Tutorial - Week 11

Please read the related material and attempt these questions before attending your allocated tutorial. Solutions are released on Friday 4pm.

Question 1

Let $\mathbb{x}_1, \ldots, \mathbb{x}_n$ be a sequence of independent random vectors from a p-dimensional normal distribution $N_p(\mu, \Sigma)$ with mean vector μ and $p \times p$ covariance matrix $\Sigma = (\sigma_{ij})$. The corresponding (population) correlation matrix $\mathbf{R}_n = (r_{ij})$ is defined by

$$r_{ii} = 1 \quad \text{and} \quad r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

for all $1 \leq i \neq j \leq p$. Given a random sample $\mathbb{x}_1, \ldots, \mathbb{x}_n$, we construct the $n \times p$ data matrix $\mathbb{X} = (x_{ij}) = (\mathbb{x}_1, \ldots, \mathbb{x}_n)'$ then the (Pearson) correlation coefficient betweeen $(x_{1i}, \ldots, x_{ni})'$ and $(x_{1j}, \ldots, x_{nj})'$ is given by

$$\hat{r}_{ij} = \frac{\sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ki} - \bar{x}_i)^2 \cdot \sum_{k=1}^n (x_{kj} - \bar{x}_j)^2}}$$

where $\bar{x}_i = \frac{1}{n}\sum_{k=1}^n x_{ki}$ and $\bar{x}_j = \frac{1}{n}\sum_{k=1}^n x_{kj}$. The *sample correlation matrix* is defined as $\hat{\mathbf{R}}_n := (\hat{r}_{ij})$. We often work with sample correlation matrices as they are invariant under scaling and shifting. Interesting results have been obtained about sample correlation matrices $\hat{\mathbf{R}}_n$ in the *high-dimensional regime* where $p, n \to \infty$ such that $p/n \to y < \infty$. Most of these results are in the case where the correlation matrix $\mathbf{R}_n = \mathbf{I}$, i.e., the p-dimensional identity matrix. In parts (a)-(e) below, assume the data vectors $\mathbb{x}_1, \ldots, \mathbb{x}_n$ are sampled from a multivariate Normal distribution with mean vector zero.

(a) Show that (in the case $\mathbf{R}_n = \mathbf{I}$) the limiting spectral distribution of the eigenvalues of $\hat{\mathbf{R}}_n$ is Marcenko-Pastur. Do this through a simulation whereby you plot a histogram of the spectral distribution for large n, p against the theoretical limiting density. Do this for the three cases $y_n = p/n \in \{0.25, 0.5, 0.75\}$.

(b) Show that (in the case $\mathbf{R}_n = \mathbf{I}$) the largest entry of $\hat{\mathbf{R}}_n$ given by

$$L_n = \max_{1 \leq i < j \leq p} |\hat{r}_{ij}|$$

satisfies

$$\lim_{n \to \infty} \sqrt{\frac{n}{\log n}} L_n = 2, \qquad \text{a.s.}$$

and the limiting cumulative distribution function is

$$P(nL_n^2 - 4\log n + \log(\log n) \leq z) \to e^{-Ke^{-z/2}}$$

as $n \to \infty$. What is the parameter K equal to? See **[A]**. Do this through a simulation whereby you plot a histogram of the largest entry for large n, p against the theoretical limiting cumulative distribution. Do this for the three cases $y_n = p/n \in \{0.25, 0.5, 0.75\}$.

(c) Show that (in the case $\mathbf{R}_n = \mathbf{I}$) the largest eigenvalue of $\hat{\mathbf{R}}_n$ satisfies a Tracy-Widom law; see **[B]**. Do this through a simulation whereby you plot a histogram of the largest eigenvalue for large n, p against the theoretical limiting density. Do this for the three cases $y_n = p/n \in \{0.25, 0.5, 0.75\}$.

(d) Show that the quantity $\log|\hat{\mathbf{R}}_n|$ satisfies a CLT result when $\mathbf{R}_n = \mathbf{I}$; See **[C]**. Do this through a simulation whereby you plot a histogram of the quantity for large n, p against the theoretical limiting density. Do this for the three cases $y_n = p/n \in \{0.25, 0.5, 0.75\}$.

(e) Show that the quantity $\log|\hat{\mathbf{R}}_n|$ satisfies a CLT result when $\mathbf{R}_n$ has a compound symmetry structure (all entries of $\mathbf{R}_n$ are equal to $\alpha \in [0, 1/2)$ except the diagonal which contains 1's) using the result of Corollary 1 of **[C]**. Do this through a simulation whereby you plot a histogram of the quantity for large n, p against the theoretical limiting density. Do this for the three cases $y_n = p/n \in \{0.25, 0.5, 0.75\}$ and $\alpha = 0.15$.

Question 2

Often real-world datasets exhibit data with heavier tails than the Gaussian distribution or that contain contaminations and this motivated researchers to introduce alternative versions of classic statistics to handle this situation. One of these is Kendall's tau that replaces the correlation between two random variables with something more robust against heavier tails, outliers, and contaminations. This idea can be extended to the multivariate setting as follows: Let $\mathbb{x}_1, \dots, \mathbb{x}_n$ be independent copies of a random vector $\mathbb{x} \in \mathbb{R}^p$ with coordinates $\mathbb{x} = (\mathbb{x}^{(1)}, \mathbb{x}^{(2)}, \dots, \mathbb{x}^{(p)})^T$. Kendall's tau matrix $\hat{\mathbf{T}}_n := (\tau_{k\ell})$ has entries given by

$$\tau_{k\ell} := \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \operatorname{sign}(\mathbb{x}_i^{(k)} - \mathbb{x}_j^{(k)}) \operatorname{sign}(\mathbb{x}_i^{(\ell)} - \mathbb{x}_j^{(\ell)}), \qquad 1 \le k, \ell \le p. \tag{1}$$

The matrix $\hat{\mathbf{T}}_n$ is a popular replacement for the sample correlation matrix $\hat{\mathbf{R}}_n$. This begs the question: how do the eigenvalues of $\hat{\mathbf{T}}_n$ behave when $n, p \to \infty$ such that $p/n \to y \in (0, 1)$? We consider this question by reproducing some recent results in the literature. In the following questions, you can use the function `cor.fk` from the `pcaPP` package for a fast implementation of (1).

(a) Show that the empirical spectral distribution of $\hat{\mathbf{T}}$ converges in probability to

$$\frac{2}{3}Y + \frac{1}{3},$$

where Y is distributed according to the standard Marchenko-Pastur disribution with parameter y; See **[D]**. Do this through a simulation whereby you plot a histogram of the spectral distribution of $\hat{\mathbf{T}}_n$ for large n, p against the theoretical limiting density. Do this for the three cases $y_n = p/n \in \{0.25, 0.5, 0.75\}$. Sample your data vectors $\mathbb{x}_1, \dots, \mathbb{x}_n$ from a multivariate Normal distribution with mean vector 0 and covariance $\Sigma = I_p$.

(b) Show that the distribution of the largest eigenvalue of $\hat{\mathbf{T}}_n$ follows a Tracy-Widom law; See **[F]** Corollary 1.4. Do this through a simulation whereby you plot a histogram of the largest eigenvalue of $\hat{\mathbf{T}}$ for large n, p against the theoretical limiting density. Do this for the three cases $y_n = p/n \in \{0.25, 0.5, 0.75\}$. Sample your data vectors $\mathbb{x}_1, \ldots, \mathbb{x}_n$ from a multivariate Normal distribution with mean vector 0 and covariance $\Sigma = I_p$.

(c) Show that $\operatorname{tr}(\hat{\mathbf{T}}_n^2)$ satisfies a CLT result; see **[E]** Theorem 3.1; Do this through a simulation whereby you plot a histogram of $\operatorname{tr}(\hat{\mathbf{T}}_n^2)$ for large n, p against the theoretical limiting density. Do this for the three cases $y_n = p/n \in \{0.25, 0.5, 0.75\}$. Sample your data vectors $\mathbb{x}_1, \ldots, \mathbb{x}_n$ from a multivariate Normal distribution with mean vector 0 and covariance $\Sigma = I_p$.

References

[A] Jiang (2004). The asymptotic distributions of the largest entries of sample correlation matrices. *Annals of Applied Probability*.

[B] Bao, Pan and Zhou (2012). Tracy-Widom law for the extreme eigenvalues of sample correlation matrices. *Electronic Journal of Probability*.

[C] Jiang (2019). Determinant of sample correlation matrix with application. *Annals of Probability*.

[D] Bandeira, Lodhia, Rigollet (2017). Marchenko-Pastur law for Kendall's tau. *Electronic Communications in Probability.*

[E] Li, Wang, Li (2021). Central Limit Theorem for Linear Spectral Statistics of Large Dimensional Kendall's Rank Correlation Matrices and its Applications. *Annals of Statistics.*

[F] Bao (2019). Tracy-Widom Limit for Kendall's Tau. *Annals of Statistics.*

[G] Leung, Drton (2018). Testing independence in high dimensions with sums of rank correlations. *Annals of Statistics.*