Data Preprocessing

Data Preprocessing

Why?

- Reduce scale effects
- Reduce noise
- Reduce redundancy

Data Preprocessing
Reducing scale effects

Normalization – map every feature to the [0,1] interval

For every feature i in data set:

- max_val = max(X_train[:,i])
- min_val = min(X_train[:,i])
- X_train_n[:,i] = (X_train[:,i] – min_val) / (max_val – min_val)
- X_test_n[:,i] = (X_test[:,i] – min_val) / (max_val – min_val)

Advantages:

- Simplicity
- Same interval for all feature values

Disadvantages:

- Sensitive to outliers

Data Preprocessing
Reducing scale effects

Standardization (a.k.a. z-normalization) – scale every feature to have zero mean and unit standard deviation

For every feature i in data set:

- m = mean(X_train[:,i])
- s = s(X_train[:,i])
- X_train_n[:,i] = (X_train[:,i] – m) / s
- X_test_n[:,i] = (X_test[:,i] – m) / s

Advantages:

- Simplicity
- Tends to produce good model performance

Disadvantages:

- Sensitive to outliers (not as much as normalization)

Data Preprocessing
Reducing noise and redundancy

Principal Component Analysis

Idea: Project data from n-dimensions (attributes) to m-dimensions (with n>m) while preserving as much information as possible

This is achieved by a simple combination of translation and rotation

Why: Features tend to be correlated – reduce redundancy

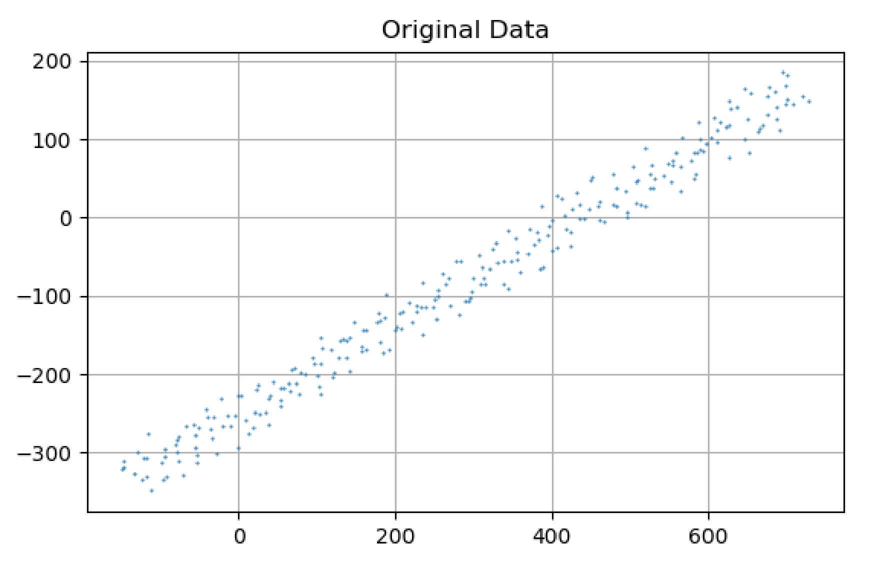
Why: Least important features in projected space approximate noise

Data Preprocessing
Reducing noise and redundancy

Principal Component Analysis

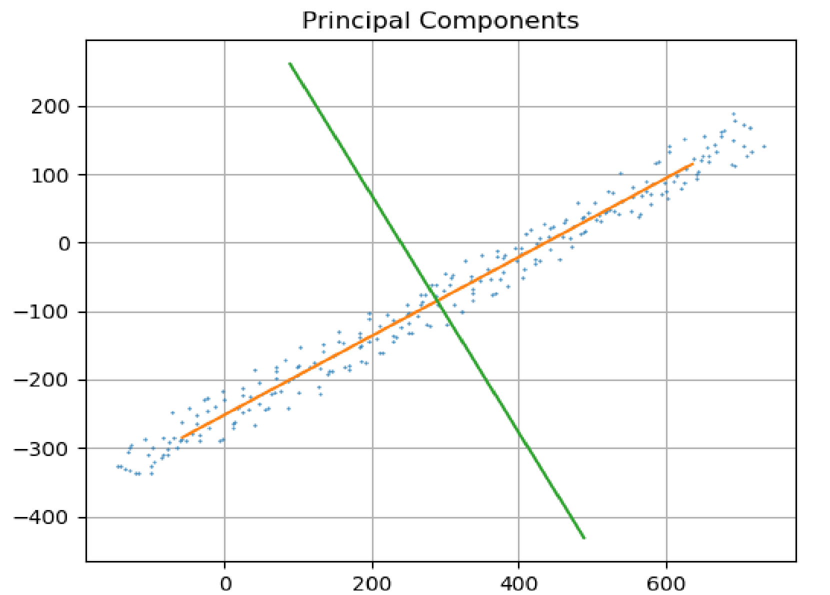
The principal components of a dataset X are the vectors $v_0, \ldots, v_n$

such that v_i is the vector that best fits X and is perpendicular to each of $v_0, \ldots, v_{i-1}$



X

PCA

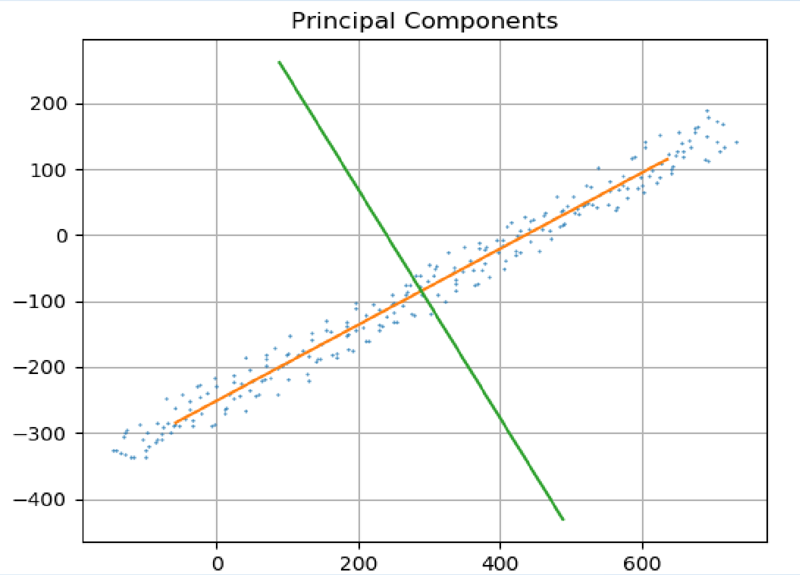


v_0 – orange, v_1 - green

Data Preprocessing
Reducing noise and redundancy

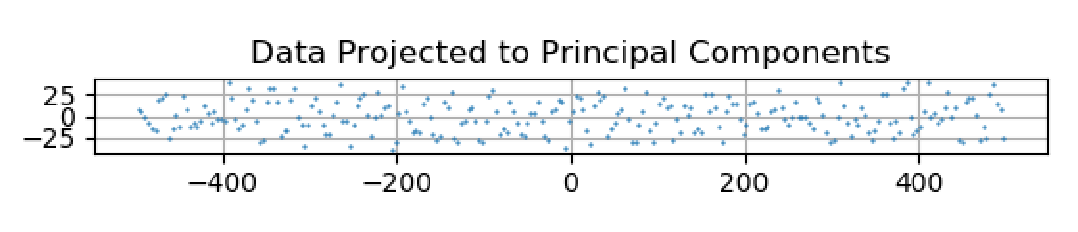
Principal Component Analysis

The principal components of a dataset X are the vectors v_0, ..., v_n
such that v_i is the vector that best fits X and is perpendicular to each of v_0,...,v_{i-1}



X in original feature space and principal components

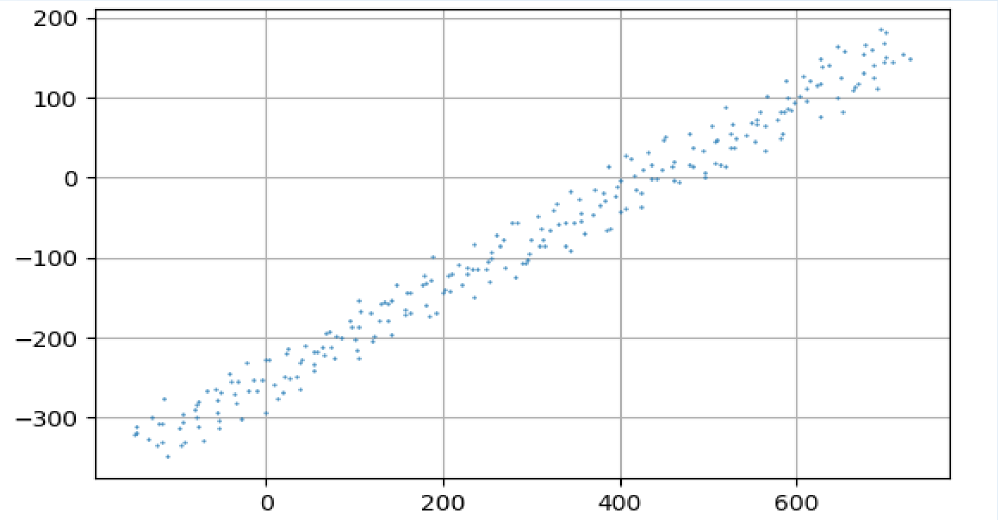
Projection



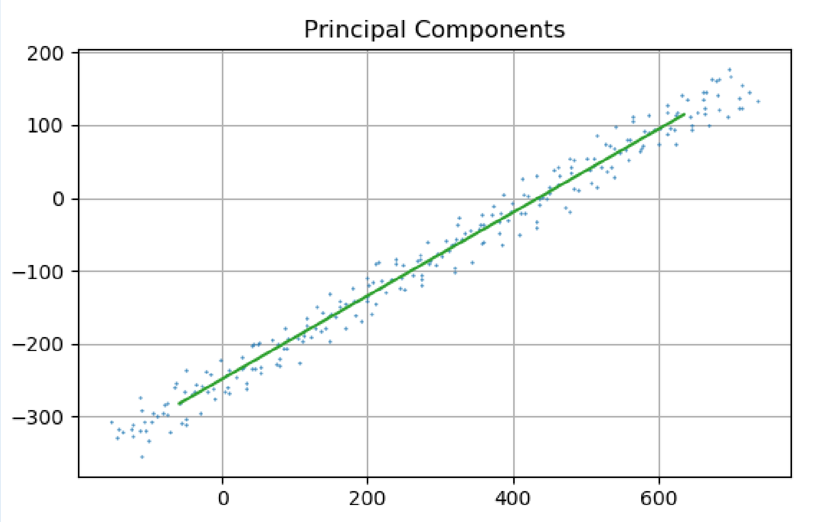
X in new feature space
Features with smallest variance can be discarded

Data Preprocessing

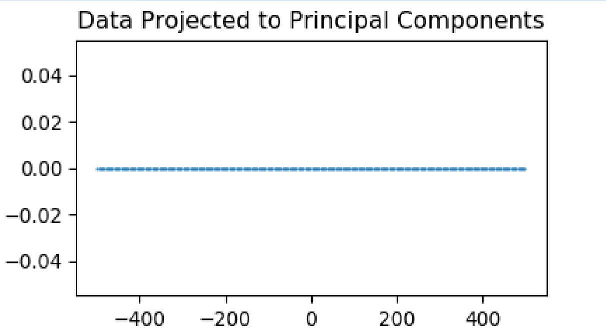
Principal Component Analysis for dimensionality reduction



pca.fit(X,n_components=1)



P = pca.transform(X)



Xf = pca.inverse_transform(P)