TAM 598 Lecture 18:

Unsupervised Learning - Dimensionality Reduction

Announcements:

- No class on Wednesday
- HW 5 covers lectures 17-20; due on Fri Apr 18

Dimensionality Reduction -

- we have observations $\underset{\sim}{x}_{1:n}$, each of which is a high-dimensional vector $\underset{\sim}{x}_i \in \mathbb{R}^D$, with $D \gg 1$.
- our goal is to describe the data set with a smaller number of dimensions without losing too much information, ie to project each $\underset{\sim}{x}_i$ to a d-dimensional vector $\underset{\sim}{z}_i$ where $d \ll D$.

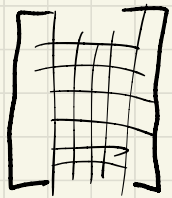
why? for visualization
to do clustering, density estimation, ...
for supervised learning tasks

Principal Component Analysis :

* we want a linear map from $\underset{\sim}{x}_i \in \mathbb{R}^D$ to $\underset{\sim}{z}_i \in \mathbb{R}^d$

* use an affine projection map $\mathbb{R}^D \to \mathbb{R}^d$

$$\underset{\sim}{z} = f(\underset{\sim}{x}) = \underset{=}{W}^T (\underset{\sim}{x} - \underset{\sim}{x}_0)$$

$\underset{=}{W}$ =  (D × d)

where
- $\underset{=}{W}$ is a $D \times d$ matrix
- $\underset{\sim}{x}_0$ is the empirical mean of the data $\underset{\sim}{x}_0 = \frac{1}{n} \sum_{i=1}^{n} \underset{\sim}{x}_i$

* use a reconstruction map $\mathbb{R}^d \to \mathbb{R}^D$

$$\underset{\sim}{x} = g(\underset{\sim}{z}) = \underset{=}{V}\ \underset{\sim}{z} + \underset{\sim}{x}_0$$

$\underset{=}{V}$ is a $D \times d$ matrix 

To find the matrices $\underline{\underline{W}}$, $\underline{\underline{V}}$ we minimize the reconstruction error.

$$L(\underline{\underline{W}}, \underline{\underline{V}}, \underset{\sim}{x}_0) = \frac{1}{n} \sum_{i=1}^{n} \left\| \underset{\sim}{x}_i - g(f(\underset{\sim}{x}_i)) \right\|^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left\| \underset{\sim}{x}_i - \underline{\underline{V}}\, \underline{\underline{W}}^T \underset{\sim}{x}_i - \underset{\sim}{x}_0 \right\|^2$$

Taking $\frac{\partial L}{\partial \underline{\underline{W}}} = 0$, $\frac{\partial L}{\partial \underline{\underline{V}}} = 0$ and solving:

(1) $\underline{\underline{W}} = \underline{\underline{V}}$

(2) $\underline{\underline{V}}$ can be constructed from eigenvectors / eigenvalues of empirical covariance matrix $\underline{\underline{C}} = \frac{1}{n} \sum_{i=1}^{n} (\underset{\sim}{x}_i - \underset{\sim}{x}_0)(\underset{\sim}{x}_i - \underset{\sim}{x}_0)^T$

D x D

$\underset{\sim}{x}_2 - \underset{\sim}{x}_0 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$

$\underset{\sim}{x}_1 - \underset{\sim}{x}_0 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$

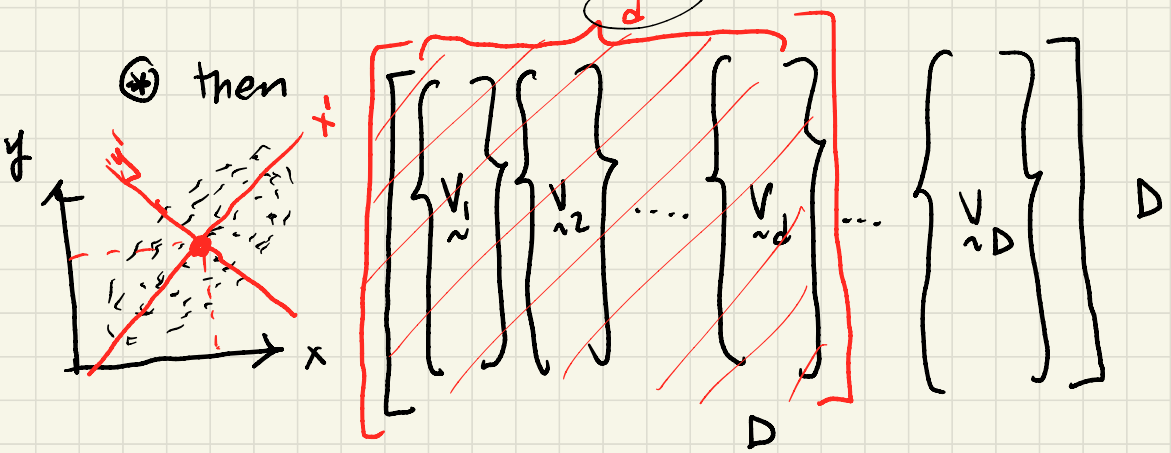
$i = 1$: $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} [1\ 2\ 3]$

Specifically:

⊛ let $\underset{\sim}{u}_i$ and λ_i be the i^{th} eigenvalue of $\underline{\underline{C}}$, sorted so

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D$$

⊛ then



column i is $\underset{\sim}{v}_i = \sqrt{\lambda_i}\,\underset{\sim}{u}_i$

$\underline{\underline{V}}$ is $D \times d$

⊛ our projection map becomes

$$\underset{\sim}{z} = f(\underset{\sim}{x}) = \sum_{i=1}^{d} \sqrt{\lambda_i}\, u_i^T \underset{\sim}{x} = \underline{\underline{V}}^T \underline{x}$$

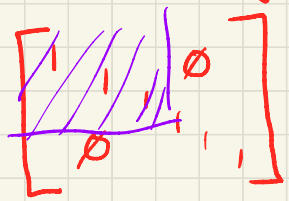
⊛ the reconstruction map is $\underset{\sim}{x} = g(\underset{\sim}{z}) = \underset{\sim}{x}_0 + \sum_{i=1}^{d} \sqrt{\lambda_i}\, \underset{\sim}{u}_i z_i$

* The sum of the first d eigenvectors

$$\sum_{i=1}^{d} \lambda_i$$

tells us how much variance is explained with a model that keeps the first d PCA components

Σ in principal coord system



* the minimum reconstruction error is

$$L(\underline{\underline{W}}, \underline{\underline{V}}, \underline{x}_0) = \frac{1}{n} \sum_{i=1}^{n} \| \underline{x}_i - g(f(\underline{x}_i)) \|^2$$

$$= \sum_{j=d+1}^{D} \lambda_j$$

probabilistic interpretation: assume our data points $\underset{\sim}{x}_{1:n}$ are generated by a linear Gaussian model

and latent variables $\underset{\sim}{z}_i$ are generated by a Gaussian prior

We can maximize the marginal likelihood:

which is a gaussian. Maximizing $\log p(\underset{\sim}{x}_{1:n})$ gives the same result as before, now with variance