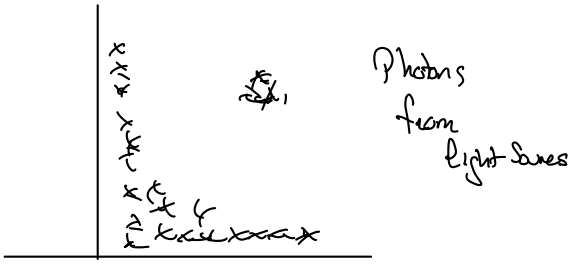
The EM Algorithm



Recall Unsupervised Algorithms

K-means & GMM

Big idea latent variable → "fraction of points from a source"

Algorithm "guess" latent variable

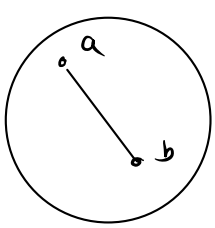
Estimate other parameters (center and shape)

Today: EM Algorithm for latent variable models

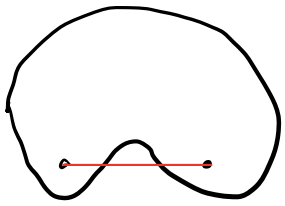
+ Technical Detour: Convexity & Jensen's inequality
+ EM Algorithm as MLE
+ GMM as EM Algorithm
+ Factor Analysis (EM for a new setting)

Detour Convexity & Jensen (This is a key result, we'll go slowly)

A set Ω is convex if for any $a, b \in \Omega$ the line joining a, b is in Ω as well.



Convex



NOT convex!

In symbols,

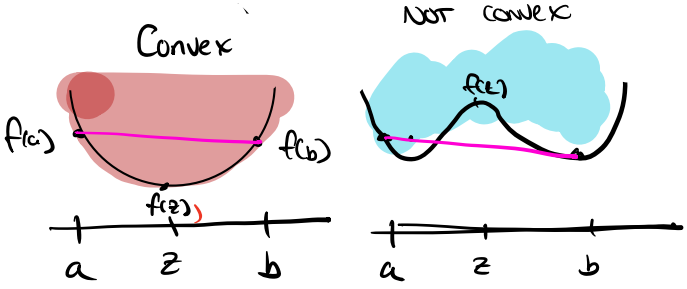
$$\forall \lambda \in [0,1], \quad a, b \in \Omega$$
$$\lambda a + (1-\lambda) b \in \Omega$$

(need to check $\forall a, b \in \Omega$)

Given a function f, the graph of f G_f is defined as

$$G_f = \{ (x, y) : y \geq f(x) \}$$

A function is <u>convex if its graph is convex (as a set)</u>



In symbols, $\forall \lambda \in [0,1]$

$$\lambda (a, f(a)) + (1-\lambda)(b, f(b)) \in \Omega$$

<u>or</u> let $z = \lambda a + (1-\lambda) b$

$$\lambda f(a) + (1-\lambda) f(b) \geq f(z)$$

"Every chord is above function"

If f twice differentiable, $\forall x$ $f''(x) \geq 0 \Rightarrow f$ is convex

Pf
$$f(a) = f(z) + f'(z)(a-z) + f''(\xi_a)(a-z)^2 \quad \exists \xi_a \in [a, z]$$
$$f(b) = f(z) + f'(z)(b-z) + f''(\xi_b)(b-z)^2 \quad \exists \xi_b \in [z, b]$$
$$\lambda f(a) + (1-\lambda) f(b) = f(z) + f'(z)(\lambda a + (1-\lambda) b - z) + C \quad C \geq 0$$

($\lambda a + (1-\lambda) b - z = 0$ by def of z.)

i.e. $\lambda f(a) + (1-\lambda) f(b) \geq f(z)$ □

We say f is strongly convex if $\forall x \in \text{Dom}(f)$ $f''(x) > 0$.

Ex: $f(x) = x^2 \Rightarrow f'(x) = 2 \Rightarrow$ Strongly convex

$f(x) = x^2(x-1)^2$: graph above (not convex)

Jensen's Inequality $\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$ for convex f.

Ex: x takes value a with prob λ

takes value b with prob $1-\lambda$

$\mathbb{E}[f(x)] = \lambda f(a) + (1-\lambda) f(b)$

$f(\mathbb{E}[x]) = f(z) \qquad z = \lambda a + (1-\lambda) b$

NB: Can prove finitely supported distribution by induction

for convex f, definition implies thm in this case!

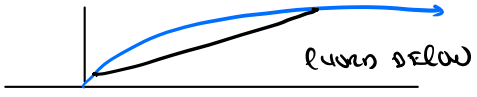
Stronger if f is strongly convex, and $\mathbb{E}[f(x)] = f(\mathbb{E}[x])$

$\Rightarrow$ X is a constant (except: almost surely)

We need **concave functions** g concave if $-g$ is convex

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

ex: $g(x) = \log(x) \Rightarrow g''(x) = -x^{-2}$ on $(0, \infty)$ negative



What about $h(x) = ax + b$ convex & concave since $h''(x) = 0$.

End detour

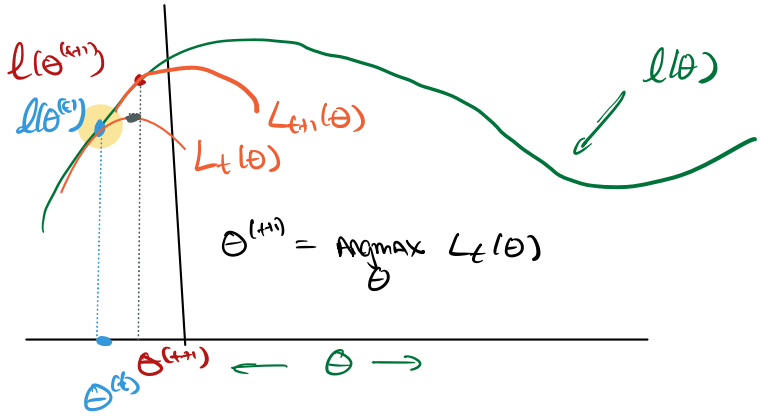
EM Algorithm as max likelihood

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

(θ: PARAMETERS; $x^{(i)}$: DATA)

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

Picture of Algorithm



$L_t(\theta) \le \ell(\theta)$ (lowerbound)

$L_t(\theta^{(t)}) = \ell(\theta^{(t)})$ (tight)

HOPE L_t easier to optimize than $\ell(\theta)$

Rough Algo (ITERATIVE)

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

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

How do we construct L_t?

(let's look at single bout term, $\log P(x;\theta)$)

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

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

$$= \log \mathop{E}_{z \sim Q(z)} \left[\frac{P(x, z; \theta)}{Q(z)} \right] \quad \text{(SYMBOL PUSHING)}$$

$$\geq \mathop{\mathbb{E}}_{z}\left[\log \frac{P(x,z;\theta)}{Q(z)}\right] \qquad \text{JENSEN! (log IS CONCAVE)}$$

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

Key step holds for any such Q: (*)

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

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!

$$P(x,z;\theta) = P(z|x;\theta)P(x;\theta)$$

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

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

NB: $Q(z)$ does depend on θ & x — we will select a $Q^{(i)}(z)$ for every point INDEPENDENTLY.

We define Evidenced-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 (*)

lower bound

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

WRAP UP

1. (E-STEP) $Q^{(i)}(z) = P(z^{(i)} | x^{(i)}; \theta)$ for $i = 1 \dots n$
2. (M-STEP) $\theta^{(t+1)} = \underset{\theta}{\text{ARGMAX}}\ L_t(\theta)$

in which $L_t(\theta) = \sum_{i=1}^{n} \text{ELBO}(x^{(i)}, Q^{(i)}; \theta)$

WHY DOES THIS TERMINATE? $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$

IS IT GLOBALLY OPTIMAL? (NOPE, SEE PICTURE)

WE DERIVED EM ALGORITHM AS TEAMS OF MLE.

EM for Mixture of Gaussians

Restate EM

(E-step) for $i = 1 \ldots n$ set $Q_i(z) = P(z^{(i)} | x^{(i)}, \theta^{(t)})$ → data & input params

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

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

Warm Up: Mixture of Gaussians. EM recovers our ad hoc algorithm

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

$z^{(i)} \sim \text{Multinomial}(\phi)$ $\phi_i \geq 0$ $\sum \phi_i = 1$ "in cluster j"

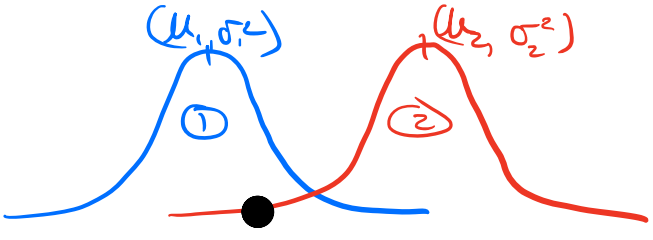
$x^{(i)} | z^{(i)} = j \sim N(\mu_j, \sigma_j^2)$ "cluster means"

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

What is EM here?

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

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



1. much more likely for (1) than (2)
2. But if we knew, $\phi_2 >> \phi_1$ maybe we'd think likely from (2)

Bayes Rule automates this reasoning

M-step: Compute Derivatives...

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

$f_i(\theta)$

↳ WRITE Θ FOR NOTATION ABOVE

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

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

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

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

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

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 CONSTRAINED $\sum_j \phi_j = 1, \ \phi_j \geq 0$, 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 $\sum \phi_j = 1$, $\quad \sum_j \phi_j = -\frac{1}{\lambda} \sum_{ij} w_j^{(i)} = -\frac{n}{\lambda}$

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

MESSAGE: EM RECOVERS GMM AUTOMATICALLY.

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

Factor Analysis

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

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

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

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

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

Problem in Gaussian likelihood

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

$|\Sigma| = 0$; Σ^{-1} is not defined.

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

Recall MLE for Gaussian

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

Equivalent

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

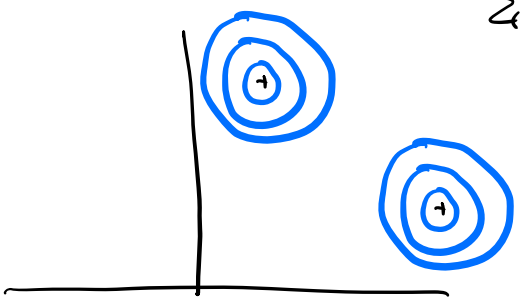
If Σ is full rank, $\nabla_\mu = \sum_{i=1}^{n} \Sigma^{-1}(x-\mu) = 0 \Rightarrow \mu = \frac{1}{n}\sum_i x^{(i)}$

WE'LL USE THIS AS PLUGIN BELOW.

Building Block 1

SUPPOSE INDEPENDENT AND IDENTICAL COVARIANCE

$\Sigma = \sigma^2 I$ (NB: PARAMETER TYING)



COVARIANCE "ARE CIRCLES"

WHAT IS MLE FOR Σ? $|\Sigma| = \sigma^{2d}$

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

Let $z = \sigma^2$ $\quad \min_z \frac{1}{z} C + d \log z$

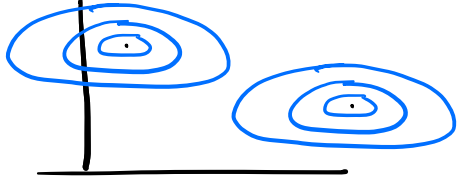
$$\Rightarrow \nabla_z = -z^{-2} C + \frac{nd}{z} = 0 \Rightarrow z = \frac{C}{nd}$$

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

"SUBTRACT MEAN AND SQUARE ALL ENTRIES."

Building Block 2

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



Axis Aligned ellipse

SET $z_i = \sigma_i^2$ (SAME IDEA AS ABOVE)

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

this is d Problems for EACH 1 dimension

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

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

Our Factor model

PARAMETERS

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

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

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

MODEL

$P(x, z) = P(x|z) P(z)$ z IS LATENT

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

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

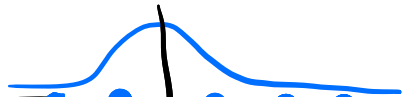
μ: MEAN IN the SPACE

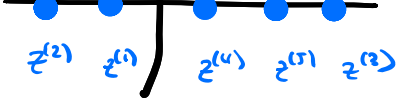
Λz: MAPS FROM SMALL LATENT SPACE TO LARGE SPACE

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

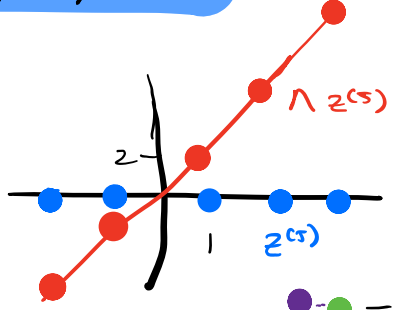
Ex: $d = 2$, $s = 1$, $n = 5$ $\quad x = \mu + \Lambda z + \epsilon$

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



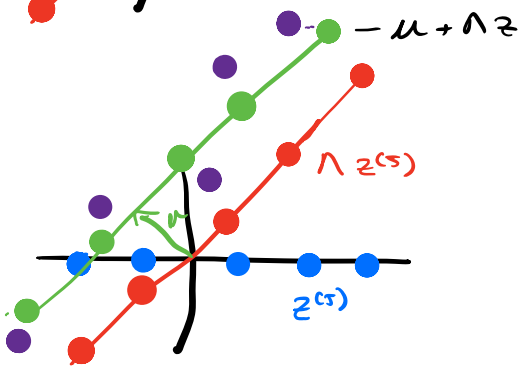


2. Suppose $\Lambda = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$



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

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

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

Notation is widely used and helpful.

FACT 1: $P(x_1) = \int_{x_2} P(x_1, x_2)$ — MARGINALIZATION

For Gaussians, $P(x_1) = N(\mu_1, \Sigma_{11})$ (not surprising)

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

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

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$ (MATRIX INVERSION LEMMA)

Proofs outline (happy to add)

Summary: Marginalization & conditioning Gaussian $\Rightarrow$ another Gaussian (closed)

we have formula for parameters.

Back to Factor Analysis $\quad x = \mu + \Lambda z + \epsilon$

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

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

what is Σ?

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

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

($\mathbb{E}[z\epsilon^T] \to 0$)

$$= \Lambda^T$$

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

$$\Sigma_{22} = \mathbb{E}[(x-\mu)(x-\mu)^T]$$

$$= \mathbb{E}[(\Lambda z + \epsilon)(\Lambda z + \epsilon)^T]$$

$$= \mathbb{E}[\Lambda z z^T \Lambda^T] + \mathbb{E}[\epsilon\epsilon^T]$$

$$= \Lambda\Lambda^T + \Phi$$

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

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