Combinatorial Hypothesis Testing

Yangxinyu Xie

April 20, 2023

1 Introduction ❖

Suppose we observe an n-dimensional vector $\mathbf{X} = (X_1, ..., X_n)$. The null hypothesis H_0 is that $\mathbf{X} \sim \mathcal{N}(0, \mathbf{I})$. We denote the probability measure and expectation under H_0 by $\mathbb{P}_0$ and $\mathbb{E}_0$, respectively.

Combinatorics kicks in as we consider the alternative hypotheses, by introducing a class $\mathcal{C}$: consider a class $\mathcal{C} = \{S_1, \ldots, S_N\}$ of N sets of indices such that $S_k \subset \{1, \ldots, n\}$ and $|S_k| = K$ for all $k = 1, \ldots, N$. Under H_1, there exists an $S \in \mathcal{C}$ such that the distribution of X_i is determined by whether i is in S:

Alternative 1. [Detection of Means [1, 4, 5]]

$$X_i \sim \begin{cases} \mathcal{N}(0, 1), & \text{if } i \notin S \\ \mathcal{N}(\mu, 1), & \text{if } i \in S \end{cases}$$

where $\mu > 0$ is a positive parameter and components of $\mathbf{X}$ are independent.

Alternative 2. [Detection of Correlations [2]] $X_i \sim \mathcal{N}(0, 1)$ for each i and

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

where $\rho \in (0, 1)$.

For each $S \in \mathcal{C}$, we denote the probability measure and expectation by $\mathbb{P}_S$ and $\mathbb{E}_S$, respectively. Many interesting combinatorial examples of $\mathcal{C}$ arises: subsets of size K, cliques, perfect matchings, spanning trees, and clusters, etc.

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 the *minimax risk*:

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

where $R^{\max}(f)$ is the worst-case risk over the class of interest $\mathcal{C}$, formally defined by

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

In this report, we discuss the techniques introduced in $[1, 2, 4]$ to derive the lower and upper bounds of $R_*^{\max}$, as well as more recent extensions. For a more detailed, yet self-contained, discussion of the proof techniques, we refer the reader to [9].

2   General Framework ❖

2.1   Lower Bounds ❋

A standard way to obtain 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. The uniform prior on $\mathcal{C}$ gives 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 the average risk over the worst-case risk is that the likelihood ratio test, denoted f^*, is optimal for the former, according to the Neyman–Pearson lemma. Introducing $L(X)$, the likelihood ratio between H_0 and H_1, the optimal test becomes

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

The (average) risk $R^* = R(f^*)$ of the optimal test is called the *Bayes risk* and it satisfies

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

2.2   Upper Bounds ❋

The analysss of the upper bounds of the risks of optimal tests are often less systematic. Even though likelihood ratio test is optimal in the Bayesian setting, it is difficult to obtain directly upper bounds for the (worst-case) risk of the likelihood ratio test. Hence, in analysing combinatorial testing problems, we often focus on obtaining upper bounds for sub-optimal tests for ease of analysis and strive for an upper bound matching the obtained lower bound.

3   Detection of Means ❖

In this section we focus on the first alternative hypothesis 1. We will first discuss simple tests are considered as we obtain upper bounds: the averaging test and the maximum test. Then, we will provide the key insights driving the results for lower bounds obtained in [1] and discuss the applications to examples such as k-**sets, spanning trees, perfect matchings, etc**.

3.1   Upper Bounds ❋

The *averaging test* has the form

$$f(\mathbf{x}) = 0 \quad \text{if and only if} \quad \sum_{i=1}^{n} X_i \leq \mu K/2.$$

Since the sum of the components of $\mathbf{X}$ is zero-mean normal under $\mathbb{P}_0$ and has mean μK under the alternative hypothesis, we obtain the following:

Proposition 3.1 ([1], Proposition 2.1)**.** *Let $\delta > 0$. The risk of the averaging test f satisfies $R(f) \leq \delta$ whenever*

$$\mu \geq \sqrt{\frac{8n}{K^2} \log \frac{2}{\delta}}.$$

The *maximum test* is based on the *scan statistic* $\max_{S \in \mathcal{C}} X_S$, and it has the form:

$$f(\mathbf{x}) = 0 \quad \text{if and only if} \quad \max_{S \in \mathcal{C}} X_S \leq \frac{\mu K + \mathbb{E}_0 \max_{S \in \mathcal{C}} X_S}{2}.$$

Observing that under the alternative hypothesis for some $S \in \mathcal{C}$, $X_S = \sum_{i \in S} X_i$ is normal $(\mu K, K)$, we obtain

Proposition 3.2 ([1], Proposition 2.2). *The risk of the maximum test f satisfies $R(f) \leq \delta$ whenever*

$$\mu \geq \frac{\mathbb{E}_0 \max_{S \in \mathcal{C}} X_S}{K} + 2\sqrt{\frac{2}{K} \log \frac{2}{\delta}}.$$

ADD SOMETHING

3.2   Lower Bounds   ❄

Let's take a closer look at the likehood ratio. If we write

$$\phi_0(\mathbf{x}) = (2\pi)^{-n/2} e^{-\sum_{i=1}^n x_i^2/2} \text{ and } \phi_S(\mathbf{x}) = (2\pi)^{-n/2} e^{-\sum_{i \in S}(x_i - \mu)^2/2 - \sum_{i \notin S} x_i^2/2}$$

for the probability densities of $\mathbb{P}_0$ and $\mathbb{P}_S$, respectively, the likelihood ratio at $\mathbf{x}$ is

$$L(\mathbf{x}) = \frac{1/N \sum_{S \in \mathcal{C}} \phi_S(\mathbf{x})}{\phi_0(\mathbf{x})} = \frac{1}{N} \sum_{S \in \mathcal{C}} e^{\mu x_S - K \mu^2/2},$$

where $x_S = \sum_{i \in S} x_i$. With the observation $R^* \geq 1 - \sqrt{1 - (\mathbb{E}_0 \sqrt{L(\mathbf{X})})^2}$ ([8]), we can apply Jensen's inequality,

$$\mathbb{E}_0 \sqrt{L(\mathbf{X})} = \int \sqrt{\frac{1/N \sum_{S \in \mathcal{C}} \phi_S(\mathbf{x})}{\phi_0(\mathbf{x})}} \phi_0(\mathbf{x}) \, d\mathbf{x} = \int \sqrt{\frac{1}{N} \sum_{S \in \mathcal{C}} \phi_S(\mathbf{x}) \phi_0(\mathbf{x})} \, d\mathbf{x} \geq \frac{1}{N} \sum_{S \in \mathcal{C}} \int \sqrt{\phi_S(\mathbf{x}) \phi_0(\mathbf{x})} \, d\mathbf{x}.$$

Because for any $S \in \mathcal{C}, \int \sqrt{\phi_S(\mathbf{x}) \phi_0(\mathbf{x})} \, d\mathbf{x} = e^{-\mu^2 K/8}$, for all classes $\mathcal{C}$, $R^* \geq 1/2$ whenever

$$\mu \leq \sqrt{(4/K)} \times \sqrt{\log(4/3)} \tag{3.2.1}$$

i.e. small risk cannot be achieved unless μ is substantially large compared to $K^{-1/2}$. But this can be substantially improved by the moment method proposed in [1], as we discuss next.

3.2.1   Moment Methods

The moment method applies the following insight: by the Cauchy–Schwarz inequality,

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

and since $\mathbb{E}_0 L(\mathbf{X}) = 1, \mathbb{E}_0 |L(\mathbf{X}) - 1|^2 = \text{Var}_0(L(\mathbf{X})) = \mathbb{E}_0 [L(\mathbf{X})^2] - 1$. We are now ready to prove the following lower bound based on overlapping pairs, which reduces the problem to studying a purely combinatorial quantity [1, 5]:

Proposition 3.3 ([1], Proposition 3.2). *Let S and S' be drawn independently, uniformly, at random from $\mathcal{C}$ and let $Z = |S \cap S'|$. Then*

$$R^* \geq 1 - \tfrac{1}{2} \sqrt{\mathbb{E} e^{\mu^2 Z} - 1}.$$

Proof. Because $L(\mathbf{X}) = \frac{1}{N} \sum_{S \in \mathcal{C}} e^{\mu X_S - K \mu^2/2}$,

$$\mathbb{E}_0 [L(\mathbf{X})^2] = \frac{1}{N^2} \sum_{S, S' \in \mathcal{C}} e^{-K \mu^2} \mathbb{E}_0 e^{\mu(X_S + X_{S'})}.$$

Meanwhile,

$$
\begin{aligned}
\mathbb{E}_0 e^{\mu(X_S + X_{S'})} &= \mathbb{E}_0[e^{\mu \sum_{i \in S \setminus S'} X_i} e^{\mu \sum_{i \in S' \setminus S} X_i} e^{2\mu \sum_{i \in S \cap S'} X_i}] \\
&= (\mathbb{E}_0 e^{\mu X})^{2(K - |S \cap S'|)} (\mathbb{E}_0 e^{2\mu X})^{|S \cap S'|} \\
&= e^{\mu^2 (K - |S \cap S'|) + 2\mu^2 |S \cap S'|}
\end{aligned}
$$

$\square$

The previous proposition allows us to obtain lower bounds by analysing the quantity $\mathbb{E} e^{\mu^2 Z}$. This allows us to exploit the combinatorial structures of the class $\mathcal{C}$ and gives us far better bounds than 3.2.1, as the folloiwng examples show:

Example 3.4 (Spanning Trees, [1], Section 4.5)**.** Let $1, 2, \ldots, n = \binom{m}{2}$ represent the edges of the complete graph K_m and let $\mathcal{C}$ be the set of all spanning trees of K_m. Thus, we have $N = m^{m-2}$ spanning trees and $K = m - 1$. With the fact $\mathbb{E}[e^{\mu^2 Z}] \leq \exp(2e^{\mu^2})$, we obtain that for any $\delta \in (0, 1)$, $R^* \geq \delta$ whenever

$$
\mu \leq \sqrt{\log\big(1 + \tfrac{1}{2} \log\big(1 + 4(1 - \delta)^2\big)\big)}.
$$

Meanwhile, the averaging test has risk $R(f) \leq \delta$ whenever $\mu \geq \sqrt{4 \log(2/\delta)}$.

We also explore two special combinatorial structures: *symmetry* and *negative association.*

Definition 3.5 (Symmetry)**.** We say that the class $\mathcal{C}$ is *symmetric* if it satisfies the following conditions. Let S, S' be drawn independently and uniformly at random from $\mathcal{C}$. Then,

1. the conditional distribution of $Z = |S \cap S'|$ given S' is identical for all values of S';

2. for any fixed $S_0 \in \mathcal{C}$ and $i \in S_0$, $\mathbb{P}\{i \in S\} = K/n$.

Via Hölder's inequality, we can obtain the following

Proposition 3.6 ([1], Proposition 3.3)**.** *Assume that $\mathcal{C}$ is symmetric. Then*

$$
\mathbb{E}[e^{\mu^2 Z}] \leq (e^{\mu^2 K} - 1)\frac{K}{n} + 1.
$$

Definition 3.7 (Negative Association)**.** A collection $Y_1, \ldots, Y_n$ of random variables is *negatively associated* if for any pair of disjoint sets $I, J \subset \{1, \ldots, n\}$ and (coordinate-wise) nondecreasing functions f and g,

$$
\mathbb{E}[f(Y_i, i \in I)g(Y_j, j \in J)] \leq \mathbb{E}[f(Y_i, i \in I)]\mathbb{E}[g(Y_j, j \in J)].
$$

Proposition 3.8 ([1], Proposition 3.4)**.** *Assume that the class $\mathcal{C}$ is symmetric. Suppose that the labels are such that $S' = \{1, 2, \ldots, K\} \in \mathcal{C}$. Let S be a randomly chosen element of $\mathcal{C}$. If the random variables $\mathbf{1}_{\{1 \in S\}}, \ldots, \mathbf{1}_{\{K \in S\}}$ are negatively associated, then*

$$
\mathbb{E}[e^{\mu^2 Z}] \leq \left((e^{\mu^2} - 1)\frac{K}{n} + 1 \right)^K.
$$

Example 3.9 (K-sets, [1], Section 4.2)**.** Consider the example when $\mathcal{C}$ contains all sets $S \subset \{1, \ldots, n\}$ of size K. Note $N = \binom{n}{K}$. This class is symmetric and satisfies the condition in the previous proposition. Hence, applying Proposition 3.3, $R^* \geq \delta$ for all μ with

$$
\mu \leq \sqrt{\log\left(1 + \frac{n \log(1 + 4(1 - \delta)^2)}{K^2}\right)}.
$$

ADD SOMETHING

3.3 Detection of an Anomalous Cluster   ❄

As a special case of the detection of means problem, we can consider the following scenario in network analysis introduced in [4]. Let $\mathbb{V}_m$ be either derived from a geometric shape or has a graph structure. For simplicity, we can think of $\mathbb{V}_m$ as the set of graphs with m nodes embedded in the Euclidean space. Let $\mathcal{K}_m$ be a class of clusters within $\mathbb{V}_m$.

Under $H_0^m, \mathbf{X} \sim \mathcal{N}(0, \mathbf{I}_m)$. Under H_1^m, there exists a cluster $K \in \mathcal{K}_m$ where

$$X_v \sim \mathcal{N}(\mu_K, 1) \qquad \forall v \in K; \qquad X_v \sim \mathcal{N}(0, 1) \qquad \forall v \notin K,$$

where $\mu_K > 0$. We choose to decompose μ_K as $\mu_K = |K|^{-1/2}\Lambda_K$, where $|K|$ denotes the number of nodes in K and Λ_K is the signal strength. We say that H_0^m and H_1^m are *asymptotically inseparable* (in the minimax sense) if

$$\liminf_{m \to \infty} R_*^{\max} = 1,$$

which is equivalent to saying that, as m becomes large, no test can perform substantially better than random guessing. A sequence of tests (T_m) is said to *asymptotically separate* H_0^m and H_1^m if

$$\lim_{m \to \infty} R^{\max}(T_m) = 0,$$

Combining the insight we gained from the moment method earlier and more delicate analysis of the signal Λ_K, [4] shows that if $\mathcal{K}_m$ is the set of mild deformations of a unit ball, then under specific conditions, H_0^m and H_1^m are asymptotically inseparable if there is $\eta_m \to 0$ slowly enough such that, for all $K \in \mathcal{K}_m$,

$$\Lambda_K \leq (1 - \eta_m)\sqrt{2\log(m/|K|)};$$

and conversely, a version of the scan statistic can asymptotically separates H_0^m and H_1^m if there is $\eta_m \to 0$ slowly enough such that, for all $K \in \mathcal{K}_m$,

$$\Lambda_K \geq (1 + \eta_m)\sqrt{2\log(m/|K|)}.$$

[4] also derived bounds for "thin clusters" and paths along a d-dimensional lattice. We refer the reader to the orignal paper for details. Notably, the authors also extended their results to exponential families.

4 Detection of Correlations   ❖

4.1 Lower Bounds   ❄

First, we note that we can rewrite the alternative hypotheses as

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

where

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

Introducing $Z_S = \exp\left(\frac{1}{2}X^T(\mathbf{I} - \mathbf{A}_S^{-1})X\right)$ for all $S \in \mathcal{C}$, the likelihood ratio between H_0 and H_1 may be written as

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

Thus the Bayes risk satisfies

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

4.1.1 Generalised Moment Method

A direct application of the momont method discussed earlier does not yield very promising lowerbounds; instead, we leverage the insight from the folloiwng representation lemma:

Lemma 4.1 ([6]; [2], Lemma 1.1). *Let $X_1, \ldots, X_k$ 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 $U, U_1, \ldots, U_k$, 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). We are now ready to reduce analysing testing problem into a combinatorial problem.

Proposition 4.2 ([2], Theorem 2.1). *For any class $\mathcal{C}$ and any $a > 0$,*

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

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

Proof. Via Lemma 4.1, we can write

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

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 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 versus $H_1(u)$. Therefore, conditioning on U,

$$R^* = R(f^*) = \mathbb{E}_U R_U(f^*) \geq \mathbb{E}_U R_U(f_U^*) = 1 - \tfrac{1}{2}\mathbb{E}_U\mathbb{E}_0|L_U(X) - 1|$$

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,

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

After some computation, we obtain

$$\mathbb{E}_0 L_u^2(X) \leq \frac{1}{N^2}\sum_{S,S' \in \mathcal{C}}\exp\left(\left(\frac{\rho u^2}{1+\rho} - \frac{1}{2}\log(1-\rho^2)\right)|S \cap S'|\right)$$

$\square$

Again, we reduce the problem to studying the purely combinatorial quantity $Z = |S \cap S'|$. We demonstrate the implications of this proposition via a few examples.

Example 4.3 (*k*-sets, [2], Section 2.3.3). Suppose $\mathcal{C}$ is the class of all sets of size k. By negative association, (see Proposition 3.8)

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

which implies that $R^* \to 1$ if either $k^2/n \to \infty$ and $\rho k^2/n \to 0$, or $(1-\rho)n^2/k^4 \to \infty$.

Example 4.4 (Spanning Trees, [2], Section 2.3.5)**.** Suppose $\mathcal{C}$ is the class of all spanning trees of a complete graph with $k + 1$ vertices. Similar to Example 3.4, notice

$$\mathbb{E}e^{\nu Z} \leq \exp 2(e^{\nu} - 1),$$

which implies that $R^* \to 1$ if $\rho \to 0$.

4.2   Upper Bounds   ❆

The *squared-sum test* is based on the observation that under $\mathbb{P}_0$, $(\sum_{i=1}^{n} X_i)^2$ is distributed as $n\chi_1^2$, while for any $S \subset \{1, \ldots, n\}$ 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 implies that the squared-sum test is near-optimal for detecting k-sets in the regime where $k^2/n \to \infty$, as stated in the following proposition.

Proposition 4.5 ([2], Proposition 3.1)**.** *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 worst-case risk converging to zero.*

Notice that in this case, i.e. $k^2/n \to \infty$, Example 4.3 implies that any test based on $(\sum_{i=1}^{n} X_i)^2$ is powerless if $\rho k^2/n \to 0$.

The *generalized likelihood ratio test* (GLRT) has the form

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

for some appropriately chosen t. This is based on the folloiwng observation: under $\mathbb{P}_0$, $X^T(\mathbf{I} - \mathbf{A}_S^{-1})X$ is distributed as

$$-\frac{\rho}{1 - \rho}\chi_{k-1}^2 + \frac{\rho(k-1)}{1 + \rho(k-1)}\chi_1^2,$$

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

$$-\rho\chi_{k-1}^2 + \rho(k-1)\chi_1^2,$$

where χ_1^2 and χ_{k-1}^2 denote independent χ^2 random variables with degrees of freedom 1 and $k - 1$, respectively. Careful analysis based on this observation gives us the folloiwng result:

Proposition 4.6 ([2], 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.*

Notice that in Proposition 4.6 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 Example 4.3 requires. In fact, if ρ is bounded away from 1, then the GLRT may be powerless ([2], Theorem 3.1).

The *goodness of fit test* (GOF test) is based on the folloiwng insight. 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 4.7. *Suppose $X_1, \ldots, X_k$ are zero-mean, unit-variance random variables satisfying $\mathrm{Cov}(X_i, X_j) \geq \rho > 0$, for all $i \neq j$. Let $\overline{X}$ denote their average. Then for any $t > 0$,*

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

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. The folloiwng result suggests that, ignoring logarithmic factors, the GOF test is near-optimal for detecting k-sets in the regime where $k^2/n \to 0$ and $k/\log n \to \infty$.

Proposition 4.8. *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 $(1-\rho)^{1/2} \ll 1/m$, the resulting test with threshold $t = n/m + \sqrt{3n\log(m)/m}$ has worst-case risk tending to zero.*

WRITE A SUMMARY

5 Extension ❖

Subsequent lines of work have found more settings where the moment methods introduced in earlier sections can be systematically applied to derive minimax lower bounds for testing problems. In this section, we focus on introducing these scenarios. On the other hand, as mentioned earlier, the upper bounds are more often obtained by considering different proposed tests, whose analyses are often much more delicate. Hence, we refer the reader to the respective papers for analyses for the upper bounds.

5.1 Dectection of positive correlations in a multivariate sample ❋

A direct extension of the dectection of correlation problem is considered in [3]. Under the alternatives, we make a slight modification of the alternative hypotheses 2, which is, if $i \neq j, i, j \in S$, we have $\text{Cov}(X_i, X_j) \geq \rho > 0$. the covariance matrix is still sparse, i.e., $k < n/2$ or $k = o(n)$, meaning that only a few variables are substantially correlated. Examples of this setting includes the block model, cliques and perfect matchings.

Via a more delicate version of the generalized moment method we saw, we can obtain the following:

Proposition 5.1 ([3], Theorem 1). *For any class $\mathcal{C}$, any $\rho \in (0,1)$, and any $a \geq \sqrt{3}$,*

$$R^* \geq \mathbb{P}\big(\chi_m^2 \leq ma^2\big)\left(1 - \frac{1}{2}\sqrt{\mathbb{E}\min\big[\exp(m\nu_a Z), \cosh^m(\xi_a Z)\big] - 1}\right), \tag{5.1.1}$$

where

$$\nu_a := \frac{\rho a^2}{1+\rho} - \frac{1}{2}\log\big(1-\rho^2\big) \quad and \quad \xi_a := \frac{\rho a^2}{1-\rho^2},$$

and where χ_m^2 has chi-squared distribution with m degrees of freedom, and $Z = |S \cap S'|$ with S, S' i.i.d. uniform from $\mathcal{C}$.

The upper bounds are obtained via squared-sum test and "localized" squared-sum test.

5.2 Dectection of sparse principal components of a high-dimensional covariance matrix ❋

Introduced in [7], we can consider the alternative hypothesis $\mathbf{X} \sim \mathcal{N}(0, \mathbf{I} + \theta vv^T)$ where $\theta > 0$ and v is k-sparse. The techniques for proving the minimax lower bound is much more technical and involves studying the χ^2 distance between probability measures. Yet, the moment methods still palys a key role in showing the following result:

Proposition 5.2 ([7], Theorem 5.1). *Fix $\nu > 0$. There exists a constant C_ν such that if*

$$\theta < \min\left(\sqrt{\frac{k\log(C_\nu p/k^2 + 1)}{n}}, \frac{1}{\sqrt{2}}\right)$$

Then $R^* \geq 0.5 - \nu$.

5.3   Combinatorial inference for graphical models   ❅

More recently, [10] introduced a new family of combinatorial inference problems for graphical models. Unlike classical statistical inference where the main interest is point estimation or parameter testing, combinatorial inference aims at testing the global structure of the underlying graph. Examples include testing the graph connectivity, the presence of a cycle of certain size, or the maximum degree of the graph.

5   References   ❖

[1] Louigi Addario-Berry, Nicolas Broutin, Luc Devroye, and Gábor Lugosi. On combinatorial testing problems. *The Annals of Statistics*, pages 3063–3092, 2010.

[2] ERY ARIAS-CASTRO, SÉBASTIEN BUBECK, and GÁBOR LUGOSI. Detection of correlations. *The Annals of Statistics*, 40(1):412–435, 2012.

[3] ERY ARIAS-CASTRO, SÉBASTIEN BUBECK, and GÁBOR LUGOSI. Detecting positive correlations in a multivariate sample. *Bernoulli*, 21(1):209–241, 2015.

[4] Ery Arias-Castro, Emmanuel J Candes, and Arnaud Durand. Detection of an anomalous cluster in a network. *The Annals of Statistics*, pages 278–304, 2011.

[5] Ery Arias-Castro, Emmanuel J Candès, Hannes Helgason, and Ofer Zeitouni. Searching for a trail of evidence in a maze. *Ann. Statist.*, 36(1):1726–1757, 2008.

[6] Simeon M Berman. Equally correlated random variables. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 155–156, 1962.

[7] Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(1):1780–1815, 2013.

[8] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[9] Gábor Lugosi. Lectures on combinatorial statistics. *47th Probability Summer School, Saint-Flour*, pages 1–91, 2017.

[10] Matey Neykov, Junwei Lu, and Han Liu. Combinatorial inference for graphical models. *The Annals of Statistics*, 47(2):795–827, 2019.