Principal Component Analysis

Thiemo Fetzer

University of Warwick

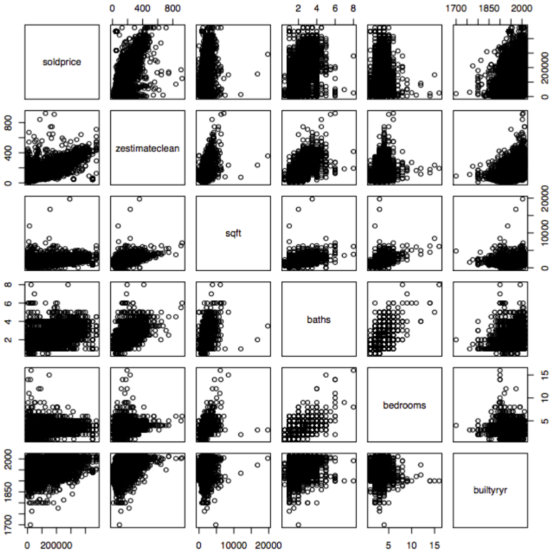
February 28, 2017

Dimensionality Reduction

The goal of dimensionality reduction or unsupervised machine learning is to summarize or simplify the data structure in a setting, where you do not have a dependent variable y to run any prediction on.

1. In the section on *Clustering*, we talked about methods and ways to find subgroups in the data that stand out by having common attributes in the feature space X.
2. This last lecture, we are briefly introduction *Principal Component Analysis*, which is a method of summarzing, visualizing and uncovering underlying relationships between covariates.

Information overflow



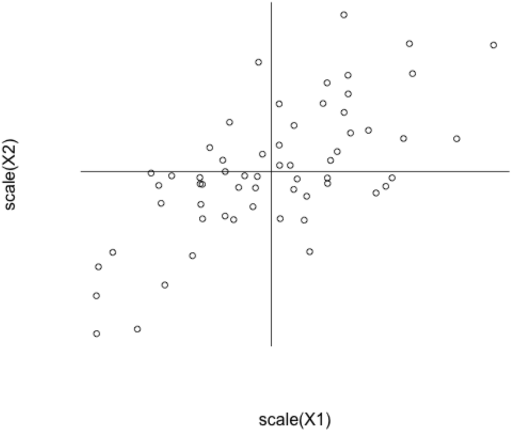
Dimensionality reduction via PCA

- The way we think of information in machine learning or data science is, that a variable X and a variable Z carry information about one another due to their correlatedness or their covariance structure.
- In many situations, you have information overflow. Suppose you have a data matrix with $p = 10$, then there is $\binom{p}{2}$, or 45 different pairs.
- Principal component analysis is a method to combine variables $X_1, ..., X_p$ into a set of variables $Z_1, ..., Z_m$ with $m < p$, without sacrificing much of the information contained in the original p variables.

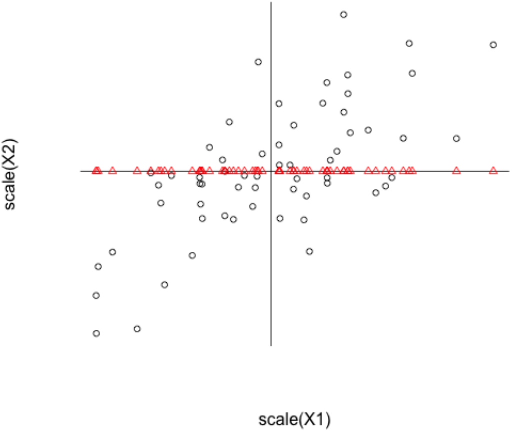
Dimensionality reduction via PCA

- The goal of PCA is to produce a low-dimensional representation of a dataset by finding a sequence of linear combbinations of the variables that have *maximal variance* and are *mutually uncorrelated*.
- In addition, PCA allows you to visualize the data in different forms. We will illustrate this using our housing price data.

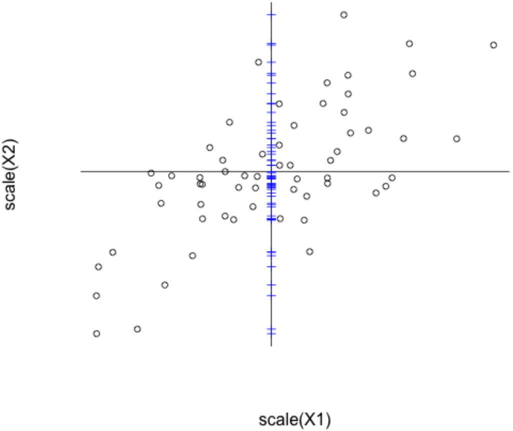
Principal Component Analysis



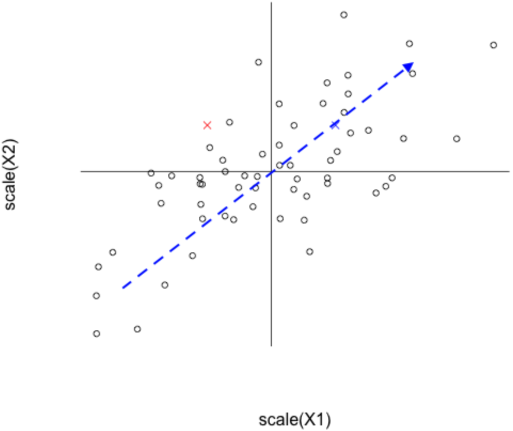
Principal Component Analysis



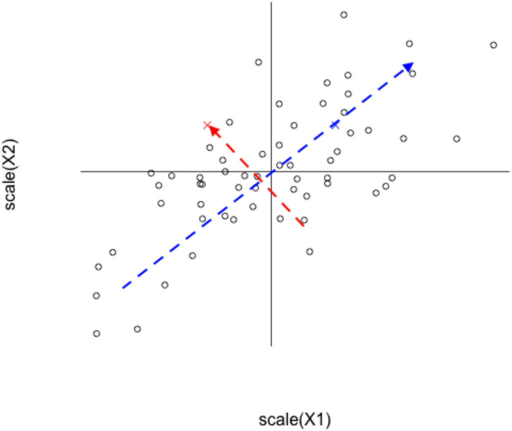
Principal Component Analysis



Principal Component Analysis



Principal Component Analysis



Dimensionality reduction via PCA

- The idea of PCA is, roughly speaking, to find the hyperplane - which is defined by a set of spanning vectors - along which the data *varies the most*.
- The requirement of *varying the most* means, higher order principal components will contain less information.
- So a bit more formally:

Principal Component Analysis

- The first principal component of a set of features $X_1, X_2, ..., X_p$ is the normalized linear combination of the features

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_{p1}X_p$$

- The normalization refers to the requirement that

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

.

- So the above expression is an expression of a hyperplane (remember we saw that in the SVM section).
- We refer to the elements $\phi_{11}, ..., \phi_{p1}$ as the loadings of the first principal component; together, the loadings make up the principal component loading vector.

Finding Hyerplane along which data has maximal variance

We can write the optimization problem as:

$$\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

where

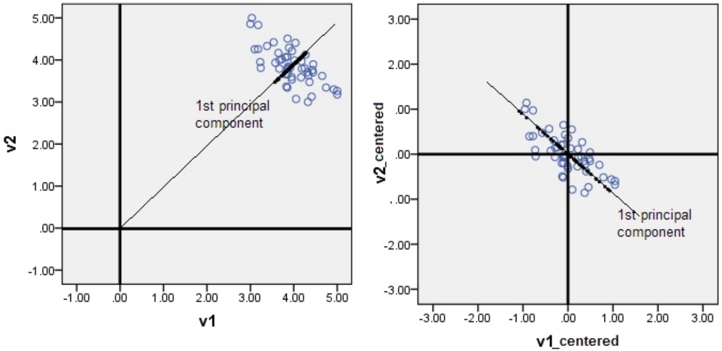
$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \ldots + \phi_{p1}x_{ip}$$

We thus choose a vector ϕ_1 to maximize the estimated sample variance $Var(Z_1)$
As with SVM, the normalization $\sum_{j=1}^{p} \phi_{j1}^2 = 1$ ensures that we obtain a hyerplane, where the function value, as we plug in the x_i's measures the orthogonal distance to that hyperplane.

Finding higher order Principal Components

- Higher order principal components need to satisfy the further constraint, that they be uncorrelated with lower order principal components.
- The second principal component of a set of features $X_1, X_2, ..., X_p$ is the linear combination that has maximal variance among all linear combinations that are uncorrelated with Z_1.
- This is identical to requiring that the loadings vector ϕ_2 is orthogonal to the loadings vector ϕ_1.
- There is a degree of arbiraryness: how many principal components to compute?
- What is the maximum number of principal components? $\min(n-1, p)$

Centering is important...



Typically, variables are standardized as well - i.e. the variances of individual variables are normalized to 1. Is this a problem?
No! Since we dont want to attribute high variability of some variable X to the fact that some variable is measured e.g. in millions of dollars, and another is measured in cents.

Applying PCA to the Housing Data

```
prcomp(HOUSES[1:1000,c("soldprice","builtyryr","bedrooms","baths","sqft"),with=F], scale=TRUE)
```

```
## Standard deviations:
## [1] 1.726 1.001 0.639 0.557 0.549
##
## Rotation:
##             PC1     PC2     PC3    PC4     PC5
## soldprice 0.472 -0.3198  0.4984  0.649  0.0699
## builtyryr 0.380 -0.6389 -0.5024 -0.161 -0.4118
## bedrooms  0.396  0.5916 -0.5516  0.432 -0.0486
## baths     0.510 -0.0275 -0.0585 -0.420  0.7479
## sqft      0.465  0.3727  0.4376 -0.435 -0.5136
```

Loading Vectors

```
prcomp(HOUSES[1:1000,c("soldprice","builtyryr","bedrooms","baths","sqft"),with=F], scale=TRUE)
```

```
## Standard deviations:
## [1] 1.726 1.001 0.638 0.557 0.549
##
## Rotation:
##               PC1     PC2     PC3    PC4     PC5
## soldprice   0.473 -0.3186  0.4929  0.654  0.0657
## builtyryr   0.380 -0.6379 -0.5003 -0.167 -0.4129
## bedrooms    0.395  0.5929 -0.5547  0.427 -0.0525
## baths       0.510 -0.0288 -0.0575 -0.415  0.7510
## sqft        0.465  0.3731  0.4425 -0.436 -0.5083
```

ϕ_1 . . . . ϕ_5

Verify whether the constraints are satisfied...

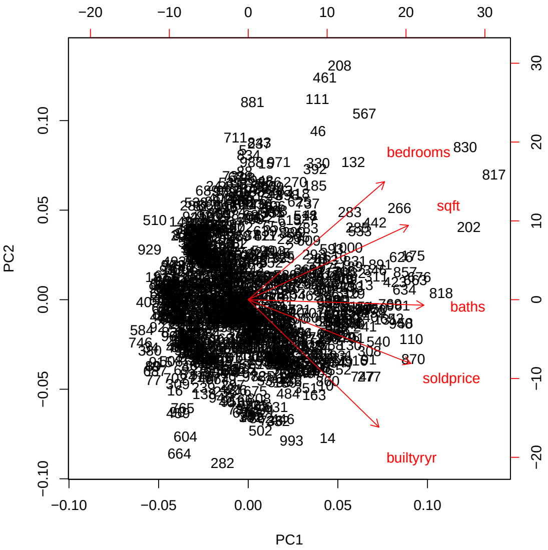
```
TEMP<-prcomp(HOUSES[1:1000,c("soldprice","builtyryr","bedrooms","baths","sqft"),with=F], scale=TRUE)
#loadings phi1 and phi2
TEMP$rotation[,1]


## soldprice builtyryr  bedrooms     baths      sqft
##     0.472     0.380     0.396     0.510     0.465


TEMP$rotation[,2]


## soldprice builtyryr  bedrooms     baths      sqft
##   -0.3198   -0.6389    0.5916   -0.0275    0.3727


##do squared loadings add to 1?
sum(TEMP$rotation[,1]^2)


## [1] 1


##what is the angle between loading vectors?
cos(TEMP$rotation[,1] %*% TEMP$rotation[,2])


##      [,1]
## [1,]    1
```

Where we remember that

$$\theta = \cos^{-1}\left(\frac{\langle \phi_1, \phi_2 \rangle}{||\phi_1||_2||\phi_2||_2}\right)$$

Visualizing Principal Components

```
biplot(TEMP)
```

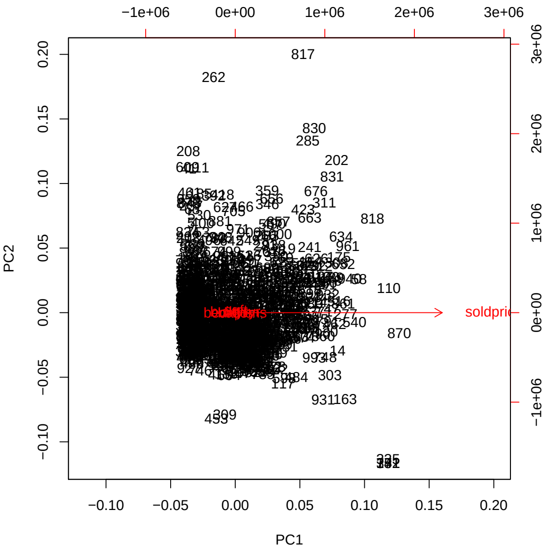


Visualizing Principal Components

- This is a bi-plot, which is plotting out the first two principal components.
- The points numbered in the background are the principal component scores for individual houses, i.e. its plotting out (z_{i1}, z_{i2}).
- The arrows indicate the loading vectors (ϕ_1, ϕ_2).
- The length of the loading tells you about the importance of that variable for a particular principal component.
  - The first PC puts very similar weight on the individual variables.
  - While PC2 has an almost zero loading for baths for the second principal component.
- Houses with large scores on the first PCA will correspond to high quality high price houses, compared to houses with negative scores.

Applying PCA to the Housing Data: Without Scaling.

```
biplot(prcomp(HOUSES[1:1000,c("soldprice","builtyryr","bedrooms","baths","sqft"),with=F], scale=FALSE))
```



Proportion of Variance Explained

The total variation in our data can be estimated as:

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

The m-th principal component has a total variation of

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2$$

Remember that $z_{im} = \phi_{1m} x_{i1} + \ldots + \phi_{pm} x_{ip}$.
One can show that for M principal components

$$\sum_{j=1}^{p} Var(X_j) = \sum_{m=1}^{M} Var(Z_m)$$

Proportion of Variance Explained

So a measure of fit may be given as:

$$\frac{\frac{1}{n}\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2}$$

Since

$$\sum_{j=1}^{p} Var(X_j) = \sum_{m=1}^{M} Var(Z_m)$$

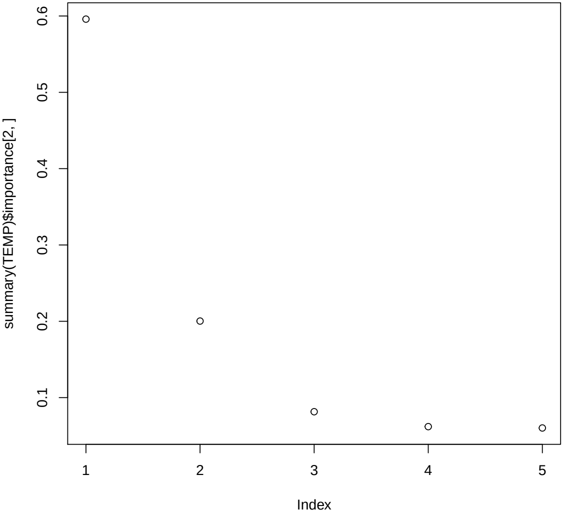
The PVEs sum to one. We sometimes display the cumulative PVEs. You can plot out the function value as a share that is explained by each principal component m.

Proportion of Variance Explained

```
summary(TEMP)

## Importance of components:
##                          PC1   PC2    PC3   PC4    PC5
## Standard deviation     1.726 1.001 0.6385 0.557 0.5488
## Proportion of Variance 0.596 0.200 0.0815 0.062 0.0602
## Cumulative Proportion  0.596 0.796 0.8778 0.940 1.0000
```

Scree Plot of Proportion of Variance Explained

```
plot(summary(TEMP)$importance[2,])
```



Some Features of PCA

- We already said, scaling matters / centering is important.
- Principal Components are unique, so as opposed to K-Means clustering, you will get an identical result on any computer as long as you have the same dataset.
- The total number of principal components is bounded as $min(n-1, p)$.
- There is no simple answer to the question as to how many principal components are adequate - typically, we look for a type of scree plot.

Latent Semantic Analysis is PCA for term-document matrices

- For term document matrices, singular value decomposition (SVD) is applied to any *mxn* matrix shape.
- Latent Semantic Analysis uses signular value decomposition to factorizes a term document matrix X into three parts $X = U\Sigma V^T X = U\Sigma V^T$
- where X is kxn (k terms in n documents).