ThoughtWorks®

Aug 2013

Principal Component Analysis

Ricardo Wendell

Agenda

Feature Engineering

(Our motivation)

Introduction to Principal Component Analysis

(And some statistical concepts)

Agile Analytics and PCA

(Helping visualization…)

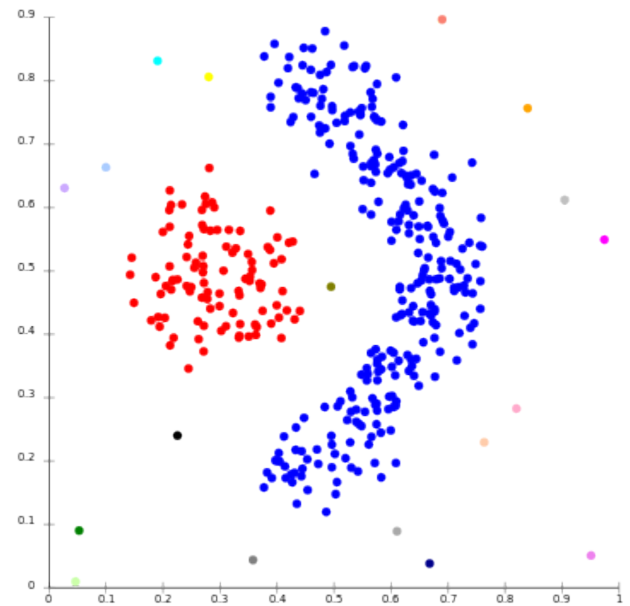
Feature Engineering

Given a classification problem...

How do we choose the right features?

Intuition fails in high dimensions

Building a classifier in two or three dimensions is relatively easy…



It’s usually possible to find a reasonable frontier between examples of different classes just by **visual inspection**.

Feature engineering

Intuitively, one might think that gathering more features never hurts, right?

At worst they provide no new information about the domain…

The curse of dimensionality

Many algorithms that work fine in low dimensions become intractable when the input is high-dimensional.

Bellman, 1961

How do we solve it?

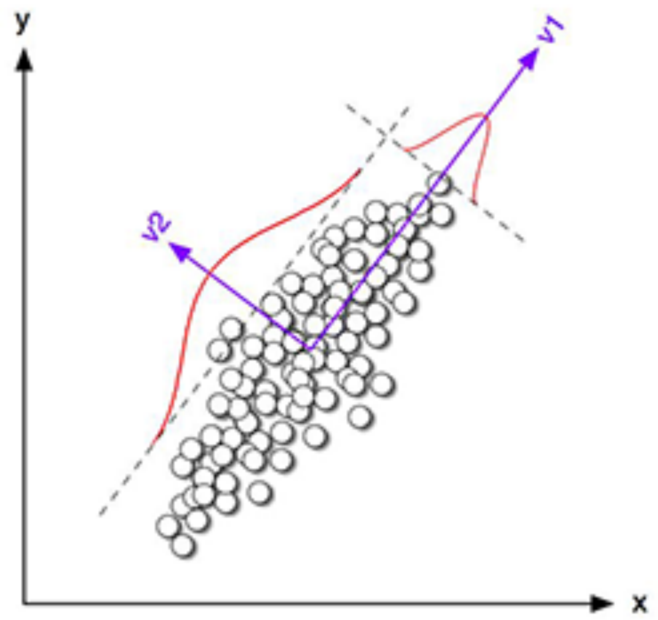
Feature Selection

Feature Extraction

Feature extraction

"In most applications examples are not spread uniformly throughout the examples space, but are concentrated on or near a lower-dimensional subspace."

Introduction to PCA



Objective of PCA

To perform dimensionality reduction while preserving as much of the randomness in the high-dimensional space as possible

Principal Component Analysis

It takes your cloud of data points, and rotates it such that the maximum variability is visible.

PCA is mainly concerned with identifying correlations in the data.

Measuring Correlation

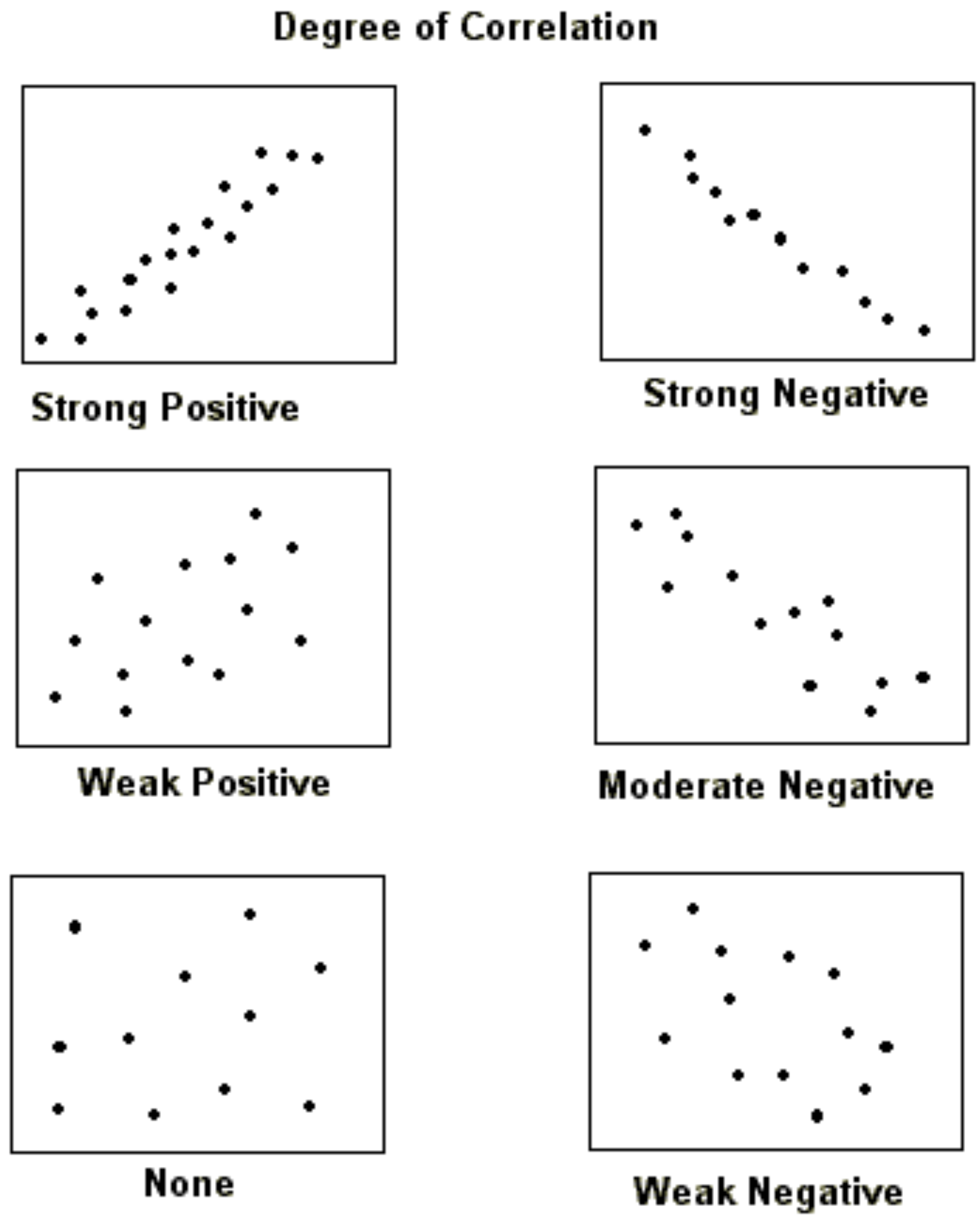
Degree and type of relationship between any two or more quantities (variables) in which they vary together over a period

Correlation can vary from +1 to -1.

Values close to +1 indicate a high-degree of **positive correlation**, and values close to -1 indicate a high degree of **negative correlation**.

Values close to zero indicate poor correlation of either kind, and 0 indicates no correlation at all

Measuring Correlation



Beware:

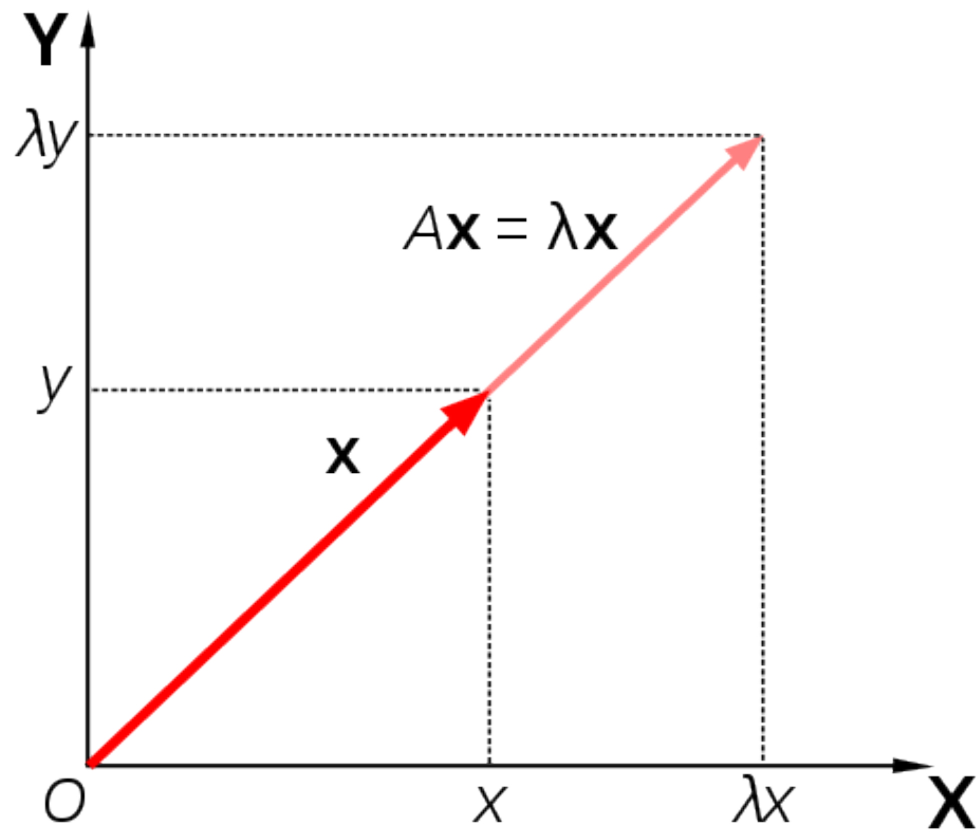
Correlation does not imply causation

Correlation matrix

It shows at a glance how variables correlate with each other

	Q1	Q2	Q3	Q4	Q5
Q1	1.00	0.77	0.95	-0.81	-0.65
Q2	0.77	1.00	0.89	-0.29	-0.84
Q3	0.95	0.89	1.00	-0.97	0.13
Q4	-0.81	-0.29	-0.97	1.00	0.35
Q5	-0.65	-0.84	0.13	0.35	1.00

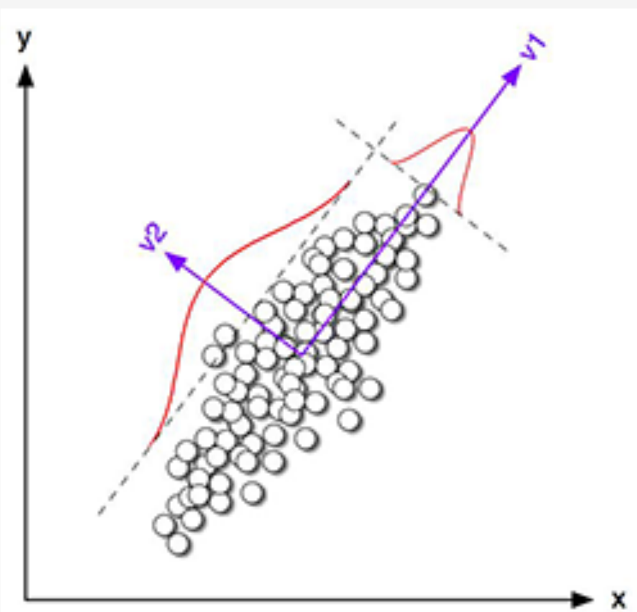
Eingenvalues and eingevectors



$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 11 \\ 5 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

Steps for PCA



1. Standardize the data

2. Calculate the covariance matrix

3. Find the eigenvalues and eingenvectors of the covariance matrix

4. Plot the eigenvectors / principal components over the scaled data

Demo with R

Let’s check the products of PCA...

Agile analytics and PCA

Agile Analytics

Machine learning and data mining tools and techniques

+

Knowledge of the domain at hand

+

Short feedback cycles

Agile Analytics

We could use PCA as a tool to quickly identify correlation between features, helping feature extraction and selection.

Reducing dimensionality using PCA or other similar technique can help us achieve better and quicker results.

ThoughtWorks®

QA & Next Steps