Lecture 13

Principal Component Analysis

Brett Bernstein

CDS at NYU

April 25, 2017

Intro Question

Question

Let $S \in \mathbb{R}^{n \times n}$ be symmetric.

1. How does **trace** S relate to the spectral decomposition $S = W \Lambda W^T$ where W is orthogonal and Λ is diagonal?
2. How do you solve $w_* = \arg\max_{\|w\|_2 = 1} w^T S w$? What is $w_*^T S w_*$?

Intro Solution

Solution

1. We use the following useful property of traces: $\textbf{trace}\, AB = \textbf{trace}\, BA$ for any matrices A, B where the dimensions allow. Thus we have

$$\textbf{trace}\, S = \textbf{trace}\, W(\Lambda W^T) = \textbf{trace}\, (\Lambda W^T)W = \textbf{trace}\, \Lambda,$$

   so the trace of S is the sum of its eigenvalues.
2. w_* is an eigenvector with the largest eigenvalue. Then $w_*^T S w_*$ is the largest eigenvalue.

Unsupervised Learning

1. Where did the y's go?
2. Try to find intrinsic structure in unlabeled data.
3. With PCA, we are looking for a low dimensional affine subspace that approximates our data well.

Centered Data

1. Throughout this lecture we will work with centered data.
2. Suppose $X \in \mathbb{R}^{n \times d}$ is our data matrix. Define

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

3. Let $\overline{X} \in \mathbb{R}^{n \times d}$ be the matrix with $\overline{x}$ in every row.
4. Define the centered data:

$$\tilde{X} = X - \overline{X}, \quad \tilde{x}_i = x_i - \overline{x}.$$

Variance Along A Direction

Definition

Let $\tilde{x}_1, \ldots, \tilde{x}_n \in \mathbb{R}^d$ be the centered data. Fix a direction $w \in \mathbb{R}^d$ with $\|w\|_2 = 1$. The sample variance along w is given by

$$\frac{1}{n-1} \sum_{i=1}^{n} (\tilde{x}_i^T w)^2.$$

This is the sample variance of the components

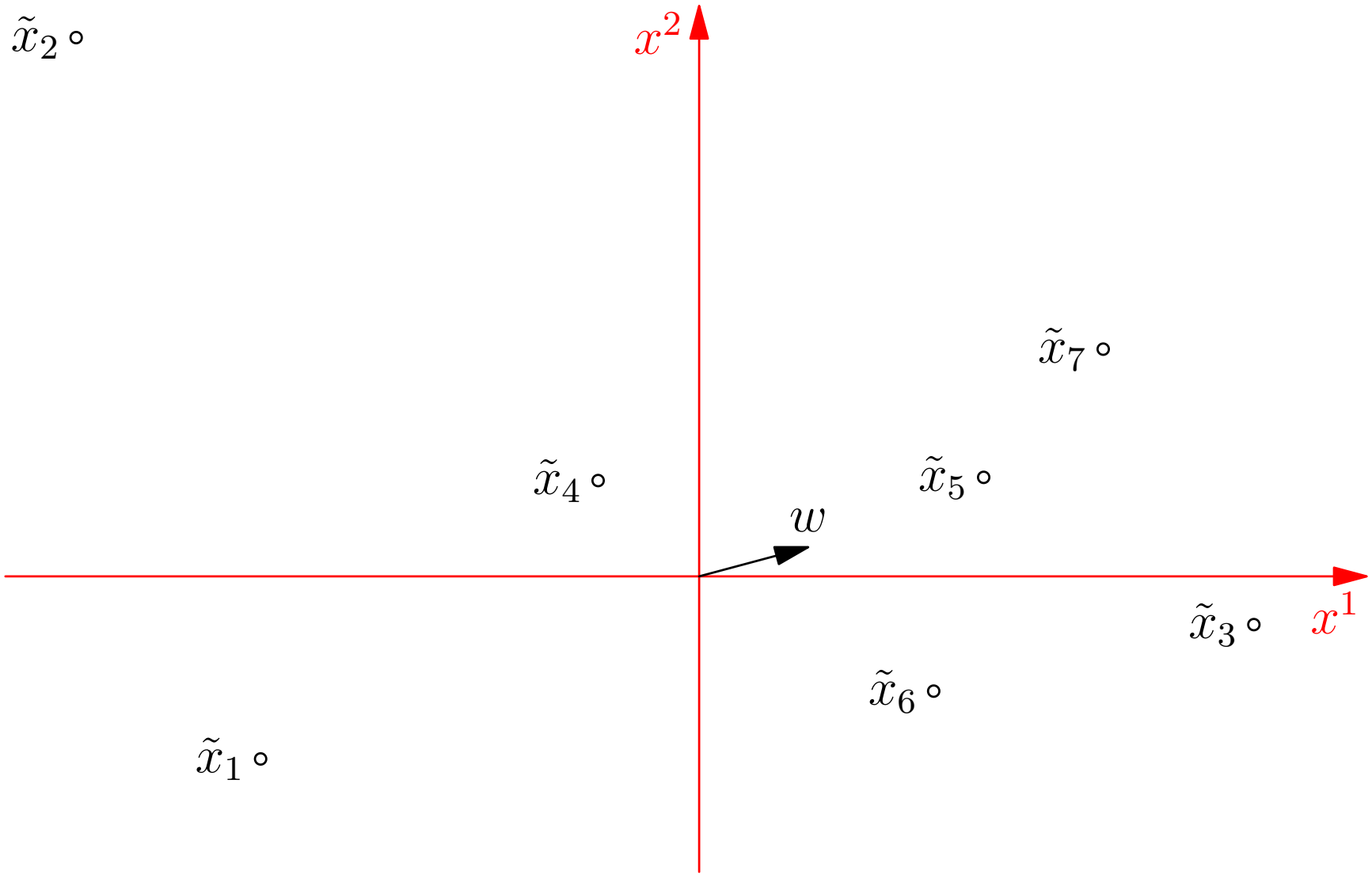
$$\tilde{x}_1^T w, \ldots, \tilde{x}_n^T w.$$

1. This is also the sample variance of

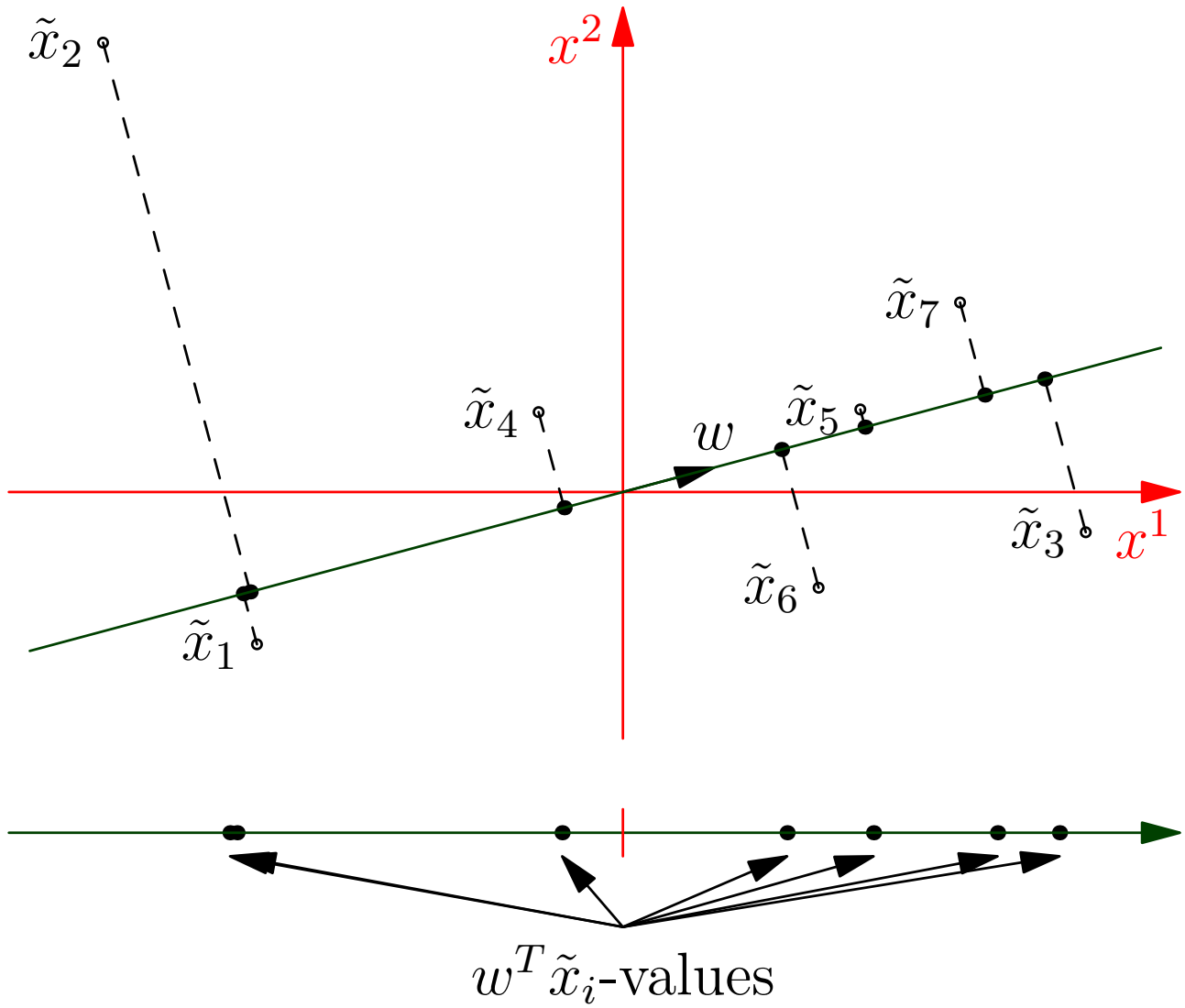
$$x_1^T w, \ldots, x_n^T w,$$

using the uncentered data.

Variance Along A Direction



Variance Along A Direction



First Principal Component

1. Define the first loading vector $w_{(1)}$ to be the direction giving the highest variance:

$$w_{(1)} = \arg\max_{\|w\|_2=1} \frac{1}{n-1} \sum_{i=1}^{n} (\tilde{x}_i^T w)^2.$$

2. Maximizer is not unique, so we choose one.

Definition

The *first principal component* of $\tilde{x}_i$ is $\tilde{x}_i^T w_{(1)}$.

Principal Components

1. Define the kth loading vector $w_{(k)}$ to be the direction giving the highest variance that is orthogonal to the first $k-1$ loading vectors:

$$w_{(k)} = \underset{\substack{\|w\|_2=1 \\ w \perp w_{(1)},\ldots,w_{(k-1)}}}{\arg\max} \frac{1}{n-1} \sum_{i=1}^{n} (\tilde{x}_i^T w)^2.$$

2. The complete set of loading vectors $w_{(1)}, \ldots, w_{(d)}$ form an orthonormal basis for $\mathbb{R}^d$.

Definition

The *kth principal component* of $\tilde{x}_i$ is $\tilde{x}_i^T w_{(k)}$.

Principal Components

1. Let W denote the matrix with the kth loading vector $w_{(k)}$ as its kth column.
2. Then $W^T \tilde{x}_i$ gives the principal components of $\tilde{x}_i$ as a column vector.
3. $\tilde{X}W$ gives a new data matrix in terms of principal components.
4. If we compute the singular value decomposition (SVD) of $\tilde{X}$ we get

$$\tilde{X} = VDW^T,$$

where $D \in \mathbb{R}^{n \times d}$ is diagonal with non-negative entries, and $V \in \mathbb{R}^{n \times n}$, $W \in \mathbb{R}^{d \times d}$ are orthogonal.

5. Then $\tilde{X}^T \tilde{X} = WD^T DW^T$. Thus we can use the SVD on our data matrix to obtain the loading vectors W and the eigenvalues $\Lambda = \frac{1}{n-1} D^T D$.

Some Linear Algebra

Recall that $w_{(1)}$ is defined by

$$w_{(1)} = \arg\max_{\|w\|_2=1} \frac{1}{n-1} \sum_{i=1}^{n} (\tilde{x}_i^T w)^2.$$

We now perform some algebra to simplify this expression. Note that

$$\begin{aligned}
\sum_{i=1}^{n} (\tilde{x}_i^T w)^2 &= \sum_{i=1}^{n} (\tilde{x}_i^T w)(\tilde{x}_i^T w) \\
&= \sum_{i=1}^{n} (w^T \tilde{x}_i)(\tilde{x}_i^T w) \\
&= w^T \left[\sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T \right] w \\
&= w^T \tilde{X}^T \tilde{X} w.
\end{aligned}$$

Some Linear Algebra

1. This shows

$$w_{(1)} = \arg\max_{\|w\|_2=1} \frac{1}{n-1} w^T \tilde{X}^T \tilde{X} w = \arg\max_{\|w\|_2=1} w^T S w,$$

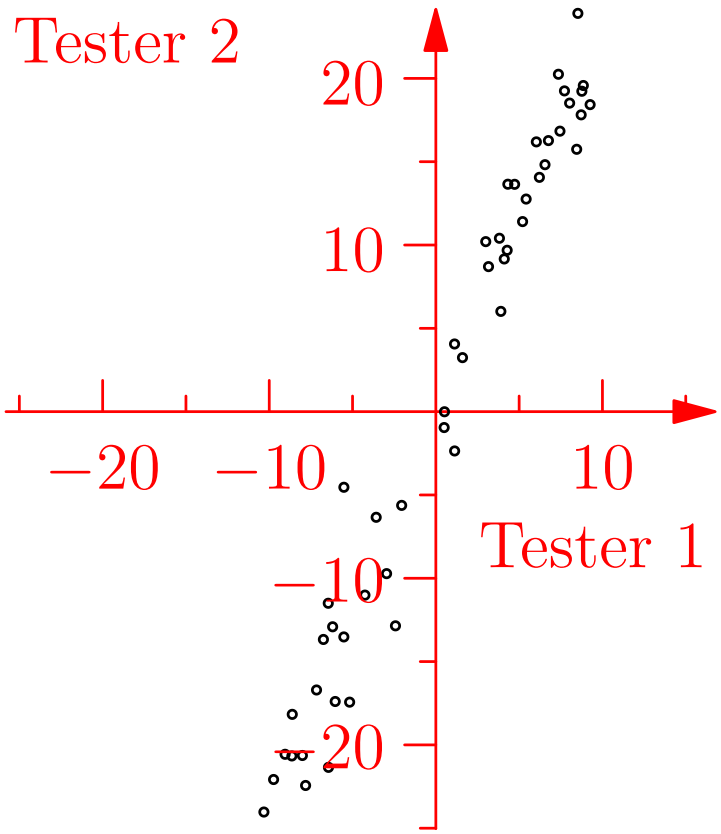
   where $S = \frac{1}{n-1} \tilde{X}^T \tilde{X}$ is the sample covariance matrix.
2. By the introductory problem this implies $w_{(1)}$ is the eigenvector corresponding to the largest eigenvalue of S.
3. We also learn that the variance along $w_{(1)}$ is λ_1, the largest eigenvalue of S.
4. With a bit more work we can see that $w_{(k)}$ is the eigenvector corresponding to the kth largest eigenvalue, with λ_k giving the associated variance.

PCA Example

Example

A collection of people come to a testing site to have their heights measured twice. The two testers use different measuring devices, each of which introduces errors into the measurement process. Below we depict some of the measurements computed (already centered).

PCA Example



1. Describe (vaguely) what you expect the sample covariance matrix to look like.
2. What do you think $w_{(1)}$ and $w_{(2)}$ are?

PCA Example: Solutions

1. We expect tester 2 to have a larger variance than tester 1, and to be nearly perfectly correlated. The sample covariance matrix is

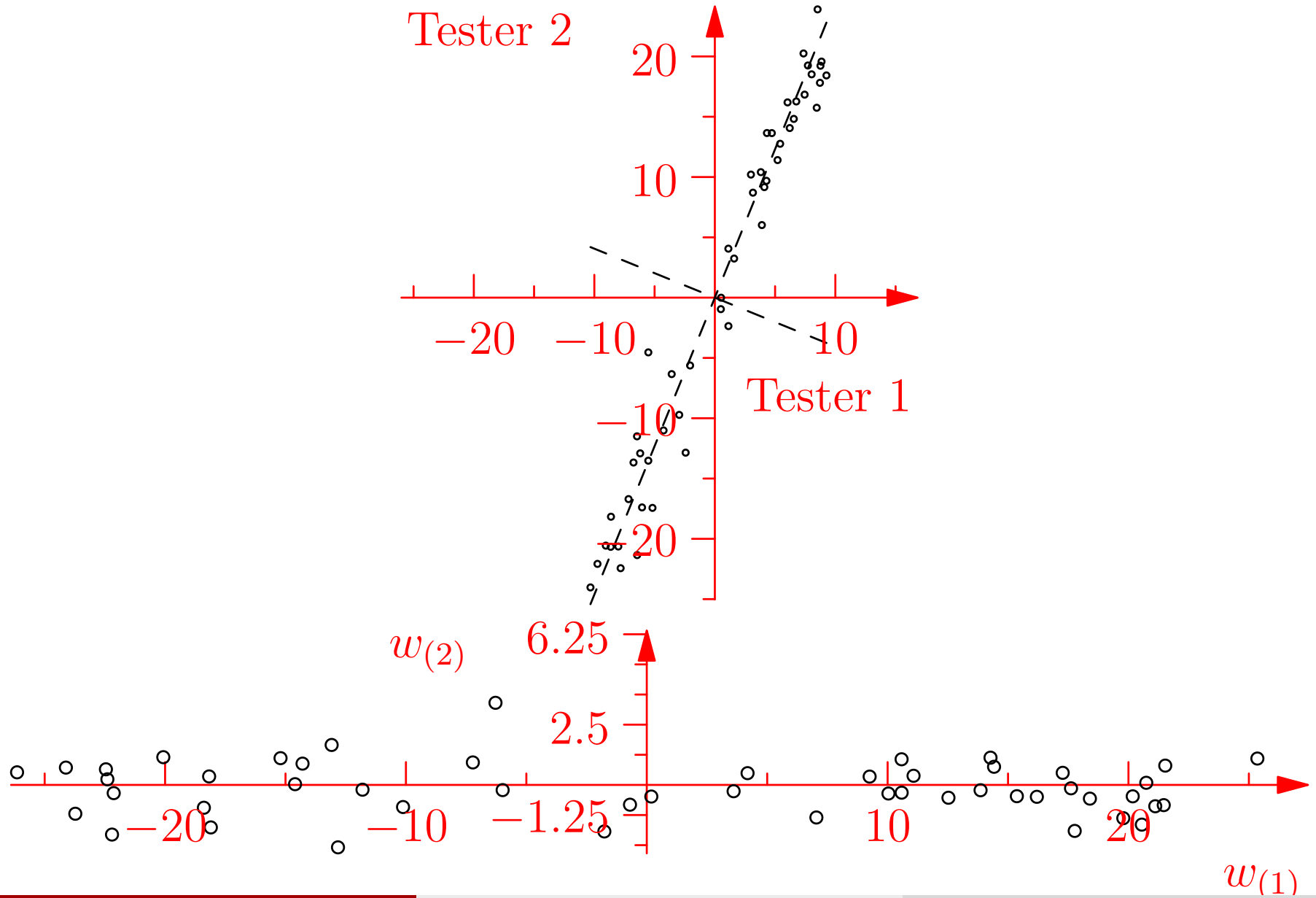
$$S = \begin{pmatrix} 40.5154 & 93.5069 \\ 93.5069 & 232.8653 \end{pmatrix}.$$

2. We have

$$S = W\Lambda W^T, W = \begin{pmatrix} 0.3762 & -0.9265 \\ 0.9265 & 0.3762 \end{pmatrix}, \Lambda = \begin{pmatrix} 270.8290 & 0 \\ 0 & 2.5518 \end{pmatrix}.$$

- Note that **trace** $\Lambda =$ **trace** S.
- Since λ_2 is small, it shows that $w_{(2)}$ is almost in the null space of S. This suggests $-.9265\tilde{x}^1 + .3762\tilde{x}^2 \approx 0$ for data points $(\tilde{x}^1, \tilde{x}^2)$. In other words, $\tilde{x}^2 \approx 2.46\tilde{x}^1$. Maybe tester 2 used centimeters and tester 1 used inches.

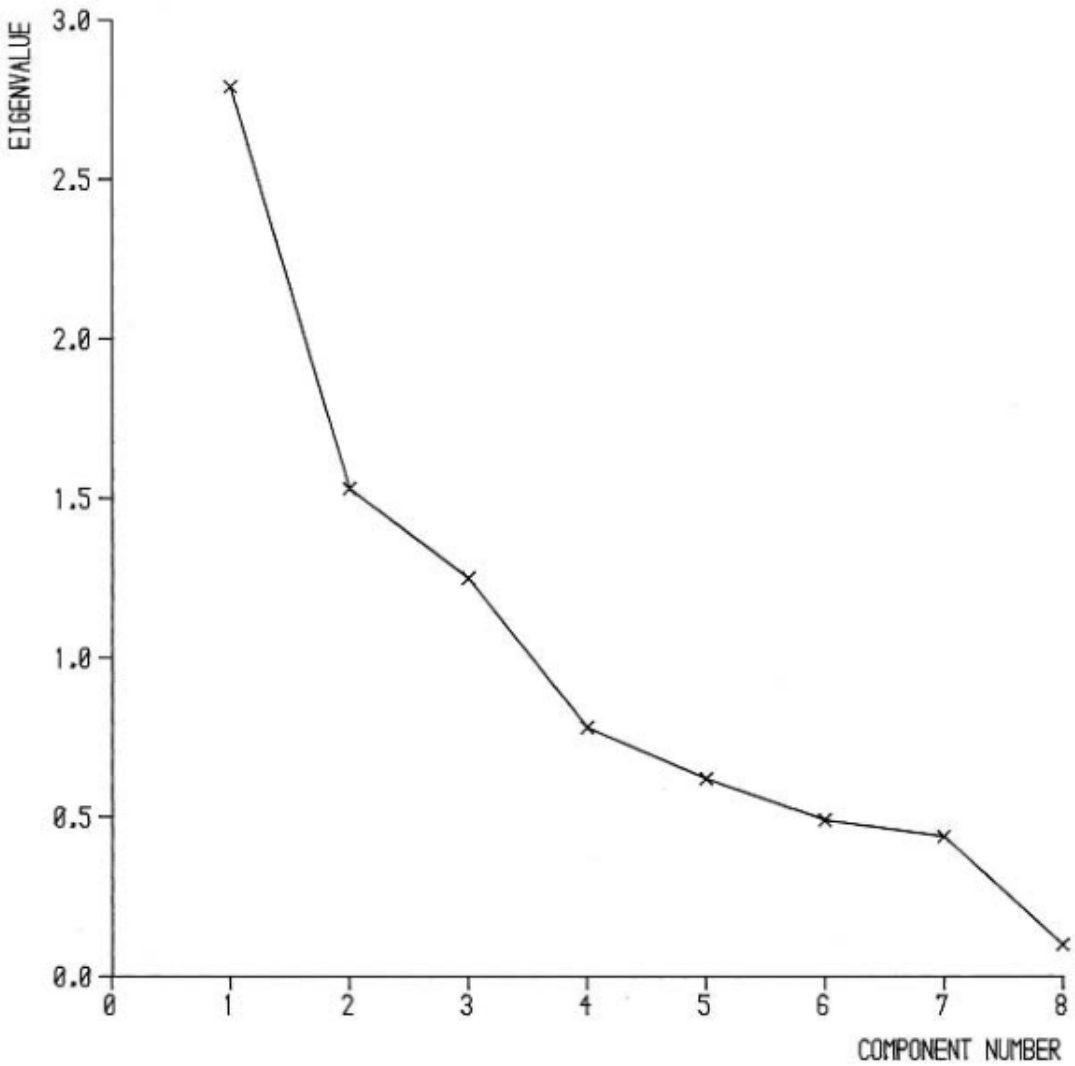
PCA Example: Plot In Terms of Principal Components



Uses of PCA: Dimensionality Reduction

1. In our height example above, we can replace our two features with only a single feature, the first principal component.
2. This can be used as a preprocessing step in a supervised learning algorithm.
3. When performing dimensionality reduction, one must choose how many principal components to use. This is often done using a scree plot: a plot of the eigenvalues of S in descending order.
4. Often people look for an "elbow" in the scree plot: a point where the plot becomes much less steep.

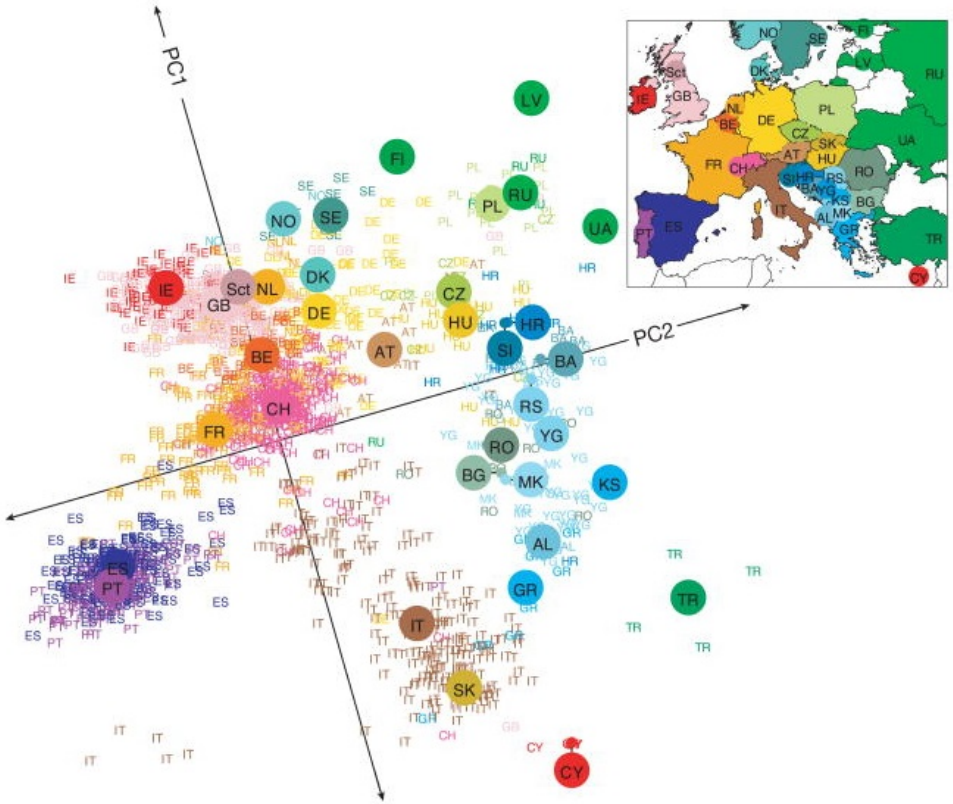
Scree Plot



[1] From Jolliffe, Principal Component Analysis

Uses of PCA: Visualization

1. Visualization: If we have high dimensional data, it can be hard to plot it effectively. Sometimes plotting the first two principal components can reveal interesting geometric structure in the data.



[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735096/

Uses of PCA: Principal Component Regression

1. Want to build a linear model with a dataset

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}.$$

2. We can choose some k and replace each $\tilde{x}_i$ with its first k principal components. Afterward we perform linear regression.
3. This is called principal component regression, and can be thought of as a discrete variant of ridge regression (see HTF 3.4.1).
4. Correlated features may be grouped together into a single principal component that averages their values (like with ridge regression). Think about the 2 tester example from before.

Standardization

1. What happens if you scale one of the features by a huge factor?

Standardization

1. What happens if you scale one of the features by a huge factor?
2. It will have a huge variance and become a dominant part of the first principal component.
3. To add scale-invariance to the process, people often standardize their data (center and normalize) before running PCA.
4. This is the same as using the correlation matrix in place of the covariance matrix.

Dispersion Of The Data

1. One measure of how dispersed our data is the following:

$$\Delta = \frac{1}{n-1}\sum_{i=1}^{n}\|x_i - \overline{x}\|_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}\|\tilde{x}_i\|_2^2.$$

2. A little algebra shows this is **trace** S, where S is the sample covariance matrix.
3. If we project onto the first k principal components, the resulting data has dispersion $\lambda_1 + \cdots + \lambda_k$.
4. We can choose k to account for a desired percentage of Δ.
5. The subspace spanned by the first k loading vectors maximizes the resulting dispersion over all possible k-dimensional subspaces.

Other Comments

1. The k-dimensional subspace V spanned by $w_{(1)}, \ldots, w_{(k)}$ best fits the centered data in the least-squares sense. More precisely, it minimizes

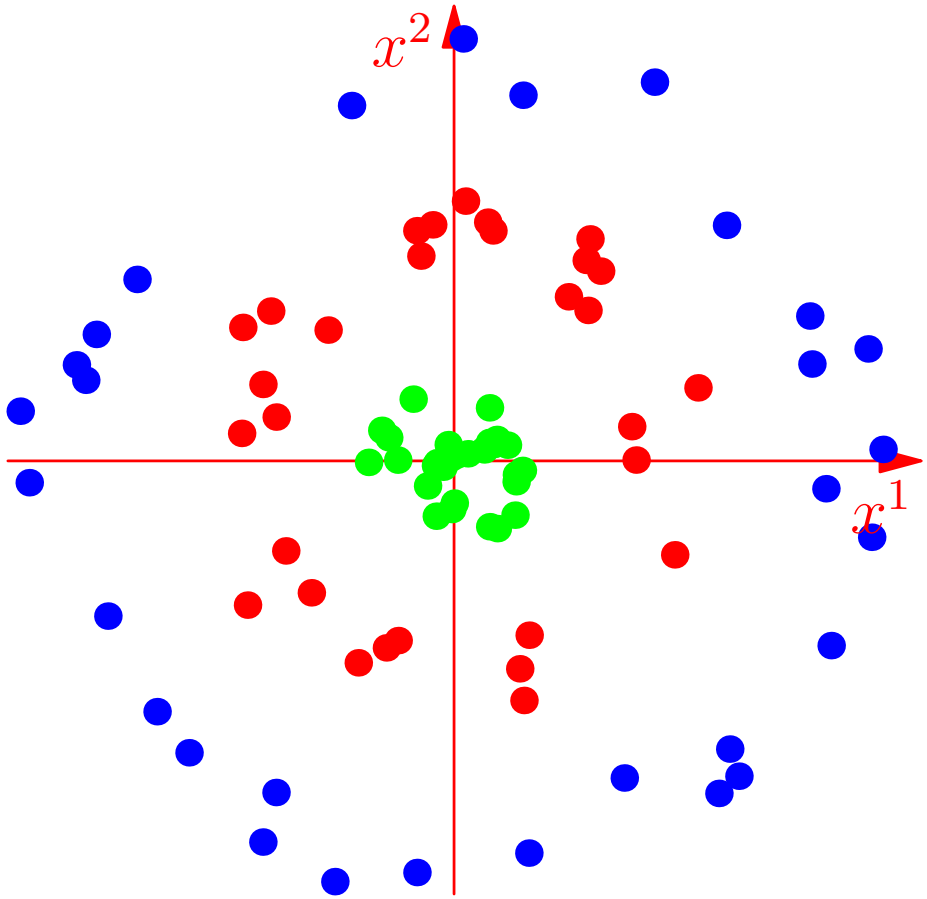
$$\sum_{i=1}^{n} \|x_i - P_V(x_i)\|_2^2$$

over all k-dimensional subspaces, where P_V orthogonally projects onto V.

2. Converting your data into principal components can sometimes hurt interpretability since the new features are linear combinations (i.e., blends or baskets) of your old features.
3. The smallest principal components, if they correspond to small eigenvalues, are nearly in the null space of X, and thus can reveal linear dependencies in the centered data.

Principal Components Are Linear

Suppose we have the following labeled data.



How can we apply PCA and obtain a single principal component that distinguishes the colored clusters?

Principal Components Are Linear

1. In general, can deal with non-linear by adding features or using kernels.
2. Using kernels results in the technique called Kernel PCA.
3. Below we added the feature $\|\tilde{x}_i\|^2$ and took the first principal component.