Assignment 2 Solutions

Ques1.

Yalefaces data set

The images are stored in a matrix called data matrix which is of dimensions 77760*6*15, where the element data_matrix(:,j,k) represents the pixel intensity vector of the image of the kth person in the jth orientation.
We can find the first principal components for each of the 15 persons and then construct the score matrix, producing a representative image. So we have to implement PCA for the data corresponding to each person (i.e a 77760*6 matrix for each person). The following steps are followed to implement PCA after obtaining the corresponding data matrix for each person.

a) The covariance matrix $S_x$is constructed after autoscaling the data.
b) The Eigen values and Eigen vectors of S_x are found out.
c) We take the Eigen vector corresponding to the largest Eigen value to obtain the principal component.
d) The representative image is obtained by multiplying the data matrix with the principal component.

Repeating the above process for all 15 persons, we would be able to obtain a score matrix of dimension 77760*15 where each column gives the representative image of the corresponding person.

Out of the 75 images 54 of them were classified correctly. (Accuracy of 72%)

Different people have obtained different answers due to number of images being selected for applying PCA and marks have been given considering possibilities. Also, different answers are obtained when auto-scaling is used.

Ques2.

(A)
OLS-
$\boldsymbol{y = 3.2827x_1 + 0.9375x_2 - 3.2725x_3 + 0.0639x_4}$

TLS-
$\boldsymbol{y = 21.0566x_1 + 3.4639x_2 - 23.4535x_3 - 0.1338x_4}$

The regression model in the original variables is given as:

$$\beta_{ols} = [0.0607 \quad 0.0059 \quad -0.1465 \quad 0.0080]; \quad \beta_{0,ols} = 11.8$$
$$\beta_{tls} = [\, 0.3895 \quad 0.0219 \quad -1.0501 \quad -0.0168]; \quad \beta_{0,tls} = 153.18$$

Coefficients correspond to CO2, CH4, N2O and Ozone respectively

An increase in concentration of any of these gases is expected to increase the global temperatures. However, the regression coefficient associated with ozone is small and even changes its sign from OLS to TLS model. A plot of annual ozone concentration with time shows that it does not increase monotonically (unlike other gases). Ozone is very reactive and therefore unlike CO2, its concentration in the atmosphere can also decrease. Thus, the atmospheric concentration may not correlate well with temperature. The negative coefficient for N_2O is an anomaly which may be difficult to account for. For the given problem, it is reasonable to consider CH_4 and CO_2 to explain the global warming trend as per the data.

(B) GWP of a gas is the amount of heat trapped by a gas as compared to unit weight of CO2. Assuming that temperature rise in atmosphere is a measure of heat trapped, we can estimate the GWP using the ratio of the regression coefficients in same units.

GWP $(CH_4)_{ols} = \frac{0.0059}{0.0607} \times 10^3 = 97.19$
GWP $(CH_4)_{tls} = \frac{0.0219}{0.3895} \times 10^3 = 56.2$

The GWP predicted using OLS model is closer to given value of 86 for CH4.

OLS model may be more reliable because three of the four coefficients are positive.

Let us use a cross validation approach to validate the performance of the two models. For that- divide the dataset into train and test data (I have used the first 24 samples for training and the rest are used as test samples- any other rational choice for the partition of data is acceptable). The prediction RMSE of temperature for the two models are 0.62 (OLS) and 0.64 (TLS). Here the OLS model seems to perform better than the TLS.

P.S
One can also scale the data to match the units and then use OLS to estimate the model. To match the units use 1ppm = 1000 ppb and 1 Dobson unit ~ 1.25 ppb. Similarly a TLS can be obtained as follows:

$A_{ols} = [0.0001 \;\; 0.0059 \;\; -0.1465 \;\; 0.0064]$
$A_{tls} = [0.0001 \;\; 0.0064 \;\; -0.1704 \;\; 0.0059]$

For the GWP computation, the estimates for CH_4 are evaluated as 59 and 64 for OLS and TLS respectively. Here the GWP (CH_4) seems to be predicted well by both models. TLS model is however marginally closer to the true value.
Take away: The choice of data pre-processing is to be made carefully considering possible error in variables. Auto-scaling can be poor choice in certain cases like in this example. The ideal scaling is the one where the scaling of the data is done according to its error variance. Since this info is not always available (The error variances can be estimated either from repeat measurements when they are available or by

IPCA under certain conditions) the **auto scaling** only acts a proxy for this info. The user is always advised to make a proper choice during modelling for good results.

Ques 3.

The question is expected to be worked out through hand calculation

The covariance matrix is given is of mean centred data. The model identified in this case corresponds to $A(z - \overline{z}) = 0$

(A)

Since one eigenvalue is given, others can be calculated by factorizing cubic polynomial into product of a quadratic and a linear polynomial with 250.4 as a root as,

$$Q(x) = (\lambda - \lambda_1)(a\lambda^2 + b\lambda + c) = 0$$

The eigenvalues are found by solving the quadratic equation. They are given as 0.08, 6.509 and 250.4

The normalized eigenvectors corresponding to these eigenvalues are,

$$v_1 = \begin{matrix} 0.9589 \\ -0.2830 \\ -0.0201 \end{matrix} \quad v_2 = \begin{matrix} 0.2330 \\ 0.8259 \\ -0.5135 \end{matrix} \quad v_3 = \begin{matrix} 0.1619 \\ 0.4877 \\ 0.8579 \end{matrix}$$

(B)

Data variance captured by first k eigenvalues are given by

$$\%var_k = \frac{\lambda_k}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \lambda_i}$$

For the given data, first eigenvalue alone contributes to 97.4 % variation, hence only one PC needs to be retained

(C)

Use eigenvector corresponding to the 2 lowest eigenvalue to obtain the linear relationship between the variables. The linear relations are given by,

$$v_1^T(Z - \overline{Z}) = 0 \quad v_2^T(Z - \overline{Z}) = 0$$

Or

$$0.9589z_1 - 0.2830z_2 - 0.0201z_3 + 13.2068 = 0 \qquad (1)$$

$$0.2330z_1 + 0.8259z_2 - 0.5135z_3 + 7.9833 = 0 \qquad (2)$$

(D)

Obtain scores by projecting the mean subtracted data sample to the largest eigenvector v_3. $t_{score} = v_3^T(z - \overline{z})$. Substituting z as $[10.1 \quad 73 \quad 135.5]$ and $\overline{z}$ as $[9 \quad 68 \quad 129]$

$$T_{score} = 8.19$$

(E)

Using the two equations (1) or (2) obtain a relation between mass (z_1) and SVL (z_2):

$$0.9496z_1 - 0.3153z_2 + 12.8943 = 0$$

For $z_2 = 73 \quad z_1 = 10.65$ g

(F)

Two linear equations are available to estimate mass and 2 measurements are available. The mass estimated needs to be consistent with both the equations given in part (c). The values given are measurements of HLS and SVL and the mass is to be estimated in TLS sense. For this, the variable for mass can be eliminated from the 2 equations to obtain the relationship between z_2 and z_3 as,

$$3.681z_2 - 2.093z_3 + 19.6462 = 0$$

To obtain reconciled estimates for z_2 and z_3, use TLS to minimize the objective function as,

$$Min_{\hat{z}} \, (z - \hat{z})^T(z - \hat{z}) \qquad (3)$$

$$s.t\ A\hat{z} = b$$

Where, z and $\hat{z}$ are the measurements and estimates respectively. A is given as $[3.61 \quad -2.093]$ and b = -19.6462

Analytical expression for the above form exists and is given as,

$$\hat{z} = z - A^T(AA^T)^{-1}(Az - b)$$

The above expression is obtained by minimizing equation (3) using lagrangian multipliers. The objective function is rewritten as an unconstrained optimization problem as,

$$Min\ \frac{1}{2}(z - \hat{z})^T(z - \hat{z}) + \lambda(A\hat{z} - b)$$

Differentiating with respect to $\hat{z}$ and λ (First order necessary condition)

$\hat{z} = z - A^T\lambda$ (4)

$A\hat{z} = b$ (5)

Substituting equation (2) in (3),

$\lambda = (AA^T)^{-1}(Az - b)$ (6)

$$\hat{z} = z - A^T(AA^T)^{-1}(Az - b)$$

Using the above relation,

$$\hat{z} = (75.9779 \quad 140.4330)$$

This estimate is consistent with both the equations simultaneously and gives the reconciled estimates of mass in addition to refining the given measurements. The redundant information available (2 variables) is efficiently used.

The mass is estimated from equation (1) or (3) as 11.117g