STAT 689: Large-scale and high-dimensional statistical inference **Fall 2024**

Lecture 3

1 Recap and motivation

Previously, we have introduced and discussed two global testing procedures, namely Bonferroni's method and the L_2 test. But each of these has its pros and cons.

- If our data exhibit strong, sparsely distributed signals, then Bonferroni's method excels and is optimal in the "needle in a haystack" setting, but the L_2 test performs very poorly.
- If our data exhibit small, widely distributed signals, the L_2 test excels and is optimal, but Bonferroni's method is powerless.

Given these facts, we hope to develop an adaptive test that combines the strengths of Bonferroni and L_2.

2 Simes test

Suppose we observe a sequence of p-values, which can be ordered as $p_{(1)} \le p_{(2)} \le \cdots \le p_{(n)}$. At level α, Bonferroni's method rejects the global null when $p_{(1)} \le \alpha/n$. The Simes considers an adaptive threshold for each $p_{(i)}$. More precisely, it rejects the global null whenever

$$p_{(i)} \le \frac{\alpha i}{n}$$

for some i. If we define

$$T_n = \min_{1\le i\le n} \left(\frac{np_{(i)}}{i}\right),$$

Simes test rejects the null if $T_n \le \alpha$. By construction, Simes test is less conservative compared to Bonferroni's method. (Why?)

Bonferron's method considers only the smallest p-value and compares it with a threshold α/n. In contrast, the Simes test considers all p-values and compares each ordered p-value with the corresponding threshold $\alpha i/n$.

Under the global null and assuming p_i's are independent, $T_n \sim \text{Unif}[0,1]$. As a result, $P(T_n \le \alpha) = \alpha$, i.e., the type I error is controlled at level α.

Proof. We prove the result using induction. If $n = 1$, the result is clearly true. Assume that $T_{n-1} \sim \text{Unif}[0,1]$, we aim to show that $T_n \sim \text{Unif}[0,1]$. To this end, we note that for any $0 < x < 1$

$$\begin{aligned}
P(T_n \le \alpha) =& P\left(\min_{1\le i\le n}\left(\frac{np_{(i)}}{i}\right) \le \alpha\right) \\
=& P\left(\min_{1\le i\le n-1}\left(\frac{np_{(i)}}{i}\right) \le \alpha, p_{(n)} > \alpha\right) + P(p_{(n)} \le \alpha) \\
=& P\left(\min_{1\le i\le n-1}\left(\frac{(n-1)p_{(i)}}{i}\right) \le \frac{(n-1)\alpha}{n}, p_{(n)} > \alpha\right) + \alpha^n.
\end{aligned}$$

The density of $p_{(n)}$ is given by $f(t) = nt^{n-1}$ for $t \in [0,1]$ as $P(p_{(n)} \le t) = t^n$. Given $p_{(n)} = t$, the other p-values are uniformly distributed over $[0,t]$ (Why?). Thus conditional on $p_{(n)} = t$, by induction, we have

$$\min_{1\le i\le n-1}\left(\frac{(n-1)p_{(i)}}{ti}\right) \sim \text{Unif}[0,1].$$

It then implies that

$$\begin{aligned}
&P\left(\min_{1\le i\le n-1}\left(\frac{(n-1)p_{(i)}}{i}\right)\le\frac{(n-1)\alpha}{n}, p_{(n)}>\alpha\right)\\
=&n\int_\alpha^1 t^{n-1}P\left(\min_{1\le i\le n-1}\left(\frac{(n-1)p_{(i)}}{ti}\right)\le\frac{(n-1)\alpha}{tn}\middle| p_{(n)}=t\right)dt\\
=&n\int_\alpha^1 t^{n-1}\frac{(n-1)\alpha}{tn}dt\\
=&\int_\alpha^1(n-1)\alpha t^{n-2}dt\\
=&\alpha(1-\alpha^{n-1})=\alpha-\alpha^n.
\end{aligned}$$

Therefore, we have $P(T_n\le\alpha)=\alpha-\alpha^n+\alpha^n=\alpha$.

3 Goodness of fit tests

Under the global null, the p-values follow the uniform distribution on $[0,1]$. Therefore, an alternative way of testing the global null is to check if the distribution of the p-values is uniform. This is essentially a goodness of fit testing problem.

3.1 Kolmogorov-Smirnov test

Suppose we observe $X_1,\dots,X_n\overset{i.i.d}{\sim}F$. We aim to test the null hypothesis that

$$H_0: F=F_0.$$

Let F_n be the empirical cdf. The Kolmogorov-Smirnoff test statistic is defined as

$$\mathrm{KS}_n=\sqrt{n}\sup_{t\in\mathbb{R}}|F_n(t)-F_0(t)|,$$

Under the null hypothesis,

$$\mathrm{KS}_n=\sqrt{n}\sup_{t\in\mathbb{R}}|F_n(t)-F(t)|=\sup_{f\in\mathcal{F}}|\sqrt{n}(P_n-P)f|$$

where $\mathcal{F}=\{1[\cdot\le t],t\in\mathbb{R}\}$.

P-Donsker. A collection $\mathcal{F}$ of functions is called P-Donsker if the process $\{\sqrt{n}(P_n-P)f\}_{f\in\mathcal{F}}$ converges to a tight limit G indexed by $\mathcal{F}$ in $L^\infty(\mathcal{F})$. Here, G is a Gaussian process. In particular,

$$\left(\sqrt{n}(P_n-P)f_1,\cdots,\sqrt{n}(P_n-P)f_k\right)\xrightarrow{d}\left(G_{f_1},\dots,G_{f_k}\right),$$

where $\mathrm{Cov}(G_{f_i},G_{f_j})=\mathrm{Cov}(f_i(X),f_j(X))$ and $X\sim P$.

Theorem. Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to $\mathbb{R}$, and let F be an envelop function of $\mathcal{F}$, (i.e. for any $x\in\mathcal{X}$ and any $f\in\mathcal{F}$, $|f(x)|\le F(x)$). Suppose $PF^2<\infty$ and

$$\int_0^\infty\sup_Q\sqrt{\log N\left(\mathcal{F},L_2(Q),\|F\|_{L_2(Q)}\epsilon\right)}\,d\epsilon<\infty,$$

where the sup is over all finitely supported measure Q. Then $\mathcal{F}$ is P-Donsker. For the proof, please check the STAT 620 notes.

We show that $\mathcal{F}=\{1[\cdot\le t],t\in\mathbb{R}\}$ is a Donsker class. To see this, we first note that the envelop function can be taken as $F\equiv1$. Second, one can show

$$\int_0^\infty\sup_Q\sqrt{\log N(\mathcal{F},L_2(Q),\epsilon\|F\|_{L_2(Q)})}\,d\epsilon<\infty.$$

Thus, we have $\sqrt{n}(P_n - P)f \to^d G_P(f)$ or equivalently

$$\sqrt{n}(F_n(t) - F(t)) \xrightarrow{d} G_F(t),$$

where G_F is a Brownian Bridge with $\text{Cov}(G_F(t), G_F(s)) = F(t \wedge s) - F(t)F(s)$. Note that the map $f \mapsto \sup_{t\in\mathbb{R}} |f(t)|$ is continuous in $\|\cdot\|_\infty$ as $|\|f\|_\infty - \|g\|_\infty| \leq \|f - g\|_\infty$. By the continuous mapping theorem

$$\text{KS}_n = \sup_{t\in\mathbb{R}} |\sqrt{n}(F_n(t) - F(t))| \to \sup_{t\in\mathbb{R}} |G_F(t)| = \sup_{t\in\mathbb{R}} |G_\lambda(F(t))| = \sup_{u\in(0,1)} |G_\lambda(u)|,$$

where λ is the uniform distribution/measure on $(0, 1)$, i.e., $\lambda(x) = x$. We can see that

$$\begin{aligned}\text{Cov}(G_\lambda(F(t)), G_\lambda(F(s))) =& \lambda(F(t) \wedge F(s)) - \lambda(F(t))\lambda(F(s)) \\ =& F(t \wedge s) - F(t)F(s) \\ =& \text{Cov}(G_F(t), G_F(s)).\end{aligned}$$

3.2 Cramér-Von Mises test

The Cramér-Von Mises statistic is defined as

$$\text{CV}_n = n \int (F_n(t) - F_0(t))^2 dF_0(t).$$

Under the null,

$$\text{CV}_n = n \int (F_n(t) - F(t))^2 dF(t) = \int \{\sqrt{n}(F_n(t) - F(t))\}^2 dF(t).$$

The map $f \mapsto \int f^2(t)dF(t)$ is continuous w.r.t. $\|\cdot\|_\infty$. By the continuous mapping theorem

$$\text{CV}_n \xrightarrow{d} \int G_F(t)^2 dF(t) = \int G_\lambda(F(t))^2 dF(t) = \int G_\lambda(u)^2 du.$$

3.3 Anderson-Darling test

Similar to the Cramér-Von Mises statistic, the Anderson-Darling statistic is another type of quadratic test statistic. Assuming that $F_0(t) = t$ and letting F_n be the empirical cdf based on $\{p_i\}_{i=1}^n$, the Anderson-Darling statistic is given by

$$\text{AD}_n = n \int \frac{(F_n(t) - t)^2}{t(1 - t)} dt.$$

It puts more weight on small and large p-values when compared with the Cramér-von Mises statistic. For statistical intuition, one can think of the Anderson-Darling statistic as "averaging" the squared z-score over t. This is because, under the global null, $nF_n(t) \sim \text{Bin}(n, t)$ and thus $\text{Var}(F_n(t)) = t(1 - t)$, so the integrand $(F_n(t) - t)^2/\{t(1 - t)\}$ is similar to a squared z-score.

Exercise 3.1: Show that we can compute the Anderson-Darling statistic by using the following expression:

$$\text{AD}_n = -n + \sum_{i=1}^n \frac{1 - 2i}{n} \log(p_{(i)}) + \sum_{i=1}^n \frac{1 - 2i}{n} \log(1 - p_{(n+1-i)}).$$

To see this, note that

$$\begin{aligned}
\mathrm{AD}_n =& n \int \frac{(F_n(t)-t)^2}{t(1-t)} dt \\
=& n \int_0^{p_{(1)}} \frac{t^2}{t(1-t)} dt + n \sum_{i=1}^{n-1} \int_{p_{(i)}}^{p_{(i+1)}} \frac{(i/n-t)^2}{t(1-t)} dt + n \int_{p_{(n)}}^{1} \frac{(1-t)^2}{t(1-t)} dt \\
=& n \int_0^{p_{(1)}} \frac{t}{1-t} dt + n \sum_{i=1}^{n-1} \int_{p_{(i)}}^{p_{(i+1)}} \frac{(i/n-t)^2}{t(1-t)} dt + n \int_{p_{(n)}}^{1} \frac{1-t}{t} dt \\
=& n \Bigg\{ - p_{(1)} - \log(1-p_{(1)}) + \sum_{i=1}^{n-1} \left[\frac{i^2}{n^2} \log \frac{p_{(i+1)}(1-p_{(i)})}{p_{(i)}(1-p_{(i+1)})} - (p_{(i+1)} - p_{(i)}) + \left(1 - \frac{2i}{n}\right) \log \frac{1-p_{(i)}}{1-p_{(i+1)}} \right] \\
& - (1-p_{(n)}) - \log(p_{(n)}) \Bigg\} \\
=& -n + n \Bigg\{ - \log(1-p_{(1)}) + \sum_{i=1}^{n-1} \left[\frac{i^2}{n^2} \log \frac{p_{(i+1)}(1-p_{(i)})}{p_{(i)}(1-p_{(i+1)})} + \left(1 - \frac{2i}{n}\right) \log \frac{1-p_{(i)}}{1-p_{(i+1)}} \right] - \log(p_{(n)}) \Bigg\} \\
=& -n + \sum_{i=1}^{n} \frac{1-2i}{n} \log(p_{(i)}) + \sum_{i=1}^{n} \frac{1-2i}{n} \log(1-p_{(n+1-i)}).
\end{aligned}$$

Recall that Fisher's test statistic is $T = -2\sum_{i=1}^n \log(p_i)$, and Pearson's test statistic is $T_{\text{Pear}} = 2\sum_{i=1}^n \log(1-p_i)$. The Anderson-Darling statistic is a combination of Fisher's test and Pearson's test. Compared to Fisher's and Pearson's tests, the Anderson-Darling test assigns greater weight to the p-values that are in the bulk because it reweights the p-values, depending on their rank, which Fisher and Pearson's tests do not do. This alleviates the high sensitivity to small p-values that Fisher's test experiences.

4 Second-level/Higher-criticism test

As we have seen, the Kolmogorov-Smirnov test looks for the maximum distance between the empirical CDF and its expected value under the global null hypothesis, while the Cramér-Von Mises test and the Anderson-Darling test integrate the differences instead. We now combine the two approaches.

As discussed before, the quantity $(F_n(t)-t)/\sqrt{t(1-t)}$ in the Anderson-Darling test can be viewed as a z-score. Instead of squaring this quantity and integrating over t as in the Anderson-Darling test, we take a maximization over t, leading to the higher-criticism statistic:

$$\mathrm{HC}_n = \sup_{0<t\le t_0} \frac{F_n(t)-t}{\sqrt{t(1-t)/n}}.$$

The higher-criticism statistic scans across the significance levels for departures from H_0. Hence, a large value of HC_n indicates the significance of an overall body of tests.

4.1 Sparse mixture models

We study a sparse mixture model to understand the power of the higher-criticism statistic and compare it to Bonferroni's method. The results presented below are based on Donoho and Jin (2004, AOS). Previously, to study the Bonferroni's method and the L_2 test, we consider n hypotheses

$$\begin{aligned}
&H_{0,i} : X_i \sim N(0,1), \\
&H_{1,i} : X_i \sim N(\mu_i, 1), \quad \mu_i > 0.
\end{aligned}$$

We are interested in the case where there is a (small) fraction of non-null hypotheses. Rather than directly assuming that there is some amount of nonzero means under the alternative, we can take a Bayesian viewpoint

by assuming that

$$\mu_i \overset{i.i.d}{\sim} (1-\varepsilon)\delta_0 + \varepsilon\delta_\mu,$$

where δ_μ denotes a point mass at μ. It thus leads to the Gaussian mixture model under the alternative:

$$\begin{aligned} &H_{0,i} : X_i \sim N(0,1), \\ &H_{1,i} : X_i \sim (1-\varepsilon)N(0,1) + \varepsilon N(\mu,1). \end{aligned}$$

The expected number of non-nulls under the alternative is $n\varepsilon$. If $\varepsilon = 1/n$, then the above would become the "needle in a haystack" problem: on average, there would be one non-zero coordinate. The likelihood ratio (LR) statistic is given by

$$L_n = \prod_{i=1}^{n} \left\{ 1 - \varepsilon + \varepsilon e^{\mu X_i - \mu^2/2} \right\}.$$

To analyze the LR test, let us consider

$$\begin{aligned} &\varepsilon = \varepsilon_n = n^{-\beta}, \quad 1/2 < \beta < 1, \\ &\mu = \mu_n = \sqrt{2r\log(n)}, \quad 0 < r < 1, \end{aligned}$$

where β controls the signal density while r specifies the signal strength. Further, define a threshold curve for the parameter r

$$\rho(\beta) = \begin{cases} \beta - 1/2 & 1/2 < \beta < 3/4, \\ (1-\sqrt{1-\beta})^2 & 3/4 \le \beta \le 1. \end{cases}$$

It has been shown that (see the thesis by Jiashun Jin)

- If $r > \rho(\beta)$, for the LR test,

$$\mathbb{P}_0(\text{Type I error}) + \mathbb{P}_1(\text{Type II error}) \to 0.$$

- If $r \le \rho(\beta)$, for any test,

$$\liminf_n \mathbb{P}_0(\text{Type I error}) + \mathbb{P}_1(\text{Type II error}) \ge 1.$$

In the above sense, the LR test is rate-optimal. However, it cannot be used in practice as μ and ε are unknown in practice. Donoho and Jin (2004, AOS) showed that HC_n achieves the same detection threshold as the LR test does based on $p_i = P(Z > X_i) = 1 - \Phi(X_i)$, where $Z \sim N(0,1)$.

Comparison with Bonferroni's method. Recall that the Bonferron's method rejects the null when

$$\max_i X_i \ge \Phi^{-1}(1-\alpha/n) = \sqrt{2\log(n)}(1+o(1)).$$

It can be shown that (exercise)

$$\max_{i:\text{non-null}} X_i \approx \sqrt{2r\log(n)} + \sqrt{2(1-\beta)\log(n)}.$$

For the Bonferron's test to be rejected, we need

$$\sqrt{2r\log(n)} + \sqrt{2(1-\beta)\log(n)} > \sqrt{2\log(n)},$$

It thus implies that $\sqrt{r} + \sqrt{1-\beta} > 1$ or equivalently $r > (1-\sqrt{1-\beta})^2$, which is optimal for $3/4 \le \beta \le 1$.

Exercise 3.2: Under the above setups, show that

$$\max_{i:\text{non-null}} X_i = \left\{ \sqrt{2r\log(n)} + \sqrt{2(1-\beta)\log(n)} \right\} (1 + o(1)).$$

Rejection threshold in the higher-criticism test. Let

$$W_n(t) = \frac{F_n(t) - t}{\sqrt{t(1-t)/n}}.$$

It can be shown that

$$\sup_{1/n \le t \le t_0} \frac{W(t)}{\sqrt{2\log\log(n)}} \xrightarrow{p} 1,$$

which suggests that the rejection threshold should be greater than $\sqrt{2\log\log(n)}$. For some $\epsilon > 0$, if we reject the null when $\text{HC}_n \ge \sqrt{2(1+\epsilon)\log\log(n)}$, Donoho and Jin (2004, AOS) showed that

$$\mathbb{P}_0(\text{Type I error}) + \mathbb{P}_1(\text{Type II error}) \to 0.$$

Variations of higher-criticism test. One variation of the original higher-criticism test is the Berk-Jones statistic, which standardizes the binomial counts using a log-likelihood ratio transformation rather than a normal approximation. Specifically, it is defined as

$$\text{BJ}_n = \max_{1 \le i \le n/2} D(p_{(i)}, i/n),$$

where $D(a,b) = a\log(a/b) + (1-a)\log((1-a)/(1-b))$ for $a, b \in [0,1]$ is the Kullback-Leibler distance between $\text{Ber}(a)$ and $\text{Ber}(b)$ distributions. Another variant is the average likelihood ratio test, defined as

$$\text{ALR} = \sum_{i=1}^{n} w_i \exp(nD(p_{(i)}, i/n)), \quad w_i = \frac{1}{2i\log(n/3)}.$$

5 Blessing of dependence

The above discussions assume that the test statistics from different hypotheses are independent of each other. In reality, the tests could be dependent on each other, e.g., genes on the same biological pathway or observations from neighboring locations. Below, we discuss how to derive the maximum/Bonferron's test statistic under dependence.

Let $X_i = (x_{i,1}, \ldots, x_{i,p})$ be a sequence of i.i.d $N(\theta, \Sigma)$ random vectors with $\theta = (\theta_1, \ldots, \theta_p)^\top$ and $\Sigma = (\sigma_{i,j})_{i,j=1}^p$. Let $\Theta_k = \{\mathbf{a} \in \mathbb{R}^p : \|\mathbf{a}\|_0 = k\}$ with $\|\mathbf{a}\|_0$ being the number of nonzero components of $\mathbf{a}$. Consider the problem:

$$H_0 : \theta = \mathbf{0} \quad \text{versus} \quad H_{1,k} : \theta \in \Theta_k.$$

Let $\Gamma = \Sigma^{-1} = (\gamma_{i,j})_{i,j=1}^p$ be the precision matrix. Up to an additive constant and scaling, the LR test is given by

$$\text{LR}_n(k) = \max_{\theta \in \Theta_k} \sum_{i=1}^{n} \left\{ X_i^\top \Gamma X_i - (X_i - \theta)^\top \Gamma (X_i - \theta) \right\}.$$

Exercise 3.3: For $S \subseteq 1, 2, \ldots, p$, let $\Gamma_{S,S}$ be the submatrix of Γ that contains the rows and columns in S. Similarly, define $\Gamma_{S,-S}$ with the rows in S and the columns in $\{1, 2 \ldots, p\} \setminus S$. Let $\bar{X}_n = (\bar{x}_{1,n}, \ldots, \bar{x}_{p,n})^\top [= n^{-1}\sum_{i=1}^n X_i$ and $Z = (z_1, \ldots, z_p)^\top = \Gamma \bar{X}_n$ Show that

$$\text{LR}_n(k) = n \max_{S:\text{card}(S)=k} Z_S^\top \Gamma_{S,S}^{-1} Z_S$$

where the maximization is over all the subset $S \subseteq \{1, 2, \ldots, p\}$ with cardinality k.

Maximum test. When $k = 1$, we have

$$\mathrm{LR}_n(1) = n \max_{1 \le j \le p} \frac{|z_j|^2}{\gamma_{j,j}},$$

which has been considered by Cai et al. (2014, JRSSB) for the two-sample problem. Cai et al. (2014, JRSSB) pointed out that the linear transformation ΓX_i magnifies the signals and the number of the signals owing to the dependence in the data.

Exercise 3.4: Run some simulations to show that LR(1) is more powerful than the usual maximum test $n \max_{1 \le j \le p} |\bar{x}_{j,n}|^2 / \sigma_{j,j}$ under dependence (i.e., $\Gamma \neq \mathbf{I}_p$). In particular, generate i.i.d samples $\{X_1, \ldots, X_n\}$ from $N(\theta, \Sigma)$, where $\Sigma = (\sigma_{ij})$ for $\sigma_{ij} = (0.5)^{|i-j|}$ and θ is a sparse vector.