EECS 545: Machine Learning, **Lecture 16**

PCA Derivation

Fall 2018

1 Problem Statement

Define $\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ to be the solution of

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

where $\mathcal{A}_k$ is the set of $d \times k$ matrices with orthonormal columns. PCA gives the least squares rank-k linear approximation to the data set. The solution to above optimization problem is given in terms of the spectral (or eigenvalue) decomposition of the sample covariance matrix

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

In particular, write

$$S = U\Lambda U^T$$

where

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

with $U^TU = UU^T = I$, $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d \geqslant 0$. Recall that $S\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j \ \forall j$.
We will show that <u>one</u> solution to above optimization problem is

$$\begin{aligned} \boldsymbol{\mu} &= \overline{\boldsymbol{x}} \\ \boldsymbol{A} &= [\boldsymbol{u}_1 \cdots \boldsymbol{u}_k] \\ \boldsymbol{\theta}_i &= \boldsymbol{A}^T(\boldsymbol{x} - \overline{\boldsymbol{x}}). \end{aligned}$$

We will also characterized the set of <u>all</u> solutions. Some terminology:

- <u>principal component transform</u>:
$$\boldsymbol{x} \mapsto \boldsymbol{A}^T(\boldsymbol{x} - \overline{\boldsymbol{x}}) \in \mathbb{R}^k$$
- <u>j^th principal component</u>:
$$\boldsymbol{\theta}^{(j)} = \boldsymbol{u}_j^T(\boldsymbol{x} - \overline{\boldsymbol{x}}) \in \mathbb{R}$$
- <u>j^th principal eigenvector</u>:
$$\boldsymbol{u}_j \in \mathbb{R}^d$$

2 Overview

We want to minimize

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

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

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

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

$$\begin{aligned} \boldsymbol{\theta}_i &= (A^T A)^{-1} A^T (\boldsymbol{x}_i - \boldsymbol{\mu}) \\ &= A^T (\boldsymbol{x}_i - \boldsymbol{\mu}) \end{aligned} \tag{2}$$

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

Holding A fixed, we wish to minimize

$$\begin{aligned} &\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu} - AA^T(\boldsymbol{x}_i - \boldsymbol{\mu})\|^2 \\ =&\sum_{i=1}^{n} \|(I - AA^T)(\boldsymbol{x}_i - \boldsymbol{\mu})\|^2 \\ =&\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})^T (I - AA^T)^T (I - AA^T)(\boldsymbol{x}_i - \boldsymbol{\mu}) \end{aligned} \tag{3}$$

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

$$\begin{aligned} (I - AA^T)^T (I - AA^T) &= (I - AA^T)(I - AA^T) \\ &= I - AA^T \end{aligned} \tag{4}$$

Note that $I - AA^T$, being a projection matrix, is PSD, and therefore

$$\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})^T (I - AA^T)(\boldsymbol{x}_i - \boldsymbol{\mu}) \tag{5}$$

is a convex function of $\boldsymbol{\mu}$. Now

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}} &= 2 \sum_{i=1}^{n} (I - AA^T)(\boldsymbol{x}_i - \boldsymbol{\mu}) \\ &= 2(I - AA^T) \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu}) \\ &= 2n(I - AA^T)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) = \mathbf{0} \end{aligned} \tag{6}$$

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

Step 3: Optimize A

It remains to solve

$$\min_{A \in \mathcal{A}_k} \sum \|\boldsymbol{x}_i - \bar{\boldsymbol{x}} - AA^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\|^2 \tag{7}$$

Assume $\bar{\boldsymbol{x}} = 0$ which can be ensured by subtracting $\bar{\boldsymbol{x}}$ from each $\boldsymbol{x}_i$. Also note that AA^T is a rank k projection matrix. Let $\mathcal{P}_k$ denote the set of $d \times d$ rank-k projection matrices. Then it remains to solve

$$\min_{P \in \mathcal{P}_k} \sum \|\boldsymbol{x}_i - P\boldsymbol{x}_i\|^2 \tag{8}$$

Introduce the data matrix

$$X = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \end{bmatrix} \qquad (d \times n) \tag{9}$$

and for an arbitary matrix $C = [c_{ij}]$, define the Frobenius norm

$$\|C\|_F := \sqrt{\sum_i \sum_j c_{ij}^2} \tag{10}$$

Then we can restate the problem as

$$\min_{P \in \mathcal{P}_k} \quad \|X - PX\|_F^2. \tag{11}$$

This is the core optimization problem at the heart of PCA. Although it is nonconvex, it is a very special nonconvex problem and has a closed form solution given by $P = AA^T$ where A consists of the first k principal eigenvectors. The derivation of this result is quite interesting in its own right, and connects with other important topics in matrix algebra. These details are given in the next section.

3 The Fundamental PCA Problem

Above we saw that PCA reduces to the following optimization problem:

$$\min_{\boldsymbol{P} \in \boldsymbol{P}_k} \|\boldsymbol{X} - \boldsymbol{P}\boldsymbol{X}\|_F \tag{PCA}$$

where $\boldsymbol{X}$ is the $d \times n$ data matrix (column mean $= 0$), $\|\cdot\|_F$ is the Frobenius norm, and $\boldsymbol{P}_k$ is the set of rank k, $d \times d$ projection matrices. In this setting, the covariance matrix of the data is

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

(verify by comparing entries).

We will derive the solution to PCA in two different ways, by connecting (PCA) to two other problems in matrix algebra.

4 Connection to the SVD

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

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

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

- $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_{d \times d}$
- $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_{n \times n}$

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

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

Theorem 1. *Let $\boldsymbol{X}$ have rank $r = k$. The solution to*

$$\min_{\boldsymbol{Z} \in \mathbb{R}^{d \times n}, rank(\boldsymbol{Z}) = k} \|\boldsymbol{X} - \boldsymbol{Z}\|_F^2 \tag{SVD}$$

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

To see the connection to (PCA), write

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

where

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

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

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

where $\boldsymbol{U}_k$ contains the first k left singular vectors. Clearly $\boldsymbol{U}_k\boldsymbol{U}_k^T \in \boldsymbol{P}_k$. Therefore $\boldsymbol{P} = \boldsymbol{U}_k\boldsymbol{U}_k^T$ gives a solution to (PCA). It remains to show that the left singular vectors are the eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$. To see this, observe that the eigenvalue decomposition of $\boldsymbol{X}\boldsymbol{X}^T$ is

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

where

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

Therefore, the SVD of $\boldsymbol{X}$ gives the (PCA) solution:

- principal eigenvectors = left singular vectors of $\boldsymbol{X}$
- $\lambda_i = \frac{1}{n}(\ i^{th}$ singular value of $\boldsymbol{X})^2$

5 Generalized Rayleigh Quotient

We now present a second solution of (PCA).

The trace of a square matrix is the sum of the diagonal entries. It satisfies the following properties:

- 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: $tr(\boldsymbol{CD}) = tr(\boldsymbol{DC})$
- the trace of a symmetric matrix is the sum of its eigenvalues
- for any matrix $\boldsymbol{C}$, $\|\boldsymbol{C}\|_F^2 = tr(\boldsymbol{C}^T\boldsymbol{C})$

These properties are easily verified. Now observe

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

where we used $\boldsymbol{P} = \boldsymbol{P}^T$ and $\boldsymbol{P}^2 = \boldsymbol{P}$. Writing $\boldsymbol{P} = \boldsymbol{AA}^T$ where $\boldsymbol{A} \in \mathcal{A}_k$, we need to maximize

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

The derivation of PCA is concluded by the following result, which I will refer to as the Generalized Rayleigh Quotient theorem (this terminology is not standard) because, in the special case $k = 1$, it relates to the Raleigh quotient

Theorem 2. *Let $\boldsymbol{C}$ be a PSD matrix with eigenvalue decompostion $\boldsymbol{C} = \boldsymbol{U}\Lambda\boldsymbol{U}^T$, where $\boldsymbol{U} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_d]$. Then a solution of*

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

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

6 Proof of Generalized Rayleigh Quotient Theorem

Introduce the change of variable

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

We know that $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ 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

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

subject to $\boldsymbol{W} = [\boldsymbol{w}_1 \cdots \boldsymbol{w}_k] \in \mathcal{A}_k$. Now

$$\begin{aligned} \sum_{i=1}^{k} \boldsymbol{w}_i^T \boldsymbol{\Lambda} \boldsymbol{w}_i \quad &= \sum_{i=1}^{k} \sum_{j=1}^{d} (w_i^{(j)})^2 \lambda_j \\ &= \sum_{j=1}^{d} \Big[\sum_{i=1}^{k} (w_i^{(j)})^2 \Big] \lambda_j \\ &= \sum_{j=1}^{d} h_j \lambda_j \end{aligned}$$

where $h_j = \sum_{i=1}^{k} (w_i^{(j)})^2$

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

Proof. The second part is easy:

$$\begin{aligned} \sum_{j=1}^{d} h_j \quad &= \textstyle\sum_{j=1}^{d} \Big(\sum_{i=1}^{k} (w_i^{(j)})^2 \Big) \\ &= \textstyle\sum_{i=1}^{k} \Big(\sum_{j=1}^{d} (w_i^{(j)})^2 \Big) \\ &= \textstyle\sum_{i=1}^{k} (1) \\ &= k \end{aligned}$$

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

$$\boldsymbol{M} = [\boldsymbol{w}_1 \cdots \boldsymbol{w}_d] \tag{12}$$

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 right inverse (a property of square matrices), meaning $\boldsymbol{M}\boldsymbol{M}^T = \boldsymbol{I}$. This implies

$$h_j = \sum_{i=1}^{k} (w_i^{(j)})^2 \leq \sum_{i=1}^{d} (w_i^{(j)})^2 = 1$$

□

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.

The GRQ theorem can be used to derive PCA from the sequential maximum variance approach discussed in class. Furthermore, instead of a sequential definition of maximum variance, we can instead ask what $\boldsymbol{A} \in \mathcal{A}_k$ maximizes the **total variance**

$$\sum_{i=1}^{k} \boldsymbol{a}_i^T \boldsymbol{S} \boldsymbol{a}_i = tr(\boldsymbol{A}^T \boldsymbol{S} \boldsymbol{A}).$$

The GRQ theorem again tells us that the principal eigenvectors of $\boldsymbol{S}$ provide the solution. We will employ the GRQ theorem again later in the course when we discuss spectral clustering.