APPLIED DATA SCIENCE

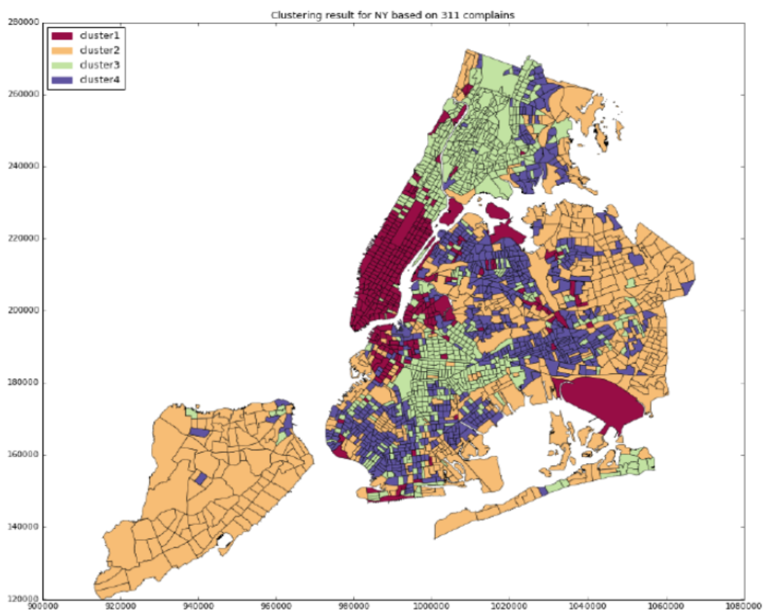
Dimensionality reduction. Principal Component Analysis

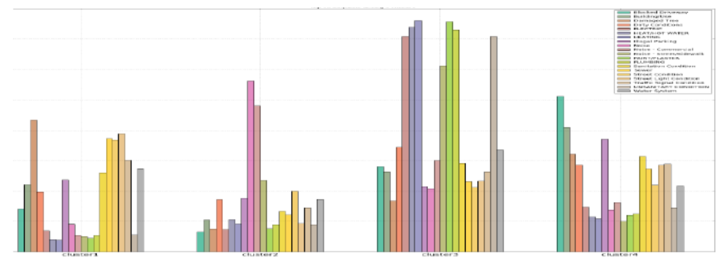
Dr. Stanislav Sobolevsky

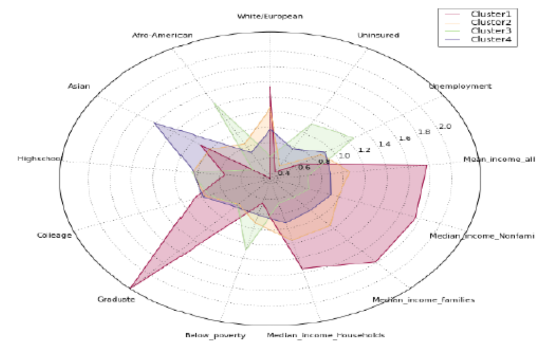
Video lectures

Principal Component Analysis

Characterize urban neighborhoods with their 311 activity
Model income, unemployment or average real estate prices

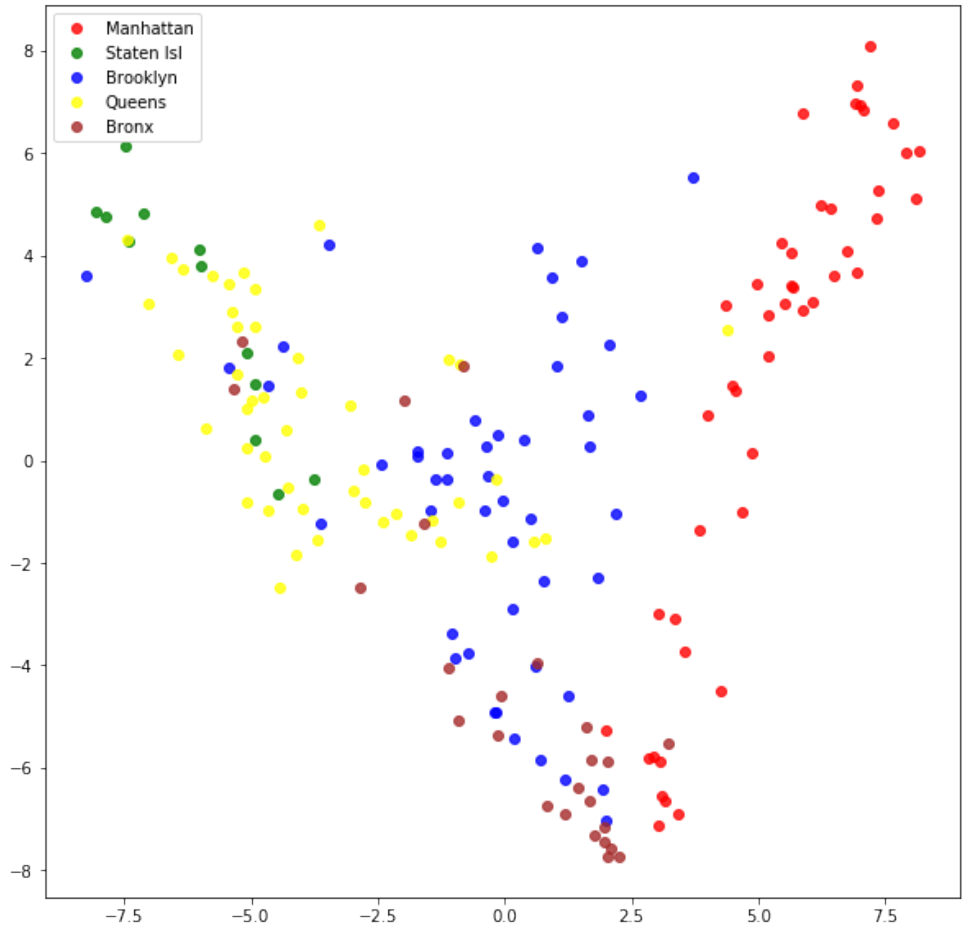






Wang L, Qian C, Kats P, Kontokosta C, Sobolevsky, S. (*corresp.) (2017) Structure of 311 service requests as a signature of urban location. PloS ONE. 12(10), e0186314.

Principal Component Analysis



Issues with multi-dimensional data

• How to analyze/visualize it?

$$x = (x_1, x_2, x_3, ..., x_n)$$

In case of regression:

$$y = f(x)$$

- complexity
- irrelevant information
- overfitting
- multi-collinearity

Feature selection vs dimensionality reduction

$$y = f(x) \qquad\qquad x = (x_1, x_2, x_3, ..., x_n)$$

feature selection reduces dimensionality of x by removing less relevant components

$$(x_1, x_2, x_3, x_4, x_5) \rightarrow (x_1, x_3, x_5)$$

dimensionality reduction looks for more general mapping

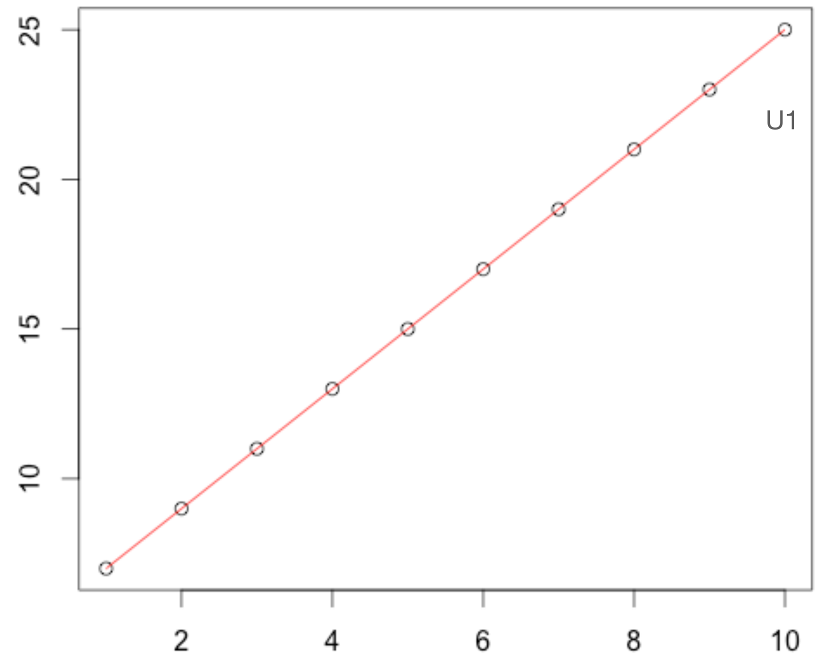
$$(x_1, x_2, x_3, ..., x_n) \rightarrow (x'_1, x'_2, x'_3, ..., x'_m),\ \ m < n$$

$$y = f(x')$$

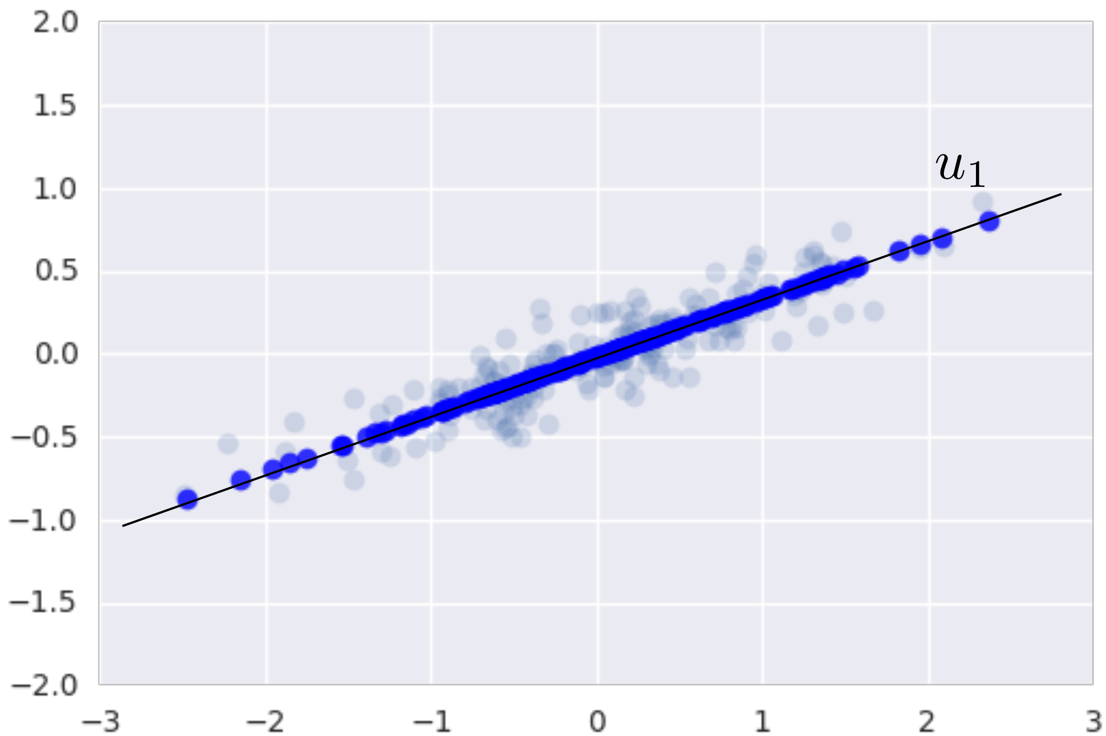
$$(x_1, x_2, x_3, x_4, x_5) \rightarrow x' = (x_1 + x_2 + x_3 + x_4 + x_5, x_1 x_2 x_3 x_4 x_5)$$

Pareto rule: 20% information often provides 80% of value

Same information with smaller number of parameters



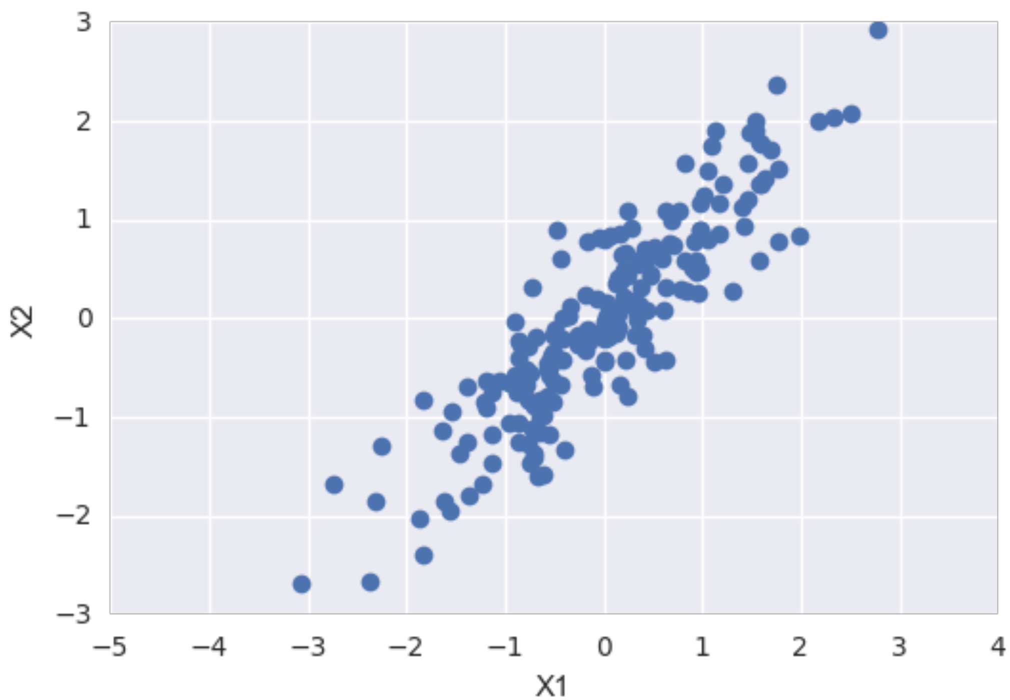
Almost the same information



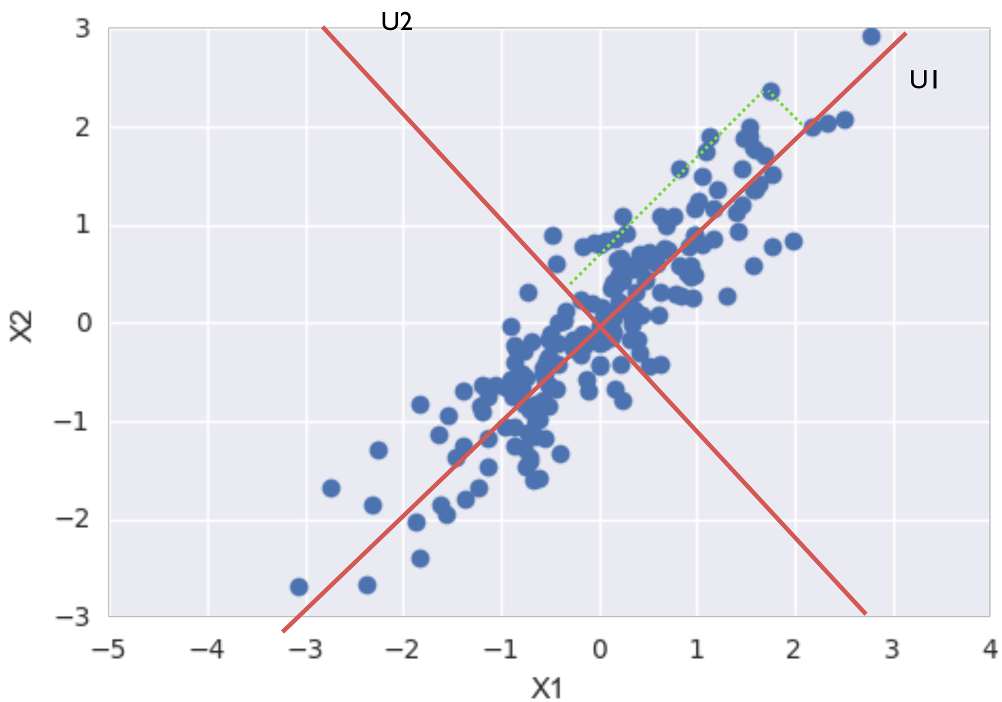
correlated features to uncorrelated

$$(x_1, x_2, x_3, ..., x_n) \rightarrow (u_1, u_2, u_3, ..., u_n)$$

Original data

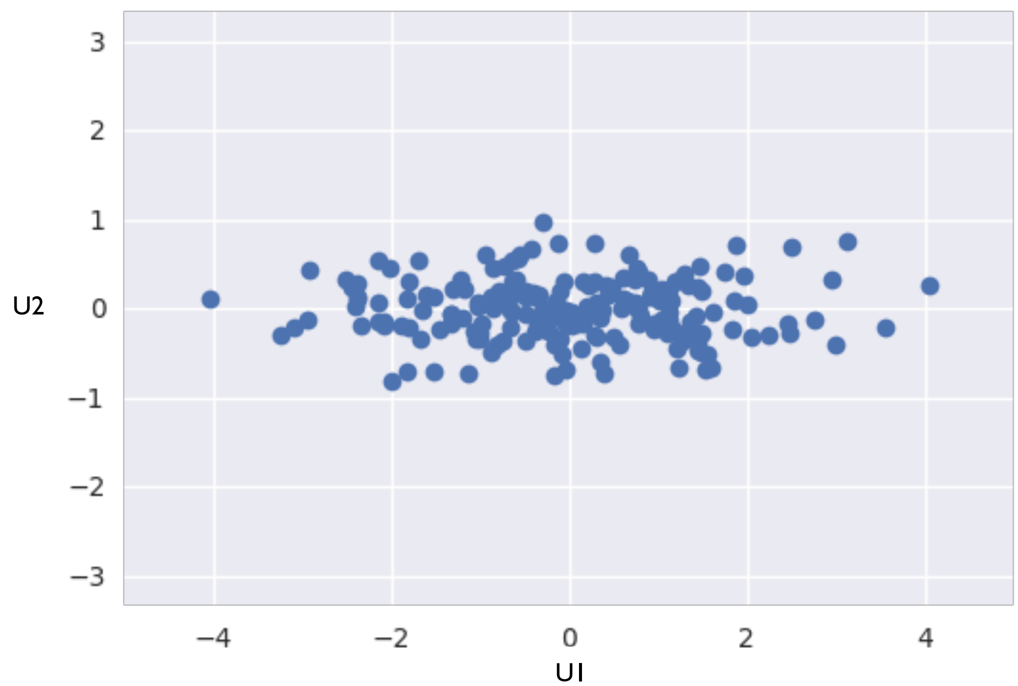


Original data - new system of coordinates

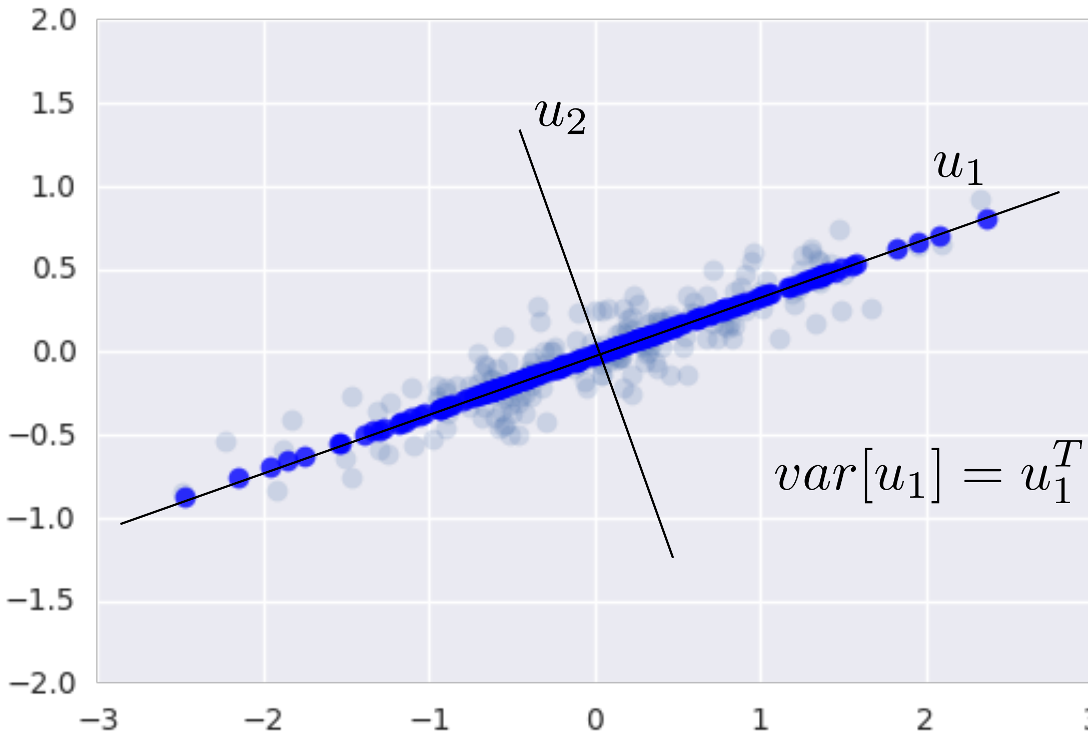


Uncorrelated data (rotation)

$$U = XV$$



Almost the same information



$$var[u_1] = u_1^T u_1 \rightarrow max$$

Principal components - maths

Given the standardized data $X = \{x_i^j, i = 1..n, j = 1..N\}$

Find uncorrelated latent features

$$u_j = x_1 v_j^1 + x_2 v_j^2 + ... + x_n v_j^n$$

$$Var[u_1] \geq Var[u_2] \geq \ldots \geq Var[u_n]$$

Learn matrix V implementing the rotation transform

$$u_i = X v_i \qquad U = XV$$

V - n x n
U - N x n

V - matrix of eigenvectors of $X^T X$

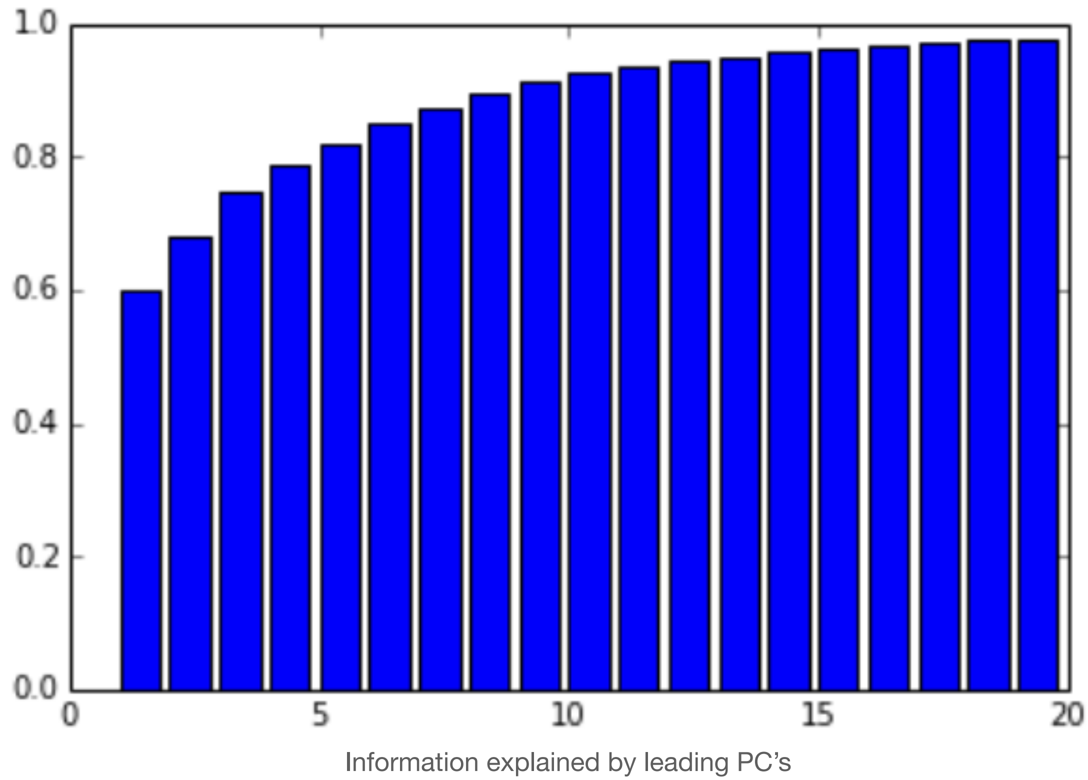
Principal components - select by variation

Information explained by a PC u_i

$$Var[u_i] = \lambda_i$$

% of info explained by leading k components

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$



Leading PCs - uncorrelated low-dimensional feature space:

- Data exploration (e.g. visualization)
- Modeling

Limitations:

- Relevance to a particular target variable
- Interpretability
- Linear