Factor Analysis & PCA — Subspace structure.

Structure	Prob.	Non Prob
"Cluster"	GMM	K-means
"Subspace"	Factor Analysis	PCA ← Today's lecture

Factor Analysis Given $x^{(1)} \dots x^{(n)} \in \mathbb{R}^d$

"$n \ll d$" #data points much smaller than dims

IDEA: Posit some structure, use to reduce dimensions.

(Spoiler: subspace)

technical Challenge: Estimating Gaussian hard when $n \ll d$

Challenge $\Longrightarrow$

$$p(x \mid \mu, \Sigma_i) = \frac{1}{\sqrt{2\pi}\,|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma_i^{-1} (x-\mu)\right\}$$

det $= 0$! undefined

look at cases (assumptions) so that Σ_i is full rank but fewer params.

Recall MLE for Gaussian is Equivalent to

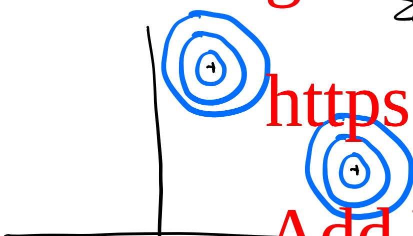
$$\min_{\mu, \Sigma_i} \sum_{i=1}^{n} (x-\mu)^T \Sigma_i^{-1} (x-\mu) + \log |\Sigma_i|$$

If Σ_i is full rank, $\nabla_\mu = \sum_{i=1}^{n} \Sigma_i^{-1}(x-\mu) = 0 \Rightarrow \mu = \frac{1}{n} \sum_i x^{(i)}$

Building Block 1

Suppose independent and identical covariance

$\Sigma_i = \sigma^2 I$ (NB: Parameter Tying)



Covariance "are circles"

What is MLE for Σ_i? $|\Sigma_i| = 2d$

$$\min_{\sigma^2} \sigma^{-2} \underbrace{\sum_{i=1}^{n} (x^{(i)}-\mu)^T (x-\mu)}_{c} + d \log \sigma^2$$

let $z = \sigma^2$ $\quad \min_{z} \frac{1}{z} c + d \log z$

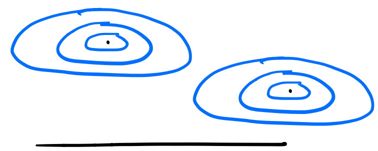
$$\Rightarrow \nabla_z = -z^{-2} c + \frac{nd}{z} = 0 \Rightarrow z = \frac{c}{nd}$$

$$\therefore \sigma^2 = \frac{1}{nd} \sum_{i=1}^{n} (x^{(i)}-\mu)^T (x^{(i)}-\mu)$$

"Substract mean and square all entries."

Building Block 2

$$\Sigma_i = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_d^2 \end{bmatrix}$$



Axis Aligned ellipse

SET $z_i = \sigma_i^2$ (SAME IDEA AS ABOVE)

$$\min_{z_1 \dots z_d} \sum_{i=1}^{n} \sum_{j=1}^{d} z_j^{-1} (x^{(i)} - \mu_j)^2 + \log z_j$$

this is d Problems for EACH 1 dimension

$$\Rightarrow \sum_{i=1}^{n} z_j^{-1} (x^{(i)} - \mu_j)^2 + \log z_j$$

$$\Rightarrow \sigma_j^2 = \frac{1}{n} \sum_i (x_j^{(i)} - \mu_j)^2$$

One FACTOR model

PARAMETERS

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

$\Lambda \in \mathbb{R}^{d \times s}$

$\Phi \in \mathbb{R}^{d \times d}$ — DIAGONAL MATRIX

MODEL

$$P(x, z) = P(x \mid z) P(z) \qquad z \text{ IS LATENT}$$

$z \sim N(0, I) \in \mathbb{R}^s$ for $s < d$ "small dim"

$x = \mu + \Lambda z + \epsilon$ OR $x \sim N(\mu + \Lambda z, \Phi)$

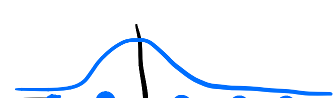
MEAN IN the SPACE

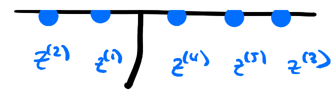
$\hookrightarrow$ MAPS FROM SMALL LATENT SPACE TO LARGE SPACE

$\epsilon \sim N(0, \Phi)$ Noisy

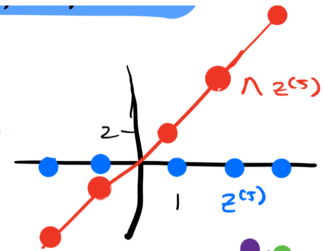
Ex: $d = 2$, $s = 1$, $n = 5$ $\qquad x = \mu + \Lambda z + \epsilon$

1. GENERATE $z^{(1)}, \dots, z^{(5)}$ from $N(0, 1)$





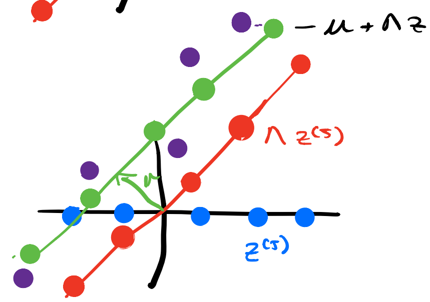
2. Suppose $\Lambda = \begin{bmatrix}1\\2\end{bmatrix}$



3. Add μ

4. Add ϵ

↳ Big space



Data we would observe are purple dots

So small latent space produces data in high dim space.

Technical tools: Block Gaussians

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad x_1 \in \mathbb{R}^{d_1},\ x_2 \in \mathbb{R}^{d_2}$$

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

$$\Sigma = \begin{array}{c c} \begin{array}{cc} d_1 & d_2 \end{array} & \\ \begin{bmatrix} \Sigma_{11} & \Sigma_{21} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix} & \begin{array}{c} d_1 \\ d_2 \end{array} \end{array} \quad \Sigma_{ij} \in \mathbb{R}^{d_i \times d_j} \quad i,j \in \{1,2\}$$

Notation is widely used and helpful.

Fact 1: $P(x_1) = \int_{x_2} P(x_1, x_2)$ MARGINALIZATION

For Gaussians, $p(x_1) = \mathcal{N}(\mu_{11}, \Sigma_{11})$ (Not surprising)

Fact 2: $P(x_1 | x_2) \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$ Conditioning

$$\mu_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2)$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\,\Sigma_{22}^{-1}\,\Sigma_{21} \qquad \text{(matrix inversion lemma)}$$

Proofs outline (happy to add)

Summary: Marginalization & conditioning Gaussian $\Rightarrow$ another Gaussian (closed)

we have formula for parameters.

Back to Factor Analysis $\qquad x = \mu + \Lambda z + \epsilon$

$$\begin{pmatrix} z \\ x \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \Sigma \right) \qquad \text{since } \mathbb{E}[z] = 0$$

$$\mathbb{E}[x] = \mu$$

What is Σ?

$$\Sigma_{11} = \mathbb{E}[zz^T] = I$$

$$\Sigma_{12} = \mathbb{E}[z(x-\mu)^T] = \mathbb{E}[zz^T\Lambda^T] + \mathbb{E}[\cancel{z\epsilon^T}]^{\,0}$$
$$= \Lambda^T$$

$$\Sigma_{21} = \Sigma_{12}^T$$

$$\begin{aligned}
\Sigma_{22} &= \mathbb{E}[(x-\mu)(x-\mu)^T] \\
&= \mathbb{E}[(\Lambda z + \epsilon)(\Lambda z + \epsilon)^T] \\
&= \mathbb{E}[\Lambda z z^T \Lambda^T] + \mathbb{E}[\epsilon\epsilon^T] \\
&= \Lambda\Lambda^T + \Phi
\end{aligned}$$

$$\Sigma = \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Phi \end{bmatrix}$$

E-STEP: $Q_i(z) = P(z^{(i)} | x^{(i)}; \theta)$ — USE CONDITIONAL!

M-STEP: WE HAVE CLOSED FORMS!

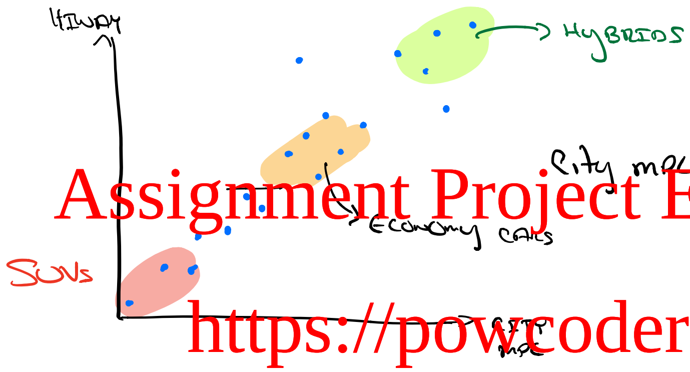
Summary of Factor Analysis

- WE LEARNED ABOUT FACTOR Analysis (LATENT low dim. STRUCTURE)
- WE SAW HOW to ESTIMATE PARAMETERS OF FA using EM.

PCA: Principal Component Analysis

Structure	Prob.	Non Prob
"Cluster"	GMM	K-means
"Subspace"	Factor Analysis	PCA

Ex: Given pairs (Hiway mpg, city mpg) of some cars

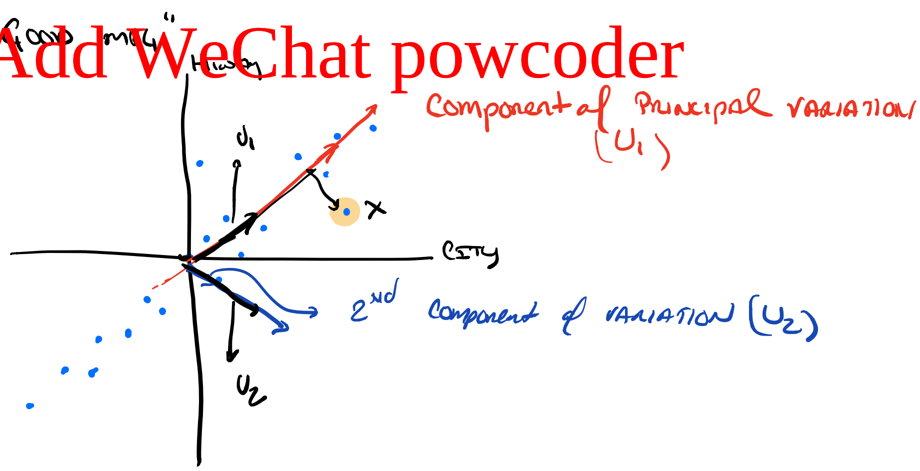


Question: "Good mpg"

① Center data

$\mu = \frac{1}{n}\sum_i x^{(i)}$

$x^{(i)} \mapsto x^{(i)} - \mu$



Now $\|u_1\| = \|u_2\| = 1$ by convention.

- u_1 is "how good is mpg"
- u_2 is 'difference between Hiway & city' (roughly)

We can write $x = \alpha_1 u_1 + \alpha_2 u_2$

↳ we may just keep this component

"Explains more variation"

TODAY: How we find these directions, and some caveats

- think about 1000s of dims $\longrightarrow$ 10s of dims
- A dimensionality reduction method

Preprocessing

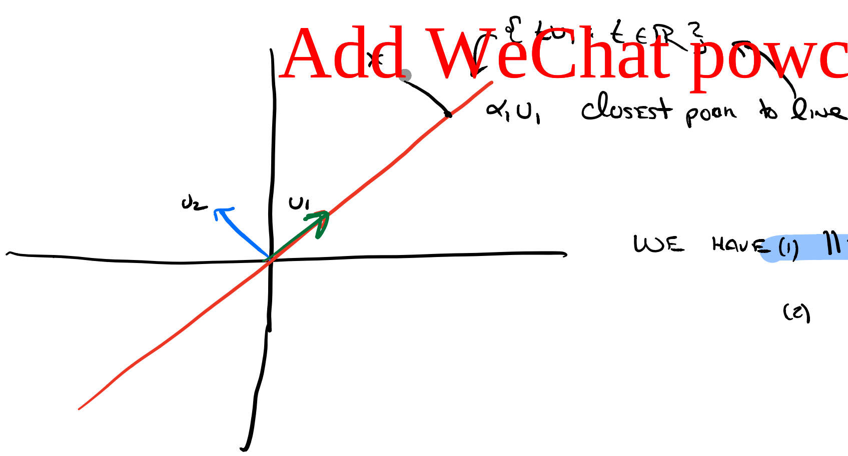
Given $x^{(1)} \dots x^{(n)} \in \mathbb{R}^d$

1. Center the data $x^{(i)} \mapsto x^{(i)} - \mu$ in which $\mu = \frac{1}{n} \sum_i x^{(i)}$
2. May need to rescale components e.g. "feet per gallon"

d, n_{PC}

We will assume data is preprocessed

PCA as Optimization



We have (1) $\|u_i\| = 1$ (unit vectors)

(2) $u_i \cdot u_j = \delta_{ij}$ (orthogonal)

How do you find closest point to the line?

$$\alpha_1 = \underset{\alpha}{\operatorname{argmin}} \|x - \alpha u_1\|^2$$

$$= \underset{\alpha}{\operatorname{Argmin}} \|x\|^2 + \alpha^2 \cancel{\|u_1\|^2}^{\,1} - 2\alpha (u_i \cdot x)$$

differentiate wrt α $\quad 2(\alpha - u_i x) = 0 \Rightarrow \alpha = u_i \cdot x$

Generalize: $u_1 \cdots u_k \in \mathbb{R}^d$ AND $x \in \mathbb{R}^d$ USE $u_i \cdot u_j = \delta_{ij}$

$$\underset{\alpha_1 \cdots \alpha_d}{\operatorname{Argmin}} \|x - \sum_{i=1}^{k} \alpha_i u_i\|^2 = \underset{\alpha}{\operatorname{argmin}} \|x\|^2 + \sum_{i=1}^{n} \alpha_i^2 \|u_i\|^2 - 2\alpha_i \langle u_i \cdot x \rangle$$

Hence $\alpha_i = u_i \cdot x$

WE CALL $\|x - \sum_{i=1}^{k} \alpha_i u_i\|^2$ the RESIDUAL

WE CAN find PCA by either

IN class ① MAXIMIZE Projected Subspace

② MINIMIZE Residual

$$\underset{\substack{U \in \mathbb{R}^d \\ \|U\| = 1}}{\max} \frac{1}{n} \sum_{i=1}^{n} (U \cdot x^{(i)})^2$$ WE NEED some facts to solve this

LET A be symmetric & square, then

$A = U \Lambda U^T$ IN which

- $UU^T = I$ (ORTHONORMAL)
- Λ is diagonal

$\Lambda_{ii} = \lambda_i$ AND $\underbrace{\lambda_1 \geq \cdots \geq \lambda_n}_{\text{eigenvalues}}$ by convention

Recall: If $x = \sum_{i=1}^{n} \alpha_i u_i$ where $[u_1 \cdots u_n] = U$

STANDARD BASIS vector

$$Ax = U\Lambda U^T x = U\Lambda \sum_{i=1}^{n} \alpha_i e_i \qquad (u_i \cdot u_j = \delta_{ij})$$

$$= U \sum_{i=1}^{n} \lambda_i \alpha_i e_i \qquad \text{diagonal } \Lambda$$

$$= \sum \lambda_i \alpha_i u_i$$

If $x = c u_i$ then x is AN eigenvector, AND $Ax = \lambda_i x$

$$\max_{x:\|x\|^2=1} x^T A x = \max_{\alpha:\|\alpha\|^2=1} \sum_{i=1}^{n} \alpha_i^2 \lambda_i$$

Hence, we set $\alpha_i = 1$, the principal eigenvalue

Which x attains it? If $\lambda_1 = \lambda_2$?

Now, back to PCA!

$$\max_{U:\|U\|^2=1} \frac{1}{n} \sum_{i=1}^{n} \underbrace{(U_i \cdot x^{(i)})^2}_{\text{the projection onto } U}$$

$$= \frac{1}{n} \sum_{i=1}^{n} U^T x^{(i)} (x^{(i)})^T U = U^T \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} x^{(i)} (x^{(i)})^T \right)}_{\text{covariance of data (we subtracted mean)}} U$$

$\therefore$ U is principal eigenvector

What if we want more dimensions? We keep top-k!

How do we represent data?

$$x^{(i)} \mapsto \sum_{j=1}^{k} \underbrace{(x^{(i)} \cdot u_j)}_{\text{we keep these } k \text{ scalars}} u_j$$

A map from $\mathbb{R}^d \to \mathbb{R}^k$

How do we choose k?

One approach "Amount of Explained Variance"

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \geq 0.9 \qquad \left(\text{aside } tr(A) = \sum_i A_{ii} = \sum_i \lambda_i\right)$$

$j=1$

NB: Only makes sense if $\lambda_j \geq 0$. Hence covariance is important

Lurking instability: Suppose $\lambda_k = \lambda_{k+1} \ldots$ what happens?

Rep is **unstable** here

Recap of PCA

- Dimensionality reduction technique (e.g. Visualization)
- Main idea is to project on a subspace, nice theory.
- cf. Factor Analysis