Factor Analysis & PCA

Continue unsupervised learning

+ Factor Analysis (many more dimensions than points $d \gg n$) EM
+ PCA (old strategy) Linear Algebra

→ up next ICA "cocktail party" (end of classical unsupervised)

—

Factor Analysis

Many fewer points than dimensions "$n \ll d$"

cf: GMMs $n \gg d$ lots of datapoints, few sources.

How does this happen?

Place sensors all over campus, record @ 1000s of locations

$\Rightarrow d \sim 1000s$

But only record for 30 days ($n < d$)

Want to fit a density, but seems hopeless.

Key idea: Assume there is some latent r.v. that is not too complex and explains behavior.

1st let's see problems w/ GMMs.... Even 1 Gaussian

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} \rightarrow \text{this is OK}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (x^{(i)}-\mu)(x^{(i)}-\mu)^T$$

$\text{Rank}(\hat{\Sigma}) \leq n < d$ — not full rank.

Problem in Gaussian likelihood (is not defined.)

$$P(x;\mu,\Sigma) = \frac{1}{2\pi |\Sigma|^{1/2}} \exp\left\{(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

$\hookrightarrow |\Sigma| = 0$

We will fix these issues by examining three models that are simpler. Spoiler: we'll combine these in the end!

Recall MLE for Gaussian

$$\max_{\mu,\Sigma} \sum_{i=1}^{n} \log \frac{1}{2\pi |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

Equivalent

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

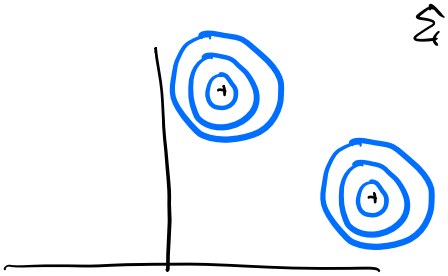
If Σ is full rank, $\nabla_\mu = \sum_{i=1}^{n} \Sigma^{-1}(x-\mu) = 0 \Rightarrow \mu = \frac{1}{n}\sum_i x^{(i)}$

we'll use this as plugin below.

Building Block 1

Suppose independent and identical covariance

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



Covariance "are circles"

What is MLE for Σ? $|\Sigma| = \sigma^{2d}$

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

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

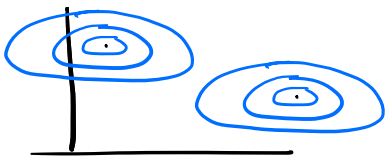
$$\Rightarrow \nabla_z = -z^{-2} C + \frac{nd}{z} = 0 \Rightarrow z = \frac{C}{nd}$$

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

"Substract mean and square all entries."

Building Block 2

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



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

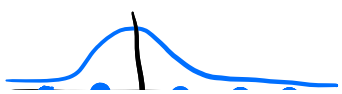
μ: MEAN IN the SPACE

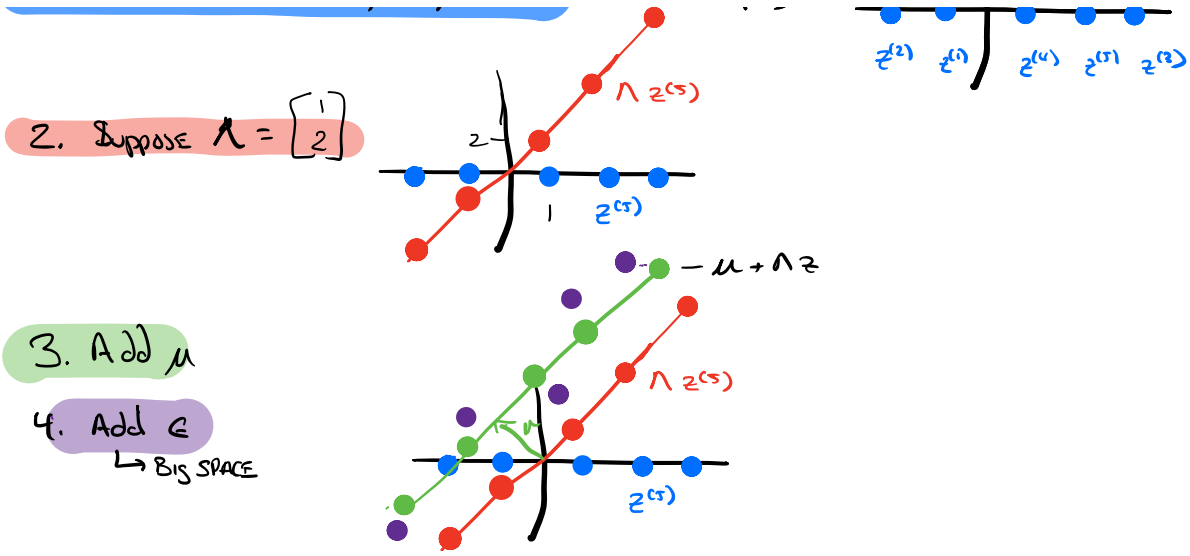
Λz: MAPS FROM SMALL LATENT SPACE TO LARGE SPACE

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

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

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

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 outline (happy to add)

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

We have formula for parameters.

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

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

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!

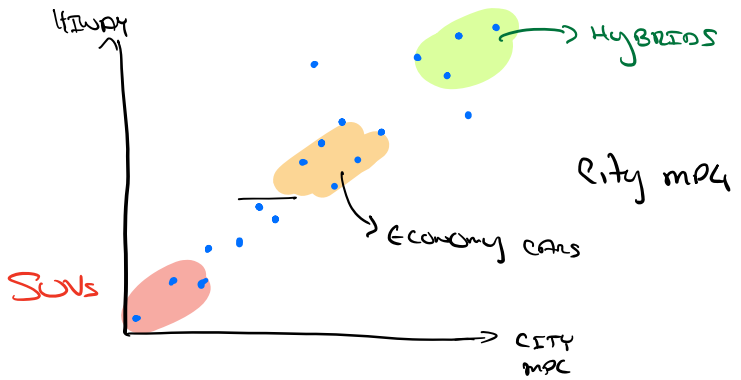
Summary:

- WE SAW *that* EM CAPTURES GMM
- 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 ← this section

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

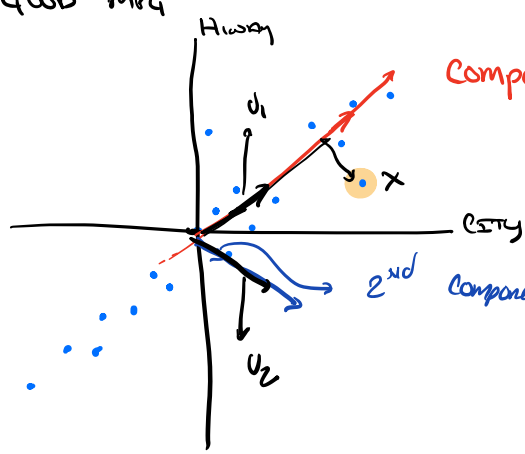


Question: "Good mpg"

1. Center data

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

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



Component of Principal variation (U_1)

2^{nd} Component of variation (U_2)

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

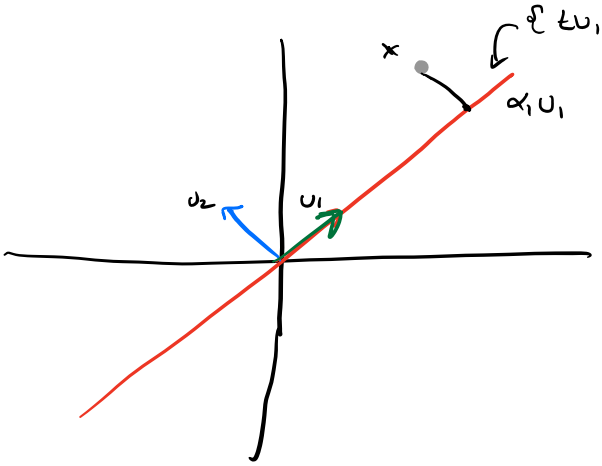
PREPROCESSING

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) $\|u_j\| = 1$ (unit vectors)

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

How do you find closest point to the line?

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

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

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

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

Hence $\alpha_i = u_i \cdot x$

WE CALL $\left\| x - \sum_{i=1}^{k} \alpha_i u_i \right\|^2$ 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-2!

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 $\text{tr}(A) = \sum_i A_{ii} = \sum_j \lambda_j$)

$j=1$

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

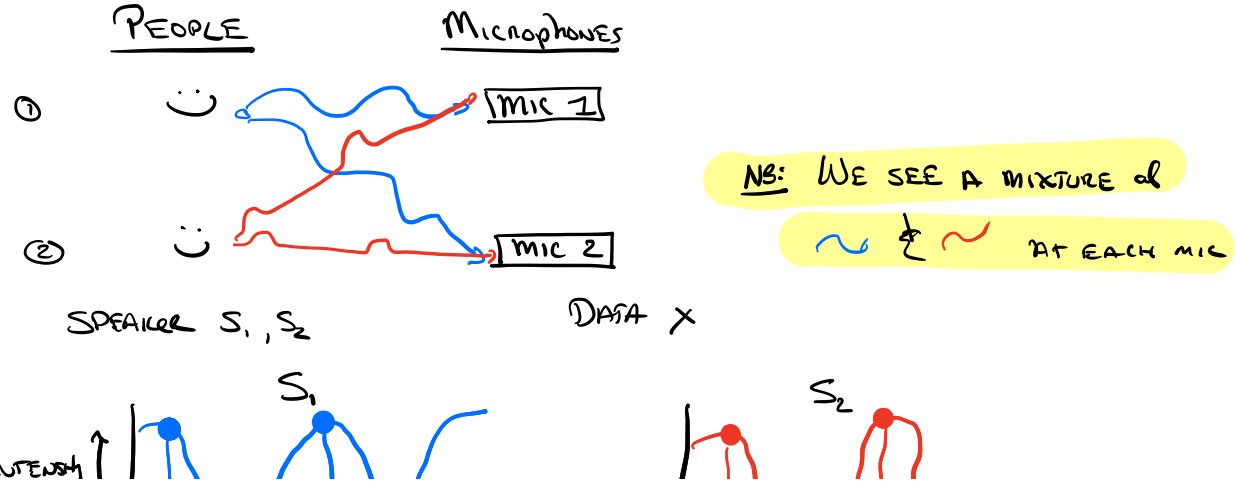
RECAP of PCA

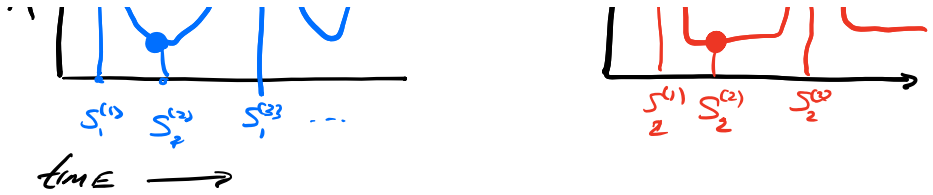
- DIMENSIONALITY REDUCTION TECHNIQUE (eg. VISUALIZATION)
- MAIN IDEA IS TO PROJECT ON A SUBSPACE, NICE THEORY.

ICA INDEPENDENT COMPONENT ANALYSIS

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

Cocktail Party Problem (IN HW!)





$s_j^{(t)}$ IS INTENSITY AT TIME t from SPEAKER j

WE DO NOT OBSERVE $s^{(t)}$ only $x^{(t)}$ – the microphones

ex model $\quad x_j^{(t)} = a_{j1} s_1^{(t)} + a_{j2} s_2^{(t)}$

"MICROPHONE j SEES A MIXTURE of $s_1^{(t)}$ & $s_2^{(t)}$"

OR $\quad x^{(t)} = A s^{(t)}$

$x^{(t)}$: OBSERVED; A: LATENT; $s^{(t)}$: LATENT

for simplicity, ASSUME # of SPEAKERS = # of MICS = d

GIVEN: $x^{(1)}, x^{(2)}, \dots, x^{(n)} \in \mathbb{R}^d$ $\quad$ d is # of microphones & speakers

DO: find $s^{(1)}, \dots, s^{(n)} \in \mathbb{R}^d$

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

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

SOME CAVEATS

- 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 SD $x^{(t)} \sim N(\mu, AA^T)$ then if $U^TU = I$ AN generates the SAME DATA.

NEVERTHELESS, WE CAN RECOVER SOMETHING MEANINGFUL!

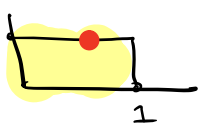
Algorithm: Just MLE, solved by GRAD DESCENT

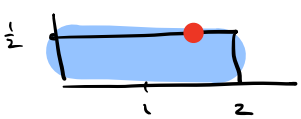
DETOUR: Density under linear transform (Key Confusion)

Ex: $S \sim \text{Uniform}[0,1]$ $U = 2S$ 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, $U = AS$

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

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

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 flow)

Now written in terms of x and A.

Key technical bit: Use non-rotationally invariant distribution

Set $P_S(x)$ or $g'(x)$ 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."