Principal Component Analysis

October 14, 2019

Raguvir Kunani

Linear Algebra Review: Vectors

A **vector** is an ordered list of numbers. An n-dimensional vector contains n numbers. A fancy math notation for saying a vector x is an n-dimensional vector is $x \in \mathbb{R}^n$.

Vectors are assumed to be column vectors unless otherwise stated.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Linear Algebra Review: Matrices

A **matrix** is a rectangular array of numbers. It often helps to think of a matrix as a collection of vectors. Every matrix has 2 dimensions: the number of rows m and the number of columns n. If a matrix A has m rows and n columns, we say that $A \in \mathbb{R}^{m \times n}$.

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix}$$

In general, $m \neq n$. But when $m = n$, we say A is a square matrix.

Linear Algebra Review: Vector Tranposes

The transpose of a column vector is a row vector and vice versa.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

column vector

row vector

$$x^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$$

Linear Algebra Review: Matrix Tranposes

The tranpose of a matrix flips the rows and columns.

first row

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix}$$

first column

$$A^T = \begin{bmatrix} A_{11} & \cdots & A_{m1} \\ \vdots & \ddots & \vdots \\ A_{1n} & \cdots & A_{mn} \end{bmatrix}$$

Linear Algebra Review: Vector and Matrix Products

Let $x, y \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m \times n}$.

Vector Inner Product

$$x^T y$$

Matrix-Vector Product

$$Ax = \underset{m \times n}{\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & \\ \vdots & & \ddots & \\ A_{m1} & & & A_{mn} \end{bmatrix}} \underset{n \times 1}{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}} = \underset{m \times 1}{\begin{bmatrix} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1n}x_n \\ A_{21}x_1 + A_{22}x_2 + \cdots + A_{2n}x_n \\ \vdots \end{bmatrix}}$$

Dimensionality Reduction

The **dimension** of some data is the number of *independent* attributes in the data. In linear algebra, we refer to dimension as the *rank* of a matrix.

What does it mean to reduce the dimensionality of our data?

- fewer "columns" to look at

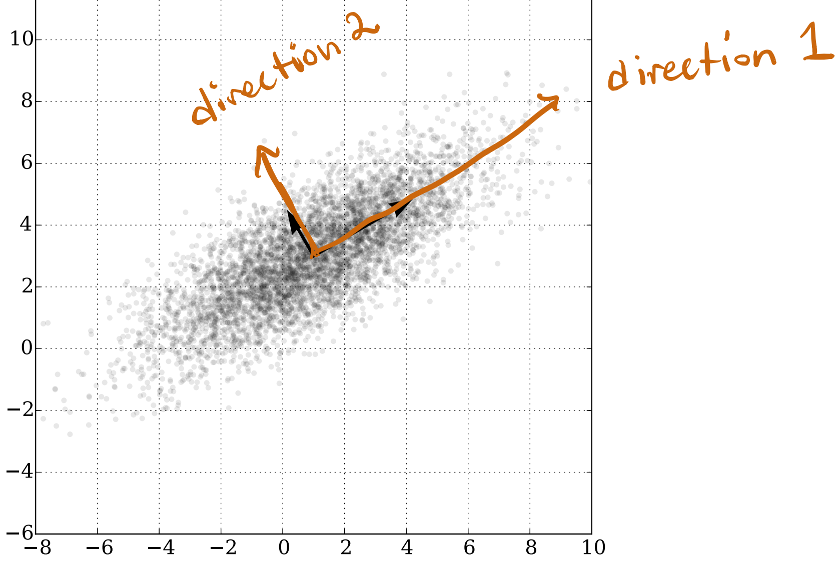
What are some ways to reduce the dimensionality of our data?

- PCA
- delete columns manually
  ↳ usually not feasible because we don't know which columns are important

Principal Component Analysis (PCA): Overview

PCA takes advantage of the fact that any data inherently lies along some set of directions. Using SVD, PCA attempts to find these directions and characterize the data using these directions.

SVD: singular value decomposition



PCA: What Are These "Directions?"

PCA uses SVD to find the "most important" directions that describe our data. What does "most important" mean? For us, the "most important" directions are the directions that our data has the most variance along.

Example: Calculating Final Course Grades

- MT1, MT2, Final Exam
- Come up with some formula for assigning final course grades
- Formula: linear combination of the attributes (columns)

Fortunately, SVD already calculates these directions for us, so we don't have to manually find them.

have orthonormal columns

PCA: How To Get the Directions From SVD

The SVD of a matrix A is $A = U\Sigma V^T$, where U, V are orthogonal and Σ is diagonal.

If A is the (centered) data matrix, then the columns of V are the "most important" directions, or **the directions of the principal components**. The columns of $U\Sigma$ contain the **principal components** themselves.

Question: In our final course grade example, what is the direction of the first principal component? What is the first principal component?

Direction of first principal component: coefficient of the linear combination that produces the largest variance

the actual final course grades

SVD: The Σ matrix

Let's a take a closer look at the Σ matrix from SVD. Let's say that after doing SVD on some data matrix A, the Σ matrix looks like:

This is wrong, see the last page for the correct version

$$\Sigma = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

What this means is that the first principal component has a variance of $3^2 = 9$, the second principal component has a variance of $2^2 = 4$, etc. You don't need to know why, just that this relationship exists.

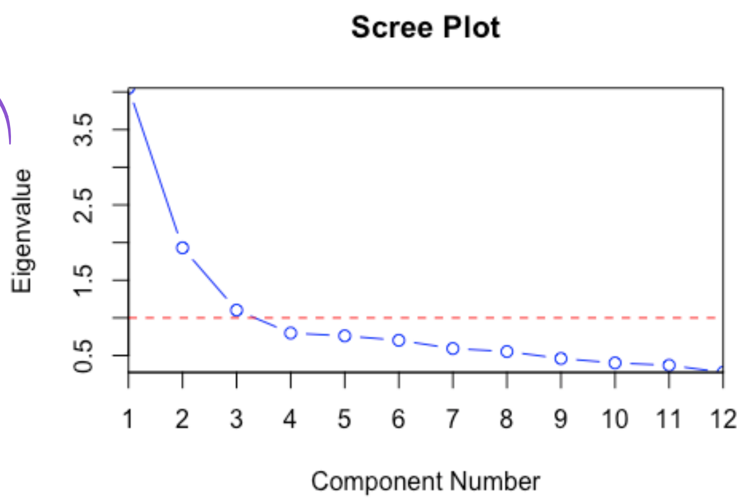
> The entries of Σ are *always* nonnegative!

Cheat sheet?

Scree Plots

The whole point of PCA is to reduce the dimensionality of our data, so let's return to that. To reduce the dimensionality of our data, we take the k principal components that account for most of the variance in our data. To make this process easier, we use a **scree plot**:

these are eigenvalues of $A^T A$, which are related to the variance of the principal components (so you can think of the y-axis as being variance)



Worksheet!

Feedback Form

This *anonymous* form is for me to learn what I can do to ensure you all get the most of discussion and lab. This form will be open all semester, and I'll be checking it regularly. Be as ruthless as you want, I promise my feelings won't get hurt.

Feedback Form: tinyurl.com/raguvirTAfeedback

Correction on Variance of Principal Components

In an earlier slide I said that the variance of PC1 is $3^2 = 9$. This is wrong. The correct statement "the amount of variance captured by PC1 is $\frac{3^2}{N}$, where N is the number of data points." When I say "the amount of variance captured by PC1," I mean that PC1 explains $\frac{3^2}{N}$ units of the total variance of the data. Similarly, the amount of variance captured by PC2 is $\frac{2^2}{N}$. In general, the amount of variance captured by PCi (the i^{th} principal component) is $\frac{\sigma_i^2}{N}$, where σ_i is the i^{th} diagonal entry of the Σ matrix