

ICA & WEAK Supervision (Brief)

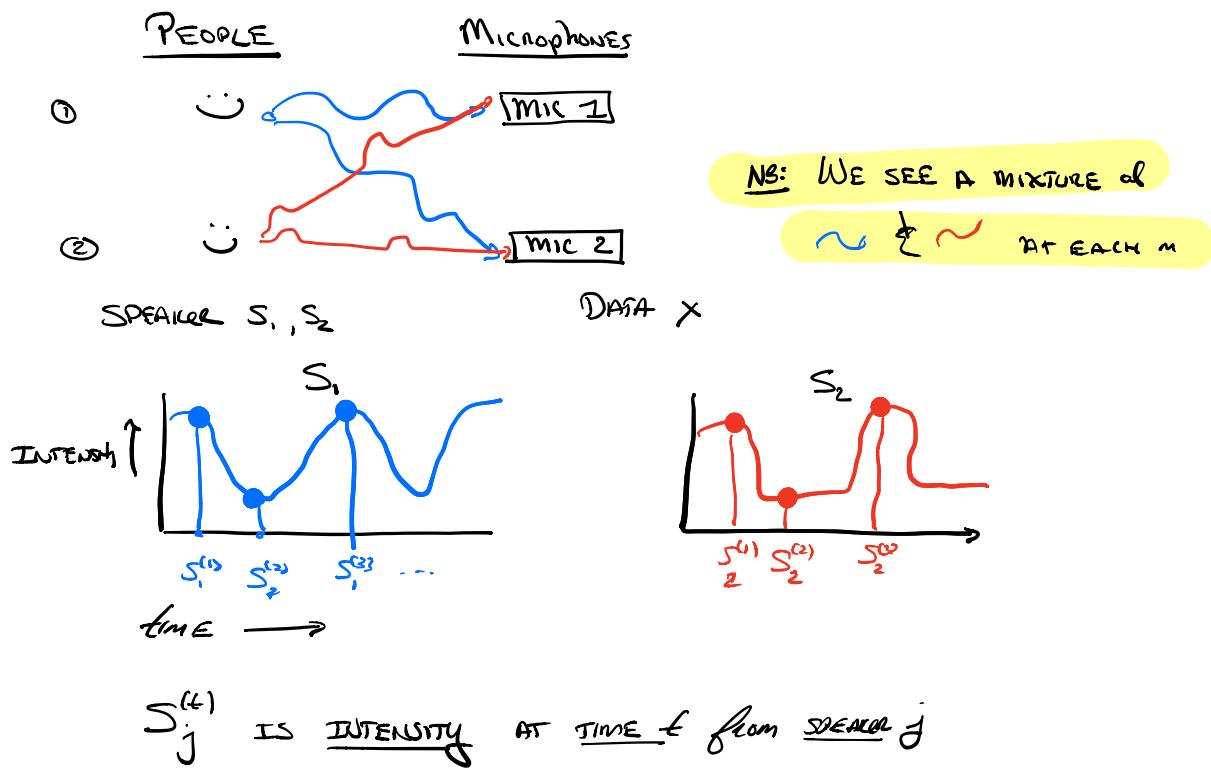
- + ICA: Source Separation (fun problem)
- + Weak Supervision

We're past midterm, lots more material than you need. More fun → happy to chat.

ICA INDEPENDENT Component Analysis

- high-level story
 - Key facts \neq likelihood
 - model

Cocktail Party Problem (in hw!)



WE DO NOT OBSERVE $S^{(x)}$ ONLY $x^{(x)}$ - THE MICROPHONES

$$\text{ex model} \quad X_j^{(t)} = \alpha_{j1} S_1^{(t)} + \alpha_{j2} S_2^{(t)}$$

"Microphone y sees a mixture of $S_1^{(+)}$ & $S_2^{(+)}$ "

→ LATENT

ON

$$X^{(t)} = AS^{(t)}$$

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

GIVEN: $X^{(1)}, X^{(2)}, \dots, X^{(n)} \in \mathbb{R}^d$ 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. $X^{(t)} = AS^{(t)}$

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 \equiv (cold swap $i \leftrightarrow j$)
 - CAN'T DETERMINE ABSOLUTE INTENSITY
 $(cA)(c^{-1}s^{(c)}) = As^{(1)}$ for any $c \neq 0$

Surprise: Speakers cannot be Gaussian

Suppose $x^{(t)} \sim N(\mu, A^T A)$ then if $U^T U = I$ AU generates the SAME data.

Nevertheless, we can recover something meaningful!

Algorithm: Just MLE, solved by grad descent

DETONE: Density under linear transform (Key Confusion)

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

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



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

THE key ISSUE is the Normalization constant

for INVERTIBLE MATRIX A , $U = As$

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

CHANGE OF VAR
formula for
integrals

$$= P_S(wx) | \det(w)| \quad \left(\frac{1}{\det(A)} = \det(A^{-1}) \right)$$

From here ICA is MLE:

$$P(s) = \prod_{j=1}^d P_s(s_j)$$

"sources are independent,"

$$P(x) = \prod_{j=1}^d P_s(w \cdot x) \cdot |\det(w)|$$

(use linear transform rule)

Now written in terms of w and A .

Key technical bit: Use non-rotationaly invariant distribution

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

$$\text{Solve } l(w) = \sum_{\ell=1}^n \sum_{j=1}^d \log g'(\omega_j \cdot x^{(\ell)}) + \log |\det(w)|$$

- $\log |\det(w)|$
- USE g' & you're done!

RECAP:

- Saw PCA. Workhorse dimensionality reduction
- ICA. Key ideas for now. Introduce "upto symmetry".

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

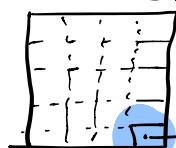
SKIP IN LECTURE

fix rectangle $U \subseteq \mathbb{R}^d$ let $AU = \{y : Ax = y, x \in U\}$

HAPPY TO RECORD

Simplified LINEAR CHANGE OF VARIABLES FOR INTEGRALS

Recall: $\int_U f(x) dx$.



x_B is a point
 B is the box

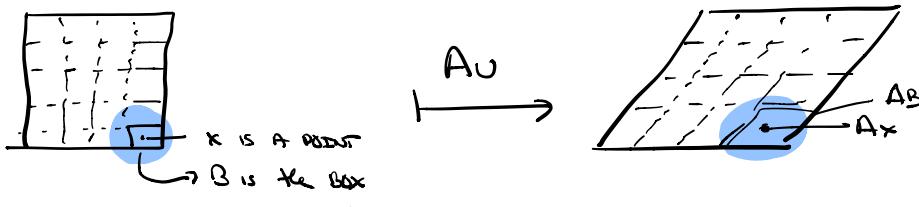
$$\sum_{B \in P} f(x_B) \text{vol}(B)$$

LET P BE A PARTITION OF U . LET x_B BE A POINT INSIDE OF B OR B

$$\int_U f(x) dx \approx \sum_{B \in P} f(x_B) \text{vol}(B) \quad \text{vol}(B) \text{ is the volume.}$$

Recall as Partition gets finer, ($\max_{B \in P} \text{vol}(B) \rightarrow 0$)
then **Apx** gets closer to integral.

Now, let's examine $\int_{AU} f(A^{-1}y) dy$. fix a partition of U



every partition of U makes a partition of AU (A is full rank)

$$\int_{AU} f(A^{-1}y) dy \approx \sum_{B \in P} f(A^{-1}Ax_B) \text{vol}(AB)$$

Key fact $\text{vol}(AB) = |\det(A)| \cdot \text{vol}(B)$

$$= \sum_{B \in P} f(x_B) \cdot \text{vol}(B) |\det(A)| = |\det(A)| \int_U f(x) dx$$

So, we have shown $\int_{AU} f(A^{-1}y) dy = |\det(A)| \int_U f(x) dx$

Recall we wanted to figure out normalization for $P_S(A^{-1}x)$

in terms of $P_U(x)$, i.e. $P_S(A^{-1}x) = c P_U(x)$

the above says, $c = |\det(A)|^{-1} = |\det(\omega)|$ as claimed
(take $f(x) = P_S(x)$)

WEAK Supervision Nuggets (to slides for review)

- INDEP CASE \rightarrow Simple Estimation trick
- Correlations \rightarrow INVERSE Covariance & Graph Structure.

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

$\lambda_1, \dots, \lambda_m : \lambda_i : \mathbb{R}^d \rightarrow \{-1, 1\} \cup \{\text{ABSTAIN}\}$

find $P(y | \lambda, x)$ $y \in \{-1, 1\}$

IDEA: λ_i is a noisy function (Inaccurate, incomplete)

λ_1 : "name in dictionary"

λ_2 : "upper case word"

... Programmatic labels ... NO Labels

Model D: No ABSTAINS INDEPENDENT Errors Class Correlated

<u>GIVEN</u>		<u>UNOBSERVED</u>
Data	$\lambda_1 \lambda_2 \lambda_3 \dots \lambda_m$	
$x^{(1)}$	1 1 1 1	?
$x^{(2)}$	1 -1 1 -1	-1
:		
$x^{(n)}$	-1 -1 -1 -1	-1

$P(y | \lambda, x)$

Each table has a hidden accuracy. $\rightarrow x \text{ AND } y$ (example)

with prob p_j

$\lambda_j(x)$	y	" λ_j is right"
$1-p_j$	$\lambda_j(x) = -y$	" λ_j is wrong"

so, we don't see y , but we do see $\lambda_j(x)$

λ_i, λ_j 's exams INDEPENDENT AND Symmetric in the following sense

$$P(\lambda_j(x)=1 | y=1) = P(\lambda_j(x)=-1 | y=-1) = p_j$$

① $E[\lambda_i \cdot y]$ observe if $\lambda_i \wedge y$ Agree value is 1
 if $\lambda_i \wedge y$ disagree value is -1

$$\begin{aligned} \therefore &= p_i \cdot 1 + (1-p_i) \cdot (-1) \\ \therefore &= 2p_i - 1 \stackrel{\Delta}{=} q_i \quad (\text{define } q_i \text{ this way}) \quad q_i \in [-1, 1] \end{aligned}$$

② $E[\lambda_i \lambda_j] = 1 \text{ if } i=j$

$$\begin{aligned} E[\lambda_i \lambda_j | y=1] &= p_i p_j \cdot 1 + (1-p_i)(1-p_j) \text{ "Agree"} \\ &\quad + (1-p_i)p_j(-1) + p_i(1-p_j)(-1) \text{ "disagree"} \\ &= a_i a_j \end{aligned}$$

Note we don't use $|y=1$, same true for $|y=-1$

$$E[\lambda_i \lambda_j] = \sum_{b \in \{-1, 1\}} E[\lambda_i \lambda_j | y=b] P(y=b) = a_i a_j \sum_b p(y=b)$$

didn't need to know $P(y=b)$.

We form a matrix $M \in \mathbb{R}^{n \times m}$ $M_{ij} = E[\lambda_i \lambda_j]$

NB: M CAN BE ESTIMATED - unlike y !

"Agreements AND Disagreements" \Rightarrow key idea don't need to see y

Simple Algorithm: for any i, j, k distinct, m_{ij}, m_{jk}, m_{ik} s.t

$$\frac{m_{ij} m_{jk}}{m_{ik}} = \frac{a_i a_j^2 a_k}{a_j a_k} = a_j^2$$

So we can solve upto sign of a_i .

Note: If we knew $\text{sign}(a_i) = s$

$$\text{then } m_{ik} = a_i a_k$$

$$\text{Sign}(m_{ik}) \text{Sign}(a_i) = \text{Sign}(a_k) \quad \therefore \text{can solve for all signs. Random}$$

so $a_i^{\frac{1}{2}} - a$ are solutions...

Assume $\sum_{i=1}^n a_i > 0$, breaks symmetry "good on average".

- What if $m_{ij} = 0$? $a_i = 0$ or $a_j = 0$.

This means $a_i = 0 \Rightarrow 2a_i - 1 = 0 \Rightarrow p_i = \frac{1}{2} \Rightarrow \text{random noise!}$

have every a_i bounded away from random noise

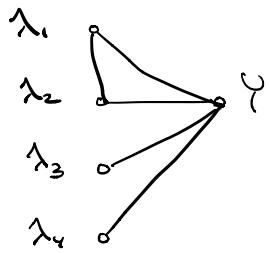
(can handle w/ fancier tricks).

Theory says "let $R = \min_{j=1..n} |p_j - \frac{1}{2}|$, NEED samples proportional
to $\frac{1}{R^2}$ " (plot of curve)

Recap: Symmetry, Simple Algebraic Learns (now called w/ Q)

e.

WHAT if variables are correlated?

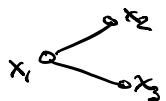


Key Concept: Probability distribution on Graphs.

$$\mathbb{E}[\lambda_i \lambda_j | \gamma] = \mathbb{E}[\lambda_i | \gamma] \mathbb{E}[\lambda_j | \gamma]$$

if (i, j) / Edge Above.

Nugget Structure of INVERSE COVARIANCE for Gaussians



$$x_1 \sim N(0, 1)$$

$$x_2 \sim N(x_1, 1) \quad \left\{ \begin{array}{l} \epsilon_2 \sim N(0, 1) \\ x_2 = x_1 + \epsilon_2 \end{array} \right.$$

$$x_3 \sim N(x_2, 1) \quad \left\{ \begin{array}{l} \epsilon_3 \sim N(0, 1) \\ x_3 = x_2 + \epsilon_3 \end{array} \right.$$

$$1. \quad \mathbb{E}[x_1] = 0 \quad \mathbb{E}[x_2] = \cancel{\mathbb{E}[x_1]}^0 + \cancel{\mathbb{E}[\epsilon_2]}^0 = 0$$

$$\mathbb{E}[x_3] = 0.$$

$$\mathbb{E}[x_1^2] = 1 \quad \mathbb{E}[x_2^2] = \mathbb{E}[(x_1 + \epsilon_2)^2] = \mathbb{E}[x_1^2] + 2\mathbb{E}[x_1 \cancel{\epsilon_2}] + \mathbb{E}[\epsilon_2^2]$$

$$= 2$$

$$\mathbb{E}[x_1 x_2] = \mathbb{E}[x_1^2 x_1 \epsilon_2] = 1$$

$$\mathbb{E}[x_2 x_3] = \mathbb{E}[(x_1 + \epsilon_2)(x_2 + \epsilon_3)] = 1$$

$$\hat{\Sigma} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

NO CLEAR STRUCTURE?

$$\hat{\Sigma}^{-1} = \begin{bmatrix} 3 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

NO EDGE

WE SAY A Probability distribution $p: \mathbb{R}^d \rightarrow [0, 1]$

agrees or factorizes w.r.t. a graph $G = (V, E)$ if

$$p(x) = \frac{1}{C} \prod_{e=(v_i, v_j) \in E} p_e(x_i, x_j) \cdot \prod_{v \in V} p_v(x_i)$$

for some functions!

Normalizing constant

Now, let's look @ Gaussians over a graph.

$$\log \exp \left\{ x^\top \hat{\Sigma}^{-1} x \right\} = \log \prod_{e \in E} p_e(x_i, x_j) \prod_{v \in V} p_v(x_i) \quad (\text{assume } c)$$

$$x^\top \hat{\Sigma}^{-1} x = \sum_{e \in E} \log p_e(x_i, x_j) + \sum_{v \in V} \log p_v(x_i)$$

let $A = \hat{\Sigma}^{-1}$ for easy notation

$$\sum_{i,j} A_{ij} x_i x_j = \underline{\hspace{10em}}$$

$$\text{for } i, j \in E \quad \nabla_{x_i} x_j = (A_{ij} + A_{ji}) = 0 \quad \text{if } (i, j) \notin E.$$

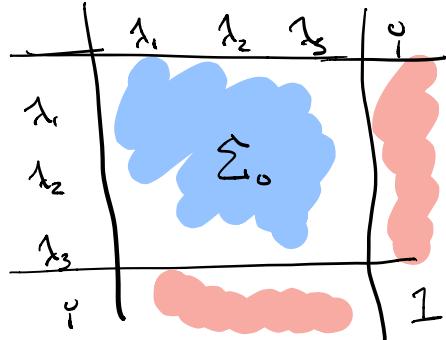
But $\hat{\Sigma}^{-1} = A$ is symmetric, so $A_{ij} = 0$.

thus, if A Gaussian factors on a graph, Entries of inverse are 0!

More complex theory for discrete R.V.s for δ Wainwright 2014
 RAJNEE et al 2018

Back to our problem

Form "covariance matrix"



WE "SEE" Σ_0 BUT NOT
 ALL OF SIGMA $\Sigma[\lambda; \gamma]$
 IS UNOBSERVED.

Let $\Theta = \{1, 2, 3\}$ \sqsubset visible teams

$$(\Sigma^{-1})_0 = (\Sigma_0 - UU^\top)^{-1} \quad \text{SOME RANK 1 VECTOR.}$$

INVERSE ON THOSE ENTRIES

let $B = \Sigma_0$

$$(B - UU^\top)^{-1} = B^{-1} + \frac{B^{-1} U U^\top B^{-1}}{1 - U^\top B^{-1} U}$$

$$Z = \frac{B^{-1} U}{\sqrt{1 - U^\top B^{-1} U}}$$

$$\text{so } (\Sigma^{-1})_0 = \Sigma_0^{-1} + Z Z^\top$$

Now, if $(i, j) \notin E$ then we $(\Sigma^{-1})_{ij} = B_{ij} = 0$.

$$\text{hence } (\Sigma_0^{-1})_{ij} = -z_i z_j$$

$$(B_{ij})^2 = z_i^2 z_j^2 \mapsto \log B_{ij}^2 = \log z_i^2 + \log z_j^2$$

This is a linear system in $z_i^2 + z_j^2$, AND WE CAN
 solve (if enough pairwise info)

IN THE WORKS:

- Higher rank versions (Handle more correlations)
- How to learn graph Structure
- How to handle sampling error (Model m , lower bands)

RECAP:

- WEAK Supervision formal theory
- Widgets about Graphs & Prob. distributions (take graphical models!)
- "Method of moments" style ALGORITHMS