MACHINE LEARNING

UNSUPERVISED LEARNING I

AGENDA

01 K-means clustering

Algorithm, applications, convergence

02 Expectation maximization

Mixture of Gaussians, Jensen's Inequality, Naïve Bayes

03 Factor Analysis

K-MEANS CLUSTERING

K-MEANS CLUSTERING

THE ALGORITHM

We are given a **training set** $\{x^{(1)}, \ldots, x^{(m)}\}$, and want to **group** the **data** into a few cohesive "**clusters**." Now, we **don't have** any **labels** $y^{(i)}$. The algorithm works as follows:

1. **Initialize** cluster **centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ **randomly.**

2. **Repeat** until convergence{

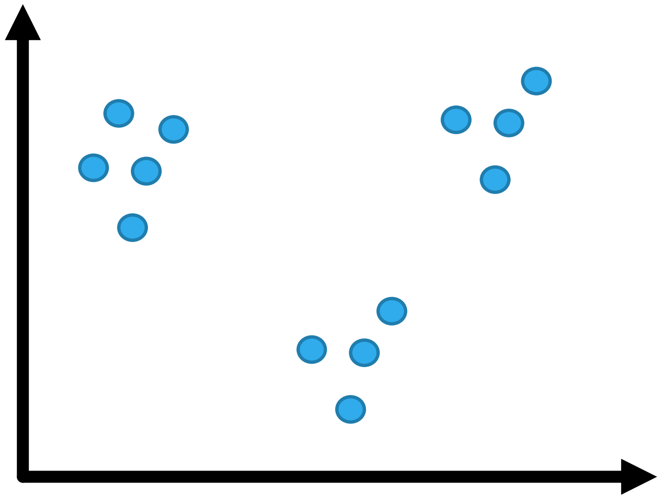
   For every i, set (**Assign point $x^{(i)}$ to cluster j**)

$$c^{(i)} := \arg\min_j \left\| x^{(i)} - \mu_j \right\|^2$$

   For every j, set (**Update cluster centroids**)

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$

}



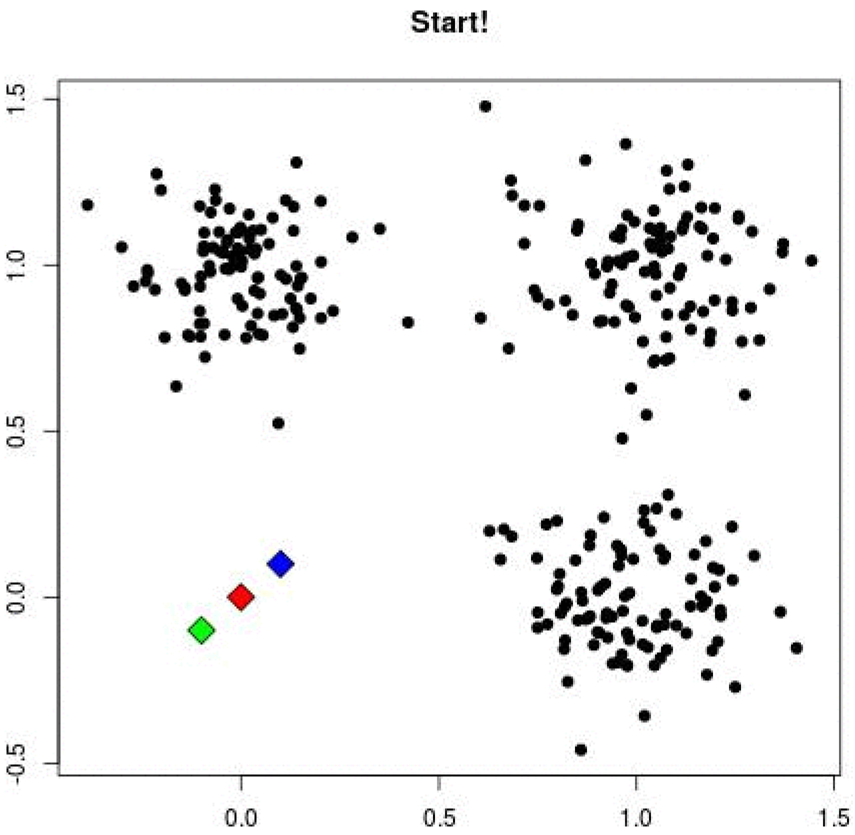
K-MEANS CLUSTERING

THE ALGORITHM

In the algorithm k is the **parameter**, which represents the **number** of **clusters** we want to find.

The **cluster centroids** μ_j represent the **estimations** we make for the **positions** of the **cluster centers**.

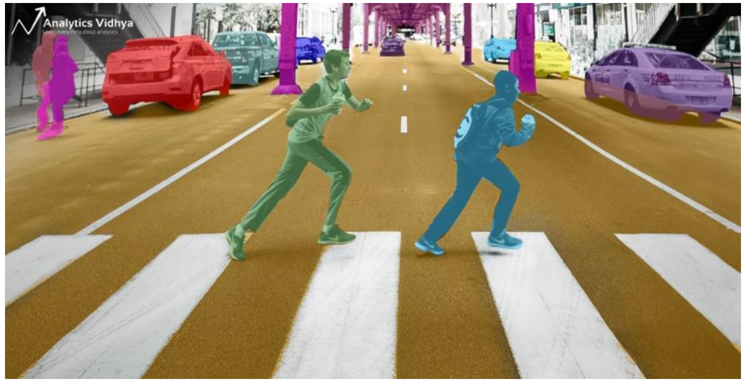
The **initialization** of the **cluster centroids** is calculated by **choosing k training examples** randomly and **setting** the **cluster centroids** to be **equal** to the **values** of these k **examples**.



K-MEANS CLUSTERING
APPLICATIONS

Some of the **applications** of **clustering algorithms are**:

- In **Biology** we need to **find clusters** of **genes**.
- In **Marketing** we would like to **segment markets**.
- In **Journalism** display **common** related **articles**.
- In **Computer Vision** we would like to do **Image Segmentation**.



K-MEANS CLUSTERING
CONVERGENCE

Let us define the **distortion function** to be:

$$J(c, \mu) = \sum_{i=1}^{m} \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2$$

Which **measures** the **sum** of **squared distances between** each **training example** $x^{(i)}$ and the **cluster centroid** $\mu_{c^{(i)}}$.

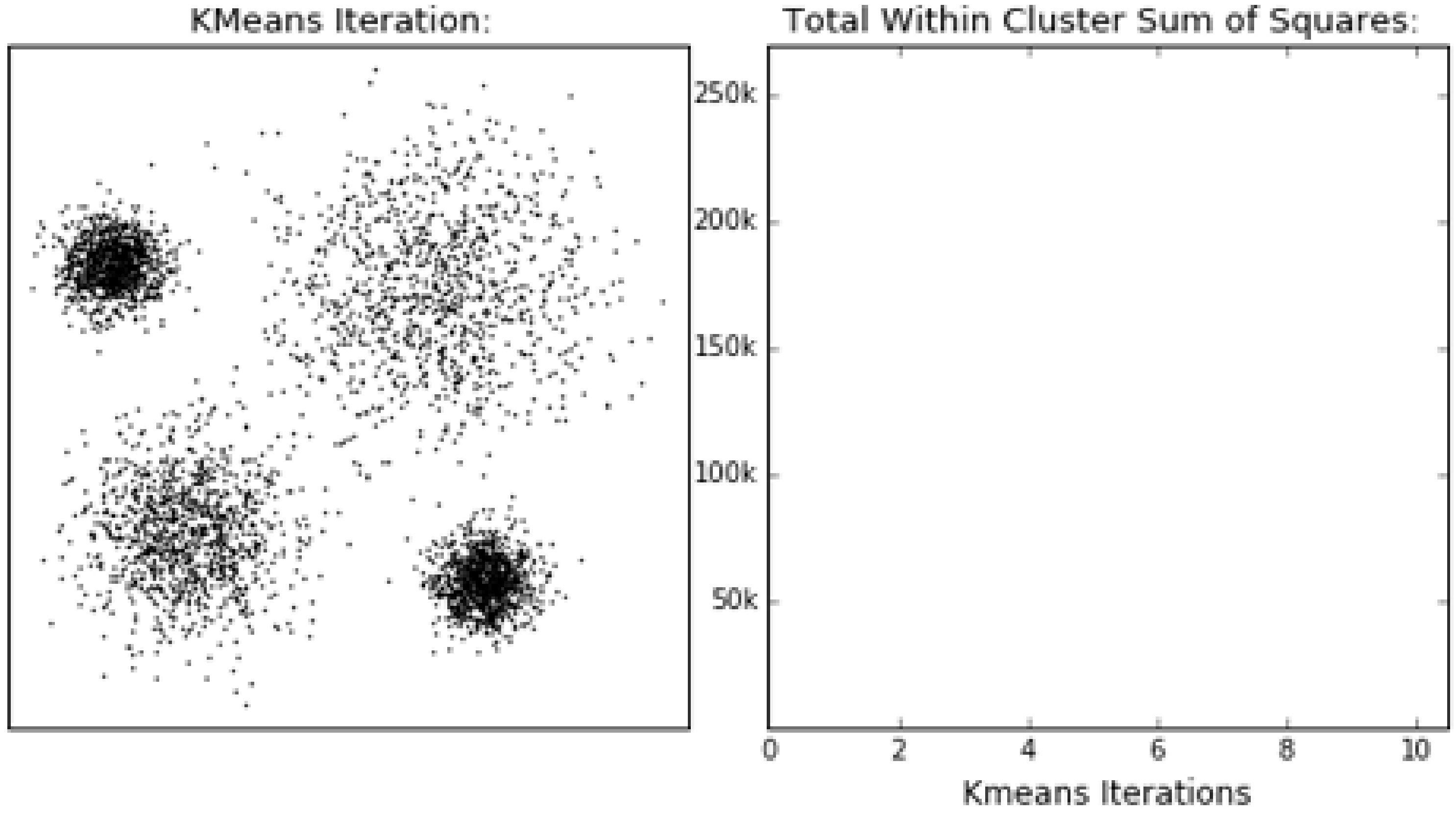
We can see that the **k-means algorithm** is **coordinate descent** on J by **minimizing** the distortion function **with respect** to c while **holding** μ **fixed**.

Thus, J must **monotonically decrease**, and the value of J **must converge**.

Even though, **because** J is a **non-convex function**, it is possible that J **doesn't converge** to a **global minimum**.

K-MEANS CLUSTERING

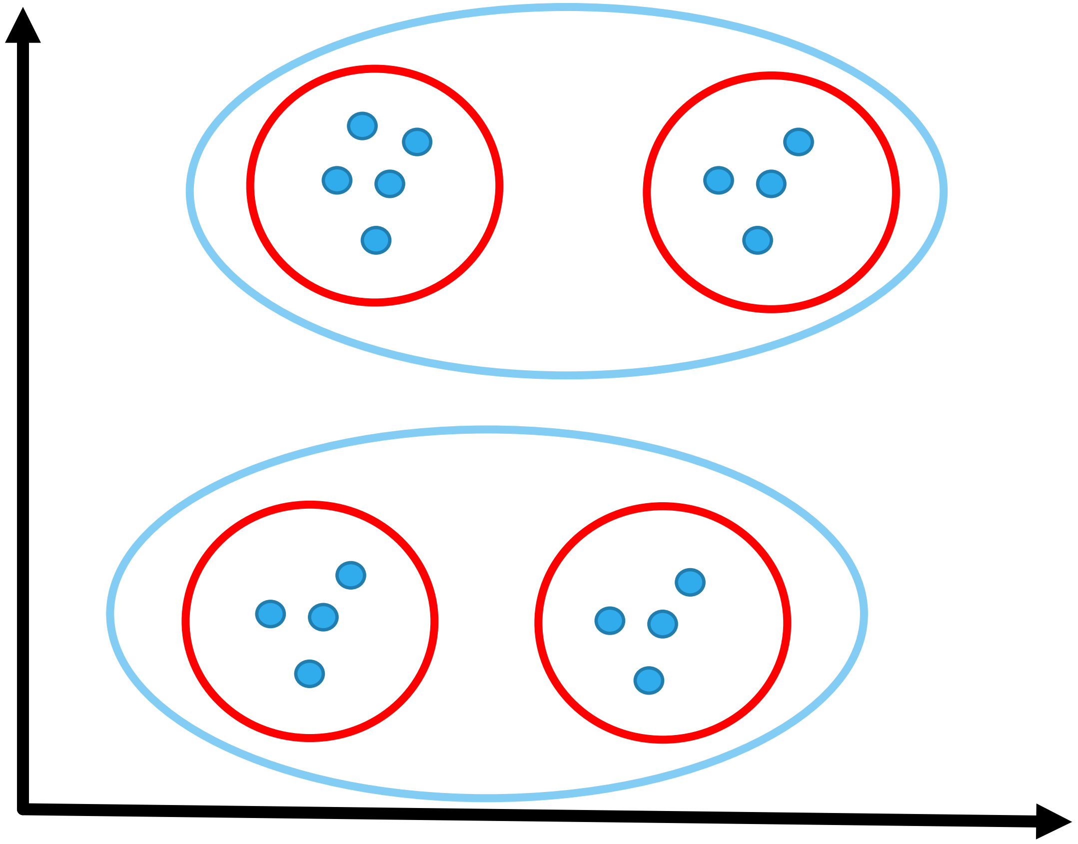
CONVERGENCE



K-MEANS CLUSTERING

PARAMETER K

Choosing the **right number** of **clusters k** may be **ambiguous**. It **depends** on the **application**.



EXPECTATION MAXIMIZATION

EXPECTATION MAXIMIZATION

MIXTURES OF GAUSSIANS

We are given a **training set** $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$. Again, we **don't have** any **labels** $\boldsymbol{y}^{(i)}$.

Now, we want to **model** the **data** by **estimating** its probability **distribution (DENSITY ESTIMATION)**: $\boldsymbol{P}(\boldsymbol{x})$

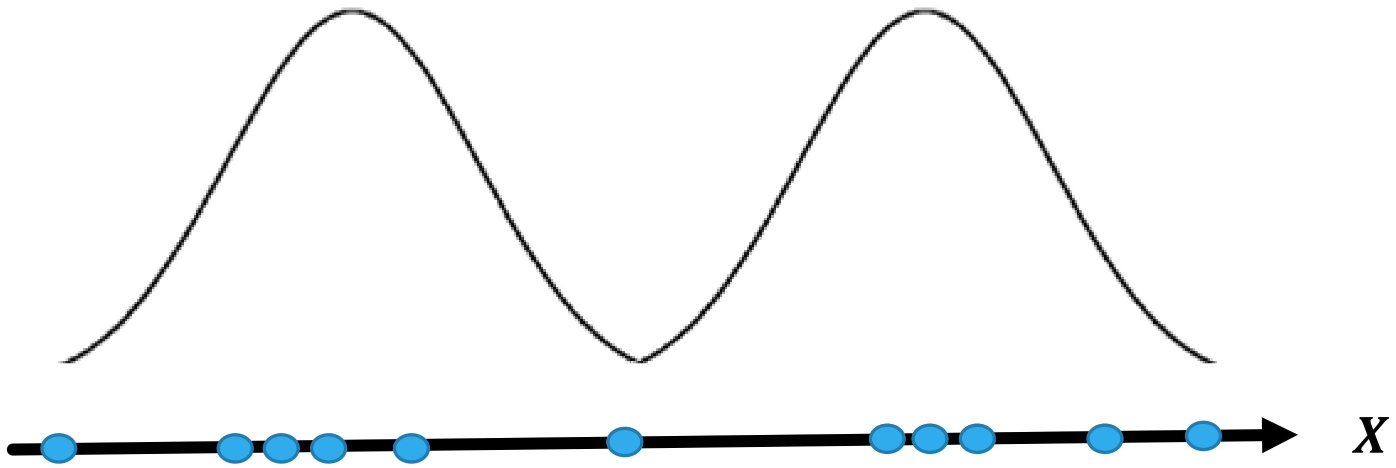
This **density estimation** will help us **detect outliers**. Thus, it **allows** to **compute** the **likelihood** of **new data points** arriving. A **problem** known as **Anomaly Detection**.

It is important to note, that **many distributions may not** be **known distributions** (Gaussian, Poisson, etc.)

EXPECTATION MAXIMIZATION

MIXTURES OF GAUSSIANS

Let us look at an **example**, **where** $\boldsymbol{x}^{(i)} \in \mathbb{R}$. The **density distribution** of our **training set** may look like the sum of two Gaussians:



We may think that the **data** set **may have come** from **two** separate **Gaussians**, but we **don't know from which Gaussian each** of the **data points came from**.

EXPECTATION MAXIMIZATION

MIXTURES OF GAUSSIANS

Let us imagine that there is a **latent** (**hidden** / unobserved) **random variable** z and $x^{(i)}$, $z^{(i)}$ have a **joint distribution:**

$$p\left(x^{(i)}, z^{(i)}\right) = p\left(x^{(i)} / z^{(i)}\right) p\left(z^{(i)}\right)$$

We will **assume** that $z^{(i)} \sim Multinomial(\phi)$ (for 2 Gaussians this will be Bernoulli)**,** where

- $\phi_j \geq 0$
- $\sum_{j=1}^{k} \phi_j = 1$
- $\phi_j = p\left(z^{(i)} = j\right)$

Also, we will **assume** that $x^{(i)} / z^{(i)} = j$ is **distributed Gaussian** $N\left(\mu_j, \Sigma_j\right)$**.**

VERY SIMILAR TO GAUSSIAN DISCRIMINANT ANALYSIS (y is known, z is not)

EXPECTATION MAXIMIZATION

MIXTURES OF GAUSSIANS

The **main difficulty** resides in the fact that we **don't know $z^{(i)}$**. Even though, let us **assume** that we know them so we can **write** the **joint log likelihood** of our **data as**:

$$l(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p\left(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma\right)$$

$$l(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p\left(x^{(i)} / z^{(i)}; \mu, \Sigma\right) p\left(z^{(i)}; \phi\right)$$

We make the **same calculations** that we have done for **maximum likelihood estimation** in **Gaussian Discriminant Analysis.**

EXPECTATION MAXIMIZATION

MIXTURES OF GAUSSIANS

The **results** of **maximum likelihood estimation** are**:**

$$\boldsymbol{\phi_j = \sum_{i=1}^{m} \frac{1(z^{(i)} = j)}{m}}$$

$$\boldsymbol{\mu_j = \frac{\sum_{i=1}^{m} 1(z^{(i)} = j) x^{(i)}}{\sum_{i=1}^{m} 1(z^{(i)} = j)}}$$

$$\boldsymbol{\Sigma j = \frac{\sum_{i=1}^{m} 1(z^{(i)} = j)\left(x^{(i)} - \mu_j\right)\left(x^{(i)} - \mu_j\right)^T}{\sum_{i=1}^{m} 1(z^{(i)} = j)}}$$

In here $\boldsymbol{z^{(i)}}$ **represent** which of the $\boldsymbol{k}$ **Gaussians each** $\boldsymbol{x^{(i)}}$ **had come from** $\boldsymbol{\{0, 1, \dots, k\}}$**.**

The problem is that we don't know $\boldsymbol{z^{(i)}}$**.**

EXPECTATION MAXIMIZATION

MIXTURES OF GAUSSIANS

The **solution** is the **EM algorithm** is an **iterative algorithm**, which has **two main steps**:

1. **E-step:** "**guess**" the **values** of the $\boldsymbol{z^{(i)}s}$.
2. **M-step: updates parameters** of the model **based** on **previous guesses**.

Repeat until convergence{

(E-step) For every i, j set

$$\boldsymbol{w_j{}^{(i)} := p\left(z^{(i)} = j / x^{(i)}; \phi, \mu, \Sigma\right)}$$

(M-step) Update the parameters

$$\boldsymbol{\phi_j = \frac{1}{m} \sum_{i=1}^{m} w_j{}^{(i)}}$$

$$\boldsymbol{\mu_j = \frac{\sum_{i=1}^{m} w_j{}^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j{}^{(i)}}}$$

$$\boldsymbol{\Sigma = \frac{\sum_{i=1}^{m} w_j{}^{(i)} \left(x^{(i)} - \mu_j\right)\left(x^{(i)} - \mu_j\right)^T}{\sum_{i=1}^{m} w_j{}^{(i)}}}$$

}

EXPECTATION MAXIMIZATION

MIXTURES OF GAUSSIANS

The "**guess**" $p\left(z^{(i)} = j / x^{(i)}; \phi, \mu, \Sigma\right)$ is calculated by **evaluating** the **density** of a **Gaussian** with mean μ_j and covariance Σ_j at $x^{(i)}$ **(the posterior).**

$$w_j^{(i)} := p\left(z^{(i)} = j / x^{(i)}; \phi, \mu, \Sigma\right) = \frac{p\left(x^{(i)} / z^{(i)} = j; \mu, \Sigma\right) p\left(z^{(i)} = j\right)}{\sum_{l=1}^{k} p\left(x^{(i)} / z^{(i)} = l; \mu, \Sigma\right) p\left(z^{(i)} = l; \phi\right)}$$

$$w_j^{(i)} := p\left(z^{(i)} = j / x^{(i)}; \phi, \mu, \Sigma\right) = \frac{Gaussian(\mu, \Sigma)\, \phi_j}{P(X)}$$

Like K-means, this algorithm is **also susceptible** to **local optima**, so **reinitializing** at several different initial **parameters may be** a **good** idea.

EXPECTATION MAXIMIZATION

JENSEN'S INEQUALITY

THEOREM

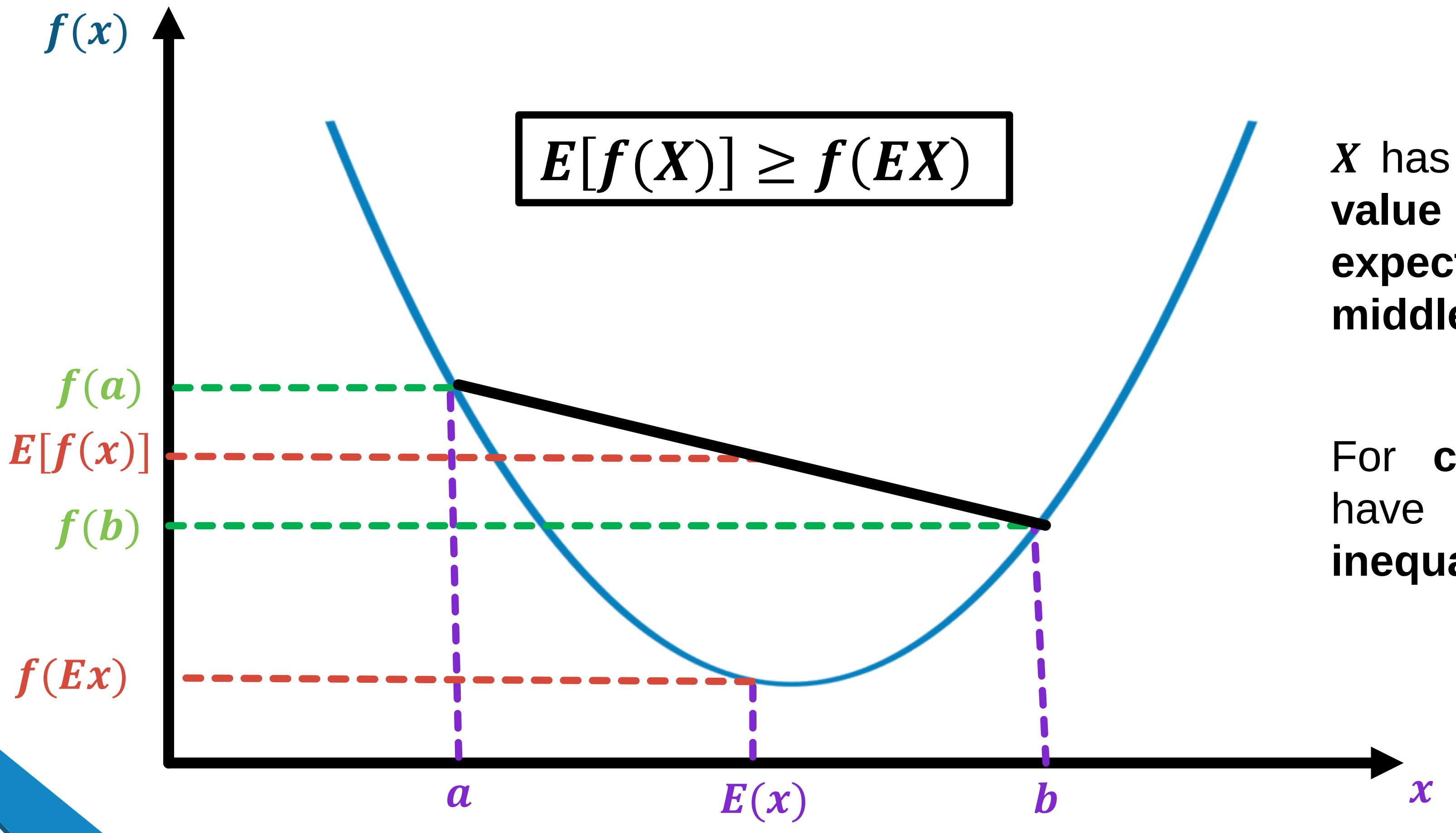
Let $\boldsymbol{f}$ be a **convex function** (if $\boldsymbol{f''(x) \geq 0}$; $\forall \boldsymbol{x} \in \boldsymbol{R}$ or its **Hessian** is **positive semi-definite** for **vector inputs**) whose domain is the set of real numbers and let $\boldsymbol{X}$ be a **random variable**. Then:

$$\boldsymbol{E[f(X)] \geq f(EX)}$$

If $\boldsymbol{f}$ is **strictly convex ($\boldsymbol{f''(x) > 0}$)**, then $\boldsymbol{E[f(X)] = f(EX)}$ holds **true if and only if** $\boldsymbol{X = E[X]}$ with **probability 1** (the **expected value** does **not change**).

EXPECTATION MAXIMIZATION

JENSEN'S INEQUALITY



X has **50%** chance of **having value** a or b, thus its **expected value** is at the **middle** of **both possibilities**.

For **concave functions** we have the **direction** of the **inequality** will be **reversed**.

EXPECTATION MAXIMIZATION

E M A L G O R I T H M

Suppose we have an **estimation problem** in which we have a **training set** $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ consisting of $\boldsymbol{m}$ **independent examples**.

The **objective** will be to **fit** the **parameters** of a **model** $\boldsymbol{p}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{w})$ **to** the **data**, where the **likelihood** is given by:

$$\boldsymbol{l}(\boldsymbol{w}) = \sum_{\boldsymbol{i}=1}^{\boldsymbol{m}} \boldsymbol{log}(\boldsymbol{p}(\boldsymbol{x}^{(\boldsymbol{i})}; \boldsymbol{w})$$

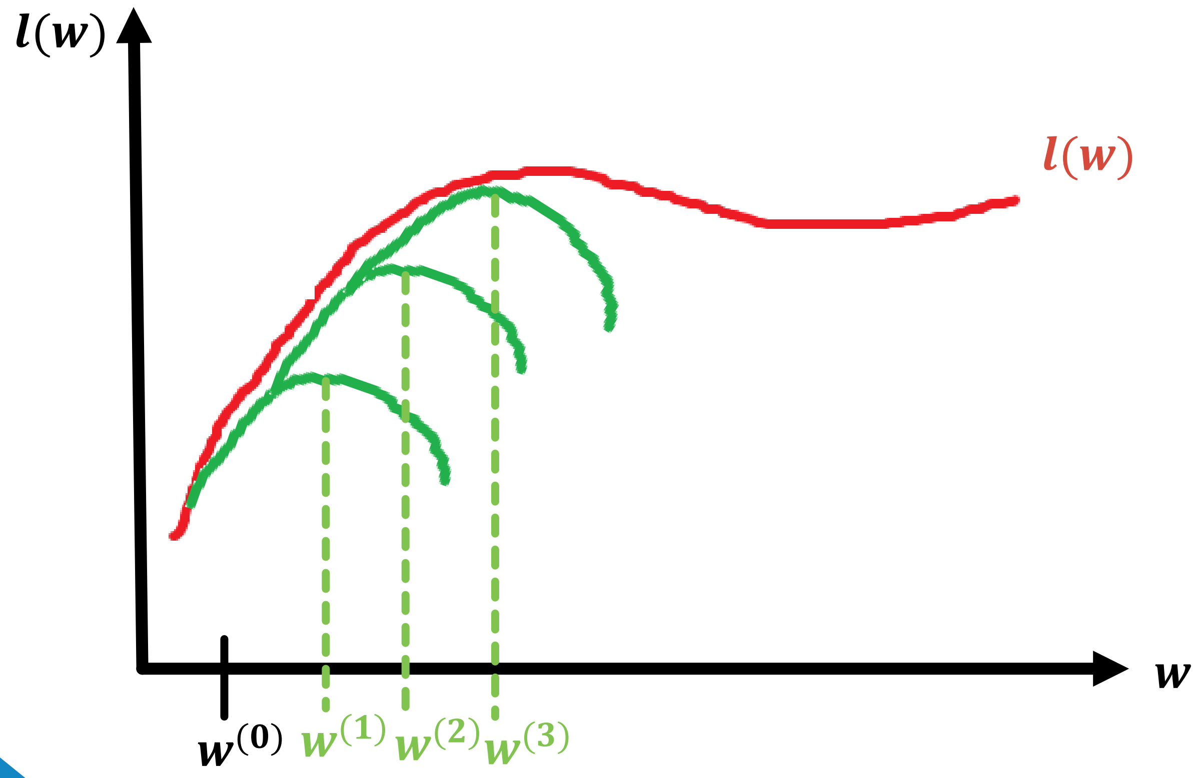
$$\boldsymbol{l}(\boldsymbol{w}) = \sum_{\boldsymbol{i}=1}^{\boldsymbol{m}} \boldsymbol{log} \sum_{\boldsymbol{z}} \boldsymbol{p}(\boldsymbol{x}^{(\boldsymbol{i})}, \boldsymbol{z}^{(\boldsymbol{i})}; \boldsymbol{w})$$

We want to **find** the **maximum likelihood estimates** of the **parameters** $\boldsymbol{w}$**.**

EXPECTATION MAXIMIZATION

EM ALGORITHM

Again, **finding** the **parameters** $\boldsymbol{w}$ with MLE is **not** an **easy task because** we **don't know** the latent $\boldsymbol{z}^{(i)}$'s. The **EM algorithm** will help us **overcome** the **problem**.



- Repeatedly **construct** a **lower-bound** on $\boldsymbol{l}$ **(E-step).**
- **Optimize** that lower-bound (**M-step**).

EXPECTATION MAXIMIZATION

E M A L G O R I T H M

Thus **we have** the following:

$$\boldsymbol{l(w) = \sum_{i=1}^{m} log \sum_{z} p(x^{(i)}, z^{(i)}; w)}$$

We will **build** a **probability distribution** $\boldsymbol{Q_i}$ over the latent variables $\boldsymbol{z^{(i)}}$, where $\boldsymbol{Q_i(z^{(i)})}$ and $\sum \boldsymbol{Q_i(z^{(i)})} = 1$**.**

$$\boldsymbol{l(w) = \sum_{i=1}^{m} log \sum_{z} \frac{Q_i(z^{(i)})\, p(x^{(i)}, z^{(i)}; w)}{Q_i(z^{(i)})}}$$

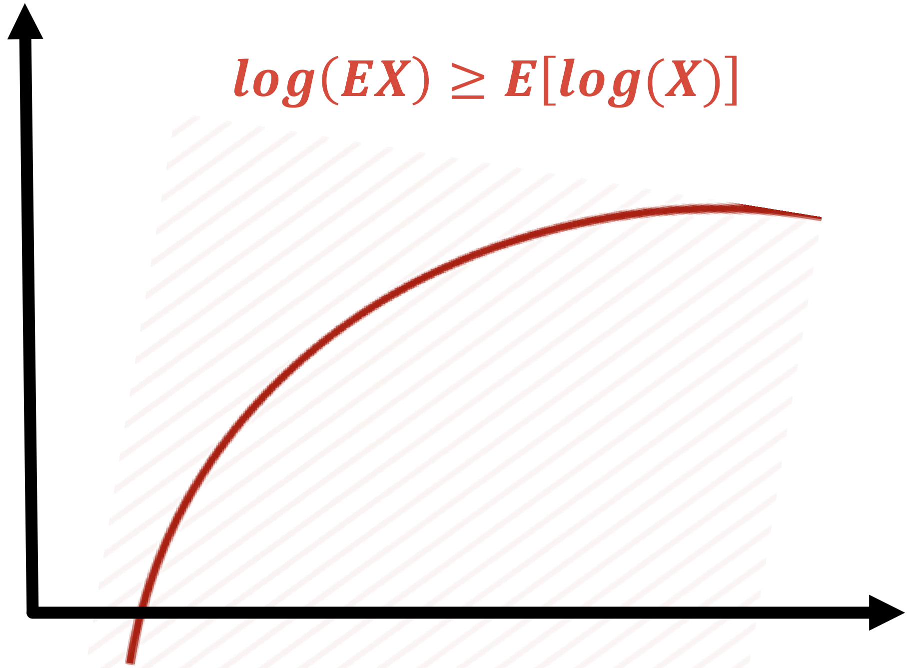
$$\boldsymbol{l(w) = \sum_{i=1}^{m} log \sum_{z} E_{z^{(i)} \sim Q_i} \left[\frac{p(x^{(i)}, z^{(i)}; w)}{Q_i(z^{(i)})} \right]}$$

EXPECTATION MAXIMIZATION

E M ALGORITHM

Now, we can see that the **log function** is **concave**, which allow us to **define** the **following**:

$$\log \sum_{z} E_{z^{(i)} \sim Q_i}\left[\frac{p(x^{(i)}, z^{(i)}; w)}{Q_i(z^{(i)})}\right] \geq \sum_{z} E_{z^{(i)} \sim Q_i}\left[\log \frac{p(x^{(i)}, z^{(i)}; w)}{Q_i(z^{(i)})}\right]$$



EXPECTATION MAXIMIZATION

EM ALGORITHM

Expanding out we have:

$$\sum_{i=1}^{m} log \sum_{z} E_{z^{(i)} \sim Q_i}\left[\frac{p(x^{(i)}, z^{(i)}; w)}{Q_i(z^{(i)})}\right] \geq \sum_{i=1}^{m} \sum_{z} E_{z^{(i)} \sim Q_i}\left[log \frac{p(x^{(i)}, z^{(i)}; w)}{Q_i(z^{(i)})}\right]$$

$$\sum_{i=1}^{m} log \sum_{z} Q_i(z^{(i)}) \left[\frac{p(x^{(i)}, z^{(i)}; w)}{Q_i(z^{(i)})}\right] \geq \sum_{i=1}^{m} \sum_{z} Q_i(z^{(i)}) \, log \frac{p(x^{(i)}, z^{(i)}; w)}{Q_i(z^{(i)})}$$

$$l(w) \geq \sum_{i=1}^{m} \sum_{z} Q_i(z^{(i)}) \, log \frac{p(x^{(i)}, z^{(i)}; w)}{Q_i(z^{(i)})}$$

We observe that we **have** a **lower bound** over the **likelihood** $l(w)$.

EXPECTATION MAXIMIZATION

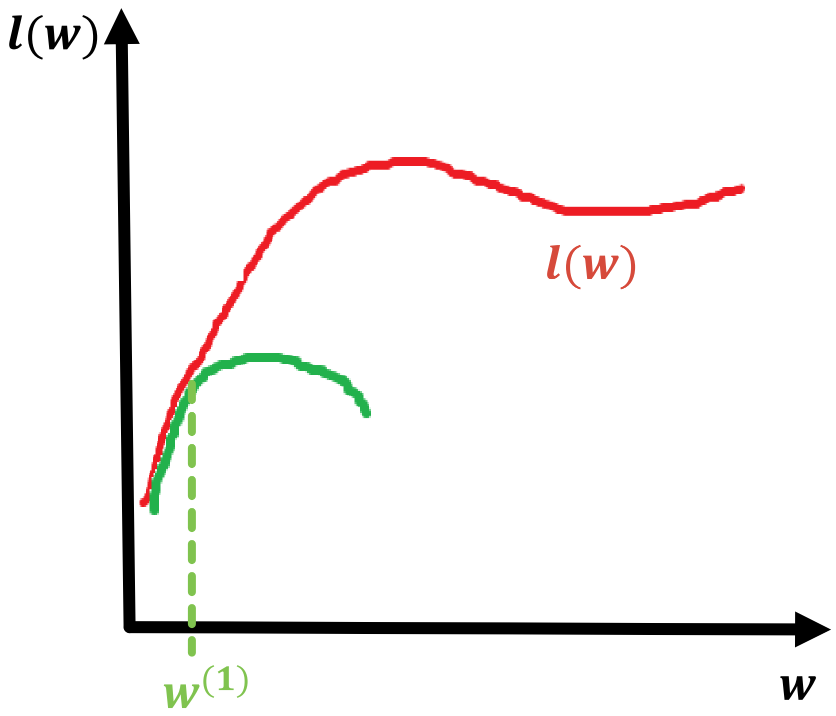
EM ALGORITHM

What we **want** is the **inequality** to **turn into** an **equality** for a current value of $\boldsymbol{w}$. Thus, when we **optimize** the **lower bound**, we are **also optimizing** the **true** $\boldsymbol{l(w)}$.

The **objective** is to **find probability distribution** $\boldsymbol{Q_i}$ that will **transform** the **inequality** to **equality**.

The only way of doing this is to take the **expectation** of a **constant value** (remembering Jensen's inequality).

$$\frac{p(x^{(i)}, z^{(i)}; w)}{Q_i(z^{(i)})} = c$$



EXPECTATION MAXIMIZATION

E M ALGORITHM

Elaborating further in the **equation**, we have:

$$Q_i\left(z^{(i)}\right) = \frac{p(x^{(i)}, z^{(i)}; w)}{c}$$

$$Q_i\left(z^{(i)}\right) \propto p(x^{(i)}, z^{(i)}; w)$$

Since we **know** that $\sum Q_i\left(z^{(i)}\right) = 1$

$$1 = \frac{p(x^{(i)}, z^{(i)}; w)}{Q_i(z^{(i)})\, c}$$

$$\frac{1}{\sum Q_i(z^{(i)})} = \frac{p(x^{(i)}, z^{(i)}; w)}{Q_i(z^{(i)})\, c}$$

$$\frac{Q_i\left(z^{(i)}\right)}{\sum Q_i(z^{(i)})} = \frac{p(x^{(i)}, z^{(i)}; w)}{c}$$

$$c = \sum_{z} p(x^{(i)}, z^{(i)}; w)$$

EXPECTATION MAXIMIZATION

E M A L G O R I T H M

In **conclusion** we have that**:**

$$Q_i\left(z^{(i)}\right) = \frac{p\left(x^{(i)}, z^{(i)}; w\right)}{\sum_z p\left(x^{(i)}, z^{(i)}; w\right)}$$

$$Q_i\left(z^{(i)}\right) = \frac{p\left(x^{(i)}, z^{(i)}; w\right)}{p\left(x^{(i)}; w\right)}$$

$$Q_i\left(z^{(i)}\right) = p\left(z^{(i)} / x^{(i)}; w\right)$$

The distribution $Q_i\left(z^{(i)}\right)$ is just the **posterior distribution** of the **latent random variables** $z^{(i)}$'s **given** we have **observed** the **data**.

EXPECTATION MAXIMIZATION

E M ALGORITHM

The **EM algorithm** therefore is as follows:

Repeat until **convergence {**

1. **(E-step):** for each i, set the lower bound

$$Q_i\left(z^{(i)}\right) := p\left(z^{(i)} / x^{(i)}; w\right)$$

2. **(M-step):** optimize the lower bound

$$w := \arg\max_{w} \sum_{i=1}^{m} \sum_{z} Q_i\left(z^{(i)}\right) \log \frac{p\left(x^{(i)}, z^{(i)}; w\right)}{Q_i\left(z^{(i)}\right)}$$

}

EXPECTATION MAXIMIZATION

E M A L G O R I T H M

We can **define**:

$$\boldsymbol{J(w,Q) = Q_i\left(z^{(i)}\right) log \frac{p(x^{(i)}, z^{(i)}; w)}{Q_i(z^{(i)})}}$$

$$\boldsymbol{l(w) \leq J(w,Q)}$$

Thus, the ***EM* algorithm** can be viewed as **coordinate ascent** on $\boldsymbol{J}$, in which the **E-step maximizes** with respect to $\boldsymbol{Q}$ and the **M-step maximizes** it with respect to $\boldsymbol{w}$.

EM and MIXTURE OF GAUSSIANS

EXPECTATION MAXIMIZATION

EM AND MIXTURE OF GAUSSIANS

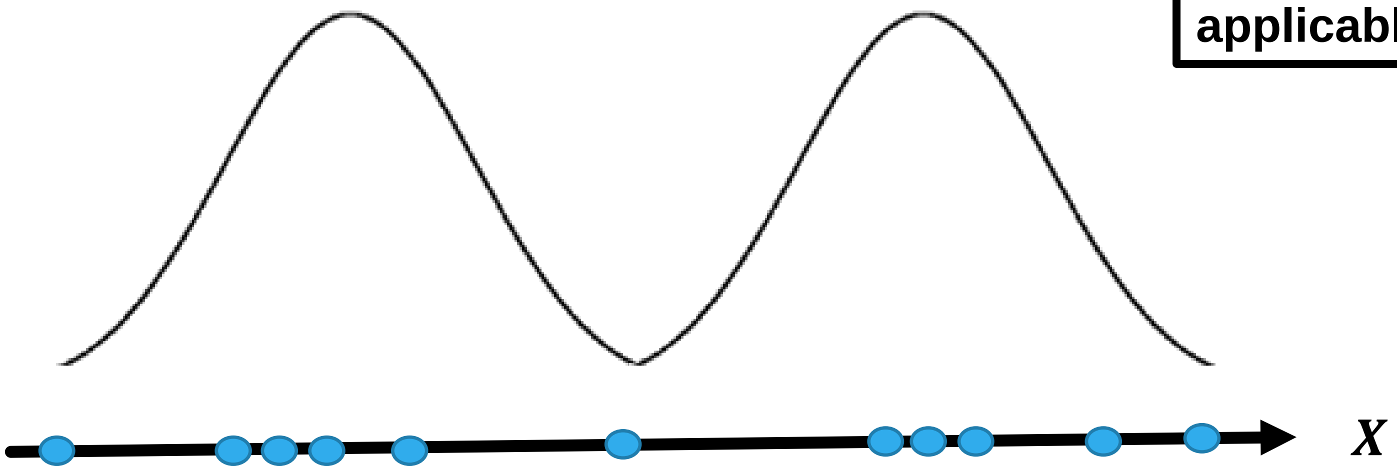
As we have seen, we can **estimate** the **probability density distribution** of a set of **data points** $\{x^{(1)},\ldots,x^{(m)}\}$ using a **mixture** of **Gaussians**.

$$p\left(x^{(i)},z^{(i)}\right)=p\left(x^{(i)}/z^{(i)}\right)p\left(z^{(i)}\right)$$

$$z^{(i)}\sim Multinomial(\phi)$$

$$x^{(i)}/z^{(i)}=j\sim N(\mu_j,\Sigma_j)$$

NOTE: the **mixture** of **Gaussians model** is **applicable** when $m \gg n$.



EXPECTATION MAXIMIZATION

EM AND MIXTURE OF GAUSSIANS

E-STEP:

Applying the **EM algorithm to** the problem of the **mixture** of **Gaussians** we can **perform** the ***E*** **step** by getting the **posterior** of the **latent variables**:

$$\boldsymbol{Q_i\left(z^{(i)}\right) := p(z^{(i)}/x^{(i)}; w)}$$

$$\boldsymbol{w^{(i)}{}_j = Q_i\left(z^{(i)} = j\right) = p(z^{(i)} = j/x^{(i)}; \phi, \mu, \Sigma)}$$

Where $\boldsymbol{Q_i\left(z^{(i)} = j\right)}$ denotes the **probability** of $\boldsymbol{z^{(i)}}$ **taking** the **value** $\boldsymbol{j}$ under the **distribution** $\boldsymbol{Q_i}$.

EXPECTATION MAXIMIZATION

EM AND MIXTURE OF GAUSSIANS

E-STEP:

Expanding the **formula** of the ***E*** step by **using Bayes' rule**:

$$w^{(i)}{}_j = Q_i\left(z^{(i)}\right) := p\left(z^{(i)}/x^{(i)}; w\right)$$

$$Q_i\left(z^{(i)} = j\right) = \frac{p\left(x^{(i)}/z^{(i)} = j\right)P\left(z^{(i)} = j\right)}{\sum_k p\left(x^{(i)}/z^{(i)} = k\right)P\left(z^{(i)} = k\right)}$$

We know that $p\left(x^{(i)}/z^{(i)} = j\right) \sim \boldsymbol{Gaussian}$ and $P\left(z^{(i)} = j\right) \sim \boldsymbol{Multinomial}$

$$Q_i\left(z^{(i)} = j\right) = \frac{\frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(x^{(i)}-\mu_j\right)^T \Sigma_j^{-1}\left(x^{(i)}-\mu_j\right)} \phi_j}{\sum_k \frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(x^{(i)}-\mu_k\right)^T \Sigma_k^{-1}\left(x^{(i)}-\mu_k\right)} \phi_k}$$

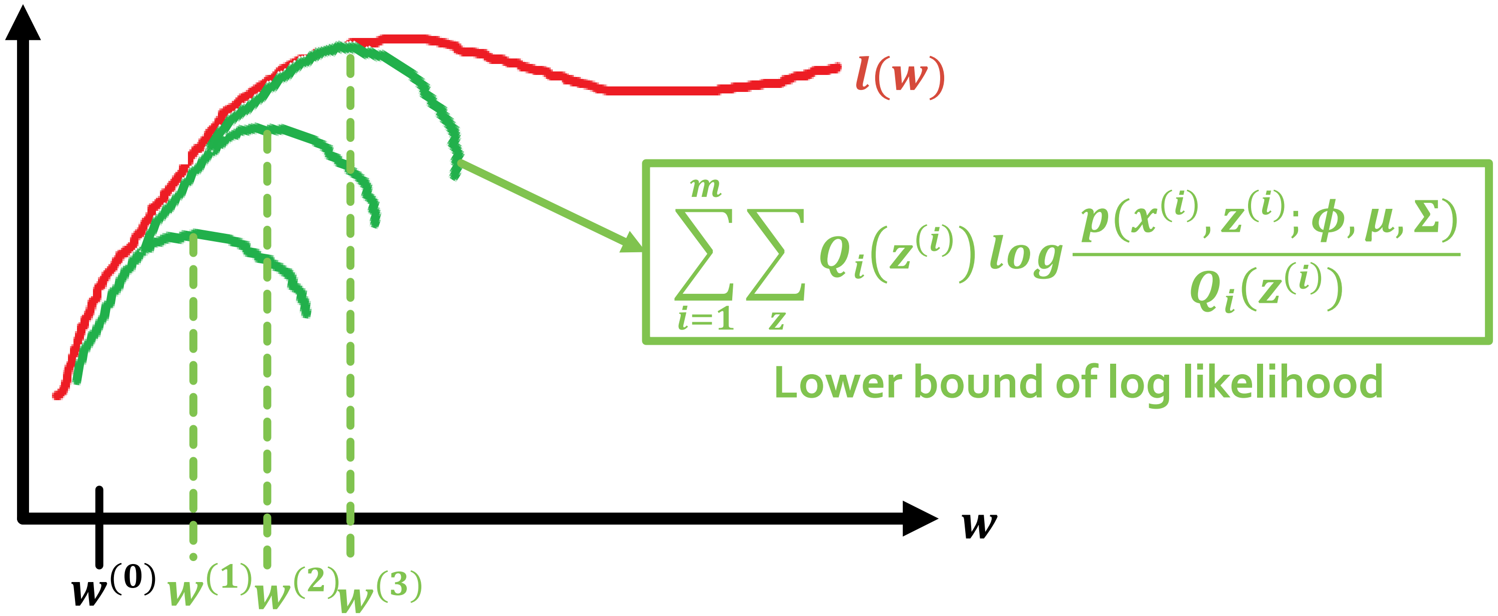
EXPECTATION MAXIMIZATION

EM AND MIXTURE OF GAUSSIANS

M-STEP:

For the ***M*** **step** we need to **maximize** with respect to the parameters $\boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ the **following**:

$$w := \arg\max_{w} \sum_{i=1}^{m} \sum_{z} Q_i\left(z^{(i)}\right) \log \frac{p\left(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma\right)}{Q_i\left(z^{(i)}\right)}$$



EXPECTATION MAXIMIZATION

EM AND MIXTURE OF GAUSSIANS

M-STEP: maximize for $\boldsymbol{\mu}$

Expanding terms out we have:

$$\sum_{i=1}^{m}\sum_{z} Q_i\left(z^{(i)}\right) log \frac{p\left(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma\right)}{Q_i\left(z^{(i)}\right)} = \sum_{i=1}^{m}\sum_{j=1}^{k} w^{(i)}{}_j \, log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(x^{(i)}-\mu_j\right)^T \Sigma_j^{-1}\left(x^{(i)}-\mu_j\right)} \phi_j}{w^{(i)}{}_j}$$

Simplifying:

$$= \sum_{i=1}^{m}\sum_{j=1}^{k} w^{(i)}{}_j \left[log \left(\frac{1}{w^{(i)}{}_j (2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} \right) + log \left(e^{-\frac{1}{2}\left(x^{(i)}-\mu_j\right)^T \Sigma_j^{-1}\left(x^{(i)}-\mu_j\right)} \phi_j \right) \right]$$

EXPECTATION MAXIMIZATION

EM AND MIXTURE OF GAUSSIANS

M-STEP: maximize for $\boldsymbol{\mu}$

Taking the derivative with respect to $\boldsymbol{\mu_l}$ we have:

$$\nabla_{\mu_l} \sum_{i=1}^{m} \sum_{j=1}^{k} w^{(i)}{}_j \left[log\left(\frac{1}{w^{(i)}{}_j (2\pi)^{n/2} |\Sigma|^{\frac{1}{2}}} \right) + log\left(e^{-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1} (x^{(i)}-\mu_j)} \phi_j \right) \right]$$

$$= \nabla_{\mu_l} \sum_{i=1}^{m} \sum_{j=1}^{k} w^{(i)}{}_j \left[-\frac{1}{2} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right]$$

$$= -\frac{1}{2} \sum_{i=1}^{m} w^{(i)}{}_l \nabla_{\mu_l} \left[-2{\mu_l}^T \Sigma_l^{-1} x^{(i)} + {\mu_l}^T \Sigma_l^{-1} \mu_l \right]$$

EXPECTATION MAXIMIZATION

EM AND MIXTURE OF GAUSSIANS

M-STEP: maximize for $\boldsymbol{\mu}$

$$= -\frac{1}{2}\sum_{i=1}^{m} w^{(i)}{}_{l}\nabla_{\mu_l}\left[-2{\mu_l}^T\Sigma_l^{-1}x^{(i)} + {\mu_l}^T\Sigma_l^{-1}\mu_l\right]$$

$$= -\frac{1}{2}\sum_{i=1}^{m} w^{(i)}{}_{l}\left[-2\Sigma_l^{-1}x^{(i)} + 2{\mu_l}^T\Sigma_j^{-1}\right]$$

$$\sum_{i=1}^{m} w^{(i)}{}_{l}\left[\Sigma_l^{-1}x^{(i)} - {\mu_l}^T\Sigma_j^{-1}\right] = 0$$

$$\sum_{i=1}^{m} w^{(i)}{}_{l}\Sigma_l^{-1}x^{(i)} - \sum_{i=1}^{m} w^{(i)}{}_{l}\,{\mu_l}^T\Sigma_j^{-1} = 0$$

EXPECTATION MAXIMIZATION

EM AND MIXTURE OF GAUSSIANS

M-STEP: maximize for $\boldsymbol{\mu}$

$$\sum_{i=1}^{m} w^{(i)}{}_l\, \mu_l{}^T \Sigma_j^{-1} = \sum_{i=1}^{m} w^{(i)}{}_l \Sigma_l^{-1} x^{(i)}$$

$$\mu_l{}^T \sum_{i=1}^{m} w^{(i)}{}_l = \Sigma_l^{-1} \sum_{i=1}^{m} w^{(i)}{}_l x^{(i)}$$

$$\mu_l{}^T = \Sigma_l \Sigma_l^{-1} \frac{\sum_{i=1}^{m} w^{(i)}{}_l x^{(i)}}{\sum_{i=1}^{m} w^{(i)}{}_l}$$

$$\mu_l{}^T = \frac{\sum_{i=1}^{m} w^{(i)}{}_l x^{(i)}}{\sum_{i=1}^{m} w^{(i)}{}_l}$$

EXPECTATION MAXIMIZATION

EM AND MIXTURE OF GAUSSIANS

M-STEP: maximize for $\boldsymbol{\phi}$

$$\sum_{i=1}^{m}\sum_{j=1}^{k} w^{(i)}{}_j \log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1}(x^{(i)}-\mu_j)} \phi_j}{w^{(i)}{}_j}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{k} w^{(i)}{}_j \left[\log\left(\frac{1}{w^{(i)}{}_j (2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}}\right) + \log\left(e^{-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1}(x^{(i)}-\mu_j)} \phi_j\right)\right]$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{k} w^{(i)}{}_j \left[\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1}(x^{(i)}-\mu_j) + \log(\phi_j)\right]$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{k} w^{(i)}{}_j \log(\phi_j)$$

EXPECTATION MAXIMIZATION

EM AND MIXTURE OF GAUSSIANS

M-STEP: maximize for $\boldsymbol{\phi}$

Because $\boldsymbol{\phi \sim Multinomial}$ we have an additional constraint that $\sum_{j=1}^{k} \boldsymbol{\phi_j} = \mathbf{1}$.
Thus we need to construct our Lagrangian:

$$L(\phi) = \sum_{i=1}^{m} \sum_{j=1}^{k} w^{(i)}{}_j \, log(\phi_j) + \beta \left(\sum_{j=1}^{k} \phi_j - 1 \right)$$

$$\frac{\partial L(\phi)}{\partial \phi} = \sum_{i=1}^{m} \frac{w^{(i)}{}_j}{\phi_j} + \beta = 0$$

$$\phi_j = \sum_{i=1}^{m} \frac{w^{(i)}{}_j}{-\beta}$$

EXPECTATION MAXIMIZATION

EM AND MIXTURE OF GAUSSIANS

M-STEP: maximize for $\boldsymbol{\phi}$

Using the constraint $\sum_{j=1}^{k} \boldsymbol{\phi}_j = \mathbf{1}$. we have:

$$\boldsymbol{\phi}_j = \sum_{i=1}^{m} \frac{\boldsymbol{w^{(i)}}_j}{-\boldsymbol{\beta}}$$

$$\sum_{j=1}^{k} \boldsymbol{\phi}_j = \sum_{j=1}^{k} \sum_{i=1}^{m} \frac{\boldsymbol{w^{(i)}}_j}{-\boldsymbol{\beta}}$$

$$\mathbf{1} = \sum_{j=1}^{k} \sum_{i=1}^{m} \frac{\boldsymbol{w^{(i)}}_j}{-\boldsymbol{\beta}}$$

$$\mathbf{1} = \sum_{i=1}^{m} \frac{\mathbf{1}}{-\boldsymbol{\beta}} \rightarrow -\boldsymbol{\beta} = \boldsymbol{m}$$

EXPECTATION MAXIMIZATION

EM AND MIXTURE OF GAUSSIANS

M-STEP: maximize for $\boldsymbol{\phi}$

Finally we have that:

$$\phi_j = \sum_{i=1}^{m} \frac{w^{(i)}{}_j}{-\beta}$$

$$\phi_j = \frac{1}{m} \sum_{i=1}^{m} w^{(i)}{}_j$$

EM and NAÏVE BAYES

EXPECTATION MAXIMIZATION

EM AND NAÏVE BAYES

As we have seen, **Naïve Bayes classifier** runs for **input** training **examples** that take **discrete values**.

Therefore, given a training set $\{x^{(1)}, \ldots, x^{(m)}\}$ where $x^{(i)} \in \{0,1\}^n$ and $x^{(i)}{}_j = 1\{word\ j\ appears\ in\ document\ i\}$.

Suppose we want to find **two clusters** $z^{(i)} = \{0,1\}$ **("spam"** or **"not spam").**

The **assumptions** are the following:

- $z^{(i)} \sim Bernoulli(\phi)$ → probability that document i comes from cluster 1 or 2.
- $x^{(i)} \sim Multinomial$**.**
- $P(x^{(i)}/z^{(i)}) = \prod_{i=1}^{n} P(x_j^{(i)}/z^{(i)})$ the appearance of words is independent from each other.
- $P\left(x_j^{(i)} = 1/z^{(i)} = 0\right) = \phi_{j/z=0}$

EXPECTATION MAXIMIZATION

EM AND NAÏVE BAYES

Computing the **EM algorithm** we **find** that:

- **E-Step:** find the posterior distribution (estimate where the document comes from).

$$\boldsymbol{w^{(i)}}_j = \boldsymbol{Q_i\left(z^{(i)}\right) := p(z^{(i)} = 1/x^{(i)}; \phi_{j/z}, \phi)}$$

- **M-Step:** maximize the lower bound .

$w^{(i)}$ captures uncertainty of cluster membership

$$\boldsymbol{\phi_{j/z=1} = \frac{\sum_{i=1}^{m} w^{(i)} 1\left\{x_j^{(i)} = 1\right\}}{\sum_{i=1}^{m} w^{(i)}} = \frac{\#\ times\ word\ j\ is\ in\ documents\ that\ we\ think\ are\ in\ cluster\ 1}{\#\ documents\ (estimated)\ we\ think\ are\ in\ cluster\ 1}}$$

$$\boldsymbol{\phi_{j/z=0} = \frac{\sum_{i=1}^{m} \left(1 - w^{(i)}\right) 1\left\{x_j^{(i)} = 1\right\}}{\sum_{i=1}^{m} (1 - w^{(i)})}}$$

$$\boldsymbol{\phi_{j/z=0} = \frac{\sum_{i=1}^{m} w^{(i)}}{m}}$$

FACTOR ANALYSIS

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

Given a training set $\{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)}\}$ where $\boldsymbol{x}^{(i)} \in \mathbb{R}^{\boldsymbol{n}}$ it would be **very difficult** to **model data** with a **mixture** of **Gaussians** when $\boldsymbol{n} \gg \boldsymbol{m}$.

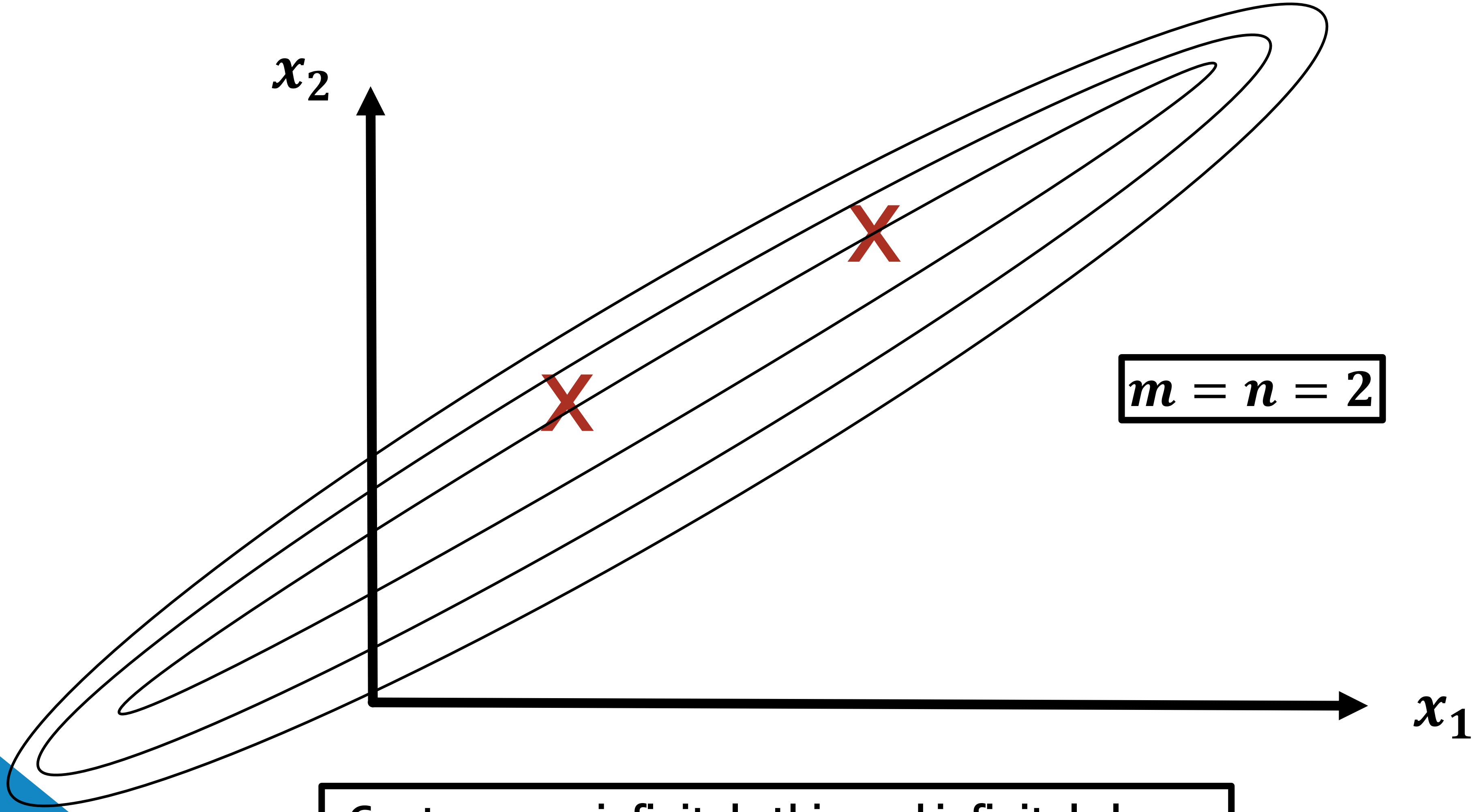
If we **model** the **data** as a **Gaussian**, and **estimate** the **mean** and **covariance** using the usual **maximum likelihood** estimators we would find that the **matrix $\boldsymbol{\Sigma}$** is **singular** $\left(\frac{\mathbf{1}}{|\boldsymbol{\Sigma}|^{\mathbf{1/2}}} = \frac{\mathbf{1}}{\mathbf{0}}\right)$.

Thus, the **maximum likelihood estimates** of the parameters **result** in a **Gaussian** that places **all** its **probability in** the affine **space spanned** by the **data,** resulting in a **singular covariance matrix**.

$$f_{\mathbf{X}}(x_1, \dots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS



$\boxed{m = n = 2}$

Contours are infinitely thin and infinitely large

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

Unless $\boldsymbol{m} \gg \boldsymbol{n}$, by a **reasonable amount**, the **maximum likelihood estimates** of the **mean** and **covariance** may be quite **poor**.

Factor analysis will be used when $\boldsymbol{m} \approx \boldsymbol{n}$ or when $\boldsymbol{n} \gg \boldsymbol{m}$. We are going to look at **two restrictions** on $\boldsymbol{\Sigma}$ that will **allow us** to **fit** $\boldsymbol{\Sigma}$ with **small amounts** of **data.**

1. **Constrain $\boldsymbol{\Sigma}$ to be diagonal.**

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}$$

2. **The diagonal entries must be equal: $\boldsymbol{\Sigma} = \sigma^2 I$.**

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

1. CONSTRAIN Σ TO BE DIAGONAL

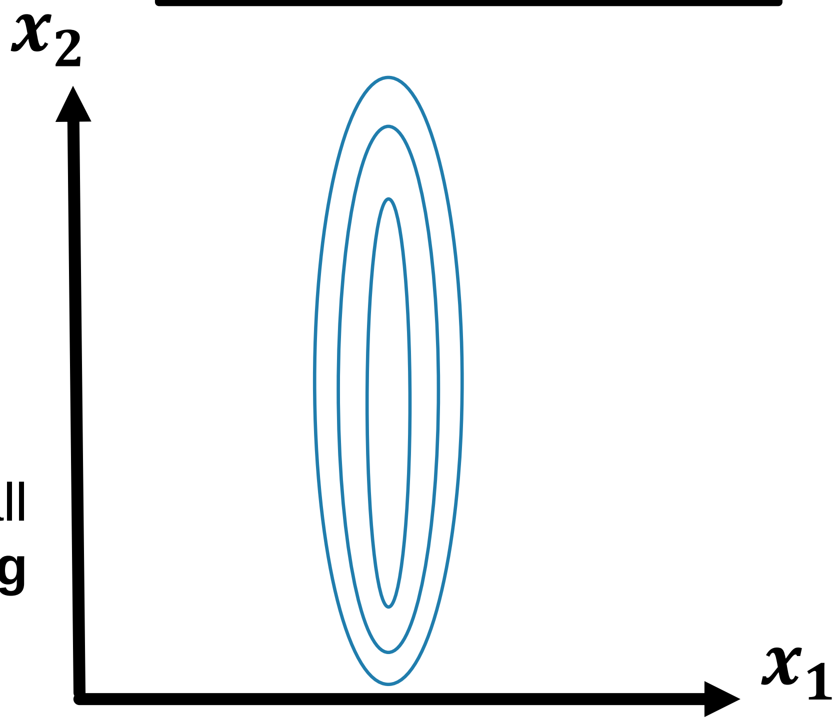
The **maximum likelihood** estimate **would be**:

$$\boldsymbol{\Sigma} = \frac{1}{m}\sum_{i=1}^{m}\left(x_j^{(i)} - \mu_j\right)\left(x_j^{(i)} - \mu_j\right)^T$$

$$\Sigma_{jj} = \frac{1}{m}\sum_{i=1}^{m}\left(x_j^{(i)} - \mu_j\right)^2$$

The **main problem** is that you are **removing** all **correlations** between **features**. → We are **assuming** that the **features** are **independent between them**.

Major axes are axis-aligned.



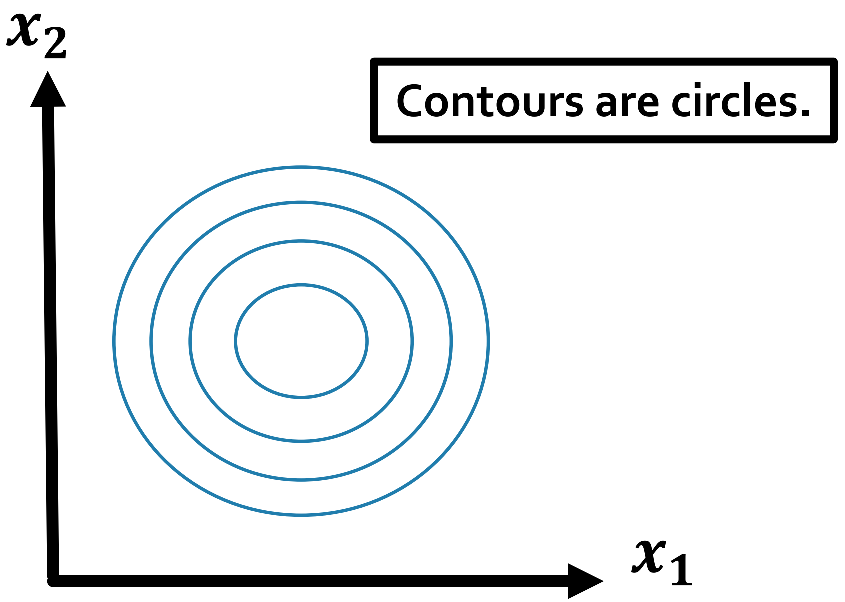
EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

2. THE DIAGONAL ENTRIES MUST BE EQUAL: $\Sigma = \sigma^2 I$.

The **maximum likelihood** estimate **would be**:

$$\sigma^2 = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\left(x_j^{(i)} - \mu_j\right)^2$$



EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

If we are **fitting** a **full**, **unconstrained**, covariance matrix $\boldsymbol{\Sigma}$ to data, it is necessary that $\boldsymbol{m} \geq \boldsymbol{n} + \mathbf{1}$ for the **maximum likelihood** estimate of $\boldsymbol{\Sigma}$ **not** to be **singular**.

Under either of the **two restrictions presented**, we may **obtain** a **non-singular** $\boldsymbol{\Sigma}$ when $\boldsymbol{n} \geq \mathbf{2}$.

The **problem** is that in **many occasions** we **want** to be able to **capture** some **interesting correlation structure** in the data.

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

MARGINALS AND CONDITIONALS OF GAUSSIANS

Before talking about the **factor analysis model**, we will discuss how to **find conditional** and **marginal distributions** of random variables with a **joint multivariate Gaussian distribution.**

Suppose we have a **vector-valued random variable**:

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \end{bmatrix}$$

Where $\boldsymbol{x_1} \in \mathbb{R}^r$, $\boldsymbol{x_2} \in \mathbb{R}^s$, and $\boldsymbol{x} \in \mathbb{R}^{r+s}$. **Suppose** $\boldsymbol{x} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{bmatrix}$$

$\boldsymbol{\mu_1} \in \mathbb{R}^r$, $\boldsymbol{\mu_2} \in \mathbb{R}^s$, $\boldsymbol{\Sigma_{11}} \in \mathbb{R}^{rxr}$, $\boldsymbol{\Sigma_{12}} \in \mathbb{R}^{rxs}$, $\boldsymbol{\Sigma_{21}} \in \mathbb{R}^{sxr}$ and $\boldsymbol{\Sigma_{22}} \in \mathbb{R}^{sxs}$.

$\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are jointly distributed multivariate Gaussian

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

MARGINALS AND CONDITIONALS OF GAUSSIANS

To **obtain** the **marginal distribution** of $\boldsymbol{x_1}$, we **can see** that:

$$\boldsymbol{E[x_1] = \mu_1}$$

$$\boldsymbol{Cov[x_1] = E[(x_1 - \mu_1)(x_1 - \mu_1)] = \Sigma_{11}}$$

To **demonstrate** the **previous statement**, we can see the **joint covariance** of $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$:

$$\boldsymbol{Cov[x] = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = E[(x - \mu)(x - \mu)^T]}$$

$$\boldsymbol{Cov[x] = E\left[\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T\right] = E\begin{bmatrix} (x_1 - \mu_1)(x_1 - \mu_1)^T & (x_1 - \mu_1)(x_2 - \mu_2)^T \\ (x_2 - \mu_2)(x_1 - \mu_1)^T & (x_2 - \mu_2)(x_2 - \mu_2)^T \end{bmatrix}}$$

Therefore the **marginal distribution** of $\boldsymbol{x_1}$ is $\boldsymbol{N(\mu_1, \Sigma_{11})}$.

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

MARGINALS AND CONDITIONALS OF GAUSSIANS

We can also **obtain** the **marginal conditional distribution** of x_1/x_2:

$$P(x_1/x_2) = \frac{P(x_1, x_2)}{P(x_2)} = \frac{N(\mu, \Sigma)}{N(\mu_2, \Sigma_{22})}$$

Substituting the **formulas** for **both gaussians**, the **joint** and the **marginal** of x_2, you would **obtain** the **following** (these **computations** are **non-trivial**):

$$x_1/x_2 \sim N\left(\mu_{1/2}, \Sigma_{1/2}\right)$$

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

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

EXPECTATION MAXIMIZATION
EM AND FACTOR ANALYSIS

FACTOR ANALYSIS MODEL

We will **create** a **joint distribution** $(\boldsymbol{x}, \boldsymbol{z})$ by **assuming** that:

- There is a **latent random variable** $\boldsymbol{z} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I})$**,** where $\boldsymbol{z} \in \mathbb{R}^{\boldsymbol{k}}$ such that $\boldsymbol{k} < \boldsymbol{m}$**.**
- $\boldsymbol{x}/\boldsymbol{z} \sim \boldsymbol{N}(\boldsymbol{\mu} + \boldsymbol{\Lambda z}, \boldsymbol{\Psi})$
- $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda z} + \boldsymbol{\varepsilon}$**, where** $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Psi})$

Therefore, the **parameters** of the **model** are:

- $\boldsymbol{\mu} \in \mathbb{R}^{\boldsymbol{n}}$**,** $\boldsymbol{\Lambda} \in \mathbb{R}^{\boldsymbol{nxk}}, \boldsymbol{\Psi} \in \mathbb{R}^{\boldsymbol{nxn}}$ and $\boldsymbol{\Psi}$ is diagonal (**usually** $\boldsymbol{k} < \boldsymbol{n}$).

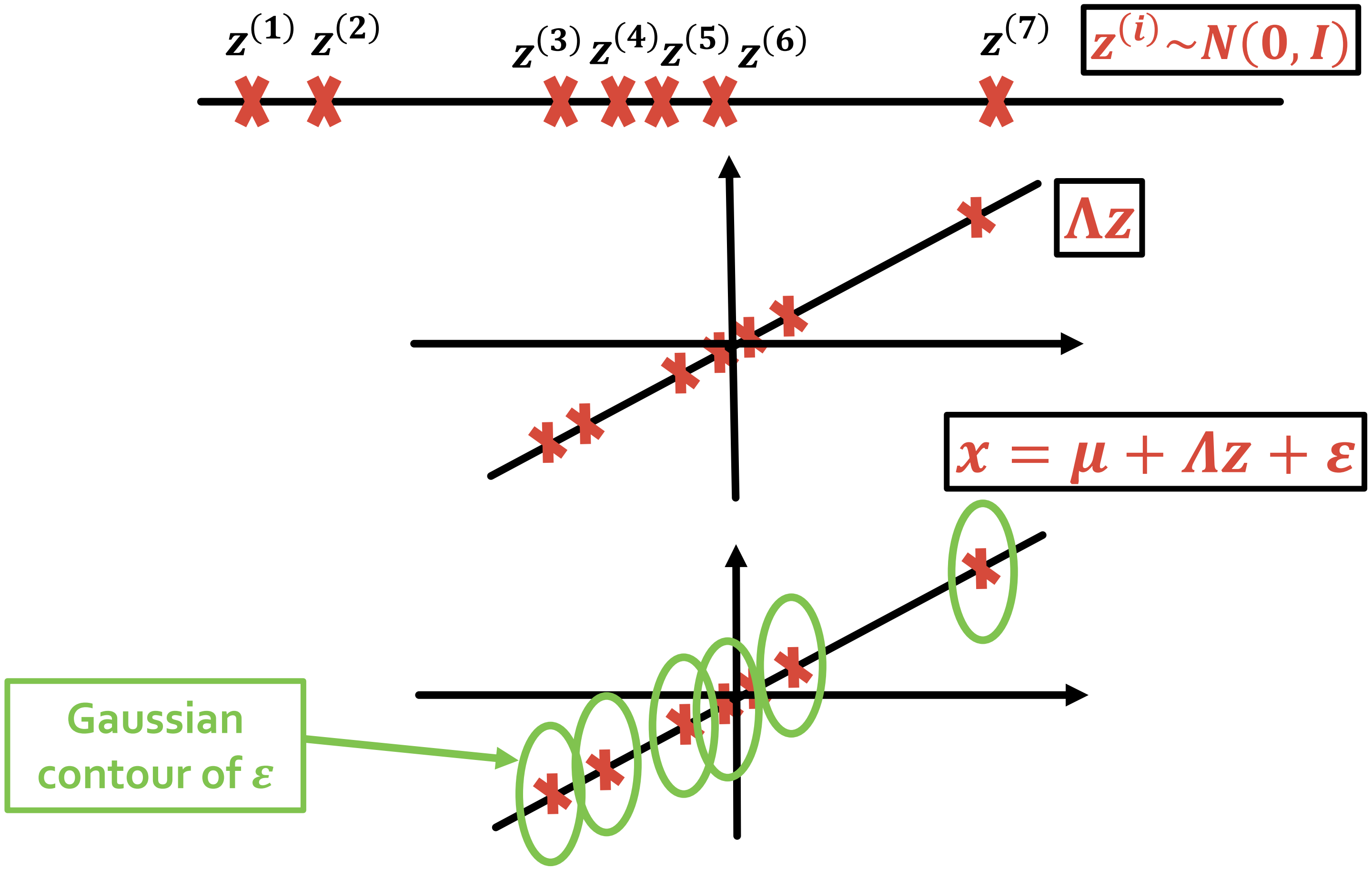
EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

FACTOR ANALYSIS MODEL

Let us give an **example**.

- $z \in \mathbb{R}^1$
- $x \in \mathbb{R}^2$
- $\Lambda = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$
- $\Psi = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$
- $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$



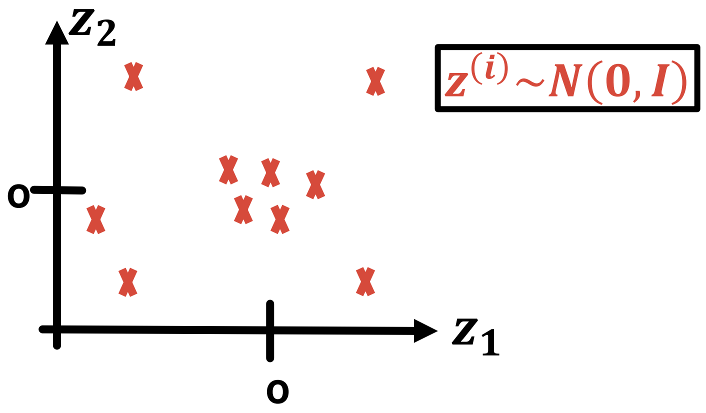
EXPECTATION MAXIMIZATION

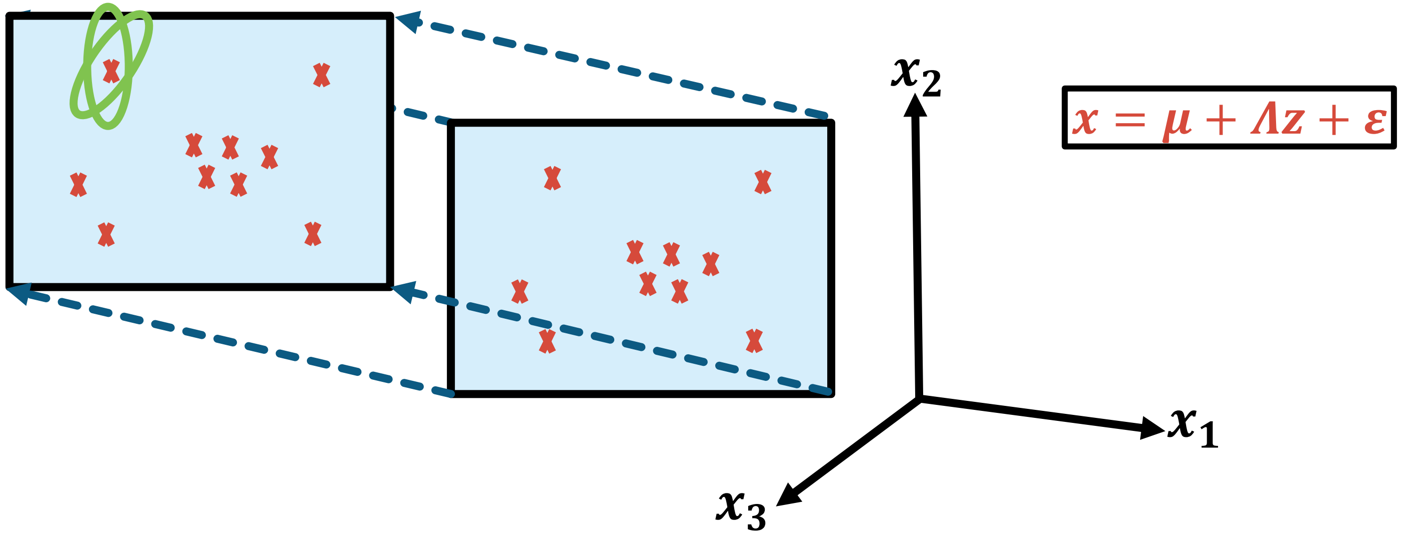
EM AND FACTOR ANALYSIS

FACTOR ANALYSIS MODEL

Let us give another **example**.

- $z \in \mathbb{R}^2$
- $x \in \mathbb{R}^3$





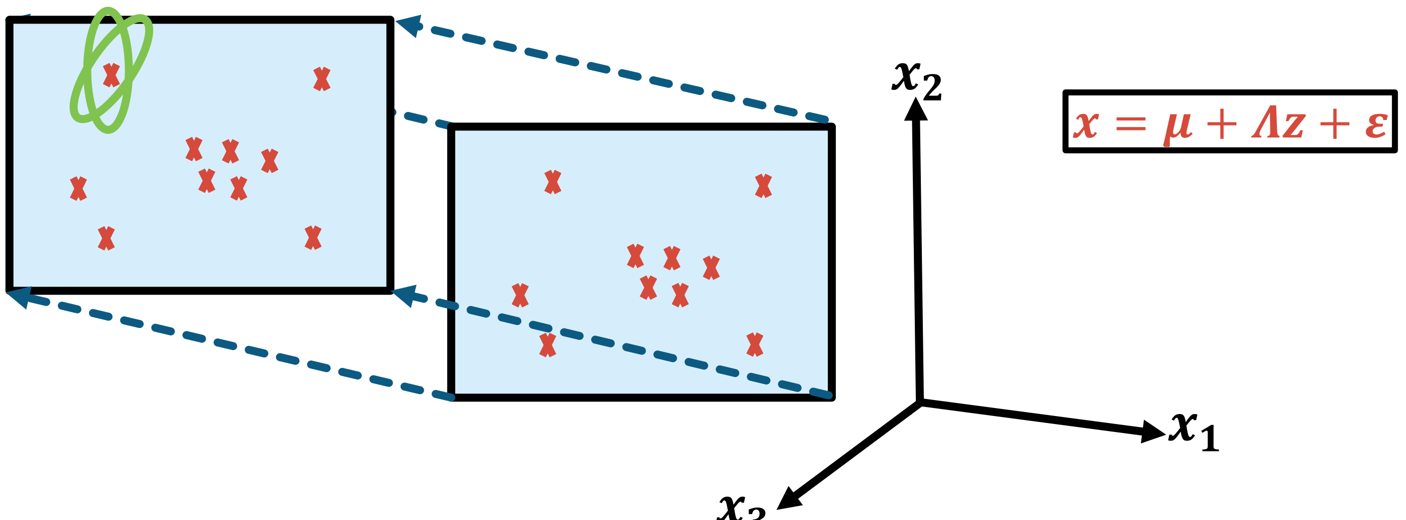
EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

FACTOR ANALYSIS MODEL

In as **summary**:

- Each **datapoint** $\boldsymbol{x}^{(i)}$ is **generated** by **sampling** a $\boldsymbol{k}$ **dimension multivariate Gaussian** $\boldsymbol{z}^{(i)}$.
- Then, it is **mapped** to a n**-dimensional affine space** of $\mathbb{R}^{\boldsymbol{n}}$ by **computing** $\mu + \boldsymbol{\Lambda z}^{(i)}$.
- Lastly, $\boldsymbol{x}^{(i)}$ is **generated** by **adding covariance** $\boldsymbol{\Psi}$ **noise** to $\mu + \boldsymbol{\Lambda z}^{(i)}$.



FACTOR ANALYSIS MODEL

We **will work out** exactly what **distribution our model defines** using **marginal** and **conditional distributions**.

Our **random variables** $\boldsymbol{z}$ and $\boldsymbol{x}$ have a **joint Gaussian distribution**:

$$\begin{bmatrix} \boldsymbol{z} \\ \boldsymbol{x} \end{bmatrix} \sim N(\boldsymbol{\mu}_{zx}, \boldsymbol{\Sigma})$$

We **want** to **find** $\boldsymbol{\mu}_{zx}$ and $\boldsymbol{\Sigma}$.

FACTOR ANALYSIS MODEL

To **find** $\boldsymbol{\mu_{zx}}$, let us **remember** that $\boldsymbol{E[z] = 0}$ because $\boldsymbol{z \sim N(0, I)}$.

Also we can **compute** the **expected value** of $\boldsymbol{x}$:

$$\boldsymbol{E[x] = E[\mu + \Lambda z + \varepsilon] = E[\mu] + \Lambda E[z] + E[\varepsilon]}$$

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

Therefore **we have** :

$$\boldsymbol{\mu_{zx} = \begin{bmatrix} E[z] \\ E[x] \end{bmatrix} = \begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}}$$

FACTOR ANALYSIS MODEL

To **find** $\boldsymbol{\Sigma}$, we need to compute:

- $\boldsymbol{\Sigma_{zz}} = \boldsymbol{E}[(\boldsymbol{z}-\boldsymbol{E}[\boldsymbol{z}])(\boldsymbol{z}-\boldsymbol{E}[\boldsymbol{z}])^{\boldsymbol{T}}]$
- $\boldsymbol{\Sigma_{zx}} = \boldsymbol{E}[(\boldsymbol{z}-\boldsymbol{E}[\boldsymbol{z}])(\boldsymbol{x}-\boldsymbol{E}[\boldsymbol{x}])^{\boldsymbol{T}}]$
- $\boldsymbol{\Sigma_{xx}} = \boldsymbol{E}[(\boldsymbol{x}-\boldsymbol{E}[\boldsymbol{x}])(\boldsymbol{x}-\boldsymbol{E}[\boldsymbol{x}])^{\boldsymbol{T}}]$
- $\boldsymbol{\Sigma_{xz}}$ (it is **not needed** because $\boldsymbol{\Sigma_{zx}}$ **and** $\boldsymbol{\Sigma_{zx}}$ are **symmetric** $\boldsymbol{\Sigma_{zx}} = \boldsymbol{\Sigma_{xz}}^{\boldsymbol{T}}$).

Since $\boldsymbol{z} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I})$ we have $\boldsymbol{\Sigma_{zz}} = \boldsymbol{I}$.

FACTOR ANALYSIS MODEL

Now we **find $\boldsymbol{\Sigma_{zx}}$**:

$$\boldsymbol{\Sigma_{zx}} = \boldsymbol{E}[(\boldsymbol{z} - \boldsymbol{E}[\boldsymbol{z}])(\boldsymbol{x} - \boldsymbol{E}[\boldsymbol{x}])^{\boldsymbol{T}}]$$

$$\boldsymbol{\Sigma_{zx}} = \boldsymbol{E}\left[\left(\boldsymbol{z} - \vec{\mathbf{0}}\right)(\boldsymbol{x} - \boldsymbol{\mu})^{\boldsymbol{T}}\right]$$

$$\boldsymbol{\Sigma_{zx}} = \mathbf{E}[(\boldsymbol{z})(\boldsymbol{x} - \boldsymbol{\mu})^{\boldsymbol{T}}]$$

$$\boldsymbol{\Sigma_{zx}} = \mathbf{E}[(\boldsymbol{z})(\boldsymbol{\mu} + \boldsymbol{\Lambda z} + \boldsymbol{\varepsilon} - \boldsymbol{\mu})^{\boldsymbol{T}}]$$

$$\boldsymbol{\Sigma_{zx}} = \boldsymbol{\Lambda}^{\boldsymbol{T}}\boldsymbol{E}[\boldsymbol{z}\boldsymbol{z}^{\boldsymbol{T}}] + \boldsymbol{E}[\boldsymbol{z}\boldsymbol{\varepsilon}^{\boldsymbol{T}}]$$

Because $\boldsymbol{z}$ and $\boldsymbol{\varepsilon}$ are **independent** $\boldsymbol{E}[\boldsymbol{z}\boldsymbol{\varepsilon}^{\boldsymbol{T}}] = \boldsymbol{E}[\boldsymbol{z}]\boldsymbol{E}[\boldsymbol{\varepsilon}^{\boldsymbol{T}}] = \mathbf{0}$. Also we have $\boldsymbol{E}[\boldsymbol{z}\boldsymbol{z}^{\boldsymbol{T}}] = \boldsymbol{Cov}(\boldsymbol{z}) = \boldsymbol{I}$:

$$\boldsymbol{\Sigma_{zx}} = \boldsymbol{\Lambda}^{\boldsymbol{T}}$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

FACTOR ANALYSIS MODEL

Now we **find $\boldsymbol{\Sigma_{xx}}$**:

$$\boldsymbol{\Sigma_{xx} = E[(\mu + \Lambda z + \varepsilon - \mu)(\mu + \Lambda z + \varepsilon - \mu)^T]}$$

$$\boldsymbol{\Sigma_{xx} = E[(\Lambda z + \varepsilon)(\Lambda z + \varepsilon)^T]}$$

$$\boldsymbol{\Sigma_{xx} = E[\Lambda z z^T \Lambda^T + \varepsilon z^T \Lambda^T + \Lambda z \varepsilon^T + \varepsilon\varepsilon^T]}$$

$$\boldsymbol{\Sigma_{xx} = E[\Lambda z z^T \Lambda^T] + E[\varepsilon z^T \Lambda^T] + E[\Lambda z \varepsilon^T] + E[\varepsilon\varepsilon^T]}$$

$$\boldsymbol{\Sigma_{xx} = E[\Lambda z z^T \Lambda^T] + \Lambda^T E[\varepsilon] E[z^T] + \Lambda E[z] E[\varepsilon^T] + E[\varepsilon\varepsilon^T]}$$

$$\boldsymbol{\Sigma_{xx} = \Lambda E[z z^T] \Lambda^T + 0 + 0 + Cov(\varepsilon)}$$

$$\boldsymbol{\Sigma_{xx} = \Lambda\Lambda^T + \Psi}$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

FACTOR ANALYSIS MODEL

With $\boldsymbol{\Sigma}_{xx}$, $\boldsymbol{\Sigma}_{zz}$, $\boldsymbol{\Sigma}_{zx}$, and $\boldsymbol{\Sigma}_{xz}$ we can now build $\boldsymbol{\Sigma}$**:**

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{zz} & \boldsymbol{\Sigma}_{zx} \\ \boldsymbol{\Sigma}_{xz} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{\Lambda}^T \\ \boldsymbol{\Lambda} & \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi} \end{bmatrix}$$

Therefore, the **joint distribution** of $(\boldsymbol{z}, \boldsymbol{x})$ is **defined** as:

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

And the **marginal distribution** of $\boldsymbol{x}$ would be:

$$\boldsymbol{x} \sim \boldsymbol{N}\left(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}\right)$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

FACTOR ANALYSIS MODEL

The **parameters** of our **model** would be $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\Psi}$.

Therefore, **given** a **training set** $\{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)}\}$ we would like to **make maximum likelihood estimation** for the **parameters**.

$$\boldsymbol{l(\mu, \Lambda, \Psi) = \log \prod_{i=1}^{m} P\left(x^{(i)}\right) = log \prod_{i=1}^{m} Gaussian\left(\mu, \Lambda\Lambda^{T} + \Psi\right)}$$

The **procedure** would be the **same**, **take derivatives** with **respect** to **each parameter**, **equal** to **0** and **solve** for **each parameter**.

But **maximizing** this **formula explicitly** is **hard**, we **have not found** an **algorithm** that **does it** in **closed-form**.

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: E-STEP

We will use the **EM algorithm** instead of **maximum likelihood estimation**.

We need to compute $Q_i(z^{(i)}) = p(z^{(i)}/x^{(i)}; \mu, \Lambda, \Psi)$. We **already know how** to **compute** a **conditional probability** for a **Gaussian**.

$$z^{(i)}/x^{(i)} \sim N\left(\mu_{z^{(i)}/x^{(i)}}, \Sigma_{z^{(i)}/x^{(i)}}\right)$$

$$\mu_{1/2} = \mu_1 + \Sigma_{12}\Sigma_{22}{}^{-1}(x_2 - \mu_2) \rightarrow \mu_{z^{(i)}/x^{(i)}} = \Lambda^T\left(\Lambda\Lambda^T + \Psi\right)^{-1}\left(x^{(i)} - \mu\right)$$

$$\Sigma_{1/2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}{}^{-1}\Sigma_{21} \rightarrow \Sigma_{z^{(i)}/x^{(i)}} = I - \Lambda^T\left(\Lambda\Lambda^T + \Psi\right)^{-1}\Lambda$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: E-STEP

Therefore, we have that distribution $\boldsymbol{Q_i}\left(\boldsymbol{z}^{(i)}\right)$ is defined as:

$$\boldsymbol{z}^{(i)}/\boldsymbol{x}^{(i)} \sim \boldsymbol{N}\left(\boldsymbol{\mu}_{\boldsymbol{z}^{(i)}/\boldsymbol{x}^{(i)}}, \boldsymbol{\Sigma}_{\boldsymbol{z}^{(i)}/\boldsymbol{x}^{(i)}}\right)$$

$$Q_i(z^{(i)}) = \frac{1}{(2\pi)^{k/2}|\Sigma_{z^{(i)}|x^{(i)}}|^{1/2}} \exp\left(-\frac{1}{2}(z^{(i)} - \mu_{z^{(i)}|x^{(i)}})^T \Sigma^{-1}_{z^{(i)}|x^{(i)}} (z^{(i)} - \mu_{z^{(i)}|x^{(i)}})\right)$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: M-STEP

Remember that the **joint log likelihood** was **expressed** as **follows.**

$$l(w) = \sum_{i=1}^{m} \sum_{z} Q_i\left(z^{(i)}\right) log\left(\frac{p\left(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi\right)}{Q_i\left(z^{(i)}\right)}\right)$$

Because now, z is a **continuous variable,** we have that the **joint log likelihood** will be **defined** as**:**

$$l(w) = \sum_{i=1}^{m} log \int_{z^{(i)}} Q_i\left(z^{(i)}\right) log\left(\frac{p\left(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi\right)}{Q_i\left(z^{(i)}\right)}\right) dz^{(i)}$$

We **want** to **maximize** this **expression.**

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: M-STEP

We will **expand** the **logarithm.**

$$l(w) = \sum_{i=1}^{m} \int_{z^{(i)}} Q_i(z^{(i)}) \log\left(\frac{p(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi)}{Q_i(z^{(i)})}\right) dz^{(i)}$$

$$l(w) = \sum_{i=1}^{m} \int_{z^{(i)}} Q_i(z^{(i)}) \left[\log\left(p(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi)\right) - \log\left(Q_i(z^{(i)})\right)\right] dz^{(i)}$$

We **know** that $p(x^{(i)}, z^{(i)}) = p(x^{(i)}/z^{(i)})p(z^{(i)})$:

$$l(w) = \sum_{i=1}^{m} \int_{z^{(i)}} Q_i(z^{(i)}) \left[\log\left(p(x^{(i)}/z^{(i)}; \mu, \Lambda, \Psi)p(z^{(i)})\right) - \log\left(Q_i(z^{(i)})\right)\right] dz^{(i)}$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: M-STEP

Expanding the **equation** we have:

$$l(w) = \sum_{i=1}^{m} \int_{z^{(i)}} Q_i\left(z^{(i)}\right) \left[log\left(p\left(x^{(i)}/z^{(i)}; \mu, \Lambda, \Psi\right) p\left(z^{(i)}\right)\right) - log\left(Q_i\left(z^{(i)}\right)\right)\right] dz^{(i)}$$

$$l(w) = \sum_{i=1}^{m} \int_{z^{(i)}} Q_i\left(z^{(i)}\right) \left[log\left(p\left(x^{(i)}/z^{(i)}; \mu, \Lambda, \Psi\right)\right) + log\left(p\left(z^{(i)}\right)\right) - log\left(Q_i\left(z^{(i)}\right)\right)\right] dz^{(i)}$$

Applying the **definition** of the **expected** value of $z^{(i)}$ **under** the **distribution** Q_i :

$$E_{z^{(i)} \sim Q_i}\left[z^{(i)}\right] = \int_{z^{(i)}} Q_i\left(z^{(i)}\right) z^{(i)} dz^{(i)}$$

$$l(w) = \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i} \left[log\left(p\left(x^{(i)}/z^{(i)}; \mu, \Lambda, \Psi\right)\right) + log\left(p\left(z^{(i)}\right)\right) - log\left(Q_i\left(z^{(i)}\right)\right)\right]$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: M-STEP

$$l(w) = \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[\log\left(p\left(x^{(i)}/z^{(i)}; \mu, \Lambda, \Psi\right)\right) + \log\left(p\left(z^{(i)}\right)\right) - \log\left(Q_i\left(z^{(i)}\right)\right)\right]$$

Substituting the **distributions** we notice that the **only term depending on** the **parameters** is:

$$\log\left(p\left(x^{(i)}/z^{(i)}; \mu, \Lambda, \Psi\right)\right).$$

Notice that $Q_i\left(z^{(i)}\right)$ is a **fixed Gaussian** (**known parameters** that **resulted from** the **previous maximization** or **initialization**).

$$l(w) = \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[\log\left(p\left(x^{(i)}/z^{(i)}; \mu, \Lambda, \Psi\right)\right) + \log\left(Gaussian(\vec{0}, I)\right) - \log\left(Gaussian(\mu_{z^{(i)}/x^{(i)}}, \Sigma_{z^{(i)}/x^{(i)}})\right)\right]$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: M-STEP

We **drop** the **terms not depending** on the **parameters:**

$$l(w) = \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[log\left(p\left(x^{(i)}/z^{(i)}; \mu, \Lambda, \Psi\right)\right) + log\left(Gaussian(\vec{0}, I)\right) - log\left(Gaussian\left(\mu_{z^{(i)}/x^{(i)}}, \Sigma_{z^{(i)}/x^{(i)}}\right)\right)\right]$$

$$l(w) = \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[log\left(p\left(x^{(i)}/z^{(i)}; \mu, \Lambda, \Psi\right)\right)\right]$$

Where **we know** that:

$$x^{(i)}/z^{(i)} \sim Gaussian\left(\mu_{x^{(i)}/z^{(i)}}, \Sigma_{x^{(i)}/z^{(i)}}\right)$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: M-STEP

We **obtain** the **mean** and **covariance matrix** of the **distribution** of $x^{(i)}/z^{(i)}$:

$$x^{(i)}/z^{(i)} \sim Gaussian(\mu_{x^{(i)}/z^{(i)}}, \Sigma_{x^{(i)}/z^{(i)}})$$

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

$$\mu_{2/1} = \mu_2 + \Sigma_{21}\Sigma_{11}{}^{-1}(x_1 - \mu_1) \rightarrow \mu_{x^{(i)}/z^{(i)}} = \mu + \Lambda I\left(z^{(i)} - \vec{0}\right) = \mu + \Lambda I z^{(i)}$$

$$\Sigma_{2/1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}{}^{-1}\Sigma_{12} \rightarrow \Sigma_{x^{(i)}/z^{(i)}} = \Lambda\Lambda^T + \Psi - \left(\Lambda I^{-1}\Lambda^T\right) = \Psi$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: M-STEP

Substituting the **mean** and **covariance matrix** in the **function** that we **want** to **maximize**:

$$l(w) = \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[log\left(p\left(x^{(i)}/z^{(i)}; \mu, \Lambda, \Psi\right)\right)\right]$$

$$l(w) = \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[log\left(Gaussian(\mu_{x^{(i)}/z^{(i)}}, \Sigma_{x^{(i)}/z^{(i)}})\right)\right]$$

$$l(w) = \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[log\left(Gaussian(\mu + \Lambda z^{(i)}, \Psi)\right)\right]$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: M-STEP

Representing the **explicit form** of the **Gaussian:**

$$l(w) = \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[log\left(Gaussian(\mu + \Lambda z^{(i)}, \Psi)\right)\right]$$

$$l(w) = \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[log\left(\frac{1}{\sqrt{2\pi}|\Psi|^{1/2}} exp\left(-\frac{1}{2}\left(x^{(i)} - \mu - \Lambda z^{(i)}\right)^T \Psi^{-1}\left(x^{(i)} - \mu - \Lambda z^{(i)}\right)\right)\right)\right]$$

$$l(w) = \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[-\frac{1}{2} log(|\Psi|) - \frac{1}{2} log(2\pi) - \frac{1}{2}\left[\left(x^{(i)} - \mu - \Lambda z^{(i)}\right)^T \Psi^{-1}\left(x^{(i)} - \mu - \Lambda z^{(i)}\right)\right]\right]$$

$$l(w) = \sum_{i=1}^{m} -E_{z^{(i)} \sim Q_i}\left[\frac{1}{2} log(|\Psi|) + \frac{1}{2}\left[\left(x^{(i)} - \mu - \Lambda z^{(i)}\right)^T \Psi^{-1}\left(x^{(i)} - \mu - \Lambda z^{(i)}\right)\right]\right]$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: M-STEP

Maximizing with **respect to Λ :**

$$\nabla_{\Lambda} \sum_{i=1}^{m} -E_{z^{(i)} \sim Q_i}\left[\frac{1}{2} log(|\Psi|) + \frac{1}{2}\left[\left(x^{(i)} - \mu - \Lambda z^{(i)}\right)^T \Psi^{-1}\left(x^{(i)} - \mu - \Lambda z^{(i)}\right)\right]\right] = 0$$

$$\nabla_{\Lambda} \sum_{i=1}^{m} -E_{z^{(i)} \sim Q_i}\left[\frac{1}{2}\left[\left(x^{(i)} - \mu - \Lambda z^{(i)}\right)^T \Psi^{-1}\left(x^{(i)} - \mu - \Lambda z^{(i)}\right)\right]\right] = 0$$

Distributing operations and getting only **results depending only** on Λ we have**:**

$$\nabla_{\Lambda} \sum_{i=1}^{m} -E_{z^{(i)} \sim Q_i}\left[\frac{1}{2}\left[-x^{(i)^T} \Psi^{-1} \Lambda z^{(i)} + \mu^T \Psi^{-1} \Lambda z^{(i)} - z^{(i)^T} \Lambda^T \Psi^{-1} x^{(i)} + z^{(i)^T} \Lambda^T \Psi^{-1} \mu + z^{(i)^T} \Lambda^T \Psi^{-1} \Lambda z^{(i)}\right]\right] = 0$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: M-STEP

Reducing terms:

$$\nabla_{\Lambda} \sum_{i=1}^{m} -E_{z^{(i)} \sim Q_i}\left[\frac{1}{2}\left[-x^{(i)^T}\Psi^{-1}\Lambda z^{(i)} + \mu^T\Psi^{-1}\Lambda z^{(i)} - z^{(i)^T}\Lambda^T\Psi^{-1}x^{(i)} + z^{(i)^T}\Lambda^T\Psi^{-1}\mu + z^{(i)^T}\Lambda^T\Psi^{-1}\Lambda z^{(i)}\right]\right] = 0$$

$$\nabla_{\Lambda} \sum_{i=1}^{m} -E_{z^{(i)} \sim Q_i}\left[\frac{1}{2}\left[z^{(i)^T}\Lambda^T\Psi^{-1}\Lambda z^{(i)} + 2z^{(i)^T}\Lambda^T\Psi^{-1}\mu - 2z^{(i)^T}\Lambda^T\Psi^{-1}x^{(i)}\right]\right] = 0$$

$$\nabla_{\Lambda} \sum_{i=1}^{m} -E_{z^{(i)} \sim Q_i}\left[\frac{1}{2}\left[z^{(i)^T}\Lambda^T\Psi^{-1}\Lambda z^{(i)} + 2z^{(i)^T}\Lambda^T\Psi^{-1}\left(\mu - x^{(i)}\right)\right]\right] = 0$$

$$\nabla_{\Lambda} \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[-\frac{1}{2}z^{(i)^T}\Lambda^T\Psi^{-1}\Lambda z^{(i)} + z^{(i)^T}\Lambda^T\Psi^{-1}\left(x^{(i)} - \mu\right)\right] = 0$$

EM FOR FACTOR ANALYSIS: M-STEP

We **apply** the **following properties** $\nabla_x b^T x = b$ and $\nabla_x x^T A x = 2Ax$ **:**

$$\nabla_\Lambda \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[-\frac{1}{2} z^{(i)^T} \Lambda^T \Psi^{-1} \Lambda z^{(i)} + z^{(i)^T} \Lambda^T \Psi^{-1}\left(x^{(i)} - \mu\right)\right] = 0$$

$$\nabla_\Lambda \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[-\frac{1}{2} \Lambda^T \Psi^{-1} z^{(i)} z^{(i)^T} \Lambda + \Psi^{-1}\left(x^{(i)} - \mu\right) z^{(i)^T} \Lambda\right] = 0$$

$$\sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[-\frac{1}{2} 2\left(\Psi^{-1} z^{(i)} z^{(i)^T} \Lambda\right) + \Psi^{-1}\left(x^{(i)} - \mu\right) z^{(i)^T}\right] = 0$$

$$\sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[-\Psi^{-1} z^{(i)} z^{(i)^T} \Lambda + \Psi^{-1}\left(x^{(i)} - \mu\right) z^{(i)^T}\right] = 0$$

EM FOR FACTOR ANALYSIS: M-STEP

We **distribute** the **expected value** and the **sum**:

$$\sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\left[-\Psi^{-1} z^{(i)} {z^{(i)}}^T \Lambda + \Psi^{-1}\left(x^{(i)} - \mu\right) {z^{(i)}}^T\right] = 0$$

$$-\sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\left[\Psi^{-1} z^{(i)} {z^{(i)}}^T \Lambda\right] + \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\left[\Psi^{-1}\left(x^{(i)} - \mu\right) {z^{(i)}}^T\right] = 0$$

$$\Psi^{-1}\Lambda \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\left[z^{(i)} {z^{(i)}}^T\right] = \Psi^{-1} \sum_{i=1}^{m} \left(x^{(i)} - \mu\right) E_{z^{(i)}\sim Q_i}\left[{z^{(i)}}^T\right]$$

$$\Lambda \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\left[z^{(i)} {z^{(i)}}^T\right] = \sum_{i=1}^{m} \left(x^{(i)} - \mu\right) E_{z^{(i)}\sim Q_i}\left[{z^{(i)}}^T\right]$$

EM FOR FACTOR ANALYSIS: M-STEP

We **solve** for $\boldsymbol{\Lambda}$:

$$\boldsymbol{\Lambda}\sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\left[z^{(i)}{z^{(i)}}^T\right] = \sum_{i=1}^{m}\left(x^{(i)}-\mu\right)E_{z^{(i)}\sim Q_i}\left[{z^{(i)}}^T\right]$$

$$\boldsymbol{\Lambda} = \left(\sum_{i=1}^{m}\left(x^{(i)}-\mu\right)E_{z^{(i)}\sim Q_i}\left[{z^{(i)}}^T\right]\right)\left(\sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\left[z^{(i)}{z^{(i)}}^T\right]\right)^{-1}$$

It is **interesting** to **note** the **close relationship between this equation** and the **normal equation** that we'd **derived** for **least squares regression:**

$$w^T = \left(y^T X\right)\left(X^T X\right)^{-1}$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: M-STEP

The **analogy** is that **here**, the **x's** are a **linear function of** the **z's** (plus noise).

Given the "**guesses**" for **z** that the **E-step has found**, we will now try to **estimate** the **unknown linearity** $\boldsymbol{\Lambda}$ **relating** the **x's** and **z's**.

$$\boldsymbol{w}^T = \left(\boldsymbol{y}^T\boldsymbol{X}\right)\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$$

$$\boldsymbol{\Lambda} = \left(\sum_{i=1}^{m}\left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}\right)E_{\boldsymbol{z}^{(i)}\sim Q_i}\left[\boldsymbol{z}^{(i)^T}\right]\right)\left(\sum_{i=1}^{m}E_{\boldsymbol{z}^{(i)}\sim Q_i}\left[\boldsymbol{z}^{(i)}\boldsymbol{z}^{(i)^T}\right]\right)^{-1}$$

The **final step** is to **work out** the **expectations** $E_{\boldsymbol{z}^{(i)}\sim Q_i}\left[\boldsymbol{z}^{(i)^T}\right]$ and $E_{\boldsymbol{z}^{(i)}\sim Q_i}\left[\boldsymbol{z}^{(i)}\boldsymbol{z}^{(i)^T}\right]$.

EM FOR FACTOR ANALYSIS: M-STEP

The **expectation** $E_{z^{(i)} \sim Q_i}\left[{z^{(i)}}^T\right]$ we **already have it** from our **previous definition**:

$$E_{z^{(i)} \sim Q_i}\left[{z^{(i)}}^T\right] = {\mu_{z^{(i)}/x^{(i)}}}^T$$

$$E_{z^{(i)} \sim Q_i}\left[{z^{(i)}}^T\right] = \left(\Lambda^T\left(\Lambda\Lambda^T + \Psi\right)^{-1}\left(x^{(i)} - \mu\right)\right)^T$$

For the **expectation** $E_{z^{(i)} \sim Q_i}\left[z^{(i)}{z^{(i)}}^T\right]$, we **have** the following **definition**:

$$Cov(Y) = E[YY^T] - E[Y]E[Y]^T$$

$$E_{z^{(i)} \sim Q_i}\left[z^{(i)}{z^{(i)}}^T\right] = \mu_{z^{(i)}/x^{(i)}}\,{\mu_{z^{(i)}/x^{(i)}}}^T + \Sigma_{z^{(i)}/x^{(i)}}$$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: M-STEP

Substituting both expectations back in the **expression** of $\boldsymbol{\Lambda}$ we have:

$$\boldsymbol{\Lambda} = \left(\sum_{i=1}^{m} (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}) E_{z^{(i)} \sim Q_i}\left[\boldsymbol{z}^{(i)^T} \right]\right)\left(\sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[\boldsymbol{z}^{(i)} \boldsymbol{z}^{(i)^T} \right]\right)^{-1}$$

$$\boldsymbol{\Lambda} = \left(\sum_{i=1}^{m} (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}) \boldsymbol{\mu}_{z^{(i)}/x^{(i)}}{}^T\right)\left(\sum_{i=1}^{m} \boldsymbol{\mu}_{z^{(i)}/x^{(i)}} \boldsymbol{\mu}_{z^{(i)}/x^{(i)}}{}^T + \boldsymbol{\Sigma}_{z^{(i)}/x^{(i)}}\right)^{-1}$$

In this **last equation** $\boldsymbol{\Sigma}_{z^{(i)}/x^{(i)}}$ is the **covariance matrix** of the **posterior**, which **represents** the **uncertainty** of our **estimations**.

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: M-STEP

Now**,** that we have **the parameter $\boldsymbol{\Lambda}$,** we want to **maximize** with **respect to $\boldsymbol{\mu}$**:

$$\nabla_{\boldsymbol{\mu}} \sum_{i=1}^{m} -E_{z^{(i)} \sim Q_i}\left[\frac{1}{2}\left[\left(x^{(i)} - \boldsymbol{\mu} - \boldsymbol{\Lambda} z^{(i)}\right)^T \left(x^{(i)} - \boldsymbol{\mu} - \boldsymbol{\Lambda} z^{(i)}\right)\right]\right] = \mathbf{0}$$

To **save time** in the **computations**, the **result** of the **maximization** for is $\boldsymbol{\mu}$:

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

An **important thing** is that **this parameter does not change as** the **parameters vary**, which is **different for** the **parameter $\boldsymbol{\Lambda}$** which we **have just calculated**.

Therefore**, $\boldsymbol{\mu}$** can be **calculated just once** and **needs not** be **further updated as** the **algorithm** is **run**.

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: M-STEP

Finally, if you **maximize** for $\boldsymbol{\Psi}$ you **obtain** the **following matrix**:

$$\boldsymbol{\Phi} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)^{T}} - \boldsymbol{x}^{(i)}\boldsymbol{\mu}_{z^{(i)}/x^{(i)}}{}^{T}\boldsymbol{\Lambda}^{T} - \boldsymbol{\Lambda}\boldsymbol{\mu}_{z^{(i)}/x^{(i)}}\boldsymbol{x}^{(i)^{T}} + \boldsymbol{\Lambda}\left(\boldsymbol{\mu}_{z^{(i)}/x^{(i)}}\boldsymbol{\mu}_{z^{(i)}/x^{(i)}}{}^{T} + \boldsymbol{\Sigma}_{z^{(i)}/x^{(i)}}\right)\boldsymbol{\Lambda}^{T}$$

The **diagonal** of $\boldsymbol{\Psi}$ is **obtained** by **setting** $\boldsymbol{\Psi}_{ii} = \boldsymbol{\Phi}_{ii}$

EXPECTATION MAXIMIZATION

EM AND FACTOR ANALYSIS

EM FOR FACTOR ANALYSIS: SUMMARY

1. **Initialize parameters** $\boldsymbol{\mu}, \boldsymbol{\Lambda},$ and $\boldsymbol{\Psi}$.

2. **Estimate** the **posterior distribution** $Q_i(z^{(i)})$:

$$\mu_{z^{(i)}/x^{(i)}} = \Lambda^T\left(\Lambda\Lambda^T + \Psi\right)^{-1}\left(x^{(i)} - \mu\right)$$

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

$$z^{(i)}/x^{(i)} \sim N\left(\mu_{z^{(i)}/x^{(i)}}, \Sigma_{z^{(i)}/x^{(i)}}\right)$$

$$Q_i(z^{(i)}) = \frac{1}{(2\pi)^{k/2}|\Sigma_{z^{(i)}|x^{(i)}}|^{1/2}} \exp\left(-\frac{1}{2}(z^{(i)} - \mu_{z^{(i)}|x^{(i)}})^T \Sigma_{z^{(i)}|x^{(i)}}^{-1} (z^{(i)} - \mu_{z^{(i)}|x^{(i)}})\right)$$

3. **Compute parameters** and **go back** to **step 2** (**repeat** until **convergence**):

$$E_{z^{(i)} \sim Q_i}\left[{z^{(i)}}^T\right] = {\mu_{z^{(i)}/x^{(i)}}}^T$$

$$E_{z^{(i)} \sim Q_i}\left[{z^{(i)}}^T\right] = \left(\Lambda^T\left(\Lambda\Lambda^T + \Psi\right)^{-1}\left(x^{(i)} - \mu\right)\right)^T$$

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

Computed only one time

$$\Lambda = \left(\sum_{i=1}^{m}(x^{(i)} - \mu)E_{z^{(i)} \sim Q_i}\left[{z^{(i)}}^T\right]\right)\left(\sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[z^{(i)}{z^{(i)}}^T\right]\right)^{-1}$$

$$\Psi_{ii} = \Phi_{ii}$$

Check previous slide

EM FOR FACTOR ANALYSIS: SUMMARY

To **compute** the **probability** of a **new sample** we would have $\boldsymbol{x^{(i)}}$**:**

$$\boldsymbol{x/z \sim N(\mu + \Lambda z, \Psi)}$$

$$\boldsymbol{p(x) = \frac{1}{\sqrt{2\pi}|\Psi|^{1/2}} exp\left(-\frac{1}{2}\left(x^{(i)} - \mu - \Lambda z^{(i)}\right)^T \Psi^{-1}\left(x^{(i)} - \mu - \Lambda z^{(i)}\right)\right)}$$