Applications of EM & Factor Analysis

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

Recall EM Algorithm

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

Last time

$\ell(\theta) = \sum_i \log \sum_z q_i(z) \frac{P(x^{(i)} | z; \theta)}{q_i(z)}$

$\forall i \in I, q_i(z) = 1, q_i(z) > 0$

$L_t(\theta) = \sum_i \sum_z q_i(z) \log \frac{P(x^{(i)} | z; \theta)}{q_i(z)}$

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

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

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

To see Property 2, we pick $q_i(z) \in o^{(i)}$

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

Verify, since $P(x^{(i)} | z; \theta) = \frac{P(z | x^{(i)}; \theta) P(x^{(i)}; \theta)}{P(z | x^{(i)}; \theta)}$

$\ell(\theta)$

1. $\sum_i \log \sum_z q_i(z) \cdot c = \log c (\text{LHS})$

2. $\sum_i q_i(z) \log q_i(z) = \log c$ (ELBO term) Property 2 holds.
**Restate EM**

(E-step) For \( i = 1 \ldots n \) set \( Q_i(x_i) = P(z^{(i)} | x_i, \theta) \)

(M-step) \( \theta^{(t+1)} = \text{argmax}_\theta \sum_i \log P(x_i; \theta) \quad \text{data & input params} \)

\( \theta^{(t+1)} = \text{argmax}_\theta \sum_i \text{ELBO}(x_i, \theta^{(t)}, \theta) \)

**Warm Up: Mixture of Gaussians.**

EM RECOVERS OUR AGG EC AlGORITHM

\[
P(x^{(i)}, z^{(i)}) = P(x^{(i)} | z^{(i)}) P(z^{(i)})
\]

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

\( x^{(i)} | z^{(i)} = j \sim N(\mu_j, \sigma^2_j) \quad \text{"Cluster means"} \)

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

**What is EM here?**

\[
Q_i(z) = P(z = j | x^{(i)}, \theta)
\]

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...

\[
\max_{\phi, \mu, \sigma^2} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)} | z^{(i)}; \theta)}{P_i(z^{(i)})}
\]
\[ w_j^{(0)} = \mathcal{O}(\mathbf{x}^{(i)} - \mu_j) \]
\[ p(x^{(i)}, z^{(i)}; \theta) = p(x^{(i)} | 1z^{(i)}) p(z^{(i)}) \]
\[ f_i(\theta) = \sum_j w_j^{(0)} \log \left( \frac{1}{\sqrt{2\pi} \sigma_j} \exp \left[ -\frac{1}{2} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right] \right) \]
\[ \nabla \theta \ E f_i(\theta) = \sum_i \nabla \phi_j \left( w_j^{(0)} - \frac{1}{2} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right) \]
\[ = -\frac{1}{2} \sum_i w_j^{(0)} \Sigma_j^{-1} (x^{(i)} - \mu_j) = -\frac{1}{2} \Sigma_j^{-1} \left( \sum_i w_j^{(0)} (x^{(i)} - \mu_j) \right) \]

Setting to 0 and using \( \Sigma_j^{-1} \) is full rank \( \Rightarrow \sum_i w_j^{(0)} (x^{(i)} - \mu_j) = 0 \)
\[ \mu_j = \frac{\sum_i w_j^{(0)} x^{(i)}}{\sum_i w_j^{(0)}} \] (as before)

\( \phi_j \) is constrained \( \sum_j \phi_j = 1 \), \( \phi_j \geq 0 \), need Lagrangian

\[ \nabla \phi_j = \sum_i w_j^{(0)} \nabla \phi_j \log \phi_j + \lambda (\sum \phi_j - 1) \]
\[ = \sum_i w_j^{(0)} \phi_j - \lambda = 0 \Rightarrow \phi_j = -\frac{1}{\lambda} \sum_i w_j^{(0)} \]

Since \( \sum_j \phi_j = 1 \), \( \sum \phi_j = -\frac{1}{\lambda} \sum_j w_j^{(0)} = -\frac{n}{\lambda} \)

\( \phi_j = \frac{1}{n} \sum_i w_j^{(0)} \)

**MESSAGE:** EM recovers GMM automatically.

**NB:** If \( z^{(i)} \) is continuous, can can replace sums with integrals
Factor Analysis

Many fewer points than dimensions, "n < d"

Could have many hypothetical sources, few sensors.

How does this happen?

Place sensors all over campus, record 100s of locations

= d ~ 100s.

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/ Gaussians... Even 1 Gaussian

\[ \mu = \frac{1}{n} \sum_{i=1}^{n} x^{(i)} \rightarrow \text{This is OK} \]

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

\[ \text{Rank } (\Sigma) \leq n < d \rightarrow \text{not full rank.} \]

Problem in Gaussian likelihood

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

\[ \Rightarrow |\Sigma| = 0 \]

We will fix these issues by examining three models

that are simpler. *Spoiler:* we'll combine these in the end!

Recall MLE for Gaussian
\[
\max_{\mu, \Sigma} \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi|\Sigma|}} \exp \left( -\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right)
\]

Equivalent
\[
\min_{\mu, \Sigma} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) + \log |\Sigma|
\]

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

We'll use this as shown below.

**Building Block 1**

**Suppose Independent and Identical Covariance**
\( \Sigma = \sigma^2 I \) (\( \sigma^2 \) is variance)

Covariance "are circles"

What is \( \mu \) for \( \Sigma \)?
\[
|\Sigma| = 2\sigma^2
\]
\[
\min_{\sigma^2} \sigma^{-2} \sum_{i=1}^{n} (x_i - \mu)^T (x_i - \mu) + d \log \sigma^2
\]
\[
\frac{\partial}{\partial \sigma^2} \sigma^{-2} \sum_{i=1}^{n} (x_i - \mu)^T (x_i - \mu) = 0 \Rightarrow \sigma^2 = \frac{c}{nd}
\]

\[
\sigma^2 = \frac{1}{nd} \sum_{i=1}^{n} (x_i - \mu)^T (x_i - \mu)
\]

"Subtract mean and square all entries."
Building Block 2

$$\mathcal{L} = \begin{bmatrix} \sigma^2 & \cdots \\ \cdots & \cdots \end{bmatrix}$$

Axis Aligned ellipse

Set $$Z_i = \mathcal{L}^2_i$$ (same idea as above)

$$\min_{Z_i \cdots Z_j} \sum_{i=1}^{d} \sum_{j=1}^{d} (x_{i}^{(j)} - \mu_{j})^2 + \log Z_j$$

This is d Problems for each d Dimension

$$\Rightarrow \sum_{i=1}^{d} \sum_{j=1}^{d} (x_{i}^{(j)} - \mu_{j})^2 + \log Z_j$$

$$\Rightarrow \sigma^2_i = \frac{1}{Z_j} (x_{i}^{(j)} - \mu_{j})^2$$

**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**

$$P(x, z) = P(x | z)P(z) \quad Z \text{ is latent}$$

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

$$x = \mu + \Lambda z + \epsilon \quad \text{or} \quad x \sim N(\mu + \Lambda z, \Phi)$$

**Mean in the space**

$$\epsilon \sim N(0, \Phi) \quad \text{noisy}$$

**Example**

$$d = 2, \ s = 1, \ n = 5$$

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

1. Generate $$z^{(1)}, \ldots, z^{(5)}$$ from $$N(0, 1)$$
2. Suppose $\Lambda = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

3. Add $\mu$

4. Add $e$ in big space

Data we would observe are purple dots.

So small latent space produces data in high dim space.

**Technical Tools**: Block Gaussians

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

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

Notation is widely used and helpful.

**Fact 1**: $P(x_1) = \int x_2 P(x_1, x_2)$

**Marginalization**

For Gaussians, $P(x) = N(\mu, \Sigma)$ (not conditioning)

**Fact 2**: $P(x_1 | x_2) \sim N(\mu_{1|2}, \Sigma_{1|2})$

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

**Conditioning**
\[ \Sigma_{12} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \]  

(Multiplication Lemma)

**Proofs Outline (temp to add)**

**Summary:** Marginalization + Conditioning Gaussian = Another Gaussian (Closed)

We have formula for parameters.

### Back to Factor Analysis

\[ x = \mu + \Lambda z + \epsilon \]

\[ (x) \sim N(\mu, \Sigma) \]

Since \( \mathbb{E}[z] = 0 \)

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

What is \( \Sigma \)?

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

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

\[ = \Lambda^T \]

\[ \Sigma_{21} = \Sigma_{12}^T \]

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

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

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

\[ = \Lambda \Lambda^T + \Phi \]

\[ \hat{\Sigma} = \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda \Lambda^T + \Phi \end{bmatrix} \]
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 (about low dim. sources)
- We saw how to estimate parameters of FA using EM