Applied Data Science
fall 2017
Session 6: Dimensionality reduction. Principle component analysis

Instructor: Prof. Stanislav Sobolevsky
Course Assistants: Tushar Ahuja, Maxim Temnogorod

Issues with multi-dimensional data

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

- complexity
- irrelevant information
- multi-collinearity
- overfitting

- not only for regression: understanding, even visualizing multi-dimensional data is hard

Skinnier data is often better

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

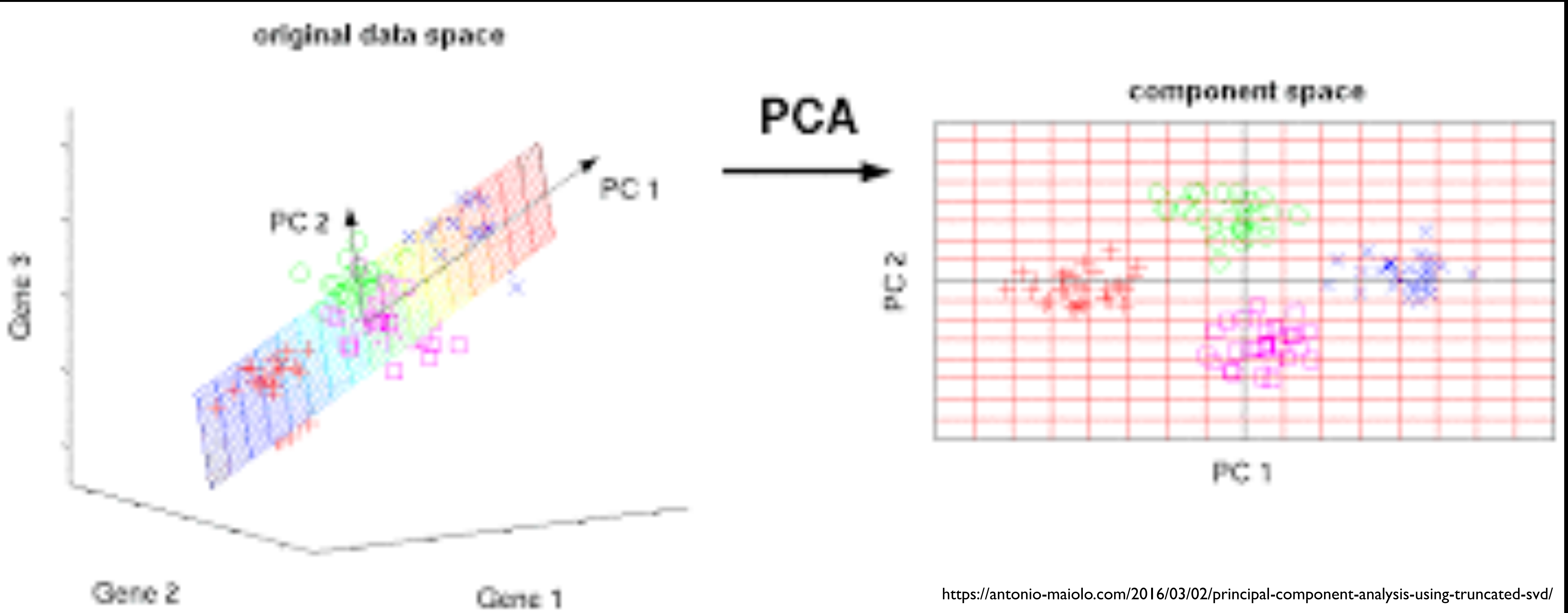
$$(x_1, x_2, x_3, x_4, x_5) \to (x_1, x_3, x_5)$$

- dimensionality reduction looks for more general mapping

$$(x_1, x_2, x_3, ..., x_n) \to (x'_1, x'_2, x'_3, ..., x'_m),\ \ m < n$$
$$y = f(x')$$
$$(x_1, x_2, x_3, x_4, x_5) \to x' = (x_1 + x_2 + x_3 + x_4 + x_5, x_1 x_2 x_3 x_4 x_5)$$

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

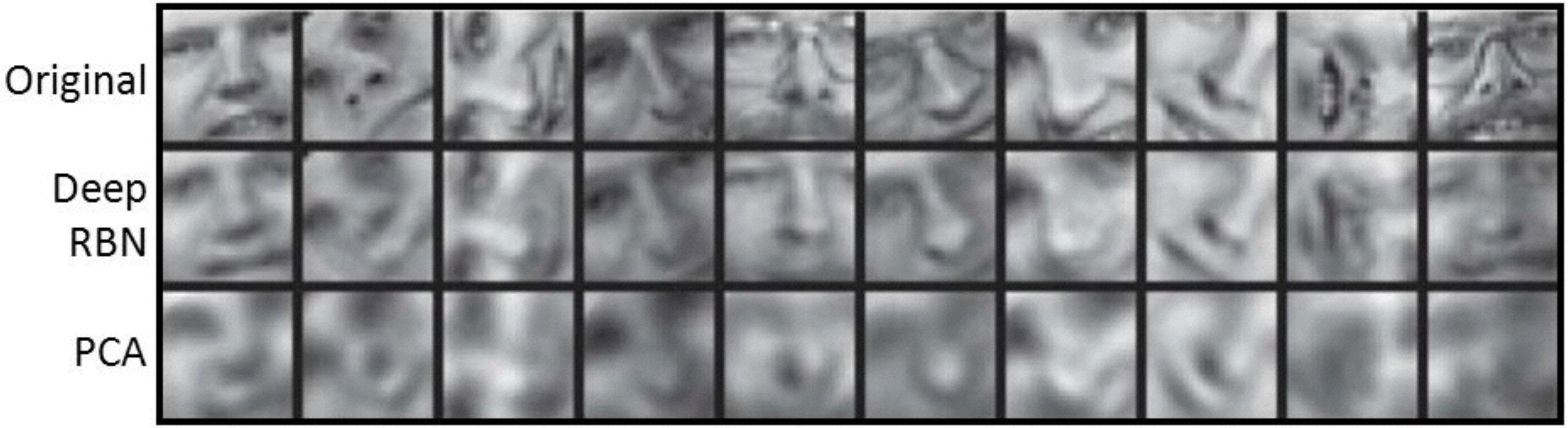
Dimensionality reduction



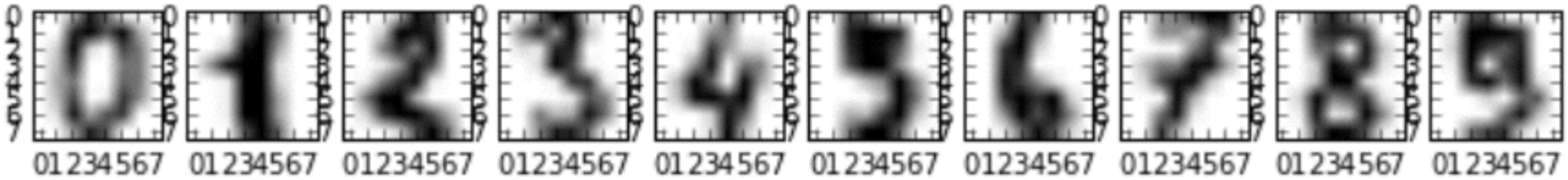
https://antonio-maiolo.com/2016/03/02/principal-component-analysis-using-truncated-svd/

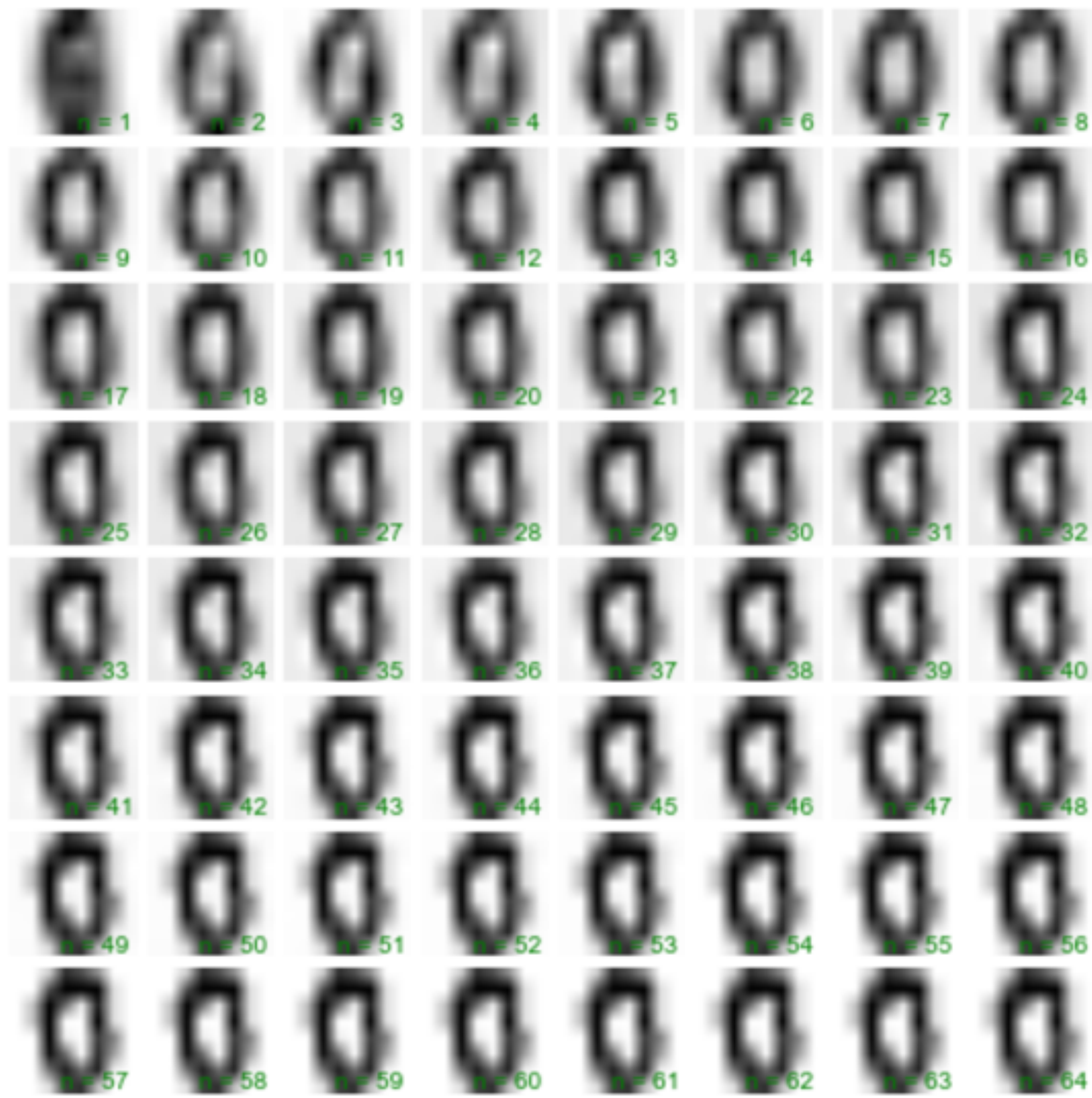
Dimensionality reduction - images

Olivetti face data, 25x25 pixel images reconstructed from 30 dimensions (625 → 30)



Principle components of an image





Example, dimensionality reduction - 311 service requests

>180 categories: 180-dimensional data

Can we use all of them for regression?

Do we need all of them to characterize the user? location?

3 parameters may largely explain 180-dimensional data

$$R_i^j = k_1^j age_i + k_2^j gender_i + k_3^j wealth_i + \varepsilon_{i,j}$$

What if we do not know demography?

Can we infer factors that matter?

Principal components

Correlation between factors is a major issue

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

Find uncorrelated latent factors U

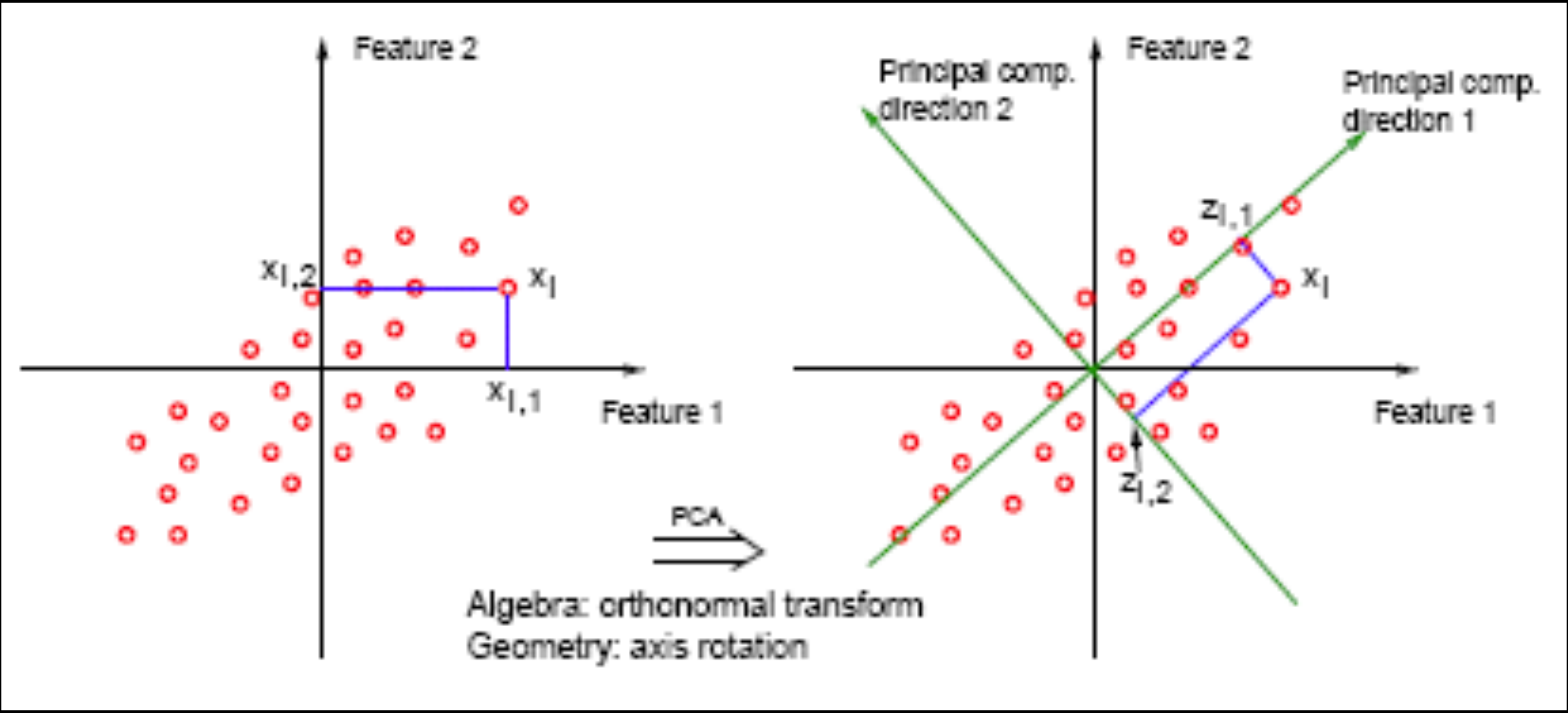
$$u_j = x_1 v_j^1 + x_2 v_j^2 + \ldots + x_n v_j^n$$

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

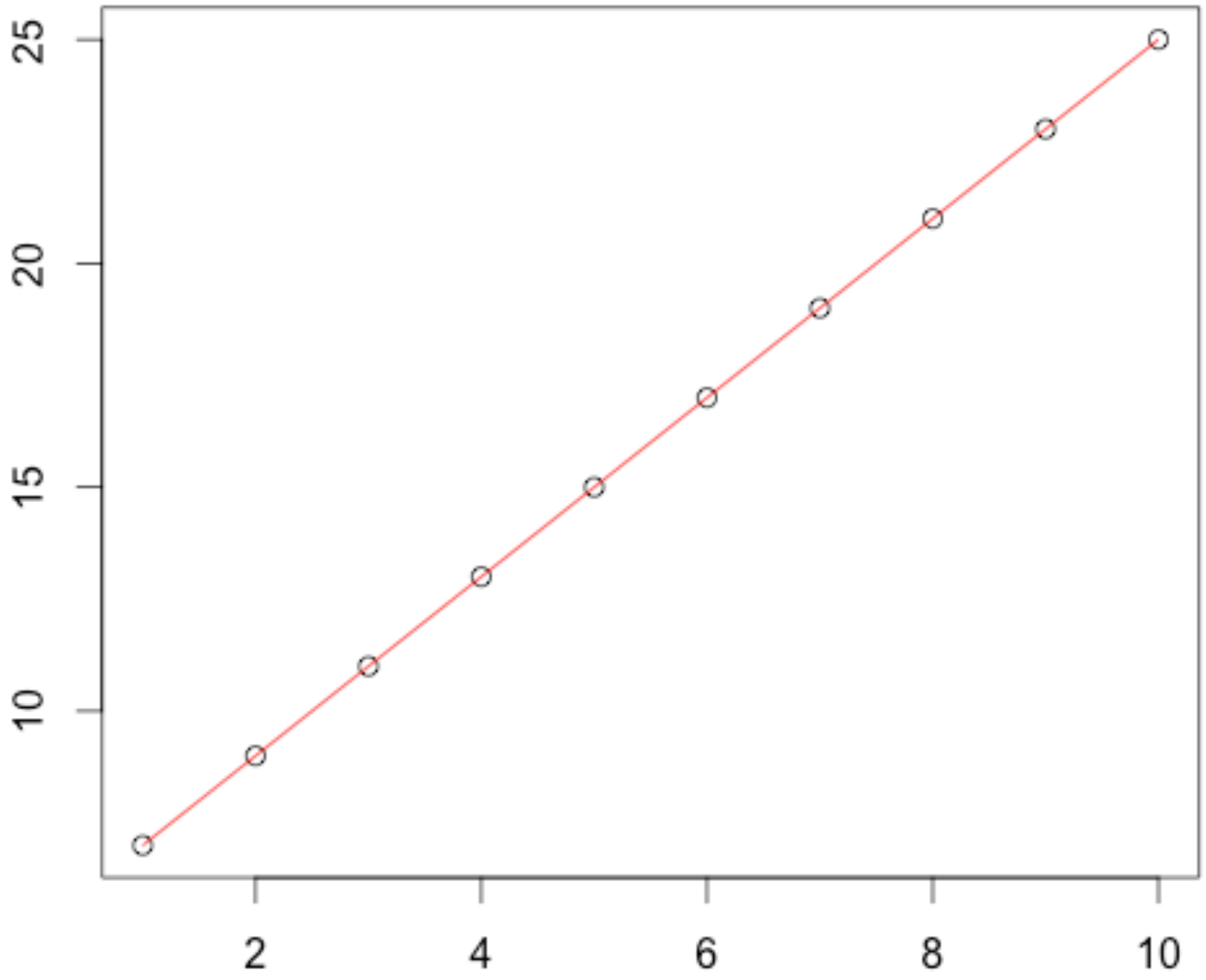
V - n x p
U - N x p

Look for linear combinations of factors one-by-one

Principle component analysis



Principal components



Principal components - technique

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

$$x_1, x_2, x_3, ..., x_n$$

$$u_1 = X v_1$$

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

$$v_1 = argmax_{v_1 : v_1^T v_1 = 1} var[u_1] = argmax_{v_1 : v_1^T v_1 = 1} u_1^T u_1 = argmax_{v_1 : v_1^T v_1 = 1} v_1^T X^T X v_1$$

$$v_i = argmax_{v_i : v_i^T v_i = 1, v_i^T v_j = 0, j < i} v_i^T X^T X v_i$$

Recall the concept of eigenvectors/eigenvalues

$$\lambda v = Av \qquad \lambda - eigenvalue, v - eigenvector$$

$$(\lambda I - A)v = 0 \qquad det(\lambda I - A) = 0$$

$$\lambda_1, \lambda_2, ...\lambda_n \qquad v_1, v_2, ...v_n \qquad v_i \to Cv_i \qquad |v_i| = 1$$

$$A^T = A \qquad \lambda_i \neq \lambda_j \;\Rightarrow\; v_i^T v_j = 0 \qquad v_i^T A v_j = \lambda_j v_i^T v_j$$

$$v_i^T A v_j = (Av_i)^T v_j = \lambda_i v_i^T v_j$$

Principal components - technique

$$v_i = argmax_{v_i : v_i^T v_i = 1, v_i^T v_j = 0, j<i} v_i^T X^T X v_i$$

Consider eigenvectors:

$$\lambda_i v_i = X^T X v_i \qquad v_i^T v_i = 1 \qquad \lambda_1 > \lambda_2 > \ldots > \lambda_n > 0$$

$$v_i^T X^T X v_i = \lambda_i v_i^T v_i = \lambda_i$$

$$w = e_1 v_1 + e_2 v_2 + \ldots + e_n v_n \qquad w^T w = e_1^2 + e_2^2 + \ldots + e_n^2 = 1$$

$$w^T X^T X w = \lambda_1 e_1^2 + \lambda_2 e_2^2 + \ldots + \lambda_n e_n^2 \rightarrow \max$$

$$w = v_1, e_1 = 1, e_2 = e_3 = \ldots = e_n = 0$$

Principal components - technique

$$v_1, v_2, \ldots v_n \qquad v_i = argmax_{v_i : v_i^T v_i = 1, v_i^T v_j = 0, j < i} v_i^T X^T X v_i$$

$$\lambda_i v_i = X^T X v_i \qquad \lambda_1 > \lambda_2 > \ldots > \lambda_n > 0$$

$$v_i^T X^T X v_i = \lambda_i v_i^T v_i = \lambda_i \qquad v_i^T v_i = 1 \qquad v_i^T v_j = 0$$

$$u_i = X v_i \qquad Var[u_i] = u_i^T u_i = v_i X^T X v_i = \lambda_i v_i^T v_i = \lambda_i$$

$$u_i^T u_j = v_i X^T X v_j = \lambda_j v_i^T v_i = 0$$

Principal components - singular value decomposition

$$diag(\lambda)V = X^T X V$$

$$V^T V = I_n$$

$$X = W \Sigma V^T$$

$$W^T W = V^T V = I_n$$

$$X^T X = V \Sigma W^T W \Sigma V^T = V \Sigma^2 V^T$$

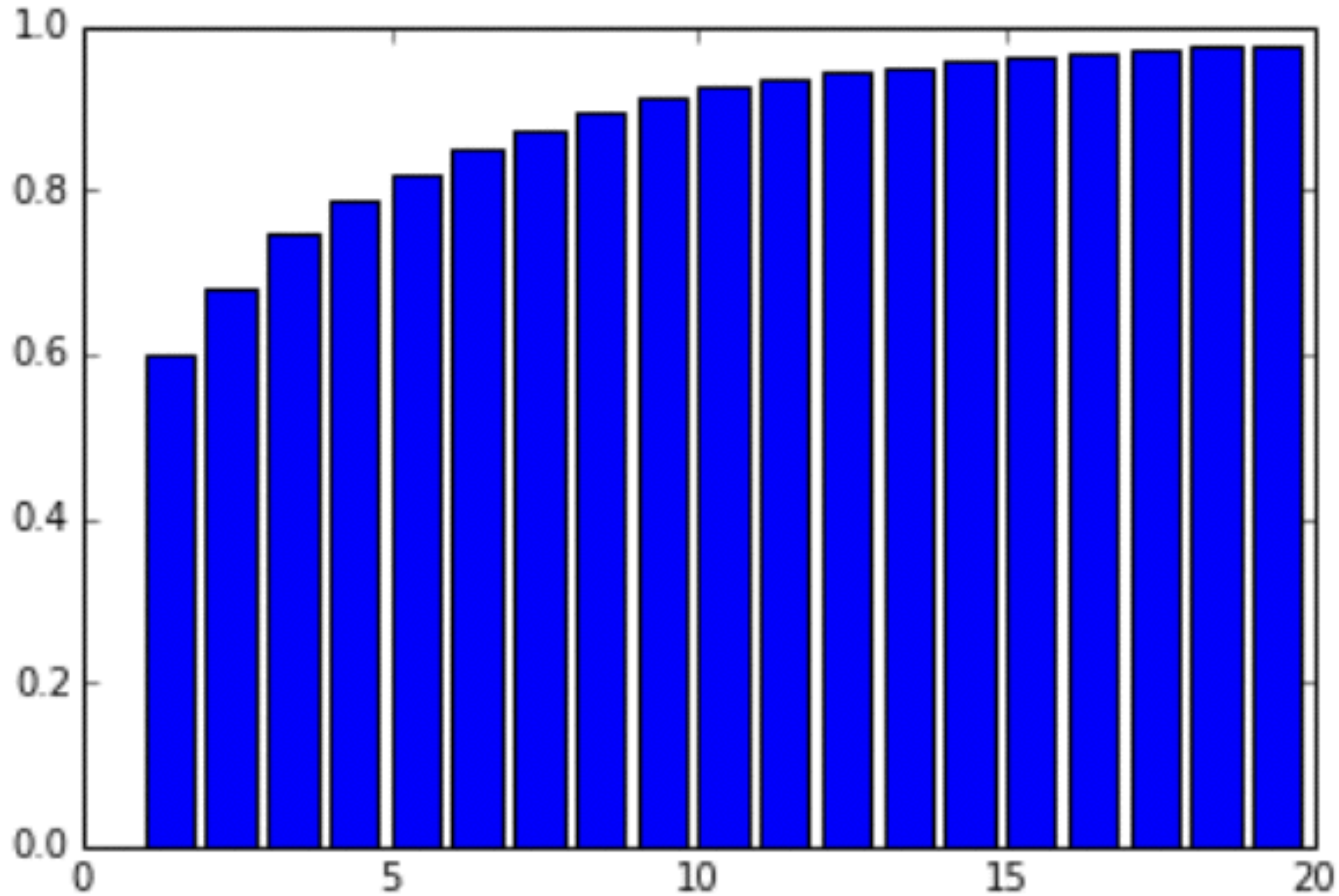
$$U = XV = W \Sigma V^T V = W \Sigma$$

Principal components - select by variation

$$Var[u_i] = \lambda_i$$

$$u_i \qquad \lambda_i / \sum_j \lambda_j$$

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



Applications of PCA - visualization

Application of PCA

- Discover patterns behind the data
  - visualize multi-dimensional data
  - clustering
  - latent variables

Applications in finance, neurobiology, healthcare, image recognition, signal processing etc

- Feature selection in regressions
  - principle component regression

Principle component regression

$$Y \sim X \qquad X \to P \qquad Y \sim P$$

PC's are intrinsic for feature space X

- Leading PC's not necessary relevant for Y (although signal-to-noise ratio is often higher)
- Feature selection after PCA (backward/forward step-wise)
- Feature selection based on p-values