Probabilistic Machine Learning

Week 49, Tuesday

Latent variable models

- Discrete
  - Mixture models
- Continuous
  - Probabilistic PCA

Discrete latent variable models: Mixture models

How do we model multimodal distributions?

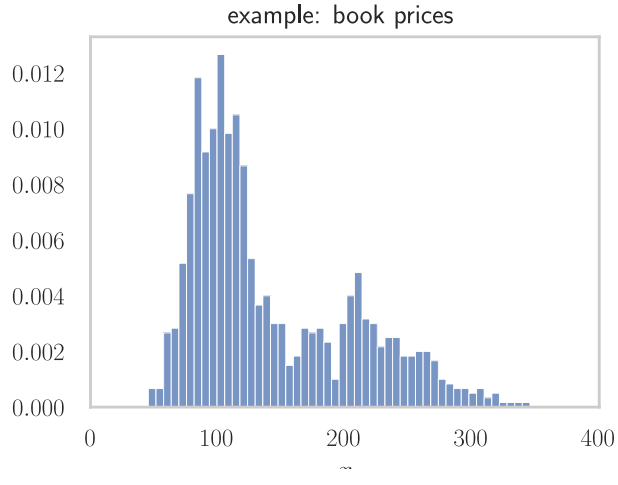
Complex density functions
can be approximated by
convex combination of
simpler distributions.

$$p(\boldsymbol{x}) = \sum_{k}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Such models are called mixture models. The simpler distributions in a mixture model are often called *components*. We will focus on Gaussian mixture models.

How do we model multimodal distributions?

Complex density functions can be approximated by convex combination of simpler distributions.

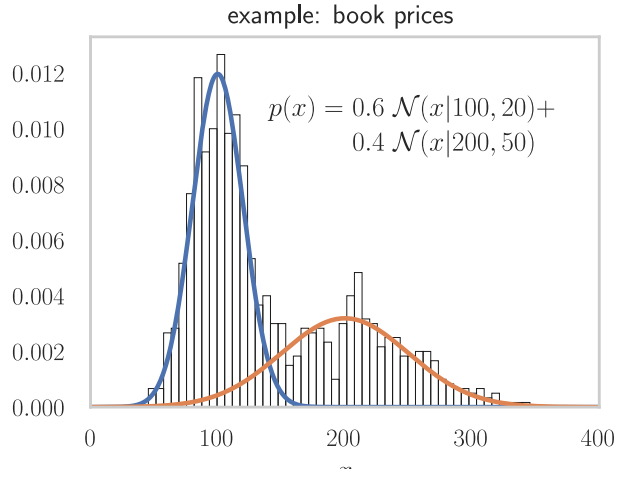


$$p(\boldsymbol{x}) = \sum_{k}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Such models are called mixture models. The simpler distributions in a mixture model are often called *components*. We will focus on Gaussian mixture models.

How do we model multimodal distributions?

Complex density functions can be approximated by convex combination of simpler distributions.

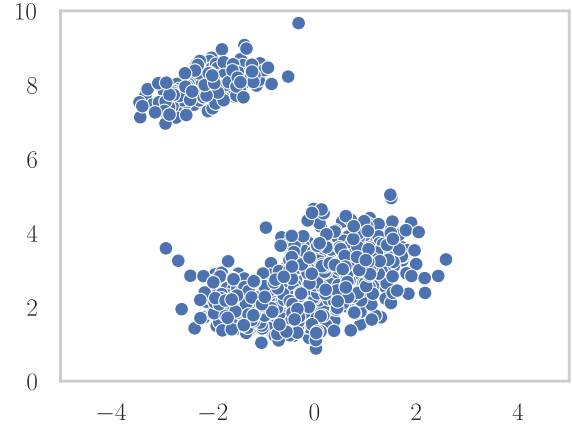


$$p(\boldsymbol{x}) = \sum_{k}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Such models are called mixture models. The simpler distributions in a mixture model are often called *components*. We will focus on Gaussian mixture models.

How do we model multimodal distributions?

Complex density functions can be approximated by convex combination of simpler distributions.

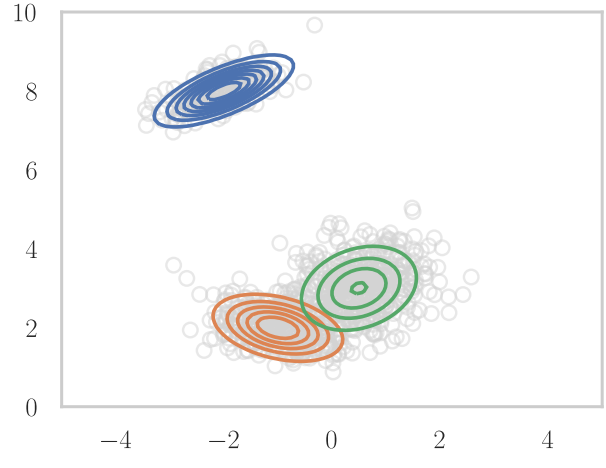


$$p(\boldsymbol{x}) = \sum_{k}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Such models are called mixture models. The simpler distributions in a mixture model are often called *components*. We will focus on Gaussian mixture models.

How do we model multimodal distributions?

Complex density functions can be approximated by convex combination of simpler distributions.



$$p(\boldsymbol{x}) = \sum_{k}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Such models are called mixture models. The simpler distributions in a mixture model are often called *components*. We will focus on Gaussian mixture models.

Mixture models are latent variable models

Idea: introduce discrete random variable z which takes K possible outcomes, representing the K components.

$$p(\boldsymbol{x}) = \sum_{k}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

As usual when representing a categorical variable, we will use a *one-hot* (or *one-out-of-K*) representation.

$$z_k \in \{0, 1\} \qquad (\text{but only one } z_k \text{ can be nonzero at a time})$$

The marginal probability of z are the mixture components weights

$$p(z_k = 1) = \pi_k$$

Since z has a one-hot representation, we can write

$$p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \qquad (\text{only one of the factors will contribute})$$

Mixture models are latent variable models (2)

Model:

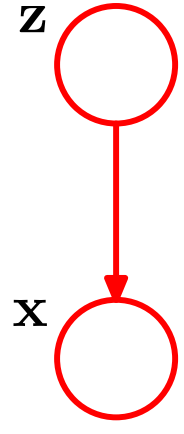
$$
\begin{aligned}
p(\boldsymbol{x}) &= \sum_{\boldsymbol{z}} p(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) \\
&= \sum_{\boldsymbol{z}} \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \pi_k^{z_k} \\
&= \sum_{k=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \underbrace{\pi_k}_{p(z_k=1)}
\end{aligned}
$$

Mixture models are latent variable models (2)

Model:

$$
\begin{aligned}
p(\boldsymbol{x}) &= \sum_{\boldsymbol{z}} p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z}) \\
&= \sum_{\boldsymbol{z}} \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \pi_k^{z_k} \\
&= \sum_{k=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \underbrace{\pi_k}_{p(z_k=1)}
\end{aligned}
$$

Graphical model:

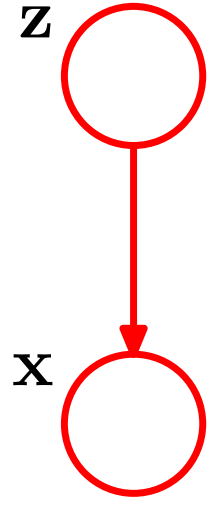


So, we recover the same expression, but with a different interpretation

Why do we need latent variables?

Reasons for introducing latent variables:

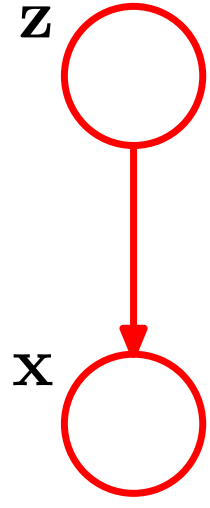
- Artificial variable introduced to facilitate modelling
  - Example: Mixture model
- Real quantity that is difficult to measure
  - Example: Underlying causes for disease



Why do we need latent variables?

Reasons for introducing latent variables:

- Artificial variable introduced to facilitate modelling
  - Example: Mixture model
- Real quantity that is difficult to measure
  - Example: Underlying causes for disease



Types of latent variables:

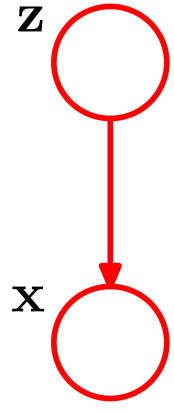
- Discrete (e.g. mixture models)
  - Structuring observations/features in groups/clusters
- Continuous (PCA, factor analysis)
  - Dimensionality reduction

Sampling from a mixture model

The factorization depicted in the graphical model suggests how we can sample from the model:

1. Draw a sample z from $p(z)$
2. Draw a sample x from $p(x|z)$

(a simple example of *ancestral sampling*)

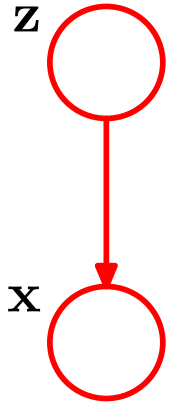


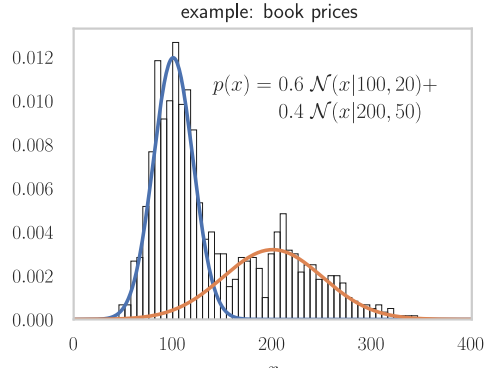
Sampling from a mixture model

The factorization depicted in the graphical model suggests how we can sample from the model:

1. Draw a sample z from $p(z)$
2. Draw a sample x from $p(x|z)$

(a simple example of *ancestral sampling*)





Which component is responsible for a given x?

We can use Bayes rule to evaluate the probability $p(z_k = 1|\boldsymbol{x})$

$$
\begin{aligned}
p(z_k = 1|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|z_k = 1)p(z_k = 1)}{p(\boldsymbol{x})} \\
&= \frac{\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k}{\sum_{j=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j}
\end{aligned}
$$

These are referred to as the *responsibilities*, $\gamma(z_k)$, of component k to an observation x.

Which component is responsible for a given x?

We can use Bayes rule to evaluate the probability $p(z_k = 1|\boldsymbol{x})$

$$\begin{aligned} p(z_k = 1|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|z_k = 1)p(z_k = 1)}{p(\boldsymbol{x})} \\ &= \frac{\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k}{\sum_{j=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j} \end{aligned}$$

These are referred to as the *responsibilities*, $\gamma(z_k)$, of component k to an observation x.

Note that we can evaluate the normalization constant $p(\boldsymbol{x})$, because z is discrete.

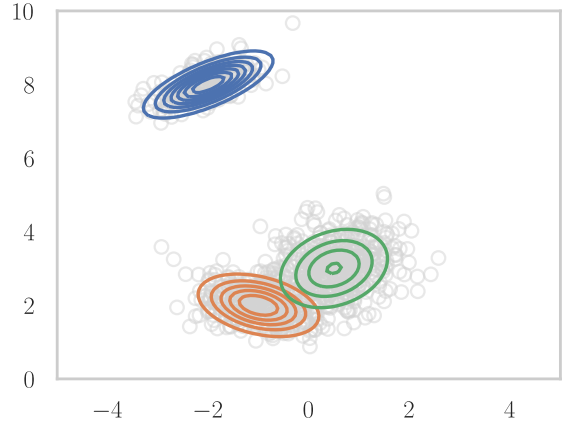
Which component is responsible for a given x?

We can use Bayes rule to evaluate the probability $p(z_k = 1|\boldsymbol{x})$

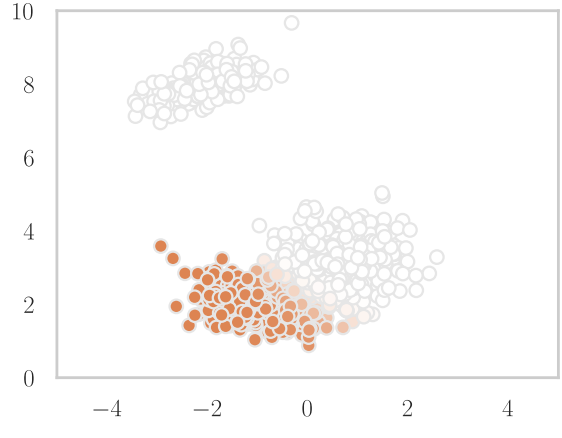
$$p(z_k = 1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|z_k = 1)p(z_k = 1)}{p(\boldsymbol{x})}$$
$$= \frac{\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k}{\sum_{j=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j}$$

These are referred to as the *responsibilities*, $\gamma(z_k)$, of component k to an observation x.

Note that we can evaluate the normalization constant $p(\boldsymbol{x})$, because z is discrete.



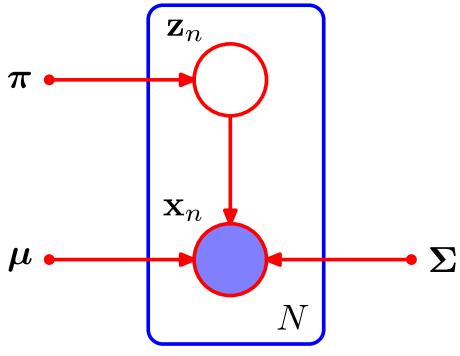
$p(z_k = 1|x)$ for $k = 1$ (the orange cluster):



Parameter estimation in mixture models

Log-likelihood for $N \times D$ data matrix $\boldsymbol{X}$:

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k \right)$$



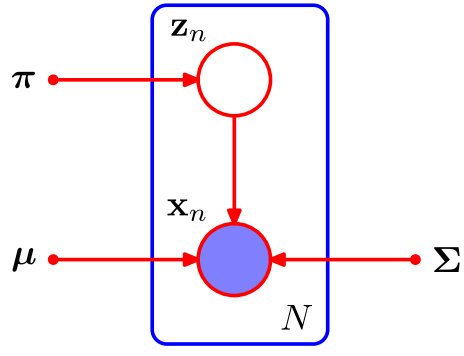
We would like to maximize this wrt $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$.
Unfortunately, unlike the Gaussian itself, we cannot optimize this in closed form (due to the sum inside the $\ln$). ?

Parameter estimation in mixture models

Log-likelihood for $N \times D$ data matrix $\boldsymbol{X}$:

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k \right)$$

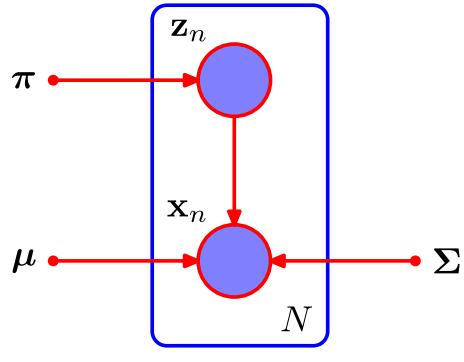
We would like to maximize this wrt $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$.
Unfortunately, unlike the Gaussian itself, we cannot optimize this in closed form (due to the sum inside the ln). ?



Life would be easier if we had observed all z_n:

$$p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}} \pi_k^{z_{nk}}$$

$$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \left(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k \right)$$



Expectation Maximization (idea)

The complete-data case had a convenient form

$$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \left(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k\right)$$

But we don't know z_{nk}.

Idea:

- For any choice of values for $\boldsymbol{\pi}_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$, we can calculate a posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\pi}_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$

Expectation Maximization (idea)

The complete-data case had a convenient form

$$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \left(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k\right)$$

But we don't know z_{nk}.

Idea:

- For any choice of values for $\boldsymbol{\pi}_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$, we can calculate a posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\pi}_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$
- We can then calculate the expectation $\mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right]$ under $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\pi}_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.

Expectation Maximization (idea)

The complete-data case had a convenient form

$$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \left(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k\right)$$

But we don't know z_{nk}.

Idea:

- For any choice of values for $\boldsymbol{\pi}_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$, we can calculate a posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\pi}_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$
- We can then calculate the expectation $\mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right]$ under $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\pi}_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.
- We can now maximize $\mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right]$ wrt $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$

Expectation Maximization

Procedure:

0. initialize $\boldsymbol{\pi}_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$ (e.g. randomly)

   $t \leftarrow 0$

   while not *converged* :

      $t \leftarrow t + 1$

1. $\underbrace{\text{calculate } \mathbb{E}_{\boldsymbol{Z}} \left[\ln p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right] \text{ under } p(\boldsymbol{Z} | \boldsymbol{X}, \boldsymbol{\pi}_{t-1}, \boldsymbol{\mu}_{t-}}_{\text{Expectation step}}$

2. $\underbrace{(\boldsymbol{\pi}_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \leftarrow \underset{(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\operatorname{argmax}} \left(\mathbb{E}_{\boldsymbol{Z}} \left[\ln p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right]\right)}_{\text{Maximization step}}$

Let's take a closer look at these steps.

E-step

The expectation goes inside the sum:

$$\mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{Z}}[z_{nk}] \ln\left(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k\right)$$

The expectation of the binary indicator variable z_{nk} is its probability:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Z}}[z_{nk}] &= p(z_{nk}=1|\boldsymbol{x}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \frac{p(\boldsymbol{x}_n|z_{nk}, \boldsymbol{\mu}, \boldsymbol{\Sigma})p(z_{nk})}{p(\boldsymbol{x}_n)} = \frac{\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k}{\sum_j^K \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j} = \gamma(z_{nk})\end{aligned}$$

So, we see that $\mathbb{E}_{\boldsymbol{Z}}[z_{nk}]$ is simply the responsibility of component k to datapoint x_n.

M-step

Now we just need to optimize:

$$\mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\underbrace{\ln(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k)}_{\ln\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)+\ln(\pi_k)} =: f(\boldsymbol{\mu}$$

M-step

Now we just need to optimize:

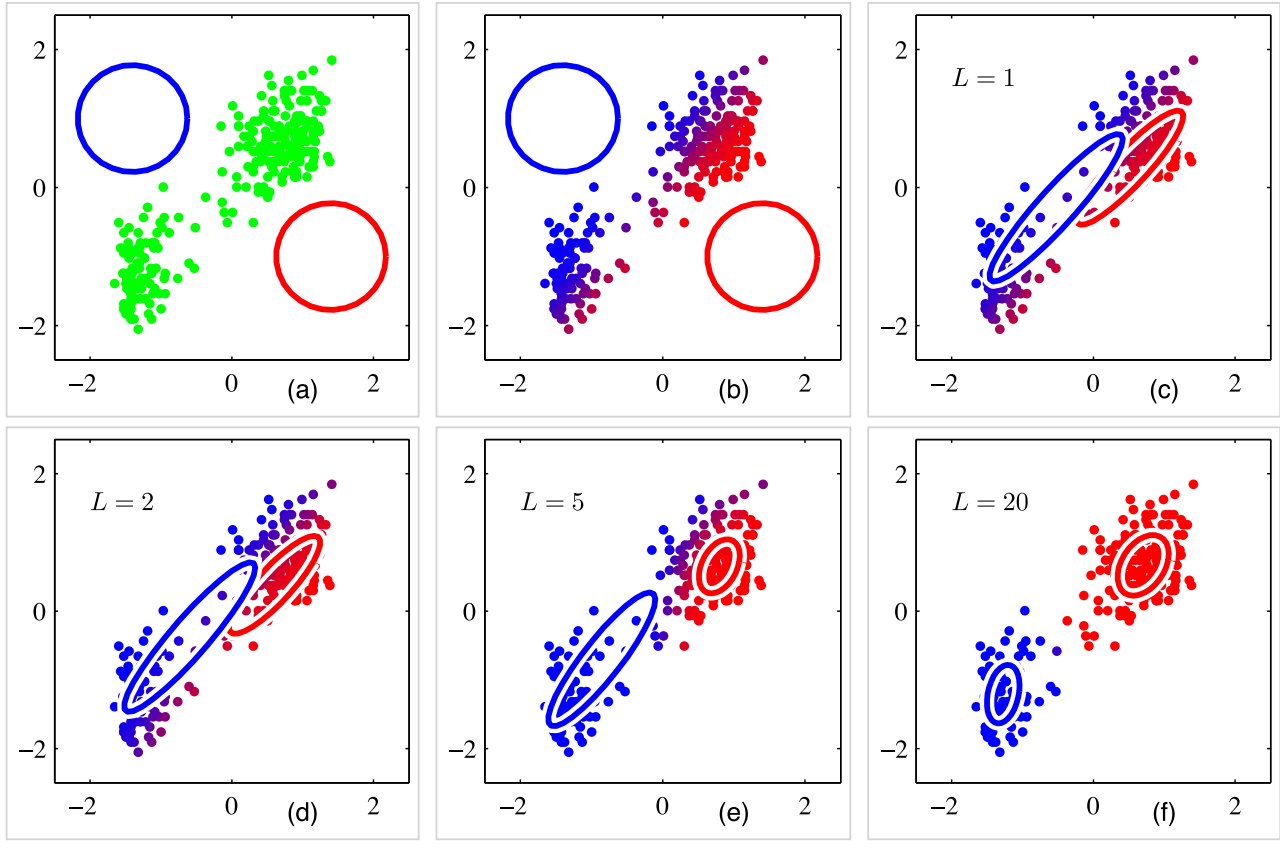
$$\mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\underbrace{\ln(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k)}_{\ln\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)+\ln(\pi_k)} =: f(\boldsymbol{\mu}$$

Considering single component μ_k. Set derivative to zero:

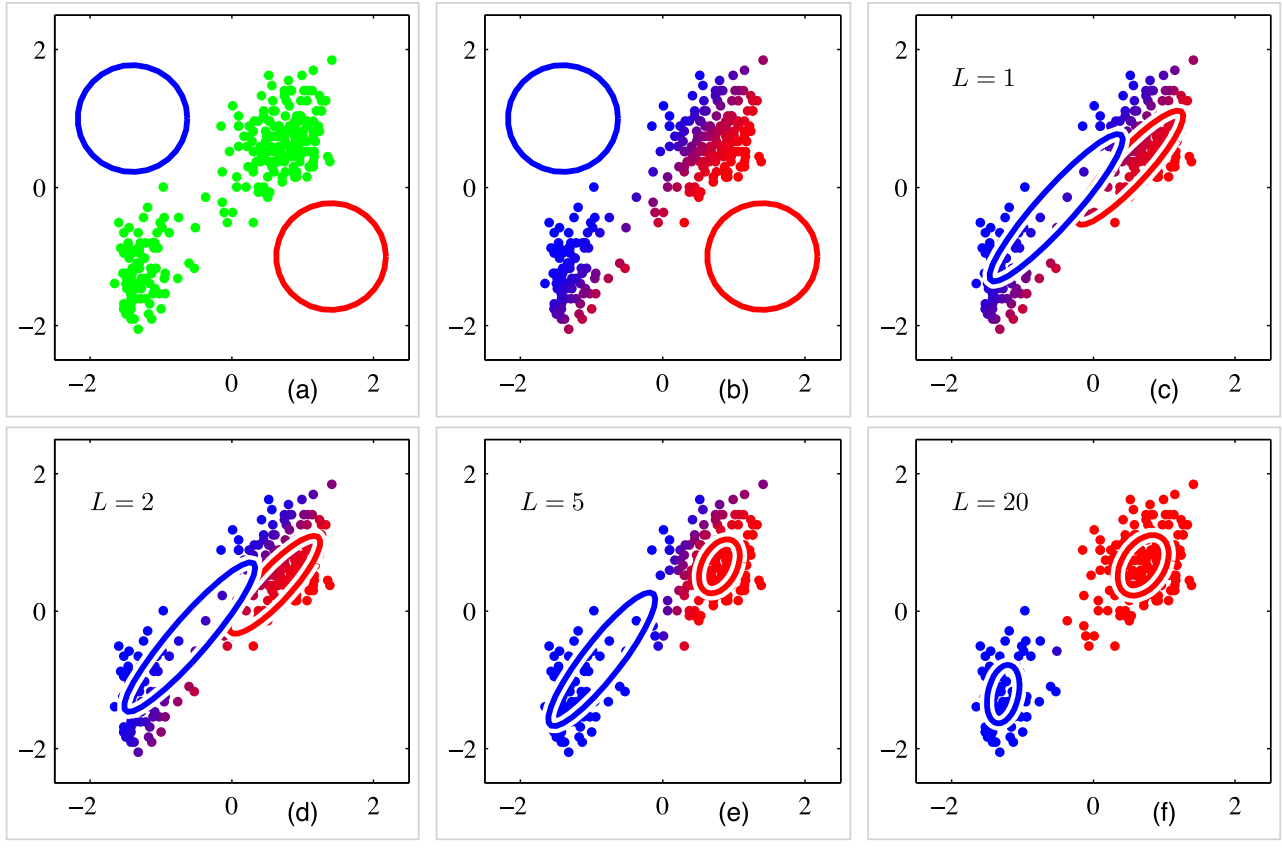
$$\begin{aligned}
\frac{\partial f}{\partial \boldsymbol{\mu}_k} &= \sum_{n=1}^{N}\gamma(z_{nk})\frac{\partial}{\partial \boldsymbol{\mu}_k}(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) \\
&= \sum_{n=1}^{N}\gamma(z_{nk})\underbrace{\frac{\partial}{\partial \boldsymbol{\mu}_k}\left(\frac{1}{2}(\boldsymbol{x}_n - \mu_k)^T\Sigma_k^{-1}(\boldsymbol{x}_n - \mu_k)\right)}_{\text{Look up in Matrix cookbook: eq 86}} \\
&= \sum_{n=1}^{N}\gamma(z_{nk})\Sigma_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \quad \Rightarrow \quad \boldsymbol{\mu}_k = \frac{1}{\sum_{n=1}^{N}\gamma(z_{nk})}\sum_{n=1}^{N}\gamma
\end{aligned}$$

similar for the other parameters

EM illustration



EM illustration



Note: you can get faster convergence using Kmeans as an initializer.

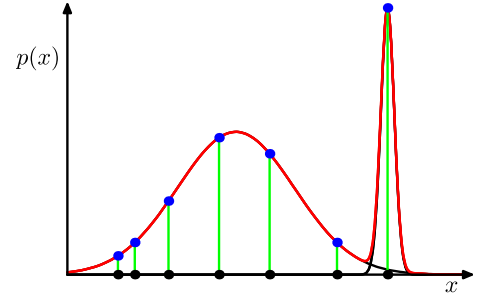
Is maximum likelihood a good idea?

Does it makes sense to optimize the likelihood?

Is maximum likelihood a good idea?

Does it makes sense to optimize the likelihood?

Not really. It's ill-posed.



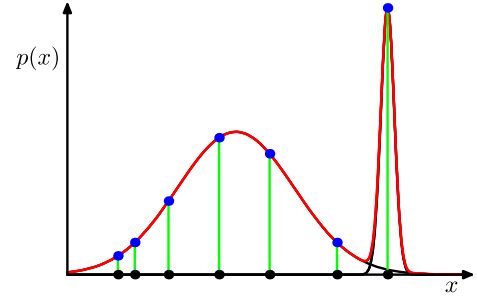
Is maximum likelihood a good idea?

Does it makes sense to optimize the likelihood?

Not really. It's ill-posed.

Why does EM work then?

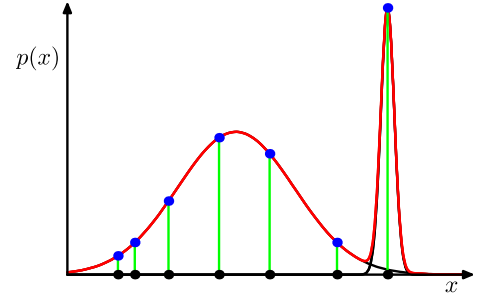
Reasonable initialization + local optimization.



Is maximum likelihood a good idea?

Does it makes sense to optimize the likelihood?

Not really. It's ill-posed.



Why does EM work then?

Reasonable initialization + local optimization.

What is the proper way to fix this?

We can apply regularization to avoid the covariances in Σ becoming too small.

In the Bayesian setting: add a prior on Σ

EM in terms of log-likelihood bound

The product rule states $p(\boldsymbol{X}) = \frac{p(\boldsymbol{X},\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})}$, so

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) - \ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$$

General setting:
1. $p(\boldsymbol{X}|\boldsymbol{\theta})$ difficult to optimize
2. $p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$ easy to optimize

EM in terms of log-likelihood bound

The product rule states $p(\boldsymbol{X}) = \frac{p(\boldsymbol{X},\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})}$, so

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) - \ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$$

Add and subtract $\ln q(\boldsymbol{Z})$ (valid for any choice of q):

General setting:
1. $p(\boldsymbol{X}|\boldsymbol{\theta})$ difficult to optimize
2. $p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$ easy to optimize

EM in terms of log-likelihood bound

The product rule states $p(\boldsymbol{X}) = \frac{p(\boldsymbol{X},\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})}$, so

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) - \ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$$

Add and subtract $\ln q(\boldsymbol{Z})$ (valid for any choice of q):

$$\begin{aligned}\ln p(\boldsymbol{X}|\boldsymbol{\theta}) &= \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) - \ln q(\boldsymbol{Z}) - (\ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}) - \ln q(\boldsymbol{Z})) \\ &= \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} - \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})}\end{aligned}$$

General setting:
1. $p(\boldsymbol{X}|\boldsymbol{\theta})$ difficult to optimize
2. $p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$ easy to optimize

EM in terms of log-likelihood bound

The product rule states $p(\boldsymbol{X}) = \frac{p(\boldsymbol{X},\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})}$, so

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) - \ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$$

General setting:
1. $p(\boldsymbol{X}|\boldsymbol{\theta})$ difficult to optimize
2. $p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$ easy to optimize

Add and subtract $\ln q(\boldsymbol{Z})$ (valid for any choice of q):

$$\begin{aligned}\ln p(\boldsymbol{X}|\boldsymbol{\theta}) &= \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) - \ln q(\boldsymbol{Z}) - (\ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}) - \ln q(\boldsymbol{Z})) \\ &= \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} - \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})}\end{aligned}$$

Since this holds for any $\boldsymbol{Z}$, it also holds for $\mathbb{E}_{\boldsymbol{Z}}$ under any distribution over $\boldsymbol{Z}$. We choose $q(\boldsymbol{Z})$:

$$\underbrace{\mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{Z})} \ln p(\boldsymbol{X}|\boldsymbol{\theta})}_{=\ln p(\boldsymbol{X}|\boldsymbol{\theta})} = \underbrace{\mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{Z})}\left(\ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right)}_{\mathcal{L}(q,\boldsymbol{\theta})} + \underbrace{\mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{Z})}\left(-\ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}}{q(\boldsymbol{Z})}\right.}_{\mathrm{KL}(q||p)}$$

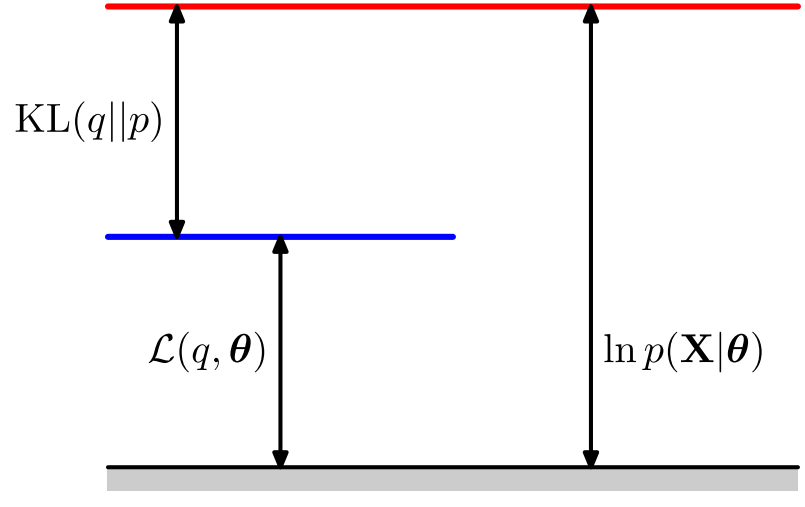
EM in terms of log-likelihood bound (2)

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \underbrace{\mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{Z})}\left(\ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right)}_{\mathcal{L}(q,\boldsymbol{\theta})} + \underbrace{\mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{Z})}\left(-\ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right)}_{\mathrm{KL}(q||p)}$$

The Kullback Leibler divergence, $\mathrm{KL}(q||p)$ specifies how "close" $q(\boldsymbol{Z})$ is to $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$.

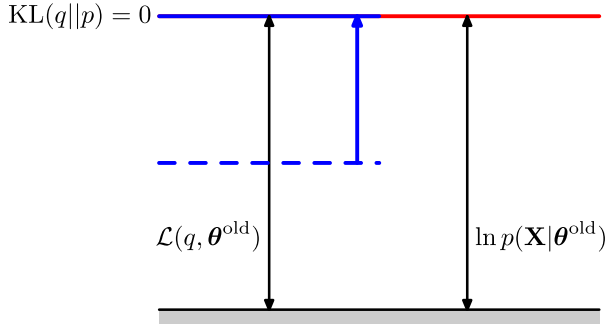
Since $\mathrm{KL}(q||p) >= 0$, $\mathcal{L}(q,\boldsymbol{\theta})$ is a lower bound on $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$.

$\mathcal{L}$ is sometimes referred to as an ELBO (evidence lower bound)



EM in terms of log-likelihood bound (3)

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \underbrace{\mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{Z})}\left(\ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right)}_{\mathcal{L}(q,\boldsymbol{\theta})} + \underbrace{\mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{Z})}\left(-\ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right)}_{\mathrm{KL}(q||p)}$$



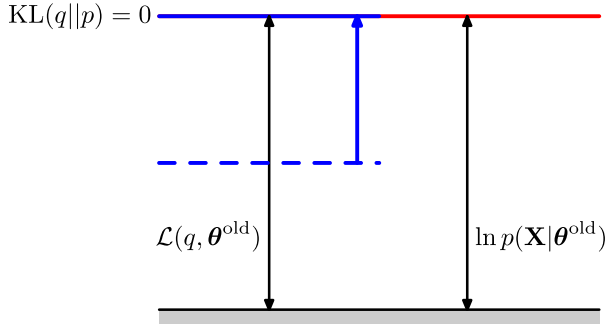
E-step (iteration t)

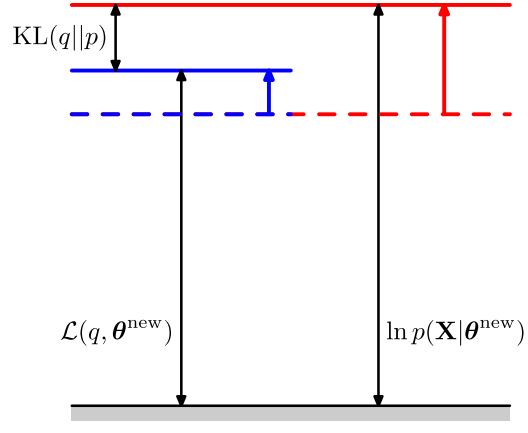
Optimize $\mathcal{L}(q, \boldsymbol{\theta}_{t-1})$ with respect to q, keeping $\boldsymbol{\theta}_{t-1}$ fixed. ?

$\Rightarrow q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}_{t-1})$

EM in terms of log-likelihood bound (3)

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \underbrace{\mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{Z})}\left(\ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right)}_{\mathcal{L}(q,\boldsymbol{\theta})} + \underbrace{\mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{Z})}\left(-\ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right)}_{\mathrm{KL}(q||p)}$$





E-step (iteration t)

Optimize $\mathcal{L}(q, \boldsymbol{\theta}_{t-1})$ with respect to q, keeping $\boldsymbol{\theta}_{t-1}$ fixed. ?

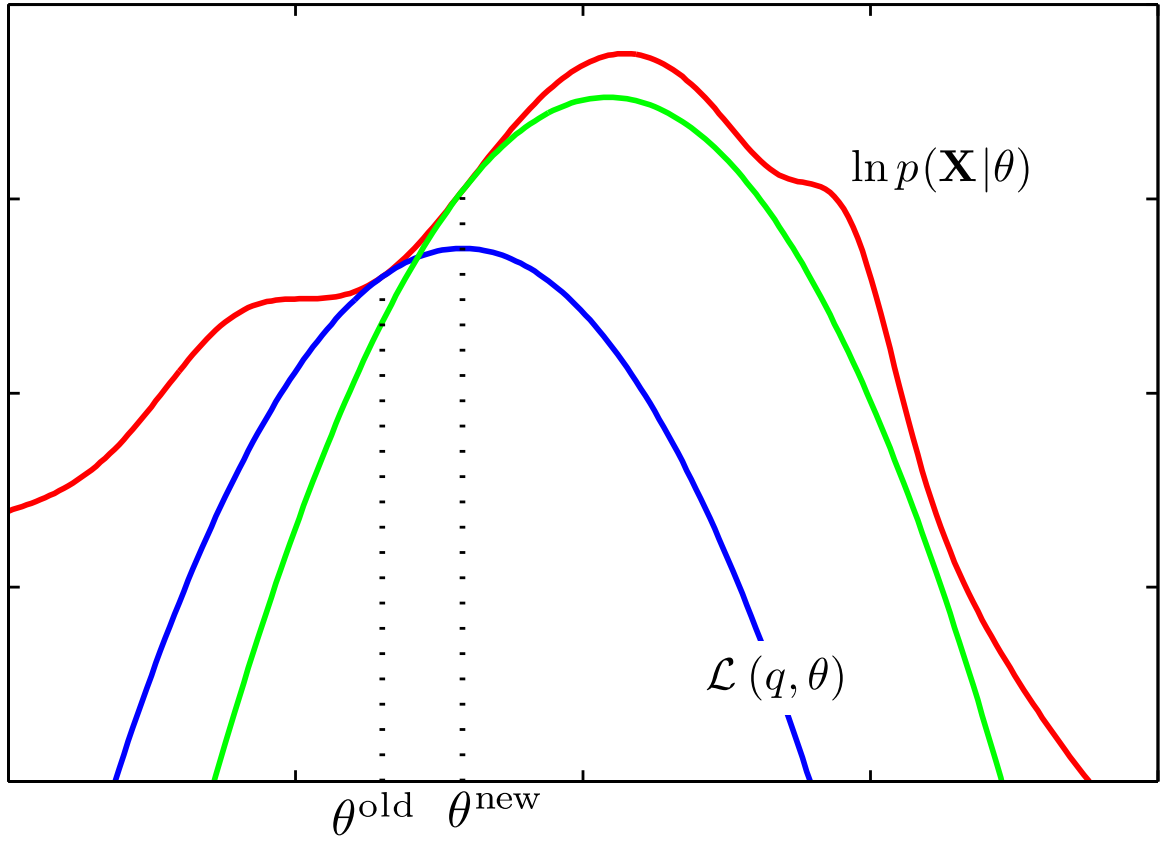
$\Rightarrow q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}_{t-1})$

M-step (iteration t)

Optimize $\mathcal{L}(q, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, keeping q fixed. ?

$\mathcal{L}(q, \boldsymbol{\theta}_t) >= \mathcal{L}(q, \boldsymbol{\theta}_{t-1}) \Rightarrow$
$p(\boldsymbol{X}|\boldsymbol{\theta}_t) >= p(\boldsymbol{X}|\boldsymbol{\theta}_{t-1})$

EM in terms of log-likelihood bound (4)



Mixture models - Concluding remarks

Mixture models allow modelling of complex distributions.

Mixture models - Concluding remarks

Mixture models allow modelling of complex distributions.

The latent variable provides structure/groups in the data

Mixture models - Concluding remarks

Mixture models allow modelling of complex distributions.

The latent variable provides structure/groups in the data

Gaussian mixtures are somewhat similar to Kmeans, but estimating (co)variances in addition to means.

Mixture models - Concluding remarks

Mixture models allow modelling of complex distributions.

The latent variable provides structure/groups in the data

Gaussian mixtures are somewhat similar to Kmeans, but estimating (co)variances in addition to means.

Mixture models can be trained using gradient descent, but the EM algorithm is often more efficient.

Continuous latent variable models

Remember PCA?

Core idea: Reduce the number of dimensions by finding linear combinations of features that capture most of the variance in the data.

Remember PCA?

Core idea: Reduce the number of dimensions by finding linear combinations of features that capture most of the variance in the data.

Two formal definitions:

- Maximize the variance of the projected data.
- Minimize the sum-of-squared of the projection errors.

Remember PCA?

Core idea: Reduce the number of dimensions by finding linear combinations of features that capture most of the variance in the data.

Two formal definitions:

- Maximize the variance of the projected data.
- Minimize the sum-of-squared of the projection errors.

These are equivalent.

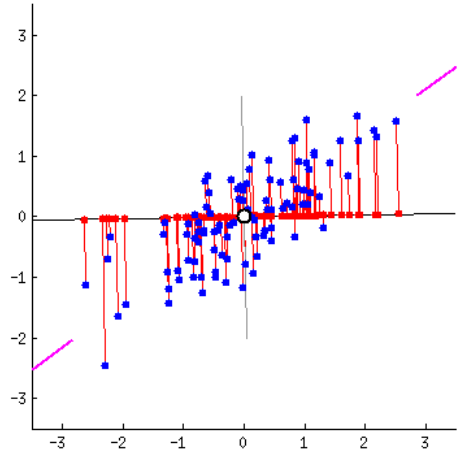


Figure from https://stats.stackexchange.com/a/140579.

Remember PCA?

Core idea: Reduce the number of dimensions by finding linear combinations of features that capture most of the variance in the data.

Two formal definitions:

- Maximize the variance of the projected data.
- Minimize the sum-of-squared of the projection errors.

These are equivalent.

PCA is calculated as the eigenvectors of the data covariance matrix.

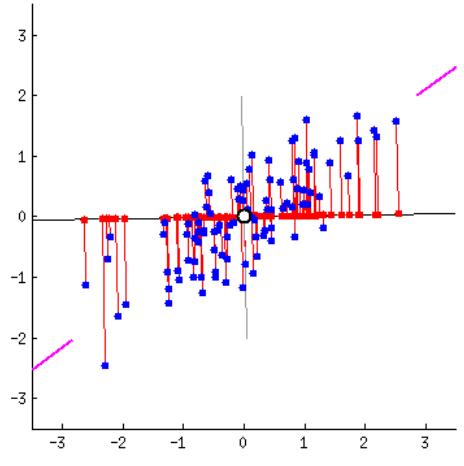


Figure from https://stats.stackexchange.com/a/140579.

Why eigenvectors of the covariance matrix?

Consider a multivariate Gaussian with covariance Σ and mean 0

Since Σ is square, symmetric, we can diagonalize it

$$\Sigma = R\tilde{\Sigma}R^T$$

$$= \begin{bmatrix} | & | & \cdots & | \\ v_1 & v_2 & \cdots & v_D \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D \end{bmatrix} \begin{bmatrix} — & v_1^T & — \\ — & v_2^T & — \\ & \vdots & \\ — & v_D^T & — \end{bmatrix}$$

where $v_1, \ldots, v_D$ are the eigenvectors and $\lambda_1, \ldots, \lambda_D$ are the eigenvalues of Σ.

Why eigenvectors of the covariance matrix? (2)

Plugging $\boldsymbol{\Sigma} = \boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T$ into the Gaussian PDF, we have

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T)^{-1}\boldsymbol{x})$$

Why eigenvectors of the covariance matrix? (2)

Plugging $\boldsymbol{\Sigma} = \boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T$ into the Gaussian PDF, we have

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto \exp(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T)^{-1}\boldsymbol{x}) \\
&\propto \exp(-\frac{1}{2}\boldsymbol{x}^T((\boldsymbol{R}^T)^{-1}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{R}^{-1})\boldsymbol{x}) \qquad \text{inverse of matrix product}
\end{aligned}
$$

Why eigenvectors of the covariance matrix? (2)

Plugging $\boldsymbol{\Sigma} = \boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T$ into the Gaussian PDF, we have

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto \exp(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T)^{-1}\boldsymbol{x}) \\
&\propto \exp(-\frac{1}{2}\boldsymbol{x}^T((\boldsymbol{R}^T)^{-1}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{R}^{-1})\boldsymbol{x}) && \text{inverse of matrix product} \\
&\propto \exp(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{R}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{R}^T)\boldsymbol{x}) && \text{since } \boldsymbol{R} \text{ is orthogonal}
\end{aligned}
$$

Why eigenvectors of the covariance matrix? (2)

Plugging $\boldsymbol{\Sigma} = \boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T$ into the Gaussian PDF, we have

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto \exp(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T)^{-1}\boldsymbol{x}) \\
&\propto \exp(-\frac{1}{2}\boldsymbol{x}^T((\boldsymbol{R}^T)^{-1}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{R}^{-1})\boldsymbol{x}) && \text{inverse of matrix product} \\
&\propto \exp(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{R}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{R}^T)\boldsymbol{x}) && \text{since } \boldsymbol{R} \text{ is orthogonal} \\
&\propto \exp(-\frac{1}{2}(\boldsymbol{R}^T\boldsymbol{x})^T\tilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{R}^T\boldsymbol{x})) && \text{since } (AB)^T = B^TA^T
\end{aligned}
$$

Why eigenvectors of the covariance matrix? (2)

Plugging $\boldsymbol{\Sigma} = \boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T$ into the Gaussian PDF, we have

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto \exp(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T)^{-1}\boldsymbol{x}) \\
&\propto \exp(-\frac{1}{2}\boldsymbol{x}^T((\boldsymbol{R}^T)^{-1}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{R}^{-1})\boldsymbol{x}) && \text{inverse of matrix product} \\
&\propto \exp(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{R}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{R}^T)\boldsymbol{x}) && \text{since } \boldsymbol{R} \text{ is orthogonal} \\
&\propto \exp(-\frac{1}{2}(\boldsymbol{R}^T\boldsymbol{x})^T\tilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{R}^T\boldsymbol{x})) && \text{since } (AB)^T = B^TA^T
\end{aligned}
$$

R can thus be interpreted as the rotation matrix which we can apply to our data such that our new feature dimensions have no covariance.

Why eigenvectors of the covariance matrix? (2)

Plugging $\boldsymbol{\Sigma} = \boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T$ into the Gaussian PDF, we have

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto \exp(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T)^{-1}\boldsymbol{x}) \\
&\propto \exp(-\frac{1}{2}\boldsymbol{x}^T((\boldsymbol{R}^T)^{-1}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{R}^{-1})\boldsymbol{x}) && \text{inverse of matrix product} \\
&\propto \exp(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{R}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{R}^T)\boldsymbol{x}) && \text{since } \boldsymbol{R} \text{ is orthogonal} \\
&\propto \exp(-\frac{1}{2}(\boldsymbol{R}^T\boldsymbol{x})^T\tilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{R}^T\boldsymbol{x})) && \text{since } (AB)^T = B^TA^T
\end{aligned}
$$

R can thus be interpreted as the rotation matrix which we can apply to our data such that our new feature dimensions have no covariance.

Note that the eigen*values* are the variances along these dimensions.

Why eigenvectors of the covariance matrix? (2)

Plugging $\boldsymbol{\Sigma} = \boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T$ into the Gaussian PDF, we have

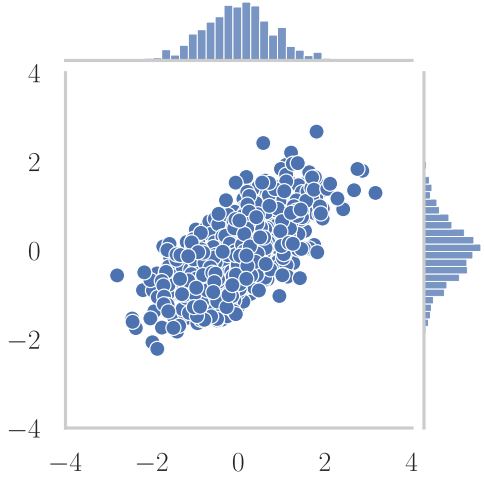
$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto \exp(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T)^{-1}\boldsymbol{x}) \\
&\propto \exp(-\frac{1}{2}\boldsymbol{x}^T((\boldsymbol{R}^T)^{-1}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{R}^{-1})\boldsymbol{x}) && \text{inverse of matrix product} \\
&\propto \exp(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{R}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{R}^T)\boldsymbol{x}) && \text{since } \boldsymbol{R} \text{ is orthogonal} \\
&\propto \exp(-\frac{1}{2}(\boldsymbol{R}^T\boldsymbol{x})^T\tilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{R}^T\boldsymbol{x})) && \text{since } (AB)^T = B^TA^T
\end{aligned}
$$

R can thus be interpreted as the rotation matrix which we can apply to our data such that our new feature dimensions have no covariance.

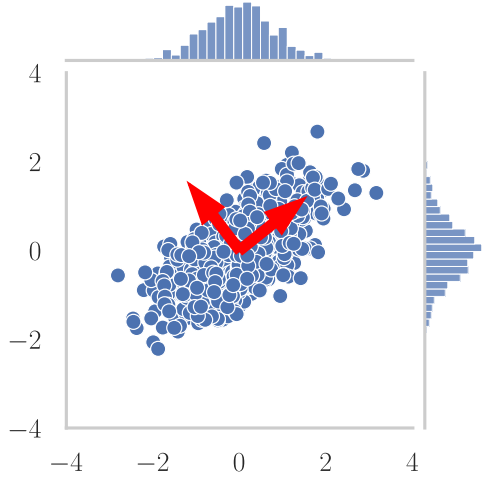
Note that the eigen*values* are the variances along these dimensions.

By sorting according to the eigenvalues (descending order) and choosing the top k, we get a dimensionality reduction scheme that explains $\frac{\sum_i^k \lambda_i}{\sum_i^D \lambda_i}$ of the total variance.

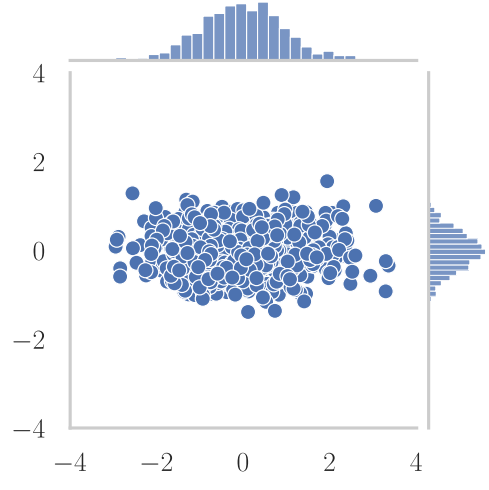
Example



$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.7 & 0.4 \\ 0.4 & 0.5 \end{bmatrix}$$
$$= \boldsymbol{R}\tilde{\boldsymbol{\Sigma}}\boldsymbol{R}^T$$



$$\boldsymbol{R} = \begin{bmatrix} 0.79 & -0.62 \\ 0.62 & 0.79 \end{bmatrix}$$
$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} 1.01 & 0 \\ 0 & 0.19 \end{bmatrix}$$



$$\boldsymbol{R}^T\boldsymbol{x}$$

Probabilistic PCA

A probabilistic equivalent of PCA can be formulated as a latent variable model - where the latent variable is now continous.

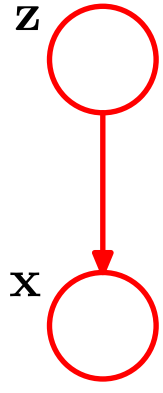
Ingredients:

Prior distribution over $\boldsymbol{z}$

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|0, \boldsymbol{I})$$

Distribution over output variable - conditioned on $\boldsymbol{z}$

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I})$$



(same graphical model as for the discrete case)

Probabilistic PCA

A probabilistic equivalent of PCA can be formulated as a latent variable model - where the latent variable is now continous.

Ingredients:

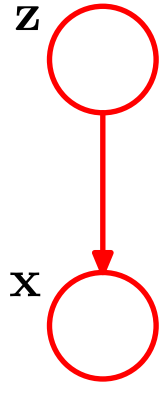
Prior distribution over $\boldsymbol{z}$

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|0, \boldsymbol{I})$$

Distribution over output variable - conditioned on $\boldsymbol{z}$

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I})$$

The latent $\boldsymbol{z}$ is M-dimensional, and $\boldsymbol{W}$ is a $D \times M$ matrix mapping linearly from $\boldsymbol{z}$ to $\boldsymbol{x}$.



(same graphical model as for the discrete case)

Probabilistic PCA

A probabilistic equivalent of PCA can be formulated as a latent variable model - where the latent variable is now continous.

Ingredients:

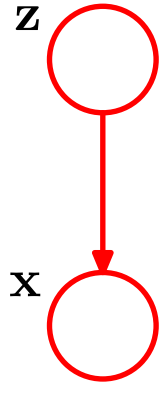
Prior distribution over $\boldsymbol{z}$

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|0, \boldsymbol{I})$$

Distribution over output variable - conditioned on $\boldsymbol{z}$

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I})$$

The latent $\boldsymbol{z}$ is M-dimensional, and $\boldsymbol{W}$ is a $D \times M$ matrix mapping linearly from $\boldsymbol{z}$ to $\boldsymbol{x}$.

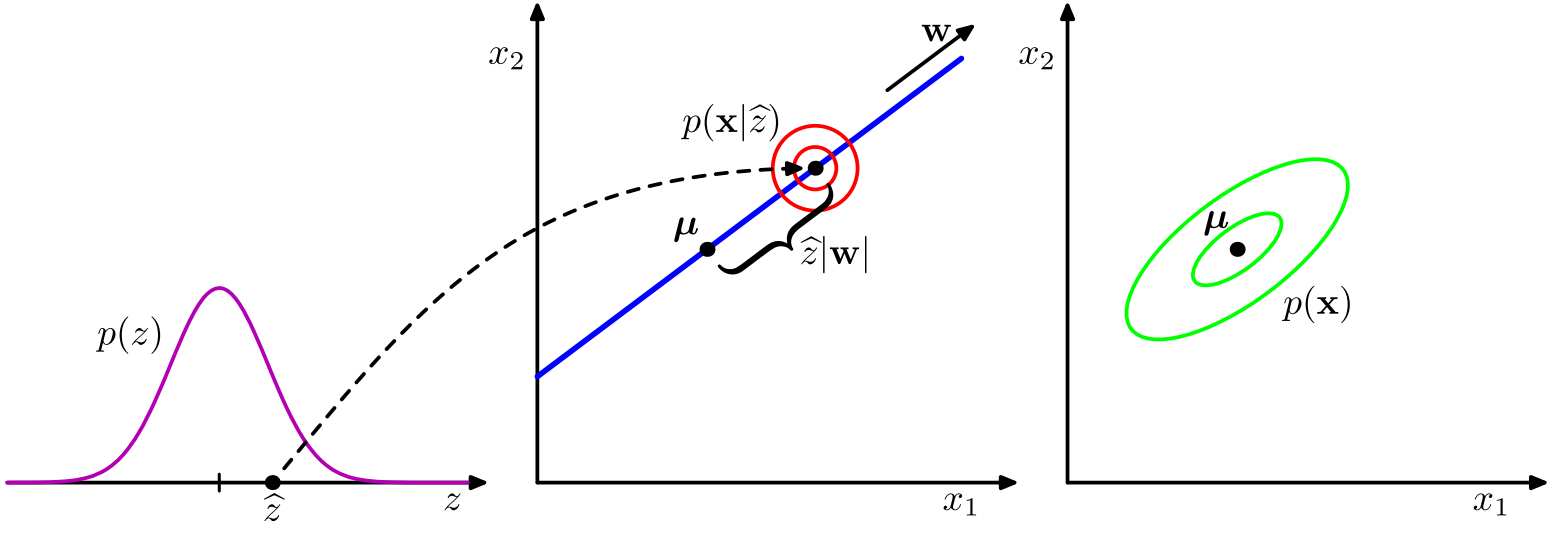


(same graphical model as for the discrete case)

Note the simplicity of the model: the elements of x are independent given $\boldsymbol{z}$, and the variances for all elements are the same.

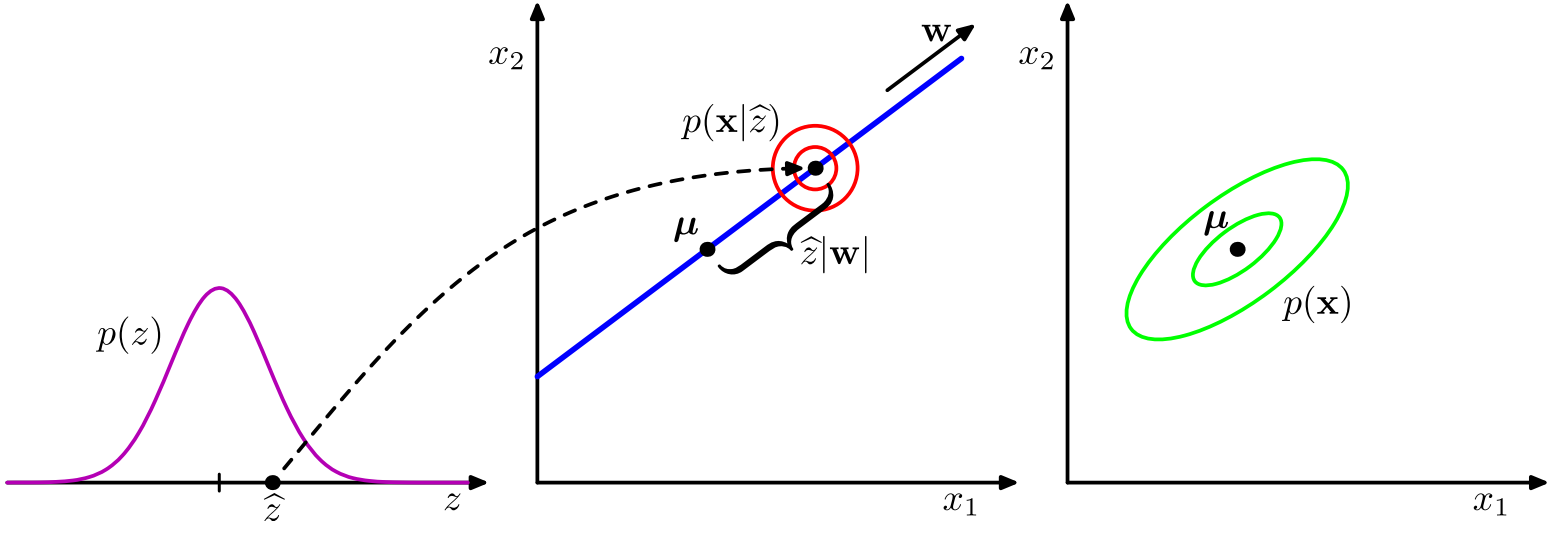
PPCA as a generative model

Similar to mixture models, we can generate samples from the model by first drawing a sample $\widehat{\boldsymbol{z}} \sim p(\boldsymbol{z})$, and then sampling $\boldsymbol{x} \sim p(\boldsymbol{x}|\widehat{\boldsymbol{z}})$:



PPCA as a generative model

Similar to mixture models, we can generate samples from the model by first drawing a sample $\widehat{\boldsymbol{z}} \sim p(\boldsymbol{z})$, and then sampling $\boldsymbol{x} \sim p(\boldsymbol{x}|\widehat{\boldsymbol{z}})$:



Note that $\boldsymbol{x} \sim p(\boldsymbol{x}|\widehat{\boldsymbol{z}})$ can be implemented as

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu} + \sigma\boldsymbol{\epsilon} \qquad \text{where } \epsilon \sim \mathcal{N}(0, \boldsymbol{I})$$

PPCA: The joint is Gaussian

A special property of the PPCA model is that the joint is Gaussian.

$$\begin{aligned}
&\ln p(\boldsymbol{x}, \boldsymbol{z}) = \ln(p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})) \\
&= \ln p(\boldsymbol{x}|\boldsymbol{z}) + \ln p(\boldsymbol{z}) = \ln\left(\mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I}) + \mathcal{N}(\boldsymbol{z}|0, \boldsymbol{I})\right) \\
&= -\left(\frac{1}{2\sigma^2}(\boldsymbol{x} - \boldsymbol{W}\boldsymbol{z} - \boldsymbol{\mu})^T(\boldsymbol{x} - \boldsymbol{W}\boldsymbol{z} - \boldsymbol{\mu}) + \frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}\right. \\
&\qquad\qquad \left. + \frac{D}{2}\ln(2\pi\sigma^2) + \frac{M}{2}\ln(2\pi)\right)
\end{aligned}$$

PPCA: The joint is Gaussian

A special property of the PPCA model is that the joint is Gaussian.

$$
\begin{aligned}
&\ln p(\boldsymbol{x}, \boldsymbol{z}) = \ln(p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})) \\
&= \ln p(\boldsymbol{x}|\boldsymbol{z}) + \ln p(\boldsymbol{z}) = \ln\left(\mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I}) + \mathcal{N}(\boldsymbol{z}|0, \boldsymbol{I})\right) \\
&= -\left(\frac{1}{2\sigma^2}(\boldsymbol{x} - \boldsymbol{W}\boldsymbol{z} - \boldsymbol{\mu})^T(\boldsymbol{x} - \boldsymbol{W}\boldsymbol{z} - \boldsymbol{\mu}) + \frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}\right. \\
&\qquad\qquad \left. + \frac{D}{2}\ln(2\pi\sigma^2) + \frac{M}{2}\ln(2\pi)\right) \\
&= -\frac{1}{2\sigma^2}\Bigg(\underbrace{\boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{W}\boldsymbol{z} - \boldsymbol{z}^T\boldsymbol{W}^T\boldsymbol{x} + \boldsymbol{z}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{z} + \sigma^2\boldsymbol{z}^T\boldsymbol{z}}_{= \begin{pmatrix}\boldsymbol{z}\\ \boldsymbol{x}\end{pmatrix}^T \begin{pmatrix}\boldsymbol{W}^T\boldsymbol{W} + \sigma^2\boldsymbol{I} & -\boldsymbol{W}^T \\ -\boldsymbol{W} & \boldsymbol{I}\end{pmatrix}\begin{pmatrix}\boldsymbol{z}\\ \boldsymbol{x}\end{pmatrix}} \\
&\qquad \underbrace{- \boldsymbol{x}^T\boldsymbol{\mu} + \boldsymbol{z}^T\boldsymbol{W}^T\boldsymbol{\mu} - \boldsymbol{\mu}^T\boldsymbol{x} + \boldsymbol{\mu}^T\boldsymbol{W}\boldsymbol{z}}_{= -2\begin{pmatrix}\boldsymbol{z}\\ \boldsymbol{x}\end{pmatrix}^T \begin{pmatrix}-\boldsymbol{W}^T\boldsymbol{\mu} \\ \boldsymbol{\mu}\end{pmatrix}} + \boldsymbol{\mu}^T\boldsymbol{\mu}\Bigg) - \frac{D}{2}\ln(2\pi\sigma^2) - \frac{M}{2}
\end{aligned}
$$

PPCA: The joint is Gaussian (2)

$$\ln p(\boldsymbol{x}, \boldsymbol{z}) = -\frac{1}{2\sigma^2} \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{x} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{W}^T\boldsymbol{W} + \sigma^2\boldsymbol{I} & -\boldsymbol{W}^T \\ -\boldsymbol{W} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{x} \end{pmatrix} + \frac{1}{\sigma^2} \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{x} \end{pmatrix}^T \begin{pmatrix} -\boldsymbol{W}^T\boldsymbol{\mu} \\ \boldsymbol{\mu} \end{pmatrix} + c$$

PPCA: The joint is Gaussian (2)

$$\ln p(\boldsymbol{x}, \boldsymbol{z}) = -\frac{1}{2\sigma^2} \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{x} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{W}^T\boldsymbol{W} + \sigma^2\boldsymbol{I} & -\boldsymbol{W}^T \\ -\boldsymbol{W} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{x} \end{pmatrix} + \frac{1}{\sigma^2} \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{x} \end{pmatrix}^T \begin{pmatrix} -\boldsymbol{W}^T\boldsymbol{\mu} \\ \boldsymbol{\mu} \end{pmatrix} + c$$

In general, we can write the exponent of a Gaussian as

$$-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = -\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}$$

PPCA: The joint is Gaussian (2)

$$\ln p(\boldsymbol{x}, \boldsymbol{z}) = -\frac{1}{2\sigma^2} \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{x} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{W}^T\boldsymbol{W} + \sigma^2\boldsymbol{I} & -\boldsymbol{W}^T \\ -\boldsymbol{W} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{x} \end{pmatrix} + \frac{1}{\sigma^2} \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{x} \end{pmatrix}^T \begin{pmatrix} -\boldsymbol{W}^T\boldsymbol{\mu} \\ \boldsymbol{\mu} \end{pmatrix} + c$$

In general, we can write the exponent of a Gaussian as

$$-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = -\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}$$

We thus recognize $\ln p(\boldsymbol{x}, \boldsymbol{z})$ as a Gaussian log density $\mathcal{N}(\mu_{z,x}, \Sigma_{z,x})$ with:

$$\boldsymbol{\Sigma}_{z,x}^{-1} = \begin{pmatrix} \frac{1}{\sigma^2}\boldsymbol{W}^T\boldsymbol{W} + \boldsymbol{I} & -\frac{1}{\sigma^2}\boldsymbol{W}^T \\ -\frac{1}{\sigma^2}\boldsymbol{W} & \frac{1}{\sigma^2}\boldsymbol{I} \end{pmatrix} \quad \Rightarrow \quad \underbrace{\boldsymbol{\Sigma}_{z,x} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{W}^T \\ \boldsymbol{W} & \sigma^2\boldsymbol{I} + \boldsymbol{W}\boldsymbol{W}^T \end{pmatrix}}_{\text{Matrix Cookbook eq 399-400)}}$$

and

$$\boldsymbol{\Sigma}_{z,x}^{-1}\boldsymbol{\mu}_{z,x} = \frac{1}{\sigma^2} \begin{pmatrix} -\boldsymbol{W}^T\boldsymbol{\mu} \\ \boldsymbol{\mu} \end{pmatrix} \quad \Rightarrow \quad \boldsymbol{\mu}_{z,x} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{\mu} \end{pmatrix}$$

PPCA: marginal $p(\boldsymbol{x})$ is also Gaussian

We now know that $p(\boldsymbol{x}, \boldsymbol{z})$ is jointly Gaussian

$$p(\boldsymbol{x}, \boldsymbol{z}) = \mathcal{N}(\boldsymbol{\mu}_{z,x}, \boldsymbol{\Sigma}_{z,x}) \qquad \boldsymbol{\mu}_{z,x} = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\mu} \end{pmatrix} \qquad \boldsymbol{\Sigma}_{z,x} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{W}^T \\ \boldsymbol{W} & \sigma^2 \boldsymbol{I} + \boldsymbol{W} \end{pmatrix}$$

That means that the marginal $p(\boldsymbol{x})$ is also Gaussian. We can actually read off its parameters directly from $\boldsymbol{\mu}_{z,x}$ and $\boldsymbol{\Sigma}_{z,x}$

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \qquad \boldsymbol{\mu}_x = \boldsymbol{\mu} \qquad \boldsymbol{\Sigma}_x = \sigma^2 \boldsymbol{I} + \boldsymbol{W}\boldsymbol{W}^T$$

PPCA: the posterior is also Gaussian

Remember - joint is gaussian:

$$p(\boldsymbol{x}, \boldsymbol{z}) = \mathcal{N}(\boldsymbol{\mu}_{z,x}, \boldsymbol{\Sigma}_{z,x}) \qquad \boldsymbol{\mu}_{z,x} = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\mu} \end{pmatrix} \qquad \boldsymbol{\Sigma}_{z,x} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{W}^T \\ \boldsymbol{W} & \sigma^2 \boldsymbol{I} + \boldsymbol{W}\boldsymbol{W}^T \end{pmatrix}$$

Since the joint is Gaussian, we can evaluate the conditional $p(\boldsymbol{z}|\boldsymbol{x})$

Using eq 353 from the matrix cookbook, we have:

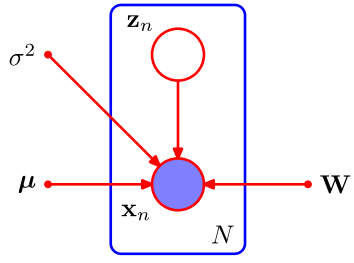
$$p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{z}|\mathbf{0} + \boldsymbol{W}^T(\sigma^2\boldsymbol{I} + \boldsymbol{W}\boldsymbol{W}^T)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \\ \boldsymbol{I} - \boldsymbol{W}^T(\sigma^2\boldsymbol{I} + \boldsymbol{W}\boldsymbol{W}^T)^{-1}\boldsymbol{W}\big)$$

We can use the Woodbury identity (cookbook eq 159) for the inversion:

$$(\sigma^2\boldsymbol{I} + \boldsymbol{W}\boldsymbol{W}^T)^{-1} = \sigma^{-2}\boldsymbol{I} - \sigma^{-2}\boldsymbol{W}\boldsymbol{M}^{-1}\boldsymbol{W}^T \qquad \text{where } \boldsymbol{M} = \boldsymbol{W}^T\boldsymbol{W} + \sigma$$

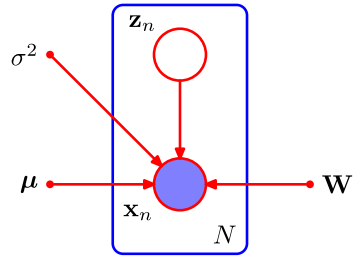
Parameter estimation - Maximum likelihood (1)

Log likelihood function:



$$
\begin{aligned}
ll(\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2) &= \ln p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2) \\
&= \sum_{n=1}^{N} \ln p(\boldsymbol{x}_n|\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) \\
&= -\frac{N}{2}\left(\frac{D}{2}\ln(2\pi) + \ln(\det(\boldsymbol{\Sigma}_x))\right) - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}_x^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})
\end{aligned}
$$

Parameter estimation - Maximum likelihood (1)



Log likelihood function:

$$
\begin{aligned}
ll(\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2) &= \ln p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2) \\
&= \sum_{n=1}^{N} \ln p(\boldsymbol{x}_n|\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) \\
&= -\frac{N}{2}\left(\frac{D}{2}\ln(2\pi) + \ln(\det(\boldsymbol{\Sigma}_x))\right) - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}_x^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})
\end{aligned}
$$

Optimize wrt $\boldsymbol{\mu}$:

$$
\begin{aligned}
\frac{\partial ll}{\partial \boldsymbol{\mu}} &= -\frac{1}{2}\sum_{n=1}^{N}\frac{\partial ll}{\partial \boldsymbol{\mu}}\left(\boldsymbol{x}_n - \boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}_x^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) \\
&= -\frac{1}{2}\sum_{n=1}^{N} -2\boldsymbol{\Sigma}_x(\boldsymbol{x}_n - \boldsymbol{\mu}) \qquad \text{Matrix cookbook eq 86} \\
&= 0 \Rightarrow \boldsymbol{\mu}_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n = \bar{\boldsymbol{x}}
\end{aligned}
$$

Parameter estimation - Maximum likelihood (2)

Optimizing $ll(\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2)$ wrt to $\boldsymbol{W}$ and σ^2 is more difficult.

Tipping and Bishop prove in a 1999 paper that the likelihood is optimized when

$$W_{ML} = \boldsymbol{U}_M(\boldsymbol{L}_M - \sigma^2\boldsymbol{I})^{1/2}\boldsymbol{R}$$

$$\sigma^2_{ML} = \frac{1}{D-M}\sum_{i=M+1}^{D}\lambda_i$$

Where $\boldsymbol{U}_M$ is a $D \times M$ matrix containing the M eigen vectors with largest eigenvalues, and $\boldsymbol{L}_M$ is a diagonal matrix with the corresponding eigenvalues.

Parameter estimation - Maximum likelihood (2)

Optimizing $ll(\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2)$ wrt to $\boldsymbol{W}$ and σ^2 is more difficult.

Tipping and Bishop prove in a 1999 paper that the likelihood is optimized when

$$W_{ML} = \boldsymbol{U}_M(\boldsymbol{L}_M - \sigma^2\boldsymbol{I})^{1/2}\boldsymbol{R}$$

$$\sigma^2_{ML} = \frac{1}{D-M}\sum_{i=M+1}^{D}\lambda_i$$

Where $\boldsymbol{U}_M$ is a $D \times M$ matrix containing the M eigen vectors with largest eigenvalues, and $\boldsymbol{L}_M$ is a diagonal matrix with the corresponding eigenvalues.

The $\boldsymbol{W}$ thus behaves exactly as in regular PCA. We can therefore also estimate it in the same way.

Parameter estimation - Maximum likelihood (2)

Optimizing $ll(\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2)$ wrt to $\boldsymbol{W}$ and σ^2 is more difficult.

Tipping and Bishop prove in a 1999 paper that the likelihood is optimized when

$$W_{ML} = \boldsymbol{U}_M(\boldsymbol{L}_M - \sigma^2\boldsymbol{I})^{1/2}\boldsymbol{R}$$

$$\sigma^2_{ML} = \frac{1}{D-M}\sum_{i=M+1}^{D}\lambda_i$$

Where $\boldsymbol{U}_M$ is a $D \times M$ matrix containing the M eigen vectors with largest eigenvalues, and $\boldsymbol{L}_M$ is a diagonal matrix with the corresponding eigenvalues.

The $\boldsymbol{W}$ thus behaves exactly as in regular PCA. We can therefore also estimate it in the same way.

$\boldsymbol{R}$ is an arbitrary orthogonal matrix, which does not affect the covariance matrix:

$$\begin{aligned}\boldsymbol{C}_{\mathrm{ML}} = \sigma^2\boldsymbol{I} + & \underbrace{\boldsymbol{W}_{\mathrm{ML}}\boldsymbol{W}_{\mathrm{ML}}^T} \\ &= (\boldsymbol{U}_M(\boldsymbol{L}_M - \sigma^2\boldsymbol{I})^{1/2}\boldsymbol{R})(\boldsymbol{U}_M(\boldsymbol{L}_M - \sigma^2\boldsymbol{I})^{1/2}\boldsymbol{R})^T \\ &= \boldsymbol{U}_M(\boldsymbol{L}_M - \sigma^2\boldsymbol{I})^{1/2}\underbrace{\boldsymbol{R}\boldsymbol{R}^T}_{=\boldsymbol{I}}(\boldsymbol{U}_M(\boldsymbol{L}_M - \sigma^2\boldsymbol{I})^{1/2})^T\end{aligned}$$

i.e. an arbitrary rotation of latent space.

Interpretation of Probabilistic PCA

Consider a direction in your output space, represented by unit vector $\boldsymbol{v}$

According to the PPCA model, the variance along this direction is $\boldsymbol{v}^T \boldsymbol{\Sigma}_x \boldsymbol{v}$.

Consider two scenarios:

1. $\boldsymbol{v}$ is orthogonal to the M selected principal component vectors

$$\boldsymbol{v}^T \boldsymbol{\Sigma}_x \boldsymbol{v} = \boldsymbol{v}^T \sigma^2 \boldsymbol{I} \boldsymbol{v} + \underbrace{\boldsymbol{v}^T (\boldsymbol{U}_M}_{=0} (\boldsymbol{L}_M - \sigma^2 \boldsymbol{I})^{1/2} (\boldsymbol{U}_M (\boldsymbol{L}_M - \sigma^2 \boldsymbol{I})^{1/2})^T \boldsymbol{v} = \sigma^2$$

2. $\boldsymbol{v}$ points along the ith eigenvector

$$\boldsymbol{v}^T \boldsymbol{\Sigma} \boldsymbol{v} = \boldsymbol{v}^T \sigma^2 \boldsymbol{I} \boldsymbol{v} + \underbrace{\boldsymbol{v}^T (\boldsymbol{U}_M (\boldsymbol{L}_M - \sigma^2 \boldsymbol{I})^{1/2}}_{(\lambda_i - \sigma^2)^{1/2}} \underbrace{((\boldsymbol{L}_M - \sigma^2 \boldsymbol{I})^{1/2})^T \boldsymbol{U}_M^T \boldsymbol{v}}_{(\lambda_i - \sigma^2)^{1/2}} = \lambda_i$$

This is quite nice! For all the principal components that we have included, the model correctly captures the variance of the data. For those directions that we leave out, it uses $\sigma^2 = \frac{1}{D-M} \sum_{i=M+1}^{D} \lambda_i$, i.e. the average variance of the left-out principal components.

Application of PPCA

When modelling continuous data in D dimensions, you will often have to make the following choice:

Use a full covariance matrix
: For high dimensions, the $D \times D$ can be difficult to estimate, requiring a lot of data.

Use a diagonal covariance matrix
: Now, you only need to estimate D variance values - but you assume that there is no covariance among the dimensions, which is probably unrealistic.

Application of PPCA

When modelling continuous data in D dimensions, you will often have to make the following choice:

Use a full covariance matrix
: For high dimensions, the $D \times D$ can be difficult to estimate, requiring a lot of data.

Use a diagonal covariance matrix
: Now, you only need to estimate D variance values - but you assume that there is no covariance among the dimensions, which is probably unrealistic.

Probabilistic PCA provides an intermediate between these extremes - allowing you to include the most important covariances while summarizing the remaining variances with an average.

EM for Probabilistic PCA

Can't we use EM to estimate the parameters of the PPCA model?

EM for Probabilistic PCA

Can't we use EM to estimate the parameters of the PPCA model?

Yes we can! But why would we want to do that when we have a closed form solution?

EM for Probabilistic PCA

Can't we use EM to estimate the parameters of the PPCA model?

Yes we can! But why would we want to do that when we have a closed form solution?

- Can be faster in higher dimensions.

EM for Probabilistic PCA

Can't we use EM to estimate the parameters of the PPCA model?

Yes we can! But why would we want to do that when we have a closed form solution?

- Can be faster in higher dimensions.
- Is necessary for other models, like Factor analysis.

EM for Probabilistic PCA

Can't we use EM to estimate the parameters of the PPCA model?

Yes we can! But why would we want to do that when we have a closed form solution?

- Can be faster in higher dimensions.
- Is necessary for other models, like Factor analysis.
- Natural way to deal with missing data.

EM for Probabilistic PCA (2)

Same procedure as for mixture models

In iteration t:

- Expectation step:
  - Calculate the expectation of the complete-data likelihood $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2)$, under the distribution $p(z|x, \boldsymbol{\mu}_{t-1}, \boldsymbol{W}_{t-1}, \sigma^2_{t-1})$
- Maximization step
  - Optimize this completed likelihood to obtain new values $\boldsymbol{\mu}_t, \boldsymbol{W}_t, \sigma^2_t$

EM for Probabilistic PCA - E-step

Complete-data log-likelihood:

$$
\begin{aligned}
ll_c = \ln p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\mu}, \boldsymbol{W}, \sigma^2) &= \sum_{n=1}^{N} \left(\ln p(\boldsymbol{x}_n | z_n) + \ln p(z_n) \right) \\
&= \sum_{n=1}^{N} \left(\ln \left(\mathcal{N}(\boldsymbol{x}_n | \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) + \mathcal{N}(z | 0, \boldsymbol{I}) \right) \right) \\
&= -\sum_{n=1}^{N} \left(\frac{D}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} ||\boldsymbol{x}_n - \boldsymbol{W}\boldsymbol{z}_n - \boldsymbol{\mu}||^2 + \frac{M}{2} \ln(2\pi) + \right. \\
&= -\sum_{n=1}^{N} \left(\frac{D}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} ||\boldsymbol{x}_n - \boldsymbol{\mu}||^2 + \frac{1}{2\sigma^2} \underbrace{(\boldsymbol{W}\boldsymbol{z}_n)^T (\boldsymbol{W}}_{= \boldsymbol{z}_n^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{z}_n = \mathrm{tr}(\boldsymbol{z}_n \boldsymbol{z}} \right. \\
&\qquad \left. \left. - \frac{1}{\sigma^2} \underbrace{(\boldsymbol{W}\boldsymbol{z}_n)^T}_{= \boldsymbol{z}_n^T \boldsymbol{W}^T} (\boldsymbol{x}_n - \boldsymbol{\mu}) + \frac{M}{2} \ln(2\pi) + \frac{1}{2} \underbrace{\boldsymbol{z}_n^T \boldsymbol{z}_n}_{= \mathrm{tr}(\boldsymbol{z}_n \boldsymbol{z}_n^T)} \right) \right.
\end{aligned}
$$

EM for Probabilistic PCA - E-step (2)

Taking the expectation wrt z:

$$\mathbb{E}_{\boldsymbol{Z}}[ll_c] = -\sum_{n=1}^{N}\Bigg(\frac{D}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}||\boldsymbol{x}_n - \boldsymbol{\mu}||^2 + \frac{1}{2\sigma^2}\text{tr}(\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{z}_n\boldsymbol{z}_n^T]\boldsymbol{W}^T\boldsymbol{W}) \\ - \frac{1}{\sigma^2}(\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{z}_n]^T\boldsymbol{W}^T)(\boldsymbol{x}_n - \boldsymbol{\mu}) + \frac{M}{2}\ln(2\pi) + \frac{1}{2}\text{tr}(\boldsymbol{z}_n\boldsymbol{z}_n^T)\Bigg)$$

EM for Probabilistic PCA - E-step (2)

Taking the expectation wrt z:

$$\mathbb{E}_{\boldsymbol{Z}}[ll_c] = -\sum_{n=1}^{N}\Bigg(\frac{D}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}||\boldsymbol{x}_n - \boldsymbol{\mu}||^2 + \frac{1}{2\sigma^2}\text{tr}(\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{z}_n\boldsymbol{z}_n^T]\boldsymbol{W}^T\boldsymbol{W}) - \frac{1}{\sigma^2}(\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{z}_n]^T\boldsymbol{W}^T)(\boldsymbol{x}_n - \boldsymbol{\mu}) + \frac{M}{2}\ln(2\pi) + \frac{1}{2}\text{tr}(\boldsymbol{z}_n\boldsymbol{z}_n^T)\Bigg)$$

Remember that all these expectations are taken wrt

$$p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{M}_{t-1}^{-1}\boldsymbol{W}_{t-1}^T(\boldsymbol{x} - \boldsymbol{\mu}_{t-1}), \sigma_{t-1}^2\boldsymbol{M}_{t-1}^{-1}) \quad \text{where } \boldsymbol{M}_{t-1} = \boldsymbol{W}_{t-1}^T\boldsymbol{W}_{t-}$$

EM for Probabilistic PCA - E-step (2)

Taking the expectation wrt z:

$$\mathbb{E}_{\boldsymbol{Z}}[ll_c] = -\sum_{n=1}^{N}\Bigg(\frac{D}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}||\boldsymbol{x}_n - \boldsymbol{\mu}||^2 + \frac{1}{2\sigma^2}\text{tr}(\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{z}_n\boldsymbol{z}_n^T]\boldsymbol{W}^T\boldsymbol{W}) - \frac{1}{\sigma^2}(\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{z}_n]^T\boldsymbol{W}^T)(\boldsymbol{x}_n - \boldsymbol{\mu}) + \frac{M}{2}\ln(2\pi) + \frac{1}{2}\text{tr}(\boldsymbol{z}_n\boldsymbol{z}_n^T)\Bigg)$$

Remember that all these expectations are taken wrt

$$p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{M}_{t-1}^{-1}\boldsymbol{W}_{t-1}^T(\boldsymbol{x} - \boldsymbol{\mu}_{t-1}), \sigma_{t-1}^2\boldsymbol{M}_{t-1}^{-1}) \quad \text{where } \boldsymbol{M}_{t-1} = \boldsymbol{W}_{t-1}^T\boldsymbol{W}_{t-}$$

So, in the E-step, all we need to calculate is:

$$\mathbb{E}_{\boldsymbol{Z}}[z_n] = \boldsymbol{M}_{t-1}^{-1}\boldsymbol{W}_{t-1}^T(\boldsymbol{x} - \boldsymbol{\mu}_{t-1})$$

$$\mathbb{E}_{\boldsymbol{Z}}[z_n z_n^T] = \sigma_{t-1}^2\boldsymbol{M}_{t-1}^{-1} + \mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{z}_n]\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{z}_n] \qquad \text{since } \text{cov}(z_n) = \mathbb{E}_{\boldsymbol{Z}}[z_n z_n^T] - \mathbb{E}_{\boldsymbol{Z}}[$$

EM for Probabilistic PCA - E-step (2)

Taking the expectation wrt z:

$$\mathbb{E}_{\boldsymbol{Z}}[ll_c] = -\sum_{n=1}^{N}\Bigg(\frac{D}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}||\boldsymbol{x}_n - \boldsymbol{\mu}||^2 + \frac{1}{2\sigma^2}\text{tr}(\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{z}_n\boldsymbol{z}_n^T]\boldsymbol{W}^T\boldsymbol{W}) - \frac{1}{\sigma^2}(\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{z}_n]^T\boldsymbol{W}^T)(\boldsymbol{x}_n - \boldsymbol{\mu}) + \frac{M}{2}\ln(2\pi) + \frac{1}{2}\text{tr}(\boldsymbol{z}_n\boldsymbol{z}_n^T)\Bigg)$$

Remember that all these expectations are taken wrt

$$p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{M}_{t-1}^{-1}\boldsymbol{W}_{t-1}^T(\boldsymbol{x} - \boldsymbol{\mu}_{t-1}), \sigma_{t-1}^2\boldsymbol{M}_{t-1}^{-1}) \quad \text{where } \boldsymbol{M}_{t-1} = \boldsymbol{W}_{t-1}^T\boldsymbol{W}_{t-}$$

So, in the E-step, all we need to calculate is:

$$\mathbb{E}_{\boldsymbol{Z}}[z_n] = \boldsymbol{M}_{t-1}^{-1}\boldsymbol{W}_{t-1}^T(\boldsymbol{x} - \boldsymbol{\mu}_{t-1})$$

$$\mathbb{E}_{\boldsymbol{Z}}[z_n z_n^T] = \sigma_{t-1}^2\boldsymbol{M}_{t-1}^{-1} + \mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{z}_n]\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{z}_n] \qquad \text{since } \text{cov}(z_n) = \mathbb{E}_{\boldsymbol{Z}}[z_n z_n^T] - \mathbb{E}_{\boldsymbol{Z}}[$$

For the M-step, we can now differentiate ll_c wrt $\boldsymbol{W}$ and σ^2 and set to zero.

Factor analysis

Factor analysis is similar to Probabilistic PCA, but with slightly more complex output covariance

$$p(x|z) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \mu, \boldsymbol{I}\boldsymbol{\sigma}^2)$$

Remember that in Probabilisic PCA we had simply:

$$p(x|z) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \mu, \boldsymbol{I}\sigma^2)$$

In Factor analysis, we thus estimate variances for all data dimensions separately.

Factor analysis

Factor analysis is similar to Probabilistic PCA, but with slightly more complex output covariance

$$p(x|z) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \mu, \boldsymbol{I}\boldsymbol{\sigma}^2)$$

Remember that in Probabilisic PCA we had simply:

$$p(x|z) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \mu, \boldsymbol{I}\sigma^2)$$

In Factor analysis, we thus estimate variances for all data dimensions separately.

With this small change, the model no longer has a closed-form maximum likelihood solution. But EM is still applicable.

Probabilistic PCA - Summary

Probabilistic PCA is a simple continuous latent variable model that brings PCA functionality into the rich framework of probabilistic modelling

Probabilistic PCA - Summary

Probabilistic PCA is a simple continuous latent variable model that brings PCA functionality into the rich framework of probabilistic modelling

Due to the particular choice of output variance, there is a closed-form maximum likelihood solution to the $\boldsymbol{W}, \boldsymbol{\mu}, \sigma$ parameters. We can also use EM, which can be more efficient in higher dimensions.

Probabilistic PCA - Summary

Probabilistic PCA is a simple continuous latent variable model that brings PCA functionality into the rich framework of probabilistic modelling

Due to the particular choice of output variance, there is a closed-form maximum likelihood solution to the $\boldsymbol{W}, \boldsymbol{\mu}, \sigma$ parameters. We can also use EM, which can be more efficient in higher dimensions.

In Factor Analysis, we change this assumption (now having separate variances for each dimension). This no longer has a closed form maximum likelihood solution.

Probabilistic PCA - Summary

Probabilistic PCA is a simple continuous latent variable model that brings PCA functionality into the rich framework of probabilistic modelling

Due to the particular choice of output variance, there is a closed-form maximum likelihood solution to the $\boldsymbol{W}, \boldsymbol{\mu}, \sigma$ parameters. We can also use EM, which can be more efficient in higher dimensions.

In Factor Analysis, we change this assumption (now having separate variances for each dimension). This no longer has a closed form maximum likelihood solution.

For both models, we should remember that the maximum likelihood only determines the latent space up to an arbitrary rotation.

Perspectives

The techniques covered today are important steps towards understanding the variational autoencoder.

- The considerations about bounds in the discussion of EM set the stage for variational inference.
- We will see that variational autoencoders arise as a natural generalization of probabilistic PCA.