Assignment 3 (3 pages)

Statistics 32950-24620 (Spring 2024)

Due 9 am Tuesday, April 9th.

Requirements

- Your answers should be typed (allowing clear handwritten between typed texts for complex math formulas).
  Started with your name, Assignment 3, STAT 32950 or 24620 on the first page, saved as `LastnameFirstnamePset3.pdf`, and uploaded to Gradescope under either 329Pset3 or 246Pset3.
  Make sure to **submit to the correct course number** you registered, and tag the pages for each question.
- When you use R (or others) to solve problems, select only relevant parts of the output to insert in your writing.
- You may discuss approaches with others. However the assignment should be devised and written by yourself.

Reference: Chapters 5, 6, and 10 of the text by Johnson and Wichern.

Problem assignments:

1. (*Hands on Hotelling's T^2, small data set*)

   (a) The data consist of four observations $\boldsymbol{x}_j' = (x_{j1}, x_{j2})$: $(2,12), (8,9), (6,9), (8,10)$. Let $\boldsymbol{\mu}_o = \begin{bmatrix} 7 \\ 11 \end{bmatrix}$.
   In the following, leave your results in integer or fraction form (instead of approximated by decimals).

   i. Calculate the sample mean vector $\bar{\boldsymbol{x}}$.
   ii. Find sample covariance matrix $\boldsymbol{S}$.
   iii. Obtain $\boldsymbol{S}^{-1}$.
   iv. Evaluate Hotelling's T^2.
   v. Specify the distribution of T^2 under $H_o : \boldsymbol{\mu} = \mu_o$.

   (b) Let $C = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$. Transform the data $\{\boldsymbol{x}_j\}$ in (a) into $\boldsymbol{y}_j = C\boldsymbol{x}_j$. Let $\boldsymbol{\mu}_o^* = C\boldsymbol{\mu}_o = \begin{bmatrix} 18 \\ 4 \end{bmatrix}$.

   i. Calculate the sample mean vector $\bar{\boldsymbol{y}}$.
   ii. Derive the new sample covariance matrix $S_y = C\boldsymbol{S}C'$.
   iii. Evaluate Hotelling's T^2 for $\{\boldsymbol{y}_j\}$ under $H_o : \boldsymbol{\mu}_y = \boldsymbol{\mu}_o^*$.

   (c) Prove that, in general, if C is $p \times p$ invertible matrix, then the transformed data $\boldsymbol{y}_j = C\boldsymbol{x}_j$ has the same Hotelling's T^2 statistic (under $H_o : \boldsymbol{\mu}_y = C\boldsymbol{\mu}_o$).

2. (*Derivation in canonical correlation analysis*)

   The 2×1 random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ have joint covariance matrix $\boldsymbol{\Sigma}$,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \text{ with } \Sigma_{11} = \begin{bmatrix} 1 & p \\ p & 1 \end{bmatrix}, \Sigma_{22} = \begin{bmatrix} 1 & q \\ q & 1 \end{bmatrix}, \Sigma_{21} = \Sigma_{12} = \begin{bmatrix} r & r \\ r & r \end{bmatrix}.$$

   where $p, q, r \in (0, 1)$.

   (a) Derive ρ_1^*, the largest canonical correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$. Show your work.

   (b) Derive the canonical variate pairs $(U_1, V_1) = (\boldsymbol{a}_1'\boldsymbol{X}, \boldsymbol{b}_1'\boldsymbol{Y})$ corresponding to ρ_1^*, with normalization $\boldsymbol{a}_1'\Sigma_{11}\boldsymbol{a}_1 = 1, \boldsymbol{b}_1'\Sigma_{22}\boldsymbol{b}_1 = 1$.

3. (*Practice canonical correlation analysis, low dimensions*)

   Download the data from stiffness.DAT (also available next to the link of this p-set in Canvas).

   (`R command to input the data: stiff = read.table("stiffness.DAT")`)

   The data were obtained by taking four different measures of stiffness, x_1, x_2, x_3 and x_4 of each of $n = 30$ boards. The first measurement involves sending a shock wave down the board, the second measurement is determined while vibrating the board, and the last two measurements are obtained from static tests. The squared distances $d_j^2 = (\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\boldsymbol{S}^{-1}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})$ are also included as the last column in the data. (ref. Table 4.3 in J&W)

   Let $\boldsymbol{X} = [X_1, X_2]'$ be the vector of variables representing the dynamic measures of stiffness, and let $\boldsymbol{Y} = [X_3, X_4]'$ be the vector of variables representing the static measures of stiffness.

   (a) Perform a canonical correlation analysis of these data. (`R command: cancor(X,Y)`)

   (b) Write the first canonical variates U_1 and V_1 as linear combinations of the components of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively.

   (c) Produce two scatterplots of the data: one in the coordinate plane of the first canonical variate pair (U_1, V_1), one in the plane of the second pair (U_2, V_2).

   (d) Based on the two plots and the values of the canonical correlations $\{\rho_1^*, \rho_2^*\}$, comment on the correlation structure "captured" by each canonical pair.

4. (*Multivariate vs univariate inference*)

   Input the dataset fly.dat (automatic download when clicked, also available next to the link of this p-set in Canvas) of 15 observations on $X_1 =$ *antenna length* (mm) and $X_2 =$ *wing length* (mm) of two species of flies.

   ```
   R command for data input:  fly = read.table("fly.dat")
   ```

   Define two univariate variables $Y_1 =$ *antenna length* + *wing length*, $Y_2 =$ *wing length*.
   Treat the data as bivariate samples from two populations (species `Af` and `Apf`) with equal covariance.

   (a) (*Property of Hotelling's T^2*)

      i. Compute a Hotelling's T^2-statistic for the hypothesis of equality of the mean vectors in the two species based on (Y_1, Y_2). Is the hypothesis of equality of the means accepted?

      ii. Should you get the same results if you use the original variables (X_1, X_2)? Why?

   (b) If you conduct (univariate) two-sample t-tests at a test level $\alpha = 0.05$ performed on each of the variables Y_1 and Y_2 separately (i.e. assuming independence of Y_1 and Y_2), would the hypothesis of equality of species means be accepted? What if the test level $\alpha = 0.01$?

   (c) Draw a scatterplot of Y_1 vs. Y_2 for the data in both species groups, marking the data points of the two species groups with different symbols, and explain how it can happen that (a) and (b) have different conclusions.

   (d) (*Comparison: Confidence region vs simultaneous confidence intervals*)

      i. Draw a 98% confidence ellipse for the species mean differences of Y_1 and Y_2 based on Hotelling's T^2.

      ii. In the same graph, draw a rectangle corresponding to univariate (marginal) 99% confidence interval for the mean differences of Y_1 and Y_2.

      iii. Explain that the rectangle is a 98% confidence region by Bonferroni method.

      iv. Is the zero vector $\mathbf{0} = (0, 0)$ in any of the two regions (ellipse by i, rectangle by ii)?
      Compare and comment on the goodness of the two regions. Which one is better?

5. (*Derivations for multivariate data and random variables*)

(a) Let $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix}$ be an $n \times p$ data matrix of n observations, the jth observation is $\boldsymbol{x}_j \in \mathbb{R}^p$, $j = 1, \cdots, n$. The sample covariance matrix can be expressed as $\boldsymbol{S} = \frac{1}{n-1}\sum_{j=1}^n (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'$. Show that

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}'\boldsymbol{H}\boldsymbol{X}, \qquad \textit{with} \quad \boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$$

where $\boldsymbol{I}_n$ is the identity matrix of dimension $n \times n$, and $\mathbf{1}_n$ is the n-vector with all elements $= 1$.

(b) Let $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}$, where $\boldsymbol{Y}$ is a p-dimensional random vector, $\boldsymbol{A}$ isa fixed, scalar matrix of dimension $k \times p$, and $\boldsymbol{c} \in \mathbb{R}^k$ is a fixed vector. Show that $Cov(\boldsymbol{W}) = \boldsymbol{A}\,Cov(\boldsymbol{Y})\boldsymbol{A}'$.

(c) (**For 32950 only**. Optional for 24620.) Let $\boldsymbol{Y}$ be a p-dimensional random vector and $\boldsymbol{W}$ be a q-dimensional random vector, $\boldsymbol{a}, \boldsymbol{b}$ are fixed vectors with dimensions p and q respectively. Show that

$$Cov(\boldsymbol{a}'\boldsymbol{Y}, \boldsymbol{b}'\boldsymbol{W}) = \boldsymbol{a}'Cov(\boldsymbol{Y}, \boldsymbol{W})\boldsymbol{b} = \boldsymbol{b}'Cov(\boldsymbol{W}, \boldsymbol{Y})\boldsymbol{a}$$