PRINCIPAL COMPONENT ANALYSIS

Dimensionality Reduction

In dimensionality reduction problems we observe

$$x_1, \ldots, x_n \in \mathbb{R}^d$$

and the goal is to transform these variables to new ones

$$x_i \longrightarrow \theta_i \in \mathbb{R}^k$$

where $k < d$, such that information loss is minimized.

Why might dimensionality reduction be useful?

1. Visualization ($k = 1, 2,$ or 3)
2. Remove noisy dimensions, to improve the performance of another algorithm
3. Decrease computational / storage requirements

Methods for dimensionality reduction can be classified

according to

1. How is "information loss" quantified?
2. Supervised or unsupervised? If outputs $y_1, \ldots, y_n$ are available, are they used?
3. Feature selection or feature extraction

$$\theta = \begin{bmatrix} x^{(2)} \\ x^{(7)} \\ \vdots \end{bmatrix}$$

select from existing features

$$\theta = \begin{bmatrix} x^{(1)} e^{-x^{(4)}} \\ (x^{(5)})^2 + 2x^{(12)} \\ \vdots \end{bmatrix}$$

create new features that are functions of old ones

In addition, we can classify DR methods by whether they are linear or nonlinear, generative or discriminative, and parametric or nonparametric.

PCA is described by

1. Sum of squared errors

2. Unsupervised
3. Feature extraction

In addition, it is linear, parametric, and can be viewed as either generative or discriminative (our perspective is discriminative).

Linear Spans and Projections

Let $a_1, \ldots, a_k \in \mathbb{R}^d$ be linearly independent column vectors. Denote

$$A = \begin{bmatrix} a_1 & \cdots & a_k \end{bmatrix} \qquad (d \times k)$$

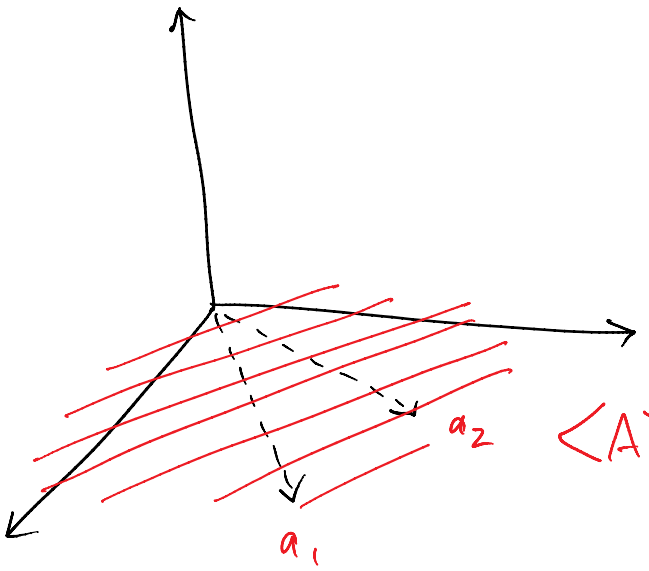
The linear span of $a_1, \ldots, a_k$ is the column span of A, written

$$\langle A \rangle = \text{colspan}(A)$$

$$= \left\{ x \in \mathbb{R}^d \;\middle|\; x = \sum_{j=1}^{k} \theta^{(j)} a_j, \; \theta = \begin{bmatrix} \theta^{(1)} \\ \vdots \\ \theta^{(k)} \end{bmatrix} \in \mathbb{R}^k \right\}.$$

$d = 3$

$k = 2$

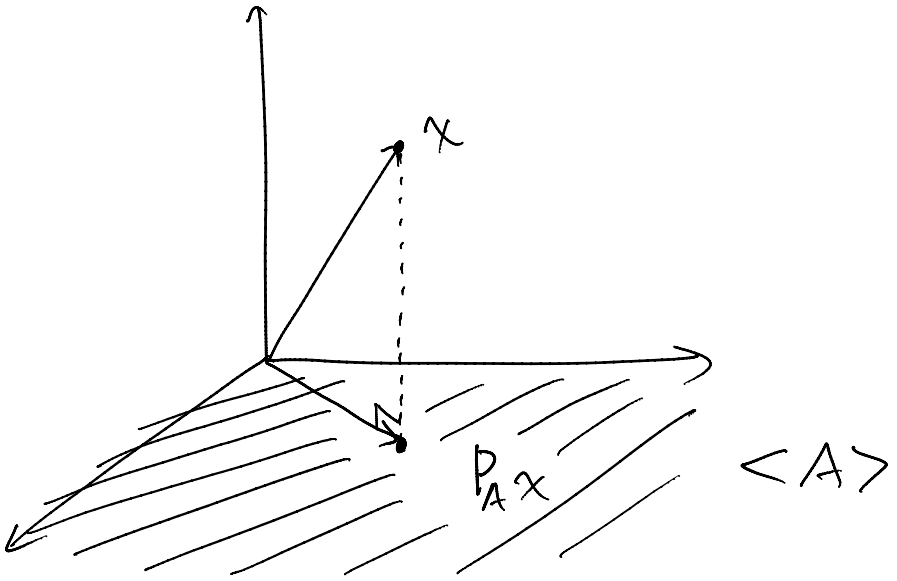


$\langle A \rangle \subseteq \mathbb{R}^d$

The projection onto $\langle A \rangle$ is the mapping

$P_A : \mathbb{R}^d \longrightarrow \langle A \rangle \subseteq \mathbb{R}^d$ given by

$P_A x$ = closest point to x in $\langle A \rangle$



Every point in $\langle A \rangle$ equals $A\theta$ for some $\theta \in \mathbb{R}^k$.

Therefore, $P_A x = A\hat{\theta}$ where

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^k} \|x - A\theta\|_2^2$$

We have already seen in our study of linear regression that the solution is

$$\hat{\theta} = \underbrace{(A^TA)^{-1}A^T}_{\text{pseudoinverse of } A} x$$

Since A has full rank, A^TA is invertible (exercise).

Therefore,

$$P_A x = \underbrace{A(A^TA)^{-1}A^T}_{\text{projection matrix } (d \times d)} x$$

If $a_1, \ldots, a_k$ are orthonormal, meaning

$$\langle a_i, a_j \rangle = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

then $P_A = A \cdot A^T$ $(d \times d)$.

The orthogonality principle states that $\forall x$, $x - P_A x$ is orthogonal to every element of $\langle A \rangle$.

(see previous figure). To see that, let $A\theta$ denote an arbitrary element of $\langle A \rangle$. Then

$$\langle A\theta, x - P_A x \rangle = \theta^T A^T \left(x - A(A^TA)^{-1}A^T x \right)$$

$$= \theta^T (A^T x - A^T x)$$

$$= 0.$$

Projection matrices are also <u>idempotent</u>, which means $P_A^2 = P_A$. This follows from

$$P_A^2 = P_A \cdot P_A$$

$$= A(A^TA)^{-1}A^T \cdot A(A^TA)^{-1}A^T$$

$$= A(A^TA)^{-1}A^T$$

$$= P_A.$$

Intuitively, the second projection has no effect because $P_A x \in \langle A \rangle$ already.

Also note that projection matrices are symmetric and positive-semidefinite.

If $b_1, \ldots, b_k \in \mathbb{R}^d$ and $B = [b_1 \cdots b_k]$ is such that $\langle B \rangle = \langle A \rangle$, then $P_B = P_A$. Therefore, we can assume $a_1, \ldots, a_k$ are orthonormal.

Let A_k denote the set of $n \times k$ matrices with orthonormal columns.

Suppose $a_1, \ldots, a_k$ are orthonormal, and let $a_{k+1}, \ldots, a_d$ extend to an orthonormal basis of $\mathbb{R}^d$. Set

$$A = \underset{d \times k}{\begin{bmatrix} a_1 & \cdots & a_k \end{bmatrix}}, \quad B = \underset{d \times (d-k)}{\begin{bmatrix} a_{k+1} & \cdots & a_d \end{bmatrix}}$$

Every $x \in \mathbb{R}^d$ has a unique representation as

$$x = u + v$$

where $u \in \langle A \rangle$ and $v \in \langle B \rangle$. Therefore

$$P_A + P_B = I$$

because $\forall x$, $P_A x + P_B x = u + v = x$.

$\langle B \rangle$ is called the orthogonal complement of $\langle A \rangle$.

PCA

The idea behind PCA is to approximate

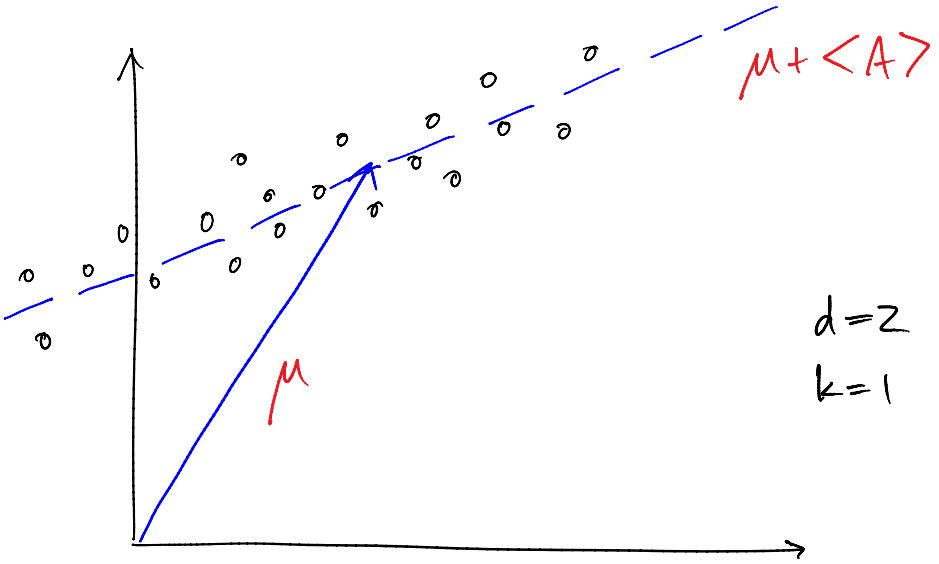
$$x_i \approx \mu + A\theta_i$$

where

$$\mu \in \mathbb{R}^d$$

$$A \in A_k$$

$$\theta_i \in \mathbb{R}^k$$



Mathematically, we define $\mu, A, \theta_1, \ldots, \theta_n$ to be the solution of

$$\star \quad \min_{\substack{\mu \in \mathbb{R}^d \\ A \in A_k \\ \theta_i \in \mathbb{R}^k}} \sum_{i=1}^{n} \|x_i - \mu - A\theta_i\|^2$$

PCA gives the least squares rank-k linear approximation to the data set.

The solution to ★ is given in terms of the spectral (or eigenvalue) decomposition of the sample covariance matrix

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T.$$

In particular, write

$$S = U \Lambda U^T$$

where

$$U = \begin{bmatrix} u_1 & \cdots & u_d \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix}$$

with $U^T U = U U^T = I$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$.

Recall that $S u_j = \lambda u_j \quad \forall j$.

We will show that a solution to ★ is

$$\mu = \bar{x}$$

$$A = \begin{bmatrix} u_1 & \cdots & u_k \end{bmatrix}$$

$$\theta_i = A^T(x - \bar{x})$$

We will also characterize the set of all solutions.

Some terminology:

- principal component transform:

$$x \longmapsto A^T(x - \bar{x}) \in \mathbb{R}^k$$

- jth principal component:

$$\theta^{(j)} = u_j^T(x - \bar{x}) \in \mathbb{R}$$

- jth principal eigenvector $\longrightarrow u_j \in \mathbb{R}^d$

Here's a picture when $d=3$, $k=2$, taken from Hastie, Tibshirani, and Friedman, The Elements of Statistical Learning.

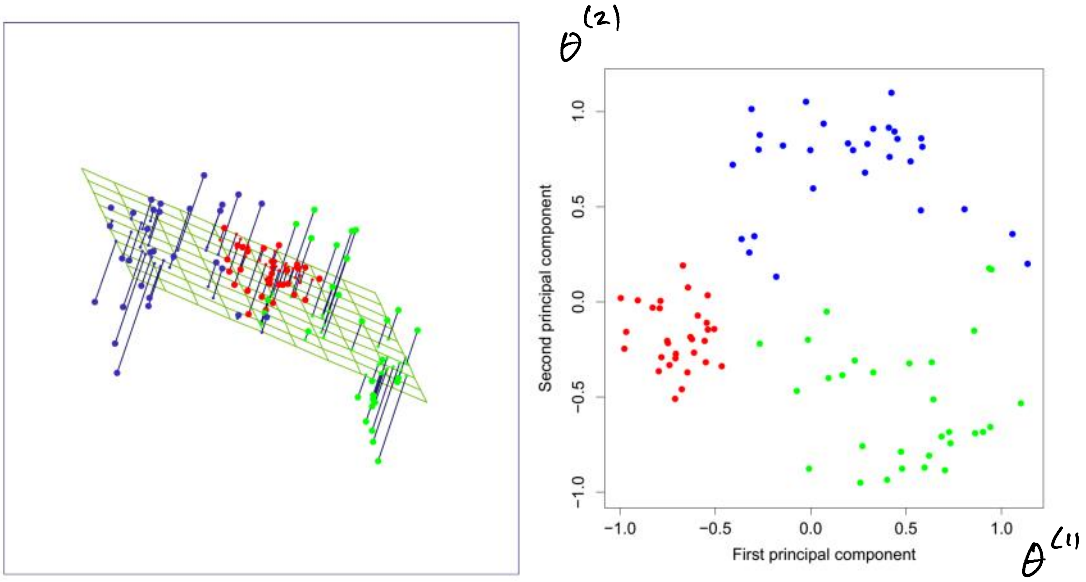


FIGURE 14.21. *The best rank-two linear approximation to the half-sphere data. The right panel shows the projected points with coordinates [redacted], the first two principal components of the data.*

PCA Derivation

PCA Derivation

We want to minimize

$$\sum_{i=1}^{n} \|x_i - \mu - A\theta_i\|^2$$

wrt to $\mu \in \mathbb{R}^d$, $A \in \mathcal{A}_k$, $\theta_i \in \mathbb{R}^k$.

Step 1: Eliminate $\{\theta_i\}$

Suppose A, μ are fixed. We can optimize each term wrt θ_i individually, yielding

$$\theta_i = (A^T A)^{-1} A^T (x - \mu)$$

$$= A^T (x - \mu)$$

Step 2: Eliminate μ

Holding A fixed, we wish to minimize

$$\sum_{i=1}^{n} \|x_i - \mu - AA^T(x_i - \mu)\|^2$$

$$= \sum_{i=1}^{n} \|(I - AA^T)(x_i - \mu)\|^2$$

$$= \sum_{i=1}^{n} (x_i - \mu)^T (I - AA^T)^T (I - AA^T)(x_i - \mu)$$

Note that $I - AA^T$ is a projection matrix onto

the orthogonal complement of $\langle A \rangle$. Therefore

$$(I - AA^T)^T (I - AA^T) = (I - AA^T)(I - AA^T)$$
$$= I - AA^T$$

Note that $I - AA^T$, being a projection matrix, is PSD, and therefore

$$\sum_{i=1}^{n} (x - \mu_i)^T (I - AA^T)(x_i - \mu)$$

is a convex function of μ. Now

$$\frac{\partial}{\partial \mu} = \sum_{i=1}^{n} (I - AA^T)(x_i - \mu)$$

$$= (I - AA^T) \sum_{i=1}^{n} (x_i - \mu)$$

$$= n (I - AA^T)(\bar{x} - \mu) = 0$$

is solved by $\mu = \bar{x}$. More generally, it suffices for $\bar{x} - \mu$ to belong to the nullspace of $I - AA^T$, which is $\langle A \rangle$. Thus any $\mu \in \bar{x} + \langle A \rangle$ is a possible solution.

Step 3 : Optimize A

It remains to solve

$$\min_{A \in \mathcal{A}_k} \sum_{i=1}^{n} \| x_i - \bar{x} - AA^T(x_i - \bar{x}) \|^2$$

Assume $\bar{x} = 0$ (otherwise we could substitute $\hat{x}_i = x_i - \bar{x}$; note that the sample covariance is not affected by subtracting the mean).

Also note that AA^T represents a rank-k projection matrix. Let $\mathcal{P}_k$ denote the set of $d \times d$, rank-k projection matricies. Then it remains to solve

$$\min_{P \in \mathcal{P}_k} \sum_{i=1}^{n} \| x_i - P x_i \|^2.$$

Introduce the data matrix

$$X = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix} \qquad (d \times n)$$

and for an arbitrary matrix $C = [c_{ij}]$, define the Frobenius norm

$$\| C \|_F := \sqrt{\sum_i \sum_j c_{ij}^2}$$

Then we can restate the problem as

Then we can restate the problem as

$$\min_{P \in \mathcal{P}_k} \| X - PX \|_F .$$

This is the core optimization problem at the heart of PCA. Although it is nonconvex, it does have a closed form solution given by $P = AA^T$ where A consists of the first k principal eigenvalues. The derivation of this result is quite interesting in its own right, and connects with other important topics in matrix algebra. These details are given in the next set of notes.

Maximum Variance Projections

PCA can be derived from a second perspective. Suppose $\bar{x} = 0$. What is the unit vector $a_1 \in \mathbb{R}^d$ $(\|a_1\| = 1)$ for which the variance of

$$\theta^{(1)} = a_1^T x$$

is maximized? The variance of $\theta^{(1)}$ is

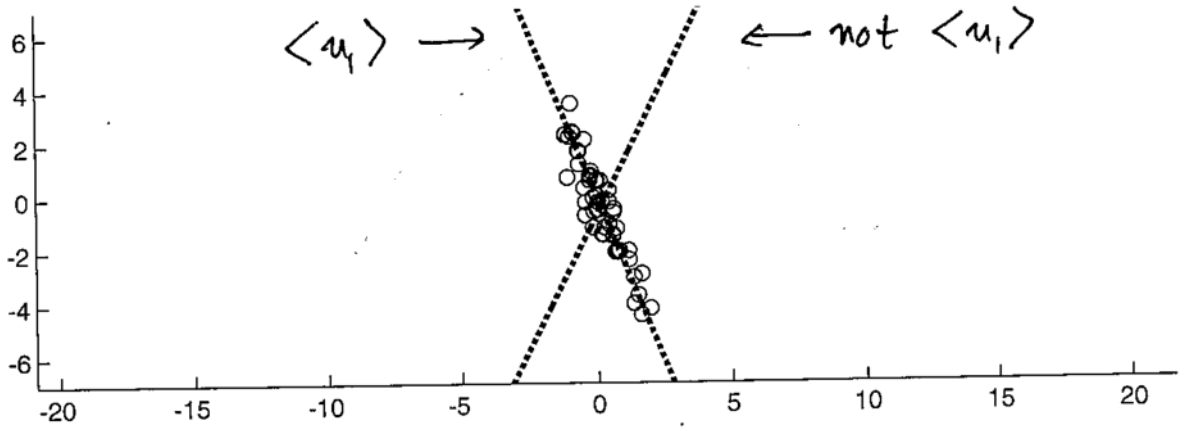
$$\text{Var}(\theta^{(1)}) = \frac{1}{n} \sum_{i=1}^{n} \left(a_1^T x_i \right)^2$$

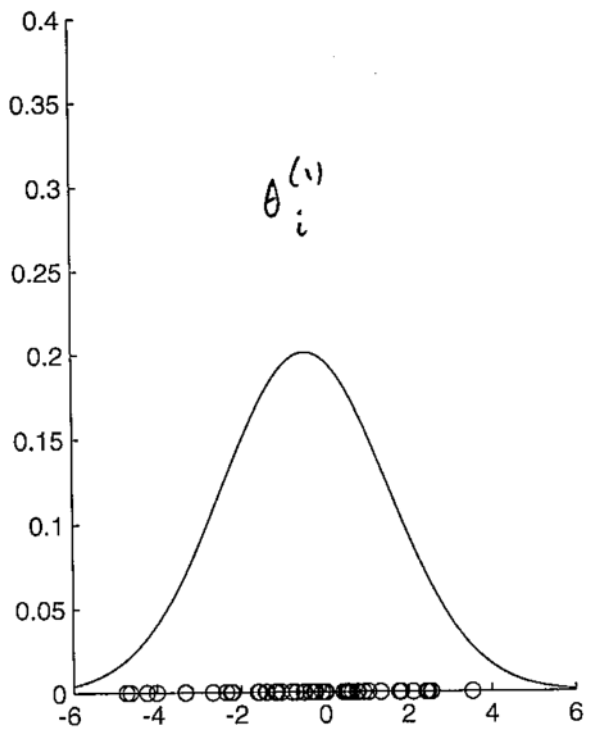
(note $\theta^{(1)}$ has zero mean since $\bar{x} = 0$). Well,

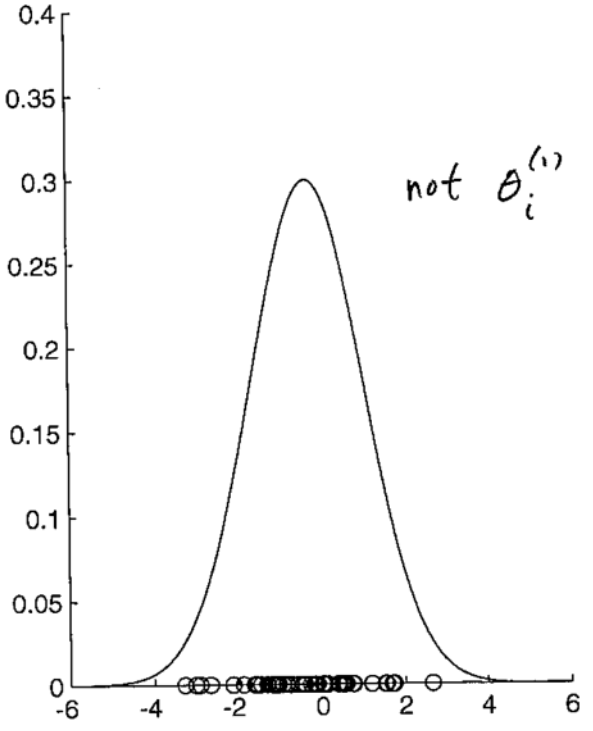
$$\begin{aligned} \mathrm{Var}(\theta^{(1)}) &= \frac{1}{n} \sum_{i=1}^{n} (a_1^T x_i)(a_1^T x_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} (a_1^T x_i)(x_i^T a_1) \\ &= a_1^T \left(\frac{1}{n} \sum x_i x_i^T \right) a_1 \end{aligned}$$

Maximizing this subject to $\|a_1\| = 1$ results in

$a_1 = u_1$, the first principal eigenvector of $S = \frac{1}{n} \sum x_i x_i^T$.







Furthermore, the remaining PC's emerge if we seek additional maximum variance directions orthogonal to the fust ones.

Theorem Let $\theta^{(k)} = a_k^T x$ and $\mathrm{Var}(\theta^{(k)}) = \frac{1}{n}\sum (a_k^T x_i)^2$.

A vector a_k that maximizes $\mathrm{Var}(\theta^{(k)})$ subject to

- $\| a_k \| = 1$
- $a_k \perp u_1, \ldots, u_{k-1}$

is $a_k = u_k$.

On the homework you will be asked to prove this result and characterize when the maximizer is unique.

and characterize when the maximizer is unique.

Selecting k

On the homework you will show that the optimal objective function value is

$$\min_{\mu, A, \theta_i} \sum \|x_i - \mu - A\theta_i\|^2 = n(\lambda_{k+1} + \cdots + \lambda_n)$$

When $k = 0$, this specializes to

$$\min_{\mu} \sum \|x_i - \mu\|^2 = n(\lambda_1 + \cdots + \lambda_n)$$

which we call the total variation of the data.

One heuristic for choosing k is to select the smallest k such that

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_n} \geq .95$$

fraction of total variation captured by PCA

arbitrary threshold

Preprocessing

It is common to center and scale data before applying PCA. This avoids problems that might arise if

different features have different units.

For $j = 1, \ldots, d$

$$\bar{x}^{(j)} \leftarrow \frac{1}{n} \sum_{i=1}^{n} x_i^{(j)}$$

$$x_i^{(j)} \leftarrow x_i^{(j)} - \bar{x}^{(j)} \qquad \forall i$$

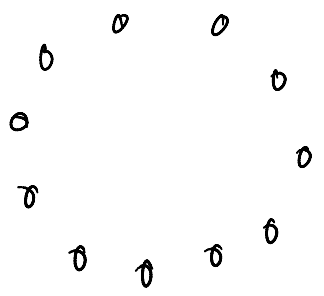
$$\sigma^{(j)} \leftarrow \frac{1}{n} \sum_{i=1}^{n} (x^{(j)})^2$$

$$x_i^{(j)} \leftarrow x_i^{(j)} / \sigma^{(j)} \qquad \forall i$$

End

Final Thoughts

PCA is a linear method, and therefore cannot capture nonlinear structure.



PCA ($k=1$) will fail to capture the circular structure.

However, PCA can be kernelized (known as "kernel PCA") which yields a method for ...

nonlinear dimensionality reduction.