ECE598: Information-theoretic methods in high-dimensional statistics	Spring 2016
Lecture 23: Functional estimation and testing	
Lecturer: Yihong Wu	*Scribe: Pengkun Yang, April 26, 2016*

Outline:

- GLM: estimating $\theta_{\max}$. More careful application of χ^2-method yields the sharp constant.
- Covariance matrix (independence testing): estimating a scalar functional can require as many samples needed as estimating the whole parameter.
- Uniformity testing: Is lottery fair?

23.1 GLM: estimating $\theta_{\max}$

The model of the observations are the same as before: $X = \theta + Z$ where $Z \sim N(0, I_p)$. We want to estimate the magnitude of θ, i.e., $T(\theta) = \theta_{\max}$. We will show the minimax risk with sharp constant in high dimensions:

$$\inf_{\hat{T}} \sup_{\theta \in \mathbb{R}^p} \mathbb{E}_\theta (\hat{T} - \theta_{\max})^2 = \left(\frac{1}{2} + o(1)\right) \log p, \quad p \to \infty.$$

Upper bound: Let's first analyze the maximum likelihood estimator, namely, $X_{\max}$. Consider $\theta = \alpha e_1$ Then $X_{\max} = \max\{\alpha + Z_1, Z_2, \ldots, Z_p\} \approx \max\left\{\alpha + Z_1, \sqrt{2 \log p}\right\}$. The picture is the blue curve in Fig. 23.1. A better idea in this case is to decrease $X_{\max}$ by $\sqrt{2 \log p}/2$, which will reduce the worst case error.

Let $\hat{T} = X_{\max} - \frac{\sqrt{2 \log p}}{2}$. WLOG, consider $\theta_{\max} = \theta_1$. Then

$$\hat{T} - \theta_{\max} = \max_i \left\{X_{\max} - \frac{\sqrt{2\log p}}{2} - \theta_{\max}\right\} \le \max_i Z_i - \sqrt{\frac{\log p}{2}} \overset{\text{w.h.p.}}{\le} \sqrt{\frac{\log p}{2}}(1 + o(1)),$$
$$\hat{T} - \theta_{\max} \ge X_1 - \frac{\sqrt{2 \log p}}{2} - \theta_{\max} = Z_1 - \sqrt{\frac{\log p}{2}} \ge O_P\left(-\sqrt{\frac{\log p}{2}}(1 + o(1))\right).$$

Lower bound: Consider two hypotheses:

$$H_0 : \theta = 0, \qquad H_1 : \theta_{\max} \ge \tau.$$

Put a prior on H_1: $\theta \sim \text{Uniform}\{\tau e_1, \tau e_2, \ldots, \tau e_p\}$. Then under H_0 the sample $X \sim P_0 = N(0, I_p)$ and under H_1 the sample $X \sim P_\pi = \frac{1}{p}\sum_{i=1}^p N(\tau e_i, I_p)$. The goal is to show that $d_{\text{TV}}(P_0, P_\pi) \to 0$ when $\tau = \sqrt{(2 - \epsilon) \log p}$ for any $\epsilon > 0$.

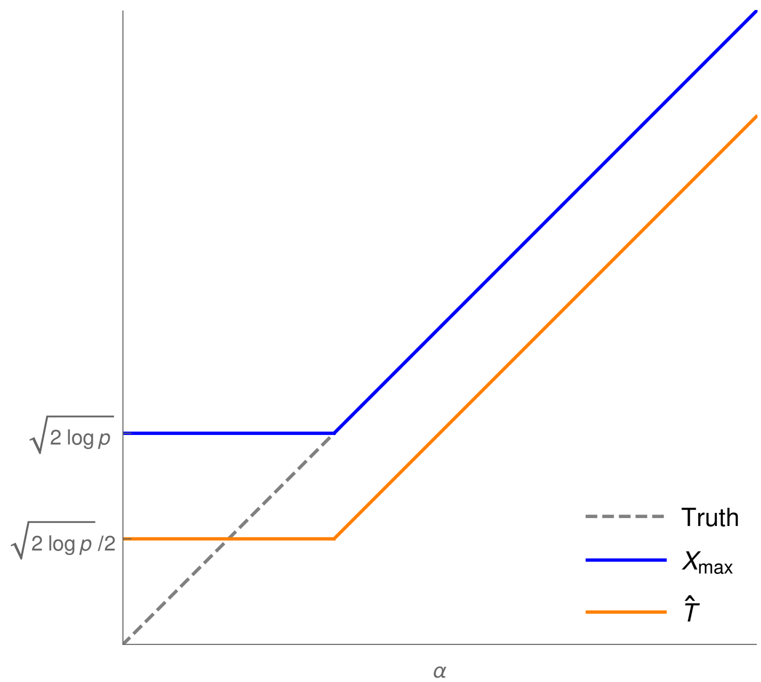


Figure 23.1: Maximum likelihood estimator and improvement via de-biasing.

In this problem, directly applying χ^2-method yields the minimax rate but not the sharp constant: Let $\theta = \tau e_I$ and $\tilde{\theta} = \tau e_{\tilde{I}}$, where $I, \tilde{I} \overset{\text{i.i.d.}}{\sim} \text{Uniform}[p]$.

$$\chi^2(P_\pi \| P_0) = \mathbb{E}\exp\langle\theta, \tilde{\theta}\rangle - 1 = \mathbb{E}\exp\left(\tau^2 \mathbf{1}_{\{I \neq \tilde{I}\}}\right) - 1 = \frac{\exp(\tau^2) - 1}{p}.$$

Therefore $\chi^2(P_\pi \| P_0) \to 0 \Leftrightarrow \frac{\tau}{\sqrt{\log p}} < 1$ and we conclude that $R^* \geq \frac{1+o(1)}{4}\log p$.

We can apply χ^2-method more carefully by conditioning on some high probability event. The main idea is that low probability event has vanishing contribution on the total variation distance but may contribute a lot to the χ^2 distance. Let $\tau = \sqrt{(2-\epsilon)\log p}$ and let

$$E = \left\{\max_i X_i \leq \sqrt{2\log p}\right\}.$$

Since $\max_i Z_i \leq \sqrt{2\log p}$ with high probability, and $Z_i = O_P(1)$ for any fixed i, E is an high probability event under *both* P_0 *and* P_π. Denote by P_0^E and P_π^E the probability measure conditioned on E, that is, $P_0^E(\cdot) = \frac{P_0(\cdot \cap E)}{P_0(E)}$. Note that

$$d_{\text{TV}}(P_0, P_0^E) = 1 - P_0(E), \quad d_{\text{TV}}(P_\pi, P_\pi^E) = 1 - P_\pi(E). \tag{23.1}$$

By triangle inequality, it suffices to show that $d_{\text{TV}}(P_0^E, P_\pi^E) \to 0$. By the formula for conditional probability, the likelihood ratio is

$$\frac{P_\pi^E}{P_0^E} = \frac{P_0(E)}{P_\pi(E)}\frac{P_\pi}{P_0}\mathbf{1}_E.$$

Applying χ^2-method on P_0^E and P_π^E, we obtain that

$$\begin{aligned}
\int \frac{P_\pi^2}{P_0}\mathbf{1}_E &= \mathbb{E}\left[\int \frac{P_\theta P_{\tilde\theta}}{P_0}\mathbf{1}_E\right] = \mathbb{E}_{\theta,\tilde\theta}\mathbb{E}_{X\sim N(\tilde\theta, I_p)}\left[\exp\left(-\frac{\|\theta\|_2^2}{2} + \langle\theta, X\rangle\right)\mathbf{1}_E\right] \\
&= \mathbb{E}_X\mathbb{E}_I\left[\exp\left(-\tau^2/2 + \tau\langle X, e_I\rangle\right)\mathbf{1}_E\right] \\
&= \left(1-\frac{1}{p}\right)\mathbb{E}\left[\exp\left(-\tau^2/2 + \tau N(0,1)\right)\mathbf{1}_E\right] + \frac{1}{p}\mathbb{E}\left[\exp\left(-\tau^2/2 + \tau X_1\right)\mathbf{1}_E\right] \\
&\le \left(1-\frac{1}{p}\right) + \frac{1}{p}\exp\left(\left(-\frac{2-\epsilon}{2} + \sqrt{2(2-\epsilon)}\right)\log p\right).
\end{aligned}$$

Note that $-(2-\epsilon)/2+\sqrt{2(2-\epsilon)} < 1$ as long as $\epsilon > 0$. Therefore $\int \frac{P_\pi^2}{P_0}\mathbf{1}_E = 1+o(1)$ and consequently

$$\begin{aligned}
\chi^2(P_\pi^E \| P_0^E) = o(1) &\implies d_{\mathrm{TV}}(P_\pi^E, P_0^E) = o(1) \\
&\overset{(23.1)}{\implies} d_{\mathrm{TV}}(P_\pi, P_0) = o(1) \\
&\overset{\text{LeCam}}{\implies} R^* \ge \frac{1+o(1)}{2}\log p,
\end{aligned}$$

where we apply LeCam's method for quadratic risk in Theorem **??**.

23.2 Covariance matrix

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(0, \Sigma)$, where Σ is the covariance matrix with size $p \times p$. A sufficient statistic for Σ is the sample covariance matrix:

$$S = \frac{1}{n}\sum_{i=1}^n X_i X_i'.$$

Let $\Theta = \left\{\Sigma : \|\Sigma\|_{op} \le \lambda\right\}$. The minimax risk for estimating Σ under the operator norm is

$$R_1^* \triangleq \inf_{\hat\Sigma}\sup_{\Sigma\in\Theta}\mathbb{E}\|\hat\Sigma - \Sigma\|_{op}^2 \asymp \lambda^2\left(1 \wedge \frac{p}{n}\right).$$

Even if we only want to estimate the operator norm, a scalar functional of Σ, the difficulty in terms of the minimax rate is the same as estimating Σ itself:

$$R_2^* \triangleq \inf_{\widehat{\|\Sigma\|_{op}}}\sup_{\Sigma\in\Theta}\mathbb{E}\left(\widehat{\|\Sigma\|_{op}} - \|\Sigma\|_{op}\right)^2 \asymp \lambda^2\left(1 \wedge \frac{p}{n}\right).$$

Note that $\|\hat\Sigma\|_{op}$ is a viable estimator for $\|\Sigma\|_{op}$. By the triangle inequality of the operator norm,

$$R_2^* \lesssim R_1^*.$$

It suffices to show an upper bound for estimating Σ and the same lower bound for estimating $\|\Sigma\|_{op}$.

Upper bound for estimating Σ: Note a trivial upper bound that $R_1^* \le \lambda^2$. It remains to show that $R_1^* \lesssim \lambda^2 p/n$ when $n \gtrsim p$. Consider the sufficient statistic S. We want to show that for any $\|\Sigma\|_{op} \le \lambda$,

$$\|S - \Sigma\|_{op} \overset{\text{w.h.p}}{\le} \lambda \sqrt{\frac{p}{n}},$$

when $n \gtrsim p$. Let $X_i = \Sigma^{1/2} Z_i$ then $Z_i \overset{\text{i.i.d.}}{\sim} N(0, I_p)$ and $S = \Sigma^{1/2}(\frac{1}{n}\sum_{i=1}^n Z_i Z_i')\Sigma'^{1/2}$. Let $\tilde{S} \triangleq \frac{1}{n}\sum_{i=1}^n Z_i Z_i'$ then

$$\|S - \Sigma\|_{op} = \|\Sigma^{1/2}(\tilde{S} - I_p)\Sigma'^{1/2}\|_{op} \le \|\Sigma^{1/2}\|_{op}\|\tilde{S} - I_p\|_{op}\|\Sigma'^{1/2}\|_{op} = \lambda\|\tilde{S} - I_p\|_{op}.$$

We use the result that, with high probability,

$$\|\tilde{S} - I_p\|_{op}^2 \lesssim \sqrt{\frac{p}{n}} + \frac{p}{n}.$$

The intuition for the above result is that

$$\|\tilde{S} - I_p\|_{op}^2 \le \sup_{\|v\|=1} \|\tilde{S}v\|^2 + 1 - 2\inf_{\|v\|=1}\|\tilde{S}v\| \approx (1 + \sqrt{p/n})^2 + 1 - 2(1 - \sqrt{p/n}) = 4\sqrt{\frac{p}{n}} + \frac{p}{n}.$$

When $n \gtrsim p$ we have $\|\tilde{S} - I_p\|_{op} \overset{\text{w.h.p.}}{\lesssim} \sqrt{p/n}$.

Lower bound for estimating $\|\Sigma\|_{op}$: Let $a, b > 0$ be two parameters to be specified in the end. Consider two hypotheses:

$$H_0 : \Sigma = \Sigma_0 = aI, \qquad H_1 : \Sigma = \Sigma_v = aI + bvv',$$

where under the alternative Σ is a rank-one perturbation from the identity matrix. Then the operator norms under H_0 and H_1 are separated by b. Put a prior on H_1 that $v \sim \text{Uniform}\left\{\frac{\pm 1}{\sqrt{p}}\right\}^p$.

Applying the χ^2-method, we obtain that

$$\begin{aligned}
\chi^2 + 1 &= \mathbb{E}_{v,\tilde{v}} \int \frac{N(0,\Sigma_v)^{\otimes n} N(0,\Sigma_{\tilde{v}})^{\otimes n}}{N(0,\Sigma_0)^{\otimes n}} = \mathbb{E}_{v,\tilde{v}}\left(\int \frac{N(0,\Sigma_v)N(0,\Sigma_{\tilde{v}})}{N(0,\Sigma_0)}\right)^n \\
&= \mathbb{E}_{v,\tilde{v}}\left(\sqrt{\frac{|\Sigma_0|}{|\Sigma_v||\Sigma_{\tilde{v}}||\Sigma_v^{-1} + \Sigma_{\tilde{v}}^{-1} - \Sigma_0^{-1}|}}\right)^n = \mathbb{E}_{v,\tilde{v}}\left(\det\left(I_p - \frac{b^2}{a^2} vv'\tilde{v}\tilde{v}'\right)\right)^{-n/2} \\
&= \mathbb{E}_{v,\tilde{v}}\left(\det\left(I_p - \frac{b^2}{a^2}\langle v', \tilde{v}\rangle v\tilde{v}'\right)\right)^{-n/2}.
\end{aligned}$$

Applying matrix determinant lemma that $\det(A + uv') = (1 + v'A^{-1}u)\det(A)$ yields that

$$\chi^2 + 1 = \mathbb{E}_{v,\tilde{v}}\left(1 - \frac{b^2}{a^2}\langle v', \tilde{v}\rangle^2\right)^{-n/2} \le \mathbb{E}_{v,\tilde{v}} \exp\left(\frac{nb^2}{2a^2}\langle v', \tilde{v}\rangle^2\right).$$

Note that the distribution of $\langle v', \tilde{v}\rangle$ is the same as $\frac{1}{p}\sum_{i=1}^p R_i$ where R_i is an i.i.d. Rademacher random variable taking values ± 1 with probability $1/2$. Then $\langle v', \tilde{v}\rangle$ is concentrated on $[-\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}]$

(this can be made rigorous through Hungarian coupling). The problem boils down to the following simple optimization:

$$\begin{aligned} &\max \ b \\ &\text{s.t. } 0 \le a \le a+b \le \lambda, \\ &\frac{nb^2}{a^2 p} \le c, \end{aligned}$$

for some constant c. The optimal solution is

$$b = \frac{\lambda}{1+\sqrt{n/cp}} \asymp \lambda \left(1 \wedge \sqrt{\frac{p}{n}}\right).$$

23.3 Uniformity testing: Is the lottery fair?

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P$ where P is a distribution on $[k]$. Consider two hypotheses:

$$H_0 : P = \text{Uniform}[k], \qquad H_1 : d_{\text{TV}}(P, \text{Uniform}[k]) \ge \epsilon.$$

A test is a function $\psi : [k]^n \to \{0,1\}$ and we want the probability of error to be

$$P_0^{\otimes n}(\psi = 1) + \sup_{P \in H_1} P^{\otimes n}(\psi = 0) \le 1\%.$$

The sample complexity $n^*(k,\epsilon)$ is defined by the minimum sample size n such that a satisfactory test exists.

Theorem 23.1 ([Pan08]).

$$n^*(k,\epsilon) \asymp \frac{\sqrt{k}}{\epsilon^2}.$$

Remark 23.1. Estimating P by $\hat{P}$ such that $d_{\text{TV}}(P,\hat{P}) \le \epsilon$ requires $\asymp k/\epsilon^2$ samples.

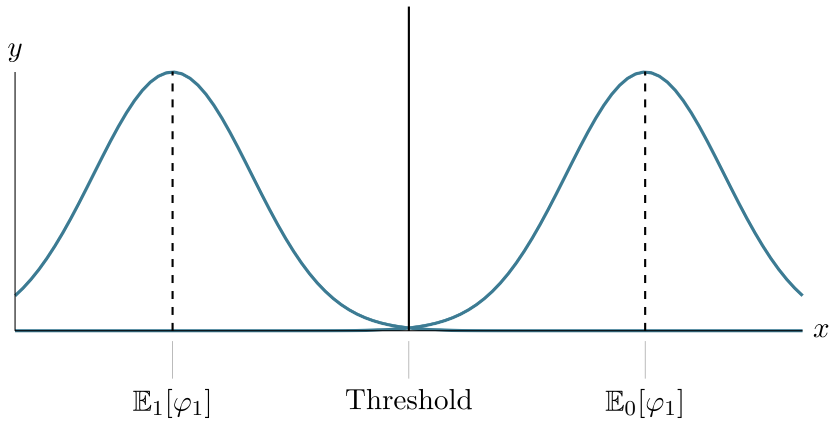
To estimate any functional of a distribution, a sufficient statistic is the histogram $(N_1, \dots, N_k)$ where N_i records the number of appearances of symbol i. Since the total variation distance is permutation invariant (symmetric), a further sufficient statistic is the profile/histogram of histogram $(\varphi_1, \dots, \varphi_n)$, where φ_i counts the number of symbols that appear exactly i times.

Upper bound: Our test statistic is φ_1. This is connected to "birthday paradox": consider k days and n people,

$$\mathbb{P}[\text{no coincident birthday}] = \frac{k}{k}\frac{k-1}{k}\cdots\frac{k-n+1}{k} = \exp\left(\sum_{i=1}^{n-1} \log(1-i/k)\right) \approx \exp(-n^2/2k).$$

When $n \lesssim \sqrt{k}$ then $\varphi_1 \approx n$. The intuition is that the coincidence is least likely under uniform distribution: φ_1 is large (close to n) under H_0 and φ_1 is small under H_1.

By definition $\varphi_1 = \sum_{i=1}^k \mathbf{1}_{N_i=1}$. We can compute that $\mathbb{E}_0[\varphi_1] - \mathbb{E}_1[\varphi_1] \gtrsim \frac{n^2\epsilon^2}{k}$ and $\text{var}_0[\varphi_1] \lesssim \frac{n^2}{k}$. If $n \gtrsim \frac{\sqrt{k}}{\epsilon^2}$ then $\sqrt{\text{var}_0[\varphi_1]} \lesssim \mathbb{E}_0[\varphi_1] - \mathbb{E}_1[\varphi_1]$. Under H_1 we can also compute that $\sqrt{\text{var}_0[\varphi_1]} \lesssim \mathbb{E}_0[\varphi_1] - \mathbb{E}_1[\varphi_1]$. The picture is shown as below and the detailed computation is referred to [Pan08].



Lower bound: Consider two hypotheses:

$$H_0 : P = \text{Uniform}[k], \qquad H_1 : P = P_I = (p_1, \dots, p_k),$$

where $I \subseteq [k]$ is of size $k/2$ and

$$p_i = \begin{cases} \frac{1+\epsilon}{k}, & i \in I, \\ \frac{1-\epsilon}{k}, & i \notin I. \end{cases}$$

Put the uniform prior on H_1 where I is chosen uniformly at random from all subsets of size $k/2$. The goal is to show that

$$d_{\text{TV}}\left(\frac{1}{\binom{k}{k/2}} \sum_{|I|=k/2} P_I^{\otimes n}, \text{Uniform}[k]^{\otimes n} \right) < c$$

for some constant $c < 1$. A sufficient condition is that

$$\chi^2 \left(\frac{1}{\binom{k}{k/2}} \sum_{|I|=k/2} P_I^{\otimes n} \middle\| \text{Uniform}[k]^{\otimes n} \right) < \infty.$$

Applying the Ingster-Suslina method (Lemma **??**):

$$\begin{aligned} \chi^2 + 1 &= \mathbb{E}_{I,\tilde{I}} \int \frac{P_I^{\otimes n} P_{\tilde{I}}^{\otimes n}}{P_0^{\otimes n}} = \mathbb{E}_{I,\tilde{I}} \left(\sum \frac{P_I P_{\tilde{I}}}{P_0} \right)^n = \mathbb{E}_{I,\tilde{I}} \left(\frac{4\epsilon^2 |I \cap \tilde{I}|}{k} + 1 - \epsilon^2 \right)^n \\ &\le \mathbb{E}_{I,\tilde{I}} \exp\left(n\epsilon^2 \left(\frac{4|I \cap \tilde{I}|}{k} - 1 \right) \right), \end{aligned}$$

where $I \cap \tilde{I} \sim \text{HyperGeometric}(k, k/2, k/2)$. Applying the convex stochastic dominance of the hypergeometric distribution over the binomial distribution, we obtain that

$$\begin{aligned} \chi^2 + 1 &\le \mathbb{E}_{I,\tilde{I}} \exp\left(n\epsilon^2 \left(\frac{4\text{Binom}(k, 1/2)}{k} - 1 \right) \right) = \left(\frac{\exp(2n\epsilon^2/k) + \exp(-2n\epsilon^2/k)}{2} \right)^{k/2} \\ &\le \exp\left(\frac{1}{2} \left(\frac{2n\epsilon^2}{k} \right)^2 \frac{k}{2} \right) < \infty, \end{aligned}$$

when $n \lesssim \frac{\sqrt{k}}{\epsilon^2}$, where we used the inequality that $\frac{e^x + e^{-x}}{2} \le e^{x^2/2}$ (by Taylor expansion).

References

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inf. Theory*, 54(10):4750–4755, 2008.