOHO, D. and JIN, J. (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4449–4470. MR2546396

[18] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields* **99** 277–303. MR1278886

[19] EFRON, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *J. Amer. Statist. Assoc.* **104** 1015–1028. MR2562003

[20] FAN, J. and FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36** 2605–2637. MR2485009

[21] FAN, J., FENG, Y. and TONG, X. (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74** 745–771. MR2965958

[22] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[23] FAN, Y., JIN, J. and YAO, Z. (2013). Supplement to “Optimal classification in sparse Gaussian graphic model.” DOI:10.1214/13-AOS1163SUPP.

[24] FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** 179–188.

[25] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

[26] GENOVESE, C. R., JIN, J., WASSERMAN, L. and YAO, Z. (2012). A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.* **13** 2107–2143. MR2956354

[27] HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. MR2662357

[28] HALL, P., PITTELKOW, Y. and GHOSH, M. (2008). Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 159–173. MR2412636

[29] HE, S. and WU, Z. (2012). Gene-based higher criticism methods for large-scale exonic SNP data. *BMC Proceedings* **5** S65.

[30] INGSTER, Y. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distributions. *Math. Methods Statist.* **6** 47–69. MR1456646

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

[32] INGSTER, Y. I., POUET, C. and TSYBAKOV, A. B. (2009). Classification of sparse high-dimensional vectors. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4427–4448. MR2546395

[33] JAGER, L. and WELLNER, J. A. (2007). Goodness-of-fit tests via phi-divergences. *Ann. Statist.* **35** 2018–2053. MR2363962

[34] JI, P. and JIN, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist.* **40** 73–103. MR3013180

[35] JIN, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **106** 8859–8864. MR2520682

[36] JIN, J. and WANG, W. (2012). Optimal spectral clustering by higher criticism thresholding. Unpublished manuscript.

[37] LI, C. and LI, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24** 1175–1182.

[38] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

[39] SABATTI, C., SERVICE, S. and HARTIKAINEN, A. T. (2008). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41** 35–46.

[40] SHAO, J., WANG, Y., DENG, X. and WANG, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.* **39** 1241–1265. MR2816353

[41] SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York. MR0838963

[42] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

[43] 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. USA* **99** 6567–6572.

[44] TUKEY, J. W. (1976). T13 N: The higher criticism. Course notes, Stat. 411, Princeton Univ.

[45] ZHONG, P., CHEN, S. and XU, M. (2012). Alternative tests to higher criticism for high dimensional means under sparsity and column-wise dependency. Unpublished manuscript.

Y. FAN
DEPARTMENT OF DATA SCIENCES AND OPERATIONS
MARSHALL SCHOOL OF BUSINESS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089
USA
E-MAIL: fanyingy@marshall.usc.edu

J. JIN
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: jiashun@stat.cmu.edu

Z. YAO
SECTION DE MATHÉMATIQUES
ECOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
EPFL STATION 8, 1015 LAUSANNE
SWITZERLAND
E-MAIL: zhigang.yao@epfl.ch