The Annals of Statistics
2012, Vol. 40, No. 1, 412–435
DOI: 10.1214/11-AOS964

DETECTION OF CORRELATIONS

BY ERY ARIAS-CASTRO[1], SÉBASTIEN BUBECK AND GÁBOR LUGOSI[2]

University of California, San Diego, Princeton University, and ICREA and Pompeu Fabra University

We consider the hypothesis testing problem of deciding whether an observed high-dimensional vector has independent normal components or, alternatively, if it has a small subset of correlated components. The correlated components may have a certain combinatorial structure known to the statistician. We establish upper and lower bounds for the worst-case (minimax) risk in terms of the size of the correlated subset, the level of correlation, and the structure of the class of possibly correlated sets. We show that some simple tests have near-optimal performance in many cases, while the generalized likelihood ratio test is suboptimal in some important cases.

1. Introduction. In this paper we consider the following statistical problem: upon observing a high-dimensional vector, one is interested in detecting the presence of a sparse, possibly structured, correlated subset of components of the vector. Such problems emerge naturally in numerous scenarios. The setting is closely related to Gaussian signal detection in Gaussian white noise, on which there is an extensive literature surveyed in [20]. In image processing, textures are modeled via Markov random fields [13], so that detecting a textured object hidden in Gaussian white noise amounts to finding an area in the image where the pixel values are correlated. Similar situations arise in remote sensing based on a variety of hardware. A related task is the detection of space–time correlations in multivariate time series, with potential applications to finance [1].

1.1. *Setting and notation.* We investigate the possibilities and limitations in problems of detecting correlations in a Gaussian framework. We may formulate this as a general hypothesis testing problem as follows. An n-dimensional Gaussian vector $X = (X_1, \dots, X_n)$ is observed. Under the null hypothesis H_0, the vector X is standard normal, that is, with zero mean vector and identity covariance matrix. To describe the alternative hypothesis H_1, let $\mathcal{C}$ be a class of subsets of

Received September 2011; revised December 2011.
[1]Supported by ONR Grant N00014-09-1-0258.
[2]Supported by the Spanish Ministry of Science and Technology Grant MTM2009-09063 and PASCAL2 Network of Excellence under EC Grant 216886.
MSC2010 subject classifications. Primary 62F03; secondary 62F05.
Key words and phrases. Sparse covariance matrix, minimax detection, Bayesian detection, scan statistic, generalized likelihood ratio test.

$\{1, \ldots, n\}$, each of size k, indexing the possible "contaminated" components. One wishes to test whether there exists an $S \in \mathcal{C}$ such that

$$\mathrm{Cov}(X_i, X_j) = \begin{cases} 1, & i = j, \\ \rho, & i \neq j, \text{ with } i, j \in S, \\ 0, & \text{otherwise,} \end{cases}$$

where $\rho > 0$ is a given parameter. Equivalently, if $X = (X_1, \ldots, X_n)$ denotes the vector of observations, then

$$H_0 : X \sim \mathcal{N}(0, \mathbf{I}) \quad \text{vs.} \quad H_1 : X \sim \mathcal{N}(0, \mathbf{A}_S) \qquad \text{for some } S \in \mathcal{C},$$

where $\mathbf{I}$ denotes the $n \times n$ identity matrix and

$$(\mathbf{A}_S)_{i,j} = \begin{cases} 1, & i = j, \\ \rho, & i \neq j, \text{ with } i, j \in S, \\ 0, & \text{otherwise.} \end{cases} \tag{1.1}$$

We write $\mathbb{P}_0$ for the probability under H_0 (i.e., the standard normal measure in $\mathbb{R}^n$) and, for each $S \subset \mathcal{C}$, $\mathbb{P}_S$ for the measure of $\mathcal{N}(0, \mathbf{A}_S)$.

The goal of this paper is to understand for what values of the parameters (n, k, ρ) reliable testing is possible. This, of course, depends crucially on the size and structure of the subset class $\mathcal{C}$. We consider the following two prototypical classes:

- *k-intervals.* In this example, we consider the class of all intervals of size k of the form $\{i, \ldots, i + k - 1\}$ modulo n—for aesthetic reasons. (We call such an interval a *k-interval*.) This class is the flagship of *parametric* classes, typical of the class of objects of interest in signal processing.
- *k-sets.* In this example, we consider the class of all sets of size k, that is, of the form $\{i_1, \ldots, i_k\}$ where the indices are all distinct in $\{1, \ldots, n\}$. (We call such a set a *k-set*.) This class is the flagship of *nonparametric* classes, and may arise in multiple comparison situations.

Our theory, however, applies more generally to other classes, such as:

- *k-hypercubes.* In this example, the variables are indexed by the d-dimensional lattice, that is, $X = (X_i : i \in \{1, \ldots, m\}^d)$, so that the sample size is $n = m^d$, and we consider the class of all hyper-rectangles of the form $\bigtimes_{s=1}^{d} \{i_s, \ldots, i_s + k_s - 1\}$—each interval modulo m—of fixed size $\prod_{s=1}^{d} k_s = k$. This class is the simplest model for objects to be detected in images (mostly $d = 2, 3$ in applications).
- *Perfect matchings.* Suppose n is a perfect square with $k^2 = n$. The components of the observed vector X correspond to edges of the complete bipartite graph on $2k$ vertices and each set in $\mathcal{C}$ corresponds to the edges of a perfect matching. Thus, $|\mathcal{C}| = k!$. In this example $\mathcal{C}$ has a nontrivial combinatorial structure.

- *Spanning trees.* In another example, $n = \binom{k+1}{2}$ and the components of X correspond to the edges of a complete graph K_{k+1} on $k+1$ vertices and every element of $\mathcal{C}$ is a spanning tree of K_{k+1}.

As usual, a *test* is a binary-valued function $f : \mathbb{R}^n \to \{0, 1\}$. If $f(X) = 0$, then the test accepts the null hypothesis H_0; otherwise H_0 is rejected by f. We measure the performance of a test based on its *worst-case risk* over the class of interest $\mathcal{C}$, formally defined by

$$R^{\max}(f) = \mathbb{P}_0\{f(X) = 1\} + \max_{S \in \mathcal{C}} \mathbb{P}_S\{f(X) = 0\}.$$

We will derive upper and lower bounds on the *minimax risk*

$$R^{\max}_* := \inf_f R^{\max}(f).$$

A standard way of obtaining lower bounds for the minimax risk is by putting a prior on the class $\mathcal{C}$ and obtaining a lower bound on the corresponding *Bayesian risk*, which never exceeds the worst-case risk. Because this is true for any prior, the idea is to find one that is hardest (often called *least favorable*). Most classes we consider here are invariant under some group action: k-intervals are invariant under translation and k-sets are invariant under permutation. Invariance considerations ([21], Section 8.4) lead us to considering the uniform prior on $\mathcal{C}$, giving rise to the following *average risk*:

$$R(f) = \mathbb{P}_0\{f(X) = 1\} + \mathbb{P}_1\{f(X) = 0\},$$

where

$$\mathbb{P}_1\{f(X) = 0\} := \frac{1}{N} \sum_{S \in \mathcal{C}} \mathbb{P}_S\{f(X) = 0\},$$

and $N := |\mathcal{C}|$ is the cardinality of $\mathcal{C}$. The advantage of considering the average risk over the worst-case risk is that we know an optimal test for the former, which, by the Neyman–Pearson fundamental lemma, is the likelihood ratio test, denoted f^*. Introducing

$$Z_S = \exp\left(\tfrac{1}{2} X^T (\mathbf{I} - \mathbf{A}_S^{-1}) X\right) \tag{1.2}$$

for all $S \in \mathcal{C}$, the likelihood ratio between H_0 and H_1 may be written as

$$L(X) = \frac{1}{N} \sum_{S \in \mathcal{C}} \frac{Z_S}{\mathbb{E}_0 Z_S}, \tag{1.3}$$

and the optimal test becomes

$$f^*(x) = 0 \quad \text{if and only if} \quad L(x) \le 1.$$

Note that $\mathbb{E}_0 Z_S = \sqrt{\det(\mathbf{A}_S)}$. The (average) risk $R^* = R(f^*)$ of the optimal test is called the *Bayes risk* and it satisfies

$$R^* = 1 - \frac{1}{2}\mathbb{E}_0|L(X) - 1| = 1 - \frac{1}{2}\mathbb{E}_0\left|\frac{1}{N}\sum_{S\in\mathcal{C}}\frac{Z_S}{\mathbb{E}_0 Z_S} - 1\right|.$$

Note that, with the only exception of the case of spanning trees, in all examples mentioned above, the minimax and Bayes risks coincide, that is, $R^* = R_*^{\max}$. This is again due to invariance ([21], Section 8.4). (The class of spanning trees is not sufficiently symmetric for this equality to hold. However, as we will see below, even in this case, R^* and $R_*^{\max}$ are of the same order of magnitude.)

We focus on the case when n is large and formulate some of the results in an asymptotic language with $n \to \infty$ though in all cases explicit nonasymptotic inequalities are available. Of course, such asymptotic statements only make sense if we define a sequence of integers $k = k_n$ and classes $\mathcal{C} = \mathcal{C}_n$. This dependency in n will be left implicit. In this asymptotic setting, we say that *reliable* detection is possible (resp., impossible) if $R_*^{\max} \to 0$ (resp., $\to 1$) as $n \to \infty$.

REMARK (Covariance structure). In this paper we assume that, under the alternative hypothesis, the correlation between any two variables in the "contaminated" set is the same. While this model has a natural interpretation (see Lemma 1.1 below), it is clearly a restrictive assumption. This simplification is convenient in understanding the fundamental limits of detection (i.e., in obtaining lower bounds on the risk). At the same time, the tests we exhibit also match these lower bounds under more general correlation structures, such as

$$(\mathbf{A}_S)_{i,j} \begin{cases} = 1, & i = j, \\ \geq \rho, & i \neq j, \text{ with } i, j \in S, \\ = 0, & \text{otherwise.} \end{cases} \tag{1.4}$$

That said, dealing with more general correlation structures remains an interesting and important challenge, relevant in the detection of textured objects in textured background, for example.

1.2. *Relation to previous work.* The vast majority of the literature on detection is concerned with the detection of a signal in additive (often Gaussian) noise, which would correspond here to an alternative where $X_i \sim \mathcal{N}(\mu, 1)$ for $i \in S$, where $\mu > 0$ is the (per-coordinate) signal amplitude. We call this the *detection-of-means* setting. The literature on this problem is quite comprehensive. Indeed, the detection of k-intervals and k-hypercubes is treated extensively in a number of papers; see, for example, [4, 6, 10, 14, 22]. A more general framework that includes the detection of perfect matchings and spanning trees is investigated in [2], and the detection of k-sets is studied in [7, 16–19]. In the literature on detection of parametric objects, the phrase "correlation detection" usually refers to the method of *matched filters*, which consists of correlating the observed signal with signals

of interest. This is not the problem we are interested in here. While the problem of *detection-of-correlations* considered here is mathematically more challenging than the detection-of-means setting, there is a close relationship between the two. The connection is established by the representation theorem of [8]—stated here for the case Gaussian random variables.

LEMMA 1.1 ([8]). *Let $X_1, \dots, X_k$ be standard normal with $\mathrm{Cov}(X_i, X_j) = \rho$ for $i \neq j$. Then there are i.i.d. standard normal random variables, denoted $U, U_1, \dots, U_k$, such that $X_i = \sqrt{\rho}U + \sqrt{1-\rho}U_i$ for all i.*

Thus, given U, the problem becomes that of detecting a subset of variables with nonzero mean (equal to $\sqrt{\rho}U$) and with a variance equal to $1-\rho$ (instead of 1). This simple observation will be very useful to us later on. When U is random, the setting is similar to that of detecting a Gaussian process (here equal to $\sqrt{\rho}U$ for $i \in S$, and equal to 0 otherwise) in additive Gaussian noise. However, the typical setting assumes that the Gaussian process affects all parts of the signal [20]. In our setting, the signal (the subset of correlated variables) will be sparse. Since we only have one instance of the signal X, the problem cannot be considered from the perspective of either multivariate statistics or multivariate time series. If indeed we had multiple copies of X, we could draw inspiration from the literature on the estimation of sparse correlation matrices [9, 12], from the literature on multivariate time series [23], or on other approaches [15]; but this is not the case as we only observe X. Closer in spirit to our goal of detecting correlations in a single vector of observation is the paper of [3], which aims at testing whether a Gaussian random field is i.i.d. or has some Markov dependency structure. Their setting models communication networks and is not directly related to ours.

It transpires, therefore, that ρ in the detection-of-correlations setting plays a role analogous to μ^2 in the detection-of-means setting. While this is true to a certain extent, the picture is quite a bit more subtle. The detection-of-means problem for parametric classes such as k-intervals is well understood. In such cases, μ^2 needs to be of order at least $(1/k)\log(n/k)$ for reliable detection of k-intervals to be possible. This remains true in the detection-of-correlations setting, and the *generalized likelihood ratio test* (*GLRT*) is near-optimal, just as in the detection-of-means problem; see, for example, [6].

Our inspiration for considering k-sets comes from the line of research on the detection of sparse Gaussian mixtures. Very precise results are known on (n, k, μ) that make detection possible [7, 18, 19] and optimal tests have been developed, such as the "higher criticism" [16, 17]. In fact, the recent paper [11] deals with heteroscedastic instances of the detection-of-means problem where the variance of the anomalous variables may be different from 1. For example, it is known that, when $n = O(k^2)$ [resp., $k^2 = o(n)$], μ^2 needs to be of order at least n/k^2 [resp., $\log(n)$] for reliable detection of k-sets to be possible, and the test based on $\sum_i X_i$ (resp., $\max_i X_i$) is near-optimal. Though more precise results are available when

$k^2 = o(n)$, these cannot be translated immediately to our case via the representation theorem of Lemma 1.1. As a bonus, we show that the GLRT is clearly suboptimal in some regimes—see Theorem 3.1. Note that in the detection-of-means problem it is not known whether the GLRT has any power.

1.3. *Contribution and content of the paper.* This paper contains a collection of positive and negative results about the detection-of-correlation problem described above. In Section 2 we derive lower bounds for the Bayes risk. The usual route of bounding the variance of the likelihood ratio, that is very successful in the detection-of-means problem, leads essentially nowhere in our case. Instead, we develop a new approach based on Lemma 1.1. We establish a general lower bound for the Bayes risk in terms of the moment generating function of the size of the overlap of two randomly chosen elements of the class $\mathcal{C}$. This quantity also plays a crucial role in the detection-of-means setting and we are able to use inequalities worked out in the literature in various examples. In Section 3 we study the performance of some simple and natural tests such as the squared-sum test—based on $(\sum_i X_i)^2$, the generalized likelihood ratio test (GLRT) and a goodness-of-fit (GOF) test, as well as some variants. We show that, in the case of parametric classes such as k-intervals and k-hypercubes, the GLRT is essentially optimal. The squared-sum test is shown to be essentially optimal in the case of k-sets when k^2/n is large, while the GLRT is clearly suboptimal in this regime. This is an interesting example where the GLRT fails miserably. When k^2/n is small, detection is only possible when ρ is very close to 1. We show that a simple GOF test is near-optimal in this case. The analysis of tests such as the squared-sum test and the GLRT involves handling quadratic forms in X. This is technically more challenging than the analogous problem for the detection-of-means setting in which only linear functions of X appear (which are normal random variables).

2. Lower bounds. In this section we investigate lower bounds on the risk, which are sometimes called information bounds. First we consider the special case when $\mathcal{C}$ contains only one element as this example will serve as a benchmark for other examples. Then we consider the standard method based on bounding the variance of the likelihood ratio under the null hypothesis, and show that it leads nowhere. We then develop a new bound based on Lemma 1.1 that has powerful implications, leading to fairly sharp bounds in a number of examples.

2.1. *The case $N = 1$.* As a warm-up, and to gain insight into the problem, consider first the simplest case where $\mathcal{C}$ contains just one set, say $S = \{1, \dots, k\}$. In this case, the alternative hypothesis is simple and the likelihood ratio (Neyman–Pearson) test may be expressed by

$$f^*(X) = 0 \quad \text{if and only if} \quad X^T(\mathbf{I} - \mathbf{A}_S^{-1})X \le \log\det(\mathbf{A}_S).$$

This follows by the fact that $\mathbb{E}Z_S = \sqrt{\det(\mathbf{A}_S)}$ which is easy to check by straightforward calculation.

The next simple lemma helps understand the behavior of the Bayes risk.

LEMMA 2.1. *Under* $\mathbb{P}_0$, $X^T(\mathbf{I} - \mathbf{A}_S^{-1})X$ *is distributed as*

$$-\frac{\rho}{1-\rho}\chi^2_{k-1} + \frac{\rho(k-1)}{1+\rho(k-1)}\chi^2_1,$$

and under the alternative $\mathbb{P}_S$, *it has the same distribution as*

$$-\rho\chi^2_{k-1} + \rho(k-1)\chi^2_1,$$

where χ^2_1 *and* χ^2_{k-1} *denote independent* χ^2 *random variables with degrees of freedom* 1 *and* $k-1$, *respectively.*

PROOF. If $Y = (Y_1, \dots, Y_n)$ denotes a standard normal vector, then under H_0, the quadratic form $X^T(\mathbf{I} - \mathbf{A}_S^{-1})X$ is distributed as $Y^T(\mathbf{I} - \mathbf{A}_S^{-1})Y$, and under the alternative, it has the distribution of $Y^T(\mathbf{A}_S - \mathbf{I})Y$, since X is distributed as $\mathbf{A}_S^{1/2}Y$.

Now, observe that for any symmetric matrix $\mathbf{B}$ with eigenvalues $\lambda_1, \dots, \lambda_n$, the quadratic form $Y^T\mathbf{B}Y$ has distribution

$$Y^T\mathbf{B}Y \sim \sum_{i=1}^n \lambda_i Y_i^2. \tag{2.1}$$

This follows simply by diagonalizing $\mathbf{B}$ and using the rotational invariance of the standard normal distribution.

The lemma follows from this simple representation and the fact that $\mathbf{A}_S$ has eigenvalue $1-\rho$ with multiplicity $k-1$, $1+\rho(k-1)$ with multiplicity 1, and the eigenvalue 1 with multiplicity $n-k$. □

Now it is straightforward to analyze the Bayes risk. In particular, we immediately have the following:

PROPOSITION 2.1. *If* $\mathcal{C}$ *is a singleton,* $\lim_{k\to\infty} R^* = 0$ *if and only if* $\rho k \to \infty$. *Similarly,* $\lim_{k\to\infty} R^* = 1$ *if and only if* $\rho k \to 0$.

PROOF. Suppose $\rho k \to \infty$. It suffices to show that there exists a threshold τ_k such that $\mathbb{P}_0\{X^T(\mathbf{I} - \mathbf{A}_S^{-1})X \geq \tau_k\} \to 0$ and $\mathbb{P}_S\{X^T(\mathbf{I} - \mathbf{A}_S^{-1})X < \tau_k\} \to 0$. We use Lemma 2.1 and the fact that, by Chebyshev's inequality,

$$\mathbf{P}\{|\chi^2_k - k| > t_k\sqrt{k}\} \to 0, \qquad k \to \infty,$$

for any sequence $t_k \to \infty$, and the fact that

$$\mathbf{P}\{t_k^{-1} < \chi^2_1 < t_k\} \to 1 \qquad \text{as } k \to \infty.$$

We choose $t_k = \log k$ and define $\tau_k := -\rho k + \rho t_k \sqrt{k} + t_k$. Then under the null,

$$\mathbb{P}_0\{X^T(\mathbf{I} - \mathbf{A}_S^{-1})X \geq \tau_k\} \to 0,$$

and under the alternative, setting $\eta_k := -\rho k - \rho t_k \sqrt{k} + \rho k t_k^{-1}$,

$$\mathbb{P}_S\{X^T(\mathbf{I} - \mathbf{A}_S^{-1})X < \eta_k\} \to 0.$$

We then conclude with the fact that, for k large enough, $\tau_k < \eta_k$.

If ρk is bounded, the densities of the test statistic under both hypotheses have a significant overlap and the risk cannot converge to 0.

The proof of the second statement is similar. □

Clearly, the role of n is immaterial in this specific example as the optimal test ignores all components whose indices are not in $S = \{1, \dots, k\}$.

2.2. *The moment method.* When the class $\mathcal{C}$ contains more than one element, the likelihood ratio with uniform prior on $\mathcal{C}$ is given by (1.3). A common approach for deriving a lower bound on the Bayes risk is via an upper bound on the *variance* of $L(X)$ under the null. Indeed, by the Cauchy–Schwarz inequality,

$$R^* = 1 - \frac{\mathbb{E}_0|L(X) - 1|}{2} \geq 1 - \frac{\sqrt{\mathbb{E}_0[L(X)^2] - 1}}{2}.$$

Therefore, an upper bound on $\mathbb{E}_0[L(X)^2] - 1 = \mathrm{Var}_0(L(X))$ leads to a lower bound on R^*.

Let $\Lambda = \det(\mathbf{A}_S) = (1-\rho)^{k-1}(1 + \rho(k-1))$, which is independent of $S \in \mathcal{C}$. By Fubini's theorem, we have

$$\mathbb{E}_0 L(X)^2 = \frac{1}{\Lambda}\frac{1}{N^2}\sum_{S,S' \in \mathcal{C}} \mathbb{E}_0(Z_S Z_{S'}),$$

where Z_S is defined in (1.2). We focus on terms of the double sum for which $S = S'$.

The following result is a straightforward consequence of the representation (2.1) and the well-known expression for the moment generating function of χ_1^2.

LEMMA 2.2. *Suppose X is a standard normal vector in $\mathbb{R}^n$ and $\mathbf{M}$ is an $n \times n$ symmetric matrix with eigenvalues strictly less than $1/2$. Then*

$$\mathbb{E}\exp(X^T\mathbf{M}X) = \det(\mathbf{I} - 2\mathbf{M})^{-1/2}.$$

If $\mathbf{M}$ has an eigenvalue exceeding $1/2$, then $\mathbb{E}\exp(X^T\mathbf{M}X) = +\infty$.

Since $\mathbf{M} := \mathbf{I} - \mathbf{A}_S^{-1}$ has eigenvalue $-\rho/(1-\rho)$ with multiplicity k, eigenvalue $\rho(k-1)/(1+\rho(k-1))$ with multiplicity 1, and eigenvalue 0 with multiplicity $n-k$, $\mathbb{E}_0[Z_S^2] = \mathbb{E}_0\exp(X^T\mathbf{M}X) = +\infty$ unless $\rho(k-1) < 1$. The implications

are rather insubstantial. It only shows that, when $\rho(k-1) \le 1-\varepsilon$ with $\varepsilon > 0$ fixed, the Bayes risk does not tend to zero. As we shall see, this lower bound is grossly suboptimal, except in the case where $\mathcal{C}$ is a singleton (as in Section 2.1) or does not grow in size with n.

A refinement of this method consists in bounding the first and second *truncated* moments of $L(X)$, again under the null hypothesis. For example, this is the approach used in [11, 18] in the detection-of-means setting for the case of k-sets to obtain sharp bounds. Unfortunately, in our case this method only provides a useful bound when the class $\mathcal{C}$ is not too large (i.e., has size polynomial in k) while it does not seem to lead anywhere in the case of k-sets. The computations are quite involved and we do not provide details here, as we were able to obtain a more powerful general bound that applies to both k-intervals and k-sets. This is presented in the next section.

2.3. *A general lower bound.* In this section we derive a general lower bound for the Bayes risk. As in the detection-of-means problem [2, 4, 5], the relevant measure of complexity is in terms of the moment generating function of the size of the overlap of two randomly chosen elements of $\mathcal{C}$. In the detection-of-means setting, this is a consequence of bounding the variance of the likelihood ratio. We saw in Section 2.2 that this method is useless here. Instead, we make a connection between the two problems using Lemma 1.1.

THEOREM 2.1. *For any class $\mathcal{C}$ and any $a > 0$,*

$$R^* \ge \mathbf{P}\{|\mathcal{N}(0,1)| \le a\}\bigl(1 - \tfrac{1}{2}\sqrt{\mathbb{E}\exp(\nu_a Z) - 1}\bigr),$$

where $\nu_a := \rho a^2/(1+\rho) - \frac{1}{2}\log(1-\rho^2)$ and $Z = |S \cap S'|$, with S, S' drawn independently, uniformly at random from $\mathcal{C}$. In particular, taking $a = 1$,

$$R^* \ge 0.6 - 0.3\sqrt{\mathbb{E}\exp(\nu_1 Z) - 1},$$

where $\nu_1 = \nu(\rho) := \rho/(1+\rho) - \frac{1}{2}\log(1-\rho^2)$.

PROOF. The starting point of the proof is Lemma 1.1,[3] which enables us to represent the vector X as

$$X_i = \begin{cases} U_i, & \text{if } i \notin S, \\ \sqrt{\rho}U + \sqrt{1-\rho}U_i, & \text{if } i \in S, \end{cases}$$

where $U, U_1, \dots, U_n$ are independent standard normal random variables.

We consider now the alternative $H_1(u)$, defined as the alternative H_1 given $U = u$. Let $R(f)$, L, f^* [resp., $R_u(f)$, L_u, f_u^*] be the risk of a test f, the likeli-

[3] In fact, we only need to assume that X is as described in distribution.

hood ratio, and the optimal (likelihood ratio) test, for H_0 versus H_1 [resp., H_0 versus $H_1(u)$]. For any $u \in \mathbb{R}$, $R_u(f_u^*) \le R_u(f^*)$, by the optimality of f_u^* for H_0 versus $H_1(u)$. Therefore, conditioning on U,

$$\begin{aligned} R^* &= R(f^*) \\ &= \mathbb{E}_U R_U(f^*) \\ &\ge \mathbb{E}_U R_U(f_U^*) \\ &= 1 - \tfrac{1}{2}\mathbb{E}_U \mathbb{E}_0 |L_U(X) - 1|. \end{aligned}$$

[$\mathbb{E}_U$ is the expectation with respect to $U \sim \mathcal{N}(0,1)$.] Using the fact that $\mathbb{E}_0|L_u(X) - 1| \le 2$ for all u, we have

$$\mathbb{E}_U \mathbb{E}_0 |L_U(X) - 1| \le 2\mathbb{P}\{|U| > a\} + \mathbb{P}\{|U| \le a\} \max_{u \in [-a,a]} \mathbb{E}_0 |L_u(X) - 1|$$

and therefore, using the Cauchy–Schwarz inequality,

$$\begin{aligned} 1 - \frac{1}{2}\mathbb{E}_U \mathbb{E}_0 |L_U(X) - 1| &\ge \mathbb{P}\{|U| \le a\}\Big(1 - \frac{1}{2} \max_{u \in [-a,a]} \mathbb{E}_0 |L_u(X) - 1|\Big) \\ &\ge \mathbb{P}\{|U| \le a\}\Big(1 - \frac{1}{2} \max_{u \in [-a,a]} \sqrt{\mathbb{E}_0 L_u^2(X) - 1}\Big). \end{aligned}$$

Since

$$\begin{aligned} L_u(x) &= \frac{1}{N} \sum_{S \in \mathcal{C}} \frac{1}{(1-\rho)^{k/2}} \exp\Big(- \sum_{i \in S} \frac{(x_i - \sqrt{\rho} u)^2}{2(1-\rho)} - \sum_{i \notin S} \frac{x_i^2}{2} \Big) \exp\Big(\sum_{i=1}^n \frac{x_i^2}{2} \Big) \\ &= \frac{1}{N} \sum_{S \in \mathcal{C}} \frac{1}{(1-\rho)^{k/2}} \exp\Big(\sum_{i \in S} \frac{x_i^2}{2} - \frac{(x_i - \sqrt{\rho} u)^2}{2(1-\rho)} \Big), \end{aligned}$$

we get

$$\begin{aligned} \mathbb{E}_0 L_u^2(X) &= \frac{1}{N^2} \sum_{S, S' \in \mathcal{C}} \frac{1}{(1-\rho)^k} \mathbb{E}_0 \exp\Big(\sum_{i \in S \cap S'} X_i^2 - \frac{(X_i - \sqrt{\rho} u)^2}{1-\rho} \\ &\qquad + \sum_{i \in S \Delta S'} \frac{X_i^2}{2} - \frac{(X_i - \sqrt{\rho} u)^2}{2(1-\rho)} \Big) \\ &= \frac{1}{N^2} \sum_{S, S' \in \mathcal{C}} \frac{1}{(1-\rho)^k (2\pi)^{n/2}} \\ &\qquad \times \int_{-\infty}^{+\infty} \exp\Big(\sum_{i \in S \cap S'} \frac{x_i^2}{2} - \frac{(x_i - \sqrt{\rho} u)^2}{1-\rho} \\ &\qquad\qquad - \sum_{i \in S \Delta S'} \frac{(x_i - \sqrt{\rho} u)^2}{2(1-\rho)} - \sum_{i \notin S \cup S'} \frac{x_i^2}{2} \Big) dx. \end{aligned}$$

It is easy to check that

$$\frac{x_i^2}{2} - \frac{(x_i - \sqrt{\rho}u)^2}{1-\rho} = \frac{\rho u^2}{1+\rho} - \frac{1+\rho}{2(1-\rho)}\left(x_i - \frac{2\sqrt{\rho}u}{1+\rho}\right)^2,$$

which implies

$$\begin{aligned}
\mathbb{E}_0 L_u^2(X) &= \frac{1}{N^2}\sum_{S,S'\in\mathcal{C}} \frac{\exp((\rho u^2/(1+\rho))|S\cap S'|)}{(1-\rho)^k(2\pi)^{n/2}} \\
&\qquad \times \int_{-\infty}^{+\infty} \exp\Bigg(-\sum_{i\in S\cap S'} \frac{1+\rho}{2(1-\rho)}\left(x_i - \frac{2\sqrt{\rho}u}{1+\rho}\right)^2 \\
&\qquad\qquad - \sum_{i\in S\Delta S'} \frac{(x_i-\sqrt{\rho}u)^2}{2(1-\rho)} - \sum_{i\notin S\cup S'} \frac{x_i^2}{2}\Bigg)\,dx \\
&= \frac{1}{N^2}\sum_{S,S'\in\mathcal{C}} \frac{\exp((\rho u^2/(1+\rho))|S\cap S'|)}{(1-\rho)^k}\left(\frac{1-\rho}{1+\rho}\right)^{|S\cap S'|/2} \\
&\qquad \times (1-\rho)^{k-|S\cap S'|} \\
&\le \frac{1}{N^2}\sum_{S,S'\in\mathcal{C}} \exp\left(\left(\frac{\rho u^2}{1+\rho} - \frac{1}{2}\log(1-\rho^2)\right)|S\cap S'|\right),
\end{aligned}$$

which concludes the proof. □

We now apply Theorem 2.1 to a few examples. The theorem converts the problem into a purely combinatorial question and [2] offers various estimates for the moment generating function of Z which we may use for our purposes.

2.3.1. *Nonoverlapping sets.* Consider first the simplest case when $\mathcal{C}$ contains N disjoint sets of size k.

COROLLARY 2.1. *Let $\mathcal{C}$ be the class of all sets of size k. If*

$$v(\rho) \le \frac{\log(N)}{k},$$

then the Bayes risk satisfies $R^* \ge 0.3$, and $R^* \to 1$ if $\rho \ll \min(1, \log(N)/k)$ or if $(1-\rho)N^{2/k} \to \infty$.

PROOF. Clearly, the size Z of the overlap of two randomly chosen elements of $\mathcal{C}$ equals zero with probability $1 - 1/N$ and k with probability $1/N$. Thus,

$$\mathbb{E}e^{vZ} - 1 = (1/N)(e^{vk} - 1) \le (1/N)e^{vk},$$

which is bounded by 1 if $\nu \le \log(N)/k$. The first part then follows from the second part of Theorem 2.1. For the second part, we need to find $a \to \infty$ such that $\nu_a k - \log N \to -\infty$. (Note that in this case the upper bound above tends to zero.) First assume that $\rho \ll \min(1, \log(N)/k)$. In that case, $\nu_a \sim \rho a^2$, so it suffices to take $a \to \infty$ slowly enough that $\rho a^2 \ll \min(1, \log(N)/k)$. Next assume that $b := \log(1-\rho) + 2\log(N)/k \to \infty$. In this case, we have $\nu_a \le a^2 - (1/2)\log(1-\rho)$, and we simply choose $a \to \infty$ slowly enough that $a^2 - b/2 \to -\infty$. □

2.3.2. *k-intervals.* Consider the class of all k-intervals. The situation is similar to that of nonoverlapping sets. (In fact, since this class of k-intervals contains $[n/k]$ nonoverlapping sets of size k, we could immediately deduce a lower bound via Corollary 2.1.)

COROLLARY 2.2. *Let $\mathcal{C}$ be the class of all k-intervals. If*

$$\nu(\rho) \le \frac{\log(n/(2k))}{k},$$

then the Bayes risk satisfies $R^* \ge 0.3$, and $R^* \to 1$ if $\rho \ll \min(1, \log(n/k)/k)$ or if $(1-\rho)(n/k)^{2/k} \to \infty$.

PROOF. For two k-intervals chosen independently and uniformly at random,

$$\mathbf{P}\{|S \cap S'| = \ell\} = \frac{2}{N} \qquad \forall \ell = 1, \dots, k.$$

Thus,

$$\mathbb{E}e^{\nu Z} - 1 = \frac{2}{N}\left(\sum_{\ell=1}^{k} e^{\nu\ell} - k\right) \le \frac{2k}{N} e^{\nu k},$$

and proceed as in the proof of Corollary 2.1, using the fact that $N \le n$. □

2.3.3. *k-sets.* Consider the class of all sets of size k.

COROLLARY 2.3. *Let $\mathcal{C}$ be the class of k-sets. If*

$$\frac{k^2}{n} \le \frac{\ln 2}{\exp(\nu(\rho)) - 1},$$

then the Bayes risk satisfies $R^* \ge 0.3$, and $R^* \to 1$ if either $k^2/n \to \infty$ and $\rho k^2/n \to 0$, or $(1-\rho)n^2/k^4 \to \infty$.

PROOF. By [2], Proposition 3.4, which uses negative association,

$$\mathbb{E}e^{\nu Z} \le \left((e^{\nu} - 1)\frac{k}{n} + 1\right)^k \le \exp\left((e^{\nu} - 1)\frac{k^2}{n}\right),$$

where the last expression is bounded by 2 under the postulated condition, and tends to 1 if either $k^2/n \to \infty$ and $\nu k^2/n \to 0$, or $k^2/n \to 0$ and $e^\nu k^2/n \to 0$. First assume that $k^2/n \to \infty$ and $\rho k^2/n \to 0$. By choosing $a \to \infty$ slowly enough that $\rho a^2 k^2/n \to 0$ we ensure that $\nu_a k^2/n \to 0$. Next assume that $b := \log(1-\rho) - 2\log(k^2/n) \to \infty$. Since $\nu_a \le a^2 - (1/2)\log(1-\rho)$, it suffices to take $a \to \infty$ slowly enough that $a^2 - b/2 \to -\infty$ to ensure that $e^\nu k^2/n \to 0$. The result then follows from Theorem 2.1. □

2.3.4. *Perfect matchings.* Consider now the example of perfect matchings described in the Introduction. Here $k = \sqrt{n}$. Once again, Theorem 2.1 applies and implies that testing is impossible for moderate values of ρ.

COROLLARY 2.4. *Let $\mathcal{C}$ be the class of all perfect matchings. If $\rho \le 1/2$, the Bayes risk satisfies $R^* \ge 0.3$. Also, $R^* \to 1$ if $\rho \to 0$.*

PROOF. The random variable Z for this class is considered by [2], who prove that

$$\mathbb{E}e^{\nu Z} \le \left((e^\nu - 1)\frac{1}{\sqrt{n}} + 1\right)^{\sqrt{n}} \le e^{e^\nu - 1}.$$

This is bounded by 2 whenever $\nu \le 1 + \ln\ln 2$, which is satisfied whenever $\rho \le 1/2$, and tends to 1 if $\nu \to 0$. We then apply Theorem 2.1. □

2.3.5. *Spanning trees.* A similar argument applies for the class of all spanning trees of a complete graph with $k+1$ vertices [and $n = (k+1)k/2$ edges] as described in the Introduction.

COROLLARY 2.5. *Let $\mathcal{C}$ be the class of all spanning trees. If $\rho \le 0.4$, then the Bayes risk satisfies $R^* \ge 0.15$. We also have $R^* \to 1$ if $\rho \to 0$.*

PROOF. It is shown in [2] that

$$\mathbb{E}e^{\nu Z} \le \left((e^\nu - 1)\frac{2}{k+1} + 1\right)^{k} \le e^{2(e^\nu - 1)},$$

which is bounded by $13/4$ whenever $\nu \le 1 + \ln((\ln(13/4))/2)$, which is satisfied whenever $\rho \le 0.4$, and tends to 1 if $\nu \to 0$. We then apply Theorem 2.1. □

3. Some near-optimal tests. We already know that the likelihood ratio test is optimal in the Bayesian setting. We study here other tests for multiple reasons. First, the likelihood ratio test seems difficult to compute in most situations. Second, the likelihood ratio test is heavily dependent on the prior we choose—here, the uniform distribution on the class. The third, and perhaps most important, reason is that it is difficult to obtain directly upper bounds for the (worst-case) risk of the

likelihood ratio test whereas the tests considered below are easier to analyze and often yield near-optimal performance. Whenever we obtain an upper bound for the risk of a test that matches the lower bounds developed in the previous section, we have a full understanding of the limitations and possibilities of detection for the particular case considered, and this is our main goal in this paper.

We consider the squared-sum test, which corresponds to the ANOVA test in the detection-of-means setting, the generalized likelihood ratio test (GLRT) and a goodness-of-fit (GOF) test, as well as some variants. We say that a test is *near-optimal* for a certain setting if it achieves the information bound for that setting to first order.

3.1. *The squared-sum test.* One of the simplest tests is based on the observation that the magnitude of the squared-sum $(\sum_{i=1}^n X_i)^2$ may be substantially different under the null and alternative hypotheses due to the higher correlation under the latter.

Indeed, under $\mathbb{P}_0$, $(\sum_{i=1}^n X_i)^2$ is distributed as $n\chi_1^2$, while for any $S \subset \{1,\dots,n\}$ with $|S| = k$, under $\mathbb{P}_S$, $(\sum_{i=1}^n X_i)^2$ has the same distribution as $(n+\rho k(k-1))\chi_1^2$; in fact, under the more general correlation model (1.4), this is a (stochastic) lower bound. This immediately leads to the following result.

PROPOSITION 3.1. *Let $\mathcal{C}$ be an arbitrary class of sets of size k and suppose that $\rho k^2/n \to \infty$ in (1.4). If t_n is such that $t_n \to \infty$ but $t_n = o(\rho k^2/n)$, then the test which rejects the null hypothesis if $(\sum_{i=1}^n X_i)^2 > nt_n$ has a worst-case risk converging to zero. However, any test based on $(\sum_{i=1}^n X_i)^2$ is powerless if $\rho k^2/n \to 0$ in (1.1).*

In Corollary 2.3, we saw that reliable detection of k-sets is impossible if $k^2/n \to \infty$ and $\rho k^2/n \to 0$. Here we see that, when $\rho k^2/n \to \infty$, the squared-sum test is asymptotically powerful. Hence, the following statement:

> *The squared-sum test is near-optimal for detecting k-sets in the regime where $k^2/n \to \infty$.*

On the other hand, in the regime $k^2/n \to 0$, the squared-sum test is powerless even if $\rho = 1$. The test does not require knowledge of ρ, though knowing ρ allows one to choose the threshold t_n in an optimal fashion; if ρ is unknown, we simply choose $t_n \to 0$ very slowly.

3.2. *The generalized likelihood ratio test.* In this section we investigate the performance of the generalized likelihood ratio test (GLRT). We show that for parametric classes such as k-intervals, the test is near-optimal. However, for the nonparametric class of k-sets, the test performs poorly in some regimes.

By definition, the GLRT rejects for large values of $\max_{S\in\mathcal{C}} Z_S/\mathbb{E}_0 Z_S$, or simply $\max_{S\in\mathcal{C}} Z_S$ when all the sets in the class $\mathcal{C}$ are of same size, since $\mathbb{E}_0 Z_S$ only depends on the size of S. Hence, the GLRT is of the form

$$f(X)=0 \quad \text{if and only if} \quad \max_{S\in\mathcal{C}} X^T(\mathbf{I}-\mathbf{A}_S^{-1})X \le t$$

for some appropriately chosen t. We immediately notice that the GLRT requires knowledge of ρ

Our analysis of the GLRT is based on Lemma 2.1, which provides the distribution of the quadratic form $X^T(\mathbf{I}-\mathbf{A}_S^{-1})X$ under the null $\mathbb{P}_0$ and under the alternative $\mathbb{P}_S$. Under the null we need to control the maximum of such quadratic forms over $S\in\mathcal{C}$, which we do using exponential concentration inequalities for chi-squared distributions.

3.2.1. *The GLRT for k-intervals and other parametric classes.* Recalling Corollary 2.2, when detecting k-intervals all tests are asymptotically powerless when $\rho \ll \min(1, \log(n/k)/k)$. We assume for concreteness that $k/\log n \to \infty$, for otherwise detecting k-intervals for very small k has more to do with detecting k-sets. We state a general result that applies for classes of small cardinality.

PROPOSITION 3.2. *Consider a class $\mathcal{C}$ of sets of size k, with cardinality $N \to \infty$ such that $\log(N)/k \to 0$. When $\rho k/\log N \to \infty$, the generalized likelihood ratio test with threshold value $t = -\rho k + \rho\sqrt{5k\log N} + 2\log N$ has worst-case risk tending to zero.*

PROOF. We first bound the probability of Type I error. Indeed, under the null, by Lemma 2.1 and its proof, we can decompose

$$X^T(\mathbf{I}-\mathbf{A}_S^{-1})X = -\frac{\rho}{1-\rho}C_S + \frac{\rho(k-1)}{1+\rho(k-1)}D_S,$$

where $C_S \sim \chi^2_{k-1}$ and $D_S \sim \chi^2_1$. Hence,

$$\max_{S\in\mathcal{C}} X^T(\mathbf{I}-\mathbf{A}_S^{-1})X \le -\rho \min_{S\in\mathcal{C}} C_S + \max_{S\in\mathcal{C}} D_S.$$

It is well known that the maximum of N standard normals is bounded by $\sqrt{2\log N}$ with probability tending to 1 as $N\to\infty$. Hence, the second term on the right-hand side is bounded by $2\log N$ with high probability. For the first term, we combine the union bound and Chernoff's bound to obtain, for all $a \le 1$,

$$\begin{aligned}\mathbb{P}_0\Big\{\min_{S\in\mathcal{C}} C_S < a(k-1)\Big\} &\le N\mathbf{P}\{\chi^2_{k-1} < a(k-1)\} \\ &\le N\exp\Big(-\frac{(k-1)}{2}(a-1-\log a)\Big).\end{aligned} \tag{3.1}$$

Using the fact that $a - 1 - \log a \sim \frac{1}{2}(1-a)^2$ when $a \to 1$, the right-hand side tends to zero when $a = 1 - \sqrt{(5/k)\log N}$. We arrive at the conclusion that the GLRT with threshold $t = -\rho k + \rho\sqrt{5k\log N} + 2\log N$ has probability of Type I error tending to zero.

Now consider the alternative under $\mathbb{P}_S$. By Lemma 2.1 and Chebyshev's inequality,

$$X^T(\mathbf{I} - \mathbf{A}_S^{-1})X \geq -\rho k - \rho s_k\sqrt{k} + \rho k/s_k$$

with high probability when $s_k \to \infty$. We then conclude by the fact that the right-hand side is larger than t when $s_k \to \infty$ sufficiently slowly. □

Comparing the performance of the GLRT in Proposition 3.2 with the lower bound for k-intervals in Corollary 2.2, we see that the GLRT is near-optimal for detecting k-intervals. This is actually the case for all parametric classes we know of.

3.2.2. *The GRLT for k-sets and other nonparametric classes.* Consider now the example of the class of all k-sets. Compared to the previous section, the situation here is different in that N, the size of the class $\mathcal{C}$, is much larger. For example, for k-sets, $N = \binom{n}{k}$, and therefore $\log(N)/k \to \infty$ with $n \to \infty$. The equivalent of Proposition 3.2 for this regime is the following:

PROPOSITION 3.3. *Consider a class $\mathcal{C}$ of sets of size k, with cardinality $N \to \infty$ such that $\log(N)/k \to \infty$. When $\eta := (1-\rho)N^{2/k}(\log N)/k \to 0$, the generalized likelihood ratio test with threshold value $t = -(\log N)/\sqrt{\eta}$ has worst-case risk tending to zero.*

PROOF. We follow the proof of Proposition 3.2. The only difference is in (3.1), where we now need $a \to 0$ and that right-hand side tends to zero when $\log a + 2(\log N)/k \to -\infty$. Choose $a = N^{-2/k}\sqrt{\eta}$, obtaining that, with high probability,

$$\max_{S\in\mathcal{C}} X^T(\mathbf{I} - \mathbf{A}_S^{-1})X \leq -\frac{\rho}{1-\rho}N^{-2/k}k\sqrt{\eta} + 2\log N. \tag{3.2}$$

As before, with high probability under $\mathbb{P}_S$,

$$X^T(\mathbf{I} - \mathbf{A}_S^{-1})X \geq -\rho k, \tag{3.3}$$

so we only need to check that the threshold t is larger than the right-hand side in (3.2) and smaller than the right-hand side in (3.3), which is the case by the assumptions we made. □

Notice that in Proposition 3.3 the condition on ρ implies that $\rho \to 1$, which is much stronger than what the squared-sum test requires when $k^2/n \to \infty$. For

k-sets, $N = \binom{n}{k}$—so that $\log N = k\log(n/k) + O(k)$—and the requirement is that $(1-\rho)(n/k)^2\log(n/k) \to 0$, which is substantially stronger than what the lower bound obtained in Corollary 2.3 requires. Moreover, if we restrict ρ to be bounded away from 1, then the GLRT may be powerless.

THEOREM 3.1. *Let $\mathcal{C}$ be the class of all k-sets. If $\rho < 0.6$ and $k = o(n^{0.7})$, the GLRT has a Bayes risk bounded away from zero.*

The proof is in the Appendix.

In view of Theorem 3.1, the GLRT is clearly suboptimal when in the situation stated there, and compares very poorly with the squared-sum test, which is asymptotically powerful if $\rho k^2/n \to \infty$ as seen in Proposition 3.1. We do not know of any other situation where the GLRT fails so miserably.

3.3. *A localized squared-sum test.* While the GLRT is near-optimal for detecting objects from a parametric class such as k-intervals, it needs knowledge of ρ. However, a simple modification solves this drawback. Indeed, consider the following "local" squared-sum test:

$$f(X) = 0 \quad \text{if and only if} \quad \max_{S\in\mathcal{C}} \Big(\sum_{i\in S} X_i\Big)^2 \le t$$

for some appropriate threshold t.

PROPOSITION 3.4. *Consider a class $\mathcal{C}$ of sets of size k, with cardinality $N \to \infty$ such that $\log(N)/k \to 0$. When $\rho \gg \log(N)/k$ in (1.4), the local squared-sum test with threshold $t = 2k\log N$ has worst-case risk tending to zero.*

PROOF. The proof is quite straightforward. Indeed, under the null, for any S of size k we have $\sum_{i\in S} X_i \sim \mathcal{N}(0,k)$ so that

$$\max_{S\in\mathcal{C}} \Big(\sum_{i\in S} X_i\Big)^2 \le t$$

with probability tending to 1. Under an alternative (1.4), S denoting the anomalous set of variables, we have

$$\mathbb{P}\Big(\Big(\sum_{i\in S} X_i\Big)^2 \ge t\Big) \ge \mathbb{P}\big((k + k(k-1)\rho)\chi_1^2 \ge t\big) \to 1,$$

when $\rho \gg \log(N)/k$. □

Specializing this result to the case of k-intervals leads to the following statement (which ignores logarithmic factors):

The localized squared-sum test is near-optimal for detecting k-intervals in the regime where $\log(n)/k \to 0$.

When k is unknown. We might only know that some interval is anomalous, without knowing the size of that interval. In that case, multiple testing at each k using the local squared-sum test yields adaptivity. Computationally, this may be done effectively by computing sums in a multiscale fashion as advocated in [6]. In fact, here it is enough to compute the sums over all *dyadic* intervals—since each interval S contains a dyadic interval of length at least $|S|/4$—and this can be done in $3n$ flops in a recursive fashion.

3.4. *A goodness-of-fit test.* By now, the parametric case is essentially solved, with the local squared-sum test being not only near-optimal but also computable in polynomial time (in n and k) for the case of k-intervals, for example. In the nonparametric case, so far, the story is not complete. We focus on the class of all k-sets. There we know that the squared-sum test is near-optimal if $k^2/n \to \infty$. If $k^2/n \to 0$, it has no power, and we only know that the GLRT works when $(1-\rho)(n/k)^2 \log(n/k) \to 0$, which does not match the rate obtained in Corollary 2.3. Worse than that, it is not clear whether computing the GLRT is possible in time polynomial in (n,k). We now show that a simple goodness-of-fit (GOF) test performs (almost) as desired.

The basic idea is the following. Let $H_i = \Phi^{-1}(X_i)$, where Φ is the standard normal distribution function. Under the null, the H_i's are i.i.d. uniform in $(0,1)$. Under an alternative with anomalous set denoted by S, the $X_i, i \in S$ are closer together, especially since we place ourselves in the regime where $\rho \to 1$. More precisely, we have the following.

LEMMA 3.1. *Suppose $X_1, \dots, X_k$ are zero-mean, unit-variance random variables satisfying* $\mathrm{Cov}(X_i, X_j) \ge \rho > 0$, *for all $i \ne j$. Let $\overline{X}$ denote their average. Then for any $t > 0$,*

$$\mathbb{P}\{\#\{i : |X_i - \overline{X}| > t\} \ge k/2\} \le \frac{2(1-\rho)}{t^2}.$$

PROOF. Let $\Lambda := \sum_{i \ne j} \mathrm{Cov}(X_i, X_j) \ge k(k-1)\rho$. Elementary calculations show that

$$\mathbb{E}\left[\frac{1}{k}\sum_i (X_i - \overline{X})^2\right] = 1 - \frac{1}{k} - \frac{\Lambda}{k^2} \le (1 - 1/k)(1-\rho) \le 1 - \rho.$$

By Markov's inequality, we then have

$$\mathbb{P}\left\{\frac{1}{k}\sum_i (X_i - \overline{X})^2 > t^2/2\right\} \le \frac{2(1-\rho)}{t^2}.$$

The statement follows from observing that

$$\#\{i : |X_i - \overline{X}| > t\} \geq k/2 \quad \Rightarrow \quad \frac{1}{k}\sum_i (X_i - \overline{X})^2 > t^2/2. \qquad \square$$

The idea, therefore, is detecting unusually high concentrations of H_i's, which is a form of GOF test for the uniform distribution. Under a general correlation model as in (1.4), with Lemma 3.1 we see that the concentration will happen over an interval of length slightly larger than $\sqrt{1-\rho}$. This is apparent from Lemma 1.1 under the simple correlation model (1.1).

Choose an integer m such that $m \gg (n/k^2)\log(n/k^2)$ and partition the interval $[0, 1]$ into m bins of length $1/m$, denoted $I_s, s = 1, \ldots, m$. Let $B_s = \#\{i : H_i \in I_s\}$ be the bin counts—thus, we are computing a histogram. Then consider the following GOF test:

$$f(X) = 0 \quad \text{if and only if} \quad \max_{s=1,\ldots,m} B_s \leq t,$$

where t is some threshold.

PROPOSITION 3.5. *Consider the class $\mathcal{C}$ of all k-sets in the case where $k^2/n \to 0$ and $k/\log n \to \infty$. In the GOF test above, choose m such that $(n/k^2)\log n \ll m \ll n/\log n$. When $(1-\rho)^{1/2} \ll 1/m$ in (1.4), the resulting test with threshold $t = n/m + \sqrt{3n\log(m)/m}$ has worst-case risk tending to zero.*

PROOF. Bernstein's inequality, applied to the binomial distribution, gives that

$$\mathbb{P}_0\{B_s > n/m + b\sqrt{n/m}\} \leq \exp\bigl[-(b^2/2)/\bigl(1 + (b/3)\sqrt{m/n}\bigr)\bigr].$$

This and the union bound imply that, indeed,

$$\mathbb{P}_0\Bigl\{\max_s B_s > t\Bigr\} \to 0.$$

Consider now an alternative of the form (1.4), with S denoting the anomalous set. Let

$$I := \{i \in S : |X_i - \overline{X}_S| \leq 1/m\}, \qquad \overline{X}_S := \frac{1}{k}\sum_{i\in S} X_i.$$

Though the set I is random, by Lemma 3.1 and the fact that $(1-\rho)^{1/2} \ll 1/m$, we have that

$$\mathbb{P}_S\{|I| \geq k/2\} \to 1.$$

Define the event $Q := \{-a \leq \overline{X}_S \leq a\}$ for some $a > 0$. Note that, since the variance of $\overline{X}_S$ is bounded by 1, $\mathbb{P}(Q^c) \leq 2(1 - \Phi(a))$. Define $\tilde{H}_S = \Phi^{-1}(\overline{X}_S)$. On Q, using a simple Taylor expansion, we have

$$|H_i - \tilde{H}_S| \leq \frac{|X_i - \overline{X}_S|}{\phi(a + 1/m)} \leq e^{a^2}/m \qquad \forall i \in I,$$

where ϕ denotes the standard normal density function and a is taken sufficiently large. Therefore, when $|I| \geq k/2$ and Q hold, at least $k/2$ of the anomalous H_i's fall in an interval of length at most $2e^{a^2}/m$. Since such an interval is covered by at most $2e^{a^2}$ bins, by the pigeonhole principle, there is a bin that contains $ke^{-a^2}/4$ anomalous H_i's. By Bernstein's inequality, the same bin will also contain at least $(n-k)/m - \sqrt{3n\log(m)/m}$ nonanomalous H_i's (with high probability), so in total this bin will contain $n/m - k/m - \sqrt{3n\log(m)/m} + ke^{-a^2}/4$ points. By our choice of m, $k \gg \sqrt{n\log(m)/m}$, so it suffices to choose $a \to \infty$ slowly enough that $ke^{-a^2} \gg \sqrt{n\log(m)/m}$ still. Then, with high probability, there is a bin with more than t points. $\square$

Ignoring logarithmic factors, we are now able to state the following:

The GOF test is near-optimal for detecting k-sets in the regime where $k^2/n \to 0$ and $k/\log n \to \infty$.

When $k/\log n \to 0$, things are somewhat different. There, the GOF test requires that $(1-\rho)n^{2k/(k-1)} \to 0$, which is still close to optimal when $k \to \infty$, but far from optimal when k is bounded (e.g., when $k=2$, the exponent is 4 instead of 2). Indeed, when $k/\log n \to 0$, m needs to be chosen larger than n, and Bernstein's inequality is not accurate. Instead, we use the simple bound

$$\mathbb{P}\big(\mathrm{Bin}(n,p) \geq \ell\big) \leq 2\frac{(np)^\ell}{\ell!} \qquad \text{when } np \leq 1/2.$$

Note that Bennett's inequality would also do. (The analysis also requires some refinement showing that, with probability tending to 1 under the alternative, one cell contains at least k points.) Note that in the remaining case, $k = O(1)$, the GLRT is optimal up to a logarithmic factor, since it only requires that $(1-\rho)n^2\log n \to 0$, as seen in Section 3.2.2. We do not know whether a comparable performance can be achieved by a test that does not have access to ρ.

When k is unknown. In essence, we are trying to detect an interval with a higher mean in a Poisson count setting. As before, it is enough to look at dyadic intervals of all sizes, which can be done efficiently as explained earlier, following the multiscale ideas in [6].

APPENDIX: PROOF OF THEOREM 3.1

The proof is divided into three steps. The first step formalizes the fact that we want to prove that (under H_1), the contaminated set has no influence (with high probability) on the GLRT statistic. The second step exhibits a useful high probability event. Finally, in the third step we show that on this high probability event, the contaminated set has no influence on the GLRT.

It can easily be seen that for every S of size k,

$$X^T(\mathbf{I} - \mathbf{A}_S^{-1})X = \frac{\rho}{(1+\rho(k-1))(1-\rho)}\Big(\sum_{i,j\in S, i\neq j} X_i X_j - \rho(k-1)\sum_{i\in S} X_i^2\Big).$$

Introduce the function $g:\mathbb{R}^k \to \mathbb{R}$ defined by

$$g(u) = \sum_{i\neq j} u_i u_j - \rho(k-1)\sum_i u_i^2 = \left(\sum_{i=1}^n u_i\right)^2 - \big(1+\rho(k-1)\big)\sum_{i=1}^n u_i^2$$

for $u = (u_1, \ldots, u_k) \in \mathbb{R}^k$. Denoting, for $x \in \mathbb{R}^n$ and $S \subset \{1, \ldots, n\}$, the vector of components of x belonging to S by $x|_S$, we may write the GLRT as

$$f(x) = 0 \quad \text{if and only if} \quad \max_{S\in\mathcal{C}} g(x|_S) < t.$$

Note that by the symmetry of $\mathcal{C}$ and the test,

$$\begin{aligned} R(f) &= \mathbb{P}_0\Big\{\max_{S\in\mathcal{C}} g(X|_S) \geq t\Big\} + \frac{1}{N}\sum_{S'\subset\mathcal{C}} \mathbb{P}_{S'}\Big\{\max_{S\in\mathcal{C}} g(X|_S) < t\Big\} \\ &= \mathbb{P}_0\Big\{\max_{S\in\mathcal{C}} g(X|_S) \geq t\Big\} + \mathbb{P}_{\{1,\ldots,k\}}\Big\{\max_{S\in\mathcal{C}} g(X|_S) < t\Big\}. \end{aligned}$$

Given $X \sim \mathcal{N}(0, \mathbf{I})$, define the coupling X' as follows: $X_i = X_i'$ for $i \notin \{1, \ldots, k\}$, and X_i, X_i' are independent for $i \in \{1, \ldots, k\}$. Note that $X' \sim \mathcal{N}(0, \mathbf{A}_{\{1,\ldots,k\}})$. Then, no matter what the threshold t is, we have

$$\begin{aligned} R(f) &= \mathbb{P}\Big\{\max_{S\in\mathcal{C}} g(X|_S) \geq t\Big\} + \mathbb{P}\Big\{\max_{S\in\mathcal{C}} g(X'|_S) < t\Big\} \\ &\geq \mathbb{P}\Big\{\max_{S\in\mathcal{C}} g(X|_S) \geq \max_{S\in\mathcal{C}} g(X'|_S)\Big\}. \end{aligned}$$

In the following we show that, with probability tending to 1, we have

$$\max_{S\in\mathcal{C}} g(X|_S) = \max_{S\in\mathcal{C}} g(X'|_S),$$

which then implies that the GLRT is asymptotically powerless.

By Lemma 1.1, there exist $U, U_1, \ldots, U_k$ independent standard normal such that for all $i \in \{1, \ldots, k\}$,

$$X_i' = \sqrt{\rho}U + \sqrt{1-\rho}U_i.$$

Using the fact that $\max_{i=1,\ldots,k} |U_i| \leq \sqrt{2\log k}$ with high probability, with probability tending to 1, we have

$$X_1', \ldots, X_k' \in [-\zeta, \zeta],$$

where $\zeta := \sqrt{2(1-\rho)\log(\omega_k k)}$ and ω_k is any sequence such that $\omega_k \to \infty$.

Fix $\gamma > 1$ to be determined later and define $p = \mathbf{P}\{\zeta \leq U \leq \gamma\zeta\}$ where $U \sim \mathcal{N}(0,1)$. By the fact that $X_1, \ldots, X_n$ are i.i.d. standard normal, $Z := \#\{i : \zeta \leq X_i \leq$

$\gamma\zeta\} \sim \mathrm{Bin}(n, p)$, so that $\mathbf{P}\{Z \geq k\} \to 1$ if $k = o(np)$. When γ is bounded away from 1, this is the case if $\sqrt{\log k}k^{2-\rho} = o(n)$.

In conclusion, we proved that the event

$$\Omega = \{X_1', \ldots, X_k' \in (-\zeta, \zeta), \exists \alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k \in \{1, \ldots, n\} \text{ distinct:} \\ X_{\alpha_1}, \ldots, X_{\alpha_k}, -X_{\beta_1}, \ldots, -X_{\beta_k} \in (\zeta, \gamma\zeta)\}$$

has a probability that tends to 1 if $\sqrt{\log k}k^{2-\rho} = o(n)$ as long as γ is bounded away from 1.

We specify $\gamma = 1/\sqrt{\rho + (\frac{1}{k-1} + \rho)^2}$. Note that, as required, γ exceeds and is bounded away from 1. Assume that we are on the event Ω. First note that

$$\begin{aligned} g(X_{\alpha_1}, \ldots, X_{\alpha_k}) &\geq k(k-1)\zeta^2 - \rho(k-1)k\gamma^2\zeta^2 \\ &= k(k-1)\zeta^2(1 - \rho\gamma^2), \end{aligned} \tag{A.1}$$

and the same holds for $g(X_{\beta_1}, \ldots, X_{\beta_k})$.

Let $S \in \mathcal{C}$ be such that $S \cap \{1, \ldots, k\} \neq \varnothing$. We want to show that there exists S' such that $g(X|_{S'}) \geq g(X'|_S)$. This entails that $\max_{S \in \mathcal{C}} g(X|_S) \geq \max_{S \in \mathcal{C}} g(X'|_S)$, since for $S \cap \{1, \ldots, k\} = \varnothing$ we have $g(X|_S) = g(X'|_S)$. First remark that we can assume that

$$\left(\sum_{i \in S} X_i'\right)^2 \geq \zeta(k-1)\sqrt{1 - \rho\gamma^2}, \tag{A.2}$$

since otherwise by (A.1) we can simply take $S' = \{\alpha_1, \ldots, \alpha_k\}$. To simplify notation, we may assume that $1 \in S \cap \{1, \ldots, k\}$. By definition of Ω and the fact that S contains at least one index in $\{1, \ldots, k\}$, there exist $u, v \in \{1, \ldots, k\}$ such that X_{α_u} and X_{β_v} do not appear in $X'|_S$. We want to show that by replacing X_1' by either X_{α_u} or X_{β_v}, in $X'|_S$, one increases the value of g. More precisely, we want to show that

$$\max\bigl(g\bigl(X_{\alpha_u}, X'|_{S\setminus\{1\}}\bigr), g\bigl(X_{\beta_v}, X'|_{S\setminus\{1\}}\bigr)\bigr) \geq g(X'|_S).$$

Then by induction one can show the existence of the S' described above.

Note that, for $x \in \mathbb{R}^k$ and $y \in \mathbb{R}$,

$$\begin{aligned} &g(x_1, \ldots, x_{j-1}, y, x_{j+1}, \ldots, x_k) - g(x) \\ &\quad = 2(y - x_j)\sum_{i \neq j} x_i - \rho(k-1)(y^2 - x_j^2) \\ &\quad = (y - x_j)\left(2\sum_{i=1}^{k} x_i - \bigl(2 + \rho(k-1)\bigr)x_j - \rho(k-1)y\right). \end{aligned}$$

Consider the case where $\sum_{i \in S} X_i' > 0$ (the case $\sum_{i \in S} X_i' < 0$ can be dealt with similarly). Since $X_{\alpha_u} \geq X_1'$, it suffices to show that $2\sum_{i \in S} X_i' \geq (2 + \rho(k-1))X_1' +$

$\rho(k-1)X_{\alpha_u}$, which follows from

$$\begin{aligned}\big(2+\rho(k-1)\big)X_1' + \rho(k-1)X_{\alpha_u} &\le (k-1)\zeta\gamma\left(\frac{2}{k-1}+2\rho\right)\\ &= 2(k-1)\zeta\sqrt{1-\rho\gamma^2}\\ &\le 2\sum_{i\in S} X_i.\end{aligned}$$

This concludes the proof.

Acknowledgments. We thank Omiros Papaspiliopoulos for his illuminating remarks and the anonymous referees for challenging us to obtain stronger results in the sparse setting and for pointing out a mistake in Proposition 3.5.

REFERENCES

[1] CBOE S&P 500®Implied Correlation Index. Available at http://www.cboe.com/micro/impliedcorrelation.

[2] ADDARIO-BERRY, L., BROUTIN, N., DEVROYE, L. and LUGOSI, G. (2010). On combinatorial testing problems. *Ann. Statist.* **38** 3063–3092. MR2722464

[3] ANANDKUMAR, A., TONG, L. and SWAMI, A. (2009). Detection of Gauss–Markov random fields with nearest-neighbor dependency. *IEEE Trans. Inform. Theory* **55** 816–827. MR2597269

[4] ARIAS-CASTRO, E., CANDÈS, E. J. and DURAND, A. (2011). Detection of an anomalous cluster in a network. *Ann. Statist.* **39** 278–304. MR2797847

[5] ARIAS-CASTRO, E., CANDÈS, E. J., HELGASON, H. and ZEITOUNI, O. (2008). Searching for a trail of evidence in a maze. *Ann. Statist.* **36** 1726–1757. MR2435454

[6] ARIAS-CASTRO, E., DONOHO, D. L. and HUO, X. (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory* **51** 2402–2425. MR2246369

[7] BARAUD, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606. MR1935648

[8] BERMAN, S. M. (1962). Equally correlated random variables. *Sankhyā Ser. A* **24** 155–156. MR0145564

[9] BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

[10] BOUTSIKAS, M. V. and KOUTRAS, M. V. (2006). On the asymptotic distribution of the discrete scan statistic. *J. Appl. Probab.* **43** 1137–1154. MR2274642

[11] CAI, T., JENG, X. and JIN, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 629–662. MR2867452

[12] CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

[13] CROSS, G. R. and JAIN, A. K. (1983). Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **5** 25–39.

[14] DESOLNEUX, A., MOISAN, L. and MOREL, J.-M. (2003). Maximal meaningful events and applications to image analysis. *Ann. Statist.* **31** 1822–1851. MR2036391

[15] DEVROYE, L., GYÖRGY, A., LUGOSI, G. and UDINA, F. (2011). High-dimensional random geometric graphs and their clique number. Unpublished manuscript.

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

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

[18] INGSTER, Y. I. (1998). Minimax detection of a signal for l^n-balls. *Math. Methods Statist.* **7** 401–428. MR1680087

[19] JIN, J. (2003). Detecting and estimating sparse mixtures. Ph.D. thesis, Stanford Univ.

[20] KAILATH, T. and POOR, H. V. (1998). Detection of stochastic processes. *IEEE Trans. Inform. Theory* **44** 2230–2259. Information theory: 1948–1998. MR1658799

[21] LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. MR2135927

[22] PERONE PACIFICO, M., GENOVESE, C., VERDINELLI, I. and WASSERMAN, L. (2004). False discovery control for random fields. *J. Amer. Statist. Assoc.* **99** 1002–1014. MR2109490

[23] RAMÍREZ, D., VÍA, J., SANTAMARÍA, I. and SCHARF, L. L. (2010). Detection of spatially correlated Gaussian time series. *IEEE Trans. Signal Process.* **58** 5006–5015. MR2722660

E. ARIAS-CASTRO
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
SAN DIEGO, CALIFORNIA 92093
USA
E-MAIL: eariasca@math.ucsd.edu

S. BUBECK
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08542
USA
E-MAIL: sbubeck@princeton.edu

G. LUGOSI
ICREA
AND
DEPARTMENT OF ECONOMICS
POMPEU FABRA UNIVERSITY
BARCELONA
SPAIN
E-MAIL: gabor.lugosi@upf.edu