Stats 102B - Introducing PCA

Miles Chen, PhD

Department of Statistics

Week 9 Monday

UCLA

Section 1

PCA

PCA: Big Picture

Principal Components Analysis

You have a dataset with many variables.

The variables in the dataset are possibly correlated.

PCA transforms the data into uncorrelated variables.

The goal is often dimension reduction. If we take advantage of the fact that many variables are correlated, can we reduce the number of variables and still capture or convey the variation that exists in the data?

Example

Housing prices: A dataset on housing prices may have several variables:

- livable sq footage
- total sq footage (includes basement)
- number of bedrooms
- number of bathrooms
- lot size
- pool: y/n
- roof style
- AC/heating type
- garage/car port/parking spaces
- date built
- kitchen quality
- floor quality
- roof quality
- location / zip code
- local school rating

Dimension reduction

The housing dataset may have many variables.

Can we reduce the number of variables?

I would argue that the many of the variables correlate together and can be grouped:

Size: sq footage, number of bedrooms, number of bathrooms generally go up or down together.

Amenities/Quality: pool, kitchen quality, floor quality, roof quality, etc.

Neighborhood desirability: local school rating, local crime rate, local income level, etc.

Dimension Reduction is useful

Other Examples of dimension reduction:

GPA - reduces an entire academic performance to a single varaible

SAT/GRE - reduces a complex notion of intelligence down to two (or three) numbers: verbal, math, writing

Credit score - reduces financial trustworthiness to a single number (takes into account payment history, debt to credit ratio, number of credit inquiries, number of negative marks, etc.). Banks will often use just two variables: credit score and annual income to make a decision on a loan.

NFL Passer rating - reduces a quarterback's entire game performance to a single value (takes into account pass attempts, pass completions, pass yards, number of touchdowns, number of interceptions)

Olympic Heptathlon - performance in seven events gets reduced to a single score

PCA: Visual Representation of Dimension Reduction

An object exists in 3 dimensions.

If you shine a light on the object, it will cast a shadow. The shadow is a 2 dimensional projection of the object.

Depending on how the shadow is projected, the shadow may resemble the object or it might not.

Similarly, you want to select a lower dimensional projection of your high dimensional data that "resembles" the original data.

PCA

Mathematically, when we say we want a lower dimensional projection that “resembles” the original data, we are saying that we want a projection that captures as much of the variance that exists in the original high-dimensional data.

If we reduce the dimensionality of the data, we probably can’t capture all of the variance that exists in the original data, but we want to find one that still has a high amount of variance.

Thus, we will search for a lower dimensional projection that maximizes the variance in the resulting projected space.

Defining our terms

The original data exists in M dimensions. The nth observation in the original data is $\mathbf{y}_n = [y_{n1}, \ldots, y_{nM}]^T$, a vector of length M. We wish to project the data down to D dimensions $(D < M)$

After projecting the nth observation down to D dimesions, we have $\mathbf{x}_n = [x_{n1}, \ldots, x_{nD}]^T$

PCA will define D vectors, $\mathbf{w}_d$, each of which is M dimensional. The dth element of the projected observation, x_{nd} is computed as:

$$x_{nd} = \mathbf{w}_d^T \mathbf{y}_n$$

That is to say that the elements x_{nd} in the projected data are linear combinations of the values in $\mathbf{y}_n$, each of which is weighted by the values in $\mathbf{w}_d$.

For example, if we project from 3 dimensions down to 2 dimensions, you will have two projection vectors each of length three: $\mathbf{w}_1 = [w_{11}, w_{12}, w_{13}]^T$ and $\mathbf{w}_2 = [w_{21}, w_{22}, w_{23}]^T$. When you multiply the nth observation $\mathbf{y}_n$ by $\mathbf{w}_1$ you'll get the first component x_{n1}. Multiply $\mathbf{y}_n$ by $\mathbf{w}_2$ to get the second component x_{n2}.

Rules for choosing $\mathbf{w}_d$

PCA will choose $\mathbf{w}_d$ so that the variance in the projected space is maximized.

The first vector, $\mathbf{w}_1$ will be set of weights so that the variance in the projected values x_{n1} is as high as possible.

The second vector, $\mathbf{w}_2$ will also maximize the variance in x_{n2}, but it must be orthogonal to $\mathbf{w}_1$. (That is $\mathbf{w}_1^T \mathbf{w}_2 = 0$)

The third vector, $\mathbf{w}_3$ will maximize the variance in x_{n3}, but must be orthogonal to all of the previous $\mathbf{w}$ vectors, and so on.

In general, $\mathbf{w}_i^T \mathbf{w}_i = 0$ for all $i \neq j$.

Also, to avoid having infinitely many solutions, we restrict the length of each $\mathbf{w}_d$ vector to be 1. $(\mathbf{w}_i^T \mathbf{w}_i = 1)$

Maximizing the variance in x_{n1}

We begin by deriving the vector $\mathbf{w}_1$ so that the variance in x_{n1} is maximized.

To make our life easier, we will center the original data Y. (Subtract the column mean from each column)

$$\bar{\mathbf{y}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{y}_n = \mathbf{0}$$

Maximizing the variance in x_{n1}

To keep the notation easier, we'll just work with one weight vector $\mathbf{w}$. Each observation in Y will be projected down to one scalar value x_n

$$\underbrace{x_n}_{1\times 1} = \underbrace{\mathbf{w}^T}_{1\times M} \underbrace{\mathbf{y}_n}_{M\times 1}$$

The variance of the projected data is

$$\sigma_x^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2$$

Simplifying things

Remember, we said the column means are 0. $\bar{\mathbf{y}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{y}_n = \mathbf{0}$

$$\begin{aligned}
\bar{x} &= \frac{1}{N}\sum_{n=1}^{N} x_n \\
&= \frac{1}{N}\sum_{n=1}^{N} \mathbf{w}^T\mathbf{y}_n \\
&= \mathbf{w}^T\left(\frac{1}{N}\sum_{n=1}^{N} \mathbf{y}_n\right) \\
&= \underbrace{\mathbf{w}^T}_{1\times M}\underbrace{\bar{\mathbf{y}}}_{M\times 1} = \underbrace{0}_{1\times 1}
\end{aligned}$$

The mean $\bar{x}$ is 0.

Our variance expression can be simplified

$$
\begin{aligned}
\sigma_x^2 &= \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}(x_n - 0)^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{y}_n)^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbf{w}^T\mathbf{y}_n\mathbf{y}_n^T\mathbf{w} \\
&= \underbrace{\mathbf{w}^T}_{1\times M}\underbrace{\left(\frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n\mathbf{y}_n^T\right)}_{M\times M}\underbrace{\mathbf{w}}_{M\times 1}
\end{aligned}
$$

Simplifying the variance

We let $\mathbf{C}$ be the covariance matrix and recall that $\bar{\mathbf{y}} = \mathbf{0}$

$$\mathbf{C} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{y}_n - \bar{\mathbf{y}})(\mathbf{y}_n - \bar{\mathbf{y}})^T = \frac{1}{N} \sum_{n=1}^{N} \mathbf{y}_n \mathbf{y}_n^T$$

The variance of x_n is then:

$$\sigma_x^2 = \mathbf{w}^T \left(\frac{1}{N} \sum_{n=1}^{N} \mathbf{y}_n \mathbf{y}_n^T \right) \mathbf{w}$$

$$\sigma_x^2 = \mathbf{w}^T \mathbf{C} \mathbf{w}$$

Maximizing the Variance

We want to maximize the variance σ_x^2 subject to the constraint that $\mathbf{w}^T\mathbf{w} = 1$.

$$\sigma_x^2 = \mathbf{w}^T\mathbf{C}\mathbf{w}$$

We will do this with a Lagrange multiplier:

$$L = \mathbf{w}^T\mathbf{C}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{w} - 1)$$

Take the derivative and set it equal to 0.

$$\frac{\partial L}{\partial \mathbf{w}} = 2\mathbf{C}\mathbf{w} - 2\lambda\mathbf{w} = 0$$

$$\mathbf{C}\mathbf{w} = \lambda\mathbf{w}$$

Finding a solution

We wish to find a solution for $\mathbf{w}$ that fits the following equation.

$$\mathbf{C}\mathbf{w} = \lambda \mathbf{w}$$

Keep in mind $\mathbf{C}$ is the covariance matrix, and λ is a scalar.

We have: matrix $\times$ vector $=$ scalar $\times$ vector

The solution is that $\mathbf{w}$ is the eigenvector of $\mathbf{C}$, and λ is the eigenvalue.

The eigenvalue/eigenvector pair with the highest eigenvalue will be the projection with the highest variance ($\mathbf{w}_1$), the second highest eigenvalue will correspond to $\mathbf{w}_2$ and so on.

Summary:

Finding the PCA to go from M dimension to D dimensions.

1. Center the data so that the column means are all 0. (Optional: scale the data so that they have variane 1)
2. Compute the covariance matrix, $\mathbf{C}$.
3. Find the M eigenvector/eigenvalue pairs of the covariance matrix.
4. Select the eigenvectors for the D highest eigenvalues
5. Compute $\mathbf{X} = \mathbf{YW}$ ($\mathbf{Y}$ is $N \times M$, $\mathbf{W}$ is $M \times D$, and $\mathbf{X}$ is $N \times D$)