MATH 271.1: Statistical Methods
Supplementary Notes on Principal Components Analysis
Jakov Ivan S. Dumbrique

3 Principal Components Analysis

Principal Components Analysis (PCA) is a technique for **dimensionality reduction** as it produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated. Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization. For an intuitive understanding of how PCA works, check out this website.

3.1 Preliminaries

Suppose we have an $n \times p$ data set $\hat{\mathbf{A}}$ (also called data matrix) consisting of n observations:

$$\begin{aligned}\hat{\mathbf{A}} &= \begin{bmatrix} \hat{a}_{11} & \hat{a}_{12} & \dots & \hat{a}_{1p} \\ \hat{a}_{21} & \hat{a}_{22} & \dots & \hat{a}_{2p} \\ \vdots & & \ddots & \vdots \\ \hat{a}_{n1} & \hat{a}_{n2} & \dots & \hat{a}_{np} \end{bmatrix} \\ &= \begin{bmatrix} \hat{\mathbf{a_1}} & \hat{\mathbf{a_2}} & \dots & \hat{\mathbf{a_p}} \end{bmatrix},\end{aligned}$$

where $\hat{\mathbf{a_1}}, \hat{\mathbf{a_2}}, \dots, \hat{\mathbf{a_p}}$ are called the feature vectors or simply *features* or *variables* of $\hat{\mathbf{A}}$. In this case, we will assume that $n > 1$–that is, our data matrix contains more than one observation.

Since we are only interested in the variance of this data matrix, we make it centered by transforming each column of $\hat{\mathbf{A}}$ to have a mean of zero. This is performed by subtracting the mean of each column $\bar{x}_j$ from the column $\hat{\mathbf{a_j}}$ itself. That is, we form the centered data matrix $\mathbf{A}$:

$$\begin{aligned}\mathbf{A} &= \hat{\mathbf{A}} - \bar{\boldsymbol{x}}^{\mathbf{T}} \\ &= \begin{bmatrix} \hat{\mathbf{a_1}} & \hat{\mathbf{a_2}} & \dots & \hat{\mathbf{a_p}} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \end{bmatrix} \\ &= \begin{bmatrix} \hat{\mathbf{a_1}} - \bar{x}_1 & \hat{\mathbf{a_2}} - \bar{x}_2 & \dots & \hat{\mathbf{a_p}} - \bar{x}_p \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{a_1} & \mathbf{a_2} & \dots & \mathbf{a_p} \end{bmatrix},\end{aligned}$$

where $\mathbf{a_1}, \mathbf{a_2}, \dots, \mathbf{a_p}$ are the features of $\mathbf{A}$.
We are interested in finding the principal components of this centered data matrix $\mathbf{A}$.

Variance-Covariance Matrix

Given an $n \times p$ data matrix $\mathbf{A}$, the sample variance-covariance matrix $\mathbf{S}$, often called

sample covariance matrix, refers to the following symmetric $p \times p$ matrix:

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ s_{21} & s_2^2 & \dots & s_{2p} \\ \vdots & & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_p^2 \end{bmatrix},$$

where

- $\bar{x}_j = \dfrac{1}{n}\sum_{i=1}^{n} a_{ij}$ is the sample mean of the jth feature $\hat{\mathbf{a_j}}$,
- $s_j^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - \bar{x}_j)^2$ is the sample variance of the jth feature $\hat{\mathbf{a_j}}$,
- $s_{jk} = \dfrac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - \bar{x}_j)(a_{ik} - \bar{x}_k)$ is the sample covariance between the jth and kth features $\hat{\mathbf{a_j}}$ and $\hat{\mathbf{a_k}}$.

Given a centered data matrix $\mathbf{A}$ where rows are observations and columns are features, it can be shown that the corresponding sample covariance matrix $\mathbf{S}$ is given by

$$\mathbf{S} = \frac{1}{n-1}\mathbf{A^T A}.$$

3.2 Finding the First Principal Component

The *first principal component* (first PC) of a set of features $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p$ of a centered data matrix $\mathbf{A}$ is the normalized linear combination of the features

$$\mathbf{z_1} = \phi_{11}\mathbf{a}_1 + \phi_{21}\mathbf{a}_2 + \ldots + \phi_{p1}\mathbf{a}_p$$

that has the largest variance. By *normalized*, we mean that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.

We refer to the elements $\phi_{11}, \dots, \phi_{p1}$ as the loadings of the first principal component (or first PC loadings); together, the loadings make up the *first principal component loading vector* $\boldsymbol{\phi_1} = [\phi_{11}\phi_{21} \dots \phi_{p1}]^T$. We constrain the loadings so that their sum of squares is equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

We note that in matrix form, the first principal component can be expressed as $\mathbf{z_1} = \mathbf{A}\boldsymbol{\phi_1}$ and the normalization constraint simplifies to $\boldsymbol{\phi_1^T}\boldsymbol{\phi_1} = 1$. Moreover, the sample variance of the first principal component, denoted by $\text{Var}(\mathbf{z_1}) = s_{\mathbf{z_1}}^2$, is given by the following theorem.

Theorem 3. Given a centered data matrix $\mathbf{A}$, the sample variance of its first principal component $\mathbf{z_1}$ is given by

$$\text{Var}(\mathbf{z_1}) = s_{\mathbf{z_1}}^2 = \boldsymbol{\phi_1^T}\mathbf{S}\boldsymbol{\phi_1},$$

where $\boldsymbol{\phi_1}$ is the first principal component loading vector and $\mathbf{S}$ is the sample covariance matrix of $\mathbf{A}$.

Proof: We first need to evaluate the sample mean of the first principal component $\mathbf{z_1} = [z_{11} z_{21} \dots z_{n1}]^T$:

$$\begin{aligned}
\bar{x}_{\mathbf{z_1}} &= \frac{1}{n}\sum_{i=1}^{n} z_{i1} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbf{a_i^T}\boldsymbol{\phi_1} \text{ (where } \mathbf{a_i} \text{ is the } i\text{th row of } \mathbf{A}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{p} a_{ij}\phi_{j1} \\
&= \frac{1}{n}\sum_{i=1}^{p}\sum_{i=1}^{n} a_{ij}\phi_{j1} \\
&= \sum_{i=1}^{p} \phi_{j1}\left[\frac{1}{n}\sum_{i=1}^{n} a_{ij}\right] \\
&= \sum_{i=1}^{p} \phi_{j1}\bar{x}_j \\
&= \sum_{i=1}^{p} \phi_{j1} 0 \\
&= 0
\end{aligned}$$

The variance of $\mathbf{z_1}$ is given by

$$\begin{aligned}
s^2_{\mathbf{z_1}} &= \frac{1}{n-1}\sum_{i=1}^{n} (z_{i1} - \bar{x}_{\mathbf{z_1}})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n} z_{i1}^2 \\
&= \frac{1}{n-1}\mathbf{z_1}^T\mathbf{z_1} \\
&= \frac{1}{n-1}(\mathbf{A}\boldsymbol{\phi_1})^{\mathbf{T}}(\mathbf{A}\boldsymbol{\phi_1}) \\
&= \frac{1}{n-1}\boldsymbol{\phi_1^T}\mathbf{A^T}\mathbf{A}\boldsymbol{\phi_1} \\
&= \boldsymbol{\phi_1^T}\left(\frac{1}{n-1}\mathbf{A^T}\mathbf{A}\right)\boldsymbol{\phi_1} \\
&= \boldsymbol{\phi_1^T}\mathbf{S}\boldsymbol{\phi_1}
\end{aligned}$$

where $\mathbf{S}$ is the sample covariance matrix of the centered data matrix $\mathbf{A}$.

Therefore, in order to find the first principal component $\mathbf{z_1} = \mathbf{A}\boldsymbol{\phi_1}$, we need to find the loadings that will maximize the sample variance $s^2_{\mathbf{z_1}} = \boldsymbol{\phi_1^T}\mathbf{S}\boldsymbol{\phi_1}$ subject to the normalization constraint $\boldsymbol{\phi_1^T}\boldsymbol{\phi_1} = 1$. We solve this constrained optimization problem using the method of Lagrange multipliers.

We start with a function $f(\boldsymbol{\phi_1})$ that we want to maximize, and an equality constraint $g(\boldsymbol{\phi_1}) = c$ for some real number c. In this case, $f(\boldsymbol{\phi_1}) = \boldsymbol{\phi_1^T S \phi_1}$, $g(\boldsymbol{\phi_1}) = \boldsymbol{\phi_1^T \phi_1}$ and $c = 1$. We rearrange the constraint equation so its RHS is zero, $g(\boldsymbol{\phi_1}) - c = 0$. We now add an extra variable to the problem, the Lagrange multiplier λ, and consider our new objective function

$$\begin{aligned} u(\boldsymbol{\phi_1}, \lambda) &= f(\boldsymbol{\phi_1}) - \lambda \left[g(\boldsymbol{\phi_1}) - c \right] \\ &= \boldsymbol{\phi_1^T S \phi_1} - \lambda \left[\boldsymbol{\phi_1^T \phi_1} - 1 \right] \end{aligned}$$

Differentiating this function with respect to its two arguments and setting the derivatives to zero will yield

$$\frac{\partial u}{\partial \boldsymbol{\phi_1}} = \frac{\partial f}{\partial \boldsymbol{\phi_1}} - \lambda \frac{\partial g}{\partial \boldsymbol{\phi_1}} = 0 \tag{1}$$

$$\frac{\partial u}{\partial \lambda} = - \left[g(\boldsymbol{\phi_1}) - c \right] = 0 \tag{2}$$

As seen in Equation 2, maximizing the objective function with respect to λ gives us back our normalization constraint. On the other hand, in order to simplify Equation 1, we have to evaluate the two partial derivatives:

$$\begin{aligned} \frac{\partial f}{\partial \boldsymbol{\phi_1}} &= \frac{\partial \left(\boldsymbol{\phi_1^T S \phi_1} \right)}{\partial \boldsymbol{\phi_1}} = \frac{\partial \boldsymbol{\phi_1}}{\partial \boldsymbol{\phi_1}} \mathbf{S} \boldsymbol{\phi_1} + \frac{\partial \boldsymbol{\phi_1}}{\partial \boldsymbol{\phi_1}} \mathbf{S^T} \boldsymbol{\phi_1} \\ &= 2 \mathbf{S} \boldsymbol{\phi_1} \\ \frac{\partial g}{\partial \boldsymbol{\phi_1}} &= \frac{\partial \left(\boldsymbol{\phi_1^T \phi_1} \right)}{\partial \boldsymbol{\phi_1}} = \frac{\partial \boldsymbol{\phi_1}}{\partial \boldsymbol{\phi_1}} \boldsymbol{\phi_1} + \frac{\partial \boldsymbol{\phi_1}}{\partial \boldsymbol{\phi_1}} \boldsymbol{\phi_1} \\ &= 2 \boldsymbol{\phi_1} \end{aligned}$$

Equation 1 then becomes

$$\begin{aligned} \frac{\partial u}{\partial \boldsymbol{\phi_1}} = 2 \mathbf{S} \boldsymbol{\phi_1} - 2 \lambda \boldsymbol{\phi_1} &= 0 \\ \mathbf{S} \boldsymbol{\phi_1} &= \lambda \boldsymbol{\phi_1} \end{aligned}$$

Thus, the desired vector $\boldsymbol{\phi_1}$ is an eigenvector of the sample covariance matrix $\mathbf{S}$. To decide which of the eigenvectors of $\mathbf{S}$ will maximize the variance of the first principal component $\mathbf{z_1}$, note that

$$s^2_{\mathbf{z_1}} = \boldsymbol{\phi_1^T S \phi_1} = \boldsymbol{\phi_1^T} \lambda \boldsymbol{\phi_1} = \lambda \boldsymbol{\phi_1^T \phi_1} = \lambda.$$

Therefore, the first principal component loading vector $\boldsymbol{\phi_1}$ is the eigenvector of the sample covariance matrix $\mathbf{S}$ that is associated with the largest eigenvalue $\lambda := \lambda_1$.

In general, the kth principal component is given by $\mathbf{z_k} = \mathbf{A} \boldsymbol{\phi_k}$, and $\text{Var}(\mathbf{z_k}) = s^2_{\mathbf{z_k}} = \lambda_k$, where λ_k is the kth largest eigenvalue of $\mathbf{S}$ and $\boldsymbol{\phi_k}$ is the corresponding eigenvector.

Exercise: Set up the constraint optimization problem of finding the second principal component $\mathbf{z_2}$ of a centered data matrix $\mathbf{A}$. Using Lagrange multipliers, show that the second principal component loading vector $\boldsymbol{\phi_2}$ is the eigenvector of the sample covariance matrix S that is associated with the second largest eigenvalue.

We have shown that finding the principal components of a centered data matrix $\mathbf{A}$ reduces to finding the eigenvalues and eigenvectors of the sample covariance matrix $\mathbf{S} = \frac{1}{n-1}\mathbf{A^T A}$. However, the eigenvectors of $\mathbf{S}$ are the same as the eigenvalues of $\mathbf{A^T A}$ and the eigenvalues of $\mathbf{S}$ are just the eigenvalues of $\mathbf{A^T A}$ scaled by a factor of $\frac{1}{n-1}$.

How do we find the eigenvalues and eigenvectors of $\mathbf{A^T A}$? If we know the SVD of A, then by Theorem 1, we already know the eigenvalues and eigenvectors of $\mathbf{A^T A}$, and hence that of $\mathbf{S}$ as well. That is, given the SVD of the centered data matrix $\mathbf{A} = \mathbf{U\Sigma V^T}$, then the first principal component loading vector $\boldsymbol{\phi_1}$ is the right singular vector corresponding to the largest singular value of A. This is simply the first right singular vector ($\mathbf{v_1}$) of A since the singular values in $\mathbf{\Sigma}$ are arranged in a non-increasing order. Also, the variance of the first principal component $\mathbf{z_1}$ is given by $\text{Var}(\mathbf{z_1}) = s^2_{\mathbf{z_1}} = \lambda_1 = \frac{1}{n-1}\sigma_1^2$. The other principal components and their loading vectors can be retrieved in a similar manner.

According to Trefethen and Bau (1997), calculating the SVD of an $n \times p$ matrix $\mathbf{A}$ ($4np^2 - \frac{4}{3}p^3$ flops) is less computationally expensive than computing the eigenvalue decomposition of $\mathbf{A^T A}$, which takes a number of flops of order $O(p^3)$. Statistical softwares use divide-and-conquer approach for solving the SVD of a matrix, which is more numerically stable than the QR algorithm used in solving for the eigenvalue decomposition of its covariance matrix.