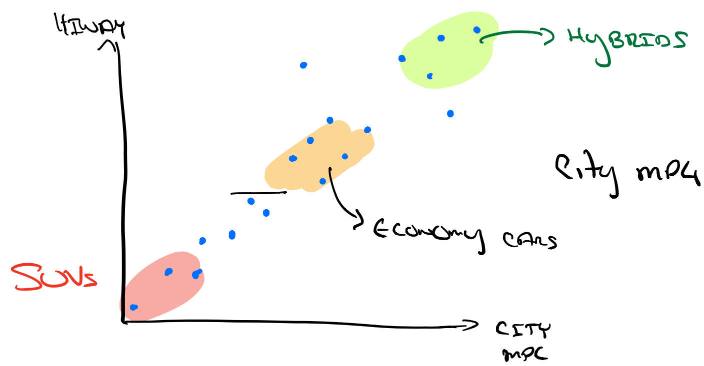
PCA: Principal Component Analysis

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

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

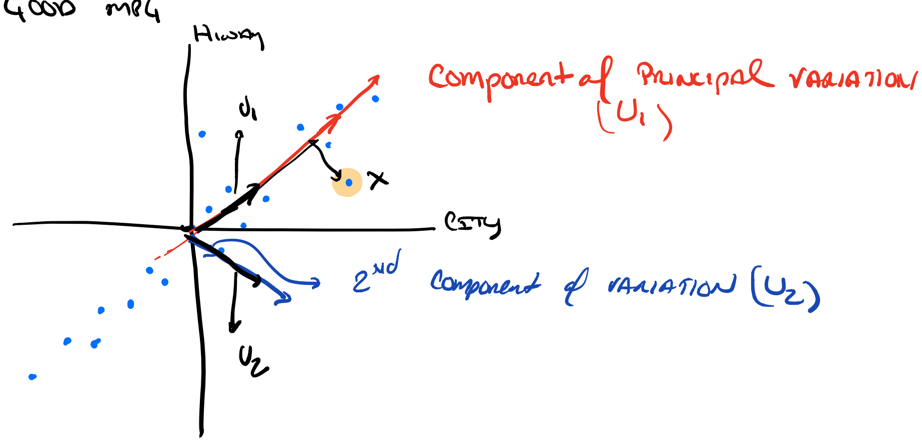


Question: "Good mpg"

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

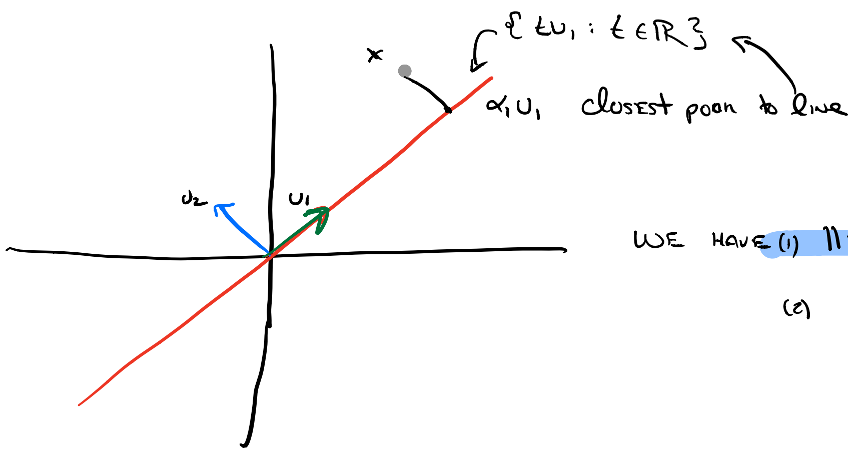
Preprocessing

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" & 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?

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

$$= \underset{\alpha}{\operatorname{argmin}} \|x\|^2 + \alpha^2 \cancelto{1}{\|u_1\|^2} - 2\alpha(u_1 \cdot x)$$

differentiate wrt α $\quad 2(\alpha - u_1 \cdot x) = 0 \implies \alpha = u_1 \cdot x$

Generalize: $u_1 \cdots u_k \in \mathbb{R}^d$ AND $x \in \mathbb{R}^d$

USE $u_i \cdot u_j = \delta_{ij}$

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

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

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

WE NEED some facts to solve this

LET A be symmetric & square, then

$A = U \Lambda U^T$ 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 $[u_1 \cdots u_n] = U$

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 $\alpha_1 = 1$, the principal eigenvalue

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

Now, back to PCA!

$$\max_{\substack{u : \|u\|^2 = 1 \\ u \in \mathbb{R}^d}} \frac{1}{n} \sum_{i=1}^{n} \underbrace{(u_i \cdot x^{(i)})^2}_{\text{the projection onto } u}$$

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

$\therefore$ 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"

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

$j=1$

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

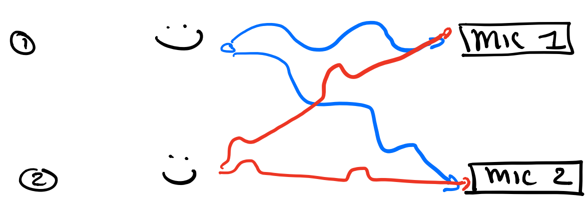
- Dimensionality reduction technique (e.g. visualization)
- Main idea is to project on a subspace, nice theory.

ICA Independent Component Analysis

- high-level story
- Key facts & likelihood
- model

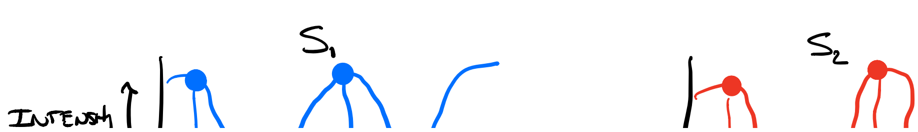
Cocktail Party Problem (in HW!)

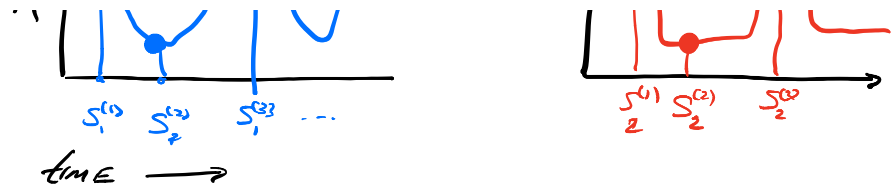
People — Microphones



NB: We see a mixture of & at each mic

Speaker S_1, S_2 — Data X





$s_j^{(t)}$ is <u>intensity</u> at <u>time t</u> from <u>speaker</u> j

We do <u>not</u> observe $s^{(t)}$ only $x^{(t)}$ – the microphones

<u>ex model</u> $$x_j^{(t)} = a_{j1} s_1^{(t)} + a_{j2} s_2^{(t)}$$

"microphone j sees a mixture of $s_1^{(t)}$ & $s_2^{(t)}$"

<u>or</u> $$\underbrace{x^{(t)}}_{\text{observed}} = \underbrace{A}_{\text{latent}}\ \overbrace{s^{(t)}}^{\text{latent}}$$

for simplicity, assume # of speakers = # of mics = d

<u>Given</u>: $x^{(1)}, x^{(2)}, \dots, x^{(n)} \in \mathbb{R}^d$ — d is # of microphones & speakers

<u>Do</u>: find $s^{(1)}, \dots, s^{(n)} \in \mathbb{R}^d$

and $A \in \mathbb{R}^{d \times d}$ s.t. $x^{(t)} = A s^{(t)}$

we call A the mixing matrix and $W = A^{-1}$ unmixing matrix

write $W = \begin{bmatrix} w_1^T \\ \vdots \\ w_d^T \end{bmatrix}$ so that $s_j^{(t)} = w_j \cdot x^{(t)}$

<u>Some Caveats</u>

- We assume A does not vary w/ time and is full rank

- THERE ARE INHERENT Ambiguity
  - WE CAN'T DETERMINE SPEAKER ID (could swap 1 & 2)
  - CAN'T DETERMINE Absolute INTENSITY

$$(cA)(c^{-1}s^{(t)}) = As^{(t)} \quad \text{for any } c \neq 0$$

- Surprising: Speakers cannot be Gaussian

Suppose so: $x^{(i)} \sim N(\mu, AA^T)$ then if $U^TU = I$, AU generates the same data.

Nevertheless, we can recover something meaningful!

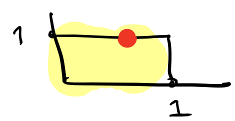
Algorithm: Just MLE, solved by Grad Descent

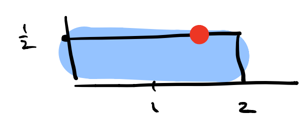
DETOUR: Density under linear transform (Key Confusion)

Ex: $S \sim \text{Uniform}[0,1]$ $\quad U = 2S$ $\quad$ what is PDF of U?

TEMPTED TO WRITE $P_U(\frac{x}{2}) = P_S(x)$

PDF of S





PDF of U

$$P_S(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{o.w} \end{cases} \qquad P_U(x) = P_S(\tfrac{x}{2}) \cdot \tfrac{1}{2}$$

THE key ISSUE IS the Normalization constant

For INVERTIBLE MATRIX A, $U = AS$

$$P_U(x) = P_S(A^{-1}x)\,|\det(A^{-1})|$$

$$= P_S(Wx)\,|\det(W)| \qquad \left(\frac{1}{\det(A)} = \det(A^{-1})\right)$$

CHANGE of VAR formula for INTEGRALS

From here ICA is MLE:

$$P(s) = \prod_{j=1}^{d} P_s(s_j)$$ "Sources are independent, and have same distribution"

$$P(x) = \prod_{j=1}^{d} P_s(W \cdot x) \cdot |\det(W)|$$ (use linear transform rule)

Now written in terms of x and A.

Key technical bit: Use non-rotationally invariant distribution

Set $P_s(x) \propto g'(x)$ for $g(x) = (1+e^{-x})^{-1}$

Solve $$\ell(W) = \sum_{i=1}^{n} \sum_{j=1}^{d} \log g'(w_j \cdot x^{(i)}) + \log|\det(W)|$$

- $\log|\det(W)|$
- Use GD & you're done!

Recap: · Saw PCA. Workhorse dimensionality reduction
· ICA. Key ideas for HW. Introduce "up to symmetry."