EECS 553: Machine Learning (ECE) **University of Michigan**

Principle Component Analysis

Winter 2023 **Clayton Scott**

Principal component analysis (PCA) is the most common method for dimensionality reduction. We will cover two different derivations.

1 Formulation 1: Linear Approximation

The idea behind PCA is to approximate

$$\boldsymbol{x}_i \approx \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{\theta}_i$$

where $\boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{A} \in \mathcal{A}_k, \boldsymbol{\theta}_i \in \mathbb{R}^k$, and $\mathcal{A}_k$ is the set of all $d \times k$ matrices with orthonormal columns. Intuitively, $\boldsymbol{\mu}$ in an affine shift, $\boldsymbol{A} \in \mathcal{A}_k$ defines a subspace, and $\boldsymbol{\theta}_i$ provide coordinates in that subspace such that $\boldsymbol{x}_i$ is optimally approximated. See Figure 1. This provides a method of dimensionality reduction, where the original data point $\boldsymbol{x}_i$ is represented by a lower-dimensional ($k < d$) point $\boldsymbol{\theta}_i$. As we will see, the method provides a function that lets us map an arbitrary point in $\mathbb{R}^d$ to $\mathbb{R}^k$.

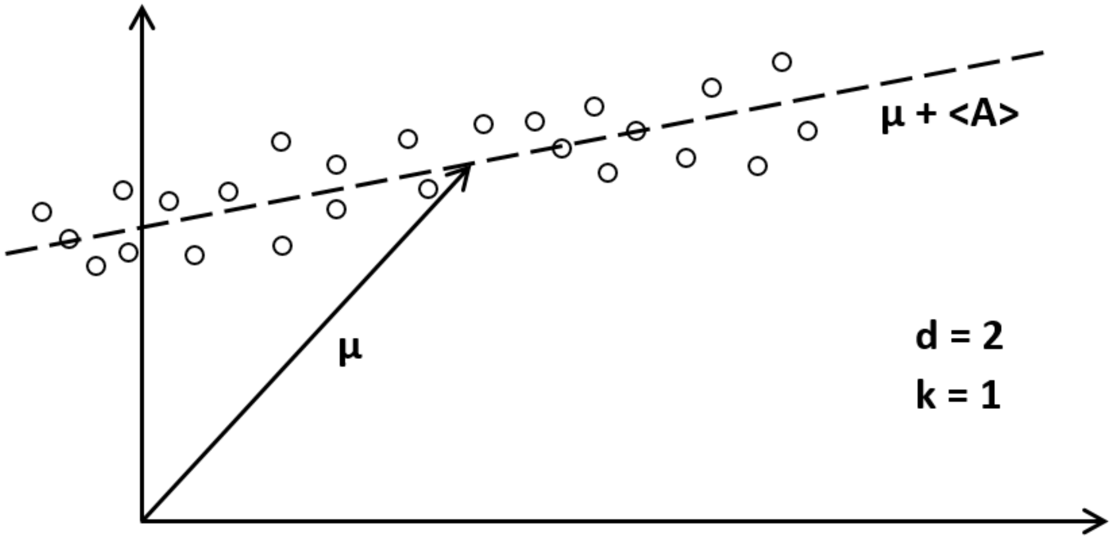

Figure 1: Optimal rank-1 approximation to a 2-d data set. Here $\boldsymbol{\mu} + \langle \boldsymbol{A} \rangle$ is a one-dimenional affine subspace, and θ_i is a scalar and determines the location along the line of the point that is closest to $\boldsymbol{x}_i$.

Our notion of optimality is the sum of squared errors, and so we select $\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ to be the solution of

$$\min_{\substack{\boldsymbol{\mu} \in \mathbb{R}^d \\ \boldsymbol{A} \in \mathcal{A}_k \\ \boldsymbol{\theta}_i \in \mathbb{R}^k}} \quad \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\theta}_i\|^2 . \tag{1}$$

PCA gives the least squares rank-k linear approximation to the data set. We will see below that the solution to the above optimization problem is given in terms of the spectral (or eigenvalue) decomposition of the sample covariance matrix

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T.$$

Note that $\boldsymbol{S}$ is symmetric and positive semi-definite. By the spectral theorem (see linear algebra review), we may write

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$$

where

$$\boldsymbol{U} = \begin{bmatrix} \\ \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_d \\ \\ \end{bmatrix} \text{ and } \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_d \end{bmatrix}$$

with $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$, $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d \geqslant 0$. Recall that $S\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j \forall j$.

We will show that *one* solution to (1) is

$$\boldsymbol{\mu} = \bar{\boldsymbol{x}}$$

$$\boldsymbol{A} = \begin{bmatrix} \\ \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_k \\ \\ \end{bmatrix}$$

$$\boldsymbol{\theta}_i = \boldsymbol{A}^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$$

We will also characterize the set of *all* solutions. Some terminology:

- The *principal component transform* is the mapping

$$\boldsymbol{x} \mapsto \boldsymbol{A}^T(\boldsymbol{x} - \bar{\boldsymbol{x}}) \in \mathbb{R}^k.$$

- The j^{th} *principal component* is the mapping

$$\boldsymbol{x} \mapsto = \boldsymbol{u}_j^T(\boldsymbol{x} - \bar{\boldsymbol{x}}) \in \mathbb{R}.$$

- The j^{th} *principal eigenvector* is

$$\boldsymbol{u}_j \in \mathbb{R}^d$$

Figure 2 illustrates PCA when $d = 3, k = 2$, taken from Hastie, Tibshirani and Friedman, *The Elements of Statistical Learning*. The principal components are the new dimensions in which the dataset is represented.

2 Formulation 2: Maximum Variance Explained

PCA can also be derived by seeking the directions that explain the most variance in the data. In this section we suppose $\bar{\boldsymbol{x}} = \boldsymbol{0}$, which can be ensured by substracting the mean from every data point. We also use $\mathbb{X}$ to denote the random vector of which $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are realizations.

2.1 Sequential formulation

In the sequential version of the maximum variance formulation, we first look for the direction in which the data has maximum variance. Thus, we seek the unit vector $\boldsymbol{a}_1 \in \mathbb{R}^d$ ($\|\boldsymbol{a}_1\| = 1$) for which the variance of

$$\theta_1 = \boldsymbol{a}_1^T \mathbb{X} \tag{2}$$

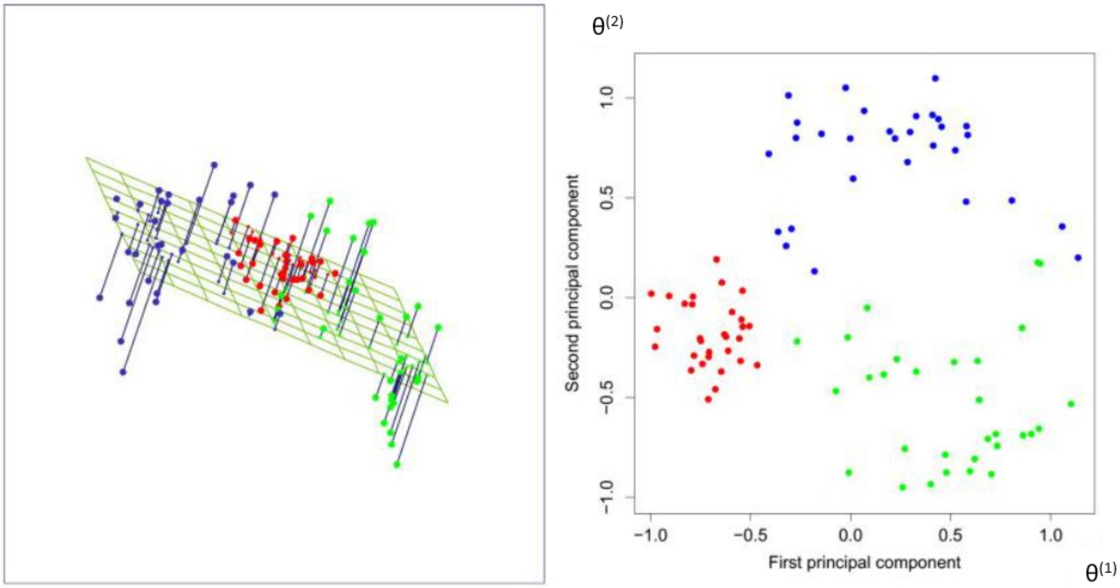

Figure 2: The best rank-two linear approximation to the half-sphere data. The right panel shows the projected points with coordinates, the first two principal components of the data.

is maximized. The sample variance of θ_1 is

$$\widehat{\mathrm{V}}(\theta_1) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{a}_1^T\boldsymbol{x}_i)^2 \tag{3}$$

(note θ_1 has zero mean since $\bar{\boldsymbol{x}} = \boldsymbol{0}$). Then

$$\begin{aligned}\widehat{\mathrm{V}}(\theta_1) &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{a}_1^T\boldsymbol{x}_i)(\boldsymbol{a}_1^T\boldsymbol{x}_i)\\ &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{a}_1^T\boldsymbol{x}_i)(\boldsymbol{x}_i^T\boldsymbol{a}_1)\\ &= \boldsymbol{a}_1^T(\frac{1}{n}\sum\boldsymbol{x}_i\boldsymbol{x}_i^T)\boldsymbol{a}_1.\end{aligned} \tag{4}$$

As we will see below, maximizing this subject to $\|\boldsymbol{a}_1\| = 1$ results in $\boldsymbol{a}_1 = \boldsymbol{u}_1$, the first principal eigenvector of $S = \frac{1}{n}\sum\boldsymbol{x}_i\boldsymbol{x}_i^T$. Note that $\boldsymbol{u}_1$ is not unique because of the sign. If $\lambda_1 = \lambda_2$, then any unit vector in the λ_1-eigenspace results in maximum variation.

Furthermore, the remaining PC's emerge if we seek additional maximum variance directions orthogonal to the first one.

Theorem 1. *Let $\theta_k = \boldsymbol{a}_k^TX$ and $\widehat{V}(\theta_k) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{a}_k^T\boldsymbol{x}_i)^2$. A vector $\boldsymbol{a}_k$ that maximizes $\widehat{V}(\theta_k)$ subject to*

- $\|\boldsymbol{a}_k\| = 1$
- $\boldsymbol{a}_k \perp \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}$

is $\boldsymbol{a}_k = \boldsymbol{u}_k$

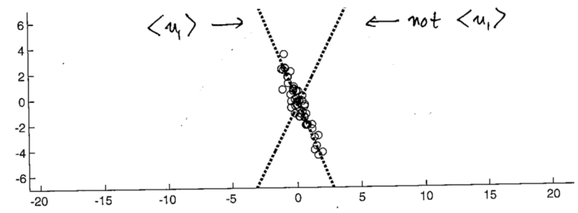

Figure 3: The direction of maximum variation, and another direction.

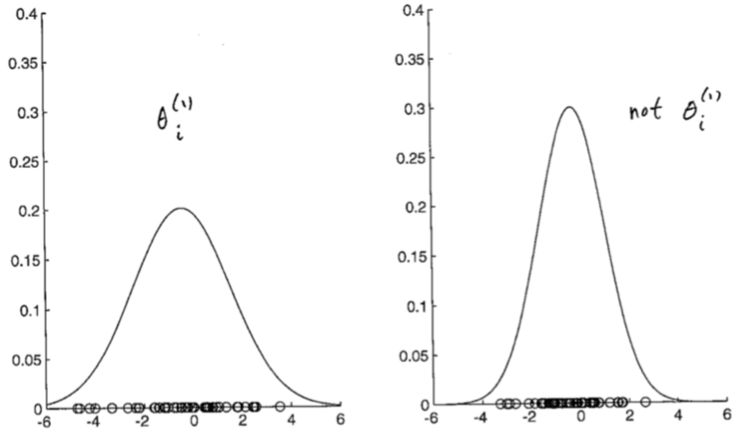

Figure 4: The first principal component of the data, and a coordinates for the suboptimal direction. The former has larger variance. The notation should be θ_{i1}, indicating the first principal component of $\boldsymbol{x}_i$.

2.2 One-step formulation

An alternative to the sequential formulation is to directly seek the matrix $\boldsymbol{A} \in \mathcal{A}_k$ such that $\boldsymbol{\theta} = \boldsymbol{A}^T\mathbb{X}$ has maximum sample variance. The sample variance of a random vector $\boldsymbol{\theta}$ is by definition $\sum_j \widehat{\mathrm{V}}(\theta_j)$, the sum of the variances of the individual components. Then we seek to solve

$$\max \sum_j \widehat{\mathrm{V}}(\boldsymbol{A}^T\mathbb{X})$$
$$\text{s.t. } \boldsymbol{A} \in \mathcal{A}_k.$$

The solution to this problem is $\boldsymbol{A} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_k]$.

3 Other Considerations

This section covers some additional facets of PCA.

3.1 Connection to the SVD

Every matrix $\boldsymbol{X}$ has a *singular value decomposition* (SVD)

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \qquad (d \times n)$$

where $\boldsymbol{U}$ is a $d \times d$ orthogonal matrix, $\boldsymbol{V}$ is a $n \times n$ orthogonal matrix, and

$$\boldsymbol{\Sigma} = \begin{cases} \begin{bmatrix} \sigma_1 & & & & & \\ & \sigma_2 & & & & \\ & & \ddots & & & \\ & & & \sigma_d & & \end{bmatrix} & \text{if } n \geq d \\ \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \\ & & & \end{bmatrix} & \text{if } n < d \end{cases}$$

The columns of $\boldsymbol{U}$ are the called *left singular vectors*, the columns of $\boldsymbol{V}$ are the *right singular vectors*, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ are called the *singular values.*

Viewing $\boldsymbol{X}$ as the data matrix, it can be shown that eigenvalues of $\boldsymbol{S}$ are related to the singular values of $\boldsymbol{X}$, and that the corresponding eigenvectors (i.e., the principal eigenvectors) are the left singular vectors of $\boldsymbol{X}$. To see this, observe that the eigenvalue decomposition of $\boldsymbol{X}\boldsymbol{X}^T$ is

$$\begin{aligned} \boldsymbol{X}\boldsymbol{X}^T \quad &= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T \\ &= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T \\ &= \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T \end{aligned}$$

where

$$\boldsymbol{\Lambda} = \begin{cases} \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_d^2 \end{bmatrix} & \text{if } n \geq d \\ \begin{bmatrix} \sigma_1^2 & & & & & \\ & \ddots & & & & \\ & & \sigma_n^2 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} & \text{if } n < d \end{cases}$$

Therefore, the left singular vector of $\boldsymbol{X}$ are the eigenvectors of $\boldsymbol{S} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T$ and the eigenvalues of $\boldsymbol{S}$ are given by

$$\lambda_i = \begin{cases} \frac{1}{n}\sigma_i^2 & i \leq \min(n, d), \\ 0 & n < i \leq d. \end{cases}$$

Therefore, PCA can be computed by simply finding the SVD of the centered data matrix.

3.2 Selecting k

As an exercise you are asked to show that the optimal objective function value is

$$\min_{\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\theta}_i} \sum \|\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\theta}_i\|^2 = n(\lambda_{k+1} + \cdots + \lambda_d).$$

When $k = 0$, this specializes to

$$\min_{\boldsymbol{\mu}} \sum \|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2 = n(\lambda_1 + \cdots + \lambda_d)$$

which we call the *total variation* of the data. One heuristic for choosing k is to select the smallest k such that

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_d} \geq 0.95$$

In words, we select the smallest k such that the optimal rank-k approximation explains 95% of the variation in the data. This threshold is arbitrary.

3.3 Preprocessing

It is common to center and scale data before applying PCA. This avoids problems that might arise if different features have different units. One common way to do this is called *standardization*. Letting x_{ij} denote the j^{th} entry of $\boldsymbol{x}_i$, a dataset is standardized by executing the following for $j = 1, \ldots, d$

$$\begin{aligned}
\bar{x}_j &\leftarrow \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\
x_{ij} &\leftarrow x_{ij} - \bar{x}_j \qquad \forall i \\
\sigma_j^2 &\leftarrow \frac{1}{n}\sum_{i=1}^{n} (x_{ij})^2 \\
x_{ij} &\leftarrow x_{ij}/\sigma_j \qquad \forall i.
\end{aligned}$$

The reason for preprocessing is that PCA finds directions of maximum variation, and we don't want it to be influenced by arbitrary differences in scale along different coordinates.

3.4 Limitations of linearity

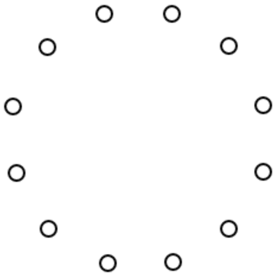

Figure 5: PCA (k=1) will fail to capture the circular structure.

PCA is a linear method, and therefore cannot capture nonlinear structure. See Figure 5. However, PCA can be kernelized (known as "kernel PCA") which yields a method for nonlinear dimensionality reduction.

4 Proof of Optimality

In this section we prove the optimality claims made above. All proofs reduce to solving a fundamental optimization problem that is quite interesting because it is nonconvex yet has a closed-form solution. This problem can be expressed in two equivalent ways. The first is

$$\min_{\boldsymbol{A}\in\mathcal{A}_k} \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{A}^T\boldsymbol{X}\|_F \tag{PCA1}$$

where $\boldsymbol{X}$ is the $d \times n$ data matrix (column mean $= \mathbf{0}$), $\mathcal{A}_k$ is the set of $d \times k$ matrices with orthonormal columns, and $\|\cdot\|_F$ is the Frobenius norm. The Frobenius norm of a matrix $\boldsymbol{B} = (b_{ij}) \in \mathbb{R}^{m\times n}$ is $\|\boldsymbol{B}\|_F := \sqrt{\sum_{i=1}^m \sum_{j=1}^n b_{ij}^2}$. This can be viewed as the standard Euclidean norm after vectorizing $\boldsymbol{B}$. Note that $\boldsymbol{A}\boldsymbol{A}^T$ is a rank-k projection matrix (see linear algebra review). Therefore (PCA1) seeks the optimal rank-k approximation of the data matrix.

The second way to express the fundamental problem is

$$\max_{\boldsymbol{A}\in\mathcal{A}_k} \operatorname{tr}(\boldsymbol{A}^T\boldsymbol{S}\boldsymbol{A}) \tag{PCA2}$$

where tr is the trace, and $\boldsymbol{S}$ is the sample covariance matrix.

To see that (PCA1) and (PCA2) are equivalent, we need the following facts about the trace operator. Recall that the trace of a square matrix is defined to be the sum of the diagonal entries.

- linearity: $tr(\boldsymbol{C}+\boldsymbol{D}) = tr(\boldsymbol{C}) + tr(\boldsymbol{D})$ for any two square matrices $\boldsymbol{C}$ and $\boldsymbol{D}$
- invariance to cyclic permutations: $tr(\boldsymbol{C}\boldsymbol{D}) = tr(\boldsymbol{D}\boldsymbol{C})$ for any matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ (not necessarily square) such that both $\boldsymbol{C}\boldsymbol{D}$ and $\boldsymbol{D}\boldsymbol{C}$ are defined.
- the trace of a symmetric matrix is the sum of its eigenvalues
- for any matrix $\boldsymbol{C}$, $\|\boldsymbol{C}\|_F^2 = tr(\boldsymbol{C}^T\boldsymbol{C})$

Proof of these properties is left as an exercise.

For $\boldsymbol{A} \in \mathcal{A}_k$, let $\boldsymbol{P} = \boldsymbol{A}\boldsymbol{A}^T$. Now observe

$$\begin{aligned}
\|\boldsymbol{X} - \boldsymbol{P}\boldsymbol{X}\|_F^2 &= tr((\boldsymbol{X} - \boldsymbol{P}\boldsymbol{X})^T(\boldsymbol{X} - \boldsymbol{P}\boldsymbol{X})) \\
&= tr(\boldsymbol{X}^T\boldsymbol{X}) - tr(\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X}) - tr(\boldsymbol{X}^T\boldsymbol{P}^T\boldsymbol{X}) + tr(\boldsymbol{X}^T\boldsymbol{P}^T\boldsymbol{P}\boldsymbol{X}) \\
&= tr(\boldsymbol{X}^T\boldsymbol{X}) - tr(\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X})
\end{aligned}$$

where we used $\boldsymbol{P} = \boldsymbol{P}^T$ and $\boldsymbol{P}^2 = \boldsymbol{P}$. Thus (PCA1) is equivalent to maximizing

$$\begin{aligned}
tr(\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X}) &= tr(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{X}) \\
&= tr(\boldsymbol{A}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{A})
\end{aligned}$$

over $\mathcal{A}_k$.

For zero-mean data, the covariance matrix may be expressed

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T,$$

This shows that (PCA1) and (PCA2) have the same solution(s).

The solution to both (PCA1) and (PCA2) is given by $\boldsymbol{A} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_k]$, the first k principal eigenvectors. We will establish this fact, which we refer to as the *fundamental PCA theorem*, in the last section. In the remainder of this section, we will apply this result to solve the various formulations of PCA above.

4.1 Proof of optimality: linear approximation formulation

We want to minimize

$$\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\theta}_i\|^2$$

with respect to $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{A} \in \mathcal{A}_k$, $\boldsymbol{\theta}_i \in \mathbb{R}^k$

Step 1: Eliminate $\boldsymbol{\theta}_i$

Suppose $\boldsymbol{A}$, $\boldsymbol{\mu}$ are fixed. We can optimize each term with respect to $\boldsymbol{\theta}_i$ individually, yielding

$$\begin{aligned}\boldsymbol{\theta}_i &= (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T(\boldsymbol{x}_i - \boldsymbol{\mu}) \\ &= \boldsymbol{A}^T(\boldsymbol{x}_i - \boldsymbol{\mu}).\end{aligned}$$

Step 2: Eliminate $\boldsymbol{\mu}$

Holding $\boldsymbol{A}$ fixed, we wish to minimize

$$\begin{aligned}&\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{A}^T(\boldsymbol{x}_i - \boldsymbol{\mu})\|^2 \\ &= \sum_{i=1}^{n} \|(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T)(\boldsymbol{x}_i - \boldsymbol{\mu})\|^2 \\ &= \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})^T(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T)^T(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T)(\boldsymbol{x}_i - \boldsymbol{\mu})\end{aligned}$$

Note that $\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T$ is a projection matrix onto the orthogonal complement of $\langle\boldsymbol{A}\rangle$. Therefore

$$\begin{aligned}(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T)^T(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T) &= (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T)(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T) \\ &= \boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T.\end{aligned}$$

Note that $\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T$, being a projection matrix, is PSD, and therefore

$$\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})^T(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T)(\boldsymbol{x}_i - \boldsymbol{\mu})$$

is a convex function of $\boldsymbol{\mu}$. The gradient (wrt $\boldsymbol{\mu}$) of this function is

$$\begin{aligned}2\sum_{i=1}^{n} (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T)(\boldsymbol{x}_i - \boldsymbol{\mu}) &= 2(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T)\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu}) \\ &= 2n(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}).\end{aligned}$$

Equating the gradient to $\mathbf{0}$ leads to the equation

$$(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) = \mathbf{0}.$$

This is solved by $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$. More generally, it suffices for $\bar{\boldsymbol{x}} - \boldsymbol{\mu}$ to belong to the nullspace of $\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^T$, which is $\langle\boldsymbol{A}\rangle$. Thus any $\boldsymbol{\mu} \in \bar{\boldsymbol{x}} + \langle\boldsymbol{A}\rangle$ is a possible solution.

Step 3: Optimize $\boldsymbol{A}$

Regardless of which solution we take for $\boldsymbol{\mu}$, it remains to solve

$$\min_{\boldsymbol{A} \in \mathcal{A}_k} \sum \|\boldsymbol{x}_i - \bar{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{A}^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\|^2$$

Assume $\bar{\boldsymbol{x}} = \mathbf{0}$ (otherwise we could substitute $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \bar{\boldsymbol{x}}$; note that the sample covariance is not changed by subtracting the mean). Also recall that $\boldsymbol{A}\boldsymbol{A}^T$ represents a rank-k projection matrix. Then it remains to solve

$$\min_{\boldsymbol{A} \in \mathcal{A}_k} \sum \|\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{x}_i\|^2 \tag{5}$$

Using the data matrix

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \end{bmatrix} \qquad (d \times n) \tag{6}$$

we can restate the problem as

$$\min_{\boldsymbol{A} \in \mathcal{A}_k} \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{X}\|_F^2. \tag{7}$$

The claimed solution now follows by the fundamental PCA theorem.

4.2 Proof of optimality: maximum variance explained

First consider the one-step formulation. The variance the j^{th} component of $\boldsymbol{\theta}$ is $\frac{1}{n}\sum_{i=1}^n \boldsymbol{a}_j^T \boldsymbol{S} \boldsymbol{a}_j$. Therefore the variance of $\boldsymbol{\theta}$ is $\sum_{j=1}^k \frac{1}{n}\sum_{i=1}^n \boldsymbol{a}_j^T \boldsymbol{S}\boldsymbol{a}_j = \frac{1}{n}\operatorname{tr}(\boldsymbol{A}^T\boldsymbol{S}\boldsymbol{A})$, where $\boldsymbol{A} = [\boldsymbol{a}_1 \cdots \boldsymbol{a}_k]$. The optimality of PCA now follows immediately from the fundamental PCA theorem (PCA2 version).

The sequential case may be proved by induction. The base case ($k = 1$) is simply the fundamental PCA theorem with $k = 1$. For the inductive step, suppose it has been established that the first $k-1$ terms in the sequence are $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}$. We seek $\boldsymbol{a}_k$ that maximizes $\frac{1}{n}\sum_{i=1}^n \boldsymbol{a}_k^T \boldsymbol{S}\boldsymbol{a}_k$ among all $\boldsymbol{a}_k$ that are orthogonal to every element of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}$. We will show that the optimal $\boldsymbol{a}_k$ is $\boldsymbol{u}_k$. Denote $\boldsymbol{U}_k = [\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_k]$ and $\boldsymbol{U}_{k-1} = [\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_{k-1}]$. From the single-step formulation, we know that $\boldsymbol{U}_k$ and maximizes $\widehat{\mathrm{V}}(\boldsymbol{A}_k^T \boldsymbol{X})$ over $\boldsymbol{A}_k \in \mathcal{A}_k$. Now consider any $\boldsymbol{a}'_k$ that is orthogonal to every element of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}$. Consider the matrix $\boldsymbol{A}'_k = [\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_{k-1} \ \boldsymbol{a}'_k] \in \mathcal{A}_k$. Then

$$\begin{aligned}
\widehat{\mathrm{V}}(\boldsymbol{u}_k^T \boldsymbol{X}) - \widehat{\mathrm{V}}((\boldsymbol{a}'_k)^T \boldsymbol{X}) &= \widehat{\mathrm{V}}(\boldsymbol{U}_{k-1}\boldsymbol{X}) + \widehat{\mathrm{V}}(\boldsymbol{u}_k^T \boldsymbol{X}) - \widehat{\mathrm{V}}(\boldsymbol{U}_{k-1}\boldsymbol{X}) - \widehat{\mathrm{V}}((\boldsymbol{a}'_k)^T \boldsymbol{X}) \\
&= \widehat{\mathrm{V}}(\boldsymbol{U}_k \boldsymbol{X}) - \widehat{\mathrm{V}}(\boldsymbol{A}'_k \boldsymbol{X}) \\
&\geq 0,
\end{aligned}$$

where the last step follows by optimality of $\boldsymbol{U}_k$. Therefore, $\boldsymbol{u}_k$ has maximal variance among all unit vectors that are orthogonal to every element of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}$.

5 Proof of Fundamental PCA Theorem

We will prove the fundamental PCA theorem in two ways. The first is associated to (PCA1), the second to (PCA2).

5.1 SVD and the Eckart-Young Theorem

The SVD arises in the following theorem due to Eckart and Young.

Theorem 2. *Let $\boldsymbol{X}$ be an arbitrary $d \times n$ matrix, with SVD gived by $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$. The solution to*

$$\min_{\boldsymbol{Z} \in \mathbb{R}^{d \times n}, \operatorname{rank}(\boldsymbol{Z}) = k} \|\boldsymbol{X} - \boldsymbol{Z}\|_F \tag{SVD}$$

is $\boldsymbol{Z}_k = \boldsymbol{U}\boldsymbol{\Sigma}_k\boldsymbol{V}^T$, where $\boldsymbol{\Sigma}_k$ is $\boldsymbol{\Sigma}$ with $\sigma_{k+1}, \sigma_{k+2}, \ldots$ set to zero.

This result and its generalizations have been used to study image compression, approximation of linear dynamical systems, and various other problems in applied mathematics.

To see the connection to (PCA1), notice that if $\boldsymbol{A} \in \mathcal{A}_k$, then $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{A}^T\boldsymbol{X}$ is $d \times k$ and has a rank of k. Therefore

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{d\times n},\operatorname{rank}(\boldsymbol{Z})=k} \|\boldsymbol{X} - \boldsymbol{Z}\|_F \leq \min_{\boldsymbol{A}\in\mathcal{A}_k} \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{A}^T\boldsymbol{X}\|_F. \tag{8}$$

To prove the fundamental PCA theorem, we will show that the $\boldsymbol{Z}$ solving (SVD) has the form $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{A}^T\boldsymbol{X}$, where the columns of $\boldsymbol{A}$ are the first k principal eigenvectors. By (8), such an $\boldsymbol{A}$ must achieve the solution of (PCA1).

By the Eckart-Young theorem, the solution of (SVD) is $\boldsymbol{Z}_k = \boldsymbol{U}\boldsymbol{\Sigma}_k\boldsymbol{V}^T$, where

$$\boldsymbol{\Sigma}_k = \boldsymbol{I}_{k,d}\boldsymbol{\Sigma},$$

and

$$\boldsymbol{I}_{k,d} = \begin{bmatrix} 1 & & & & & & & \\ & 1 & & & & & & \\ & & \ddots & & & & & \\ & & & 1 & & & & \\ & & & & 0 & & & \\ & & & & & 0 & & \\ & & & & & & \ddots & \\ & & & & & & & 0 \end{bmatrix}$$

with k 1's, in a $d \times d$ matrix. Then

$$\begin{aligned} \boldsymbol{Z}_k \quad &= \boldsymbol{U}\boldsymbol{\Sigma}_k\boldsymbol{V}^T \\ &= \boldsymbol{U}\left(\boldsymbol{I}_{k,d}\boldsymbol{\Sigma}\right)\boldsymbol{V}^T \\ &= \boldsymbol{U}\left(\boldsymbol{I}_{k,d}\boldsymbol{U}^T\boldsymbol{X}\boldsymbol{V}\right)\boldsymbol{V}^T \\ &= \boldsymbol{U}_k\boldsymbol{U}_k^T\boldsymbol{X} \end{aligned}$$

where $\boldsymbol{U}_k$ contains the first k left singular vectors. Clearly $\boldsymbol{U}_k \in \mathcal{A}_k$. Therefore $\boldsymbol{U}_k$ gives a solution to (PCA1). In Section (3.1), we showed that the left singular vectors of $\boldsymbol{X}$ are the eigenvectors of $\boldsymbol{S}$. This completes the proof.

5.2 The Generalized Rayleigh Quotient

The (PCA2) version of the fundamental PCA theorem is an immediate corollary of a more general result that I will refer to as the Generalized Rayleigh Quotient theorem (this terminology is not standard) because, in the special case $k = 1$, it is equivalent to the Raleigh quotient. This result characterizes the solution of a very general type of "trace minimization problem," and is also used to derive several other unsupervised learning methods including spectral clustering and various forms of nonlinear dimensionality reduction.

Theorem 3. *Let $\boldsymbol{C}$ be a symmetric matrix with eigenvalue decomposition $\boldsymbol{C} = \boldsymbol{U}\Lambda\boldsymbol{U}^T$, where $\boldsymbol{U} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_d]$. Then a solution of*

$$\max_{\boldsymbol{A}\in\mathcal{A}_K} tr(\boldsymbol{A}^T\boldsymbol{C}\boldsymbol{A}) \tag{GRQ}$$

is $\boldsymbol{A} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_k]$.

To prove the GRQ theorem introduce the change of variables

$$\boldsymbol{w}_i = \boldsymbol{U}^T\boldsymbol{a}_i,$$

where $\boldsymbol{a}_i$ is the i^{th} column of $\boldsymbol{A}$. Denoting $\boldsymbol{W} = [\boldsymbol{w}_1 \ \cdots \ \boldsymbol{w}_k]$, we have that $\boldsymbol{A} \in \mathcal{A}_k$ iff $\boldsymbol{W} \in \mathcal{A}_k$ since $\boldsymbol{U}$ is orthogonal. In particular, $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ are orthonormal bacause

$$\boldsymbol{w}_i^T \boldsymbol{w}_j = \boldsymbol{a}_i^T \boldsymbol{U}\boldsymbol{U}^T \boldsymbol{a}_j = \boldsymbol{a}_i^T \boldsymbol{a}_j.$$

Then we need to maximize

$$\begin{aligned} tr(\boldsymbol{A}^T\boldsymbol{C}\boldsymbol{A}) \qquad &= \sum_{i=1}^{k} \boldsymbol{a}_i^T \boldsymbol{C} \boldsymbol{a}_i \\ &= \sum_{i=1}^{k} \boldsymbol{w}_i^T \boldsymbol{\Lambda} \boldsymbol{w}_i \end{aligned}$$

subject to $\boldsymbol{W} \in \mathcal{A}_k$. Now

$$\begin{aligned} \sum_{i=1}^{k} \boldsymbol{w}_i^T \boldsymbol{\Lambda} \boldsymbol{w}_i \qquad &= \sum_{i=1}^{k}\sum_{j=1}^{d} \lambda_j w_{ij}^2 \\ &= \sum_{j=1}^{d} \lambda_j \Big[\sum_{i=1}^{k} w_{ij}^2\Big] \\ &= \sum_{j=1}^{d} \lambda_j h_j \end{aligned}$$

where $h_j = \sum_{i=1}^{k} w_{ij}^2$

Lemma 1. *$0 \le h_j \le 1 \ \forall j$ and $\sum_{j=1}^{d} h_j = k$.*

Proof. The second part is easy:

$$\begin{aligned} \sum_{j=1}^{d} h_j &= \sum_{j=1}^{d} \Big(\sum_{i=1}^{k} w_{ij}^2\Big) \\ &= \sum_{i=1}^{k} \Big(\sum_{j=1}^{d} w_{ij}^2\Big) \\ &= \sum_{i=1}^{k} (1) \\ &= k. \end{aligned}$$

$h_j \ge 0$ is also obvious. To show $h_j \le 1$, let $\boldsymbol{w}_{k+1}, \ldots, \boldsymbol{w}_d$ extend $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ to an orthonormal basis of $\mathbb{R}^d$. Consider the $d \times d$ matrix

$$\boldsymbol{M} = [\boldsymbol{w}_1 \ \cdots \ \boldsymbol{w}_d].$$

We know $\boldsymbol{M}^T\boldsymbol{M} = \boldsymbol{I}$ by orthonormality. Therefore $\boldsymbol{M}^T$ is a right inverse of $\boldsymbol{M}$, and so must also be a left inverse (a property of square matrices), meaning $\boldsymbol{M}\boldsymbol{M}^T = \boldsymbol{I}$. This implies

$$h_j = \sum_{i=1}^{k} w_{ij}^2 \le \sum_{i=1}^{d} w_{ij}^2 = 1.$$

□

We need to maximize

$$\sum_{j=1}^{d} h_j \lambda_j$$

with respect to the constraints imposed by the lemma.
Since $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$, this is accomplished by

$$h_j = \begin{cases} 1 & \text{if } 1 \leq j \leq k \\ 0 & \text{otherwise} \end{cases}$$

which in turn is achieved by

$$\boldsymbol{W} = \begin{bmatrix} I_{k \times k} \\ ------- \\ 0 \end{bmatrix}.$$

Therefore $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{W} = [\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_k]$. Note that the optimal $\boldsymbol{A}$ is not unique. Indeed if

$$\boldsymbol{W} = \begin{bmatrix} \text{any set of length } k \\ \text{orthonormal vectors} \\ ---------- \\ 0 \end{bmatrix},$$

then $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{W}$also achieves the maximum.

An interesting question is when is $\langle \boldsymbol{A} \rangle$ unique. This is left as an exercise.

Exercises

1. (★) Verify the properties of the trace operator stated in Section 4.
2. (★★) Let $k \in \{0, 1, \ldots, d\}$ be arbitrary. Show that

   $$\min_{\boldsymbol{\mu}, \boldsymbol{A}, \{\boldsymbol{\theta}_i\}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\theta}_i\|^2 = n \sum_{j=k+1}^{d} \lambda_j,$$

   where $\boldsymbol{A}$ ranges over all $d \times k$ matrices with orthonormal columns. *Hint:* Use properties of the trace operator.
3. (★★) Give a simple condition involving the spectral decomposition of the sample covariance matrix that is both necessary and sufficient for the subspace $\langle \boldsymbol{A} \rangle$ in PCA to be unique.
4. (★★) Show that the principal components are uncorrelated, i.e., that the sample covariance matrix of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ is diagonal
5. (★★) Interpret the *right* singular vectors of the data matrix in the context of PCA. The identity

   $$\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \sum_{i=1}^{d} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$$

   may be helpful.