Detection of correlations *

Ery Arias-Castro† Sébastien Bubeck‡ Gábor Lugosi§

Abstract

We consider the hypothesis testing problem of deciding whether an observed high-dimensional normal vector has independent 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 Bayes 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 settings, while the generalized likelihood ratio test is suboptimal in some important cases.

Keywords: Sparse covariance matrix; minimax detection; Bayesian detection; scan statistic; generalized likelihood ratio test.

1 Introduction

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 including signal processing, remote sensing, finance, image processing, etc.

1.1 Setting and notation

Here 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

*The first author's work was partially supported by ONR grant N00014-09-1-0258. The third author is supported of the Spanish Ministry of Science and Technology grant MTM2009-09063 and PASCAL2 Network of Excellence under EC grant no. 216886.

†Department of Mathematics, University of California, San Diego

‡Centre de Recerca Matemática, Barcelona

§ICREA and Department of Economics, Universitat Pompeu Fabra

$\{1, \ldots, n\}$, 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

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

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

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

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 $\mathcal{N}(0, \boldsymbol{A}_S)$.

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 emblematic 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 esthetic 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. We have a fairly complete understanding of this case as described below in this paper.

- **k-sets.** In this example, we consider the class of all sets of size k, i.e., 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, for example, in genetics. We show that when $\rho \in (0, 1)$ is a constant (e.g., $\rho = 1/2$) then it is possible to have a risk converging to zero (as $n \to \infty$) when $k^2/n \to \infty$ but no test has a vanishing risk when $\limsup_{n\to\infty} k^2/n < \infty$.

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, i.e., $X = (X_i : i \in \{1, \ldots, m\}^d)$, so that the sample size is $n = m^d$, and we consider the class of all hyperrectangles of the form $\times_{s=1}^d \{i_s, \ldots, i_s + k_s - 1\}$—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 non-trivial combinatorial structure, which makes the problem more interesting. Our results imply that it is impossible to detect the existence of a "correlated" perfect matching with a risk converging to zero unless $\rho \to 1$ as $n \to \infty$.

- **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 we will see, the situation is quite similar to that of the previous example.

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. There are (at least) two natural ways of measuring the risk of a test f. The first corresponds to a *Bayesian* point of view in which under H_1, an element $S \in \mathcal{C}$ is selected at random and X is drawn according to $\mathbb{P}_S$. (We restrict ourselves to the uniform prior on $\mathcal{C}$.) In this case, the *average* risk of the test is

$$R(f) = \mathbb{P}_0\{f(X) = 1\} + \frac{1}{N}\sum_{S\in\mathcal{C}} \mathbb{P}_S\{f(X) = 0\} = \mathbb{P}_0\{f(X) = 1\} + \mathbb{P}_1\{f(X) = 0\}$$

where we write $N = |\mathcal{C}|$ for the cardinality of $\mathcal{C}$ and $\mathbb{P}_1 = (1/N)\sum_{S\in\mathcal{C}} \mathbb{P}_S$ for the distribution of X under the alternative hypothesis, which is a mixture over $\mathcal{C}$. The last expression for the Bayesian risk effectively shows that we are testing $\mathbb{P}_0$ versus $\mathbb{P}_1$, a simple versus simple hypothesis testing problem, so that the likelihood ratio test f^* is optimal by the Neyman-Pearson fundamental lemma. Introducing

$$Z_S = \exp\left(\frac{1}{2}X^T(\boldsymbol{I} - \boldsymbol{A}_S^{-1})X\right), \tag{1.1}$$

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

$$L(X) = \frac{1}{N}\sum_{S\in\mathcal{C}} \frac{Z_S}{\mathbb{E}_0 Z_S}, \tag{1.2}$$

and the optimal test becomes

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

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

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

A natural alternative measure of the risk corresponds to a "worst-case" point of view. We may define

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

Clearly, for any test, $R(f) \leq R^{\max}(f)$ and in many cases when $\mathcal{C}$ and f are sufficiently symmetric, the two risk measures coincide. In this paper we restrict our attention to $R(f)$ but most results carry through without any difficulty to $R^{\max}(f)$. In fact, the uniform prior is the least favorable for k-sets, and nearly so for k-intervals.

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 non-asymptotic 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 $R^* \to 0$ (resp. $\to 1$) as $n \to \infty$.

Remark. (MONOTONICITY OF THE BAYES RISK.) One may intuitively expect that the larger the class $\mathcal{C}$, the more difficult the detection problem becomes. This is obvious when one considers the worst-case risk $R^{\max}(f)$ but it is not necessarily true for the average risk $R(f)$ in general. However, it is the case for sufficiently symmetric classes as the following little argument shows. To emphasize the dependence on the class $\mathcal{C}$, write $R_{\mathcal{C}}(f)$ and $R^{\max}_{\mathcal{C}}(f)$ for the average and worst-case risks of a test f and let $f^*_{\mathcal{C}}$ denote the optimal test for $\mathcal{C}$. Let $\mathcal{A} \subset \mathcal{C}$ and suppose that $\mathcal{C}$ is symmetric in the sense that $R_{\mathcal{C}}(f^*_{\mathcal{C}}) = R^{\max}_{\mathcal{C}}(f^*_{\mathcal{C}})$. In this case, we have

$$R_{\mathcal{C}}(f^*_{\mathcal{C}}) = R^{\max}_{\mathcal{C}}(f^*_{\mathcal{C}}) \geq R^{\max}_{\mathcal{A}}(f^*_{\mathcal{C}}) \geq R_{\mathcal{A}}(f^*_{\mathcal{C}}) \geq R_{\mathcal{A}}(f^*_{\mathcal{A}})$$

and therefore the Bayes risk is indeed monotone. See (Addario-Berry et al., 2010), which displays an example of a non-symmetric class for which monotonicity does not hold in the detection-of-means setting.

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 below), it is clearly a restrictive assumption. Most results of the paper have an easy generalization. Let $0 < \rho_{\min} < \rho_{\max} \leq 1$. Then we may modify the description alternative hypothesis such that tho covariance matrix is any matrix such that

$$(\boldsymbol{A}_S)_{i,j} \begin{cases} = 1 & i = j \\ \in (\rho_{\min}, \rho_{\max}) & i \neq j, \text{ with } i, j \in S; \\ = 0 & \text{otherwise.} \end{cases}$$

Then for any prior on the class of such alternatives, the lower bound that we establish for the Bayes risk (Theorem 1) applies when ρ is replaced by $\rho_{\max}$, and the detection performances we derive for the various tests we consider remain valid when ρ is replaced with $\rho_{\min}$. To keep the notation and the framework simple, we do not detail these trivial modifications. However, dealing with more general correlation structures remains an interesting and important challenge.

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, (Arias-Castro et al., 2011, 2005; Boutsikas and Koutras, 2006; Desolneux et al., 2003; Perone Pacifico et al., 2004). A more general framework that includes the detection of perfect matchings and spanning trees is investigated in (Addario-Berry et al., 2010), and the detection of k-sets is studied in a number of papers, e.g., (Baraud, 2002; Donoho and Jin, 2004; Hall and Jin, 2010; Ingster, 1999; Jin, 2003). 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. That said, there is a close relationship between the detection-of-means setting and our *detection-of-correlations* setting, particularly in view of the representation theorem of Berman (1962)—stated here for the case Gaussian random variables.

Lemma 1 (Berman (1962)) *Let $X_1, \dots, X_k$ be standard normal with $Cov(X_i, X_j) = \rho$ for $i \neq j$. Then there are i.i.d. standard normal random variables $U, U_1, \dots, U_k$ 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 $\sqrt{\rho}\, U$) 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 (Kailath and Poor, 1998). 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 (Bickel and Levina, 2008; Cai et al., 2010), from the literature on multivariate time series (Ramírez et al., 2010), or on other approaches (Devroye et al., 2011); 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 Anandkumar et al. (2009), 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 are 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. On the other hand, the *generalized likelihood ratio test (GLRT)*, which rejects for large values of $\max_{S \in \mathcal{C}} \sum_{i \in S} X_i$, is near-optimal for the detection-of-means problem, see, for example, (Arias-Castro et al., 2005). In this paper, we show that this statement is no longer true in the detection-of-correlations setting, see Theorem 2.

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 (Baraud, 2002; Ingster, 1999; Jin, 2003) and optimal tests have been developed, such as the "higher criticism" (Donoho and Jin, 2004; Hall and Jin, 2010). 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.

1.3 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. 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 $(\sum_i X_i)^2$) and the generalized likelihood ratio 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. The fine details of this case are not completely understood yet, though the case $\rho \approx 1$ is perhaps less interesting from the point of view of applications. The analysis of tests such as the squared-sum test and the GLRT involve handling quadratic forms in X. This is technically more challenging than the analogous problem for the detection-of-means setting in which only linear function of X appear (which are normal random variables).

2 Lower bounds for the Bayes risk

In this section we investigate lower bounds on the Bayes risk. 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. However, this approach involves rather complicated computations and it results in rather weak conclusions. We then develop a new bound based on Lemma 1 that has powerful implications, leading to 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 $S = \{1, \ldots, k\}$. In this case, the alternative hypothesis is simple and the likelihood ratio (Neyman-Pearson) test may be expressed by

$$f^*(X) = 0 \quad \text{if and only if} \quad X^T(\boldsymbol{I} - \boldsymbol{A}_S^{-1})X \leq \frac{1}{2}\log\det(\boldsymbol{A}_S) \ .$$

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

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

Lemma 2 *Under $\mathbb{P}_0$, $X^T(\boldsymbol{I} - \boldsymbol{A}_S^{-1})X$ 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 $Y = (Y_1, \ldots, Y_n)$ denotes a standard normal vector, then under H_0, the quadratic form $X^T(\boldsymbol{I} - \boldsymbol{A}_S^{-1})X$ is distributed as $Y^T(\boldsymbol{I} - \boldsymbol{A}_S^{-1})Y$, and under the alternative, it has the distribution of $Y^T(\boldsymbol{A}_S - \boldsymbol{I})Y$, since X is distributed as $\boldsymbol{A}_S^{1/2}Y$.

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

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

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

The lemma follows from this simple representation and the fact that $\boldsymbol{A}_S$ has eigenvalue $1-\rho$ with multiplicity k, $1+\rho(k-1)$ 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 1 *In the setting where $\mathcal{C}$ 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 $\mathbb{P}_0\{X^T(\boldsymbol{I} - \boldsymbol{A}_S^{-1})X \geq \tau_k\} \to 0$ and $\mathbb{P}_S\{X^T(\boldsymbol{I} - \boldsymbol{A}_S^{-1})X < \tau_k\} \to 0$. We use Lemma 2 and the fact that, by Chebyshev's inequality,

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

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

$$\mathbf{P}\left\{t_k^{-1} < \chi_1^2 < t_k\right\} \to 1, \quad \text{as } k \to \infty.$$

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

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

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

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

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 $S = \{1, \ldots, k\}$.

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 (1.2). 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 $\Lambda = \det(\boldsymbol{A}_S) = (1-\rho)^{k-1}(1 + \rho(k-1))$, 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 (1.1). 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 3 *Suppose X is standard normal vector in $\mathbb{R}^n$ and $\boldsymbol{M}$ is an $n \times n$ symmetric matrix with eigenvalues strictly less than $1/2$. Then*

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

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

Since $\boldsymbol{M} := \boldsymbol{I} - \boldsymbol{A}_S^{-1}$ has eigenvalue $-\rho/(1-\rho)$ with multiplicity k, eigenvalue $\rho(k-1)/(1+\rho(k-1))$ with multiplicity 1, and eigenvalue 0 with multiplicity $n-k$, $\mathbb{E}_0[Z_S^2] = \mathbb{E}_0 \exp\left(X^T \boldsymbol{M} X\right) = +\infty$ unless $\rho(k-1) < 1$. The implications are rather insubstantial. It only shows that, when $\rho(k-1)$ is bounded away from 1 from above, the Bayes risk does not tend to zero. 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. This trick is used in (Ingster, 1999; Jin, 2003) 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. In any case, the computations are rather intricate 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 (Addario-Berry et al., 2010; Arias-Castro et al., 2011, 2008), 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.

Theorem 1 *For any class $\mathcal{C}$ and any $a > 0$,*

$$R^* \geq \mathbf{P}\left\{|\mathcal{N}(0,1)| \leq a\right\} \left(1 - \frac{1}{2}\sqrt{\mathbb{E}\exp\left(\nu_a Z\right) - 1}\right),$$

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 $a = 1$,

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

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

Proof. The starting point of the proof is Lemma 1[1], 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 $U, U_1, \ldots, U_n$ 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 likelihood ratio, and the

[1] In fact, we only need to assume that X is as described, since what matters is the distribution of X.

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

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

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

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

and therefore, using the Cauchy-Schwarz inequality,

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

Since

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

we get

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

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

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

which concludes the proof. □

We apply Theorem 1 to a few examples. The theorem converts the problem into a purely combinatorial question and Addario-Berry et al. (2010) offer various estimates for the moment generating function of Z which we may use for our purposes.

2.3.1 Non-overlapping sets

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

Corollary 1 *Let $\mathcal{C}$ be the class of all sets of size k. If*

$$\nu = \rho/(1+\rho) - \frac{1}{2}\log(1-\rho^2) \leq \frac{\log(N)}{k},$$

then the Bayes risk satisfies $R^* \geq 0.3$, and $R^* \to 1$ if $\rho = o((1/k)\log N)$ and is bounded away from 1.

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,

$$\mathbb{E}e^{\nu Z} - 1 = (1/N)(e^{\nu k} - 1) \leq (1/N)e^{\nu k},$$

which is bounded by 1 if $\nu \leq (1/k)\log(N)$. The first part then follows from the second part of Theorem 1. For the second part, we need to find $a \to \infty$ such that $\nu_a = o((1/k)\log N)$. (note that in this case the upper bound above tends to zero). First note that under the assumptions that $\rho = o((1/k)\log N)$ and is bounded away from 1, one always has $-\frac{1}{2}\log(1-\rho^2) = o((1/k)\log N)$ (consider separately the cases when $(1/k)\log N)$ remains bounded and when it is unbounded). Thus one can simply take $a \to \infty$ such that $\rho a^2 = o((1/k)\log N)$, and the result follows from the first part of Theorem 1. □

2.3.2 k-intervals

Consider the class of all k-intervals. The situation is similar to that of non-overlapping sets. (In fact, since this class of k-intervals contains $[n/k]$ non-overlapping sets of size k and the class is sufficiently symmetric, we could use the monotonicity property of the Bayes risk, mentioned at the end of Section 1.1, to immediately deduce a lower bound on the worst-case risk via Corollary 1.)

Corollary 2 *Let $\mathcal{C}$ be the class of all k-intervals. If*

$$\nu = \rho/(1+\rho) - \log(1-\rho^2) \leq \frac{\log(n/(2k))}{k}$$

then the Bayes risk satisfies $R^* \geq 0.3$, and $R^* \to 1$ if $\rho = o((1/k)\log(n/k))$ and is bounded away from 1.

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

$$\mathbf{P}\left\{|S \cap S'| = \ell\right\} = \frac{2}{N}, \quad \forall \ell = 1, \ldots, k.$$

Thus,

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

which is bounded by 1 if $\nu \leq (1/k)\log(N/(2k))$ and tends to zero if $\nu = o((1/k)\log(N/(2k)))$. Reasoning as in the proof of Corollary 1, the result follows from Theorem 1 and the fact that $N \leq n$. □

2.3.3 k-sets

Consider the class of all sets of size k.

Corollary 3 *Let $\mathcal{C}$ be the class of k-sets. Denote $\nu = \rho/(1+\rho) - \log(1-\rho^2)$. If*

$$\frac{k^2}{n} \leq \frac{\ln 2}{e^\nu - 1}$$

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

The result implies that, if $k^2/n \to 0$, reliable detection is impossible is ρ is remains bounded away from 1.

Proof. By (Addario-Berry et al., 2010, Proposition 3.4), which uses negative association,

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

which is bounded by 2 under the postulated condition, and tends to one if either $k^2/n \to \infty$ and $\nu k^2/n \to 0$, or $k^2/n \to 0$ and $e^\nu k^2/n \to 0$. The result follows from Theorem 1—for the case $k^2/n \to 0$ and $e^\nu k^2/n \to 0$, we use the fact that $\nu \sim -(1/2)\log(1-\rho)$ as $\rho \to 1$. □

2.3.4 Perfect matchings

Consider now the example of perfect matchings described in the introduction. Here $k = \sqrt{n}$. Once again, Theorem 1 applies and implies that testing is impossible for moderate values of ρ.

Corollary 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 Addario-Berry et al. (2010), 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 one if $\nu \to 0$. We then apply Theorem 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 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 (Addario-Berry et al., 2010) 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 one if $\nu \to 0$. We then apply Theorem 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 two 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. We consider the squared-sum test, which corresponds to the ANOVA test in the detection-of-means setting, and the generalized likelihood ratio 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 test 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 $(n + \rho k(k-1))\chi_1^2$. This immediately leads to the following result.

Proposition 2 *Let $\mathcal{C}$ be an arbitrary class of sets of size k and suppose that $\rho k^2/n \to \infty$. 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 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 Corollary 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 squared-norm test is near-optimal for detecting k-sets in the regime $k^2/n \to \infty$. A similar phenomenon occurs in the detection-of-means setting, where the test based on $\sum_{i=1}^n X_i$ is optimal for detecting k-sets in the same regime. In the regime $k^2/n \to 0$, the squared-sum test is powerless even if $\rho = 1$.

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 nonparametric classes such as the class of all 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

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

for some appropriately chosen t.

Our analysis of the GLRT is based on Lemma 2, which provides the distribution of the quadratic form $X^T(\boldsymbol{I} - \boldsymbol{A}_S^{-1})X$ 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-square distributions.

3.2.1 The GLRT for k-intervals and other parametric classes

Recalling Corollary 2, when detecting k-intervals all tests are asymptotically powerless when $\rho k/\log(n/k) \to 0$, while ρ remains bounded away from 1. We assume for concreteness that $k/\log n \to \infty$, for otherwise detecting k-intervals for very small k is almost equivalent to detecting k-sets—and exactly the same when $k = 1$. We state a general result that applies for classes of small cardinality.

Proposition 3 *Consider a class $\mathcal{C}$ of size $N \to \infty$ such that $(1/k)\log N \to 0$. When $\rho k/\log N \to \infty$, the GLRT with threshold value $t = -\rho k + \rho\sqrt{5k\log N} + 3\log N$ has Bayes risk tending to zero.*

Proof. We first bound the probability of type I error. Indeed, under the null, by Lemma 2 and its proof, we can decompose

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

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

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

It is well-known that the maximum of N standard normals is bounded by $\sqrt{3\log N}$ with probability tending to one as $N \to \infty$. Hence, the second term on the right-hand side is bounded by $3\log N$ with high probability. For the first term, we combine the union bound and Chernoff's bound to obtain, for all $a \le 1$,

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

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 $t = -\rho k + \rho\sqrt{5k\log N} + 3\log N$ has probability of type I error tending to zero.

Now consider the alternative under $\mathbb{P}_S$. By Lemma 2 and Chebyshev's inequality,

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

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 with the lower bound for k-intervals in Corollary 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 $(1/k)\log N \to \infty$ with $n \to \infty$. The equivalent of Proposition 3 for this regime is the following:

Proposition 4 *Consider a class $\mathcal{C}$ of size $N \to \infty$ such that $(1/k)\log N \to \infty$. When $\eta := (1-\rho)N^{2/k}(\log N)/k \to 0$, the GLRT with threshold value $t = -(\log N)/\sqrt{\eta}$ has Bayes risk tending to zero.*

Proof. We follow the proof of Proposition 3. 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,

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

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

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

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 4 the condition on ρ implies that $\rho \to 1$, an assumption that may not be realistic in most applications. For k-sets, the requirement is that $1-\rho = o(n/k)^2$, which does not match the lower bound obtained in Corollary 3. If we restrict ρ to be bounded away from 1, then the GLRT may be powerless, as we show next.

Theorem 2 *Let $\mathcal{C}$ be the class of all k-sets. If $\rho < 0.6$ and $k = o(n^{0.7})$, the GLRT has a risk bounded away from zero.*

Proof. The proof is divided in 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,

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

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

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

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

$$f(x) = 1 \quad \text{if and only if} \quad \max_{S \in \mathcal{C}} g(x|_S) > t .$$

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

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

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

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

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 the Lemma 1, there exists $U, U_1, \dots, U_k$ independent standard normal such that for all $i \in \{1, \dots, k\}$,

$$X'_i = \sqrt{\rho}\, U + \sqrt{1-\rho}\, U_i.$$

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

$$X'_1, \dots, X'_k \in [-\zeta, \zeta]\,,$$

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 $X_1, \dots, X_n$ 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

$$\begin{aligned} \Omega = \{&X'_1, \dots, X'_k \in (-\zeta, \zeta)\,, \text{ and } \exists \alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k \in \{1, \dots, n\} \text{ distinct} : \\ &X_{\alpha_1}, \dots, X_{\alpha_k}, -X_{\beta_1}, \dots, -X_{\beta_k} \in (\zeta, \gamma\zeta)\} \end{aligned}$$

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

$$\begin{aligned} g\left(X_{\alpha_1}, \dots, X_{\alpha_k}\right) &\geq k(k-1)\zeta^2 - \rho(k-1)k\gamma^2\zeta^2 \\ &= k(k-1)\zeta^2\left(1 - \rho\gamma^2\right), \end{aligned} \tag{3.4}$$

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

Let $S \in \mathcal{C}$ be such that $S \cap \{1, \dots, k\} \neq \emptyset$. 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 $S \cap \{1, \dots, k\} = \emptyset$ we have $g(X|_S) = g(X'|_S)$. First remark that we can assume that

$$\left(\sum_{i\in S} X'_i\right)^2 \geq \zeta(k-1)\sqrt{1-\rho\gamma^2}\,, \tag{3.5}$$

since otherwise by (3.4) 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

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

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

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

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

Consider the case where $\sum_{i \in S} X'_i > 0$ (the case $\sum_{i \in S} X'_i < 0$ can be dealt 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

$$\begin{aligned} (2 + \rho(k-1)) X'_1 + \rho(k-1) X_{\alpha_u} &\leq (k-1)\zeta\gamma \left(\frac{2}{k-1} + 2\rho \right) \\ &= 2(k-1)\zeta \sqrt{1 - \rho\gamma^2} \\ &\leq 2 \sum_{i \in S} X_i. \end{aligned}$$

This concludes the proof. □

In view of Theorem 2, 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 2. We do not know of any other situation where the GLRT fails so miserably.

3.3 Other tests

In this section we discuss two other natural tests, both comparable to the GLRT in terms of, but with computational advantages. Indeed, note that the GLRT may be difficult to compute in many cases such as the class of k-sets.

3.3.1 A localized squared-sum test

The first variant that we propose is a "localized" version of the squared-sum test, based on the test statistic $T = \max_{S \in \mathcal{C}} (\sum_{i \in S} X_i)^2$. An advantage of this statistic is that it does not depend on knowledge of ρ. (Naturally, the threshold value of T under which the null hypothesis is accepted must depend on ρ.)

It is not difficult to see, by an argument parallel to that of Proposition 3, that this test is nearly optimal for the detection of small (parametric) classes. Also, if $\mathcal{C}$ is the class of all k-sets, T is easy to compute, since $T = \max\left((\sum_{i=1}^k X_{(i)})^2, (\sum_{i=1}^k X_{(n-i+1)})^2\right)$, where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the order statistics of X. However, in the case of k-sets this test suffers from the same drawback as the GLRT: it can be powerless in situations where the basic square-sum test is powerful. In fact, one can see that Theorem 2 also holds true for this test (with a similar proof).

3.3.2 Tests based on pairwise distances

Another natural idea when designing tests is based on the observation that (positively) correlated variables tend to be closer to each other than uncorrelated ones. Here we briefly present such a test, only for the case of k-sets. Unfortunately, such a test can only be powerful for values of ρ very close to 1. In that regime the test has a similar performance as the GLRT which has some power for detecting k-sets in the regime where $k^2/n \to 0$. The advantage of the distance-based test is that it is computationally tractable.

The proposed test rejects for small values of D, defined as

$$D = \min_{i=1,\ldots,n} |X_i - X_i^{(k-1)}|$$

where $X_i^{(k-1)}$ denotes the $(k-1)$-st nearest neighbor of X_i among $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$).

Observe first that under $\mathbb{P}_S$, if $i \in S$, then for all $j \in S$, by Lemma 1, $|X_i - X_j|$ is normally distributed with mean 0 and variance $2(1-\rho)$. Thus, under the alternative, D is bounded from above by the maximum of $\binom{k}{2}$ normal random variables with variance $2(1-\rho)$. This implies that $D \leq \sqrt{9(1-\rho)\log k}$ with high probability.

On the other hand, under the null hypothesis, D is at least k/n with high probability. To see this, note first that since the standard normal distribution function Φ is Lipschitz with constant $1/\sqrt{2\pi}$,

$$D \geq D_U \stackrel{\text{def}}{=} \sqrt{2\pi} \min_{i=1,\ldots,n} |\Phi(X_i) - \Phi(X_i^{(k-1)})| \geq \sqrt{2\pi} \min_{i=1,\ldots,n} |U_i - U_i^{(k-1)}| \ .$$

where the $U_i = \Phi(X_i)$ are independent uniform random variables in $(0,1)$ and $U_i^{(k-1)}$ is the $(k-1)$-st nearest neighbor of U_i. By conditioning on the value of X_i, we see that for any $i = 1, \ldots, n$ and $t < 1/2$,

$$\mathbb{P}\left\{|U_i - U_i^{(k-1)}| < t\right\} \leq \mathbf{P}\left\{\mathrm{Bin}(n-1, 2t) \geq k\right\} \ .$$

Taking, for example, $t = k/(4(n-1))$, and using Bernstein's inequality, we have

$$\mathbb{P}\left\{|U_i - U_i^{(k-1)}| < t\right\} \leq e^{-k/4}$$

and therefore, by the union bound,

$$\mathbb{P}\left\{D_U < (\sqrt{2\pi}/4)\frac{k}{n-1}\right\} \leq ne^{-k/4} \ .$$

Thus, under the null hypothesis, $D \geq (\sqrt{2\pi}/4)(k/(n-1))$ with high probability, whenever $k \geq 4 \log n$. Hence, the test based on D is powerful when

$$1 - \rho \geq \frac{1}{\log k} \left(\frac{\sqrt{2\pi}}{12} \frac{k}{n-1} \right)^2 .$$

Acknowledgements

We thank Omiros Papaspiliopoulos for his illuminating remarks. EAC was partially supported by a grant from the Office of Naval Research (N00014-09-1-0258).

References

Addario-Berry, L., N. Broutin, L. Devroye, and G. Lugosi (2010). On combinatorial testing problems. *Ann. Statist. 38*(5), 3063–3092.

Anandkumar, A., L. Tong, and A. Swami (2009). Detection of Gauss-Markov random fields with nearest-neighbor dependency. *IEEE Trans. Inform. Theory 55*(2), 816–827.

Arias-Castro, E., E. J. Candès, and A. Durand (2011). Detection of an anomalous cluster in a network. *Ann. Statist. 39*(1), 278–304.

Arias-Castro, E., E. J. Candès, H. Helgason, and O. Zeitouni (2008). Searching for a trail of evidence in a maze. *Ann. Statist. 36*(4), 1726–1757.

Arias-Castro, E., D. Donoho, and X. Huo (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory 51*(7), 2402–2425.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli 8*(5), 577–606.

Berman, S. M. (1962). Equally correlated random variables. *Sankhyā Ser. A 24*, 155–156.

Bickel, P. J. and E. Levina (2008). Covariance regularization by thresholding. *Ann. Statist. 36*(6), 2577–2604.

Boutsikas, M. V. and M. V. Koutras (2006). On the asymptotic distribution of the discrete scan statistic. *J. Appl. Probab. 43*(4), 1137–1154.

Cai, T. T., C.-H. Zhang, and H. H. Zhou (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist. 38*(4), 2118–2144.

Desolneux, A., L. Moisan, and J.-M. Morel (2003). Maximal meaningful events and applications to image analysis. *Ann. Statist. 31*(6), 1822–1851.

Devroye, L., A. György, G. Lugosi, and F. Udina (2011). High-dimensional random geometric graphs and their clique number. Submitted.

Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist. 32*(3), 962–994.

Hall, P. and J. Jin (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist. 38*(3), 1686–1732.

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

Jin, J. (2003). *Detecting and Estimating Sparse Mixtures*. Ph. D. thesis, Stanford University.

Kailath, T. and H. V. Poor (1998). Detection of stochastic processes. *IEEE Trans. Inform. Theory 44*(6), 2230–2259. Information theory: 1948–1998.

Perone Pacifico, M., C. Genovese, I. Verdinelli, and L. Wasserman (2004). False discovery control for random fields. *J. Amer. Statist. Assoc. 99*(468), 1002–1014.

Ramírez, D., J. Vía, I. Santamaría, and L. L. Scharf (2010). Detection of spatially correlated Gaussian time series. *IEEE Trans. Signal Process. 58*(10), 5006–5015.