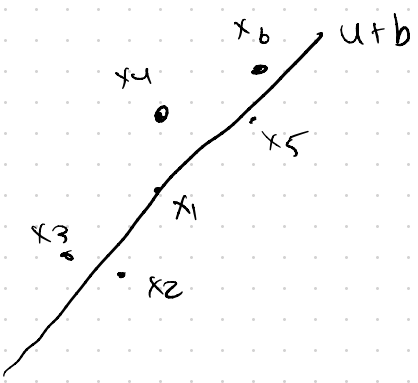
Principal Component Analysis (PCA)

Method for dimensionality reduction

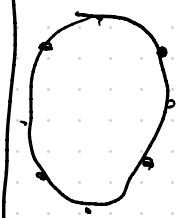
We focus now on input variables $x_1, \dots, x_n \in \mathbb{R}^D$.

Why PCA?
- data compression
- Learn shape of data (when $x_1, \dots, x_n$ are geometric)

Goal of PCA: find linear space $U \subseteq \mathbb{R}^D$ with $\dim(U) = d \ll D$ and a vector $b \in \mathbb{R}^D$ so that $U + b$ is "close to" $x_1, \dots, x_n$.



What about lower dimensional data that is not linear?



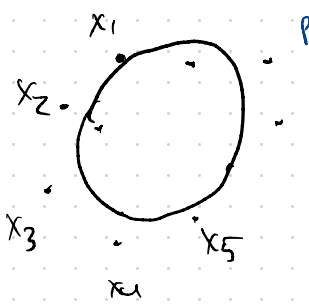
Compose with a feature map

$\phi : \mathbb{R}^D \to \mathbb{R}^M$

$z_i = \phi(x_i) \quad 1 \le i \le n$

ex $\phi: \mathbb{R}^2 \to \mathbb{R}^6$ $\phi(x_i) = z_i$

$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto (1, x_1, x_2, x_1 x_2, x_1^2, x_2^2)$



perhaps quadratic, so map to all monomials degree ≤ 2.

Image under ϕ very close to $z_5 + z_6 = 1$

The x_i live close to a zero set of a polynomial ⟸ The z_i live close to a linear space.

Called an "algebraic variety"

Suppose $x_1, \dots, x_n \overset{iid}{\sim} x \in \mathbb{R}^D$ (↑ random variable)

ϕ = feature map $\mathbb{R}^D \to \mathbb{R}^M$

$\Rightarrow z_1, \dots, z_n \overset{iid}{\sim} z \in \mathbb{R}^M$

Let $z = (z^{(1)} \dots z^{(M)})$, $\mu = \mathbb{E} z \in \mathbb{R}^M$ (expected value)

covariance matrix $\Sigma := \left[\mathrm{Cov}(z^{(i)}, z^{(j)}) \right]_{i,j=1}^{M} \in \mathbb{R}^{M \times M}$

$\mathrm{Cov}(X, Y) = \mathbb{E}XY - \mathbb{E}X\,\mathbb{E}Y$

Fact Σ is positive semidefinite

$\Leftrightarrow$ diagonalizable with orthogonal matrix P.

i.e. $P \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_M \end{pmatrix} P^T = \Sigma$

with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_M \geq 0$

We don't know distribution of Z, so we want to approximate Σ.

Definition $z_1, \ldots, z_n \in \mathbb{R}^M$

1. <u>empirical average</u> is $\bar{z} = \frac{1}{n}(z_1 + \cdots + z_n)$

2. <u>empirical covariance matrix</u> is $S = (s_{ij})_{i,j=1}^{M}$

with

$$s_{ij} = \frac{1}{n} \sum_{k=1}^{n} \left[(z_k)_i - \bar{z}_i \right] \cdot \left[(z_k)_j - \bar{z}_j \right]$$

equivalently for Ω the feature matrix

$$\Omega = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} \in \mathbb{R}^{n \times M}$$

Then $\bar{z} = \frac{1}{n} \Omega^T e$

$$e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$$

$$S = \frac{1}{n} \left(\Omega - e\bar{z}^T \right)^T \left(\Omega - e\bar{z}^T \right)$$

Notion of distance for $1 \le d < M$,

A space of maximal variance is a linear space $U \subseteq \mathbb{R}^M$ of $\dim(U) = d$ so

$$U \in \underset{U : \dim(U)=d}{\operatorname{argmax}} \; \mathbb{E} \left|\left| P_U(z-\mu) \right|\right|^2$$

where P_U is orthogonal projection onto U.

Thm 3.39 $\lambda_1 \ge \cdots \ge \lambda_M \ge 0$ eigenvalues of Σ, u_i eigenvectors associated to λ_i so that $\langle u_i, u_j \rangle = \delta_{ij}$. Then

$$U = \operatorname{span}\{u_1, \ldots, u_d\}$$

is a space of maximal variance of dimension d.

Moreover when $\lambda_d > \lambda_{d+1}$, U is uniquely determined.

Problem Σ hard to know, so we need to use S.

Note on choosing d

Can choose $\lambda_{d+1} \approx 0$

or maximize $\frac{\lambda_d}{\lambda_{d+1}}$ or $\lambda_d - \lambda_{d+1}$ } art of Data science.

Definition $u_1, \dots, u_d$ are called <u>principal components</u> of Σ.

From Thm 3.39 we compress data by mapping

$$z_i \longmapsto P_U (z_i - \bar{z}) + \bar{z}$$

To mak computationally viable, we minimize distance to data points:

$$\text{minimize} \quad \sum_{i=1}^{n} \| (z_i - \bar{z}) - P_U (z_i - \bar{z}) \| \qquad (*)$$

Theorem Let $\lambda_1 \geq \cdots \geq \lambda_M \geq 0$ be eigenvectors of S.

Let u_i orthonormal basis of eigenvectors λ_i

So $\langle u_i, u_j \rangle = \delta_{ij}$. Then $U = \text{Span}\{u_1, \dots, u_d\}$

minimizes $(*)$ If $\lambda_d > \lambda_{d+1}$, U is uniquely determined.

Pf Let u_i be any orthonormal basis of $\mathbb{R}^M$.

<u>Center data</u> set $w_i = z_i - \bar{z}$

Let $A = [u_1 \cdots u_d] \in \mathbb{R}^{M \times d}$ thus $\begin{bmatrix} u_1^T \\ \vdots \\ u_d^T \end{bmatrix}$

$$P_U = AA^T = [u_1 \cdots u_d] \sum_{i=1}^{d} u_i u_i^T$$

Set $W = (\Sigma - e\bar{z}^T)^T \in \mathbb{R}^{M \times n}$

So $nS = WW^T$

Since $I_M = \sum_{i=1}^{M} u_i u_i^T$,

$$W - P_d W = (I_M - AA^T) W = \left(\sum_{i=d+1}^{M} u_i u_i^T \right) W$$

So (*) minimizes $\sum_{i=1}^{n} \|w_i - P_d(w_i)\|^2$

$$= \text{Trace}\,(W - P_d W)^T (W - P_d W)$$

$$= \text{Trace}\left(W^T \left(\sum_{i=d+1}^{M} u_i u_i^T \right)^T \left(\sum_{i=d+1}^{M} u_i u_i^T \right) W \right)$$

$$= \text{Trace}\left(\sum_{i=d+1}^{M} \sum_{j=d+1}^{M} u_i u_i^T u_j u_j^T \right) WW^T$$

(Since trace is cyclic invariant

$$= \text{Trace}\left(\sum_{i=d+1}^{M} u_i u_i^T \, WW^T \right)$$

(since $\langle u_i, u_j \rangle = 1 \Leftrightarrow i = j$

$$(\text{linearity of trace}) = \sum_{i=d+1}^{M} \text{Trace}\,(u_i u_i^T WW^T)$$

$\text{trace}(u^T u)$
where $u = [u_1 \cdots u_n]$

$u^T u \in \mathbb{R}^{n \times n}$

$$\begin{bmatrix} u_1^T \\ \vdots \\ u_n^T \end{bmatrix} \begin{bmatrix} u_1 \cdots u_n \end{bmatrix} = \begin{bmatrix} u_1^T u_1 & & \\ & \ddots & \\ & & u_n^T u_n \end{bmatrix}$$

$u_i^T u_i = \|u_i\|^2$

$(nS = WW^T)$

$$= n \sum_{i=d+1}^{M} \text{Trace}\left(u_i^T S u_i\right)$$

$Su_i = \lambda_i u_i$ If u_i eigenvectors of S

$$= n \sum_{i=d+1}^{M} \lambda_i \, \text{Trace}\left(u_i u_i^T\right)$$

$$= \boxed{n \sum_{i=d+1}^{M} \lambda_i}$$

This is the squared distance! Minimized by taking smallest eigenvalues!

Proof of minimization using Lagrange multipliers

$$\text{minimize} \quad n \sum_{i=d+1}^{M} u_i^T S u_i$$

Subject to constraints $\langle u_i, u_j \rangle = \delta_{ij}$

Add parameter l_{ij} for $\underset{d+1 \le i \le j}{j = d+1, \ldots, M}$ Then

$$\mathcal{L}\left(u_{d+1}, \ldots, u_M, l_{ij}\right) = \sum_{i=d+1}^{M} u_i^T S u_i - \sum_{d+1 \le i \le j \le M} (u_i^T u_j - \delta_{ij}) l_{ij}$$

Optimized when $\frac{\partial \mathcal{L}}{u_j} = 0$ and $\frac{\partial \mathcal{L}}{\partial l_{ij}} = 0$

$$\frac{\partial \mathcal{L}}{\partial u_j} = 2Su_j - 2l_{jj}u_j - \sum_{d+1 \le i < j} u_i l_{ij} = 0$$

$$d+1 \le j \le M$$

$$\frac{\partial \mathcal{L}}{\partial l_{ij}} = u_i^T u_j - \delta_{ij} = 0$$

When $j = d+1$

$$2Su_{d+1} = 2u_{d+1}l_{d+1,d+1}$$

$$\Rightarrow \quad l_{d+1,d+1} = \lambda_{d+1}$$

When $j = d+2$

$$2Su_{d+2} - 2l_{d+2,d+2}u_{d+2} - \left(u_{d+1}l_{d+1,d+2}\right) = 0$$

multiply by u_{d+1}^T on left

$$2u_{d+1}^T S u_{d+2} - \boxed{2l_{d+2,d+2}u_{d+1}^T u_{d+2}} - l_{d+1,d+2} = 0$$

$$\boxed{\quad = 0 \quad}$$

$$2u_{d+1}^T S u_{d+2} = l_{d+1,d+2}$$

$$\underline{\underline{l_{d+1,d+2}}} = \left(u_{d+1}^T S u_{d+2}\right)^T = u_{d+2}^T S^T u_{d+1}$$

$$\left(\begin{array}{l}\text{Since}\\ \text{symmetric}\end{array}\right) = u_{d+2}^T S u_{d+1}$$

$$= \lambda_{d+1}\, u_{d+2}^T u_{d+1} = 0$$

$$\Rightarrow l_{d+1,d+2} = 0 \quad \text{and thus}$$

$$S u_{d+2} = l_{d+2,d+2}\, u_{d+2} \Rightarrow l_{d+2,d+2} = \lambda_{d+2}$$

$\rightsquigarrow$ And so on to conclude $l_{i,j} = \begin{cases} \lambda_j & i = j \\ 0 & i \neq j \end{cases}$

So the eigenvalues give optimal solutions (i.e. derivative zero), the smallest eigenvalue is the minimum.

$\approx$ Note $\lambda_d > \lambda_{d+1}$ makes U unique since there are no other choices for U.

Next time using SVD to get eigenvalues & Notebook 6.