MAT8034: Machine Learning

Independent components analysis

Fang Kong

https://fangkongx.github.io/Teaching/MAT8034/Spring2025/index.html

Part of slide credit: Stanford CS229

Motivation

- Consider the cocktail party problem
  - *d* speakers are talking simultaneously in a room
  - Place *d* microphones at different locations
  - Each microphone records a different combination of the speakers' voices
- Can we recover the original speech signals of each speaker?

Problem formulation

- Source $s \in \mathbb{R}^d$
- Observation $x \in \mathbb{R}^d$

- Model the observation and source

$$x = As$$

  - A is the mixing matrix

Problem formulation (cont'd)

- Now we have multiple observations

$$\{x^{(i)}; i = 1, \ldots, n\}$$

  - The i-the data satisfies $x^{(i)} = As^{(i)}$

- Illustration
  - x_j^i is the acoustic reading recorded by microphone j at time i
  - s_j^i is the sound that speaker j was uttering at time i

Objective

- Given observation x^i, can we recover the sources?

- How?
  - $s = A^{-1}x := Wx$ Unmixing matrix
  - $$W = \begin{bmatrix} — & w_1^T & — \\ & \vdots & \\ — & w_d^T & — \end{bmatrix}$$ *then* $s_j^i = W_j^\top x^i$

ICA ambiguities

To what degree can W be recovered?

- Only given x, are there cases where W is impossible to recover?

To what degree can W be recovered?

- Only given x, are there cases where W is impossible to recover?

- How about the permutation?

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

  - Given an observation x, can you distinguish between Ws and PWs'?

To what degree can W be recovered?

- Only given x, are there cases where W is impossible to recover?

- How about the scaling?
  - Given an observation x, can you distinguish between Ws and $(2W)(0.5s)$?

- Permutation and scaling do not matter for most applications

To what degree can W be recovered?

- Only given x, are there cases where W is impossible to recover?

- How about the rotational symmetry?
  - Consider an example with $n = 2, s \sim N(0, I)$
  - Now we observe $x = As$

$$\mathrm{E}_{s \sim \mathcal{N}(0,I)}[x] = \mathrm{E}[As] = A\mathrm{E}[s] = 0$$

$$\mathrm{Cov}[x] = \mathrm{E}_{s \sim \mathcal{N}(0,I)}[xx^T] = \mathrm{E}[Ass^T A^T] = A\mathrm{E}[ss^T]A^T = A \cdot \mathrm{Cov}[s] \cdot A^T = AA^T$$

  - Thus $x \sim N(0, AA^\top)$

To what degree can W be recovered?

- Only given x, are there cases where W is impossible to recover?

- How about the rotational symmetry?
  - Consider an example with $n = 2, s \sim N(0, I)$
  - Now we observe $x = As$
  - Thus $x \sim N(0, AA^\top)$

  - Consider another generation $x' = A's$
  - We can construct $A' = AR$ with $RR^\top = R^\top R = I$
  - Can we distinguish x' from x?

To what degree can W be recovered?

- Only given x, are there cases where W is impossible to recover?
    - So long as the data is not Gaussian, it is possible to recover the d independent sources with enough data

ICA algorithm

Maximum likelihood

- Construct a joint distribution of the sources

$$p(s) = \prod_{j=1}^{d} p_s(s_j)$$

Imply independence

- Recall that the observation follows $x = As$, $s = A^{-1}x \coloneqq Wx$

- What's the probability of x?
  - Is $p_x(x) = p_s(Wx)$?

Counterexample

- Is $p_x(x) = p_s(Wx)$?
  - Suppose:
    $s \sim \text{Uniform}[0,1] \Rightarrow p_s(s) \ = 1_{[0\leq\, x \leq 1]}$
    $A = 2$, so $x = 2s$
  - Then $x \sim \text{Uniform}[0,2]$
  - $p_x(x) = 0.5 \cdot 1_{[0\leq\, x \leq\, 2]}$
  - But:
    $p_s(Wx) = p_s(0.5x) = 1$, which is incorrect
- Intuition: This ignores how the distribution stretches or compresses in space

Densities and linear transformations

- The correct formulation

$$\boxed{p_x(x) = p_s(Wx) \cdot |W|} \quad \text{where} \quad W = A^{-1}$$

  - Accounts for scaling/stretching of the space

- For multi-dimensional vectors

$$\boxed{p_x(x) = p_s(A^{-1}x) \cdot |\det(A^{-1})| = p_s(Wx) \cdot |W|}$$

Intuition

- Let $C_1 = [0,1]^d$ (unit hypercube)
- Let $C_2 = \{As : s \in C_1\}$
- Then:
  - $Vol(C_2) = |\det(A)|$

- If $s \sim$Uniform(C_1), then:
  - $p_x(x) = \frac{1}{Vol(C_2)} = \frac{1}{\det(A)} = |\det(W)|$

Back to maximum likelihood

- Construct a joint distribution of the sources

$$p(s) = \prod_{j=1}^{d} p_s(s_j)$$

- Recall that the observation follows $x = As$, $s = A^{-1}x := Wx$

- What's the probability of x?
  - $p_x(x) = p_s(Wx)|W|$?

$$\Longrightarrow \quad p(x) = \prod_{j=1}^{d} p_s(w_j^T x) \cdot |W|$$

How to specify a density for s?
Cannot be gaussian

Specify a density for sources

- The density function is the derivative of the cumulative distribution function (cdf)

$$F(z_0) = P(z \leq z_0) = \int_{-\infty}^{z_0} p_z(z) dz$$

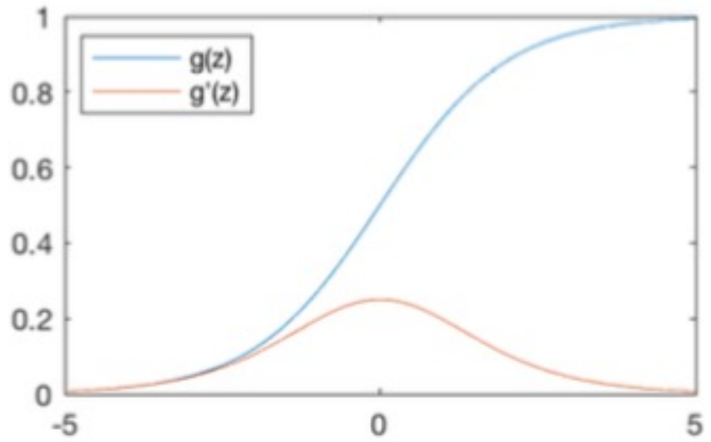
$$p_z(z) = F'(z)$$

- We can first specify a cdf (a monotonic function that increases from zero to one)
  - Sigmoid?

Selecting Sigmoid

- Log-likelihood becomes

$$\ell(W) = \sum_{i=1}^{n} \left(\sum_{j=1}^{d} \log g'(w_j^T x^{(i)}) + \log |W| \right)$$



$$g(z) = \frac{1}{1 + e^{-z}}$$

$$g'(z) = g(z)(1 - g(z))$$

- Using stochastic gradient ascent to optimize

Summary

- Independent components analysis (ICA)
  - Motivation: detect independent source feature
  - ICA ambiguities (permutation, scale, rotational symmetry)
  - Algorithm: maximum likelihood to find the unmixing matrix