MATH 6020 – HW 3

Due back on Sunday, 11th April [1].

-1- Suppose $\mathrm{E}\{\mathrm{X}_{q\times 1}\} = \boldsymbol{\mu}$ and $\mathrm{VarCov}\{\mathrm{X}_{q\times 1}\} = \boldsymbol{\Sigma}_{q\times q}$. Let $\mathbf{C}_{q\times q}$ be a diagonal matrix of constants. Derive the $\mathrm{VarCov}\{\mathbf{C}(\mathrm{X}-\boldsymbol{\mu})\}$ under OFM.

-2- Suppose $\mathrm{E}\{\mathrm{X}_{q\times 1}\} = \boldsymbol{\mu}$ and $\mathrm{VarCov}\{\mathrm{X}_{q\times 1}\} = \boldsymbol{\Sigma}_{q\times q} = \mathrm{diag}(\sigma_{11}, \ldots, \sigma_{qq})$. Show that OFM reduces to $\boldsymbol{\Sigma} = \boldsymbol{\Psi}$ immaterial of the number of latent variables, m.

-3- Let $\mathbf{S}$ be a sample covariance matrix:

$$\mathbf{S} = 10^{-3}\begin{bmatrix} 11.072 & 8.019 & 8.160 \\ 8.019 & 6.417 & 6.005 \\ 8.160 & 6.005 & 6.773 \end{bmatrix}$$

Consider the Factor model ($m = 1$) based on $\mathrm{X}_{q\times 1} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose the maximum likelihood estimator of $\mathbf{L}$ is

$$\widehat{\mathbf{L}} = \begin{bmatrix} .1022 \\ .0752 \\ .0765 \end{bmatrix}.$$

Based on this MLE for $\mathbf{L}$, find the MLEs for

-a- the specific variances ψ_i

-b- the communalities $l_{i1}^2 + \ldots l_{im}^2$

-c- the proportion of variance explained by each factor

-4- In `R` use the commands `install.packages("HSAUR3")` followed by `data("USairpollution", package="HSAUR3")` to download air pollution information for 41 cities in the United States. You can find documentation on this data set at

`https://www.rdocumentation.org/packages/HSAUR3/versions/1.0-8/topics/USairpollution`

Find a principle component solution to a factor model with $m = 1$ and $m = 2$. What fraction of the variance does each one explain? Is there a rotation of the $m = 2$ model that gives an interpretation to the factors?

-5- Use the file `HW3_spca.csv` for this problem. Construct a sparse collection of PCs for this dataset and report results. State your choices of the parameters and the reasons for those choices.

-6- This question requires implementing linear discriminant analysis in `R`. For a good exposition of how to do this see

`https://rstudio-pubs-static.s3.amazonaws.com/35817_2552e05f1d4e4db8ba87b334101a43da.html`

[1] HW Version: 2021-04-10 at 02:18

There is also a file called `discriminant analysis.r` on Canvas that contains most of the relevant commands which you can modify to answer these questions.

-a- Install the `MASS` package in `R`, which has a built in function called `lda` for linear discriminant analysis. Type `?lda` to bring up the help file, and read it to understand how the `lda` function works.

-b- In `R` use the commands `install.packages("boot")` and then `data("urine", package="boot")` to download urine analysis data for 77 patients. You can find documentation on this data set at

`https://stat.ethz.ch/R-manual/R-devel/library/boot/html/urine.html`

Note that it consists of 79 rows and 7 columns. The first column is categorical (indicates if calcium oxalate crystals were found in the urine or not), while the remaining six are numerical. Note that rows 1 and 55 each have a missing entry in them (recorded as an NA). Delete these rows entirely.

-c- Use the `lda` function to compute the linear discriminant for the remaining data set of 47 points. What is the dominant feature in the discriminant?

-d- Use the `predict` function to compute the linear discriminants for the 77 data points. The output has a `class` function that tells you the predicted value of each data point (i.e. the prediction of whether it does or does not have crystals). What percentage of data points are predicted correctly?

-e- Use the `ldahist` function to plot a stacked histogram of the two linear discriminant values, grouped according to the presence or absence of crystals. Do the points seem to be well separated by this linear discriminant?