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: March 9, 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

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

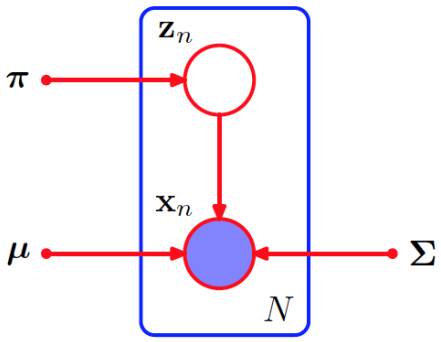

Figure 1: Gaussian Mixture Model.

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

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

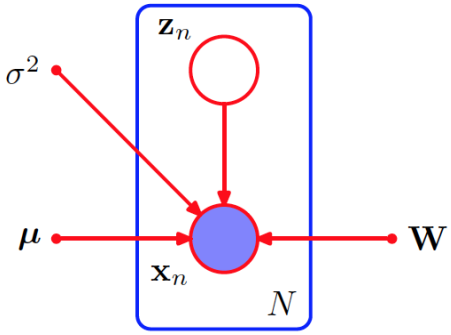

Figure 2: Gaussian Mixture Model.

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:

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

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:

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

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

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

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

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

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

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

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

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

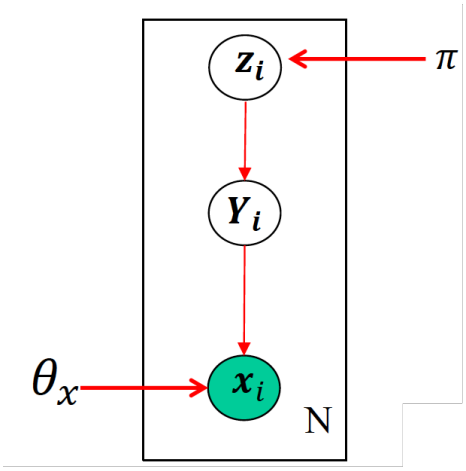

Figure 3: PPCA Mixture Model.

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:

$$
p(\boldsymbol{x}_i | \boldsymbol{z}_i, \boldsymbol{Y}_i, \boldsymbol{\theta}_x) = \prod_{k=1}^{K} p(\boldsymbol{x}_i | \boldsymbol{z}_{ik} = 1, \boldsymbol{y}_{ik}, \boldsymbol{W}_k, \boldsymbol{\mu}_k, \sigma_k^2 \boldsymbol{I})^{z_{ik}}
$$

$$= \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_{ik}|\boldsymbol{W}_k\boldsymbol{y}_{ik} + \boldsymbol{\mu}_k, \sigma_k\boldsymbol{I})^{z_{ik}}$$

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

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

4. Full marginal distributions:

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

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:

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

2. Taking the expectation yields

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

3. The maximization step:

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

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

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 ($a_{ij}(t) > 0$)
   - Aperiodicity: $d(i) = gcd\{t : a_{ii}(t) > 0\} = 1, \forall i$. 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

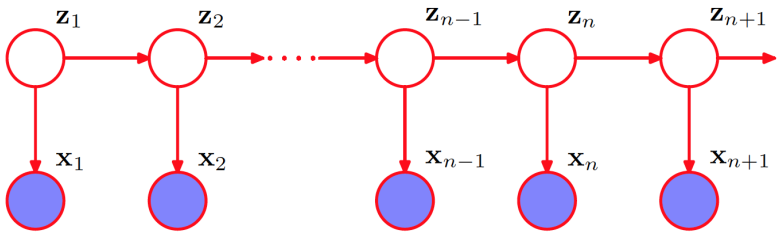

Figure 4: State Space Model. If z are discrete and markovian, then it is a Hidden Markov model.

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 $p(\boldsymbol{z}_t|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)$ offline, given all the evidence. Reducing the uncertainty by conditioning on the past and future data.
- **Fixed lag smoothing**: Compute $p(\boldsymbol{z}_{t-\ell}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$, where $\ell > 0$ is the lag.
- **MAP estimation**: Compute $\arg\max_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n} p(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n|\boldsymbol{x_1},\ldots,\boldsymbol{x_n})$. This is known as Viterbi Decoding.
- **Evaluation**: Find the probability of the evidence $p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = \sum_{\boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z})$.
- **Prediction**: Predicting the future given the past, we compute $p(\boldsymbol{z}_{n+\delta}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ and $p(\boldsymbol{x}_{n+\delta}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$.
- **Parameter Estimation**: This is know as the Baum-Welch algoritm. We estimate the parameter $\boldsymbol{A}$, $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$.

5.2.2 Formulation (Discrete Case)

Let there be K possible states for the latent variable $\boldsymbol{z}$ and S possible states for the observed variable $\boldsymbol{x}$.

$$p(\boldsymbol{z}_1|\boldsymbol{\pi}) = \prod_{c=1}^{K} \pi_c^{z_{1,c}}$$

$$p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1},\boldsymbol{A}) = \prod_{i=1}^{K}\prod_{j=1}^{K} a_{ij}^{z_{n-1,i}z_{n,j}}$$

$$p(\boldsymbol{x}_n|\boldsymbol{z}_n) = \prod_{i=1}^{S}\prod_{j=1}^{K} b_{ij}^{x_{n,i}z_{n,j}}$$

If $\boldsymbol{x}$ is continuous, we just replace the last equation as

$$p(\boldsymbol{x}_n|\boldsymbol{z}_n) = \prod_{j=1}^{K} p(\boldsymbol{x}_n|\boldsymbol{z}_{n,j})^{z_{n,j}}$$

Given the M observations of sequence D with size N

$$\begin{aligned}
p(\boldsymbol{D},\boldsymbol{Z}|\boldsymbol{\theta}) &= \prod_{l=1}^{M} p(\boldsymbol{x}_1^l,\ldots,\boldsymbol{x}_N^l,\boldsymbol{z}_1^l,\ldots,\boldsymbol{z}_N^l|\theta) \\
&= \prod_{l=1}^{M}\left(\prod_{n=1}^{N} p(\boldsymbol{x}_n^l|\boldsymbol{z}_n^l)p(\boldsymbol{z}_1^l)\prod_{n=2}^{N} p(\boldsymbol{z}_n^l|\boldsymbol{z}_{n-1}^l)\right) \\
&= \prod_{l=1}^{M}\left(\prod_{n=1}^{N}\prod_{i=1}^{S}\prod_{j=1}^{K} b_{ij}^{x_{n,i}^l z_{n,j}^l}\prod_{c=1}^{K}\pi_c^{z_{1,c}^l}\prod_{n=2}^{N}\prod_{i=1}^{K}\prod_{j=1}^{K} a_{ij}^{z_{n-1,i}^l z_{n,j}^l}\right)
\end{aligned}$$

Taking the log of the likelihood function

$$L = \sum_{l=1}^{M}\sum_{n=1}^{N}\sum_{i=1}^{S}\sum_{j=1}^{K} x_{n,i}^{l} z_{n,j}^{l} \ln b_{ij} + \sum_{l=1}^{M}\sum_{c=1}^{K} z_{1,c}^{l} \ln \pi_c + \sum_{l=1}^{M}\sum_{n=2}^{N}\sum_{i=1}^{K}\sum_{j=1}^{K} z_{n-1,i}^{l} z_{n,j}^{l} \ln a_{ij}$$

Taking expectation of the likelihood function

$$\mathbb{E}[L] = \sum_{l=1}^{M}\sum_{n=1}^{N}\sum_{i=1}^{S}\sum_{j=1}^{K} x_{n,i}^{l} \mathbb{E}[z_{n,j}^{l}] \ln b_{ij} + \sum_{l=1}^{M}\sum_{c=1}^{K} \mathbb{E}[z_{1,c}^{l}] \ln \pi_c + \sum_{l=1}^{M}\sum_{n=2}^{N}\sum_{i=1}^{K}\sum_{j=1}^{K} \mathbb{E}[z_{n-1,i}^{l} z_{n,j}^{l}] \ln a_{ij}$$

$$\begin{aligned}
\mathbb{E}\left[z_{1,c}^{l}\right] &= \sum_{z_{1,c}^{l}} z_{1,c}^{l} p\left(z_{1,c}^{l} | \boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_T^l\right) = p\left(z_{1,c}^{l} = 1 | \boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_T^l\right) \\
&= \frac{p\left(\boldsymbol{x}_1^l, z_{1,c}^{l} = 1\right) p\left(\boldsymbol{x}_2^l, \ldots, \boldsymbol{x}_N^l | z_{1,c}^{l} = 1\right)}{p\left(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right)} = \frac{\alpha\left(z_{1,c}^{l}\right)\beta\left(z_{1,c}^{l}\right)}{p\left(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right)} \\
\mathbb{E}\left[z_{n,j}^{l}\right] &= \sum_{z_{n,j}^{l}} z_{n,j}^{l} p\left(z_{n,j}^{l} | \boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right) = p\left(z_{n,j}^{l} = 1 | \boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right) \\
&= \frac{p\left(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_n^l, z_{n,j}^{l} = 1\right) p\left(\boldsymbol{x}_{t+1}^l, \ldots, \boldsymbol{x}_N^l | z_{n,j}^{l} = 1\right)}{p\left(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right)} = \frac{\alpha\left(z_{n,j}^{l}\right)\beta\left(z_{n,j}^{l}\right)}{p\left(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right)} \\
\mathbb{E}\left[z_{n-1,i}^{l} z_{n,j}^{l}\right] &= \sum_{z_{n-1,i}^{l}} \sum_{z_{n,j}^{l}} z_{n-1,i}^{l} z_{n,j}^{l} p\left(z_{n-1,i}^{l} z_{n,j}^{l} | \boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right) = p\left(z_{n-1,i}^{l} = 1, z_{n,j}^{l} = 1 | \boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right) \\
&= \frac{\alpha\left(z_{n-1,i}^{l}\right) \prod_{r=1}^{S} b_{j,r}^{x_{n,r}^{l}} a_{ij} \beta\left(z_{n,j}^{l}\right)}{p\left(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right)}
\end{aligned}$$

Formulating the Lagrangian and solving for the parameters we get

$$\begin{aligned}
\mathcal{L} = &\sum_{l=1}^{M}\sum_{n=1}^{N}\sum_{i=1}^{S}\sum_{j=1}^{K} x_{n,i}^{l} \mathbb{E}[z_{n,j}^{l}] \ln b_{ij} + \sum_{l=1}^{M}\sum_{c=1}^{K} \mathbb{E}[z_{1,c}^{l}] \ln \pi_c + \sum_{l=1}^{M}\sum_{n=2}^{N}\sum_{i=1}^{K}\sum_{j=1}^{K} \mathbb{E}[z_{n-1,i}^{l} z_{n,j}^{l}] \ln a_{ij} \\
&+ \left(\sum_{j=1}^{K} \lambda_j \sum_{i=1}^{S} b_{ij} - 1\right) + \mu\left(\sum_{c=1}^{K} \pi_c - 1\right) + \phi\left(\sum_{k=1}^{K} a_{jk} - 1\right)
\end{aligned}$$

Taking the derivative w.r.t. to the parameters, we get

$$\begin{aligned}
b_{ij} &= \frac{\sum_{l=1}^{M}\sum_{n=1}^{N} \mathbb{E}\left[z_{n,j}^{l}\right] x_{n,i}^{l}}{\sum_{l=1}^{M}\sum_{n=1}^{N} \mathbb{E}\left[z_{n,j}^{l}\right]} \\
\pi_c &= \frac{\sum_{l=1}^{M} \mathbb{E}\left[z_{1,c}^{l}\right]}{\sum_{l=1}^{M}\sum_{r=1}^{K} \mathbb{E}\left[z_{1,r}^{l}\right]} \\
a_{ij} &= \frac{\sum_{l=1}^{M}\sum_{n=2}^{N} \mathbb{E}\left[z_{n-1,i}^{l} z_{n,j}^{l}\right]}{\sum_{r=1}^{K}\sum_{l=1}^{M}\sum_{n=2}^{N} \mathbb{E}\left[z_{n-1,i}^{l} z_{n,r}^{l}\right]}
\end{aligned}$$

5.2.3 Formulation (Continuous Case)

Let there be K possible states for the latent variable $\boldsymbol{z}$ and $p(\boldsymbol{x}_n|\boldsymbol{z}_n)$ is a normal distribution.

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

$$p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}, \boldsymbol{A}) = \prod_{i=1}^{K}\prod_{j=1}^{K} a_{ij}^{z_{n-1,i}z_{n,j}}$$

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

Given the M observations of sequence D with size N

$$\begin{aligned} p(\boldsymbol{D}, \boldsymbol{Z}|\boldsymbol{\theta}) &= \prod_{l=1}^{M} p(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l, \boldsymbol{z}_1^l, \ldots, \boldsymbol{z}_N^l|\theta) \\ &= \prod_{l=1}^{M}\left(\prod_{n=1}^{N} p(\boldsymbol{x}_n^l|\boldsymbol{z}_n^l)p(\boldsymbol{z}_1^l)\prod_{n=2}^{N} p(\boldsymbol{z}_n^l|\boldsymbol{z}_{n-1}^l)\right) \\ &= \prod_{l=1}^{M}\left(\prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)^{z_{n,k}^l}\prod_{k=1}^{K}\pi_k^{z_{1,k}^l}\prod_{n=2}^{N}\prod_{i=1}^{K}\prod_{j=1}^{K} a_{ij}^{z_{n-1,i}^l z_{n,j}^l}\right) \end{aligned}$$

Taking log of the likelihood,

$$L = \sum_{l=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K} z_{n,k}^l \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k) + \sum_{l=1}^{M}\sum_{k=1}^{K} z_{1,k}^l \ln \pi_k + \sum_{l=1}^{M}\sum_{n=2}^{N}\sum_{i=1}^{K}\sum_{j=1}^{K} z_{n-1,i}^l z_{n,j}^l \ln a_{ij}$$

$$\mathbb{E}[L] = \sum_{l=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}\left[z_{n,k}^l\right] \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k) + \sum_{l=1}^{M}\sum_{k=1}^{K} \mathbb{E}\left[z_{1,k}^l\right] \ln \pi_k + \sum_{l=1}^{M}\sum_{n=2}^{N}\sum_{i=1}^{K}\sum_{j=1}^{K} \mathbb{E}\left[z_{n-1,i}^l z_{n,j}^l\right] \ln a_{ij}$$

$$\mathbb{E}\left[z_{1,c}^l\right] = \frac{p\left(\boldsymbol{x}_1^l, z_{1,c}^l = 1\right) p\left(\boldsymbol{x}_2^l, \ldots, \boldsymbol{x}_N^l | z_{1,c}^l = 1\right)}{p\left(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right)} = \frac{\alpha\left(z_{1,c}^l\right)\beta\left(z_{1,c}^l\right)}{p\left(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right)}$$

$$\mathbb{E}\left[z_{n,j}^l\right] = \frac{p\left(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_n^l, z_{n,j}^l = 1\right) p\left(\boldsymbol{x}_{t+1}^l, \ldots, \boldsymbol{x}_N^l | z_{n,j}^l = 1\right)}{p\left(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right)} = \frac{\alpha\left(z_{n,j}^l\right)\beta\left(z_{n,j}^l\right)}{p\left(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right)}$$

$$\mathbb{E}\left[z_{n-1,i}^l z_{n,j}^l\right] = \frac{\alpha\left(z_{n-1,i}^l\right)\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k) a_{ij} \beta\left(z_{n,j}^l\right)}{p\left(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right)}$$

Taking the derivative w.r.t. to the parameters, we get

$$\boldsymbol{\mu}_k = \frac{\sum_{l=1}^{M}\sum_{n=1}^{N} \mathbb{E}[z_{n,k}^l]\boldsymbol{x}_n^l}{\sum_{l=1}^{M}\sum_{n=1}^{N} \mathbb{E}[z_{n,k}^l]}$$

$$\Sigma_k = \frac{\sum_{l=1}^{M}\sum_{n=1}^{N}\mathbb{E}[z_{n,k}^l](\boldsymbol{x}_n^l - \boldsymbol{\mu}_k)(\boldsymbol{x}_n^l - \boldsymbol{\mu}_k)^\top}{\sum_{l=1}^{M}\sum_{n=1}^{N}\mathbb{E}[z_{n,k}^l]}$$

$$\pi_k = \frac{\sum_{l=1}^{M}\mathbb{E}\left[z_{1,k}^l\right]}{\sum_{l=1}^{M}\sum_{r=1}^{K}\mathbb{E}\left[z_{1,r}^l\right]}$$

$$a_{ij} = \frac{\sum_{l=1}^{M}\sum_{n=2}^{N}\mathbb{E}\left[z_{n-1,i}^l z_{n,j}^l\right]}{\sum_{r=1}^{K}\sum_{l=1}^{M}\sum_{n=2}^{N}\mathbb{E}\left[z_{n-1,i}^l z_{n,r}^l\right]}$$

5.2.4 Inference

1. **Filtering and Smoothing: Forwards and Backwards Algorithm**

   (a) We can write the posterior probabilities of the latent variables as below. After that, we will formulate the computation for $\alpha(\boldsymbol{z}_n)$ (forward probabilities) and $\beta(\boldsymbol{z}_n)$ (backward probabilities).

$$\begin{aligned}\gamma(\boldsymbol{z}_n) = p(\boldsymbol{z}_n|\boldsymbol{X}) &= \frac{p(\boldsymbol{X}|\boldsymbol{z}_n)p(\boldsymbol{z}_n)}{p(\boldsymbol{X})}\\ &= \frac{p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n|\boldsymbol{z}_n)p(\boldsymbol{x}_{n+1},\ldots,\boldsymbol{x}_N|\boldsymbol{z}_n)p(\boldsymbol{z}_n)}{p(\boldsymbol{X})}\\ &= \frac{p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\boldsymbol{z}_n)p(\boldsymbol{x}_{n+1},\ldots,\boldsymbol{x}_N|\boldsymbol{z}_n)}{p(\boldsymbol{X})}\\ &= \frac{\alpha(\boldsymbol{z}_n)\beta(\boldsymbol{z}_n)}{p(\boldsymbol{X})}\end{aligned}$$

   (b) The forward probabilities can be computed recursively based on the below:

$$\begin{aligned}\alpha(\boldsymbol{z}_n) &= p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\boldsymbol{z}_n)\\ &= p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n|\boldsymbol{z}_n)p(\boldsymbol{z}_n)\\ &= p(\boldsymbol{x}_n|\boldsymbol{z}_n)p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n-1}|\boldsymbol{z}_n)p(\boldsymbol{z}_n)\\ &= p(\boldsymbol{x}_n|\boldsymbol{z}_n)p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n-1},\boldsymbol{z}_n)\\ &= p(\boldsymbol{x}_n|\boldsymbol{z}_n)\sum_{\boldsymbol{z}_{n-1}}p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n-1},\boldsymbol{z}_n|\boldsymbol{z}_{n-1})p(\boldsymbol{z}_{n-1})\\ &= p(\boldsymbol{x}_n|\boldsymbol{z}_n)\sum_{\boldsymbol{z}_{n-1}}p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n-1}|\boldsymbol{z}_{n-1})p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})p(\boldsymbol{z}_{n-1})\\ &= p(\boldsymbol{x}_n|\boldsymbol{z}_n)\sum_{\boldsymbol{z}_{n-1}}p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n-1},\boldsymbol{z}_{n-1})p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})\\ &= p(\boldsymbol{x}_n|\boldsymbol{z}_n)\sum_{\boldsymbol{z}_{n-1}}\alpha(\boldsymbol{z}_{n-1})p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})\end{aligned}$$

The initial condition of the recursive formula is given by:

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

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

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

(c) The backward probabilities can be computed as follows:

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

Initial condition for the backward probabilities is given by:

$$p(\boldsymbol{z}_N|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \frac{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{z}_N)}{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)} = \frac{\alpha(\boldsymbol{z}_N)}{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)} \Rightarrow \beta(\boldsymbol{z}_N) = 1$$

2. **Prediction**

$$\begin{aligned}
p(\boldsymbol{x}_{n+2}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) &= \sum_{\boldsymbol{z}_{n+2}} p(\boldsymbol{x}_{n+2}|\boldsymbol{z}_{n+2})p(\boldsymbol{z}_{n+2}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \\
p(\boldsymbol{z}_{n+2}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) &= \sum_{\boldsymbol{z}_{n+1}} \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_{n+2}, \boldsymbol{z}_{n+1}, \boldsymbol{z}_n|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \\
&= \sum_{\boldsymbol{z}_{n+1}} \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_{n+2}, \boldsymbol{z}_{n+1}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_n)p(\boldsymbol{z}_n|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \\
&= \sum_{\boldsymbol{z}_{n+1}} \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_{n+2}, \boldsymbol{z}_{n+1}|\boldsymbol{z}_n)\tilde{\alpha}(\boldsymbol{z}_n) \\
&= \sum_{\boldsymbol{z}_{n+1}} \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_{n+2}|\boldsymbol{z}_{n+1})p(\boldsymbol{z}_{n+1}|\boldsymbol{z}_n)\tilde{\alpha}(\boldsymbol{z}_n) \\
&= \boldsymbol{A}^\top \boldsymbol{A}^\top \tilde{\boldsymbol{\alpha}}
\end{aligned}$$

3. **Smooth Transition**

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

4. **Baum-Welch Algorithm / EM**

   (a) Initialization: We set all the parameters to random values. For $\boldsymbol{A}$ or $\boldsymbol{\pi}$, if any of their elements are initialized to 0, they will remain to be 0.

   (b) Compute both the smoothed and filtered posterior probability recursively

   (c) Compute the expectation $\mathbb{E}[z^l_{n,k}]$ for all M number of sequences.

   (d) Compute the smooth transition probability

5. **Scaling Issue**

$$p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = \prod_{k=1}^{n} c_k$$
$$c_k = p(\boldsymbol{x}_k|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{k-1})$$

We can then define $\hat{\alpha}(\boldsymbol{z}_n)$ and $\hat{\beta}(\boldsymbol{z}_n)$

$$\alpha(\boldsymbol{z}_n) = p(\boldsymbol{z}_n|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = \hat{\alpha}(\boldsymbol{z}_n)\prod_{k=1}^{n} c_k$$
$$\beta(\boldsymbol{z}_n) = \hat{\beta}(\boldsymbol{z}_n)\prod_{k=n+1}^{N} c_k$$

The recursive formulae to use are then:

$$c_n\hat{\alpha}(\boldsymbol{z}_n) = p(\boldsymbol{x}_n|\boldsymbol{z}_n)\sum_{\boldsymbol{z}_{n-1}}\hat{\alpha}(\boldsymbol{z}_{n-1})p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})$$
$$c_{n+1}\hat{\beta}(\boldsymbol{z}_n) = \sum_{\boldsymbol{z}_{t+1}}\hat{\beta}(\boldsymbol{z}_{n+1})p(\boldsymbol{x}_{n+1}|\boldsymbol{z}_{n+1})p(\boldsymbol{z}_{n+1}|\boldsymbol{z}_n)$$

Hence, the posteriors become:

$$p(\boldsymbol{z}_n|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = \hat{\alpha}(\boldsymbol{z}_n)\hat{\beta}(\boldsymbol{z}_n)$$
$$p(\boldsymbol{z}_{n-1},\boldsymbol{z}_n|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = c_n^{-1}\hat{\alpha}(\boldsymbol{z}_{n-1})p(\boldsymbol{x}_n|\boldsymbol{z}_n)p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})\hat{\beta}(\boldsymbol{z}_n)$$

6. **Vertebi Algorithm**

   We define the probability of ending in a state j at time n, given that we take the most probable path

$$\begin{aligned}\delta_n(j) &= \max_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_{n-1}} p(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_{n-1},\boldsymbol{z}_{n,j}=1|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)\\ &= c_n^{-1} p(\boldsymbol{x}_n|\boldsymbol{z}_{n,j}=1)\max_i \delta_{n-1}(i)p(\boldsymbol{z}_{n,j}=1|\boldsymbol{z}_{n-1,i}=1)\\ a_n(j) &= \arg\max_i p(\boldsymbol{x}_n|\boldsymbol{z}_{n,j}=1)\delta_{n-1}(i)p(\boldsymbol{z}_{n,j}=1|\boldsymbol{z}_{n-1,i}=1)\end{aligned}$$

6 Linear Dynamical System (Kalman Filter)

6.1 Definition of a Linear Dynamical System

The model and the transition model are given as follows

$$\begin{aligned}\boldsymbol{x}_n &= \boldsymbol{C}\boldsymbol{z}_n + \boldsymbol{v}_n \\ \boldsymbol{z}_n &= \boldsymbol{A}\boldsymbol{z}_{n-1} + \boldsymbol{w}_n \\ \boldsymbol{z}_1 &= \boldsymbol{\mu}_0 + \boldsymbol{u}\end{aligned}$$

The distributions of the noise are given by

$$\begin{aligned}\boldsymbol{v} &\sim \mathcal{N}(\boldsymbol{v}|\boldsymbol{0}, \Sigma) \\ \boldsymbol{w} &\sim \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \boldsymbol{\Gamma}) \\ \boldsymbol{u} &\sim \mathcal{N}(\boldsymbol{u}|\boldsymbol{0}, \boldsymbol{V}_0)\end{aligned}$$

The emission probability and the transition probability is given by

$$\begin{aligned}p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}) &= \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{A}\boldsymbol{z}_{n-1}, \boldsymbol{\Gamma}) \\ p(\boldsymbol{x}_n|\boldsymbol{z}_n) &= \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{C}\boldsymbol{z}_n, \Sigma) \\ p(\boldsymbol{z}_1) &= \mathcal{N}(\boldsymbol{z}_1|\boldsymbol{\mu}_0, \boldsymbol{V}_0)\end{aligned}$$

6.2 Inference for LDS

6.2.1 Filtering

We can denote the probability $p(\boldsymbol{z}_n|\boldsymbol{x}_{1:n})$ as below:

$$\hat{\alpha}(\boldsymbol{z}_n) = p(\boldsymbol{z}_n|\boldsymbol{x}_{1:n}) = \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{\mu}_n, \boldsymbol{V}_n)$$

Deriving a recursive relation:

$$\begin{aligned}c_n\hat{\alpha}(\boldsymbol{z}_n) &= p(\boldsymbol{x}_n|\boldsymbol{z}_n)\int_{\boldsymbol{z}_{n-1}} \hat{\alpha}(\boldsymbol{z}_{n-1})p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})d\boldsymbol{z}_{n-1} \\ c_n\mathcal{N}(\boldsymbol{z}_n|\boldsymbol{\mu}_n, \boldsymbol{V}_n) &= \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{C}\boldsymbol{z}_n, \Sigma)\int_{\boldsymbol{z}_{n-1}} \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{A}\boldsymbol{z}_{n-1}, \boldsymbol{\Gamma})\mathcal{N}(\boldsymbol{z}_{n-1}|\boldsymbol{\mu}_{n-1}, \boldsymbol{V}_{n-1})d\boldsymbol{z}_{n-1}\end{aligned}$$

We can then either complete the square or use the identity of Gaussian multiplication for the integral:

$$\int_{\boldsymbol{z}_{n-1}} \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{A}\boldsymbol{z}_{n-1}, \boldsymbol{\Gamma})\mathcal{N}(\boldsymbol{z}_{n-1}|\boldsymbol{\mu}_{n-1}, \boldsymbol{V}_{n-1})d\boldsymbol{z}_{n-1} = \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{A}\boldsymbol{\mu}_{n-1}, \boldsymbol{P}_{n-1})$$

where,

$$\boldsymbol{P}_{n-1} = \boldsymbol{A}\boldsymbol{V}_{n-1}\boldsymbol{A}^\top + \boldsymbol{\Gamma}$$

Hence, we have

$$\begin{aligned} c_n\hat{\alpha}(\boldsymbol{z}_n) &= \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{C}\boldsymbol{z}_n, \Sigma)\mathcal{N}(\boldsymbol{z}_n|\boldsymbol{A}\boldsymbol{\mu}_{n-1}, \boldsymbol{P}_{n-1}) \\ c_n\mathcal{N}(\boldsymbol{z}_n|\boldsymbol{\mu}_n, \boldsymbol{V}_n) &= \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{C}\boldsymbol{z}_n, \Sigma)\mathcal{N}(\boldsymbol{z}_n|\boldsymbol{A}\boldsymbol{\mu}_{n-1}, \boldsymbol{P}_{n-1}) \end{aligned}$$

We can then use the product of Gaussians identity and the Woodbury identity to derive the update rule

$$\begin{aligned} \boldsymbol{\mu}_n &= \boldsymbol{A}\boldsymbol{\mu}_{n-1} + \boldsymbol{K}_n(\boldsymbol{x}_n - \boldsymbol{C}\boldsymbol{A}\boldsymbol{\mu}_{n-1}) \\ \boldsymbol{V}_n &= (\boldsymbol{I} - \boldsymbol{K}_n\boldsymbol{C})\boldsymbol{P}_{n-1} \\ c_n &= \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{C}\boldsymbol{A}\boldsymbol{\mu}_{n-1}, \boldsymbol{C}\boldsymbol{P}_{n-1}\boldsymbol{C}^\top + \Sigma) \\ \boldsymbol{K}_n &= \boldsymbol{P}_{n-1}\boldsymbol{C}^\top(\boldsymbol{C}\boldsymbol{P}_{n-1}\boldsymbol{C}^\top + \Sigma)^{-1} \end{aligned}$$

The initial condition for the recursion is given by

$$\begin{aligned} c_1\hat{\alpha}(\boldsymbol{z}_1) &= p(\boldsymbol{z}_1)p(\boldsymbol{x}_1|\boldsymbol{z}_1) \\ \boldsymbol{\mu}_1 &= \boldsymbol{\mu}_0 + \boldsymbol{K}_1(\boldsymbol{x}_1 - \boldsymbol{C}\boldsymbol{\mu}_0) \\ c_1 &= \mathcal{N}(\boldsymbol{x}_1|\boldsymbol{C}\boldsymbol{\mu}_0, \boldsymbol{C}\boldsymbol{V}_0\boldsymbol{C}^\top + \Sigma) \\ \boldsymbol{K_1} &= \boldsymbol{V}_0\boldsymbol{C}^\top(\boldsymbol{C}\boldsymbol{V}_0\boldsymbol{C}^\top + \Sigma)^{-1} \end{aligned}$$

6.2.2 Smoothing

$$\gamma(\boldsymbol{z}_n) = p(\boldsymbol{z}_n|\boldsymbol{x}_{1:T}) = \hat{\alpha}(\boldsymbol{z}_n)\hat{\beta}(\boldsymbol{z}_n) = \mathcal{N}(\boldsymbol{z}_n|\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{V}}_n)$$

$$c_{n+1}\hat{\beta}(\boldsymbol{z}_n) = \int_{\boldsymbol{z}_{n+1}} \hat{\beta}(\boldsymbol{z}_{n+1})p(\boldsymbol{x}_{n+1}|\boldsymbol{z}_{n+1})p(\boldsymbol{z}_{n+1}|\boldsymbol{z}_n)d\boldsymbol{z}_{n+1}$$

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \boldsymbol{\mu}_n + \boldsymbol{J}_n(\hat{\boldsymbol{\mu}}_{t+1} - \boldsymbol{A}\boldsymbol{\mu}_n) \\ \hat{\boldsymbol{V}}_n &= \boldsymbol{V}_n + \boldsymbol{J}_n(\hat{\boldsymbol{V}}_{n+1} - \boldsymbol{P}_n)\boldsymbol{J}_n^\top \\ \boldsymbol{J}_n &= \boldsymbol{V}_n\boldsymbol{A}^\top\boldsymbol{P}_n^{-1} \end{aligned}$$

$$\begin{aligned} \xi(\boldsymbol{z}_{n-1}, \boldsymbol{z}_n) &= P(\boldsymbol{z}_n, \boldsymbol{z}_{n-1}|\boldsymbol{x}_{1:T}) = c_n^{-1}\hat{\alpha}(\boldsymbol{z}_{n-1})p(\boldsymbol{x}_n|\boldsymbol{z}_n)P(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})\hat{\beta}(\boldsymbol{z}_n) \\ &= \frac{\mathcal{N}(\boldsymbol{z}_{n-1}|\boldsymbol{\mu}_{n-1}, \boldsymbol{V}_{n-1})\mathcal{N}(\boldsymbol{z}_n|\boldsymbol{A}\boldsymbol{z}_{n-1}, \boldsymbol{\Gamma})\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{C}\boldsymbol{z}_n, \Sigma)\mathcal{N}(\boldsymbol{z}_n|\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{V}}_n)}{c_n\hat{\alpha}(\boldsymbol{z}_n)} \\ &= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{z}_{n-1} \\ \boldsymbol{z}_n\end{bmatrix}\middle\|\begin{bmatrix}\hat{\boldsymbol{\mu}}_{n-1} \\ \hat{\boldsymbol{\mu}}_n\end{bmatrix}, \begin{bmatrix}\hat{\boldsymbol{V}}_{n-1} & \boldsymbol{J}_{n-1}\hat{\boldsymbol{V}}_n \\ (\boldsymbol{J}_{n-1}\hat{\boldsymbol{V}}_n)^\top & \hat{\boldsymbol{V}}_n\end{bmatrix}\right) \end{aligned}$$

6.3 EM Algorithm

First, we can compute the expectations that we will need in the EM algorithm

$$\begin{aligned}\mathbb{E}[\boldsymbol{z}_n] &= \hat{\boldsymbol{\mu}}_n \\ \mathbb{E}[\boldsymbol{z}_n \boldsymbol{z}_{n-1}^\top] &= \boldsymbol{J}_{n-1}\hat{\boldsymbol{V}}_n + \hat{\boldsymbol{\mu}}_n \hat{\boldsymbol{\mu}}_{n-1}^\top \\ \mathbb{E}[\boldsymbol{z}_n \boldsymbol{z}_n^\top] &= \hat{\boldsymbol{V}}_n + \hat{\boldsymbol{\mu}}_n \hat{\boldsymbol{\mu}}_n^\top\end{aligned}$$

Deriving the log joint likelihood

$$\begin{aligned}p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) &= p(\boldsymbol{x}_{1:N}, \boldsymbol{z}_{1:N}|\boldsymbol{\theta}) \\ &= \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{z}_n) p(\boldsymbol{z}_1) \prod_{n=2}^{N} p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})\end{aligned}$$

6.3.1 Expectation Step

$$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) = \ln p(\boldsymbol{z}_1|\boldsymbol{\mu}_0, \boldsymbol{V}_0) + \sum_{n=2}^{N} \ln p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}, \boldsymbol{A}, \boldsymbol{\Gamma}) + \sum_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{C}, \Sigma)$$

$$\mathbb{E}[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})] = \mathbb{E}[\ln p(\boldsymbol{z}_1|\boldsymbol{\mu}_0, \boldsymbol{V}_0)] + \mathbb{E}\left[\sum_{n=2}^{N} \ln p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}, \boldsymbol{A}, \boldsymbol{\Gamma})\right] + \mathbb{E}\left[\sum_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{C}, \Sigma)\right]$$

6.3.2 Maximization Step

To find $\boldsymbol{\mu}_0$ and $\boldsymbol{V}_0$:

$$\begin{aligned}\mathbb{E}[\ln p(\boldsymbol{z}_1|\boldsymbol{\mu}_0, \boldsymbol{V}_0)] &= -\frac{1}{2}\ln|\boldsymbol{V}_0| - \frac{1}{2}\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{\theta}^{old}}\left[(\boldsymbol{z}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{V}_0^{-1}(\boldsymbol{z}_1 - \boldsymbol{\mu}_0)\right] + \text{const.} \\ \boldsymbol{\mu}_0^{new} &= \mathbb{E}[\boldsymbol{z}_1] \\ \boldsymbol{V}_0^{new} &= \mathbb{E}[\boldsymbol{z}_1 \boldsymbol{z}_1^\top] - \mathbb{E}[\boldsymbol{z}_1]\mathbb{E}[\boldsymbol{z}_1]^\top\end{aligned}$$

To find $\boldsymbol{A}$ and $\boldsymbol{\Gamma}$:

$$\mathbb{E}\left[\sum_{n=2}^{N} \ln p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}, \boldsymbol{A}, \boldsymbol{\Gamma})\right] = -\frac{N-1}{2}\ln|\boldsymbol{\Gamma}| - \frac{1}{2}\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{\theta}^{old}}\left[\sum_{n=2}^{N}(\boldsymbol{z}_n - \boldsymbol{A}\boldsymbol{z}_{n-1})^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{z}_n - \boldsymbol{A}\boldsymbol{z}_{n-1})\right]$$

$$\boldsymbol{A}^{new} = \left(\sum_{n=2}^{N} \mathbb{E}[\boldsymbol{z}_n \boldsymbol{z}_{n-1}^\top]\right)\left(\sum_{n=2}^{N} \mathbb{E}[\boldsymbol{z}_{n-1} \boldsymbol{z}_{n-1}^\top]\right)^{-1}$$

$$\begin{aligned}\boldsymbol{\Gamma}^{new} = &\frac{1}{N-1}\sum_{n=2}^{N}\left\{\mathbb{E}[\boldsymbol{z}_n \boldsymbol{z}_n^\top] - \boldsymbol{A}^{new}\mathbb{E}[\boldsymbol{z}_{n-1}\boldsymbol{z}_n^\top]]\right\} \\ &- \frac{1}{N-1}\sum_{n=2}^{N}\left\{\mathbb{E}[\boldsymbol{z}_n \boldsymbol{z}_{n-1}^\top]\boldsymbol{A}^{new} - \boldsymbol{A}^{new}\mathbb{E}[\boldsymbol{z}_{n-1}\boldsymbol{z}_{n-1}^\top(\boldsymbol{A}^{new})^\top]\right\}\end{aligned}$$

To find $\boldsymbol{C}$ and Σ

$$\mathbb{E}\left[\sum_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{z}_n,\boldsymbol{C},\Sigma)\right] = -\frac{N}{2}\ln|\Sigma| - \frac{1}{2}\mathbb{E}_{Z|\theta^{old}}\left[\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{C}\boldsymbol{z}_n)^\top \Sigma^{-1}(\boldsymbol{x}_n - \boldsymbol{C}\boldsymbol{z}_n)\right] + \text{const.}$$

$$\boldsymbol{C}^{new} = \left(\sum_{n=1}^{N}\boldsymbol{x}_n\mathbb{E}[\boldsymbol{z}_n^\top]\right)\left(\sum_{n=1}^{N}\boldsymbol{x}_n\mathbb{E}[\boldsymbol{z}_n\boldsymbol{z}_n^\top]\right)^{-1}$$

$$\Sigma^{new} = \frac{1}{N}\sum_{n=1}^{N}\left\{\boldsymbol{x}_n\boldsymbol{x}_n^\top - \boldsymbol{C}^{new}\mathbb{E}[\boldsymbol{z}_n]\boldsymbol{x}_n^\top - \boldsymbol{x}_n\mathbb{E}[\boldsymbol{z}_n^\top]\boldsymbol{C}^{new} + \boldsymbol{C}^{new}\mathbb{E}[\boldsymbol{z}_n\boldsymbol{z}_n^\top]\boldsymbol{C}^{new}\right\}$$

7 Markov Random Fields

7.1 Graph Preliminaries

1. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a finite collection of nodes $\mathcal{V} = \{n_1, \dots, n_N\}$ and set of edges $\mathcal{E} \subset \begin{bmatrix} \mathcal{V} \\ 2 \end{bmatrix}$

2. Neighbor: Two nodes $n_i, n_j \in \mathcal{V}$ are neighbors $\Leftrightarrow (n_i, n_j) \in \mathcal{E}$. The neighbor of a node is denoted by $\mathcal{N}(n_i) = \{n_j : (n_i, n_j) \in \mathcal{E}\}$

3. Neighbor is a symmetric relation (undirected graph): $n_i \in \mathcal{N}(n_j) \Leftrightarrow n_j \in \mathcal{N}(n_i)$

4. Complete graph: $\forall n_i \in \mathcal{V}, \mathcal{N}(n_i) = \{(n_i, n_j), j = \{1, \dots, N\} \backslash \{i\}\}$

5. A clique is a complete subgraph of $\mathcal{G}$. Maximal clique is a clique with maximal number of nodes, i.e. cannot add any other node while still retaining complete connectedness.

7.2 Definition and Properties

1. Markov Random Field: Undirected graphical model in which each node corresponds to a random variable or a collection of random variables, and the edges identify conditional dependencies

2. Let A, B, C be three disjoint subsets of $\mathcal{V}$. Let x_A denote the collection of random variables in A. Conditional independence is defined as

$$p(x_A, x_B | x_C) = p(x_A | x_C) p(x_B | x_C)$$

3. Pairwise Markovianity: $(n_i, n_j) \notin \mathcal{E} \Rightarrow x_i$ and x_j are independent when conditioned on all other variables

$$p\left(x_i, x_j | \boldsymbol{x}_{\backslash \{i,j\}}\right) = p\left(x_i | \boldsymbol{x}_{\backslash \{i,j\}}\right) p\left(x_j | \boldsymbol{x}_{\backslash \{i,j\}}\right)$$

4. Local Markovianity: Given its neighborhood, a variable is independent on the rest of the variables

$$p(x_i | \boldsymbol{x}_{\mathcal{V} \backslash \{i\}} = p(x_i | \boldsymbol{x}_{\mathcal{N}(i)})$$

5. Global Markovianity: Let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ be three disjoint subsets of $\mathcal{V}$. If $\boldsymbol{C}$ separates $\boldsymbol{A}$ from $\boldsymbol{B}$, $\Rightarrow p(x_A, x_B | x_C) = p(x_B | x_C) p(x_B | x_C)$, then $p(x_A, x_B | x_C)$ is global markov w.r.t to graph $\mathcal{G}$.

6. Consider a random field $\boldsymbol{x}$ on a graph $\mathcal{G}$, s.t. $p(\boldsymbol{x} > 0)$. Let $\mathcal{C}$ denote the set of all maximal cliques of the graph.

- If the field has the local Markov property, then $p(\boldsymbol{x})$ can be written as a Gibbs distribution

$$p(\boldsymbol{x}) = \frac{1}{Z}\prod_{C}\Psi_C(\boldsymbol{x}_C) = \frac{1}{Z}\exp\left(-\sum_{C}V_C(\boldsymbol{x}_C)\right)$$
$$\Psi_C(\boldsymbol{x}_C) = \exp\left(-V_C(\boldsymbol{x}_C)\right) \text{ (potential functions)}$$
$$Z = \sum_{\boldsymbol{x}}\prod_{C}\Psi_C(\boldsymbol{x}_C) \text{ (partition functions)}$$

If $p(\boldsymbol{x})$ can be written as Gibbs distribution for the cliques of some graph, then it has the global Markov property.

7.3 The Ising Model

1. The Ising model considers an idealized system of interacting particles, arranged onto a regular planar grid.

2. Each particle can have only two possible states and it interacts only with its nearest neighbors.

3. The contribution of each particle to the total energy of the system depends upon the orientation of its spin compared to its neighbors.

4. If $\boldsymbol{x}_i$ denotes the possible states of the particle, assuming no external force, the total energy of the system is given by

$$E(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = -\frac{\beta}{2}\sum_{k=1}^{N}\sum_{j\in N(k)}\boldsymbol{x}_k^\top\boldsymbol{x}_j$$

The probability of the configuration is given by

$$p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = \frac{1}{Z(\beta)}\exp\left(-E(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)\right)$$

5. If there is external force

$$E(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = -\frac{\beta}{2}\sum_{k=1}^{N}\sum_{j\in N(k)}\boldsymbol{x}_k^\top\boldsymbol{x}_j - \lambda\sum_{k=1}^{N}\boldsymbol{x}_k^\top\boldsymbol{v}_k$$
$$p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = \frac{1}{Z(\beta)}\exp\left(-E(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)\right)$$

where $\boldsymbol{v}_k$ is a different constant per site.

7.3.1 Application: Change Detection

Let $\boldsymbol{z}$ be the latent variable and $\boldsymbol{x}$ be the pixels,

$$\boldsymbol{z} = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$$

The joint likelihood is given by

$$\begin{aligned} p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) &= p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) \\ &= \left(\prod_{i=1}^{N} p(\boldsymbol{x}_i | \boldsymbol{z}_i) \right) \left(\frac{1}{Z(\beta)} \exp\left(\frac{1}{2} \sum_{k=1}^{N} \sum_{j \in N(k)} \boldsymbol{z}_k^\top \boldsymbol{z}_j \right) \right) \end{aligned}$$

Assuming the $p(\boldsymbol{x}|\boldsymbol{z})$ is Gaussian,

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

$$\begin{aligned} p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) &= p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) \\ &= \left(\prod_{i=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \Sigma_k)^{\boldsymbol{z}_{i,k}} \right) \left(\frac{1}{Z(\beta)} \exp\left(\frac{1}{2} \sum_{k=1}^{N} \sum_{j \in N(k)} \boldsymbol{z}_k^\top \boldsymbol{z}_j \right) \right) \end{aligned}$$

However, $Z(\beta)$ cannot be computed. Typical solution to resolve this is to use mean field approximation

$$\begin{aligned} p(\boldsymbol{Z}) &= \frac{1}{Z(\beta)} \exp\left(\frac{1}{2} \beta \sum_{k=1}^{N} \sum_{j \in N(k)} \boldsymbol{z}_k^\top \boldsymbol{z}_j \right) \\ &\approx \prod_{k=1}^{N} p\left(\boldsymbol{z}_k \left| \mathbb{E}[\boldsymbol{z}_{j, \forall j \in N(k)}] \right. \right) \\ p\left(\boldsymbol{z}_k \left| \mathbb{E}[\boldsymbol{z}_{j, \forall j \in N(k)}] \right. \right) &= \frac{\exp\left(\frac{1}{2} \beta \sum_{k=1}^{N} \sum_{j \in N(k)} \boldsymbol{z}_k^\top \mathbb{E}[\boldsymbol{z}_j] \right)}{\sum_{\boldsymbol{z}_k} \exp\left(\frac{1}{2} \beta \sum_{k=1}^{N} \sum_{j \in N(k)} \boldsymbol{z}_k^\top \mathbb{E}[\boldsymbol{z}_j] \right)} \end{aligned}$$

7.4 Energy Minimization for MRFs

Generally we want to minimize the energy function

$$\begin{aligned} p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) &= p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) \\ &= \left(\prod_{i=1}^{N} p(\boldsymbol{x}_i | \boldsymbol{z}_i) \right) \left(\frac{1}{Z(\beta)} \exp\left(\frac{1}{2} \sum_{k=1}^{N} \sum_{j \in N(k)} \boldsymbol{z}_k^\top \boldsymbol{z}_j \right) \right) \\ &= \frac{1}{Z(\beta)} \exp\left(\frac{1}{2} \sum_{k=1}^{N} \sum_{j \in N(k)} \boldsymbol{z}_k^\top \boldsymbol{z}_j - \sum_{k=1}^{N} \ln p(\boldsymbol{x}_k | \boldsymbol{z}_k) \right) \end{aligned}$$

There are two constraints

1. Data constraint: Labeling should reflect the observation
2. Smoothness constraint: Labeling should reflect spatial consistency

7.5 Gaussian Markov Random Fields

1. A Gaussian random field $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ that satisfies

$$p(\boldsymbol{x}_i|\{\boldsymbol{x}_j : j \neq i\}) = p(\boldsymbol{x}_i|\{\boldsymbol{x}_j : j \in N(i)\})$$

   is a Gaussian Markov random field.

2. Another definition: A random vector $\boldsymbol{x} = (x_1, \ldots, x_N)^\top$ is called a Gaussian Markov random field with respect to graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with mean $\boldsymbol{\mu}$ and a positive semi-definite precision matrix $\boldsymbol{Q}$ if its density has the form

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{N}{2}}|\boldsymbol{Q}|\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

   and

$$q_{ij} \neq 0 \Leftrightarrow n_i \in \mathcal{N}(n_j)$$
$$q_{ij} = 0 \Leftrightarrow \boldsymbol{x}_i \perp \boldsymbol{x}_j|\boldsymbol{x}_{\setminus\{i,j\}}$$

3. Construction of $\boldsymbol{Q}$: Let A be the adjacency matrix of the graph (we assume that there's a recurrent edge to each node) and D is the degree matrix of the graph,

$$\boldsymbol{A} = [a_{ij}] \Rightarrow a_{ij} = \begin{cases} 1 & n_i \in \mathcal{N}(j) \\ 1 & i = j \\ 0 & \text{otherwise} \end{cases}$$
$$\boldsymbol{D} = [D_{ij}] \Rightarrow a_{ij} = \begin{cases} \sum_j a_{ij} & i = j \\ 0 & \text{otherwise} \end{cases}$$
$$\boldsymbol{Q} = \boldsymbol{D} - \boldsymbol{A}$$

   Note that $\boldsymbol{Q}$ is a sparse matrix.

4. We can also specify a Gaussian Markov random field via clique potentials:

$$V_C(\boldsymbol{x}_C) = \frac{1}{2}\left(\sum_{i\in C}\alpha_i^C x_i\right)^2 = \frac{1}{2}\left(\sum_{i\in\mathcal{V}}\alpha_i^C x_i\right)^2$$

   where

$$i \notin C \Rightarrow \alpha_i^C = 0$$

Then the exponent of the GMRF density becomes

$$\begin{aligned}
-\sum_{C\in\mathcal{C}} V_C(\boldsymbol{x}_C) &= -\frac{1}{2}\sum_{C\in\mathcal{C}}\left(\sum_{i\in\mathcal{V}} \alpha_i^C x_i\right)^2 \\
&= -\frac{1}{2}\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{V}}\left(\sum_{C\in\mathcal{C}} \alpha_i^C \alpha_j^C\right) x_i x_j \\
&= -\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}
\end{aligned}$$

5. Example of a GMRF: AR(1)

$$\begin{aligned}
\boldsymbol{x}_1 &\sim \mathcal{N}\left(0, (1-\phi^2)^{-1}\right) \\
\boldsymbol{x}_t|\boldsymbol{x}_{1:t-1} &\sim \mathcal{N}\left(\phi\boldsymbol{x}_{t-1}, 1\right) \\
\boldsymbol{Q} &= \begin{bmatrix} 1 & \phi & 0 & \cdots & 0 \\ -\phi & 1+\phi^2 & -\phi & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\phi & 1 \end{bmatrix}
\end{aligned}$$

7.5.1 Application: Image Denoising

1. Consider an image to be a rectangular lattice with first-order pixel neighbourhoods:
   - GMRF is used as the smoothing prior
   - Cliques: pairs of vertically or horizontally adjacent pixels
   - Clique potentials: squares of first-order differences

$$V_{\{(i,j),(i,j-1)\}}(x_{i,j}, x_{i,j-1}) = \frac{1}{2}\left(x_{i,j} - x_{i,j-1}\right)^2$$

   - $\boldsymbol{Q}$ will be a matrix with tridiagonal blocks.

2. The model is given by

$$\begin{aligned}
\boldsymbol{x} &= \boldsymbol{H}\boldsymbol{y} + \boldsymbol{u}, \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{u}|\boldsymbol{0}, \sigma^2\boldsymbol{I}) \text{ (observation model)} \\
p(\boldsymbol{x}) &= (2\pi)^{-\frac{N}{2}}|\boldsymbol{Q}|\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x}\right) \text{ (GMRF prior)} \\
\mathbb{E}[\boldsymbol{y}] &= [\sigma^2\boldsymbol{Q} + \boldsymbol{H}^\top\boldsymbol{H}]^{-1}\boldsymbol{H}^\top\boldsymbol{y} \text{ (posterior mean)}
\end{aligned}$$

3. The model is good for deblurring. However, when it comes to denoising, it oversmooth the image as the edge discontinuities are smoothed out. Preservation of discontinuities can be solved by
   - Other prior models

- Hidden / latent binary random variables v to turn off clique potentials

$$V(x_{i,j}, x_{i,j-1}, v_{i,j}) = \frac{1}{2}(1 - v_{i,j})(x_{i,j} - x_{i,j-1})^2$$

- Robust potential functions (L_2 vs L_1 norm)

4. Some other possible potentials

- Convex potentials

$$V(x) = |x^p|, p \in [1, 2] \qquad \text{Generalized Gaussians}$$

$$V(x) = \begin{cases} x^2 & |x| < a \\ 2a|x| - a^2 & |x| \geq a \end{cases} \qquad \text{Stevenson}$$

$$V(x) = 2a^2 \log\cosh(x/a) \qquad \text{Green}$$

- Non-convex potentials

$$V(x) = [\min(|x|, a)]^2 \qquad \text{Blake, Zisserman}$$

$$V(x) = \frac{x^2}{x^2 + a^2} \qquad \text{Geman, McClure}$$

8 Useful Identities:

- Woodbury Identity:

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

- Matrix Inversion Identities:

$$(\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1})^{-1} = \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{A}$$

- 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

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

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

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

- Derivative of inverse

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

- Derivative of trace

$$\frac{\partial \boldsymbol{X}\boldsymbol{A}}{\partial \boldsymbol{X}} = \boldsymbol{A}^\top$$
$$\frac{\partial \boldsymbol{X}^\top\boldsymbol{A}}{\partial \boldsymbol{X}} = \boldsymbol{A}$$
$$\frac{\partial \boldsymbol{A}\boldsymbol{X}\boldsymbol{B}}{\partial \boldsymbol{X}} = \boldsymbol{A}^\top\boldsymbol{B}^\top$$

- Derivative of determinant

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

- Matrix identities for trace

$$\boldsymbol{a}^{\top} \boldsymbol{a} = \text{tr} \left(\boldsymbol{a} \boldsymbol{a}^{\top} \right)$$
$$\text{tr}(\boldsymbol{ABC}) = \text{tr}(\boldsymbol{BCA}) = \text{tr}(\boldsymbol{CAB})$$