ICA & Weak Supervision (Brief)

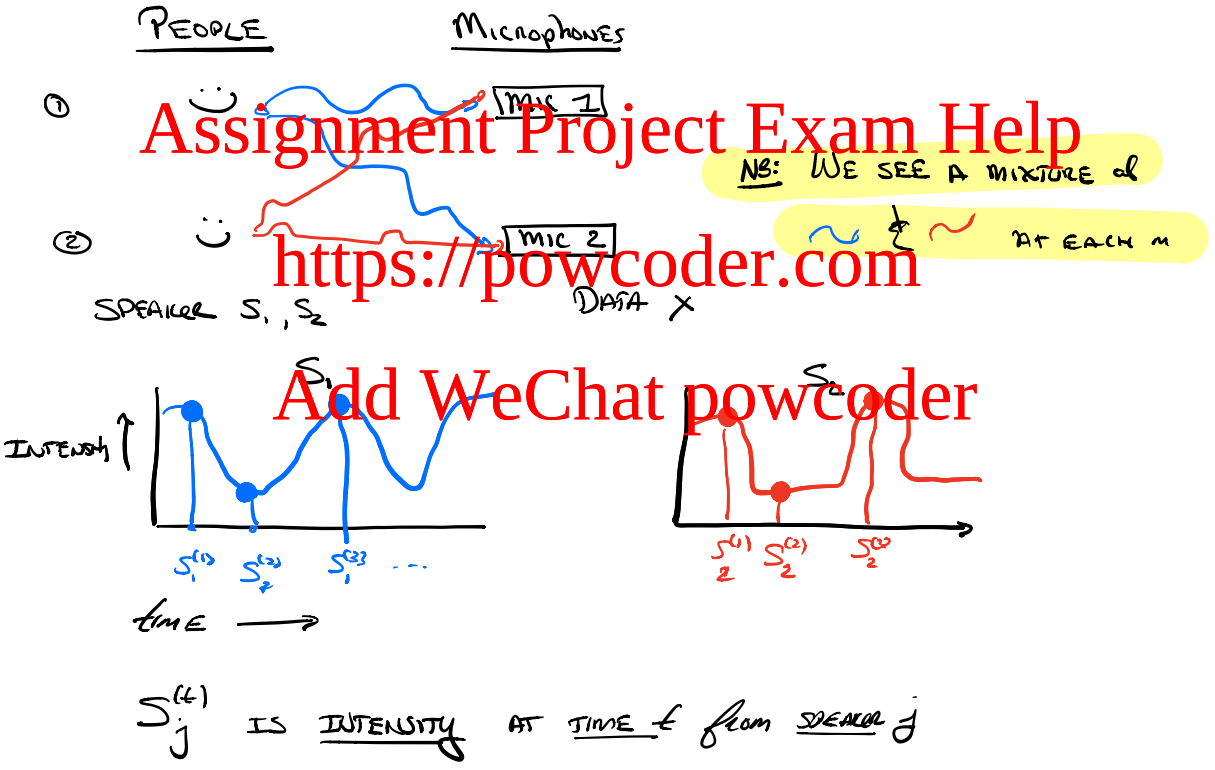
+ ICA: Source Separation (fun problem)
+ Weak Supervision

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

ICA INDEPENDENT Component Analysis

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

Cocktail Party Problem (IN HW!)



NB: WE SEE A MIXTURE of ~ & ~ AT EACH m

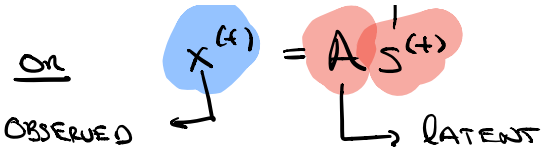
$S_j^{(t)}$ IS INTENSITY AT TIME t from SPEAKER j

WE DO NOT Observe $S^{(t)}$ only $x^{(t)}$ - the microphones

ex model $x_j^{(t)} = a_{j1} S_1^{(t)} + a_{j2} S_2^{(t)}$ ...

"MICROPHONE j SEES A MIXTURE of $S_1^{(t)}$ & $S_2^{(t)}$"

↳ LATENT



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

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

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

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

Some Caveats

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

- THERE ARE INHERENT AMBIGUITY
  - WE CAN'T DETERMINE SPEAKER ID (cold 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 $x^{(t)} \sim N(\mu, AA^T)$ 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 $P_U(\frac{x}{2}) = P_S(x)$

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

THE KEY ISSUE IS the NORMALIZATION CONSTANT

FOR INVERTIBLE MATRIX A, $U = AS$

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

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

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 $P_S(x)$ 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"

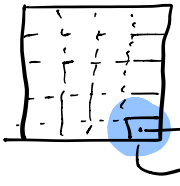
Extra: $P_U(x) = P_S(A^{-1}x) \cdot |\det(A^{-1})|$

SKIP IN lecture HAPPY to RECORD

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

Simplified linear change of variables for integrals

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



$\sum_{B \in \mathcal{P}} f(x_B)\,\mathrm{vol}(B)$

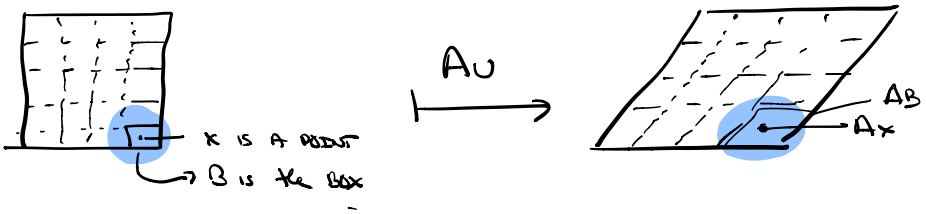
Let $\mathcal{P}$ be a partition of U. Let x_B be a point inside a box B

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

Recall as partition gets finer, $\left(\max_{B \in P} \mathrm{vol}(B) \to 0\right)$

then RHS 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)

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

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

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

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 derived)

- Indep case → Simple estimation trick
- Correlations → Inverse Covariance & Graph Structure.

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

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

find $P(y \mid \bar{\lambda}, x^{(i)})$ $\quad y \in \{-1, 1\}$

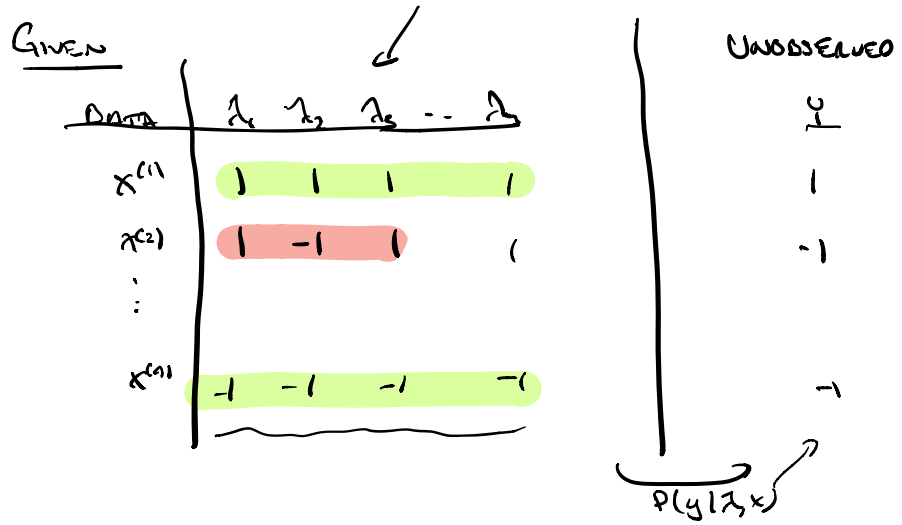
IDEA: λ_i is a noisy function (inaccurate, incomplete)

λ_1: "Name in dictionary"

λ_2: "Upper case word"

... Programmatic labels ... No labels

Model 0: No abstains, independent, balanced, clean, correlated



Each labeler has a HIDDEN ACCURACY. — x 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"

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

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

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

① $\mathbb{E}[\lambda_i y]$ OBSERVE if λ_i & y Agree VALUE is 1

λ_i & y disagree, value is -1

$$\therefore = p_i \cdot 1 + (1-p_i)(-1)$$

$$= 2p_i - 1 \triangleq a_i \quad \text{(defines } a_i \text{ this way)} \qquad a_i \in [-1, 1]$$

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

$$\mathbb{E}[\lambda_i \lambda_j \mid y=1] = p_i p_j + (1-p_i)(1-p_j) \quad \text{"Agree"}$$

$$+ (1-p_i) p_j (-1) + p_i (1-p_j)(-1) \quad \text{"disagree"}$$

$$= a_i a_j$$

NOTE we don't use $\mid y=1$, same true for $\mid y=-1$

$$\mathbb{E}[\lambda_i \lambda_j] = \sum_{b \in \{-1,1\}} \mathbb{E}[\lambda_i \lambda_j \mid y=b] P(y=b) = a_1 a_2 \sum_b P(y=b) = 1$$

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

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

NB: M CAN BE ESTIMATED — unlike y!

"Agreements and disagreements" ⇒ Key idea: don't need to see y

Simple Algorithm: for any i,j,k distinct, $M_{ij}, M_{jk}, M_{ik} \neq 0$

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

So we can solve upto sign of a_i.

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

then $M_{ik} = a_i a_k$

$\text{sign}(M_{ik})\text{sign}(a_i) = \text{sign}(a_k)$ $\therefore$ can solve for all signs. from one

So $a \neq -a$ 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 $P_i = 0 \Rightarrow \sum_{j=1}^{n} M_{ij} = 0 \Rightarrow \sum_{j=1}^{n} a_i a_j = 0 \Rightarrow a_i \sum_{j=1}^{n} a_j = 0$ RANDOM NOISE!

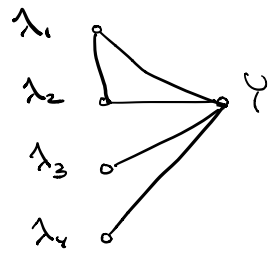
have every a_i bounded away from random noise

(can handle w/ fancier tricks).

Theory says "let $\xi = \min_{j=1..n} |P_j - \frac{1}{2}|$, need samples proportional to $\frac{1}{\xi^2}$" (lot of work)

RECAP: Symmetry, Simple Algebraic Learn (aka called the GCD)

What if labels are correlated?

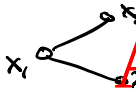


Key Concept: Probability distribution on Graphs.

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

if (i,j) $\not\in$ Edge above.

Nugget Structure of Inverse Covariance for Gaussians



$X \sim N(0, \Lambda)$

$x_2 \sim N(x_1, 1)$ $\quad \epsilon_2 \sim N(0,1)$ $\quad x_2 = x_1 + \epsilon_2$

$x_3 \sim N(x_2, 1)$ $\quad \epsilon_3 \sim N(0,1)$ $\quad x_3 = x_1 + \epsilon_3$

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

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

$\mathbb{E}[x_1^2] = 1$ $\quad \mathbb{E}[x_2^2] = \mathbb{E}[(x_1 + \epsilon_2)^2] = \mathbb{E}[x_1^2] + 2\mathbb{E}[x_1 \epsilon_2]^{\nearrow 0} + \mathbb{E}[\epsilon_2^2]$

$= 2$

$\mathbb{E}[x_1 x_2] = \mathbb{E}[x_1^2 + x_1 \epsilon_1] = 1$

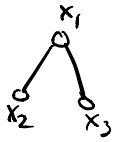
$\mathbb{E}[x_2 x_3] = \mathbb{E}[(x_1 + \epsilon_1)(x_1 + \epsilon_3)] = 1$

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

NO CLEAR STRUCTURE?

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

NO Edge



We say A probability distribution $p: \mathbb{R}^d \to [0,1]$ agrees or factorizes wrt a graph $G = (V, E)$ if

$$p(x) = \frac{1}{Z} \prod_{e=(i,j) \in E} p_e(x_i, x_j) \cdot \prod_{v \in V} p_v(x_i)$$

for some functions!

Normalization Constant

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

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

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

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

$$\sum_{i,j} A_{ij} x_i x_j = \quad [\text{redacted}]$$

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

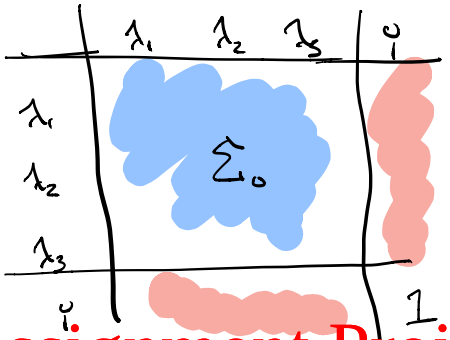
But $\Sigma^{-1} = A$ is symmetric, so $A_{ij} = 0$.

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

More complex theory for discrete R.V.s for ℓ₁ Wainwright 2014
Ravikumar et al 2018

Back to our Problem

Form "covariance matrix"



We "see" Σ_O but not all of Sym $\mathbb{E}[\lambda_i Y]$ is unobserved.

let $O = \{1,2,3\}$ ← visible terms

$(\Sigma^{-1})_O = (\Sigma_O - UU^T)^{-1}$ ← Sparse + rank 1 update.

↳ inverse on those entries

let $B = \Sigma_O$

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

$$z = \frac{B^{-1}U}{\sqrt{1 - U^T B^{-1} U}}$$

So $(\Sigma^{-1})_O = \Sigma_O^{-1} + zz^T$

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

Hence $(\Sigma_O^{-1})_{ij} = -z_i z_j$

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

This is a linear system in z_i^2 & z_j^2, and we can solve (if enough pairwise indep)

IN the NOTES:

- Higher RANK VERSIONS (HANDLE MORE CORRELATIONS)
- How to learn graph Structure
- How to Handle sampling error (modern ML, power tools)

RECAP:

- WEAK Supervision framework
- Nuggets about Concepts & Prob. distributions (take graphical models!)
- "Method of moments" style ALGORITHMS