Homework 1

The total points on this group homework is 100.

Objective

The objective of this exercise is to perform a thorough simulation study in multivariate analysis and report the results in a way that can be verified. The key aspects involve writing wrapper R code to evaluate the different methods and to write a concise report on your evaluations.

Background

The assumption of normality is a cornerstone to statistical inference that justifies concepts from simple linear regression up to more advanced statistical techniques, such as linear discriminant analysis. While there are many good test of normality for univariate samples, it can be difficult and computationally intensive to determine normality in samples with high dimensionality.

There are a number of methods available for evaluating multivariate normality. Beyond the methods covered in class, there is also the BHEP method of Tenreiro (2017) whose R code is limited to work for up to 10 dimensions.

The Problem

Our task is to perform a reproducible comprehensive simulation study to compare the different tests in terms of their power. The objective is to study the power and how well these tests can distinguish departures from normality.

In general, such studies are commonly done by evaluating power in different scenarios where generating distributions are not very close to the normal. However, we will evaluate performance in distinguishing the multivariate normal distribution from the multivariate-t distribution which, like the multivariate normal distribution, has ellipsoidal contours and converges to the normal distribution as the degrees of freedom converges to infinity.

We will obtain realizations from ellipsoidally symmetric distributions with thicker tails than the multivariate normal distribution. Although there are many such options, we will use the multivariate t-distribution $t(\nu; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. We will generate $\boldsymbol{\Sigma}$ as before and set $\boldsymbol{\mu} = \mathbf{0}$. We will evaluate the power of the different tests as ν increases and as p increases. Specifically, we want to evaluate performance at each setting for $n \in \{100, 250\}$ and $p \in \{5, 10, 50, 100, 500\}$. Also, for $\nu \in \{3, 5, 10, 30, 100\}$. Once again, we will evaluate power based on 1,000 realizations each.

Observations from the multivariate t distribution can be obtained in R using the `rmvt` function in the package `mvtnorm` package.

For choosing $\boldsymbol{\Sigma}$ for each replication, I suggest using the in-built `rWishart(n, df, Sigma)` function in R, with the `df` set at p, `Sigma` set at $\boldsymbol{I}_p$. For evaluation with a singular matrix, simply premultiply with a lower rank matrix, and postmultiply by its transpose.

Setting Seed: For reproducibility, it is a good idea to set and store the seed. However, it is not a good idea to set the seed in a way that defeats the purpose of having a randomized experiment. For this reason, I suggest that your use R to generate a seed (use `1e8*runif(1)`, say) and store this value, and start from here.

What to turn in

1. The annotated R code. You can exclude the functions that I have provided, unless you find errors.
2. A short writeup with a brief 1-page description of the experiments and your findings and 2-pages of plots, tables and figures.