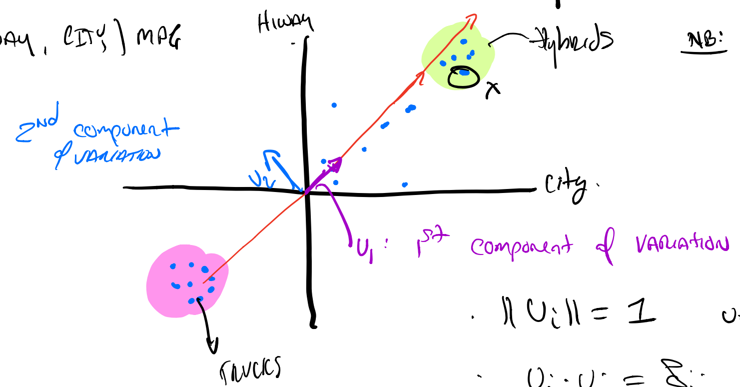
PCA & ICA

PCA: Principal Component Analysis

Given (Hiway, City) MPG



NB: we centered data

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

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

· $\|u_i\| = 1$ unit length

· $u_i \cdot u_j = \delta_{ij}$ (orthonormal)

· u_1 – "How good is the mpg"

· u_2 – "variation in city/hiway from "good""

$x = \alpha_1 u_1 + \alpha_2 u_2$

$x^{(i)} = \alpha_1^{(i)} u_1 + \alpha_2^{(i)} u_2$

Today: How we find directions

Think about dimension 1000s → 10

Preprocessed

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

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

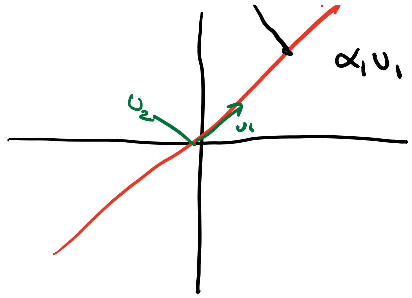
   "Whiten" compute sample variance ..

   $$x_j^{(i)} \mapsto \frac{(x^{(i)} - \mu)_j}{\sigma_j}$$

We will assume data is preprocessed.

PCA as Optimization

u_1 $\{t u_1 : t \in \mathbb{R}\}$ line corresponds to u_1



How do you find closest point to the line generated by u_1?

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

$$= \underset{\alpha}{\operatorname{argmin}} \underbrace{\|x\|^2 + \alpha^2 \|u_1\|^2 - 2\alpha(u_1 \cdot x)}_{g(\alpha)}$$

1

$$\nabla_\alpha g(\alpha) = 2\alpha - 2(u_1 \cdot x) = 0 \Rightarrow \alpha = u_1 \cdot x$$

<u>Generalize:</u> $u_1 \dots u_k \in \mathbb{R}^d$ and $x \in \mathbb{R}^d$ $\quad (u_i \cdot u_j = \delta_{i,j})$

$$\underset{\alpha_1 \dots \alpha_d}{\operatorname{Argmin}} \|x - \sum_{j=1}^{k} \alpha_j u_j\|^2 \dots$$

Hence $\alpha_j = u_j \cdot x$

$\|x - \sum_{j=1}^{k} \alpha_j u_j\|^2 \leftarrow$ RESIDUAL

We can find PCA by either:

1. Maximize Projected subspace
2. Minimize Residuals

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

Solve this optimization problem we need some facts.

Let A be symmetric & square then

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

- $UU^T = U^T U = I$ (ORTHONORMAL BASIS)
- Λ IS DIAGONAL MATRIX

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

Recall: If $x = \sum_{i=1}^{d} \alpha_i u_i$ where $[u_1 \dots u_d] = U$

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

$$= U \sum_{i=1}^{d} \alpha_i \lambda_i e_i \qquad \text{if } i=j \Rightarrow \delta_{ij} = 1$$

$$= \sum_{i=1}^{d} \alpha_i \lambda_i u_i \qquad \text{else } 0$$

fix i, and let $c \in \mathbb{R} \neq 0$

$x = c u_i$ eigenvectors $Ax = \lambda_i x$

$$\max_{\substack{u \in \mathbb{R}^d \\ \|u\| = 1}} \underline{\frac{1}{n} \sum_{i=1}^{n} (x^{(i)} \cdot u)^2} \iff \max_{\substack{u \in \mathbb{R}^d \\ \|u\| = 1}} u^T A u \qquad A = \frac{1}{n} \sum_{i=1}^{n} x^{(i)} x^{(i)T}$$

$$\iff \max_{\substack{\alpha \in \mathbb{R}^d \\ \|\alpha\|^2 = 1}} \sum_{i=1}^{d} \alpha_i^2 \lambda_i$$

WHAT SHOULD WE PICK TO MAXIMIZE? $\alpha_1 = 1$ $\alpha_2 = \alpha_3 = \dots = \alpha_d = 0$

IS IT UNIQUE? $\lambda_1 = \lambda_2$ WHAT HAPPENS? (PCA "lucky INSTABILITY")

$\lambda_1 > \lambda_2 \Rightarrow$ UNIQUE

u_1 IS the PRINCIPAL eigenvector

WHAT IF WE WANT THE TOP-K such vectors?

$u_1 \dots u_k$ BECAUSE $\lambda_1 \geq \dots \geq \lambda_k$

$$x^{(i)} \mapsto \sum_{i}^{k} (x^{(i)} \cdot u_i) u_i$$

$j=1$

→ keep these numbers.

How do we pick k?

One approach "Amount of Explained Variance"

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{d} \lambda_j} \geq 0.9 \qquad \lambda_i \geq 0$$

Lurking instability $\lambda_{k-1} = \lambda_k = \lambda_{k+1}$ Are you were looking for top k?

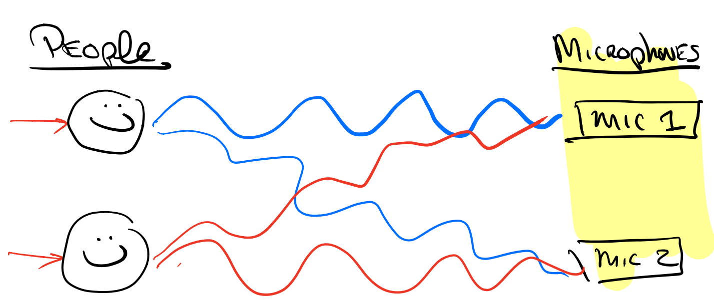
Pick any 2 of three! $\Rightarrow$ different bases

PCA · Dimensionality reduction technique

- Main idea is project on subspace
- Nice theory! Contrast w/ EM

ICA Independent Component Analysis

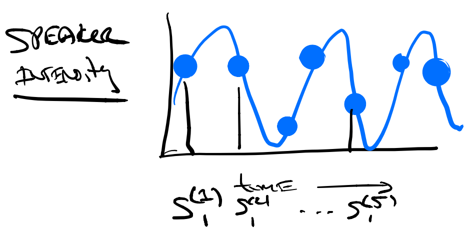
- High-level story
- Key facts and likelihood
- model

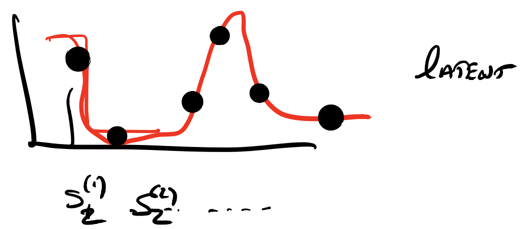


We see a mixture of and

Speaker separation.

Speaker $s_1^{(t)}, s_2^{(t)}$ Data $x_1^{(t)}, x_2^{(t)}$





$s_j^{(t)}$ is the intensity at time t of speaker j

We do **not** observe $s_j^{(t)}$ only $x_j^{(t)}$ — the microphones

model $x_j^{(t)} = a_{j1} s_1^{(t)} + a_{j2} s_2^{(t)}$ ← Hidden

"microphone j sees a mixture of the speaker intensities"

observed ← $x^{(t)} = A s^{(t)}$ → latent

for simplicity, d is the number of speakers & microphones

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

Do: 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 $s_j^{(t)} = w_j \cdot x^{(t)}$

- We center the data $x^{(i)} \mapsto x^{(i)} - \mu$ $\quad \mu = \frac{1}{n} \sum x^{(i)}$
- A does not vary w/ time, A is full rank
- There are some inherent ambiguity:
  - We can't determine speaker id
  - Can't determine intensity (absolute)

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

· <u>Surprising</u> Speakers cannot <u>be Gaussian</u>

$$x^{(i)} = As^{(i)} \qquad s^{(i)} \sim N(0, I)$$
$$\Rightarrow x^{(i)} \sim N(0, AA^T) \qquad UU^T = I$$
$$AUU^TA^T = AA^T$$

Nevertheless, we can recover something meaningful

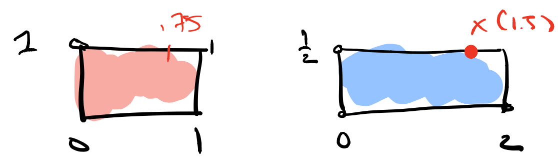
<u>Algorithm</u>: Just MLE w/ Grad Descent.

<u>Detour</u>: Density under linear transform

<u>Ex</u>: $S \sim$ Uniform $[0,1]$ $U = 2S$ what's the pdf of U?

tempted $P_U(x) = P_S(\frac{x}{2})$ (not right)

PDF of S



Pdf of U

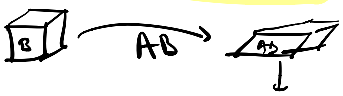
$$P_S(x) \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{o.w.} \end{cases} \qquad P_U(x) = P_S(\tfrac{x}{2}) \cdot \tfrac{1}{2}$$

The key issue is the normalization constant

A square & invertible, $U = As$ $s \sim$ PDF of P_S

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

Change of variables formula



volume vol(B) Vol(AB) = |det(A)| vol(B)

FROM HERE ICA IS MLE!

LATENT $\longrightarrow$ $P(s) = \prod_{j=1}^{d} P_s(s_j)$ "SOURCES INDEPENDENT AND HAVE DISTRIBUTION"

OBSERVED $\longrightarrow$ $P(x) = \prod_{j=1}^{d} P_s(w_j \cdot x)\ |\det(W)|$

Key technical trick NOT ROTATIONALLY SYMMETRIC

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

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

- log det ok
- USE GD & you're done

Recap:
- SAW PCA. WORKHORSE dimensionality REDUCTION
- ICA - Key IDEAS. HANDLING SYMMETRY.