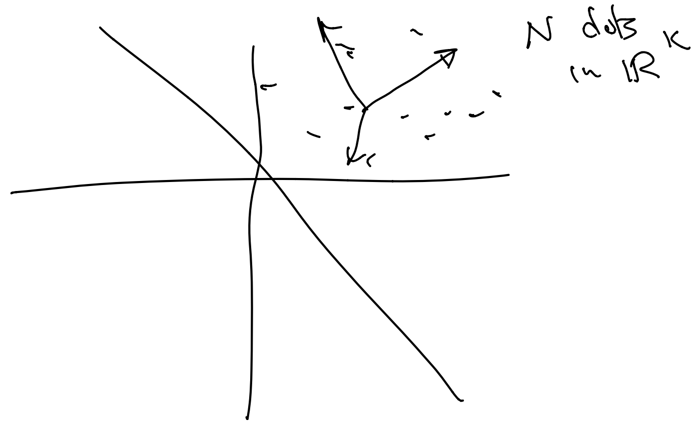
Dimensionality Reduction

Dimensionality Reduction – Preliminaries

- Typical dataset can be represented as N points in k-dimensional space, where N and k are large.
- Difficult to visualize
- Hard to extract meaningful information

Principal Component Analysis identifies "directions" in $\mathbf{R}^k$ that most effectively spread out the data points by maximizing the variance in that direction.



Principal Directions

- Given data X_0 with covariance matrix D_0, where the number of samples is N and the number of features is k.

$D_0 = \frac{1}{N} X_0^T X_0$

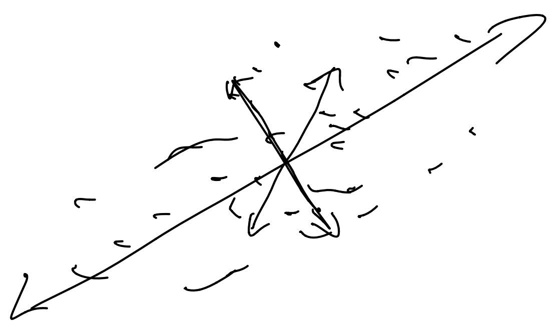
$\uparrow_{k \times k}$

- The *principal directions* in the data are the orthonormal eigenvectors $u_1, \ldots, u_k$ of D_0 and the variance in the u_i direction is $\sigma_i^2 = \underline{\lambda_i}$ where λ_i is the eigenvalue corresponding to u_i. We assume that

$$\underline{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0}$$

$u_1 \qquad \sigma_1^2 = \lambda_1$

- From our earlier work we know that σ_1^2 is the largest variance associated to any score S of unit norm, and σ_k is the smallest.

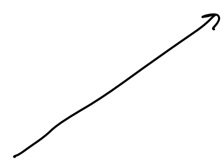


u_1

$\sigma_1^2 = \lambda_1$

Subspaces of maximal variance

Theorem: Let U be the span of eigenvectors corresponding to s of the largest eigenvalues of D_0. (Since the eigenvalues need not be distinct, there may be several choices for U). Then the total variance σ_U^2 of the data projected into U is $\sum_{i=1}^{s} \lambda_i$, and this is the largest total variance among all subspaces of dimension s.



Projection into subspaces of maximal variance

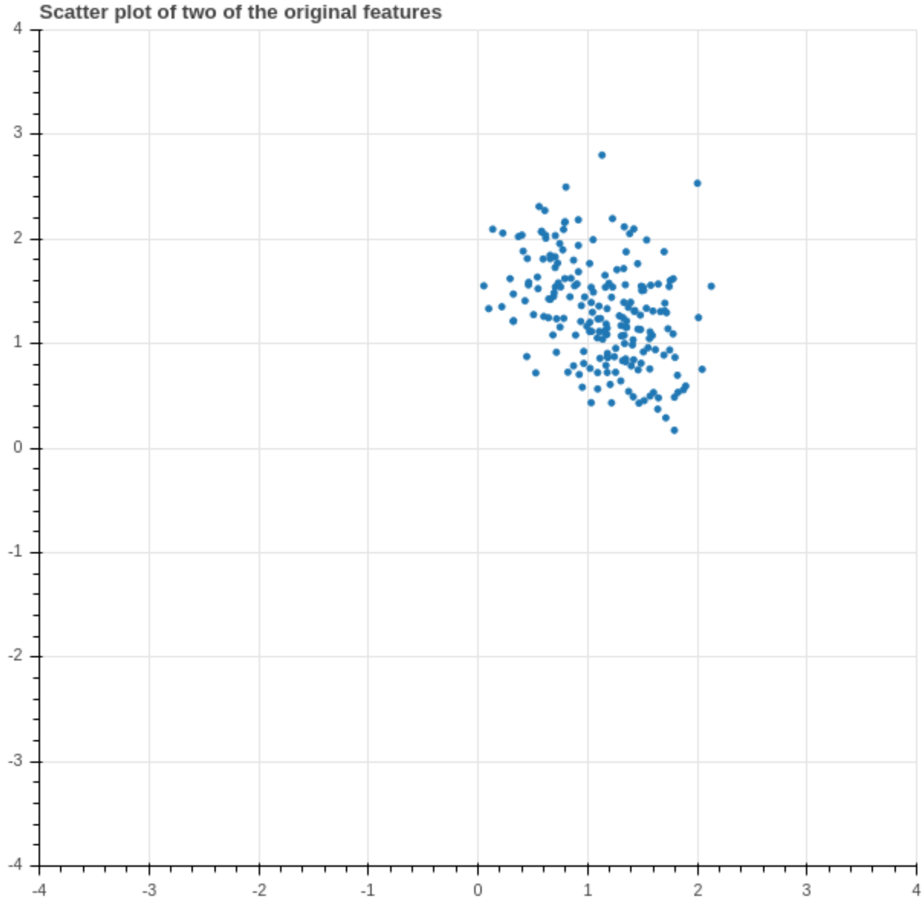
Strategy: Given data in a high dimensional space, project it into a much lower dimensional space that still captures a high percentage of the total variance.

Example

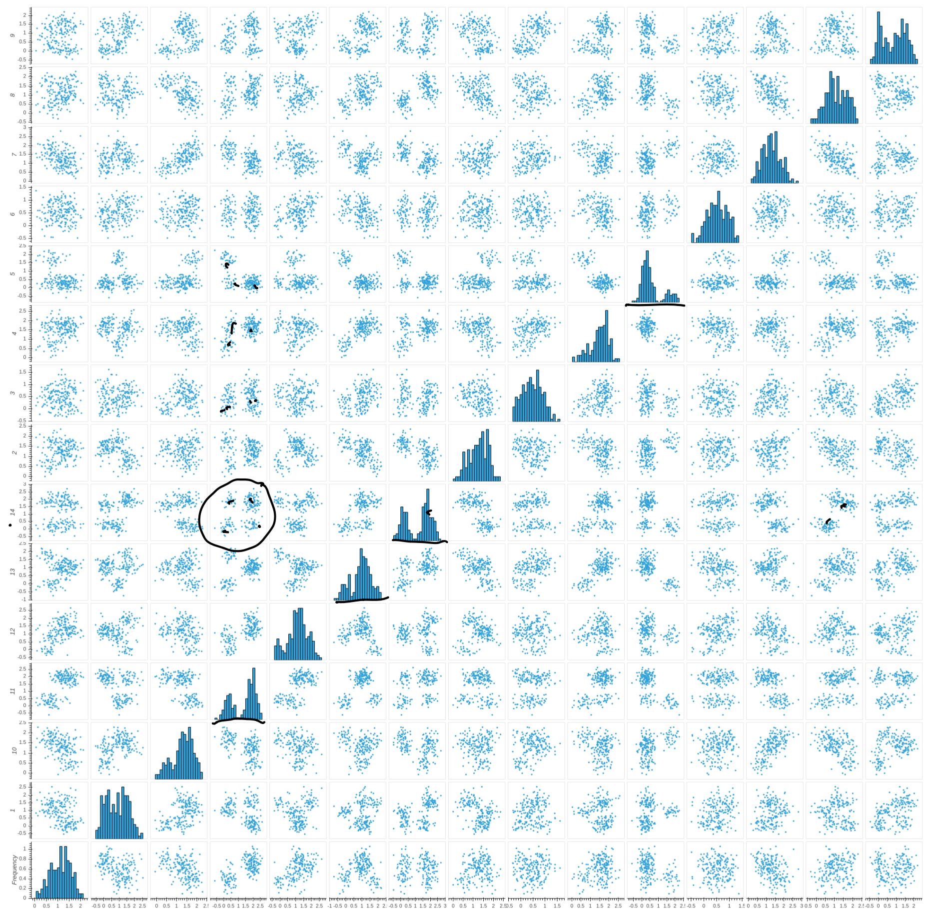
We have a 200 points with 15 features, so a 200×15 matrix with column sums equal to zero. 3000 numbers total. How to make sense of it?

$N = 200$
$K = 15$
3000 pb in total

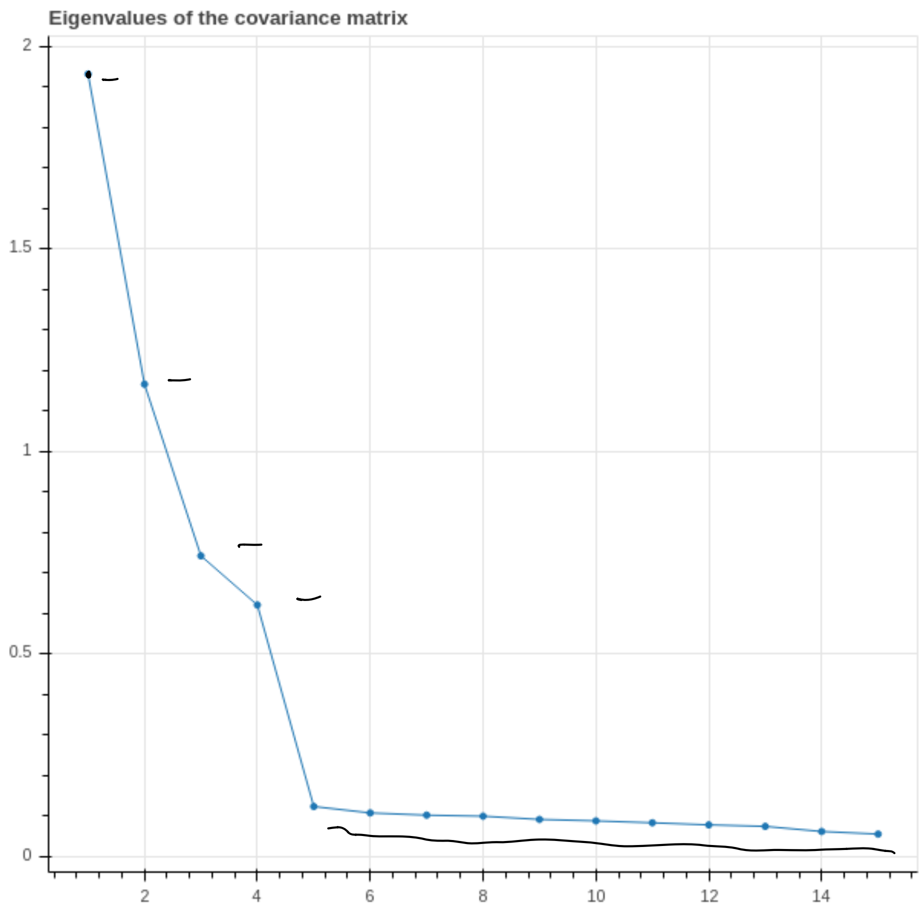
First try. Scatter plot of two features.



Density Plot



Eigenvalues



The first four eigenvalues account for 80% of the variance.

X_0 centered

$D_0 = \frac{1}{N} X_0^T X_0$

Compute

$\lambda_1, \ldots, \lambda_K$

each u is $K \times 1$ vector

u_1, u_2, u_3, u_4 $\|u_i\|^2 = 1$

U = columns

are u_1, u_2, u_3, u_4

$K \times 4$ matrix

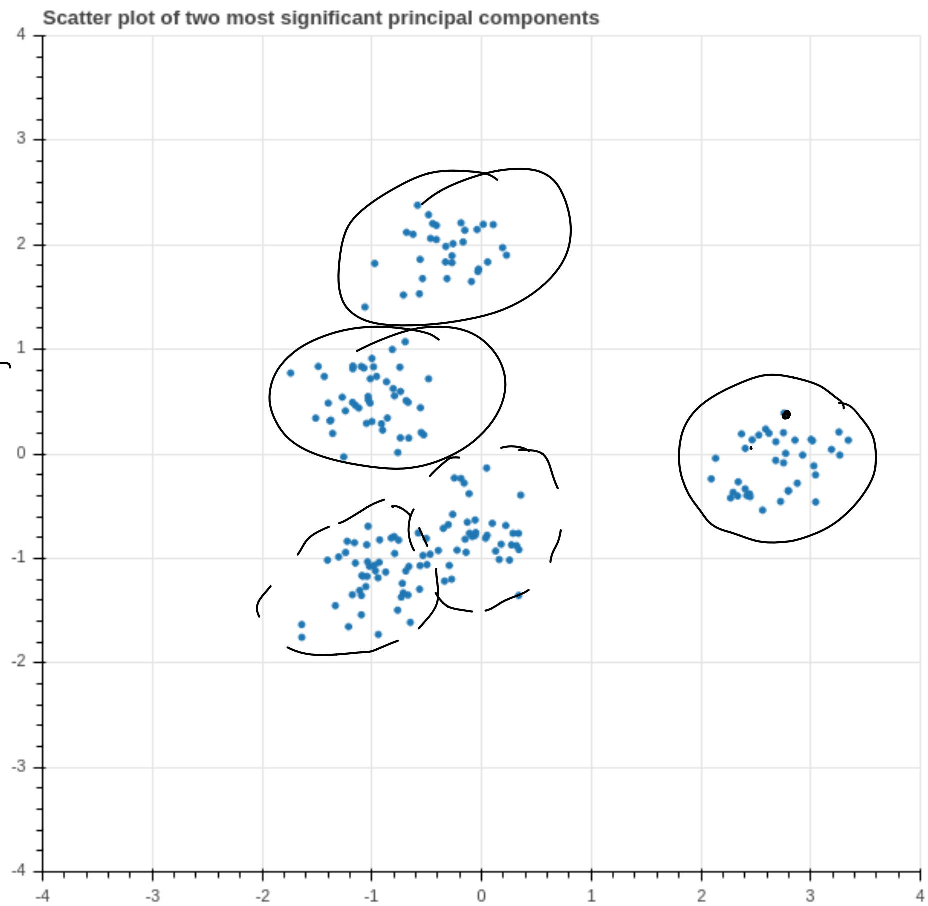
$X_0 U$ yields

an

$N \times 4$ matrix

$$\frac{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}{\sum_{i=1}^{15} \lambda_i} \sim .8$$

Two principal directions



2nd PC

units of PC_1

2 largest eigenvalues

$\lambda_1 \; \lambda_2$

u_1, u_2

$X_0 U$

U columns are u_1, u_2.

data has 4 or 5 clusters hidden in it.