Lecture 4
Global testing in high-dimensions

23 April 2020

Aldo Solari
University of Milano-Bicocca
Statistical Inference II
PhD in Economics and Statistics

High-dimensional statistics

Classical theory

- It concerns the behaviour when the *sample size* $n \to \infty$
- Suppose $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \underset{m \times 1}{Y}$ with mean $\mu = \mathbb{E}(Y)$ and finite variance $\Sigma = \mathbb{V}\mathrm{ar}(Y)$
- *Law of large numbers*: the sample mean $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$ converges in probability to μ
- *Central limit theorem*: the rescaled deviation $\sqrt{n}(\hat{\mu}_n - \mu)$ converges in distribution to a centered Gaussian with covariance matrix Σ
- *Consistency of maximum likelihood estimation*
- Etc.

Suppose that we are given $n = 1000$ samples from a statistical model in $m = 500$ dimensions

Will theory that requires $n \to \infty$ with the dimension m remaining fixed provide useful predictions?

High-dimensional data

- The data sets arising in many parts of modern science have a “high-dimensional flavor”, with m on the same order as, or possibly larger than n

$$m \gg n$$

- Classical “large n, fixed m” theory fails to provide useful predictions
- Classical methods can break down dramatically in high-dimensional regimes

Reference
Wainwright (2019)
High-Dimensional Statistics: A Non-Asymptotic Viewpoint
Cambridge University Press

Linear discriminant analysis in high-dimensions

Two classes

- Hypothesis testing can be used to determine whether an observed vector $x = (x_1, \ldots, x_m)^\mathsf{T} \in \mathbb{R}^m$ has been drawn from one of two possible densities $f_A \equiv \mathrm{P}(X|Y = A)$ and $f_B \equiv \mathrm{P}(X|Y = B)$, corresponding to two possible classes A and B
- Consider testing $H_A : X_A \sim f_A$ vs $H_B : X_B \sim f_B$, where $X_A \equiv (X|Y = A)$ and $X_B \equiv (X|Y = B)$
- When these two distributions are known, then the Neyman-Pearson lemma says that the optimal decision rule is based on thresholding the log-likelihood ratio

$$\log \frac{f_B(x)}{f_A(x)}$$

- By testing H_A vs H_B and H_B vs H_A the conclusion is that the observed data x is consistent with A (H_B rejected), with B (H_A rejected), with both (no rejections), or with neither (both rejected)

Classification problem

- Let's turn to the classification problem involving the allocation of the observed unit x to one of two classes A and B
- For a Bayesian analysis suppose that the prior probabilities are $\pi_A \equiv \mathrm{P}(Y = A)$ and $\pi_B \equiv \mathrm{P}(Y = B)$ with $\pi_A + \pi_B = 1$. Then the posterior probabilities satisfy

$$\frac{\mathrm{P}(Y = B|X = x)}{\mathrm{P}(Y = A|X = x)} = \frac{\pi_B}{\pi_A}\frac{f_B(x)}{f_A(x)}$$

  giving the class with the higher posterior probability
- As a special case, suppose that the two classes are distributed as multivariate Gaussians $X_A \sim N(\mu_A, I_m)$ and $X_B \sim N(\mu_B, I_m)$, with $\pi_A = \pi_B = 1/2$

Optimal decision

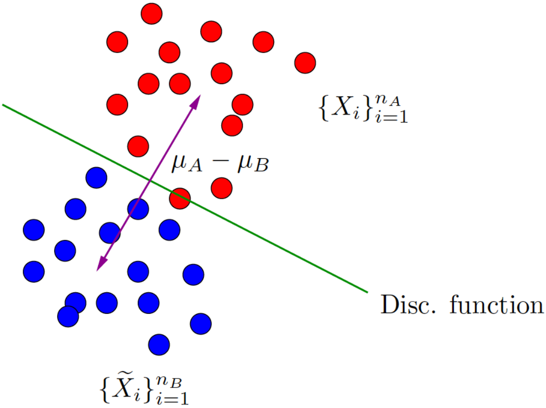
- The optimal decision rule is to threshold the log-likelihood ratio

$$\Psi(x) = \langle \mu_A - \mu_B, \left(x - \frac{\mu_A + \mu_B}{2}\right) \rangle$$

  where $\langle x, z \rangle = x^\top z = \sum_{j=1}^m x_j z_j$ denotes the Euclidean inner product in $\mathbb{R}^m$
- If $\Psi(x) > 0$ then classify A, otherwise B
- Error probability of the optimal rule:

$$\mathrm{Err}(\Psi) = \frac{1}{2}\mathrm{P}(\Psi(X_A) < 0) + \frac{1}{2}\mathrm{P}(\Psi(X_B) \geq 0) = \Phi\left(-\frac{\gamma}{2}\right)$$

  where $\gamma = \|\mu_A - \mu_B\|_2$, $\|\mu\|_2 = \sqrt{\mu^\top \mu}$, and Φ is the cdf of a standard normal variable



$$\langle \mu_A - \mu_B, \left(x - \frac{\mu_A + \mu_B}{2}\right) \rangle = 0$$

source: Wainwright

Linear Discriminant Analysis

- Fisher's LDA: uses the plug-in principle based on n_A samples from class A and n_B samples from class B

$$\hat{\Psi}(x) = \langle \hat{\mu}_A - \hat{\mu}_B, x - \frac{\hat{\mu}_A + \hat{\mu}_B}{2} \rangle$$

- Error probability of LDA (is itself a random variable)

$$\mathrm{Err}(\hat{\Psi}) = \frac{1}{2}\mathrm{P}(\hat{\Psi}(X_A) < 0) + \frac{1}{2}\mathrm{P}(\hat{\Psi}(X_B) \geq 0)$$

- Classical theory: if $(n_A, n_B) \to \infty$ and m remains fixed, then $\hat{\mu}_A \overset{prob.}{\to} \mu_A$, $\hat{\mu}_B \overset{prob.}{\to} \mu_B$ and the asymptotic error probability is $\mathrm{Err}(\hat{\Psi}) \overset{prob.}{\to} \mathrm{Err}(\Psi) = \Phi(-\gamma/2)$

High-Dimensional Theory

- What happens if $(n_A, n_B, m) \to \infty$ with
  - $m/n_A \to \delta$ with $\delta \geq 0$
  - $m/n_B \to \delta$
  - $\|\mu_A - \mu_B\|_2 \to \gamma > 0$
- Kolmogorov (1960) showed that

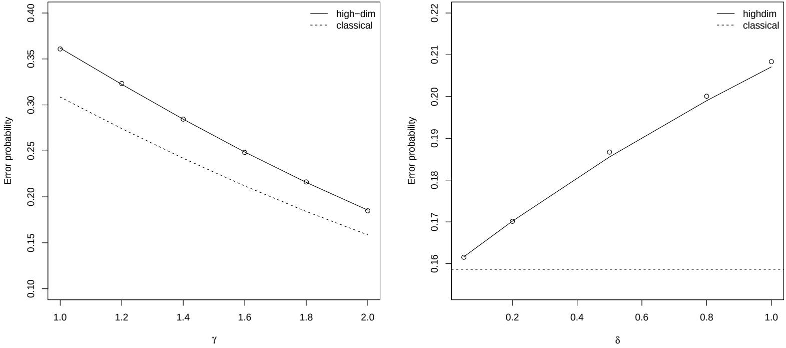
$$\mathrm{Err}(\hat{\Psi}) \overset{prob.}{\to} \Phi\Big(- \frac{\gamma^2}{2\sqrt{\gamma^2 + 2\delta}}\Big)$$

- If $m/n \to 0$, then the asymptotic error probability is $\Phi(-\gamma/2)$ as is predicted by classical theory
- If $m/n \to \delta > 0$, then the asymptotic error probability is strictly larger than $\Phi(-\gamma/2)$

The error probability of $\hat{\Phi}$, for the finite triple

$$(m, n_A, n_B) = (400, 800, 800)$$

is better described by the classical $\Phi(-\gamma/2)$, or the high-dimensional analog $\Phi(-\gamma^2/(2\sqrt{\gamma^2 + 2\delta}))$?



circles correspond to the empirical error probabilities, averaged over 10 trials

What can help us in high dimensions?

- An important fact is that high-dimensional phenomena are unavoidable
- If the ratio m/n stays bounded strictly above zero, then it is not possible to achieve the optimal classification rate
- Our only hope is that the data is endowed with some form of *low-dimensional structure*

- What is the underlying cause of the inaccuracy of the prediction for the LDA in high-dimensions?
- The squared Euclidean error

$$\|\hat{\mu} - \mu\|_2^2 = \sum_{j=1}^{m} (\hat{\mu}_j - \mu_j)^2$$

concentrates sharply around m/n, i.e. for $t \in (0, 1)$

$$\mathrm{P}\left(\left|\|\hat{\mu} - \mu\|_2^2 - \frac{m}{n}\right| \geq \frac{m}{n} t\right) = \mathrm{P}\left(\left|\frac{1}{m} \sum_{j=1}^{m} Z_j^2 - 1\right| \geq t\right) \leq 2e^{-\frac{mt^2}{8}}$$

where $Z_j = \sqrt{n}(\hat{\mu}_j - \mu_j) \sim N(0, 1)$; for the upper bound see Wainwright (2019), Example 2.11

Sparsity

- Suppose that the m-vector μ is *sparse*, with only s of its m entries being nonzero, for some sparsity parameter $s \ll m$
- If sparsity holds, we can obtain a better estimator by thresholding the sample means

$$\tilde{\mu}_j = \hat{\mu}_j \mathbb{1}\{|\hat{\mu}_j| > \lambda\}$$

where

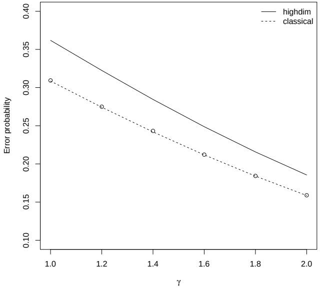
$$\lambda = \sqrt{\frac{2 \log m}{n}}$$

Thresholded mean

Suppose to replace $\hat{\mu}$ by the thresholded mean $\tilde{\mu}$, then

$$\tilde{\Psi}(x) = \langle \tilde{\mu}_A - \tilde{\mu}_B, x - \frac{\tilde{\mu}_A + \tilde{\mu}_B}{2} \rangle$$

approaches the optimal $\mathrm{Err}(\Psi)$ if $\log \binom{m}{s} / n \to 0$. For $s = 5$:



circles correspond to the empirical error probabilities, averaged over 10 trials

Inference for the mean vector

- Random sample of n observations from $y \sim N_m(\mu, \Sigma)$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \sim N_m\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdot & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \cdot & \sigma_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{m1} & \cdot & \cdot & \sigma_m^2 \end{bmatrix}\right)$$

- The parameter of interest is $\mathbb{E}(y) = \mu$, where $\mu_j = 0$ means “no effect” and $\mu_j \neq 0$ means “effect” in the jth component
- The nuisance parameter is the variance/covariance matrix $\mathbb{V}\mathrm{ar}(y) = \Sigma$

Three questions

1. *Detecting effects*: There is at least one μ_j different from 0?
2. *Counting effects*: How many μ_j are different from 0?
3. *Identifying effects*: Which μ_j are different from 0?

Global null hypothesis

Testing the global null hypothesis aims at detecting any effect

$$H_0 : \mu = 0, \text{ i.e. } \bigcap_{j=1}^{m} \{\mu_j = 0\}$$

$$H_1 : \mu \neq 0, \text{ i.e. } \bigcup_{j=1}^{m} \{\mu_j \neq 0\}$$

One-sided alternative

$$H_0 : \bigcap_{j=1}^{m} \{\mu_j = 0\}$$

$$H_1 : \bigcup_{j=1}^{m} \{\mu_j > 0\}$$

MaxT and SumT

- For simplicity, assume $\Sigma = I_m$ and $n = 1$ and consider the one-sided alternative
- $T_j = y_j \sim N(\mu_j, 1)$ for $j = 1, \ldots, m$
- $(T_1, \ldots, T_m)' \overset{H_0}{\sim} N_m(0, I_m)$
- MaxT
$$T_{\max} = \max(T_1, \ldots, T_m)$$
- SumT
$$T_{\text{sum}} = \sum_{j=1}^{m} T_j$$

MaxT

- The critical value $t_{1-\alpha}$ of $T_{\max}$ is

$$P_0(T_{\max} \geq t_{1-\alpha}) = \alpha$$

where $t_{1-\alpha}$ is the $1-\alpha$ quantile of the distribution of the maximum of m independent standard normal variables

$$\int_{t_{1-\alpha}}^{\infty} m\phi(y)\Phi(y)^{m-1} dy = \alpha$$

where ϕ and Φ are the density and cdf of $N(0, 1)$

Critical value approximation

– We can replace $t_{1-\alpha}$ by $z_{1-\frac{\alpha}{m}}$

$$
\begin{aligned}
P_0(T_{\max} \geq z_{1-\frac{\alpha}{m}}) &= P_0\Big(\bigcup_{j=1}^{m}\{T_j \geq z_{1-\frac{\alpha}{m}}\}\Big) \\
&\leq \sum_{j=1}^{m} P_0(T_j \geq z_{1-\frac{\alpha}{m}}) = m\frac{\alpha}{m} = \alpha
\end{aligned}
$$

– The union bound might seem crude, but with independent T_js the size of the test is very near α

$$
\begin{aligned}
P_0(T_{\max} \geq z_{1-\frac{\alpha}{m}}) &= 1 - \prod_{j=1}^{m} P_0(T_j < z_{1-\frac{\alpha}{m}}) \\
&= 1 - \Big(1 - \frac{\alpha}{m}\Big)^m \stackrel{m\to\infty}{\to} 1 - e^{-\alpha}
\end{aligned}
$$

For $\alpha = 0.05$, $1 - e^{-\alpha} = 0.0487$

Magnitude of the critical value

– How large is the threshold $z_{1-\frac{\alpha}{m}}$? For large m

$$\begin{aligned} z_{1-\frac{\alpha}{m}} &\approx \sqrt{2\log m} - \frac{\log(2\log m) + \log 2\pi}{2\sqrt{2\log m}} \\ &\approx \sqrt{2\log m} \end{aligned}$$

with no dependence on α

Needle in a haystack problem

$$H_0 : \mu_j = 0 \text{ for all } j = 1, \ldots, m$$
$$H_1 : \mu_j = c_m > 0,\ \mu_k = 0 \text{ for } k \neq j$$

- What is the limiting power of the MaxT test?

$$\lim_{m\to\infty} \mathrm{P}_1(T_{\max} > z_{1-\frac{\alpha}{m}})$$

It depends on the limiting ratio

$$\lim_{m\to\infty} \frac{c_m}{\sqrt{2\log m}} \lesseqgtr 1$$

where c_m is the value of the single non-zero mean, which depends on m

Two cases:

- Assume without loss of generality that $\mu_1 = c_m$. Suppose $c_m > (1+\epsilon)\sqrt{2\log m}$. Then, for $m \to \infty$

$$P_1(T_{\max} > z_{1-\frac{\alpha}{m}}) \geq P_1(T_1 > z_{1-\frac{\alpha}{m}}) = P(N(0,1) > z_{1-\frac{\alpha}{m}} - c) \to 1$$

- Suppose $c_m < (1-\epsilon)\sqrt{2\log m}$. Then for $m \to \infty$

$$\begin{aligned} P_1(T_{\max} > z_{1-\frac{\alpha}{m}}) &\leq P(T_1 > z_{1-\frac{\alpha}{m}}) + P(\max_{j>1} T_j > z_{1-\frac{\alpha}{m}}) \\ &= P(N(0,1) > z_{1-\frac{\alpha}{m}} - c) + P(\max_{j>1} T_j > z_{1-\frac{\alpha}{m}}) \\ &\to 0 + (1 - e^{-\alpha}) \end{aligned}$$

and the MaxT test has no power

SumT

$$T_{\text{sum}} = \sum_{j=1}^{m} T_j \sim N(\sum_{j=1}^{m} \mu_j, m)$$

- $\frac{T_{\text{sum}}}{\sqrt{m}} \overset{H_0}{\sim} N(0,1)$; $\frac{T_{\text{sum}}}{\sqrt{m}} \overset{H_1}{\sim} N(\theta_m, 1)$ with

  $$\theta_m = \frac{\sum_{j=1}^{m} \mu_j}{\sqrt{m}}$$

  but if $\theta_m \to 0$ when $m \to \infty$, then T_{sum} has no power
- By the Neyman-Pearson lemma, T_{sum} is the UMP test for

  $H_0 : \mu_j = 0$ for all j
  $H_1 : \mu_j = c_m > 0$ for all j

  where $\theta_m = \sqrt{m} c_m$, but if $c_m = \frac{1}{m}$ the UMP test has no power

Comparison

- **Few strong effects**:
  $m^{1/4}$ of the μ_js are equal to $\sqrt{2\log m}$, the rest 0.
  E.g. when $m = 10^6$, $m^{1/4} \approx 36$ and $\sqrt{2\log m} \approx 5.3$. In this setting T_{max} has full power, but T_{sum} has no power because

$$\theta_m = \frac{m^{1/4}\sqrt{2\log m}}{\sqrt{m}} \to 0$$

- **Small, distributed effects**:
  $\sqrt{2m}$ of the μ_js are equal to 3, the rest 0.
  The T_{sum} has (almost) full power, but T_{max} has no power because when m is large it's very likely that the largest y_j value comes from a null μ_j

MinP

- Let $p_j = 1 - \Phi(T_j)$ be the jth p-value
- $p_1, \ldots, p_m \overset{i.i.d.}{\sim} U(0,1)$ under H_0
- The MinP test is based on the minimum p-value

$$p_{\min} = \min(p_1, \ldots, p_m) \overset{H_0}{\sim} \mathrm{Beta}(1, m)$$

- The MinP test rejects H_0 if $p_{\min} \leq 1 - (1-\alpha)^{\frac{1}{m}}$ and has size α:

$$\begin{aligned} \mathrm{P}_0(p_{\min} \leq 1 - (1-\alpha)^{\frac{1}{m}}) &= 1 - \mathrm{P}_0\Big(\bigcap_{i=1}^{m} \{p_i > 1 - (1-\alpha)^{\frac{1}{m}}\} \Big) \\ &= 1 - \big[(1-\alpha)^{\frac{1}{m}}\big]^m = \alpha \end{aligned}$$

Simes test

- $p_1, \ldots, p_m \overset{i.i.d.}{\sim} U(0,1)$ under H_0
- Sort the p-values

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$$

- The null distribution of jth ordered p-value is

$$p_{(j)} \overset{H_0}{\sim} \mathrm{Beta}(j, m-j+1)$$

- The Simes test p-value

$$p_{\mathrm{s}} = \min_{j=1,\ldots,m} \left\{ p_{(j)} \frac{m}{j} \right\} \overset{H_0}{\sim} U(0,1)$$

- It rejects H_0 if

$$\exists\, j : p_{(j)} \leq \frac{\alpha j}{m}$$

Fisher combination

- $p_1, \ldots, p_m \overset{i.i.d.}{\sim} U(0,1)$ under H_0
- Fisher's method of combining p-values

$$T_{\mathrm{f}} = \sum_{j=1}^{m} 2 \log\left(\frac{1}{p_j}\right) \overset{H_0}{\sim} \chi^2_{2m}$$

Higher criticism

- Empirical cdf $\hat{F}_m(t) = \dfrac{\sum_{j=1}^{m} \mathbb{1}\{p_j \leq t\}}{m}$ for $t \in [0, 1]$.
- Since $p_1, \ldots, p_m \overset{i.i.d.}{\sim} U(0, 1)$ under H_0, then

$$m\hat{F}_m(t) \overset{H_0}{\sim} \text{Binomial}(m, t)$$

- The higher criticism test is

$$T_{\text{hc}} = \sup_{t\in[0,1]} \frac{\hat{F}_m(t) - t}{\sqrt{t(1-t)/m}}$$

or equivalently

$$T_{\text{hc}} = \max_{j=1,\ldots,m} \sqrt{m}\frac{(i/m) - p_{(i)}}{\sqrt{p_{(i)}(1 - p_{(i)})}}$$

- For $m \to \infty$, $b_m T_{\mathrm{hc}} - c_m$ converges weakly to the standard Gumbel distribution, where $b_m = \sqrt{2 \log \log m}$ and $c_m = \frac{1}{2}(\log \log \log(m) - 4\pi)$
- For any fixed α and $m \to \infty$, its critical value is

$$t_{1-\alpha} \approx (1 + a)\sqrt{2 \log \log m}$$

for some $a > 0$, e.g. $a = 1.08$ for $m \approx 10^6$ and $\alpha = 0.05$

Mixture distribution

- We assume that our samples follow a mixture of $N(0,1)$ and $N(\mu,1)$ distributions

$$
\begin{aligned}
H_0 &: \quad y_j \overset{i.i.d}{\sim} N(0,1) \\
H_1 &: \quad y_j \overset{i.i.d}{\sim} \pi_0 N(0,1) + \pi_1 N(\mu,1)
\end{aligned}
$$

  where $\pi_1 = 1 - \pi_0$

- To carry out asymptotic analysis, we must specify the dependence scheme of $\pi_1 = \pi_1(m)$ and $\mu = \mu(m)$ on m:

$$
\begin{aligned}
\pi_1 &= m^{-\beta} \qquad && \frac{1}{2} < \beta < 1 \\
\mu &= \sqrt{2r \log m} \qquad && 0 < r < 1
\end{aligned}
$$

- The needle in a haystack problem: $\beta = 1$ and $r = 1$; small distributed effects: $\beta = 1/2$

Threshold curve

Consider the following threshold curve for r

$$\rho_{\text{hc}}(\beta) = \begin{cases} \beta - \frac{1}{2} & \text{if } \frac{1}{2} < \beta \leq \frac{3}{4} \\ (1 - \sqrt{1-\beta})^2 & \text{if } \frac{3}{4} \leq \beta \leq 1 \end{cases}$$

– If $r > \rho_{\text{hc}}(\beta)$ the Neyman-Pearson optimal test achieves

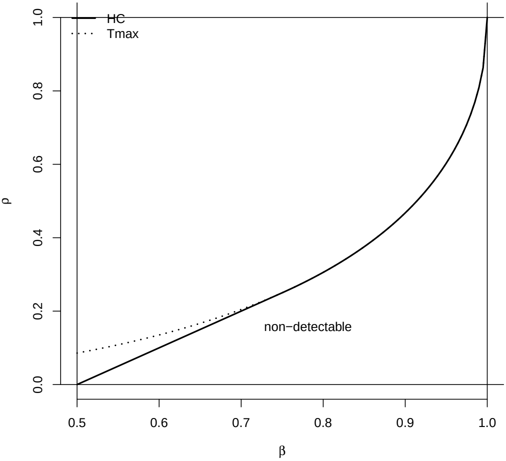
$$\text{P}_0(\text{Type I Error}) + \text{P}_1(\text{Type II Error}) \to 0$$

The Higher Criticism is asymptotically equivalent to the optimal test without knowledge of π_1 and/or μ

– If $r < \rho_{\text{hc}}(\beta)$ then for *any* test

$$\liminf_{m \to \infty} \text{P}_0(\text{Type I Error}) + \text{P}_1(\text{Type II Error}) \geq 1$$

Detectable region



High-dimensional linear model

-
$$y = X\beta + \varepsilon$$

  with response vector $\underset{n\times 1}{y}$, design matrix $\underset{n\times m}{X}$, vector of parameters $\underset{m\times 1}{\beta}$, gaussian errors $\underset{n\times 1}{\varepsilon} \sim N_n(0, \sigma^2 I_n)$ and $m > n$
- For testing $H_0 : \beta = 0$, the global test of Goeman (2006)

$$T_{\mathrm{g}} = y' X X' y$$

- In low dimensions $m < n$, the F statistic is $\propto y' X (X' X)^{-1} X' y$

Reference
Goeman et al. (2006)
Testing against a high dimensional alternative
JRSSB