The material from this lesson is drawn from "Linear Algebra and its Application", 5th Edition, by Lay, "A First Course in Linear Model Theory" by Ravishanker and Dey, from Dr. Steve MacEachern's Spring 2014 course "STAT 6860: Foundations of the Linear Model" at The Ohio State University, and from "A Tutorial on Principal Components Analysis," accessed at http://faculty.iiit.ac.in/ mkrishna/PrincipalComponents.pdf.

I. Eigenvalues and Eigenvectors

Recall that linear transformations are transformations that preserve additivity and scalar multiplication. More importantly, we can write any linear transformation $f(\mathbf{x})$ as the matrix multiplication $\mathbf{Ax}$.

A particular class of linear transformations of interest are those satisfying the equation $\mathbf{Ax} = \lambda\mathbf{x}$, where λ is a scalar. This implies that the matrix $\mathbf{A}$ merely stretches or shrinks $\mathbf{x}$ but does not otherwise change $\mathbf{x}$.

If $\mathbf{Ax} = \lambda\mathbf{x}$, we say that λ is an eigenvalue of $\mathbf{A}$ and that $\mathbf{x}$ is the eigenvector of $\mathbf{A}$ that corresponds to λ.

- Given a matrix $\mathbf{A}$, we showed that we can find λ by solving $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ for λ.
- Given a matrix $\mathbf{A}$ and a nonzero eigenvalue λ, we showed that we can find the corresponding eigenvector $\mathbf{x}$ by solving $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$ for $\mathbf{x}$.

II. Decomposition of Matrices

We have briefly discussed before the idea of decomposing one matrix $\mathbf{A}$ into multiple matrices.

Formally, a matrix $\mathbf{A}$ can be decomposed into matrices $\mathbf{X}$ and $\mathbf{Y}$ if $\mathbf{A} = \mathbf{XY}$.

While this seems counterintuitive, there are particular benefits to decomposing matrices.

- When solving the system of equations $\mathbf{Ax} = \mathbf{b}$, applying the LU decomposition to $\mathbf{A}$ decomposes $\mathbf{A}$ into $\mathbf{L}$ and $\mathbf{U}$, where $\mathbf{L}$ is a lower-triangular matrix and $\mathbf{U}$ is an upper-triangular matrix. (This exists for many square matrices; a slight variant called the LUP decomposition exists for all square matrices.)
  - Computers can use the properties of triangular matrices to invert $\mathbf{LU}$ much more quickly than $\mathbf{A}$.
- The Cholesky decomposition of $\mathbf{A}$ into $\mathbf{VV^T}$ where $\mathbf{V} = \mathbf{LD}^{1/2}$, $\mathbf{L}$ is lower-triangular and $\mathbf{D}$ is a diagonal matrix, can be used for any square, symmetric, positive definite matrix $\mathbf{A}$.
  - A matrix $\mathbf{A}$ is positive definite if $\mathbf{xAx^T} > 0$ for all nonzero vectors $\mathbf{x}$. Positive definiteness ensures nice properties like invertibility of certain matrices.
  - While the Cholesky decomposition can only be applied to a subset of the matrices to which the LU decomposition can be applied, the Cholesky decomposition is roughly twice as computationally efficient.
- The spectral decomposition (also called the eigendecomposition) of a matrix $\mathbf{A}$ is where $\mathbf{A} = \mathbf{PDP}^{-1}$ with $\mathbf{D} = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ and $\mathbf{P}$ consisting of the eigenvectors corresponding to the eigenvalues in $\mathbf{D}$.
  - Suppose we want to find $\mathbf{A}^k$.

$$\begin{aligned}
\mathbf{A}^k &= (\mathbf{PDP}^{-1})^k \\
&= (\mathbf{PDP}^{-1})(\mathbf{PDP}^{-1})\cdots(\mathbf{PDP}^{-1}) \\
&= \mathbf{PD}^k\mathbf{P}^{-1} \\
&= \mathbf{P} \times \text{diag}\{\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k\} \times \mathbf{P}^{-1}
\end{aligned}$$

  - Spectral decomposition will also be very important in principal component analysis.

  - This works for any diagonalizable (also called "diagonable") matrix, which means there are n independent eigenvectors.

III. Review: Covariance Matrix

Recall that, for any two random variables, we can calculate the covariance.

$$\text{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n}$$

When working with a vector of random variables, we can construct a covariance matrix comparing the covariances of the different random variables. Consider the matrix $\mathbf{X} = [\mathbf{X}_1 \;\; \mathbf{X}_2 \;\; \cdots \;\; \mathbf{X}_p]$.

$$\begin{aligned}
\text{Cov}(\mathbf{X}) &= \begin{bmatrix} \text{Cov}(\mathbf{X}_1,\mathbf{X}_1) & \text{Cov}(\mathbf{X}_1,\mathbf{X}_2) & \cdots & \text{Cov}(\mathbf{X}_1,\mathbf{X}_p) \\ \text{Cov}(\mathbf{X}_2,\mathbf{X}_1) & \text{Cov}(\mathbf{X}_2,\mathbf{X}_2) & \cdots & \text{Cov}(\mathbf{X}_2,\mathbf{X}_p) \\ \vdots & \vdots & \vdots & \vdots \\ \text{Cov}(\mathbf{X}_p,\mathbf{X}_1) & \text{Cov}(\mathbf{X}_p,\mathbf{X}_2) & \cdots & \text{Cov}(\mathbf{X}_p,\mathbf{X}_p) \end{bmatrix} \\
&= \begin{bmatrix} \text{Var}(\mathbf{X}_1) & \text{Cov}(\mathbf{X}_1,\mathbf{X}_2) & \cdots & \text{Cov}(\mathbf{X}_1,\mathbf{X}_p) \\ \text{Cov}(\mathbf{X}_2,\mathbf{X}_1) & \text{Var}(\mathbf{X}_2) & \cdots & \text{Cov}(\mathbf{X}_2,\mathbf{X}_p) \\ \vdots & \vdots & \vdots & \vdots \\ \text{Cov}(\mathbf{X}_p,\mathbf{X}_1) & \text{Cov}(\mathbf{X}_p,\mathbf{X}_2) & \cdots & \text{Var}(\mathbf{X}_p) \end{bmatrix}
\end{aligned}$$

Note that the covariance matrix is diagonalizable.

IV. Principal Component Analysis

1. Take the matrix of independent variables $\mathbf{X}$ and, for each column $\mathbf{X}_1,\ldots,\mathbf{X}_p$, subtract the mean of that column from each entry.
   i. This ensures that the data set is centered at $\mathbf{0}$.
2. For the matrix of independent variables $\mathbf{X}$, calculate the covariance matrix $\text{Cov}(\mathbf{X})$.
3. Calculate the eigenvalues and eigenvectors of the covariance matrix.
   i. The highest eigenvalue corresponds to the eigenvector that explains the most variance in the data.
      a. Cloud example.
   ii. This eigenvector is called the "principal component" or the "first principal component." Thus, order the eigenvectors based on the eigenvalues from largest to smallest.
4. Determine how many principal components you want or what proportion of the variance you want to explain with your model. Then keep the proper number of principal components by taking the matrix of eigenvectors and dropping the eigenvectors corresponding to the smallest eigenvalues.
   i. This is called a "feature vector" despite being a bit of a misnomer. (It is a vector of vectors, which is ultimately a matrix.) We'll denote this $\mathbf{F}$.
      a. The columns of $\mathbf{F}$ should be a subset of the columns of the matrix of eigenvectors of $\text{Cov}(\mathbf{X})$.
5. Transpose the feature vector and multiply it by the original matrix of independent variables. We'll call this final result $\mathbf{Z}$.

$$\mathbf{Z} = \mathbf{F}^T\mathbf{X}$$

   i. The result here is the original data, but only in terms of the most important eigenvectors.

ii. Because each eigenvector is orthogonal to the others, each column of our new dataset $\mathbf{Z}$ is orthogonal to the other columns of $\mathbf{Z}$. This ensures that the assumption of independence of our features in a linear regression model is satisfied.

6. Fit your model by regressing your dependent variable $\mathbf{Y}$ on the transformed independent variables $\mathbf{Z}$.