E.M. Algorithm

MLE in presence of latent variables

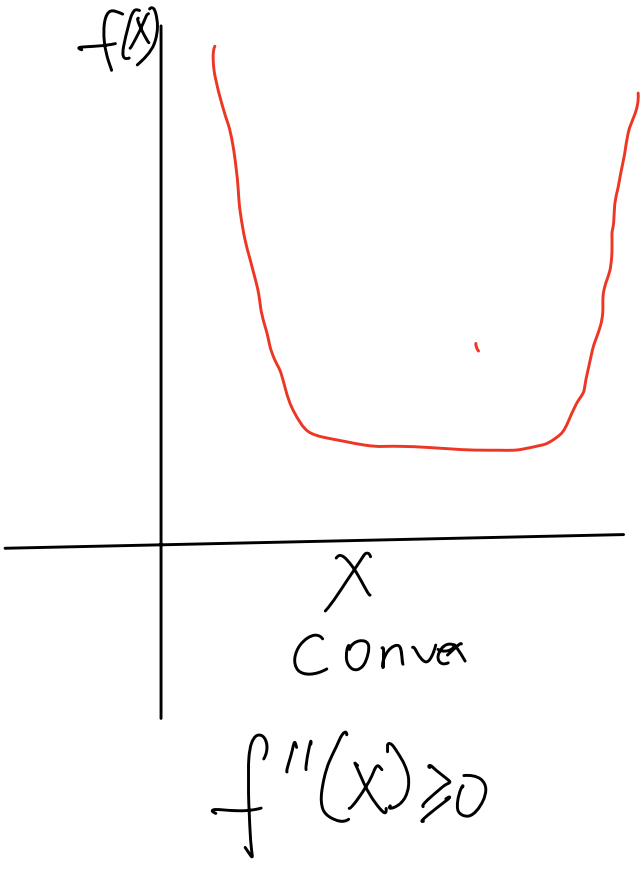
True Model : $p(x, z; \theta)$

when z is unobserved.

Maximization $l(\theta) = \log p(x; \theta)$

Strictly Convex

$f''(x) > 0$



Conve

$f''(x) \geq 0$

Jensen Inequality

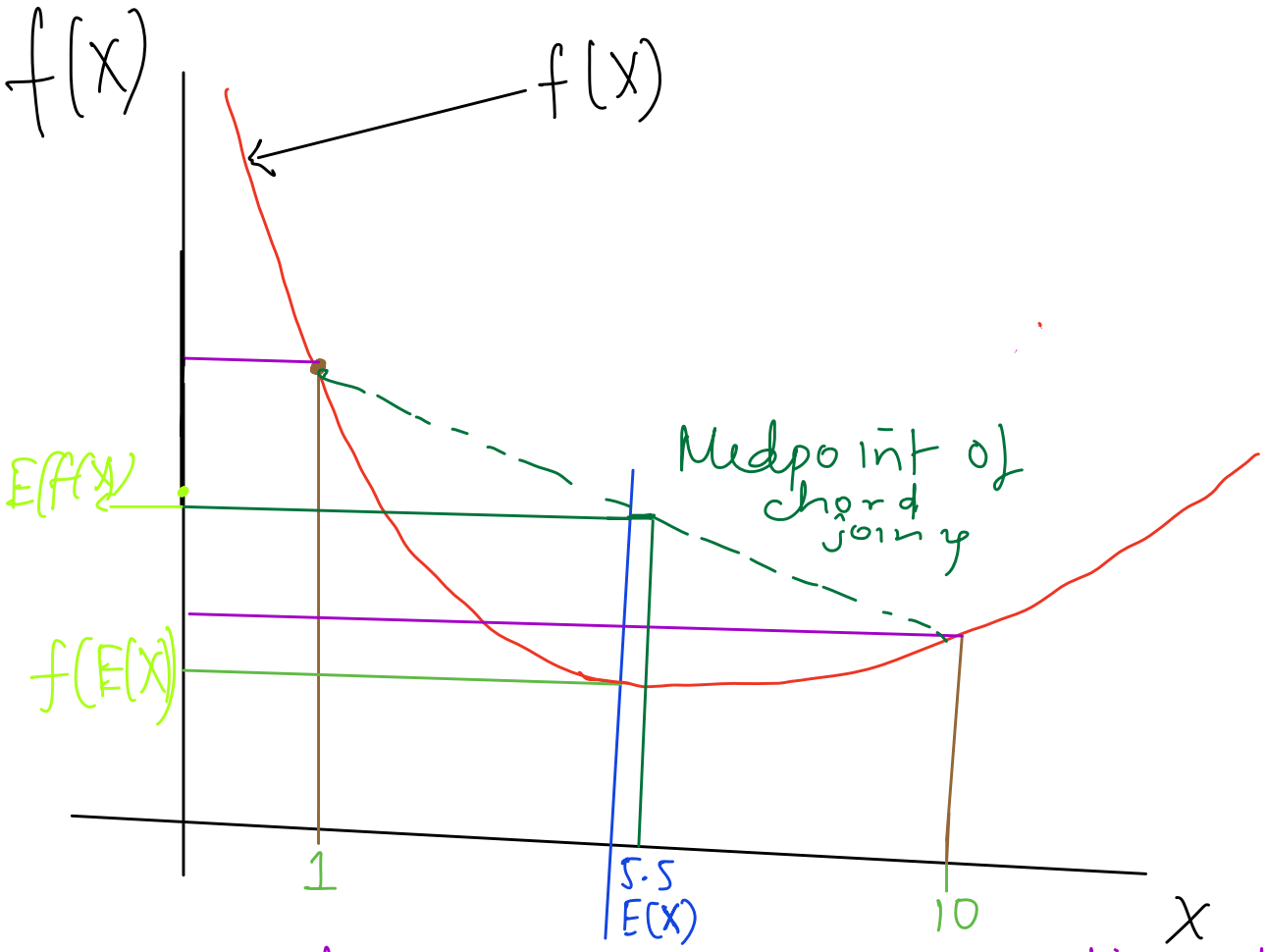
$E[f(X)] \geq f(E[X])$ where f is convex

If f is **strictly convex**, then,

if $E[f(X)] = f(E(X))$

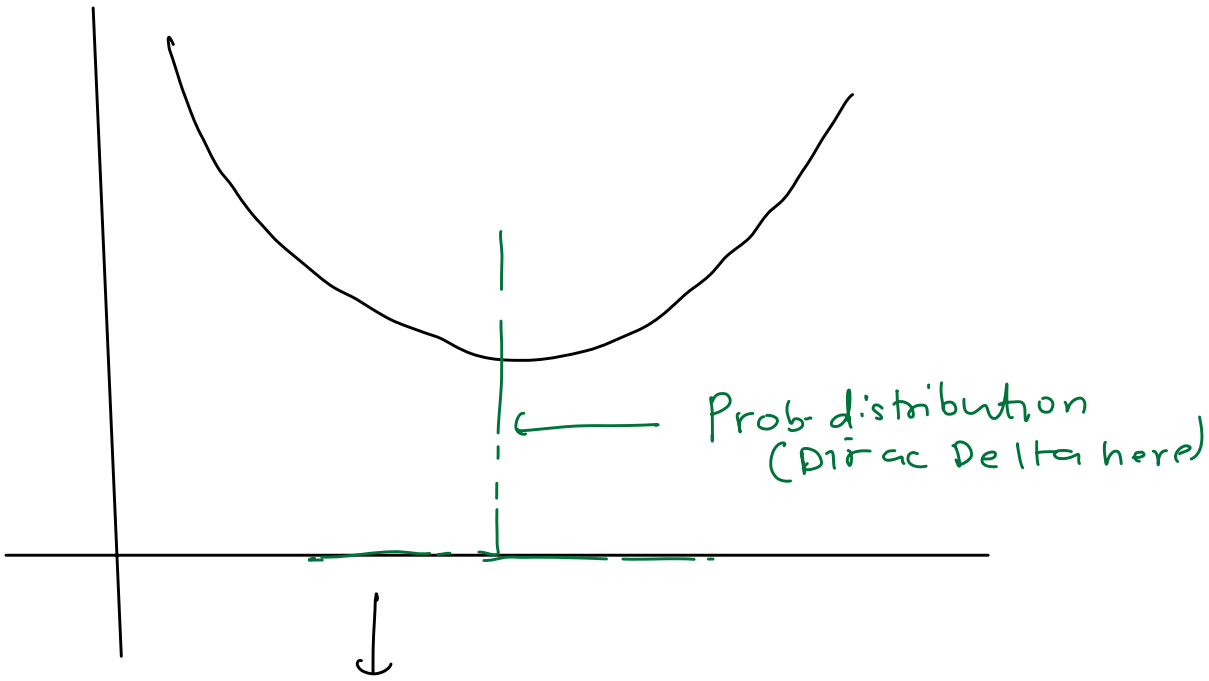
then $X = E(X)$ w.p. 1

i.e. $X \to$ constant



x is drawn from probability dist^n

p(x) which is discrete i.e. can be 1 or 10 with $p = 1/2$



$E(f(x)) = f(E(x))$

as x evaluates to same value each time & length of chord connecting is zero

$\rightarrow f''(x) = 0$ so both convex & concave

Convex	Concave	Strict
x^2	$-x^2$	✓
$mx+c$	$mx+c$	X
e^x	$-e^x$	✓
$-\log x$	$\log x$	✓

If f is concave

eg. $f(x) = \log(x)$

$$E[\log(X)] \leq \log E[X]$$

Goal is max

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

z → unobserved

$$\max \log p(x; \theta)$$

$$\log \sum_z p(x, z; \theta)$$

$$\log \sum_z \frac{Q(z)\, p(x, z; \theta)}{Q(z)}$$

[where $Q(z)$ is arbitrary prob. dist. over z)

$Q(z) > 0$ for all z

$$\log E_{z \sim Q}\left[\frac{p(x, z; \theta)}{Q(z)}\right]$$

$$\geq \mathbb{E}_{z \sim q}\left[\log \frac{P(x, z; \theta)}{q(z)}\right]$$

$$:= ELBO(x; q, \theta)$$

Evidence lower bound

so if we find max ELBO (i.e. setting ↓ good θ), then $\log p(x)$ will be greater

Are there cases $\log p(x; \theta) = ELBO(x; q, \theta)$?

→ Yes when $\frac{P(x, z; \theta)}{q(z)}$ → constant

constant wrt z

$$q(z) = \frac{1}{C} p(x, z; \theta)$$

$$q(z) \propto p(x, z; \theta)$$

Since $q(z)$ is prob. distribution
it should be summed to 1

$$q(z) = \frac{p(x, z; \theta)}{\sum_z p(x, z; \theta)}$$

$$= \frac{p(x, z; \theta)}{p(x; \theta)}$$

$$= p(z | x; \theta)$$

$$q(z) = p(z | x; \theta) \Rightarrow \log p(x) = ELBO$$

EM Algo

Random θ

E Step

For each i, set

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

M Step

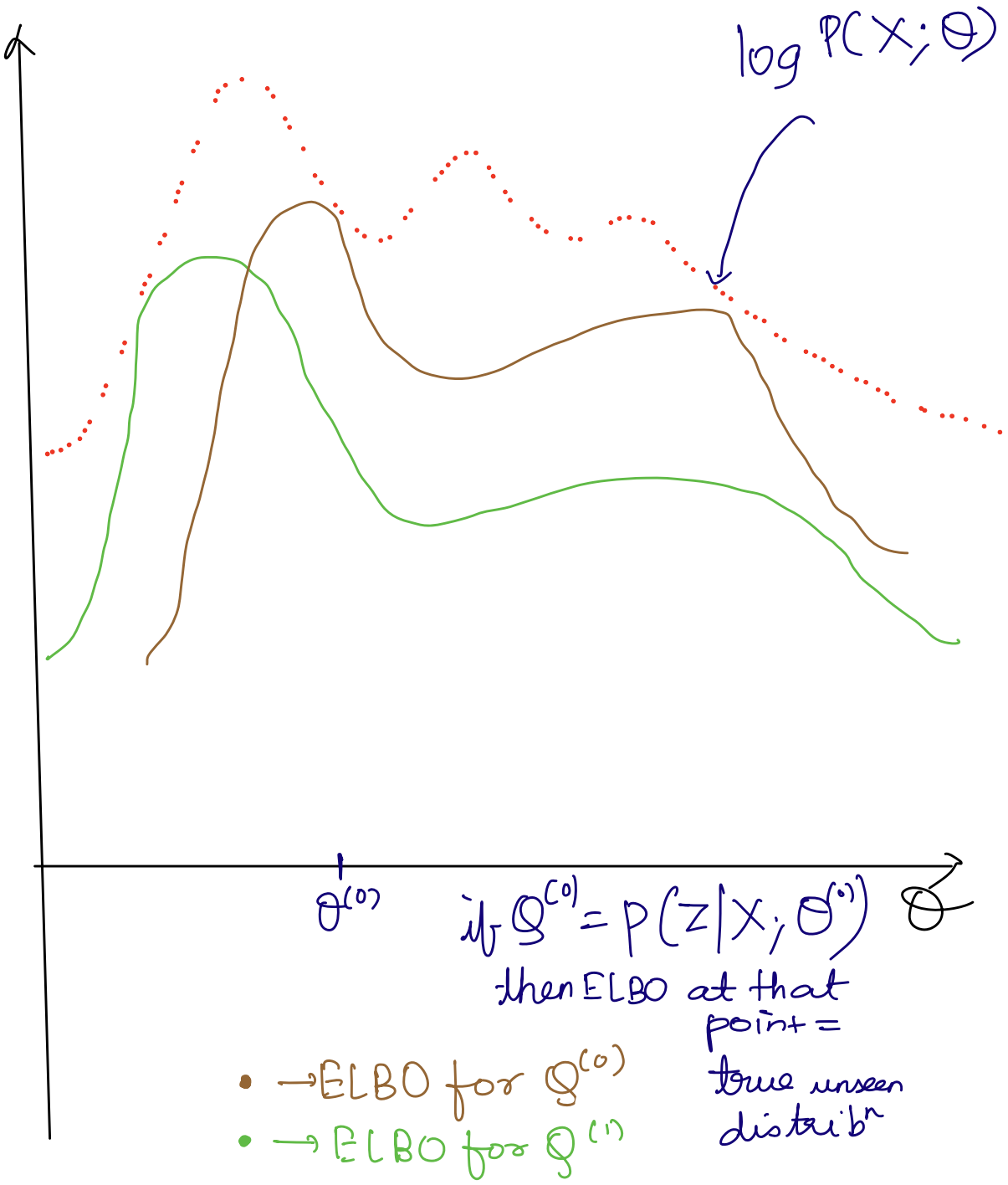
Set

$$\theta := \arg\max_{\theta} \sum_{i=1}^{n} ELBO(x^{(i)}; Q_i, \theta)$$

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

$Q \rightarrow$ is constant

Although it has θ, we assume it to be from previter.

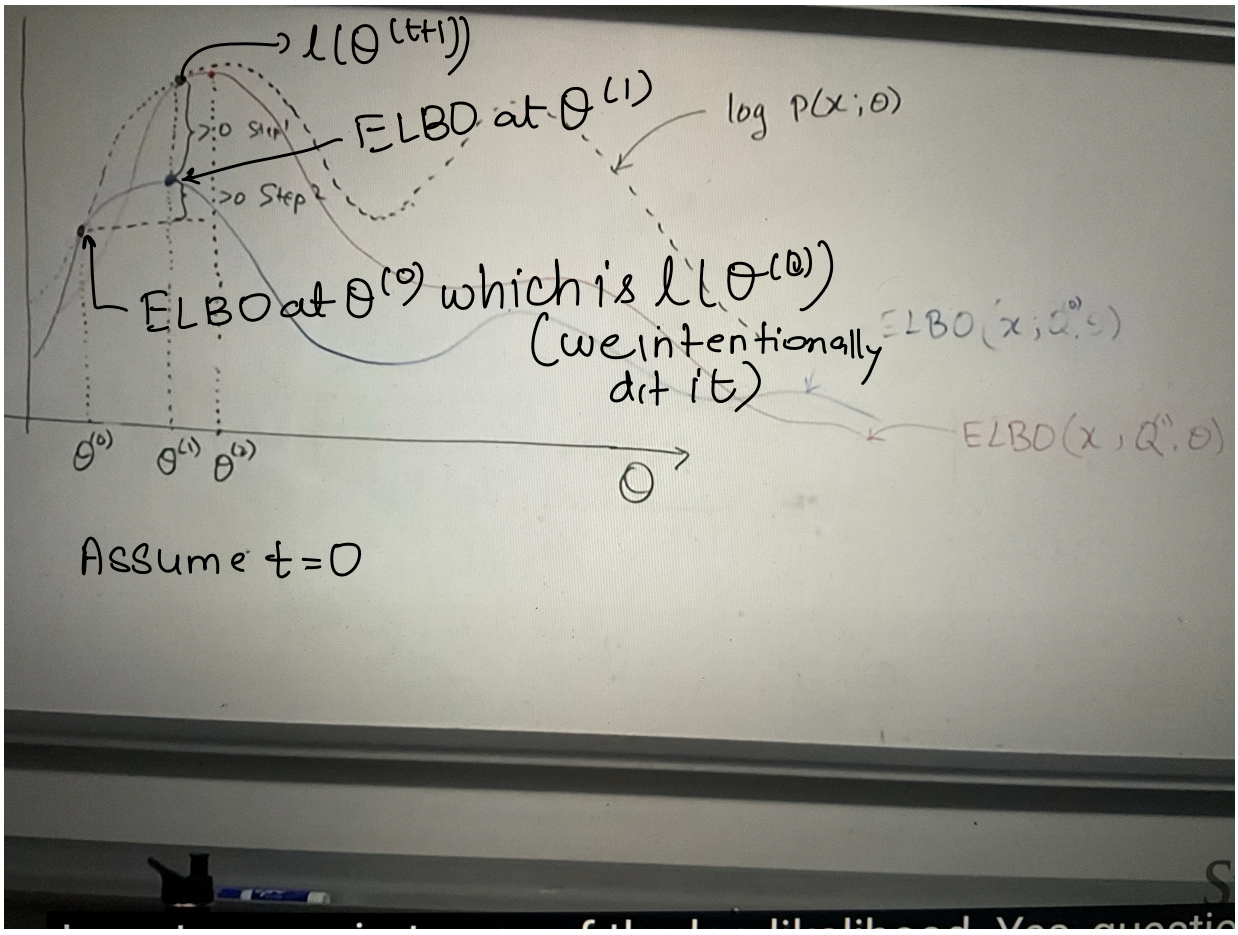


using ELBO, we indirectly max log p(X;θ)
& find θ which is a local optima

$$\text{Q}[z^{(3)}] = \begin{bmatrix} 0.1 \\ \vdots \\ 0.9 \\ 0.01 \end{bmatrix} \begin{matrix} \text{- Cluster 1} \\ \\ \\ \text{- Cluster 12} \end{matrix}$$

3^{rd} ex

Proof that EM converges



For every t

$$\ell(\theta) = \log P(x;\theta)$$

$$\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$$

$$\ell(\theta^{(t+1)}) \geq ELBO(x; Q^{(t)}, \theta^{(t+1)}) \rightarrow \text{Jensen}$$

$$\geq ELBO(x; Q^{(t)}, \theta^{(t)}) \rightarrow \text{M-step}$$

$$= \ell(\theta^{(t)}) \rightarrow \text{Jensen}$$

GMM via EM

① First write out model $P(X, Z; \theta)$

$Z \sim \text{Multinomial}$

$X|Z \sim N(\mu_Z, \Sigma_Z)$

② Then write what is latent & what are observed

x - Evidence

z - latent

[Unknown $z \rightarrow$ example specific

Unknown parameters $\rightarrow$ global]

③ In E-step, we estimate z, for each eg. holding parameter fixed

④ In M-step, we update parameters holding z fixed

E-Step

For each i,

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

$$= \frac{p(x|z)\,p(z)}{p(x)}$$

$$= \frac{p(x|z)\,p(z)}{\sum_z p(x|z=j)\,p(z=j)}$$

$$= \frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left\{(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right\} \times \phi_j}{\sum_j (\quad)}$$

M-Step

$$\mu, \Sigma, \phi = \arg\max_{\mu,\Sigma,\phi} \sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} \log \frac{p(x^{(i)}, z^{(i)}; \mu, \Sigma, \phi)}{w_j^{(i)}}$$

$$= \arg\max_{\mu,\Sigma,\phi} \sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} \log \frac{p(x^{(i)} | z^{(i)}=j)\, p(z^{(i)}=j)}{w_j^{(i)}}$$

[Here for M-step hold w constant]

Solve it then take gradient solve for 0 then you get G.M.M. update rules

[k diff μ, Σ]

[If we take mode instead of expection, we get K-Means]

FACTOR ANALYSIS

$x^{(i)} \in \mathbb{R}^d$

$S = \{x^{(i)}\}_{i=1}^n \qquad d >> n > k$

$\Sigma \in \mathbb{R}^{d \times d}$

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

$\Sigma \to$ rank n

but actually it is rank d

$k < n$

k-low dimensional subspace

$z^{(i)} \sim N(0, I) \qquad I - k$ dim

$z^{(i)} \in \mathbb{R}^k$

$x^{(i)} | z^{(i)} \sim N(\mu + Lz, \Psi)$

$L \in \mathbb{R}^{d \times k}$

$\Psi \in \mathbb{R}^{d \times d} \to$ diagonal matrix

[In GMM, each x is discrete and has seperate μ, Σ
Here μ, Σ constant across eqs]

Here, we are not doing classification

We want to find a lower dimensional subspace in which Z resides and find X for each z

Suppose there are d temp. sensor & we get temp values at time t

$$x^{(t)} \in \mathbb{R}^d$$

d is very high & all values are not independent. In one room, value by sensor is almost same.

So we try to estimate a few k factors which decide temp. & then make inf.

$$z^{(t)} \in \mathbb{R}^k$$

Model

$$Z \sim N(0, I)$$
$$X|Z \sim N(\mu + LZ, \Psi)$$

$$Z \sim N(0, I)$$
$$\varepsilon \sim N(0, \Psi)$$
$$X = \mu + LZ + \varepsilon \rightarrow \text{Location scale Transform}$$

$$E(X) = \mu$$

Joint distribution

$$\begin{bmatrix} Z \\ X \end{bmatrix} \sim N(\mu_{2x}, \Sigma)$$

$$\mu_{2x} = \begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \Sigma_Z & \Sigma_{ZX} \\ \Sigma_{ZX}^T & \Sigma_X \end{bmatrix} = \begin{bmatrix} I & L^T \\ L & LL^T + \Psi \end{bmatrix}$$

Observe - x

Latent - z

Parameters : μ, L, Ψ

NOTE:

$$\ell(\mu, L, \Psi) = \log p(x; \_\_)$$

$$\vdots$$

$$X \sim N(\mu, LL^T + \Psi)$$

L: low rank

Ψ: diagonal matrix

$$\log P(X) = \frac{1}{(2\pi)^{d/2} |LL^T + \Psi|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T (LL^T + \Psi)^{-1}(x-\mu)\right)$$

This is correct but you won't get closed form soln here.

E Step

$$p(z|x) \sim N\Big[\, L^T(LL^T+\Psi)^{-1}(x-\mu),\ \underbrace{I - L^T(LL^T+\Psi)^{-1}L}_{\text{schurr complement}} \,\Big]$$

$$Q_i(z^{(i)}) = N\Big[\ \text{''}\ ,\ \text{''}\ \Big]$$

M-Step

$$\mu, L, \Psi = \arg\max_{\mu,L,\Psi} \sum_{i=1}^{\wedge} \int_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \mu, L, \Psi)}{Q_i(z^{(i)})}$$

$$= \sum_{i=1}^{\wedge} E\left[\log \frac{P(x^{(i)}, z^{(i)}; \mu, L, \Psi)}{Q_i(z^{(i)})}\right]$$

$$= \sum_{i=1}^{\wedge} E\left[\log p(x^{(i)}|z^{(i)}) + \log p(z^{(i)}) - \log Q_i(z^{(i)})\right]$$

constant always

In this case z holds no inf about paramaters

$$= \sum_{i=1}^{n} \mathbb{E}_{z^{(i)} \sim Q_i} \left[\log \frac{1}{(2\pi)^{d/2} |\Psi|^{1/2}} \exp\left\{ -\frac{1}{2} (x^{(i)} - \mu - L z^{(i)}) \Psi^{-1} (x^{(i)} - \mu - L z^{(i)}) \right\} \right]$$

$$= \sum_{i=1}^{n} \mathbb{E}_{z^{(i)} \sim Q_i} \left[\log \frac{1}{(2\pi)^{1/2}} - \frac{1}{2} \log |\Psi| - \frac{1}{2} (x^{(i)} - \mu - L z^{(i)}) \Psi^{-1} (x^{(i)} - \mu - L z^{(i)}) \right]$$

$\frac{1}{(2\pi)^{1/2}}$ → no z or no parameters

$-\frac{1}{2}\log|\Psi|$ → comes out of exp

$-\frac{1}{2}(x^{(i)} - \mu - L z^{(i)}) \Psi^{-1} (x^{(i)} - \mu - L z^{(i)})$ → expand it

$$L = \sum_{i=1}^{n} (x^{(i)} - \mu) \mu_{z|x}^{T} \left(\sum_{i=1}^{n} \mu_{z|x} \mu_{z|x}^{T} + \Sigma_{z|x} \right)$$

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

$$\underline{\Phi} = \frac{1}{n} \sum_{i=1}^{n} x^{(i)} x^{(i)T} - x^{(i)} \mu_{z|x}^{T} L^{T} - L \mu_{z|x} x^{(i)T} + L \left(\mu_{z|x} \mu_{z|x}^{T} + \Sigma_{z|x} \right) L^{T}$$

$\Psi_{ii} = \underrightarrow{\Phi}_{ii}$

FACTOR ANALYSIS

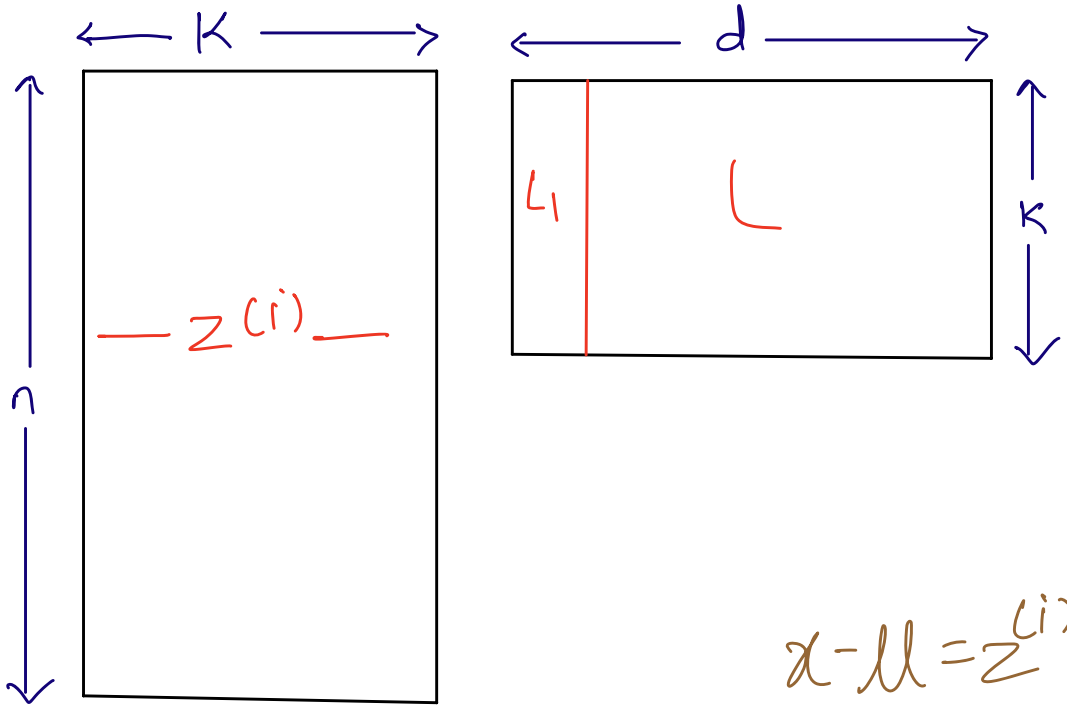
Model: $z \sim N(0, I) \in \mathbb{R}^k$

$x|z \sim N(\mu + Lz, \Psi) \in \mathbb{R}^d$

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

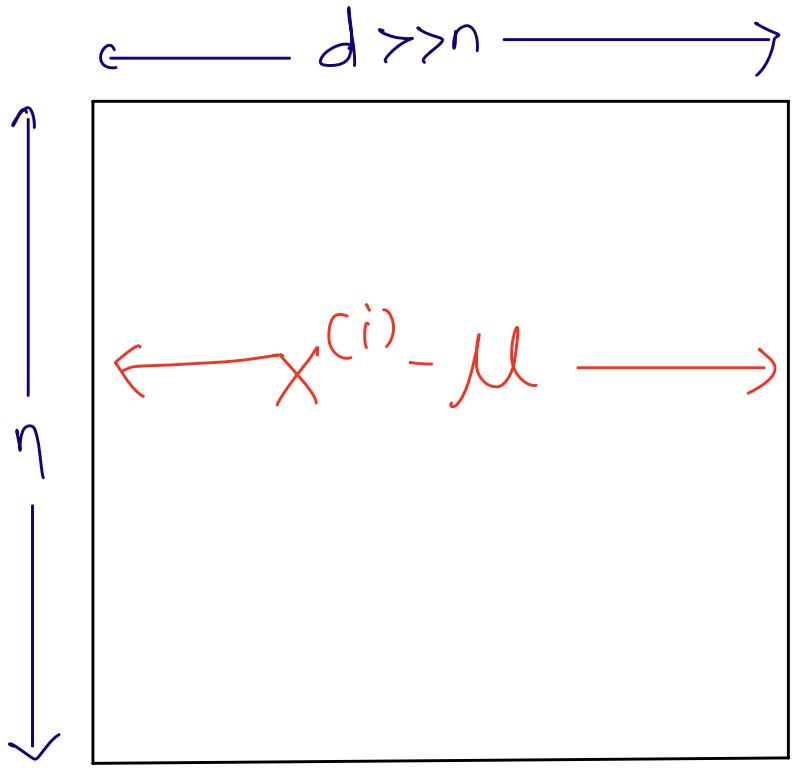
$L \in \mathbb{R}^{d \times k}$

$\Psi \in \mathbb{R}^{d \times d}$ → diagonal matrix (each dimension in d has diff noise added to it)



$x - \mu = z^{(i)T} L + \Psi_{ii}$

$\approx$



Think of it as d linear regression happening simultaneously

$\mu_{z^{(i)}|x^{(i)}} \approx \hat{z}^{(i)}$ best estimate of $z^{(i)}$

$$L \mu_{z^{(i)}|x^{(i)}} = L\hat{z}^{(i)} \approx x^{(i)} - \mu$$
$$:= \hat{x}^{(i)} - \mu$$

$$L = \left[\sum_{i=1}^{n} y^{(i)} z^{(i)T}\right]\left[z^T z + \Sigma\right]^{-1}$$

$$L^T = \left(Z^T Z + \sim\right)^{-1} Z^T y$$

L: Regularised L.R.

$$\Psi_{ii} \approx \left[\sum_{i=1}^{n} \underline{(x^{(i)} - \hat{x}^{(i)})(x^{(i)} - \hat{x}^{(i)})^T} + L \Sigma_{2k} L^T \right]_{ii}$$

$$y = \theta^T x + \varepsilon$$

$$\varepsilon = \frac{1}{n} \sum_{i=1}^{n} (y - \theta^T v)^2$$

We take ii because we don't care about covariance of noise bet' 1 LR & other. We are evaluating each seperately

Given x, we have to estimate z & L.

z & x have affine relationship

P.C.A.