Multivariate Analysis - Homework 5

Please upload your homework on SPOC before 8:00pm, May 4, including all details needed. For R exercises, R markdown is highly encouraged; for other parts, try to use LaTex.

1. Verify that for λ_i, $i = 1, \ldots, s$, the (ordered) eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$,
   (a) The Wilks' Lambda test statistic $\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i} = |\mathbf{E}|/|\mathbf{E}+\mathbf{H}|$.
   (b) The Pillai's statistic $V^{(s)} = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i} = tr\{(\mathbf{H}+\mathbf{E})^{-1}\mathbf{H}\}$.
   (c) The Lawley-Hotelling statistic $U^{(s)} = \sum_{i=1}^{s} \lambda_i = tr(\mathbf{E}^{-1}\mathbf{H})$.

2. For the multivariate regression model, using the same notation as the class slides, derive the distribution of the predicted vector $\hat{\mathbf{y}}_0$ of a new response $\mathbf{y}_0$, with the observed $\mathbf{x}_0$.

3. For the covariance matrix of $\mathbf{y} = (Y_1, Y_2)'$,

$$\mathbf{\Sigma} = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix},$$

   (a) Determine the population principal components $\mathbf{z} = (Z_1, Z_2)'$ by hand.
   (b) Compute the proportion of total population variance explained by the first principal component Z_1.
   (c) Suppose the original variables follows bivariate normal distribution with mean vector $(1, 2)'$. Sketch the constant density ellipse and indicate the principal components on your graph.
   (d) Convert the covariance matrix to a correlation matrix $\mathbf{P}$. Determine the principal components $\mathbf{v} = (V_1, V_2)'$ from $\mathbf{P}$, and compute the proportion of total population variance explained by the first principal component V_1.
   (e) Compare the components calculated in (d) with those obtained in (b). Are they the same? Should they be?
   (f) Find the correlation matrix $CORR(\mathbf{z}, \mathbf{y})$, $CORR(\mathbf{v}, \mathbf{y})$ and $CORR(\mathbf{z}, \mathbf{v})$, respectively. Comment.

4. Measurements of properties of pulp fibers and the paper made from them are contained in the following table. (Data attached as fiber.DAT). There are $n = 62$ observations of the pulp fiber characteristics, X_l = arithmetic fiber length, X_2 = long fiber fraction, X_3 = fine fiber fraction, X_4 = zero span tensile, and the paper properties, Y_l = breaking length, Y_2 = elastic modulus, Y_3 = stress at failure, Y_4 = burst strength.

y_1 BL	y_2 EM	y_3 SF	y_4 BS	x_1 AFL	x_2 LFF	x_3 FFF	x_4 ZST
21.312	7.039	5.326	.932	−.030	35.239	36.991	1.057
21.206	6.979	5.237	.871	.015	35.713	36.851	1.064
20.709	6.779	5.060	.742	.025	39.220	30.586	1.053
19.542	6.601	4.479	.513	.030	39.756	21.072	1.050
20.449	6.795	4.912	.577	−.070	32.991	36.570	1.049
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
16.441	6.315	2.997	−.400	−.605	2.845	84.554	1.008
16.294	6.572	3.017	−.478	−.694	1.515	81.988	.998
20.289	7.719	4.866	.239	−.559	2.054	8.786	1.081
17.163	7.086	3.396	−.236	−.415	3.018	5.855	1.033
20.289	7.437	4.859	.470	−.324	17.639	28.934	1.070

(a) Perform regression analysis using each of the 4 response variables $Y_1, \ldots, Y_4$ and all the four independent variables $X_1, \ldots, X_4$.

(i) Suggest and fit appropriate linear regression models.

(ii) Analyze the residuals. Check for outliers and observations with high leverages.

(iii) Perform the tests for the overall significance of independent variables. Give your conclusions.

(iv) Obtain the 95% confidence interval for each mean response and prediction interval for each individual response when $X_1 = 0.330$, $X_2 = 45.500$, $X_3 = 20.375$ and $X_4 = 1.010$ based on each regression.

(b) Perform a multivariate regression analysis using all 4 response variables $Y_1, \ldots, Y_4$ and all the four independent variables $X_1, \ldots, X_4$.

(i) Suggest and fit an appropriate linear regression model. Specify the matrix of estimated coefficients $\mathbf{B}$ and the estimated error covariance matrix $\mathbf{\Sigma}$.

(ii) Perform the multivariate tests for the overall significance of independent variables. Give your conclusions.

(iii) Perform a lack-of-fit test of X_3 and X_4. Conclude.

(iv) Obtain the 95% simultaneous confidence interval for each mean response and 95% simultaneous prediction interval for each individual response for the same settings of X's as (a)(iv).

5. (R exercise.) Consider the air-pollution data (attached in a separate .dat file.) Conduct principal component analysis of the data using both the covariance and correlation matrix. What have you learned? Give the detail of analysis, your conclusion remarks and the interpretations.