National Taiwan University
Institute of Industrial Engineering
Jakey BLUE

Data Analytics

111-2 Homework #05

Due at 23h59, April 2, 2023; files uploaded to NTU-COOL

1. Find the proper libraries/packages in your coding environment to perform the LASSO and Ridge regressions on the ORL face dataset (use the same gender labels created in your HW02).
   a. (10%) Select the lambda associated with the minimal MSE fit and compare the results with that of your stepwise regression in HW02.
   b. (5%) Plot the chosen pixels from LASSO regression on a 46×56 canvas.

2. The following table, provided by Dr. Philip Israelovich of the Federal Reserve Bank, gives the information on capital, labor, and value added of the economics of transportation equipment. (Ashish Sen, and Muni Srivastava, *Regression Analysis*)

Year	Capital	Labor	Value Added
72	1209188	1259142	11150.0
73	1330372	1371795	12853.6
74	1157371	1263084	10450.8
75	1070860	1118226	9318.3
76	1233475	1274345	12097.7
77	1355769	1369877	12844.8
78	1351667	1451595	13309.9
79	1326248	1328683	13402.3
80	1089545	1077207	8571.0
81	1111942	1056231	8739.7
82	988165	947502	8140.0
83	1069651	1057159	10958.4
84	1191677	1169442	10838.9
85	1246536	1195255	10030.5
86	1281262	1171664	10836.5

   a. (5%) Consider the model
   $$V_t = \alpha K_t^{\beta_1} L_t^{\beta_2} \eta_t \,,$$
   where the subscript t indicates the year, V_t is value added, K_t is capital, L_t is labor, and η_t is the error term, with $\mathrm{E}[\log(\eta_t)] = 0$ and $\mathrm{V}[\log(\eta_t)]$ a constant. Assuming the errors are independent across the years, estimate β_1 and β_2.
   b. (10%) The model in (a) is said to be of the Cobb-Douglas form. It is easier to interpret if $\beta_1 + \beta_2 = 1$. Estimate β_1 and β_2 under this constraint.

3. Implement a PCA function without using the available packages/libraries in R/Python. The input parameters of this function are the data matrix **X** and a Boolean flag "isCorrMX." The Boolean flag allows users to choose if the correlation matrix is used when set TRUE; otherwise, the covariance matrix would be decomposed. You can start with the function of Spectral Decomposition or Singular Value Decomposition.
   a. (15%) Necessary outputs are:
      - the loading matrix;
      - the eigenvalue value vector;
      - the score matrix, i.e., the matrix of principal components; and
      - the scree plot where eigenvalues are shown as bars and cumulative variance explained is drawn as a line (similar to the one on p. 36 of the slides DA04).
   b. (5%) Demonstrate your PCA function using the AutoMPG dataset. By comparing the results of "isCorrMX == TRUE" and "isCorrMX == FALSE", do you think PCA is scale-invariant?

   Note: directly applying any existed PCA libraries/packages in your function loses all the 20 points in this exercise.

4. Transpose the ORL face dataset to let **X** be a 2576 × 400 data matrix. Apply PCA to **X**, using the PCA function you created in EX3 above.
   a. (10%) How many principal components are needed to explain 50%, 60%, 70%, 80%, and 90% of the total variance?
   b. (10%) Rescale the first principal component (PC) into the range of [0, 255]. Reshape the first PC (initially an 2576 × 1 vector) into a 46 × 56 matrix. Plot an image from the 46 × 56 matrix using the rescaled PC scores as the grayscale values.