Modern Inference (Statistical Learning) - University of Milano-Bicocca

Homework 2

To submit via e-mail by 10/04/2020 h 14:00.
You will have to submit your solution by the deadline. Feel free to choose the format for your solution (.txt, .tex, .pdf etc.), the nicer the better, and the number of files in attachment (one or more, but not too many). Answer with clarity and precision. All R code must be reproducible. For theoretical questions, try to provide a well-reasoned mathematical argument. Simulations can help form your intuition; but a purely empirical answer will only receive partial points. It is encouraged to discuss the problem sets with others, but every group needs to turn in a unique write-up. Use of sources (people, books, internet and so on) without citing them in homework sets results in failing grade.

1 LDA in high-dimensions

Reference:
Wainwright (2019)
High-Dimensional Statistics: A Non-Asymptotic Viewpoint
Cambridge University Press
Chapter 1.2.1

Reproduce the 4 Figures displayed in the slides "high-dimensional statistics" (page 13 and page 18) by computing LDA empirical probability error (averaged over 10 repetitions) and comparing it to what is described by the classical and high-dimensional theories:

- $(m, n_A, n_B) = (400, 800, 800)$ with $\delta = 1/2$ and $\gamma = 1, 1.2, 1.4, 1.6, 1.8, 2$, by using the sample averages $\hat{\mu}_A$ and $\hat{\mu}_B$
- $m = 400$, $n_A = n_B = m/\delta$ with $\gamma = 2$ and $\delta = .05, .2, .5, .8, 1$, by using the sample averages $\hat{\mu}_A$ and $\hat{\mu}_B$
- $(m, n_A, n_B) = (400, 800, 800)$ with $\delta = 1/2$, $\gamma = 1, 1.2, 1.4, 1.6, 1.8, 2$ and sparsity $s = 5, 50$, by using the thresholded means $\tilde{\mu}_A$ and $\tilde{\mu}_B$ with $\lambda = \sqrt{\frac{2 \log m}{n}}$

The R code should be reproducible.

2 Covariance estimation in high-dimensions

Reference:
Wainwright (2019)
High-Dimensional Statistics: A Non-Asymptotic Viewpoint
Cambridge University Press
Chapter 1.2.2

Suppose $x_1, \ldots, x_n$ are i.i.d. $N_m(0, \Sigma)$. A natural estimator for Σ is the sample covariance matrix

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^{\mathsf{T}}$$

(you can also consider the usual estimator `cov()`)

A classical analysis considers the behavior of the sample covariance matrix Σ as the sample size n increases while the dimension m stays fixed. The sample covariance $\hat{\Sigma}$ is a consistent estimate of Σ in the classical setting. Is this type of consistency preserved if we also allow the dimension m to tend to infinity?

Using n samples $x_1, \ldots, x_n$ i.i.d. $N_m(0, I_m)$, obtain $\hat{\Sigma}$ and then compute its vector of eigenvalues $\lambda(\hat{\Sigma}) \in \mathbb{R}^m$ arranged in non-increasing order:

$$\lambda_1(\hat{\Sigma}) \geq \lambda_2(\hat{\Sigma}) \geq \ldots \geq \lambda_m(\hat{\Sigma})$$

If the sample covariance matrix $\hat{\Sigma}$ were converging to the identity matrix $\Sigma = I_m$, then the vector of eigenvalues should converge to the all-ones vector:

$$\lambda_1(I_m) = \lambda_2(I_m) = \ldots = \lambda_m(I_m) = 1$$

Perform a simulation with $(m, n) = (10, 2000)$ and $(m, n) = (1000, 2000)$: repeat the estimation of $\lambda(\hat{\Sigma})$ many times, and plot the histogram of the estimated vector of eigenvalues. Comments on the results.

3 Fisher's method of combining p-values

Fisher's combination its a global null test, but with combinations of p-values. It assumes independence of p-values. Under the global null H_0, $p_1, \ldots, p_m$ are i.i.d. $U(0, 1)$. The test statistic is

$$T_{\text{Fisher}} = \sum_{i=1}^{m} 2 \log\left(\frac{1}{p_i}\right)$$

Find the null distribution for this test statistic.

4 Comparison of global tests

Compare the power of the tests $T_{\max}$, T_{sum} and T_{Fisher} by simulations. Generate data from $y \sim N_m(\mu, I_m)$. Marginally, $y_i \sim N(\mu_i, 1)$ and $p_i = 1 - \Phi(y_i)$. Let $m = 10^6$ and $\alpha = 0.05$.

- Sparse strong effects: $\mu_i = 5$ for $1 \leq i \leq 4$ and 0 otherwise.
- Distributed weak effects: $\mu_i = 1.1$ for $1 \leq i \leq 2400$ and 0 otherwise.

The R code should be reproducible. Comment on the results.