2 **(PCA)** As discussed in class, show that PCA can be viewed as minimizing the approximation error; that is, the optimization problem (following the notation in our lecture note)

Suppose $\bar{X}=0$

$$\min_{\mu,X,U:U^\top U=I_p} \sum_{i=1}^{N} \|Y_i - \mu - UX_i\|^2$$

has solution $\mu = \bar{Y}$, $U = [U_1, \ldots, U_p]$ being the top p eigenvectors of the sample covariance of Y, and $X_i = U^\top(Y_i - \bar{Y})$.

Denote $L = \sum_{i=1}^{N} \|Y_i - \mu - UX_i\|^2$

① $\frac{\partial L}{\partial \mu} = \sum_{i=1}^{N} -2(Y_i - \mu - UX_i) = 0 \Rightarrow \sum_{i=1}^{N} Y_i - N\mu - U\sum_{i=1}^{N} X_i = 0 \Rightarrow \mu = \frac{1}{N}\sum_{i=1}^{N} Y_i - U(\frac{1}{N}\sum_{i=1}^{N} X_i)$

$\Rightarrow \hat{\mu} = \bar{Y} \Rightarrow L = \sum_{i=1}^{N} \|Y_i - \bar{Y} - UX_i\|^2$, denote $Y_i - \bar{Y} = \overline{Y_i}$ $\Rightarrow L = \sum_{i=1}^{N} \|\overline{Y_i} - UX_i\|^2$

② $\|\overline{Y_i} - UX_i\|^2 = (\overline{Y_i} - UX_i)^T(\overline{Y_i} - UX_i) = \overline{Y_i}^T\overline{Y_i} - 2X_i^T U^T \overline{Y_i} + X_i^T U^T U X_i = \overline{Y_i}^T \overline{Y_i} - 2X_i^T U^T \overline{Y_i} + X_i^T X_i$

$\frac{\partial L}{\partial X_i} = -2U^T\overline{Y_i} + 2X_i = 0 \Rightarrow \hat{X_i} = U^T\overline{Y_i} = U^T(Y_i - \bar{Y})$

③ $\|\overline{Y_i} - UX_i\|^2 = \overline{Y_i}^T \overline{Y_i} - 2X_i^T U^T \overline{Y_i} + X_i^T X_i = \overline{Y_i}^T \overline{Y_i} - 2X_i^T X_i + X_i^T X_i = \overline{Y_i}^T \overline{Y_i} - X_i^T X_i = \|\overline{Y_i}\|^2 - \|X_i\|^2$

$\min L = \max \sum_{i=1}^{N} \|X_i\|^2 = \max \sum_{i=1}^{N} X_i^T X_i$

$X_i = U^T\overline{Y_i} \Rightarrow \sum_{i=1}^{N} X_i^T X_i = \sum_{i=1}^{N} \overline{Y_i}^T U U^T \overline{Y_i} = \sum_{i=1}^{N} tr(\overline{Y_i}^T U U^T \overline{Y_i}) = tr(U^T(\sum_{i=1}^{N} \overline{Y_i}\,\overline{Y_i}^T)U) = tr(U^T S U)$

here $S = \sum_{i=1}^{N} (Y_i - \bar{Y})(Y_i - \bar{Y})^T$

$\Rightarrow \min L = \max tr(U^T S U)$, $U^T U = I_p$

according to Lagrange multiplier: $\mathcal{L}(U, \Lambda) = tr(U^T S U) - tr(\Lambda(U^T U - I))$

here $\frac{\partial \mathcal{L}}{\partial U} = 0 \Rightarrow 2SU - 2U\Lambda = 0 \Rightarrow SU = U\Lambda \Rightarrow \Lambda = U^T S U$, for decorrelation, Λ should be diagonal

$\Rightarrow SU = \Lambda U \Rightarrow \Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix}$, λ_i is eigenvalue of S, U is stacked eigenvector of S

To minimize L, which is to maximize $tr(U^T S U)$, the eigenvectors should be the top p vectors