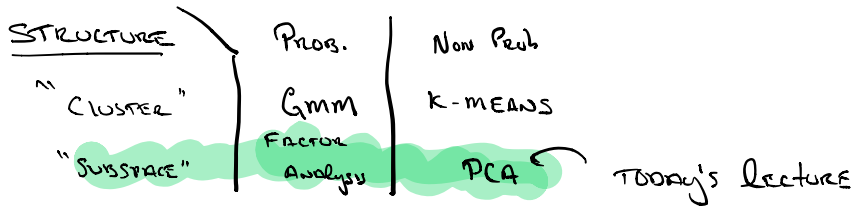
Factor Analysis & PCA Subspace structure.



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 need

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

$\det = 0$! undefined

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

Recall MLE for Gaussian is equivalent to

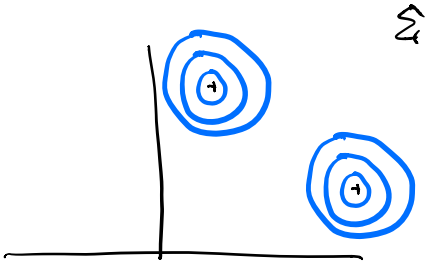
$$\min_{\mu,\Sigma_i} \sum_{i=1}^{n} (x-\mu)^T \Sigma_i^{-1} (x-\mu) + \log|\Sigma_i|$$

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

$\hat{\Sigma}_i = \sigma^2 I$ (NB: Parameter Tying)



Covariance "are circles"

What is MLE for Σ_i? $|\Sigma_i| = \sigma^{2d}$

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

Let $z = \sigma^2$ $\min_z \frac{1}{z} c + d \log z$

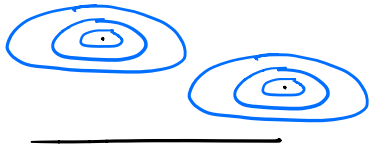
$$\Rightarrow \nabla_z = -z^{-2} c + \frac{nd}{z} = 0 \Rightarrow z = \frac{c}{nd}$$

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

"Subtract 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)

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

this is d Problems for EACH 1 dimension

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

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

Our 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

$P(x, z) = P(x|z) P(z)$ z IS LATENT

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

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

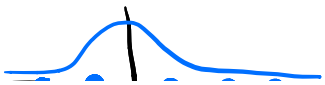
μ: MEAN IN the SPACE

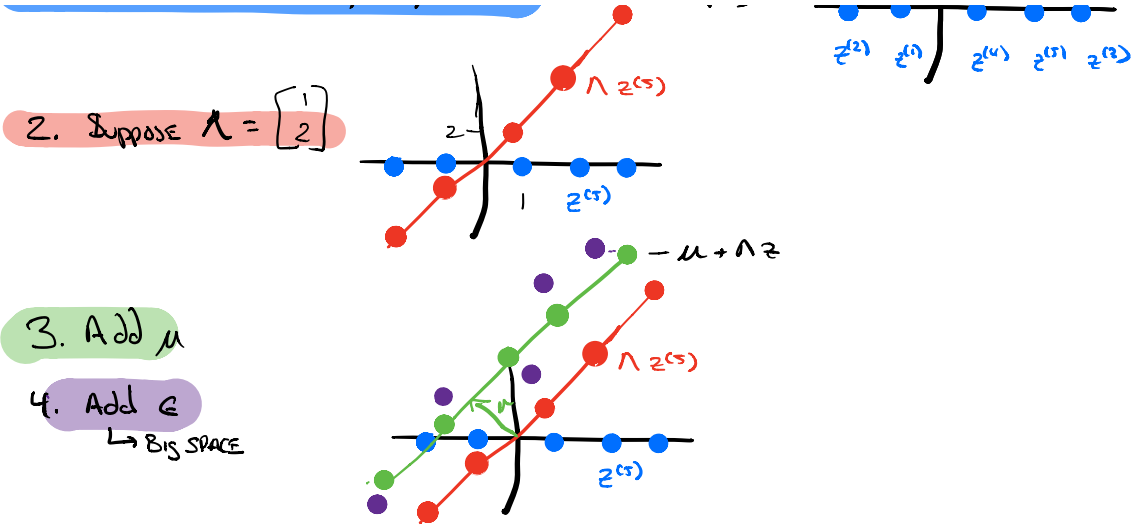
Λz: MAPS FROM SMALL LATENT SPACE TO LARGE SPACE

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

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

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





DATA WE WOULD OBSERVE ARE PURPLE DOTS

SO SMALL LATENT SPACE PRODUCES DATA IN high dim SPACE.

TECHNICAL TOOLS : Block GAUSSIANS

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

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

NOTATION IS WIDELY USED AND helpful.

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

FOR GAUSSIANS, $P(x_1) = N(\mu_{11}, \Sigma_{11})$ (NOT SURPRISING)

FACT 2: $P(x_1 | x_2) \sim N(\mu_{1|2}, \Sigma_{1|2})$ 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 online (happy to add)

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

We have formula for parameters.

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

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

What is Σ?

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

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

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

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

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

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

M-STEP : WE HAVE CLOSED FORMS!

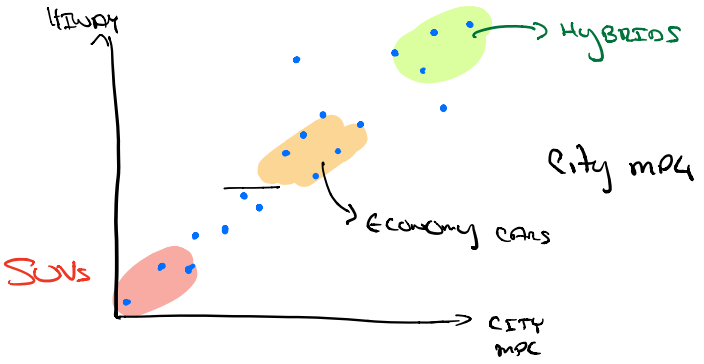
Summary of Factor Analysis

- 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

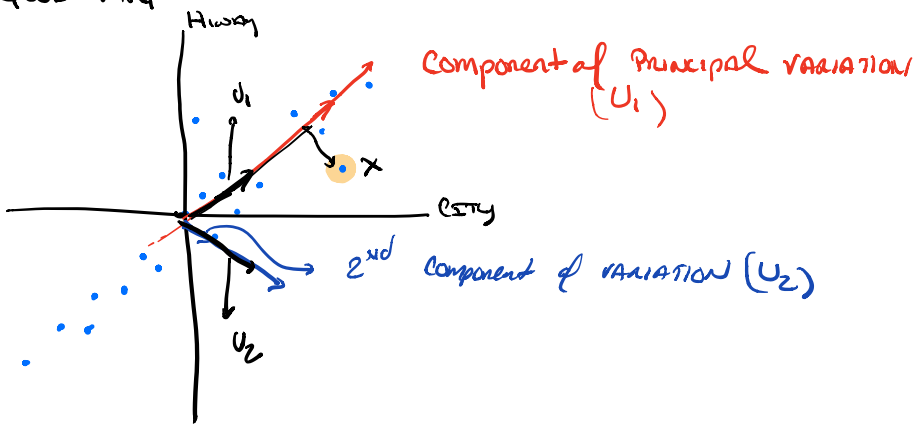


Question: "Good mpg"

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 → 10s of dims
- A dimensionality REDUCTION METHOD

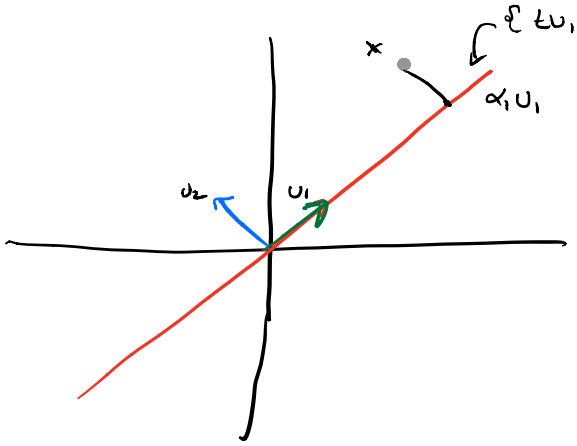
PREPROCESSING

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" ? MPG

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?

$$\alpha_1 = \arg\min_{\alpha} \|x - \alpha u_1\|^2$$

$$= \arg\min_{\alpha} \|x\|^2 + \alpha^2 \underbrace{\|u_1\|^2}_{1} - 2\alpha (u_1 \cdot x)$$

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 $U_i \cdot U_j = \delta_{ij}$

$$\underset{\alpha_1 \cdots \alpha_d}{\text{ARGMIN}} \left\| x - \sum_{i=1}^{k} \alpha_i U_i \right\|^2 = \underset{\alpha}{\text{argmin}} \|x\|^2 + \sum_{i=1}^{k} \alpha_i^2 \|u_i\|^2 - 2\alpha_i \langle u_i \cdot x \rangle$$

Hence $\alpha_i = U_i \cdot x$

WE CALL $\left\| x - \sum_{i=1}^{k} \alpha_i x_i \right\|^2$ THE RESIDUAL

WE CAN find PCA by either

In class ① MAXIMIZE Projected Subspace

② MINIMIZE Residual

$$\max_{\substack{U \in \mathbb{R}^d \\ \|U\| = 1}} \frac{1}{n} \sum_{i=1}^{n} (U \cdot x^{(i)})^2$$

WE NEED SOME facts to solve this

LET A BE symmetric & SQUARE, then

$$A = U \Lambda U^T \text{ IN which}$$

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

$\Lambda_{ii} = \lambda_i$ AND $\lambda_1 \geq \cdots \geq \lambda_n$ by convention

eigenvalues

Recall: If $x = \sum_{i=1}^{n} \alpha_i U_i$ where $[U_1 \cdots U_n] = U$

$$Ax = U \Lambda U^T x = U \Lambda \sum_{i=1}^{n} \alpha_i e_i \quad (U_i \cdot U_j = \delta_{ij})$$

(e_i: STANDARD BASIS VECTOR)

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

$$= \sum \lambda_i \alpha_i U_i$$

If $x = cU_i$ 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!

$$\max_{u:\|u\|^2=1,\ u\in\mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (u \cdot x^{(i)})^2$$

(the projection onto u)

$$= \frac{1}{n} \sum_{i=1}^{n} u^T x^{(i)} (x^{(i)})^T u = u^T \left(\frac{1}{n} \sum_i x^{(i)} (x^{(i)})^T \right) u$$

(covariance of data (we subtracted mean))

$\therefore$ u is principal eigenvector

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

How do we represent data?

$$x^{(i)} \mapsto \sum_{j=1}^{k} (x^{(i)} \cdot u_j) u_j$$

(we keep these k scalars)

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

How do we choose k?

One approach "Amount of Explained Variance"

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \geq 0.9$$

(aside $tr(A) = \sum_i A_{ii} = \sum_i \lambda_i$)

$j=1$

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 (eg. VISUALIZATION)
- MAIN IDEA IS TO project ON A SUBSPACE, NICE THEORY.
- of Factor Analysis