Principal Component Analysis

Which features can we ignore?

Constant: e.g., number of cabins: 1, 1, 1, 1, 1

Constant with Noise: e.g., hair thickness: .008, .003, .005

Linearly Dependent: e.g., weight and height

Variance

Covariance

$$\operatorname{var}(X) = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)}{(n-1)}$$

$$\operatorname{cov}(X, Y) = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{(n-1)}$$

Covariance Matrix

$$C = \begin{bmatrix} \operatorname{cov}(X,X) & \operatorname{cov}(X,Y) & \operatorname{cov}(X,Z) \\ \operatorname{cov}(Y,X) & \operatorname{cov}(Y,Y) & \operatorname{cov}(Y,Z) \\ \operatorname{cov}(Z,X) & \operatorname{cov}(Z,Y) & \operatorname{cov}(Z,Z) \end{bmatrix}$$

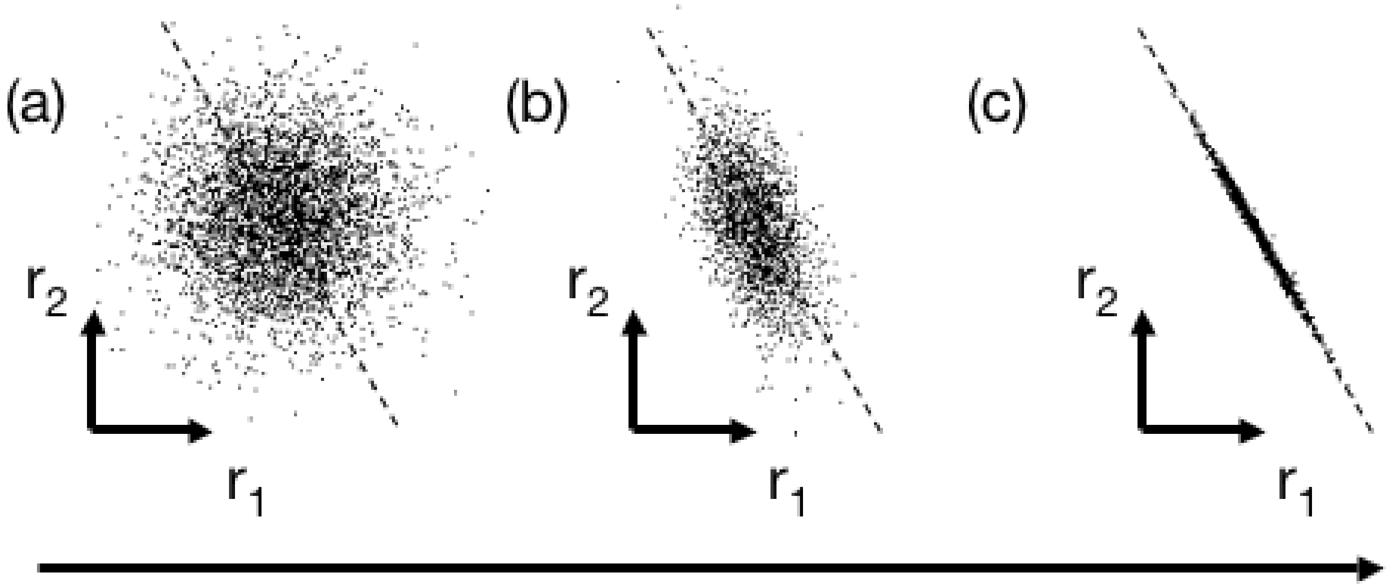
Covariance

Positive: both features move together

Negative: when one increases the other decreases

Independent: no relationship (covariance is 0)

Covariance



Low/0 covariance

High covariance

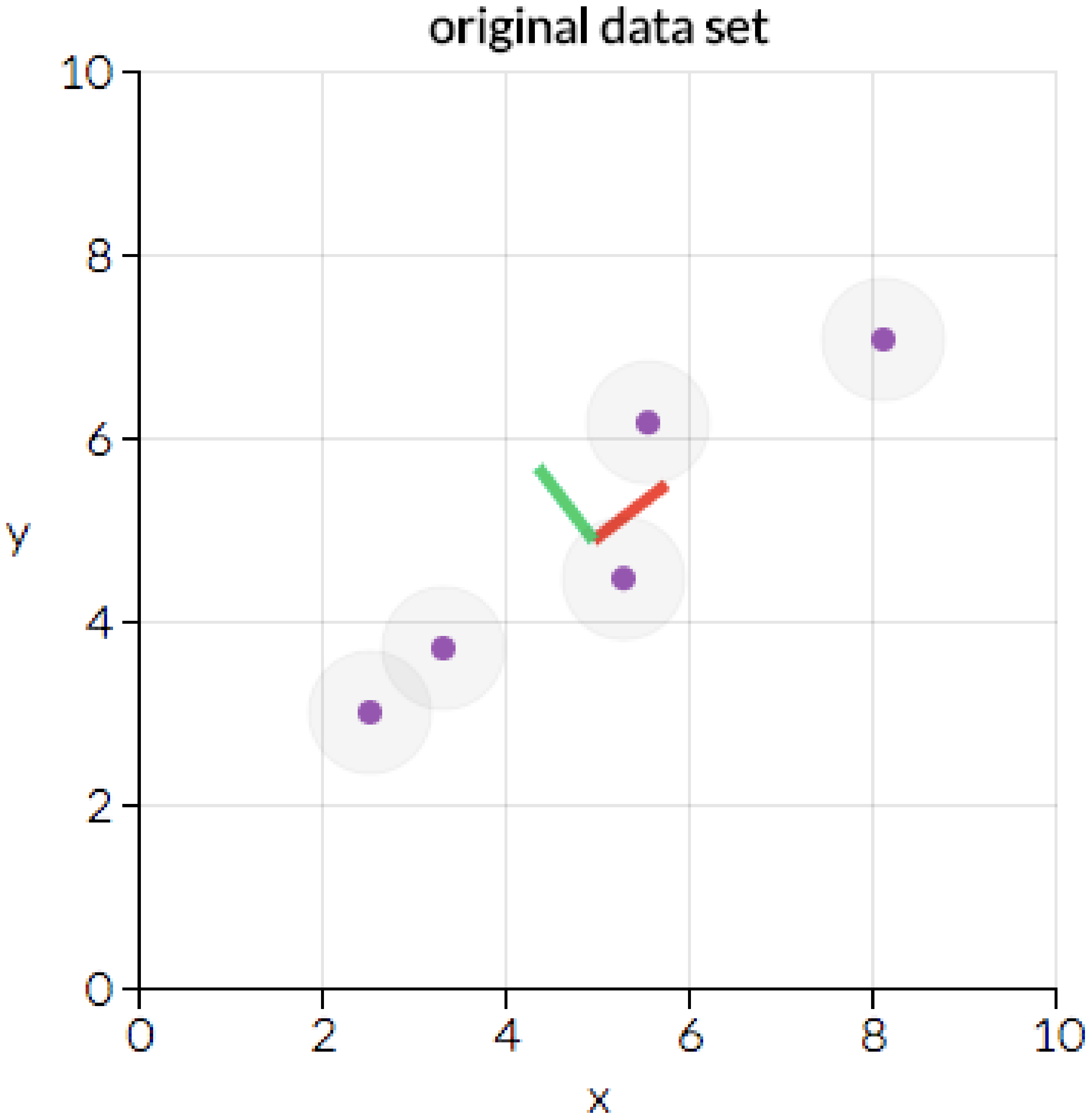
Which features do we want to keep?

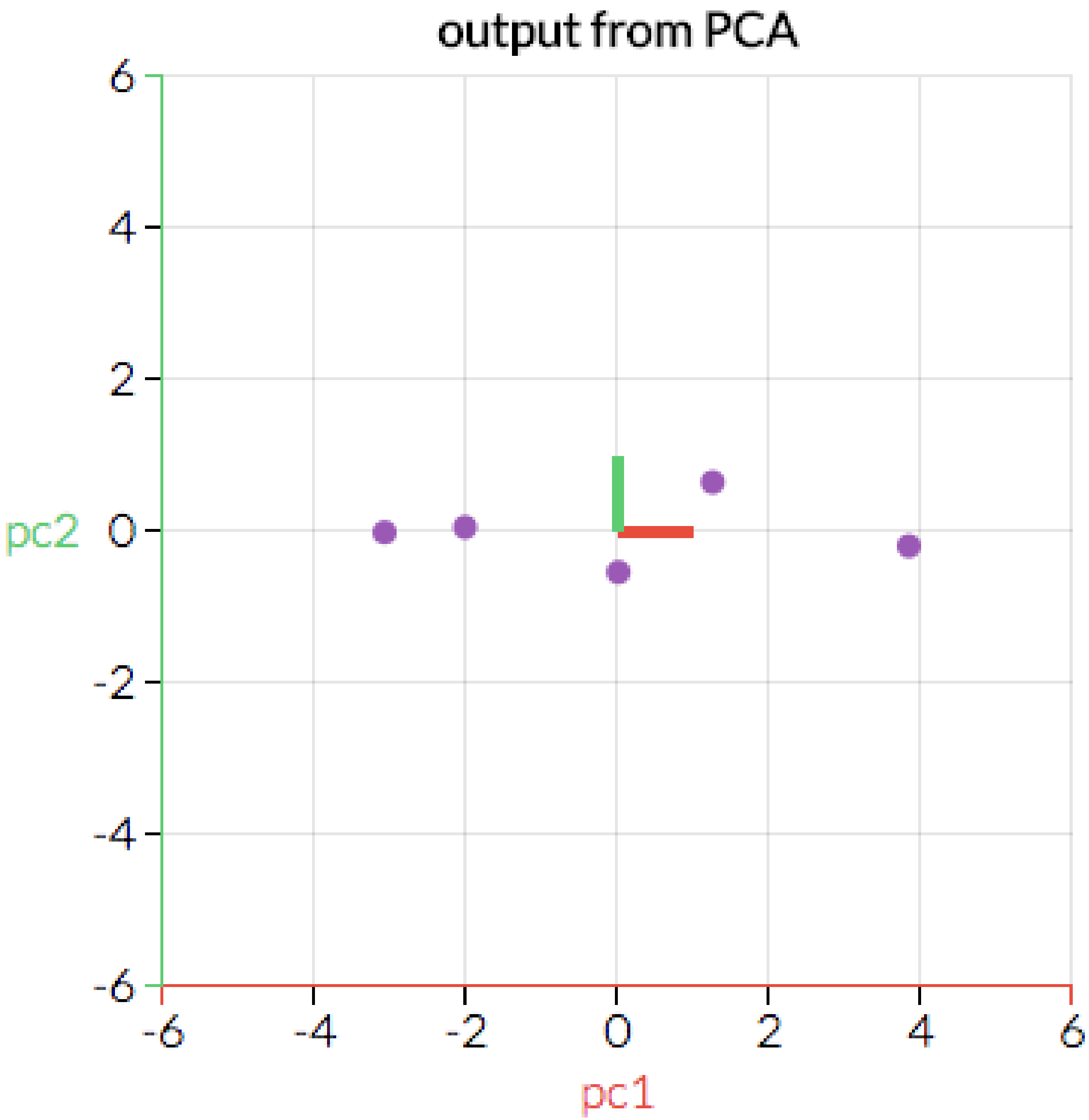
Low covariance: e.g., hours of sleep the previous night

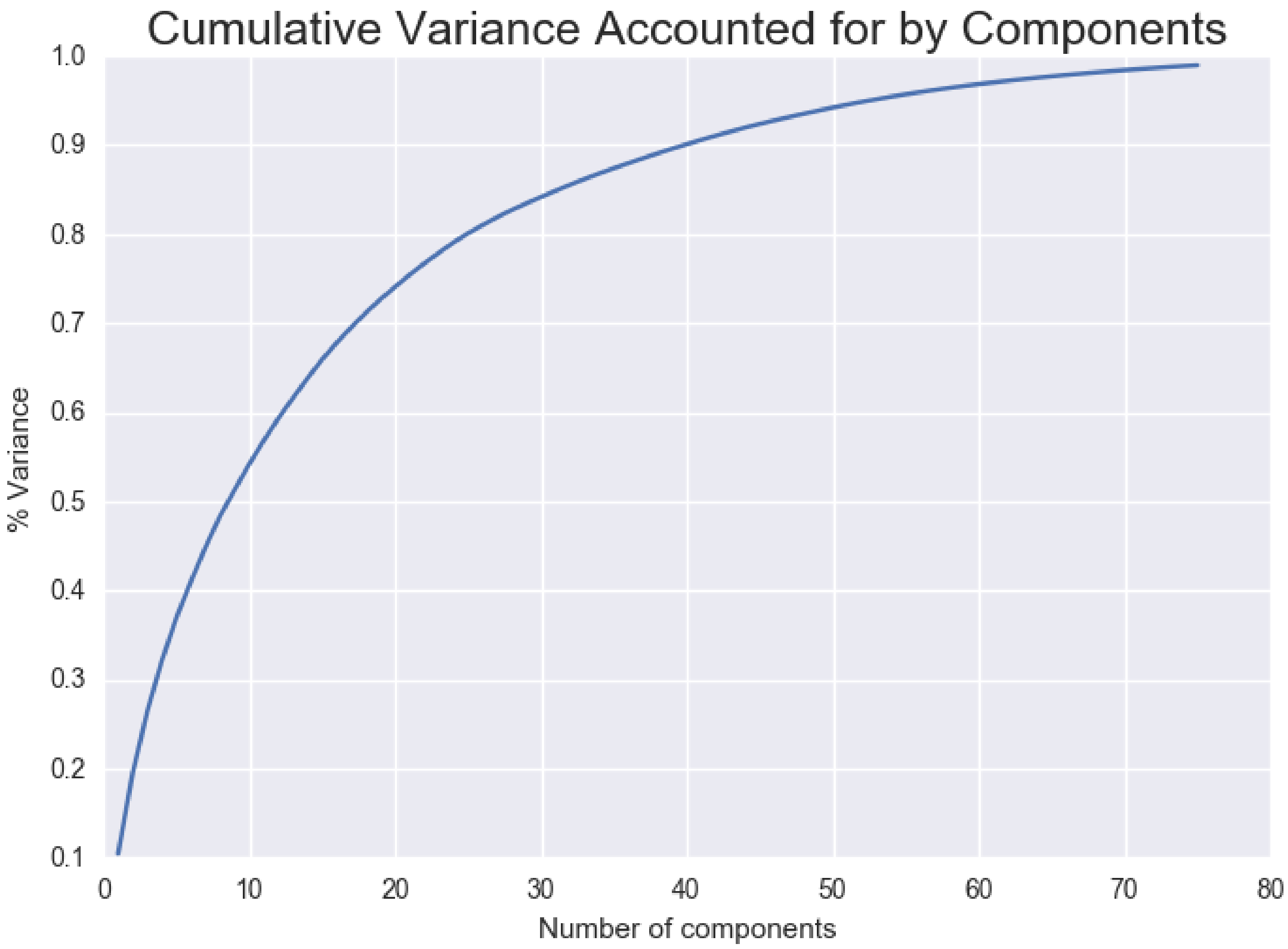
High variance: e.g., minutes spent inside cabin

Eigenvalues

Eigenvectors


original data set
10
8
6
4
2
0
0
2
4
6
8
10
y
x


output from PCA
6
4
2
0
-2
-4
-6
pc2
-6
-4
-2
0
2
4
6
pc1


Cumulative Variance Accounted for by Components
% Variance
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10
20
30
40
50
60
70
80
Number of components

PCA is:

Covariance Matrix

A measure of how each variable is associated with one another.

Eigenvectors

The directions in which our data are dispersed.

Eigenvalues

The relative importance (magnitude) of these different directions.