Jared Yu
Module 9 Assignment

1. Fisher's linear discriminant is

$$\hat{\mathbf{w}}^* = \underset{\hat{\mathbf{w}}}{\operatorname{argmax}}\, J(\hat{\mathbf{w}}) = \underset{\hat{\mathbf{w}}}{\operatorname{argmax}} \frac{\hat{\boldsymbol{w}}^\top \mathbf{S}_b \hat{\boldsymbol{w}}}{\hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}}},$$

where $\mathbf{S}_b = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^\top$ and $\mathbf{S}_w = \sum_j \sum_\alpha (\mathbf{x}_\alpha - \mathbf{m}_j)(\mathbf{x}_\alpha - \mathbf{m}_j)^\top$.

   a. By writing $\frac{\partial J}{\partial \hat{\mathbf{w}}} = 0$, show that

$$\mathbf{S}_w^{-1}\mathbf{S}_b\hat{\mathbf{w}} = J(\hat{\mathbf{w}})\hat{\mathbf{w}}.$$

<u>Ans:</u> Reference: [1], [2]

$$\frac{\partial J}{\partial \hat{\mathbf{w}}} = \frac{\partial}{\partial \hat{\mathbf{w}}} \cdot \frac{\hat{\boldsymbol{w}}^\top \mathbf{S}_b \hat{\boldsymbol{w}}}{\hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}}} = \frac{\left(\frac{\partial}{\partial \hat{\mathbf{w}}} \hat{\boldsymbol{w}}^\top \mathbf{S}_b \hat{\boldsymbol{w}}\right) \hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}} - \left(\frac{\partial}{\partial \hat{\mathbf{w}}} \hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}}\right) \hat{\boldsymbol{w}}^\top \mathbf{S}_b \hat{\boldsymbol{w}}}{(\hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}})^2}$$

$$= \frac{(2\mathbf{S}_b\hat{\boldsymbol{w}})\hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}} - (2\mathbf{S}_w\hat{\boldsymbol{w}})\hat{\boldsymbol{w}}^\top \mathbf{S}_b \hat{\boldsymbol{w}}}{(\hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}})^2} = 0$$

$$\Rightarrow (2\mathbf{S}_b\hat{\boldsymbol{w}})\hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}} - (2\mathbf{S}_w\hat{\boldsymbol{w}})\hat{\boldsymbol{w}}^\top \mathbf{S}_b \hat{\boldsymbol{w}} = 0$$

$$\Rightarrow \hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}}(\mathbf{S}_b\hat{\boldsymbol{w}}) - \hat{\boldsymbol{w}}^\top \mathbf{S}_b \hat{\boldsymbol{w}}(\mathbf{S}_w\hat{\boldsymbol{w}}) = 0$$

$$\Rightarrow \frac{\hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}}(\mathbf{S}_b\hat{\boldsymbol{w}})}{\hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}}} - \frac{\hat{\boldsymbol{w}}^\top \mathbf{S}_b \hat{\boldsymbol{w}}(\mathbf{S}_w\hat{\boldsymbol{w}})}{\hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}}} = 0$$

$$\Rightarrow \mathbf{S}_b\hat{\boldsymbol{w}} - J(\hat{\mathbf{w}})\mathbf{S}_w\hat{\boldsymbol{w}} = 0$$

$$\Rightarrow \mathbf{S}_b\hat{\boldsymbol{w}} = J(\hat{\mathbf{w}})\mathbf{S}_w\hat{\boldsymbol{w}}$$

$$\Rightarrow \mathbf{S}_w^{-1}\mathbf{S}_b\hat{\mathbf{w}} = J(\hat{\mathbf{w}})\hat{\mathbf{w}} \;\; \blacksquare$$

The last step of taking the inverse is possible if $\mathbf{S}_w$ is full rank and thus invertible.

   b. Explain why $\hat{\mathbf{w}}^*$ is the eigenvector for which $J(\hat{\mathbf{w}})$ is the maximum eigenvalue of $\mathbf{S}_w^{-1}\mathbf{S}_b$.

<u>Ans:</u>
Looking at the result $\mathbf{S}_w^{-1}\mathbf{S}_b\hat{\mathbf{w}} = J(\hat{\mathbf{w}})\hat{\mathbf{w}}$ from part a), we can see that is a square matrix $\mathbf{S}_w^{-1}\mathbf{S}_b$, $J(\hat{\mathbf{w}})$ is a scalar, and $\hat{\mathbf{w}}$ is a vector. Combining these facts, we have an equation in the form of $\mathbf{Av} = \lambda\mathbf{v}$, which when solved for $\lambda$ is the eigenvalue and $\mathbf{v}$ is the eigenvector. Therefore, we have here that $\hat{\mathbf{w}}^*$ in turn becomes the eigenvector.

   c. Explain why $\mathbf{S}_b\hat{\mathbf{w}}$ is always in the direction of $\mathbf{m}_1 - \mathbf{m}_2$ and thus show that

$$\hat{\mathbf{w}}^* = \text{const.} \cdot \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2).$$

<u>Ans:</u> Reference: [1]
For any vector $\mathbf{x}$, $\mathbf{S}_b\mathbf{x}$ will point in the same direction as $\mathbf{m}_1 - \mathbf{m}_2$. This will be shown below:

$$\mathbf{S}_b\mathbf{x} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^\top\mathbf{x} = \alpha(\mathbf{m}_1 - \mathbf{m}_2)$$

where $\alpha = (\mathbf{m}_1 - \mathbf{m}_2)^\top\mathbf{x}$. Thus, it follows that:

$$\Rightarrow \mathbf{S}_w^{-1}\mathbf{S}_b\hat{\mathbf{w}} = J(\hat{\mathbf{w}})\hat{\mathbf{w}}$$

$$\Rightarrow \alpha\mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2) = J(\hat{\mathbf{w}})\hat{\mathbf{w}}$$

$$\Rightarrow \hat{\mathbf{w}}^* = \text{const.} \cdot \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

where const. $= \frac{\alpha}{J(\hat{\mathbf{w}})}$.

2. Another way to optimize Fisher's linear discriminant (suggested by Barry Fridling):
   a. Show that for any two real vectors **x** and **y**

$$(\mathbf{x}^\top\mathbf{y})^2 \leq (\mathbf{x}^\top\mathbf{x})(\mathbf{y}^\top\mathbf{y}), \qquad (\text{Cauchy} - \text{Schwarz}).$$

<u>Ans:</u> References: [3], [4]

First let the two real vectors **x** and **y** be nonzero vectors, since the inequality is trivially true when either or both are the zero vector (i.e., $0 \leq 0$). Notice that for all $t \in \mathbb{R}$,

$$(\mathbf{x} + t\mathbf{y}) \cdot (\mathbf{x} + t\mathbf{y}) \geq 0$$

$$\mathbf{x} \cdot \mathbf{x} + \mathbf{x} \cdot (t\mathbf{y}) + (t\mathbf{y}) \cdot \mathbf{x} + (t\mathbf{y}) \cdot (t\mathbf{y}) \geq 0$$

$$\underbrace{\|\mathbf{x}\|^2}_{c} + \underbrace{2(\mathbf{x} \cdot \mathbf{y})}_{b}\, t + t^2 \underbrace{\|\mathbf{y}\|^2}_{a} \geq 0$$

The above is a quadratic equation around $t$. Using the vertex formula, we find that

$$t = -\frac{b}{2a} = -\frac{2(\mathbf{x} \cdot \mathbf{y})}{2\|\mathbf{y}\|^2} = -\frac{(\mathbf{x} \cdot \mathbf{y})}{\|\mathbf{y}\|^2}.$$

Plugging this into the formula we have,

$$\|\mathbf{x}\|^2 + 2(\mathbf{x} \cdot \mathbf{y})\left(-\frac{(\mathbf{x} \cdot \mathbf{y})}{\|\mathbf{y}\|^2}\right) + \left(-\frac{(\mathbf{x} \cdot \mathbf{y})}{\|\mathbf{y}\|^2}\right)^2 \|\mathbf{y}\|^2 \geq 0$$

$$\frac{(\mathbf{x} \cdot \mathbf{y})^{\mathbf{2}}}{\|\mathbf{y}\|^2} - 2\frac{(\mathbf{x} \cdot \mathbf{y})^2}{\|\mathbf{y}\|^2} + \|\mathbf{x}\|^2 \geq 0$$

$$(\mathbf{x} \cdot \mathbf{y})^{\mathbf{2}} - 2(\mathbf{x} \cdot \mathbf{y})^{\mathbf{2}} + \|\mathbf{x}\|^2\|\mathbf{y}\|^2 \geq 0$$

$$-(\mathbf{x} \cdot \mathbf{y})^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2 \geq 0$$

$$\|\mathbf{x}\|^2\|\mathbf{y}\|^2 \geq (\mathbf{x} \cdot \mathbf{y})^{\mathbf{2}}$$

$$(\mathbf{x} \cdot \mathbf{y})^2 \leq (\mathbf{x}^\top\mathbf{x})(\mathbf{y}^\top\mathbf{y}) \;\blacksquare$$

   b. Show that if the $\lambda_k$ are positive,

$$\left(\sum_{k=1}^{N} x_k y_k\right)^2 \leq \left(\sum_{k=1}^{N} \lambda_k x_k^2\right)\left(\sum_{k=1}^{N} \frac{y_k^2}{\lambda_k}\right).$$

<u>Ans:</u> Reference: [5]

We can denote $\tilde{x}_k = x_k\sqrt{\lambda_k}$ and $\tilde{y}_k = \frac{y_k}{\sqrt{\lambda_k}}$. The Cauchy-Schwarz inequality from part a) can also be rewritten as follows,

$$(\mathbf{x} \cdot \mathbf{y})^{\mathbf{2}} \leq (\mathbf{x}^\top\mathbf{x})(\mathbf{y}^\top\mathbf{y})$$

$$\left(\sum_{k=1}^{N} x_k y_k\right)^2 \leq \left(\sum_{k=1}^{N} x_k^2\right)\left(\sum_{k=1}^{N} y_k^2\right)$$

Then, plugging in $\tilde{x}_k$ and $\tilde{y}_k$ into the above Cauchy-Schwarz inequality yields:

$$\left(\sum_{k=1}^{N} \tilde{x}_k \tilde{y}_k\right)^2 \leq \left(\sum_{k=1}^{N} \tilde{x}_k^2\right)\left(\sum_{k=1}^{N} \tilde{y}_k^2\right)$$

$$\left(\sum_{k=1}^{N} x_k\sqrt{\lambda_k}\frac{y_k}{\sqrt{\lambda_k}}\right)^2 \leq \left(\sum_{k=1}^{N} \left(x_k\sqrt{\lambda_k}\right)^2\right)\left(\sum_{k=1}^{N} \left(\frac{y_k}{\sqrt{\lambda_k}}\right)^2\right)$$

$$\left(\sum_{k=1}^{N} x_k y_k\right)^2 \leq \left(\sum_{k=1}^{N} \lambda_k x_k^2\right)\left(\sum_{k=1}^{N} \frac{y_k^2}{\lambda_k}\right)$$

which we know holds based on the proof for Cauchy-Schwarz seen in part a). ■

c. Thus, show that for $\mathbf{A}$ positive definite

$$(\mathbf{x}^\top \mathbf{y})^2 \leq (\mathbf{x}^\top \mathbf{A}\mathbf{x})(\mathbf{y}^\top \mathbf{A}^{-1}\mathbf{y}).$$

Ans: Reference: [5], [6], [7], [8]
First let the two real vectors $\mathbf{x}$ and $\mathbf{y}$ be nonzero vectors, since the inequality is trivially true when either or both are the zero vector (i.e., $0 \leq 0$). It is stated that $\mathbf{A}$ is positive definite, therefore it contains at least one matrix square root. Furthermore, the inverse of a positive definite matrix is also positive definite. Let $\mathbf{A}^{\frac{1}{2}}$ and $\mathbf{A}^{-\frac{1}{2}}$ then be the square root matrices of $\mathbf{A}$ and $\mathbf{A}^{-1}$ respectively. Some other properties are that $\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}} = \mathbf{I}$, $\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{A}$, and $\mathbf{A}^{-\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}} = \mathbf{A}^{-1}$. Another important note is that since $\mathbf{A}$ is positive definite, we also have that $\mathbf{A}^{\frac{1}{2}} = \left(\mathbf{A}^{\frac{1}{2}}\right)^\top$.

From this it follows that for vectors $\mathbf{x}$ and $\mathbf{y}$,

$$\mathbf{x}^\top \mathbf{y} = \mathbf{x}^\top \mathbf{I}\mathbf{y} = \mathbf{x}^\top \mathbf{A}^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}\mathbf{y} = \left(\mathbf{A}^{\frac{1}{2}}\mathbf{x}\right)^\top \left(\mathbf{A}^{-\frac{1}{2}}\mathbf{y}\right).$$

We can then apply the Cauchy-Schwarz inequality to the vectors $\mathbf{A}^{\frac{1}{2}}\mathbf{x}$ and $\mathbf{A}^{-\frac{1}{2}}\mathbf{y}$. To simplify notation, we can let $\tilde{\mathbf{x}} = \mathbf{A}^{\frac{1}{2}}\mathbf{x}$ and $\tilde{\mathbf{y}} = \mathbf{A}^{-\frac{1}{2}}\mathbf{y}$.

Therefore, it follows that,

$$(\tilde{\mathbf{x}}^\top \tilde{\mathbf{y}})^2 \leq (\tilde{\mathbf{x}}^\top \tilde{\mathbf{x}})(\tilde{\mathbf{y}}^\top \tilde{\mathbf{y}})$$
$$(\mathbf{x}^\top \mathbf{y})^2 \leq \left[\left(\mathbf{A}^{\frac{1}{2}}\mathbf{x}\right)^\top \mathbf{A}^{\frac{1}{2}}\mathbf{x}\right]\left[\left(\mathbf{A}^{-\frac{1}{2}}\mathbf{y}\right)^\top \mathbf{A}^{-\frac{1}{2}}\mathbf{y}\right]$$
$$(\mathbf{x}^\top \mathbf{y})^2 \leq \left(\mathbf{x}^\top \mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\mathbf{x}\right)\left(\mathbf{y}^\top \mathbf{A}^{-\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}\mathbf{y}\right)$$
$$(\mathbf{x}^\top \mathbf{y})^2 \leq (\mathbf{x}^\top \mathbf{A}\mathbf{x})(\mathbf{y}^\top \mathbf{A}^{-1}\mathbf{y}) \ ■$$

d. By letting $\mathbf{A} = \mathbf{S}_w$ in the expression above, and writing

$$J(\hat{\mathbf{w}}) = \frac{|\hat{\boldsymbol{w}}^\top(\mathbf{m}_1 - \mathbf{m}_2)|^2}{\hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}}},$$

show again that

$$\hat{\mathbf{w}}^* = \text{const.} \cdot \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2).$$

Ans: Reference: (office hours 11/9/20)
We have here in part d) that $\mathbf{A} = \mathbf{S}_w$, $\mathbf{x} = \hat{\boldsymbol{w}}$, and $\mathbf{y} = \mathbf{m}_1 - \mathbf{m}_2$. Applying this to the inequality from part c), we get the following,

$$(\mathbf{x}^\top \mathbf{y})^2 \leq (\mathbf{x}^\top \mathbf{A}\mathbf{x})(\mathbf{y}^\top \mathbf{A}^{-1}\mathbf{y})$$
$$|\hat{\boldsymbol{w}}^\top(\mathbf{m}_1 - \mathbf{m}_2)|^2 \leq (\hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}})\left[(\mathbf{m}_1 - \mathbf{m}_2)^\top \mathbf{S}_w{}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)\right] \qquad (1)$$

Furthermore, we are given that,

$$J(\hat{\mathbf{w}}) = \frac{|\hat{\boldsymbol{w}}^\top(\mathbf{m}_1 - \mathbf{m}_2)|^2}{\hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}}}$$

or equivalently,

$$|\hat{\boldsymbol{w}}^\top(\mathbf{m}_1 - \mathbf{m}_2)|^2 = \hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}} J(\hat{\mathbf{w}}).$$

This implies that from equation (1),

$$\hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}} J(\hat{\mathbf{w}}) \leq (\hat{\boldsymbol{w}}^\top \mathbf{S}_w \hat{\boldsymbol{w}})\left[(\mathbf{m}_1 - \mathbf{m}_2)^\top \mathbf{S}_w{}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)\right]$$

$$J(\widehat{\mathbf{w}}) \leq (\mathbf{m}_1 - \mathbf{m}_2)^\top \mathbf{S}_w{}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

An important note about the Cauchy-Schwarz inequality (i.e. $(\mathbf{x}^\top \mathbf{y})^2 \leq (\mathbf{x}^\top \mathbf{A}\mathbf{x})(\mathbf{y}^\top \mathbf{A}^{-1}\mathbf{y})$) is that it becomes an equality iff $\mathbf{A}\mathbf{x} = \text{const.}\,\mathbf{y}$ or equivalently $\mathbf{x} = \text{const.}\,\mathbf{A}^{-1}\mathbf{y}$. We can show that this is the case here in part d) with the following steps. Let $\mathbf{A} = \mathbf{S}_w = \mathbf{B}\mathbf{B}^\top$. Then let $\xi = \mathbf{B}^\top \mathbf{x}$ and $\eta = \mathbf{B}^{-1}\mathbf{y}$. From this it follows that $\xi = \text{const.}\,\eta$. By looking at $\mathbf{B}^\top \mathbf{x} = \mathbf{B}^{-1}\mathbf{y} \cdot \text{const.}$, we further get that $\mathbf{A}\mathbf{x} = \text{const.}\,\mathbf{y}$ if we multiply both sides from the left by $\mathbf{B}$. In part d), $\mathbf{A}\mathbf{x} = \text{const.}\,\mathbf{y}$ corresponds to $\mathbf{S}_w \cdot \widehat{\mathbf{w}} = \text{const.}\,(\mathbf{m}_1 - \mathbf{m}_2)$.

This implies then that our inequality is an equality. In other words,

$$|\widehat{\boldsymbol{w}}^\top (\mathbf{m}_1 - \mathbf{m}_2)|^2 = (\widehat{\boldsymbol{w}}^\top \mathbf{S}_w \widehat{\boldsymbol{w}})\left[(\mathbf{m}_1 - \mathbf{m}_2)^\top \mathbf{S}_w{}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)\right]$$

$$\frac{|\widehat{\boldsymbol{w}}^\top (\mathbf{m}_1 - \mathbf{m}_2)|^2}{(\widehat{\boldsymbol{w}}^\top \mathbf{S}_w \widehat{\boldsymbol{w}})} = \left[(\mathbf{m}_1 - \mathbf{m}_2)^\top \mathbf{S}_w{}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)\right]$$

$$J(\widehat{\mathbf{w}}) = \left[(\mathbf{m}_1 - \mathbf{m}_2)^\top \mathbf{S}_w{}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)\right]$$

iff

$$\mathbf{S}_w \cdot \widehat{\mathbf{w}} = \text{const.}\,(\mathbf{m}_1 - \mathbf{m}_2),$$

which we have just shown to be the case, due to the positive definite property of $\mathbf{S}_w$ (for convenience, it is also being assumed that it is a diagonal matrix). From there it follows that,

$$\widehat{\mathbf{w}}^* = \text{const.} \cdot \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2).\ \blacksquare$$

3. Using the Optidigits dataset from the UCI repository, implement PCA. Reconstruct the digit images and calculate the reconstruction error $E(n) = \sum_j \left\|\hat{\mathbf{x}}_j - \mathbf{x}\right\|^2$ for various values of $n$, the number of eigenvectors. Plot $E(n)$ versus $n$.

<u>Ans:</u> Reference: [9], [10]

The data was first normalized by first centering each column by the corresponding sample mean and then dividing it by the corresponding sample standard deviation. Two exceptions are the 1st and 40th columns, where they consist only of zeros, therefore they cannot be normalized and left as is. Based on the normalized dataset, the sample variance-covariance matrix is calculated. Following this, the eigenvectors are then computed, leading to a $64 \times 64$ principal component (PC) matrix, $\mathbf{W}$. To construct the projection matrix, we can calculate $\mathbf{Z} = \mathbf{X} \times \mathbf{W}$, where $\mathbf{X}$ is the original dataset of 3,823 observations and 64 attributes. However, we can also limit the number of PC's from the PC matrix to use in the projection, using between 1 and 64 (i.e., limit the number of PC's by limiting the number of columns in $\mathbf{W}$, where we take either the first, or the first $p$ columns for $p$ PC's to consider).

The resulting plot can be seen below in Figure 1. It is interesting to see also that at the end, that $E(n)$ flattens for the last three added PC's. This seems to be due to the fact that 2 of the 64 columns consist only of zeroes.

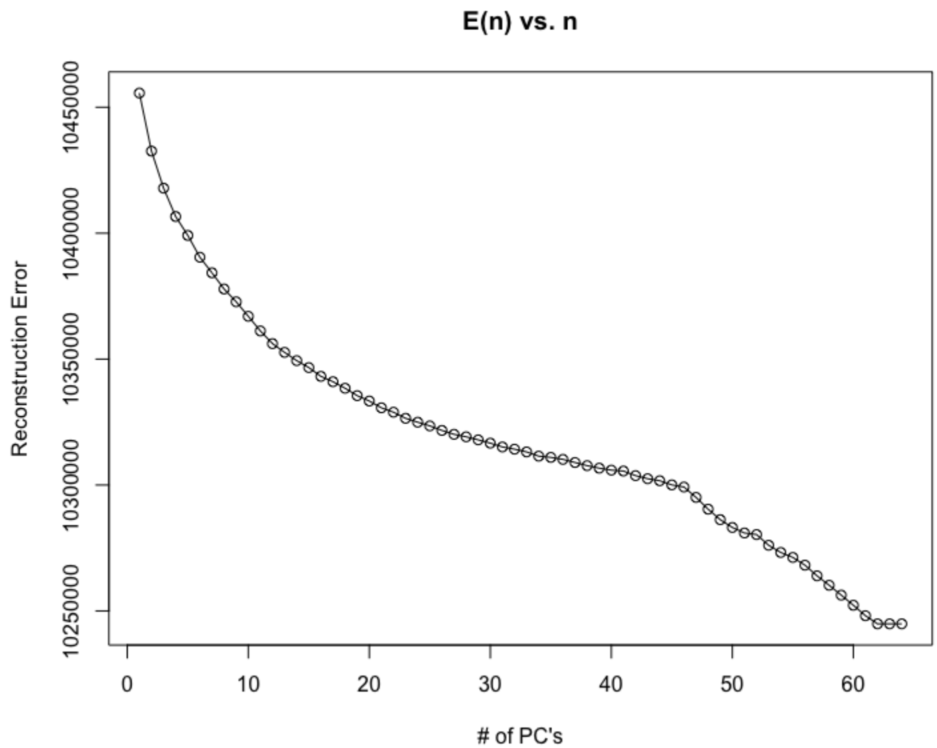


*Figure 1 The above figure shows a plot of the reconstruction error E(n) vs. the corresponding number of PC's used to calculate E(n). It thus is showing the E(n) for when either 1,..,64 PC's are used.*

Reference:

[1] https://www.csd.uwo.ca/~oveksler/Courses/CS434a_541a/Lecture8.pdf

[2] https://piazza.com/class/kc0jkwru805u1?cid=181

[3] https://www.youtube.com/watch?v=wECTos-t_EQ

[4] https://www.youtube.com/watch?v=SPCYCVa5DmM

[5] https://piazza.com/class/kc0jkwru805u1?cid=184

[6] https://math.stackexchange.com/questions/1226455/what-does-a-positive-definite-matrix-have-to-do-with-cauchy-schwarz-inequality

[7] https://mathworld.wolfram.com/PositiveDefiniteMatrix.html

[8] https://math.stackexchange.com/questions/3268470/when-is-square-root-of-transpose-and-transpose-of-square-root-of-a-matrix-are-eq

[9] https://stats.stackexchange.com/questions/194278/meaning-of-reconstruction-error-in-pca-and-lda

[10] http://www.cs.cornell.edu/courses/cs4786/2016sp/lectures/lec03.pdf

## Code Appendix:

```r
# Load optidigits data
train <- read.table('optdigits.tra', sep = ',')
table(train$V65)
# test <- read.table('optdigits.tes', sep = ',')
# table(test$V65)

# V1-V64 are features, V65 is target vector
# attributes are ranged 0:16
# class are ranged 0:9
# no N/A
X <- train[,1:(ncol(train)-1)]
sample_means <- colMeans(X)
sample_sd <- sqrt(diag(cov(X)))
X_list <- as.list(X)

# Reference: https://stackoverflow.com/questions/39731068/how-to-let-a-matrix-minus-vector-by-row-rather-t
han-by-column
# Reference: https://stackoverflow.com/questions/3444889/how-to-use-the-sweep-function
X_centered <- sweep(X, 2, colMeans(X))
X_standardized <- sweep(X_centered, 2, sample_sd, FUN = "/")
### The first column is all zeros, cannot be standardized
X_standardized$V1 <- 0
X_standardized$V40 <- 0
sum(is.na(X_standardized))
X <- X_standardized

### Normalization stats
colMeans(X) # roughly zero
diag(cov(X)) # all ones except cols 1, 40
### Normalization stats end

S <- cov(X)
eigens <- eigen(S)
W <- eigens$vectors # W is the PC matrix that is 64x64
Z <- as.matrix(X) %*% as.matrix(W) # projection matrix

# Reconstruction
# x_hat <- t(Z[1,]) %*% t(W) + sample_means # single observation
X_hat <- Z %*% t(W) + sample_means # reconstruction matrix

# Create an X-hat for 1-64 PC's
X_hat_levels <- lapply(1:64, function(x) {
  Z <- as.matrix(X) %*% as.matrix(W[,1:x])
  X_hat <- Z %*% t(W[,1:x]) + sample_means
  X_hat
})

reconstruction_error_list <- lapply(X_hat_levels, function(x) {
  sum(rowSums((x - X)^2))
})
reconstruction_error_df <- do.call(rbind, reconstruction_error_list)
plot(1:64, reconstruction_error_df, type = 'l',
     main = 'E(n) vs. n', xlab = '# of PC\'s', ylab = 'Reconstruction Error')
points(1:64, reconstruction_error_df, pch = 1)

# Reconstruction error
a = X_hat - X
total_reconstruction_error <- rowSums((X_hat - X)^2)
```