# Applied Data Mining

Rockhurst University

Curse of Dimensionality

Section 1

# How many is too many?

# Clustering revisited

Consider the issues we raised with clustering and categorical variables

- We know the algorithm can't handle categorical (factor) variables.
- One solution is to "one hot" dummy all the relevant variables.
- This can expand the number of variables **signficantly**.
- Example: If you have one categorical data for "State", it becomes 50 dummy variables.
- This can lead to statistical problems and drastically increase computational time and resource requirements.

# Curse of dimensionality

- When dimensionality increases, data become increasingly sparse (e.g. lots of zeros in those dummy variables)
- Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
- The possible combinations of subspaces will grow exponentially

# Dimensionality reduction

What if we could obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results? We could:

- Avoid the curse of dimensionality
- Help eliminate irrelevant features and reduce noise
- Reduce time and space required in data mining
- Allow easier visualization

Section 2

# PCA

# Principal component analysis (PCA)

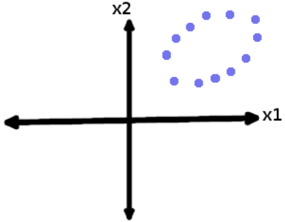
Goal: retain information with fewer variables.

- Find a projection that captures the largest amount of variation in data.
- The original data are projected onto a much smaller space, resulting in dimensionality reduction.
- We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space

Projection? Eigenvectors?
Help?

# Pictures: original dataset

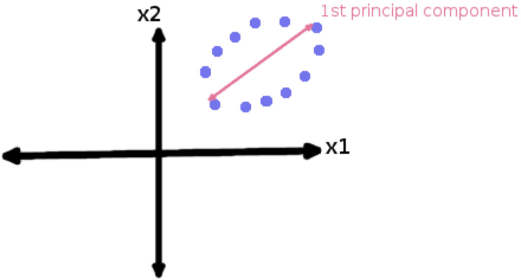
Let's say our dataset has two variables $x_1$ and $x_2$[1]:



[1] Many thanks to http://www.lauradhamilton.com/introduction-to-principal-component-analysis-pca

# Pictures: First component

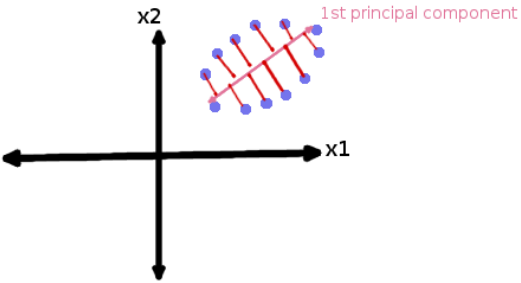
Now, we want to identify the first principal component that has explains the highest amount of variance. Graphically, if we draw a line that splits the oval lengthwise, that line signifies the component that explains the most variance:



# Pictures: One dimension

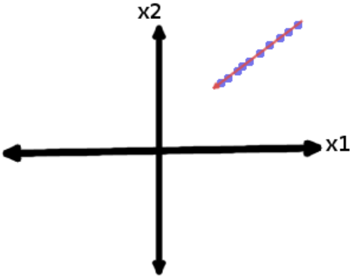
Let's say we just wanted to project the data onto the first principal component only. In other words, we wanted to use PCA to reduce our two-dimensional dataset onto a one-dimensional dataset.

Basically, we would collapse our dataset onto a single line (by projecting it onto that line). The single line is the first principal component.



# Pictures: Just the line

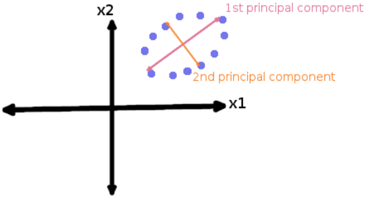
This means we can represent our data by using just the line. We lose some of the information but not all of it. We kept information about both $x_1$ and $x_2$



# Pictures: second component

But we can still use more of the information than just the first component. The second principle component explains more of the variation — by definition, it explains variation not captured in the first component. (Technical speak: it's **orthogonal** to the first component)

For our simple 2-dimensional example, there can only be 2 principle components:

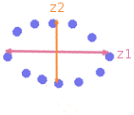


# Pictures: two components

The two principal components are perpendicular to each other. They capture independent elements of the dataset.

If we actually performed PCA on this dataset, we would have two components that are linear combinations of $x_1$ and $x_2$ but don't represent the exact same points.

Essentially we rotate the data to use new dimensions:



# PCA - steps

Official nerd description: Given $N$ data vectors from $n$-dimensions, find $k \leq n$ orthogonal vectors (principal components) that can be best used to represent data

- Normalize input data: Each attribute falls within the same range
- Compute k orthonormal (unit) vectors, i.e., principal components
- Each input data (vector) is a linear combination of the k principal component vectors
- The principal components are sorted in order of decreasing "significance" or strength
- Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)

# How many principal components?

By definition, you cannot have more components than you have variables. (You can't need more components to explain all the variance than you have original data that defines the variance.)

In general, you want to choose the smallest number of components that explain the most variance. The first component will explain the most variation by construction. The subsequent components explain smaller proportions each.

If you are lucky, a very small number of components will explain a large amount of variation. You get to choose how much information you want to lose to gain a smaller number of dimensions.

# What are the components?

Mathematically, the principal components are the eigenvectors of the covariance matrix of the original dataset.

They are linear combinations of the original data that capture the most variation in n-dimensional space. Remember how linear regression fits the line that captures the most variance? Same idea.

Yes but what *are* they? How do I interpret them? Unfortunately, they don't often have business interpretation. You may be able to see how the linear combination represents something relative to your business but it doesn't happen very often.

# Caveats

A few things to remember:

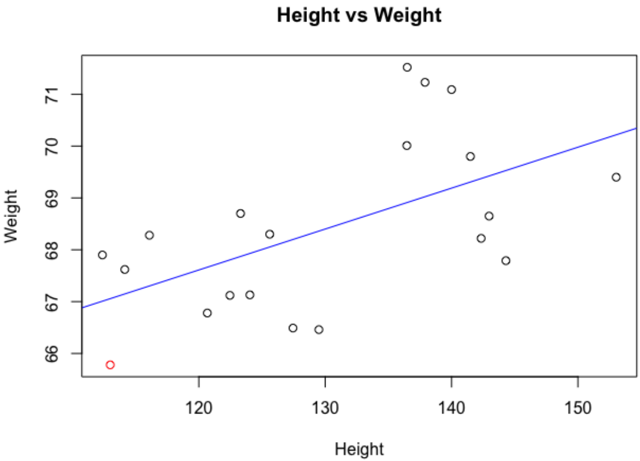
- PCA only works on numeric data. If you have categorical variables, they need to be converted to dummy variables. (Just like with linear regression).
- Data need to be normalized: because PCA is trying to find the combinations that capture the most variance, if the data are not on the same scale then the variables with the largest scale will dominate the first component which may or may not be appropriate.

# Simple Example

Height and weight for 20 people:

| Height | Weight | Height | Weight |
|---|---|---|---|
| 65.78 | 112.99 | 66.49 | 127.45 |
| 71.52 | 136.49 | 67.62 | 114.14 |
| 69.4 | 153.03 | 68.3 | 125.61 |
| 68.22 | 142.34 | 67.12 | 122.46 |
| 67.79 | 144.3 | 68.28 | 116.09 |
| 68.7 | 123.3 | 71.09 | 140 |
| 69.8 | 141.49 | 66.46 | 129.5 |
| 70.01 | 136.46 | 68.65 | 142.97 |
| 67.9 | 112.37 | 71.23 | 137.9 |
| 66.78 | 120.67 | 67.13 | 124.04 |

# Picture of data



# PCA

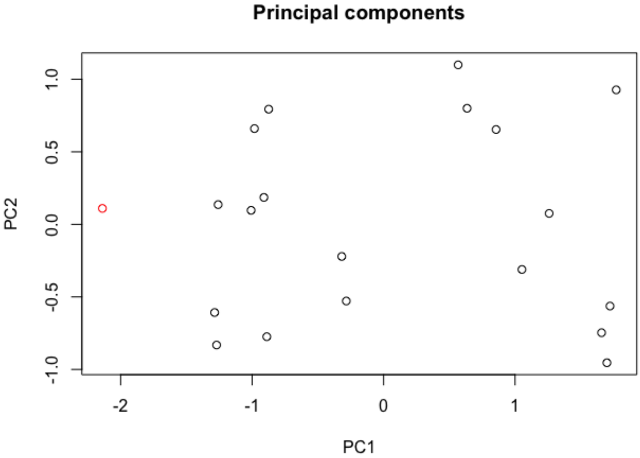
If we wanted to reduce the dimensionality of these data from two variables to one, what would that look like?

The principal components are:

- PC1: 0.7071068 * Height + 0.7071068* Weight
  Proportion of Variance: 0.7855
- PC2: - 0.7071068 * Height + 0.7071068* Weight
  Proportion of Variance: 0.2145

PC1 shows that we can use a linear combination of Height and Weight to create a single variable that captures 79% of the variation in the data.

# Picture of principle components



## Terminology

- **Loadings**: the weights used to project the original data points onto the first and second principal component direction. [0.7071068 , 0.7071068] are the loadings for the first principal component, [-0.7071068 , 0.7071068 ] are the loadings for the second.
- **Score**: the The first calculation shows the projection of the data onto the first principal component line. For observation 1, where Height = 65.78 and Weight = 112.99 , the score for the first principal component is: -1.59004784 * 0.7071068 -1.43446151* 0.7071068 = -2.138651. For PC2: 1.59004784 * 0.7071068 -1.43446151* 0.7071068 = 0.1100162.

# Working example: KC Fed's LMCI

- KC Fed economists wanted a simple way to capture many different aspects of the labor market
- Labor Market Conditions Indicators (LMCI): takes 24 input indicators (unemployment rate, quits rate, hires rate, aggregate weekly hours, etc.) and reduces them to 2 principal components
- It happens in this case that these linear combinations of the 24 variables have economic interpretations (level of activity and momentum) but this is not always the case.

Section 3

# Feature selection

# LASSO

PCA creates new variables from existing ones to try to capture as much variation as possible with the fewest variables as possible.

There are other ways to select variables (feature selection) to include/exclude to reduce dimensionality in your model.

One prominent approach which is covered in detail in Predictive Models is **Least Absolute Shrinkage and Selection Operator** or LASSO regression.

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean or zero.

The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).

# A few details

- Lasso regression performs $L1$ regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients.
- Some coefficients can become zero and eliminated from the model.
- Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models.

Goal is to minimize: $\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$

E.g. minimize the sum of squares with constraint $\sum |\beta_j| \leq s$. Some of the $\beta$s are shrunk to exactly zero.

# Another parameter choice

A tuning parameter, $\lambda$ controls the strength of the L1 penalty. $\lambda$ is basically the amount of shrinkage:

- When $\lambda = 0$, no parameters are eliminated. The estimate is equal to the one found with linear regression.
- As $\lambda$ increases, more and more coefficients are set to zero and eliminated. Theoretically, when $\lambda = \infty$, all coefficients are eliminated.
- As $\lambda$ increases, bias increases.
- As $\lambda$ decreases, variance increases.

# Sources

- http://www.lauradhamilton.com/introduction-to-principal-component-analysis-pca
- http://www.statisticshowto.com/lasso-regression/