

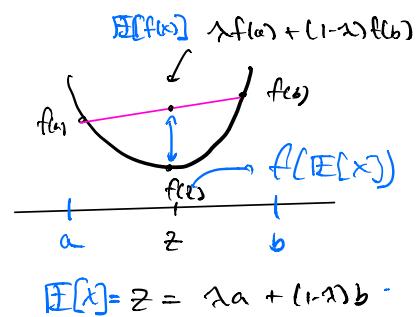
Em Algorithm

- + Derive GMM AS AN EM ALGORITHM
- + Factor Analysis (high dimensional data w/ Em)

Recall: JENSEN'S INEQUALITY

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$$

Canonical convex fn. x^2



Def g Concave if -g convex

e.g. $\log(x)$ on $(0, \infty)$



or $\frac{d^2}{dx^2} \log(x) = \frac{d}{dx} x^{-1} = -x^{-2} < 0$ on $(0, \infty)$

for concave g, JENSEN says

$$\therefore \mathbb{E}[-g(x)] \geq -g(\mathbb{E}[x])$$

or $\mathbb{E}[g(x)] \leq g(\mathbb{E}[x])$ (INEQUALITY REVERSES)

NB: if x is constant $\Rightarrow \mathbb{E}[g(x)] = g(\mathbb{E}[x])$

for strongly concave g ($g''(x) < 0$) then, \Leftarrow holds a.s.

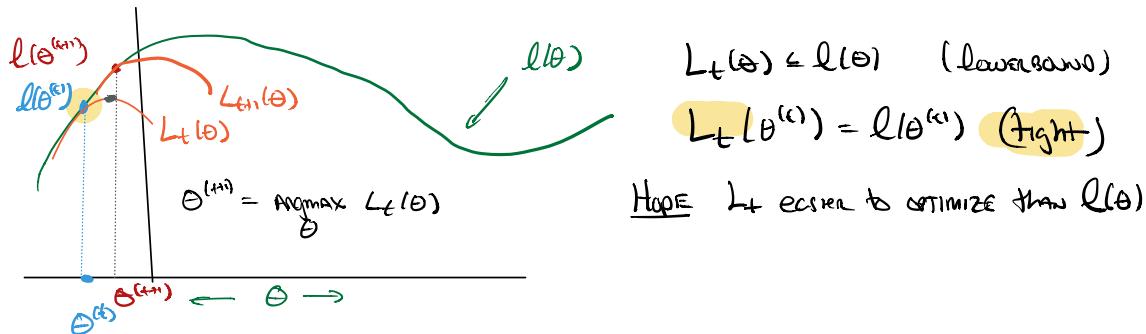
EM Algorithm as max likelihood

$$l(\theta) = \sum_{i=1}^n \log P(x^{(i)}; \theta)$$

↑ DATA ↑ PARAMETERS

WE ASSUME $P(x_j; \theta) = \sum_z P(x, z; \theta)$ \leftarrow GMM LATENT VARIABLE

Picture of Algorithm



Rough Algo

(E-STEP) 1. GIVEN $\theta^{(t)}$ FIND L_t

(M-STEP) 2. GIVEN L_t , SET $\theta^{(t+1)} = \operatorname{argmax}_{\theta} L_t(\theta)$

How do we construct L_t ? (let's look at single DATA point $x^{(i)}$)

$$\log \sum_z P(x, z; \theta) = \log \sum_z \frac{Q(z) P(x, z; \theta)}{Q(z)} \quad \text{for any } Q(z)$$

WE PICK $Q(z)$ S.T. $\sum_z Q(z) = 1$ AND $Q(z) = 0 \Leftrightarrow$

$$= \log \mathbb{E}_z \left[\frac{P(x, z; \theta)}{Q(z)} \right] \quad (\text{Symbol Rushing})$$

$$\geq \mathbb{E}_z \left[\log \frac{P(x, z; \theta)}{Q(z)} \right] \text{ JENSEN! } (\log \text{ is concave})$$

$$= \sum_z Q(z) \log \frac{P(x, z; \theta)}{Q(z)} \text{ (DEF of } \mathbb{E} \text{)}$$

Key step holds for any such Q : (a)

This gives a family of lower bounds, one for each choice of Q ($D_F \leq l$)

How do we make it tight? Select Q to make inequality tight

What if... $\log \frac{P(x, z; \theta)}{Q(z)} = c$ for some constant, then JENSEN's is Equality!

$$\text{So, } Q(z) = P(z|x; \theta) \text{ then } P(x, z; \theta) = P(z|x; \theta) P(x; \theta)$$

$c = \log P(x; \theta)$ does not depend on z , so constant!

NB: $Q(z)$ does depend on $\theta + x$ - we will select a $Q^{(i)}(z)$

for every point independently. (i.e. for every i)

Comment Since $\log(\cdot)$ is strongly concave, this was our only choice

WE DEFINE Evidence-based Lower Bound (ELBO), sum over z

$$\text{ELBO}(x, Q, z) = \sum_z Q(z) \log \frac{P(x, z; \theta)}{Q(z)}$$

WE'VE SHOWN $\ell(\theta) \geq \sum_{i=1}^n \text{ELBO}(x^{(i)}, Q^{(i)}; \theta)$ for any $Q^{(i)}$ satisfying (a)
lower bound

$$\ell(\theta^{(i)}) = \sum_{i=1}^n \text{ELBO}(x^{(i)}, Q^{(i)}; \theta^{(i)}) \text{ for choice of } Q^{(i)} \text{ above.}$$

RESTATE EM

$$\begin{aligned}
 (\text{E-STEP}) \quad & \text{for } i=1..n \quad \text{SET } Q_i(z) = P(z^{(i)} | x^{(i)}, \theta^{(t)}) \\
 (\text{M-STEP}) \quad & \theta^{(t+1)} = \underset{\theta}{\operatorname{Argmax}} \mathcal{L}_t(\theta) \\
 & = \underset{\theta}{\operatorname{Argmax}} \sum_{i=1}^n ELBO(x^{(i)}, Q^{(t)}, \theta)
 \end{aligned}$$

DATA & INPUT PARAMS

WARM UP: Mixture of Gaussians. EM RECOVERS OUR AD-HOC ALGORITHM

$$P(x^{(i)}, z^{(i)}) = P(x^{(i)} | z^{(i)}) P(z^{(i)})$$

$$z^{(i)} \sim \text{Multinomial}(\mathbf{Q}) \quad \phi_j \geq 0 \quad \sum_{j=1}^k \phi_j = 1 \quad \text{"in cluster j"}$$

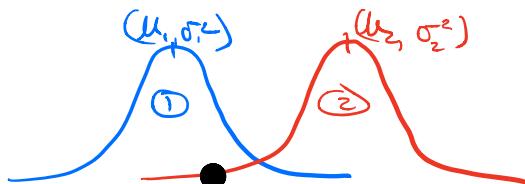
$$x^{(i)} | z^{(i)} = j \sim N(\mu_j, \sigma_j^2) \quad \text{"source means"}$$

$z^{(i)}$ is our LATENT VARIABLE. K SOURCES

WHAT IS EM HERE?

$$Q_i(z) = P(z^{(i)} = j | x^{(i)}; \theta)$$

WE SAW THAT COULD COMPUTE VIA Bayes Rule $P(x^{(i)} | z^{(i)} = j)$



1. much more likely for ① than ②
2. BUT if we knew $\phi_2 \gg \phi_1$, maybe we'd think likely from ②

Bayes Rule AUTOMATES this REASONING

M-STEP: Compute Derivatives ..

$$\begin{aligned}
 \text{MAX}_{\phi, \mu, \Sigma} \quad & \sum_{i=1}^n \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \\
 & f(\theta)
 \end{aligned}$$

WRITE Θ FOR NOTATION ABOVE

$$\text{Rewrite } \omega_j^{(i)} \triangleq Q_i(z=j)$$

$$P(x^{(i)}, z^{(i)}; \Theta) = P(x^{(i)} | z^{(i)}) P(z^{(i)})$$

$$f_i(\Theta) = \sum_j \omega_j^{(i)} \log \left(\frac{\frac{1}{2\pi |\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right\} \cdot \phi_j}{\omega_j^{(i)}} \right) \quad \text{GAUSSIAN } (\mu, \Sigma)$$

$$\nabla_{\mu_j} \hat{\sum}_i f_i(\Theta) = \sum_i \nabla_{\mu_j} \left(\omega_j^{(i)} - \frac{1}{2} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right)$$

$$= -\frac{1}{2} \sum_i \omega_j^{(i)} \Sigma_j^{-1} (x^{(i)} - \mu_j) = -\frac{1}{2} \sum_j \left(\sum_i \omega_j^{(i)} (x^{(i)} - \mu_j) \right)$$

SETTING TO 0 AND USING Σ_j^{-1} IS FULL RANK $\Rightarrow \sum_i \omega_j^{(i)} (x^{(i)} - \mu_j) = 0$

$$\therefore \mu_j = \frac{\sum_i \omega_j^{(i)} x^{(i)}}{\sum_i \omega_j^{(i)}} \quad \text{(AS BEFORE)}$$

ϕ_j is CONSTRAINED $\sum_j \phi_j = 1, \phi_j \geq 0$, NEED LAGRANGIAN

$$\nabla \phi_j = \sum_{i=1}^n \omega_j^{(i)} \nabla_{\phi_j} \log \phi_j + \nabla_{\phi_j} \lambda \left(\sum_j \phi_j - 1 \right)$$

$$= \sum_{i=1}^n \frac{\omega_j^{(i)}}{\phi_j} + \lambda = 0 \Rightarrow \phi_j = -\frac{1}{\lambda} \sum_{i=1}^n \omega_j^{(i)}$$

$$\text{SINCE } \sum \phi_j = 1, \quad \sum_j \phi_j = -\frac{1}{\lambda} \sum_i \omega_j^{(i)} = -\frac{n}{\lambda}$$

$$\therefore \phi_j = \frac{1}{n} \sum_i \omega_j^{(i)}$$

MESSAGE: EM RECOVERS GMM AUTOMATICALLY.

NB: IF $z^{(i)}$ IS CONTINUOUS, ONE CAN REPLACE SUMS w/ INTERVALS

Factor Analysis

MANY fewer parts than dimensions "n < d"

cf: GMMs n ≫ d lots of neurons, few sources.

How does this happen?

PLACE SENSORS ALL OVER CAMPUS, RECORD @ 1000s of locations
 $\Rightarrow d \approx 1000s$

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

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

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

RANK(Σ) ≤ n < d - not full rank.

Problem IN GAUSSIAN likelihood

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

IS NOT DEFINED.

$$\hookrightarrow |\Sigma| = 0$$

WE will fix these issues by examining three models

that are simpler. Spoiler: we'll combine these in the end!

RECALL MLE for Gaussian

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

Equivalent

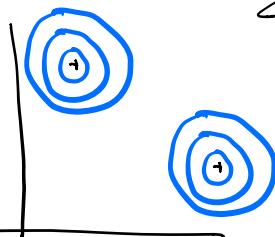
$$\underset{\mu, \Sigma}{\text{MIN}} \sum_{i=1}^n (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) + \log |\Sigma|$$

If Σ is full rank, $\nabla_\mu = \sum_{i=1}^n \Sigma^{-1} (\mathbf{x} - \mu) = 0 \Rightarrow \mu = \frac{1}{n} \sum_i \mathbf{x}^{(i)}$
we'll use this as plugin below.

Building Block 1

Suppose INDEPENDENT AND IDENTICAL COVARIANCE

$$\hat{\Sigma} = \sigma^2 \mathbb{I} \quad (\text{NB: PARAMETER Tying})$$



COVARIANCE "ARE CIRCLES"

WHAT IS MLE FOR $\hat{\Sigma}$?

$$|\Sigma| = 2\delta$$

$$\underset{\sigma^2}{\text{MIN}} \sigma^{-2} \underbrace{\sum_{i=1}^n (\mathbf{x}^{(i)} - \mu)^T (\mathbf{x}^{(i)} - \mu)}_C + d \log \sigma^2$$

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

$$\Rightarrow \frac{1}{z} = -z^{-2} C + \frac{nd}{z} = 0 \Rightarrow z = \frac{C}{nd}$$

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

"SUBTRACT MEAN AND SQUARE ALL ENTRIES."

Building Block 2

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



Axis Aligned ellipse

SET $z_i = \sigma_i^2$ (same idea as above)

$$\min_{z_1 \dots z_d} \sum_{i=1}^n \sum_{j=1}^d z_j^{-1} (x^{(i)} - \mu_j)^2 + \log z_j$$

This is d problems for each 1 dimension

$$\Rightarrow \sum_{i=1}^n z_j^{-1} (x^{(i)} - \mu_j)^2 + \log z_j$$

$$\Rightarrow \sigma_j^2 = \frac{1}{n} \sum_i (x_j^{(i)} - \mu_j)^2$$

Our FACTOR model

PARAMETERS

$$\mu \in \mathbb{R}^d$$

$$\Lambda \in \mathbb{R}^{d \times s}$$

$$\Phi \in \mathbb{R}^{d \times d} \text{ - DIAGONAL MATRIX}$$

MODEL

$$P(x, z) = P(x|z) P(z) \quad z \text{ IS LATENT}$$

$$z \sim N(0, I) \in \mathbb{R}^s \text{ for } s < d \text{ "small dim"}$$

$$x = \underbrace{\mu}_{\substack{\text{MEAN} \\ \text{IN} \\ \text{the space}}} + \underbrace{\Lambda z}_{\substack{\text{MAPS FROM SMALL LATENT SPACE TO LARGE SPACE}}} + \epsilon \quad \text{or} \quad x \sim N(\mu + \Lambda z, \Phi)$$

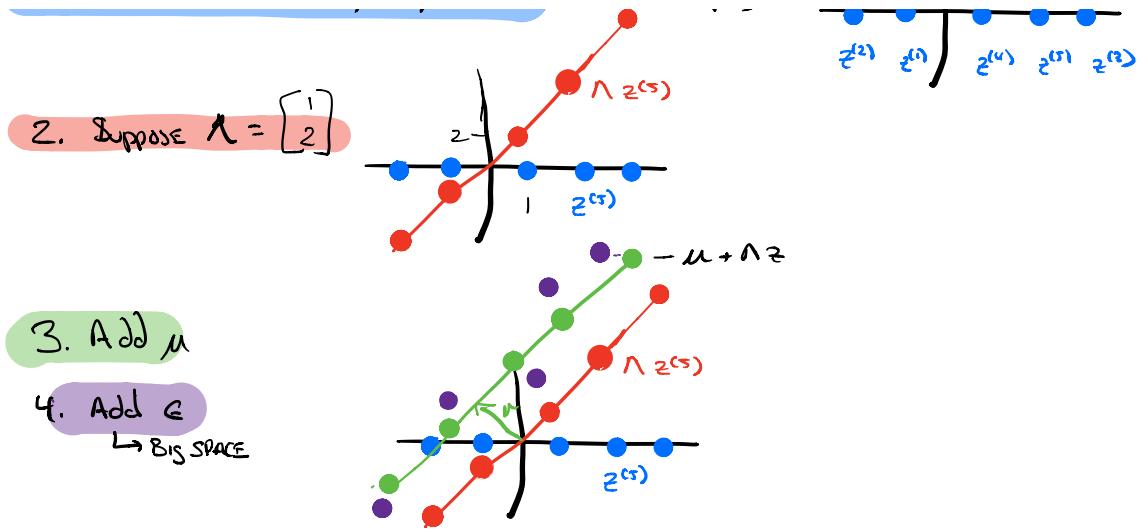
$$\epsilon \sim N(0, \Phi) \quad \text{Noisy}$$

$$\underline{\text{Ex:}} \quad d=2, \quad s=1, \quad n=5$$

$$x = \underbrace{\mu}_{\text{mean}} + \underbrace{\Lambda z}_{\text{latent variable}} + \epsilon$$

1. GENERATE $z^{(1)}, \dots, z^{(s)}$ from $N(0, I)$





DATA WE WOULD OBSERVE ARE PURPLE DOTS

SO SMALL LATENT SPACE PRODUCES DATA IN HIGH DIM SPACE.

TECHNICAL TOOLS: BLOCK GAUSSIANS

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \quad \mathbf{x}_1 \in \mathbb{R}^{d_1}, \mathbf{x}_2 \in \mathbb{R}^{d_2}$$

$$\mathbf{x} \in \mathbb{R}^d$$

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

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 SWAPPING)

FACT 2: $P(x_1 | x_2) \sim N(\mu_{1|2}, \Sigma_{1|2})$ CONDITIONING

$$\mu_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2)$$

$$\hat{\Sigma}_{12} = \hat{\Sigma}_{11} - \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1} \hat{\Sigma}_{21} \quad (\text{matrix inversion lemma})$$

Proofs outline (apply to ABB)

Summary: Marginalization $\not\equiv$ Conditioning Gaussian \Rightarrow
Another GAUSSIAN (CLOSED)
WE HAVE formula for PARAMETERS.

Back to Factor Analysis

$$x = \mu + \Lambda z + \epsilon$$

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

$$\mathbb{E}[x] = \mu$$

WHAT IS Σ ?

$$\hat{\Sigma}_{11} = \mathbb{E}[zz^T] = \mathbf{I}$$

$$\begin{aligned} \hat{\Sigma}_{12} &= \mathbb{E}[z(x-\mu)^T] = \mathbb{E}[zz^T \Lambda^T] + \mathbb{E}[z\epsilon^T] \\ &= \Lambda^T \end{aligned}$$

$$\hat{\Sigma}_{21} = \hat{\Sigma}_{12}^T$$

$$\begin{aligned} \hat{\Sigma}_{22} &= \mathbb{E}[(x-\mu)(x-\mu)^T] \\ &= \mathbb{E}[(\Lambda z + \epsilon)(\Lambda z + \epsilon)^T] \\ &= \mathbb{E}[\Lambda z z^T \Lambda^T] + \mathbb{E}[\epsilon \epsilon^T] \\ &= \Lambda \Lambda^T + \Phi \end{aligned}$$

$$\hat{\Sigma} = \begin{bmatrix} \mathbf{I} & \Lambda^T \\ \Lambda & \Lambda \Lambda^T + \Phi \end{bmatrix}$$

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.