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

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$$

- 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.

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.

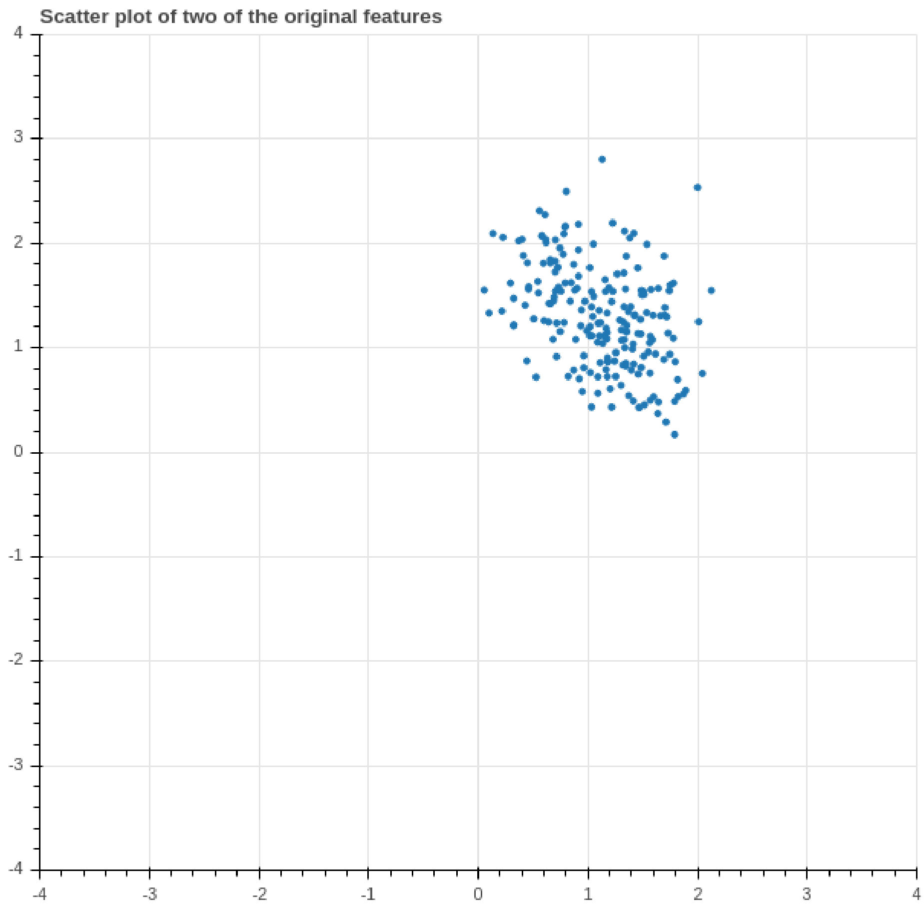
Projection into subspaces of maximal variance

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.

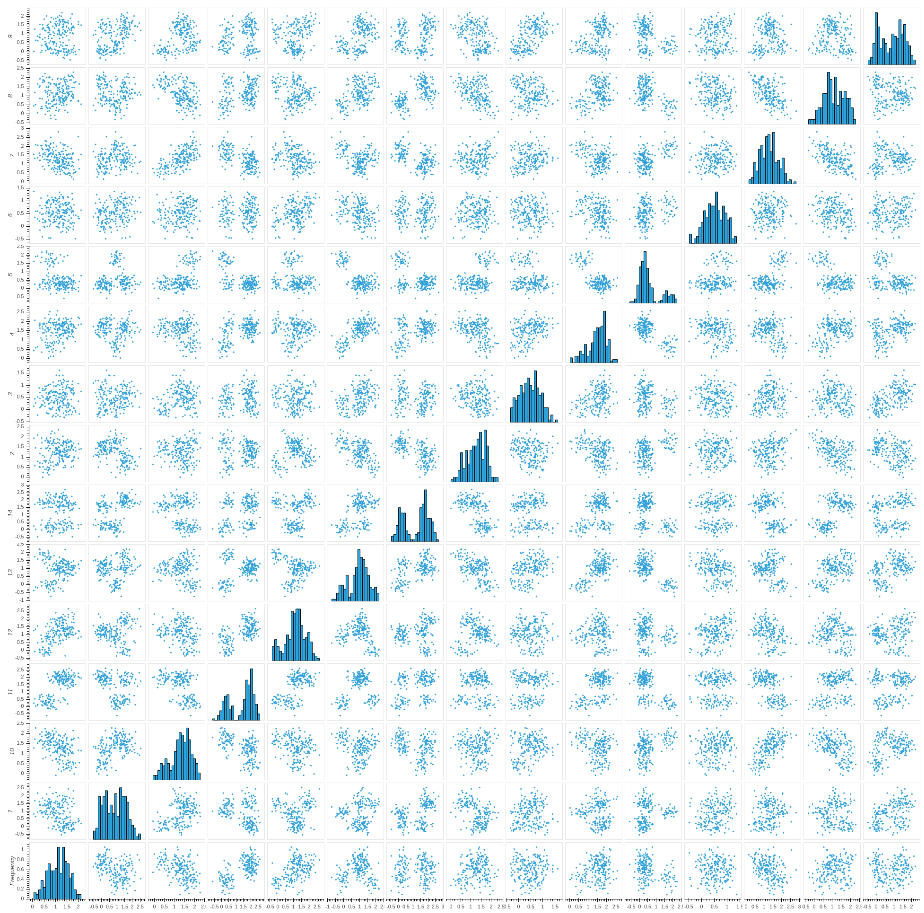
Example

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?

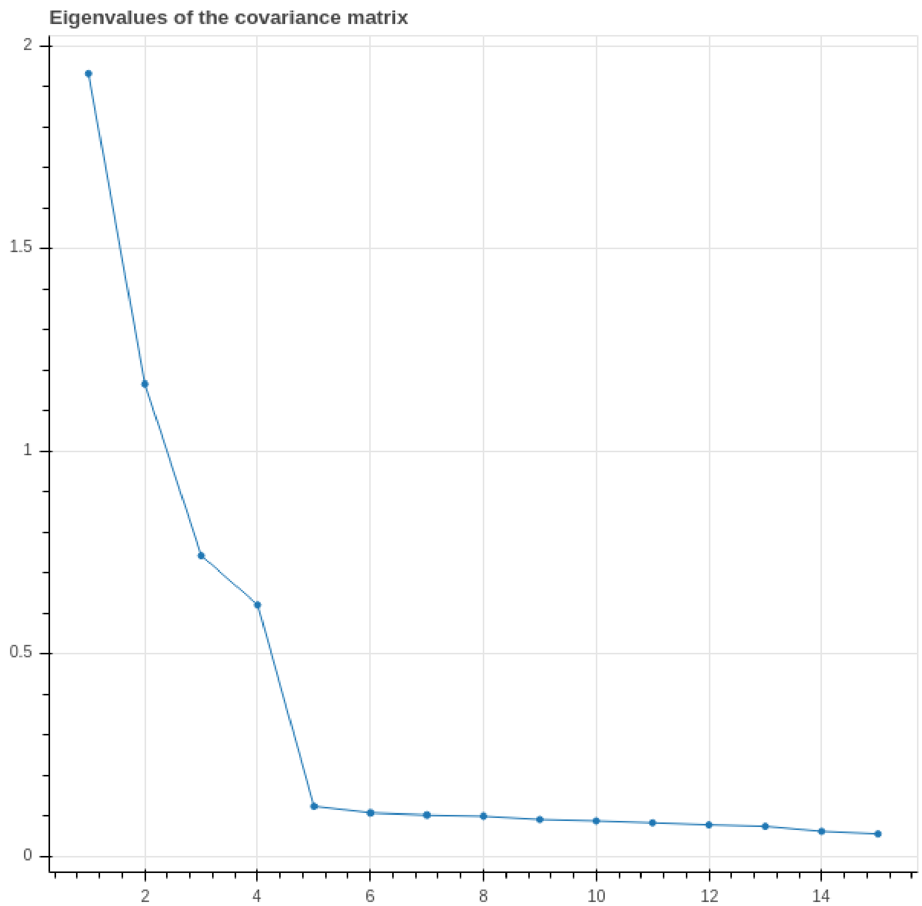
First try. Scatter plot of two features.



Density Plot



Eigenvalues



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

Two principal directions