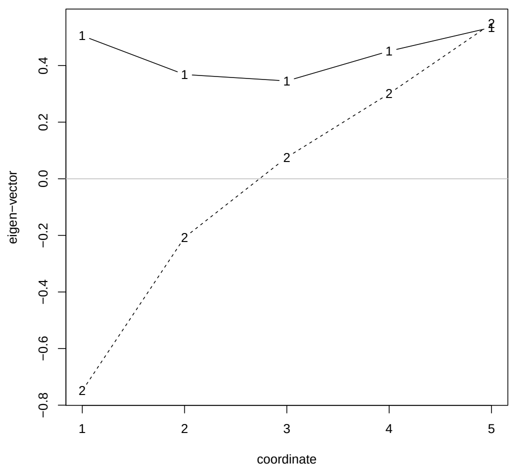
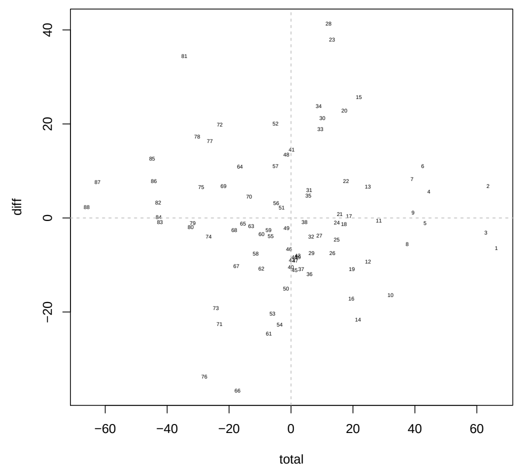
ISYE 7405

Homework 5

Jacob Aguirre *Email: aguirre@gatech.edu*
Instructor: Dr. Shihao Yang

1. Repeat the student score data PCA calculations and reproduce the following figures that we saw in class.





Solution. R code is shown below.

```
data_hw3   = read.table("scoredata.txt", header = FALSE)
data_hw3   = as.matrix(data_hw3)
data_hw3   = scale(data_hw3, center = TRUE, scale = FALSE)
S = cov(data_hw3)
eigen.S = -eigen(S, symmetric = TRUE)$vectors
eigen.S[, 1:2]

pdf("fg1_hw3.pdf", height = 6, width = 6)
plot(eigen.S[,2], type = "b", lty = 2, pch = "2",
     xlab = "coordinate", ylab = "eigen-vector")
lines(eigen.S[,1], type = "b", pch = "1")
abline(h=0, col = "grey")
dev.off()

pdf("fg2_hw3.pdf", height = 6, width = 6)
total = data_hw3%*%eigen.S[,1]
diff  = data_hw3%*%eigen.S[,2]
plot(diff~total, type = "n")
text(total, diff, label = 1:dim(data_hw3)[1], cex = 0.4)
abline(h = 0, col = "grey", lty = 2)
abline(v = 0, col = "grey", lty = 2)
dev.off()
```

□

2. Let $X_{p\times n}$ be a data matrix. Assume that X has row means 0. Let $Y_{(j)} = L_{(j)}^T X$ (recall we introduced $L_{(j)}$ through the SVD of X).

1. Calculate the $(p+j)\times(p+j)$ matrix $\begin{pmatrix} X \\ Y_{(j)} \end{pmatrix} \begin{pmatrix} X^T & Y_{(j)}^T \end{pmatrix}$

   Solution. We have

$$\begin{pmatrix} X \\ Y_{(j)} \end{pmatrix} \begin{pmatrix} X^T & Y_{(j)}^T \end{pmatrix} = \begin{pmatrix} XX^\top & XY_{(j)}^\top \\ Y_{(j)}X^\top & Y_{(j)}Y_{(j)}^\top \end{pmatrix} = \begin{pmatrix} XX^\top & XX^\top L_{(j)} \\ L_{(j)}^\top XX^\top & L_{(j)}^\top XX^\top L_{(j)} \end{pmatrix}.$$

   Since

$$XX^\top L_{(j)} = LC^2L^\top L_{(j)} = LC^2 \begin{pmatrix} I_j \\ 0 \end{pmatrix} = L \begin{pmatrix} C_{(j)}^2 \\ 0 \end{pmatrix} = L_{(j)}C_{(j)}^2,$$

   we have

$$\begin{pmatrix} X \\ Y_{(j)} \end{pmatrix} \begin{pmatrix} X^T & Y_{(j)}^T \end{pmatrix} = \begin{pmatrix} LC^2L^\top & L_{(j)}C_{(j)}^2 \\ C_{(j)}^2 L_{(j)}^\top & C_{(j)}^2 \end{pmatrix}.\square$$

2. Calculate $\widehat{X}$, the projection of X row by row into $L_{row}(Y_{(j)})$

   Solution. The projection of X row by tow into $L_{row}(Y_{(j)})$ is

$$\begin{aligned} \widehat{X} &= XY_{(j)}^\top (Y_{(j)}Y_{(j)}^\top)^{-1}Y_{(j)} = XX^\top L_{(j)}(L_{(j)}^\top XX^\top L_{(j)})^{-1}L_{(j)}^\top X \\ &= XX^\top L_{(j)}(L_{(j)}^\top LC^2L^\top L_{(j)})^{-1}L_{(j)}^\top X \\ &= LC^2L^\top L_{(j)}C_{(j)}^{-2}L_{(j)}^\top LCR^\top \\ &= LC^2 \begin{pmatrix} I_j \\ 0 \end{pmatrix} C_{(j)}^{-2} \begin{pmatrix} I_j & 0 \end{pmatrix} CR^\top \\ &= L \begin{pmatrix} C_{(j)} & 0 \\ 0 & 0 \end{pmatrix} R^\top \\ &= L_{(j)}C_{(j)}R_{(j)}^\top \\ &= \sum_{k=1}^{j} c_k l_l \gamma_k^\top .\square \end{aligned}$$

3. Calculate $X^\perp (X^\perp)^T$, where $X^\perp = X - \widehat{X}$

   Solution. According to SVD of X:

$$X = LCR^\top = \sum_{k=1}^{r} c_k l_l \gamma_k^\top,$$

   we have

$$X^\perp = X - \widehat{X} = \sum_{k=j+1}^{r} c_k l_k \gamma_k^\top.$$

   Therefore, we have

$$X^\perp (X^\perp)^\top = \left(\sum_{k=j+1}^{r} c_k l_k \gamma_k^\top \right) \left(\sum_{k=j+1}^{r} c_k \gamma_k l_k^\top \right) = \sum_{k=j+1}^{r} c_k^2 l_k \gamma_k^\top \gamma_k l_k^\top = \sum_{k=j+1}^{r} c_k^2 l_k l_k^\top .\square$$

3. (Prove Theorem A that we discussed in class.) Suppose $X \sim [0, \Sigma]$, $\Sigma = \Gamma\Lambda\Gamma^\top$ with all $\lambda_i > 0$. Let $\Gamma_{(j)} = (\gamma_1, \gamma_2, \ldots, \gamma_j)$. Then

1. The best linear predictor of X in terms of $\Gamma_{(j)}$ is the projection of X onto the column space of $\Gamma_{(j)}$:

$$\widehat{X} = \Gamma_{(j)}\Gamma_{(j)}^\top X = \sum_{i=1}^{j} y_i\gamma_i,$$

where $Y_{(j)} = \Gamma_{(j)}^\top X$.

Solution. The best linear predictor of X in terms of $\Gamma_{(j)}$ is the projection of X onto the column space of $\Gamma_{(j)}$:

$$\widehat{X} = \Gamma_{(j)}(\Gamma_{(j)}^\top\Gamma_{(j)})^{-1}\Gamma_{(j)}^\top X = \Gamma_{(j)}\Gamma_{(j)}^\top X = \Gamma_{(j)}Y_{(j)} = (\gamma_1, \ldots, \gamma_j)\begin{pmatrix} y_1 \\ \vdots \\ y_j \end{pmatrix} = \sum_{i=1}^{j} y_i\gamma_i.\square$$

2. The residual $X^\perp = X - \widehat{X}$ has covariance matrix

$$\Sigma_{(j)}^\perp = \sum_{i=j+1}^{p} \lambda_i\gamma_i\gamma_i^\top$$

with $\mathrm{tr}\Sigma_{(j)}^\perp = \sum_{i=j+1}^{p} \lambda_i$.

Solution. Assume $\Gamma = (\Gamma_{(j)}, \Gamma_{(-j)})$ is the orthogonal matrix in the spectral decomposition of Σ. The residual

$$X^\perp = X - \widehat{X} = \sum_{i=1}^{p} y_i\gamma_i - \sum_{i=1}^{j} y_i\gamma_i = \sum_{i=j+1}^{p} y_i\gamma_i = \Gamma_{(-j)}Y_{(-j)} = \Gamma_{(-j)}\Gamma_{(-j)}^\top X$$

has covariance matrix

$$\Sigma_{(j)}^\perp = \Gamma_{(-j)}\Gamma_{(-j)}^\top\Sigma\Gamma_{(-j)}\Gamma_{(-j)}^\top = \Gamma_{(-j)}\Gamma_{(-j)}^\top\Gamma\Lambda\Gamma^\top\Gamma_{(-j)}\Gamma_{(-j)}^\top = \Gamma_{(-j)}\begin{pmatrix} 0 & 0 \\ 0 & \Lambda_{(j)} \end{pmatrix}\Gamma_{(-j)}^\top = \sum_{i=j+1}^{p} \lambda_i\gamma_i\gamma_i^\top,$$

with

$$\mathrm{tr}\ \Sigma_{(j)}^\perp = \mathrm{tr}\sum_{i=j+1}^{p} \lambda_i\gamma_i\gamma_i^\top = \sum_{i=j+1}^{p} \lambda_i \mathrm{tr}(\gamma_i^\top\gamma_i) = \sum_{i=j+1}^{p} \lambda_i.\square$$

3. For any matrix $A_{j\times p}$, let $z = AX$ and $X_z^\perp = X - \Sigma_{Xz}\Sigma_{zz}^{-1}z$. Show that

$$\mathrm{tr}\Sigma_{(j)}^\perp = \mathrm{tr}\ \mathrm{cov}(X^\perp) \geq \sum_{i=j+1}^{p} \lambda_i.$$

Solution. Since

$$\Sigma_{Xz} = E(XZ^\top) = E(XX^\top A^\top) = \Sigma A^\top, \quad \Sigma_{zz} = E(ZZ^\top) = E(AXX^\top A^\top) = A\Sigma A^\top,$$

the covariance of the residual is

$$\Sigma_{(j)}^\perp = \mathrm{cov}(X - \Sigma_{Xz}\Sigma_{zz}^{-1}z) = \Sigma - \Sigma_{Xz}\Sigma_{zz}^{-1}\Sigma_{zX} = \Sigma - \Sigma A^\top(A\Sigma A^\top)^{-1}A\Sigma.$$

In order to show that

$$\mathrm{tr}\Sigma_{(j)}^\perp = \mathrm{tr}\ \mathrm{cov}(X^\perp) \geq \sum_{i=j+1}^{p} \lambda_i,$$

we only need to show that

$$\sum_{i=1}^{j} \lambda_i \geq \mathrm{tr}\{\Sigma A^\top (A\Sigma A^\top)^{-1} A\Sigma\} = \mathrm{tr}\{\Gamma\Lambda\Gamma^\top A^\top (A\Gamma\Lambda\Gamma^\top A^\top)^{-1} A\Gamma\Lambda\Gamma^\top\} = \mathrm{tr}\{(A\Gamma\Lambda\Gamma^\top A^\top)^{-1} A\Gamma\Lambda^2\Gamma^\top A^\top\}.$$

Define $C = A\Gamma\Lambda^{1/2}$, and the above inequality reduces to

$$\sum_{i=1}^{j} \lambda_i \geq \mathrm{tr}\{(CC^\top)^{-1}(C\Lambda C^\top)\} = \mathrm{tr}\{C^\top (CC^\top)^{-1} C\Lambda\} = \mathrm{tr}(P_C\Lambda),$$

where $P_C = C^\top (CC^\top)^{-1} C$ is a projection matrix of rank j. The projection matrix has spectral decomposition $P_C = \sum_{i=1}^{j} \delta_i \delta_i^\top$, where δ_i's are unit vectors that are orthogonal. Therefore, the above inequality further reduces to

$$\sum_{i=1}^{j} \lambda_i \geq \mathrm{tr}\left(\sum_{i=1}^{j} \delta_i\delta_i^\top \Lambda\right) = \sum_{i=1}^{j} \delta_i^\top \Lambda \delta_i.$$

Let $\Delta_{p\times p} = (\Delta_1, \Delta_2)^T = (\delta_1, \ldots, \delta_j, \delta_{j+1}, \ldots, \delta_p)^T = (\delta_{ij})$ orthogonal matrix. (Adding $p-j$ orthogonal row vectors to complement $\delta_1, \ldots, \delta_j$ to form orthogonal basis). Then.

$$\sum_{i=1}^{j} \delta_i^\top \Lambda \delta_i = \sum_{k=1}^{p} (\lambda_k \sum_{i=1}^{j} \delta_{ik}^2)$$

Also,

$$0 \leq \sum_{i=1}^{j} \delta_{ik}^2 \leq 1, \sum_{k=1}^{p}\sum_{i=1}^{j} \delta_{ik}^2 = \sum_{i=1}^{j} ||\delta_i||^2 = j.$$

So the maximum is taken when $\sum_{i=1}^{j} \delta_{ik}^2 = 1$, for $k \leq j$, equivalently, maximum is $\sum_{i=1}^{j} \lambda_i$

According to the fundamental lemma, $\delta^\top \Lambda \delta$ is maximize at e_1 with value λ_1; among the unit vectors orthogonal to e_1, $\delta^\top \Lambda \delta$ is maximize at e_2 with value λ_2; and so on. Consequently, the right hand side has maximum value $\sum_{i=1}^{j} \lambda_i$, corresponding to $(\delta_1, \ldots, \delta_j) = (e_1, \ldots, e_j)$. The conclusion follows.□

4. (Ridge regression) Hoerl and Kennard (1970) have proposed the method of ridge regression to improve the accuracy of the parameter estimates in the regression model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \mu\mathbf{1} + \boldsymbol{u}, \quad \boldsymbol{u} \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{I}).$$

Suppose the columns of $\boldsymbol{X}$ have been standardized to have mean 0 and variance 1. The ridge estimate of $\boldsymbol{\beta}$ is defined by

$$\boldsymbol{\beta}^* = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

where for given $\boldsymbol{X}$, $k \geq 0$ is a small fixed number.

1. Show that $\boldsymbol{\beta}^*$ reduces to the OLS estimate $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ when $k = 0$.

   Solution. When $k = 0$, we have $\boldsymbol{\beta}^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, and we need to show that this is the OLS estimator for $\boldsymbol{\beta}$.

   Since $\boldsymbol{X}$ is standardized to have mean 0 and variance 1, we have $\mathbf{1}^\top\boldsymbol{X} = 0$. Therefore, the OLS estimator for $(\mu, \boldsymbol{\beta}^\top)^\top$ is

$$\begin{pmatrix}\widehat{\mu}\\ \widehat{\boldsymbol{\beta}}\end{pmatrix} = \left\{\begin{pmatrix}\mathbf{1}^\top\\ \boldsymbol{X}^\top\end{pmatrix}\begin{pmatrix}\mathbf{1} & \boldsymbol{X}\end{pmatrix}\right\}^{-1}\begin{pmatrix}\mathbf{1}^\top\\ \boldsymbol{X}^\top\end{pmatrix}\boldsymbol{y} = \begin{pmatrix}n^{-1} & 0\\ 0 & (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{1}^\top\\ \boldsymbol{X}^\top\end{pmatrix}\boldsymbol{y} = \begin{pmatrix}\bar{y}\\ (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\end{pmatrix}.\square$$

2. Let $\boldsymbol{X'X} = \boldsymbol{GLG'}$ be a spectral decomposition of $\boldsymbol{X'X}$ and let $\boldsymbol{W} = \boldsymbol{XG}$ be the principal component transformation given in (8.8.2). If $\boldsymbol{\alpha} = \boldsymbol{G'\beta}$ represents the parameter vector for the principal components, show that the ridge estimate $\boldsymbol{\alpha}^*$ of $\boldsymbol{\alpha}$ can be simply related to the OLS estimate $\widehat{\boldsymbol{\alpha}}$ by

$$\alpha_j^* = \frac{l_j}{l_j + k}\widehat{\alpha}_j, \quad j = 1, \dots, p,$$

and hence

$$\boldsymbol{\beta}^* = \boldsymbol{GDG'}\widehat{\boldsymbol{\beta}}, \text{ where } \boldsymbol{D} = \text{diag}\{l_i/(l_i + k)\}.$$

Solution. Denote $\boldsymbol{\gamma} = \boldsymbol{W'}y = \boldsymbol{G'X'y}$. The ridge estimator of α is

$$\boldsymbol{\alpha}^* = \boldsymbol{G'}(\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'y} = \boldsymbol{G'}(\boldsymbol{GLG'} + k\boldsymbol{GG'})^{-1}\boldsymbol{X'y} = (\boldsymbol{L} + k\boldsymbol{I})^{-1}\boldsymbol{G'X'y} = \begin{pmatrix} \gamma_1/(l_1 + k) \\ \vdots \\ \gamma_p/(l_p + k) \end{pmatrix} X'y.$$

The OLS estimator of α is the ridge estimator at $k = 0$, i.e.,

$$\widehat{\alpha} = \begin{pmatrix} \gamma_1/l_1 \\ \vdots \\ \gamma_p/l_p \end{pmatrix}.$$

Therefore, we have

$$\alpha_j^* = \frac{l_j}{l_j + k}\widehat{\alpha}_j, \quad j = 1, \dots, p,$$

or, equivalently, $\boldsymbol{\alpha}^* = \boldsymbol{D}\widehat{\boldsymbol{\alpha}}$. We have

$$\boldsymbol{G\alpha}^* = \boldsymbol{GDG'G}\widehat{\boldsymbol{\alpha}},$$

and by definition the $\boldsymbol{\alpha} = \boldsymbol{G'\beta}$ we further have

$$\boldsymbol{\beta}^* = \boldsymbol{GDG'}\widehat{\boldsymbol{\beta}}.\square$$

3. One measure of the accuracy of $\boldsymbol{\beta}^*$ is given by the trace mean square error,

$$\phi(k) = \text{tr}E\{(\boldsymbol{\beta}^* - \boldsymbol{\beta})(\boldsymbol{\beta}^* - \boldsymbol{\beta})'\} = \sum_{i=1}^{p} E(\beta_i^* - \beta_i)^2.$$

Show that we can write $\phi(k) = \gamma_1(k) + \gamma_2(k)$, where

$$\gamma_1(k) = \sum_{i=1}^{p} V(\beta_i^*) = \sigma^2 \sum_{i=1}^{p} \frac{l_i}{(l_i + k)^2}$$

represents the sum of the variances of β_i^*, and

$$\gamma_2(k) = \sum_{i=1}^{p}\{E(\beta_i^* - \beta_i)\}^2 = k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(l_i + k)^2}$$

represents the sum of the squared biases of β_i^*.

Solution. We have the following bias^2-variance decomposition:

$$\phi(k) = \text{tr}E\{(\boldsymbol{\beta}^* - \boldsymbol{\beta})(\boldsymbol{\beta}^* - \boldsymbol{\beta})'\} = \text{tr}\{(E\boldsymbol{\beta}^* - \boldsymbol{\beta})(E\boldsymbol{\beta}^* - \boldsymbol{\beta})'\} + \text{tr cov}(\boldsymbol{\beta}^*),$$

where the first term is the bias^2, i.e.,

$$\gamma_2(k) = \text{tr}\{(E\boldsymbol{\beta}^* - \boldsymbol{\beta})(E\boldsymbol{\beta}^* - \boldsymbol{\beta})'\} = \sum_{i=1}^{p}\{E(\beta_i^* - \beta_i)\}^2,$$

and the second term is the variance, i.e.,

$$\gamma_1(k) = \text{tr cov}(\boldsymbol{\beta}^*) = \sum_{i=1}^{p} V(\beta_i^*).$$

Since

$$E\boldsymbol{\beta}^* - \boldsymbol{\beta} = \boldsymbol{GD\alpha} - \boldsymbol{G\alpha} = \boldsymbol{G}(\boldsymbol{D} - \boldsymbol{I})\boldsymbol{\alpha} = -k\boldsymbol{G}\begin{pmatrix}\alpha_1/(l_1+k)\\ \vdots \\ \alpha_p/(l_p+k)\end{pmatrix},$$

we have

$$\gamma_2(k) = k^2 \text{tr}\left\{\begin{pmatrix}\alpha_1/(l_1+k) & \cdots & \alpha_p/(l_p+k)\end{pmatrix}\begin{pmatrix}\alpha_1/(l_1+k)\\ \vdots \\ \alpha_p/(l_p+k)\end{pmatrix}\right\} = k^2\sum_{i=1}^{p}\frac{\alpha_i}{(l_i+k)^2}.$$

Since

$$\text{cov}(\boldsymbol{\beta}^*) = \sigma^2\boldsymbol{GDG}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{GDG}' = \sigma^2\boldsymbol{GDG}'(\boldsymbol{GLG}')^{-1}\boldsymbol{GDG}' = \sigma^2\boldsymbol{G}\text{diag}\left\{\frac{l_1}{(l_1+k)^2},\cdots,\frac{l_p}{(l_p+k)^2}\right\}\boldsymbol{G}',$$

we have

$$\gamma_1(k) = \sigma^2\text{tr}\left[\boldsymbol{G}\text{diag}\left\{\frac{l_1}{(l_1+k)^2},\cdots,\frac{l_p}{(l_p+k)^2}\right\}\boldsymbol{G}'\right] = \sigma^2\sum_{i=1}^{p}\frac{l_i}{(l_i+k)^2}.$$

4. Show that the first derivative of $\gamma_1(k)$ and $\gamma_2(k)$ at 0 are

$$\gamma_1'(0) = -2\sigma^2\sum 1/l_i^2, \quad \gamma_2'(0) = 0.$$

Hence there exist values of $k > 0$ for which $\phi(k) < \phi(0)$, that is for which $\boldsymbol{\beta}^*$ has smaller trace mean square error than $\widehat{\boldsymbol{\beta}}$. Note that the increase in accuracy is most pronounced when some of the eigenvalues l_i are near 0, that is, when the columns of $\boldsymbol{X}$ are nearly colinear. However, the optimal choice for k depends on the unknown value of $\boldsymbol{\beta} = \boldsymbol{G\alpha}$.

Solution. The first derivative is $\gamma_1(k)$ is

$$\gamma_1'(k) = -2\sigma^2\sum_{i=1}^{p}\frac{l_i}{(l_i+k)^3},$$

and therefore,

$$\gamma_1'(0) = -2\sigma^2\sum_{i=1}^{p} l_i^{-2}.$$

The first derivative of $\gamma_2(k)$ is

$$\gamma_2'(k) = -2k^2\sum_{i=1}^{p}\frac{\alpha_i}{(l_i+k)^3} + 2k\sum_{i=1}^{p}\frac{\alpha_i}{(l_i+k)^2},$$

and therefore,

$$\gamma_2'(0) = 0.$$

Other conclusions follow straightforwardly. □