Combinatorial testing problems

the problem

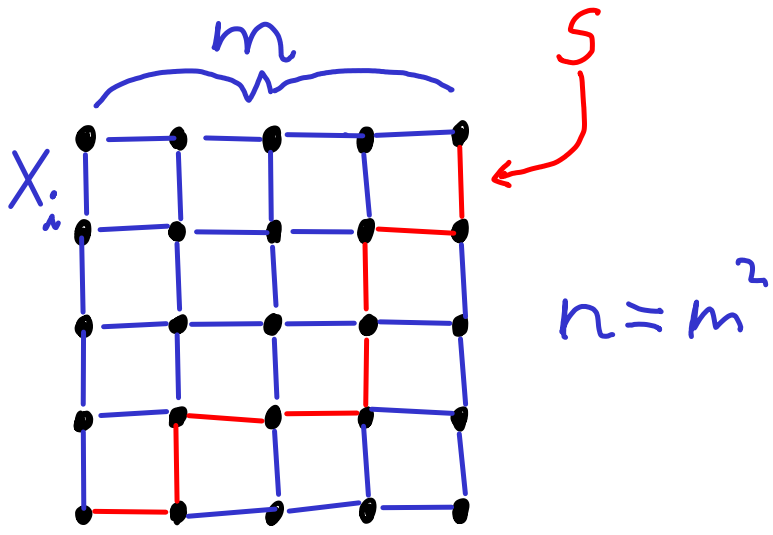
A hypothesis testing problem.
One observes $X = (X_1, \ldots, X_n)$.
Null hypothesis: the components of X are i.i.d. standard normal.
Alternative hypothesis: We have a class $\mathcal{C} = \{S_1, \ldots, S_N\}$ of sets $S_k \subset \{1, \ldots, n\}$.
Under H_1, there exists an $S \in \mathcal{C}$ such that

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

The components of X are independent.
The distribution of X is denoted by $\mathbb{P}_S$.
Assume that $|S| = k$ for every $S \in \mathcal{C}$.

examples

- the set of all subsets $S \subset \{1, \ldots, n\}$ of size k (classical multiple testing, e.g., Ingster (1999), Baraud (2002), Donoho and Jin (2004))
- set of all paths in a square grid (Arias-Castro et al., 2008)



examples

- set of all clusters of a certain structure on a grid (Arias-Castro, Candès, and Durand, 2009)
- the set of all cliques of a given size in a complete graph;
- the set of all bicliques of a given size in a complete bipartite graph; (exploratory analysis of microarray data, see Shabalin, Weigman, Perou, and Nobel, 2009)
- the set of all spanning trees of a complete graph;
- the set of all perfect matchings in a complete bipartite graph;

the risk

A test is a function $f : \mathbb{R}^n \to \{0, 1\}$.
If $f(X) = 0$ then the test accepts the null hypothesis.
The risk of f is

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

The optimal test is easy to determine: if $x_S = \sum_{i \in S} x_i$,

$$f^*(x) = \begin{cases} 0 & \text{if } \frac{1}{N} \sum_{S \in \mathcal{C}} e^{\mu x_S - k\mu^2/2} \leq 1 \\ 1 & \text{otherwise.} \end{cases}$$

The optimal risk is

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

where $L(X) = \frac{1}{N} \sum_{S \in \mathcal{C}} e^{\mu X_S - K\mu^2/2}$ is the likelihood ratio where $X_S = \sum_{i \in S} X_i$.

averaging test

Compute the average and threshold:

$$f(x) = \mathbb{1}_{\{\sum_{i=1}^{n} x_i > \mu k/2\}} .$$

Then for any $\delta > 0$, $R(f) \leq \delta$ whenever

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

Surprisingly, this simple test is nearly optimal in many cases.

scan statistic

Another simple test is

$$f(x) = 1 \quad \text{if and only if} \quad \max_{S \in \mathcal{C}} X_S \geq \frac{\mu k + \mathbb{E}_0 \max_{S \in \mathcal{C}} X_S}{2} .$$

Then $R(f) \leq \delta$ whenever

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

This is an easy consequence of Tsirelson's Gaussian concentration inequality.

lower bounds

Let S and S' be drawn independently, uniformly, at random from $\mathcal{C}$ and let $Z = |S \cap S'|$. Then

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

Proof.

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

Since $\mathbb{E}_0 L(X) = 1$,

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

a lower bound for gaussian processes

Let $\mu_{\mathcal{C}}$ be such that

$$\mathbb{E} \exp\left(\mu_{\mathcal{C}}^2 Z\right) = 2 \ .$$

Then for all $\mu \leq \mu_{\mathcal{C}}$,

$$R^* \geq \frac{1}{2} \ .$$

But recall from the scan statistic test that $R^* < 1/2$ whenever

$$\mu \geq \frac{\mathbb{E}_0 \max_{S \in \mathcal{C}} X_S}{k} + \frac{4}{\sqrt{k}} \ .$$

This implies that

$$\mathbb{E}_0 \max_{S \in \mathcal{C}} X_S \geq k\mu_{\mathcal{C}} - 4\sqrt{k} \ .$$

negative association

Assume that the class $\mathcal{C}$ is symmetric. Suppose that $S' = \{1, 2, \ldots, k\} \in \mathcal{C}$ and let S be a randomly chosen from $\mathcal{C}$. If the random variables $\mathbb{1}_{\{1 \in S\}}, \ldots, \mathbb{1}_{\{k \in S\}}$ are negatively associated then $R^* \geq 1/2$ whenever

$$\mu \leq \sqrt{\log\left(1 + \frac{n \log 2}{k^2}\right)}\,.$$

These tools allow us to determine the order of the critical μ in various applications.

examples: k-sets

Consider the class $\mathcal{C}$ of all sets of size k. Thus, $N = \binom{n}{k}$.
This example is well–understood, see, e.g., Ingster (1999), Baraud (2002), Donoho and Jin (2004).
We have negative association, so $R^* \geq 1/2$ whenever

$$\mu \leq \sqrt{\log\left(1 + \frac{n \log 2}{k^2}\right)}\,.$$

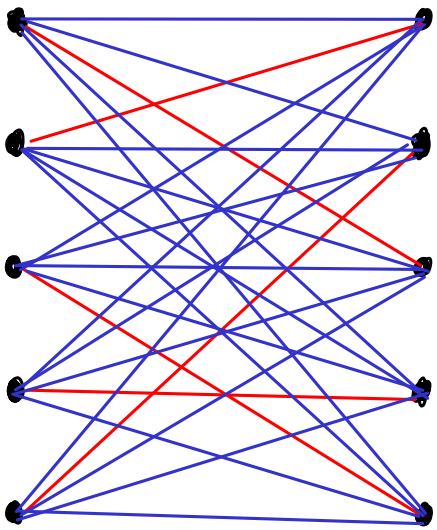
Various regimes. Since

$$\frac{\mathbb{E}_0 \max_{S \in \mathcal{C}} X_S}{k} \leq \frac{\sqrt{2k \log \binom{n}{k}}}{k} \leq \sqrt{2 \log\left(\frac{ne}{k}\right)}\,,$$

the scan test shows that for $k = O\left(n^{(1-\epsilon)/2}\right)$ for some $\epsilon > 0$, then the threshold value is of the order of $\sqrt{\log n}$.
When k^2/n, is bounded away from zero, then the averaging test provides a matching upper bound.

perfect matchings

Let $\mathcal{C}$ be the set of all perfect matchings of the complete bipartite graph $K_{m,m}$.



$n = m^2$

$K = m$

$N = m!$

By the the averaging test, $R(f) \leq 1/2$ whenever $\mu \geq \sqrt{8 \log 4}$.

perfect matchings

$Z = |S \cap S'|$ is the number of fixed points in a random permutation.
Then $R^* \geq 1/2$ whenever

$$\mu \leq \sqrt{\log(1 + \log 2)} \; .$$

The optimal test f^* can be computed efficiently: computing

$$\frac{1}{N} \sum_{S \in \mathcal{C}} e^{\mu X_S} = \frac{1}{m!} \sum_{\sigma} \prod_{j=1}^{m} e^{\mu X_{(i,\sigma(i))}}$$

is equivalent to computing the permanent of a non-negative $m \times m$ matrix. Jerrum, Sinclair, and Vigoda (2004) show that this may be done by a polynomial-time randomized approximation.

spanning trees

Let $\mathcal{C}$ be the class of all spanning trees of the complete graph K_m.

$$n = \binom{m}{2} \quad k = m - 1 \quad N = m^{m-2} \; .$$

Negative association holds by a result of Feder and Mihail (1992).

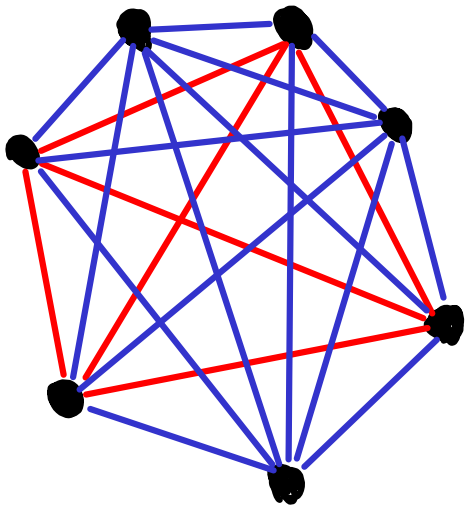
$R^* \geq 1/2$ whenever

$$\mu \leq \sqrt{\log\left(1 + \frac{\log 2}{2}\right)} \; .$$

The likelihood ratio $(1/N) \sum_{S \in \mathcal{C}} e^{\mu X_S}$ may be computed by the Propp-Wilson (1998) algorithm.

cliques

Let $\mathcal{C}$ contain all cliques of size k of the complete graph K_m.



$$n = \binom{m}{2}$$

$$K = \binom{k}{2}$$

$$N = \binom{n}{k}$$

Negative association doesn't hold anymore.

cliques

Assume $k \leq \sqrt{m(\log 2)/e}$. Then
(i) $R^* \leq 1/2$ whenever

$$\mu \geq 2\sqrt{\frac{1}{k-1}\log\left(\frac{me}{k}\right)} + 4\sqrt{\frac{\log 4}{k(k-1)}}\,,$$

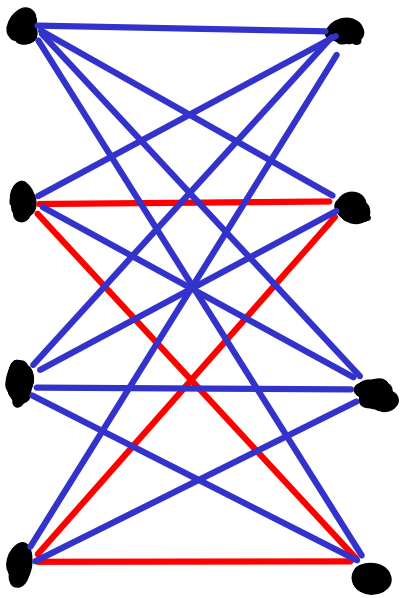
(ii) $R^* \geq 1/2$ whenever

$$\mu \leq \sqrt{\frac{1}{k}\log\left(\frac{m}{2k}\right)}\,.$$

Upper bound holds by scan test. Both scan test and optimal test are difficult to compute.

bi-cliques

A similar model, relevant in the exploratory analysis of microarray data (Shabalin, Weigman, Perou, and Nobel, 2009):
$\mathcal{C}$ contains all bi-cliques of size k of the complete bipartite graph $K_{m,m}$.
Then $n = m^2$, $K = k^2$, and $N = \binom{m}{k}$.



The analysis and results are similar to the case of cliques.

research problem: sharp threshold?

The optimal risk $R^* = R^*_{\mathcal{C}}(\mu)$ is a decreasing function of μ.

For what classes is there a sharp threshold?

If μ_a is such that $R^*_{\mathcal{C}}(\mu) = a$, is it true that

$$\mu_\epsilon - \mu_{1-\epsilon} = o(\mu_{1/2}) \quad \text{as } k \to \infty?$$

detection of correlations

n sensors receive noisy signals.

Sensor i receives $(X_{i,1}, \ldots, X_{i,d})$.

If there is no signal, $X_{i,t}$ are i.i.d. standard normal.

In the presense of an event, a small subset $S \subset \{1, \ldots, n\}$ of sensors receives a common signal covered by noise:

$$X_{i,t} = \begin{cases} N_{i,t} & \text{if } i \notin S \\ (N_{i,t} + Y_t)/\sqrt{1+\rho} & \text{if } i \in S \end{cases}$$

where $N_{i,t}$ are standard normal and Y_t are normal $(0, \rho)$ independent of $(\rho \ll 1)$.

hypothesis testing problem

One observes d i.i.d vectors $X_1, \ldots, X_d \in \mathbb{R}^n$.

Under the null hypothesis, the components of $X_1, \ldots, X_d$ are independent.

Under the alternative hypothesis, a subset S of components ($|S| = k \ll n$) has correlation $\rho > 0$, the rest are independent.

One does not know the (possibly) "contaminated" subset.

tests and their risk

A test is a function $f : \mathbb{R}^{nd} \to \{0, 1\}$.

$f(X_1, \ldots, X_d) = 0$ means that the null is accepted.

The risk of a test is

$$\begin{aligned} R(f) &= \mathbb{P}_0\{f(X_1, \ldots, X_d) = 1\} \\ &\quad + \frac{1}{\binom{n}{k}} \sum_S \mathbb{P}_S\{f(X_1, \ldots, X_d) = 0\} \end{aligned}$$

where $\mathbb{P}_0$ is the probability under the null.
$\mathbb{P}_S$ is the probability if the set of contaminated components is S.
Another measure of risk:

$$\begin{aligned} \overline{R}(f) &= \mathbb{P}_0\{f(X_1, \ldots, X_d) = 1\} \\ &\quad + \max_S \mathbb{P}_S\{f(X_1, \ldots, X_d) = 0\} \end{aligned}$$

optimal test

There exists a test f^*, (likelihood ratio test) with optimal risk $R^* = R(f^*)$.

This test is difficult to compute. Requires calculating a sum of $\binom{n}{k}$ terms.

a lower bound

$$R^* \geq \frac{1}{2} - \frac{1}{4}\sqrt{\exp\left((e^{\rho d} - 1)\frac{k^2}{n}\right) - 1}\ .$$

(Arias-Castro, Bubeck, Lugosi, 2015)
Proof idea: The bound

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

does not work. Instead, we condition on the value of the common "signal" and treat the problem as shifted means.

If the class of possible "contaminated sets" is not necessarily the class of all sets of size k, then again R^* maybe bounded from below in terms of the moment generating function of $Z = |S \cap S'|$.

a lower bound

$$R^* \geq \frac{1}{2} - \frac{1}{4}\sqrt{\exp\left((e^{\rho d} - 1)\frac{k^2}{n}\right) - 1}\,.$$

If $d \ll (1/\rho)\log n$, it is impossible to have $R^* \to 0$ unless $k = n^{1/2 - o(1)}$.

However, if $d \geq (9/\rho)\log n$, small risk is possible even if k is a constant.

This can be achieved by a simple test based on a "scan statistic".

scan statistic

$$T_{n,k,d} = \max_S \sum_{t=1}^{d} \sum_{i,j \in S, i \neq j} X_{i,t} X_{j,t} \ .$$

The test f_n that accepts the null hypothesis if and only if

$$T_{n,k,d} \leq dk(k-1)\rho/4$$

has risk $R(f_n) \to 0$ if $(k-1)\rho > 8$ and $d \geq (9/\rho) \log n$.

The proof is based on the analysis of Gaussian quadratic forms.

random correlation graph

Let $X_i = (X_{i,1}, \ldots, X_{i,d})$.
For the testing problem, it is natural to calculate pairwise correlations

$$\frac{(X_i, X_i)}{\|X_i\| \cdot \|X_j\|}$$

and define a graph by connecting i and j if the correlation is large enough (e.g., > 0).

In the presence of a signal, one expects a large clique.

What is large?

Under the null hypothesis, the $Z_i = X_i/\|X_i\|$ are uniformly distributed on the sphere and we have a random geometric graph in $\mathbb{R}^d$.

random geometric graph

Model: Let $X_1, \ldots, X_n$ be independent vectors, uniform on $S_{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$.

For a given $p \in (0, 1)$, we define the random geometric graph $\overline{G}(n, d, p)$

Vertex set $V = \{1, \ldots, n\}$.

i and j are connected by an edge if an only if

$$(X_i, X_j) \geq t_{p,d}$$

where $t_{p,d}$ is such that

$$\mathbb{P}\left\{(X_i, X_j) \geq t_{p,d}\right\} = p \ .$$

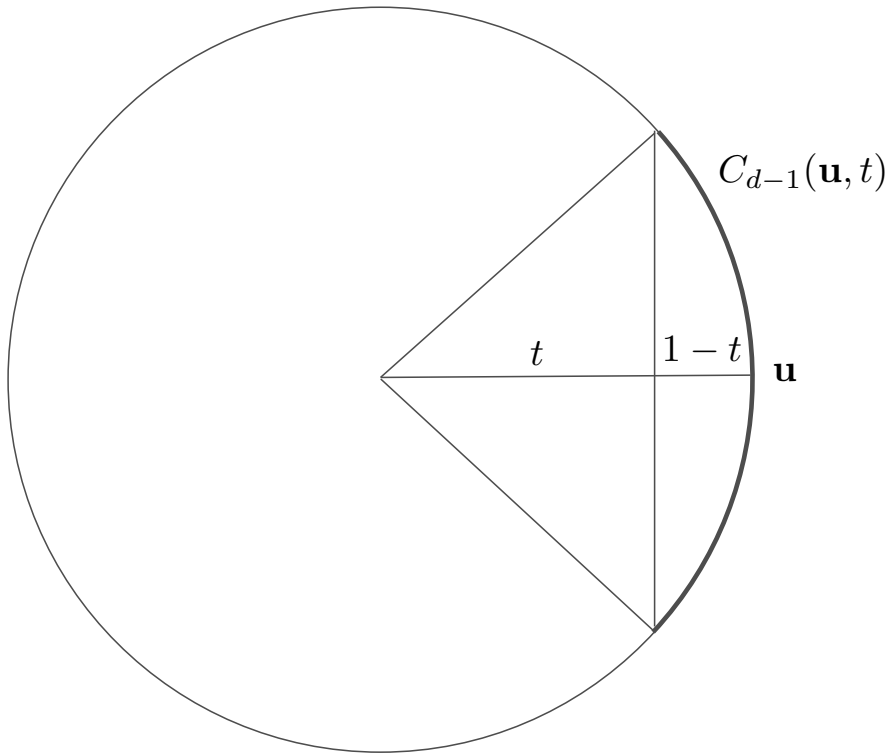
Equivalently, $i \sim j$ if and only if $\|X_i - X_j\| \leq \sqrt{2(1 - t_{p,d})}$.

edge probability

For $p = 1/2$, $t_{p,d} = 0$.

Let μ_{d-1} be the uniform probability measure over S_{d-1}.
For $u \in S_{d-1}$ and $0 \leq t \leq 1$, a spherical cap of height $1 - t$ around u is

$$C_{d-1}(u, t) = \{x \in \mathbb{R}^d : x \in S_{d-1}, (x, u) \geq t\}$$



edge probability

$p = \mu_{d-1}(C_{d-1}(e, t_{p,d}))$ is the normalized surface area of a spherical cap of height $1 - t_{p,d}$.

It is useful to represent

$$X = \frac{Z}{\|Z\|}$$

with $Z \in \mathbb{R}^d$ standard normal.

Clearly, $\mathbb{E}\|Z\|^2 = d$.

Since $\|Z\|$ is a Lipschitz function of Z, $\mathbf{var}(\|Z\|) \leq 1$.

In particular, $\|Z\|/\sqrt{d} \to 1$ in probability.

This implies $X_1\sqrt{d}$ is approximately standard normal.

edge probability

Consequence: for any $s > 0$,

$$\mu_{d-1}(C_{d-1}(e, s/\sqrt{d})) = \mathbb{P}\{X_1 > s/\sqrt{d}\} \to 1 - \Phi(s)$$

as $d \to \infty$.

For any fixed $p \in (0, 1)$,

$$\lim_{d\to\infty} t_{p,d}\sqrt{d} = \Phi^{-1}(1 - p) \ .$$

very large dimension

$G(n,p)$ denotes the Erdős-Rényi random graph. (n vertices, egdes are present independently, with probability p.)

Total variation distance between two random graphs G and G':

$$d_{TV}(G,G') = \max_{\mathcal{G}} |\mathbb{P}\{G \in \mathcal{G}\} - \mathbb{P}\{G' \in \mathcal{G}\}|$$

where the maximum is over all $2^{\binom{n}{2}}$ sets of graphs over n vertices.

THEOREM. Fix n and p. Then

$$\lim_{d\to\infty} d_{TV}(\overline{G}(n,d,p), G(n,p)) = 0 \ .$$

very large dimension

For the proof, write

$$\mathbb{P}\{\overline{\boldsymbol{G}}(n,d,p)=\boldsymbol{g}\} = \mathbb{P}\left\{\bigcap_{1\le i<j\le n} \mathbb{1}_{\{(\boldsymbol{X}_i,\boldsymbol{X}_j)>t_{p,d}\}}=\boldsymbol{g}_{i,j}\right\}$$

$$= \mathbb{P}\left\{\bigcap_{1\le i<j\le n} \mathbb{1}_{\{\sum_t \sqrt{d}\boldsymbol{X}_{i,t}\boldsymbol{X}_{j,t}>\sqrt{d}t_{p,d}\}}=\boldsymbol{g}_{i,j}\right\}$$

and

$$\mathbb{P}\{\boldsymbol{G}(n,p)=\boldsymbol{g}\} = \mathbb{P}\left\{\bigcap_{1\le i<j\le n} \mathbb{1}_{\{\boldsymbol{N}_{i,j}>\Phi^{-1}(1-p)\}}=\boldsymbol{g}_{i,j}\right\}$$

where the $\boldsymbol{N}_{i,j}$ are independent standard normal. The proof follows from a multivariate central limit theorem.

clique number of $\overline{G}(n, d, 1/2)$

For fixed d, the clique number $\omega(n, d, 1/2)$ grows linearly with n.
For $d = \infty$ the behavior is very different:

$$\omega(n, \infty, 1/2) = 2\log_2 n - 2\log_2 \log_2 n + O(1).$$

How fast does $\omega(n, d, 1/2)$ approach the clique number of

$G(n, 1/2)$?

How large does d need to be for similar behavior?

clique number bounds

if $d \sim$ const., then $\omega(n, d, 1/2) = \Omega_p(n)$

if $d \to \infty$, then $\omega(n, d, 1/2) = o_p(n)$

if $d = o(\log n)$, then $\omega(n, d, 1/2) \geq n^{1-o_p(1)}$

if $d \sim \log^2 n$, then $\omega(n, d, p) = O_p(\log^3 n)$ when $p < 1/2$.

if $d \gg \log^3 n$, then $\omega(n, d, 1/2) = (2 + o_p(1)) \log_2 n$

if $d \sim \log^5 n$, then
$\omega(n, d, 1/2) = 2 \log_2 n - 2 \log_2 \log_2 n + O_p(1)$

if $d \geq \log((2 \log 2)n)$,

$$\omega(n, d, 1/2) \geq \exp\left(\log^2(n)/(20d)\right) - 1 \, .$$

proof ideas

first three statements are easy (from area estimates of a spherical cap)

fourth follows from Jung's theorem and Vapnik-Chervonenkis inequality

Jung's theorem (1901): For every set $A \subset \mathbb{R}^d$ of diameter at most 1 there exists a closed ball of radius $\sqrt{d/(2(d+1))}$ that includes A.

clique number bounds

the points X_i in a clique K form a set of diameter at most $\sqrt{2(1-t_{p,d})}$.

by Jung's theorem, K is contained in a spherical cap $C_{d-1}(u, s_p)$ for some $u \in S_{d-1}$ and $s_p = \sqrt{(dt_{p,d}+1)/(d+1)}$.

$$\omega(d,n,p) \leq n \sup_{C=C_{d-1}(u,s_p)} \mu_{d-1,n}(C) \ ,$$

where $\mu_{d-1,n}$ is the empirical measure.

by Vapnik and Chervonenkis, with probability at least $1-\eta$, for all $C = C_{d-1}(u, s_p)$,

$$\mu_{d-1,n}(C) \leq 2\mu_{d-1}(C) + \frac{4(d+1)}{n} \ln \frac{2ne}{d+1} + \frac{4}{n} \ln \frac{4}{\eta} \ .$$

and

$$\mu_{d-1,n}(C) \leq \frac{1}{s_p\sqrt{d}} e^{-s_p^2(d-1)/2} + \frac{4(d+1)}{n} \ln \frac{2ne}{d+1} + \frac{4}{n} \ln \frac{4}{\eta} \ .$$

clique number bounds

if $d \sim$ const., then $\omega(n, d, 1/2) = \Omega_p(n)$

if $d \to \infty$, then $\omega(n, d, 1/2) = o_p(n)$

if $d = o(\log n)$, then $\omega(n, d, 1/2) \geq n^{1-o_p(1)}$

if $d \sim \log^2 n$, then $\omega(n, d, p) = O_p(\log^3 n)$ when $p < 1/2$.

if $d \gg \log^3 n$, then $\omega(n, d, 1/2) = (2 + o_p(1)) \log_2 n$

if $d \sim \log^5 n$, then
$\omega(n, d, 1/2) = 2 \log_2 n - 2 \log_2 \log_2 n + O_p(1)$

if $d \geq \log((2 \log 2)n)$,

$$\omega(n, d, 1/2) \geq \exp\left(\log^2(n)/(20d)\right) - 1 \ .$$

upper bound for $d \gg \log^3 n$

N_k is the number of cliques of size k. For $G(n, 1/2)$,

$$\mathbb{E} N_k = \binom{n}{k} 2^{-\binom{k}{2}}$$

Let $\delta > 0$ and $K > 2$. If

$$d \geq \frac{K^3}{\delta^2} \ ,$$

then, for $1 \leq k \leq K$,

$$\mathbb{E} N_k(n, d, 1/2) \leq \binom{n}{k} \Phi(\delta)^{\frac{(k-1)(k-2)}{2}} \ .$$

Follows from an inductive argument, using approximate orthogonality of $X_1, \ldots, X_n$.

clique number estimates

The upper bounds for $\omega(n, d, p)$ follow from the first moment method:

$$\mathbb{P}\{\omega(n, d, 1/2) \geq k\} = \mathbb{P}\{N_k \geq 1\} \leq \mathbb{E}N_k \ ,$$

Lower bounds for $\omega(n, d, p)$ follow from the second moment method.
First we prove a similar lower bound for $\mathbb{E}N_k$ and then show

$$\frac{\mathbf{var}(N_k)}{(\mathbb{E}N_k)^2} \to 0$$

for the relevant values of k.

clique number bounds

if $d \sim$ const., then $\omega(n, d, 1/2) = \Omega_p(n)$

if $d \to \infty$, then $\omega(n, d, 1/2) = o_p(n)$

if $d = o(\log n)$, then $\omega(n, d, 1/2) \geq n^{1-o_p(1)}$

if $d \sim \log^2 n$, then $\omega(n, d, p) = O_p(\log^3 n)$ when $p < 1/2$.

if $d \gg \log^3 n$, then $\omega(n, d, 1/2) = (2 + o_p(1)) \log_2 n$

if $d \sim \log^5 n$, then
$\omega(n, d, 1/2) = 2 \log_2 n - 2 \log_2 \log_2 n + O_p(1)$

if $d \geq \log((2 \log 2)n)$,

$$\omega(n, d, 1/2) \geq \exp\left(\log^2(n)/(20d)\right) - 1 \ .$$

lower bound

The lower bound follows from the lower bound for the risk of any test.

If the largest clique was small then there would exist a test with small risk.

lower bound

Suppose $0 < \rho \leq 1/2$. Under the alternative hypothesis, with probability at least $1 - \delta$, the graph contains a clique of size k whenever

$$\binom{k}{2} \leq \delta \exp\left(\frac{d\sigma^4}{10}\right) ,$$

where $\sigma^2 = \rho/(1-\rho)$.

references

G. Lugosi. Lectures on Combinatorial Statistics.
`.../SaintFlour.pdf`

L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi (2010). Combinatorial testing problems. *Annals of Statistics.*

L. Devroye, A. György, G. Lugosi, and F. Udina (2011). High-dimensional random geometric graphs and their clique number. *Electronic Journal of Probability.*

E. Arias-Castro, S. Bubeck, and G. Lugosi, (2012). Detection of correlations. *Annals of Statistics.*

E. Arias-Castro, S. Bubeck, and G. Lugosi, (2015). Detecting positive correlations in a multivariate sample. *Bernoulli.*