Combinatorial Hypothesis Testing

April 18, 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, 3, 4]]

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

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

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

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 [1–3] to derive the lower and upper bounds of $R^{\max}_*$, as well as more recent extensions.

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:

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

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 $R(f) \le \delta$ whenever*

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

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 $R^* \ge 1 - \sqrt{1 - (\mathbb{E}_0\sqrt{L(\mathbf{X})})^2}$ ([6]), we can apply Jensen's inequality,

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

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 $\mathcal{C}$, $R^* \ge 1/2$ whenever

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

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,

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

and since $\mathbb{E}_0 L(\mathbf{X}) = 1, \mathbb{E}_0|L(\mathbf{X})-1|^2 = \mathrm{Var}_0(L(\mathbf{X})) = \mathbb{E}_0[L(\mathbf{X})^2] - 1$. 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, 4]:

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*

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

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,

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

□

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 (Disjoint Sets, [1], Section 4.1)**.** Suppose all $S \in \mathcal{C}$ are disjoint (and therefore $KN \le n$). Fix $\delta \in (0,1)$. Let $Z = K$ with probability $1/N$ and $Z = 0$ otherwise. Thus,

$$\mathbb{E}e^{\mu^2 Z} - 1 = \frac{1}{N}(e^{\mu^2 K} - 1) \le \frac{1}{N}e^{\mu^2 K}$$

and therefore $R^* \ge \delta$ whenever

$$\mu \le \sqrt{\frac{\log(4N(1-\delta)^2)}{K}}.$$

Notice that with the bound the bound $\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S \le \sqrt{2K\log N}$, maximum test f gives us $R^* \le R(f) \le \delta$ whenever

$$\mu \ge \sqrt{\frac{2\log N}{K}} + 2\sqrt{\frac{2\log(2/\delta)}{K}}.$$

So in this case the critical transition occurs when μ is of the order of $\sqrt{(1/K)\log N}$.

Example 3.5 (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 $\mathbb{E}[e^{\mu^2 Z}] \le \exp(2e^{\mu^2})$, we obtain that for any $\delta \in (0,1)$, $R^* \ge \delta$ whenever

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

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

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

Definition 3.6 (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 3.7 ([1], Proposition 3.3)**.** *Assume that $\mathcal{C}$ is symmetric. Then*

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

Definition 3.8 (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,

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

Proposition 3.9 ([1], Proposition 3.4)**.** *Let $\delta \in (0,1)$ and 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 3.10 (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 [3]. 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, [3] 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 $K \in \mathcal{K}_m$,

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

[3] 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 $Z_S = \exp\big(\frac{1}{2} X^T (\mathbf{I} - \mathbf{A}_S^{-1}) X\big)$ 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 Representation Lemma:

Lemma 4.1 ([5]; [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).

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

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

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

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

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

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,

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

After some computation, we obtain

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

□

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 (Disjoint Sets, [2], Section 2.3.1)**.** Suppose all $S \in \mathcal{C}$ are disjoint (and therefore $KN \leq n$). Let $Z = K$ with probability $1/N$ and $Z = 0$ otherwise. Thus,

$$\mathbb{E}e^{\nu Z} - 1 = \frac{1}{N}(e^{\nu K} - 1) \leq \frac{1}{N}e^{\nu K}$$

which is bounded by 1 if $\nu \leq \log(N)/k$, in which case $R^* \geq 0.3$.

Example 4.4 (k-sets, [2], Section 2.3.3)**.** Suppose $\mathcal{C}$ is the class of all sets of size k. By negative association, (see Proposition 3.9)

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

which is bounded by 2 when

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

in which case $R^* \geq 0.3$.

Example 4.5 (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 3.5, notice

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

which is bounded by 13/4 when $\nu \leq 1 + \ln((\ln(13/4))/2)$, in which case $R^* \geq 0.15$.

4.2 Upper Bounds ❄

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

Proposition 4.6. *Let $\mathcal{C}$ be an arbitrary class of sets of size k and suppose that $\rho k^2/n \to \infty$ in (**??**). 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 (**??**).*

In Corollary **??**, 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.

4.3 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 **??**, 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.

4.3.1 The GLRT for k-intervals and other parametric classes

Recalling Corollary **??**, 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 4.7. *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 **??** 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 $a \le 1$,

$$\begin{aligned} \mathbb{P}_0\Big\{\min_{S\in\mathcal{C}} C_S < a\ (k-1)\Big\} &\le N\mathbf{P}\{\chi^2_{k-1} < a(k-1)\} \\ &\le N\exp\Big(-\frac{(k-1)}{2}(a\ -1-\log a)\Big). \end{aligned} \tag{4.3.1}$$

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 ¶$_S$. By Lemma **??** and Chebyshev's inequality,

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

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 4.7 with the lower bound for k-intervals in Corollary **??**, we see that the GLRT is near-optimal for detecting k-intervals. This is actually the case for all parametric classes we know of.

4.3.2 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 4.7 for this regime is the following:

Proposition 4.8. *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 4.7. The only difference is in (4.3.1), 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,

$$\max_{S\in\mathcal{C}} X^T(\mathbf{I} - \mathbf{A}_S^{-1})X \le -\frac{\rho}{1-\rho}N^{-2/k}k\sqrt{\eta} + 2\log N. \tag{4.3.2}$$

As before, with high probability under $\P_S$,

$$X^T(\mathbf{I} - \mathbf{A}_S^{-1})X \ge -\rho k, \tag{4.3.3}$$

so we only need to check that the threshold t is larger than the right-hand side in (4.3.2) and smaller than the right-hand side in (4.3.3), which is the case by the assumptions we made. □

Notice that in Proposition 4.8 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 **??** requires. Moreover, if we restrict ρ to be bounded away from 1, then the GLRT may be powerless.

Theorem 4.9. *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.*

In view of Theorem 4.9, 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 4.6. We do not know of any other situation where the GLRT fails so miserably.

4.4 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:

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

for some appropriate threshold t.

Proposition 4.10. *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 (**??**), 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

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

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

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

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 [**?**]. 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.

4.5 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 **??**. 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 4.11. *Suppose $X_1, \dots, X_k$ 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 $t > 0$,*

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

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

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

By Markov's inequality, we then have

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

The statement follows from observing that

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

□

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 (**??**), with Lemma 4.11 we see that the concentration will happen over an interval of length slightly larger than $\sqrt{1-\rho}$. This is apparent from Lemma 4.1 under the simple correlation model (**??**).

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 $I_s, s = 1, \dots, m$. 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:

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

where t is some threshold.

Proposition 4.12. *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 (**??**), 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

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

This and the union bound imply that, indeed,

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

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

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

Though the set I is random, by Lemma 4.11 and the fact that $(1-\rho)^{1/2} \ll 1/m$, we have that

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

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

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

where ϕ denotes the standard normal density function and a is taken sufficiently large. Therefore, when $|I| \ge k/2$ 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. □

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

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

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 4.3.2. 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 [**?**].

5 Extension ❖

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, Emmanuel J Candes, and Arnaud Durand. Detection of an anomalous cluster in a network. *The Annals of Statistics*, pages 278–304, 2011.

[4] 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.

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

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