Imperial College
London

NOTES

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

495 Advanced Statistical Machine Learning and Pattern Recognition

Author:
Thomas Teh (CID: 0124 3008)

Date: February 10, 2017

1 Expectation Maximization

1.1 General Approach for Expectation Maximization

1.1.1 Notes

1. The goal of the Expectation-Maximization is to find maximum likelihood solutions for models having latent variables
2. The general concept is that since our knowledge of the latent variables in $\boldsymbol{Z}$ is given by the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{Z},\boldsymbol{\theta})$, we use the expectation of the latent variables instead of the actual values.
3. EM algorithm can be used to find MAP solutions models in which a prior is defined over the parameters.

1.1.2 Algorithm

Given a joint distribution $p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$ over observed variables $\boldsymbol{X}$ and latent variables $\boldsymbol{Z}$, governed by parameter $\boldsymbol{\theta}$ the goal is to maximize the likelihood function $p(\boldsymbol{X}|\boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$.

1. Choose an initial for the parameters $\boldsymbol{\theta}^{old}$.
2. E Step: Evaluate $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old})$
3. M Step: Evaluate $\boldsymbol{\theta}^{new}$ given by

   $$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$$

   where

   $$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$$

4. Check for convergence of either the log-likelihood or the parameter values. If convergence is not satisfied

   $$\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$$

   and return to step 2

1.2 Gaussian Mixture Models

The Gaussian mixture distribution can be written as a linear superposition of Gaussians

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k,\Sigma_k)$$

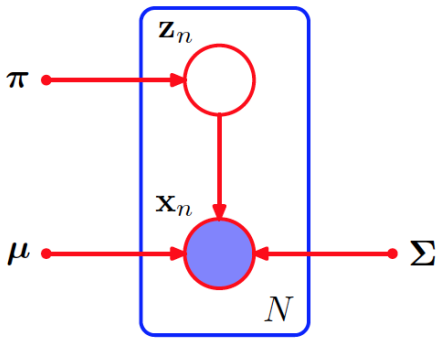

Figure 1: Gaussian Mixture Model.

1.2.1 Formulation of the Gaussian Mixture Models

Let $\boldsymbol{z}$ be a K-dimensional binary random variable with 1-of-K representation.

$$p(z_k = 1) = \pi_k \Rightarrow p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

Conditional probability of $\boldsymbol{x}$ given a particular value for latent variable $\boldsymbol{z}$:

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)^{z_k}$$

Using Bayes theorem and marginalize the latent variable $\boldsymbol{z}$

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

Similarly, the posterior probability of $\boldsymbol{z}$ is given by

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

1.2.2 Maximum Likelihood

The log-likelihood function is given by

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

The maximum likelihood method will yield the following:

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top}{\sum_{n=1}^{N} \gamma(z_{nk})}$$
$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}$$

Issues with maximum likelihood:

1. Presence of singularities: When we have at least two components in the mixture, one of them can have a finite variance and assign finite probability to all the data points, while the other component can shrink onto one specific data point and therefore contribute to an ever increasing additive value to the log likelihood.
2. Identifiability: Solutions may not be unique, hence it may be hard to interpret the parameter values discovered by a model.
3. The log likelihood equation is difficult to optimize over.

1.2.3 Expectation Maximization Formulation

Supposed that in addition to $\boldsymbol{X}$, we were also given the values of the latent variables $\boldsymbol{Z}$, the likelihood and the log likelihood function are given by

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

1. **Expectation Step**:

   Taking the expectation on the log likelihood

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})} [\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \Sigma, \boldsymbol{\pi})] &= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}[z_{nk}] \Big(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k) \Big) \\ &= G(\boldsymbol{\theta})\end{aligned}$$

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

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

2. **Maximization Step**: By taking the derivative of $(\boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$ and set them to 0, the parameters can be found to be

$$\begin{aligned}
\boldsymbol{\mu}_k &= \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \\
\Sigma_k &= \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top}{\sum_{n=1}^{N} \gamma(z_{nk})} \\
\pi_k &= \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}
\end{aligned}$$

1.2.4 Expectation Maximization Algorithm

Given a Gaussian mixture model, the goal is to maximize the likelihood functions w.r.t to the parameters:

1. Initialize the means $\boldsymbol{\mu}_k$, covariances Σ_k and mixing coefficients π_k and evaluate the initial of the log likelihood.

2. E Step: Evaluate the responsibilities using the current parameter values

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

3. M Step: Re-estimate the parameters using the current responsibilities

$$\begin{aligned}
\boldsymbol{\mu}_k^{new} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n \\
\Sigma_k^{new} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \\
\pi_k^{new} &= \frac{N_k}{N}
\end{aligned}$$

   where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

4. Evaluate the log likelihood

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

   and check for convergence of either parameter of the log likelihood. If the convergence criterion is not satisfied, return to step 2

1.3 Bernoulli Mixture Models

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\mu}) &= \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{1-x_i} \\
\mathbb{E}[\boldsymbol{x}] &= \boldsymbol{\mu} \\
\mathbb{V}[\boldsymbol{x}] &= diag\{\mu_i(1-\mu_i)\}
\end{aligned}
$$

The mixture of the Bernoulli distributions is given by

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\pi}) &= \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k) \\
p(\boldsymbol{x}|\boldsymbol{\mu}_k) &= \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki}^{x_i})^{1-x_i} \\
\mathbb{E}[\boldsymbol{x}] &= \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k \\
\mathbb{V}[\boldsymbol{x}] &= \sum_{k=1}^{K} \pi_k \left(\Sigma_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top\right) - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\top \\
\Sigma_k &= diag\{\mu_{ki}(1-\mu_{ki})\}
\end{aligned}
$$

Let $\boldsymbol{z}$ be the one hot representation, we have

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\mu}) &= \prod_{k=1}^{K} p(\boldsymbol{x}|\boldsymbol{\mu}_k)^{z_k} \\
p(\boldsymbol{z}|\boldsymbol{\pi}) &= \prod_{k=1}^{K} \pi_k^{z_k}
\end{aligned}
$$

The log likelihood function is given by

$$
\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\mu},\boldsymbol{\pi}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left\{\ln \pi_k + \sum_{i=1}^{D}[x_{ni}\ln \mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})]\right\}
$$

The E step: we then take the expectation of the log likelihood above

$$
\mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\mu},\boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})\left\{\ln \pi_k + \sum_{i=1}^{D}[x_{ni}\ln \mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})]\right\}
$$

$$
\gamma(z_{nk}) = \frac{\sum_{z_{nk}} z_{nk}[\pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k)]^{z_{nk}}}{\sum_{z_{nj}}[\pi_j p(\boldsymbol{x}_n|\boldsymbol{\mu}_j)]^{z_{nj}}} = \frac{\pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k)}{\sum_{j=1}^{K}\pi_j p(\boldsymbol{x})_{\boldsymbol{n}}|\boldsymbol{\mu}_j)}
$$

The M step: we then maximize the log likelihood wrt to the parameters

$$\begin{aligned}\boldsymbol{\mu}_k &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\\ \pi_k &= \frac{N_k}{N}\\ N_k &= \sum_{n=1}^{N}\gamma(z_{nk})\end{aligned}$$

1.4 Convergence of the EM Algorithm

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) \Rightarrow \ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + KL(q \parallel p)$$

$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln\left[\frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right]$$

$$KL(q \parallel p) = -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln\left[\frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right]$$

In the E step, the lower bound $\mathcal{L}(q,\boldsymbol{\theta}^{old})$ is maximized with respect to $q(\boldsymbol{Z})$. The solution to this maximization problem can be done by setting $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old})$ such that $KL(q \parallel p) = 0$.

In the M step, the lower bound $\mathcal{L}(q,\boldsymbol{\theta}^{old})$ is maximized with respect to $\boldsymbol{\theta}$. However, the KL divergence will no longer be zero. Hence, the increase in the upper bound of the log likelihood function is greater than the increase in the lower bound.

By substituting $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old})$, we get the lower bound to be the following after the E step:

$$\begin{aligned}\mathcal{L}(q,\boldsymbol{\theta}) &= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) - \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old})\\ &= \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) + const\end{aligned}$$

2 Probabilistic PCA

$$
\begin{aligned}
\boldsymbol{x} &= \boldsymbol{W}\boldsymbol{y} + \boldsymbol{\mu} + \boldsymbol{e} \\
\boldsymbol{e} &\sim \mathcal{N}(\boldsymbol{e}|\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \\
\boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{I})
\end{aligned}
$$

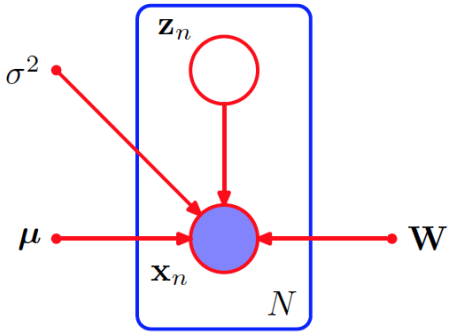

Figure 2: Gaussian Mixture Model.

2.1 Maximum Likelihood

1. We are interested in the distribution of $\boldsymbol{x}$ and the parameters $\boldsymbol{\theta}$. The marginal distribution of $\boldsymbol{x}$ can be obtained by integrating $\boldsymbol{y}$ out of the joint distribution.

$$
\begin{aligned}
p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N|\boldsymbol{\theta}) &= \int_{y_i} p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N|\boldsymbol{\theta}) d\boldsymbol{y}_1 \ldots d\boldsymbol{y}_N \\
&= \int_{y_i} \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma) p(\boldsymbol{y}_i) d\boldsymbol{y}_1 \ldots d\boldsymbol{y}_N \\
&= \prod_{i=1}^{N} \int_{y_i} p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma) p(\boldsymbol{y}_i) d\boldsymbol{y}_i
\end{aligned}
$$

2. The joint distribution can be obtained by applying Bayes' Rule.

$$
\begin{aligned}
p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N|\boldsymbol{\theta}) &= \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma) p(\boldsymbol{y}_i) \\
p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma) p(\boldsymbol{y}_i) &= \left[(2\pi\sigma^2)^{-\frac{F}{2}} e^{-\frac{1}{2\sigma^2}(\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{y}_i)^\top (\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{y}_i)} \right] \left[(2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\boldsymbol{y}_i^\top \boldsymbol{y}_i} \right]
\end{aligned}
$$

3. By collecting the $\boldsymbol{y}$ in the exponents, completing the squares and then applying the Woodbury identity (more details please refer to notes in 496):

$$
p(\boldsymbol{x}_i|\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) = \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{D})
$$

$$\boldsymbol{D} = \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}$$
$$p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) = \mathcal{N}(\boldsymbol{y}_i|\boldsymbol{M}^{-1}\boldsymbol{W}^\top(\boldsymbol{x}_i - \boldsymbol{\mu}), \sigma^2\boldsymbol{M}^{-1})$$
$$\boldsymbol{M} = \sigma^2\boldsymbol{I} + \boldsymbol{W}^\top\boldsymbol{W}$$

4. Maximizing the likelihood and solve for the parameters

$$\boldsymbol{S_t} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top \text{ ($\boldsymbol{S_t}$ is the covariance matrix)}$$
$$\sigma^2 = \frac{1}{F-d}\sum_{j=d+1}^{F}\lambda_j$$
$$\boldsymbol{W_d} = \boldsymbol{U_d}(\boldsymbol{\Lambda} - \sigma^2\boldsymbol{I})^{\frac{1}{2}}\boldsymbol{V}^\top$$

5. Hence we no longer have a projection but:

$$\mathbb{E}_{\boldsymbol{p}(\boldsymbol{y}_i|\boldsymbol{x}_i)}[\boldsymbol{y}_i] = \boldsymbol{M}^{-1}\boldsymbol{W}^\top(\boldsymbol{x}_i - \boldsymbol{\mu})$$
$$\widehat{\boldsymbol{x_i}} = \boldsymbol{W}\mathbb{E}_{\boldsymbol{p}(\boldsymbol{y}_i|\boldsymbol{x}_i)}[\boldsymbol{y}_i] + \boldsymbol{\mu}$$

2.2 EM PPCA

1. Write the log likelihood of the joint distribution of the observed and latent variables

$$p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{\theta})p(\boldsymbol{y}_i)$$
$$\ln p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta}) = \sum_{i=1}^{N}\Big(\ln p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{\theta}) + \ln p(\boldsymbol{y}_i)\Big)$$
$$\ln p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{\theta}) = -\frac{F}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{y}_i - \boldsymbol{\mu})^\top(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{y}_i - \boldsymbol{\mu})$$
$$\ln p(\boldsymbol{y}_i) = -\frac{1}{2}\boldsymbol{y}_i^\top\boldsymbol{y}_i - \frac{D}{2}\ln 2\pi$$

2. **E-Step:** Take the expectation on the log likelihood on the joint distribution:

$$\ln p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta}) = \sum_{i=1}^{N}\Big[-\frac{F}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{y}_i - \boldsymbol{\mu})^\top(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{y}_i - \boldsymbol{\mu}) - \frac{1}{2}\boldsymbol{y}_i^\top\boldsymbol{y}_i - \frac{D}{2}\ln 2\pi\Big]$$

Expanding the above and use the identities $\mathrm{tr}\Big[\boldsymbol{y}_i^\top\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{y}_i\Big] = \mathrm{tr}\Big[\boldsymbol{y}_i\boldsymbol{y}_i^\top\boldsymbol{W}\boldsymbol{W}^\top\Big]$

$$\mathbb{E}_{p(\boldsymbol{Y}|\boldsymbol{X})}[\ln p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta})] = -\frac{NF}{2}\ln 2\pi\sigma^2 - \frac{ND}{2}\ln 2\pi$$
$$-\sum_{i=1}^{N}\Big\{\frac{1}{2\sigma^2}\Big[(\boldsymbol{x}_i - \boldsymbol{\mu})^\top(\boldsymbol{x}_i - \boldsymbol{\mu}) - 2(\boldsymbol{x}_i - \boldsymbol{\mu})^\top\boldsymbol{W}\mathbb{E}[\boldsymbol{y}_i]$$

$$+ \operatorname{tr}\left[\mathbb{E}(\boldsymbol{y}_i \boldsymbol{y}_i^\top)\boldsymbol{W}^\top \boldsymbol{W}\right]\Big] + \frac{1}{2}\operatorname{tr}\left[\mathbb{E}(\boldsymbol{y}_i^\top \boldsymbol{y}_i)\right]\Big\}$$

We can obtain the moments of $\boldsymbol{y}_i$ from the earlier derivation of the PPCA

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) = \mathcal{N}(\boldsymbol{y}_i|\boldsymbol{M}^{-1}\boldsymbol{W}^\top(\boldsymbol{x}_i - \boldsymbol{\mu}), \sigma^2\boldsymbol{M}^{-1})$$
$$\boldsymbol{M} = \sigma^2\boldsymbol{I} + \boldsymbol{W}^\top\boldsymbol{W}$$
$$\mathbb{E}_{p(\boldsymbol{Y}|\boldsymbol{X})}[\boldsymbol{y}_i] = \boldsymbol{M}^{-1}\boldsymbol{W}^\top(\boldsymbol{x}_i - \boldsymbol{\mu})$$
$$\mathbb{E}_{p(\boldsymbol{Y}|\boldsymbol{X})}[\boldsymbol{y}_i\boldsymbol{y}_i^\top] = \sigma^2\boldsymbol{M}^{-1} + \mathbb{E}[\boldsymbol{y}_i]\mathbb{E}[\boldsymbol{y}_i]^\top$$

3. **M-Step:** Maximize the log likelihood

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{W}\mathbb{E}[\boldsymbol{y}_i])$$
$$\boldsymbol{W} = \left[\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})\mathbb{E}[\boldsymbol{y}_i]^\top\right]\left[\sum_{i=1}^{N}\mathbb{E}[\boldsymbol{y}_i\boldsymbol{y}_i^\top]\right]^{-1}$$
$$\sigma^2 = \frac{1}{NF}\sum_{i=1}^{N}\left\{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2 - 2\mathbb{E}[\boldsymbol{y}_i]^\top\boldsymbol{W}^\top(\boldsymbol{x}_i - \boldsymbol{\mu}) + \operatorname{tr}(\mathbb{E}[\boldsymbol{y}_i\boldsymbol{y}_i^\top]\boldsymbol{W}^\top\boldsymbol{W})\right\}$$

4. Advantages of EM PPCA:
   - A complexity of $O(NFD)$ can be significantly smaller than $O(NF^2)$ (from the computation of the covariance)
   - The EM procedure can be extended to factor analysis model
   - EM allows us to deal with missing values

3 PPCA Mixture Model

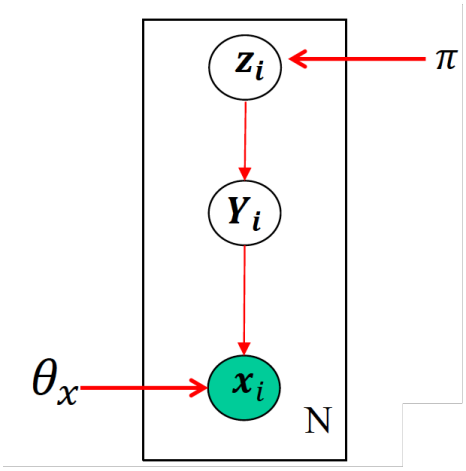

Figure 3: PPCA Mixture Model.

The model is given by:

$$\begin{aligned}
\boldsymbol{x}_i &= \boldsymbol{\mu}_k + \boldsymbol{W}_k \boldsymbol{y}_{ik} + \boldsymbol{e}_{ik} \\
\boldsymbol{y}_{ik} &\sim \mathcal{N}(\boldsymbol{y}_{ik}|\boldsymbol{0}, \boldsymbol{I}) \\
\boldsymbol{e}_{ik} &\sim \mathcal{N}(\boldsymbol{e}_{ik}|\boldsymbol{0}, \sigma_k^2 \boldsymbol{I}), \forall k = 1, \ldots, K \\
p(\boldsymbol{x}_i|\boldsymbol{z}_{ik} = 1, \boldsymbol{y}_{ik}, \boldsymbol{\mu_k}, \boldsymbol{W_k}, \sigma_k^2 \boldsymbol{I}) &= \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k + \boldsymbol{W}_k \boldsymbol{y}_{ik}) \\
p(\boldsymbol{x}_i|\boldsymbol{\theta}) &= \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{D}_k) \\
\boldsymbol{D}_k &= \boldsymbol{W}_k \boldsymbol{W}_k^\top + \sigma_k^2 \boldsymbol{I}
\end{aligned}$$

3.1 Probability Distributions for EM Formulation

1. Priors distributions:

$$\begin{aligned}
p(\boldsymbol{z}_i|\boldsymbol{\theta}_z) &= \prod_{k=1}^{K} \pi_k^{z_{ik}} \\
p(\boldsymbol{Y}_i|\boldsymbol{z}_i, \boldsymbol{\theta}_z) &= \prod_{k=1}^{K} p(y_{ik})^{z_{ik}} = \prod_{k=1}^{K} \mathcal{N}(y_{ik}|\boldsymbol{0}, \boldsymbol{I})^{z_{ik}}
\end{aligned}$$

2. Conditional distribution:

$$\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{Y}_i, \boldsymbol{\theta}_{\boldsymbol{x}}) &= \prod_{k=1}^{K} p(\boldsymbol{x}_i|\boldsymbol{z}_{ik} = 1, \boldsymbol{y}_{ik}, \boldsymbol{W}_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2 \boldsymbol{I})^{z_{ik}} \\
&= \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_{ik}|\boldsymbol{W}_k \boldsymbol{y}_{ik} + \boldsymbol{\mu}_k, \sigma_k \boldsymbol{I})^{z_{ik}}
\end{aligned}$$

3. Marginal distributions per cluster is found by integrating $\boldsymbol{y}$ out:

$$\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{z}_{ik} = 1, \boldsymbol{W}_k, \boldsymbol{\mu}_k, \sigma_k^2) &= \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{D}_k) \\
\boldsymbol{D}_k &= \boldsymbol{W}_k \boldsymbol{W}_k^\top + \sigma_k^2 \boldsymbol{I}
\end{aligned}$$

4. Full marginal distributions:

$$\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{\theta}) &= \sum_{k=1}^{K} p(z_{ik} = 1) p(\boldsymbol{x}|z_{ik} = 1, \theta_x) \\
&= \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{D}_k)
\end{aligned}$$

5. Posteriors on $\boldsymbol{y}_{ik}$ (we will take expectation on this in the E step):

$$p(\boldsymbol{y}_{ik}|\boldsymbol{x}_{ik}, z_{ik} = 1, \boldsymbol{W}_k, \boldsymbol{\mu}_k, \sigma_k^2) = \mathcal{N}\left(\boldsymbol{y}_{ik}|\boldsymbol{M}_k^{-1} \boldsymbol{W}_k^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_k), \sigma_k^2 \boldsymbol{M}_k^{-1}\right)$$

$$\mathbb{E}(\boldsymbol{y}_{ik}) = \boldsymbol{M}_k^{-1}\boldsymbol{W}_k^\top(\boldsymbol{x}_i - \boldsymbol{\mu}_k)$$
$$\mathbb{E}(\boldsymbol{y}_{ik}\boldsymbol{y}_{ik}^\top) = \sigma_k^2\boldsymbol{M}_k^{-1} + \mathbb{E}(\boldsymbol{y}_{ik})\mathbb{E}(\boldsymbol{y}_{ik})^\top$$

6. Posteriors on $\boldsymbol{z}_i$ (we will take expectation on this in the E step):

$$p(\boldsymbol{z}_i|\boldsymbol{x}_i,\theta) = \frac{p(\boldsymbol{x}_i,\boldsymbol{z}_i|\theta)}{p(\boldsymbol{x}_i|\theta)} = \frac{p(\boldsymbol{x}_i|\boldsymbol{z}_i,\theta)p(\boldsymbol{z}_i|\theta)}{p(\boldsymbol{x}_i|\theta)} = \frac{\prod_k^K\left(\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{D}_k)\pi_k\right)^{z_{nk}}}{\sum_{l=1}^K \pi_l\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{D}_l,\boldsymbol{\mu}_l)}$$
$$\mathbb{E}[z_{ik}] = \sum_{z_{ik}=0,1} z_{ik}p(\boldsymbol{z_i}|\boldsymbol{x}_i,\theta) = \frac{\pi_k\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{D}_k)}{\sum_{l=1}^K \pi_l\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l,\boldsymbol{D}_l)}$$

3.2 Formulation of EM

1. Setting the joint likelihood (decomposition below can be guided by graphical model):

$$p(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}|\theta) = p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{Y},\theta_x)p(\boldsymbol{Z},\boldsymbol{Y}|\theta_Z) = p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{Y},\theta_x)p(\boldsymbol{Y}|\boldsymbol{Z})p(\boldsymbol{Z}|\theta_Z)$$
$$p(\boldsymbol{Z}|\theta_Z) = \prod_{i=1}^N\prod_{k=1}^K \pi_k^{z_{ik}}$$
$$p(\boldsymbol{Y}|\boldsymbol{Z}) = \prod_{i=1}^N p(\boldsymbol{Y}_i|\boldsymbol{z}_i,\theta_Z) = \prod_{i=1}^N\prod_{k=1}^3 p(\boldsymbol{y}_{ik})^{z_{ik}} = \prod_{i=1}^N\prod_{k=1}^3 \mathcal{N}(\boldsymbol{y}_{ik}|\boldsymbol{0},\boldsymbol{I})^{z_{ik}}$$
$$p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{Y},\theta_x) = \prod_{i=1}^N p(\boldsymbol{x}_i\boldsymbol{z}_i,\boldsymbol{Y}_i,\theta_x = \prod_{i=1}^N\prod_{k=1}^K \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}_k\boldsymbol{y}_{ik}+\boldsymbol{\mu}_k,\sigma_k^2)^{z_{ik}}$$

Hence the log likelihood of the joint distribution:

$$\begin{aligned}\ln p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{Y}|\theta) &= \sum_{i=1}^N\sum_k^K z_{ik}\left[\ln\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}_k\boldsymbol{y}_{ik}+\boldsymbol{\mu}_k,\sigma_k^2) + \ln\mathcal{N}(\boldsymbol{y}_{ik}|\boldsymbol{0},\boldsymbol{I}) + \ln\pi_k\right]\\ &= \sum_{i=1}^N\sum_k^K z_{ik}\left[-\frac{1}{2\sigma_k^2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k-\boldsymbol{W}_k\boldsymbol{y}_k)^\top(\boldsymbol{x}_i-\boldsymbol{\mu}_k-\boldsymbol{W}_k\boldsymbol{y}_k) - \frac{F}{2}\ln 2\pi - F\ln\sigma_k\right]\\ &+ \sum_{i=1}^N\sum_k^K z_{ik}\left[-\frac{1}{2}\boldsymbol{y}_{ik}^\top\boldsymbol{y}_{ik} - \frac{d}{2}\ln 2\pi\right] + \sum_{i=1}^N\sum_k^K z_{ik}\ln\pi_k\end{aligned}$$

2. Taking the expectation yields

$$\begin{aligned}&\mathbb{E}[\ln p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{Y}|\theta)]\\ &= -\frac{1}{2\sigma_k^2}\sum_{i=1}^N\sum_k^K \mathbb{E}[z_{ik}]\left[\|\boldsymbol{x}_i-\boldsymbol{\mu}_k\|^2 - \mathbb{E}[\boldsymbol{y}_{ik}]^\top\boldsymbol{W}_k^\top(\boldsymbol{x}_i-\boldsymbol{\mu}_k) + \mathrm{tr}(\boldsymbol{W}\boldsymbol{W}^\top\mathbb{E}[\boldsymbol{y}_{ik}\boldsymbol{y}_{ik}^\top])\right]\\ &+ \sum_{i=1}^N\sum_k^K\left[\frac{F}{2}\ln 2\pi - F\ln\sigma_k\right] + \sum_{i=1}^N\sum_k^K \mathbb{E}[z_{ik}]\left[-\frac{1}{2}\mathrm{tr}(\mathbb{E}[\boldsymbol{y}_{ik}\boldsymbol{y}_{ik}^\top]) - \frac{d}{2}\ln 2\pi\right] + \sum_{i=1}^N\sum_k^K \mathbb{E}[z_{ik}]\ln\pi_k\end{aligned}$$

3. The maximization step:

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[z_{ik}]$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N}\mathbb{E}[z_{ik}](\boldsymbol{x}_i - \boldsymbol{W}_k\mathbb{E}[\boldsymbol{y}_{ik}])}{\sum_{i=1}^{N}\mathbb{E}[z_{ik}]}$$

$$\boldsymbol{W}_k = \left[\sum_{i=1}^{N}\mathbb{E}[z_{ik}](\boldsymbol{x}_i - \boldsymbol{\mu})\mathbb{E}[\boldsymbol{y}_{ik}^\top]\right]\left[\sum_{i=1}^{N}\mathbb{E}[z_{ik}]\mathbb{E}[\boldsymbol{y_{ik}}\boldsymbol{y}_{ik}^\top])\right]^{-1}$$

$$\sigma_k^2 = \frac{1}{F\sum_{i=1}^{N}\mathbb{E}[z_{ik}]}\sum_{i=1}^{N}\mathbb{E}[z_{ik}]\left[\| \boldsymbol{x}_i - \boldsymbol{\mu}_k \|^2 - \mathbb{E}[\boldsymbol{y}_{ik}]^\top \boldsymbol{W}_k^\top(\boldsymbol{x}_i - \boldsymbol{\mu}_k) + \mathrm{tr}\left(\boldsymbol{W}_k^\top \boldsymbol{W}_k\mathbb{E}[\boldsymbol{y}_{ik}\boldsymbol{y}_{ik}^\top]\right)\right]$$

4 PPCA with Missing Values

Given the PPCA model:

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{W}\boldsymbol{y} + \boldsymbol{\mu} + \boldsymbol{e} \\ \boldsymbol{e} &\sim \mathcal{N}(\boldsymbol{e}|\boldsymbol{0}, \sigma^2\boldsymbol{I}) \\ \boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{I}) \end{aligned}$$

$$\begin{aligned} p(\boldsymbol{x}_i|\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) &= \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{D}) \\ p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) &= \mathcal{N}(\boldsymbol{y}_i|\boldsymbol{M}^{-1}\boldsymbol{W}^\top(\boldsymbol{x}_i - \boldsymbol{\mu}), \sigma^2\boldsymbol{M}^{-1}) \\ \boldsymbol{D} &= \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I} \\ \boldsymbol{M} &= \boldsymbol{W}^\top\boldsymbol{W} + \sigma^2\boldsymbol{I} \end{aligned}$$

1. We can reformulate the PPCA model to take into account of missing data:

$$\begin{aligned} \boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}^o \\ \boldsymbol{x}^u \end{bmatrix} \Rightarrow p(\boldsymbol{x}^o, \boldsymbol{x}^u) &= \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}^o \\ \boldsymbol{x}^u \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu}^o \\ \boldsymbol{\mu}^u \end{bmatrix}, \begin{bmatrix} \boldsymbol{D}^{oo} & \boldsymbol{D}^{ou} \\ \boldsymbol{D}^{uo} & \boldsymbol{D}^{uu} \end{bmatrix} \right) \\ p(\boldsymbol{x}^u|\boldsymbol{x}^o) &= \mathcal{N}\left(\boldsymbol{x}^u \middle| \boldsymbol{\mu}^u + \boldsymbol{D}_{uo}\boldsymbol{D}_{oo}^{-1}(\boldsymbol{x}^o - \boldsymbol{\mu}^o), \boldsymbol{D}_{uu} - \boldsymbol{D}_{uo}\boldsymbol{D}_{oo}^{-1}\boldsymbol{D}_{ou} \right) \end{aligned}$$

2. For convenience, we can rewrite this formulation as derive its first order and second order moments:

$$\begin{aligned} p(\boldsymbol{x}_i|\boldsymbol{x}_i^o) &= \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{Q}) \\ \boldsymbol{z}_i &= \begin{bmatrix} \boldsymbol{x}_i^o \\ \boldsymbol{\mu}^u + \boldsymbol{D}_{uo}\boldsymbol{D}_{oo}^{-1}(\boldsymbol{x}^o - \boldsymbol{\mu}^o) \end{bmatrix} \\ \boldsymbol{Q} &= \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{D}_{uu} - \boldsymbol{D}_{uo}\boldsymbol{D}_{oo}^{-1}\boldsymbol{D}_{ou} \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \mathbb{E}_{p(\boldsymbol{x}_i|\boldsymbol{x}_i^o)}[\boldsymbol{x}_i] &= \boldsymbol{z}_i \\ \mathbb{E}_{p(\boldsymbol{x}_i|\boldsymbol{x}_i^o)}[\boldsymbol{x}_i\boldsymbol{x}_i^\top] &= \boldsymbol{Q} + \boldsymbol{z}_i\boldsymbol{z}_i^\top \\ \mathbb{E}_{p(\boldsymbol{x}_i|\boldsymbol{x}_i^o)}[(\boldsymbol{x}_i - \boldsymbol{a})(\boldsymbol{x}_i - \boldsymbol{a})^\top] &= \boldsymbol{Q} + (\boldsymbol{z}_i - \boldsymbol{a})(\boldsymbol{z}_i - \boldsymbol{a})^\top \end{aligned}$$

3. Writing down the log likelihood for the observed variable X and the latent variable Y

$$\begin{aligned} \ln p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta}) = &-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left\{\mathrm{tr}\left[(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top\right] - 2\mathrm{tr}\left[\boldsymbol{y}_i(\boldsymbol{x}_i - \boldsymbol{\mu})^\top\boldsymbol{W}\right] + \mathrm{tr}\left[\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{y}_i\boldsymbol{y}_i^\top\right]\right\} \\ &- \frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}[\boldsymbol{y}_i\boldsymbol{y}_i^\top] - \frac{NF}{2}\ln 2\pi - NF\ln\sigma - N\ln 2\pi \end{aligned}$$

4. Take the expectation with regards to the probability distribution $p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{X}^o)$

$$
\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{x}_i,\boldsymbol{y}_i|\boldsymbol{x}_i^o)}\left[\ln p(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta})\right] \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left\{\mathrm{tr}\left(\mathbb{E}\left[(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\right]\right) - 2\mathrm{tr}\left(\mathbb{E}\left[\boldsymbol{y}_i(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\right]\boldsymbol{W}\right) + \mathrm{tr}\left(\boldsymbol{W}^\top\boldsymbol{W}\mathbb{E}\left[\boldsymbol{y}_i\boldsymbol{y}_i^\top\right]\right)\right\} \\
&- \frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}\left(\mathbb{E}\left[\boldsymbol{y}_i\boldsymbol{y}_i^\top\right]\right) - \frac{NF}{2}\ln 2\pi - NF\ln\sigma - N\ln 2\pi
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_{p(\boldsymbol{x}_i,\boldsymbol{y}_i|\boldsymbol{x}_i^o)}[(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top] &= \mathbb{E}_{p(\boldsymbol{x}_i,\boldsymbol{y}_i|\boldsymbol{x}_i^o)}[(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top] \\
&= \boldsymbol{Q} + (\boldsymbol{z}_i-\boldsymbol{\mu})(\boldsymbol{z}_i-\boldsymbol{\mu})^\top \\
\mathbb{E}_{p(\boldsymbol{x}_i,\boldsymbol{y}_i|\boldsymbol{x}_i^o)}[\boldsymbol{y}_i(\boldsymbol{x}_i-\boldsymbol{\mu})^\top] &= \mathbb{E}_{p(\boldsymbol{x}_i|\boldsymbol{x}_i^o)}\left[\mathbb{E}_{p(\boldsymbol{y}_i|\boldsymbol{x}_i^o)}\left[\boldsymbol{y}_i(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\right]\right] \\
&= \mathbb{E}_{p(\boldsymbol{x}_i|\boldsymbol{x}_i^o)}\left[\boldsymbol{M}^{-1}\boldsymbol{W}^\top(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\right] \\
&= \boldsymbol{M}^{-1}\boldsymbol{W}^\top\mathbb{E}_{p(\boldsymbol{x}_i|\boldsymbol{x}_i^o)}\left[(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\right] \\
\mathbb{E}_{p(\boldsymbol{x}_i,\boldsymbol{y}_i|\boldsymbol{x}_i^o)}[\boldsymbol{y}_i\boldsymbol{y}_i^\top] &= \mathbb{E}_{p(\boldsymbol{x}_i|\boldsymbol{x}_i^o)}\left[\mathbb{E}_{p(\boldsymbol{y}_i|\boldsymbol{x}_i^o)}\left[\boldsymbol{y}_i\boldsymbol{y}_i^\top\right]\right] \\
&= \mathbb{E}_{p(\boldsymbol{x}_i|\boldsymbol{x}_i^o)}\left[\sigma^2\boldsymbol{M}^{-1} + \boldsymbol{M}^{-1}\boldsymbol{W}^\top(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\boldsymbol{W}\boldsymbol{M}^{-1}\right] \\
&= \sigma^2\boldsymbol{M}^{-1} + \boldsymbol{M}^{-1}\boldsymbol{W}^\top\mathbb{E}_{p(\boldsymbol{x}_i|\boldsymbol{x}_i^o)}\left[(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\right]\boldsymbol{W}\boldsymbol{M}^{-1}
\end{aligned}
$$

5. Maximization step:

$$
\begin{aligned}
\boldsymbol{\mu} &= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{z}_i \\
\boldsymbol{W} &= \left[\sum_{i=1}^{N}\mathbb{E}\left[(\boldsymbol{x}_i-\boldsymbol{\mu})\boldsymbol{y}_i^\top\right]\right]\left[\sum_{i=1}^{N}\mathbb{E}\left[\boldsymbol{y}_i\boldsymbol{y}_i^\top\right]\right]^{-1} \\
\sigma^2 &= \frac{1}{NF}\sum_{i=1}^{N}\left\{\mathrm{tr}\left(\mathbb{E}\left[(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\right]\right) - 2\mathrm{tr}\left(\mathbb{E}\left[\boldsymbol{y}_i(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\right]\boldsymbol{W}\right) + \mathrm{tr}\left(\mathbb{E}\left[\boldsymbol{y}_i\boldsymbol{y}_i^\top\right]\boldsymbol{W}^\top\boldsymbol{W}\right)\right\} \\
\boldsymbol{D} &= \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}
\end{aligned}
$$

5 Hidden Markov Model

5.1 Preliminary: Markov Chains

1. Markov Property:

$$p(\boldsymbol{x_i}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{i-1}) = p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})$$

2. Markov Model: Given the observations $\{\boldsymbol{x}_t\}_{t=1}^N$

$$p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = p(\boldsymbol{x}_1)\prod_{t=2}^{N} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \qquad \text{bigram model}$$

$$p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = p(\boldsymbol{x}_1)p(\boldsymbol{x}_2|\boldsymbol{x}_1)\prod_{t=3}^{N} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_{t-2}) \qquad \text{trigram model}$$

$$p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = p(\boldsymbol{x}_1)p(\boldsymbol{x}_2|\boldsymbol{x}_1)p(\boldsymbol{x}_3|\boldsymbol{x}_1,\boldsymbol{x}_2)\prod_{t=4}^{N} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_{t-2},\boldsymbol{x}_{t-3}) \qquad \text{n-gram model}$$

3. A transition matrix $\boldsymbol{A}$ specifies the probabilities of getting from state i (*row*) to state j (*column*) in one step. Note that $\boldsymbol{A}$ is a stochastic matrix, i.e. the row must sum to 1.

4. Stationary distribution is the long term distribution over the states

$$\boldsymbol{\pi}^\top = \boldsymbol{\pi}^\top \boldsymbol{A}$$

   Solving for the stationary distribution is the same as eigenanalysis, where $\boldsymbol{\pi}$ is an eigenvector with eigenvalue 1.

$$\begin{aligned}\boldsymbol{\pi}^\top(\boldsymbol{I}-\boldsymbol{A}) &= \boldsymbol{0}\\ \boldsymbol{\pi}^\top\boldsymbol{1} &= \boldsymbol{1}\end{aligned}$$

5. For a stationary distribution to exist, the markov chain has to be irreducible and aperiodic.

   - Irreducibility: The state transition diagram must be singly connected, i.e. it is possible to move from one state to another ($a_{ij}(t) > 0$)
   - Aperiodicty: $d(i) = gcd\{t : a_{ii}(t) > 0\} = 1, \forall i$. At some point in time, the probability of a recurrent connection is greater than 0.

6. **Detailed balance condition**: A markov process is called a reversible Markov process if it satisfies the detailed balance equations.

$$\pi_i P_{ij} = \pi_j P_{ji}$$

7. Estimation of the transition matrix from training data (language model)

   (a) The probability of a character is given by

$$p(\boldsymbol{x}_1|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{x_{1k}}$$

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \prod_{j=1}^{K}\prod_{k=1}^{K} a_{jk}^{x_{(t-1)j}x_{tk}}$$

   (b) Probability of a sequence D_l with length T

$$P(D_l|\theta) = p(\boldsymbol{x}_1^l,\ldots,\boldsymbol{x}_T^l) = p(\boldsymbol{x}_1^l)\prod_{t=2}^{T} p(\boldsymbol{x}_t^l|\boldsymbol{x}_{t-1}^l) = \prod_{k=1}^{K} \pi_k^{x_{1k}^l}\prod_{t=2}^{T}\prod_{j=1}^{K}\prod_{k=1}^{K} a_{jk}^{x_{(t-1)j}^l x_{tk}^l}$$

   (c) The likelihood function is then

$$p(D_1,\ldots,D_N|\theta) = \prod_{l=1}^{N} p(D_l|\theta) = \prod_{l=1}^{N}\left\{\prod_{k=1}^{K} \pi_k^{x_{1k}^l}\prod_{t=2}^{T}\prod_{j=1}^{K}\prod_{k=1}^{K} a_{jk}^{x_{(t-1)j}^l x_{tk}^l}\right\}$$

$$\begin{aligned}
\ln p(D_1,\ldots,D_N|\theta) &= \sum_{l=1}^{N}\sum_{k=1}^{K} x_{1k}^l \ln \pi_k + \sum_{l=1}^{N}\sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} x_{(t-1)j}^l x_{tk}^l \ln a_{jk} \\
&= \sum_{k=1}^{K}\left(\sum_{l=1}^{N} x_{1k}^l\right)\ln \pi_k + \sum_{j=1}^{K}\sum_{k=1}^{K}\left(\sum_{l=1}^{N}\sum_{t=2}^{T} x_{(t-1)j}^l x_{tk}^l\right)\ln a_{jk} \\
&= \sum_{k=1}^{K} N_k^1 \ln \pi_k + \sum_{j=1}^{K}\sum_{k=1}^{K} N_{jk} \ln a_{jk}
\end{aligned}$$

   where

$$N_k^1 = \sum_{l=1}^{N} x_{1k}^l \qquad\qquad N_{jk} = \sum_{l=1}^{N}\sum_{t=2}^{T} x_{(t-1)j}^l x_{tk}^l$$

   (d) We just need to solve the optimization below by formulating the Lagrangian:

$$\begin{aligned}
&\max_{\theta} \sum_{k=1}^{K} N_k^1 \ln \pi_k + \sum_{j=1}^{K}\sum_{k=1}^{K} N_{jk} \ln a_{jk} \\
&\text{s.t.} \sum_{k=1}^{K} \pi_k = 1
\end{aligned}$$

$$\sum_{k=1}^{K} a_{jk} = 1$$

$$\pi_k = \frac{N_k^1}{\sum_{k=1}^{K} N_k^1} \qquad a_{jk} = \frac{N_{jk}}{\sum_{k=1}^{K} N_{jk}}$$

5.2 Hidden Markov Model

5.2.1 Model Description and Important Properties

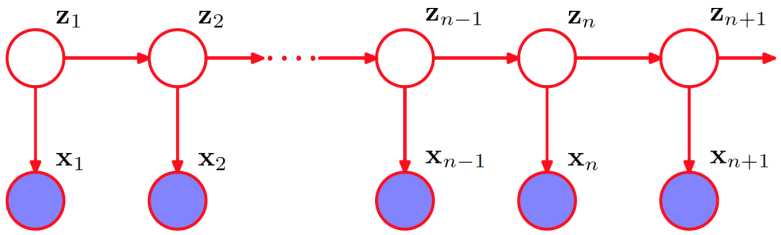

Figure 4: State Space Model. If z are discrete and markovian, then it is a Hidden Markov model.

1. The HMM is a specific instance of the state space model in Figure 4 in which the latent variables are discrete.
2. If we view a single time slice of the model, it corresponds to a mixture distribution densities given by $p(\boldsymbol{x}|\boldsymbol{z})$, where $p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})$.
3. Properties of conditional independence

$$\begin{aligned}
p(\boldsymbol{X}|\boldsymbol{z}_n) &= p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n|\boldsymbol{z}_n)p(\boldsymbol{x}_{n+1},\ldots,\boldsymbol{x}_N|\boldsymbol{z}_n) \\
p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n-1}|\boldsymbol{x}_n,\boldsymbol{z}_n) &= p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n-1}|\boldsymbol{z}_n) \\
p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n-1}|\boldsymbol{z}_{n-1},\boldsymbol{z}_n) &= p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n-1}|\boldsymbol{z}_{n-1}) \\
p(\boldsymbol{x}_{n+1},\ldots,\boldsymbol{x}_N|\boldsymbol{z}_n,\boldsymbol{z}_{n+1}) &= p(\boldsymbol{x}_{n+1},\ldots,\boldsymbol{x}_N|\boldsymbol{z}_{n+1}) \\
p(\boldsymbol{x}_{n+2},\ldots,\boldsymbol{x}_N|\boldsymbol{z}_{n+1},\boldsymbol{x}_{n+1}) &= p(\boldsymbol{x}_{n+2},\ldots,\boldsymbol{x}_N|\boldsymbol{z}_{n+1}) \\
p(\boldsymbol{X}|\boldsymbol{z}_{n-1},\boldsymbol{z}_n) &= p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n-1}|\boldsymbol{z}_{n-1})p(\boldsymbol{x}_n|\boldsymbol{z}_n)p(\boldsymbol{x}_{n+1},\ldots,\boldsymbol{x}_N|\boldsymbol{z}_n) \\
p(\boldsymbol{x}_{N+1}|\boldsymbol{X},\boldsymbol{z}_{N+1}) &= p(\boldsymbol{x}_{N+1}|\boldsymbol{z}_{N+1}) \\
p(\boldsymbol{z}_{N+1}|\boldsymbol{z}_N,\boldsymbol{X}) &= p(\boldsymbol{z}_{N+1}|\boldsymbol{z}_N)
\end{aligned}$$

4. Inferences using the Hidden Markov Model
   - **Filtering**: Compute the belief state $p(\boldsymbol{z}_n|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ online as the data streams in. It is known as filtering to reduce the noise in the estimation of the hidden state.

- **Smoothing**: Compute $p(z_t|x_1,\dots,x_N)$ offline, given all the evidence. Reducing the uncertainty by conditioning on the past and future data.
- **Fixed lag smoothing**: Compute $p(z_{t-\ell}|x_1,\dots,x_n)$, where $\ell > 0$ is the lag.
- **MAP estimation**: Compute $\arg\max_{z_1,\dots,z_n} p(z_1,\dots,z_n|\boldsymbol{x_1},\dots,\boldsymbol{x_n})$. This is known as Viterbi Decoding.
- **Evaluation**: Find the probability of the evidence $p(x_1,\dots,x_N) = \sum_{\boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z})$.
- **Prediction**: Predicting the future given the past, we compute $p(z_{n+\delta}|x_1,\dots,x_n)$ and $p(x_{n+\delta}|x_1,\dots,x_n)$.
- **Parameter Estimation**: This is know as the Baum-Welch algoritm. We estimate the parameter $\boldsymbol{A}$, $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$.

5.2.2 Filtering and Smoothing: Forwards and Backwards Algorithm

1. We can write the posterior probabilities of the latent variables as below. After that, we will formulate the computation for $\alpha(z_n)$ (forward probabilities) and $\beta(z_n)$ (backward probabilities).

$$\begin{aligned}\gamma(z_n) = p(z_n|\boldsymbol{X}) &= \frac{p(\boldsymbol{X}|z_n)p(z_n)}{p(\boldsymbol{X})} \\ &= \frac{p(x_1,\dots,x_n|z_n)p(x_{n+1},\dots,x_N|z_n)p(z_n)}{p(\boldsymbol{X})} \\ &= \frac{p(x_1,\dots,x_n,z_n)p(x_{n+1},\dots,x_N|z_n)}{p(\boldsymbol{X})} \\ &= \frac{\alpha(z_n)\beta(z_n)}{p(\boldsymbol{X})}\end{aligned}$$

2. The forward probabilities can be computed recursively based on the below:

$$\begin{aligned}\alpha(z_n) &= p(x_1,\dots,x_n,z_n) \\ &= p(x_1,\dots,x_n|z_n)p(z_n) \\ &= p(x_n|z_n)p(x_1,\dots,x_{n-1}|z_n)p(z_n) \\ &= p(x_n|z_n)p(x_1,\dots,x_{n-1},z_n) \\ &= p(x_n|z_n)\sum_{z_{n-1}} p(x_1,\dots,x_{n-1},z_n|z_{n-1})p(z_{n-1}) \\ &= p(x_n|z_n)\sum_{z_{n-1}} p(x_1,\dots,x_{n-1}|z_{n-1})p(z_n|z_{n-1})p(z_{n-1}) \\ &= p(x_n|z_n)\sum_{z_{n-1}} p(x_1,\dots,x_{n-1},z_{n-1})p(z_n|z_{n-1}) \\ &= p(x_n|z_n)\sum_{z_{n-1}} \alpha(z_{n-1})p(z_n|z_{n-1})\end{aligned}$$

The initial condition of the recursive formula is given by:

$$\alpha(\boldsymbol{z}_1) = p(\boldsymbol{x}_1, \boldsymbol{z}_1) = p(\boldsymbol{z}_1)p(\boldsymbol{x}_1|\boldsymbol{z}_1) = \prod_{k=1}^{K} \{\pi_k p(\boldsymbol{x}_1|z_{1k})\}^{z_{1k}}$$

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_n) = \sum_{\boldsymbol{z}_n} \alpha(\boldsymbol{z}_n)$$

$$p(\boldsymbol{z}_n|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \frac{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_n)}{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)} = \frac{\alpha(\boldsymbol{z}_n)}{\sum_{\boldsymbol{z}_n} \alpha(\boldsymbol{z}_n)} = \tilde{\alpha}(\boldsymbol{z}_n)$$

3. The backward probabilities can be computed as follows:

$$\begin{aligned}
\beta(\boldsymbol{z}_n) &= p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N|\boldsymbol{z}_n) \\
&= \sum_{\boldsymbol{z}_{n+1}} p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N, \boldsymbol{z}_{n+1}|\boldsymbol{z}_n) \\
&= \sum_{\boldsymbol{z}_{n+1}} p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N|\boldsymbol{z}_n, \boldsymbol{z}_{n+1}) p(\boldsymbol{z}_{n+1}|\boldsymbol{z}_n) \\
&= \sum_{\boldsymbol{z}_{n+1}} p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N|\boldsymbol{z}_{n+1}) p(\boldsymbol{z}_{n+1}|\boldsymbol{z}_n) \\
&= \sum_{\boldsymbol{z}_{n+1}} p(\boldsymbol{x}_{n+2}, \ldots, \boldsymbol{x}_N|\boldsymbol{z}_{n+1}) p(\boldsymbol{x}_{n+1}|\boldsymbol{z}_{n+1}) p(\boldsymbol{z}_{n+1}|\boldsymbol{z}_n) \\
&= \sum_{\boldsymbol{z}_{n+1}} \beta(\boldsymbol{z}_{n+1}) p(\boldsymbol{x}_{n+1}|\boldsymbol{z}_{n+1}) p(\boldsymbol{z}_{n+1}|\boldsymbol{z}_n)
\end{aligned}$$

5.2.3 Prediction

5.2.4 Baum-Welch Algorithm

5.2.5 Vertebi Algorithm

6 Useful Identities:

- Woodbury Identity:

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{C}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})\boldsymbol{V}\boldsymbol{A}^{-1}$$

- Law of Total Expectation:

$$\mathbb{E}_{p(\boldsymbol{X})}[\boldsymbol{X}] = \mathbb{E}_{p(\boldsymbol{Y})}[\mathbb{E}_{p(\boldsymbol{X}|\boldsymbol{Y})}[\boldsymbol{X}|\boldsymbol{Y}]]$$

- Conditional and Margin of Block Distributions Given the distribution:

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix}\right)$$

We have

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}) &= \mathcal{N}\left(\boldsymbol{\mu}_{x|y}, \Sigma_{x|y}\right) \\ \boldsymbol{\mu}_{x|y} &= \boldsymbol{\mu}_x + \Sigma_{xy}\Sigma_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y) \\ \Sigma_{x|y} &= \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \end{aligned}$$

- Product of two Gaussians (note that the underlying random variable must be the same):

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{a}, \boldsymbol{A})\mathcal{N}(\boldsymbol{x}|\boldsymbol{b}, \boldsymbol{B}) = c\mathcal{N}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{C})$$

$$\begin{aligned} \boldsymbol{C} &= (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1} \\ \boldsymbol{c} &= \boldsymbol{C}(\boldsymbol{A}^{-1}\boldsymbol{a} + \boldsymbol{B}^{-1}\boldsymbol{b}) \\ c &= \mathcal{N}(\boldsymbol{a}|\boldsymbol{b}, \boldsymbol{A} + \boldsymbol{B}) = \mathcal{N}(\boldsymbol{b}|\boldsymbol{a}, \boldsymbol{A} + \boldsymbol{B}) \end{aligned}$$

- Derivative of inverse

$$\begin{aligned} \frac{\partial \boldsymbol{Y}^{-1}}{\partial x} &= -\boldsymbol{Y}^{-1}\frac{\partial \boldsymbol{Y}}{\partial x}\boldsymbol{Y}^{-1} \\ \frac{\partial \boldsymbol{a}^\top \boldsymbol{X}^{-1}\boldsymbol{B}}{\partial \boldsymbol{X}} &= -\boldsymbol{X}^{-\top}\boldsymbol{a}\boldsymbol{b}^\top\boldsymbol{X}^{-\top} \end{aligned}$$

- Derivative of determinant

$$\frac{\partial |\boldsymbol{X}|}{\partial \boldsymbol{X}} = |\boldsymbol{X}|\boldsymbol{X}^{-\top}$$

$$\boldsymbol{a}^\top\boldsymbol{a} = \mathrm{tr}\,(\boldsymbol{a}\boldsymbol{a}^\top)$$