SOLUTIONS – ASSIGNMENT 2

1.

(a) The regression model is obtained as:

$B_{ols} = [0.0607 \quad 0.0059 \quad -0.1465 \quad 0.0080]$; b0 = 11.7998

An increase in concentration of any of these gases is expected to increase the global temperatures, which is indicated by the OLS model for CO_2 CH_4 and ozone. The negative coefficient for N_2O is an anomaly.

(b) Estimated error variance in temperature = 0.0166 (deg C)^2
95% Confidence intervals for coefficients are [0.0463, 0.0752]; [0.0025, 0.0094]; [-0.1907, -0.1024]; [-0115, 0.0275];

All the standardized residuals are between the limits [-2, 2]. Therefore, there are no outliers.

(c) The coefficient for ozone is insignificant and can be dropped. The regression coefficients are $B_{ols} = [0.0652 \quad 0.0058 \quad -0.1567]$ b0 = 0.8512

(d) Since unit of CO2 is in ppm and units of Methane is in ppb, the GWP of CH4 in terms of CO2 is 0.0058*1000/0.0652 = 89 (which is close to the literature value). For N2O we cannot estimate since coefficient is negative.

2.

(a) Estimated value of A' = 4.7607 and B' = 37.896 (using y = ln Psat and x = 1/T). Maximum error between estimated and measured pressure = 39.82 kPa

(b) Estimated values of A = 14.1018, B = 2821.4 and C = 228.7553, where sum of squared error between measured and estimated Psat, and not sum squared difference between ln(Psat) and ln(Psatest). Maximum error between estimated and measured pressure = 4.48 kPa

(c) This is a difficult optimization problem since we have to solve for the parameters A, B, C as well as the estimates of both temperature and pressure for all samples. We can replace the estimates of pressure in terms of temperature estimates using the model equation (similar to what is done is TLS for two variables), but this implies we have to solve for N+3 decision variables where N is the number of samples. It is also a nonlinear optimization problem. In order to solve this we can consider an outer optimization loop where the parameters A, B, and C are the decision variables while the inner optimization problem solves the WTLS problem for estimating the temperature and pressure for every sample for the given guess of the parameters. Note that if A, B, and C are given then the problem of estimating temperature and pressure for every sample gets decoupled and can be solved independently (each of which is a one variable nonlinear unconstrained problem). This was coded in MATLAB and the solution obtained for A = 14.1271; B = 2835.2 and C = 229.414

Maximum error between predicted and measured pressure is 4.2112 kPa

(d) Maximum absolute difference predicted and estimated pressures for all three approaches are given in respective parts

(3) The question is expected to be solved manually

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,

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

b) Data variance captured by first k eigenvalues are given by

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

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 smallest eigenvalues to obtain the linear relationships between the variables. The linear relations are given by,

$$v_1^T\left(Z - \overline{Z}\right) = 0$$

$$v_2^T\left(Z - \overline{Z}\right) = 0$$

Or

$$0.9589z_1 - 0.2830z_2 - 0.0201z_3 + 13.2067 = 0 \qquad (1)$$

$$0.2330z_1 + 0.8259z_2 - 0.5135z_3 + 7.9800 = 0 \qquad (2)$$

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) and (2) estimate the values of mass (z_1) and HLS (z_3), using the measured value of z_3. We get the following estimates

$z_1 = 10.65$ 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,

$$Min_{\hat{z}}\ (z - \hat{z})^T(z - \hat{z}) \tag{3}$$
$$s.t\ A\hat{z} = b$$

Where, z and $\hat{z}$ are the measurements and estimates respectively. A is given as $[3.681 \quad -2.093]$ 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)

$$\hat{z} = z - A^T\lambda \tag{4}$$
$$A\hat{z} = b \tag{5}$$

Substituting equation (2) in (3),

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

Using the above relation,

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

Substituting the above estimates in either of the original two equations will give the estimate for mass as 11.117g

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