Machine Learning – Brett Bernstein

Lecture 13: Principal Component Analysis

Intro Question

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

1. How does $\textbf{trace}\, S$ relate to the spectral decomposition $S = W\Lambda W^T$ where W is orthogonal and Λ is diagonal?

   Solution. 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. How do you solve $w_* = \arg\max_{\|w\|_2=1} w^T S w$? What is $w_*^T S w_*$?

   Solution. Suppose S was diagonal:

   $$S = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \lambda_n \end{pmatrix},$$

   with $\lambda_1 \geq \cdots \geq \lambda_n$. Then $w_* = e_1$, the first standard basis vector since

   $$v^T S v = \sum_{i=1}^n \lambda_i v_i^2 \leq \lambda_1 \sum_{i=1}^n v_i^2 = \lambda_1 = e_1^T S e_1.$$

   In general, we have $S = W\Lambda W^T$, so we want $W^T w = e_1$ (using the fact that $\|W^T v\|_2 = \|v\|_2$). But then the answer is the first column of W, i.e., the eigenvector corresponding to largest eigenvalue λ_1.

Principal Component Analysis (PCA)

This will be our first topic of *unsupervised learning*. Simply put, in unsupervised learning we have no y-values (i.e., no labels). As such, our goal is to find and exploit intrinsic structure in the training data. With PCA, we are trying to find a low dimensional affine subspace that explains most of the variance in our dataset.

Definition of Principal Components

When studying PCA, we will always work with centered data. As such, we define the centered data matrix

$$\tilde{X} := X - \overline{X},$$

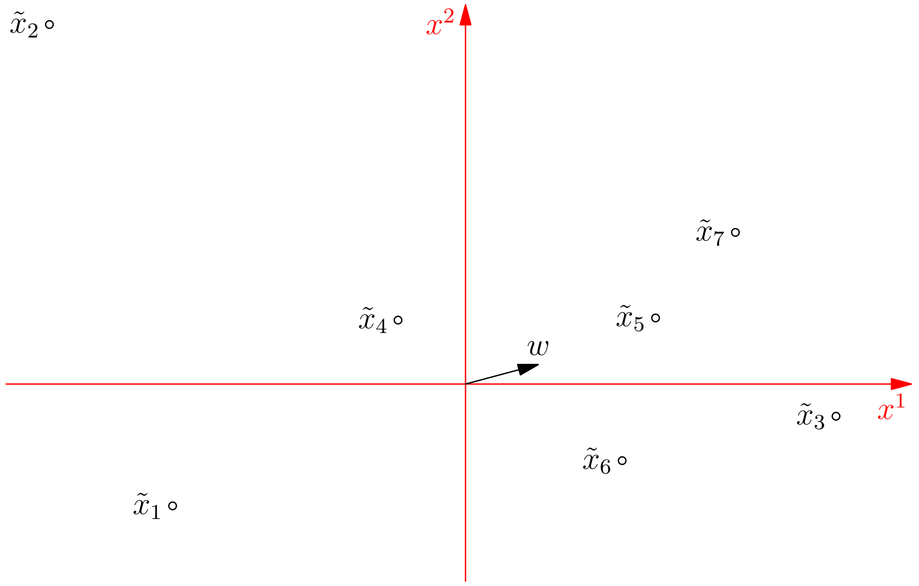
with rows $\tilde{x}_i^T = (x_i - \overline{x})^T$. Next we need the concept of *variance along a direction.*

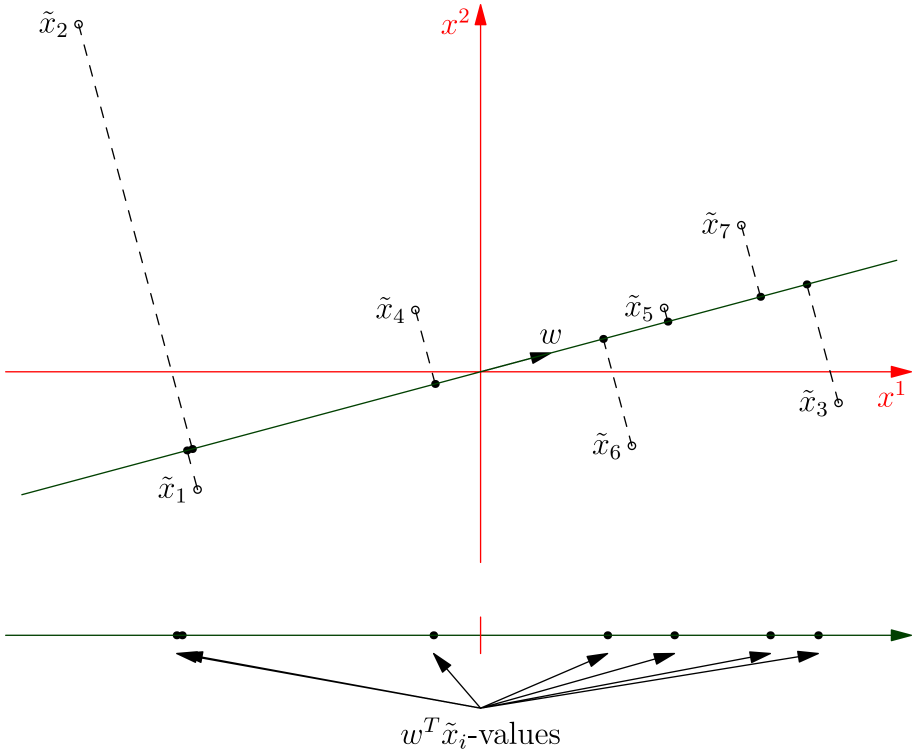
Definition 1 (Variance Along a Direction)**.** Let $\tilde{x}_1, \ldots, \tilde{x}_n$ be the centered data with $\tilde{x}_i \in \mathbb{R}^d$. Fix a direction $w \in \mathbb{R}^d$ with $\|w\|_2 = 1$. The sample variance along the direction w is given by

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

This is the sample variance of the components

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





As an aside, note that the variance along w is also the sample variance of

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

where we haven't centered the data.

Define the first loading vector $w_{(1)}$ to be the direction along which the sample variance is maximized:

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

The maximizer will not be unique, so we arbitrarily choose one of the maximizers. We then define $\tilde{x}_i^T w_{(1)}$ to be the *first principal component* of the centered data point $\tilde{x}_i$. That is, it is the component of $\tilde{x}_i$ in the direction $w_{(1)}$.

The kth loading vector $w_{(k)}$ maximizes the variance along it while being orthogonal to the first $k-1$ loading vectors:

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

Taken together, $w_{(1)}, \ldots, w_{(d)}$ form an orthonormal basis of $\mathbb{R}^d$. Analgously, we define $\tilde{x}_i^T w_{(k)}$ to be the *kth principal component* of the centered data point $\tilde{x}_i$. If W is a matrix whose kth column is $w_{(k)}$ then $W^T \tilde{x}_i$ expresses $\tilde{x}_i$ in terms of its principal components. We can write $\tilde{X}W$ to express the entire data matrix in terms of principal components.

Computing Principal Components

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

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

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

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

This shows

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

where $S = \frac{1}{n-1}\tilde{X}^T\tilde{X}$ is the sample covariance matrix. From the introductory questions, we know that the maximizer is the eigenvector corresponding to the largest eigenvalue of S, and the maximum value attained is the sample variance along $w_{(1)}$.

In fact, we can take this further. Suppose we compute the spectral decomposition of S. That is,

$$S = W\Lambda W^T$$

where W is orthogonal and

$$\Lambda = \begin{pmatrix}\lambda_1 & 0 & 0 & 0\\ 0 & \lambda_2 & 0 & 0\\ 0 & 0 & \ddots & 0\\ 0 & 0 & 0 & \lambda_d\end{pmatrix}$$

with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ (all are non-negative since S is PSD). Then the ith column of W is $w_{(i)}$ and λ_i is the variance along $w_{(i)}$.

As an aside, we sketch the proof idea.

Proof sketch. By the spectral theorem we have

$$S = \sum_{i=1}^{n}\lambda_i W_{:,i}W_{:,i}^T,$$

where $W_{:,i}$ is the ith column of W. Note that

$$\begin{aligned}
w_{(k)} &= \underset{\substack{\|w\|_2=1 \\ w\perp w_{(1)},\ldots,w_{(k-1)}}}{\arg\max} \quad w^T S w \\
&= \underset{\substack{\|w\|_2=1 \\ w\perp w_{(1)},\ldots,w_{(k-1)}}}{\arg\max} \quad w^T \left[\sum_{i=1}^{n} \lambda_i W_{:,i} W_{:,i}^T\right] w \\
&= \underset{\substack{\|w\|_2=1 \\ w\perp w_{(1)},\ldots,w_{(k-1)}}}{\arg\max} \quad w^T \left[\sum_{i=k}^{n} \lambda_i W_{:,i} W_{:,i}^T\right] w.
\end{aligned}$$

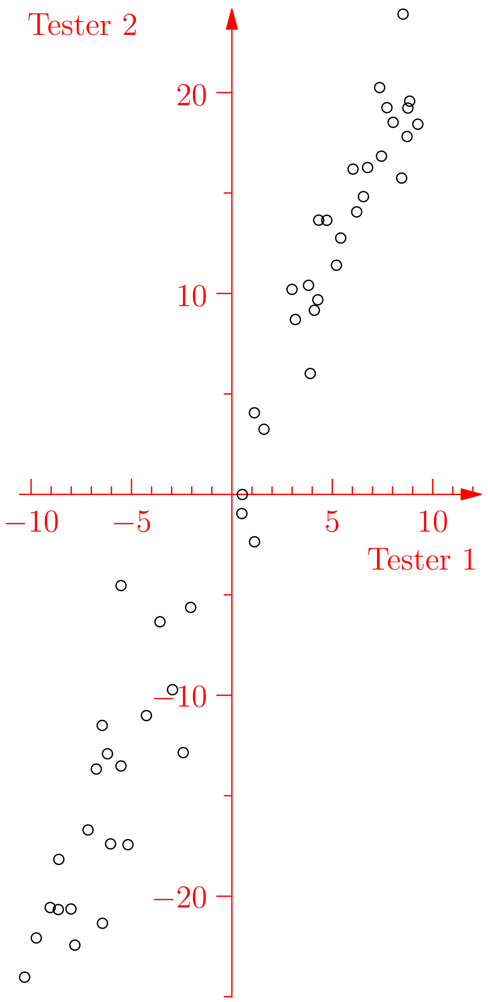
Noting that the maximizer w must be in the span of $W_{:,k}, \ldots, W_{:,d}$, we see the maximizer is $w_{(k)} = W_{:,k}$:

$$\begin{aligned}
&\left(\sum_{i=k}^{n} \alpha_i W_{:,i}\right)^T \left[\sum_{i=k}^{n} \lambda_i W_{:,i} W_{:,i}^T\right] \left(\sum_{i=k}^{n} \alpha_i W_{:,i}\right) \\
&= \sum_{i=k}^{n} \alpha_i^2 \lambda_i \\
&\leq \lambda_k \sum_{i=k}^{n} \alpha_i^2 \\
&= \lambda_k \\
&= W_{:,k}^T S W_{:,k}.
\end{aligned}$$

□

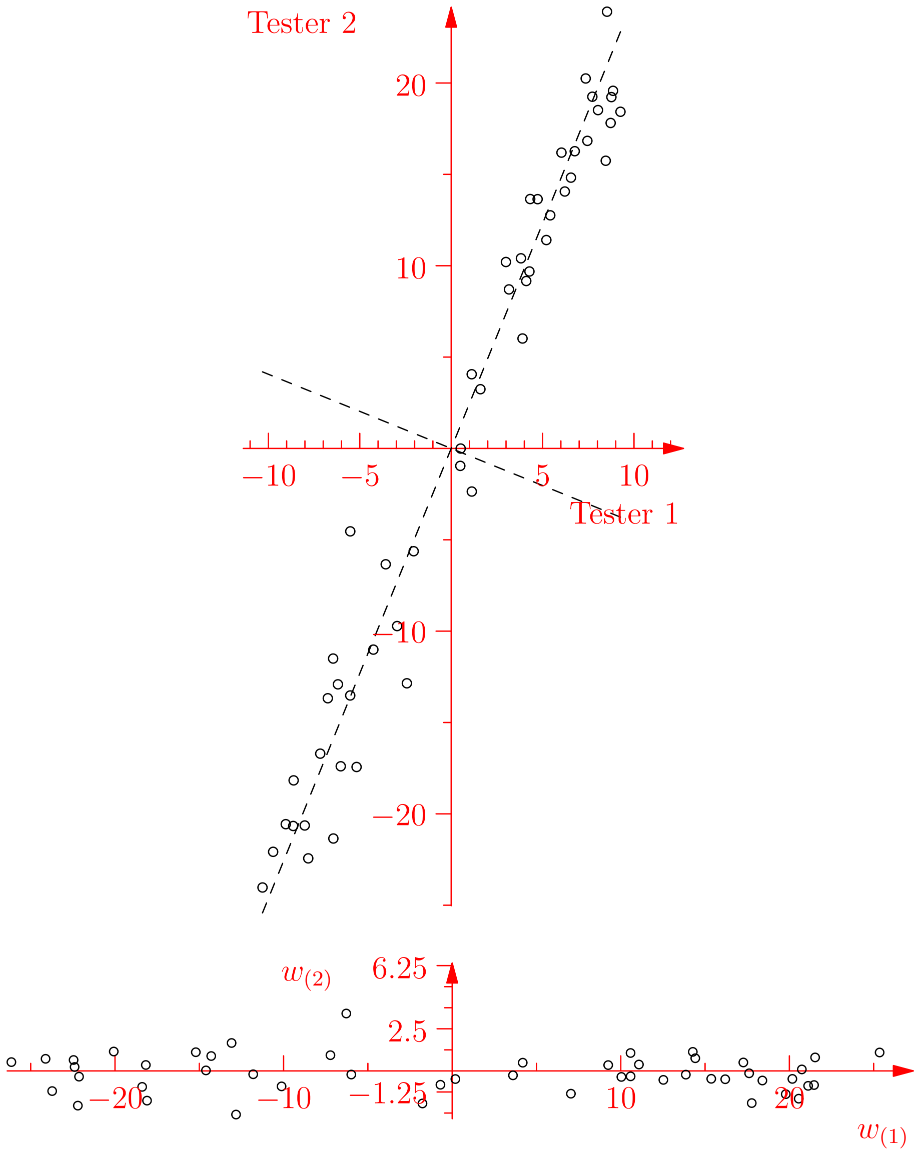
Let's illustrate the ideas thus far using an example.

Example 2. 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).



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

We can now plot the data in terms of the principal components (i.e., we plot $\tilde{X}W$).



Uses of Principal Component Analysis

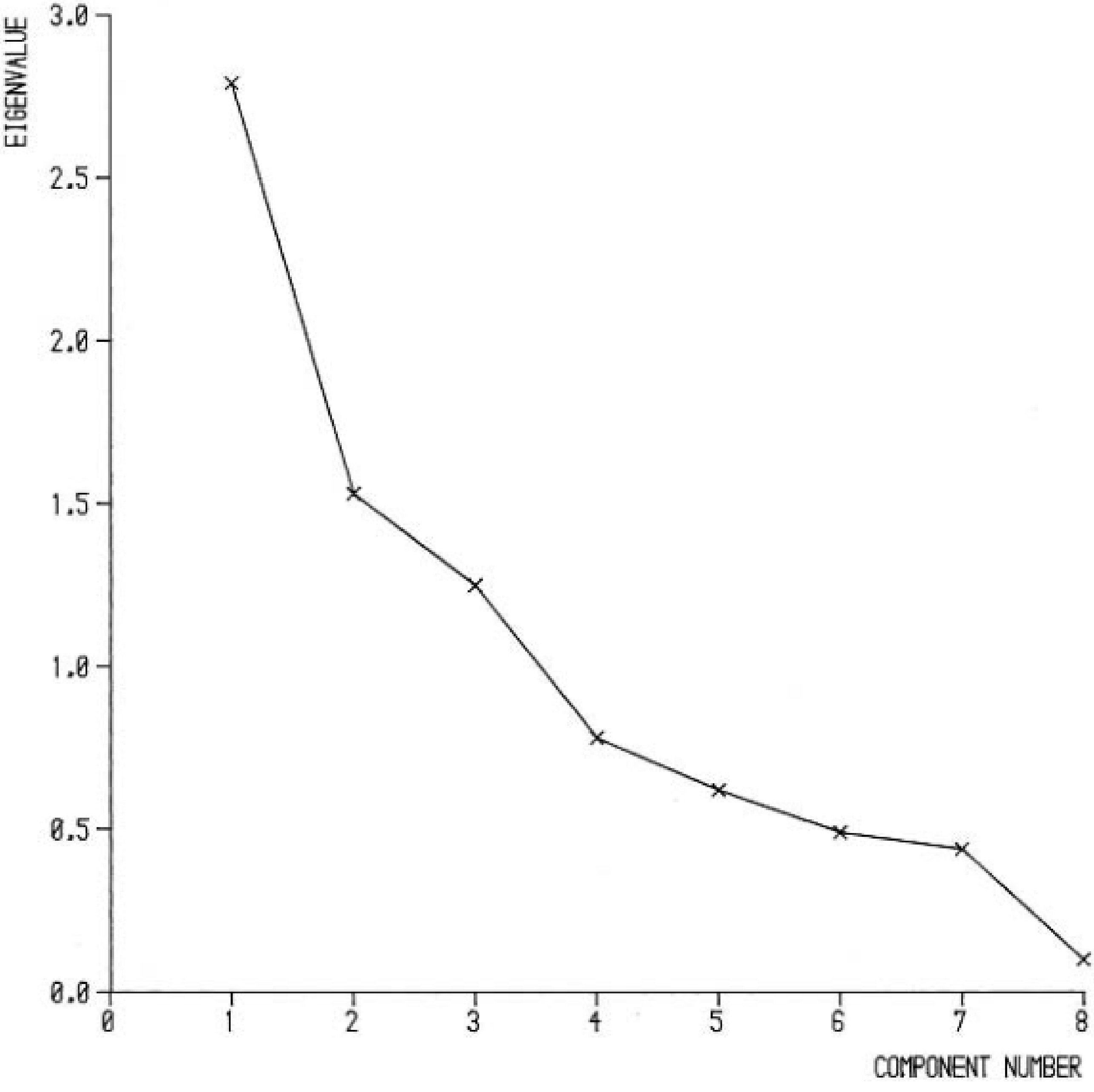
1. Dimensionality reduction: In our height example above, we can replace our two features with only a single feature, the first principal component. This can be used as a preprocessing step in a supervised learning algorithm. More about this in a moment.

2. 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.

3. Principal Component Regression: Building on dimensionality reduction, suppose we begin with a dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ and want to build a linear model. We can choose some k and replace each $\tilde{x}_i$ with its first k principal components. Afterward

we perform linear regression. This is called principal component regression, and can be thought of as a discrete variant of ridge regression (see HTF 3.4.1).

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.



Scree plot taken from Jolliffe's Principal Component Analysis. Often people look for an "elbow" in the scree plot: a point where the plot becomes much less steep.

Other Comments About PCA

1. Often people standardize their data before running PCA to add scale-invariance. Stated different, if a feature is scaled (maybe by choice of measurement unit) by a large factor we arbitrarily increase its variance, and thus can incorrectly force it to be a large part of the first principal component.

2. Define the dispersion of the data by

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

   Projecting the centered data onto the k-dimensional subspace spanned by $w_{(1)}, \ldots, w_{(k)}$ maximizes the resulting dispersion over all possible k-dimensional subspaces.

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

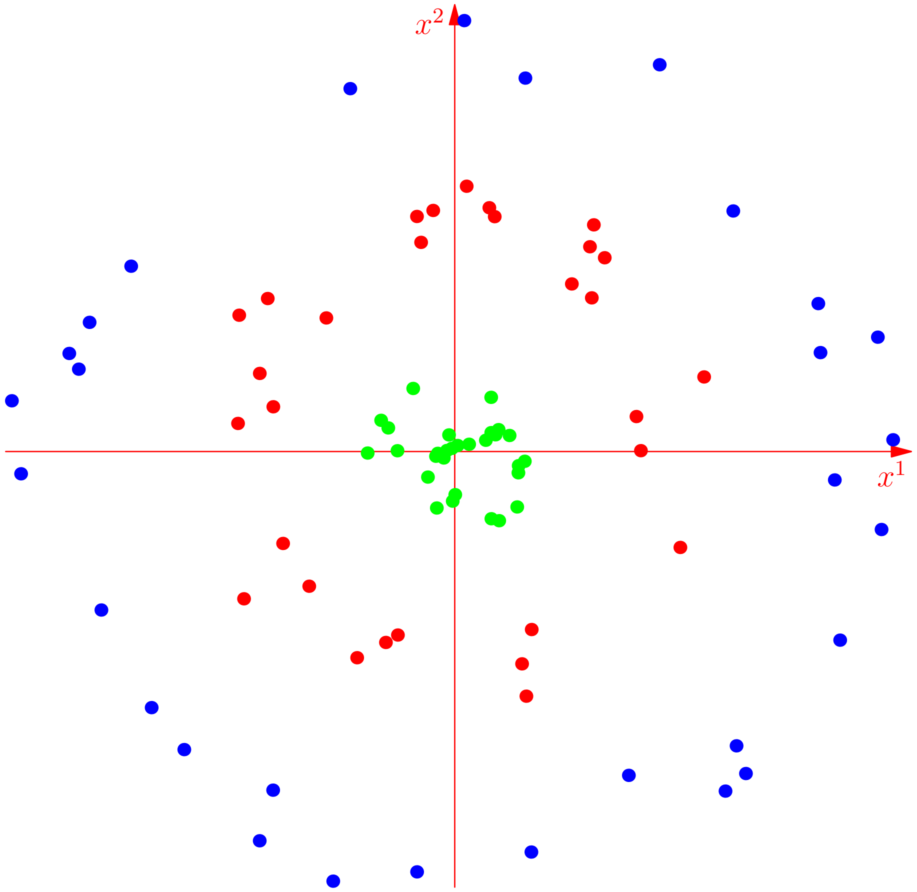
   $$\sum_{i=1}^{n} \|x_i - P_V(x_i)\|_2^2$$

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

4. Converting your data into principal components can hurt interpretability since the new features are linear combinations (i.e., blends or baskets) of your old features.

5. The smallest principal components, if they correspond to small eigenvalues, are nearly in the null space of X, and thus reveal linear dependencies in the centered data.

Example 3. Suppose you have the following data:



1. How can we get the first principal component to properly distinguish the rings above?

   Solution. Add features or use kernels. Below we added the feature $\|\tilde{x}_i\|^2$ and took the first principal component.