Tutorial - Week 3

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

Question 1

"*Every minute of every data, everywhere on the planet, dozens of companies — largely unregulated, little scrutinized — are logging the movements of tens of millions of people with mobile phones and storing the information in gigantic data files*", see **[A]**.

Start by reading the paper **[B]** about errors in GPS movement data. This paper is located in the 'Readings' folder on Wattle under the name *2016-RanacherBrunauerEtAl.pdf*.

(a) Consider the movement data in the file 'path.txt' that contains 100 positions over time and store it in the variable x of dimensions 100 x 2. Plot the path, and calculate the distance between the start of the path **P** and the end of the path **Q**.

(b) Perform a simulation study whereby you first assume your movement data x contains no measurement error, then add measurement noise ε (to each measurement) drawn from a bivariate normal with covariance $\Sigma = \sigma I$ where I is a 2×2 identity matrix and $\sigma > 0$. Vary σ and plot the distance $d(P, Q)$ as a function of σ. What can you conclude?

(c) Consider the p-dimensional case of Theorem 3.1, that is, assume $\mathbf{P} = 0 \in \mathbb{R}^p$ and $\mathbf{Q} = (d_0, 0, \ldots, 0) \in \mathbb{R}^p$. Reprove the result. What can you conclude?

(d) Consider distances in a p-dimensional space, take the starting point to be $P = (0, 0, 0, \ldots, 0)$ and end point to be $Q = (d_0, 0, \ldots, 0)$. Add noise to this path and consider how the mean length changes as p increases. Consider the simplified path $P^m + \varepsilon$ to $Q^m = Q + \eta$. Do this study for various choices of $p = 2, 10, 50, 100$. What can you conclude?

See **[C]**, for further real-world data.

Question 2

Illustrate numerically that the spectral density of large symmetric matrices formed from independent identically distributed random variables with zero mean and finite variance converges to the density of the Wigner Semicircle distribution. That is:

(a) Take $p = 100$ and write a function that generates a $p \times p$ symmetric matrix with entries sampled from the standard Normal distribution. Hint: generate a $p \times p$ matrix `A` with Normal entries and then symmetrise using `A[lower.tri(A)] <- t(A)[lower.tri(A)]`.

(b) Write a simulation that generates n of these matrices, calculates the eigenvalues of each of these matrices and plots the histogram of all these eigenvalues together (i.e., obtained from all the matrices).

(c) Plot a matching Wigner semicircle distribution over this histogram (you'll need to guess the appropriate parameters of the distribution).

(d) Repeat the experiment three times with p and n larger and larger with the ratio p/n

fixed. What do you observe?

(e) Now, repeat the experiment with the entries sampled from a Student-T distribution with parameter $1 < \nu \leq 2$. What do you observe and what can you conclude?

References

[A] https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html

[B] Ranachera, Brunauer, Trutschnig, Van der Spek, and Reich (2016).*Why GPS makes distances bigger than they are*. International Journal of Geographical Information Science. Vol 30, No 2, 316 – 333.

[C] https://archive.ics.uci.edu/ml/datasets/GPS+Trajectories