Homework 5 – Due 11:59 pm CST, 28 March 2021

The total points on this assignment is 100.

1. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ and $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m$ be samples from two independent p-variate populations. Let $\boldsymbol{U}_i^{(\boldsymbol{\lambda})} = \boldsymbol{\psi}(\boldsymbol{X}_i; \boldsymbol{\lambda})$ and $\boldsymbol{V}_i^{(\boldsymbol{\lambda})} = \boldsymbol{\psi}(\boldsymbol{Y}_i; \boldsymbol{\lambda})$, where $\boldsymbol{\psi}(\boldsymbol{w}; \boldsymbol{\lambda}) \equiv (\psi(w_1; \lambda_1), \psi(w_2; \lambda_2), \ldots, \psi(w_p; \lambda_p)$ is a function from $\mathbb{R}_+^p$ to $\mathbb{R}^p$ involving the parameter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)$ and defined through individual univariate real-valued functions as follows:

$$\psi(w_j; \lambda_j) = \begin{cases} \frac{w_j^{\lambda_j} - 1}{\lambda_j}, & \text{for } \lambda_j \neq 0 \\ \log w_j, & \text{for } \lambda_j = 0. \end{cases}$$

In other words, the univariate functions are the Box-Cox transformation with coordinate specific parameter λ_j.

Suppose that for some $\boldsymbol{\lambda}$ (which we do not know) we assume that $\boldsymbol{U}_1^{(\boldsymbol{\lambda})}, \boldsymbol{U}_2^{(\boldsymbol{\lambda})}, \ldots, \boldsymbol{U}_n^{(\boldsymbol{\lambda})} \sim N_p(\boldsymbol{\mu}_{(\boldsymbol{\lambda})}, \boldsymbol{\Sigma}_{(\boldsymbol{\lambda})})$ and that $\boldsymbol{V}_1^{(\boldsymbol{\lambda})} \boldsymbol{V}_2^{(\boldsymbol{\lambda})}, \ldots, \boldsymbol{V}_m^{(\boldsymbol{\lambda})} \sim N_p(\boldsymbol{\nu}_{(\boldsymbol{\lambda})}, \boldsymbol{\Sigma}_{(\boldsymbol{\lambda})})$.

   (a) For a given $\boldsymbol{\lambda}$, write down the maximized loglikelhood function given the observations $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ and $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m$. [*10 points*]

   (b) We will analyze the dataset on the attributes of the liberal arts colleges and public universities as provided in the `Colleges.txt` dataset.

      i. Because the variables in each of the two populations may not be multivariate normally distributed, we will find the $\boldsymbol{\lambda}$ which transforms the data such that they are so. For a grid of $\boldsymbol{\lambda}$ values, where each component λ_j takes values in $\{0, 1/4, 1/3, 1/2, 1, 2, 3, 4\}$, find the $\boldsymbol{\lambda}$ which maximizes the joint likelihood of $\boldsymbol{\lambda}$ (from among the grid) give the observations. [*10 points*]

      ii. With the transformed data, compare the mean values of the (transformed) SAT, % acceptance, cost per student, per cent of students in top 10 per cent of HS graduating class, per cent faculty with Ph.D.s and graduation rate, for the liberal arts vis-a-vis public universities? Are any of these means equal? [*10 points*]

      iii. Setting the False Discovery Rate at $q = 0.05$,which of the six variables have a significant difference between the liberal arts colleges and public universities. Interpret the results. *10 points*]

2. A researcher has collected data on three psychological variables, four academic variables (standardized test scores), and the type of educational program the student is in for 600 high school students. She is interested in how the set of psychological variables relate to the academic variables and gender. In particular, the researcher is interested in how many dimensions are necessary to understand the association between the two sets of variables.

   The SAS 7 data object `psych.sas7bdat` available at the usual place and which may be read in `R` using, for instance, the package `sas7bdat`. The dataset is of 600 observations on three psychological response variables, namely, locus of control, self-concept and motivation. The academic variables are standardized tests scores in reading, writing, and science, as well as a categorical variable giving the type of program the student is in (general, academic, or vocational). Note that the last variable is categorical, however, the SAS data object does not appear to read in these categorical variables as such (and should be converted).

   (a) Fit a linear model to the above and all the variables. Ignore interactions for now. Assume that the first level in the categorical variable has no additional effect (*i.e.* $\tau_1 = 0$) in the contrast. Summarize the results. [*10 points*]

   (b) Refit the model but after dropping the dependent variables on the test scores of writing and science. Summarize the results. [*8 points*]

(c) Is there a significant evidence that the writing and science test scores are related to the psychological profiles? [*2 points*]

(d) From the model in your results in (c) above, test simultaneously for whether there is a difference in psychological profiles between Program 1 and 2 and between Program 2 and 3. [*10 points*]

(e) Test the null hypothesis that the coefficient for the written test scores with locus of control as the outcome is equal to the corresponding coefficient with self concept as the outcome. [*10 points*]

(f) Now, test the null hypothesis that the coefficient for science scores for locus of control is equal to the corresponding coefficient for science for the self concept variable, and that the coefficient for the written scores for locus of control is equal to the coefficient for the written scores for self concept. [*10 points*]

(g) Depending on the results from (c), fit a linear model with all interactions included. Interpret the results. [*10 points*]