CS 229 Lecture Thirteen
Unsupervised Learning:
Gaussian Mixture Models as EM

Chris Ré

May 14, 2023

EM for GMM and Factor Analysis

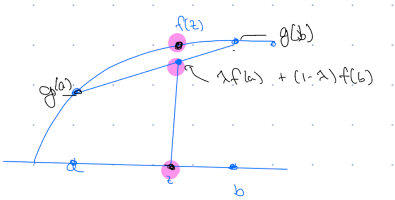
- EM recovers our ad hoc algorithm GMM.
- Factor Analysis: What happens when many fewer points than dimension "$n \ll d$", need even more structure (but also EM).

Recall: GMM from Last Time

- We saw an iterative method for GMM:
  - We estimate the distribution of the latent variable $z^{(i)}$ i.e., a $P(z^{(i)} = j)$, which is a probabilistic assignment of each point $x^{(i)}$ to a source j.
  - We then refit parameters (the mean and shape of each source (μ_j, σ_j^2), and the fraction of points each sees, ϕ_j) for $j = 1, \ldots, k$.
  - We repeat.

Jensen's Inequality and Concave Functions

The canonical concave function is $g(x) = \log x$ on $(0, \infty)$.



For any $z \in [a, b]$ we can write $z = \lambda a + (1 - \lambda)b$. Then, the *"chord is below"* picture means

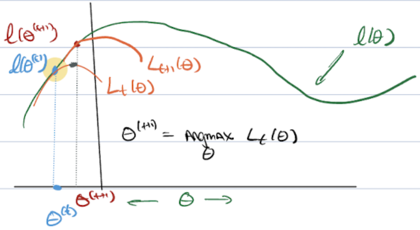
$$g(z) \geq \lambda g(a) + (1 - \lambda)g(b).$$

Last time we saw, *"a with prob λ and b with prob $1 - \lambda$"* which leads to Jensen's Inequality for concave g

$$g(\mathbb{E}[z]) \geq \mathbb{E}[g(z)] \text{ specifically } \log \mathbb{E}[z] \geq \mathbb{E}[\log z]$$

Hopefully, drawing this picture helps you remember!

Picture of EM Algorithm



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

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

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

- **E-Step** Given $\theta^{(t)}$ find a curve L_t
- **M-Step** Given L_t, set $\theta^{(t+1)} = \mathrm{argmax}_\theta\, L_t(\theta)$.

How do we construct L_t?

We examine a single data point (and drop scripts). First a trick,

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

We pick $Q(z)$ s.t. $\sum_z Q(z) = 1$ and $Q(z) \geq 0$ then,

$$= \log \mathbb{E}_{z \sim Q(z)} \left[\frac{P(x, z; \theta)}{Q(z)} \right] \qquad \text{Def of } \mathbb{E}$$

How do we construct L_t?

We examine a single data point (and drop scripts). First a trick,

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

We pick $Q(z)$ s.t. $\sum_z Q(z) = 1$ and $Q(z) \geq 0$ then,

$$
\begin{aligned}
&= \log \mathbb{E}_{z \sim Q(z)} \left[\frac{P(x, z; \theta)}{Q(z)} \right] && \text{Def of } \mathbb{E} \\
&\geq \mathbb{E}_{z \sim Q(z)} \left[\log \frac{P(x, z; \theta)}{Q(z)} \right] && \text{Jensen, since log is concave.}
\end{aligned}
$$

How do we construct L_t?

We examine a single data point (and drop scripts). First a trick,

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

We pick $Q(z)$ s.t. $\sum_z Q(z) = 1$ and $Q(z) \geq 0$ then,

$$\begin{aligned}
&= \log \mathbb{E}_{z \sim Q(z)} \left[\frac{P(x, z; \theta)}{Q(z)} \right] && \text{Def of } \mathbb{E} \\
&\geq \mathbb{E}_{z \sim Q(z)} \left[\log \frac{P(x, z; \theta)}{Q(z)} \right] && \text{Jensen, since log is concave.} \\
&= \sum_z Q(z) \log \frac{P(x, z; \theta)}{Q(z)} && \text{Def of } \mathbb{E}
\end{aligned}$$

How do we construct L_t?

We examine a single data point (and drop scripts). First a trick,

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

We pick $Q(z)$ s.t. $\sum_z Q(z) = 1$ and $Q(z) \geq 0$ then,

$$
\begin{aligned}
&= \log \mathbb{E}_{z \sim Q(z)} \left[\frac{P(x, z; \theta)}{Q(z)} \right] && \text{Def of } \mathbb{E} \\
&\geq \mathbb{E}_{z \sim Q(z)} \left[\log \frac{P(x, z; \theta)}{Q(z)} \right] && \text{Jensen, since log is concave.} \\
&= \sum_z Q(z) \log \frac{P(x, z; \theta)}{Q(z)} && \text{Def of } \mathbb{E}
\end{aligned}
$$

This lower bound holds for *any* such choice of Q–a family of lower bounds. We can select Q *per point*.

How do we make it tight?

Select each Q to make tight for its term...

$$\frac{P(x, z; \theta)}{Q(z)} = c \text{ is constant wrt } z\text{, then Jensen is an equality.}$$

That is, if the random variable's distribution doesn't depend on z.

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

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

They are equal, that is, Jensen's inequality is tight.

How do we make it tight?

Select each Q to make tight for its term...

$$\log \frac{P(x, z; \theta)}{Q(z)} = c \text{ is constant wrt } z\text{, then Jensen is an equality.}$$

How do we make it tight?

Select each Q to make tight for its term...

$\log \frac{P(x,z;\theta)}{Q(z)} = c$ is constant wrt z, then Jensen is an equality.

So what if $Q(z) = P(z \mid x;\theta)$ then

$$\log \frac{P(x,z;\theta)}{P(z \mid x;\theta)} = \log P(x;\theta)$$

If we examine the argument above, the only inequality is now equality so with this choice of Q we are tight!

Note: $Q(z)$ *depends* on θ and x–but not z–so we will select a $Q^{(i)}(z)$ for each point $x^{(i)}$ for $i = 1, \ldots, n$.

ELBO!

We define the Evidence Lower Bound (ELBO) as:

$$\mathrm{ELBO}(x, Q, \theta) = \sum_z Q(z) \log \frac{P(x, z; \theta)}{Q(z)}.$$

So now, we've shown:

$$\ell(\theta) \geq \sum_{i=1}^{n} \mathrm{ELBO}(x^{(i)}, Q^{(i)}, \theta) \qquad \text{for any } Q^{(i)} \text{ and } \theta$$

ELBO!

We define the Evidence Lower Bound (ELBO) as:

$$\mathrm{ELBO}(x, Q, \theta) = \sum_{z} Q(z) \log \frac{P(x, z; \theta)}{Q(z)}.$$

So now, we've shown:

$$\ell(\theta) \geq \sum_{i=1}^{n} \mathrm{ELBO}(x^{(i)}, Q^{(i)}, \theta) \qquad \text{for any } Q^{(i)} \text{ and } \theta$$

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

ELBO!

We define the Evidence Lower Bound (ELBO) as:

$$\mathrm{ELBO}(x, Q, \theta) = \sum_{z} Q(z) \log \frac{P(x, z; \theta)}{Q(z)}.$$

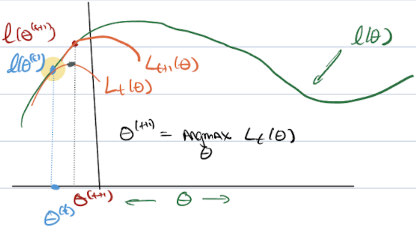
So now, we've shown:

$$\ell(\theta) \geq \sum_{i=1}^{n} \mathrm{ELBO}(x^{(i)}, Q^{(i)}, \theta) \qquad \text{for any } Q^{(i)} \text{ and } \theta$$

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

We've shown lower bound and tight, deriving the picture!

Wrap-up of EM!



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

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

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

- **E-Step** $Q^{(i)}(z) = P(z^{(i)} \mid x^{(i)}; \theta)$ for $i = 1, \dots, n$.
- **M-Step** $\theta^{(t+1)} = \mathrm{argmax}_\theta L_t(\theta)$ in which

$$L_t(\theta) = \sum_{i=1}^{n} \mathrm{ELBO}(x^{(i)}, Q^{(i)}, \theta).$$

Some comments:

- Why does this terminate? $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$
- Is it globally optimal? Nope! See the picture.

EM for Mixture of Gaussians

Generic EM algorithm.

- **E-Step**. for $i = 1, \ldots, n$, estimate the latent variable $z^{(i)}$. Set

$$Q^{(i)}(z) = P(z^{(i)} \mid x^{(i)}; \theta^{(t)}).$$

EM for Mixture of Gaussians

Generic EM algorithm.

- **E-Step**. for $i = 1, \ldots, n$, estimate the latent variable $z^{(i)}$. Set

$$Q^{(i)}(z) = P(z^{(i)} \mid x^{(i)}; \theta^{(t)}).$$

- **M-Step** Update the parameters, given our estimate of $z^{(i)}$

$$\theta^{(t+1)} = \operatorname*{argmax}_{\theta} L_t(\theta) = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \mathrm{ELBO}(x^{(i)}, Q^{(i)}, \theta).$$

EM for Mixture of Gaussians

Generic EM algorithm.

- **E-Step**. for $i = 1, \ldots, n$, estimate the latent variable $z^{(i)}$. Set

$$Q^{(i)}(z) = P(z^{(i)} \mid x^{(i)}; \theta^{(t)}).$$

- **M-Step** Update the parameters, given our estimate of $z^{(i)}$

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, L_t(\theta) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \mathrm{ELBO}(x^{(i)}, Q^{(i)}, \theta).$$

Mixture of Gaussians P s.t. for each data point $(i = 1, \ldots, n)$ and each source $(j = 1, \ldots, k)$ we find a *soft assignment* $P(z^{(i)} = j)$

$$P(x^{(i)}, z^{(i)}) = P(x^{(i)}|z^{(i)})P(z^{(i)}) \qquad \text{Bayes' Rule}$$

EM for Mixture of Gaussians

Generic EM algorithm.

- **E-Step**. for $i = 1, \ldots, n$, estimate the latent variable $z^{(i)}$. Set

$$Q^{(i)}(z) = P(z^{(i)} \mid x^{(i)}; \theta^{(t)}).$$

- **M-Step** Update the parameters, given our estimate of $z^{(i)}$

$$\theta^{(t+1)} = \operatorname*{argmax}_{\theta} L_t(\theta) = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \mathrm{ELBO}(x^{(i)}, Q^{(i)}, \theta).$$

Mixture of Gaussians P s.t. for each data point $(i = 1, \ldots, n)$ and each source $(j = 1, \ldots, k)$ we find a *soft assignment* $P(z^{(i)} = j)$

$$P(x^{(i)}, z^{(i)}) = P(x^{(i)}|z^{(i)})P(z^{(i)}) \qquad \text{Bayes' Rule}$$

$$z^{(i)} \sim \text{Multinomial}(\phi) \qquad \text{Mixture of sources}$$

EM for Mixture of Gaussians

Generic EM algorithm.

- **E-Step**. for $i = 1, \ldots, n$, estimate the latent variable $z^{(i)}$. Set

$$Q^{(i)}(z) = P(z^{(i)} \mid x^{(i)}; \theta^{(t)}).$$

- **M-Step** Update the parameters, given our estimate of $z^{(i)}$

$$\theta^{(t+1)} = \operatorname*{argmax}_{\theta} L_t(\theta) = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \mathrm{ELBO}(x^{(i)}, Q^{(i)}, \theta).$$

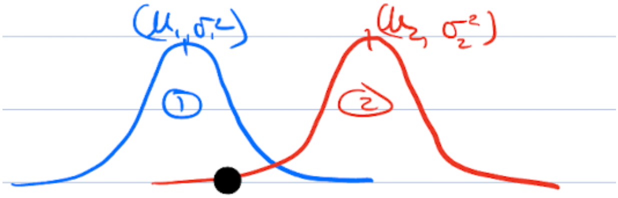
Mixture of Gaussians P s.t. for each data point $(i = 1, \ldots, n)$ and each source $(j = 1, \ldots, k)$ we find a *soft assignment* $P(z^{(i)} = j)$

$$\begin{aligned} P(x^{(i)}, z^{(i)}) &= P(x^{(i)}|z^{(i)})P(z^{(i)}) && \text{Bayes' Rule} \\ z^{(i)} &\sim \text{Multinomial}(\phi) && \text{Mixture of sources} \\ x^{(i)} \mid z^{(i)} = j &\sim \mathcal{N}(\mu_j, \sigma_j^2) && \text{Gaussian in each source} \end{aligned}$$

The E-Step for Mixture of Gaussians

Given $\theta^{(t)}$ and the data $x^{(1)}, \ldots, x^{(n)}$ estimate:

$$Q^{(i)}(z) = P(z^{(i)} \mid x^{(i)}; \theta^{(t)}).$$



Recall. We did this in detail. Bayes Rule to automate reasoning two factors in source of a point?

- Did more points come from source 1 or 2? (i.e. ϕ_1 v. ϕ_2).
- How likely is this to be generated by that source? (i.e., likelihood of $\mathcal{N}(\mu_1, \sigma_1^2)$ v. $\mathcal{N}(\mu_2, \sigma_2^2)$.

M-Step for Mixture of Gaussians

Given $P(z^{(i)} = j)$ for $i = 1, \ldots, n$ estimate

$$\theta = (\phi, \mu_1, \Sigma_1, \ldots, \mu_n, \Sigma_j).$$

Note: Here, the dimension is greater than 1, i.e., $d \geq 1$ so:

$$\mu_j \in \mathbb{R}^d \text{ and } \Sigma_j \in \mathbb{R}^{d \times d}$$

Perhaps, confusingly but conventionally, if $d = 1$, $\Sigma_1 = \sigma_1^2$. Σ_j is called the covariance matrix, and it's symmetric positive definite, i.e., $\Sigma_j \succeq 0$ and $\Sigma_j^T = \Sigma_j$.

M-Step for Mixture of Gaussians

Given $P(z^{(i)} = j)$ for $i = 1, \dots, n$ estimate θ

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

It's all computing derivatives, first recall the model and Bayes rule:

$$P(x^{(i)}, z^{(i)}; \theta) = P(x^{(i)} | z^{(i)}) P(z^{(i)}) \text{ in which}$$
$$P(x^{(i)} \mid z^{(i)} = j) \sim \mathcal{N}(\mu_j, \Sigma_j) \text{ and } \phi_j = P(z^{(i)} = j)$$

M-Step for Mixture of Gaussians

Given $P(z^{(i)} = j)$ for $i = 1, \dots, n$ estimate θ

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

It's all computing derivatives, first recall the model and Bayes rule:

$$P(x^{(i)}, z^{(i)}; \theta) = P(x^{(i)}|z^{(i)})P(z^{(i)}) \text{ in which}$$
$$P(x^{(i)} \mid z^{(i)} = j) \sim \mathcal{N}(\mu_j, \Sigma_j) \text{ and } \phi_j = P(z^{(i)} = j)$$

In previous lecture, we wrote $w_j^{(i)} = Q^{(i)}(z_i = j)$. Then, says:

$$f_i(\theta) = \sum_j w_j^{(i)} \log \left(\frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\{-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1}(x^{(i)} - \mu_j)\}}{w_j^{(i)}} \right).$$

Let's find the parameters!

So we have:

$$f_i(\theta) = \sum_j w_j^{(i)} \log\left(\frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}\exp\{-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\}}{w_j^{(i)}}\right).$$

- To find the value of $\operatorname{argmax}_\theta$ we look for critical points of some parameter, say μ_j. That is, we look for solutions of

$$\nabla_{\mu_j}\sum_{i=1}^{n} f_i(\theta) = 0.$$

- For μ_j we *expect* it has a really nice form from lecture:

$$\mu_j = \frac{\sum_{i=1}^{n} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{n} w_j^{(i)}}.$$

- Let's derive it formally!

So we have:

$$f_i(\theta) = \sum_j w_j^{(i)} \log \left(\frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\{-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\}}{w_j^{(i)}} \right).$$

So we have:

$$f_i(\theta) = \sum_j w_j^{(i)} \log \left(\frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\{-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\}}{w_j^{(i)}} \right).$$

Sum over all $i = 1, \ldots, n$ and let's take the derivative wrt μ_j:

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

So we have:

$$f_i(\theta) = \sum_j w_j^{(i)} \log \left(\frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\{-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\}}{w_j^{(i)}} \right).$$

Sum over all $i = 1, \ldots, n$ and let's take the derivative wrt μ_j:

$$\begin{aligned}
\nabla_{\mu_j} \sum_{i=1}^n f_i(\theta) &= -\sum_{i=1}^n w_j^{(i)} \frac{1}{2} \nabla_{\mu_j} \left((x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right) \\
&= -\sum_{i=1}^n w_j^{(i)} \Sigma_j^{-1} (x^{(i)} - \mu_j) = -\Sigma_j^{-1} \sum_{i=1}^n w_j^{(i)} (x^{(i)} - \mu_j)
\end{aligned}$$

So we have:

$$f_i(\theta) = \sum_j w_j^{(i)} \log \left(\frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\{-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\}}{w_j^{(i)}} \right).$$

Sum over all $i = 1, \ldots, n$ and let's take the derivative wrt μ_j:

$$\begin{aligned}
\nabla_{\mu_j} \sum_{i=1}^n f_i(\theta) &= -\sum_{i=1}^n w_j^{(i)} \frac{1}{2} \nabla_{\mu_j} \left((x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right) \\
&= -\sum_{i=1}^n w_j^{(i)} \Sigma_j^{-1} (x^{(i)} - \mu_j) = -\Sigma_j^{-1} \sum_{i=1}^n w_j^{(i)} (x^{(i)} - \mu_j)
\end{aligned}$$

So we have:

$$f_i(\theta) = \sum_j w_j^{(i)} \log\left(\frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}\exp\{-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\}}{w_j^{(i)}}\right).$$

Sum over all $i = 1, \ldots, n$ and let's take the derivative wrt μ_j:

$$\begin{aligned}\nabla_{\mu_j}\sum_{i=1}^n f_i(\theta) &= -\sum_{i=1}^n w_j^{(i)}\frac{1}{2}\nabla_{\mu_j}\left((x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\right)\\ &= -\sum_{i=1}^n w_j^{(i)}\Sigma_j^{-1}(x^{(i)}-\mu_j) = -\Sigma_j^{-1}\sum_{i=1}^n w_j^{(i)}(x^{(i)}-\mu_j)\end{aligned}$$

Since we want a critical point, we set to 0.

$$\nabla_{\mu_j}\sum_{i=1}^n f_i(\theta) = 0 \iff \Sigma_j^{-1}\sum_{i=1}^n w_j^{(i)}(x^{(i)}-\mu_j) = 0$$

Now, Σ_j is full rank so it must be that the inner term is 0 and,

$$\sum_{i=1}^n w_j^{(i)}(x^{(i)}-\mu_j) = 0 \implies \mu_j = \frac{\sum_{i=1}^n w_j^{(i)}x^{(i)}}{w_j^{(i)}}.$$

A few comments

- *Reminder for what's next*: We used that Σ_j was full rank.
- Same process for each of the parameters: $\mu_j, \sigma_j^2, \phi_j$.
- **Detail:** If a parameter has *constraints*, e.g. ϕ we have $\sum_j \phi_j = 1$, need to use *Lagrange multipliers*.

$$\nabla_{\phi_j} \sum_{i=1}^{n} f_i(\theta) + \lambda \left(\sum_j \phi_j - 1 \right).$$

If this is not familiar, please check in the TA notes (and I prepped some notes for this!)

Message: EM recovers our ad hoc algorithm!

Factor Analysis

So far more data n than parameters, what happens when it's the other way? In symbols, $d \gg n$?

Factor Analysis

Suppose we place 1000 temperature sensors all over campus, each gives us a reading hourly reading for a day. We have $n = 24$ readings and the sensors are a value $d = 1000$.

Factor Analysis

Suppose we place 1000 temperature sensors all over campus, each gives us a reading hourly reading for a day. We have $n = 24$ readings and the sensors are a value $d = 1000$.

We want to fit a density to $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$ with $d \gg n$, but it seems *hopeless*. Let's examine why that is...

Factor Analysis: Technical Motivation

Suppose we want to fit a Gaussian to

$$x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d \text{ with } d \gg n.$$

Let's see where we get stuck with $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$.

Factor Analysis: Technical Motivation

Suppose we want to fit a Gaussian to

$$x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d \text{ with } d \gg n.$$

Let's see where we get stuck with $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$. Let's examine the covariance:

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

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

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

Factor Analysis: Technical Motivation

Suppose we want to fit a Gaussian to

$$x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^d \text{ with } d \gg n.$$

Let's see where we get stuck with $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$. Let's examine the covariance:

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

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

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

Divide by 0 in the first occurrence of Σ, and an inverse of a rank deficient matrix.

The main technical idea is to *restrict* the model in some way. To build that model, we'll examine building blocks (Gaussians) where we *can* estimate the parameters.

Spoiler: We use these building blocks in our final model.

The Key Property

The key property we use is to estimate MLE for Gaussian $\mathcal{N}(\mu, \Sigma)$.

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

Throughout, we will use the equivalent min form taking a log.

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

The Key Property

The key property we use is to estimate MLE for Gaussian $\mathcal{N}(\mu, \Sigma)$.

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

Throughout, we will use the equivalent min form taking a log.

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

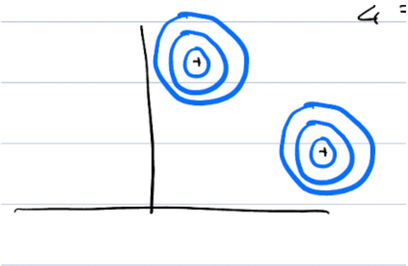
We use this property repeatedly. If Σ is full rank, we can find the mean μ by averaging. That is, take ∇_μ and set equal to 0:

$$\sum_{i=1}^{n} \Sigma^{-1}(x^{(i)} - \mu) = 0 \implies \mu = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}.$$

Building Block 1

Suppose components are *independent* with *identical* covariance.

$$\Sigma = \sigma^2 I \text{ where } \sigma \in \mathbb{R} \text{ and } \Sigma^{d \times d}$$



Visualize these Gaussians as *circles* centered at different points. Each has a center, $\mu \in \mathbb{R}^d$, and a single scalar standard deviation, $\sigma \in \mathbb{R}$.

Building Block 1

Suppose components are *independent* with *identical* covariance.

$$\Sigma = \sigma^2 I \text{ where } \sigma \in \mathbb{R} \text{ and } \Sigma^{d \times d}$$

Our MLE equation *simplifies* since $\Sigma = \sigma^2 I$

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

Let $z = \sigma^2$ for notation, we have an equation:

$$\nabla_z \frac{C}{z} + d \log z = 0 \implies -z^{-2} C + n \frac{d}{z} = 0.$$

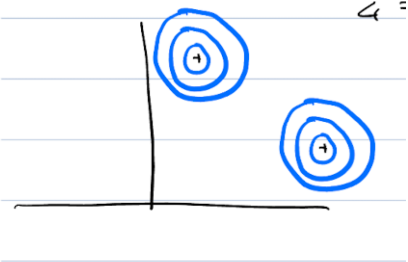
Thus, we have $z = \frac{C}{nd}$ or in original notation:

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

Building Block 1

Suppose components are *independent* with *identical* covariance.

$$\Sigma = \sigma^2 I \text{ where } \sigma \in \mathbb{R} \text{ and } \Sigma^{d \times d}$$



Visualize these Gaussians as *circles* centered at different points. Each has a center, $\mu \in \mathbb{R}^d$, and a scalar standard deviation, $\sigma \in \mathbb{R}$.

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

Not surprisingly, the variance is the sum of the variance of each individual component.

Building Block 2

Suppose components are independent but with possibly *different* covariances:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3^2 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & \sigma_d^2 \end{pmatrix} \text{ for } \sigma_i^2 \in \mathbb{R}_+.$$

Diagonal Matrix. Axis-aligned isoclines with each principal axis can be different. There are d different numbers here (still less than roughly d^2) possible.

MLE for data

set $z_i = \sigma_i^2$ for $i = 1, \ldots, d$ and plug in, we get the equation:

$$\min_{z_1,\ldots,z_d} \sum_{i=1}^{n} \sum_{j=1}^{d} z_j^{-1}(x^{(i)} - \mu_j)^2 + \log z_j.$$

Notice this is d one dimensional problems (not surprising, since independent)–and so,

$$\min_{z_j} \sum_{i=1}^{n} z_j^{-1}(x_j^{(i)} - \mu_j)^2 + \log z_j \implies \sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_j^{(i)} - \mu_j)^2.$$

we average over each of the d components independently.

Building Block Wrap-up

- We saw two forms of estimation with a single free parameter and d free parameters.
  - We reduced dramatically from the nearly d^2 free parameters in a general covariance matrix.
- We assumed we were given μ. If we have to estimate μ too at the same time, only minor changes.

Our Factor Model

Our Factor Model

Let d the "big dimension" and s the "small dimension" i.e. $s < d$

$$\mu \in \mathbb{R}^d \text{ and } \Lambda \in \mathbb{R}^{d \times s} \text{ and a diagonal matrix } \Phi \in \mathbb{R}^{d \times d}.$$

The model is given as a latent model with variable $z \in \mathbb{R}^s$.

$$P(x, z) = P(x|z)P(z) \text{ with } z \in \mathcal{N}(0, I_s) \text{ and } \varepsilon \sim \mathcal{N}(0, \Phi)$$

Then,

$$x = \mu + \Lambda z + \varepsilon \text{ or } x \sim \mathcal{N}(\mu + \Lambda z, \Phi).$$

The latent is in the small dimension and the observed dimension is the larger dimension

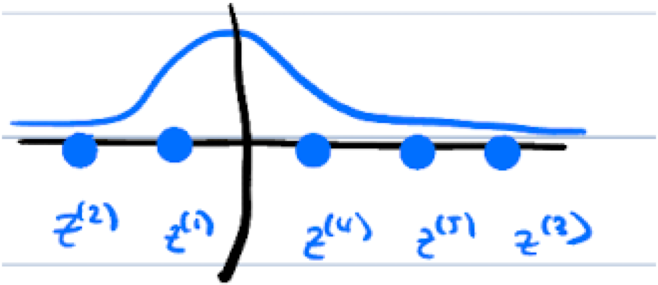
Let's unpack this!

Understanding the Factor as a Sampling procedure

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

Example: Let the big dimension, $d = 2$, small dimension $s = 1$, and number of points $n = 5$. Let's draw some pictures!

1. We first generate $z^{(1)}, \ldots, z^{(n)}$ with $\mathcal{N}(0, 1)$

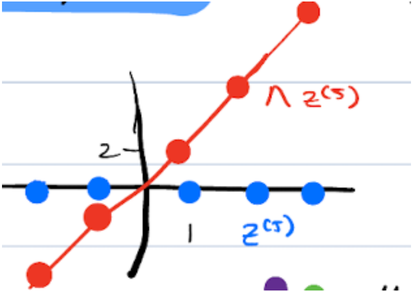


Understanding the Factor as a Sampling procedure: Step 2

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

Example: Let the big dimension, $d = 2$, small dimension $s = 1$, and number of points $n = 5$. Let's draw some pictures!

1. We first generate $z^{(1)}, \ldots, z^{(n)}$ with $\mathcal{N}(0,1)$
2. Suppose $\Lambda = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ then, construct $\Lambda z^{(i)}$ for $i = 1, \ldots, n$.

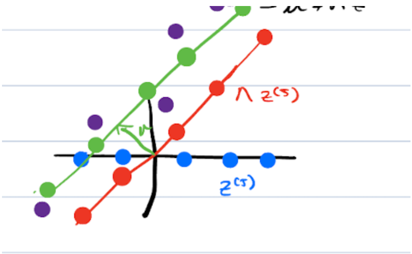


Understanding the Factor as a Sampling procedure: Step 2

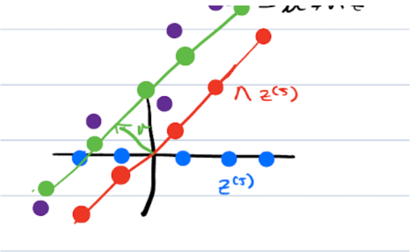
$$x = \mu + \Lambda z + \varepsilon$$

Example: Let the big dimension, $d = 2$, small dimension $s = 1$, and number of points $n = 5$. Let's draw some pictures!

1. We first generate $z^{(1)}, \ldots, z^{(n)}$ with $\mathcal{N}(0, 1)$
2. Suppose $\Lambda = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ then, construct $\Lambda z^{(i)}$ for $i = 1, \ldots, n$.
3. Add the mean μ.
4. Generate $\varepsilon^{(i)} \in \mathbb{R}^d$ to give full dimensional noise



Our Learning Goal



- We observe the end result of the process, here the *purple dots*.
- We estimate the likelihood using this model, and the smaller latent space (For example, $s < n < d$).
- Key point: Even though the data is in a high dimensional space, we can fit it.

Detour for Useful Tool Block Gaussians

Given $d_1 + d_2 = d$, we break vectors into blocks

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \text{ for } x \in \mathbb{R}^{d_1+d_2}, x_1 \in \mathbb{R}^{d_1}, \text{ and } x_2 \in \mathbb{R}^{d_2}$$

Detour for Useful Tool Block Gaussians

Given $d_1 + d_2 = d$, we break vectors into blocks

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \text{ for } x \in \mathbb{R}^{d_1+d_2}, x_1 \in \mathbb{R}^{d_1}, \text{ and } x_2 \in \mathbb{R}^{d_2}$$

and matrices into blocks:

$$\Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix} \text{ for } \Sigma \in \mathbb{R}^{d \times d}, \Sigma_{i,j} \in \mathbb{R}^{d_i \times d_j} \text{ for } i, j \in \{1, 2\}.$$

This notation is widely used and helpful. Note that they are *compatible* so that we can write:

$$\Sigma x = \begin{pmatrix} \Sigma_{1,1} x_1 + \Sigma_{1,2} x_2 \\ \Sigma_{2,1} x_1 + \Sigma_{2,2} x_2 \end{pmatrix}$$

Facts about Gaussians

Suppose $x = (x_1, x_2) \sim \mathcal{N}(\mu, \Sigma)$.

- **Marginalization is Gaussian**:

$$P(x_1) = \int_{x_2} P(x_1, x_2)\, dx_2 \text{ for Gaussians } p(x_1) = \mathcal{N}(\mu_1, \Sigma_{1,1})$$

Facts about Gaussians

Suppose $x = (x_1, x_2) \sim \mathcal{N}(\mu, \Sigma)$.

- **Marginalization is Gaussian**:

$$P(x_1) = \int_{x_2} P(x_1, x_2)\, dx_2 \text{ for Gaussians } p(x_1) = \mathcal{N}(\mu_1, \Sigma_{1,1})$$

- **Conditioning is also Gaussian**:

$$\begin{aligned} p(x_1 \mid x_2) \sim& \mathcal{N}(\mu_{1|2}, \Sigma_{1|2}) \text{ in which} \\ \mu_{1|2} =& \mu_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(x_2 - \mu_2). \\ \Sigma_{1|2} =& \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1} \end{aligned}$$

- Key point: there are explicit formulas and they are *still* Gaussians. Gaussians are special!

Proofs online (happy to add!) this uses Matrix Inversion Lemma.

Back to Factor Analysis with Our Tools

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

Our model can be written (since $\mathbb{E}[z] = 0$ and $\mathbb{E}[x] = \mu$)

$$\begin{pmatrix} z \\ x \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \Sigma\right)$$

We have to compute Σ, but we can solve with EM. Since:

- **E-Step** $Q^{(i)}(z) = P(z^{(i)} \mid x^{(i)}; \theta)$ – use the conditional!
- **M-Step** We have closed forms and can solve!

Now, just an application of EM to learn Factor Analysis.

Time Permitting: Deriving Σ

The model is $x = \mu + \Lambda z + \varepsilon$ and we derive:

$$\begin{pmatrix} z \\ x \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \Sigma\right) \text{ with } \Sigma = \begin{pmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Phi \end{pmatrix}$$

Time Permitting: Deriving Σ

The model is $x = \mu + \Lambda z + \varepsilon$ and we derive:

$$\begin{pmatrix} z \\ x \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \Sigma\right) \text{ with } \Sigma = \begin{pmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Phi \end{pmatrix}$$

Recall $z \sim N(0, I_d)$ and $\varepsilon \sim \mathcal{N}(0, \Phi)$ so $\mathbb{E}[x] = \mu$.

$$\begin{aligned}
\Sigma_{1,1} =& \mathbb{E}[zz^T] = I. \\
\Sigma_{1,2} =& \mathbb{E}[z(x-\mu)^T] = \mathbb{E}[z(\Lambda z + \varepsilon)^T] \\
=& \mathbb{E}[zz^T\Lambda^T] + \mathbb{E}[z\varepsilon^T] = \Lambda^T \\
\Sigma_{2,1} =& \Sigma_{1,2}^T = \Lambda \\
\Sigma_{2,2} =& \mathbb{E}[(x-\mu)(x-\mu)^T] = \mathbb{E}[(\Lambda z + \varepsilon)(\Lambda z + \varepsilon)^T] \\
=& \mathbb{E}[\Lambda zz^T\Lambda^T] + \mathbb{E}[\varepsilon\varepsilon^T] \\
=& \Lambda\Lambda^T + \Phi
\end{aligned}$$

Summary of Today

- EM can used to derive our algorithm for GMM
- Factor Analysis (latent low dimensional space).
- How to estimate the parameters of FA using EM.
- Introduced useful notation for your homework and ML!