Homework 2 – Due 12 am CST, 27 February 2021

The total points on this homework is 100.

1. The 109th US Congress, comprising the Senate and the House of Representatives, was the legislative branch of the US government from January 3, 2005 to January 3, 2007. During this period, 542 bills were voted on by the US Senate. Each of 100 Senators either voted in favor or against or failed to record their vote on each of these bills. Details voting preferences of the 100 senators on these bills is provided in the MS Excel file `senate_voting_data.xls` available on the Datasets folder.

   (a) Read in the file: you may use `readxl` or `XLConnect` or some other package. (I usually use `readxl` but you should use whatever package you find most comfortable to work with.) Store this dataset into a file called `senators`. [*5 points*]

   (b) Note that the columns contain the senators' votes – the names header include the states that they represent and the party affiliation. You may use the following code to access the last name and party affiliation of each senator:

   ```
   senators.names <- names(senators)
   rev
   .party.state.names <- lapply(X = strsplit(gsub(pattern = "[.]",
      replacement = " ", x = senators.names),split = " "), FUN = rev)
   senators.party <- lapply(X = rev.party.state.names, FUN = function(x)(unlist(x)[1]))
   senators.last.names <- lapply(X = rev.party.state.names, FUN = function(x)(unlist(x)[7]))
   ```

   Using Andrews' curves, display the voting preferences of each senator. Use different colors for "Republican", "Democratic" and "Independent". Comment. [*10 points*]

2. The file `tornf1p.txt` contains the number of F1+ tornadoes per month in the US from 1954-2014. The columns are the Year and the 12 months from January through December.

   (a) Formulate a correlation plot and discuss the general relationships, if any, between the month-to-month frequency of the number of tornadoes in the US. Are there any patterns? How about the variances? [*10 points*]

   (b) Divide the period into the three 20 year periods (1955-1974, 1975-1994, 1995-2014). Call these periods I, II and III respectively.

      i. For each period, construct a parallel-coordinates plot (use color for the different periods). Superimpose the mean for each group. Are there any patterns that you can see? [*10 points*]

      ii. Next, display these observations using a survey plot. Are there any patterns that we can see? You may try out patterns with different months as the ordering variable and see which one provides the clearest separation. [*10 points*]

      iii. Display, using Chernoff faces and stars, the three group means. Are there any differences between these three means. [*5 points*]

3. Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Answer the following questions:

   (a) Show that the moment-generating function of $\boldsymbol{X}$ is given by

   $$M_{\boldsymbol{X}}(\boldsymbol{t}) \equiv E[\exp{(\boldsymbol{t}'\boldsymbol{X})}] = \exp{(\boldsymbol{t}'\boldsymbol{\mu} + \boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}/2)}.$$

   Alternatively, if you would rather derive the characteristic function, show that that is given by

   $$\phi_{\boldsymbol{X}}(\boldsymbol{t}) \equiv E[\exp{(i\boldsymbol{t}'\boldsymbol{X})}] = \exp{(i\boldsymbol{t}'\boldsymbol{\mu} - \boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}/2)}.$$

   [*10 points*]

(b) Partition the vectors $\boldsymbol{X}$ into q- and $(p-q)$-variate vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ and (correspondingly) $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}$ as follows:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}; \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}; \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix};$$

Show that the marginal distribution of $\boldsymbol{X}_1$ is $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$. [*5 points*]

(c) Show that the conditional distribution of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2$ is $N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$. [*15 points*]

4. Let $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Let $\boldsymbol{\Gamma}$ be an orthogonal matrix such that $\boldsymbol{\Gamma}\boldsymbol{\Gamma}' = \boldsymbol{I}_p$. Then show that $\boldsymbol{W} = \boldsymbol{\Gamma}\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. [*5 points*]

5. Using results from the spectral decomposition theorem for positive-definite matrices and multivariable calculus, show that the identity matrix $\boldsymbol{I}_p$ is the one among all $p \times p$-dimensional positive definite matrices $\boldsymbol{B}$ that maximizes

$$f(\boldsymbol{B}) = |n\boldsymbol{B}|^{n/2} \exp\left\{-\frac{n}{2}\text{trace}(\boldsymbol{B})\right\}.$$

(*Hint: Consider using the spectral decomposition of $\boldsymbol{B}$ and write $f(\boldsymbol{B})$ in terms of the eigenvalues of $\boldsymbol{B}$*). [*5 points*]

6. Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Using the method of derivatives of symmetric positive definite matrices and vectors, find the MLE of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$, given a sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. [*10 points*]