Statistical Inference II - lecture 3

12 April 2021 09:13

REFERENCES

WAINWRIGHT (2019) SECTIONS 1.1, 1.2, 1.3

CANDÈS, LECTURE NOTES: STATS 300C - THEORY OF STATISTICS

$\boxed{\text{HIGH-DIMENSIONAL STATISTICS}}$

CLASSICAL THEORY

- CONCERNS THE BEHAVIOUR WHEN SAMPLE SIZE $n \to \infty$
- $Y_1, \dots, Y_n$ IID $\underset{m \times 1}{Y}$ WITH $\underset{m\times 1}{\mu} = \mathbb{E}(Y)$ AND $\underset{m \times m}{\Sigma} = \mathbb{V}ar(Y)$ FINITE.
- LAW OF LARGE NUMBERS

  SAMPLE MEAN $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i \xrightarrow{p} \mu$
- CLT $\quad \sqrt{n}\,(\hat{\mu}_n - \mu) \longrightarrow N(0, \Sigma)$
- CONSISTENCY OF MLE
- ...

SUPPOSE DATA $n = 1000$ SAMPLE SIZE

$m = 500$ DIMENSION

QUESTION: IS CLASSICAL THEORY $\left(\begin{array}{l}\text{REQUIRES } n \to \infty \\ m \text{ FIXED}\end{array}\right)$ PROVIDING USEFUL INFORMATION?

HD DATA $m >> n$

CLASSICAL "LARGE n, FIXED m" FAILS IN HD

" STATISTICAL METHODS BREAK DOWN IN HD.

CLASSIFICATION PROBLEM

- $Y \in \{A, B\}$ CLASSES
- DETERMINE WHETHER

DENSITY FUNCTION OF X FROM CLASS A

$\underline{x = (x_1, \dots, x_m)' \in \mathbb{R}^m}$

... FROM $\{$ A: $f_A(x) = P(X \mid Y = A)$

$P(\ \mid Y = B)$

$B: \quad f_B(X) = \mathbb{P}(X \mid Y = B)$

PRIOR PROBAB. $\quad \pi_A = \mathbb{P}(Y = A)$

$\pi_B = \mathbb{P}(Y = B)$

POSTERIOR PROB

$$\frac{\mathbb{P}(Y = B \mid X = x)}{\mathbb{P}(Y = A \mid X = x)} = \frac{\pi_B}{\pi_A} \underbrace{\frac{f_B(x)}{f_A(x)}}_{LR}$$

ALLOCATE TO THE CLASS WITH HIGHER POSTERIOR PROB.

$$X_A \equiv X \mid Y = A \sim N(\mu_A, I_m)$$
$$X_B \equiv X \mid Y = B \sim N(\mu_B, I_m)$$

$\pi_A = \pi_B = 1/2$

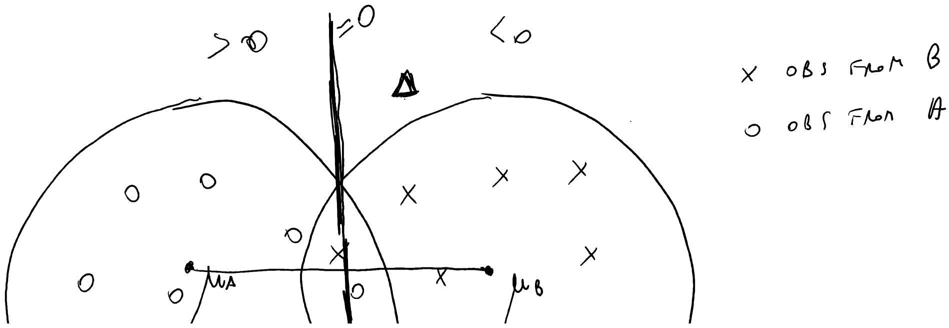
$$\log \frac{f_B(x)}{f_A(x)} = \Psi(x) = \alpha_0 + \alpha^T x$$

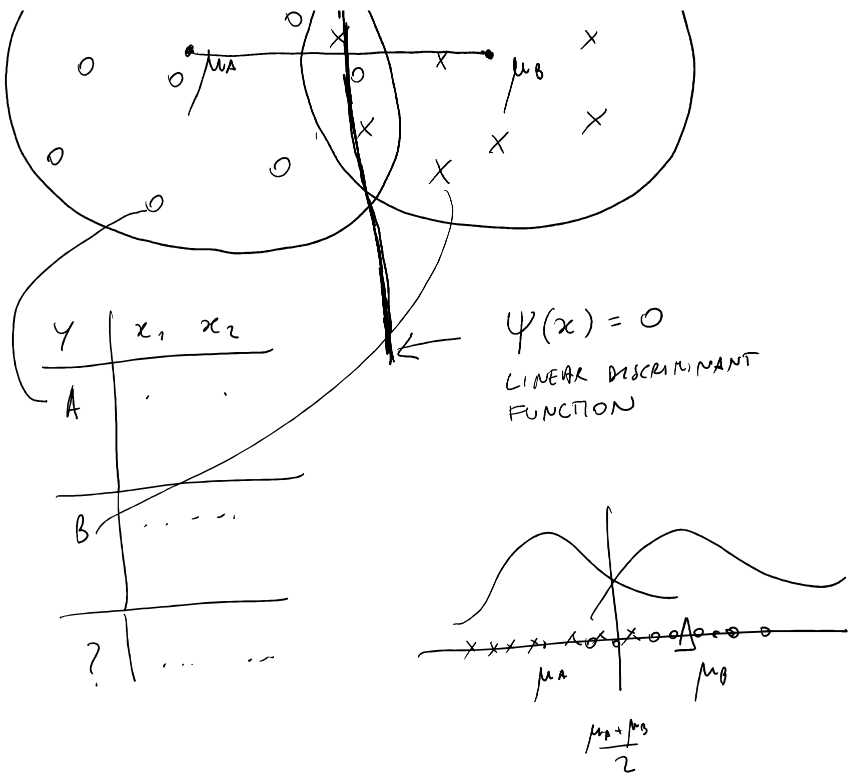
OPTIMAL DECISION IS TO THRESHOLD THE LOG-LIK. RATIO

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

WHERE $\langle x, z \rangle = x^T z = \sum_{j=1}^{m} x_j z_j$ EUCLIDEAN INNER PRODUCT.

$$\text{CLASSIFY} \begin{cases} A & \text{IF } \Psi(x) > 0 \\ B & \text{IF } \Psi(x) \leq 0 \end{cases}$$





ERROR PROBABILITY OF THE OPTIMAL RULE

$$\underline{Err(\psi)} = \frac{1}{2}\,\mathbb{P}\left(\psi(X_A) < 0\right) + \frac{1}{2}\,\mathbb{P}\left(\psi(X_B) \geq 0\right)$$

$$= \Phi\left(-\frac{\gamma}{2}\right) \qquad \Phi \text{ CDF OF } N(0,1)$$

WHERE $\gamma = \| \mu_A - \mu_B \|_2$ $\qquad \|\mu\|_2 = \sqrt{\mu^T \mu}$

EUCL. NORM

FISHER'S LDA.

PLUG-IN PRINCIPLE BASED ON $\begin{cases} n_A \text{ OBS FROM } A \\ n_B \text{ OBS FROM } B \end{cases}$

$$\hat{\psi}(x) = \left\langle \hat{\mu}_A - \hat{\mu}_B,\; x - \frac{\hat{\mu}_A + \hat{\mu}_B}{2} \right\rangle$$

ERROR PROB. OF FISHER'S LDA

$$\underbrace{Err\left(\hat{\psi}(X)\right)} = \frac{1}{2}\,\mathbb{P}\left(\hat{\psi}(X_A) < 0\right) + \frac{1}{2}\,\mathbb{P}\left(\hat{\psi}(X_B) \geq 0\right)$$

IS A RANDOM VARIABLE

CLASSICAL THEORY

$(n_A, n_B) \to \infty \Rightarrow \hat{\mu}_A \to \mu_A \quad \mathrm{Err}(\hat{\psi}) \to \mathrm{Err}(\psi)$

m FIXED $\quad \hat{\mu}_B \to \mu_B \quad = \Phi(-\gamma/2)$

HIGH-DIM THEORY

$(n_A, n_B, m) \to \infty$

$\frac{m}{n_A} \to \delta \geq 0$

$\frac{m}{n_B} \to \delta$

$\| \mu_A - \mu_B \|_2 \to \gamma > 0$

KOLMOGOROV (1960)

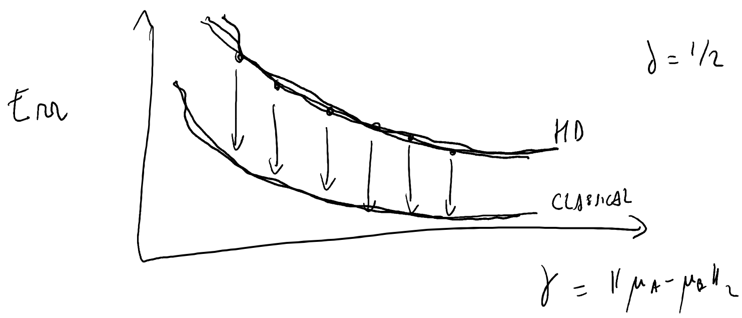
$$\mathrm{Err}(\hat{\psi}) \xrightarrow{P} \Phi\left(- \frac{\gamma^2}{2\sqrt{\gamma^2 + 2\delta}} \right)$$

$$\geq \Phi(-\gamma/2)$$

WHEN $\delta = 0$

EXAMPLE

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



WHAT CAN HELP US IN HIGH-DIMENSIONS ?

HOPE THAT THE "EFFECTIVE" DIMENSION OF THE PROBLEM IS LOW-DIMENSIONAL

SPARSITY

$$\mu = (\underbrace{\mu_1, \dots,}_{\substack{s \\ \neq 0}} \underbrace{\dots, \mu_m}_{\substack{m-s \\ = 0}})^T$$ IS SPARSE WHEN ONLY s ENTRIES ARE $\neq 0$

ESTIMATOR

THRESHOLDED MEAN

$$\tilde{\mu}_j = \hat{\mu}_j \, \mathbb{1}\{ |\hat{\mu}_j| > \lambda \} = \begin{cases} \hat{\mu}_j & \text{IF } |\hat{\mu}_j| > \lambda \\ 0 & \text{O/W} \end{cases}$$

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

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

$$\mathrm{Err}(\tilde{\Psi}(x)) \xrightarrow{P} \mathrm{Err}(\Psi) \quad \text{IF} \quad \log \binom{m}{s} \Big/ n \to 0$$

TESTING THE MEAN VECTOR

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}}_{y} \sim N\left(\underbrace{\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix}}_{\mu}, \Sigma \right)$$

PARAMETER OF INTEREST : $\mu = \mathbb{E}(y)$

NUISANCE PAR. : $\Sigma = \mathbb{V}ar(y)$

$\mu_j = 0$ NO EFFECT

$\mu_j \neq 0$ EFFECT

QUESTIONS

1) DETECT IF THERE IS AT LEAST ONE EFFECT : AT LEAST ONE $\mu_j \neq 0$?

2) COUNTING THE NUMBER OF EFFECTS : HOW MANY $\mu_j \neq 0$?

3) IDENTIFY EFFECTS : WHICH $\mu_j \neq 0$?

GLOBAL NULL HYPOTHESIS

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

$$H_1 : \mu \neq 0 \iff \bigcup_{j=1}^{m} \{\mu_j \neq 0\}$$

DATA

$y =$

3	1	3	2
2	2	y_i	**9**
4	1	**7**	**5**
3	6	2	1

$\xrightarrow{\text{INFERENCE}}$

TRUE PARAMETERS

0	0	0	0
0	0	μ_i	**7**
0	0	**5**	**6**
0	0	0	0

$= \mu$

ASSUMPTIONS

- $\Sigma = I_m$
- $n = 1$
- $y \sim N(\mu, I_m)$

$H_0 : \mu = 0$

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

$\mu \gneq 0$

- $T_j = y_j \sim N(\mu_j, 1) \quad \forall j = 1, \dots, m$

$$(T_1, \dots, T_m)' \overset{H_0}{\sim} N(0, I_m)$$

- MAX T $\qquad T_{MAX} = \max(T_1, \dots, T_m)$
- SUM T $\qquad T_{SUM} = \sum_{j=1}^{m} T_j$

MAX T

$$\mathbb{P}_0 \left(T_{max} > t_{1-\alpha} \right) = \alpha$$

$t_{1-\alpha}$ IS $(1-\alpha)$ QUANTILE OF THE DISTRIBUTION OF THE MAXIMUM OF m INDEPENDENT $N(0,1)$

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

APPROXIMATION OF THE CRITICAL VALUE

$$\mathbb{P}_0 \left(T_{max} \geq z_{1-\frac{\alpha}{m}} \right) = \mathbb{P}_0 \left(\bigcup_{j=1}^{m} \left\{ T_j \geq z_{1-\frac{\alpha}{m}} \right\} \right)$$

$\left(1 - \frac{\alpha}{m}\right)$ QUANTILE OF $N(0,1)$

$$\underset{\text{BOOLE INEQ.}}{\leq} \sum_{j=1}^{m} \mathbb{P}_0 \left(\underbrace{T_j}_{N(0,1)} \geq z_{1-\frac{\alpha}{m}} \right)$$

$$= \cancel{m} \cdot \frac{\alpha}{\cancel{m}} = \alpha$$

$$\mathbb{P}_0 \left(T_{max} \geq z_{1-\frac{\alpha}{m}} \right) = 1 - \prod_{j=1}^{m} \mathbb{P}_0 \left(T_j < z_{1-\frac{\alpha}{m}} \right)$$

$$= 1 - \left(1 - \frac{\alpha}{m} \right)^m \xrightarrow[m \to \infty]{} 1 - e^{-\alpha}$$

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

MAGNITUDE OF THE CRITICAL VALUE $z_{1-\frac{\alpha}{m}}$

FOR LARGE m

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

$$\underset{m \to \infty}{\approx} \sqrt{2 \log m}$$ ← IT DOESN'T DEPEND ON α

NEEDLE IN A HAYSTACK PROBLEM

$H_0 : \mu_j = 0 \quad \forall j$

$H_1 : \mu_1 = C_m > 0, \quad \mu_2 = \dots = \mu_m = 0$

WHAT IS THE LIMITING POWER OF MAXT ?

$$\lim_{m \to \infty} \mathbb{P}_1 \left(T_{MAX} > z_{1-\frac{\alpha}{m}} \right) = ?$$

IT DEPENDS ON $\displaystyle \lim_{m \to \infty} \frac{C_m}{\sqrt{2 \log m}}$ → > 1 POWER 1 ; → < 1 NO POWER

Proof

① $\boxed{C_m > (1+\varepsilon)\sqrt{2 \log m}}$

$$\lim_{m \to \infty} \mathbb{P}_1 \left(T_{max} > z_{1-\frac{\alpha}{m}} \right) \geq \mathbb{P}_1 \left(T_1 > z_{1-\frac{\alpha}{m}} \right)$$

$$= \mathbb{P}_1 \left(N(C_m, 1) > \sqrt{2 \log m} \right)$$

$$= \mathbb{P}_1 \left(N(0,1) > \underbrace{\sqrt{2 \log m} - C_m}_{\to -\infty} \right)$$

$$= 1$$

② $\boxed{c_m < (1-\varepsilon)\sqrt{2\log m}}$

$$\lim_{m\to\infty} \mathbb{P}_1\left(T_{max} > z_{1-\frac{\alpha}{m}}\right) \underset{\text{BOOLE}}{\leq} \mathbb{P}_1\left(T_1 > z_{1-\frac{\alpha}{m}}\right) + \mathbb{P}_1\left(\max_{j>1} T_j > z_{1-\frac{\alpha}{m}}\right)$$

$$= \underbrace{\mathbb{P}_1\left(N(0,1) > \underbrace{\sqrt{2\log m} - c_m}_{\to\infty}\right)}_{0} + \left(1 - e^{-\alpha}\right)$$

$$\leq 1 - e^{-\alpha} \simeq \alpha$$

SUM T

$$T_{SUM} = \sum_{j=1}^{m} T_j \sim N\left(\sum_{j=1}^{m} \mu_j, m\right)$$

$$\frac{T_{SUM}}{\sqrt{m}} \overset{H_0}{\sim} N(0,1) \quad , \quad \frac{T_{SUM}}{\sqrt{m}} \overset{H_1}{\sim} N(\vartheta_m, 1)$$

WITH

$$\vartheta_m = \frac{\sum_{j=1}^{m} \mu_j}{\sqrt{m}} \quad \begin{cases} \xrightarrow{m\to\infty} 0 & \text{NO POWER} \\ \to \infty & \text{POWER } 1 \end{cases}$$

$H_0 : \mu_j = 0 \quad \forall j$

$H_1 : \mu_j = c_m > 0 \quad \forall j$

$$\vartheta_m = \sqrt{m}\, c_m \xrightarrow{m\to\infty} 0 \quad \text{IF} \quad c_m = \frac{1}{m}$$

$$\xrightarrow{m\to\infty} \infty \quad \text{IF} \quad c_m = 1$$

Min P

- $p_j = 1 - \Phi(T_j)$

- $p_1, \dots, p_m$ IID $U(0,1)$ UNDER H_0

- $p_{min} = \min(p_1, \dots, p_m)$

$$p_{min} \overset{H_0}{\sim} \text{Beta}(1, m)$$

- THE TEST REJECTS IF $p_{min} \leq \underline{1-(1-\alpha)^{\frac{1}{m}}}$

$$\mathbb{P}_0\left(p_{min} \leq 1-(1-\alpha)^{1/m}\right) = 1 - \mathbb{P}_0\left(\bigcap_{j=1}^{m}\left\{p_j > 1-(1-\alpha)^{\frac{1}{m}}\right\}\right)$$

$$= 1 - \left[(1-\alpha)^{1/m}\right]^m = \alpha \quad \square$$

$$p_{min} \leq \underset{\substack{\uparrow \\ \text{APPROXIMATED} \\ \text{CRITICAL} \\ \text{VALUE}}}{\frac{\alpha}{m}} < \underset{\substack{\uparrow \\ \text{CRITICAL} \\ \text{VALUE}}}{1-(1-\alpha)^{1/m}}$$

FISHER'S COMBINATION METHOD

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

SIMES' TEST

$$p_{(1)} \leq \dots \leq p_{(m)}$$

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

THE SIMES P-VALUE IS

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

REJECTS H_0 IF $\exists j : p_{(j)} \leq \frac{\alpha j}{m}$
