Chapter 2

Principal Component Analysis and Its Extensions

In this chapter, we give a brief review of principal component analysis (PCA), i.e., the method for finding a dominant affine subspace to fit a set of data points. The solution to PCA has been well established in the literature and it has become one of the most useful tools for data modeling, compression, and visualization. In this section, we first show that the singular value decomposition (SVD) provides an optimal solution to PCA. Both the geometric and statistical formulation of PCA will be introduced and their equivalence will be established. When the dimension of the subspace is unknown, we introduce some conventional model selection methods to determine the number of principal components. When the samples contain outliers and incomplete data points, we review some robust statistical techniques that help resolve these difficulties. Finally, some nonlinear extensions to PCA such as nonlinear PCA and kernel PCA will also be reviewed.

2.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) refers to the problem of fitting a low-dimensional affine subspace S to a set of points $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ in a high-dimensional space $\mathbb{R}^D$, the ambient space. Mathematically, this problem can be formulated as either a statistical problem or a geometric one, and they both lead to the same solution, as we will show in this section.

2.1.1 A Geometric Approach to PCA

We first examine the more intuitive geometric approach to PCA. That is, one tries to find an (affine) subspace that fits the given data points. Let us assume for now that the dimension of the subspace d is known. Then every point $\boldsymbol{x}_i$ on a d-dimensional affine subspace in $\mathbb{R}^D$ can be represented as

$$\boldsymbol{x}_i = \boldsymbol{x}_0 + U_d \boldsymbol{y}_i, \quad i = 1, \ldots, N \tag{2.1}$$

where $\boldsymbol{x}_0 \in \mathbb{R}^D$ is a(ny) fixed point in the subspace, U_d is a $D \times d$ matrix with d orthonormal column vectors, and $\boldsymbol{y}_i \in \mathbb{R}^d$ is simply the vector of new coordinates of $\boldsymbol{x}_i$ in the subspace. Notice that there is some redundancy in the above representation due to the arbitrariness in the choice of $\boldsymbol{x}_0$ in the subspace. More precisely, for any $\boldsymbol{y}_0 \in \mathbb{R}^d$, we can re-represent $\boldsymbol{x}_i$ as $\boldsymbol{x}_i = (\boldsymbol{x}_0 + U_d \boldsymbol{y}_0) + U_d(\boldsymbol{y}_i - \boldsymbol{y}_0)$. Therefore, we need some additional constraints in order to end up with a unique solution to the problem of finding an affine subspace to fir the data. A common constraint is to impose that the mean of $\boldsymbol{y}_i$ is zero:[1]

$$\bar{\boldsymbol{y}} \doteq \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{y}_i = 0. \tag{2.2}$$

In general the given points are imperfect and have noise. We define the "optimal" affine subspace to be the one that minimizes the sum of squared error between $\boldsymbol{x}_i$ and its projection on the subspace, i.e.,

$$\min_{\boldsymbol{x}_0, U_d, \{\boldsymbol{y}_i\}} \sum_{i=1}^{N} \left\| \boldsymbol{x}_i - \boldsymbol{x}_0 - U_d \boldsymbol{y}_i \right\|^2, \quad \text{s.t. } U_d^T U_d = I \text{ and } \bar{\boldsymbol{y}} = 0. \tag{2.3}$$

Differentiating this function with respect to $\boldsymbol{x}_0$ and $\boldsymbol{y}_i$ (assuming U_d is fixed) and setting the derivatives to be zero,[2] we obtain the relations:

$$\hat{\boldsymbol{x}}_0 = \bar{\boldsymbol{x}} \doteq \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i; \quad \hat{\boldsymbol{y}}_i = U_d^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}). \tag{2.4}$$

The vector $\hat{\boldsymbol{y}}_i \in \mathbb{R}^d$ is simply the coordinates of the projection of $\boldsymbol{x}_i \in \mathbb{R}^D$ in the subspace S. We may call such $\hat{\boldsymbol{y}}$ the "geometric principal components" of $\boldsymbol{x}$.[3]

Then the original objective becomes one of finding an orthogonal matrix $U_d \in \mathbb{R}^{D \times d}$ that minimizes

$$\min_{U_d} \sum_{i=1}^{N} \left\| (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) - U_d U_d^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right\|^2. \tag{2.5}$$

---

[1] In the statistical setting, $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ will be samples of two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Then this constraint is equivalent to setting their means to be zero.

[2] which are the necessary conditions for the minima.

[3] As we will soon see in the next section, it coincides with the traditional principal components defined in a statistical sense.

Note that this is a restatement of the original problem with the mean $\bar{\boldsymbol{x}}$ subtracted from each of the sample points. Therefore, from now on, we will consider only the case in which the data points have zero mean. If not, simply subtract the mean from each point and the solution for U_d remains the same. The following theorem gives a constructive solution to the optimal solution $\hat{U}_d$.

Theorem 2.1 (PCA via SVD). *Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N] \in \mathbb{R}^{D\times N}$ be the matrix formed by stacking the (zero-mean) data points as its column vectors. Let $\boldsymbol{X} = U\Sigma V^T$ be the singular value decomposition (SVD) of the matrix $\boldsymbol{X}$. Then for any given $d < D$, a solution to PCA, $\hat{U}_d$ is exactly the first d columns of U; and $\hat{\boldsymbol{y}}_i$ is the ith column of the top $d \times N$ submatrix $\Sigma_d V_d^T$ of the matrix ΣV^T.*

Proof. Note that the problem

$$\min_{U_d} \sum_{i=1}^{N} \left\| \boldsymbol{x}_i - U_d U_d^T \boldsymbol{x}_i \right\|^2 \tag{2.6}$$

is equivalent to

$$\begin{aligned} & \min_{U_d} \sum_{i=1}^{N} \text{trace}\left[\left(\boldsymbol{x}_i - U_d U_d^T \boldsymbol{x}_i\right)\left(\boldsymbol{x}_i - U_d U_d^T \boldsymbol{x}_i\right)^T \right] \\ \Leftrightarrow \quad & \min_{U_d} \text{trace}\left[(I - U_d U_d^T)\boldsymbol{X}\boldsymbol{X}^T \right], \end{aligned}$$

where, for the second equivalence, we use the facts $\text{trace}(AB) = \text{trace}(BA)$, $U_d U_d^T U_d U_d^T = U_d U_d^T$, and $\boldsymbol{X}\boldsymbol{X}^T = \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T$ to simplify the expression. Substitute $\boldsymbol{X} = U\Sigma V^T$ into the above expression, the problem becomes

$$\min_{U_d} \text{trace}\left[(I - U^T U_d U_d^T U)\Sigma^2 \right].$$

Let $\sum_{i=1}^{D} \sigma_i^2 e_i e_i^T$ be the dyadic decomposition of the diagonal matrix Σ^2.[4] Since $U_d^T U$ is an orthogonal matrix, the above minimization is the same as

$$\begin{aligned} & \min_{U_d} \sum_{i=1}^{D} \text{trace}\left[\left(\sigma_i e_i - U^T U_d U_d^T U \sigma_i e_i\right)\left(\sigma_i e_i - U^T U_d U_d^T U \sigma_i e_i\right)^T \right] \\ \Leftrightarrow \quad & \min_{U_d} \sum_{i=1}^{D} \sigma_i^2 \left\| \left(I - U^T U_d U_d^T U\right) e_i \right\|^2. \end{aligned}$$

Because U_d is an orthogonal matrix of rank d so is $U_d^T U$ so that $I - U^T U_d U_d^T U$ is an idempotent matrix of rank $D - d$, so that the D terms $\left\| \left(I - U^T U_d U_d^T U\right) e_i \right\|^2$ always sum up to a constant $D - d$, and $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \sigma_D^2$ are ordered. Therefore, the minimum is achieved when the d terms associated with the higher weights $\sigma_1^2, \dots, \sigma_d^2$ become zero. This happens only when $\hat{U}_d$ consists of the first d columns of U. The rest of the theorem then easily follows.

[4] Here $e_i \in \mathbb{R}^D$ is the standard ith base vector of $\mathbb{R}^D$, i.e., its ith entry is 1 and others are 0.

When there are repeated singular values with $\sigma_d = \sigma_{d+1}$, there is a loss of uniqueness of the solution corresponding to the principal components. □

According to the theorem, the SVD gives an optimal solution to the PCA problem. The resulting matrix $\hat{U}_d$ (together with the mean $\bar{\boldsymbol{x}}$ if the data is not zero-mean) provides a geometric description of the dominant subspace structure for all the points;[5] and the columns of the matrix $\Sigma_d V_d^T = [\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_N] \in \mathbb{R}^{d \times N}$, i.e., the principal components, give a more compact representation for the points $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}$, as d is typically much smaller than D.

2.1.2 A Statistical View of PCA

Historically PCA was first formulated in a statistical setting: to estimate the principal components of a multivariate random variable $\boldsymbol{x}$ from given sample points $\{\boldsymbol{x}_i\}$ [Hotelling, 1933]. For a multivariate random variable $\boldsymbol{x} \in \mathbb{R}^D$ and any $d < D$, the d "principal components" are defined to be d *uncorrelated* linear components of $\boldsymbol{x}$:

$$y_i = u_i^T \boldsymbol{x} \ \in \mathbb{R}, \quad i = 1, \ldots, d \tag{2.7}$$

for some $u_i \in \mathbb{R}^D$ such that the variance of y_i is maximized subject to

$$u_i^T u_i = 1, \quad \text{Var}(y_1) \geq \text{Var}(y_2) \geq \cdots \geq \text{Var}(y_d).$$

For example, to find the first principal component, we seek a vector $u_1^* \in \mathbb{R}^D$ such that

$$u_1^* = \arg \max_{u_1 \in \mathbb{R}^D} \text{Var}(u_1^T \boldsymbol{x}), \quad \text{s.t.} \quad u_1^T u_1 = 1. \tag{2.8}$$

Without loss of generality, we will, in what follows assume $\boldsymbol{x}$ has zero-mean.

Theorem 2.2 (Principal Components of a Random Variable). *The first d principal components of a multivariate random variable $\boldsymbol{x}$ are given by the d leading eigenvectors of its covariance matrix $\Sigma_{\boldsymbol{x}} \doteq E[\boldsymbol{x}\boldsymbol{x}^T]$.*

Proof. Notice that for any $u \in \mathbb{R}^D$,

$$\text{Var}(u^T \boldsymbol{x}) = E[(u^T \boldsymbol{x})^2] = E[u^T \boldsymbol{x}\boldsymbol{x}^T u] = u^T \Sigma_{\boldsymbol{x}} u.$$

Then to find the first principal compoent, the above minimization (2.8) is equivalent to

$$\max_{u_1 \in \mathbb{R}^D} u_1^T \Sigma_{\boldsymbol{x}} u_1, \quad \text{s.t.} \quad u_1^T u_1 = 1. \tag{2.9}$$

Solving the above constrained minimization problem using s Lagrange multiplier method, we obtain the necessary condition for u_1 to be an extrema:

$$\Sigma_{\boldsymbol{x}} u_1 = \lambda u_1 \tag{2.10}$$

[5] From a statistical standpoint, the column vectors of U_d give the directions in which the data X has the largest variance, hence the name "principal components."

for some Lagrange multiplier $\lambda \in \mathbb{R}$, and the the associated extremum value is $u_1^T \Sigma_{\boldsymbol{x}} u_1 = \lambda$. Obviously, the optimal solution u_1^* is exactly the eigenvector associated with the largest eigenvalue of $\Sigma_{\boldsymbol{x}}$.

To find the remaining principal components, since $u_1^T \boldsymbol{x}$ and $u_i^T \boldsymbol{x}$ ($i > 1$) need to be uncorrelated, we have

$$E[(u_1^T \boldsymbol{x})(u_i^T \boldsymbol{x})] = E[u_1^T \boldsymbol{x} \boldsymbol{x}^T u_i] = u_1^T \Sigma_{\boldsymbol{x}} u_i = \lambda_1 u_1^T u_i = 0.$$

That is, $u_2, \ldots, u_d$ are all orthogonal to u_1. Following the proof for the optimality of u_1, u_2 is then the leading eigenvector of $\Sigma_{\boldsymbol{x}}$ restricted to the orthogonal complement of u_1.[6] Overall, u_2 is the second leading eigenvector of $\Sigma_{\boldsymbol{x}}$. Inductively, one can show for the rest of the principal components. □

Normally, we do not know $\Sigma_{\boldsymbol{x}}$ and can only estimate it from the given N samples $\boldsymbol{x}_i$. It is known from statistics that

$$\hat{\Sigma}_{\boldsymbol{x}} \doteq \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T = \frac{1}{N} \boldsymbol{X} \boldsymbol{X}^T \tag{2.11}$$

is an asymptotically unbiased estimate of the covariance matrix $\Sigma_{\boldsymbol{x}}$. The eigenvectors of $\hat{\Sigma}_{\boldsymbol{x}}$, or equivalently those of $\boldsymbol{X}\boldsymbol{X}^T$, lead to the "sample principal components":

$$\hat{y}_i = \hat{u}_i^T \boldsymbol{x}, \quad \text{s.t.} \quad \hat{\Sigma}_{\boldsymbol{x}} \hat{u}_i = \lambda \hat{u}_i \text{ and } \hat{u}_i^T \hat{u}_i = 1. \tag{2.12}$$

One can show that, if $\boldsymbol{x}$ is Gaussian, then every eigenvector u of $\hat{\Sigma}_{\boldsymbol{x}}$ is an asymptotically unbiased estimate for the corresponding eigenvector of $\Sigma_{\boldsymbol{x}}$ [Jollife, 1986].

Theorem 2.3 (Equivalence of Geometric and Sample Principal Components). *Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}$ be the data matrix (with $\bar{\boldsymbol{x}} = 0$). The vectors $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_d \in \mathbb{R}^D$ associated with the d sample principal components for $\boldsymbol{X}$ are exactly the columns of the matrix $\hat{U}_d \in \mathbb{R}^{D \times d}$ that minimizes the least-squares error (2.6).*

Proof. The proof is simple. Notice that if $\boldsymbol{X}$ has the singular value decomposition $\boldsymbol{X} = U \Sigma V^T$, then $\boldsymbol{X}\boldsymbol{X}^T = U \Sigma^2 U^T$ is the eigenvalue decomposition of $\boldsymbol{X}$. If Σ is ordered, then the first d columns of U are exactly the leading d eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$, which give the d sample principal components. □

Therefore, both the geometric and statistical formulation of PCA lead to exactly the same solutions/estimates of the principal components. The geometric formulation allows us to apply PCA to data even if the statistical nature of the data is unclear; the statistical formulation allows to quantitatively evaluate the quality of the estimates. For instance, for Gaussian random variables, one can derive explicit formulae for the mean and covariance of the estimated principal components. For

[6] The reason for this is that both u_1 and its orthogonal complement $u_1^\perp$ are invariant subspaces of $\Sigma_{\boldsymbol{x}}$.

a more thorough analysis of the statistical properties of PCA, we refer the reader to the classical book [Jollife, 1986].

2.1.3 Determining the Number of Principal Components

Notice that SVD does not only give a solution to PCA for a particular d, but also the solutions to all $d = 1, 2, \ldots, D$. This has an important side-benefit: if the dimension d is *not* known or specified a priori, one may have to look at the entire spectrum of solutions to decide on the "best" estimate $\hat{d}$ for the dimension and hence the subspace S for the given data.

As we have discussed in the introduction of the book, the conventional wisdom is to strike a good balance between the complexity of the chosen model and the data fidelity (to the model). The dimension d of the subspace S can be viewed as a natural measure of the complexity of the model; and the sum of squares of the remaining singular values $\sum_{i=d+1}^{D} \sigma_i^2$ is exactly the modeling error $\sum_{i=1}^{N} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2$. The leading singular value σ_{d+1}^2 of the remaining ones is a good index of the modeling error. Therefore, one can seek for a model that balances between d and σ_{d+1}^2 by minimizing an objective function of the form:

$$J_{PCA}(S) \doteq \alpha \cdot \sigma_{d+1}^2 + \beta \cdot d \tag{2.13}$$

for some proper positive weights $\alpha, \beta > 0$. In general, the ordered singular values of the data matrix $\boldsymbol{X}$ versus the dimension d of the subspace resemble a plot as in Figure 2.1. In the statistics literature, this is known as the "Scree graph." We

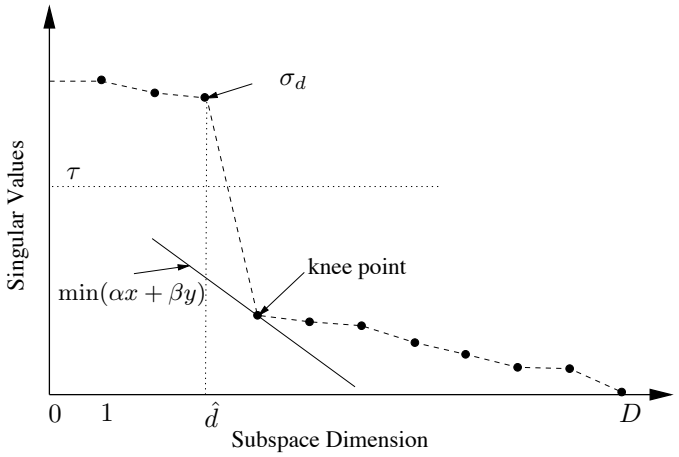

Figure 2.1. Singular value as a function of the dimension of the subspace.

will see a significant drop in the singular value right after the "correct" dimension $\hat{d}$, which is sometimes called the "knee" or "elbow" point of the plot. Obviously, such a point is a stable minimum as it optimizes the above objective function (2.13) for a range of values for α and β.