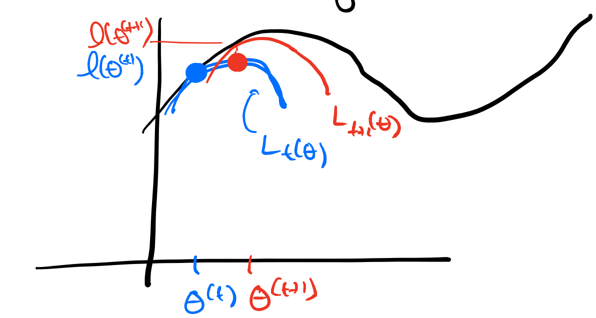
Applications of EM & Factor Analysis

- Finish EM properties
- Gaussian mixture model as EM
- Factor Analysis

Recall EM Algorithm



1. $L_t(\theta) \le \ell(\theta)$ lower bound
2. $L_t(\theta^{(t)}) = \ell(\theta^{(t)})$ "tight"

Last time

$$\ell(\theta) = \sum_{i=1}^{n} \log \sum_{z} Q_i(z) \frac{P(x^{(i)}, z; \theta)}{Q_i(z)}$$

$\forall i \quad \sum_z Q_i(z) = 1, \quad Q_i(z) \geq 0$

$$L_t(\theta) = \sum_{i=1}^{n} \sum_{z} Q_i(z) \log \frac{P(x^{(i)}, z; \theta)}{Q_i(z)}$$

We showed Property 1, $\ell(\theta) \geq L_t(\theta)$ key step is Jensen

$$\log \sum_z Q_i(z) \frac{P(x^{(i)}, z; \theta)}{Q_i(z)} \geq \sum_z Q_i(z) \log \frac{P(x^{(i)}, z; \theta)}{Q_i(z)}$$

$\text{ELBO}(x^{(i)}, z, \theta)$

To set Property 2, we pick $Q_i(z)$ @ $\theta^{(t)}$

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

Verify, since $\frac{P(x^{(i)}, z; \theta)}{Q_i(z)} = \frac{P(z \mid x^{(i)}; \theta) P(x^{(i)}; \theta)}{\cancel{P(z \mid x^{(i)}; \theta)}}$ does not depend on z

(LHS) $\log \sum_z Q_i(z) \cdot c = \log c$

(ELBO(x…)) $\sum Q_i(z) \log c = \log c$, Property 2 holds.

Restate EM

(E-step) for $i = 1 \ldots n$ set $Q_i(z) = P(z^{(i)} \mid x^{(i)}, \theta^{(t)})$

$\hookrightarrow$ Data & input params

(M-step) $\theta^{(t+1)} = \underset{\theta}{\text{Argmax}}\ \mathcal{L}_t(\theta)$

$$= \underset{\theta}{\text{Argmax}} \sum_{i=1}^{n} \text{ELBO}(x^{(i)}, Q^{(i)}, \theta)$$

Warm Up: Mixture of Gaussians.

EM recovers our ad hoc algorithm

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

$z^{(i)} \sim \text{Multinomial}(\Phi)$ $\quad \phi_i \geq 0 \quad \sum_i \phi_i = 1$ "In cluster j"

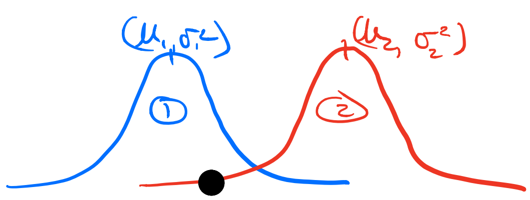
$x^{(i)} \mid z^{(i)} = j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ "Cluster means"

$z^{(i)}$ is our latent variable.

What is EM here?

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

We saw that could compute via Bayes Rule $P(x^{(i)} \mid z^{(i)} = j; \theta)$



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...

$$\max_{\phi, \mu, \Sigma} \sum_{i=1}^{n} \underbrace{\sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}}_{f_i(\theta)}$$

↳ WRITE Θ FOR NOTATION ABOVE

<u>RECALL</u> $\quad w_j^{(i)} \triangleq Q_i(z=j)$

$$P(x^{(i)}, z^{(i)}; \theta) = P(x^{(i)} \mid z^{(i)})\, P(z^{(i)})$$

GAUSSIAN (μ, Σ)

$$f_i(\theta) = \sum_j w_j^{(i)} \log\left(\frac{\frac{1}{2\pi |\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1} (x^{(i)}-\mu_j)\right\} \cdot \phi_j}{w_j^{(i)}} \right)$$

$$\nabla_{\mu_j} \sum_i f_i(\theta) = \sum_i \nabla_{\mu_j} \left(w_j^{(i)} \; -\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1} (x^{(i)}-\mu_j) \right)$$

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

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

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

ϕ_j IS <u>CONSTRAINED</u> $\sum_j \phi_j = 1,\ \phi_j \geq 0$, NEED LAGRANGIAN

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

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

SINCE $\sum \phi_j = 1$, $\quad \sum_j \phi_j = -\frac{1}{\lambda} \sum_{i,j} w_j^{(i)} = -\frac{n}{\lambda}$

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

<u>MESSAGE</u>: EM RECOVERS GMM AUTOMATICALLY.

NB: IF $z^{(i)}$ IS CONTINUOUS, CAN CAN REPLACE SUMS W/ INTEGRALS

Factor Analysis

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

cf: GMMs $n \gg d$ lots of electrons, 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

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

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

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

Problem in Gaussian likelihood — is not defined.

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

$\hookrightarrow |\Sigma| = 0$

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

Recall MLE for Gaussian

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

Equivalent

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

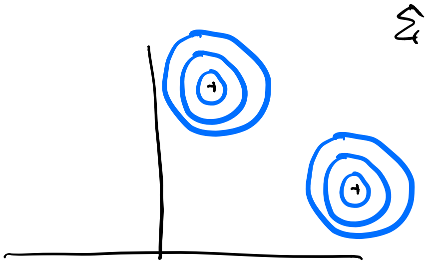
If Σ_i is full rank, $\nabla_\mu = \sum_{i=1}^{n} \Sigma^{-1}(x-\mu) = 0 \Rightarrow \mu = \frac{1}{n}\sum_i x^{(i)}$

we'll use this as plugin below.

Building Block 1

Suppose independent and identical covariance

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



Covariance "are circles"

What is MLE for Σ_i? $|\Sigma_i| = 2d$

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

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

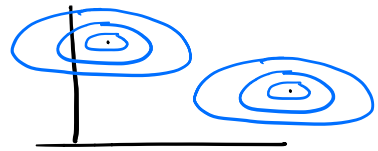
$$\Rightarrow \nabla_z = -z^{-2} c + \frac{nd}{z} = 0 \Rightarrow z = \frac{c}{nd}$$

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

"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)

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

this is d Problems for Each i dimension

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

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

One 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

$$P(x, z) = P(x \mid z) P(z)$$ z IS LATENT

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

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

MEAN IN the SPACE

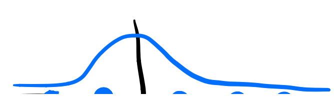
↳ MAPS FROM SMALL LATENT SPACE TO LARGE SPACE

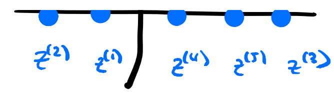
$\epsilon \sim N(0, \Phi)$ NOISY

Ex: $d = 2$, $s = 1$, $n = 5$

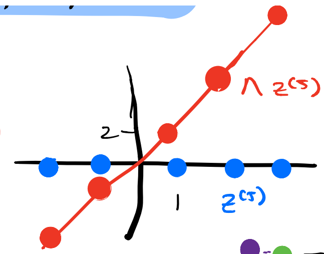
$x = \mu + \Lambda z + \epsilon$

1. GENERATE $z^{(1)}, \ldots, z^{(5)}$ from $N(0, 1)$



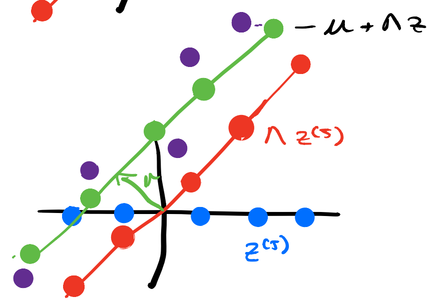


2. Suppose $\Lambda = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$



3. Add μ

4. Add ϵ
↳ Big space



Data we would observe are purple dots

So small latent space produces data in high dim space.

Technical tools: Block Gaussians

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

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

Notation is widely used and helpful.

Fact 1: $P(x_1) = \int_{x_2} P(x_1, x_2)$ Marginalization

For Gaussians, $p(x_1) = N(\mu_{1}, \Sigma_{11})$ (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 ⇒
ANOTHER GAUSSIAN (CLOSED)
WE HAVE FORMULA FOR PARAMETERS.

BACK TO FACTOR ANALYSIS $\quad x = \mu + \Lambda z + \epsilon$

$$\begin{pmatrix} z \\ x \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \Sigma\right)$$

SINCE $\mathbb{E}[z] = 0$

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

WHAT IS Σ?

$$\Sigma_{11} = \mathbb{E}[zz^T] = I$$

$$\Sigma_{12} = \mathbb{E}[z(x-\mu)^T] = \mathbb{E}[zz^T\Lambda^T] + \mathbb{E}[\cancel{z\epsilon^T}]^{\,0}$$
$$= \Lambda^T$$

$$\Sigma_{21} = \Sigma_{12}^T$$

$$\Sigma_{22} = \mathbb{E}[(x-\mu)(x-\mu)^T]$$
$$= \mathbb{E}[(\Lambda z + \epsilon)(\Lambda z + \epsilon)^T]$$
$$= \mathbb{E}[\Lambda z z^T \Lambda^T] + \mathbb{E}[\epsilon\epsilon^T]$$
$$= \Lambda\Lambda^T + \Phi$$

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

E-STEP: $Q_i(z) = P(z^{(i)} \mid x^{(i)}; \theta)$ — 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.