Probabilisitic and Unsupervised Learning

Phu Sakulwongtana

1 Probability Basis

1.1 Statistics Introduction

Remark 1. Before we start anything, let's recall the definition of all distributions (that would be used):

Distribution	Sample Space	Probability Density Function	$\mathbb{E}[\boldsymbol{x}]$	$\text{var}(\boldsymbol{x})$	Parameter
Bernoulli	$\{0,1\}$	$\theta^x(1-\theta)^{1-x}$	θ	$\theta(1-\theta)$	$\theta \in [0,1]$
Bernoulli(+)	$\begin{cases}\boldsymbol{x} \in \{0,1\}^D \\ \sum_{i=1}^D x_i = 1\end{cases}$	$\prod_{i=1}^D \theta_i^{x_i}$	θ_i	$\text{var}[x_i] = \theta_i(1-\theta_i)$	$\begin{cases}0 \le \boldsymbol{\theta} \le 1 \\ \sum_{i=1}^D \theta_i = 1\end{cases}$
Binomial	$[N]$	$\binom{N}{x}\theta^x(1-\theta)^{1-x}$	$N\theta$	$N\theta(1-\theta)$	$\theta \in [0,1]$
Multinormal	$\begin{cases}\boldsymbol{x} \in [N]^D \\ \sum_{i=1}^D x_i = N\end{cases}$	$\frac{N!}{x_1 x_2 \cdots x_K}\prod_{i=1}^D \theta_i^{x_i}$	$N\theta_i$	$N\theta_i(1-\theta_i)$	$\begin{cases}0 \le \boldsymbol{\theta} \le 1 \\ \sum_{i=1}^D \theta_i = 1\end{cases}$
Gaussian	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$	μ	σ^2	$\begin{cases}\mu \in \mathbb{R} \\ \sigma \in \mathbb{R}_{\ge 0}\end{cases}$
Multinormal	$\mathbb{R}^D$	$\frac{1}{\sqrt{\lvert 2\pi\boldsymbol{\Sigma}\rvert}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	$\begin{cases}\boldsymbol{\mu} \in \mathbb{R}^D \\ \boldsymbol{\Sigma} \in \mathbb{S}_+^{D\times D}\end{cases}$
Beta	$[0,1]$	$\frac{\Gamma}{\Gamma(a)+\Gamma(b)}x^{a-1}(1-x)^{b-1}$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	$\begin{cases}a > 0 \\ b > 0\end{cases}$
Dirichlet	$\begin{cases}\boldsymbol{x} \in \mathbb{R}^D \\ 0 \le \boldsymbol{x} \le 1 \\ \sum_{i=1}^D x_i = 1\end{cases}$	$\frac{\Gamma(\hat{\theta})}{\Gamma(\theta_1)\cdots\Gamma(\theta_D)}\prod_{i=1}^D x_i^{\theta_i - 1}$	$\theta_i/\hat{\theta}$	$\begin{cases}\text{var}[x_i] = \frac{\theta_i(\hat{\theta}-\theta_i)}{\hat{\theta}^2(\hat{\theta}+1)} \\ \text{cov}[x_i x_j] = \frac{-\theta_i\theta_j}{\hat{\theta}^2(\hat{\theta}+1)}\end{cases}$	$\theta_i > 0$
Gamma	$x > 0$	$\frac{1}{\Gamma(a)}b^a x^{a-1}\exp(-bx)$	$\frac{a}{b}$	$\frac{a}{b^2}$	$\begin{cases}a > 0 \\ b > 0\end{cases}$
Wishart*	$\boldsymbol{\Lambda}^{-1} \in \mathbb{S}_+^{D\times D}$	$B(\boldsymbol{W},\nu)\lvert\boldsymbol{\Lambda}\rvert^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\text{Tr}(\boldsymbol{W}^{-1}\boldsymbol{\Lambda})\right)$	$\nu\boldsymbol{W}$	–	$\begin{cases}\boldsymbol{W} \in \mathbb{S}_+^{D\times D} \\ \nu > D-1\end{cases}$
Poisson	$\mathbb{N}_0$	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ	$\lambda > 0$

Table 1: *1D Wishart is Gamma with $a = \nu/2$ and $1/2W$

where $\mathbb{S}_+^{D\times D}$ is the set of positive definite matrix of size $D\times D$, also we have the following addition definitions:

$$\Gamma(z)=\int_0^\infty x^{z-1}\exp(-x)\ \mathrm{d}x \qquad \hat{\theta}=\sum_{i=1}^{D}\theta_i$$
$$B(\boldsymbol{W},\nu)=|\boldsymbol{W}|^{-\nu/2}\left(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1}$$

Remark 2. **(Basic Quantities)** Now we will consider the following probability facts, which would be useful in the future:

$$\mathbb{E}[f(\boldsymbol{x})]=\int p(\boldsymbol{x})f(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} \qquad \begin{aligned}\mathrm{cov}(\boldsymbol{x},\boldsymbol{y})&=\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])(\boldsymbol{y}-\mathbb{E}[\boldsymbol{y}])^T]\\&=\mathbb{E}_{\boldsymbol{xy}}[\boldsymbol{x}\boldsymbol{y}^T]-\mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}^T]\end{aligned}$$
$$\begin{aligned}\mathrm{var}[f(\boldsymbol{x})]&=\mathbb{E}[(f(\boldsymbol{x})-\mathbb{E}[f(\boldsymbol{x})])^2]\\&=\mathbb{E}[f(\boldsymbol{x})^2]-\mathbb{E}[f(\boldsymbol{x})]^2\end{aligned}$$

Remark 3. **(Additional Quantities)** We have the following equality:

$$\mathbb{E}_{\boldsymbol{x}}[f(\boldsymbol{x})]=\mathbb{E}_{\boldsymbol{z}}\left[\mathbb{E}_{\boldsymbol{x}|\boldsymbol{z}}[f(\boldsymbol{x})]\right] \qquad \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}]=\mathbb{E}_{\boldsymbol{z}}\left[\mathbb{V}[\boldsymbol{x}|\boldsymbol{z}]\right]+\mathbb{V}_{\boldsymbol{z}}[\mathbb{E}[\boldsymbol{x}|\boldsymbol{z}]]$$

1.2 Linear Algebra

Proposition 1.1. ***(Woodbury Identity)*** *This following identity can helps the computation as follows:*

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

This identity is useful when $\boldsymbol{A}\in\mathbb{R}^{a\times a}$ is large and diagonal (easy to invert), while $\boldsymbol{B}\in\mathbb{R}^{a\times b}$ has many rows but few columns ($a>b$) conversely for $\boldsymbol{C}\in\mathbb{R}^{b\times a}$. The RHS is simplier than LHS.

Proof. This can be proven easily as:

$$\begin{aligned}&\Big[\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D}+\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}\Big](\boldsymbol{A}+\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})\\
&\quad=\boldsymbol{A}^{-1}(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})-\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D}+\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})\\
&\quad=\boldsymbol{A}^{-1}\boldsymbol{A}+\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}-\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D}+\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{A}\\
&\qquad\qquad-\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D}+\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\\
&\quad=\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}-\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D}+\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{D}\boldsymbol{D}^{-1}\boldsymbol{C}\\
&\qquad\qquad-\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D}+\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\\
&\quad=\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}-\boldsymbol{A}^{-1}\boldsymbol{B}\Big[(\boldsymbol{D}+\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{D}+(\boldsymbol{D}+\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\Big]\boldsymbol{D}^{-1}\boldsymbol{C}\\
&\quad=\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}-\boldsymbol{A}^{-1}\boldsymbol{B}\Big[(\boldsymbol{D}+\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}(\boldsymbol{D}+\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})\Big]\boldsymbol{D}^{-1}\boldsymbol{C}\\
&\quad=\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}-\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}=\boldsymbol{I}\end{aligned}$$

□

Proposition 1.2. ***(Another Identity)*** *Another useful identity can be stated as:*

$$(\boldsymbol{P}^{-1}+\boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{R}^{-1}=\boldsymbol{P}\boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{P}\boldsymbol{B}^T+\boldsymbol{R})^{-1}$$

Proof. This can be proven easily as:

$$\begin{aligned}
&(\boldsymbol{P}^{-1} + \boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{R}^{-1}(\boldsymbol{B}\boldsymbol{P}\boldsymbol{B}^T + \boldsymbol{R}) \\
&\quad = (\boldsymbol{P}^{-1} + \boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B}\boldsymbol{P}\boldsymbol{B}^T + (\boldsymbol{P}^{-1} + \boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{R} \\
&\quad = (\boldsymbol{P}^{-1} + \boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B}\boldsymbol{P}\boldsymbol{B}^T + (\boldsymbol{P}^{-1} + \boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^T \\
&\quad = (\boldsymbol{P}^{-1} + \boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B})^{-1}\left[\boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B}\boldsymbol{P} + \boldsymbol{P}^{-1}\boldsymbol{P}\right]\boldsymbol{B}^T \\
&\quad = (\boldsymbol{P}^{-1} + \boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B})^{-1}\left[\boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B} + \boldsymbol{P}^{-1}\right]\boldsymbol{P}\boldsymbol{B}^T = \boldsymbol{P}\boldsymbol{B}^T
\end{aligned}$$

□

Remark 4. **(More Matrix Identities)** We have the following:

$$\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = \mathrm{Tr}(\boldsymbol{C}\boldsymbol{A}\boldsymbol{B}) = \mathrm{Tr}(\boldsymbol{B}\boldsymbol{C}\boldsymbol{A}) \qquad \left|\boldsymbol{A}^{-1}\right| = \frac{1}{|\boldsymbol{A}|} \qquad |\boldsymbol{A}\boldsymbol{B}| = |\boldsymbol{A}|\,|\boldsymbol{B}|$$

Remark 5. **(Additional Identity)**

$$\left|\boldsymbol{I}_N + \boldsymbol{A}\boldsymbol{B}^T\right| = \left|\boldsymbol{I}_M + \boldsymbol{A}^T\boldsymbol{B}\right|$$

This also implies that $\left|\boldsymbol{I}_N + \boldsymbol{a}\boldsymbol{b}^T\right| = 1 + \boldsymbol{a}^T\boldsymbol{b}$

Remark 6. **(Matrix Derivative - Basics)** We still use the following facts:

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}^T\boldsymbol{a}) = \frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{a}^T\boldsymbol{x}) = \boldsymbol{a} \qquad \frac{\partial}{\partial x}\boldsymbol{A}\boldsymbol{B} = \frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{B} + \frac{\partial \boldsymbol{B}}{\partial x}\boldsymbol{A} \qquad \frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x}$$

Proposition 1.3. *The derivative of the inverse matrix is given by:*

$$\frac{\partial}{\partial x}(\boldsymbol{A}^{-1}) = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1}$$

Proof. We consider differetiate the following eqation $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$ as we have:

$$\boldsymbol{0} = \left(\frac{\partial}{\partial x}\boldsymbol{I}\right)\boldsymbol{A}^{-1} = \left(\frac{\partial}{\partial x}\boldsymbol{A}^{-1}\boldsymbol{A}\right)\boldsymbol{A}^{-1} = \left(\frac{\partial \boldsymbol{A}^{-1}}{\partial x}\boldsymbol{A} + \boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\right)\boldsymbol{A}^{-1} = \frac{\partial \boldsymbol{A}^{-1}}{\partial x} + \boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1}$$

With algebraic manipulation the proposition is proven. □

Remark 7. **(Additional Matrix Derivative)**

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{A}}\,\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}) = \boldsymbol{B}^T \qquad \frac{\partial}{\partial \boldsymbol{A}}\,\mathrm{Tr}(\boldsymbol{A}^T\boldsymbol{B}) = \boldsymbol{B} \\
&\frac{\partial}{\partial \boldsymbol{A}}\,\mathrm{Tr}(\boldsymbol{A}) = \boldsymbol{I} \qquad \frac{\partial}{\partial \boldsymbol{A}}\,\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^T) = \boldsymbol{A}(\boldsymbol{B} + \boldsymbol{B}^T)
\end{aligned}$$

Proposition 1.4. *The matrix can be diagonalization as:*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T = \sum_{i=1}^{N}\lambda_i\boldsymbol{u}_i\boldsymbol{u}_i^T$$

$\boldsymbol{U}$ to be the matrix constrcuted that has the column as the eigenvectors $\boldsymbol{u}_i$. The matrix $\boldsymbol{\Lambda}$ is the diagonal matrix, whose diagonal element is the eigenvalue λ_i.

Proof. Recall that the square matrix $\boldsymbol{A} \in \mathbb{R}^{M\times M}$'s eigenvalue and eigenvector, which are given by:

$$\boldsymbol{A}\boldsymbol{u}_i = \lambda_i\boldsymbol{u}_i$$

For $i = 1, \ldots, M$ where $\boldsymbol{u}_i$ is the eigenvector and λ_i is the corresponding eigenvalue. This means that $\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda}$; furthermore, $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ as we have $|\boldsymbol{U}| = 1$. The identity follows by right multiplying with $\boldsymbol{U}^T$. □

Proposition 1.5. *We can show that the determinant and trace of the matrix to be:*

$$|\boldsymbol{A}| = \prod_{i=1}^{M} \lambda_i \qquad \text{Tr}(\boldsymbol{A}) = \sum_{i=1}^{M} \lambda_i$$

This follows from the identity of determinant and the cyclic properties of trace.

Proposition 1.6. *We now going to show very useful identity*

$$\frac{\partial}{\partial x} \ln |\boldsymbol{A}| = \text{Tr}\left(\boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial x}\right)$$

Proof. We consider the LHS first, as we have:

$$\frac{\partial}{\partial x} \ln |\boldsymbol{A}| = \frac{\partial}{\partial x} \ln \left(\prod_{i=1}^{M} \lambda_i\right) = \frac{\partial}{\partial x} \sum_{i=1}^{M} \ln \lambda_i = \sum_{i=1}^{M} \frac{1}{\lambda_i} \frac{\partial \lambda_i}{\partial x}$$

Now, consider the RHS, which we have:

$$\begin{aligned}
\text{Tr}\left(\left[\sum_{i=1}^{M} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^T\right]\left[\sum_{i=1}^{M} \frac{\partial \lambda_i}{\partial x} \boldsymbol{u}_i \boldsymbol{u}_i^T\right]\right) &= \text{Tr}\left(\sum_{i=1}^{M}\sum_{j=1}^{M} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^T \frac{\partial \lambda_j}{\partial x} \boldsymbol{u}_j \boldsymbol{u}_j^T\right) \\
&= \sum_{i=1}^{M}\sum_{j=1}^{M} \frac{1}{\lambda_i} \frac{\partial \lambda_j}{\partial x} \text{Tr}\left(\boldsymbol{u}_i \boldsymbol{u}_i^T \boldsymbol{u}_j \boldsymbol{u}_j^T\right) \\
&= \sum_{i=1}^{M}\sum_{j=1}^{M} \frac{1}{\lambda_i} \frac{\partial \lambda_j}{\partial x} \text{Tr}\left(\boldsymbol{u}_i^T \boldsymbol{u}_j \boldsymbol{u}_j^T \boldsymbol{u}_i\right) = \sum_{i=1}^{M} \frac{1}{\lambda_i} \frac{\partial \lambda_i}{\partial x}
\end{aligned}$$

The equality is proven. □

Corollary 1.1. *The above proposition implies that:*

$$\frac{\partial}{\partial \boldsymbol{A}} \ln |\boldsymbol{A}| = (\boldsymbol{A}^{-1})^T$$

Proof. Consider the following partial derivative, we have:

$$\frac{\partial}{\partial a_{cd}} \ln |\boldsymbol{A}| = \text{Tr}\left(\boldsymbol{A}^{-1} \frac{\partial A}{\partial a_{cd}}\right) = a_{dc}^{-1}$$

where we have a_{ij}^{-1} be the (i,j)-th element of the matrix $\boldsymbol{A}^{-1}$. Thus we have proven the equality. □

Proposition 1.7. *We can show that:*

$$\frac{\partial}{\partial \boldsymbol{A}} \text{Tr}[\boldsymbol{A}^T \boldsymbol{B} \boldsymbol{A} \boldsymbol{C}] = \boldsymbol{B}\boldsymbol{A}\boldsymbol{C} + \boldsymbol{B}^T \boldsymbol{A} \boldsymbol{C}^T$$

Proof. We will use the identity mapping $F_1(\cdot)$ and $F_2(\cdot)$ to make the differetiation easier:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{A}} \text{Tr}[\boldsymbol{A}^T \boldsymbol{B} \boldsymbol{A} \boldsymbol{C}] &= \frac{\partial}{\partial \boldsymbol{A}} \text{Tr}[F_1(\boldsymbol{A})^T \boldsymbol{B} F_2(\boldsymbol{A}) \boldsymbol{C}] \\
&= \frac{\partial}{\partial \boldsymbol{F}_1} \text{Tr}[\boldsymbol{F}_1^T \boldsymbol{B} \boldsymbol{F}_2 \boldsymbol{C}] \frac{\partial F_1}{\partial \boldsymbol{A}} + \frac{\partial}{\partial \boldsymbol{F}_2} \text{Tr}[\boldsymbol{F}_1^T \boldsymbol{B} \boldsymbol{F}_2 \boldsymbol{C}] \frac{\partial F_2}{\partial \boldsymbol{A}} \\
&= \frac{\partial}{\partial \boldsymbol{F}_1} \text{Tr}[\boldsymbol{F}_1^T \boldsymbol{B} \boldsymbol{F}_2 \boldsymbol{C}] \frac{\partial F_1}{\partial \boldsymbol{A}} + \frac{\partial}{\partial \boldsymbol{F}_2} \text{Tr}[\boldsymbol{C} \boldsymbol{F}_1^T \boldsymbol{B} \boldsymbol{F}_2] \frac{\partial F_2}{\partial \boldsymbol{A}} \\
&= \frac{\partial}{\partial \boldsymbol{F}_1} \text{Tr}[\boldsymbol{F}_1^T \boldsymbol{B} \boldsymbol{F}_2 \boldsymbol{C}] \frac{\partial F_1}{\partial \boldsymbol{A}} + \frac{\partial}{\partial \boldsymbol{F}_2} \text{Tr}[\boldsymbol{F}_2^T \boldsymbol{B}^T \boldsymbol{F}_1 \boldsymbol{C}^T] \frac{\partial F_2}{\partial \boldsymbol{A}} \\
&= \boldsymbol{B}\boldsymbol{F}_2\boldsymbol{C} + \boldsymbol{B}^T \boldsymbol{F}_1 \boldsymbol{C}^T = \boldsymbol{B}\boldsymbol{A}\boldsymbol{C} + \boldsymbol{B}^T \boldsymbol{A} \boldsymbol{C}^T
\end{aligned}$$

where we have $F_1(\boldsymbol{A}) = F_2(\boldsymbol{A}) = \boldsymbol{F}_1 = \boldsymbol{F}_2 = \boldsymbol{A}$ □

Remark 8. **(Notes on Symmetric Matrix)** We can construct the symmetric matrix from any kind of matrix $\boldsymbol{A}$ using the following formula:

$$\boldsymbol{M} = \frac{\boldsymbol{A} + \boldsymbol{A}^T}{2}$$

One can show that its eigenvalue is real (given real symmetric matrix), as we consider the complex of, assuning $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{M}$ with its eigenvalue to be λ:

$$\langle \boldsymbol{M}\boldsymbol{x}, \boldsymbol{M}\boldsymbol{x} \rangle = \boldsymbol{x}^* \boldsymbol{M}^* \boldsymbol{M}\boldsymbol{x} = \boldsymbol{x}^* \boldsymbol{M}\boldsymbol{M}\boldsymbol{x} = \boldsymbol{x}^* \lambda^2 \boldsymbol{x} = \lambda^2 \|\boldsymbol{x}\|^2$$

Where $\boldsymbol{M}^* = \bar{\boldsymbol{M}}^T$, and so λ^2 is real a non-negative, thus being a real number.

Remark 9. **(Notes on Square-Root of Matrix)** Given positive semi-definite matrix $\boldsymbol{A}$, one can show that there a matrix $\boldsymbol{B}$ such that $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}$ (or $\boldsymbol{B}^T\boldsymbol{B}$ as $\boldsymbol{B}$ is symmetric as we will shown later). Given the eigendecomposition of $\boldsymbol{A}$ to be $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, matrix $\boldsymbol{B}$ is $\boldsymbol{U}\sqrt{\boldsymbol{\Lambda}}\boldsymbol{U}^T$, where $\sqrt{\boldsymbol{\Lambda}}$ is the matrix contains square root of the diagonal of $\boldsymbol{\Lambda}$:

$$\boldsymbol{B}\boldsymbol{B} = (\boldsymbol{U}\sqrt{\boldsymbol{\Lambda}}\boldsymbol{U}^T)(\boldsymbol{U}\sqrt{\boldsymbol{\Lambda}}\boldsymbol{U}^T) = \boldsymbol{U}\sqrt{\boldsymbol{\Lambda}}\sqrt{\boldsymbol{\Lambda}}\boldsymbol{U}^T = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T = \boldsymbol{A}$$

Please note that, since $\boldsymbol{A}$ is positive semi-definite, the eigenvalue is non-negative implies that $\boldsymbol{B}$ has real value eigenvalue (and non-negative), and so $\boldsymbol{B}$ is symmetric. Finally, if $\boldsymbol{A}$ is positive define, then there is a *unique* $\boldsymbol{B}$.

Proposition 1.8. ***(Partition Matrix)*** *The block matrix can be inversed as:*

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix} = \begin{pmatrix} \boldsymbol{M} & -\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1} \end{pmatrix}$$

where $\boldsymbol{M} = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}$

1.3 Optimization

Definition 1.1. (Constraint Optimization) The constraint optimization is the optimization problem is in the form of:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f_0(\boldsymbol{x}) \\ \text{s.t} \quad & f_i(\boldsymbol{x}) \leq 0 \quad i \in \mathcal{I} = [m] \\ & h_i(\boldsymbol{x}) = 0 \quad i \in \mathcal{E} = [p] \end{aligned}$$

Definition 1.2. (KKT Condition) The constraint optimization problem given above can be solved using the KKT condition. Before that, we consider the Lagragian to be defined as:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f_0(\boldsymbol{x}) + \sum_{i \in \mathcal{I}} \lambda_i f_i(\boldsymbol{x}) + \sum_{i \in \mathcal{E}} \mu_i h_i(\boldsymbol{x})$$

The KKT condition is given by:

$$\begin{aligned} & \nabla \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = 0 \\ & f_i(\boldsymbol{x}) \leq 0 \quad \text{for} \quad i \in \mathcal{I} \\ & h_i(\boldsymbol{x}) = 0 \quad \text{for} \quad i \in \mathcal{E} \\ & \lambda_i \geq 0 \quad \text{for} \quad i \in \mathcal{I} \\ & \lambda_i f_i(\boldsymbol{x}) = 0 \quad \text{for} \quad i \in \mathcal{I} \end{aligned}$$

1.4 What are we going to do ?

Theorem 1.1. ***(Bayes' Theorem)*** *One can show that:*

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Remark 10. There are many ways to learn the parameter given the dataset $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^N$, as we have:

- *Maximum Likelihood:* Find the parameter $\boldsymbol{\theta}_{\text{ML}}$ such that it maximizes the log-likelihood as we have:

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \ \log p(\mathcal{D}|\boldsymbol{\theta})$$

- *Bayesian Inference*: Find the distribution over the parameter $\boldsymbol{\theta}$ using Bayes' Theorem:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta})}{p(\mathcal{D})}$$

- *Maximum A Posteriori*: Find the mode of the posterior distribution over parameter

$$\boldsymbol{\theta}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \ \log P(\boldsymbol{\theta}|\mathcal{D})$$

The main problems/solutions of this works is simply trying to get better estimate of $p(\boldsymbol{\theta}|\mathcal{D})$ as it maybe intractable to calculate.

Definition 1.3. (Bayesian Model) The model is $\mathcal{M} = \{P(\cdot|\theta) : \theta \in \mathcal{T}\}$, where they are the distribution of a single random variable $X \sim P(\cdot|\theta)$. Given the prior, we also have a prior π on parameter space $\mathcal{T}$. The data is generated by the following process:

$$\Theta \sim \pi \qquad X_1, \ldots, X_n|\Theta \sim_{\text{iid}} p(\cdot|\Theta)$$

The tuple $(\mathcal{M}, \pi)$ is the Bayesian model.

Remark 11. **(Model Selection)** Given various kinds of model $\mathcal{M}_1, \mathcal{M}_2, \ldots$. The following set of likelihood associated with $\mathcal{M}_i$ is

$$\{p(\boldsymbol{x}|\boldsymbol{\theta}_i, \mathcal{M}_i) : \boldsymbol{\theta}_i \in \mathcal{T}_i\}$$

We are interested in selecting the $\mathcal{M}_i$. Starting with the prior $p(\mathcal{M}_i)$ and the prior probability of parameter, given the model $\mathcal{M}_i$, is $p(\boldsymbol{\theta}_i|\mathcal{M}_i)$. Finally, the data probability is, where we assume the iid of the dataset:

$$p(\mathcal{D}|\boldsymbol{\theta}_i, \mathcal{M}_i) = \prod_{i=1}^N p(\boldsymbol{x}_i|\boldsymbol{\theta}_i, \mathcal{M}_i)$$

Now, we can find the posterior of the parameter given the model $\mathcal{M}_i$ together with dataset evidence:

$$p(\boldsymbol{\theta}_i|\mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D}|\boldsymbol{\theta}_i, \mathcal{M}_i)P(\boldsymbol{\theta}_i|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_i)} \qquad p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\boldsymbol{\theta}_i, \mathcal{M})p(\boldsymbol{\theta}_i|\mathcal{M}_i) \ \mathrm{d}\boldsymbol{\theta}_i$$

We can perform Bayesian inference over the model $\mathcal{M}_i$

$$p(\mathcal{M}_i|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)}{p(\mathcal{D})}$$

Now, we have the distribution over possible models.

1.5 Exponential Family and Friends

Definition 1.4. (Exponential Family) The set of probability distribution $\{p(\cdot|\theta) : \theta \in \mathcal{T}\}$, where $\mathcal{T}$ is the parameter space, is exponential family if we have the distribution of the form:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = f(\boldsymbol{x})g(\boldsymbol{\theta}) \exp\left(\boldsymbol{\phi}(\boldsymbol{\theta})^T \boldsymbol{T}(\boldsymbol{x})\right)$$

where each components are given (and named) as:

- Sufficient Statistics: $\boldsymbol{T} : \mathcal{X} \to \mathbb{R}^m$
- Natural Parameter: $\phi : \mathcal{T} \to \mathbb{R}^m$
- Auxilliary Functions $f : \mathcal{X} \to \mathbb{R}_{\geq 0}$ and $g : \mathcal{T} \to \mathbb{R}_{>0}$ (normalizing factor)

Please note that the function g has the following properties:

$$g(\boldsymbol{\theta}) \int f(\boldsymbol{x}) \exp\left(\boldsymbol{\phi}(\boldsymbol{\theta})^T \boldsymbol{T}(\boldsymbol{x})\right) \, \mathrm{d}\boldsymbol{x} = 1$$

Remark 12. Let's consider the example of exponential families as:

Distribution	$\boldsymbol{\phi}(\boldsymbol{\theta})$	$\boldsymbol{T}(\boldsymbol{x})$
Bernoulli/Binomial	$\ln\left(\frac{\theta}{1-\theta}\right)$	x
Bernoulli(+)/Multinormal	$[\ln\theta_1, \ln\theta_2, \ldots, \ln\theta_D]$	$[x_1, x_2, \ldots, x_D]$
Gaussian	$[\mu/\sigma^2, -1/2\sigma^2]$	$[x, x^2]$
Multinormal	$\left[-\frac{1}{2}\mathrm{Vec}(\boldsymbol{\Sigma}^{-1}),\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]$	$\left[\mathrm{Vec}(\boldsymbol{x}\boldsymbol{x}^T),\ \boldsymbol{x}\right]$
Beta	$[a-1, b-1]$	$[\ln x, \ln(1-x)]$
Dirichlet	$[a_1-1, a_2-1, \ldots, a_D-1]$	$[\ln x_1, \ln x_2, \ldots, \ln x_D]$
Gamma	$[a-1, -b]$	$[\ln x, x]$
Poisson	$\ln\lambda$	x

where we have the following opeartion $\mathrm{Vec} : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\cdot n\times 1}$ is defined as:

$$\mathrm{Vec}(\boldsymbol{X}) = [X_{11}, \ldots, X_{n1}, \ldots, X_{1m}, \ldots, X_{nm}]^T$$

Proposition 1.9. *The normal distribution can be written as:*

$$\begin{aligned}
&\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\
&\qquad = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right\} \exp\left\{\left[-\frac{1}{2}\,\mathrm{Vec}\left(\Sigma^{-1}\right), \Sigma^{-1}\boldsymbol{\mu}\right]^T \left[\mathrm{Vec}\left(\boldsymbol{x}\boldsymbol{x}^T\right), \boldsymbol{x}\right]\right\}
\end{aligned}$$

Please note that $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{vec}(\boldsymbol{A})^T\,\mathrm{vec}(\boldsymbol{B})$*, which is proven by the expanding the equation.*

Proof. Now, we expand the normal distribution:

$$\begin{aligned}
&\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\
&\qquad = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1} - \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1})(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\
&\qquad = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}\Big(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\Big)\right\} \\
&\qquad = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right\} \exp\left\{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right\}
\end{aligned}$$

Now, we consider the quadratic $\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}$, as we have:

$$\text{Tr}(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}) = \text{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\boldsymbol{x}^T) = \text{Vec}(\boldsymbol{\Sigma}^{-1})^T\,\text{Vec}(\boldsymbol{x}\boldsymbol{x}^T)$$

And, so we have the following:

$$\begin{aligned}
&\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right\}\exp\left\{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right\}\\
&\qquad=\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right\}\exp\left\{-\frac{1}{2}\,\text{Vec}(\boldsymbol{\Sigma}^{-1})^T\,\text{Vec}(\boldsymbol{x}\boldsymbol{x}^T)+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right\}
\end{aligned}$$

and so we have proven the proposition. □

Remark 13. Now, we can consider the iid observations $\{\boldsymbol{x}_i\}_{i=1}^N$ of exponential family and we have:

$$\prod_{i=1}^N\left[f(\boldsymbol{x}_i)g(\boldsymbol{\theta})\exp\left(\boldsymbol{\phi}(\boldsymbol{\theta})^T\boldsymbol{T}(\boldsymbol{x}_i)\right)\right]=g(\boldsymbol{\theta})^N\left(\prod_{i=1}^N f(\boldsymbol{x}_i)\right)\exp\left(\sum_{i=1}^N\boldsymbol{\phi}(\boldsymbol{\theta})^T\boldsymbol{T}(\boldsymbol{x}_i)\right)$$

Definition 1.5. (Conjugate Prior) The conjugate prior of the exponential family is the probability distribution of the form of:

$$p(\boldsymbol{\theta}|\boldsymbol{\tau},\nu)=f(\boldsymbol{\tau},\nu)g(\boldsymbol{\theta})^\nu\exp\left(\boldsymbol{\phi}(\boldsymbol{\theta})^T\boldsymbol{\tau}\right)$$

where $\nu>0$ and $\boldsymbol{\tau}\in\mathbb{R}^m$. It is designed so that the posterior given this prior will have the same distribution as the conjugate prior.

Remark 14. The conjugate prior allow use to find the posterior with ease as we don't have to find the normalization of the Bayes' theorem:

$$\begin{aligned}
p(\boldsymbol{\theta}|\mathcal{D})&\propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\tau},\nu)\propto g(\boldsymbol{\theta})^{N+\nu}\exp\left(\boldsymbol{\phi}(\boldsymbol{\theta})^T\left(\boldsymbol{\tau}+\sum_{i=1}^N\boldsymbol{T}(\boldsymbol{x}_i)\right)\right)\\
&=F\left(\boldsymbol{\tau}+\sum_{i=1}^N T(\boldsymbol{x}_i),\ N+\nu\right)g(\boldsymbol{\theta})^{N+\nu}\exp\left(\boldsymbol{\phi}(\boldsymbol{\theta})^T\left(\boldsymbol{\tau}+\sum_{i=1}^N\boldsymbol{T}(\boldsymbol{x}_i)\right)\right)
\end{aligned}$$

And so we have $\boldsymbol{\tau}$ quantify the pseudo-observations via the sufficient statistics. Furthermore, ν is the pseudo-count of the pseudo-observation (can also be seen as the weight of prior belief).

Remark 15. Now, we have the following list of conjugate prior, where it is shown in the table below:

Distribution	**Conjugate**	**Distribution (2)**	**Conjugate (2)**
Bernoulli	Beta	Multinormal (unknown $\boldsymbol{\mu}$)	Multinormal
Poisson	Gamma	Multinormal (unknown $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1}$)	Wishart
Multinormal	Dirichlet	Multinormal (unknown $\boldsymbol{\Lambda}$)	Normal-Wishart
Normal (unknown μ)	Normal		
Normal (unknown $\tau=\sigma^{-1}$)	Gamma		
Normal (unknown μ,τ)	Normal-Gamma		

Definition 1.6. (Normal-Gamma Distribution) The Normal-Gamma is defined as:

$$p(x,\tau|\mu,\lambda,\alpha,\beta)=\frac{\beta^\alpha\sqrt{\lambda}}{\Gamma(\alpha)\sqrt{2\pi}}\tau^{\alpha-1/2}\exp(-\beta\tau)\exp\left(-\frac{\lambda\tau(x-\mu)^2}{2}\right)$$

similarly the Normal-Wishart distribution is given by $\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\lambda\boldsymbol{\Lambda})^{-1})\mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{W}, \nu)$, where $\mathcal{W}$ is the Wishart distribution, where we consider $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \text{NW}(\boldsymbol{\mu}_0, \lambda, \boldsymbol{W}, \mu)$, given as:

$$\begin{aligned}
&\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\lambda\boldsymbol{\Lambda})^{-1})\mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{W}, \nu) \\
&\quad = B(\boldsymbol{W}, \nu)\,|\boldsymbol{\Lambda}|^{(\nu-D-1)/2} \exp\left(-\frac{1}{2}\text{Tr}(\boldsymbol{W}^{-1}\boldsymbol{\Lambda})\right) \frac{1}{\sqrt{|2\pi(\lambda\boldsymbol{\Lambda})^{-1}|}} \exp\left\{-\frac{\lambda}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{\Lambda}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right\} \\
&\quad = \frac{B(\boldsymbol{W}, \nu)\,|\boldsymbol{\Lambda}|^{(\nu-D-1)/2}}{\sqrt{|2\pi(\lambda\boldsymbol{\Lambda})^{-1}|}} \exp\left(-\frac{1}{2}\,\text{Tr}\left[\boldsymbol{W}^{-1}\boldsymbol{\Lambda}\right] - \frac{\lambda}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{\Lambda}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\right)
\end{aligned}$$

1.6 Everything You Always Wanted to Know About Gaussian But Were Afraid to Ask..

Proposition 1.10. ***(Gaussian Integration)*** *We can show that:*

$$\int_{-\infty}^{\infty} \exp(-x^2)\ \mathrm{d}x = \sqrt{\pi}$$

Remark 16. **(Shape of Gaussian)** Recalling the definition of multivariate Gaussian distribution:

$$\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

where we will define the Mahalanobis distance of relating to the Gaussian to be $\Delta^2 = (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$. Now, it is clear that the shape of the Gaussian depends on this value. Now, using the eigendecomposition on the covariance function, we have:

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i}\boldsymbol{u}_i\boldsymbol{u}_i^T \implies \Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$$

where $(\lambda_i, \boldsymbol{u}_i)$ is the eigenvalue/eigenvector pair of $\boldsymbol{\Sigma}$, and $y_i = \boldsymbol{u}_i^T(\boldsymbol{x}-\boldsymbol{\mu})$. Now for the full vector $\boldsymbol{y}$ is equal to $U(\boldsymbol{x}-\boldsymbol{\mu})$. Therefore, the shape of Gaussian is charaterized as:

- Ellipsoids with the center at $\boldsymbol{\mu}$
- Axis is in the direction of eigenvector $\boldsymbol{u}_i$.
- Scaling of each direction is the eigenvector λ_i associated with $\boldsymbol{u}_i$

Remark 17. **(Normalizing Factor of Gaussian)** To show that the multivariate Gaussian indeed is normalized, we consider a change of coordinate to the eigen-basis consider above. To do this, we consider Jacobina matrix $\boldsymbol{J}$ as:

$$J_{ij} = \frac{\partial x_i}{\partial y_i} = U_{ji}$$

Since $\boldsymbol{U}$ is orthonormal, we have: $|\boldsymbol{J}|^2 = |\boldsymbol{U}^T|^2 = |\boldsymbol{U}^T||\boldsymbol{U}| = |\boldsymbol{U}^T\boldsymbol{U}| = |\boldsymbol{I}| = 1$. Now, using the definition of determinant: $|\boldsymbol{\Sigma}|^{1/2} = \prod_{i=1}^{D}|\lambda_j|^{1/2}$. And, so one can perform the integration over as:

$$\int p(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} = \int p(\boldsymbol{y})|\boldsymbol{J}|\ \mathrm{d}\boldsymbol{y} = \int \prod_{i=1}^{D} \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left\{-\frac{y_i^2}{2\lambda_i}\right\}\ \mathrm{d}\boldsymbol{y} = \prod_{i=1}^{D} \int \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left\{-\frac{y_i^2}{2\lambda_i}\right\}\ \mathrm{d}y_i = 1$$

The final equality integration follows from the Gaussian integration.

Proposition 1.11. *Consider the following multivariate Gaussian distribution:*

$$\begin{pmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \right)$$

where $\boldsymbol{\Lambda}^{-1} = \boldsymbol{\Sigma}$*, then we can show that:*

$$p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{a|b}, \Lambda_{aa}^{-1}) \qquad \textit{where} \qquad \boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)$$

One can simply the equaltion by consider the value $\boldsymbol{K} = \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}$ *and since* $\boldsymbol{\Sigma}_{ab} = \boldsymbol{\Sigma}_{ba}$*, then we have:*

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{K}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{K}\boldsymbol{\Sigma}_{bb}\boldsymbol{K}^T \\ &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \end{aligned}$$

The above equation follows from the block-matrix inverse.

Proof. We will consider only the quadratic term inside Gaussian, as we have:

$$\begin{aligned} -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = \\ -\frac{1}{2}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{aa}(\boldsymbol{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \\ -\frac{1}{2}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{bb}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \end{aligned}$$

Now, we are interested to find the condition distribution as it will have the form of:

$$-\frac{1}{2}(\boldsymbol{x}_a - \boldsymbol{\mu}_{a|b})^T\boldsymbol{\Sigma}_{a|b}^{-1}(\boldsymbol{x}_a - \boldsymbol{\mu}_{a|b}) = \underbrace{-\frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Sigma}_{a|b}^{-1}\boldsymbol{x}_a}_{①} + \underbrace{\boldsymbol{x}_a^T\boldsymbol{\Sigma}_{a|b}^{-1}\boldsymbol{\mu}_{a|b}}_{②} + \text{const}$$

Now, let's consider the term for full Gaussian that have the form that matches the condition distribution:

1. The first one is simple as we have: $-\frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a$, and so, one can conclude that $\boldsymbol{\Sigma}_{a|b}^{-1} = \boldsymbol{\Lambda}_{aa}$.
2. For the second term, we consider equation with $\boldsymbol{x}_a^T(\cdots)$, as we have (from term 2 and 3):

$$\begin{aligned} &\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{x}_b + \frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b + \frac{1}{2}\boldsymbol{\mu}_b^T\boldsymbol{\Lambda}_{ba}\boldsymbol{x}_a - \frac{1}{2}\boldsymbol{x}_b^T\boldsymbol{\Lambda}_{ba}\boldsymbol{x}_a \\ &= \boldsymbol{x}_a^T\Big[\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)\Big] \end{aligned}$$

   For the second equality, we use the fact that $\boldsymbol{\Lambda}_{ba}^T = \boldsymbol{\Lambda}_{ab}$. Now we simply apply the inverse of $\boldsymbol{\Lambda}_{aa}$ to get the mean, which means:

$$\begin{aligned} \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_{a|b} &= \boldsymbol{x}_a^T\Big[\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)\Big] \\ \implies \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \end{aligned}$$

   as required.

Please note that one can use the block inverse above to calculate $\boldsymbol{\Lambda}_{aa}$ and $\boldsymbol{\Lambda}_{ab}$ in terms of $\boldsymbol{\Sigma}$. □

Proposition 1.12. *Consider the following multivariate Gaussian distribution:*

$$\begin{pmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \right)$$

where $\boldsymbol{\Lambda}^{-1} = \boldsymbol{\Sigma}$*, then we can show that:* $p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$

Proof. We will use the full expansion of Gaussian like above proof. However, we will consider the term with $\boldsymbol{x}_b$, first as we have (similar to the conditional case):

$$
\begin{aligned}
&-\frac{1}{2}\boldsymbol{x}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{x}_b + \boldsymbol{x}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \frac{1}{2}\boldsymbol{x}_b^T\boldsymbol{\Lambda}_{ba}\boldsymbol{x}_a + \frac{1}{2}\boldsymbol{x}_b^T\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a - \frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{x}_b + \frac{1}{2}\boldsymbol{\mu}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{x}_b \\
&\qquad = -\frac{1}{2}\boldsymbol{x}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{x}_b + \boldsymbol{x}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{x}_b^T\boldsymbol{\Lambda}_{ba}\boldsymbol{x}_a + \boldsymbol{x}_b^T\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a \\
&\qquad = -\frac{1}{2}\Big[\boldsymbol{x}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{x}_b + 2\boldsymbol{x}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{\Lambda}_{bb}^{-1}\underbrace{\big(\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)\big)}_{\boldsymbol{m}}\Big] \\
&\qquad = -\frac{1}{2}\Big[\boldsymbol{x}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{x}_b - 2\boldsymbol{x}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m} + (\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m})^T\boldsymbol{\Lambda}_{bb}(\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m})\Big] + \frac{1}{2}\boldsymbol{m}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m} \\
&\qquad = -\frac{1}{2}(\boldsymbol{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m})^T\boldsymbol{\Lambda}_{bb}(\boldsymbol{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m}) + \frac{1}{2}\boldsymbol{m}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m}
\end{aligned}
$$

Now, we can integrate out the quantities (that depends on $\boldsymbol{x}_b$) i.e:

$$
\int \exp\left\{-\frac{1}{2}(\boldsymbol{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m})^T\boldsymbol{\Lambda}_{bb}(\boldsymbol{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m})\right\}\,\mathrm{d}\boldsymbol{x}_b
$$

Since it is a Gaussian integration, we didn't have to perform any thing further as this would yields similar value for normalization factor. Now, consider the vales related to $\boldsymbol{x}_a$ (without $\boldsymbol{x}_b$) and the leftout value from above:

$$
\begin{aligned}
&\frac{1}{2}\boldsymbol{m}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m} - \frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a + \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b + \frac{1}{2}\boldsymbol{\mu}_b^T\boldsymbol{\Lambda}_{ba}\boldsymbol{x}_a \\
&\qquad = \frac{1}{2}\Big[\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)\Big]^T\boldsymbol{\Lambda}_{bb}^{-1}\Big[\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)\Big] \\
&\qquad\quad - \frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a + \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b \\
&\qquad = \frac{1}{2}\Big[\boldsymbol{\mu}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - (\boldsymbol{x}_a - \boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{ba}^T\boldsymbol{\mu}_b \\
&\qquad\quad - \boldsymbol{\mu}_b^T\boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a - \boldsymbol{\mu}_a) + (\boldsymbol{x}_a - \boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{ba}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)\Big] \\
&\qquad\quad - \frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a + \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b \\
&\qquad = \frac{1}{2}\Big[\boldsymbol{\mu}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ba}^T\boldsymbol{\mu}_b + \boldsymbol{\mu}_a^T\boldsymbol{\Lambda}_{ba}^T\boldsymbol{\mu}_b - \boldsymbol{\mu}_b^T\boldsymbol{\Lambda}_{ba}\boldsymbol{x}_a + \boldsymbol{\mu}_b^T\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a \\
&\qquad\quad + \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ba}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{x}_a - \boldsymbol{\mu}_a^T\boldsymbol{\Lambda}_{ba}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{x}_a - \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ba}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a + \boldsymbol{\mu}_a^T\boldsymbol{\Lambda}_{ba}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a\Big] \\
&\qquad\quad - \frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a + \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b \\
&\qquad = \frac{1}{2}\Big[-2\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ba}^T\boldsymbol{\mu}_b + \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ba}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{x}_a - 2\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ba}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a\Big] \\
&\qquad\quad - \frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a + \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b + \text{ const} \\
&\qquad = \frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ba}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{x}_a - \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{ba}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a - \frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a + \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \text{ const} \\
&\qquad = -\frac{1}{2}\boldsymbol{x}_a^T\Big[\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ba}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\Big]\boldsymbol{x}_a + \boldsymbol{x}_a^T\Big[\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ba}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\Big]\boldsymbol{\mu}_a + \text{ const}
\end{aligned}
$$

Now, we can compare this to the form:

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

and we have $\boldsymbol{\Sigma}_a^*$ (marginalized) is equal to $(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ba}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})^{-1}$, and so we have $\boldsymbol{\mu}_a^* = \boldsymbol{\mu}_a$. Furthermore, from the partition inverse of $(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ba}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})^{-1} = \boldsymbol{\Sigma}_{aa}$. Thus, we completed the proof. □

Proposition 1.13. ***(Linear Gaussian Model)*** *Consider the following distributions:*

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \qquad p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1})$$

Now, we can show that the marginal distribution and the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$ is given by:

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T) \qquad p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{x} \middle| \boldsymbol{\Sigma}\left\{\boldsymbol{A}^T\boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}\right)$$

where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})^{-1}$

Proof. Let's consider the joint distribution first, where we denote $\boldsymbol{z} = \begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}$, and consider the inside of exponential as we have:

$$\begin{aligned}
&-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{x} - \boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T\boldsymbol{L}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) + \text{ const} \\
&\quad = -\frac{1}{2}\Big[\boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu}\Big] \\
&\qquad\quad - \frac{1}{2}\Big[\boldsymbol{y}^T\boldsymbol{L}\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{y} - \boldsymbol{b}^T\boldsymbol{L}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{L}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{L}\boldsymbol{A}\boldsymbol{x} \\
&\qquad\quad - \boldsymbol{y}^T\boldsymbol{L}\boldsymbol{b} + \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{L}\boldsymbol{b}\Big] + \text{ const} \\
&\quad = -\frac{1}{2}\Big[\boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{\mu} + \boldsymbol{y}^T\boldsymbol{L}\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{y} - \boldsymbol{b}^T\boldsymbol{L}\boldsymbol{y} \\
&\qquad\quad - \boldsymbol{y}^T\boldsymbol{L}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{L}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}^T\boldsymbol{L}\boldsymbol{b} + \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{b}\Big] + \text{ const} \\
&\quad = -\frac{1}{2}\Big[\boldsymbol{x}^T\Big(\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A}\Big)\boldsymbol{x} + \boldsymbol{y}^T\boldsymbol{L}\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{L}\boldsymbol{A}\boldsymbol{x} \\
&\qquad\quad - 2\boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{\mu} - 2\boldsymbol{b}^T\boldsymbol{L}\boldsymbol{y} + 2\boldsymbol{b}^T\boldsymbol{L}\boldsymbol{A}\boldsymbol{x}\Big] + \text{ const} \\
&\quad = -\frac{1}{2}\boldsymbol{x}^T\Big(\boldsymbol{\Lambda} - \frac{1}{2}\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A}\Big)\boldsymbol{x} - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{L}\boldsymbol{y} + \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{L}\boldsymbol{A}\boldsymbol{x} \\
&\qquad + \boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{\mu} + \boldsymbol{b}^T\boldsymbol{L}\boldsymbol{y} - \boldsymbol{b}^T\boldsymbol{L}\boldsymbol{A}\boldsymbol{x} + \text{ const} \\
&\quad = -\frac{1}{2}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}^T\begin{pmatrix}\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{L}\\ -\boldsymbol{L}\boldsymbol{A} & \boldsymbol{L}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix} + \begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}^T\begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu} - \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{b}\\ \boldsymbol{L}\boldsymbol{b}\end{pmatrix} + \text{ const}
\end{aligned}$$

Now, using the same pattern matching, we can see that the covariance of $\boldsymbol{z}$ is equal to:

$$\begin{pmatrix}\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{L}\\ -\boldsymbol{L}\boldsymbol{A} & \boldsymbol{L}\end{pmatrix}^{-1} = \begin{pmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T\\ \boldsymbol{A}\boldsymbol{\Lambda}^{-1} & \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T\end{pmatrix}$$

Similarly, the mean is equal to:

$$\begin{pmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T\\ \boldsymbol{A}\boldsymbol{\Lambda}^{-1} & \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T\end{pmatrix}\begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu} - \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{b}\\ \boldsymbol{L}\boldsymbol{b}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}\end{pmatrix}$$

Now, we can use results above (conditional and marginalized) to get the result. □

Remark 18. **(Conjugate Prior of Multinormal)** The proof follows from here. We are now consider the likelihood of Multinormal distribution given the dataset:

$$\begin{aligned}
&\prod_{i=1}^{N}\frac{1}{\sqrt{\left|2\pi\boldsymbol{\Lambda}^{-1}\right|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{x}_i - \boldsymbol{\mu})\right\} \\
&\qquad = \frac{1}{\left|2\pi\boldsymbol{\Lambda}^{-1}\right|^{N/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{x}_i - \boldsymbol{\mu})\right\} \\
&\qquad \propto |\boldsymbol{\Lambda}|^{N/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{x}_i - \boldsymbol{\mu})\right\}
\end{aligned}$$

Now, we will consider the Normal-Wishart distribution in the similar form as:

$$\frac{B(\boldsymbol{W},\nu)\,|\boldsymbol{\Lambda}|^{(\nu-D-1)/2}}{\sqrt{|2\pi(\lambda\boldsymbol{\Lambda})^{-1}|}}\exp\left(-\frac{1}{2}\operatorname{Tr}\left[\boldsymbol{W}^{-1}\boldsymbol{\Lambda}\right]-\frac{\lambda}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{\Lambda}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\right)$$
$$\propto |\boldsymbol{\Lambda}|^{(\nu-D)/2}\exp\left(-\frac{1}{2}\operatorname{Tr}\left[\boldsymbol{W}^{-1}\boldsymbol{\Lambda}\right]\right)\exp\left(-\frac{\lambda}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{\Lambda}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\right)$$

Now, we simply have to multiply the conjugate prior and the normal distribution, which gives us:

$$\begin{aligned}
&|\boldsymbol{\Lambda}|^{N/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{x}_i-\boldsymbol{\mu})\right\}\\
&\qquad |\boldsymbol{\Lambda}|^{(\nu-D)/2}\exp\left(-\frac{1}{2}\operatorname{Tr}\left[\boldsymbol{W}^{-1}\boldsymbol{\Lambda}\right]\right)\exp\left(-\frac{\lambda}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{\Lambda}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\right)\\
&=|\boldsymbol{\Lambda}|^{(\nu-D+N)/2}\exp\left(-\frac{1}{2}\operatorname{Tr}\left[\boldsymbol{W}^{-1}\boldsymbol{\Lambda}\right]\right)\\
&\qquad\qquad \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{N}\left(\boldsymbol{x}_i^T\boldsymbol{\Lambda}\boldsymbol{x}_i\right)+N\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu}-2N\bar{\boldsymbol{x}}^T\boldsymbol{\Lambda}\boldsymbol{\mu}\right]\right\}\\
&\qquad\qquad \exp\left\{-\frac{\lambda}{2}\left[\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu}-2\boldsymbol{\mu}_0^T\boldsymbol{\Lambda}\boldsymbol{\mu}+\boldsymbol{\mu}_0^T\boldsymbol{\Lambda}\boldsymbol{\mu}_0\right]\right\}\\
&=|\boldsymbol{\Lambda}|^{(\nu-D+N)/2}\exp\left(-\frac{1}{2}\left[\operatorname{Tr}\left[\boldsymbol{W}^{-1}\boldsymbol{\Lambda}\right]+\sum_{i=1}^{N}\left(\boldsymbol{x}_i^T\boldsymbol{\Lambda}\boldsymbol{x}_i\right)+(N+\lambda)\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu}\right.\right.\\
&\qquad\qquad\left.\left.-2(N\bar{\boldsymbol{x}}^T+\lambda\boldsymbol{\mu}_0^T)\boldsymbol{\Lambda}\boldsymbol{\mu}+\lambda\boldsymbol{\mu}_0^T\boldsymbol{\Lambda}\boldsymbol{\mu}_0\right]\right)
\end{aligned}$$

Let's consider exclusively for the expression in the square bracket:

$$\begin{aligned}
&\operatorname{Tr}\left[\boldsymbol{W}^{-1}\boldsymbol{\Lambda}\right]+\sum_{i=1}^{N}\left(\boldsymbol{x}_i^T\boldsymbol{\Lambda}\boldsymbol{x}_i\right)+\lambda\boldsymbol{\mu}_0^T\boldsymbol{\Lambda}\boldsymbol{\mu}_0+(N+\lambda)\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu}-2(N\bar{\boldsymbol{x}}^T+\lambda\boldsymbol{\mu}_0^T)\boldsymbol{\Lambda}\boldsymbol{\mu}\\
&=\operatorname{Tr}\left[\boldsymbol{W}^{-1}\boldsymbol{\Lambda}\right]+\sum_{i=1}^{N}\left(\boldsymbol{x}_i^T\boldsymbol{\Lambda}\boldsymbol{x}_i\right)+\lambda\boldsymbol{\mu}_0^T\boldsymbol{\Lambda}\boldsymbol{\mu}_0-\frac{1}{\lambda+N}(\lambda\boldsymbol{\mu}_0+N\bar{\boldsymbol{x}})^T\boldsymbol{\Lambda}(\lambda\boldsymbol{\mu}_0+N\bar{\boldsymbol{x}})\\
&\qquad+(N+\lambda)\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu}-2\boldsymbol{\mu}^T\boldsymbol{\Lambda}(N\bar{\boldsymbol{x}}+\lambda\boldsymbol{\mu}_0)+\frac{1}{\lambda+N}(\lambda\boldsymbol{\mu}_0+N\bar{\boldsymbol{x}})^T\boldsymbol{\Lambda}(\lambda\boldsymbol{\mu}_0+N\bar{\boldsymbol{x}})
\end{aligned}$$

Please note that the blue part, whichis equal to:

$$(N+\lambda)\left(\boldsymbol{\mu}-\frac{\lambda\boldsymbol{\mu}_0+N\bar{\boldsymbol{x}}}{\lambda+N}\right)^T\boldsymbol{\Lambda}\left(\boldsymbol{\mu}-\frac{\lambda\boldsymbol{\mu}_0+N\bar{\boldsymbol{x}}}{\lambda+N}\right)$$

Now for the red part, we consider adding and substracting $2N\bar{\boldsymbol{x}}^T\Lambda\bar{\boldsymbol{x}}$, which we have:

$$\begin{aligned}
&\sum_{i=1}^{N}\left(\boldsymbol{x}_i^T\boldsymbol{\Lambda}\boldsymbol{x}_i\right)-2\sum_{i=1}^{N}\boldsymbol{x}_i^T\Lambda\bar{\boldsymbol{x}}+\sum_{i=1}^{N}\bar{\boldsymbol{x}}\Lambda\bar{\boldsymbol{x}}+N\bar{\boldsymbol{x}}^T\Lambda\bar{\boldsymbol{x}}\\
&\qquad=\sum_{i=1}^{N}\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}\right)^T\boldsymbol{\Lambda}\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}\right)+N\bar{\boldsymbol{x}}^T\Lambda\bar{\boldsymbol{x}}
\end{aligned}$$

For the brown equation (together with $N\bar{\boldsymbol{x}}^T\Lambda\bar{\boldsymbol{x}}$), we have:

$$\begin{aligned}
&\lambda\boldsymbol{\mu}_0^T\boldsymbol{\Lambda}\boldsymbol{\mu}_0 - \frac{1}{\lambda+N}(\lambda\boldsymbol{\mu}_0 + N\bar{\boldsymbol{x}})^T\boldsymbol{\Lambda}(\lambda\boldsymbol{\mu}_0 + N\bar{\boldsymbol{x}}) + N\bar{\boldsymbol{x}}^T\Lambda\bar{\boldsymbol{x}} \\
&= \lambda\boldsymbol{\mu}_0^T\boldsymbol{\Lambda}\boldsymbol{\mu}_0 + N\bar{\boldsymbol{x}}^T\Lambda\bar{\boldsymbol{x}} - \frac{1}{\lambda+N}\left[\lambda^2\boldsymbol{\mu}_0^T\Lambda\boldsymbol{\mu}_0 + 2N\lambda\bar{\boldsymbol{x}}^T\boldsymbol{\Lambda}\boldsymbol{\mu}_0 + N^2\bar{\boldsymbol{x}}^T\boldsymbol{\Lambda}\bar{\boldsymbol{x}}\right] \\
&= \frac{\lambda+N}{\lambda+N}\lambda\boldsymbol{\mu}_0^T\boldsymbol{\Lambda}\boldsymbol{\mu}_0 + \frac{\lambda+N}{\lambda+N}N\bar{\boldsymbol{x}}^T\Lambda\bar{\boldsymbol{x}} - \frac{1}{\lambda+N}\left[\lambda^2\boldsymbol{\mu}_0^T\Lambda\boldsymbol{\mu}_0 + 2N\lambda\bar{\boldsymbol{x}}^T\boldsymbol{\Lambda}\boldsymbol{\mu}_0 + N^2\bar{\boldsymbol{x}}^T\boldsymbol{\Lambda}\bar{\boldsymbol{x}}\right] \\
&= \frac{1}{\lambda+N}\left[\lambda^2\boldsymbol{\mu}_0^T\boldsymbol{\Lambda}\boldsymbol{\mu}_0 + N\lambda\boldsymbol{\mu}_0^T\boldsymbol{\Lambda}\boldsymbol{\mu}_0 + N\lambda\bar{\boldsymbol{x}}^T\Lambda\bar{\boldsymbol{x}} + N^2\bar{\boldsymbol{x}}^T\Lambda\bar{\boldsymbol{x}}\right] \\
&\qquad - \frac{1}{\lambda+N}\left[\lambda^2\boldsymbol{\mu}_0^T\Lambda\boldsymbol{\mu}_0 + 2N\lambda\bar{\boldsymbol{x}}^T\boldsymbol{\Lambda}\boldsymbol{\mu}_0 + N^2\bar{\boldsymbol{x}}^T\boldsymbol{\Lambda}\bar{\boldsymbol{x}}\right] \\
&= \frac{1}{\lambda+N}\left[N\lambda\boldsymbol{\mu}_0^T\boldsymbol{\Lambda}\boldsymbol{\mu}_0 + N\lambda\bar{\boldsymbol{x}}^T\Lambda\bar{\boldsymbol{x}} + 2N\lambda\bar{\boldsymbol{x}}^T\boldsymbol{\Lambda}\boldsymbol{\mu}_0\right] \\
&= \frac{N\lambda}{\lambda+N}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T\boldsymbol{\Lambda}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)
\end{aligned}$$

Combining the first part (black, red, and brown), and together with the trace trick:

$$\mathrm{Tr}\left(\boldsymbol{W}^{-1}\boldsymbol{\Lambda} + \sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\boldsymbol{\Lambda} + \frac{N\lambda}{\lambda+N}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T\boldsymbol{\Lambda}\right)$$

Now, the distribution becomes:

$$\begin{aligned}
|\boldsymbol{\Lambda}|^{(\nu-D+N)/2}\exp\Bigg(&-\frac{1}{2}\mathrm{Tr}\left(\left[\boldsymbol{W}^{-1} + \sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T + \frac{N\lambda}{\lambda+N}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T\right]\boldsymbol{\Lambda}\right) \\
&- \frac{N+\lambda}{2}\left(\boldsymbol{\mu} - \frac{\lambda\boldsymbol{\mu}_0 + N\bar{\boldsymbol{x}}}{\lambda+N}\right)^T\boldsymbol{\Lambda}\left(\boldsymbol{\mu} - \frac{\lambda\boldsymbol{\mu}_0 + N\bar{\boldsymbol{x}}}{\lambda+N}\right)\Bigg)
\end{aligned}$$

Which is simply a Normal-Wishart Distribution as required.

Proposition 1.14. ***(Maximum Likelihood of Mean)*** *We can show that for Gaussian distribution, the maximum likelihood of $\boldsymbol{\mu}$, given the dataset $\{\boldsymbol{x}_i\}_{i=1}^N$ is*

$$\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$$

Proof. Starting with the log-likelihood to be:

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log\prod_{i=1}^{N}\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{N}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})$$

And, so we can consider the derivative over $\boldsymbol{\mu}$ as:

$$\begin{aligned}
\frac{\partial(-l)}{\partial\boldsymbol{\mu}} &= \frac{\partial}{\partial\boldsymbol{\mu}}\left[\frac{N}{2}\log|2\pi\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right] \\
&= \frac{1}{2}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\mu}}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) \\
&= \frac{1}{2}\sum_{i=1}^{N}\left(\frac{\partial}{\partial\boldsymbol{\mu}}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right] - 2\frac{\partial}{\partial\boldsymbol{\mu}}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i\right]\right) = N\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}\boldsymbol{x}_i
\end{aligned}$$

Setting the derivative to zero and we yields the result. □

Proposition 1.15. ***(Maximum Likelihood of Covariance)*** *We can show that for Gaussian distribution, the maximum likelihood of* $\boldsymbol{\Sigma}$*, given the dataset* $\{\boldsymbol{x}_i\}_{i=1}^N$ *is*

$$\hat{\boldsymbol{\Sigma}}_{ML} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

Proof. We consider the derivative over $\boldsymbol{\Sigma}^{-1}$ as (please note that we have to constraint of the covariance to be positive definite but it turn out we don't have to as the result already satisfies the constriant):

$$\begin{aligned}
\frac{\partial(-l)}{\partial\boldsymbol{\Sigma}^{-1}} &= \frac{\partial}{\partial\boldsymbol{\Sigma}^{-1}}\left[\frac{N}{2}\log|2\pi\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right] \\
&= -\frac{\partial}{\partial\boldsymbol{\Sigma}^{-1}}\left[\frac{N}{2}\log\left|\boldsymbol{\Sigma}^{-1}\right|\right] + \frac{1}{2}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\Sigma}^{-1}}\left[(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right] \\
&= -\frac{N}{2}\boldsymbol{\Sigma}^T + \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^T
\end{aligned}$$

Setting the derivative to zero and we yields the result. □

Definition 1.7. (Linear Regression) Now, given the data: $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y})\}_{i=1}^N$. We have $\boldsymbol{y}_i$ is conditionally independent given $\boldsymbol{x}_i$. Now, we consider the supervised learning as we have a linear function $\boldsymbol{x}$ together with Gaussian noise:

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{\Sigma}_y) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_y|}}\exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{x})^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{x})\right\}$$

Proposition 1.16. ***(Maximum Likelihood of Linear Regression)*** *The maximum likelihood is given by the following:*

$$\hat{\boldsymbol{W}} = \sum_{i=1}^{N}\boldsymbol{y}_i\boldsymbol{x}_i^T\left(\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1}$$

Proof. Consider the log-likelihood of the linear regression:

$$\begin{aligned}
l &= \log\prod_{i=1}^{N}\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_y|}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_i)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_i)\right\} \\
&= \sum_{i=1}^{N} -\frac{1}{2}\log|2\pi\boldsymbol{\Sigma}_y| - \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_i)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_i) \\
&= -\frac{N}{2}\log|2\pi\boldsymbol{\Sigma}_y| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_i)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_i)
\end{aligned}$$

We consider the derivative, as we have:

$$\begin{aligned}
\frac{\partial(-l)}{\partial\boldsymbol{W}} &= \frac{\partial}{\partial\boldsymbol{W}}\left[\frac{N}{2}\log|2\pi\boldsymbol{\Sigma}_y| + \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_i)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_i)\right] \\
&= \frac{1}{2}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{W}}\left[(\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_i)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_i)\right] \\
&= \frac{1}{2}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{W}}\left[\boldsymbol{y}_i^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{y}_i + \boldsymbol{x}_i^T\boldsymbol{W}^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{W}\boldsymbol{x}_i - 2\boldsymbol{x}_i^T\boldsymbol{W}^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{y}_i\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{i=1}^{N}\left[\frac{\partial}{\partial \boldsymbol{W}}\operatorname{Tr}[\boldsymbol{x}_i^T\boldsymbol{W}^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{W}\boldsymbol{x}_i] - 2\frac{\partial}{\partial \boldsymbol{W}}\operatorname{Tr}[\boldsymbol{x}_i^T\boldsymbol{W}^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{y}_i]\right] \\
&= \frac{1}{2}\sum_{i=1}^{N}\left[\frac{\partial}{\partial \boldsymbol{W}}\operatorname{Tr}[\boldsymbol{W}^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{W}\boldsymbol{x}_i\boldsymbol{x}_i^T] - 2\frac{\partial}{\partial \boldsymbol{W}}\operatorname{Tr}[\boldsymbol{W}^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{y}_i\boldsymbol{x}_i^T]\right] \\
&= \frac{1}{2}\sum_{i=1}^{N}\left[2\boldsymbol{\Sigma}_y^{-1}\boldsymbol{W}\boldsymbol{x}_i\boldsymbol{x}_i^T - 2\boldsymbol{\Sigma}_y^{-1}\boldsymbol{y}_i\boldsymbol{x}_i^T\right]
\end{aligned}
$$

Please noted that we use the derivative of the trace above, and setting the derivative to zero and we have the answer as we wanted. □

Proposition 1.17. *The MAP estimate given prior over the weight $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{-1})$ is:*

$$
\boldsymbol{w}^{MAP} = \underbrace{\left(\boldsymbol{A} + \frac{\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T}{\sigma_y^2}\right)^{-1}}_{\boldsymbol{\Sigma}_w}\frac{\sum_{i=1}^{N}\boldsymbol{y}_i\boldsymbol{x}_i}{\sigma_y^2} = \left(A\sigma_y^2 + \sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1}\sum_{i=1}^{N}y_i\boldsymbol{x}_i
$$

We will denote the a Please note that we consider the prediction in one-dimensional, as we have

$$
p(\mathcal{D}|\left\{\boldsymbol{x}_i\right\}_{i=1}^{N}, \boldsymbol{w}, \sigma_y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left\{-\frac{1}{2\sigma_y^2}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2\right\}
$$

Proof. We start with the log-posterior on $\boldsymbol{w}$ is given as:

$$
\begin{aligned}
\log p(\boldsymbol{w}|\mathcal{D}, A, \sigma_y) &= \log p(\mathcal{D}|\left\{\boldsymbol{x}_i\right\}_{i=1}^{N}, \boldsymbol{w}, \sigma_y) + \log p(\boldsymbol{w}|A) + \text{ const} \\
&= -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w} - \frac{1}{2\sigma_y^2}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2 + \text{ const} \\
&= -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w} - \frac{1}{2\sigma_y^2}\sum_{i=1}^{N}\left[y_i^2 - 2y_i\boldsymbol{w}^T\boldsymbol{x}_i + (\boldsymbol{w}^T\boldsymbol{x}_i)^2\right] + \text{ const} \\
&= -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w} - \frac{1}{2\sigma_y^2}\sum_{i=1}^{N}2y_i\boldsymbol{w}^T\boldsymbol{x}_i - \frac{1}{2}\boldsymbol{w}^T\left(\sigma_y^{-2}\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\boldsymbol{w} + \text{ const} \\
&= -\frac{1}{2}\boldsymbol{w}^T\left(\boldsymbol{A} + \sigma_y^{-2}\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\boldsymbol{w} - \boldsymbol{w}^T\sum_{i=1}^{N}(y_i\boldsymbol{x}_i)\sigma_y^{-2} + \text{ const} \\
&= -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_w\sum_{i=1}^{N}(y_i\boldsymbol{x}_i)\sigma_y^{-2} + \text{ const} \\
&= \log\ \mathcal{N}\left(\boldsymbol{\Sigma}_w\sum_{i=1}^{N}(y_i\boldsymbol{x}_i)\sigma_y^{-2}, \boldsymbol{\Sigma}_w\right)
\end{aligned}
$$

As the Gaussian gives the maximum probability at its mean; therefore, we have completed the prove. □

2 Latent Variable Model

Definition 2.1. (Latent Variable Model) The latent variable model can be seen as:

$$\begin{aligned} \boldsymbol{z} &\sim p(\boldsymbol{\theta}_z) \\ \boldsymbol{x}|\boldsymbol{z} &\sim p(\boldsymbol{\theta}_z) \\ p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}_x, \boldsymbol{\theta}_z) &= p(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\theta}_x) p(\boldsymbol{z}; \boldsymbol{\theta}_z) \\ p(\boldsymbol{x}; \boldsymbol{\theta}_x, \boldsymbol{\theta}_z) &= \int p(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\theta}_x) p(\boldsymbol{z}; \boldsymbol{\theta}_z) \, \mathrm{d}\boldsymbol{z} \end{aligned}$$

Note that $p(\boldsymbol{z}), p(\boldsymbol{x}|\boldsymbol{z})$ and $p(\boldsymbol{x}, \boldsymbol{z})$ are exponential family but $p(\boldsymbol{x})$ doesn't have to be an exponential family.

2.1 PCA Formulation

Remark 19. We will consider the family of PCA formulation. We will start with PCA definition, which can be formulated in 2 ways: Maximal Variance and Average Projection Cost. Then, we will consider the PPCA a probabilistic version of PCA.

Definition 2.2. (Maximal Variance) Consider the dataset $\{\boldsymbol{x}_i\}_{i=1}^N$ where $\boldsymbol{x}_i \in \mathbb{R}^D$. We want to project the data onto space with dimension $M < D$. To do this, we want to find a subspace orthonormal basis $\boldsymbol{u}_i$ for $i = 1, \cdots, M$, such that $\boldsymbol{u}_i^T \boldsymbol{u}_j = \delta_{ij}$. So that the empirical projected variance, given as:

$$\frac{1}{N} \sum_{j=1}^{N} (\boldsymbol{u}_i^T \boldsymbol{x}_j - \boldsymbol{u}_i^T \bar{\boldsymbol{x}})^2$$

is maximized for $i = 1, \ldots, M$ where $\boldsymbol{u}_1$ gives the highest variance, where $\bar{\boldsymbol{x}} = 1/N \sum_{i=1}^N \boldsymbol{x}_i$

Proposition 2.1. *The maximum projected variance direction $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M$ is the M eigenvectors of the data-covariance matrix:*

$$\boldsymbol{S} = \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}_i})^T$$

associated with the following eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M \geq \cdots \geq \lambda_D$.

Proof. Let's start with the first direction $\boldsymbol{u}_1$ as we can show that the empirical projected variance:

$$\frac{1}{N} \sum_{j=1}^{N} (\boldsymbol{u}_1^T \boldsymbol{x}_j - \boldsymbol{u}_1^T \bar{\boldsymbol{x}})^2 = \boldsymbol{u}_1^T \boldsymbol{S} \boldsymbol{u}_1$$

We will consider the following contraint optimization problem as we have the following Lagrange multiplier and set the derivative to be 0:

$$\frac{\partial}{\partial \boldsymbol{u}} \Big[\boldsymbol{u}_1^T \boldsymbol{S} \boldsymbol{u}_1 + \lambda_1 (1 - \boldsymbol{u}_1^T \boldsymbol{u}_1) \Big] = 2\boldsymbol{S}\boldsymbol{u}_1 - 2\lambda_1 \boldsymbol{u}_1 = 0$$

Please note that the matrix $\boldsymbol{S}$ is symmetric. This leads us to the following, equation, which is:

$$\boldsymbol{S}\boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1$$

Furthermore, if we multiply $\boldsymbol{u}_1^T$ on the LHS together with the contraint $\boldsymbol{u}_1^T \boldsymbol{u}_1 = 1$, then we have $\boldsymbol{u}_1^T \boldsymbol{S} \boldsymbol{u}_1 = \lambda_1$. This means that first maximal variance projection direction is the eigenvector $\boldsymbol{u}_1$ that has the highest assocated eigenvalue. Furthermore, the orthogonal comes from the properties of eigenvectors. □

Definition 2.3. (Minimum-Error Formulation) We get the formulation PCA where we have orthonormal set D-dimensional basis vector $\{\boldsymbol{u}_i\}$ where $i = 1, \ldots, D$ and $\boldsymbol{u}_i^T \boldsymbol{u}_j = \delta_{ij}$. This means that the data point can be represented by the basis:

$$\boldsymbol{x}_i = \sum_{j=1}^{D} \alpha_{ij} \boldsymbol{u}_j = \sum_{j=1}^{D} (\boldsymbol{x}_i^T \boldsymbol{u}_j) \boldsymbol{u}_j$$

The value of α_{ij} can be found by the inner product $\alpha_{ij} = \boldsymbol{x}_i^T \boldsymbol{u}_j$ as above (by the orthogonal properties). Now, we will consider the approximation using the projection over linear subspace $M < D$:

$$\tilde{\boldsymbol{x}}_i = \sum_{j=1}^{M} z_{ij} \boldsymbol{u}_j + \sum_{j=M+1}^{D} b_j \boldsymbol{u}_j$$

where the $\{b_j\}$ are component that is same for all data points. Now, we are free to find the $\{b_j\}, \{z_{ij}\}$ and $\{\boldsymbol{u}_j\}$ given the following objective:

$$\frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|^2$$

Proposition 2.2. *The solution to the minimum error formulation is the same as the maximal variance formulation. This gives us the difference interpretation of the PCA.*

Proof. Let's start with finding the value $\{z_{ij}\}$, first. As we want to find derivative of z_{ij} to be zero:

$$\begin{aligned}
&\frac{1}{N} \sum_{i=1}^{N} \left\| \boldsymbol{x}_i - \sum_{j=1}^{M} z_{ij} \boldsymbol{u}_j - \sum_{j=M+1}^{D} b_j \boldsymbol{u}_j \right\|^2 \\
&\quad = \frac{1}{N} \sum_{i=1}^{N} \left[\left(\boldsymbol{x}_i - \sum_{j=1}^{M} z_{ij} \boldsymbol{u}_j - \sum_{j=M+1}^{D} b_j \boldsymbol{u}_j \right)^T \left(\boldsymbol{x}_i - \sum_{j=1}^{M} z_{ij} \boldsymbol{u}_j - \sum_{j=M+1}^{D} b_j \boldsymbol{u}_j \right) \right] \\
&\quad = \frac{1}{N} \sum_{i=1}^{N} \left[\boldsymbol{x}_i^T \boldsymbol{x}_i - \sum_{j=1}^{M} z_{ij} \boldsymbol{u}_j^T \boldsymbol{x}_i - \sum_{j=M+1}^{D} b_j \boldsymbol{u}_j^T \boldsymbol{x}_i - \sum_{j=1}^{M} z_{ij} \boldsymbol{x}_i^T \boldsymbol{u}_j \right. \\
&\qquad\qquad + \left(\sum_{j=1}^{M} z_{ij} \boldsymbol{u}_j^T \right) \left(\sum_{j=1}^{M} z_{ij} \boldsymbol{u}_j \right) + \left(\sum_{j=M+1}^{D} b_j \boldsymbol{u}_j^T \right) \left(\sum_{j=1}^{M} z_{ij} \boldsymbol{u}_j \right) \\
&\qquad\qquad \left. - \sum_{j=M+1}^{D} b_j \boldsymbol{x}_i^T \boldsymbol{u}_j + \left(\sum_{j=1}^{M} z_{ij} \boldsymbol{u}_j^T \right) \left(\sum_{j=M+1}^{D} b_j \boldsymbol{u}_j \right) + \left(\sum_{j=M+1}^{D} b_j \boldsymbol{u}_j^T \right) \left(\sum_{j=M+1}^{D} b_j \boldsymbol{u}_j \right) \right] \\
&\quad = \frac{1}{N} \sum_{i=1}^{N} \left[\boldsymbol{x}_i^T \boldsymbol{x}_i - \sum_{j=1}^{M} z_{ij} \boldsymbol{u}_j^T \boldsymbol{x}_i - \sum_{j=M+1}^{D} b_j \boldsymbol{u}_j^T \boldsymbol{x}_i - \sum_{j=1}^{M} z_{ij} \boldsymbol{x}_i^T \boldsymbol{u}_j - \sum_{j=M+1}^{D} b_j \boldsymbol{x}_i^T \boldsymbol{u}_j \right. \\
&\qquad\qquad + \left(\sum_{a=1}^{M} z_{ia} \boldsymbol{u}_a^T \right) \left(\sum_{b=1}^{M} z_{ib} \boldsymbol{u}_b \right) + \left(\sum_{a=M+1}^{D} b_a \boldsymbol{u}_a^T \right) \left(\sum_{b=1}^{M} z_{ib} \boldsymbol{u}_b \right) \\
&\qquad\qquad \left. + \left(\sum_{a=1}^{M} z_{ia} \boldsymbol{u}_a^T \right) \left(\sum_{b=M+1}^{D} b_b \boldsymbol{u}_b \right) + \left(\sum_{a=M+1}^{D} b_a \boldsymbol{u}_a^T \right) \left(\sum_{b=M+1}^{D} b_b \boldsymbol{u}_b \right) \right] \\
&\quad = \frac{1}{N} \sum_{i=1}^{N} \left[\boldsymbol{x}_i^T \boldsymbol{x}_i - 2 \sum_{j=1}^{M} z_{ij} \boldsymbol{u}_j^T \boldsymbol{x}_i - 2 \sum_{j=M+1}^{D} b_j \boldsymbol{u}_j^T \boldsymbol{x}_i + \sum_{a=1}^{M} \sum_{b=1}^{M} z_{ia} z_{ib} \boldsymbol{u}_a^T \boldsymbol{u}_b \right. \\
&\qquad\qquad \left. + 2 \sum_{a=M+1}^{D} \sum_{b=1}^{M} b_a z_{ib} \boldsymbol{u}_a^T \boldsymbol{u}_b + \sum_{a=M+1}^{D} \sum_{b=M+1}^{D} b_a b_b \boldsymbol{u}_a^T \boldsymbol{u}_b \right]
\end{aligned}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\boldsymbol{x}_i^T\boldsymbol{x}_i - 2\sum_{j=1}^{M} z_{ij}\boldsymbol{u}_j^T\boldsymbol{x}_i - 2\sum_{j=M+1}^{D} b_j\boldsymbol{u}_j^T\boldsymbol{x}_i + \sum_{j=1}^{M} z_{ij}^2 + \sum_{j=M+1}^{D} b_j^2 + 2\cancel{\sum_{a=M+1}^{D}\sum_{b=1}^{M} b_a z_{ib}\boldsymbol{u}_a^T\boldsymbol{u}_b}\right]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left[\boldsymbol{x}_i^T\boldsymbol{x}_i - 2\sum_{j=1}^{M} z_{ij}\boldsymbol{u}_j^T\boldsymbol{x}_i - 2\sum_{j=M+1}^{D} b_j\boldsymbol{u}_j^T\boldsymbol{x}_i + \sum_{j=1}^{M} z_{ij}^2 + \sum_{j=M+1}^{D} b_j^2\right]$$

Now, let's consider the derivative with respected to z_{ab} as we now have:

$$\begin{aligned}
&\frac{\partial}{\partial z_{ab}}\ \frac{1}{N}\sum_{i=1}^{N}\left[\boldsymbol{x}_i^T\boldsymbol{x}_i - 2\sum_{j=1}^{M} z_{ij}\boldsymbol{u}_j^T\boldsymbol{x}_i - 2\sum_{j=M+1}^{D} b_j\boldsymbol{u}_j^T\boldsymbol{x}_i + \sum_{j=1}^{M} z_{ij}^2 + \sum_{j=M+1}^{D} b_j^2\right] \\
&\quad = -\frac{\partial}{\partial z_{ab}}\ \frac{2}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} z_{ij}\boldsymbol{u}_j^T\boldsymbol{x}_i + \frac{\partial}{\partial z_{ab}}\ \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} z_{ij}^2 \\
&\quad = -\frac{2}{N}\boldsymbol{u}_b^T\boldsymbol{x}_a + \frac{2}{N}z_{ab} = 0
\end{aligned}$$

And so, we have the $z_{ij} = \boldsymbol{x}_i^T\boldsymbol{u}_j$ for $j = 1, \ldots, M$. Now, we consider the derivative of the with respected to b_a as we have:

$$\begin{aligned}
&\frac{\partial}{\partial b_a}\ \frac{1}{N}\sum_{i=1}^{N}\left[\boldsymbol{x}_i^T\boldsymbol{x}_i - 2\sum_{j=1}^{M} z_{ij}\boldsymbol{u}_j^T\boldsymbol{x}_i - 2\sum_{j=M+1}^{D} b_j\boldsymbol{u}_j^T\boldsymbol{x}_i + \sum_{j=1}^{M} z_{ij}^2 + \sum_{j=M+1}^{D} b_j^2\right] \\
&\quad = -\frac{\partial}{\partial b_a}\ \frac{2}{N}\sum_{i=1}^{N}\sum_{j=M+1}^{M} b_j\boldsymbol{u}_j^T\boldsymbol{x}_i + \frac{\partial}{\partial b_a}\ \frac{1}{N}\sum_{i=1}^{N}\sum_{j=M+1}^{M} b_j^2 \\
&\quad = -\frac{\partial}{\partial b_a}\ 2\sum_{j=M+1}^{M} b_j\boldsymbol{u}_j^T\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\right) + \frac{\partial}{\partial b_a}\sum_{j=M+1}^{M} b_j^2 \\
&\quad = -\frac{\partial}{\partial b_a}\ 2\sum_{j=M+1}^{M} b_j\boldsymbol{u}_j^T\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\right) + \frac{\partial}{\partial b_a}\sum_{j=M+1}^{M} b_j^2 \\
&\quad = -2\boldsymbol{u}_a^T\bar{\boldsymbol{x}} + 2b_a
\end{aligned}$$

And so, we have $b_j = \bar{\boldsymbol{x}}^T\boldsymbol{u}_j$ for $j = M+1, \ldots, D$. To find the $\boldsymbol{u}_i$, we have the following:

$$\begin{aligned}
\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i &= \boldsymbol{x}_i - \sum_{j=1}^{M} z_{ij}\boldsymbol{u}_j - \sum_{j=M+1}^{D} b_j\boldsymbol{u}_j \\
&= \sum_{j=1}^{M}(\boldsymbol{x}_i^T\boldsymbol{u}_j)\boldsymbol{u}_j + \sum_{j=M+1}^{D}(\boldsymbol{x}_i^T\boldsymbol{u}_j)\boldsymbol{u}_j - \sum_{j=1}^{M}(\boldsymbol{x}_i^T\boldsymbol{u}_j)\boldsymbol{u}_j - \sum_{j=M+1}^{D}(\bar{\boldsymbol{x}}^T\boldsymbol{u}_j)\boldsymbol{u}_j \\
&= \sum_{j=M+1}^{D}(\boldsymbol{x}_i^T\boldsymbol{u}_j)\boldsymbol{u}_j - \sum_{j=M+1}^{D}(\bar{\boldsymbol{x}}^T\boldsymbol{u}_j)\boldsymbol{u}_j \\
&= \sum_{j=M+1}^{D}\left\{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\boldsymbol{u}_j\right\}\boldsymbol{u}_j
\end{aligned}$$

And, so we now have, the following objective:

$$\frac{1}{N}\sum_{i=1}^{N}\left[\left(\sum_{j=M+1}^{D}\left\{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\boldsymbol{u}_j\right\}\boldsymbol{u}_j\right)^T\left(\sum_{j=M+1}^{D}\left\{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\boldsymbol{u}_j\right\}\boldsymbol{u}_j\right)\right]$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N}\left[\left(\sum_{a=M+1}^{D}\boldsymbol{u}_a^T\left\{\boldsymbol{u}_a^T(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\right\}\right)\left(\sum_{b=M+1}^{D}\left\{(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T\boldsymbol{u}_b\right\}\boldsymbol{u}_b\right)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\left(\sum_{a=M+1}^{D}\sum_{b=M+1}^{D}\left\{\boldsymbol{u}_a^T(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T\boldsymbol{u}_b\right\}\ \boldsymbol{u}_a^T\boldsymbol{u}_b\right)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{a=M+1}^{D}\boldsymbol{u}_a^T(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T\boldsymbol{u}_a = \sum_{a=M+1}^{D}\boldsymbol{u}_a^T\boldsymbol{S}\boldsymbol{u}_a
\end{aligned}
$$

Now, we have the same objective to the maximal variance formulation. And the proposition is proven. □

2.2 Probabilistic PCA

Definition 2.4. (PPCA) We consider the following system of probabilities:

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{I}) \qquad p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z}+\boldsymbol{\mu},\Psi)$$

which we can consider $p(\boldsymbol{z})$ to be the PCA projection, while $p(\boldsymbol{x}|\boldsymbol{z})$ is the reconstruction of the PCA. This means that we can perform PPCA of it.

Proposition 2.3. *We consider the marginalization of Gaussian to be:*

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{C}) \qquad \textit{where} \qquad \boldsymbol{C} = \boldsymbol{\Psi}+\boldsymbol{W}\boldsymbol{W}^T$$

Proof. We consider the linear Gaussian model in this case, where we use the marginalization (see above) to get the value of $p(\boldsymbol{x})$ □

Proposition 2.4. *We consider the inference of the latent and we have:*

$$p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^T\boldsymbol{\Psi}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}),\boldsymbol{\Sigma}^{-1}\Big)$$

where we have $\boldsymbol{\Sigma} = I+\boldsymbol{W}^T\Psi^{-1}\boldsymbol{W}$.

Remark 20. We have the projection to be:

$$\hat{\boldsymbol{x}}_i = \boldsymbol{W}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^T\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})$$

As we have the PCA projection that also take noise into consideration. Furthermore, if $\boldsymbol{\Psi} = \psi^2\boldsymbol{I}$ and $\psi\to 0$, then it leads to the PCA estimation (given the correct $\boldsymbol{W}$, which we will explore later).

Remark 21. **(Likelihood of PPCA)** Now, we are left to find the actual value of $\boldsymbol{W}$ and we will assume that we are aware of $\boldsymbol{\mu}$ (which is usually $\boldsymbol{0}$), while the covariance matrix is assumed to be $\boldsymbol{\Psi} = \psi^2\boldsymbol{I}$. The log-likelihood of this PPCA is (using the marginalized):

$$
\begin{aligned}
l = \log p(\{\boldsymbol{x}_i\}_{i=1}^N|\boldsymbol{\mu},\boldsymbol{C}) &= \log\prod_{i=1}^{N}\frac{1}{\sqrt{|2\pi\boldsymbol{C}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right\} \\
&= -\frac{N}{2}\log|\boldsymbol{C}| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}) + \text{const.} \\
&= -\frac{N}{2}\log|\boldsymbol{C}| - \frac{N}{2}\operatorname{Tr}\left(\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right) + \text{const.} \\
&= -\frac{N}{2}\log|\boldsymbol{C}| - \frac{N}{2}\operatorname{Tr}\left(\boldsymbol{C}^{-1}\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T\right) + \text{const.} \\
&= -\frac{N}{2}\log|\boldsymbol{C}| - \frac{N}{2}\operatorname{Tr}\left(\boldsymbol{C}^{-1}\boldsymbol{S}\right) + \text{const.}
\end{aligned}
$$

Now $\boldsymbol{C} = \boldsymbol{W}\boldsymbol{W}^T+\psi^2\boldsymbol{I}$ is given above.

Proposition 2.5. *Non-Trivial Solution of Maximum Likelihood estimate of $\boldsymbol{W}$ is equal to:*

$$\boldsymbol{W}_{ML} = \boldsymbol{U}(\boldsymbol{\Lambda} - \psi^2\boldsymbol{I})^{1/2}\boldsymbol{V}^T$$

where $\boldsymbol{U} \in \mathbb{R}^{D\times M}$ is the first M eigenvector of $\boldsymbol{S}$ the empirical variance matrix and $\boldsymbol{\Lambda} \in \mathbb{R}^{M\times M}$ be the matrix, which is the eigenvalues of $\boldsymbol{S}$. Finally, $\boldsymbol{V} \in \mathbb{R}^{M\times M}$ is an arbitrary orthogonal matrix.

Proof. Let's consider the derivative of the log-likelihood with respect to $\boldsymbol{W}$ as we now have:

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{W}} &= \frac{\partial}{\partial \boldsymbol{W}}\left[-\frac{N}{2}\log|\boldsymbol{C}| - \frac{N}{2}\operatorname{Tr}\left(\boldsymbol{C}^{-1}\boldsymbol{S}\right)\right] \\ &= -N\left[-\boldsymbol{C}^{-1}\boldsymbol{W} + \boldsymbol{C}^{-1}\boldsymbol{S}\boldsymbol{C}^{-1}\boldsymbol{W}\right]\end{aligned}$$

Setting this to $\mathbf{0}$, and we can see that, we have the following equation: $\boldsymbol{S}\boldsymbol{C}^{-1}\boldsymbol{W} = \boldsymbol{W}$, there are 2 outcome to the solution: $\boldsymbol{W} = 0$, which can be shown to be minimum. Or, Consider $\boldsymbol{W}$ in the SVD form i.e $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^T$ for orthogonal matrix $\boldsymbol{U}$ and $\boldsymbol{V}$, while $\boldsymbol{L}$ is diagonal matrix. This would entail:

$$\begin{aligned}&\boldsymbol{S}\Big(\boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{L}^T\boldsymbol{U}^T + \psi^2\boldsymbol{I}\Big)^{-1}\boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^T = \boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^T \\ \implies &\boldsymbol{S}\Big(\boldsymbol{U}\boldsymbol{L}^2\boldsymbol{U}^T + \psi^2\boldsymbol{I}\Big)^{-1}\boldsymbol{U} = \boldsymbol{U} \\ \implies &\boldsymbol{S}\boldsymbol{U}\Big(\boldsymbol{L}^2 + \psi^2\boldsymbol{I}\Big)^{-1} = \boldsymbol{U} \\ \implies &\boldsymbol{S}\boldsymbol{U} = \boldsymbol{U}\Big(\boldsymbol{L}^2 + \psi^2\boldsymbol{I}\Big)\end{aligned}$$

For the second implication, we have:

$$\boldsymbol{U}\Big(\boldsymbol{L}^2 + \psi^2\boldsymbol{I}\Big) = \Big(\boldsymbol{U}\boldsymbol{L}^2\boldsymbol{U}^T + \psi^2\boldsymbol{I}\Big)\boldsymbol{U} \implies \Big(\boldsymbol{U}\boldsymbol{L}^2\boldsymbol{U}^T + \psi^2\boldsymbol{I}\Big)^{-1}\boldsymbol{U} = \boldsymbol{U}\Big(\boldsymbol{L}^2 + \psi^2\boldsymbol{I}\Big)^{-1}$$

We can see that the $\boldsymbol{U}$ is the eigenvector of $\boldsymbol{S}$, where the corresponding eigenvalues are $\lambda_i = l_i^2 + \psi^2$, and so we can rewrite the weight to be:

$$\boldsymbol{W} = \boldsymbol{U}(\boldsymbol{\Lambda} - \psi^2\boldsymbol{I})^{1/2}\boldsymbol{V}^T$$

where $\boldsymbol{\Lambda}$ be the matrix, which is the eigenvalues of $\boldsymbol{S}$. □

2.3 Other Related Models

Definition 2.5. (Factor Analysis) The factor analysis is PPCA:

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\mathbf{0}, \boldsymbol{I}) \qquad p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \Psi)$$

but we consider the matrix $\boldsymbol{\Psi}$ to be $D \times D$ diagonal matrix. This means that the inferences still holds. However, the training is much harder now as we may not find the $\boldsymbol{W}$ from the data in closed form.

Definition 2.6. (Canonical Correlation Analysis) The data vector $\mathcal{D} = \{(\boldsymbol{u}_1, \boldsymbol{v}_1), (\boldsymbol{u}_2, \boldsymbol{v}_2), \ldots\}$ where $\boldsymbol{u}_i \in \mathcal{U}$ and $\boldsymbol{v}_i \in \mathcal{V}$. We want to find the correlation:

- We find te unti vector $\boldsymbol{a} \in \mathcal{U}$ and $\boldsymbol{b} \in \mathcal{V}$ such that the correlation $\boldsymbol{u}_i^T\boldsymbol{a}$ and $\boldsymbol{v}_i^T\boldsymbol{b}$ is the maximum the covariance between them.
- This also requires some in the orthogonal subspace.

Now, the probabilistic CCA is the generative model with latent $\boldsymbol{z}_i \in \mathbb{R}^K$ such that:

$$\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \qquad \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{\Upsilon}\boldsymbol{z}, \boldsymbol{\Psi}_u) \qquad \boldsymbol{v} \sim \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{z}, \boldsymbol{\Psi}_u)$$

where we have $\boldsymbol{\Psi}_{\boldsymbol{u}} \succeq 0$ and $\boldsymbol{\Psi}_{\boldsymbol{v}} \succeq 0$. This is block diagonal noise. There are certain restriction of Gaussian FA and PCA as it is modelled the distribution that is too restrictive.

Definition 2.7. (Mixture Distribution) The mixture distribution has simple discrete latent variable:

$$s_i \sim \text{Discrete}[\boldsymbol{\pi}] \qquad \boldsymbol{x}_i|\boldsymbol{s}_i \sim P_{s_i}[\theta_{s_i}]$$

The mixture can be seen as a mixture of multiple sources of data. The probability density of the single data point is given as:

$$p(\boldsymbol{x}_i) = \sum_{i=1}^{k} p(\boldsymbol{x}_i|s_i = m)p(s_i = m) = \sum_{i=1}^{k} \pi_m p(s_i = m)$$

The most notable mixture distribution is the mixture of Gaussian distribution.

Remark 22. Please note that once can perform a Baysian inference to infer the probability that particular point $\boldsymbol{x}$ belongs to the cluster m of the mixture distribution:

$$p(s_i = m|\boldsymbol{x}) = \frac{p_m(\boldsymbol{x})\pi_m}{\sum_{i=1}^{k} p_i(\boldsymbol{x})\pi_i}$$

Remark 23. **(Mixture of Gaussian)** Let's consider the mixture of Gaussian, where we have the following mixture distribution:

$$p(\{\boldsymbol{x}_i\}_{i=1}^{N} \mid \{\boldsymbol{\mu}_i\}_{i=1}^{k}, \{\boldsymbol{\Sigma}_i\}_{i=1}^{k}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \sum_{m=1}^{k} \pi_m \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_m|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right\}$$

Again it is hard to find the solution to the problem, and so we will consider the method to solve such the problem, which is called Expectation-Maximization (EM).

Remark 24. **(Mixture of Factor Analyzers)** Now, we consider the clustering and dimensionality reduction:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{m=1}^{k} \pi_m \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{W}_k\boldsymbol{W}_k^T + \boldsymbol{\Psi})$$

where π_k is the mixing proportion, while the parameter are $\boldsymbol{\theta} = \left\{\{\pi_m, \boldsymbol{\mu}_m, \boldsymbol{W}_m\}_{m=1}^{k}, \boldsymbol{\Psi}\right\}$. Please note that this model has 2 kinds of latent variables, which are:

- Cluster indicator variable π_m for $m \in \{1, \cdots, k\}$
- Continuous factor $\boldsymbol{z}_{im} \in \mathbb{R}^M$

Together giving us the following data generating distribution:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{m=1}^{k} p(\pi_m) \int p(\boldsymbol{z})p(\boldsymbol{x}_m|\boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{z}$$

We can use EM to perform an optimization over it.

3 Expectation-Maximization

3.1 Methods

Remark 25. **(General Form of The Objective)** Now, we shall consider the latent variable model to be:

$$p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\theta}_x) = \frac{f_x(\boldsymbol{x}) \exp(\boldsymbol{\phi}(\boldsymbol{\theta}_x, \boldsymbol{z})^T \boldsymbol{T}_x(\boldsymbol{x}))}{Z_x(\boldsymbol{\phi}(\boldsymbol{\theta}_x, \boldsymbol{z}))} \qquad p(\boldsymbol{z}|\boldsymbol{\theta}_x) = \frac{f_z(\boldsymbol{z}) \exp(\boldsymbol{\theta}_z^T \boldsymbol{T}_z(\boldsymbol{z}))}{Z_z(\boldsymbol{\theta}_z)}$$

and so the marginalization of the latent variable model is equal to:

$$p(\boldsymbol{x}|\boldsymbol{\theta}_x, \boldsymbol{\theta}_z) = \int \frac{f_x(\boldsymbol{x}) \exp(\boldsymbol{\phi}(\boldsymbol{\theta}_x, \boldsymbol{z})^T \boldsymbol{T}_x(\boldsymbol{x}))}{Z_x(\boldsymbol{\phi}(\boldsymbol{\theta}_x, \boldsymbol{z}))} \cdot \frac{f_z(\boldsymbol{z}) \exp(\boldsymbol{\theta}_z^T \boldsymbol{T}_z(\boldsymbol{z}))}{Z_z(\boldsymbol{\theta}_z)} \, \mathrm{d}\boldsymbol{z}$$

And so, the log-likelihood over the dataset $\{\boldsymbol{x}_i\}_{i=1}^N$ is equal to:

$$l(\boldsymbol{\theta}_x, \boldsymbol{\theta}_z) = \sum_{i=1}^N \log \int \frac{f_x(\boldsymbol{x}_i) \exp(\boldsymbol{\phi}(\boldsymbol{\theta}_x, \boldsymbol{z})^T \boldsymbol{T}_x(\boldsymbol{x}_i))}{Z_x(\boldsymbol{\phi}(\boldsymbol{\theta}_x, \boldsymbol{z}))} \cdot \frac{f_z(\boldsymbol{z}) \exp(\boldsymbol{\theta}_z^T \boldsymbol{T}_z(\boldsymbol{z}))}{Z_z(\boldsymbol{\theta}_z)} \, \mathrm{d}\boldsymbol{z}$$

Theorem 3.1. ***(Jensen's Inequality)*** *Getting the set of weights $\{\alpha_i\}_{i=1}^N$ where $\sum \alpha_i = 1$ and $\{x_i > 0\}_{i=1}^N$, then we can show that:*

$$\log\left(\sum_{i=1}^N \alpha_i x_i\right) \geq \sum_{i=1}^N \alpha_i \log(x_i)$$

for the concave f (and α_i is the probability measure). This also implies that: $f(\mathbb{E}_\alpha[x]) \geq \mathbb{E}_\alpha[f(x)]$ with the equality iff $f(x)$ is almost surely constant or linear support of α.

Definition 3.1. (Free Energy) Consider the latent variable model again: Given the observed data $\mathcal{X} = \{\boldsymbol{x}_i\}$ and the set of latent variables $\mathcal{Z} = \{\boldsymbol{z}_i\}$ with the parameter $\boldsymbol{\theta} = \{\boldsymbol{\theta}_x, \boldsymbol{\theta}_z\}$:

$$l(\boldsymbol{\theta}) = \log p(\mathcal{X}|\boldsymbol{\theta}) = \log \int p(\mathcal{Z}, \mathcal{X}|\boldsymbol{\theta}) \, \mathrm{d}\mathcal{Z}$$

We will consider the Jensen's inequality for arbitrary distribution $q(\mathcal{Z})$. We can find the lower bound:

$$l(\boldsymbol{\theta}) = \log \int q(\mathcal{Z}) \frac{p(\mathcal{Z}, \mathcal{X}|\boldsymbol{\theta})}{q(\mathcal{Z})} \, \mathrm{d}\mathcal{Z} \geq \int q(\mathcal{Z}) \log \frac{p(\mathcal{Z}, \mathcal{X}|\boldsymbol{\theta})}{q(\mathcal{Z})} \, \mathrm{d}\mathcal{Z} = F(q, \boldsymbol{\theta})$$

We denote $F(q, \boldsymbol{\theta})$ to be a free energy.

Remark 26. **(Other Form of Free Energy)** We can consider the free energy to be:

$$\begin{aligned}\int q(\mathcal{Z}) \log \frac{p(\mathcal{Z}, \mathcal{X}|\boldsymbol{\theta})}{q(\mathcal{Z})} \, \mathrm{d}\mathcal{Z} &= \int q(\mathcal{Z}) \log p(\mathcal{Z}, \mathcal{X}|\boldsymbol{\theta}) \, \mathrm{d}\mathcal{Z} - \int q(\mathcal{Z}) \log q(\mathcal{Z}) \, \mathrm{d}\mathcal{Z} \\ &= \int q(\mathcal{Z}) \log p(\mathcal{Z}, \mathcal{X}|\boldsymbol{\theta}) \, \mathrm{d}\mathcal{Z} + H[q]\end{aligned}$$

and so the free-energy is equal to $F(q, \boldsymbol{\theta}) = \langle \log p(\mathcal{Z}, \mathcal{X}|\boldsymbol{\theta}) \rangle_{q(\mathcal{Z})} + H[q]$

Definition 3.2. (Expectation Maximization) The EM-Step follows the following step to be:

- *E-Step*: Optimize $F(q, \theta)$ with respected to $q(\mathcal{Z})$ as:

$$q^{(k)}(\mathcal{Z}) = \arg\max_{q(\mathcal{Z})} F(q(\mathcal{Z}), \boldsymbol{\theta}^{(k-1)})$$

- *M-Step*: Optimize $F(q, \theta)$ with respected to $\boldsymbol{\theta}$ as:

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}} F(q^{(k)}(\mathcal{Z}), \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \langle \log P(\mathcal{Z}, \mathcal{X}|\boldsymbol{\theta}) \rangle_{q^{(k)}(\mathcal{Z})}$$

Remark 27. **(Simplification of E-Step)** To consider the E-step, we have the following free-energy principle:

$$\begin{aligned}F(q, \boldsymbol{\theta}) &= \int q(\mathcal{Z}) \frac{p(\mathcal{Z}, \mathcal{X}|\boldsymbol{\theta})}{q(\mathcal{Z})} \, \mathrm{d}\mathcal{Z} \\ &= \int q(\mathcal{Z}) \frac{p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}) p(\mathcal{X}|\boldsymbol{\theta})}{q(\mathcal{Z})} \, \mathrm{d}\mathcal{Z} \\ &= \int q(\mathcal{Z}) \log p(\mathcal{X}|\theta) \, \mathrm{d}\mathcal{Z} + \int q(\mathcal{Z}) \log \frac{p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta})}{q(\mathcal{Z})} \, \mathrm{d}\mathcal{Z} \\ &= l(\boldsymbol{\theta}) - \mathrm{KL}[q(\mathcal{Z}) \| p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta})]\end{aligned}$$

where $l(\boldsymbol{\theta})$ is the log-likelihood and to minimize the $q(\mathcal{Z})$ by making KL-divergence equal to 0, when setting:

$$q^{(k)}(\mathcal{Z}) = p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^{(k-1)})$$

Furthermore, we can see that after E-step the free energy is equal to likelihood.

Theorem 3.2. ***(Improvement of EM)*** *The results θ given by EM step never decrease in the likelihood.*

Proof. We have the following chain of inequalities:

$$l(\boldsymbol{\theta}^{(k-1)}) =_1 F(q^{(k)}, \boldsymbol{\theta}^{(k-1)}) \leq_2 F(q^{(k)}, \boldsymbol{\theta}^{(k)}) \leq_3 l(\boldsymbol{\theta}^{(k)})$$

Let's consider each step of the EM as we have:

1. The E-step gives free energy to the likelihood.
2. The M-step maximizes the free energy with respected to $\boldsymbol{\theta}$
3. $F \leq l$ comes from the lower-bound property of the free energy.

Thus the theorem is proved □

Theorem 3.3. ***(Convergence of EM to Optimum)*** *The fixed point of the EM algorithm is the stationary point of log-likelihood $l(\boldsymbol{\theta})$ and it is maxima.*

Proof. Please note that the fixed point of the EM is when (M-step with complete E-step):

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left\langle \log p(\mathcal{X}, \mathcal{Z}|\boldsymbol{\theta}) \right\rangle_{p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^*)} \Big|_{\boldsymbol{\theta}^*} = \mathbf{0}$$

Consider the log-likelihood, which we have:

$$\begin{aligned} l(\boldsymbol{\theta}) &= \log p(\mathcal{X}|\boldsymbol{\theta}) = \left\langle \log p(\mathcal{X}|\boldsymbol{\theta}) \right\rangle_{p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^*)} \\ &= \left\langle \log \frac{p(\mathcal{Z}, \mathcal{X}|\boldsymbol{\theta})}{p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta})} \right\rangle_{p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^*)} \\ &= \left\langle p(\mathcal{Z}, \mathcal{X}|\boldsymbol{\theta}) \right\rangle_{p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^*)} - \left\langle p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}) \right\rangle_{p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^*)} \end{aligned}$$

Consider the derivative with respected to $\boldsymbol{\theta}$ of the log-likelihood evaluated at $\boldsymbol{\theta}$ and we have:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}^*} &= \frac{\partial}{\partial \boldsymbol{\theta}} \left\langle p(\mathcal{Z}, \mathcal{X}|\boldsymbol{\theta}) \right\rangle_{p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^*)} \Big|_{\boldsymbol{\theta}^*} - \frac{\partial}{\partial \boldsymbol{\theta}} \left\langle p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}) \right\rangle_{p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^*)} \Big|_{\boldsymbol{\theta}^*} \\ &= \mathbf{0} - \frac{\partial}{\partial \boldsymbol{\theta}} \left\langle p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}) \right\rangle_{p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^*)} \Big|_{\boldsymbol{\theta}^*} = \mathbf{0} \end{aligned}$$

The second term is equal to $\mathbf{0}$ as the KL-divergence is zero. Now, we are left to show that the second derivative is "negative", where we have:

$$\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} l(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}^*} = \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \left\langle p(\mathcal{Z}, \mathcal{X}|\boldsymbol{\theta}) \right\rangle_{p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^*)} \Big|_{\boldsymbol{\theta}^*} - \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \left\langle p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}) \right\rangle_{p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^*)} \Big|_{\boldsymbol{\theta}^*}$$

The first term is negative due to the improvement of EM, now the second term is positive because it is the minimum of KL-divergence. And, the point $\boldsymbol{\theta}^*$ is minima. □

Definition 3.3. (Partial M-Step and E-Step) Let's start with the *M-Step* first, as long as we increase the $F(q, \boldsymbol{\theta})$ in the M-step, the theorem above still holds. And so, for partial M-step, after E-step, we have the following gradient update:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}^{(k-1)}} \left\langle \log p(\mathcal{Z}, \mathcal{X}|\boldsymbol{\theta}) \right\rangle_{p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta})} = \frac{\partial}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}^{(k-1)}} \log p(\mathcal{X}|\boldsymbol{\theta})$$

For *E-Step*, we can use the typical gradient based scheme on $q(\mathcal{Z})$ but the theorem above may not hold.

3.2 Practical

Remark 28. **(Useful Derivative Results)** Now, we are going to consider EM applied to the mixture of Gaussian that we have described above. There are some derivatives that would be useful in the future. Starting wthe the log-likelihood:

$$\begin{aligned} l(\{\boldsymbol{\mu}_i\}_{i=1}^k, \{\boldsymbol{\Sigma}_i\}_{i=1}^k, \boldsymbol{\pi}) &= \sum_{i=1}^N \log \sum_{m=1}^k \pi_m \underbrace{\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_m|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right\}}_{p_m(\boldsymbol{x}_i;\boldsymbol{\theta}_m)} \\ &= \sum_{i=1}^N \log \sum_{m=1}^k \pi_m p_m(\boldsymbol{x}_i;\boldsymbol{\theta}_m) \end{aligned}$$

where we denote $\boldsymbol{\theta}_m = \left\{\{\boldsymbol{\mu}_i\}_{i=1}^k, \{\boldsymbol{\Sigma}_i\}_{i=1}^k\right\}$. Now consider the derivatives of the log-likelihood with respected to both $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$: Starting with outer derivatives, we have:

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\theta}_m}\sum_{i=1}^N \log\sum_{m=1}^k \pi_m p_m(\boldsymbol{x}_i;\boldsymbol{\theta}_m) &= \sum_{i=1}^N \frac{\pi_m}{\sum_{m=1}^k \pi_m p_m(\boldsymbol{x}_i;\boldsymbol{\theta}_m)} \frac{\partial p_m(\boldsymbol{x}_i;\boldsymbol{\theta}_m)}{\partial\boldsymbol{\theta}_m} \\ &= \sum_{i=1}^N \underbrace{\frac{\pi_m p_m(\boldsymbol{x}_i;\boldsymbol{\theta}_m)}{\sum_{m=1}^k \pi_m p_m(\boldsymbol{x}_i;\boldsymbol{\theta}_m)}}_{\gamma_{im}} \frac{\partial \log p_m(\boldsymbol{x}_i;\boldsymbol{\theta}_m)}{\partial\boldsymbol{\theta}_m} \end{aligned}$$

Please remeber γ_{im} as it is going to be useful later on. We are left to find the derivatives on the RHS, starting with the mean $\boldsymbol{\mu}_m$, which we can use the results from before:

$$\frac{\partial \log p_m(\boldsymbol{x}_i;\boldsymbol{\theta}_m)}{\partial\boldsymbol{\mu}_m} = \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_m)$$

Similarly for $\boldsymbol{\Sigma}^{-1}$ we have:

$$\frac{\partial \log p_m(\boldsymbol{x}_i;\boldsymbol{\theta}_m)}{\partial\boldsymbol{\Sigma}_m} = \frac{1}{2}\left[\boldsymbol{\Sigma}_m - (\boldsymbol{x}_i - \boldsymbol{\mu}_m)(\boldsymbol{x}_i - \boldsymbol{\mu}_m)^T\right]$$

We can plug both into get the full derivative.

Proposition 3.1. ***(E-Step for mixture of Gaussian)*** *The E-step for mixture of Gaussian is done by setting:*

$$\gamma_{im}^{(t)} = \frac{\pi_m p_m(\boldsymbol{x}_i;\boldsymbol{\theta}_m)}{\sum_{m=1}^k \pi_m p_m(\boldsymbol{x}_i;\boldsymbol{\theta}_m)}$$

This follows directly from the Bayes' rule.

Proposition 3.2. ***(M-Step for mixture of Gaussian)*** *The M-step for mixture of Gaussian is done by setting:*

$$\begin{aligned} \boldsymbol{\mu}_m^{(t)} &= \frac{1}{N_m}\sum_{i=1}^N \gamma_{im}^{(t)}\boldsymbol{x}_i \\ \boldsymbol{\Sigma}_m^{(t)} &= \frac{1}{N_m}\sum_{i=1}^N \gamma_{im}^{(t)}(\boldsymbol{x}_i - \boldsymbol{\mu}_m)(\boldsymbol{x}_i - \boldsymbol{\mu}_m)^T \\ \pi_m^{(t)} &= \frac{N_m}{N} \end{aligned}$$

where $N_m = \sum_{i=1}^N \gamma_{im}^{(t)}$.

Proof. Setting the derivative of log-likelihood to zero for $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$, which are:

$$\frac{\partial l}{\partial \boldsymbol{\mu}_m} = \sum_{i=1}^{N} \gamma_{im} \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_m) \qquad \frac{\partial l}{\partial \boldsymbol{\Sigma}_m} = \frac{1}{2}\sum_{i=1}^{N} \gamma_{im}\Big[\boldsymbol{\Sigma}_m - (\boldsymbol{x}_i - \boldsymbol{\mu}_m)(\boldsymbol{x}_i - \boldsymbol{\mu}_m)^T\Big]$$

with simple rearrangement, we have the required update. We are left with $\pi_m^{(t)}$ as we required the constriant so that $\sum_{m=1}^{k} \pi_m^{(t)} = 1$. We have the following Lagragian:

$$\begin{aligned}
&\frac{\partial}{\partial \pi_m}\left(\sum_{i=1}^{N} \log \sum_{m=1}^{k} \pi_m p_m(\boldsymbol{x}_i; \boldsymbol{\theta}_m) - \lambda\left[\sum_{m=1}^{k} \pi_m - 1\right]\right) \\
&\qquad = \sum_{i=1}^{N} \frac{p_m(\boldsymbol{x}_i; \boldsymbol{\theta}_m)}{\sum_{m=1}^{k} \pi_m p_m(\boldsymbol{x}_i; \boldsymbol{\theta}_m)} + \lambda
\end{aligned}$$

Setting this to zero, and we have:

$$\begin{aligned}
&\sum_{i=1}^{N} \frac{p_m(\boldsymbol{x}_i; \boldsymbol{\theta}_m)}{\sum_{m=1}^{k} \pi_m p_m(\boldsymbol{x}_i; \boldsymbol{\theta}_m)} + \lambda = 0 \\
\implies &\sum_{m=1}^{k}\sum_{i=1}^{N} \frac{p_m(\boldsymbol{x}_i; \boldsymbol{\theta}_m)\pi_m}{\sum_{m=1}^{k} \pi_m p_m(\boldsymbol{x}_i; \boldsymbol{\theta}_m)} + \lambda \sum_{m=1}^{k} \pi_m = 0 \\
\implies &N + \lambda = 0
\end{aligned}$$

and so, we have $\lambda = -N$. Now we simply times π_k on both side of the original derivative and we get the result. □

Remark 29. **(Connection between K-Mean and Gaussian Mixture Model)** If we consider $\phi_m = 1/k$ and $\boldsymbol{\Sigma}_m = \sigma^2 I$ where $\sigma^2 \to 0$, the leads us to the responsibility:

$$r_{im} = \delta\Big(m, \arg\min_{l} \ \|\boldsymbol{x}_i - \boldsymbol{\mu}_l\|^2\Big)$$

We find the data in which it is closest to the mean, which we allocate the weight to be 1 and 0 for all others. Now, we simply have to update the mean (M-step), which is given by:

$$\boldsymbol{\mu}_m = \frac{\sum_i \gamma_{im} \boldsymbol{x}_i}{\sum_i \boldsymbol{x}_i}$$

This is exactly the K-mean algorithm.

Remark 30. **(EM for Factor Analysis)** Now, we are interested in training the factor analysis using the EM algorithm. Recall that factor analysis consists of the following model:

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

Furthermore, we are going to assume that $\boldsymbol{\Psi} \in \mathbb{R}^{D\times D}$ where it is a diagonal matrix with values $\{\psi_{ii}\}_{i=1}^{D}$. And so recall that the marginal distribution, which is:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \Psi + \boldsymbol{W}\boldsymbol{W}^T)$$

We can see that the model parameter is $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{\Psi}\}$. Now let's recall the EM algorithm, and see what we have to do:

- E-Step: We have to find the posterior distribution of each latent variable that corresponds to each data point in dataset $\{\boldsymbol{x}_i\}_{i=1}^{N}$:

$$q_i^{(t)}(\boldsymbol{z}_i) = p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})$$

- M-Step: We have to find $\boldsymbol{\theta}^{(t)}$ as we have:

$$\arg\max_{\boldsymbol{\theta}} F(q_i^{(t)}, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \int q_i^{(t)}(\boldsymbol{z}_i) \Big[\log p(\boldsymbol{z}_i|\boldsymbol{\theta}) + \log p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) \Big] \, \mathrm{d}\boldsymbol{z}_i$$

Proposition 3.3. ***(E-Step for Factor Analysis)*** *Finding the posterior of $\boldsymbol{z}_i$ for the dataset $\{\boldsymbol{x}_i\}_{i=1}^N$, is:*

$$p(\boldsymbol{z}_i|\boldsymbol{x}_i) = \mathcal{N}(\boldsymbol{z}_i|\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^T\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i, \boldsymbol{\Sigma}^{-1})$$

where $\boldsymbol{\Sigma} = \boldsymbol{I} + \boldsymbol{W}^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}$. This is the old result from PPCA that we have derived eariler. We denote the mean to be $\boldsymbol{\mu}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{W}^T\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i)$.

Proposition 3.4. ***(M-Step for Factor Analysis)*** *The M-Step update of the factor analysis is given by:*

$$\boldsymbol{W} = \left(\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{\mu}_i^T \right) \left(\sum_{i=1}^{N} \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T + N\boldsymbol{\Sigma} \right)^{-1} \qquad \boldsymbol{\Psi} = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T + \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{\mu}_i)(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{\mu}_i)^T$$

Proof. M-step of the Factor analysis, we are interested into find the variables $\boldsymbol{\theta}$ such that

$$\arg\max_{\boldsymbol{\theta}} F(q, \boldsymbol{\theta}) = \sum_{i=1}^{N} \langle \log p(\boldsymbol{z}_i|\boldsymbol{\theta}) + \log p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) \rangle_{q_i(\boldsymbol{z}_i)}$$

Now let's consoder the sum of the log, and so we have:

$$\begin{aligned}
\log p(\boldsymbol{z}_i|\boldsymbol{\theta}) + \log p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) &= -\frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i)^T\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i) + \text{const.} \\
&= -\frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\Big[\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{x}_i - 2\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{z}_i^T\boldsymbol{W}^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}\boldsymbol{z}_i\Big] + \text{const.} \\
&= -\frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\Big[\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{x}_i - 2\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}\boldsymbol{z}_i + \mathrm{Tr}[\boldsymbol{W}^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}\boldsymbol{z}_i\boldsymbol{z}_i^T]\Big] + \text{const.}
\end{aligned}$$

Now, we consider the expectation over $q_i^{(t)}(\boldsymbol{z}_i)$ as we now have:

$$\begin{aligned}
&\langle \log p(\boldsymbol{z}_i|\boldsymbol{\theta}) + \log p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) \rangle_{q_i^{(t)}(\boldsymbol{z}_i)} \\
&\quad= \left\langle -\frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\Big[\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{x}_i - 2\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}\boldsymbol{z}_i + \mathrm{Tr}[\boldsymbol{W}^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}\boldsymbol{z}_i\boldsymbol{z}_i^T]\Big] \right\rangle_{q_i^{(t)}(\boldsymbol{z}_i)} \\
&\quad= -\frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\Big[\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{x}_i - 2\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{W} \left\langle \boldsymbol{z}_i \right\rangle_{q_i^{(t)}(\boldsymbol{z}_i)} + \mathrm{Tr}[\boldsymbol{W}^T\boldsymbol{\Psi}^{-1}\boldsymbol{W} \left\langle \boldsymbol{z}_i\boldsymbol{z}_i^T \right\rangle_{q_i^{(t)}(\boldsymbol{z}_i)}]\Big] \\
&\quad= -\frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\Big[\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{x}_i - 2\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}\boldsymbol{\mu}_i + \mathrm{Tr}[\boldsymbol{W}^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T + \boldsymbol{\Sigma})]\Big]
\end{aligned}$$

Using the expectation over the Gaussian distribution to be $\langle \boldsymbol{z}_i \rangle = \boldsymbol{\mu}_i$ and $\left\langle \boldsymbol{z}_i\boldsymbol{z}_i^T \right\rangle = \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T + \boldsymbol{\Sigma}$. Now, let's consider the derivative with respected to $\boldsymbol{W}$ and $\boldsymbol{\Psi}^{-1}$:

$$\begin{aligned}
&\frac{1}{2}\frac{\partial}{\partial \boldsymbol{W}} \sum_{i=1}^{N} \Big[2\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}\boldsymbol{\mu}_i - \mathrm{Tr}[\boldsymbol{W}^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T + \boldsymbol{\Sigma})]\Big] \\
&\qquad= \frac{1}{2}\sum_{i=1}^{N} \left[2\frac{\partial}{\partial \boldsymbol{W}} \mathrm{Tr}[\boldsymbol{W}\boldsymbol{\mu}_i\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}] - \frac{\partial}{\partial \boldsymbol{W}} \mathrm{Tr}[\boldsymbol{W}^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T + \boldsymbol{\Sigma})] \right] \\
&\qquad= \frac{1}{2}\sum_{i=1}^{N} \left[2\boldsymbol{\Psi}^{-1}\boldsymbol{x}_i\boldsymbol{\mu}_i^T - 2\left(\boldsymbol{\Psi}^{-1}\boldsymbol{W}(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T + \boldsymbol{\Sigma})\right) \right] \\
&\qquad= \boldsymbol{\Psi}^{-1} \left[\sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{\mu}_i^T - \boldsymbol{W} \left(\sum_{i=1}^{N} \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T + N\boldsymbol{\Sigma} \right) \right]
\end{aligned}$$

Setting the derivative to zero and we yields:

$$\boldsymbol{W} = \left(\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{\mu}_i^T\right)\left(\sum_{i=1}^{N} \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T + N\boldsymbol{\Sigma}\right)^{-1}$$

Let's consider the case for $\boldsymbol{\Psi}^{-1}$, which we have:

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Psi}^{-1}} \sum_{i=1}^{N} \left(-\frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\Big[\boldsymbol{x}_i^T \boldsymbol{\Psi}^{-1}\boldsymbol{x}_i - 2\boldsymbol{x}_i^T \boldsymbol{\Psi}^{-1}\boldsymbol{W}\boldsymbol{\mu}_i + \mathrm{Tr}[\boldsymbol{W}^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T + \boldsymbol{\Sigma})]\Big]\right) \\
&\quad = \frac{1}{2}\sum_{i=1}^{N}\Bigg(-\frac{\partial}{\partial \boldsymbol{\Psi}^{-1}}\log|\boldsymbol{\Psi}| - \frac{\partial}{\partial \boldsymbol{\Psi}^{-1}}\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{x}_i + 2\frac{\partial}{\partial \boldsymbol{\Psi}^{-1}}\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}\boldsymbol{\mu}_i \\
&\qquad\qquad - \frac{\partial}{\partial \boldsymbol{\Psi}^{-1}}\mathrm{Tr}[\boldsymbol{W}^T\boldsymbol{\Psi}^{-1}\boldsymbol{W}(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T + \boldsymbol{\Sigma})]\Bigg) \\
&\quad = \frac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{\Psi} - \boldsymbol{x}_i\boldsymbol{x}_i^T + 2\boldsymbol{x}_i\boldsymbol{\mu}_i^T\boldsymbol{W}^T - \boldsymbol{W}\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T\boldsymbol{W}^T - \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T\right)
\end{aligned}$$

Setting the derivative to zero, and we have:

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{\Psi} - \boldsymbol{x}_i\boldsymbol{x}_i^T + 2\boldsymbol{x}_i\boldsymbol{\mu}_i^T\boldsymbol{W}^T - \boldsymbol{W}\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T\boldsymbol{W}^T - \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T\right) = 0 \\
\implies & N\boldsymbol{\Psi} - N\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T - \sum_{i=1}^{N}\Big[\boldsymbol{x}_i\boldsymbol{x}_i^T - 2\boldsymbol{x}_i\boldsymbol{\mu}_i^T\boldsymbol{W}^T + \boldsymbol{W}\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T\boldsymbol{W}^T\Big] = 0 \\
\implies & \boldsymbol{\Psi} = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T + \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{\mu}_i)(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{\mu}_i)^T
\end{aligned}$$

Which is the result as required. □

3.3 Additional EM Methods

Remark 31. This is going to be more generalized version of EM as we may dealing with the exponential family, which occures for both latent variable and the observable.

Proposition 3.5. *Given the exponential family of the form:*

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{f(\boldsymbol{x})\exp(\boldsymbol{\theta}^T\boldsymbol{T}(\boldsymbol{x}))}{Z(\boldsymbol{\theta})}$$

We can show that the

$$\frac{\partial}{\partial \boldsymbol{\theta}}\log Z(\boldsymbol{\theta}) = \langle T(\boldsymbol{x})\rangle_{p(\boldsymbol{x}|\boldsymbol{\theta})}$$

Proof.

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\phi}}\log Z(\boldsymbol{\phi}) &= \frac{1}{Z(\boldsymbol{\phi})}\frac{\partial}{\partial \boldsymbol{\phi}}Z(\boldsymbol{\phi}) \\
&= \frac{1}{Z(\boldsymbol{\phi})}\frac{\partial}{\partial \boldsymbol{\phi}}\iint f(\boldsymbol{x},\boldsymbol{z})\exp\left\{\boldsymbol{\theta}^T\boldsymbol{T}(\boldsymbol{x})\right\}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z} \\
&= \iint \frac{1}{Z(\boldsymbol{\phi})}f(\boldsymbol{x},\boldsymbol{z})\frac{\partial}{\partial \boldsymbol{\phi}}\exp\left\{\boldsymbol{\theta}^T\boldsymbol{T}(\boldsymbol{x})\right\}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z} \\
&= \iint \frac{1}{Z(\boldsymbol{\phi})}f(\boldsymbol{x},\boldsymbol{z})\exp\left\{\boldsymbol{\theta}^T\boldsymbol{T}(\boldsymbol{x})\right\}\boldsymbol{T}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z} \\
&= \langle T(\boldsymbol{x})\rangle_{p(\boldsymbol{x}|\boldsymbol{\theta})}
\end{aligned}$$

□

Definition 3.4. (EM For Exponential Family) We now consider the *joint* probability distribution over $\boldsymbol{x}$ and $\boldsymbol{z}$ to be an exponential family, now we have:

$$p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\phi}) = \frac{f(\boldsymbol{x}, \boldsymbol{z}) \exp\left(\boldsymbol{\phi}(\boldsymbol{\theta}_x, \boldsymbol{\theta}_z)^T \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{z})\right)}{Z(\boldsymbol{\phi})}$$

Let's consider the free-energy (consider only terms that depends on $\boldsymbol{\phi}$), using the previous result:

$$\begin{aligned} F(q, \boldsymbol{\phi}) &= \int q(\boldsymbol{z}) \log P(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\phi}) \ \mathrm{d}\boldsymbol{z} + H[q] \\ &= \int q(\boldsymbol{z}) \Big[\boldsymbol{\phi}(\boldsymbol{\theta}_x, \boldsymbol{\theta}_z)^T \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{z}) - \log Z(\boldsymbol{\phi})\Big] \ \mathrm{d}\boldsymbol{z} + \text{const.} \\ &= \boldsymbol{\phi}(\boldsymbol{\theta}_x, \boldsymbol{\theta}_z)^T \left\langle T(\boldsymbol{x}, \boldsymbol{z})\right\rangle_{q(\boldsymbol{z})} - \log Z(\boldsymbol{\phi}) \end{aligned}$$

For the EM algorithm, we have the following:

- For *E-Step*, we only need to compute the sufficient statistics under the distribution $q(\boldsymbol{z})$.
- For *M-step*, we solve for the following equation, where we use the derivative from above:

$$\frac{\partial F}{\partial \boldsymbol{\phi}} = \left\langle T(\boldsymbol{x}, \boldsymbol{z})\right\rangle_{q(\boldsymbol{z})} - \left\langle T(\boldsymbol{x}, \boldsymbol{z})\right\rangle_{p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\phi})} = \boldsymbol{0}$$

Remark 32. **(EM For Exponential Family Mixture)** We consider a short-hand version where we consider the 1-hot vector i.e for the mixture component m, we have:

$$s_i = m \iff \boldsymbol{s}_i = [0, 0, \ldots, \underbrace{1}_{\text{m-th position}}, \ldots, 0]^T$$

To compute the components, we consider the "list" of parameters, which is a matrix $\boldsymbol{\Theta} = [\boldsymbol{\theta}_m]$ (has the size of $\mathbb{R}^{D \times M}$). The log-likelihood is the given by:

$$\begin{aligned} \log p\left(\{\boldsymbol{x}_i\}_{i=1}^N, \{\boldsymbol{s}_i\}_{i=1}^N\right) &= \sum_{i=1}^N \log\left(\pi_{is_i}\left[\frac{f(\boldsymbol{x}_i)\exp\left(\boldsymbol{\theta}_{s_i}^T \boldsymbol{T}(\boldsymbol{x}_i)\right)}{Z(\boldsymbol{\theta}_{s_i})}\right]\right) \\ &= \sum_{i=1}^N \Big[\log \pi_{is_i} + \boldsymbol{\theta}_{s_i}^T \boldsymbol{T}(\boldsymbol{x}_i) - \log Z(\boldsymbol{\theta}_{s_i})\Big] + \text{const} \\ &= \sum_{i=1}^N \left[(\log \boldsymbol{\pi})^T \boldsymbol{s}_i + (\boldsymbol{\Theta}\boldsymbol{s}_i)^T \boldsymbol{T}(\boldsymbol{x}_i) - \boldsymbol{s}_i^T \log Z(\boldsymbol{\Theta})\right] + \text{const} \end{aligned}$$

where $Z(\boldsymbol{\Theta})$ is the normalizing factor of each component, so it has the size of M. If we were to consider the EM-algorithm for this, we have:

- E-step, we calculate the expectation over the latent variable, which gives us:

$$\gamma_{im} = \sum_{i=1}^N \left\langle \boldsymbol{s}_i\right\rangle_p \qquad \sum_{i=1}^N \boldsymbol{T}(\boldsymbol{x}_i) \left\langle \boldsymbol{s}_i^T\right\rangle_p$$

- M-step, maximizing the log-joint probability, we have the following solutions (to the derivative):

$$\boldsymbol{\pi}^{(k+1)} \propto \sum_{i=1}^N \left\langle \boldsymbol{s}_i\right\rangle_p \qquad \left\langle \boldsymbol{T}(\boldsymbol{x}_i) \,\middle|\, \boldsymbol{\theta}_m^{(k+1)}\right\rangle = \frac{\sum_{i=1}^N \boldsymbol{T}(\boldsymbol{x}_i) \left\langle [\boldsymbol{s}_i]_m\right\rangle_p}{\sum_{i=1}^N \left\langle [\boldsymbol{s}_i]_m\right\rangle_p}$$

Thinking of this as the Gaussian mixture model, where we also have similar form.

Remark 33. **(EM for MAP)** Recall the probability distribution over the exponential model; however, we also added the prior over the parameters:

$$p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\phi}) = \frac{f(\boldsymbol{x}, \boldsymbol{z}) \exp\left(\boldsymbol{\phi}(\boldsymbol{\theta}_x, \boldsymbol{\theta}_z)^T \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{z})\right)}{Z(\boldsymbol{\phi})} \qquad p(\boldsymbol{\phi}|\nu, \boldsymbol{\tau}) = \frac{F(\nu, \boldsymbol{\tau}) \exp\left(\boldsymbol{\phi}(\boldsymbol{\theta}_x, \boldsymbol{\theta}_z)^T \boldsymbol{\tau}\right)}{Z(\boldsymbol{\phi})^\nu}$$

To consider the free-energy over the dataset $\mathcal{X}$ and set of latent variables $\mathcal{Z}$, we have:

$$\begin{aligned} F(q, \boldsymbol{\phi}) &= \int q(\boldsymbol{z}) \log P(\mathcal{X}, \mathcal{Z}, \boldsymbol{\phi}) \, \mathrm{d}\boldsymbol{z} + H[q] \\ &= \int q(\mathcal{Z}) \left[\boldsymbol{\phi}(\boldsymbol{\theta}_x, \boldsymbol{\theta}_z)^T \left(\boldsymbol{\tau} + \sum_{i=1}^N \boldsymbol{T}(\boldsymbol{x}_i, \boldsymbol{z}_i) \right) - (N + \nu) \log Z(\boldsymbol{\phi}) \right] \mathrm{d}\mathcal{Z} + \text{const.} \\ &= \boldsymbol{\phi}(\boldsymbol{\theta}_x, \boldsymbol{\theta}_z)^T \left(\boldsymbol{\tau} + \sum_{i=1}^N \langle \boldsymbol{T}(\boldsymbol{x}_i, \boldsymbol{z}_i) \rangle_{q(\boldsymbol{z}_i)} \right) - (N + \nu) \log Z(\boldsymbol{\phi}) + \text{const.} \end{aligned}$$

Now, we carry on with the EM step as usual.

4 Latent Variable Models for Time Series

4.1 Problems

Definition 4.1. (Time Series with Latent Variable) The latent variable model for time series (with first order markov model) is given by:

$$p(\boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}) = p(\boldsymbol{z}_1) p(\boldsymbol{x}_1|\boldsymbol{z}_1) \prod_{i=2}^T p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) p(\boldsymbol{x}_t|\boldsymbol{z}_t)$$

Definition 4.2. (Discrete Model) The discrete model is described using the following set of parameters as there are K number of latent discrete values:

- Initial State Probability $p(\boldsymbol{z}_1)$: is given as $\pi_z = p(\boldsymbol{z}_1 = j)$
- Transition Matrix $p(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t)$: is given as $\boldsymbol{\Phi}_{ij} = p(\boldsymbol{z}_{t+1} = j|\boldsymbol{z}_t = i)$
- Emission Distribution $p(\boldsymbol{x}_t|\boldsymbol{z}_t)$:
  - Continuous variable: $\boldsymbol{A}_j(\boldsymbol{x}) = p(\boldsymbol{x}_t = \boldsymbol{x}|\boldsymbol{z}_t = j)$
  - Discrete variable: $\boldsymbol{A}_{jk}(\boldsymbol{x}) = p(\boldsymbol{x}_t = k|\boldsymbol{z}_t = j)$

And we have the following process for a discrete model:

$$\boldsymbol{z}_1 \sim \boldsymbol{\pi} \qquad \boldsymbol{z}_{t+1}|\boldsymbol{z}_t \sim \boldsymbol{\Phi}_{\boldsymbol{z}_t} \qquad \boldsymbol{x}_t|\boldsymbol{z}_t \sim \boldsymbol{A}_{\boldsymbol{z}_t}$$

Definition 4.3. (Continuous Variable) We consider the continuous variable model, which we model the time series with Gaussian and Linear transformation, which we have the following process:

$$\boldsymbol{z}_1 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{Q}_0) \qquad \boldsymbol{z}_t|\boldsymbol{z}_{t-1} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{z}_{t-1}, \boldsymbol{Q}) \qquad \boldsymbol{x}_t|\boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{C}\boldsymbol{z}_t, \boldsymbol{R})$$

We still use the same joint distribution as above. This has the other name called Linear Gaussian State Space Model (LGSSM).

Definition 4.4. (Inference Problems) We are interesting to find the following joint distributions:

$$p(\boldsymbol{z}_t|\boldsymbol{x}_{1:t}) \qquad p(\boldsymbol{z}_t|\boldsymbol{x}_{1:T})$$

The first one is called filtering distribution, while the second one is called smoothing distribution.

Remark 34. **(Problem for the Problem)** It is clear that we can integrate out the posterior to get the answer for filtering (and smoothing), but it is very hard:

$$p(\boldsymbol{z}|\boldsymbol{x}_1,\dots,\boldsymbol{x}_t) = \int \cdots \int p(\boldsymbol{z}_1,\dots,\boldsymbol{z}_t|\boldsymbol{x}_1,\dots,\boldsymbol{x}_t)\ \mathrm{d}\boldsymbol{z}_1 \cdots\ \mathrm{d}\boldsymbol{z}_t$$

For the discrete variable, we can have sum instead of integration.

4.2 Solving Inference Problem

Lemma 4.1. ***(Baysian Filtering)*** *One can compute the filtering recursively as:*

$$p(\boldsymbol{z}_t|\boldsymbol{x}_{1:t}) \propto \int p(\boldsymbol{x}_t|\boldsymbol{z}_t)p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})p(\boldsymbol{z}_{t-1}|\boldsymbol{x}_{1:t-1})\ \mathrm{d}\boldsymbol{z}_{t-1}$$

We have the same treatment for discrete variable.

Proof. We have the following:

$$\begin{aligned}
p(\boldsymbol{z}_t|\boldsymbol{x}_{1:t}) &= \int p(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_{1:t-1})\ \mathrm{d}\boldsymbol{z}_{t-1} \\
&= \int \frac{p(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}|\boldsymbol{x}_{1:t-1})}{p(\boldsymbol{x}_t|\boldsymbol{x}_{1:t-1})}\ \mathrm{d}\boldsymbol{z}_{t-1} \\
&\propto \int p(\boldsymbol{x}_t|\boldsymbol{z}_t, \boldsymbol{z}_{t-1}, \boldsymbol{x}_{1:t-1})p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, \boldsymbol{x}_{1:t-1})p(\boldsymbol{z}_{t-1}|\boldsymbol{x}_{1:t-1})\ \mathrm{d}\boldsymbol{z}_{t-1} \\
&= \int p(\boldsymbol{x}_t|\boldsymbol{z}_t)p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})p(\boldsymbol{z}_{t-1}|\boldsymbol{x}_{1:t-1})\ \mathrm{d}\boldsymbol{z}_{t-1}
\end{aligned}$$

And we have the values as required. □

Definition 4.5. (Forward Message) We consider the quantity $\alpha_t(i)$ for which it is the joint probability of observation and the current latent variable $\boldsymbol{z}_t = i$ i.e:

$$\alpha_t(i) = p(\boldsymbol{x}_1,\dots,\boldsymbol{x}_t, \boldsymbol{z}_t = i|\boldsymbol{\theta})$$

We call this a forward message.

Proposition 4.1. *The forward message can be calculated recursively as:*

$$\alpha_{t+1}(i) = \left(\sum_{j=1}^{K} \alpha_t(j)\boldsymbol{\Phi}_{ij}\right) \boldsymbol{A}_i(\boldsymbol{x}_{t+1})$$

where at the start: $\alpha_1(i) = \pi_i \boldsymbol{A}_i(\boldsymbol{x})$

Proof. We consider the following:

$$\begin{aligned}
\alpha_{t+1}(i) &= \sum_{j=1}^{K} p(\boldsymbol{x}_1,\dots,\boldsymbol{x}_t, \boldsymbol{x}_{t+1}, s_t = j|s_{t+1} = i) \\
&= \sum_{j=1}^{K} p(\boldsymbol{x}_1,\dots,\boldsymbol{x}_t|s_t = j)p(s_{t+1} = i|s_t = j)p(\boldsymbol{x}_{t+1}|s_{t+1} = i) \\
&= \left(\sum_{j=1}^{K} \alpha_t(j)\boldsymbol{\Phi}_{ij}\right) \boldsymbol{A}_i(\boldsymbol{x}_{t+1})
\end{aligned}$$

□

Remark 35. **(Filtering For Discrete Model)** One can solve the filtering problem, as we have:

$$p(\boldsymbol{s}_t = i|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \boldsymbol{\theta}) = \frac{\alpha_t(i)}{\sum_{j=1}^K \alpha_t(j)}$$

Remark 36. **(Usefulness of Filtering Method)** Given the values of $\alpha_T(i)$, we can calculate the likelihood of the parameter $\boldsymbol{\theta}$ in $\mathcal{O}(TK^2)$ time instead of exponential time via the marginalization:

$$p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T|\theta) = \sum_{k=1}^K \alpha_T(k)$$

Proposition 4.2. ***(First Step Filter)*** *We can show that the $p(\boldsymbol{z}_1|\boldsymbol{x}_{1:1})$ is equal to:*

$$p(\boldsymbol{z}_1|\boldsymbol{x}_{1:1}) = \mathcal{N}\Big(\boldsymbol{z}_1 \Big| \underbrace{\hat{\boldsymbol{z}}_1^0 + \boldsymbol{K}_1(\boldsymbol{x}_1 - \boldsymbol{C}\hat{\boldsymbol{z}}_1^0)}_{\boldsymbol{z}_1^1}, \underbrace{\hat{\boldsymbol{V}}_1^0 - \boldsymbol{K}_1\boldsymbol{C}\hat{\boldsymbol{V}}_1^0}_{\hat{\boldsymbol{V}}_1^1}\Big)$$

where we set $\boldsymbol{K}_1 = \hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T\Big[\boldsymbol{R} + \boldsymbol{C}\hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T\Big]^{-1}$.

Proof. For the state space model, we start with $p(\boldsymbol{z}_1|\boldsymbol{x}_1)$. We denote $\hat{\boldsymbol{z}}_1^0 = \boldsymbol{\mu}_0$ and $\hat{\boldsymbol{V}}_1^0 = \boldsymbol{Q}_0$. Now, we apply linear Gaussian model above to get the inference. Recall that

$$p(\boldsymbol{z}_1) = \mathcal{N}(\boldsymbol{z}_1|\hat{\boldsymbol{z}}_1^0, \hat{\boldsymbol{V}}_1^0) \qquad p(\boldsymbol{x}_1|\boldsymbol{z}_1) = \mathcal{N}(\boldsymbol{x}_1|\boldsymbol{C}\boldsymbol{z}_1, \boldsymbol{R})$$

Now, we have the linear Gaussian model as

$$p(\boldsymbol{z}_1|\boldsymbol{x}_1) = \mathcal{N}\Big(\boldsymbol{z}_1 \Big| \boldsymbol{\Sigma}\Big[\boldsymbol{C}^T\boldsymbol{R}^{-1}\boldsymbol{x}_1 + (\hat{\boldsymbol{V}}_1^0)^{-1}\hat{\boldsymbol{z}}_1^0\Big],\ \boldsymbol{\Sigma}\Big)$$

where $\boldsymbol{\Sigma}$ is simplified using Woodbury identity:

$$\begin{aligned}
\boldsymbol{\Sigma} &= \Big[(\hat{\boldsymbol{V}}_1^0)^{-1} + \boldsymbol{C}^T\boldsymbol{R}^{-1}\boldsymbol{C}\Big]^{-1} \\
&= \hat{\boldsymbol{V}}_1^0 - \hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T\Big[\boldsymbol{R} + \boldsymbol{C}\hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T\Big]^{-1}\boldsymbol{C}\hat{\boldsymbol{V}}_1^0 \\
&= \hat{\boldsymbol{V}}_1^0 - \boldsymbol{K}_1\boldsymbol{C}\hat{\boldsymbol{V}}_1^0
\end{aligned}$$

Now, we consider the mean, as we now have:

$$\begin{aligned}
&\Big[\hat{\boldsymbol{V}}_1^0 - \boldsymbol{K}_1\boldsymbol{C}\hat{\boldsymbol{V}}_1^0\Big]\Big[\boldsymbol{C}^T\boldsymbol{R}^{-1}\boldsymbol{x}_1 + (\hat{\boldsymbol{V}}_1^0)^{-1}\hat{\boldsymbol{z}}_1^0\Big] \\
&\qquad = \hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T\boldsymbol{R}^{-1}\boldsymbol{x}_1 + \hat{\boldsymbol{z}}_1^0 - \boldsymbol{K}_1\boldsymbol{C}\hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T\boldsymbol{R}^{-1}\boldsymbol{x}_1 - \boldsymbol{K}_1\boldsymbol{C}\hat{\boldsymbol{z}}_1^0 \\
&\qquad = \hat{\boldsymbol{z}}_1^0 - \boldsymbol{K}_1\boldsymbol{C}\hat{\boldsymbol{z}}_1^0 + (\hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T\boldsymbol{R}^{-1} - \boldsymbol{K}_1\boldsymbol{C}\hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T\boldsymbol{R}^{-1})\boldsymbol{x}_1
\end{aligned}$$

Let's consider the value, which we can show that it is indeed $\boldsymbol{K}_1$, where we will consider $\boldsymbol{B} = \boldsymbol{C}\hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T$:

$$\begin{aligned}
&\hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T\boldsymbol{R}^{-1} - \boldsymbol{K}_1\boldsymbol{B}\boldsymbol{R}^{-1} = \boldsymbol{K}_1 \\
\iff &\hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T\boldsymbol{R}^{-1} = \boldsymbol{K}_1 + \boldsymbol{K}_1\boldsymbol{B}\boldsymbol{R}^{-1} \\
\iff &\hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T\boldsymbol{R}^{-1} = \boldsymbol{K}_1(\boldsymbol{I} + \boldsymbol{B}\boldsymbol{R}^{-1}) \\
\iff &\hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T\boldsymbol{R}^{-1}(\boldsymbol{I} + \boldsymbol{B}\boldsymbol{R}^{-1})^{-1} = \boldsymbol{K}_1 \\
\iff &\hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T\boldsymbol{R}^{-1}(\boldsymbol{I} + \boldsymbol{B}\boldsymbol{R}^{-1})^{-1} = \hat{\boldsymbol{V}}_1^0\boldsymbol{C}^T\Big[\boldsymbol{R} + \boldsymbol{B}\Big]^{-1} \\
\iff &\boldsymbol{R}^{-1}(\boldsymbol{I} + \boldsymbol{B}\boldsymbol{R}^{-1})^{-1} = \Big[\boldsymbol{R} + \boldsymbol{B}\Big]^{-1} \\
\iff &\Big[\boldsymbol{R}(\boldsymbol{I} + \boldsymbol{B}\boldsymbol{R}^{-1})\Big]^{-1} = \Big[\boldsymbol{R} + \boldsymbol{B}\Big]^{-1}
\end{aligned}$$

And so equation the is proven. □

Proposition 4.3. ***(General Time Filtering)*** *We want to find the value $p(\boldsymbol{z}_t|\boldsymbol{x}_{1:t})$, which we can show that it is equal to, given $p(\boldsymbol{z}_{t-1}|\boldsymbol{x}_{1:t-1}) = \mathcal{N}(\boldsymbol{z}_{t-1}|\hat{\boldsymbol{z}}_{t-1}^{t-1}, \hat{\boldsymbol{V}}_{t-1}^{t-1})$:*

$$\mathcal{N}\Big(\boldsymbol{z}_t\Big|\hat{\boldsymbol{z}}_{t-1}^{t} + \boldsymbol{K}_t(\boldsymbol{x}_t - \boldsymbol{C}\hat{\boldsymbol{z}}_{t-1}^{t}),\ \hat{\boldsymbol{V}}_t^{t-1} - \boldsymbol{K}_t\boldsymbol{C}\hat{\boldsymbol{V}}_t^{t-1}\Big)$$

where $\hat{\boldsymbol{z}}_{t-1}^{t} = \boldsymbol{A}\hat{\boldsymbol{z}}_{t-1}^{t-1}$ and $\hat{\boldsymbol{V}}_t^{t-1} = \boldsymbol{Q} + \boldsymbol{A}\hat{\boldsymbol{V}}_{t-1}^{t-1}\boldsymbol{A}^T$ and $\boldsymbol{K}_t = \hat{\boldsymbol{V}}_t^{t-1}\boldsymbol{C}^T\left[\boldsymbol{R} + \boldsymbol{C}\hat{\boldsymbol{V}}_t^{t-1}\boldsymbol{C}^T\right]^{-1}$

Proof. Now, we want to find the probability $p(\boldsymbol{z}_t|\boldsymbol{x}_{1:t-1})$, first, which we can use the marginal distribution from the linear Gaussian model between $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ and $p(\boldsymbol{z}_{t-1}|\boldsymbol{x}_{1:t-1})$, which we marginalize out $\boldsymbol{z}_{t-1}$, which we have:

$$\begin{aligned}
&\int \mathcal{N}(\boldsymbol{z}_{t-1}|\hat{\boldsymbol{z}}_{t-1}^{t-1}, \hat{\boldsymbol{V}}_{t-1}^{t-1})\mathcal{N}(\boldsymbol{z}_t|\boldsymbol{A}\boldsymbol{z}_{t-1}, \boldsymbol{Q})\ \mathrm{d}\boldsymbol{z}_{t-1} \\
&\qquad = \mathcal{N}\Big(\boldsymbol{z}_t\Big|\boldsymbol{A}\hat{\boldsymbol{z}}_{t-1}^{t-1},\ \boldsymbol{Q} + \boldsymbol{A}\hat{\boldsymbol{V}}_{t-1}^{t-1}\boldsymbol{A}^T\Big) \\
&\qquad = \mathcal{N}\Big(\boldsymbol{z}_t\Big|\hat{\boldsymbol{z}}_{t-1}^{t},\ \hat{\boldsymbol{V}}_t^{t-1}\Big)
\end{aligned}$$

This follows directly from linear Gaussian model. Now, we follows the same method above (as we now need to find distribution of $\boldsymbol{z}_t$ given *addiitonal* information of $\boldsymbol{x}_t$), and so we have:

$$\mathcal{N}\Big(\boldsymbol{z}_t\Big|\hat{\boldsymbol{z}}_{t-1}^{t} + \boldsymbol{K}_t(\boldsymbol{x}_t - \boldsymbol{C}\hat{\boldsymbol{z}}_{t-1}^{t}),\ \hat{\boldsymbol{V}}_t^{t-1} - \boldsymbol{K}_t\boldsymbol{C}\hat{\boldsymbol{V}}_t^{t-1}\Big)$$

where $\boldsymbol{K}_t = \hat{\boldsymbol{V}}_t^{t-1}\boldsymbol{C}^T\left[\boldsymbol{R} + \boldsymbol{C}\hat{\boldsymbol{V}}_t^{t-1}\boldsymbol{C}^T\right]^{-1}$. Thus complete the prove. □

Remark 37. **(Baysian Smoothing - For Discrete Model)** To find the smoothing, we consider calculating marginal posterior as:

$$p(\boldsymbol{z}_t|\boldsymbol{x}_{1:T}) = \frac{p(\boldsymbol{z}_t, \boldsymbol{x}_{t+1:T}|\boldsymbol{x}_{1:t})}{p(\boldsymbol{x}_{t+1:T}|\boldsymbol{x}_{1:t})} = \frac{p(\boldsymbol{x}_{t+1:T}|\boldsymbol{z}_t)p(\boldsymbol{z}_t|\boldsymbol{x}_{1:t})}{p(\boldsymbol{x}_{t+1:T}|\boldsymbol{x}_{1:t})} = \frac{p(\boldsymbol{x}_{t+1:T}|\boldsymbol{z}_t)p(\boldsymbol{z}_t, \boldsymbol{x}_{1:t})}{p(\boldsymbol{z}_{1:T})}$$

We can see that the probability distribution in blue are the forward passing, while the probability distribution in red is the backward distribution.

Definition 4.6. (Backward Message) Consider the value

$$\beta_t(i) = p(\boldsymbol{x}_{t+1:T}|\boldsymbol{z}_t = i)$$

We call it a backward message.

Proposition 4.4. *Backward message can be calculated in recursive manner:*

$$\beta_t(i) = \sum_{j=1}^{K} \boldsymbol{\Phi}_{ij}\boldsymbol{A}_j(\boldsymbol{x}_{t+1})\beta_{t+1}(j)$$

where $\beta_T(i) = 1/K$, we consider a uniform distribution over the state.

Proof. We consider the following

$$\begin{aligned}
\beta_t(i) &= \sum_{j=1}^{K} p(s_{t+1} = j, \boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+2:T}|s_t = i) \\
&= \sum_{j=1}^{K} p(s_{t+1} = j|s_t = i)p(\boldsymbol{x}_{t+1}|s_t = i)p(\boldsymbol{x}_{t+2:T}|s_{t+1} = j) \\
&= \sum_{j=1}^{K} \boldsymbol{\Phi}_{ij}\boldsymbol{A}_j(\boldsymbol{x}_{t+1})\beta_{t+1}(j)
\end{aligned}$$

And so it proposition is proven. □

Definition 4.7. (Forward-Backward Algorithm) The smooth distribution can be calculated as:

$$\gamma_t(i) = p(s_t = i|\boldsymbol{x}_{1:T}) = \frac{p(s_t = i, \boldsymbol{x}_{1:t})p(\boldsymbol{x}_{t+1:T}|s_t = i)}{p(\boldsymbol{x}_{1:T})} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{k}\alpha_t(j)\beta_t(j)}$$

Remark 38. **(Problem with Choosing State)** We consider the value of $\gamma_t(i)$ is computed using forward and backward algorithm. Choosing the state i_t^* with the largest $\gamma_t(i)$ is the best way since the path has the maximum expected number of correct state. However, this may leads to the path, which is impossible.

Definition 4.8. (Viterbi Decoding) The best path algorithm is called Viterbi decoding as we can use the Bellman's dynamics programming. This is done by compute the most probable state sequence (instead of expected state):

$$\arg\max_{\boldsymbol{z}_{1:T}} p(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$$

And so, we use the same recursion max instead of $\sum$, just like the Bellman equation algorithm.

Proposition 4.5. ***(Smoothing for LGSSM)*** *We receives the value $p(\boldsymbol{z}_{t+1}|\boldsymbol{x}_{1:T}) = \mathcal{N}(\boldsymbol{z}_{t+1}|\hat{\boldsymbol{z}}_{t+1}^T, \hat{\boldsymbol{V}}_{t+1}^T)$ from the privious time step. Then we can show the following recursive calculation:*

$$p(\boldsymbol{z}_t|\boldsymbol{x}_{1:T}) = \mathcal{N}\left(\boldsymbol{z}_t \middle| \hat{\boldsymbol{z}}_t + \boldsymbol{J}_t(\boldsymbol{z}_{t+1} - \boldsymbol{A}\hat{\boldsymbol{z}}_t^t),\ \hat{\boldsymbol{V}}_t^t + \boldsymbol{J}_t(\hat{\boldsymbol{V}}_{t+1}^T - \hat{\boldsymbol{V}}_{t+1}^t)\boldsymbol{J}_t^T\right)$$

where $\boldsymbol{J}_t = \hat{\boldsymbol{V}}_t^t\boldsymbol{A}^T(\hat{\boldsymbol{V}}_{t+1}^t)^{-1}$, and the first time step can be calculated from result of filtering.

Proof. We consider the following steps:

$$\begin{aligned}
p(\boldsymbol{z}_t|\boldsymbol{x}_{1:T}) &= \int p(\boldsymbol{z}_t, \boldsymbol{z}_{t+1}|\boldsymbol{x}_{1:T})\ \mathrm{d}\boldsymbol{z}_{t+1} \\
&= \int p(\boldsymbol{z}_t|\boldsymbol{z}_{t+1}, \boldsymbol{x}_{1:T})p(\boldsymbol{z}_{t+1}|x_{1:T})\ \mathrm{d}\boldsymbol{z}_{t+1} \\
&= \int p(\boldsymbol{z}_t|\boldsymbol{z}_{t+1}, \boldsymbol{x}_{1:t})p(\boldsymbol{z}_{t+1}|x_{1:T})\ \mathrm{d}\boldsymbol{z}_{t+1} \\
&= \int \left[\frac{p(\boldsymbol{z}_{t+1}, \boldsymbol{z}_t|\boldsymbol{x}_{1:t})}{p(\boldsymbol{z}_{t+1}|\boldsymbol{x}_{1:t})}\right] p(\boldsymbol{z}_{t+1}|x_{1:T})\ \mathrm{d}\boldsymbol{z}_{t+1}
\end{aligned}$$

The third equality comes from the markov property. Let's consider finding the value $p(\boldsymbol{z}_t|\boldsymbol{z}_{t+1}, \boldsymbol{x}_{1:t})$ using linear Gaussian model as we have proven above: Noted that the value $p(\boldsymbol{z}_t|\boldsymbol{x}_{1:t}) = \mathcal{N}(\boldsymbol{z}_t|\hat{\boldsymbol{z}}_t^t, \hat{\boldsymbol{V}}_t^t)$. We consider the mean vector $[\boldsymbol{z}_t, \boldsymbol{z}_{t+1}]^T$, we will have to consider the covariance:

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{z}_t, \boldsymbol{z}_{t+1}|\boldsymbol{x}_{1:t}) &= \mathbb{E}\left[\left(\boldsymbol{z}_t - \hat{\boldsymbol{z}}_t^t\right)\left(\boldsymbol{z}_{t+1} - \hat{\boldsymbol{z}}_{t+1}^t\right)^T \middle| \boldsymbol{x}_{1:t}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{z}_t - \hat{\boldsymbol{z}}_t^t\right)\left(\boldsymbol{A}\boldsymbol{z}_t + \boldsymbol{w} - \boldsymbol{A}\hat{\boldsymbol{z}}_t^t\right)^T \middle| \boldsymbol{x}_{1:t}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{z}_t - \hat{\boldsymbol{z}}_t^t\right)\left(\boldsymbol{A}\boldsymbol{z}_t - \boldsymbol{A}\hat{\boldsymbol{z}}_t^t\right)^T \middle| \boldsymbol{x}_{1:t}\right] + \mathbb{E}\left[\left(\boldsymbol{z}_t - \hat{\boldsymbol{z}}_t^t\right) \middle| \boldsymbol{x}_{1:t}\right]\mathbb{E}[\boldsymbol{w}^T] \\
&= \mathbb{E}\left[\left(\boldsymbol{z}_t - \hat{\boldsymbol{z}}_t^t\right)\left(\boldsymbol{z}_t - \hat{\boldsymbol{z}}_t^t\right)^T \middle| \boldsymbol{x}_{1:t}\right]\boldsymbol{A}^T = \hat{\boldsymbol{V}}_t^t\boldsymbol{A}^T
\end{aligned}$$

where $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$. Now, we have the following normal distirbution for the joint :

$$\mathcal{N}\left(\begin{bmatrix}\boldsymbol{z}_t \\ \boldsymbol{z}_{t+1}\end{bmatrix} \middle| \begin{bmatrix}\hat{\boldsymbol{z}}_t^t \\ \hat{\boldsymbol{z}}_{t+1}^t\end{bmatrix}, \begin{bmatrix}\hat{\boldsymbol{V}}_t^t & \hat{\boldsymbol{V}}_t^t\boldsymbol{A}^T \\ \boldsymbol{A}\hat{\boldsymbol{V}}_t^t & \hat{\boldsymbol{V}}_{t+1}^t\end{bmatrix}\right)$$

Please note that the value $\hat{\boldsymbol{V}}_{t+1}^t$ can be found in the intermediate value of the filtering. Now, we use conditional Gaussian result from above, and we have:

$$\begin{aligned}
p(\boldsymbol{z}_t|\boldsymbol{z}_{t+1}, \boldsymbol{x}_{1:t}) &= \mathcal{N}\left(\boldsymbol{z}_t \middle| \hat{\boldsymbol{z}}_t + \boldsymbol{J}_t(\boldsymbol{z}_{t+1} - \hat{\boldsymbol{z}}_{t+1}^t),\ \hat{\boldsymbol{V}}_t^t - \boldsymbol{J}_t\hat{\boldsymbol{V}}_{t+1}^t\boldsymbol{J}_t^T\right) \\
&= \mathcal{N}\left(\boldsymbol{z}_t \middle| \hat{\boldsymbol{z}}_t + \boldsymbol{J}_t(\boldsymbol{z}_{t+1} - \boldsymbol{A}\hat{\boldsymbol{z}}_t^t),\ \hat{\boldsymbol{V}}_t^t - \boldsymbol{J}_t\hat{\boldsymbol{V}}_{t+1}^t\boldsymbol{J}_t^T\right)
\end{aligned}$$

where $\boldsymbol{J}_t = \hat{\boldsymbol{V}}_t^t \boldsymbol{A}^T (\hat{\boldsymbol{V}}_{t+1}^t)^{-1}$. Now consider the marginalization using a normal Gaussian linear model together with $p(\boldsymbol{z}_{t+1}|\hat{\boldsymbol{z}}_{t+1}^T, \hat{\boldsymbol{V}}_{t+1}^T)$, as we now have:

$$\begin{aligned} p(\boldsymbol{z}_t|\boldsymbol{x}_{1:T}) &= \mathcal{N}\left(\boldsymbol{z}_t \,\middle|\, \hat{\boldsymbol{z}}_t + \boldsymbol{J}_t(\boldsymbol{z}_{t+1} - \boldsymbol{A}\hat{\boldsymbol{z}}_t^t),\ \hat{\boldsymbol{V}}_t^t - \boldsymbol{J}_t \hat{\boldsymbol{V}}_{t+1}^t \boldsymbol{J}_t^T + \boldsymbol{J}_t \hat{\boldsymbol{V}}_{t+1}^T \boldsymbol{J}_t\right) \\ &= \mathcal{N}\left(\boldsymbol{z}_t \,\middle|\, \hat{\boldsymbol{z}}_t + \boldsymbol{J}_t(\boldsymbol{z}_{t+1} - \boldsymbol{A}\hat{\boldsymbol{z}}_t^t),\ \hat{\boldsymbol{V}}_t^t + \boldsymbol{J}_t(\hat{\boldsymbol{V}}_{t+1}^T - \hat{\boldsymbol{V}}_{t+1}^t)\boldsymbol{J}_t^T\right) \end{aligned}$$

Thus the prove is now complete. □

4.3 Learning Parameter

Definition 4.9. (EM-Algorithm for SSM) We have the following free-energy:

$$F(q, \theta) = \int q(\boldsymbol{z}_{1:T}) \Big[\log p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}|\boldsymbol{\theta}) - \log q(\boldsymbol{z}_{1:T}) \Big] \, \mathrm{d}\boldsymbol{z}_{1:T}$$

Now, we consider the following EM-step, as we have:

- *E-Step*: We find $q^*(\boldsymbol{z}_{1:T}) = p(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$, which is already done in the Kalman smoothing.
- *M-Step*: We want to find $\boldsymbol{\theta}$ by maximizing the $\langle \log p(\boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}|\boldsymbol{\theta}) \rangle_{q(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})}$, where we have:

$$p(\boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}|\boldsymbol{\theta}) = p(\boldsymbol{z}_1) p(\boldsymbol{x}_1|\boldsymbol{z}_1) \prod_{t=2}^{T} p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) p(\boldsymbol{x}_t|\boldsymbol{z}_t)$$

  which are all Gaussian, this leads to least square problem as we will see. This also works with discrete case.

Proposition 4.6. ***(M-Step for $\boldsymbol{C}$)*** *The update for $\boldsymbol{C}$ is:*

$$\boldsymbol{C} = \boldsymbol{R}^{-1} \sum_{t=1}^{T} \langle \boldsymbol{z}_t \rangle_q \boldsymbol{x}_t^T \left(\sum_{t=1}^{T} \left\langle \boldsymbol{z}_t \boldsymbol{z}_t^T \right\rangle_q \right)^{-1}$$

Proof. We have the following log-likelihood (with removed unnecessary terms):

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{C}} \log p(\boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}|\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{C}} \left[\sum_{t=1}^{T} \langle \boldsymbol{P}(\boldsymbol{x}_t|\boldsymbol{z}_t) \rangle_q \right] \\ &= \frac{\partial}{\partial \boldsymbol{C}} \left[-\frac{1}{2} \sum_{t=1}^{T} \left\langle (\boldsymbol{x}_t - \boldsymbol{C}\boldsymbol{z}_t)^T \boldsymbol{R}^{-1} (\boldsymbol{x}_t - \boldsymbol{C}\boldsymbol{z}_t) \right\rangle_q \right] \\ &= \frac{\partial}{\partial \boldsymbol{C}} \left[-\frac{1}{2} \sum_{t=1}^{T} \left\langle \boldsymbol{x}_t^T \boldsymbol{R}^{-1} \boldsymbol{x}_t^T - 2\boldsymbol{x}_t^T \boldsymbol{R}^{-1} \boldsymbol{C}\boldsymbol{z}_t + \boldsymbol{z}_t^T \boldsymbol{C}^T \boldsymbol{R}^{-1} \boldsymbol{C}\boldsymbol{z}_t \right\rangle_q \right] \\ &= \frac{\partial}{\partial \boldsymbol{C}} \left[\operatorname{Tr}\left[\boldsymbol{C} \sum_{t=1}^{T} \langle \boldsymbol{z}_t \rangle_q \boldsymbol{x}_t^T \boldsymbol{R}^{-1} \right] - \frac{1}{2} \operatorname{Tr}\left[\boldsymbol{C}^T \boldsymbol{R}^{-1} \boldsymbol{C} \left\langle \sum_{t=1}^{T} \boldsymbol{z}_t \boldsymbol{z}_t^T \right\rangle_q \right] \right] \\ &= \boldsymbol{R}^{-1} \sum_{t=1}^{T} \langle \boldsymbol{z}_t \rangle_q \boldsymbol{x}_t^T - \boldsymbol{R}^{-1} \boldsymbol{C} \left\langle \sum_{t=1}^{T} \boldsymbol{z}_t \boldsymbol{z}_t^T \right\rangle_q \end{aligned}$$

Setting this to 0, and we have the update. □

Proposition 4.7. ***(M-Step for A)*** *The update for* $\boldsymbol{A}$ *is:*

$$\boldsymbol{A} = \boldsymbol{R}^{-1} \sum_{t=1}^{T} \left\langle \boldsymbol{z}_t \boldsymbol{z}_{t+1}^T \right\rangle_q \left(\sum_{t=1}^{T} \left\langle \boldsymbol{z}_t \boldsymbol{z}_t^T \right\rangle_q \right)^{-1}$$

Proof. We see that the we have the same situation as the $\boldsymbol{C}$, and so the update is the same. Please note that we replace $\boldsymbol{x}_t$ with $\boldsymbol{z}_{t+}$. □

Proposition 4.8. ***(M-Step for HMM)*** *We can take the free-energy with respected to the parameter* $\boldsymbol{\theta}$*, and we have:*

- *Initial State Distribution* $\pi_i = \gamma_i(i)$*, which is expected number of times in state* i *at the start.*
- *Define the expected transition from state* $i \to j$ *to be:*

$$\xi_t(i \to j) = p(z_t = i, z_{t+1} = j | \boldsymbol{x}_{1:T}) = \frac{\alpha_t(i) \boldsymbol{\Phi}_{ij} \boldsymbol{A}_j(\boldsymbol{x}_{t+1}) \beta_{t+1}(j)}{p(\boldsymbol{x}_{1:T})}$$

- *We have the transition probability and output probability to be:*

$$\hat{\Phi}_{ij} = \sum_{t=1}^{T-1} \xi_t(i \to j) \Big/ \sum_{t=1}^{T} \gamma_t(i) \qquad \hat{A}_{ik} = \sum_{t:\boldsymbol{x}_t = k} \gamma_t(i) \Big/ \sum_{t=1}^{T} \gamma_t(i)$$

Proof. As we have computed the E-step, now we consider the free energy, as M-Step trying to maximize the following quantity

$$\begin{aligned}
&\langle \log p(z_{1:T}, \boldsymbol{x}_{1:T}) \rangle_{q(z_{1:T} | \boldsymbol{x}_{1:T}, \boldsymbol{\theta})} \\
&\quad = \mathbb{E}_{q(z_{1:T} | \boldsymbol{x}_{1:T})} \left[\log p(z_1 | \boldsymbol{\pi}) + \sum_{t=2}^{T} \log p(z_t | z_{t-1}, \boldsymbol{\Phi}_{ij}) + \sum_{t=1}^{T} \log p(\boldsymbol{x}_t | z_t, A_{z_t}(\boldsymbol{x}_t)) \right] \\
&\quad = \mathbb{E}_{q(z_1 | \boldsymbol{x}_{1:T})} \left[\log p(z_1 | \boldsymbol{\pi}) \right] + \sum_{t=2}^{T} \mathbb{E}_{q(z_t, z_{t-1} | \boldsymbol{x}_{1:T})} \left[\log p(z_t | z_{t-1}, \boldsymbol{\Phi}_{ij}) \right] + \sum_{t=1}^{T} \mathbb{E}_{q(z_t)} [\log p(\boldsymbol{x}_t | z_t, A_{z_t}(\boldsymbol{x}_t))] \\
&\quad = \sum_{i=1}^{L} \gamma_1(i) \pi_i + \sum_{t=2}^{T} \sum_{i=1}^{L} \sum_{j=1}^{L} \xi(z_i = i, z_{t-1} = j) \Phi_{ji} + \sum_{t=1}^{T} \sum_{i=1}^{L} \gamma_t(i) A_{i \boldsymbol{x}_i}
\end{aligned}$$

Recall that the number of state to be L. We will denote the emission matrix $\boldsymbol{A} \in \mathbb{R}^{L \times O}$ where O is the number of possible observable. Furthermore, we recall that we have the following constraint to be:

$$\sum_j \Phi_{ij} = 1 \qquad \sum_{\boldsymbol{x} \in \mathcal{X}} A_i(\boldsymbol{x}) \qquad \sum_i \pi_i = 1$$

for all $i = 1, \ldots, L$ in the first and second constraint. Now, we have the following Lagrange multiplier:

$$\begin{aligned}
\mathcal{L} = &\sum_{i=1}^{L} \gamma_1(i) \pi_i + \sum_{t=2}^{T} \sum_{i=1}^{L} \sum_{j=1}^{L} \xi(z_i = i, z_{t-1} = j) \Phi_{ji} + \sum_{t=1}^{T} \sum_{i=1}^{L} \gamma_t(i) A_{i \boldsymbol{x}_i} \\
&- \sum_{i=1}^{L} \lambda_i \left(\sum_{j=1}^{L} \Phi_{ij} - 1 \right) - \sum_{i=1}^{L} \nu_i \left(\sum_{j=1}^{L} A_{ij} - 1 \right) - \eta \left(\sum_{i=1}^{L} \pi_i - 1 \right)
\end{aligned}$$

We have the following derivative for π_a:

$$\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial \pi_a} = \frac{\partial}{\partial \pi_a} \sum_{i=1}^{L} \gamma_1(i) \log \pi_i - \frac{\partial}{\partial \pi_a} \eta \left(\sum_{i=1}^{L} \pi_i - 1 \right) = \frac{\gamma_i(a)}{\pi_a} - \eta = 0 \\
\iff &\gamma_1(a) = \nu \pi_a \\
\iff &\sum_{a=1}^{L} \gamma_1(a) = \nu \sum_{a=1}^{L} \pi_a = \nu
\end{aligned}$$

Solving algebra yields to solution for π_a. Let's consider the value of Φ_{ab}, as we have:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \Phi_{ab}} &= \frac{\partial \mathcal{L}}{\partial \Phi_{ab}} \sum_{t=2}^{T}\sum_{i=1}^{L}\sum_{j=1}^{L} \xi(z_i = i, z_{t-1} = j)\Phi_{ji} - \frac{\partial \mathcal{L}}{\partial \Phi_{ab}} \sum_{i=1}^{L} \lambda_i \left(\sum_{j=1}^{L} \Phi_{ij} - 1 \right) \\
&= \sum_{T=1}^{T} \xi(z_t = b, z_{t-1} = b) \frac{1}{\Phi_{ab}} - \lambda_a = 0 \\
\iff \lambda_a \Phi_{ab} &= \sum_{t=1}^{T} \xi(z_t = b, z_{t-1} = a) \\
\iff \sum_{j=1}^{L} \lambda_a \Phi_{aj} &= \sum_{j=1}^{L}\sum_{t=1}^{T} \xi(z_t = j, z_{t-1} = a) \\
\iff \lambda_a &= \sum_{j=1}^{L}\sum_{t=1}^{T} \xi(z_t = j, z_{t-1} = a) = \sum_{t=1}^{T} \gamma_{t-1}(a)
\end{aligned}$$

And, so the equality is proven. Note that there are difference in time notation but they are equivalent. The proof for A_{ij} is the same, so we will not go into details. □

Remark 39. **(Practical Numerical Method)** The message passing algorithm can be implode i.e $\alpha_t(i) = p(\boldsymbol{x}_{1:t}, \boldsymbol{z}_t = i) \to 0$ to fix this, we do the following rescale:

$$\bar{\alpha}_t(i) = \boldsymbol{A}_i(\boldsymbol{x}_t) \sum_{j=1}^{L} \tilde{\boldsymbol{\alpha}}_{t-1}(j) \boldsymbol{\Phi}_{ij} \qquad \rho_t = \sum_{i=1}^{L} \bar{\alpha}_t(i) \qquad \tilde{\alpha}(i) = \bar{\alpha}_t(i)/\rho_t$$

Proposition 4.9. *We can show that:* $\rho_t = p(\boldsymbol{x}_t | \boldsymbol{x}_{1:t-1}, \boldsymbol{\theta})$

Proof. We will consider the quantity $p(z_i | \boldsymbol{x}_{1:i})$ to be equal to:

$$\tilde{\alpha}_t(z) = p(z_t | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t) = \frac{\alpha_t(i)}{p(\boldsymbol{x}_{1:t})}$$

As, we will denote $\rho_t = p(\boldsymbol{x}_t | \boldsymbol{x}_{1:t-1})$, and from the product rule, we have $p(\boldsymbol{x}_{1:t}) = \prod_{i=1}^{t} \rho_i$, and so, we can recover:

$$\alpha_t(i) = \left(\prod_{i=1}^{t} \rho_i \right) \tilde{\alpha}_t(i) = \left(\prod_{i=1}^{t-1} \rho_i \right) \rho_t \tilde{\alpha}_t(i)$$

Let's consider the recursive property of the α, recall its formula:

$$\begin{aligned}
\alpha_{t+1}(i) &= \left(\sum_{j=1}^{L} \alpha_t(j) \boldsymbol{\Phi}_{ij} \right) \boldsymbol{A}_i(\boldsymbol{x}_{t+1}) \\
\iff \left(\prod_{i=1}^{n} \rho_i \right) \rho_{n+1} \tilde{\alpha}_{t+1}(i) &= \left(\sum_{j=1}^{L} \left[\left(\prod_{i=1}^{n} \rho_i \right) \tilde{\alpha}(\boldsymbol{z}_n) \right] \boldsymbol{\Phi}_{ij} \right) \boldsymbol{A}_i(\boldsymbol{x}_{t+1}) \\
\iff \rho_{n+1} \tilde{\alpha}_{t+1}(i) &= \left(\sum_{j=1}^{L} \tilde{\alpha}(\boldsymbol{z}_n) \boldsymbol{\Phi}_{ij} \right) \boldsymbol{A}_i(\boldsymbol{x}_{t+1})
\end{aligned}$$

Thus, we have proven the identity. For $\beta_t(i)$. we can consider the fact that $\hat{\beta}_t(i)$ is:

$$\beta_t(i) = \left(\prod_{i=t+1}^{N} c_i \right) \hat{\beta}_t(i)$$

□

Remark 40. **(Decoding/Derivation of Viterbi)** We are interested in finding the best hidden state path given the observations $\boldsymbol{o}_{1:t}$, together with the following quantity $\delta_t(i)$:

$$\{z_1^*, \dots, z_t^*\} = \arg\max_{z_1,\dots,z_t} p(z_1, \dots, z_t, \boldsymbol{x}_{1:t} = \boldsymbol{o}_{1:t}|\boldsymbol{\theta})$$
$$\delta_t(i) = p(z_1, \dots, z_{t-1}, z_t = i, \boldsymbol{x}_{1:t} = \boldsymbol{o}_{1:t}|\boldsymbol{\theta})$$

We can see that we can use the induction step to find this value:

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i)\Phi_{ij}\right] A_i(\boldsymbol{o}_{t+1})$$

This recursion gives us the Viterbi algorithm.

Definition 4.10. (Viterbi Algorithm) The algorithm to find the best path is given by, where the sequence of hidden state is stored in ψ:

- Initialization:
$$\delta_1(i) = \pi_i b_i(\boldsymbol{o}_1) \qquad \psi_1(i) = 0$$
- Induction Step:
$$\delta_t(j) = \left[\max_{i\in[L]} \delta_{t-1}(i)a_{ij}\right] b_j(\boldsymbol{o}_t) \qquad \psi_t(j) = \arg\max_{i\in[L]} \delta_{t-1}(i)a_{ij}$$
- Termination:
$$P^* = \max_{i\in[L]}[\delta_T(i)] \qquad q_T^* = \arg\max_{i\in[L]} [\delta_T(i)]$$

Remark 41. **(Multiple sequences)** We now consider the update for HMM but with multiple sequences $l \in [L]$, which we can show that the batch update is:

$$\pi_i = \frac{1}{L}\sum_{l=1}^{L} \gamma_1^{(l)}(i) \qquad \boldsymbol{\Phi}_{ij} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T^{(i)}-1} \xi_t^{(l)}(i \to j)}{\sum_{l=1}^{L}\sum_{t=1}^{T^{(i)}} \gamma_t^{(l)}(i)} \qquad \boldsymbol{A}_{ik} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T^{(i)}-1} \delta(\boldsymbol{x}_t = k)\gamma_t^{(l)}(i \to j)}{\sum_{l=1}^{L}\sum_{t=1}^{T^{(i)}} \gamma_t^{(l)}(i)}$$

5 Markov Chain (Monte Carlo)

5.1 Markov Chain

Definition 5.1. (Transition Matrix) We consider the probability of state $i \in [d]$ will move to state $j \in [d]$ i.e $P_{i\to j}$, where d is the number of states. This information for the markov chain is stored in $d \times d$ transtion matrix:

$$\boldsymbol{P} = \begin{pmatrix} P_{1\to 1} & \cdots & P_{1\to d} \\ \vdots & \ddots & \vdots \\ P_{d\to 1} & \cdots & P_{d\to d} \end{pmatrix}$$

Definition 5.2. (Initial Distribution) At the start, we consider the distribution over the states:

$$\boldsymbol{p}_{\text{init}} = \Big(p(x_0 = 1), \cdots, p(x_0 = d)\Big)^T$$

Lemma 5.1. *Given the state distribution p_t, we can find the probability distribution over the next state after transtion $\boldsymbol{p}_{t+1}$ is:*

$$\boldsymbol{p}_{t+1} = \boldsymbol{P}\boldsymbol{p}_t$$

Note that $\boldsymbol{p}_t \in \mathbb{R}^{d\times 1}$ for any $t \in \mathbb{N}$

Proof. Consider the distribution of the first state:

$$\boldsymbol{p}_{t+1}(s) = p(x_{t+1} = s) = \sum_{i=1}^{d} p(x_{t+1} = s | x_t = i) p(x_t = i) = \sum_{i=1}^{d} P_{si} \boldsymbol{p}_t[i]$$

And so it is proven. □

Definition 5.3. (Stationary) Stationary markov chain is when the transtion matrix $\boldsymbol{P}$ doesn't depends on time step.

Definition 5.4. (Equilibrium Distribution) Let $\boldsymbol{P}$ be transtion matrix, then a distribution $\boldsymbol{p}$ such that: $\boldsymbol{P}\boldsymbol{p} = \boldsymbol{p}$ is called equilibrium distribution.

Remark 42. Please note that if we consider applying transtion matrix to any starting distribution:

$$\boldsymbol{p}_\infty = \lim_{n\to} \boldsymbol{P}^n \boldsymbol{p}_{\text{init}}$$

assuming that it exists, then we can show that it gives rise to (somewhat) equilibrium distribution:

$$\boldsymbol{P}\boldsymbol{p}_\infty = \boldsymbol{P} \lim_{n\to} \boldsymbol{P}^n \boldsymbol{p}_{\text{init}} = \lim_{n\to} \boldsymbol{P}^n \boldsymbol{p}_{\text{init}} = \boldsymbol{p}_\infty$$

Definition 5.5. (Aperiodic MC) A stationary MC is aperiodic if: $\boldsymbol{P}_{ii} = 0$ for $i \in [d]$.

Definition 5.6. (Irreducible) MC is irreducible if there is a path from each state to every other state in the transtion.

Theorem 5.1. *If first-order (depends only last time step), stationary MC with finite state space is a periodic and irreducible, then:*

- *Limit distribution $\boldsymbol{p}_\infty$ exists.*
- *Limit distribution is also the equilibrium distribution.*
- *Equilibrium distribution is unique.*
- *Equilibrium distribution doesn't depends on initial distribution.*

Definition 5.7. (Aperiodic State) The state i is a aperiodic if there exits a time t such that for all $t \geq t_0$ with positive integer t_0, the state comes back to itself with positive probability.

Definition 5.8. (Positive Recurrence State) The state i is positive recurrence if the expected number of times that it will return to itself is finite.

Definition 5.9. (Ergodicity) There exists a positive integer t_0, such that for all pair of states i, j, if markov chain starts at time 0 at state i, then for all $t \geq t_0$, the probability of being in state j at time t is more than 0. Note that this implies all states are positive recurrence and aperiodic.

Theorem 5.2. ***(Result of Ergodicity)*** *The ergodicity implies the existence of unique stationary distribution $\boldsymbol{p}_\infty$ given any initialized probability distribution.*

Definition 5.10. (Detailed Balance wrt. $\boldsymbol{p}^*$) The markov chain is detailed balance with respect to a stationary distribution $\boldsymbol{p}^*$ if for all states i, j:

$$p_i^* P_{ij} = p_j^* P_{ji}$$

This also means that the markov chain is *reversible* under this stationary distribution.

Remark 43. **(Solving Detailed Balance)** Detailed balance implies that we have to solve $|\mathcal{S}|^2$ equation, which is more than the simple stationary equation, which require only $|\mathcal{S}|$ equations.

Proposition 5.1. *The stationary distribution of detailed balance markov chain with respect to stationary distribution $\boldsymbol{p}^*$ is $\boldsymbol{p}^*$*

Proof. We consider the summation over the detailed balance equation, as we have:

$$\sum_{i=1}^{|\mathcal{S}|} p_i^* P_{ij} = \sum_{i=1}^{|\mathcal{S}|} p_j^* P_{ji} = p_j^* \sum_{i=1}^{|\mathcal{S}|} P_{ji} = p_j^*$$

This complete the proof. □

Remark 44. **(Ergodicity and Detailed Balance)** Eventhough the detailed balance with respect to stationary distribution $\boldsymbol{p}^*$, will have the stationary distribution to be p^*, this doesn't means that it is unique. For the uniqueness to hold, ergodicity is required.

Remark 45. **(Finding Equilibrium)** There are 2 ways we can find the equilibrium: power method, and eigenvalue. Both of them is based on finding the underlying eigenvalue and eigenvectors.

Definition 5.11. (Random Walk) Given a directed graph G with d vertices, a random walk generates a sequence of nodes: $x_0, x_1, \cdots$ by first select the vertex x_0 at random and, at time t, we uniformly at random select the children of x_{t-1} to get to vertex x_t.

Remark 46. Random walk can be defined as a markov chain, where:

$$\boldsymbol{p}_{\text{init}} = \left(\frac{1}{p}, \cdots, \frac{1}{p}\right) \qquad P_{ij} = \begin{cases} \dfrac{1}{\text{number of edges}} & \text{if } i \text{ links to j} \\ 0 & \text{otherwise} \end{cases}$$

Remark 47. PageRank algorithm applies random walk to a graph, in order to make the original markov chain defined by a transtion matrix $\boldsymbol{T}$ to be irreducible and aperiodic:

$$\boldsymbol{P}_{ij} = (1-\alpha)\boldsymbol{T}_{ij} + \frac{\alpha}{d}$$

where $\boldsymbol{\alpha} \in (0, 1)$

5.2 Sampling Algorithm

Definition 5.12. (Sampling Algorithm) Given a distribution p, the sampling algorithm samples to get the random point $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ such that the marginal distribution is p. Ideally, the drawning are independent.

Remark 48. **(Usefulness of Sampling Algorithm)** There are many use-cases of sampling algorithm notably:

- Compute expectation. Given the output $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ to be independent from distribution p, the expectation is:
$$\mathbb{E}[f(\boldsymbol{x})] \approx \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}_i)$$
- We can approximate the distribution using the sample to perform a density estimation.

Definition 5.13. (Sampling in Baysian Model) Sometimes the posterior over parameter $\hat{q}(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})$ can't be computed analytically, we can still sample from it to get $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$ given the posterior $\hat{q}$. We can now use it for computing expectation like above or perform a prediction:

$$p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_{1:n}) = \int_{\Theta} p(\boldsymbol{x}_{n+1}|\boldsymbol{\theta})\hat{q}(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) \, \mathrm{d}\boldsymbol{\theta} \approx \frac{1}{m} \sum_{i=1}^{m} p(\boldsymbol{x}_{n+1}|\boldsymbol{\theta}_i)$$

Definition 5.14. (Boxed Rejection Sampling) We perform the following procedure, as we want to sample from distribution $\tilde{p}$:

- We generates the $X_i \sim \mathcal{U}[a,b], Y_i \sim \mathcal{U}[0,c]$ uniformly to define the box.
- If $Y_i \leq \tilde{p}(X_i)$, then we keep the sample X_i.
- Otherwise, we reject the sample and repeat the process.

Remark 49. **(Invariance to Scale)** Please note that rejection sampling still works if we have $Y_i \sim \mathcal{U}[0, kc]$ given the distribution $kp(\cdot)$ for $k > 0$. This means that rejection sampling only require us to know the shape of the distribution, as long as kc is big enough to cover the whole distribution p. This also means that, in Baysian inference, we doesn't have to know the evidence in order to sample from the posterior:

$$p(\boldsymbol{\theta}|\mathcal{X}) = \frac{p(\mathcal{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{Z} \qquad \text{where} \qquad Z = \int p(\mathcal{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \ \mathrm{d}\boldsymbol{\theta}$$

We only need $\tilde{p} = p(\mathcal{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ to do rejection sampling.

Definition 5.15. (Rejection Sampling) One doesn't have to use the uniform distribution to sample proposed point. We simply have to find the distribution r such that $\tilde{p} < r$ everywhere:

- Sample $X_i \sim r$
- Sample $Y_i|X_i \sim \mathcal{U}[0, r(X_i)]$
- If $Y_i \leq \tilde{p}(X_i)$, then we keep the sample X_i.
- Otherwise, we reject the sample and repeat the process.

The scaling properties still holds. It will generates iid sample $\theta_1, \ldots, \theta_m$, where the samples $1/m \sum_{i=1}^m f(x_i)$ is an unbiased estimate of $\mathbb{E}_p[f(x)]$.

Remark 50. Given the hight of r, $|A|$ is too high, and the hight of $\tilde{p}$, $|B|$. The rejection sampling will accepts sample with probability $|B|/|A|$, so if $|A|$ is too high, the algorithm may be inefficient.

Definition 5.16. (Improtance Sampling) Importance sampling is given distribution $p = 1/Z_p \tilde{p}$ that we want to sample to and arbitary proposal distribution $q = 1/Z_q \tilde{q}$ (that we can sample), then:

- Draw $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ iid from proposal q.
- Expectation $\mathbb{E}_p[f(\boldsymbol{x})]$ is approximated as:

$$\frac{1}{m} \sum_{i=1}^m \frac{f(\boldsymbol{x}_i)[\tilde{p}(\boldsymbol{x}_i)/\tilde{q}(\boldsymbol{x}_i)]}{\sum_{j=1}^m \tilde{p}(\boldsymbol{x}_j)/\tilde{q}(\boldsymbol{x}_j)}$$

Proposition 5.2. *The importance sampling gives unbiased estimate of* $\mathbb{E}_p[f(\boldsymbol{x})]$.

Proof. We have the following sampling:

$$\mathbb{E}_p[f(\boldsymbol{x})] = \int f(\boldsymbol{x})p(\boldsymbol{x}) \ \mathrm{d}\boldsymbol{x} = \int f(\boldsymbol{x}) \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x}) \ \mathrm{d}\boldsymbol{x} = \mathbb{E}\left[f(\boldsymbol{x}) \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \right] \approx \frac{1}{m} \sum_{i=1}^m \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \sim q$. Now, consider the ratio of normalized factor:

$$\frac{Z_p}{Z_q} = \frac{\int \tilde{p}(\boldsymbol{x}) \ \mathrm{d}\boldsymbol{x}}{Z_q} = \frac{\int \tilde{p}(\boldsymbol{x}) \frac{q(\boldsymbol{x})}{q(\boldsymbol{x})} \ \mathrm{d}\boldsymbol{x}}{Z_q} = \int \tilde{p}(\boldsymbol{x}) \frac{q(\boldsymbol{x})}{Z_q q(\boldsymbol{x})} \ \mathrm{d}\boldsymbol{x} = \mathbb{E}_q \left[\frac{\tilde{p}(\boldsymbol{x})}{\tilde{q}(\boldsymbol{x})} \right] \approx \frac{1}{m} \sum_{i=1}^m \frac{\tilde{p}(\boldsymbol{x}_i)}{\tilde{q}(\boldsymbol{x}_i)}$$

Now, the estimator of f can be found as

$$\mathbb{E}_p[f(\boldsymbol{x})] \approx \frac{1}{m} \sum_{i=1}^m \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} = \frac{1}{m} \sum_{i=1}^m \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} \frac{\tilde{p}(\boldsymbol{x}_i)}{\tilde{q}(\boldsymbol{x}_i)} \frac{Z_q}{Z_p} = \frac{1}{m} \sum_{i=1}^m \frac{f(\boldsymbol{x}_i)[\tilde{p}(\boldsymbol{x}_i)/\tilde{q}(\boldsymbol{x}_i)]}{\sum_{j=1}^m \tilde{p}(\boldsymbol{x}_j)/\tilde{q}(\boldsymbol{x}_j)}$$

□

5.3 More Probabilistic Models

Definition 5.17. (Random Field) Given a weighted undirected graph $\mathcal{N} = (v_{\mathcal{N}}, w_{\mathcal{N}})$ where $v_{\mathcal{N}}$ is the vertex set, and $w_{\mathcal{N}}$ is the set of edge weights. The edge weights are scalar $w_{ij} \in \mathbb{R}$. An edge weight w_{ij} means to edge between i and j. Each vertex v_i is associated with random variable Θ_i. The neighbourhood of vertex v_i is the set:

$$\partial(i) = \{j : w_{ij} \neq 0\}$$

The set $\{\Theta_j : j \in \partial(i)\}$ of random variable associated with neighbourhood is called Markov blanket of Θ_i

Definition 5.18. (Markov Property) The Markov properties is when:

$$p(\theta_i|\theta_j, j \neq i) = p(\theta_i|\theta_j, j \in \partial(i))$$

Each Θ is conditionally independent of remaining field given its Markov blanket. A Markov random field is a random field that is Markov.

Definition 5.19. (Energy Function) Any (strictly positive) probability or density can be rewritten as:

$$p(x) = \frac{1}{Z}\exp(-H(x)) \qquad \text{where} \qquad H : \mathcal{X} \to \mathbb{R}_+ \text{ and } Z = \int \exp(-H(x))\,\mathrm{d}x$$

where H is called potential.

Remark 51. The Markov random field (MRF) density for random variable $\Theta_{1:n}$ can be written as:

$$p(\theta_1, \ldots, \theta_n) = \frac{1}{Z}\exp(-H(\theta_1, \ldots, \theta_n))$$

Definition 5.20. (Potts Model) Let $\mathcal{N}$ be a neighbourhood of graph with weight w_{ij} and $\beta > 0$. The MRF, where we have:

$$p(\theta_1, \ldots, \theta_n) = \frac{1}{Z(\beta)}\exp\left(\beta \sum_{ij} w_{ij}\mathbb{I}\left\{\theta_i = \theta_j\right\}\right)$$

Note that the energy is additive over pairs. Positive weight encorate smoothness:

- $w_{ij} > 0$, this means that $\theta_i = \theta_j$ increases probability.
- $w_{ij} < 0$, this means that $\theta_i = \theta_j$ decreases probability.
- $w_{ij} = 0$, this means that no interaction between θ_i and θ_j.

Definition 5.21. (Ising Model) We have the following distribution:

$$p(\boldsymbol{\theta}_{1:n}) = \frac{1}{Z(\beta)}\exp\left(\sum_{(i,j)\in \text{edge}} \beta\mathbb{I}[\theta_i = \theta_j]\right)$$

Given rejection sampling, we can sample the distribution without the normalizing factor $Z(\beta)$, where $\theta_i \in \{-1, +1\}$ and $w_{ij} \in \{0, 1\}$. This model is on the d-dimensional grid.

Remark 52. **(Usage of Markov Random Field)** Given the problem with observation x_i for each i location on a grid. If we want to model the observation with a distribution $p(x_i|\Theta_i)$ for each location, with its own parameter Θ_i. We can use MRF as a prior distribution, while we represent $p(x_i|\Theta_i)$ as the emission probability:

- Define a joint $(\Theta_1, \ldots, \Theta_n)$ as an MRF on the grid graph.
- Given a positive weight, the MRF will encorate the model to explain a neighbourhood of x_i by similar paramter value. This leads to smoothing in the result.

Definition 5.22. (Baysian Mixture Model) The model of the form:

$$\pi(\boldsymbol{x}) = \sum_{k=1}^{K} c_k p(\boldsymbol{x}|\Theta_k)$$

is called Baysian model if $p(\boldsymbol{x}|\theta)$ is an exponetial model and:

- $\Theta_1, \ldots, \Theta_K \sim q$ where q is a prior over Θ
- $(c_1, \ldots, c_K)$ is sampled from K-dimensional dirichlet distribution.

Definition 5.23. (Posterior Baysian Mixture Model) Another intractable model that may need sampling is:

$$\hat{q}_n(c_{1:k}, \boldsymbol{\theta}|\boldsymbol{x}_{1:n}) \propto \prod_{i=1}^{n} \left(\sum_{k=1}^{K} c_k p(\boldsymbol{x}_i|\boldsymbol{\theta}_n) \right) \left(\prod_{k=1}^{K} q(\boldsymbol{\theta}_k) \right) q_{\text{dir}}(\boldsymbol{c}_{1:K})$$

individual evaluate of non-normalized $\tilde{q}$ is numerically unstable but given specific value of $c, \boldsymbol{x}$ or $\boldsymbol{\theta}$, this collapse to $\sum_{k=1}^{K} c_k p(\boldsymbol{x}_i|\boldsymbol{\theta}_k)$ making it tractable. Furthermore, please note that: we can multiple the Baysian Mixture model $\prod_{k=1}^{K} q(\boldsymbol{\theta}_k)$ with MRF prior to get smoothing effect.

Remark 53. **(Problems and Solution)** If MRF is used as prior, we have to compute or approximate the posterior distribution. The solution of MRF distribution on grids are not analytically tractable, so we have to perform sampling and inference using Markov chain sampling algorithm.

5.4 Markov Chain Monte Carlo

Definition 5.24. (MCMC) We want to sample from distribution with density p. Suppose, we can define Markov chain with invariance distribution i.e $P_{\text{inv}} \equiv P$. If we sample $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ from the chain, then once it has converged and we obtain the sample.

Definition 5.25. (Continuous MC) A continuous Markov chain is defined by an initial p_{init} and conditional probability $t(y|x)$ is transtion probability or transtion kernel. For example, markov chain on $\mathbb{R}^2$, we can define Markov chain by:

$$x_{i+1}|x_i = x_i \sim g(\cdot|x_i, \sigma_i^2)$$

where g is the spherical Gaussian with fixed variance. Suppose the state $\mathcal{X}$ is uncountable, so the transtion matrix is substuited by conditional probability t. The distribution p_{inv} with density p_{inv} is invariance if:

$$\int_{\mathcal{X}} t(y|x) p_{\text{inv}}(x) \, \mathrm{d}x = p_{\text{inv}}(y)$$

Remark 54. **(Several Problems)** There are several problems that we have to solve. We have to construct MC with invariance distribution p. We can't actually start sampling with $x_1 \sim p$ as if we know how to sample, our method would be pointless. Furthermore, each point x_i is marginally distribution as $x_i \sim p$ but the points are not iid.

Definition 5.26. (Metropolis-Hasting) We define the conditional probability $q(y|x)$ on $\mathcal{X}$. We then define a rejection kernel A as we have:

$$A(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i) = \min \left\{ 1, \frac{q(\boldsymbol{x}_i|\boldsymbol{x}_{i+1}) p(\boldsymbol{x}_{i+1})}{q(\boldsymbol{x}_{i+1}|x_i) p(\boldsymbol{x}_i)} \right\}$$

Knowing $\tilde{p}$ (unnormalized) is enough for the rejection kernel $A(\cdot|\cdot)$ as the normalizing factor is cancelled. We define the trasntion probability of the chain as:

$$t(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i) = q(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i) A(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i) + \delta_x(\boldsymbol{x}_{i+1}) c(\boldsymbol{x}_i) \quad \text{where} \quad c(\boldsymbol{x}_i) = \int q(\boldsymbol{y}|\boldsymbol{x}_i)(1 - A(\boldsymbol{y}|\boldsymbol{x}_i)) \, \mathrm{d}\boldsymbol{y}$$

To sample from the MH algorithm, at each step $i+1$, generate a proposal $\boldsymbol{x}^* \sim q(\cdot|\boldsymbol{x}_i)$ and $\boldsymbol{u}_i \sim \mathcal{U}[0,1]$:

- If $u_i > A(\boldsymbol{x}^*|\boldsymbol{x}_i)$ reject a proposal, and we set $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i$
- If $u_i \leq A(\boldsymbol{x}^*|\boldsymbol{x}_i)$, accept proposal, and we set $\boldsymbol{x}_{i+1} = \boldsymbol{x}^*$

Remark 55. **(Derivation of Metropolis-Hasting)** Now, we will consider how Metropolis-Hasting can to be. Starting with the detailed balance equation, as we have:

$$p(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i)p(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{x}_{i+1})p(\boldsymbol{x}_{i+1}) \qquad \iff \qquad \frac{p(\boldsymbol{x}_{i+1}|\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{x}_{i+1})} = \frac{p(\boldsymbol{x}_{i+1})}{p(\boldsymbol{x})}$$

Now, we will separate the transition step to be:

- Proposal Distribution $q(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i)$
- Acceptance Distribution $A(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i)$

Now, we have:

$$\frac{A(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i)}{A(\boldsymbol{x}_i|\boldsymbol{x}_{i+1})} = \frac{p(\boldsymbol{x}_{i+1})}{p(\boldsymbol{x}_i)} \frac{q(\boldsymbol{x}_i|\boldsymbol{x}_{i+1})}{q(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i)}$$

Now, if we use the Metropolis-Hasting rejection kernel, we now have:

$$\begin{aligned} & A(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i)p(\boldsymbol{x}_i)q(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i) = A(\boldsymbol{x}_i|\boldsymbol{x}_{i+1})p(\boldsymbol{x}_{i+1})q(\boldsymbol{x}_i|\boldsymbol{x}_{i+1}) \\ \iff & \min\Big\{P(\boldsymbol{x}_i)q(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i), p(\boldsymbol{x}_{i+1})q(\boldsymbol{x}_i|\boldsymbol{x}_{i+1})\Big\} = \min\Big\{q(\boldsymbol{x}_{i+1}|x_i)p(\boldsymbol{x}_i), p(\boldsymbol{x}_{i+1})q(\boldsymbol{x}_i|\boldsymbol{x}_{i+1})\Big\} \end{aligned}$$

Thus the detailed balance is satisfied.

Remark 56. There are several observations on the Metropolis-Hasting as we have:

- We accept if the second term is larger than 1: $q(\boldsymbol{x}_i|\boldsymbol{x}_{i+1})p(\boldsymbol{x}_{i+1}) > q(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i)p(\boldsymbol{x}_i)$. We accept if the proposal increases the probability under p.
- If it decreases the probability, we still accept with a probability which depends on the difference to the current probability.
- We can see this as noisy hill-climbing as it tends to move to the probability under p.
- However, there are some probability that the sampling can move down-hill with certain probability. Finally, we can show that it won't stuck at the local maxima.

Definition 5.27. (Burn-in and Mixing-Time) The first m samples are called burn-in phase. The first m samples are discarded. And, we have:

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{m-1}, \boldsymbol{x}_m, \boldsymbol{x}_{m+1}, \ldots$$

We don't know the m but we can use a certain heuristic called convergence dianostic.

Definition 5.28. (Sequential Dependence) Even after burn-in MC are not iid, we can use the following strategy as we have to consider the when the sample can be used:

- Estimate empirically how many steps L are needed for $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i+L}$ to be independent. We keep every L-th sample and discard in the between.
- The most common method uses is called auto-correlation function as we have:

  $$\text{Auto}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\mathbb{E}[(\boldsymbol{x}_i - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_j)]}{\sigma_i \sigma_j}$$

  where $\boldsymbol{\mu}_i$ is the mean and σ_i is the standard deviation of $\boldsymbol{x}_i$. We can use this auto-correlation to calculate the value L.

Definition 5.29. (Gelman-Rubin Criterion) We can also start several chain at random for chain k the sample x_i^k has the marginal of p_k^i:

- The distribution has to converge to all $p_i = p_{\text{inv}}$, which are all idetical.
- We can use hypothesis testing to compare p_i^k for difference k.
- Once the test doesn't reject anymore, assume that the chain has passed the burn-in phase

Remark 57. **(Rules for Selecting a Proposal Distribution)** Selecting a proposal distribution, we have to be aware of the tradeoff, where if $\text{var}(q)$ is too large will overstep p and leads to rejection. If $\text{var}(q)$ is too small as many steps will be needed to achives a good converge of the domain.

- If p is unimodal and can be roughly approximate by Gaussian, $\text{var}(q)$ should be choosen to be smaller than the covariance of p.
- With complex posterior, choosing q is difficult but it is important to convergence speed. If we know some information about the posterior, this might help choosing q.

There are mnay other ways to sample for example mixture of proposal.

Definition 5.30. (Gibbs Sampling) Suppose $p(\boldsymbol{x})$ is a distribution on $\mathbb{R}^d$ and so $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_D)$. The full conditional probability of the entry $\boldsymbol{x}_d$ is given by the other entries to be:

$$p(\boldsymbol{x}_d | \boldsymbol{x}_1, \dots, \boldsymbol{x}_{d-1}, \boldsymbol{x}_{d+1}, \dots, \boldsymbol{x}_D)$$

The Gibbs sampler is the special case of MH-algorithm where the proposal of $\boldsymbol{x}_d$ is the full conditional over $\boldsymbol{x}_d$

Remark 58. **(Connection to HM)** Suppose p is a distribution on $\mathbb{R}^D$ so each sampler is of the form $\boldsymbol{x}_i = (\boldsymbol{x}_{i,1}, \dots, \boldsymbol{x}_{i,D})$. We generate a proposal $\boldsymbol{x}_{i+1}$ as:

$$\begin{aligned}
\boldsymbol{x}_{i+1,1} &\sim p(\cdot | \boldsymbol{x}_{i,2}, \dots, \boldsymbol{x}_{i,D}) \\
&\vdots \\
\boldsymbol{x}_{i+1,d} &\sim p(\cdot | \boldsymbol{x}_{i+1,1}, \dots, \boldsymbol{x}_{i+1,d-1}, \boldsymbol{x}_{i,d+1}, \dots, \boldsymbol{x}_{i,D}) \\
&\vdots \\
\boldsymbol{x}_{i+1,D} &\sim p(\cdot | \boldsymbol{x}_{i,1}, \dots, \boldsymbol{x}_{i,D-1})
\end{aligned}$$

This is like Metropolis-Hasting as we use the proposal distribution and accepting probability to be 1.

Remark 59. **(MRF with Gibbs Sampling)** Consider D nodes, each dimension d and markov property:

$$\begin{aligned}
&\tilde{p}(\theta_d | \theta_1, \dots, \theta_{d-1}, \theta_{d+1}) \\
&\quad = \exp\left(\beta(\mathbb{I}\left\{\theta_d = \theta_{\text{left}}\right\} + \mathbb{I}\left\{\theta_d = \theta_{\text{right}}\right\}), \mathbb{I}\left\{\theta_d = \theta_{\text{up}}\right\}, \mathbb{I}\left\{\theta_d = \theta_{\text{down}}\right\})\right)
\end{aligned}$$

And so we can sample it.