Combinatorial Hypothesis Testing

April 13, 2023

1 Introduction ❖

Suppose we observe an n-dimensional vector $\mathbf{X} = (X_1, ..., X_n)$. The null hypothesis H_0 is that the components of $\mathbf{X}$ are independent and identically distributed (i.i.d.) standard normal random variables. 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 $\mathcal{C}$ with some combinatorial structure: consider a class $\mathcal{C} = \{S_1, \ldots, S_N\}$ of N sets of indices such that $S_k \subset \{1, \ldots, n\}$ for all $k = 1, \ldots, N$. Under H_1, there exists an $S \in \mathcal{C}$ such that X_i has a distribution determined by whether i is in S:

1. In its simplest form, as discussed in [1, 3, 4], we consider

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

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

2. In testing correlations [2], we consider

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

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

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 asymptotic upper and lower bounds of $R_*^{\max}$, as well as more recent extensions.

2 Lower Bounds ❖

2.1 Moment Methods ❄

A standard way of obtaining lower bounds for the minimax risk is by putting a prior on the class $\mathcal{C}$ and obtaining a lower bound on the corresponding *Bayesian risk*, which never exceeds the worst-case risk. Because this is true for any prior, the idea is to find one that is hardest (often called *least favorable*). Consider the uniform prior on $\mathcal{C}$, giving rise to the following *average risk*:

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

where

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

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

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

3 Clusters ❖

4 Extension ❖

4 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.