EECS 16B Designing Information Devices and Systems II
Summer 2020 UC Berkeley

Note 18

1 Overview

In this note, we will be taking a look at another application of the SVD called **Principal Component Analysis**. It is a commonly used technique to reduce the number of dimensions in our data.

Let's imagine we have a large dataset of noisy, redundant, and intuitively intractable data. We **know** that this data should have some inherent meaning, but we just don't know it. Each data point may consist of hundreds or thousands of attributes and Principal Component Analysis or PCA will help us find trends in this data.

To do this, we will be looking at two perspectives of PCA. The first perspective will be to find the directions of maximal variance in the data while the second is to look at how the SVD can approximate a dataset. In either case, we will look at how to transform our data into a new coordinate system which better represents the trends in our data.

2 Problem Statement

Lets say we have a collected m observations of n variable features $x_1, x_2, \ldots, x_n$. We can then aggregate our data into an $m \times n$ data matrix X where the rows of X represent a single sample $(x_1, x_2, \ldots, x_n)$. We will call the i^{th} sample $\vec{x}_i^T$ where $\vec{x}_i$ is a vector in $\mathbb{R}^n$.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} = \begin{bmatrix} \leftarrow \vec{x}_1^T \rightarrow \\ \leftarrow \vec{x}_2^T \rightarrow \\ \vdots \\ \leftarrow \vec{x}_m^T \rightarrow \end{bmatrix} \tag{1}$$

We would like to find a new basis $\{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k\}$ with $k < n$, that can approximate the n features we currently have. This new basis spans some k dimensional subspace of $\mathbb{R}^n$ and we can project all of our datapoints $\vec{x}_i^T$ onto this subspace to approximate the features. In the next sections, we look at two different persepctives on how this basis is formed and can approximate our datapoints.

3 Variance Maximization Perspective

The first perspective we will be taking a look at is that of variance maximization. Given a dataset, certain variables will be more correlated than others. However, how could we compute the direction which the correlation between our variables?

3.1 Preliminary Notation

To start off, let's look at the empirical variance of a random variable X with m samples $\begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix}$. Recall from CS70 that we can define the mean μ and variance $\text{Var}(X)$ of this random variable as follows [1]

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{Var}(X) = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)^2 \tag{2}$$

Now let us take a look at the vector $X\vec{w}$ where $\vec{w}$ is a unit vector of some arbitrary direction in $\mathbb{R}^n$.

$$X\vec{w} = \begin{bmatrix} \leftarrow \vec{x}_1^T \rightarrow \\ \leftarrow \vec{x}_2^T \rightarrow \\ \vdots \\ \leftarrow \vec{x}_m^T \rightarrow \end{bmatrix} \vec{w} = \begin{bmatrix} \vec{x}_1^T\vec{w} \\ \vec{x}_2^T\vec{w} \\ \vdots \\ \vec{x}_m^T\vec{w} \end{bmatrix}$$

Since $\vec{w}$ is a unit vector, we can in fact show that $\vec{x}_i^T\vec{w}$ is the weight of the projection of the i^{th} datapoint onto the vector $\vec{w}$.

$$\text{proj}_{\vec{w}}\vec{x}_i = \langle \vec{x}_i, \vec{w}\rangle \vec{w} = (\vec{x}_i^T\vec{w})\vec{w}$$

Therefore we conclude that the vector $X\vec{w}$ represents the projection of each individual datapoint onto a unit direction $\vec{w}$. Now before we look at the variance of these projections, let's create a new matrix A where we subtract the mean of each column. We do this so that the origin $\vec{0}$ represents the center of our data.

$$A = X - \frac{1}{m}\vec{1}\vec{1}^T X \tag{3}$$

As an example, if we have the matrix X, the demeaned matrix A is as follows

$$X = \begin{bmatrix} 1 & 2 \\ -1 & 3 \\ 3 & 4 \end{bmatrix} \implies A = \begin{bmatrix} 1 & 2 \\ -1 & 3 \\ 3 & 4 \end{bmatrix} - \begin{bmatrix} 1 & 3 \\ 1 & 3 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ -2 & 0 \\ 2 & 3 \end{bmatrix} \tag{4}$$

Note how the columns of X now sum to 0. As a quick lemma, we can show that if X has columns that sum to 0, then the entries of $A\vec{w}$ will also sum to 0. We won't show this here, but we leave it as an exercise.

3.2 Variance Optimization

Now that we have defined some preliminary notation, let us look at the variance of $A\vec{w}$ which represents the projection of each datapoint. Since the matrix A has zero mean, we know that $A\vec{w}$ has zero mean. Therefore, we can write out the formula for variance as

$$\text{Var}(A\vec{w}) = \frac{1}{m}\sum_{i=1}^{m}(\vec{x}_i^T\vec{w})^2 = \frac{1}{m}\|A\vec{w}\|^2 \tag{5}$$

The goal behind PCA is to find a new basis which best represents our data. One way to think about this is to maximize the variance of $A\vec{w}$ over all unit vectors $\vec{w} \in \mathbb{R}^n$. We can phrase this as the following optimization problem

$$\max_{\|\vec{w}\|=1} \text{Var}(A\vec{w}) \implies \max_{\|\vec{w}\|=1} \|A\vec{w}\|^2 \tag{6}$$

We can remove the $\frac{1}{m}$ term since it does not affect our goal of searching for the optimal $\vec{w}$.

[1] Don't worry if you haven't taken CS70 yet. We won't be talking much more about mean and variance.

3.2.1 Spectral Optimization

So how can we solve the following optimization problem stated above? One way to do this is by looking at the SVD of A.

$$\max_{\|\vec{w}\|=1} \left\| U\Sigma V^T \vec{w} \right\|^2$$

The matrix U has orthonormal columns, so it will not change the length of the vector $\Sigma V^T \vec{w}$. We can view $V^T \vec{w}$ as another rotation of the vector $\vec{w}$ into a new basis represented by the columns of V. Searching over all vectors $\vec{w}$ with norm 1 is equivalent to searching over all $\vec{z} = V^T \vec{w}$ with norm 1.

Therefore, we can rephrase our optimization problem as

$$\max_{\|\vec{z}\|=1} \|\Sigma \vec{z}\|^2 \tag{7}$$

It follows that the optimal solution is $\vec{z} = \vec{e}_1$ since the Σ values are ordered from largest to smallest. Changing coordinates back to the standard basis, $\vec{w} = V\vec{e}_1 = \vec{v}_1$.

3.2.2 Spectral Norm

As an aside, let us take a look at the **Spectral Norm** of a matrix and see how it is related to the optimziation problem above. The spectral norm of a matrix A is denoted as $\|A\|_2$ and can be thought of as the maximum factor A can scale the norm of a vector $\vec{x}$.

$$\|A\|_2 = \max_{\vec{x} \neq \vec{0}} \frac{\|A\vec{x}\|}{\|\vec{x}\|} = \sigma_1 \tag{8}$$

Note that this norm is defined over the vector space of $n \times n$ matrices. Here a vector is an $n \times n$ matrix A. In either case, the vector that maximizes the norm $A\vec{x}$ is $\vec{v}_1$ or the eigenvector of largest eigenvalue of $A^T A$. The spectral norm $\vec{A}_2$ will always be equal to the largest singular value of the matrix A. Try to verify that all of the properties of norms do indeed hold for the spectral norm of a matrix!

3.3 Principal Components

We have solved our variance maximization problem to find our first vector $\vec{v}_1$ which turned out to be the eigenvector of largest eigenvalue of $A^T A$. Now how can we pick our remaining vectors $\{\vec{v}_2, \ldots, \vec{v}_k\}$?

We will continue to look at vectors that maximize the variance of our datapoints. Since we have already found the direction of maximal variance, we will now try to find the maximum variance across all directions orthogonal to the vector $\vec{v}_1$.

As an optimization problem, we can phrase this as the following

$$\max_{\|\vec{w}\|=1} \text{Var}(A\vec{w}) \text{ subject to } \vec{w}^T \vec{v}_1 = 0 \tag{9}$$

The solution to this problem is $\vec{w} = \vec{v}_2$.

If we were to continue doing this, it turns out that the eiegenvectors $\vec{v}_i$ of $A^T A$ form our PCA basis. These basis vectors are called **principal components** and often times, we will pick $k < n$ vectors meaning $\{\vec{v}_1, \ldots, \vec{v}_k\}$ are the principal components of our data.

4 Low Rank Approximation Perspective

An alternate way of viewing Principal Component Analysis is through a low rank approximation using the SVD. We will be using the same data matrix X, and demeaned matrix A from the previous section.

Given some dataset A with rank r, we would like to find a matrix B_k of rank k that best approximates A. Recall that the SVD can be used to make low-rank approximations! An $m \times n$ matrix A can be approximated as follows

$$A \approx A_k = \sum_{i=1}^{k} \sigma_i \vec{u}_i \vec{v}_i^T \tag{10}$$

While this may be a good rank k approximation of the matrix A, how do we know that it's best approximation?

4.1 The Size of a Matrix

We should naturally question how to measure the size of a matrix A. One way to define a norm for a matrix is through the **spectral norm.**

$$\|A\|_2 = \max_{\vec{x} \neq \vec{0}} \frac{\|A\vec{x}\|}{\|\vec{x}\|} = \sigma_1 \tag{11}$$

An alternate norm that we could define for a matrix is the **Frobenius Norm.**

$$\|A\|_F^2 = \sum_{i=1}^{r} \sigma_i^2 \tag{12}$$

While Frobenius norm may align more naturally with the true size of a matrix, we will be sticking with the spectral norm for this note.

4.2 Optimization Problem

We would like to find a matrix B_k of rank k that best approximates A. As an optimization problem, this would look like the following

$$\min_{B_k} \|A - B_k\|_2$$
$$\text{subject to } \text{Rank}(B_k) = k$$

We will show that the optimal B_k is in fact the rank k SVD approximation of A.

$$A_k = \sum_{i=1}^{k} \sigma_i \vec{u}_i \vec{v}_i^T \tag{13}$$

4.3 Proof

This is quite a difficult proof so we will break it up into multiple parts and lemmas.

4.3.1 Goal of the Proof

Let's first look at the spectral norm of $A - A_k$. Since $A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i \vec{u}_i \vec{v}_i^T$,

$$A - A_k = \sum_{i=k+1}^{n} \sigma_i \vec{u}_i \vec{v}_i^T \tag{14}$$

The spectral norm of $A - A_k$ is its largest singular value, which in this case will be σ_{k+1}. Therefore, if we can show that $\|A - B_k\|_2 \geq \sigma_{k+1}$ for all matrices B_k, then we will have shown that A_k is optimal. We now make it our goal to show that $\|A - B_k\|_2 \geq \sigma_{k+1}$.

4.3.2 Constructing a Vector

Since the spectral norm of a matrix $A - B_k$ is defined as

$$\|A - B_k\|_2 = \max_{\vec{x} \neq \vec{0}} \frac{\left\|(A - B_k)\vec{x}\right\|}{\|\vec{x}\|} = \sigma_{max} \tag{15}$$

If we consider a vector $\vec{x}$ of norm 1, then $\|A\|_2 \geq \|A\vec{x}\|$.[2]

Now using the $\vec{v}$ vectors from the SVD of A, we will define a subspace $S = \text{span}\{\vec{v}_1, \ldots, \vec{v}_{k+1}\}$. In addition, we will pick a vector $\vec{y} \in S$ with norm 1 that is also in the $\text{Nul}(B_k)$. Why we have chosen such a $\vec{y}$ will be clear by the end of the proof, but for the time being notice that if $\vec{y} \in \text{Nul}(B_k)$, then $\left\|(A - B_k)\vec{y}\right\| = \|A\vec{y}\|$.

4.3.3 Existence of $\vec{y}$.

To show that a $\vec{y} = \alpha_1 \vec{v}_1 + \ldots + \alpha_{k+1} \vec{v}_{k+1}$ must exist, we will use a dimensionality argument. The subspace S is of rank $k+1$ and $\text{Nul}(B_k)$ must be of rank $n-k$ since B_k is of rank k.

If $R = \{\vec{v}_1, \ldots, \vec{v}_{k+1}\}$ is a basis for S and $B = \{\vec{w}_1, \ldots, \vec{w}_{n-k}\}$ is a basis for $\text{Nul}(B_k)$, then the union of the two bases is a set of $n+1$ vectors in $\mathbb{R}^n$ which must be linearly dependent.

From the definition of linear dependence, we can write $\vec{w}_{n-k}$ as a linear combination of the remaining vectors:

$$\vec{w}_{n-k} = \alpha_1 \vec{v}_1 + \ldots + \alpha_{k+1} \vec{v}_{k+1} + \beta_1 \vec{w}_1 + \ldots + \beta_{n-k-1} \vec{w}_{n-k-1}.$$

Subtracting over the $\vec{w}_i$ vectors, we see that

$$\alpha_1 \vec{v}_1 + \ldots + \alpha_{k+1} \vec{v}_{k+1} = -\beta_1 \vec{w}_1 - \ldots - \beta_{n-k-1} \vec{w}_{n-k-1} + \vec{w}_{n-k}.$$

As a result, $\vec{y} = \alpha_1 \vec{v}_1 + \ldots + \alpha_{k+1} \vec{v}_{k+1}$ is a linear combination of the vectors in R so it must be in S. It is also however, a linear combination of the vectors in B so it must also be in $\text{Nul}(B_k)$.

[2] Here $\|\cdot\|_2$ refers to the Spectral Norm of a matrix while $\|\cdot\|$ refers to the norm of a vector. Make sure you understand the distinction in notation from here onward.

4.3.4 Simplifying our Goal

Our original goal was to show that $\|A - B_k\|_2 \geq \sigma_{k+1}$. However, if we can show that for some vector $\vec{y}$ with norm 1 that $\left\|(A - B_k)\vec{y}\right\| \geq \sigma_{k+1}$, then by the transitive property,

$$\|A - B_k\|_2 \geq \left\|(A - B_k)\vec{y}\right\| \geq \sigma_{k+1} \tag{16}$$

Our new updated goal is to show that $\left\|(A - B_k)\vec{y}\right\| \geq \sigma_{k+1}$.

4.3.5 Finishing the Proof

If $\vec{y} = \alpha_1\vec{v}_1 + \ldots + \alpha_{k+1}\vec{v}_{k+1}$, is a vector with norm 1, then what would the norm of $(A - B_k)\vec{y}$ look like? We know that $\vec{y}$ is in the $\text{Nul}(B_k)$ so

$$\left\|(A - B_k)\vec{y}\right\| = \|A\vec{y} - B_k\vec{y}\| = \|A\vec{y}\| \tag{17}$$

By picking a $\vec{y}$ in the $\text{Nul}(B_k)$, we have removed the influence of B_k on our matrix $A - B_k$ and it remains to compute $\|A\vec{y}\|$.

Let's start by writing out $\|\vec{y}\|$ in terms of α_i. Remember that $\vec{v}_i$ are all orthonormal.

$$\begin{aligned}\|\vec{y}\|^2 &= \langle\vec{y},\vec{y}\rangle = \langle\alpha_1\vec{v}_1 + \ldots + \alpha_{k+1}\vec{v}_{k+1}, \alpha_1\vec{v}_1 + \ldots + \alpha_{k+1}\vec{v}_{k+1}\rangle \\ &= \sum_{i=1}^{k+1}\langle\alpha_i\vec{v}_i, \alpha_i\vec{v}_i\rangle = \sum_{i=1}^{k+1}\alpha_i^2\langle\vec{v}_i,\vec{v}_i\rangle = \sum_{i=1}^{k+1}\alpha_i^2\end{aligned}$$

Then we compute $\|A\vec{v}_i\|^2$ since $\vec{y}$ is a linear combination of $\vec{v}_i$.

$$\begin{aligned}\|A\vec{v}_i\|^2 &= \langle A\vec{v}_i, A\vec{v}_i\rangle = (A\vec{v}_i)^T(A\vec{v}_i) \\ &= \vec{v}_i^T A^T A\vec{v}_i = \vec{v}_i^T(\lambda_i\vec{v}_i) = \lambda_i\vec{v}_i^T\vec{v}_i \\ &= \lambda_i = \sigma_i^2\end{aligned}$$

Therefore, we can write out $\|A\vec{y}\|^2$ as

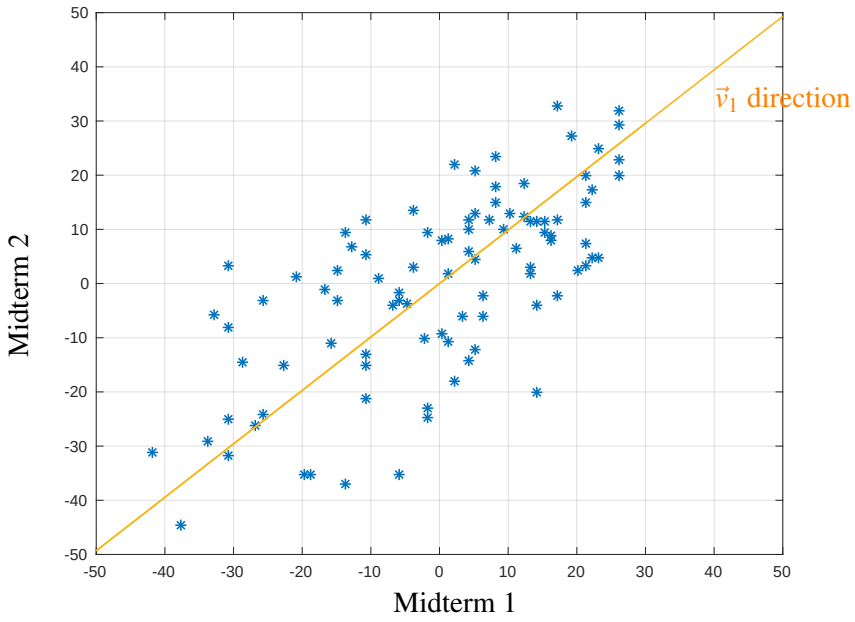
$$\begin{aligned}\|A\vec{y}\|^2 &= \langle\alpha_1 A\vec{v}_1 + \ldots + \alpha_{k+1}A\vec{v}_{k+1}, \alpha_1 A\vec{v}_1 + \ldots + \alpha_{k+1}A\vec{v}_{k+1}\rangle \\ &= \alpha_1^2\langle A\vec{v}_1, A\vec{v}_1\rangle + \ldots + \alpha_{k+1}^2\langle A\vec{v}_{k+1}, A\vec{v}_{k+1}\rangle \\ &= \alpha_1^2\sigma_1^2 + \ldots + \alpha_{k+1}^2\sigma_{k+1}^2\end{aligned}$$

However, since $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{k+1}$ and $\alpha_1^2 + \ldots + \alpha_{k+1}^2 = 1$, this implies that $\|A\vec{y}\|^2$ is a weighted sum with nonnegative scalars α_i^2. Its minimum value will be when $\alpha_{k+1} = 1$ and all other terms are 0. Therefore, we conclude by saying that $\|A\vec{y}\|^2 \geq \sigma_{k+1}^2$ finishing our proof!

5 Example

Here is an example of PCA from Professor Arcak's Reader:

As an illustration, the scatter plot below shows $n = 2$ midterm scores in a class of $m = 94$ students that I taught in the past. The data points are centered around zero because the class average is subtracted from the test scores. Each data point corresponds to a student and those in the first quadrant (both midterms ≥ 0) are those students who scored above average in each midterm. You can see that there were students who scored below average in the first and above average in the second, and vice versa.



For this data set the covariance matrix is:

$$\frac{1}{93}A^TA = \begin{bmatrix} 297.69 & 202.53 \\ 202.53 & 292.07 \end{bmatrix}$$

where the diagonal entries correspond to the squares of the standard deviations 17.25 and 17.09 for Midterms 1 and 2, respectively. The positive sign of the $(1,2)$ entry implies a positive correlation between the two midterm scores as one would expect.

The singular values of A, obtained from the square roots of the eigenvalues of A^TA, are $\sigma_1 = 215.08$, $\sigma_2 = 92.66$, and the corresponding eigenvectors of A^TA are:

$$\vec{v}_1 = \begin{bmatrix} 0.7120 \\ 0.7022 \end{bmatrix} \quad \vec{v}_2 = \begin{bmatrix} -0.7022 \\ 0.7120 \end{bmatrix}.$$

The principal component $\vec{v}_1$ is superimposed on the scatter plot and we see that the data is indeed clustered around this line. Note that it makes an angle of $\tan^{-1}(0.7022/0.7120) \approx 44.6^\circ$ which is skewed slightly towards the Midterm 1 axis because the standard deviation in Midterm 1 was slightly higher than in Midterm 2. We may interpret the points above this line as students who performed better in Midterm 2 than in Midterm 1, as measured by their scores relative to the class average that are then compared against the factor $\tan(44.6^\circ)$ to account for the difference in standard deviations.

The $\vec{v}_2$ direction, which is perpendicular to $\vec{v}_1$, exhibits less variation than the $\vec{v}_1$ direction ($\sigma_2 = 92.66$ vs. $\sigma_1 = 215.08$), but enough to convince you that you can do better on the final!

Contributors:

- Taejin Hwang.
- Murat Arcak.
- Saavan Patel.
- Utkarsh Singhal.

