The Annals of Statistics
2010, Vol. 38, No. 3, 1686–1732
DOI: 10.1214/09-AOS764

INNOVATED HIGHER CRITICISM FOR DETECTING SPARSE SIGNALS IN CORRELATED NOISE

BY PETER HALL AND JIASHUN JIN[1]

University of Melbourne and University of California at Davis, and Carnegie Mellon University

Higher criticism is a method for detecting signals that are both sparse and weak. Although first proposed in cases where the noise variables are independent, higher criticism also has reasonable performance in settings where those variables are correlated. In this paper we show that, by exploiting the nature of the correlation, performance can be improved by using a modified approach which exploits the potential advantages that correlation has to offer. Indeed, it turns out that the case of independent noise is the most difficult of all, from a statistical viewpoint, and that more accurate signal detection (for a given level of signal sparsity and strength) can be obtained when correlation is present. We characterize the advantages of correlation by showing how to incorporate them into the definition of an optimal detection boundary. The boundary has particularly attractive properties when correlation decays at a polynomial rate or the correlation matrix is Toeplitz.

1. Introduction. Donoho and Jin [18] developed Tukey's [52] proposal for "higher criticism" (HC), showing that a method based on the statistical significance of a large number of statistically significant test results could be used very effectively to detect the presence of very sparsely distributed signals. They demonstrated that HC is capable of optimally detecting the presence of signals that are so weak and so sparse that the signal cannot be consistently estimated. Applications include the problem of signal detection against cosmic microwave background radiation (Cayon, Jin and Treaster [10], Cruz et al. [16], Jin [36–38], Jin et al. [44]). Related work includes that of Cai, Jin and Low [8], Hall, Pittelkow and Ghosh [29] and Meinshausen and Rice [45].

The context of Donoho and Jin's [18] work was that where the noise is white, although a small number of investigations have been made of the case of correlated noise (Hall, Pittelkow and Ghosh [29], Hall and Jin [30], Delaigle and Hall [17]). However, that research has focused on the ability of standard HC, applied in the form that is appropriate for independent data, to accommodate the nonindependent case. In this paper we address the problem of how to modify HC by developing

Received August 2009; revised October 2009.

[1]Supported in part by NSF CAREER Award DMS-0908613.

AMS 2000 subject classifications. Primary 62G10, 62M10; secondary 62G32, 62H15.

Key words and phrases. Adding noise, Cholesky factorization, empirical process, innovation, multiple hypothesis testing, sparse normal means, spectral density, Toeplitz matrix.

innovated higher criticism (iHC) and showing how to optimize performance for correlated noise.

Curiously, it turns out that when using the iHC method tuned to give optimal performance, the case of independence is the most difficult of all, statistically speaking. To appreciate why this result is reasonable, note that if the noise is correlated then it does not vary so much from one location to a nearby location, and so is a little easier to identify. In an extreme case, if the noise is perfectly correlated at different locations then it is constant, and in this instance it can be easily removed.

On the other hand, standard HC does not perform well in the case of correlated noise, because it utilizes only the marginal information in the data without much attention to the correlation structure. Innovated HC is designed to exploit the advantages offered by correlation and gives good performance across a wide range of settings.

The concept of the "detection boundary" was introduced by Donoho and Jin [18] in the context of white noise. In this paper, we extend it to the correlated case. In brief, the detection boundary describes the relationship between signal sparsity and signal strength that characterizes the boundary between cases where the signal can be detected and cases where it cannot. In the setting of dependent data, this watershed depends on the correlation structure of the noise as well as on the sparsity and strength of the signal. When correlation decays at a polynomial rate we are able to characterize the detection boundary quite precisely. In particular, we show how to construct concise lower/upper bounds to the detection boundary, based on the diagonal components of the inverse of the correlation matrix, Σ_n. A special case is where Σ_n is Toeplitz; there the upper and the lower bounds to the detection boundary are asymptotically the same. In the Toeplitz case, the iHC is optimal for signal detection but standard HC is not.

There is a particularly extensive literature on multiple hypothesis testing under conditions of dependence. It includes contributions to the control of familywise error rate and false discovery rate, and work of Abramovich et al. [1], Benjamini and Hochberg [2], Benjamini and Yekutieli [3], Brown and Russel [7], Cai and Sun [9], Clarke and Hall [12], Cohen, Sackrowitz and Xu [13], Donoho and Jin [19], Dunnett and Tamhane [22], Efron [23], Finner and Roters [24], Genovese and Wasserman [25], Jin and Cai [40], Olejnik et al. [46], Rom [47], Sarkar and Chang [48] and Wu [54]. Work of Kuelbs and Vidyashankar [41] is also related. Our contributions differ from those of these authors in that we point to the advantages, rather than the disadvantages, of dependence, and show how the advantages can be exploited. In particular, as noted above, the problem of denoising dependent data is actually simpler than in the case of independence. We show how to exploit dependence and obtain improvements in performance relative to what is possible in the context of independence and also relative to the inferior performance that is obtained if a method that is designed for the case of independence is applied inappropriately to dependent data. In contrast, earlier work has tended to try to minimize the problems caused by dependence rather than to capitalise on the advantages that are available.

The paper is organized as follows. Section 2 introduces the sparse signal model followed by a brief review of the uncorrelated case. Section 3 establishes lower bounds to the detection boundary in correlated settings. Section 4 introduces innovated HC and establishes an upper bound to the detection boundary. Section 5 applies the main results in Sections 3 and 4 to the case where the Σ_n's are Toeplitz. In this case, the lower bound coincides with the upper bound and innovated HC is optimal for detection. Section 6 discusses a case where the signals have a more complicated structure. Section 7 investigates a case of strong dependence. Simulations are given in Section 8, and discussion is given in Section 9. Section 10 and the Appendix give proofs of theorems and lemmas, respectively.

2. Sparse signal model and review of HC.

2.1. *Model.* Consider an n-dimensional Gaussian vector,

$$X = \mu + Z \qquad \text{where } Z \sim \mathrm{N}(0, \Sigma), \tag{2.1}$$

with the mean vector μ unknown and the dimension n large. In most parts of the paper, we assume that $\Sigma = \Sigma_n$ is known and has unit diagonal elements (the case where Σ_n is unknown is discussed in Section 4.4 and Section 9). We are interested in testing whether no signal exists (i.e., $\mu = 0$) or there is a sparse and faint signal.

Formulae (2.2) and (2.4), below, introduce quantities m and A_n that represent signal sparsity and signal strength, respectively. In particular, as m increases the amount of sparsity decreases, and as A_n increases the strength of the signal increases. Of course, an increase in either m or A_n leads to an increase in the ease with which the signal can be detected and read. It would be possible to connect m and A_n by a formula, and use that relationship to adjust the signal, but we feel that the influence of the key elements of sparsity and strength are most clearly presented by treating them separately. In particular, we model the number of nonzero entries of μ as

$$m = n^{1-\beta} \qquad \text{where } \beta \in (1/2, 1). \tag{2.2}$$

This is a very sparse case, for the proportion of signals is much smaller than $1/\sqrt{n}$. We suppose that the signals appear at m different locations—$\ell_1 < \ell_2 < \cdots < \ell_m$—that are randomly drawn from $\{1, 2, \dots, n\}$ without replacement,

$$P\{\ell_1 = n_1, \ell_2 = n_2, \dots, \ell_m = n_m\} = \binom{n}{m}^{-1} \qquad \text{for all } 1 \le n_1 < n_2 < \cdots < n_m \le n, \tag{2.3}$$

and that they have a common magnitude of

$$A_n = \sqrt{2r \log n} \qquad \text{where } r \in (0, 1). \tag{2.4}$$

These assumptions are made throughout the paper, in cases where Σ is relatively general as well as in cases (see Sections 2.2 and 2.3, below) where the noise variables are assumed uncorrelated and so Σ is the identity. Variations of this model give similar results. For example, if we take the jth nonzero signal to equal $W_j\sqrt{2\log n}$, where the W_j's are independent random variables with a common, nonnegative distribution that has an upper endpoint $r^{1/2}$ satisfying $P(W \le r^{1/2}) = 1$ and $P(W > r^{1/2} - \varepsilon) > 0$ for all $\varepsilon > 0$, then the results are identical to their counterparts when signal strength is given by (2.4).

We are interested in testing which of the following two hypotheses is true:

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1^{(n)} : \mu \text{ is a sparse vector as above.} \tag{2.5}$$

This testing problem was found to be delicate even in the uncorrelated case where $\Sigma_n = I_n$. See [18] (also [8, 32, 33, 36, 45]) for details.

The case where Σ_n is not the identity can arise when signals are recorded at time points that are closely spaced in time or space. See Section 4.4 for discussion. An example of a different type is that of *global testing* in linear models. Here we consider a model $Y \sim \mathrm{N}(M\mu, I_n)$, where the matrix M has many rows and columns, and we are interested in testing whether $\mu = 0$. The setting is closely related to model (2.1), since the least squares estimator of μ is distributed as $\mathrm{N}(\mu, (M'M)^{-1})$. The global testing problem is important in many applications. One is that of testing whether a clinical outcome is associated with the expression pattern of a pre-specified group of genes (Goeman et al. [26, 27]) where M is the expression profile of the specified group of genes. Another is expression quantitative Trait Loci (eQTL) analysis where M is related to the numbers of common alleles for different genetic markers and individuals (Chen, Tong and Zhao [11]). In both examples, M is either observable or can be estimated. Also, it is frequently seen that only a small proportion of genes is associated with the clinical outcome, and each gene contributes weakly to the clinical outcome. In such a situation, the signals are both sparse and faint.

2.2. *Detection boundary in the uncorrelated case* ($\Sigma_n = I_n$). The testing problem is characterized by the curve $r = \rho^*(\beta)$ in the β–r plane where

$$\rho^*(\beta) = \begin{cases} \beta - 1/2, & 1/2 < \beta \le 3/4, \\ (1 - \sqrt{1-\beta})^2, & 3/4 < \beta < 1, \end{cases} \tag{2.6}$$

and we call $r = \rho^*(\beta)$ the *detection boundary*. The detection boundary partitions the β–r plane into two sub-regions: the *undetectable region* below the boundary and the *detectable region* above the boundary (see Figure 1). In the interior of the undetectable region, the signals are so sparse and so faint that no test is able to successfully separate the alternative hypothesis from the null hypothesis in (2.5): the sum of types I and II errors of any test tends to 1 as n diverges to infinity. In the interior of the detectable region, it is possible to have a test such that as n

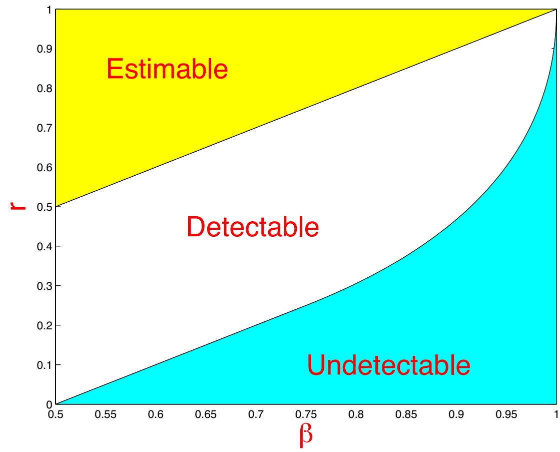

FIG. 1. *Phase diagram for the detection problem in the uncorrelated case. The detection boundary separates the β–r plane into the detectable region and the undetectable region. In the estimable region, it is not only possible to reliably tell the existence of nonzero coordinates, but is also possible to identify them individually.*

diverges to infinity, the type I error tends to zero and the power tends to 1. [In fact, Neyman–Pearson's Likelihood Ratio Test (LRT) is such a test.] See [18, 32, 36], for example.

The drawback of LRT is that it needs detailed information about the unknown parameters (β, r). In practice, we need a test that does not need such information; this is where HC comes in.

2.3. *Higher criticism and its optimal adaptivity in the uncorrelated case* ($\Sigma_n = I_n$). A notion that goes back to Tukey [52], higher criticism was first proposed in [18] to tackle the aforementioned testing problem in the uncorrelated case. To apply higher criticism, let $p_j = P\{|\mathrm{N}(0, 1)| \geq |X_j|\}$ be the p-value associated with the jth observation unit, and let $p_{(j)}$ be the jth p-value after sorting in ascending order. The higher criticism statistic is defined as

$$\mathrm{HC}_n^* = \max_{j\,:\,1/n \leq p_{(j)} \leq 1/2} \left\{ \frac{\sqrt{n}(j/n - p_{(j)})}{\sqrt{p_{(j)}(1 - p_{(j)})}} \right\}. \tag{2.7}$$

There are also other versions of HC (see, e.g., [18, 20, 21]). When H_0 is true, HC_n^* equals in distribution to the maximum of the standardized uniform stochastic process [18]. Therefore, by a well-known result for empirical processes [49],

$$\frac{\mathrm{HC}_n^*}{\sqrt{2 \log\log n}} \to 1 \qquad \text{in probability}. \tag{2.8}$$

Consider the higher criticism test which rejects the null hypothesis when

$$\mathrm{HC}_n^* \geq (1 + a)\sqrt{2 \log\log n} \qquad \text{where } a > 0 \text{ is a constant}. \tag{2.9}$$

It follows from (2.8) that the type I error tends to zero as n diverges to infinity. For any parameters (β, r) that fall in the interior of the detectable region, the type II error also tends to zero. This is the following theorem.

THEOREM 2.1. *Consider the higher criticism test that rejects H_0 when $\mathrm{HC}_n^* \geq (1+a)\sqrt{2\log\log n}$. For every alternative $H_1^{(n)}$ where the associated parameters (r, β) satisfy $r > \rho^*(\beta)$, the HC test has asymptotically full power for detection*:

$$P_{H_1^{(n)}}\{\textit{Reject } H_0\} \to 1 \qquad \textit{as } n \to \infty.$$

That is, the higher criticism test adapts to unknown parameters (β, r) and yields asymptotically full power for detection throughout the entire detectable region. We call this the *optimal adaptivity* of higher criticism [18].

Theorem 2.1 is closely related to [18], Theorem 1.2, where a mixture model is used. The mixture model reduces approximately to the current model if we randomly shuffle the coordinates of X. However, despite its appealing technical convenience, it is not clear how to generalize the mixture model from the uncorrelated case to general correlated settings. Theorem 2.1 is a special case of Theorem 4.2.

We now turn to the correlated case. In this case, the exact "detection boundary" may depend on Σ_n in a complicated manner, but it is possible to establish both a tight lower bound and a tight upper bound. We discuss the lower bound first.

3. Lower bound to detectability.

To establish the lower bound, a key element is the theory in comparison of experiments (e.g., [50]) where a useful guideline is that adding noise always makes the inference more difficult. Thus we can alter the model by either adding or subtracting a certain amount of noise so that the difficulty level (measured by the Hellinger distance, or the χ^2-distance, etc., between the null density and the alternative density) of the original problem is sandwiched by those of the two adjusted models. The correlation matrices in the latter have a simpler form and hence are much easier to analyze. Another key element is the recent development of matrix characterizations based on polynomial off-diagonal decay where it shows that the inverse of a matrix with this property shares the same rate of decay as the original matrix.

3.1. *Comparison of experiments*: *Adding noise makes inference harder.*

We begin by comparing two experiments that have the same mean, but where the data from one experiment are more noisy than those from the other. Intuitively, it is more difficult to make inference in the first experiment than in the other. Specifically, consider the two Gaussian models

$$X = \mu + Z, \qquad Z \sim \mathrm{N}(0, \Sigma) \quad \text{and} \tag{3.1}$$
$$X^* = \mu + Z^*, \qquad Z^* \sim \mathrm{N}(0, \Sigma^*),$$

where μ is an n-vector that is generated according to some distribution $G = G_n$. The second model is more noisy than the first, in the sense that $\Sigma^* \geq \Sigma$. Here, given two matrices, A and B, we write $A \geq B$ if $A - B$ is positive semi-definite.

The second model in (3.1) can be viewed as the result of adding noise to the first. Indeed, defining $\Delta = \Sigma^* - \Sigma$, taking ξ to be $\mathrm{N}(0, \Delta)$ (independently of Z), and noting that $Z + \xi \sim \mathrm{N}(0, \Sigma + \Delta)$, the second model is seen to be equivalent to $X + \xi = \mu + (Z + \xi)$. Intuitively, adding noise makes inference more difficult because it reduces the distance between between X and Z. To make this point concisely, let $\mathrm{Hd}(X, Z; \mu, \Sigma)$ and $\mathrm{Hd}(X^*, Z^*; \mu, \Sigma^*)$ denote the Hellinger distance between (the distributions of) X and Z, and between X^* and Z^*, respectively. Then we claim that the first of these distances exceeds the second

$$\text{if } \Sigma^* \geq \Sigma \text{ in (3.1)} \qquad \text{then } \mathrm{Hd}(X, Z; \mu, \Sigma) \geq \mathrm{Hd}(X^*, Z^*; \mu, \Sigma^*). \tag{3.2}$$

See Section 10 for a proof. [The Hellinger distance between distributions with densities f and g equals $\frac{1}{2}\int (f^{1/2} - g^{1/2})^2$.]

3.2. *Matrices having polynomial off-diagonal decay.* Next, we review results concerning matrices with polynomial off-diagonal decay. The main message is that, under mild conditions, if a matrix has polynomial off-diagonal decay, then its inverse as well as its Cholesky factorization (which is unique if we require the diagonal entries to be positive) also have polynomial off-diagonal decay, and with the same rate. This beautiful result was recently obtained by Jaffard [34] (see also [28, 51]).

In detail, writing Θ_n for the set of $n \times n$ correlation matrices, we introduce, for $\lambda > 1$,

$$\Theta_n^*(\lambda, c_0, M) = \{\Sigma_n \in \Theta_n : |\Sigma_n(j,k)| \leq M(1 + |j - k|)^{-\lambda}, \|\Sigma_n\| \geq c_0\}. \tag{3.3}$$

This is the set of matrices which have a given rate of polynomial off-diagonal decay and where the operator norm is uniformly bounded from below. Consider a sequence of matrices $\{\Sigma_n\}_{n=1}^\infty$ such that $\Sigma_n \in \Theta_n^*(\lambda, c_0, M)$ for each n. It turns out that the inverses (as well as the Cholesky factorizations) of such sequences enjoy polynomial off-diagonal decay with the same rate as that of the matrices themselves. See the Appendix for the proof.

We are now ready for the lower bound.

3.3. *Lower bound to detectability.* Consider a sequence of matrices $\{\Sigma_n\}_{n=1}^\infty$ such that $\Sigma_n \in \Theta_n^*(\lambda, c_0, M)$ for each n. Suppose the extreme diagonal entries of Σ_n^{-1} have an upper limit $\bar{\gamma}_0$ in the range $0 < \bar{\gamma}_0 < \infty$; that is,

$$\overline{\lim_{n\to\infty}} \Big(\max_{\sqrt{n} \leq k \leq n - \sqrt{n}} \Sigma_n^{-1}(k,k) \Big) = \bar{\gamma}_0. \tag{3.4}$$

Recall that the detection boundary in the uncorrelated case is defined by $r < \rho^*(\beta)$. The following theorem asserts that, in the presence of correlation, if we change the definition to $r < \bar{\gamma}_0^{-1} \cdot \rho^*(\beta)$, then we obtain at least a lower bound to the detection boundary.

THEOREM 3.1. *Fix $\beta \in (1/2, 1)$, $r \in (0, 1)$, $\lambda > 1$, $c_0 > 0$, and $M > 0$. Consider a sequence of correlation matrices $\Sigma_n \in \Theta_n^*(\lambda, c_0, M)$ that satisfy* (3.4). *If $r < \bar{\gamma}_0^{-1}\rho^*(\beta)$, then the null hypothesis and alternative hypothesis in* (2.5) *merge asymptotically, and the sum of types* I *and* II *errors of any test converges to* 1 *as n diverges to infinity.*

We now turn to the upper bound. The key is to adapt the higher criticism to correlated noise and form a new statistic—*innovated higher criticism.*

4. Innovated higher criticism, upper bound to detectability. Originally designed for the independent case, standard HC is not really appropriate for dependent data for the following reasons. First, HC only summarizes the information that resides in the marginal effects of each coordinate and neglects the correlation structure of the data. Second, HC remains the same if we randomly shuffle different coordinates of X. Such shuffling does not have an effect if $\Sigma_n = I_n$, but does otherwise. In this section we build the correlation into the standard higher criticism and form a new statistic—innovated higher criticism (iHC). We then use iHC to establish an upper bound to detectability. The iHC is intimately connected to the well-known notion of innovation in time series [6] [see (4.1) below], hence the name innovated higher criticism.

Below, we begin by discussing the role of correlation in the detection problem.

4.1. *Correlation among different coordinates*: *Curse or blessing*? Consider model (2.1) in the two cases $\Sigma_n = I_n$ and $\Sigma_n \neq I_n$. Which is the more difficult detection problem?

Here is one way to look at it. Since the mean vectors are the same in the two cases, the problem where the noise vector contains more "uncertainty" is more difficult than the other. In information theory, the *total amount of uncertainty* is measured by the *differential entropy*, which in the Gaussian case is proportional to the determinant of the correlation matrix [15]. As the determinant of a correlation matrix is largest when and only when it is the identity matrix, the uncorrelated case contains the largest amount of "uncertainty" and therefore gives the most difficult detection problem. In a sense, the correlation is a "blessing" rather than a "curse" as one might have expected.

Here is another way to look at it. For any positive definite matrix Σ_n, denote the inverse of its Cholesky factorization by U_n, a function of Σ_n (so that $U_n\Sigma_n U_n' = I_n$). Model (2.1) is equivalent to

$$U_n X = U_n\mu + U_n Z \qquad \text{where } U_n Z \sim \mathrm{N}(0, I_n). \tag{4.1}$$

(In the literature of time series [6], $U_n X$ is intimately connected to the notion of innovation.) Compared to the uncorrelated case, that is,

$$X = \mu + Z \qquad \text{where } Z \sim \mathrm{N}(0, I_n).$$

It turns out that the noise vectors have the same distribution, but the signals in the former are stronger. In fact, let $\ell_1 < \ell_2 < \cdots < \ell_m$ be the m locations where μ is nonzero. Recalling that $\mu_j = A_n$ if $j \in \{\ell_1, \ell_2, \dots, \ell_m\}$, $\mu_j = 0$ otherwise, and that U_n is a lower triangular matrix,

$$(U_n\mu)_{\ell_k} = A_n \sum_{j=1}^{k} U_n(\ell_k, \ell_i) = A_n U_n(\ell_k, \ell_k) + A_n \sum_{j=1}^{k-1} U_n(\ell_j, \ell_k). \tag{4.2}$$

Two key observations are as follows. First, since Σ_n has unit diagonal entries, every diagonal entry of U_n is greater than or equal to 1, especially

$$U_n(\ell_k, \ell_k) \geq 1. \tag{4.3}$$

Second, recall that $m \ll n$, and $\{\ell_1, \ell_2, \dots, \ell_m\}$ are randomly generated from $\{1, 2, \dots, n\}$, so different ℓ_j are far apart from each other. Therefore, under mild decay conditions on U_n,

$$U_n(\ell_j, \ell_k) \approx 0, \qquad j = 1, 2, \dots, k-1. \tag{4.4}$$

Inserting (4.3) and (4.4) into (4.2), we expect that $(U_n\mu)_{\ell_k} \gtrsim A_n$ for $k = 1, 2, \dots, m$. Therefore, "on average," $U_n\mu$ has at least m entries each of which is at least as large as A_n. This says that, first, the correlated case is easier for detection than the uncorrelated case. Second, applying standard HC to U_nX yields a larger power than applying it to X directly.

Next we make the argument more precise. Fix a positive sequence $\{\delta_n : n \geq 1\}$ that tends to zero as n diverges to infinity, and a sequence of integers $\{b_n : n \geq 1\}$ that satisfy $1 \leq b_n \leq n$. Recall that U_n is the function of Σ_n defined by $U_n \Sigma_n U_n' = I_n$, and let

$$\tilde{\Theta}_n^*(\delta_n, b_n) = \Bigg\{\Sigma_n \in \Theta_n, \sum_{j=1}^{k-b_n} |U_n(k, j)| \leq \delta_n, \\ \text{for all } k \text{ satisfying } b_n + 1 \leq k \leq n\Bigg\}.$$

Introducing $\tilde{\Theta}_n^*$ seems a digression from our original plan of focusing on Θ_n^* (the set of matrices with polynomial off-diagonal decay), but it is interesting in its own right. In fact, compared to Θ_n^*, $\tilde{\Theta}_n^*$ is much broader as it does not impose much of a condition on $\Sigma_n(j, k)$ for $|j - k| \leq b_n$. This helps to illustrate how broadly the aforementioned phenomenon holds. The following theorem is proved in Section 10.

THEOREM 4.1. *Fix $\beta \in (1/2, 1)$ and $r \in (\rho^*(\beta), 1)$. Let $b_n = n^\beta/3$, and let δ_n be a positive sequence that tends to zero as n diverges to infinity. Suppose we apply standard higher criticism to U_nX and we reject H_0 if and only if the resulting score* *exceeds* $(1+a)\sqrt{2\log\log n}$ *where* $a>0$. *Then*, *uniformly in all sequences of* Σ_n *satisfying* $\Sigma_n \in \tilde{\Theta}_n^*(\delta_n, b_n)$,

$$P_{H_0}\{\textit{Reject } H_0\} + P_{H_1^{(n)}}\{\textit{Accept } H_0\} \to 0, \qquad n\to\infty.$$

Generally, directly applying standard HC to X does not yield the same result (e.g., [30]).

4.2. *Innovated higher criticism*: *Higher criticism based on innovations*. We have learned that applying standard HC to U_nX yields better results than applying it to X directly. Is this the best we can do? No, there is still space for improvement. In fact, HC applied to U_nX is a special case of innovated higher criticism to be elaborated in this section. Innovated higher criticism is even more powerful in detection.

To begin, we revisit the vector $U_n\mu$ via an example. Fix $n=100$; let Σ_n be a symmetric tri-diagonal matrix with 1 on the main diagonal, 0.4 on two sub-diagonals and zero elsewhere; and let μ be the vector with 1 at coordinates 27, 50, 71 and zero elsewhere. Figure 2 compares μ and $U_n\mu$. Especially, the nonzero coordinates of $U_n\mu$ appear in three visible clusters, each of which corresponds to a different nonzero entry of μ. Also, at coordinates 27, 50, 71, $U_n\mu$ approximately equals to 1.2, but μ equals 1. To interpret the figure caption, recall that U_n is the function of Σ_n defined by $U_n\Sigma_nU_n' = I_n$.

Now we can either simply apply standard HC to U_nX as before, or we can first linearly transform each cluster of signals to a singleton and then apply the standard HC. Note that in the second approach, we may have fewer signals, but each of them is much stronger than those in U_nX. Since the HC test is more sensitive to signal

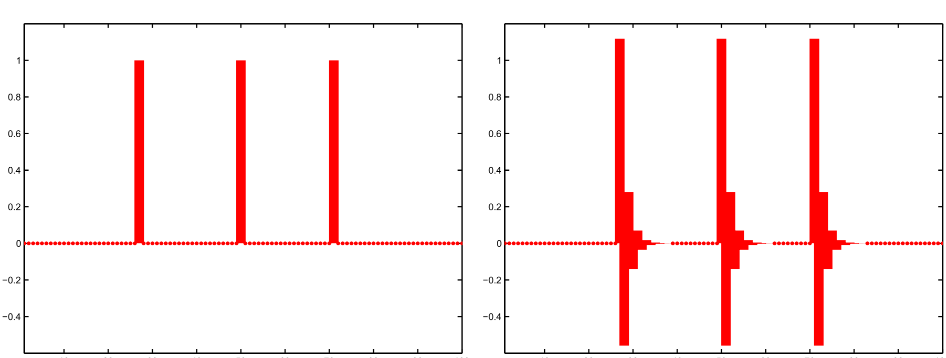

FIG. 2. *Comparison of* μ (*left*) *and* $U_n\mu$ (*right*). *Here* $n=100$ *and* Σ_n *is a symmetric tri-diagonal matrix with* 1 *on the main diagonal*, 0.4 *on two sub-diagonals and zero elsewhere*. *Also*, μ *is* 1 *at coordinates* 27, 50, *and* 71 *and* 0 *elsewhere*. *In comparison*, *the nonzero entries of* $U_n\mu$ *appear in three visible clusters*, *each of which corresponds to a nonzero coordinate of* μ.

strength than to the number of signals, we expect that the second approach yields greater power for detection than the first.

In light of this we propose the following approach. Write $U_n = (u_{kj})_{\{1 \le k, j \le n\}}$. We pick a bandwidth $1 \le b_n \le n$, and construct a matrix $\tilde{U}_n(b_n) = U_n(\Sigma_n, b_n)$ by banding U_n [4]

$$(4.5)\qquad \tilde{U}(b_n) \equiv (\tilde{u}_{kj})_{1 \le j,k \le n}, \qquad \tilde{u}_{kj} = \begin{cases} u_{kj}, & k - b_n + 1 \le j \le k, \\ 0, & \text{otherwise.} \end{cases}$$

We then normalize each column of $\tilde{U}_n(b_n)$ by its own ℓ^2-norm, and call the resulting matrix $\bar{U}_n(b_n)$. Next, defining

$$(4.6)\qquad V_n(b_n) = V_n(b_n; \Sigma_n) = \bar{U}_n'(b_n; \Sigma_n) \cdot U_n,$$

we transform model (2.1) into

$$(4.7)\qquad X \longmapsto V_n(b_n)X = V_n(b_n)\mu + V_n(b_n)Z.$$

Finally, we apply standard higher criticism to $V_n(b_n)X$, and call the resulting statistic *innovated higher criticism*,

$$(4.8)\qquad \begin{aligned} \mathrm{iHC}_n^*(b_n) &= \mathrm{iHC}_n^*(b_n; \Sigma_n) \\ &= \frac{1}{\sqrt{2b_n - 1}} \sup_{j\,:\,1/n \le p_{(j)} \le 1/2} \left\{ \sqrt{n} \cdot \frac{j/n - p_{(j)}}{\sqrt{p_{(j)}(1 - p_{(j)})}} \right\}. \end{aligned}$$

Note that standard HC applied to $U_n X$ is a special case of iHC_n^* with $b_n = 1$.

We briefly comment on the selection of the bandwidth parameter b_n. First, for each $k \in \{\ell_1, \ell_2, \ldots, \ell_m\}$, direct calculations show that $(V_n(b_n)\mu)_k \approx A_n \cdot \sqrt{\sum_{j=1}^{b_n} u_{k,k-j+1}^2} \ge A_n$. Second, $V_n(b_n)Z \sim \mathrm{N}(0, \bar{U}_n'(b_n)\bar{U}_n(b_n))$, where $\bar{U}_n'(b_n) \times \bar{U}_n(b_n)$ is a banded correlation matrix with bandwidth $2b_n - 1$. Therefore, choosing b_n involves a trade-off: a larger b_n usually means stronger signals but also means stronger correlation among the noise. While it is hard to give a general rule for selecting the best b_n, we must mention that in many cases, the choice of b_n is not very critical. For example, when Σ_n has polynomial off-diagonal decay, a logarithmically large b_n is usually appropriate.

4.3. *Upper bound to detectability.* We now establish an upper bound to detectability. Suppose the diagonal entries of Σ_n^{-1} have a lower limit as follows:

$$(4.9)\qquad \varliminf_{n \to \infty} \Big(\min_{\sqrt{n} \le k \le n - \sqrt{n}} \Sigma_n^{-1}(k, k) \Big) = \underline{\gamma_0}.$$

Recall that the nonzero coordinates of μ are modeled as $A_n = \sqrt{2r \log n}$. If we let $b_n = \log n$ then it can be proved that the vector $V_n(b_n) \cdot X$ has at least m nonzero coordinates, each of which is as large as $\sqrt{\underline{\gamma_0}} A_n = \sqrt{2\underline{\gamma_0} \cdot r \cdot \log n}$. (See

Lemma A.3.) Note that a larger b_n cannot improve the signal strength significantly, but may yield a much stronger correlation in $V_n(b_n)Z$. Therefore, a smaller bandwidth is preferred. The choice $b_n = \log n$ is mainly for convenience, and can be modified.

We now turn to the behavior of $\mathrm{iHC}_n^*(b_n)$ under the null hypothesis. In the independent case, iHC_n^* reduces to HC_n^* and is approximately equal to $\sqrt{2\log\log n}$. In the current situation, iHC_n^* is comparably larger due to the correlation. However, since the selected bandwidth is relatively small, iHC_n^* remains logarithmically large. See Lemma A.5 for details. The following theorem elaborates on the upper bound, and is proved in Section 10.

THEOREM 4.2. *Fix* $c_0 > 0$, $\lambda > 1$, *and* $M > 0$, *and set* $b_n = \log n$. *Suppose* $\underline{\gamma_0} \cdot r > \rho^*(\beta)$. *If we reject* H_0 *when* $\mathrm{iHC}_n^*(b_n; \Sigma_n) \geq (\log n)^2$, *then, uniformly in all* $\Sigma_n \in \Theta_n^*(\lambda, c_0, M)$,

$$P_{H_0}\{\mathit{Reject}\ H_0\} + P_{H_1^{(n)}}\{\mathit{Accept}\ H_0\} \to 0 \qquad \text{as } n \to \infty.$$

The cut-off value $(\log n)^2$ can be replaced by other logarithmically large terms that tend to infinity faster than $(\log n)^{3/2}$. For finite n, this cut-off value may be conservative. In Section 8 [i.e., experiment (a)], we suggest an alternative where we select the cut-off value by simulation.

In summary, a lower bound and an upper bound are established as $r = \bar{\gamma}_0^{-1}\rho^*(\beta)$ and $r = \underline{\gamma_0}^{-1}\rho^*(\beta)$, respectively, under reasonably weak off-diagonal decay conditions. When $\bar{\gamma}_0 = \underline{\gamma_0}$, the gap between the two bounds disappears, and iHC is optimal for detection. Below in Sections 5–7, we investigate several Toeplitz cases, ranging from weak dependence to strong dependence; for these cases, iHC is optimal in detection.

4.4. *Effect of estimating* Σ_n *and related issues.* So far, we have assumed that the covariance matrix Σ_n is known. When Σ_n is unknown, we could still use iHC if Σ_n could be estimated. We now briefly comment on the effect of estimating Σ_n.

In practical problems where iHC methodology would be used, noise could reasonably be represented as a time series, and its characteristics estimated from data. In particular, the time series might be an autoregression, and data over a longer period than that for which the current dataset was recorded could be used to deduce properties of the noise. Examples include detection of xenon byproducts as evidence of a nuclear explosion, early detection of bioweapons and detection of covert communications.

If data are gathered over a time period of length p, if the signal is present at no more than $m = n^{1-\beta}$ points where $\beta \in (1/2, 1)$ and if the maximum size of the signal is no greater than a constant multiple of $(\log n)^{1/2}$, then it is typically possible to estimate the components of U_n at rate $(p^{-1}\log p)^{1/2}$ uniformly in all components. From this property it can be proved that the difference between $U_n\mu$

and its empirical form equals $O_p[\{n^{2-\beta}p^{-1}(\log n)(\log p)\}^{1/2}]$, uniformly in all components. Similarly, if the noise process is conventional (e.g., an autoregression) then the distance between $U_n Z$ and its empirical form can be shown to equal $O_p(n^{1+\varepsilon}/p)$ for all $\varepsilon > 0$. Therefore the effects of variance estimation will be asymptotically negligible if, for some $\varepsilon > 1 - \beta$, $n^{1+\varepsilon}/p \to 0$ converges to zero as $n \to \infty$.

To appreciate the extent to which this condition is restrictive, consider the case where the signals are particularly sparse, that is, β is close to 1; say, $\beta = 1 - \eta_1$ where $\eta_1 > 0$ is small. Then the condition holds if p is at least as large as $n^{1+\eta_2}$ for some $\eta_2 > \eta_1$. That is, the amount of time for which data have to be acquired in order to estimate Σ_n with sufficient accuracy need only be a factor n^{ε} greater than n, for $\varepsilon > 0$ relatively small. As the prevalence of the signal increased, the size of ε would have to too.

Application of our methods to other problems, such as those involving genomic data, can be inhibited by the difficulty of estimating Σ_n without information from outside the dataset. However, while there is sometimes evidence of strong dependence in genomic data, from other viewpoints the overall level of correlation is often quite low. For example, Messer and Arndt [43] argue that correlation decays from about 0.08, at a separation of approximately two base pairs, to about 0.01 for a separation of ten base pairs. Work of Mansilla et al. [42] corroborates these figures. Results such as these, together with the upper tail independence property which is generally available for light-tailed distributions, suggest that for genomic data it is possible to work effectively under the assumption that expression levels are statistically independent, even when they are not. Details are given by Delaigle and Hall [17], who use the fact that in the case of genomic data the variables are typically t-statistics.

More generally, cases where the signals are distributed nonrandomly can be compared readily with the case of independent, randomly-distributed signals, noted just below (2.2), as follows. Let us take as our benchmark the classical problem $\mathcal{P}(n_0, m_0)$ where there are n_0 independent noise variables, and m_0 signals are distributed randomly among the n_0 locations. We shall compare it with the more general problem where the noise variables are d-dependent with the integer d depending on n. To quantify the effects of nonrandomness we assume that $m = n^{1-\beta}$ signals are distributed among m/K clumps of length $K = K(n)$, and that the points in clumps that are furthest to the right are distributed sequentially among the integers $K, K+1, \ldots, n-1, n$, with each placement being conditional on the clump not overlapping any pre-existing clumps. We make no other assumption about the dependence structure of the process for placing the clumps, only that it be independent of the noise variables; and we assume that $d \le K$. If $K = O(n^{\eta})$ for all $\eta > 0$ then, for each η, the difficulty of the signal detection problem is bounded above by that of $\mathcal{P}(n^{1-\eta}, mn^{-\eta})$, and below by that of $\mathcal{P}(n, m)$. Since η here is arbitrary then it can be deduced that the effect of clustering has asymptotically negligible effect. On the other hand, if $K = n^{\eta}\ell_n$ for a fixed $\eta > 0$ and a quantity

ℓ_n that satisfies $\ell_n = O(n^{\varepsilon})$ and $n^{\varepsilon} = O(\ell_n)$ for all $\varepsilon > 0$, then the problem can be asymptotically as difficult as $\mathcal{P}(n^{1-\eta}, mn^{-\eta})$ for the given value of η.

5. Application in the Toeplitz case. In this section, we discuss the case where Σ_n is a (truncated) Toeplitz matrix that is generated by a spectral density f defined over $(-\pi, \pi)$. In detail, let $a_k = (2\pi)^{-1} \int_{|\theta|<\pi} f(\theta) e^{-ik\theta}\, d\theta$ be the kth Fourier coefficient of f. The nth truncated Toeplitz matrix generated by f is the matrix $\Sigma_n(f)$ of which the (j,k)th element is a_{j-k}, for $1 \le j, k \le n$.

We assume that f is symmetric and positive, that is,

$$c_0(f) \equiv \operatorname*{essinf}_{-\pi \le \theta \le \pi} f(\theta) > 0. \tag{5.1}$$

First, note that f is a density, so $a_0 = 1$ and $\Sigma_n(f)$ has unit diagonal entries. Second, from the symmetry of f, it can be seen that $\Sigma_n(f)$ is a real-valued symmetric matrix. Last, it is well known [5] that the smallest eigenvalue of $\Sigma_n(f)$ is no smaller than $c_0(f)$, so $\Sigma_n(f)$ is positive definite. Putting all these together, $\Sigma_n(f)$ is seen to be a correlation matrix.

Toeplitz matrices enjoy convenient asymptotic properties. In detail, let $\lambda > 1$ and suppose that additionally f has at least λ bounded derivatives [meaning, if λ is a positive integer, that $|f^{(j)}|$ is bounded for $0 \le j \le \lambda$, and, if λ is not an integer, that $|f^{(j)}|$ is bounded for $0 \le j < \lambda$ and $|f^{(\lambda')}(\theta_1) - f^{(\lambda')}(\theta_2)|/|\theta_1 - \theta_2|^{\lambda-\lambda'}$ is bounded, where λ' denotes the largest integer less than λ]. Then by elementary Fourier analysis, there is a constant $M_0 = M_0(f) > 0$ such that

$$|a_k| \le M_0(f)(1+k)^{-\lambda} \qquad \text{for } k = 0, 1, 2, \dots. \tag{5.2}$$

Comparing (5.1) and (5.2) with the definition of Θ_n^*, we conclude that

$$\Sigma_n \in \Theta_n^*(\lambda, c_0(f), M_0(f)). \tag{5.3}$$

In addition, it is known that the inverse of $\Sigma_n(f)$ is typically asymptotically equivalent to the Toeplitz matrix generated by $1/f$. The diagonal entries of $\Sigma_n(1/f)$ are the well-known *Wiener interpolation rates* [53],

$$C(f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{f(\theta)}\, d\theta. \tag{5.4}$$

From this property and a result of [5], Theorem 2.15, it can be proved that

$$\max_{\sqrt{n} \le k \le n - \sqrt{n}} |\Sigma_n^{-1}(f)(k,k) - C(f)| = o(1).$$

Comparing this with (3.4) and (4.9) we deuce that

$$\bar{\gamma}_0 = \underline{\gamma_0} = C(f). \tag{5.5}$$

Combining (5.3) and (5.5), the following theorem is a direct result of Theorems 3.1 and 4.1 (the proof is omitted).

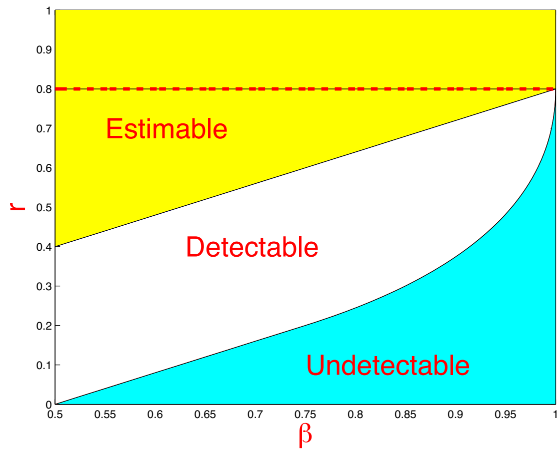

FIG. 3. *Phase diagram in the case where Σ_n is a Toeplitz matrix generated by a spectral density f. Similarly, in Figure 1, the β–r plane is partitioned into three regions—undetectable, detectable, estimable—each of which can be viewed as the corresponding region in Figure 1 squeezed vertically by a factor of $1/C(f)$. In the rectangular region on the top, the largest signals in $V_n(b_n) \cdot X$ [see (4.6)] are large enough to stand out by themselves.*

THEOREM 5.1. *Fix $\lambda > 1$, and let $\Sigma_n(f)$ be the Toeplitz matrix generated by a symmetric spectral density f that satisfies (5.1) and (5.2). When $C(f) \cdot r < \rho^*(\beta)$, the null and alternative hypotheses merge asymptotically, and the sum of types I and II errors of any test converges to 1 as n diverges to infinity. When $C(f) \cdot r > \rho^*(\beta)$, suppose we apply iHC with bandwidth $b_n = \log n$ and reject the null hypothesis when $\mathrm{iHC}_n^*(b_n, \Sigma_n(f)) \geq (\log n)^2$. Then the type I error of iHC converges to zero, and its power converges to 1.*

The curve $r = C(f)^{-1}\rho^*(\beta)$ partitions the β–r plane into the undetectable region and the detectable region, similarly to the uncorrelated case. The regions of the current case can be viewed as the corresponding regions in the uncorrelated squeezed vertically by a factor of $1/C(f)$. See Figure 3. [Note that $C(f) \geq 1$, with equality if and only if $f \equiv 1$, which corresponds to the uncorrelated case.]

6. Extension: When signals appear in clusters. In the preceding sections [see, e.g., (2.3) in Section 2], the m locations of signals were generated randomly from $\{1, 2, \ldots, n\}$. Since $m \ll \sqrt{n}$, the signals appear as singletons with overwhelming probabilities. In this section we investigate an extension where the signals may appear in clusters.

We consider a setting where the signals appear in a total of m clusters, whose locations are randomly generated from $\{1, 2, \ldots, n\}$. Each cluster contains a total of K consecutive signals, whose strengths are $g_0 A_n, g_1 A_n, \ldots, g_{K-1} A_n$, from right to left. Here, $A_n = \sqrt{2r \log n}$ as before, $K \geq 1$ is a fixed integer and g_i are constants. Approximately, the signal vector can be modeled as follows.

As before, let $\ell_1, \ell_2, \ldots, \ell_m$ be indices that are randomly sampled from $\{1, 2, \ldots, n\}$. Let $\mu = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$, where $\mu_j = A_n$ if $j \in \{\ell_1, \ell_2, \ldots, \ell_m\}$, and $\mu_j = 0$ otherwise. Let $B = B_n$ denote the "backward shift" matrix with 0 in every position except that it has 1 in position $(j+1, j)$ for $1 \le j \le n-1$. Thus, $B\mu$ differs from μ in that the components are shifted one position backward, with 0 added at the bottom. We model the signal vector as

$$\nu = g_0\mu + g_2 B\mu + \cdots + g_k B^{K-1}\mu = \left(\sum_{k=0}^{K-1} g_k B^k\right)\mu.$$

Thus ν is comprised of m clusters, each of which contains K consecutive signals. Let g be the function $g(\theta) = \sum_{0 \le k \le K-1} g_k e^{-ik\theta}$. We note that $\sum_{0 \le k \le K-1} g_k B^k$ is the lower triangular Toeplitz matrix generated by g. With the same spectral density f, we consider an extension of that in Section 5 by considering the following model:

$$X = \Sigma_n(g)\mu + Z \qquad \text{where } Z \sim \mathrm{N}(0, \Sigma_n(f)), \tag{6.1}$$

with f denoting the spectral density in Section 5.

We note that the model can be equivalently viewed as

$$\tilde{X} = \mu + \tilde{Z} \qquad \text{where } \tilde{Z} \sim \mathrm{N}(0, \tilde{\Sigma}_n) \quad \text{and} \quad \tilde{\Sigma}_n = \Sigma_n^{-1}(g) \cdot \Sigma_n(f) \cdot \Sigma_n^{-1}(\bar{g}),$$

with $\bar{g}$ denoting the complex conjugate of g. Asymptotically,

$$\tilde{\Sigma}_n^{-1} \sim \Sigma_n(\bar{g}) \cdot \Sigma_n^{-1}(f) \cdot \Sigma_n(g) \sim \Sigma_n(|g|^2/f),$$

where the diagonal entries of $\Sigma_n(|g|^2/f)$ are

$$C(f, g) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|g(\theta)|^2}{f(\theta)} \, d\theta.$$

If $\bar{\gamma}_0$ and $\underline{\gamma_0}$ are as defined in (3.4) and (4.9), then $\underline{\gamma_0} = \bar{\gamma}_0 = C(f, g)$, and we expect the detection boundary to be $r = C(f, g)^{-1} \cdot \rho^*(\beta)$. This is affirmed by the following theorem which is proved in Section 10.

THEOREM 6.1. *Fix $\lambda > 1$. Suppose $g_0 \neq 0$ and let f be a symmetric spectral density that satisfies* (5.1) *and* (5.2). *When $C(f, g) \cdot r < \rho^*(\beta)$, the null and alternative hypotheses merge asymptotically, and the sum of types* I *and* II *errors of any test converges to* 1 *as n diverges to infinity. When $C(f, g) \cdot r > \rho^*(\beta)$, if we apply iHC to $\Sigma_n^{-1}(g)X$ with bandwidth $b_n = \log n$ and reject the null hypothesis when* $\mathrm{iHC}_n^*(b_n, \Sigma_n^{-1}(g)\Sigma_n(f)\Sigma_n^{-1}(\bar{g})) \ge (\log n)^2$, *then the type* I *error converges to zero, and the power converges to* 1.

7. The case of strong dependence. So far, we have only discussed weakly dependent cases. In this section, we investigate the case of strong dependence.

Suppose we observe an n-variate Gaussian vector $X = \mu + Z$, where μ contains a total of m signals, of equal strength to be specified, whose locations are randomly drawn from $\{1, 2, \dots, n\}$ without replacement, and $Z \sim \mathrm{N}(0, \Sigma_n)$ where we assume that Σ_n displays slowly decaying correlation,

$$\Sigma_n(j,k) = \max\{0, 1 - |j-k|^{\alpha} n^{-\alpha_0}\}, \qquad 1 \le j, k \le n, \tag{7.1}$$

with $\alpha > 0$ and $0 < \alpha_0 \le \alpha$. The range of dependence can be calibrated in terms of $k_0 = k_0(n; \alpha, \alpha_0)$, denoting the largest integer by $k < n^{\alpha_0/\alpha}$. Clearly, $k_0 \approx n^{\alpha_0/\alpha}$. Seemingly, the most interesting range is $0 < \alpha_0 \le \alpha \le 1$.

Condition (7.1) is more restrictive than similar assumptions in other places in this paper. There are at least two reasons. First, the constants in the definition of the detection boundary turn out to depend intimately on the value of α used in the definition of Σ_n at (7.1), and so we need to make an assumption which is driven by that parameter. Secondly, a significantly more general definition of Σ_n would need to satisfy the positive definiteness property which (as can be seen from Lemma A.12) is somewhat delicate.

Model (7.1) has been studied in detail by Hall and Jin [30] who showed that the detectability of standard HC is seriously damaged by strong dependence. However, it remains open as to what is the detection boundary, and how to adapt HC to overcome the strong dependence and obtain optimal detection. This is what we address in the current section.

The key idea is to decompose the correlation matrix as the product of three matrices each of which is relatively easy to handle. To begin with we introduce a spectral density,

$$f_\alpha(\theta) = 1 - \sum_{k=1}^{\infty} [(k+1)^\alpha + (k-1)^\alpha - 2k^\alpha] \cos(k\theta). \tag{7.2}$$

[Note that the Fourier coefficients of $f_\alpha(\theta)$ satisfy the decay condition in (5.2) with $\lambda = 2 - \alpha$.] Next, let

$$g_0(\theta) = 1 - e^{-i\theta}, \qquad a_n = a_n(\alpha_0) = n^{\alpha_0}/2.$$

The Toeplitz matrix $\Sigma_n(g_0)$ is a lower triangular matrix with 1's on the main diagonal, -1's on the sub-diagonal and 0's elsewhere. Additionally, let D_n be the diagonal matrix where on the diagonal the first entry is 1 and the remaining entries are $\sqrt{a_n}$. Let $\tilde{X} = D_n \cdot \Sigma_n(g_0) \cdot X$. Then model (7.1) can be rewritten equivalently as

$$\tilde{X} = \tilde{\mu} + \tilde{Z} \qquad \text{where } \tilde{\mu} = D_n \cdot \Sigma_n(g_0) \cdot \mu \text{ and } \tilde{Z} \sim \mathrm{N}(0, \tilde{\Sigma}_n) \tag{7.3}$$

with $\tilde{\Sigma}_n = D_n \cdot \Sigma_n(g_0) \cdot \Sigma_n \cdot \Sigma_n(\bar{g}_0) \cdot D_n$. The key is that $\tilde{\Sigma}_n$ is asymptotically equivalent to the Toeplitz matrix generated by f_α. In detail, introduce

$$\bar{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & \Sigma_{n-1}(f_\alpha) \end{pmatrix}.$$

It follows from Lemma A.6 that the spectral norm of $\tilde{\Sigma}_n - \bar{\Sigma}_n$ converges to zero as n diverges.

Note that $\tilde{\mu} = \sqrt{a_n} \cdot \Sigma_{n-1}(g) \cdot \mu$ except for the first coordinate. Therefore, we expect model (7.3) to be approximately equivalent to

$$\tilde{X} = \sqrt{a_n} \cdot \Sigma_n(g_0) \cdot \mu + \tilde{Z} \qquad \text{where } \tilde{Z} \sim \mathrm{N}(0, \Sigma_n(f_\alpha)).$$

This is a special case of the cluster model we considered in Section 6 with $f = f_\alpha$ and $g = g_0$, except that the signal strength has been re-scaled by $\sqrt{a_n}$. Therefore, if we calibrate the nonzero entries in μ as

$$a_n^{-1/2} \cdot A_n = a_n^{-1/2} \cdot \sqrt{2r \log n}, \tag{7.4}$$

then the detection boundary for the model is succinctly characterized by

$$r = \frac{1}{C(f_\alpha, g_0)} \cdot \rho^*(\beta),$$

$$C(f_\alpha, g_0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|g_0(\theta)|^2}{f_\alpha(\theta)} \, d\theta = \frac{1}{\pi} \int_{-\pi}^{\pi} \frac{1 - \cos(\theta)}{f_\alpha(\theta)} \, d\theta.$$

See Figure 4 for the display of $C(f_\alpha, g_0)$. The following theorem is proved in Section 10.

THEOREM 7.1. *Let $0 < \alpha_0 \leq \alpha < \frac{1}{2}$, $\beta \in (\frac{1}{2}, 1)$, and $r \in (0, 1)$. Assume X is generated according to model (7.1), with signal strength re-scaled as in (7.4). When $C(f_\alpha, g_0) \cdot r < \rho^*(\beta)$, the null and alternative hypotheses merge asymptotically, and the sum of types* I *and* II *errors of any test converges to* 1 *as n diverges to infinity. When $C(f_\alpha, g_0) \cdot r > \rho^*(\beta)$, if we apply the iHC to X with bandwidth $b_n = \log n$ and reject the null when $\mathrm{iHC}_n^*(b_n, \Sigma_n) \geq (\log n)^2$, then the type* I *error converges to zero, and its power converges to* 1.

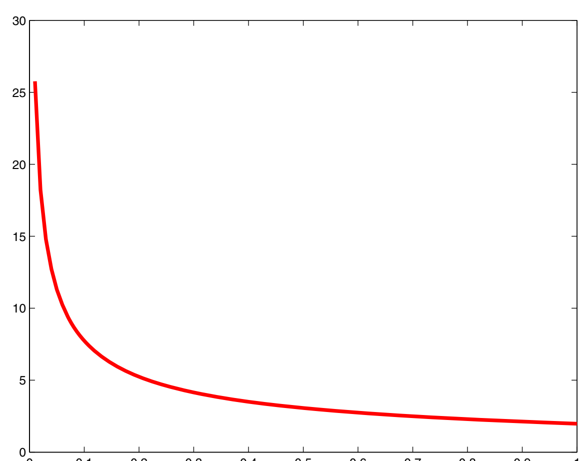

FIG. 4. *Display of $C(f_\alpha, g_0)$. x-axis: α. y-axis: $C(f_\alpha, g_0)$.*

8. Simulation study. We conducted a small-scale empirical study to compare the performance of iHC and standard HC. For iHC, we investigate two choices of bandwidth: $b_n = 1$ and $b_n = \log n$. In this section, we denote standard HC, iHC with $b_n = 1$, and iHC with $b_n = \log n$ by HC, HC-a and HC-b correspondingly.

The algorithm for generating data included the following four steps: (1) Fix n, β, and r, let $m = n^{1-\beta}$ and $A_n = \sqrt{2r \log n}$. (2) Given a correlation matrix Σ_n, generate a Gaussian vector $Z \sim \mathrm{N}(0, \Sigma_n)$. (3) Randomly draw m integers $\ell_1 < \ell_2 < \cdots < \ell_m$ from $\{1, 2, \ldots, n\}$ without replacement, and let μ be the n-vector such that $\mu_j = A_n$ if $j \in \{\ell_1, \ell_2, \ldots, \ell_m\}$ and 0 otherwise. (4) Let $X = \mu + Z$. Using data generated in this manner we explored three parameter settings, (a)–(c), which we now describe.

In experiment (a), we took $n = 1000$ and $\Sigma_n(\rho)$ as the tri-diagonal Toeplitz matrix generated by $f(\theta) = 1 + 2\rho\cos(\theta)$, $|\rho| < 1/2$. The corresponding detection boundary was $r = \rho^*(\beta)/C(f)$ with $C(f) = (2\pi)^{-1}\int_{-\pi}^{\pi}[1 - 2\rho\cos(\theta)]^{-1}\,d\theta$. Consider all ρ that range from -0.45 to 0.45 with an increment of 0.05, and four pairs of parameters $(\beta, r) = (0.5, 0.2)$, $(0.5, 0.25)$, $(0.55, 0.2)$ and $(0.55, 0.25)$. [Note that the corresponding parameters (m, A_n) are $(32, 1.66)$, $(32, 2.63)$, $(22, 1.66)$ and $(22, 2.63)$]. For each triple (β, r, ρ), we generated data according to (1)–(4), applied HC, HC-a and HC-b to both Z and X and repeated the whole process independently 500 times. As a result, for each triple (β, r, ρ) and each procedure, we got 500 HC scores that corresponded to the null hypothesis and 500 HC scores that corresponded to the alternative hypothesis.

We report the results in two different ways. First, we report the minimum sum of types I and II errors (i.e., the minimum of the sum across all possible cut-off values) (see Figure 5). Second, we pick the upper 10% percentile of the 500 HC scores corresponding to the null hypothesis as a threshold (for later references, we call this threshold the *empirical threshold*) and calculate the empirical power of the test (i.e., the fraction of HC scores corresponding to the alternative hypothesis that exceeds the threshold). The empirical thresholds are displayed in Table 1 (to save space, only part of the thresholds are reported), and the power is displayed in Figure 6. Recall that in Theorem 4.2 we recommend $(\log n)^2$ as a cut-off point in the asymptotic setting. For moderately large n, this cut-off point is conservative, and we recommend the empirical threshold instead.

The results suggest that (1) iHC-b outperforms iHC-a, and iHC-a outperforms HC. (2) As $|\rho|$ increases (note that a larger $|\rho|$ means a stronger correlation), the detection problem is increasingly easier, and the advantage of iHC is increasingly prominent. (3) Under the null hypothesis, the HC-b scores are usually smaller than those of HC and HC-a. This is mainly due to the normalization term $\sqrt{2b_n - 1}$ in the definition of iHC [see (4.8)].

We set the cut-off value as the 10% percentile only for convenience. Replacing 10% by other percentage gives similar conclusion. See Figure 7 for details.

In experiment (b), we took Σ_n to be the Toeplitz matrix generated by $f(\theta) = 1 + \frac{1}{2}\cos(\theta) + 2\rho\cos(2\theta)$ where ρ ranged from -0.2 to 0.45 with an increment of

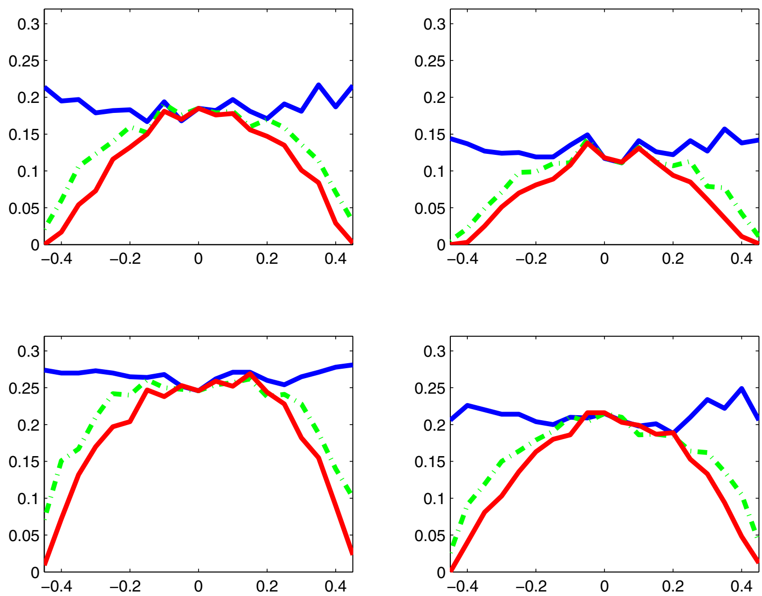

FIG. 5. *Sum of types* I *and* II *errors as described in experiment* (a). *From top to bottom then from left to right*, $(\beta, r) = (0.5, 0.2)$, $(0.5, 0.25)$, $(0.55, 0.2)$, $(0.55, 0.25)$. *In each panel*, *the x-axis displays ρ*, *and three curves* (*blue*, *dashed-green*, *and red*) *display the sum of errors corresponding to HC*, *HC-a and HC-b.*

0.05. (The matrix Σ_n is positive definite when ρ is in this range.) Other parameters are the same as in experiment (a). The minimum sums of types I and II errors are reported in Figure 8. The results suggest similarly that HC-b outperforms HC-a, and HC-a outperforms HC.

In experiment (c), we investigated the behavior of HC-a/HC-b/HC for larger n. We took $(\beta, r) = (0.5, 0.25)$, $n = 500 \times (1, 2, 3, 4, 5)$ and Σ_n as the tri-diagonal matrix in experiment (a) with $\rho = 0.4$. The sum of types I and II errors is reported in Table 2. The results suggest that the performance of HC-a/HC-h/HC improve when n gets larger. (Investigation of the case where n was much larger than 2500 needed much greater computer memory, and so we omitted it.)

TABLE 1
Display of empirical thresholds in experiment (a) *for different* ρ

ρ	**−0.45**	**−0.35**	**−0.25**	**−0.15**	**−0.05**	**0.05**	**0.15**	**0.25**	**0.35**	**0.45**
HC	3.059	2.851	2.913	2.892	2.722	2.835	2.742	2.858	2.834	3.032
HC-a	2.919	2.858	2.924	2.837	2.723	2.899	2.713	2.826	2.677	2.758
HC-b	0.890	0.847	0.806	0.773	0.769	0.775	0.772	0.761	0.832	0.859

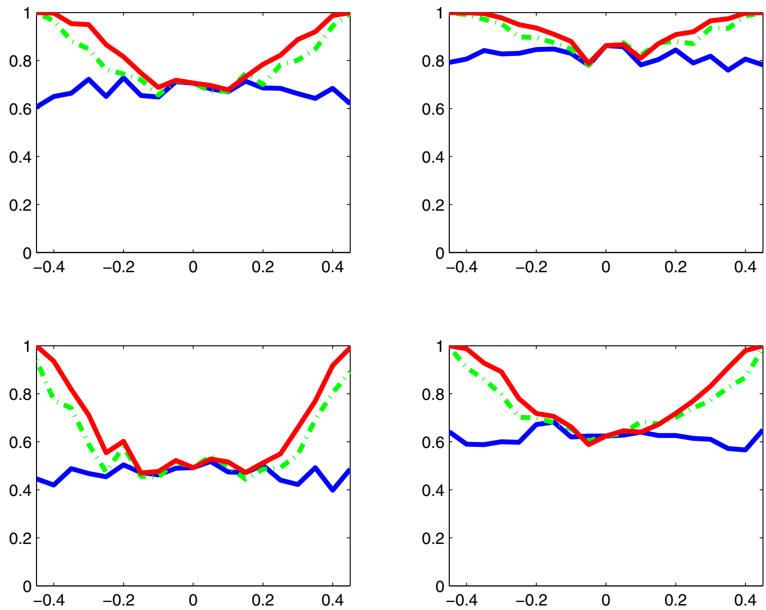

FIG. 6. *Power as described in experiment* (a). *From top to bottom then from left to right*, $(\beta, r) = (0.5, 0.2)$, $(0.5, 0.25)$, $(0.55, 0.2)$, $(0.55, 0.25)$. *In each panel*, *the* x*-axis displays* ρ, *and three curves* (*blue*, *dashed-green and red*) *display the power of HC*, *HC-a and HC-b*.

9. Discussion. We have extended standard HC to innovated HC by building in the correlation structure. The extreme diagonal entries of Σ_n^{-1} play a key role in the testing problem. If the extreme value has finite upper and lower limits, $\bar{\gamma}_0$ and $\underline{\gamma_0}$, then in the β–r plane, the detection boundary is bounded by the curves $r = \underline{\gamma_0}^{-1} \cdot \rho^*(\beta)$ from above and $r = \bar{\gamma}_0^{-1} \cdot \rho^*(\beta)$ from below. When the correlation matrix is Toeplitz, the upper and lower limits merge and equal the Wiener interpolation rate $C(f)$. The detection boundary is therefore $r = C(f)^{-1} \cdot \rho^*(\beta)$. The detection boundary partitions the β–r plane into a detectable region and an undetectable region. Innovated HC has asymptotically full power for detection whenever (β, r) falls into the interior of the detectable region (we note, however, neither β nor r is used to construct iHC). We call this the optimally adaptivity of innovated higher criticism.

9.1. *Connection to recent literature.* The work complements that of Donoho and Jin [18] and Hall and Jin [30]. The focus of [18] is standard HC and its performance in the uncorrelated case. The focus of [30] is how strong dependence may harm the effectiveness of standard HC; what could be a remedy was, however, not explored. The innovated HC proposed in the current paper is optimal for both the model in [18] and that in [30].

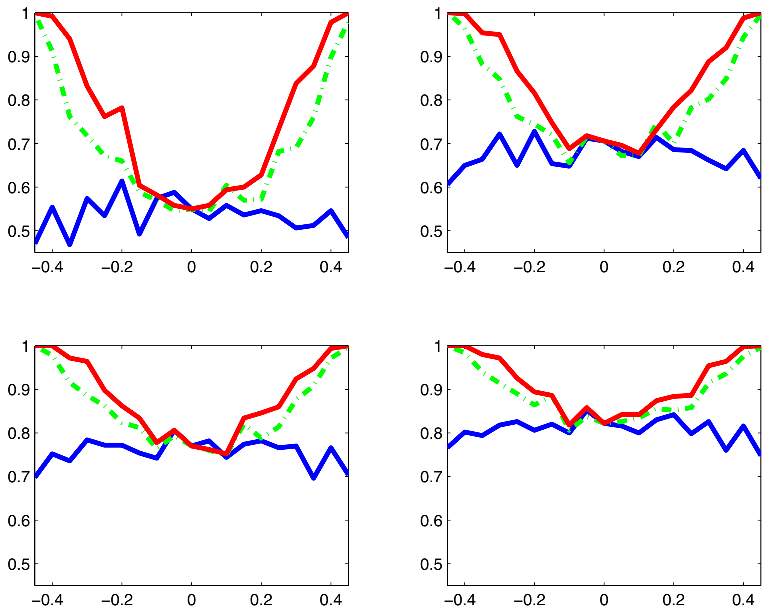

FIG. 7. *Display of powers for different choices of cut-off value. Fix* $(\beta, r) = (0.5, 0.2)$ *as in experiment* (a). *From top to bottom then from left to right, the cut-off values are the* 5%, 10%, 15% *and* 20% *percentile of the* 500 *HC scores corresponding to the null hypothesis. In each panel, the* x*-axis displays* ρ, *and three curves* (*blue, dashed-green and red*) *display the power of HC, HC-a and HC-b. The display suggest that, for different choices of cut-off value, HC-b consistently outperforms HC-a, and HC-a consistently outperforms HC.*

The work is related to that of Jager and Wellner [35] where the authors proposed a family of goodness-of-fit statistics for detecting sparse normal mixtures. The work is also related to that of Meinshausen and Rice [45] and of Cai, Jin and Low [8], where the authors focused on how to estimate ε_n—the proportion of nonnull effects.

Recently, HC was also found to be useful for feature selection in high-dimensional classification. See Donoho and Jin [20, 21], Hall, Pittelkow and Ghosh [29] and Jin [39]. The work concerned the situation where there are relatively few samples containing a very large number of features, out of which only a small fraction is useful, and each useful feature contributes weakly to the classification problem. In a related setting, Delaigle and Hall [17] investigated HC for classification when the data is non-Gaussian or dependent.

9.2. *Future work.* The work is also intimately connected to recent literature on estimating covariance matrices. While the study is focused more on situations where the correlation matrices can be estimated using other approaches (e.g., [11, 26, 27]), it can be generalized to cases where the correlation matrix is unknown but can be estimated from data. Cases where data on the covariance structure are

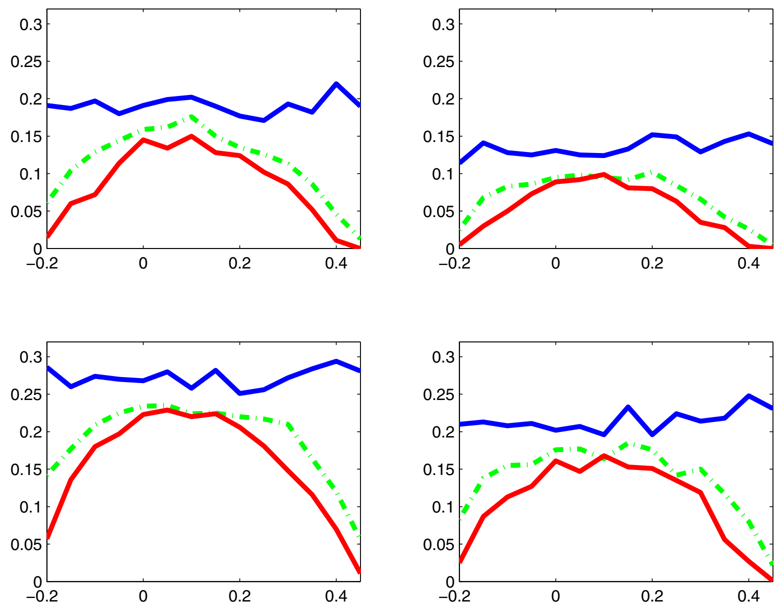

FIG. 8. *Sum of types* I *and* II *errors as described in experiment* (b). *From top to bottom then from left to right*, $(\beta, r) = (0.5, 0.2)$, $(0.5, 0.25)$, $(0.55, 0.2)$, $(0.55, 0.25)$. *In each panel*, x*-axis displays* ρ, *and three curves* (*blue*, *dashed-green and red*) *display the sum of errors corresponding to HC*, *HC-a and HC-b.*

available from other time periods were discussed in Section 4.4, but even if we stay within the confines of the current data, progress can be made. In particular, it is noteworthy that it was shown in Bickel and Levina [4] that when the correlation matrix has polynomial off-diagonal decay, the matrix and its inverse can be estimated accurately in terms of the spectral norm. In such situations we expect the proposed approach to perform well once we combine it with that in [4].

Another interesting direction is to explore cases where the correlation matrix does not have polynomial off-diagonal decay, but is sparse in an unspecified pattern. This is a more challenging situation as relatively little is known about the inverse of the correlation matrix.

TABLE 2
Display of the sum of types I *and* II *errors in experiment* (c) *for different* n

n	500	1000	1500	2000	2500
HC	0.201	0.144	0.115	0.123	0.098
HC-a	0.073	0.035	0.017	0.017	0.021
HC-b	0.033	0.005	0.004	0.003	0.002

Our study also opens opportunities for improving other recent procedures. Take the aforementioned work on classification [20, 21, 29, 39], for example. The approach derived in this paper suggests ways of incorporating correlation structure into feature selection, and therefore raises hopes for better classifiers. For reasons of space, we leave explorations along these directions to future study.

10. Proofs of main results. In this section we prove all theorems in preceding sections, except Theorems 2.1 and 5.1. These two theorems are the direct result of Theorems 3.1 and 4.2, so we omit the proofs. For simplicity, we drop the subscript n whenever there is no confusion.

10.1. *Proof of* (3.2). rewrite the second model in (3.1) as $X + \xi = \mu + \xi + Z$, where independently, $Z \sim \mathrm{N}(0, \Sigma)$, $\xi \sim \mathrm{N}(0, \Delta)$, $\mu \sim G$ for $\Delta = \Sigma_n^* - \Sigma$ and some distribution G. It suffices to show the monotonicity in the Hellinger distance. Denote the density function of $\mathrm{N}(0, \Sigma)$ by $f(x) = f(x_1, x_2, \ldots, x_n)$, and write $dx_1\, dx_2 \cdots dx_n$ as dx for short. Then the Hellinger distance corresponding to the second model in (3.1) can be written as

$$h(\Sigma, \Delta, G) \equiv \int \sqrt{E_\Delta\bigl(E_G\bigl(f(x-\mu-\xi)\bigr)\bigr) \cdot E_\Delta\bigl(f(x-\xi)\bigr)}\, dx.$$

Note that by Hölder's inequality, $\sqrt{E[f]E[g]} \geq E[\sqrt{fg}]$ for any positive and integrable functions f and g. Using Fubini's theorem, $h(\Sigma, \Delta, G)$ is not less than

$$\begin{aligned}&\int \bigl[E_\Delta \sqrt{\bigl(E_G f(x-\mu-\xi)\bigr) \cdot \bigl(f(x-\xi)\bigr)}\bigr]\, dx \\ &\quad = E_\Delta\biggl[\int \sqrt{E_G f(x-\mu-\xi) f(x-\xi)}\, dx\biggr].\end{aligned}$$

Note that $\int \sqrt{E_G f(x-\mu-\xi) f(x-\xi)}\, dx \equiv \int \sqrt{E_G f(x-\mu) f(x)}\, dx$ for any fixed ξ. It follows that

$$\begin{aligned}h(\Sigma, \Delta, G) &\geq E_\Delta\biggl[\int \sqrt{E_G f(x-\mu-\xi) f(x-\xi)}\, dx\biggr] \\ &= \int \sqrt{E_G f(x-\mu) f(x)}\, dx,\end{aligned}$$

where the last term is the Hellinger distance corresponds to the first model of (3.1). Combining these results gives the claim.

10.2. *Proof of Theorem* 3.1. It is sufficient to show that the Hellinger distance between the joint density of X and Z converges to zero as n diverges to infinity. By the assumption $\bar{\gamma}_0 r < \rho^*(\beta)$, we can choose a sufficiently small constant $\delta = \delta(r, \beta, \gamma_0)$ such that $\bar{\gamma}_0(1-\delta)^{-2} r < \rho^*(\beta)$. Let $\tilde{\mu} = \mu/\sqrt{1-\delta}$, let U be the inverse of the Cholesky factorization of Σ, and let $\tilde{U}$ be the banded version of U,

$$\tilde{U}(i, j) = \begin{cases} U(i, j), & |i - j| \leq \log^2(n), \\ 0, & \text{otherwise}. \end{cases}$$

Model (2.1) can be equivalently written as

$$X = \tilde{\mu} + Z \qquad \text{where } Z \sim \mathrm{N}\big(0, (1-\delta)^{-1} \cdot \Sigma\big). \tag{10.1}$$

The key to the proof is to compare model (10.1) with the following model:

$$X = \tilde{\mu} + Z \qquad \text{where } Z \sim \mathrm{N}(0, (\tilde{U}'\tilde{U})^{-1}). \tag{10.2}$$

In fact, by (3.2), to establish the claim it suffices to prove that (i) $\tilde{U}'\tilde{U} \le (1-\delta)^{-1}\Sigma$ for sufficiently large n, and (ii) the Hellinger distance between the joint density of X and that of Z associated with model (10.2) tends to zero as n diverges to infinity.

To prove the first claim, noting that $\Sigma = (U'U)^{-1}$, it suffices to show $(1-\delta)U'U \le \tilde{U}'\tilde{U}$. Define $W = U - \tilde{U}$ and observe that there is a generic constant $C > 0$ such that $\|\tilde{U}\| \le C$ and $\|W\| \le C$, whence $\|\tilde{U}'\tilde{U} - \tilde{U}'\tilde{U}\| = \|W'W + \tilde{U}'W + W'\tilde{U}\| \le C\|W\|$. Moreover, by [31], Theorem 5.6.9, for any symmetric matrix, the spectral norm is no greater than the ℓ^1-norm. In view of the definitions of W and $\Theta_n^*(\lambda, c_0, M)$, the ℓ^1-norm of W is no greater than $(\log n)^{-2(\lambda-1)}$. Therefore, $\|\tilde{U}'\tilde{U} - \tilde{U}'\tilde{U}\| \le C\|W\| \le C(\log n)^{-2(\lambda-1)}$. This, and the fact that all eigenvalues of $\tilde{U}'\tilde{U}$ are bounded from below by a positive constant, imply the claim.

We now consider the second claim. Model (10.2) can be equivalently written as $X = \tilde{U}\tilde{\mu} + Z$ where $Z \sim \mathrm{N}(0, I_n)$. The key to the proof is that $\tilde{U}$ is a banded matrix and μ is a sparse vector where with probability converging to 1, the inter-distances of nonzero coordinates are no less than $3(\log n)^2$ (see Lemma A.8 for the proof). As a result the nonzero coordinates of $\tilde{U}\tilde{\mu}$ are disjoint clusters of sizes $O(\log^2 n)$ which simplifies the calculation of the Hellinger distance. The derivation of the claim is summarized in Lemma A.7 which is stated and proved in the Appendix.

10.3. *Proof of Theorem* 4.1. Recall that U_n is the function of Σ_n defined by $U_n \Sigma_n U_n' = I_n$. Put $Y = U_n X$, $\nu = U_n \mu$ and $Z = U_n z$. Model (4.1) reduces to

$$Y = \nu + Z, \qquad Z \sim \mathrm{N}(0, I_n). \tag{10.3}$$

Recalling that $\mathrm{HC}_n^*/\sqrt{2\log\log n} \to 1$ in probability under H_0, it follows that $P_{H_0}\{\text{Reject } H_0\}$ tends to zero as n diverges to infinity, and it suffices to show $P_{H_1^{(n)}}\{\text{Accept } H_0\} \to 0$.

The key to the proof is to compare model (10.3) with

$$Y^* = \nu^* + Z \qquad \text{where } Z \sim \mathrm{N}(0, I_n), \tag{10.4}$$

with ν^* having m nonzero entries of equal strength $(1-\delta_n)A_n$ whose locations are randomly drawn from $\{1, 2, \dots, n\}$ without replacement. By (4.2) and (4.3) and the way $\tilde{\Theta}_n^*(\delta_n, b_n)$ is defined, we note that $\nu_j \ge (1-\delta_n)A_n$ for all $j \in \{\ell_1, \ell_2, \dots, \ell_m\}$. Therefore,

$$\text{signals in } \nu \text{ are both denser and stronger than those in } \nu^*. \tag{10.5}$$

Intuitively, standard HC applied to model (10.3) is no "less" than that applied to model (10.4).

We now establish this point. Let $\bar{F}_0(t)$ be the survival function of the central χ^2-distribution $\chi_1^2(0)$, and let $\bar{F}_n(t)$ and $\bar{F}_n^*$ be the empirical survival function of $\{Y_k^2\}_{k=1}^n$ and $\{(Y_k^*)^2\}_{k=1}^n$, respectively. Using arguments similar to those of Donoho and Jin [18] it can be shown that standard HC applied to models (10.3) and (10.4), denoted by $\mathrm{HC}_n^{(1)}$ and $\mathrm{HC}_n^{(2)}$ for short, can be rewritten as

$$\mathrm{HC}_n^{(1)} = \sup_{t:\,1/n\le\bar{F}_0(t)\le 1/2}\left\{\frac{\sqrt{n}(\bar{F}_n(t)-\bar{F}_0(t))}{\sqrt{\bar{F}_0(t)F_0(t)}}\right\},$$

$$\mathrm{HC}_n^{(2)} = \sup_{t:\,1/n\le\bar{F}_0(t)\le 1/2}\left\{\frac{\sqrt{n}(\bar{F}_n^*(t)-\bar{F}_0(t))}{\sqrt{\bar{F}_0(t)F_0(t)}}\right\},$$

respectively. The key fact is now that the family of noncentral χ^2-distribution $\{\chi_1^2(\delta), \delta\ge 0\}$ is a monotone likelihood ratio family (MLR), that is, for any fixed x and $\delta_2\ge\delta_1\ge 0$, $P\{\chi_1^2(\delta_2)\ge x\}\ge P\{\chi_1^2(\delta_1)\ge x\}$. Consequently, it follows from (10.5) and mathematical induction that for any x and t, $P\{\bar{F}_n^*(t)\ge x\}\ge P\{\bar{F}_n(t)\ge x\}$. Therefore, for any fixed $x>0$,

$$P\{\mathrm{HC}_n^{(1)}<x\}\le P\{\mathrm{HC}_n^{(2)}<x\}. \tag{10.6}$$

Finally, by an argument similar to that of Donoho and Jin [18], Section 5.1, the second term in (10.6) with $x=(1+a)\sqrt{2\log\log n}$ tends to zero as n diverges to infinity. This implies the claim.

10.4. *Proof of Theorem* 4.2. In view of Lemma A.5, it suffices to show that $P_{H_1^{(n)}}\{\text{Accept } H_0\}\to 0$. Put $\bar{U}=\bar{U}(b_n)$, $V=V_n(b_n)$, $Y=VX$, $\nu=V\mu$, $\tilde{Z}=VZ$. Model (4.7) reduces to

$$Y=\nu+\tilde{Z} \qquad \text{where } \tilde{Z}\sim \mathrm{N}(0,\bar{U}'\bar{U}). \tag{10.7}$$

Let $\bar{F}_n(t)$ and $\bar{F}_0(t)$ be the empirical survival function of $\{Y_k^2\}_{k=1}^n$ and the survival function of $\chi_1^2(0)$, respectively. Let $q=q(\beta,r)=\min\{(\beta+\bar{\gamma}_0 r)^2/(4\bar{\gamma}_0 r), 4\bar{\gamma}_0 r\}$ and set $t_n^*=\sqrt{2q\log n}$. Since $\bar{\gamma}_0 r<\rho^*(\beta)$, then it can be shown that $0<q<1$ and $n^{-1}\le\bar{F}_0(t_n^*)\le 1/2$ for sufficiently large n. Using an argument similar to that in the proof of Theorem 4.1,

$$\begin{aligned}\mathrm{iHC}_n^* &= \sup_{s:\,1/n\le\bar{F}_0(s)\le 1/2}\frac{\sqrt{n}(\bar{F}_n(s)-\bar{F}_0(s))}{\sqrt{(2b_n-1)\bar{F}_0(s)(1-\bar{F}_0(s))}}\\ &\ge \frac{\sqrt{n}(\bar{F}_n(t_n^*)-\bar{F}_0(t_n^*))}{\sqrt{(2b_n-1)\bar{F}_0(t_n^*)(1-\bar{F}_0(t_n^*))}}\end{aligned}$$

and it follows that

$$P\{\mathrm{iHC}_n^* \le \log^{3/2}(n)\} \le P\left\{\frac{\sqrt{n}(\bar{F}_n(t_n^*) - \bar{F}_0(t_n^*))}{\sqrt{(2b_n-1)\bar{F}_0(t_n^*)(1-\bar{F}_0(t_n^*))}} \le \log^{3/2}(n)\right\}. \tag{10.8}$$

It remains to show that the right-hand side of (10.8) is algebraically small. The proof needs detailed calculations summarized in Lemma A.11 which is stated and proved in the Appendix.

10.5. *Proof of Theorem* 6.1. Inspection of the proof of Theorems 3.1 and 4.2 reveals that the condition that Σ_n is a correlation matrix and that $\Sigma_n \in \Theta_n^*(\lambda, c_0, M)$ in those theorems can be relaxed. In particular, Σ_n need not have equal diagonal entries, and the decay condition on Σ_n can be replaced by a weaker condition that concerns the decay of U_n (the inverse of the Cholesky factorization of Σ_n), specifically

$$|U_n(i,j)| \le M(1 + |i-j|^\lambda)^{-1}.$$

Let $U_n(f)$ be the inverse of the Cholesky factorization of $\Sigma_n(f)$, and define $\tilde{U}_n = U_n(f)\Sigma_n(g)$. Since $\Sigma_n(g)$ is a lower triangular matrix with positive diagonal entries, then it is seen that $\tilde{U}_n$ is the inverse of the Cholesky factorization of $\tilde{\Sigma}_n$. By Lemma A.1, $U_n(f)$ has polynomial off-diagonal decay with the parameter λ. It follows that $\tilde{U}_n$ decays at the same rate. Applying Theorems 3.1 and 4.2, we see that all that remains to prove is that

$$\max_{\sqrt{n} \le k \le n - \sqrt{n}} |\tilde{\Sigma}_n^{-1}(k,k) - C(f,g)| \to 0. \tag{10.9}$$

By [5], Theorem 2.15, for any $\sqrt{n} \le k \le n - \sqrt{n}$, $k - K \le j \le k + K$ and $1 \le \lambda' < \lambda$,

$$\left|\Sigma_n^{-1}(f)(k,j) - \left(\Sigma_n(1/f)\right)(k,j)\right| = o\left(n^{-(1-\lambda')/2}\right).$$

Since $\tilde{\Sigma}_n^{-1} = \Sigma_n(\bar{g}) \cdot \Sigma_n^{-1}(f) \cdot \Sigma_n(g)$, it follows that $\sup_{\sqrt{n} \le k \le n-\sqrt{n}} |\tilde{\Sigma}_n^{-1}(k,k) - (\Sigma_n(\bar{g}) \cdot \Sigma_n(1/f) \cdot \Sigma_n(g))(k,k)| \to 0$. Moreover, direct calculations show that $(\Sigma_n(\bar{g}) \cdot \Sigma_n(1/f) \cdot \Sigma_n(g))(k,k) = C(f,g)$, $\sqrt{n} \le k \le n - \sqrt{n}$. Combining these results gives (10.9) and concludes the proofs.

10.6. *Proof of Theorem* 7.1. Consider the first claim. It suffices to show that the Hellinger distance between $\tilde{X}$ and $\tilde{Z}$ in model (7.3) tends to zero as n diverges to infinity. Since $C(f_\alpha, g_0) \cdot r < \rho^*(\beta)$, there is a small constant $\delta > 0$ such that $(1-\delta)^{-1} \cdot C(f_\alpha, g_0) \cdot r < \rho^*(\beta)$. Using Lemma A.13, we see that $\Sigma_{n-1}(f_\alpha)$ is a positive matrix the smallest eigenvalue of which is bounded away from zero. It follows from Lemma A.6 and basic algebra that $\tilde{\Sigma} \ge (1-\delta)\bar{\Sigma}_n$ for sufficiently large n. Compare model (7.3) with

$$X^* = \tilde{\mu} + Z^* \qquad \text{where } Z^* \sim \mathrm{N}\left(0, (1-\delta)\bar{\Sigma}\right). \tag{10.10}$$

By the monotonicity of Hellinger distance at (3.2), it suffices to show that the Hellinger distance between X^* and Z^* tends to zero as n diverges to infinity.

Now, by the definition of $\tilde{\mu}$, $\tilde{\mu} - \sqrt{a_n} \cdot \Sigma_n(g_0) \cdot \mu = (\mu_n, \sqrt{a_n} \cdot \mu_n, 0, \dots, 0)'$. Since $P\{\mu_n \neq 0\} = o(1)$ then, except for an event with negligible probability, $\tilde{\mu} = \bar{\mu}$. Therefore, replacing $\tilde{\mu}$ by $\sqrt{a_n} \cdot \Sigma_n(g_0) \cdot \mu$ in model (10.10) alters the Hellinger distance only negligibly. Note that the first coordinate of X^* is uncorrelated with all other coordinates, and its mean equals zero with probability converging to 1, so removing it from the model only has a negligible effect on the Hellinger distance. Combining these properties, model (10.10) reduces to the following with only a negligible difference in the Hellinger distance:

$$X^*(2:n) = \Sigma_{n-1}(g_0)\bigl(\sqrt{a_n} \cdot \mu(2:n)\bigr) + Z^*(2:n),$$
$$Z^*(2:n) \sim \mathrm{N}\bigl(0, (1-\delta)\Sigma_{n-1}(f_\alpha)\bigr),$$

where $X(2:n)$ denotes the vector X with the first entry removed. Dividing both sides by $\sqrt{1-\delta}$, this reduces to the following model:

$$\tilde{X}(2:n) = \Sigma_{n-1}(g_0)\frac{\sqrt{a_n} \cdot \mu(2:n)}{\sqrt{1-\delta}} + \tilde{Z}(2:n),$$
$$\tilde{Z}(2:n) \sim \mathrm{N}(0, \Sigma_{n-1}(f_\alpha)), \tag{10.11}$$

which is in fact model (6.1) considered in Section 6. It follows from (7.4) that $\sqrt{a_n} \cdot \mu(2:n)/\sqrt{1-\delta}$ has m nonzero coordinates each of which equals $\sqrt{2(1-\delta)^{-1} r \log n}$. Comparing model (10.11) with model (6.1) and recalling that $(1-\delta)^{-1} \cdot r \cdot C(f_\alpha, g_0) < \rho^*(\beta)$, the claim follows from Theorem 6.1.

Consider the second claim. Since $C(f_\alpha, g_0) \cdot r > \rho^*(\beta)$, then there is a small constant $\delta > 0$ such that $(1-\delta) \cdot r \cdot C(f_\alpha, g_0) > \rho^*(\beta)$. Let U_n be the inverse of the Cholesky factorization of Σ_n, and let $\bar{U}_n(b_n)$ and $V_n(b_n)$ be as defined right below (4.5). Write model (7.1) equivalently as

$$VX = V\mu + VZ \qquad \text{where } VZ \sim \mathrm{N}(0, \bar{U}'(b_n)\bar{U}(b_n)).$$

Recall that $\bar{U}'(b_n)\bar{U}(b_n)$ is a banded correlation matrix with bandwidth $2b_n - 1$. Let $\ell_1, \ell_2, \dots, \ell_m$ be the m locations of nonzero means of μ. By an argument similar to that in the proof of Theorem 4.2, all remains to show is that, except for an event with negligible probability,

$$(V\mu)_k \geq \sqrt{2r' \log n}$$
$$\text{for some constant } r' > \rho^*(\beta) \text{ and all } k \in \{\ell_1, \ell_2, \dots, \ell_k\}. \tag{10.12}$$

We now show (10.12). First, by Lemma A.3 and (7.4), except for an event with negligible probability,

$$(V\mu)_k \geq (1-\delta)^{1/4} \cdot \bigl(a_n \cdot \Sigma_n(k,k)\bigr)^{-1/2} \cdot A_n, \qquad k \in \{\ell_1, \ell_2, \dots, \ell_m\}.$$

Second, by the way $\tilde{\Sigma}_n$ is defined,

$$(a_n \Sigma_n^{-1})(k,k) = \big(\Sigma_n(g_0) \cdot \tilde{\Sigma}_n^{-1} \cdot \Sigma_n(\bar{g}_0)\big)(k,k) \qquad \text{for all } k \geq 2,$$

and by the way $\bar{\Sigma}_n$ is defined and Lemma A.6, for sufficiently large n,

$$\tilde{\Sigma}_n^{-1} \geq (1-\delta)^{-1/2} \bar{\Sigma}_n^{-1},$$

and so

$$\Sigma_n(g_0) \tilde{\Sigma}_n^{-1} \Sigma_n(\bar{g}_0) \geq (1-\delta)^{1/2} \Sigma_n(g) \bar{\Sigma}_n^{-1} \Sigma_n(\bar{g}).$$

Last, by [5], Theorem 2.15, $|(\Sigma_n(g_0) \cdot \bar{\Sigma}_n^{-1} \cdot \Sigma_n(\bar{g}_0))(k,k) - C(f_\alpha, g_0)| = o(1)$ when $\min\{k, n-k\}$ is sufficiently large. Combining these results gives (10.12) with $r' = (1-\delta) \cdot r \cdot C(f_\alpha, g_0)$, and the claim follows directly.

APPENDIX

A.1. Statement and proof of Lemma A.1.

LEMMA A.1. *Fix $\lambda > 1$, $c_0 > 0$, and $M > 0$. For any sequence of matrices Σ_n, $n \geq 1$, such that $\Sigma_n \in \Theta_n^*(\lambda, c_0, M)$, let U_n be the inverse of the Cholesky factorization of Σ_n. Then there is a constant $C = C(\lambda, c_0, M) > 0$ such that, for any n and any $1 \leq j, k \leq n$,*

$$|\Sigma_n^{-1}(j,k)| \leq C \cdot (1 + |j-k|)^{-\lambda}, \qquad |U_n(j,k)| \leq C \cdot (1 + |j-k|)^{-\lambda}.$$

PROOF. When $\lambda = 1$, the first inequality continues to hold, and the second holds if we adjoin a $\log n$ factor to the right-hand side.

As a prelude to giving the proof we state the following result, taken directly from [51]. Let $\mathbb{Z}$ be the set of all integers. Write ℓ^2 for the set of summable sequences $x = \{x_k\}_{k \in \mathbb{Z}}$, and let $A = (A(j,k))_{j,k \in \mathbb{Z}}$ be an infinite matrix. Also, let $|x|_2$ be the ℓ^2-vector norm of x, and $\|A\|$ be the operation norm of A: $\|A\| = \sup_{x : |x|_2 = 1} |Ax|_2$. Fixing positive constants λ, M and c_0, we define the class of matrices

$$\Theta_\infty(\lambda, c_0, M) = \Big\{ A = (A(j,k))_{j,k \in \mathbb{Z}} : |A(j,k)| \leq \frac{M}{(1 + |j-k|)^\lambda}, \|A\| \geq c_0 \Big\}. \tag{A.1}$$

LEMMA A.2. *Fix $\lambda > 1$, $c_0 > 0$, and $M > 0$. For any matrix $A \in \Theta_\infty(\lambda, M)$, there is a constant $C > 0$, depending only on λ, M and c_0, such that $|A^{-1}(j,k)| \leq C \cdot (1 + |j-k|)^{-\lambda}$.*

Next we consider the first claim in Lemma A.1. Construct an infinite matrix Σ_∞ by arranging the finite matrices along the diagonal, and note that the inverse of Σ_∞ is the matrix formed by arranging the inverse of the finite matrices along the diagonal. Since $\Sigma_\infty(i,j) \le M(1+|i-j|^\lambda)^{-1}$, then applying Lemma A.2 gives the claim.

Consider the second claim. It suffices to show that $|U_n(k,j)| \le C/(1+|k-j|^\lambda)$ for all $1 \le j < k \le n$. Denote the first $k \times k$ main diagonal sub-matrix of Σ_n by $\Sigma_{(k)}$, the kth row of $\Sigma_{(k)}$ by $(\xi'_{k-1}, 1)$, and the kth row of U_n by u'_k. It follows from direct calculations that

$$u'_k = \bigl(1 - \xi'_{k-1}\Sigma^{-1}_{(k-1)}\xi_{k-1}\bigr)^{-1/2} \cdot \bigl(\xi'_{k-1}\Sigma^{-1}_{(k-1)}, 1\bigr). \tag{A.2}$$

At the same time, by (A.2) and basic algebra,

$$\bigl(1 - \xi'_{k-1}\Sigma^{-1}_{(k)}\xi_{k-1}\bigr)^{-1} \le u'_k u_k = \Sigma^{-1}_{(k)}(k,k). \tag{A.3}$$

Combining (A.2) and (A.3) gives

$$|U_n(k,j)| = |u_k(j)| \le C\bigl|\bigl(\Sigma^{-1}_{(k-1)}\xi_{k-1}\bigr)_j\bigr|, \qquad 1 \le j \le k-1. \tag{A.4}$$

Now, by Lemma A.2, $|\Sigma^{-1}_{(k-1)}(j,s)| \le C(1+|j-s|^\lambda)^{-1}$ for all $1 \le i, j \le k-1$. Note that $|\xi_{k-1}(s)| \le C(1+|s-k|^\lambda)^{-1}$, $1 \le s \le n$ and $\lambda > 1$. It follows from basic algebra that

$$\bigl|\bigl(\Sigma^{-1}_{(k-1)}\xi_{k-1}\bigr)_j\bigr| \le \sum_{s=1}^n \frac{C}{(1+|j-s|^\lambda)(1+|s-k|^\lambda)} \le \frac{C}{1+|k-j|^\lambda}. \tag{A.5}$$

Inserting (A.5) into (A.4) gives the claim. □

A.2. Statement and proof of Lemma A.3.

LEMMA A.3. *Fix $c_0 > 0$, $\lambda \ge 1$, and $M > 0$. Consider a sequence of bandwidths b_n that tends to infinity. Let $\{\ell_1, \ell_2, \dots, \ell_m\}$ be the m random locations of signals in μ, arranged in the ascending order. For sufficiently large n, there is a constant $C = C(c_0, \lambda, M)$ such that, except for an event with asymptotically vanishing probability,*

$$(V_n(b_n)\mu)_k \ge \bigl(1 - Cb_n^{1/2-\lambda} + o(1)\bigr) \cdot \sqrt{\Sigma_n^{-1}(k,k)} \cdot A_n \qquad \forall k \in \{\ell_1, \ell_2, \dots, \ell_m\},$$

for all $\Sigma_n \in \Theta_n^*(\lambda, c_0, M)$, where $o(1)$ tends to zero algebraically fast.

PROOF. To derive the lemma, note that we may assume without loss of generality that $\ell_1 < \ell_2 < \cdots < \ell_m$. By Lemma A.8, except for an event with negligible probability, $\ell_1 \ge b_n$, $\ell_m \le n - b_n$, and the inter-ℓ_j distances are not less than

$C \log n \cdot n^{2\beta-1}$. For any $k \in \{\ell_1, \ell_2, \dots, \ell_m\}$, let $d_k = (\sum_{j=k}^{k+b_n-1} u_{jk}^2)^{-1/2}$. By the way $\bar{U}(b_n)$ is defined,

$$(\text{A.6}) \qquad \begin{aligned}(\bar{U}'(b_n)U\mu)_k &= d_k \sum_{s,j=1}^{n} \tilde{u}_{ks} u_{sj} \mu_j \\ &= d_k \Bigg[\sum_{s,j=1}^{n} u_{ks} u_{sj} \mu_j - \sum_{s,j=1}^{n} (u_{ks} - \tilde{u}_{ks}) u_{sj} \mu_j \Bigg].\end{aligned}$$

Consider d_k first. Write

$$1/d_k^2 = \sum_{j=k}^{k+b_n-1} u_{jk}^2 = \sum_{j=k}^{n} u_{jk}^2 - \sum_{j=k-b_n}^{n} u_{jk}^2.$$

First, $U'U = \Sigma^{-1}$, $\sum_{j=k}^{n} u_{jk}^2 = (U'U)(k,k) = (\Sigma^{-1})(k,k)$. Second, by the polynomial off-diagonal decay of U and basic calculus,

$$\sum_{j=k+b_n}^{n} u_{jk}^2 \le C \sum_{j=k+b_n}^{n} \frac{1}{1 + |j-k|^{\lambda}} \le C b_n^{1-2\lambda}.$$

Last, note that the quantities $\Sigma^{-1}(k,k)$ are uniformly bounded away from zero and infinity. Combining these results gives

$$(\text{A.7}) \qquad \left| d_k - \sqrt{\Sigma^{-1}(k,k)} \right| \le C b_n^{1-2\lambda}.$$

Consider $\sum_{s,j=1}^{n} u_{ks} u_{sj} \mu_j$ next. Recall that $\mu_j = A_n$ when $j \in \{\ell_1, \ell_2, \dots, \ell_m\}$ and $\mu_j = 0$ otherwise. Since $U'U = \Sigma^{-1}$,

$$\sum_{s,j=1}^{n} u_{ks} u_{sj} \mu_j = \sum_{j=1}^{n} (\Sigma^{-1})(k,j) \mu_j = A_n \Sigma^{-1}(k,k) + A_n \sum_{\ell_s \ne k} \Sigma^{-1}(k, \ell_s).$$

Define $L_n = n^{\beta - 1/2}$. By Lemma A.8, except for an event with negligible probability, the inter-distance of ℓ_j is no less than L_n. So by the polynomial off-diagonal decay of Σ^{-1}, the second term is algebraically small. Therefore,

$$(\text{A.8}) \qquad \sum_{s,j=1}^{n} u_{ks} u_{sj} \mu_j = A_n [(\Sigma^{-1})(k,k) + o(b_n^{1-\lambda})].$$

Last, we consider $\sum_{s,j=1}^{n} (u_{ks} - \tilde{u}_{ks}) u_{sj} \mu_j$. Direct calculations show that

$$\left| ((U - \tilde{U})'U)(k,j) \right| \le \begin{cases} \dfrac{C}{1 + |k-j|^{\lambda}}, & \lambda > 1, \\[2ex] \dfrac{C \log n}{1 + |k-j|^{\lambda}}, & \lambda = 1, \end{cases}$$

so by a similar argument,

$$\left|\sum_{s,j=1}^{n}(u_{ks}-\tilde{u}_{ks})u_{sj}\mu_j\right| = \left|\sum_{j=1}^{n}\bigl((U-\tilde{U})'U\bigr)(k,j)\mu_j\right|$$
$$\le A_n\cdot\bigl((U-\tilde{U})'U\bigr)(k,k)+o(1),$$

where $o(1)$ is algebraically small. Moreover, by the inequality,

$$\bigl((U-\tilde{U})'U\bigr)(k,k)\le\sum_{s=1}^{n}|(u_{ks}-\tilde{u}_{ks})u_{sk}|\le b_n^{1/2-\lambda}$$

and the claim follows. □

A.3. Statement and proof of Lemma A.4. Let $p_1,\dots,p_N$ be N independent and identically distributed data from $U(0,1)$, and $F_N(t)$ be the empirical cdf. The normalized uniform stochastic process is defined as

$$\mathbb{W}_N(t)=\sqrt{N}[F_N(t)-t]/\sqrt{t(1-t)}.$$

LEMMA A.4. *There is a generic constant $C>0$ such that for sufficiently large n,*

$$P\Bigl\{\sup_{1/n\le t\le 1/2}|\mathbb{W}_N(t)|\ge C(\log n)^{3/2}\Bigr\}\le Cn^{-C}.$$

PROOF. To derive this result, note that by the Hungarian construction [14], there is a Brownian bridge $\mathbb{B}(t)$ such that

$$P\Bigl\{\sup_{1/n\le t\le 1/2}\bigl|\sqrt{N}\bigl(F_N(t)-t\bigr)-\mathbb{B}(t)\bigr|\ge\frac{C(\log N+x)}{\sqrt{N}}\Bigr\}\le Ce^{-Cx},$$

where $C>0$ are generic constants. Noting that $1/\sqrt{t(1-t)}\le\sqrt{n}\le C\sqrt{N\log N}$ when $1/n\le t\le 1/2$, it follows that

$$P\Bigl\{\sup_{1/n\le t\le 1/2}\left|\frac{\sqrt{N}(F_N(t)-t)-\mathbb{B}(t)}{\sqrt{t(1-t)}}\right|\ge C(\log N)^{1/2}(\log N+x)\Bigr\}$$
$$\le Ce^{-Cx}. \tag{A.9}$$

At the same time, by [49], page 446,

$$P\Bigl\{\sup_{1/n\le t\le 1/2}\left|\frac{\mathbb{B}(t)}{\sqrt{t(1-t)}}\right|\ge C(\log N)^{1/2}x\Bigr\}\le C\log N\cdot e^{-Cx}. \tag{A.10}$$

Combining (A.9) and (A.10), taking $x=C\log N$ and using the triangle inequality, we deduce the lemma. □

A.4. Statement and proof of Lemma A.5.

LEMMA A.5. *Take the bandwidth to be $b_n = \log n$ and suppose H_0 is true. Then, except for an algebraically small probability,* $\mathrm{iHC}_n^*(b_n) \le C(\log n)^{3/2}$ *for some constant $C > 0$, uniformly for all correlation matrices.*

PROOF. To derive the lemma, note that we may assume without loss of generality that n is divisible by $2b_n - 1$, and let $N = N(n, b_n) = n/(2b_n - 1)$ in Lemma A.4. Define $Y = \bar{U}'UX$. Under the null hypothesis, $Y \sim \mathrm{N}(0, \bar{U}'\bar{U})$ and the coordinates Y_k are block-wise dependent with a bandwidth $\le 2b_n - 1$. Split the Y_k's into $2b_n - 1$ different subsets $\Omega_j = \{Y_k : k \equiv j \bmod (2b_n - 1)\}$, $1 \le j \le 2b_n - 1$. Note that the Y_k's in each subset are independent, and that $|\Omega_j| = N$, $1 \le j \le 2b_n - 1$.

Let $\bar{F}_n(t)$ and $\bar{F}_0(t)$ be as in the proof of Theorem 4.1, and let

$$\bar{F}_{n,j} = \frac{2b_n - 1}{n} \sum_{k=1}^{n} 1_{\{Y_k^2 \ge t, Y_k \in \Omega_j\}}, \qquad 1 \le j \le 2b_n - 1.$$

Note that $\bar{F}_n(t) = \frac{1}{2b_n - 1} \sum_{j=1}^{2b_n - 1} \bar{F}_{n,j}(t)$. By arguments similar to that of Donoho and Jin [18] and basic algebra, it follows that

$$\mathrm{iHC}_n^* = \sup_t \frac{\sqrt{n}(\bar{F}_n(t) - \bar{F}_0(t))}{\sqrt{(2b_n - 1)\bar{F}_0(t)F_0(t)}} \le \sum_{j=1}^{2b_n - 1} \sup_t \frac{\sqrt{N}(\bar{F}_{n,j}(t) - \bar{F}_0(t))}{\sqrt{\bar{F}_0(t)F_0(t)}},$$

and so for any $x > 0$,

$$P\{\mathrm{iHC}_n^* \ge x\} \le \sum_{j=1}^{2b_n - 1} P\left\{\sup_t \frac{\sqrt{N}(\bar{F}_{n,j}(t) - \bar{F}_0(t))}{\sqrt{\bar{F}_0(t)F_0(t)}} \ge x\right\}.$$

Finally, since $\bar{F}_{n,j}$'s are the empirical survival functions of N independent samples from $\chi_1^2(0)$, then

$$\sup_{t : 1/n \le \bar{F}_0(t) \le 1/2} \frac{\sqrt{N}(\bar{F}_{n,j}(t) - \bar{F}_0(t))}{\sqrt{\bar{F}_0(t)F_0(t)}} = \sup_{1/n \le t \le 1/2} \mathbb{W}_N(t) \text{ in distribution.}$$

Therefore,

$$P\{\mathrm{iHC}_n^* \ge x\} \le (2b_n - 1) P\left\{\sup_{1/n \le t \le 1/2} \mathbb{W}_N(t) \ge x\right\}.$$

Taking $x = C(\log n)^{3/2}$, the claim follows from Lemma A.4. $\square$

A.5. Statement and proof of Lemma A.6.

LEMMA A.6. *The spectral norm of $\tilde{\Sigma}_n - \bar{\Sigma}_n$ tends to zero as n tends to infinity.*

PROOF. To establish the lemma, note that by direct calculations and the way $\tilde{\Sigma}$ is defined, we have

$$\tilde{\Sigma} = \begin{pmatrix} \Sigma^* & \xi_{n-1} \\ \xi'_{n-1} & 1 \end{pmatrix}, \tag{A.11}$$

where

$$\begin{aligned} \xi'_{n-1} = \sqrt{2n^{-\alpha}} \times \bigl(0, \ldots, n^{\alpha_0} - k_0(n)^\alpha, \\ k_0(n)^\alpha - \bigl(k_0(n) - 1\bigr)^\alpha, \ldots, 2^\alpha - 1, 1\bigr), \end{aligned} \tag{A.12}$$

and Σ^* is a symmetric matrix with unit diagonal entries and with the following on the kth sub-diagonal:

$$\frac{1}{2} \cdot \begin{cases} 2k^\alpha - (k+1)^\alpha - (k-1)^\alpha, & k \le k_0(n) - 1, \\ 1 + \bigl((k-1)^\alpha - 2k^\alpha\bigr)/n^{\alpha_0} = O(n^{-\alpha_0/\alpha}), & k = k_0(n), \\ -\bigl(1 - (k-1)^\alpha/n^{\alpha_0}\bigr) = O(n^{-\alpha_0/\alpha}), & k = k_0(n) + 1, \\ 0, & k \ge k_0(n) + 2. \end{cases}$$

Note that $\Sigma_{n-1}(g_0)$ and Σ^* share the $2k_0(n) - 1$ sub-diagonals that are closest to the main diagonal (including the main diagonal). Let H_1 be the matrix containing all other sub-diagonals of $\Sigma_{n-1}(g_0)$, and let H_2 be the matrix which contains the $k_0(n)$th and the $(k_0(n) + 1)$th diagonals (upper and lower) of Σ^*. It is seen that

$$\tilde{\Sigma} - \bar{\Sigma} = \begin{pmatrix} H_1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} H_2 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & \xi'_{n-1} \\ \xi'_{n-1} & 0 \end{pmatrix} \equiv B_1 + B_2 + B_3.$$

Let $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the ℓ^1 matrix norm and the ℓ^2 matrix norm, respectively. First, by direct calculations, since $\alpha < 1/2$, $\|B_1 + B_2\|_1 \le Cn^{\alpha_0(\alpha-1)/\alpha} \le Cn^{-\alpha_0}$. At the same time, by (A.12) and since $\alpha < 1/2$,

$$\|B_3\|^2 \le \frac{C}{n^{\alpha_0}} \sum_{k=1}^{n} [k^\alpha - (k+1)^\alpha]^2 \le \frac{C}{n^{\alpha_0}} \sum_{k=1}^{n} k^{2\alpha - 2} \le C/n^{\alpha_0}.$$

Since the spectral norm is no greater than the ℓ^1-matrix norm and the ℓ^2-matrix norm, the spectral norm of $B_1 + B_2 + B_3$ is no greater than $Cn^{-\alpha_0/2}$, and the claim follows. □

A.6. Statement and proof of Lemma A.7.

LEMMA A.7. *Fix $\beta \in (\frac{1}{2}, 1)$, $r \in (0, 1)$ and $\delta \in (0, 1)$ such that $\bar{\gamma}_0(1 - \delta)^{-2} r < \rho^*(\beta)$. As n tends to infinity the Hellinger distance associated with model (10.2) tends to zero.*

PROOF. To derive the lemma, let $a = \sqrt{(1-\delta)/\bar{\gamma}_0}$, $r' = \bar{\gamma}_0(1-\delta)^{-2}r$, $U_1 = a\tilde{U}$, and $\tilde{\tilde{\mu}} = \frac{1}{a}\tilde{\mu}$. Model (10.2) can be equivalently written as

$$X = \tilde{U}\tilde{\mu} + Z = U_1\tilde{\tilde{\mu}} + Z \qquad \text{where } Z \sim \mathrm{N}(0, I_n). \tag{A.13}$$

Using the argument in the first paragraph of the proof of Theorem 3.1 it is not difficult to verify that (I) $\tilde{\tilde{\mu}}$ has $m = n^{1-\beta}$ nonzero entries; each of which is equal to $\sqrt{2r'\log n}$ with $r' < \rho^*(\beta)$, and whose locations are randomly sampled from $(1, 2, \ldots, n)$; (II) U_1, where $U_1(k, j) = 0$ if $|k-j| > (\log n)^2$, is a banded lower triangular matrix and (III) $\overline{\lim}_{n\to\infty} \max_{\sqrt{n}\le k\le n-\sqrt{n}} (U_1'U_1)(k,k) = (1-\delta) < 1$.

Below, write $\mu = \tilde{\tilde{\mu}}$ and $r = r'$ for short. Note that the Hellinger distance associated with model (10.2) is $E_0(\sqrt{W_n^*})$, where E_0 denotes the law $Z \sim \mathrm{N}(0, I_n)$, and

$$W_n^* = W_n^*(r, \beta; Z_1, Z_2, \ldots, Z_n) = \binom{n}{m}^{-1} \sum_{\ell=(\ell_1,\ell_2,\ldots,\ell_m)} e^{\mu_\ell' U_1' Z - \|U_1\mu_\ell\|^2/2}.$$

Introduce the set of indices

$$\begin{aligned} S_n = \Big\{&\ell = (\ell_1, \ell_2, \ldots, \ell_m), \\ &\min_{1\le j\le m-1} |\ell_{j+1} - \ell_j| \ge 3(\log n)^2, \ell_1 \ge \sqrt{n}, n - \ell_m \ge \sqrt{n}\Big\}. \end{aligned} \tag{A.14}$$

The following lemma is proved in Section A.7.

LEMMA A.8. *Let $\ell_1 < \ell_2 < \cdots < \ell_m$ be m distinct indices randomly sampled from $(1, 2, \ldots, n)$ without replacement. Then for any $1 \le K \le n$, (a) $P\{\ell_1 \le K\} \le Km/n$, (b) $P\{\ell_m \ge n - K\} \le Km/n$ and (c) $P\{\min_{1\le i\le m-1}\{|\ell_{i+1} - \ell_i| \le K\} \le Km(m+1)/n$. As a result, $P\{\ell = (\ell_1, \ell_2, \ldots, \ell_m) \notin S_n\} = O\{(\log n)^2 n^{1-2\beta}\} = o(1)$.*

Applying Lemma A.8, we make only a negligible difference by restricting ℓ to S_n and defining

$$W_n = \frac{1}{\binom{n}{m}} \sum_{\ell=(\ell_1,\ell_2,\ldots,\ell_m)\in S_n} e^{\mu_\ell' U_1' Z - \|U\mu_\ell\|^2/2}, \tag{A.15}$$

in which case

$$E(W_n^{1/2}) = E(W_n^{*1/2}) + o(1). \tag{A.16}$$

Define $Y = U_1'Z$,

$$\sigma_j^2 = \mathrm{var}(Y_j) \equiv (U_1'U_1)(j, j), \qquad 1 \le j \le n, \tag{A.17}$$

and the event

$$D_n = \{Y_j/\sigma_j \le \sqrt{2\log n}, 1 \le j \le n\}.$$

By direct calculation, $P\{D_n^c\} = o(1)$, and so by Hölder's inequality,

$$E\big(W_n^{1/2} 1_{\{D_n^c\}}\big) = E(W_n^{1/2}) + o(1).$$

Combining this result and (A.16) we deduce that $E(W_n^{*1/2}) = E(W_n^{1/2} 1_{\{D_n\}}) + o(1)$, and comparing this property with the desired result we see that it is is sufficient to show that

$$E\big(W_n^{1/2} 1_{\{D_n\}}\big) = 1 + o(1). \tag{A.18}$$

The key to (A.18) is the following lemma, which is proved in Section A.8.

LEMMA A.9. *Consider the model* (A.13) *where* U_1 *and* μ *satisfy* (I)–(III). *As* $n \to \infty$, $E(W_n 1_{\{D_n\}}) = 1 + o(1)$, *and* $E(W_n^2 1_{\{D_n\}}) = 1 + o(1)$.

Since

$$\big|W_n^{1/2} 1_{\{D_n\}} - 1\big| \le \frac{|W_n 1_{\{D_n\}} - 1|}{1 + W_n^{1/2} 1_{\{D_n\}}} \le \big|W_n 1_{\{D_n\}} - 1\big|,$$

then by Hölder's inequality,

$$\begin{aligned} \big(E\big|W_n^{1/2} 1_{\{D_n\}} - 1\big|\big)^2 &\le \big|W_n 1_{\{D_n\}} - 1\big|^2 \\ &= E\big(W_n^2 1_{\{D_n\}}\big) - 2E\big(W_n 1_{\{D_n\}}\big) + 1. \end{aligned} \tag{A.19}$$

Combining (A.19) with Lemma A.9 gives (A.18). □

A.7. Proof of Lemma A.8. The last claim follows once (a)–(c) are proved. Consider (a)–(b) first. Fixing $K \ge 1$, we have

$$P\{\ell_1 = K\} = \frac{\binom{n-K}{m-1}}{\binom{n}{m}} = m \frac{(n-m)(n-m-1)\cdots(n-m-K+2)}{n(n-1)\cdots(n-K+1)} \le m/n,$$

so $P\{\ell_1 \le K\} \le Km/n$. Similarly, $P\{n - \ell_m \le K\} \le Km/n$. This gives (a) and (b).

Next we prove (c). Denote the minimum inter-distance of $\ell_1, \ell_2, \ldots, \ell_m$ by

$$L(\ell) = L(\ell; m, n) = \min_{1 \le i \le m-1} |\ell_{i+1} - \ell_i|.$$

Note that

$$P\{L(\ell) = K\} \le \sum_{j=1}^{m-1} P\{\ell_{j+1} - \ell_j = K\} \le \sum_{j=1}^{m-1} \sum_{k=1}^{n} P\{\ell_j = k, \ell_{j+1} = k + K\}.$$

Writing $P\{\ell_j = k, \ell_{j+1} = k + K\} = \binom{n}{m}^{-1}\binom{k-1}{j-1}\binom{n-k-K}{m-j-1}$, we have

$$P\{L(\ell) = K\} \le \frac{1}{\binom{n}{m}} \sum_{j=1}^{m-1} \sum_{k=j}^{n} \binom{k-1}{j-1} \binom{n-k-K}{m-j-1}$$
$$= \frac{1}{\binom{n}{m}} \sum_{k=1}^{n} \sum_{j=1}^{k} \binom{k-1}{j-1} \binom{n-k-K}{m-j-1},$$

where the last term is no greater than

$$\frac{1}{\binom{n}{m}} \sum_{k=1}^{n} \binom{n-K-1}{m-2} \le \frac{n}{\binom{n-2}{m}} \binom{n}{m-2} \le m^2/n.$$

The claim follows.

A.8. Proof of Lemma A.9. We need the following lemma, proved in Section A.9.

LEMMA A.10. *Consider a bivariate zero mean normal variable $(X, Y)'$ that satisfies* $\operatorname{Var}(X) = \sigma_1^2$, $\operatorname{Var}(Y) = \sigma_2^2$ *and* $\operatorname{corr}(X, Y) = \varrho$, *where $c_0 \le \sigma_1, \sigma_2 \le 1$ for some constant $c_0 \in (0, 1)$. Then there is a constant $C > 0$ such that, for sufficiently large n,*

$$E\big[\exp(A_n X - \sigma_1^2 A_n^2/2) \cdot 1_{\{Y > \sigma_2 T_n\}}\big] \le C \cdot n^{-(1-\varrho\sqrt{r})^2}$$
$$\le C n^{-(1-\sqrt{r})^2},$$

$$E\left[\exp\left(A_n(X + Y) - \frac{\sigma_1^2 + \sigma_2^2}{2} A_n^2\right) \cdot 1_{\{X \le \sigma_1 T_n, Y \le \sigma_2 T_n\}}\right] \le C n^{-d(r)},$$

where $d(r) = \min\{2r, 1 - 2(1 - \sqrt{r})^2\}$.

Now we proceed with the derivation of Lemma A.9. Consider the first claim. Note that for any $\ell = (\ell_1, \ell_2, \dots, \ell_m) \in S_n$, the minimum inter-distance of ℓ_i is no less than $3(\log n)^2$, and so

$$\|U_1 \mu_\ell\|^2 = A_n^2 \sum_{i=1}^{m} (U_1' U_1)(\ell_i, \ell_i) = A_n^2 \sum_{i=1}^{m} \sigma_{\ell_i}^2.$$

In view of the definition of Y_j and σ_j [see (A.17)], we can rewrite W_n as

$$W_n = \frac{1}{\binom{n}{m}} \sum_{\ell = (\ell_1, \ell_2, \dots, \ell_m) \in S_n} \exp\left(A_n \sum_{i=1}^{m} Y_{\ell_i} - \frac{A_n^2}{2} \sum_{i=1}^{m} \sigma_{\ell_i}^2\right). \tag{A.20}$$

Note that

$$1_{\{D_n^c\}} \le \sum_{j=1}^{n} 1_{\{Y_j/\sigma_j > T_n\}}. \tag{A.21}$$

Combining (A.20) and (A.21) gives

$$\begin{aligned} &E\big(W_n \cdot 1_{\{D_n^c\}}\big) \\ &\quad \le \frac{1}{\binom{n}{m}} \sum_{\ell=(\ell_1,\ldots,\ell_m)\in S_n} \sum_{k=1}^{n} E\Bigg[\exp\Bigg(A_n \sum_{j=1}^{m} Y_{\ell_j} - \frac{A_n^2}{2}\sum_{j=1}^{m}\sigma_{\ell_j}^2\Bigg) \\ &\qquad\qquad \times 1_{\{Y_k/\sigma_k > T_n\}}\Bigg]. \end{aligned} \tag{A.22}$$

We shall say that two indices j and k are near each other if $|j-k| \le (\log n)^2$. In this notation, for each $1 \le k \le n$, when k is near one ℓ_j, say ℓ_{j_0}, Y_k must be independent of all other Y_{ℓ_j} with $j \ne j_0$. It follows that

$$\begin{aligned} &E\Bigg[\exp\Bigg(A_n \sum_{j=1}^{m} Y_{\ell_j} - \frac{A_n^2}{2}\sum_{j=1}^{m}\sigma_{\ell_j}^2\Bigg) \cdot 1_{\{Y_k/\sigma_k > T_n\}}\Bigg] \\ &\qquad = E\big[\exp(A_n Y_{\ell_{j_0}} - \sigma_{j_0}^2 A_n^2/2) \cdot 1_{\{Y_k/\sigma_k > T_n\}}\big]. \end{aligned}$$

By Lemma A.10, the right-hand side is no greater than $Cn^{-(1-\sqrt{r})^2}$. Therefore,

$$E\Bigg[\exp\Bigg(A_n \sum_{j=1}^{m} Y_{\ell_j} - \frac{A_n^2}{2}\sum_{j=1}^{m}\sigma_{\ell_j}^2\Bigg) \cdot 1_{\{Y_k/\sigma_k > T_n\}}\Bigg] \le Cn^{-(1-\sqrt{r})^2}. \tag{A.23}$$

Moreover, for each fixed $\ell = (\ell_1, \ldots, \ell_m) \in S_n$, there are at most $2m(\log n)^2$ different indices k that can be near some of the ℓ_j's; and when they are, they can be near only one such ℓ_j. Combining these results gives

$$\begin{aligned} E\big[W_n \cdot 1_{\{D_n^c\}}\big] &\le \frac{1}{\binom{n}{m}} \sum_{\ell=(\ell_1,\ldots,\ell_m)\in S_n} C(\log n)^2 m n^{-(1-\sqrt{r})^2} \\ &\le C(\log n)^2 n^{(1-\beta)-(1-\sqrt{r})^2}. \end{aligned} \tag{A.24}$$

By the definition of $\rho^*(\beta)$ and the assumption of the lemma, $r < \rho^*(\beta) \le (1 - \sqrt{1-\beta})^2$, and so the first claim follows directly from (A.24).

We now consider the second claim. Fix $0 \le N \le m$, and let $\tilde{S}_N(\ell)$ denote the set of all $k = (k_1, k_2, \ldots, k_m) \in S_n$ such that there are exactly N k_j's that are near to one ℓ_i. (Clearly, any k_j can be near to at most one ℓ_i.) The two sets of indices $(\ell_1, \ell_2, \ldots, \ell_m)$ and $(k_1, k_2, \ldots, k_m)$ form exactly N pairs where each contains one candidate from the first set and one candidate from the second. These pairs are not near to each other and not near to any remaining indices outside the pairs. Using (A.20), we write

$$
\begin{aligned}
E\bigl[W_n^2 \cdot 1_{\{D_n\}}\bigr]
&= \binom{n}{m}^{-2} \sum_{\ell=(\ell_1,\ell_2,\dots,\ell_m)\in S_n} \\
&\quad \times \sum_{N=0}^{m} \sum_{k=(k_1,k_2,\dots,k_m)\in \tilde{S}_N(\ell)} E\Biggl[\exp\Biggl(A_n \sum_{i=1}^{m}(Y_{\ell_i}+Y_{k_i}) \\
&\qquad - \frac{A_n^2}{2}\sum_{i=1}^{m}(\sigma_{\ell_i}^2+\sigma_{k_i}^2)\Biggr)\cdot 1_{\{D_n\}}\Biggr].
\end{aligned}
\tag{A.25}
$$

For any fixed ℓ and $k \in \tilde{S}_N(\ell)$, by symmetry, and without loss of generality, we suppose the N pairs are (ℓ_1, k_1), $(\ell_2, k_2), \dots, (\ell_N, k_N)$. By independence of the pairs with other indices, and also by independence among the pairs,

$$
\begin{aligned}
&E\Biggl[\exp\Biggl(A_n \sum_{j=1}^{m}(Y_{\ell_j}+Y_{k_j}) - \frac{A_n^2}{2}\sum_{j=1}^{m}(\sigma_{\ell_j}^2+\sigma_{k_j}^2)\Biggr)\cdot 1_{\{D_n\}}\Biggr] \\
&\quad \le E\Biggl[\exp\Biggl(A_n \sum_{j=1}^{m}(Y_{\ell_j}+Y_{k_j}) - \frac{A_n^2}{2}\sum_{j=1}^{m}(\sigma_{\ell_j}^2+\sigma_{k_j}^2)\Biggr) \\
&\qquad\qquad \times 1_{\{Y_{\ell_j}/\sigma_{\ell_j}\le T_n,\, Y_{k_j}/\sigma_{k_j}\le T_n,\ \text{for all } 1\le j\le N\}}\Biggr] \\
&\quad \le E\Biggl[\exp\Biggl(A_n\Biggl\{\sum_{j=1}^{N}(Y_{\ell_j}+Y_{k_j}) - \frac{A_n^2}{2}\sum_{j=1}^{N}(\sigma_{\ell_j}^2+\sigma_{k_j}^2)\Biggr\}\Biggr) \\
&\qquad\qquad \times 1_{\{Y_{\ell_j}/\sigma_{\ell_j}\le T_n,\, Y_{k_j}/\sigma_{k_j}\le T_n,\ \text{for all } 1\le j\le N\}}\Biggr] \\
&\quad = \prod_{j=1}^{N}\Biggl(E\Biggl[\exp\Biggl\{A_n(Y_{\ell_j}+Y_{k_j}) - \frac{A_n^2}{2}(\sigma_{\ell_j}^2+\sigma_{k_j}^2)\Biggr\} \\
&\qquad\qquad \times 1_{\{Y_{\ell_j}/\sigma_{\ell_j}\le T_n,\, Y_{k_j}/\sigma_{k_j}\le T_n\}}\Biggr]\Biggr).
\end{aligned}
\tag{A.26}
$$

Here, in the first inequality, we have used the fact that

$$1_{\{D_n\}} \le 1_{\{Y_{\ell_j}/\sigma_{\ell_j}\le T_n,\, Y_{k_j}/\sigma_{k_j}\le T_n,\ \text{for all } 1\le j\le N\}};$$

in the second inequality, we have utilized the independence and the fact that

$$E[\exp(A_n Y_j - \sigma_j^2 A_n^2/2)] = 1 \qquad \text{for all } j = 1, \dots, n,$$

and in the third equality, we have used again the independence. Moreover, in view of the definition of U_1, and Lemma A.1, there is a constant $c_0 \in (0,1)$ such that $\sigma_j \in [c_0, 1]$. Using Lemma A.10, for sufficiently large n and each $1 \le j \le N$,

$$E\biggl[\exp\biggl(A_n(Y_{\ell_j}+Y_{k_j}) - \frac{A_n^2}{2}(\sigma_{\ell_j}^2+\sigma_{k_j}^2)\biggr) \times 1_{\{Y_{\ell_j}/\sigma_{\ell_j}\le T_n, Y_{k_j}/\sigma_{k_j}\le T_n\}}\biggr] \le Cn^{d(r)} \tag{A.27}$$

with $d(r)$ being as in Lemma A.10. Combining (A.26) and (A.27) gives

$$E[W_n^2 \cdot 1_{\{D_n\}}] \le \binom{n}{m}^{-2} \sum_{\ell=(\ell_1,\dots,\ell_m)} \sum_{N=0}^{m} (Cn^{d(r)})^N |\tilde{S}_N(\ell)|, \tag{A.28}$$

where $|\tilde{S}_N(\ell)|$ denotes the cardinality of $\tilde{S}_N(\ell)$. By elementary combinatorics,

$$\begin{aligned} |\tilde{S}_N(\ell)| &\le \binom{m}{N}(2\log^2 n)^N \binom{n-N}{m-N} \\ &\le (2\log^2 n)^N \binom{m}{N}\binom{n}{m-N}. \end{aligned} \tag{A.29}$$

Direct calculations show that

$$\frac{\binom{m}{N}\binom{n}{m-N}}{\binom{n}{m}} = \frac{1}{N!}\biggl(\frac{m!}{(m-N)!}\biggr)^2 \frac{(n-m)!}{(n-m+N)!} \lesssim \frac{1}{N!}\biggl(\frac{m^2}{n}\biggr)^N. \tag{A.30}$$

Substituting (A.29) and (A.30) into (A.28) and recalling that $m = n^{1-\beta}$, we deduce that

$$\begin{aligned} &E[W_n^2 \cdot 1_{\{D_n\}}] \\ &\quad \le \binom{n}{m}^{-1} \sum_{\ell=(\ell_1,\ell_2,\dots,\ell_m)\in S_n} \sum_{N=0}^{m} \frac{1}{N!}\biggl(\frac{m^2}{n}\biggr)^N ((C\log^2 n)n^{d(r)})^N, \end{aligned} \tag{A.31}$$

where the last term does not exceed $\sum_{N=0}^{\infty}(N!)^{-1}[C(\log^2 n)n^{1+d(r)-2\beta}]^N$. By the assumption of the lemma,

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

thus it can be seen that $1 + d(r) - 2\beta < 0$ for all fixed β and $r \in (0, \rho^*(\beta))$. Combining this with (A.31) gives the second claim.

A.9. Proof of Lemma A.10. Denote the density, cdf and survival function of N(0, 1) by ϕ, Φ and $\bar{\Phi}$. For the first claim, define $W = X/\sigma_1$ and $V = Y/\sigma_2$ if

$\rho \geq 0$ and $V = -Y/\sigma_2$ otherwise. The proofs for two cases $\rho \geq 0$ and $\rho < 0$ are similar, so we only show the first one. In this case, it suffices to show that

$$E\big[\exp(\sigma_1 A_n W - \sigma_1^2 A_n^2/2) \cdot 1_{\{V > T_n\}}\big] \leq C \cdot n^{-(1-\varrho\sqrt{r})^2}.$$

Write $W = (W - \rho V) + \rho V$, and note that $(1-\rho)^2 + \rho^2 \leq 1$. It is seen that

$$\begin{aligned} \sigma_1 A_n W - \sigma_1^2 A_n^2/2 &\leq [\sigma_1 A_n (W - \rho V) - \sigma_1^2 (1-\rho)^2 A_n^2/2] \\ &\quad + [\sigma_1 A_n \rho V - \sigma_1^2 \rho^2 A_n^2/2]. \end{aligned} \tag{A.32}$$

Since W and V have unit variance and correlation ρ, then $(W - \rho V)$ is independent of V and is distributed as $\mathrm{N}(0, (1-\rho)^2)$. Therefore, $E[\exp(\sigma_1 A_n (W - \rho V) - \sigma_1^2 (1-\rho)^2 A_n^2/2)] = 1$. Combining this with (A.32) gives

$$\begin{aligned} &E\big[\exp(\sigma_1 A_n W - \sigma_1^2 A_n^2/2) \cdot 1_{\{V > T_n\}}\big] \\ &\qquad = E\big[\exp(\sigma_1 \rho A_n V - \sigma_1^2 \rho^2 A_n^2/2) \cdot 1_{\{V > T_n\}}\big]. \end{aligned}$$

Now, by direct calculation,

$$E\big[\exp(A_n V - A_n^2/2) \cdot 1_{\{V > T_n\}}\big] = \int_{T_n}^{\infty} \phi(x - \sigma_1 \rho A_n)\, dx = \bar{\Phi}(T_n - \sigma_1 \rho A_n).$$

Since $\bar{\Phi}(x) \leq C\phi(x)$ for all $x > 0$,

$$\bar{\Phi}(T_n - \sigma_1 \rho A_n) \leq C\phi(T_n - \sigma_1 \rho A_n) = Cn^{-(1-\rho\sqrt{r})^2}.$$

Combining these results gives the claim.

We now establish the second claim. By Hölder's inequality, it suffices to show that

$$E\big[\exp(2A_n X - \sigma_1^2 A_n^2) \cdot 1_{\{X \leq \sigma_1 T_n\}}\big] \leq Cn^{-d(r)}.$$

Recalling that $W = X/\sigma_1$, we have

$$E\big[\exp(2A_n X - \sigma_1 A_n^2) \cdot 1_{\{X \leq \sigma_1 T_n\}}\big] = E\big[\exp(2\sigma_1 A_n W - \sigma_1^2 A_n^2) \cdot 1_{\{W \leq T_n\}}\big].$$

By direct calculation,

$$\begin{aligned} E\big[\exp(2\sigma_1 A_n W - \sigma_1^2 A_n^2) \cdot 1_{\{W \leq T_n\}}\big] &= e^{\sigma_1^2 A_n^2} \int_{-\infty}^{T_n} \phi(x - 2\sigma_1 A_n)\, dx \\ &= e^{\sigma_1^2 A_n^2} \Phi(T_n - 2\sigma_1 A_n). \end{aligned}$$

Since $\Phi(x) \leq C\phi(x)$ for all $x < 0$ and $\Phi(x) \leq 1$ for all $x \geq 0$,

$$\begin{aligned} &e^{\sigma_1^2 A_n^2} \Phi(T_n - 2\sigma_1 A_n) \\ &\qquad \leq \begin{cases} Ce^{\sigma_1^2 A_n^2} = Cn^{2\sigma_1^2 r}, & \sigma_1^2 r \leq 1/4, \\ e^{\sigma_1^2 A_n^2} \phi(T_n - 2\sigma_1 A_n) = Cn^{1-2(1-\sigma_1\sqrt{r})^2}, & \sigma_1^2 r > 1/4. \end{cases} \end{aligned}$$

In view of the definition of $d(r)$, $e^{\sigma_1^2 A_n^2} \Phi(T_n - 2\sigma_1 A_n) \leq Cn^{d(\sigma_1^2 r)}$. Since that $\sigma_1 \leq 1$ and that $d(r)$ is a monotonely increasing function, we have $d(\sigma_1^2 r) \leq d(r)$. Combining these results gives the claim.

A.10. Statement and proof of Lemma A.11.

LEMMA A.11. *Under the conditions of Theorem 4.2, the right-hand side of (10.8) converges to zero algebraically fast as n diverges to infinity.*

PROOF. The key observation needed to establish the lemma is that there is a sequence of positive numbers δ_n that tends to zero as n diverges to infinity such that $\nu_k \geq (1-\delta_n)A_n$ for all $k \in \{\ell_1, \ell_2, \ldots, \ell_m\}$, so it is natural to compare model (10.7) with the following model:

$$Y^* = \nu^* + Z, \qquad Z \sim \mathrm{N}(0, I_n), \tag{A.33}$$

where ν^* has m nonzero entries of equal strength $(1-\delta_n)A_n$ whose locations are randomly drawn from $\{1, 2, \ldots, n\}$ without replacement.

For short, write $t = t_n^*$ and

$$h_n(t) = \frac{\sqrt{n}(\bar{F}_n(t) - \bar{F}_0(t))}{\sqrt{(2b_n - 1)\bar{F}_0(t)(1 - \bar{F}_0(t))}}.$$

Let $\bar{F}_n^*(t)$ be the empirical survival function of $\{(Y_k^*)^2\}_{k=1}^n$, and let $\bar{F}(t) = E[\bar{F}_n(t)]$ and $\bar{F}^*(t) = E[\bar{F}_n^*(t)]$. Recall that the family of noncentral χ^2-distributions has monotone likelihood ratio. Then $\bar{F}(t) \geq \bar{F}^*(t) \geq \bar{F}_0(t)$. Now, first, since the Y_k's are block-wise dependent with a block size $\leq 2b_n - 1$, it follows by direct calculations that

$$\mathrm{Var}(h_n(t)) \leq C\bar{F}(t)/\bar{F}_0(t).$$

Second, by $\bar{F}(t) \geq \bar{F}_n^*(t)$,

$$\begin{aligned} E[h_n(t)] &= \frac{\sqrt{n}(\bar{F}(t) - \bar{F}_0(t))}{\sqrt{(2b_n - 1)\bar{F}_0(t)(1 - \bar{F}_0(t))}} \\ &\geq \frac{\sqrt{n}(\bar{F}^*(t) - \bar{F}_0(t))}{\sqrt{(2b_n - 1)\bar{F}_0(t)(1 - \bar{F}_0(t))}}, \end{aligned} \tag{A.34}$$

where the right-hand side diverges to infinity algebraically fast by an argument similar to that in [18]. Combining Chebyshev's inequality, the identity $b_n = \log n$ and calculations of the mean and variance of $h_n(t)$, we deduce that

$$P\{h_n(t) \leq (\log n)^2\} \leq C(\log n)\frac{\bar{F}(t)}{n(\bar{F}(t) - \bar{F}_0(t))^2}. \tag{A.35}$$

It remains to show that the last term in (A.35) is algebraically small. We discuss separately the cases $\bar{F}(t)/\bar{F}_0(t) \geq 2$ and $\bar{F}(t)/\bar{F}_0(t) < 2$. For the first case,

$$\frac{\bar{F}(t)}{n(\bar{F}(t) - \bar{F}_0(t))^2} \leq \frac{C}{n\bar{F}(t)} \leq \frac{C}{n\bar{F}_0(t)},$$

which is algebraically small since $t = \sqrt{2q \log n}$ and $0 < q < 1$. For the second case,

$$\frac{\bar{F}(t)}{n(\bar{F}(t) - \bar{F}_0(t))^2} \le \frac{C\bar{F}_0(t)}{n(\bar{F}(t) - \bar{F}_0(t))^2} \le \frac{C\bar{F}_0(t)}{n(\bar{F}^*(t) - \bar{F}_0(t))^2}, \tag{A.36}$$

which is seen to be algebraically small by comparing it to the right-hand side of (A.34). □

A.11. Statement and proof of Lemma A.12.

LEMMA A.12. *Let Σ_n be as in* (7.1). *For sufficiently large n, necessary and sufficient conditions for Σ_n to be positive definite are*, *respectively*, $0 \le \alpha \le 2$ *and* $0 < \alpha_0 \le \alpha \le 1$.

PROOF. We begin by establishing the first claim. Suppose such an autoregressive structure exists for $\alpha \ge \alpha_0 > 0$. Let

$$Y_k = \sqrt{a_n} \cdot (X_{k+1} - X_k)/d, \qquad a_n = n^{\alpha_0}/2, k = 1, 2, \dots, n-1.$$

Clearly, $\operatorname{var}(Y_k) = 1$. At the same time, direct calculation shows that the correlation between Y_1 and Y_{j+1} equals to $[(j+1)^\alpha + (j-1)^\alpha - 2j^\alpha]/2$ for all $1 \le j \le n-2$, which is no larger than 1. Taking $j = 2$ yields $(3^\alpha + 1 - 2 \cdot 2^\alpha)/2 \le 1$, and hence $\alpha \le 2$.

Consider the second claim. For any $k \ge 1$, define the partial sum $S_k(t) = 1 + 2\sum_{j=1}^k (1 - \frac{j^\alpha}{n^{\alpha_0}})^+ \cos(kt)$. By a well-known result in trigonometry [55], to establish the positive-definiteness of Σ_n, it suffices to show that

$$\begin{aligned} &S_{k_0+1}(t) \ge 0 \qquad \text{for all } t \in [-\pi, \pi] \quad \text{and} \\ &S_{k_0+1}(t) > 0 \qquad \text{except for a set of measure zero.} \end{aligned} \tag{A.37}$$

Here, $k_0 = k_0(n; \alpha, \alpha_0)$ is the largest integer k such that $k^\alpha \le n^{\alpha_0}$.

We now derive (A.37). Using a result from [55], page 183, if we let $a_0 = 2$, and $a_j = 2(1 - \frac{j^\alpha}{n^{\alpha_0}})^+$, $1 \le j \le n-1$, then $S_{k_0+1}(t) = \sum_{j=0}^{k_0-1} (j+1)\Delta^2 a_j K_j(t) + (k_0 + 1)K_{k_0}(t)\Delta a_{k_0} + D_n(t) a_{k_0+1}$. Here, $\Delta a_j = a_j - a_{j+1}$, $\Delta^2 a_j = a_j + a_{j+2} - 2a_{j+1}$, and $D_j(t)$ and $K_j(t)$ are the Dirichlet's kernel and the Fejér's kernel, respectively,

$$\begin{aligned} D_j(t) &= \frac{\sin((j+1/2)t)}{2\sin(t/2)}, \\ K_j(t) &= \frac{2}{j+1}\left(\frac{\sin((j+1)/2t)}{2\sin(t/2)}\right)^2, \qquad j = 0, 1, \dots. \end{aligned} \tag{A.38}$$

In view of the definition of k_0, $a_{k_0+1} = (1 - \frac{(k_0+1)^\alpha}{n^{\alpha_0}})^+ = 0$. Also, by the monotonicity of $\{a_j\}$, $\Delta a_{k_0} = a_{k_0} - a_{k_0+1} \ge 0$. Therefore, $S_{k_0+1}(t) \ge \sum_{j=0}^{k_0-1} (j + 1)\Delta^2 a_j K_j(t)$.

We claim that the sequence $\{a_0, a_1, \dots, a_{n-1}\}$ is convex. In detail, since $\alpha \le 1$, the sequence $\{j^\alpha\}$ is concave. As a result, the sequence $\{(1 - \frac{j^\alpha}{n^{\alpha_0}})\}$ is convex, and so is the sequence $\{(1 - j^\alpha/n^{\alpha_0})^+\}$. In view of the definition of a_j, the claim follows directly. The convexity of the a_j's implies that $\Delta^2 a_j \ge 0$, $0 \le j \le n-2$. Therefore, $S_{k_0+1}(t) \ge 0$. This proves the first part of (A.37).

We now prove the second part of (A.37), and discuss separately the two cases $\alpha < 1$ and $\alpha = 1$. In the first case, $\Delta a_0 = n^{-\alpha_0}(2 - 2^\alpha) > 0$ and $K_0(t) = \frac{1}{2}$. As a result, $S_{k_0+1}(t) \ge (2 - 2^\alpha)/(2n^{\alpha_0}) > 0$, and the claim follows. In the second case, $\Delta a_j = n^{-\alpha_0}(2j - j - (j+2)) = 0$, and $\Delta a_{k_0-1} = [1 - n^{-\alpha_0}(k_0 - 1)] - 2(1 - n^{-\alpha_0}k_0) = n^{-\alpha_0}(k_0+1) - 1 > 0$. Therefore, $S_{k_0+1}(t) \ge (k_0+1)[(k_0+1)n^{-\alpha_0} - 1]K_{k_0}(t)$. Clearly, $S_{k_0+1}(t)$ can only assume 0 when $\frac{1}{2}(k_0+1)t$ is a multiple of π. Since the set of such t has measure zero, the claim follows directly. $\square$

A.12. Statement and proof of Lemma A.13.

LEMMA A.13. *For* $0 < \alpha < 1$, *we have* $\operatorname{essinf}_{-\pi \le \theta \le \pi}\{f_\alpha(\theta)\} > 0$.

PROOF. To derive the lemma, let $a_0 = 2$, and $a_k = 2k^\alpha - (k+1)^\alpha - (k-1)^\alpha$, $1 \le k \le n-1$. Clearly, $a_k > 0$ for all k, so $f_\alpha(0;\alpha) > 0$. Furthermore, when $\theta \ne 0$, by [55], equation 1.7, page 183,

$$f_\alpha(\theta) = \sum_{\nu=0}^{\infty}(\nu+1)[a_{\nu+2} + a_\nu - 2a_{\nu+1}]a_\nu K_\nu(\theta), \tag{A.39}$$

where $K_\nu(\theta)$ is the Fejér's kernel as in (A.38). By the positiveness of the Fejér's kernel, all remains to show is that $a_{k+1} + a_{k-1} - 2a_k > 0$, for all $k \ge 2$.

Define $h(x) = (1+2x)^\alpha + (1-2x)^\alpha - 4(1+x)^\alpha - 4(1-x)^\alpha + 6$, $0 \le x \le 1/2$. By direct calculations, for all $k \ge 2$,

$$\begin{aligned} &a_{k+1} + a_{k-1} - 2a_k \\ &\quad = -k^\alpha\left[\left(1+\frac{2}{k}\right)^\alpha + \left(1-\frac{2}{k}\right)^\alpha - 4\left(1+\frac{1}{k}\right)^\alpha - 4\left(1-\frac{1}{k}\right)^\alpha + 6\right] \\ &\quad = -k^\alpha h(1/k). \end{aligned} \tag{A.40}$$

Also, by basic calculus,

$$h''(x) = 4\alpha(\alpha-1)[(1+2x)^{\alpha-2} + (1-2x)^{\alpha-2} - (1+x)^{\alpha-2} - (1-x)^{\alpha-2}].$$

Since $0 < \alpha < 1$, $x^{\alpha-2}$ is a convex function. It follows that $h''(x) < 0$ for all $x \in (0, 1/2)$, and $h(x)$ is a strictly concave function. At the same time, note that $h(0) = h'(0) = 0$, so $h(x) < 0$ for $x \in (0, 1/2]$. Combining this with (A.40) gives the claim. $\square$

Acknowledgments. Jiashun Jin would like to thank Christopher Genovese and Larry Wasserman for extensive discussion, and Aad van der Vaart for help on the proof of (3.2). He would also like to thank Peter Bickel, Emmanuel Candés, David Donoho, Karlheinz Gröchenig, Michael Leinert, Joel Tropp and Zepu Zhang for encouragement and pointers.

REFERENCES

[1] ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. and JOHNSTONE, I. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584–653. MR2281879

[2] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

[3] BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

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

[5] BÖTTCHER, A. and SILBERMANN, B. (1998). *Introduction to Large Truncated Toeplitz Matrices*. Springer, New York.

[6] BROCKWELL, P. and DAVIS, R. (1991). *Time Series and Methods*, 2nd ed. Springer, New York. MR1093459

[7] BROWN, B. W. and RUSSELL, K. (1997). Methods correcting for multiple testing: Operating characteristics. *Stat. Med.* **16** 2511–2528.

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

[9] CAI, T. and SUN, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc.* **104** 1467–1481.

[10] CAYON, L., JIN, J. and TREASTER, A. (2005). Higher criticism statistic: Detecting and identifying non-Gaussianity in the WMAP first year data. *Monthly Notes of the Royal Astronomical Society* **362** 826–832.

[11] CHEN, L., TONG, T. and ZHAO, H. (2005). Considering dependence among genes and markers for false discovery control in eQTL mapping. *Bioinformatics* **24** 2015–2022.

[12] CLARKE, S. and HALL, P. (2009). Robustness of multiple testing procedures against dependence. *Ann. Statist.* **37** 332–358. MR2488354

[13] COHEN, A., SACKROWITZ, H. B. and XU, M. (2009). A new multiple testing method in the dependent case. *Ann. Statist.* **37** 1518–1544. MR2509082

[14] CSÖRGÖ, M., CSÖRGÖ, S., HORVATH, L. and MASON, D. (1986). Weighted empirical and quantile processes. *Ann. Probab.* **14** 31–85. MR0815960

[15] COVER, T. M. and THOMAS. J. A. (2006). *Elementary Information Theory*. Wiley, Hoboken, NJ. MR2239987

[16] CRUZ, M., CAYON, L., MARTÍNEZ-GONZÁLEZ, E., VIEVA, P. and JIN, J. (2007). The non-Gaussian cold spot in the 3-year WMAP data. *Astrophys. J.* **655** 11–20.

[17] DELAIGLE, A. and HALL, J. (2009). Higher criticism in the context of unknown distribution, non-independence and classification. In *Platinum Jubilee Proceedings of the Indian Statistical Institute* (N. S. Narasimha Sastry, T. S. S. R. K. Rao, M. Delampady and B. Rajeev, eds.) 109–138. World Scientific, Hackensack, NJ.

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

[19] DONOHO, D. and JIN, J. (2006). Asymptotic minimaxity of False Discovery Rate thresholding for sparse exponential data. *Ann. Statist.* **34** 2980–3018. MR2329475

[20] 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.

[21] DONOHO, D. and JIN, J. (2009). Higher criticism thresholding achieves optimal phase diagram. *Phil. Trans. R. Soc. A* **367** 4449–4470.

[22] DUNNETT, C. W. and TAMHANE, A. C. (1995). Step-up testing of parameters with unequally correlated estimates. *Biometrics* **51** 217–227.

[23] EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302

[24] FINNER, H. and ROTERS, M. (1998). Asymptotic comparison of step-down and step-up multiple test procedures based on exchangeable test statistics. *Ann. Statist.* **26** 505–524. MR1626043

[25] GENOVESE, C. and WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32** 1035–1061. MR2065197

[26] GOEMAN, J., VAN DE GEER, S., DE KORT, F. and VAN HOUWELINGEN, H. (2004). A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics* **20** 93–99.

[27] GOEMAN, J., VAN DE GEER, S. and VAN HOUWELINGEN, H. (2006). Testing against a high dimensional alternative. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 477–493. MR2278336

[28] GRÖCHENIG, K. and LEINERT, M. (2006). Symmetry and inverse-closedness of matrix algebra and functional calculus for infinite matrices. *Trans. Amer. Math. Soc.* **358** 2695–2711. MR2204052

[29] 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

[30] HALL, P. and JIN, J. (2008). Properties of higher criticism under strong dependence. *Ann. Statist.* **36** 381–402. MR2387976

[31] HORN, R. A. and JOHNSON, C. R. (2006). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR1084815

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

[33] INGSTER, Y. I. (1999). Minimax detection of a signal for l_n^p-balls. *Math. Methods Statist.* **7** 401–428. MR1680087

[34] JAFFARD, S. (1990). Propriétés des matrices "bien localisées" près de leur diagonale et quelques applications. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **7** 461–476. MR1138533

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

[36] JIN, J. (2004). Detecting a target in very noisy data from multiple looks. In *A Festschrift to Honor Herman Rubin* (A. Dasgupta, ed.). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **45** 255–286. IMS, Beachwood, OH. MR2126903

[37] JIN, J. (2006). Higher criticism statistic: Theory and applications in non-Gaussian detection. In *Statistical Problems in Particle Physics, Astrophysics And Cosmology* (L. Lyons and M. K. Ünel, eds.). Imperial College Press, London. MR2270215

[38] JIN, J. (2007). Proportion of nonzero normal means: Universal oracle equivalences and uniformly consistent estimators. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 461–493. MR2420411

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

[40] JIN, J. and CAI, T. (2007). Estimating the null and the proportion of non-null effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 496–506. MR2325113

[41] KUELBS, J. and VIDYASHANKAR, A. N. (2009). Asymptotic inference for high dimensional data. *Ann. Statist.* To appear.

[42] MANSILLA, R., DE CASTILLO, N., GOVEZENSKY, T., MIRAMONTES, P., JOSÉ, M. and COHO, G. (2004). Long-range correlation in the whole human genome. Available at http://arxiv.org/pdf/q-bio/0402043v1.

[43] MESSER, P. W. and ARNDT, P. F. (2006). CorGen-measuring and generating long-range correlations for DNA sequence analysis. *Nucleic Acids Research* **34** W692–W695.

[44] JIN, J., STARCK, J.-L., DONOHO, D., AGHANIM, N. and FORNI, O. (2005). Cosmological non-Gaussian signature detection: Comparing performance of different statistical tests. *EURASIP J. Appl. Signal Process.* **15** 2470–2485. MR2210857

[45] MEINSHAUSEN, M. and RICE, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* **34** 373–393. MR2275246

[46] OLEJNIK, S., LI, J. M., SUPATTATHUM, S. and HUBERTY, C. J. (1997). Multiple testing and statistical power with modified Bonferroni procedures. *J. Educ. Behav. Statist.* **22** 389–406.

[47] ROM, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* **77** 663–665. MR1087860

[48] SARKAR, S. K. and CHANG, C. K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Amer. Statist. Assoc.* **92** 1601–1608. MR1615269

[49] SHROACK, G. and WELLNER, J. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York. MR0838963

[50] STRASSER, H. (1998). Differentiability of statistical experiments. *Statist. Decisions* **16** 113–130. MR1621858

[51] SUN, Q. (2005). Wiener's lemma for infinite matrices with polynomial off-diagonal decay. *C. R. Math. Acad. Sci. Paris* **340** 567–570. MR2138705

[52] TUKEY, J. W. (1989). Higher criticism for individual significances in several tables or parts of tables. Internal working paper, Princeton Univ.

[53] WIENER, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley, New York.

[54] WU, W. B. (2008). On false discovery control under dependence. *Ann. Statist.* **36** 364–380. MR2387975

[55] ZYGMUND, A. (1959). *Trigonometric Series*, 2nd ed. Cambridge Univ. Press, New York. MR0107776

DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF MELBOURNE
PARKVILLE, VIC, 3010
AUSTRALIA
AND
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
DAVIS, CALIFORNIA 95616
USA

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