EECS 16B Designing Information Devices and Systems II
Fall 2018 Elad Alon and Miki Lustig

Discussion 9B

Notes

Principal Component Analysis

Toy example

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.

As a motivating example, let's say you are a physicist attempting to measure the motion of an oscillating mass on a spring. You know the mass is oscillating in a particular direction, you just don't know what direction that is, and you want to figure it out. You set up 3 cameras to view what is going on.

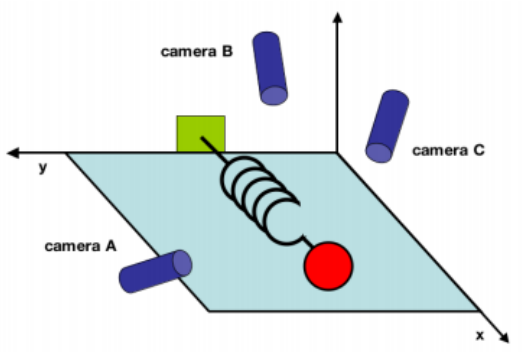

Figure 1: A diagram of the toy example.

Each camera is able to measure the x and y distance on the screen of the mass/spring system. If you had known the direction of the oscillations in the first place, you would only need one value at a given time point, but you have 6 now. You put all of this data into a large matrix of values representing where the spring is at each time point, and you want to do some linear algebra magic to figure out what direction the mass was moving, and how it was moving as a function of x, it's lineary distance. What kind of linear algebra magic can we do to make this happen?

PCA and the SVD

It turns out we have developed all of the linear algebra to do solve this problem. With some manipulation, the SVD actually lets us solve this problem. Each row of the A matrix above represents a "observation" and each column is a "feature". What we are looking for is a subspace of the A matrix that captures the largest variation in the dataset. In the case above, the largest variation in the dataset would be the direction that the mass is moving in (the x direction).

$$A = \begin{bmatrix} x_{11} & y_{11} & x_{12} & y_{12} & x_{13} & y_{13} \\ x_{21} & y_{21} & x_{22} & y_{22} & x_{23} & y_{23} \\ x_{31} & y_{31} & x_{32} & y_{32} & x_{33} & y_{33} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & y_{n1} & x_{n2} & y_{n2} & x_{n3} & y_{n3} \end{bmatrix}$$

Figure 1: Matrix representing various x and y values. Each row represents 1 time point, with the x and y points of the ball for each camera

The SVD can give us this subspace. The singular (σ_i) values in the Σ matrix represent the square root of the variation captured in the corresponding u and v vectors.

$$A = U\Sigma V^T = \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \dots & \vec{u}_r \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots & \\ 0 & \sigma_2 & 0 & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix} \begin{bmatrix} \vec{v}_1 & \vec{v}_2 & \dots & \vec{v}_r \end{bmatrix}^T = \sigma_1 \vec{u}_1 \vec{v}_1^T + \sigma_2 \vec{u}_2 \vec{v}_2^T + \cdots + \sigma_r \vec{u}_r \vec{v}_r^T$$

If the individual singular values σ_i represent the variation captured in data in that direction, then a large singular value would represent a large amount of variation captured by that direction. If our "features" are in the columns, we look at the V matrix; if our "features" are in the rows we look at the U matrix. The steps to actually calculate the PCA transform for a given set of data are given below.

(a) Demean the matrix along the terms. If the terms are in the column, then the mean of each column should be zero. If the terms are in the rows, the mean of each row should be zero.

(b) Calculate the SVD for your matrix.

(c) Look at your Σ matrix. To do PCA analysis, you need to see singular values that are much larger than the next greatest singular values. This implies that there is indeed some large amount of correlation in a few directions.

(d) Truncate your U and V matrices to the number of significant singular values and transform your original data points to the new basis of V vectors (for terms in columns) or U vectors (for terms in rows). Analyze the data.

Questions

1. PCA and Financial Markets

(a) See iPython notebook