CS 181 LECTURE 3/21/24

Scribe Notes

Contents

1 Introduction to PCA

Principal Component Analysis (PCA) is a powerful statistical technique used for dimensionality reduction and data visualization. It's commonly used in data analysis, pattern recognition, and machine learning.

At its core, PCA helps to simplify complex datasets by finding and highlighting the most important patterns or relationships among variables. It does this by transforming the original variables into a new set of uncorrelated variables, called principal components. These principal components are ordered in such a way that the first few components capture the maximum amount of variation in the data.

PCA works by finding the directions, or axes, along which the data varies the most. These directions are the principal components. By representing the data in terms of these components, which are ordered by their importance, PCA can effectively reduce the dimensionality of the dataset while preserving most of its information.

One of the key benefits of PCA is that it allows us to visualize high-dimensional data in a lower-dimensional space, making it easier to understand and interpret. It's widely used in fields such as image processing, genetics, finance, and many others where understanding complex datasets is essential.

2 Story

2.1 Cellular Dynamics

By using PCA, scientists can identify which genes are important for cells to change from one state to another. It's like figuring out which switches need to be flipped to move from one place to another on the epigenic landscape. This helps us understand how cells make decisions and change during development or in response to signals.

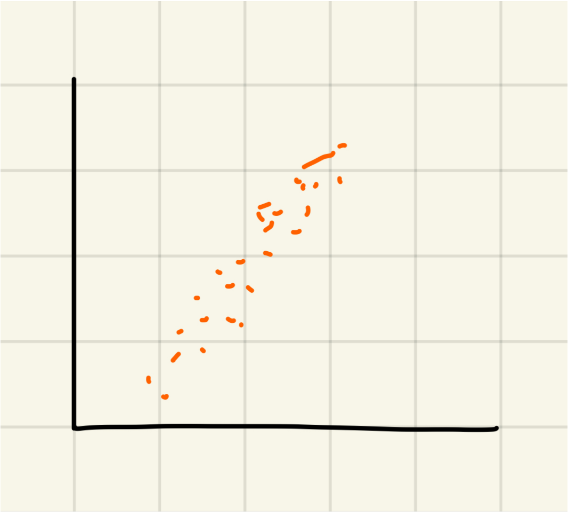
2.2 Ancestry

Another example is understanding ancestry. Find a low dimensional representation of the data. Say we know the rough ancestry origin. You can see the geography of Europe showing up. We are using something that doesn't know anything about geo-location, but location influences genomes so we are able to see geographical space. So when we project in certain dimensions, new things show up.

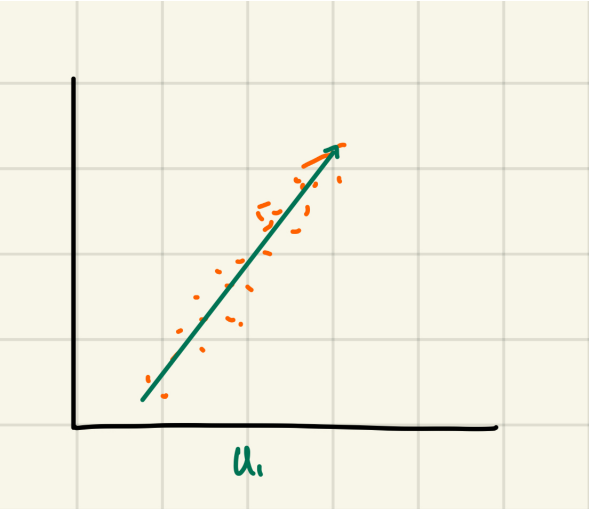
3 Finding hidden "directions"

Suppose we want to find D-dimensional vectors $\{u_1, ..., u_k\}$ such that $X \approx z_1 u_1 + ... + z_k u_k$. We are expressing the data as a linear combination. The u's give us the direction and the z's are scalars, which can be understood as the magnitude.

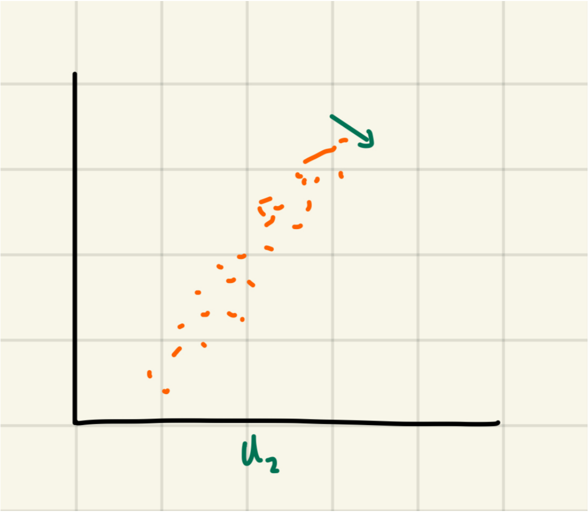
Let us say we have a dataset that looks like the following.



We see that if we had to choose one dimension to look at the data with, we would choose this vector, because it captures the most amount of information about the data.



But this other dimension also exists. It is possible for us to represent the data using both these dimensions. The number of dimensions we want to use is up to us.



$$\mathcal{L}(z,u) = \frac{1}{n}\sum_{n=1}^{N} ||x_n - \cup z_n||_2^2$$

Our goal is to minimize reconstruction loss.

Let us say we find a solution that minimizes this equation. Would the solution be unique? No.

What makes a good U? Orthogonal.

Let U be orthonormal: which means that $\langle u_k, u_{k'}\rangle = 0$ if $k \neq k'$ and $\langle u_k, u_k\rangle = ||u_k||^2 = 1$.

Given $X = UZ$. Then, $Z = U^\top X$.

The vectors start from the origin, and if we first move the data and center it, then we are back to using our vectors. So, we center the data by subtracting the mean, and we can do this in a much more interpretable way. We define an $X = \bar{X} + z_1u_1 + ...z_ku_k$. We use the mean as a frame of reference and express everything with respect to the mean. So now we are approximating $X - \bar{X}$ which we can call $\tilde{X}$.

$$\tilde{X} = X - \bar{X} = z_1u_1 + ...z_ku_k$$

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} ||\tilde{X_n} - UZ_n||_2^2$$

An observation is that we can complete basis.

$$\tilde{X} = \sum_{d=1}^{D} Z_{nd}U_d$$

Every data point we can express like this. We have our truncated expression. and we compare to see how far are we from the true datapoint.

We plug in our approximate datapoint for the true data point in our loss.

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} ||\sum_{d=1}^{D} Z_{nd}U_d - \sum_{d=1}^{K} Z_{nd}U_d||$$

Use orthonormality in two steps.

$$\begin{aligned}
\mathcal{L} &= \frac{1}{N}\sum_{n=1}^{N} || \sum_{d=1}^{D} Z_{nd}U_d - \sum_{d=1}^{K} Z_{nd}U_d||^2 \\
&= \frac{1}{N}\sum_{n=1}^{N} || \sum_{d=k+1}^{D} Z_{kd}U_d|| \\
&= \frac{1}{N}\sum_{n=1}^{N} (\sum_{d=k+1}^{D} Z_{kd}^2 \langle U_d, U_d \rangle) \\
&= \frac{1}{N}\sum_{n=1}^{N} \sum_{d=k+1}^{D} Z_{kd}^2
\end{aligned}$$

Since $Z_{nd} = U_d^\top \tilde{X_n}$, then we have

$$= \sum_{d=k+1}^{D} [\frac{1}{N}\sum_{n=1}^{N} (X_n - \bar{X})(X_n - \bar{X})^\top]U_d$$

4 Statistical Interpretation

So our loss becomes

$$= \sum_{d=k+1}^{D} U_d^\top \Sigma U_d$$

Now we can formalize in a statistical perspective. We want the U, z that minimizes the loss.

We want to minimize the variance along $U_{k+1}, ...U_D$, which is equivalent to maximizing the variance along the vectors that we did use. From our earlier example, U_2 would be a leftover vector. We would want to minimize the variance along the U_2 direction.

5 Linear Algebra Interpretation

Suppose first that we just want to find a single U from the leftover vectors.

Given the optimization problem

$$\min_U U^\top \Sigma U$$

subject to:

$$U^\top U = 1$$

We can use the method of Lagrange multipliers to solve this problem. The Lagrangian is formulated as follows

$$\mathcal{L}(U, \lambda) = U^\top \Sigma U - \lambda(U^\top U - 1)$$

Taking the derivative of $\mathcal{L}$ with respect to U, we get

$$\frac{\partial \mathcal{L}}{\partial U} = 2\Sigma U - 2\lambda U$$

Setting this expression to zero to find the critical points, we have

$$2\Sigma U = 2\lambda U$$

Which simplifies to

$$\Sigma U = \lambda U$$

For many U's, we have

$$\min_{U} \sum_{d=k+1}^{D} U_d^\top \Sigma U_d = \sum_{d=k+1}^{D} U_d^\top \lambda U_d = \sum_{d=k+1}^{D} \lambda$$

Therefore, we can rewrite our reconstruction loss minimization problem to be:

1. Find eigenvectors of $Sigma$ with smallest λ and discard.
2. Find eigenvectors of $Sigma$ with largest λ and discard.

5.1 Scree Plots

5.1.1 Purpose

The main purpose of a scree plot is to visualize the eigenvalues or variance explained by each factor or component in a factor analysis or PCA, respectively. By examining the plot, you can identify the point at which the eigenvalues or explained variance begin to level off, indicating a potential cutoff point for retaining factors or components.

To create a scree plot, you plot the eigenvalues (in the case of factor analysis) or the variance explained (in the case of PCA) against the number of factors or components. The eigenvalues/variance explained are typically plotted on the y-axis, while the number of factors/components is plotted on the x-axis.

5.1.2 Interpretation

When you examine a scree plot, you look for the "elbow" or point where the plot begins to level off. This point represents a natural cutoff where the eigenvalues/variance explained by additional factors/components decrease substantially. Factors/components before the elbow are considered significant and are typically retained, while those after the elbow may be discarded.

The decision of how many factors/components to retain based on a scree plot is somewhat subjective and may depend on the specific context of the analysis, the goals of the study, and other considerations. Some common rules of thumb include retaining factors/components with eigenvalues/variance above 1 or above the point of inflection in the plot.

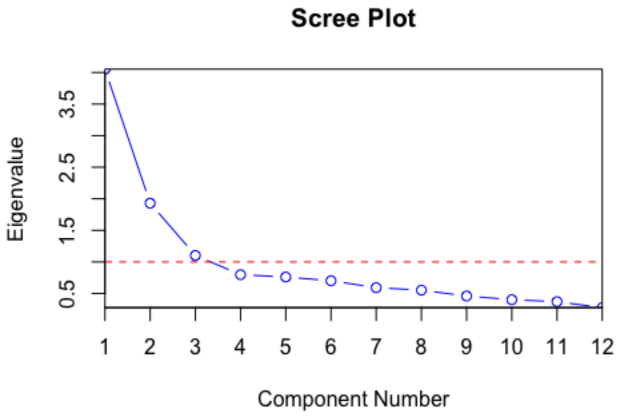


Figure 1: Image from Wikipedia.

6 Summary

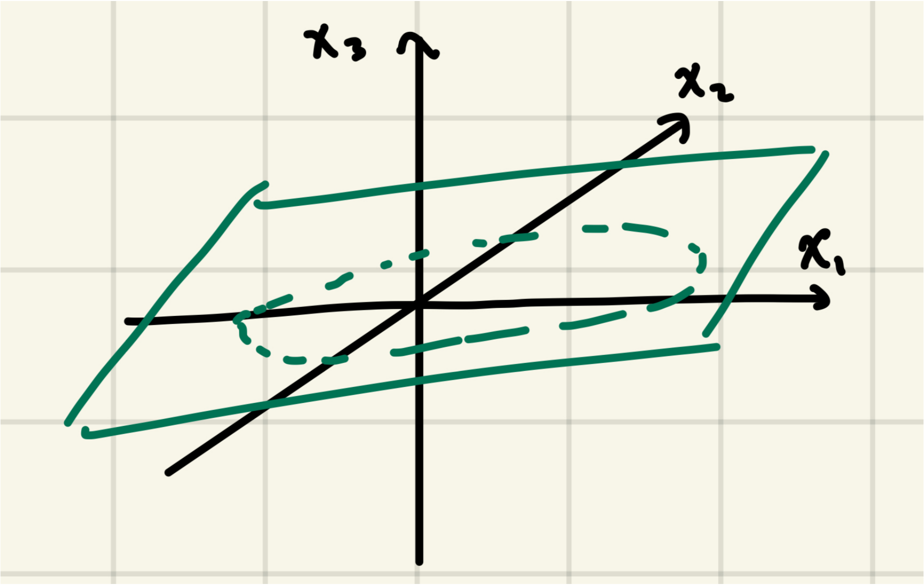
We found the PC decomposition to be $\tilde{X} = UZ$.
The dimensions of $\tilde{X}$ is D x N, dimensions of U is DxD, and the dimensions of Z is DxN. The columns of U are the eigenvectors of Σ. given by $\sqrt{\lambda_k} u_k$, where λ_k is the eigenvalue corresponding to the k-th eigenvector. The loadings are useful because they capture the contribution of each original variable (feature) to the principal component. They represent the coefficients of the linear combination of the original variables that make up each principal component. The eigenvalue λ_k captures the amount of variance explained by the corresponding principal component, and the eigenvector u_k determines the direction (pattern) of that component in the original feature space.

6.0.1 Non-uniqueness of U

Because of the rotation and scaling properties of PCA, there can be multiple sets of eigenvectors (principal components) that are equally valid for representing the data. This means that different choices of eigenvectors can lead to the same amount of explained variance or reconstruction loss. While PCA provides a useful way to reduce the dimensionality of data and capture its variability, the specific choice of principal components is not unique. Different choices can still capture similar patterns in the data, but they may vary in terms of the ordering and sign of the components. This non-uniqueness is a characteristic of PCA and is important to keep in mind when interpreting the results.

7 Concept Check

1) Order PC with repsect to λ.
2) Project into 2, we use UQ where Q is orthonormal. If we use UQ, what is the loss?
3) Will UQ reflect variance in data?



8 Concept Check Answer

1) x_1, x_2, x_3

2) The loss remains the same.

3) No.