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

Note 20

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 $x_1, x_2, \dots, x_n$. We can then aggregate our data into an $m \times n$ data matrix X where the rows of X represent a single sample $(x_1, x_2, \dots, x_n)$. 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 basis $\{\vec{v}_1, \vec{v}_2, \dots, \vec{v}_k\}$ for $k < n$, 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.

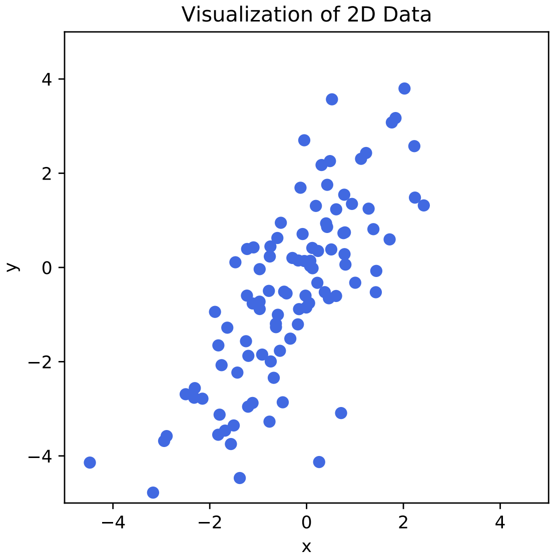
3 Variance Maximization Perspective

The first perspective we will be looking at is that of variance maximization. In a given dataset, certain variables will be show more correlation than other variables and will show larger signs of variability. What makes our data "special" is the directions in which this variability occurs and the magnitude of the correlation between variables.

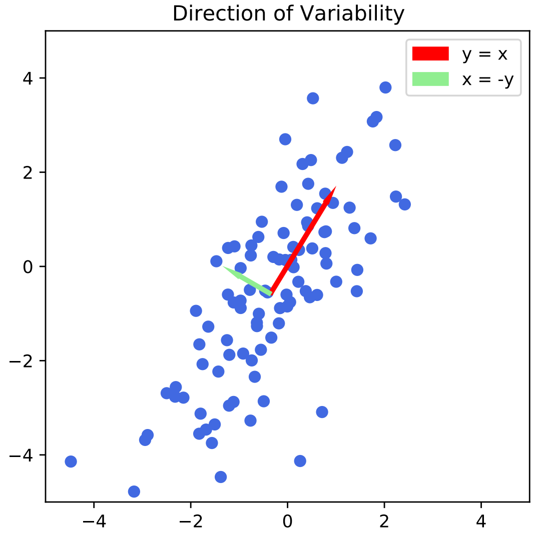
We will analyze the meaning behind variability in a dataset and try to understand how certain directions in the data can capture more variability than others. Using these most important directions, we will be able to perform dimensionality reduction through PCA.

3.1 Variability in Data

Let's suppose we had a sample dataset of $n = 2$ dimensions and $m = 100$ points.



Between the x and y variables, we are able to say that there is more variability across the line $y \approx x$ than its perpendicular complement $x \approx -y$.



If we wanted to capture our data with just a single dimension, then one way to do this would be to take the line $y \approx x$ and project all of our data-points onto that line. We would no longer be able to capture variability across the perpendicular axis, but we are doing the best we can with a single dimension.

The next section formalizes the observations we've made from this example into the familiar language of Linear Algebra with a pinch of Probability.

3.2 Preliminary Notation

To start off, let's introduce a term called **variance** which captures the amount of variability in a random variable X. Given m samples $x_1, \ldots, x_m$ from a random variable X, we define the **mean**, μ and **variance**, $\mathrm{Var}(X)$, from the sample observations:[1]

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

From our model defined in Section 2, $\vec{x}_i^T$ represents a single data-point and all of the data-points are aggregated as the rows of the data matrix X. Now let's a look at the vector $X\vec{w}$ where $\vec{w}$ is a unit vector of arbitrary direction in $\mathbb{R}^n$.

$$X\vec{w} = \begin{bmatrix} \leftarrow \vec{x}_1^T \rightarrow \\ \leftarrow \vec{x}_2^T \rightarrow \\ \vdots \\ \leftarrow \vec{x}_m^T \rightarrow \end{bmatrix} \vec{w} = \begin{bmatrix} \vec{x}_1^T \vec{w} \\ \vec{x}_2^T \vec{w} \\ \vdots \\ \vec{x}_m^T \vec{w} \end{bmatrix} \tag{3}$$

This follows the idea from the previous section where we are searching for a direction in the data with most variability. Since $\vec{w}$ is a unit vector, $\vec{x}_i^T \vec{w}$ is the weight of the projection of the i^{th} datapoint onto $\vec{w}$.

$$\mathrm{proj}_{\vec{w}} \vec{x}_i = \langle \vec{x}_i, \vec{w} \rangle \vec{w} = (\vec{x}_i^T \vec{w})\vec{w} \tag{4}$$

The projection signifies how strongly the data-point $\vec{x}_i$ aligns with the direction of $\vec{w}$. Therefore, our goal is to find the direction $\vec{w}$ that maximizes the variability in these projections as much as possible.

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.

$$A = X - \frac{1}{m}\vec{1}\vec{1}^T X \tag{5}$$

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

$$X = \begin{bmatrix} 1 & 2 \\ -1 & 3 \\ 3 & 4 \end{bmatrix} \implies A = \begin{bmatrix} 1 & 2 \\ -1 & 3 \\ 3 & 4 \end{bmatrix} - \begin{bmatrix} 1 & 3 \\ 1 & 3 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ -2 & 0 \\ 2 & 3 \end{bmatrix} \tag{6}$$

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.

3.3 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

$$\mathrm{Var}(A\vec{w}) = \frac{1}{m}\sum_{i=1}^{m}(\vec{x}_i^T \vec{w})^2 = \frac{1}{m}\|A\vec{w}\|^2 \tag{7}$$

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

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

$$\max_{\|\vec{w}\|=1} \operatorname{Var}(A\vec{w}) \implies \max_{\|\vec{w}\|=1} \|A\vec{w}\|^2 \tag{8}$$

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

3.3.1 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

$$\max_{\|\vec{z}\|=1} \|\Sigma \vec{z}\|^2 \tag{9}$$

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, $\vec{w} = V\vec{e}_1 = \vec{v}_1$.

3.3.2 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 $\vec{x}$.

$$\|A\|_2 = \max_{\vec{x} \neq \vec{0}} \frac{\|A\vec{x}\|}{\|\vec{x}\|} = \sigma_1 \tag{10}$$

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 $\|A\|_2$ 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!

3.4 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

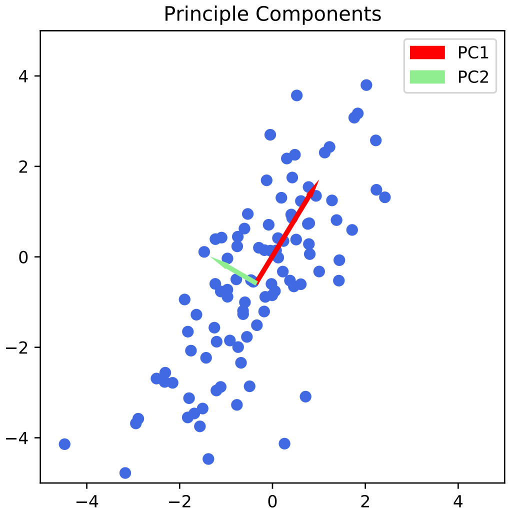
$$\max_{\|\vec{w}\|=1} \operatorname{Var}(A\vec{w}) \text{ subject to } \vec{w}^T\vec{v}_1 = 0 \tag{11}$$

The solution to this problem is $\vec{w} = \vec{v}_2$. If we were to continue doing this, it turns out that the orthonormal eigenvectors $\vec{v}_i$ of A^TA form our PCA basis. These basis vectors are called **principal components** and in practice, we pick $k < n$ vectors to perform our dimensionality reduction.

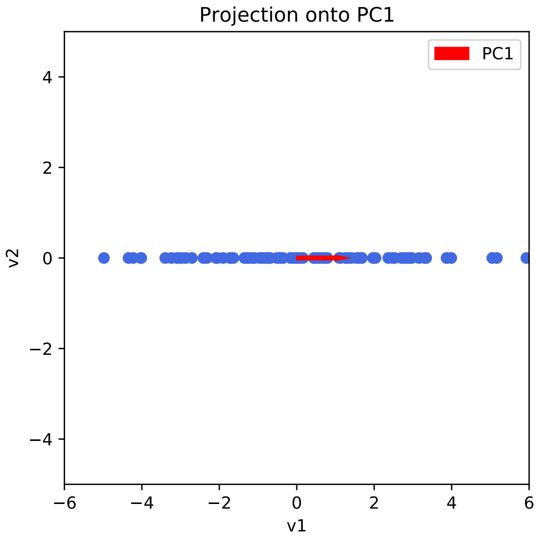
As a final remark, the variance in each direction will be $\frac{\sigma_i^2}{m}$ and the standard deviation, sometimes referred to as the **weights** of each principal component will be $\frac{\sigma_i}{\sqrt{m}}$.

3.5 Back to the Visuals

Let's revisit the visual example and connect it back to PCA. We center the data and compute the SVD to find the principal components $\vec{v}_1, \vec{v}_2$. The figure below plots the principal components $\vec{v}_i$ scaled by the weights.



Since $\vec{v}_1$ is the direction with maximal variance in the data, we can project all of the (x, y) data-points onto the first principal component $\vec{v}_1$ to perform dimensionality reduction.



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 data matrix A with rank r, we would like to find a matrix B_k of rank k that best approximates A. Recall that we can use the truncated SVD to make a low-rank approximation!

$$A \approx A_k = \sum_{i=1}^{k} \sigma_i \vec{u}_i \vec{v}_i^T \tag{12}$$

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

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.**

$$||A||_2 = \max_{\vec{x} \neq \vec{0}} \frac{||A\vec{x}||}{||\vec{x}||} = \sigma_1 \tag{13}$$

An alternate norm that we could define for a matrix is the **Frobenius Norm** which is defined as the square root of the sum of all singular values of A squared.

$$||A||_F^2 = \sum_{i=1}^{r} \sigma_i^2 \tag{14}$$

Whichever norm we use, this lets us represent the error of our rank k matrix as the norm of a matrix ε.

$$||\varepsilon||_F^2 = ||A - B_k||_F^2 \tag{15}$$

4.2 Optimization Problem

The goal is to find a matrix B_k that best approximates the original data A. As an optimization problem, this would look like the following

$$\begin{gathered} \min_{B_k} ||A - B_k||_F \\ \text{subject to } \text{Rank}(B_k) = k \end{gathered}$$

It turns out that the **Eckart-Young Mirsky Theorem** states that the optimal B_k is in fact the rank k SVD approximation of A.

$$A_k = \sum_{i=1}^{k} \sigma_i \vec{u}_i \vec{v}_i^T \tag{16}$$

Amazingly, the Eckart-Young Theorem holds for both the Spectral and Frobenius norm, but the proof is quite difficult and is deemed out of scope.

To summarize the result, the best rank k approximation of the matrix A comes from the truncated SVD. The principal components will be the basis that best approximates this data. Since our data was arranged as the rows of A, the subspace that best approximates $\text{Col}(A^T)$ will be the span of the first k right-singular vectors $\{\vec{v}_1, \ldots, \vec{v}_k\}$.

5 Computing the Principal Components

Having introduced two perspectives on Principal Component Analysis, let us take a look at two ways to compute the principal components.

5.1 Covariance Matrices

Through the variance maximization perspective, we can compute the principal components by finding the eigenvectors of the covariance matrix $C = \frac{1}{m}A^T A$ of our data.

1 **Compute the covariance matrix**:

- Center the matrix along the attributes (columns in this case), so that $A = X - \frac{1}{m}\vec{1}\vec{1}^T X$.
- Find the covariance matrix $C = \frac{1}{m}A^T A$ or $C = \frac{1}{m-1}A^T A$ depending on whether your data is a sample from a population or the full population.[2]

2 **Diagonalize the covariance matrix** C:

- The covariance matrix is symmetric so it is orthogonally diagonalizable: $C = P\Lambda P^{-1}$.
- Since P has orthonormal columns, we know $P^T = P^{-1}$.
- The columns of P are the principal components.
- The square root of the eigenvalues $\sqrt{\lambda_i}$ are the weights.

5.2 Low-Rank SVD

While the approach through the covariance matrix is mathematically sound, for reasons outside of the scope of this class, computing eigenvectors of the matrix $A^T A$ can become numerically unstable. Therefore, in practice, the SVD is better suited to compute the principal components.

1 **Normalize the data matrix**:

- Center the matrix along the attributes (columns in this case), so that $A = X - \frac{1}{m}\vec{1}\vec{1}^T X$.
- Scale the matrix A by $\frac{1}{\sqrt{m}}$ so that $S = \frac{1}{\sqrt{m}}A$. This is done in order to compute the correct weights.

2 **Compute the SVD of the matrix** S:

- The SVD of the matrix S will be of the form $S = U\Sigma V^T$.
- The columns of V are the principal components.
- The singular values σ_i are the weights.

As a sanity check, we can see that $C = S^T S$ meaning the eigenvectors and singular values align properly. You might ask why we use the SVD over covariance matrices when the SVD also involves computing eigenvectors of the matrix $A^T A$. It turns out that numerical tools have developed efficient methods to compute the SVD that don't involve computing the eigenvectors of $A^T A$.

[2] This distinction is called Bessel's correction which takes into account of the bias when sampling from a population. However in practice, with a lot of data, m will be large so this distinction will be very small.

Contributors:

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