Lecture-15

⊛ EM algorithm and mixture of Gaussian model:-

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

$z^{(i)} \sim$ Multinomial (ϕ_j)

$[P(z^{(i)} = j) = \phi_j]$.

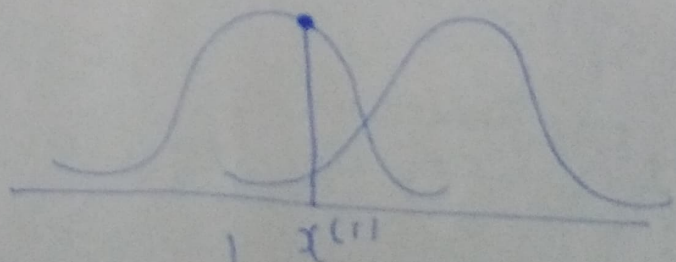
$x^{(i)} | z^{(i)} \sim N(\mu_j, \Sigma_j)$

→ Deriving:-

① E-step:-

$w_j^{(i)} = Q_i(z^{(i)} = j) = P(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$.

" $P(z^{(i)} = j)$



↳ $w_j^{(i)}$ from where $x^{(i)}$ is likely to come (which gaussian)

② M-Step:-

$$\max_{\phi, \mu, \Sigma} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \left(\frac{P(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)}{Q_i(z^{(i)})} \right)$$

$Q_i(z^{(i)})$

$$= \sum_i \sum_j \boxed{w_j^{(i)}} \log \frac{1}{(2\pi)^{1/2}|\Sigma_j|^{1/2}} \cdot \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)\Sigma_j^{-1}(x^{(i)}-\mu \right. \quad \left. P(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)\right.$$

ϕ_j

$\boxed{w_j^{(i)}}$

$Q_i(z^{(i)})$.

→ This formula

$$\nabla_{\mu_j}(\dots) = 0 \Longrightarrow \mu_j = \boxed{\frac{\sum_i w_j^{(i)} x^{(i)}}{\sum_i w_j^{(i)}}}$$

→ Remeber $w_j^{(i)}$ is the strength with which $x^{(i)}$ is a signal to Gaussian j ($P(z_j^{(i)} | x^{(i)}; \dots)$).

↳ Formula for mixture of gaussian model derived from EM algorithm.

Same Formula above

$$\nabla_{\phi}(\dots) = 0 \Rightarrow \phi_j = \frac{1}{m}\sum_i w_j^{(i)}$$

Same Formula above

$$\nabla_{\Sigma}(\dots) = 0 \Rightarrow \Sigma_j = \frac{\sum_{i=1}^{m}(x-\mu_j)^T(x-\mu_j)}{\sum_{i=1}^{m} w_j^{(i)}}$$

→ Factor analysis model:-

→ In this cases $z^{(i)}$ is continous random variable rather than discrete random variable.

→ • One equvilent view of EM-algorithm:-

$$J(\theta, Q) = \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \left(\frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right).$$

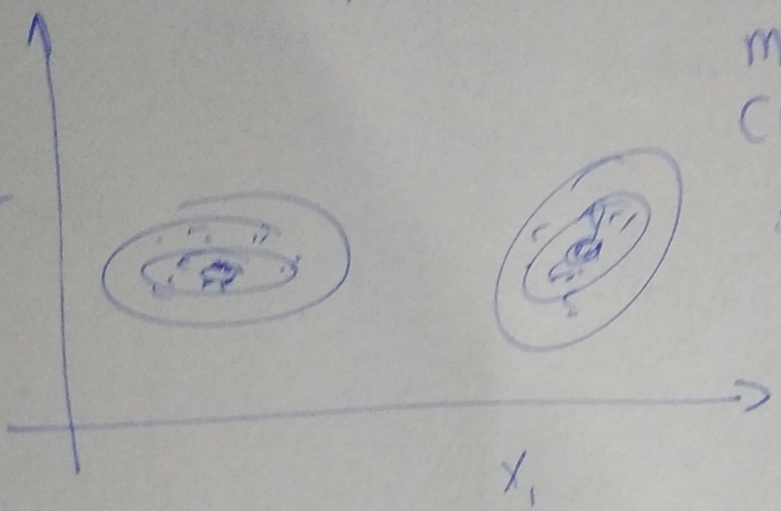
⊙ We know $\ell(\theta) \geq J(\theta, Q)$ for any θ, Q.

E-step ⇒ Maximize J w.r.t Q
M-step ⇒ Maximime J w.r.t. θ.

→ This procedure is called coordinate ascent because you are maximizing $J(\theta, Q)$ w.r.t to one parameter at a time. (EM is a coordinate ascent algorithm).

→ What is Factor Analysis and comparsion with mixture of gaussian?

$n = 2$ $m = 100$



$m >> n$
(Mixture of gaussian seems good).

$m \geq n$, or $m << n$
(Factor Analysis is better in this case as compare to mixture of gaussian).

Say $m = 30,\ n = 100$ (100 dimension with 30 examples).

↳ Problem with applying Gaussian is as follow in this case:

→ You can model it as single gaussian ($m = 30,\ n = 100$).

$x \sim \mathcal{N}(\mu, \Sigma)$.

MLE:

$$\mu = \frac{1}{m} \sum_i x^{(i)}$$

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

if $m \leq n$ then Σ matrix will be singular / non-invert

(*)

$$\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\cdots)\Sigma^{-1}(\cdots)\right)$$

Being singular → this term will be zero.

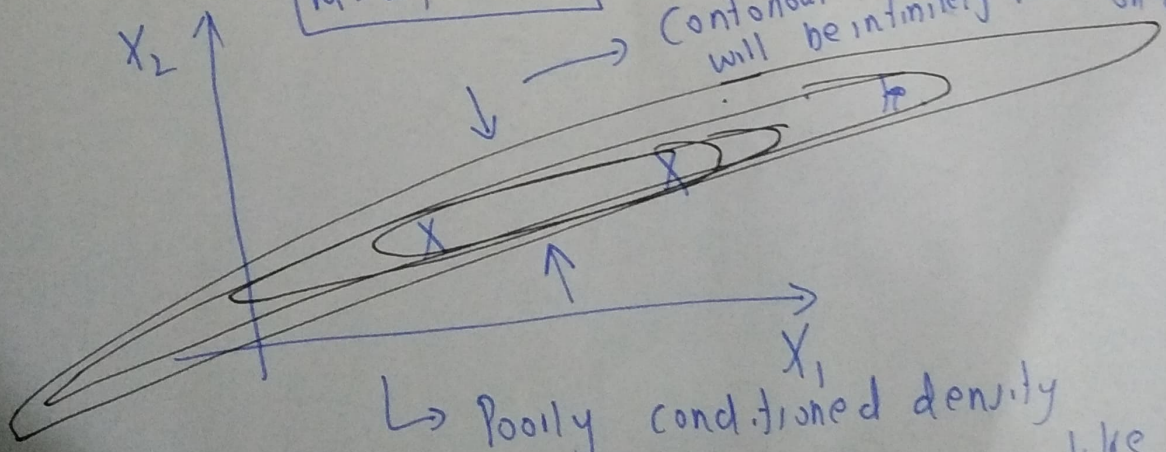
Being singular → this will be undefined.

(when $m << n$).

(*) So the problem in applying mixture of gaussian model is that you will end up with Σ that is singular hence non-invertible.

→ What it looks like graphically:

$M = 2,\ n = 2$



↳ Poorly conditioned density mall as the contonour is like a line segment (Not a good model).

⇝ **Factor Analysis** to the rescue in this case: $(m \ll n)$

← [→ 100 psychological attributes
  → 30 persons

one case of $m \ll n$.

100 ↑ ↓

$$\begin{bmatrix} & \\ & \\ & \end{bmatrix}_{X}$$

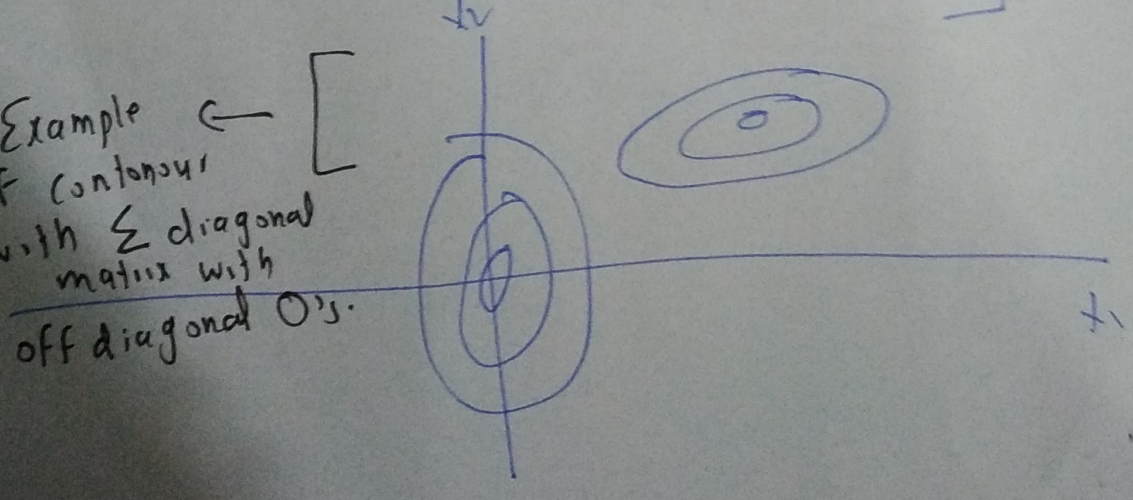
Modelling $p(x)$.

⇝ Some other alternatives for the case $m \ll n$:-

⊙ **Option 1:** Constraint Σ to be diagonal

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \ddots \\ & & & \sigma_n^2 \end{bmatrix}$$

Example of contours with Σ diagonal matrix with off diagonal 0's. ←



• With Σ as diagonal matrix:

↳ MLE:

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

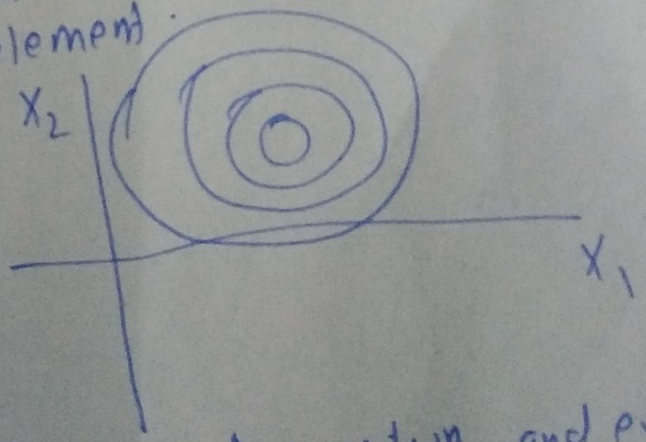
→ Σ has n diagonal parameters

→ The problem with this option is that it assumes that features are uncorrelated.

⊙ Option 2:- Constraint Σ to be $\Sigma = \sigma^2$]

i.e $\Sigma = \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix}$

→ Constraining Σ matrix not only to be diagonal but also have the same diagonal element.



→ Very Strong Assumption and even

$$\sigma^2 = \frac{1}{m}\frac{1}{n}\sum_i \sum_j (x_j^{(i)} - \mu_j^{(i)})^2$$

→ We won't use any of this options because of their strong assumptions but instead go with **Factor Analysis** which capture correlation and does not go in problem of invertibility.

→ Factor Analysis explanation:

$P(x, z) = P(x|z) \cdot P(z)$

z is a hidden variable

$d = 3$
$n = 10$,
$m = 30$.

$z \sim \mathcal{N}(0, I)$. $z \in \mathbb{R}^d$ $(d < n)$.

$$x = \mu + \Lambda z + \varepsilon \quad (1)$$

where $\varepsilon \sim \mathcal{N}(0, \Psi)$

Parameters:

$\mu \in \mathbb{R}^n$ $\Lambda \in \mathbb{R}^{n \times d}$ $\Psi \in \mathbb{R}^{n \times n}$ diagonal.

→ The expression (1) can be equivalently written as:

$$x|z \sim \mathcal{N}(\mu + \Lambda z, \Psi)$$

→ z is basically the power full which we assume is deriving the features of x.

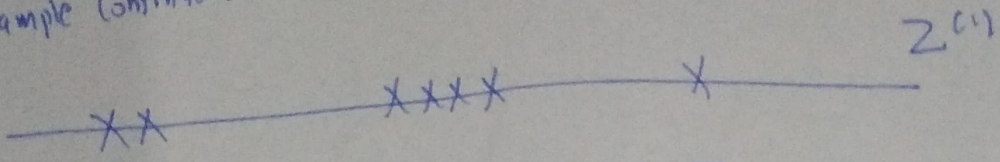
→ So, x is basically the linear combination of mean and the power ful which we assume is deriving it.

→ Ψ being a diagonal covriance matrix assumes that noise added at each feature of x is independent of the noise of other feature. (Each feature has independent noise).

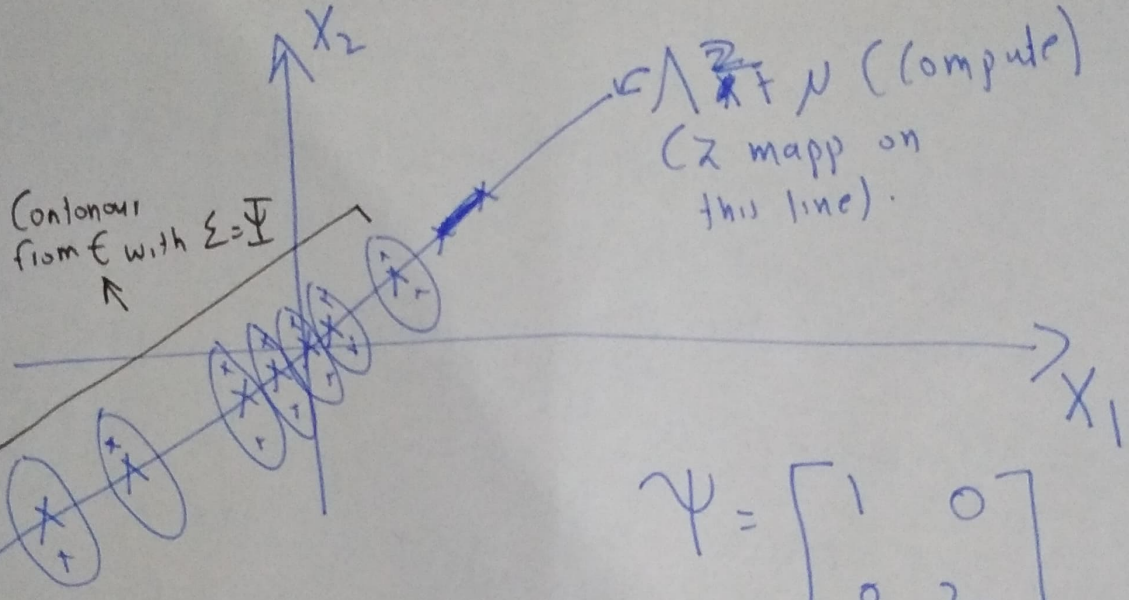
~> Examples of data which factor analysis can model:-

① $z \in \mathbb{R}^1$ $x \in \mathbb{R}^2$ $d=1$, $n=2$, $m=7$.

Example Continued



Say $\Lambda = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ $\quad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

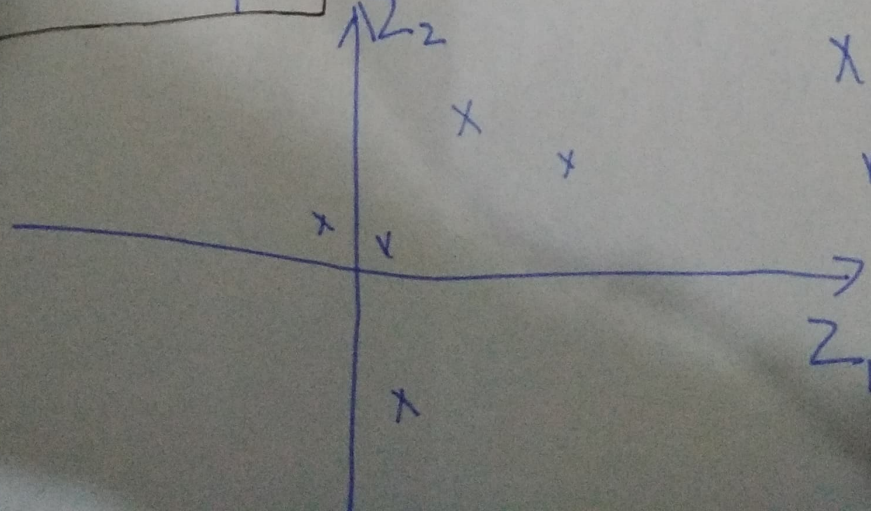


$\Psi = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$

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

→ Samples from each of the Contour (Normal distribution $x|z$).

→ Another Example:



$z \in \mathbb{R}^2$ $\quad d=2$

$x \in \mathbb{R}^3$ $\quad n=3$

$m=5$ $\quad m=5$

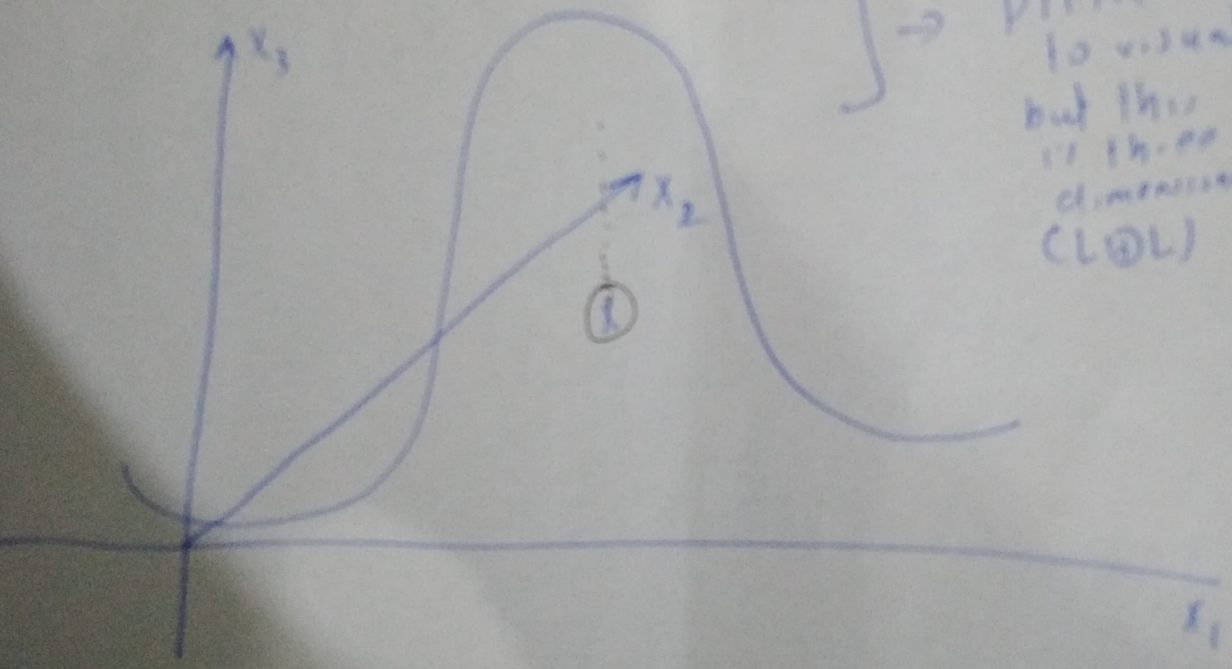
→ Now we will compute → Mapping to higher dimensional space

$\Lambda z + \mu$

3x2 2x1 3x1

→ Now we are going to take 2 dimensional space and map to higher dimensional space.

→ Now each point in three dimensional i.e all five points would have a gaussian bump:



] → Difficult to visualize but this is three dimensional. (L@L)

⑧ Factor analysis can take very high dimensional data and model the data roughly lying on lower dimension subspace with a little bit of fuzz on that lower dimension

(X)

How to derive E-step and M-step of Factor Analysis model (Going to be tricky):

→ Multivariate Gaussian: (Proposition)

Partitioned Vector!!!

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{matrix} \updownarrow r \\ \updownarrow s \end{matrix} \qquad X_1 \in \mathbb{R}^r, X_2 \in \mathbb{R}^s \qquad X \in \mathbb{R}^{r+s}$$

So:

$$X \sim \mathcal{N}(\mu, \Sigma) \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \begin{matrix} \updownarrow r \\ \updownarrow s \end{matrix} \rightarrow \text{Partitioned}$$

$$\Sigma \overset{\begin{matrix} r & s \\ \longleftrightarrow & \longleftrightarrow \end{matrix}}{=} \begin{matrix} r \updownarrow \\ s \updownarrow \end{matrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Partitioned matrix

Marginal $(P(X_1)) = ?$

$$P(x) = P(x_1, x_2)$$

$$\int_{x_2} P(x_1, x_2)\, dx_2 = P(x_1)$$

$$\int_{x_2} \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right)$$

→ Of-Course not going to integrate but results will be as follows:-

$$P(X_1) = X \sim N(\mu_1, \Sigma_{11})$$

(Not shocking)

Conditional: $P(X_1 | X_2) = ?$

$$\frac{P(X_1, X_2)}{P(X_2)} = P(X_1 | X_2)$$

$$\leftarrow X_1 | X_2 \sim N(\mu_{1|2}, \Sigma_{1|2})$$

Simplifying the above expression

where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

(*) Using these above properties of multivariate gaussian, we will ~~we EM~~ go through how you derive EM algorithm for factor analysis :-

① Derive $P(X, Z)$

← Stacking them up $\left[\begin{pmatrix} Z \\ X \end{pmatrix} \sim \mathcal{N}(\mu_{X,Z}, \Sigma)\right.$

$$\left.\begin{aligned} Z &\sim \mathcal{N}(0, I) \\ X &= \mu + \Lambda Z + \varepsilon \end{aligned}\right\} \rightarrow \text{Orginal defination}$$

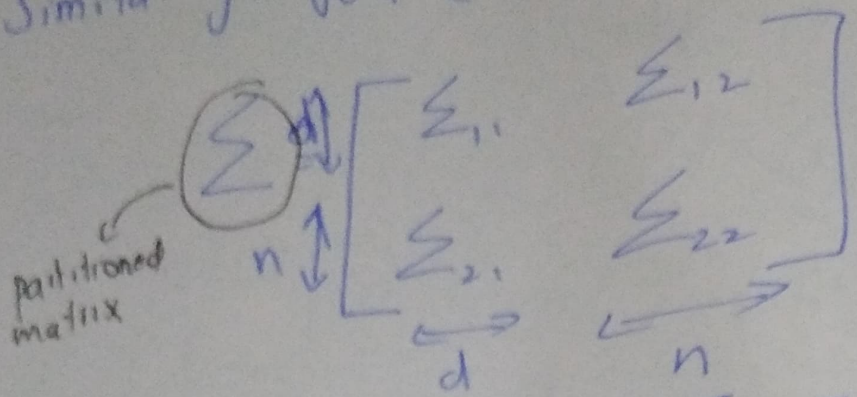
$$E_Z = 0 \,.\, E_X = \mu + \Lambda \overset{\downarrow 0}{Z} + \overset{\downarrow 0}{\varepsilon}$$

$$= \mu$$

therefore

$$\mu_{X,Z} = \begin{bmatrix} 0 \\ \mu \end{bmatrix} \begin{matrix} \updownarrow d \\ \updownarrow n \end{matrix}.$$

Similarly: You can compute:-



$$\Sigma = \begin{bmatrix} E(z-E_z)(z-E_z)^T & E(z-E_z)(x-E_x)^T \\ E(x-E_x)(z-E_z) & E(x-E_x)(x-E_x)^T \end{bmatrix}$$

→ From above you can derive $\Sigma_{11}, \Sigma_{12}, \Sigma_{21}, \Sigma_{22}$ one by one:- Lets do Σ_{22} (rest are similar)

$$\Sigma_{22} \overset{E}{=} (x-E_x)(x-E_x)^T$$

$$= E\left[(\Lambda z + \mu + \epsilon - \mu)(\Lambda z + \mu + \epsilon - \mu)^T\right] \quad \text{(From previous)}$$

= Doing the Quadratic expansion

$$= E\left[\Lambda z z^T \Lambda^T + \Lambda z \epsilon^T + \epsilon z^T \Lambda^T + \epsilon \epsilon^T\right]$$

(the $\Lambda z \epsilon^T$ and $\epsilon z^T \Lambda^T$ terms → 0)

$$= E\left[\Lambda z z^T \Lambda^T + \epsilon \epsilon^T\right]$$

$$= E[\Lambda z z^T \Lambda^T] + E[\varepsilon \varepsilon^T]$$

$$= \Lambda E[z z^T] \Lambda^T + \Psi$$

$\Sigma_{22} = \Lambda \Lambda^T + \Psi$ (Lower Right Block).

→ Similary other blocks are as follow::

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

~So we have figure out

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix} \right)$$

$P(x^{(i)})$ ↳ You could write down $P(x^{(i)})$ and find derivative of log-likelihood of above expression then set it to zero and then find the parameter but turns out there is no closed form solution.

→ In order to fit we are going to resort to EM:

→ E-step

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

(But $z^{(i)}$ is a continous density).

So

$$z^{(i)} | x^{(i)} \sim \mathcal{N}(\mu_{z^{(i)}|x^{(i)}}, \Sigma_{z^{(i)}|x^{(i)}})$$

Through inf where

$$\mu_{z^{(i)}|x^{(i)}} = \vec{0} + \Lambda^T (\Lambda\Lambda^T + \Psi)^{-1} (x^{(i)} - \mu)$$

$$\Sigma_{z^{(i)}|x^{(i)}} = I - \Lambda^T (\Lambda\Lambda^T + \Psi)^{-1} \Lambda$$

From previous equation

↳ In e-step you compute these.

→ M-step:

→ Derivation of M-step is quite long & complicated but I would mention just a key algebric trick you need to use when deriving M-step.

$$Q_i(z^{(i)}) = \frac{1}{2(\pi)^{d/2}|\Sigma|^{1/2}(\dots)} \exp\left(-\frac{1}{2}(\dots)\right)$$

(from above expressions) (from above expressions)

$$= \int_{z^{(i)}} Q_i(z^{(i)}) z^{(i)} dz^{(i)}$$ (There will be steps where you need to compute this).

$$= \int_{z^{(i)}} \frac{1}{2\pi^{d/2}|\Sigma|} \exp\left(-\frac{1}{2}(\dots)\right) z^{(i)} dz^{(i)}$$ (There is much simple

⇒ There much simple way to compute this integral:-

$$\int_{z^{(i)}} \underbrace{Q_i(z^{(i)}) z^{(i)} dz^{(i)}}_{\text{Integral this is expected value}}$$

$$E_{z^{(i)} \sim Q_i}[z^{(i)}] = \boxed{\mu_{z^{(i)}|x^{(i)}}}$$ → We have already computed this

→ M-step after derivations would be as follow

$$\theta := \arg\max_{\theta} \sum_i \int_{z^{(i)}} Q(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)})}{Q_i(z^{(i)})} dz^{(i)}$$

$$= \sum_i E_{z^{(i)} \sim Q(z^{(i)})} \left[\log \frac{P(x^{(i)}, z^{(i)})}{Q(z^{(i)})} \right]$$

Plug in gaussian density.

After Plugging in gaussian density and then you take the derivate of above expression with respect to the parameter set to zero and find the parameters i.e

$$\nabla_{\mu}(\dots) = 0 \quad \text{(and solve)}$$