Tutorial - Week 12

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

Question 1

We are going to consider the topic of *factor analysis* where the aim is to describe the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called *factors*. Consider n daily returns from 1 Jan 2018 to 1 Jan 2019 for $p = 11$ stocks: BHP, RIO, ANZ, NAB, CBA, WBC, GXY, NUF, CGC, CGF, WSA. You can use the `Rmd` file I've provided in last week's workshop to download this data. Now implement the "Principal Component Method" (PCM) found in Section 9.3 of **[A]** using the correlation matrix $\mathbf{R}$ of the daily returns. That is:

(a) Determine the number of factors m using a screeplot and print out $\widetilde{\mathbf{L}}$ as a table showing stock names as row labels and factor numbers as column headers.

(b) Print out the communalities h_i^2, the estimated specific variances $\widetilde{\psi}_i$, and the proportion of total sample variance due to the jth factor. Use these values to argue that you've made the correct choice of m.

Now implement the "Maximum Likelihood Method" (see p.495 in **[A]**) using the correlation matrix $\mathbf{R}$ of the daily returns. For this part, you are allowed to use the inbuilt `facanal` command in `R` that inplements the ML method by default.

(c) First, perform the Maximum Likelihood Method (MLM) method *without rotation* using something like `fit = factanal(daily.returns, factors = m, rotation="none", covmat=R)` where `m` is the value of m you found previously and `R` is the correlation matrix. Print out the loadings with `print(fit)` and compare them to those found using the principal component method. What do you notice? Can you give a label to one (or more) of these factors?

(d) We are now going to perform some *factor rotations*, see Section 9.4 in **[A]**. Perform a varimax orthogonal rotation using `varimax(tilde.L)` where `tilde.L` was derived using PCM. Now extract the MLM estimated loading matrix $\widehat{L}$ from `fit` using `hat.L = loadings(fit)`. Now perform a varimax orthogonal rotation using `varimax(hat.L)` and an oblique rotation using `promax(hat.L)`. After rotation, do the loadings group the stocks in the same manner? Which rotation do you prefer? For your favourite rotation, can you give labels to the factors?

(e) Now perform a large sample test for the number of common factors by testing the hypothesis $H_0 : \Sigma = \mathbf{L}\mathbf{L}^T + \Psi$ with your choice of m at level $\alpha = 0.05$. The test is given in Eq. (9-39) of **[A]** and, since we are using the correlation matrix $\mathbf{R}$, it is based on the determinant of the matrix

$$\mathbf{R}^{-1}(\widehat{\mathbf{L}}\widehat{\mathbf{L}}^T + \widehat{\Psi}) \tag{1}$$

and uses a chi-square approximation to the sampling distribution. Implement this test in R. For your choice of m, do you accept or reject the null hypothesis?

(f) Considering the theory we've learnt this semester, comment on the form of (1) and the use of the chi-square approximation for the sampling distribution for the situation where the number of stocks p became large and $y_n := p/n = 0.5$. What might be a better alternative for this high-dimensional case?

Question 2

Suppose you have a dataset consisting of n vectors of the form $\mathbb{x} \in \mathbb{R}^p$ with coordinates $\mathbb{x} = (\mathbb{x}^{(1)}, \mathbb{x}^{(2)}, \ldots, \mathbb{x}^{(p)})^T$. You are tasked with determining if the coordinates $\mathbb{x}^{(i)}$ for $i = 1, \ldots, p$ are independent. That is, you want to accept or reject the hypothesis

$$H_0 : \mathbb{x}^{(1)}, \mathbb{x}^{(2)}, \ldots, \mathbb{x}^{(p)} \text{ independent},$$

versus the alternate (coordinates are not independent). You suspect that the data contain some contamination so you are inclined to use Kendall's tau $\hat{\mathbf{T}}_n$ from last week's tutorial for testing this hypothesis.

Let Z_α be the upper-α quantile of a standard normal distribution, we shall compare the following procedures for rejecting the null hypothesis H_0:

(P1) Reject H_0 if

$$Q_{\tau,2} - p - \frac{4p^2}{9n} > \frac{8p}{9n} Z_\alpha + \frac{14p^2}{9n^2} - \frac{4p}{9n},$$

where $Q_{\tau,2} = \operatorname{tr}(\hat{\mathbf{T}}_n^2)$; See **[F]** page 1576.

(P2) Reject H_0 if

$$Q_{\tau,\log} + \frac{b}{a}\sqrt{pn} - (p+n)\log(a) + (n-p)\log(a - b\sqrt{p/n}) < \hat{\sigma}_{\tau,\log} Z_{1-\alpha} + \hat{\mu}_{\tau,\log},$$

where $Q_{\tau,\log} = \log|\hat{T}_n|$; See **[F]** pages 1575–1576 for the definition of a, b, $\hat{\mu}_{\tau,\log}$ and $\hat{\sigma}_{\tau,\log}$.

When generating test data, follow the procedure: sample random vectors $\mathbb{x}_1, \ldots, \mathbb{x}_n \in \mathbb{R}^p$ from a multivariate Normal distribution with mean vector zero and pentadiagonal covariance matrix. That is, $\mathbb{x}_i \sim N_p(0, \Sigma_{\mathrm{b}})$ where $\Sigma_b = (\sigma_{ij})$ has diagonal entries $\sigma_{ii} = 1$ and entry $\sigma_{ij} = 0.1$ if $1 \le |i-j| \le 2$ and $\sigma_{ij} = 0$ if $|i-j| \ge 3$. Form the data matrix $\mathbb{X}$ from these samples and given your choice of (n, p) values, randomly select 5% of the $n \times p$ entries of the data matrix $\mathbb{X}$ to be contaminated. Each selected value to contaminate is replaced by an independent univariate sample from $N(2.5, 0.1)$ multiplied by a random sign.

(a) Perform a simulation experiment to evaluate procedures (P1) and (P2) with the test data you generated. That is, compare their empirical sizes for the following pairs of (p, n): $(150, 300)$, $(250, 500)$, $(200, 200)$, $(400, 400)$, $(300, 150)$, $(500, 250)$. Present this neatly in a table. What can you comment?

(b) (Bonus) Can you do the same for the empirical powers? Present this neatly in a table. What can you comment?

References

[A] Johnson, Wichern (2007). Applied Multivariate Statistical Analysis. Pearson Prentice Hall.

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

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

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

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

[F] 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.*

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

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