Australian National University

COMP4670 Introduction to Machine Learning

Assignment 02

Edited by LaTeX

College of Engineering and Computer Science

©*Author*
Jimmy Lin
u5223173

¶ *Lecturer*
Christfried Webers

¶ *Tutor*
Wen Shao

† *Release Date*
April. 22 2013

‡ *Due Date*
May. 19 2013

τ *Time Spent*
40 hours

May 18, 2013

Contents

1 Sampling from a Multivariate Normal Distribution

1.1 Explain how create vectorial data conforming to multivariate Gaussian

First and foremost, we assume the **independent property** of this random number generator while producing scalar value for the vectorial data $\boldsymbol{x} \in R^n$(n is the dimensionality of the resulting data object). Formally, we describe the vectorial data $\boldsymbol{x}$ and the finally resulted data matrix $\boldsymbol{X}$ as follows,

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_m^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{11} & \cdots & x_{1n} \\ x_{21} & x_{21} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \tag{1}$$

where n is the dimensionality of each single vectorial data we aim to construct, m is the number of vectorial object we are intended to collect. Note that since the random number generator only produce scalar value, the number of times to apply this generator is $n \times m$ if the number of such n-dimensional data objects is m.

The way to construct the data matrix. For each dimension, we use the random number generator with specific μ and σ to produce a set of specific scalar value (specifically, the number of such scalar value is m for each dimension). And then we apply the column appending for each set of scalar values, as what we formualted in the (1).

In this way, we are able to formualte the following probability distribution for each dimension $i = 1, .., n$.

$$p(x_i) = \mathcal{N}(x_i|\mu_i, \sigma_i^2) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma_i} exp\Big(- \frac{(x_i - \mu_i)^2}{2\sigma_i^2}\Big) \tag{2}$$

where x_i is the i-th element of vectorial data $\boldsymbol{x}$, μ_i is the mean value of the i-th dimension's univariate gaussian distribution, and σ_i is variance of i-th dimension's univariate gaussian distribution.

Show $\boldsymbol{x}$ conforms to Multivariate Gaussian Distribution. The following content discuss the proof for the vectorial data $\boldsymbol{x}$ generated in the way we presented above conforms to multivariate gaussian diatribution.

Due to the independent assumption we have made at the beginning,

$$\begin{aligned}
p(\boldsymbol{x}) &= \prod_{i=1}^{n} p(x_i) && \text{by i.d assumption} && (3)\\
&= \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma_i} exp\Big(- \frac{(x_i - \mu_i)^2}{2\sigma_i^2}\Big) && \text{by (2)} && (4)\\
&= \frac{1}{(2\pi)^{\frac{n}{2}} \prod_{i=1}^{n}\sigma_i} exp\Big(- \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{2\sigma_i^2}\Big) && \text{put } \textstyle\prod \text{ inside} && (5)\\
&= \frac{1}{(2\pi)^{\frac{n}{2}} \prod_{i=1}^{n}\sigma_i} exp\Big(- \frac{1}{2}\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}\Big) && \text{put } \tfrac{1}{2} \text{ outside} && (6)
\end{aligned}$$

Then we define mean vector $\boldsymbol{\mu} \in R^{n\times 1}$ and covariance matrix $\mathbf{C} \in R^{n\times n}$ of the generated vectorial object to be,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix} \tag{7}$$

where μ_i is the mean value of the i-th dimension's random variable x_i and $\mathbf{C}$ is the covariance matrix of $\boldsymbol{x}$, whose entry c_{ij} , by definition, is the covariance between x_i and x_j.

Then let's focus on deriving the form of covariance matrix $\mathbf{C}$ and the specific value of each entry c_{ij}(i is the row index and j denotes column index) based on the characteristics of data collection.

First, look at the case of $i \neq j$ as follows.

$$c_{ij(i\neq j)} = \int\int \Big(x_i - E(x_i)\Big)\Big(x_j - E(x_j)\Big)P(x_i, x_j)dx_i dx_j \qquad \text{definition of covariance} \tag{8}$$

$$= \int\int \Big(x_i - E(x_i)\Big)\Big(x_j - E(x_j)\Big)P(x_i)P(x_j)dx_i dx_j \qquad \text{by i.i.d when } i \neq j \tag{9}$$

$$= \int \Big(x_i - E(x_i)\Big)P(x_i)dx_i \int \Big(x_j - E(x_j)P(x_j)\Big)dx_j \qquad \text{independence of integration} \tag{10}$$

$$= \Big(\int x_i P(x_i)dx_i - \int E(x_i)P(x_i)dx_i\Big)$$
$$\Big(\int x_j P(x_j)dx_j - \int E(x_j)P(x_j)dx_j\Big) \qquad \text{linearity of integration} \tag{11}$$

$$= \Big(\int x_i P(x_i)dx_i - E(x_i)\int P(x_i)dx_i\Big) \qquad E_{x_i} \text{ independent on } x_i$$
$$\Big(\int x_j P(x_j)dx_j - E(x_j)\int P(x_j)dx_j\Big) \qquad E_{x_j} \text{ independent on } x_j \tag{12}$$

where $P(x_i, x_j)$ is the joint probability of two independent random variables from $\boldsymbol{x}$. $P(x_i)$ is the marginal probability distribution of x_i.

By definition of the expectation, we have

$$E(x_i) = \int x_i P(x_i)dx_i \tag{13}$$

$$E(x_j) = \int x_j P(x_j)dx_j \tag{14}$$

By normalisation property of the probability distribution, we have

$$\int P(x_i)dx_i = 1 \tag{15}$$

$$\int P(x_j)dx_j = 1 \tag{16}$$

Continue the previous derivation with regard to σ_{ij} in (12), we have

$$c_{ij(i\neq j)} = \Big(\int x_i P(x_i)dx_i - E(x_i)\int P(x_i)dx_i\Big)$$
$$\Big(\int x_j P(x_j)dx_j - E(x_j)\int P(x_j)dx_j\Big) \qquad \text{from (12)} \tag{17}$$

$$= \Big(E(x_i) - E(x_i)\int P(x_i)dx_i\Big) \qquad \text{apply (13)}$$
$$\Big(E(x_j) - E(x_j)\int P(x_j)dx_j\Big) \qquad \text{apply (14)} \tag{18}$$

$$= \Big(E(x_i) - E(x_i)\Big)\Big(E(x_j) - E(x_j)\Big) \qquad \text{apply (15) and (16)} \tag{19}$$

$$= 0 \tag{20}$$

And the analysis the other case is, since the covariance of one random variable to itself is its variance (by definition). if $i = j$, we have

$$c_{ij(i=j)} = \sigma_i^2 \tag{21}$$

where the σ_i^2 is the variance of the scalar random variable x_i, which is shown in the univariate gaussian distribution (2) at the very beginning.

Summarise the c_{ij}, which is the entry of covariance matrix of multivariate gaussian distribution.

$$c_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ \sigma^2 & \text{if } i = j \end{cases} \tag{22}$$

That is to say, $\mathbf{C}$ is a diagonal matrix with its diagonal entry being $\sigma_i^2 (i = 1..n)$.

$$\mathbf{C} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} \tag{23}$$

From the form of $\mathbf{C}$ we have derived in (23), the determinant of $\mathbf{C}$ can be figured out

$$|\mathbf{C}| = \prod_{i=1}^{n} \sigma_i^2 \tag{24}$$

Obviously, all factors in (24) is non-negative, we can take square root for them

$$|\mathbf{C}|^{\frac{1}{2}} = \prod_{i=1}^{n} \sigma_i \tag{25}$$

Another point we can conclude from the diagonal property of covariance matrix $\mathbf{C}$ is its inverse. Since we assume that all the variance of univariate gaussian distribution we used here is non-zero, which means the covariance matrix $\mathbf{C}$ is full-rank and thus invertible.

$$\mathbf{C}^{-1} = \begin{pmatrix} \sigma_1^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^{-2} \end{pmatrix} \tag{26}$$

Next, present the form of $\boldsymbol{x} - \boldsymbol{\mu}$ before we turn to derive the scalar form of $(\boldsymbol{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$

$$\boldsymbol{x} - \boldsymbol{\mu} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_n - \mu_n \end{pmatrix} \tag{27}$$

where n is dimensionality of the vectorial data we wish to construct.
According to (27) and (26), we can easily have the following deduction,

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \begin{pmatrix} x_1 - \mu_1, x_2 - \mu_2, \cdots, x_n - \mu_n \end{pmatrix} \begin{pmatrix} \sigma_1^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^{-2} \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_n - \mu_n \end{pmatrix} \tag{28}$$

$$= \begin{pmatrix} \frac{x_1 - \mu_1}{\sigma_1^2}, \frac{x_2 - \mu_2}{\sigma_2^2}, \cdots, \frac{x_n - \mu_n}{\sigma_n^2} \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_n - \mu_n \end{pmatrix} \tag{29}$$

$$= \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} + \cdots + \frac{(x_n - \mu_n)^2}{\sigma_n^2} \tag{30}$$

$$= \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \tag{31}$$

Therefore, we have

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \tag{32}$$

Manipulate we have derived before for $p(\boldsymbol{x})$,

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \prod_{i=1}^{n} \sigma_i} exp\Big(-\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \Big) \qquad \text{from (6)} \tag{33}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{C}|^{\frac{1}{2}}} exp\Big(-\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \Big) \qquad \text{replace } |\mathbf{C}|^{\frac{1}{2}} \text{ by (25)} \tag{34}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{C}|^{\frac{1}{2}}} exp\Big(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \Big) \qquad \text{apply (32)} \tag{35}$$

Therefore, the probability distribution of one single object $\boldsymbol{x}$ collected by our approach, which is illustrated in the very beginning, can be represented as follows,

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{C}|^{\frac{1}{2}}} exp\Big(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \Big) \tag{36}$$

which apprently is a multivariate Gaussian distribution with two semantic parameters, one is mean vector $\boldsymbol{\mu}$ of $\boldsymbol{x}$ and the other one covariance matrix $\mathbf{C}$ of $\boldsymbol{x}$. Hence, it can be concluded that

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}) \tag{37}$$

2 Reverse Prediction

2.1 Complete optimisation problem for $\boldsymbol{w}^*$

Given the regression hypothesis,

$$\widetilde{t_i} = \boldsymbol{w}^T \boldsymbol{x_i} \tag{38}$$

where the $\widetilde{t_i}$ is prediction by our hypothesis, $\boldsymbol{w}$ is the coefficient for each dimension, and $\boldsymbol{x_i}$ is the i-th input object in given set of n training data.

Since the optimal $\boldsymbol{w}$ minimising the quadratic regression error of all data for a linear model, we can write down the optimisation problem as follows,

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{n} (\widetilde{t_i} - t_i)^2 \qquad \text{minimise quadratic training error} \tag{39}$$

$$= \arg\min_{\boldsymbol{w}} \sum_{i=1}^{n} (\boldsymbol{w}^T \boldsymbol{x_i} - t_i)^2 \qquad \text{replace by regression hypothesis} \tag{40}$$

$$= \arg\min_{\boldsymbol{w}} \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{w}^T \boldsymbol{x_i} - t_i)^2 \qquad \text{ultimate optimisation objective} \tag{41}$$

Note that the final step is valid because the newly added coefficient $\frac{1}{2}$ is not dependent on the $\boldsymbol{w}$ and doing so is to enhance the convenience for taking derivative.

2.2 Solution for $\boldsymbol{w}^*$ and condition of Uniqueness

Define error function to be the term we wish to minimise

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{w}^T \boldsymbol{x_i} - t_i)^2 \tag{42}$$

Directional derivative of $E(\boldsymbol{w})$ to zero,

$$\mathcal{D}(E(\boldsymbol{w}))(\xi) = \langle grad\ E(\boldsymbol{w}), \xi \rangle,\ \forall \xi \tag{43}$$

where ξ is defined in the space R^d.

Take gradient of the error function $E(\boldsymbol{w})$ with regard to $\boldsymbol{w}$,

$$grad\ E(\boldsymbol{w}) = \sum_{i=1}^{n} (\boldsymbol{w}^T \boldsymbol{x_i} - t_i) \boldsymbol{x_i} \tag{44}$$

From the first-order derivative (44) of error function $E(\boldsymbol{w})$, it is obvious that error funciton is convex function with regard to parameter $\boldsymbol{w}$. (this is because the $\frac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}}$ is linear to $\boldsymbol{w}^T$.) Therefore, setting the first derivative to zero would help us derive the global minima of error function.

Write the equation above (44) in the form of matrix,

$$grad\ E(\boldsymbol{w}) = \boldsymbol{X}^T (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{t}) \tag{45}$$

where, $\boldsymbol{X}$ is a matrix, each row of whom is $\boldsymbol{x_i}$, and $\boldsymbol{t}$ is a column vector whose element is t_i.

Set $E(\boldsymbol{w}) = 0$ to derive the global minimum for this convex function,

$$\boldsymbol{X}^T (\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{t}) = 0 \tag{46}$$

$$\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{X}^T \boldsymbol{t} = 0 \tag{47}$$

$$\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{X}^T \boldsymbol{t} \tag{48}$$

$$\boldsymbol{w}^* = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{t} \tag{49}$$

In the derivation from (48) to (49), we have to specify condition that $\boldsymbol{X}^T \boldsymbol{X}$ is invertible.

2.3 Optimisation problem for $\boldsymbol{u}$

The squared error for between target variable $\boldsymbol{x_i}$ and $\widetilde{\boldsymbol{x_i}}$

$$\arg\min_{\boldsymbol{u}} \frac{1}{2}\sum_{i=1}^{n} ||t_i\boldsymbol{u} - \boldsymbol{x_i}||^2 \tag{50}$$

which can be written as

$$\arg\min_{\boldsymbol{u}} \frac{1}{2}\sum_{i=1}^{n} (t_i\boldsymbol{u} - \boldsymbol{x_i})^T(t_i\boldsymbol{u} - \boldsymbol{x_i}) \tag{51}$$

2.4 Solution for $\boldsymbol{w}^*$ and condition of Uniqueness

Define the error function $E_2(\boldsymbol{u})$ as follows,

$$E_2(\boldsymbol{u}) = \frac{1}{2}\sum_{i=1}^{n} (t_i\boldsymbol{u} - \boldsymbol{x_i})^T(t_i\boldsymbol{u} - \boldsymbol{x_i}) \tag{52}$$

Directional derivative of $E_2(\boldsymbol{u})$ to zero,

$$\mathcal{D}(E_2(\boldsymbol{u}))(\xi) = \langle grad\ E(\boldsymbol{w}), \xi\rangle,\ \forall\xi \tag{53}$$

where ξ is defined to be in the space R^d.
Take gradient of $E_2(\boldsymbol{u})$, we have

$$grad\ E_2(\boldsymbol{u}) = \sum_{i=1}^{n} (t_i\boldsymbol{u} - \boldsymbol{x_i})t_i \tag{54}$$

The formula above can be written as,

$$grad\ E_2(\boldsymbol{u}) = (\boldsymbol{t}\boldsymbol{u}^T - \boldsymbol{X})^T\boldsymbol{t} \tag{55}$$

where $\boldsymbol{t}$ is the column vector, each element of whom is t_i, and $\boldsymbol{X}$ is matrix whose row is $\boldsymbol{x_i}$.
Set $grad\ E_2(\boldsymbol{u}) = 0$, we have

$$(\boldsymbol{t}(\boldsymbol{u}^*)^T - \boldsymbol{X})^T\boldsymbol{t} = 0 \tag{56}$$
$$(\boldsymbol{u}^*\boldsymbol{t}^T - \boldsymbol{X}^T))\boldsymbol{t} = 0 \tag{57}$$
$$\boldsymbol{u}^*\boldsymbol{t}^T\boldsymbol{t} - \boldsymbol{X}^T\boldsymbol{t} = 0 \tag{58}$$
$$\boldsymbol{u}^*\boldsymbol{t}^T\boldsymbol{t} = \boldsymbol{X}^T\boldsymbol{t} \tag{59}$$
$$\boldsymbol{u}^* = \boldsymbol{X}^T\boldsymbol{t}(\boldsymbol{t}^T\boldsymbol{t})^{-1} \tag{60}$$

When $\boldsymbol{t}^T\boldsymbol{t}$ is invertible, we have unique solution for $\boldsymbol{u}^*$. However, it is obvious that the $\boldsymbol{t}^T\boldsymbol{t}$ is a scalar value, more specifically, invertibility corresponds to whether that scalar is zero. Therefore, the condition for uniqueness of $\boldsymbol{u}^*$ is

$$\boldsymbol{t}^T\boldsymbol{t} \text{ is non-zero.} \tag{61}$$

which can be interpreted as

$$\sum_{i=1}^{n} t_i^2 \neq 0 \tag{62}$$

2.5 Relation between unique $\boldsymbol{w}^*$ and $\boldsymbol{u}^*$

From (49) and (60), we have

$$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t} \tag{63}$$

$$\boldsymbol{u}^* = \boldsymbol{X}^T\boldsymbol{t}(\boldsymbol{t}^T\boldsymbol{t})^{-1} \tag{64}$$

Accompanied with the following conditions for unique $\boldsymbol{w}^*$ and $\boldsymbol{u}^*$

$$\boldsymbol{X}^T\boldsymbol{X} \text{ is invertible.} \tag{65}$$

$$\boldsymbol{t}^T\boldsymbol{t} \text{ is non-zero.} \tag{66}$$

From the expression of $\boldsymbol{w}^*$ and $\boldsymbol{u}^*$, it can be concluded that

$$\boldsymbol{w}^* = \frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}}{(\boldsymbol{t}^T\boldsymbol{t})^{-1}}\boldsymbol{u}^* \tag{67}$$

which can be simplified since $(\boldsymbol{t}^T\boldsymbol{t})^{-1}$ is scalar

$$\boldsymbol{w}^* = (\boldsymbol{t}^T\boldsymbol{t})(\boldsymbol{X}^T\boldsymbol{X})^{-1} \cdot \boldsymbol{u}^* \tag{68}$$

3 Conditional Probability and Variance via Parzen Estimator

3.1 Conditional density, expectations and variance

As prescribed by the question,

$$p(x,t) = \frac{1}{N}\sum_{n=1}^{N} f(x - x_n, t - t_n) \tag{69}$$

where, the $f(\cdot)$ is bivariate gaussian distribution centered at $(0,0)$ with covariance matrix $\sigma \boldsymbol{I}^2$. Therefore, we can further extend the $p(x,t)$

$$p(x,t) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\Big(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\Big) \tag{70}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \frac{1}{2\pi\sigma^2} exp\Big(-\frac{(x - x_n)^2 + (t - t_n)^2}{2\sigma^2}\Big) \tag{71}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \frac{1}{2\pi\sigma^2} exp\Big(-\frac{(x - x_n)^2}{2\sigma^2}\Big) exp\Big(-\frac{(t - t_n)^2}{2\sigma^2}\Big) \tag{72}$$

$$= \frac{1}{2\pi\sigma^2 N}\sum_{n=1}^{N} exp\Big(-\frac{(x - x_n)^2}{2\sigma^2}\Big) exp\Big(-\frac{(t - t_n)^2}{2\sigma^2}\Big) \tag{73}$$

Note that the derivation from (70) to (71) is valid in that the x_n and t_n are both scalar value and the density estimator is a bivariate gaussian distribution. In the last step of derivation, we harness the observation that $2\pi\sigma^2$ is independent on n.

Work out the marginal probability $p(x)$,

$$p(x) = \int_t p(x,t)dt \tag{74}$$

$$= \int_t \frac{1}{2\pi\sigma^2 N}\sum_{n=1}^{N} exp\Big(-\frac{(x - x_n)^2}{2\sigma^2}\Big) exp\Big(-\frac{(t - t_n)^2}{2\sigma^2}\Big) dt \tag{75}$$

$$= \frac{1}{2\pi\sigma^2 N}\sum_{n=1}^{N} \int_t exp\Big(-\frac{(x - x_n)^2}{2\sigma^2}\Big) exp\Big(-\frac{(t - t_n)^2}{2\sigma^2}\Big) dt \tag{76}$$

$$= \frac{1}{2\pi\sigma^2 N}\sum_{n=1}^{N} exp\Big(-\frac{(x - x_n)^2}{2\sigma^2}\Big) \int_t exp\Big(-\frac{(t - t_n)^2}{2\sigma^2}\Big) dt \tag{77}$$

$$= \frac{1}{2\pi\sigma^2 N}\sum_{n=1}^{N} exp\Big(-\frac{(x - x_n)^2}{2\sigma^2}\Big) (2\pi)^{\frac{1}{2}}\sigma \tag{78}$$

$$= \frac{(2\pi)^{\frac{1}{2}}\sigma}{2\pi\sigma^2 N}\sum_{n=1}^{N} exp\Big(-\frac{(x - x_n)^2}{2\sigma^2}\Big) \tag{79}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N} exp\Big(-\frac{(x - x_n)^2}{2\sigma^2}\Big) \tag{80}$$

Note that to derive (74), sum rule is utilised. From (74) to (75), we cite the previous statement in (73). Derivation from (75) to (76) is based on linearity of integration. The validity of derivation from (76) to (77) comes from the observation that $\frac{1}{2\pi\sigma^2} exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big)$ is independent on t, on which integration is applied. The support for integration exponential term (from (77) to (78)) is presented in the following derivations. And we make use of the observation that $(2\pi)^{\frac{1}{2}}\sigma$ is independent on the summation variable n.

$$\int_t exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt = \int_t (2\pi)^{\frac{1}{2}}\sigma\frac{1}{(2\pi)^{\frac{1}{2}}\sigma}exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt \qquad \text{basic algebra} \tag{81}$$

$$= (2\pi)^{\frac{1}{2}}\sigma\int_t \frac{1}{(2\pi)^{\frac{1}{2}}\sigma}exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt \qquad \sigma \text{ independent on } t \tag{82}$$

$$= (2\pi)^{\frac{1}{2}}\sigma\cdot 1 \qquad \text{normalisation} \tag{83}$$

$$= (2\pi)^{\frac{1}{2}}\sigma \tag{84}$$

Note that from (82) to (83), we use the normalisation property of gaussian distribution (integration over the whole space of its argument equals to 1.)

Then we derive $p(t|x)$, which is the probability of target variable t conditioned on x

$$p(t|x) = \frac{p(x,t)}{p(x)} \tag{85}$$

$$= \frac{\frac{1}{2\pi\sigma^2 N}\sum_{n=1}^{N} exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big)exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)}{\frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N} exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big)} \tag{86}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\frac{\sum_{n=1}^{N} exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big)exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)}{\sum_{n=1}^{N} exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big)} \tag{87}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N}\left(\frac{exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big)}{\sum_{n=1}^{N} exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big)}\cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)\right) \tag{88}$$

Note that first step of derivation (85) above is based on product rule. Then to derive (86) we cited (73) for and (80) for $p(x)$ respectively. The most important step is the third one, which lies in the obeservation that the whole term $\sum_{n=1}^{N} exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big)$ is independent on summation variable n so that it can put into the summation on the numerator.

Then we define $k(x, x_n)$ to be,

$$k(x,x_n) = \frac{exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big)}{\sum_{n=1}^{N} exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big)} \tag{89}$$

Therefore, $p(t|x)$ can be represented with the help of $k(x, x_n)$ as follows,

$$p(t|x) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N}\left(k(x,x_n)\cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)\right) \tag{90}$$

Next derive the expectation $\mathbb{E}(t|x)$ by using probability above,

$$\mathbb{E}(t|x) = \int_t t\ p(t|x)dt \tag{91}$$

$$= \int_t t\ \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N}\left(k(x,x_n)\cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)\right)dt \tag{92}$$

$$= \int_t \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N}\left(k(x,x_n)\cdot t\cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)\right)dt \tag{93}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N}\left(\int_t k(x,x_n)\cdot t\cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)\ dt\right) \tag{94}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N}\left(k(x,x_n)\int_t t\cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)\ dt\right) \tag{95}$$

Note that we make use of the definition of expectation for starting up this series of derivations. For second step, we refer to (90) to substitute $p(t|x)$ in (92). Then since t is independent on summation variable n, we place the t into the term within summation of n. Next (94) is because of the linearity of integration operator. And since $k(x, x_n)$ independent on t we can place $k(x, x_n)$ out of the integration over t. In the following derivation, we seek to work out the integration of $t\ exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)$ over t.

$$\int_t t \cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt = \int_t (2\pi)^{\frac{1}{2}}\sigma \cdot \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \cdot t \cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt \tag{96}$$

$$= (2\pi)^{\frac{1}{2}}\sigma \int_t t \cdot \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt \tag{97}$$

$$= (2\pi)^{\frac{1}{2}}\sigma t_n \tag{98}$$

Thus, we have

$$\mathbb{E}(t|x) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N} \cdot \sum_{n=1}^{N}\Big(k(x, x_n)(2\pi)^{\frac{1}{2}}\sigma t_n\Big) \tag{99}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N} \cdot (2\pi)^{\frac{1}{2}}\sigma \cdot \sum_{n=1}^{N}\Big(k(x, x_n)t_n\Big) \tag{100}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\Big(k(x, x_n)t_n\Big) \tag{101}$$

Finally, derive the conditional variance,

$$\mathbb{VAR}(t|x) = \int_t \Big(t - \mathbb{E}(t|x)\Big)^2 p(t|x)\ dt \tag{102}$$

$$= \int_t \Big(t^2 - 2t\mathbb{E}(t|x) + \mathbb{E}^2(t|x)\Big) \cdot p(t|x)\ dt \tag{103}$$

$$= \int_t t^2\ \cdot p(t|x)\ dt + \int_t -2t\mathbb{E}(t|x) \cdot p(t|x)\ dt + \int_t \mathbb{E}^2(t|x)p(t|x)\ dt \tag{104}$$

$$= \int_t t^2\ \cdot p(t|x)\ dt - 2\mathbb{E}(t|x)\int_t t \cdot p(t|x)\ dt + \mathbb{E}^2(t|x)\int_t p(t|x)\ dt \tag{105}$$

$$= \int_t t^2\ \cdot p(t|x)\ dt - 2\mathbb{E}^2(t|x) + \mathbb{E}^2(t|x)\int_t p(t|x)\ dt \tag{106}$$

$$= \int_t t^2\ \cdot p(t|x)\ dt - 2\mathbb{E}^2(t|x) + \mathbb{E}^2(t|x) \tag{107}$$

$$= \int_t t^2\ \cdot p(t|x)\ dt - \mathbb{E}^2(t|x) \tag{108}$$

It seems that computation for first term is more involved. However, we can use some tricks to transform the integration of first term.

$$\int_t t^2\ \cdot p(t|x)\ dt = \int_t t^2 \cdot \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N}\left(k(x, x_n) \cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)\right)\ dt \tag{109}$$

$$= \int_t \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N}\left(k(x, x_n) \cdot t^2 \cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)\right)\ dt \tag{110}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N}\left(\int_t k(x, x_n) \cdot t^2 \cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)\ dt\right) \tag{111}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N}\left(k(x, x_n)\int_t t^2 \cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)\ dt\right) \tag{112}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N}\left(k(x, x_n)\int_t \Big((t-t_n)^2 + 2tt_n - t_n^2\Big) \cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)\ dt\right) \tag{113}$$

Then we focus on manipulating the integration in (113)

$$\int_t \Big((t-t_n)^2 + 2tt_n - t_n^2\Big) \cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)\ dt \tag{114}$$

$$= \int_t (t-t_n)^2 exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt + \int_t 2tt_n \cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt + \int_t -t_n^2 \cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt \tag{115}$$

$$= \int_t (t-t_n)^2 exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt + 2t_n \int_t t \cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt - t_n^2 \int_t exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt \tag{116}$$

$$= \int_t (t-t_n)^2 exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt + 2t_n \cdot \big((2\pi)^{\frac{1}{2}}\sigma t_n\big) - t_n^2((2\pi)^{\frac{1}{2}}\sigma) \tag{117}$$

$$= \int_t (t-t_n)^2 exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt + 2t_n^2(2\pi)^{\frac{1}{2}}\sigma - t_n^2(2\pi)^{\frac{1}{2}}\sigma \tag{118}$$

$$= (2\pi)^{\frac{1}{2}}\sigma^3 + t_n^2(2\pi)^{\frac{1}{2}}\sigma \tag{119}$$

$$= (2\pi)^{\frac{1}{2}}\sigma\big(\sigma^2 + t_n^2\big) \tag{120}$$

Explain the derivation for $exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)$ in above manipulation,

$$\int_t \Big(t-t_n\Big)^2 exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt = \int_t (2\pi)^{\frac{1}{2}}\sigma \frac{1}{(2\pi)^{\frac{1}{2}}\sigma}\Big(t-t_n\Big)^2 exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt \tag{121}$$

$$= (2\pi)^{\frac{1}{2}}\sigma \int_t \Big(t-t_n\Big)^2 \cdot \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)dt \tag{122}$$

$$= (2\pi)^{\frac{1}{2}}\sigma \cdot \sigma^2 \tag{123}$$

$$= (2\pi)^{\frac{1}{2}}\sigma^3 \tag{124}$$

Finally, turn back to conditional variance (108),

$$\mathbb{VAR}(t|x) = \int_t t^2\ \cdot p(t|x)\ dt - \mathbb{E}^2(t|x) \tag{125}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N}\left(k(x,x_n)\int_t \Big((t-t_n)^2 + 2tt_n - t_n^2\Big)\cdot exp\Big(-\frac{(t-t_n)^2}{2\sigma^2}\Big)\ dt\right) - \mathbb{E}^2(t|x) \tag{126}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N}\left(k(x,x_n)\cdot(2\pi)^{\frac{1}{2}}\sigma\big(\sigma^2 + t_n^2\big)\right) - \mathbb{E}^2(t|x) \tag{127}$$

$$= \frac{(2\pi)^{\frac{1}{2}}\sigma}{(2\pi)^{\frac{1}{2}}\sigma N}\sum_{n=1}^{N}\left(k(x,x_n)\big(\sigma^2 + t_n^2\big)\right) - \mathbb{E}^2(t|x) \tag{128}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(k(x,x_n)\big(\sigma^2 + t_n^2\big)\right) - \left(\frac{1}{N}\sum_{n=1}^{N}\Big(k(x,x_n)t_n\Big)\right)^2 \tag{129}$$

$$= \frac{1}{N}\sum_{n=1}^{N}k(x,x_n)\sigma^2 + \frac{1}{N}\sum_{n=1}^{N}k(x,x_n)t_n^2 - \frac{1}{N^2}\Big(\sum_{n=1}^{N}k(x,x_n)t_n\Big)^2 \tag{130}$$

As we will derive in the next question, $\sum_{n=1}^{N}k(x,x_n) = 1$ will hold. But we use it here first,

$$\mathbb{VAR}(t|x) = \frac{1}{N}\sum_{n=1}^{N}k(x,x_n)\sigma^2 + \frac{1}{N}\sum_{n=1}^{N}k(x,x_n)t_n^2 - \frac{1}{N^2}\Big(\sum_{n=1}^{N}k(x,x_n)t_n\Big)^2 \tag{131}$$

3.2 Show that the function $\sum_{n=1}^{N} k(x, x_n) = 1$

In last section, we have defined the $k(x, x_n)$ in (89) as

$$k(x, x_n) = \frac{exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big)}{\sum_{n=1}^{N} exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big)}$$

Now we start to prove that

$$\sum_{n=1}^{N} k(x, x_n) = 1. \tag{132}$$

Based on the definition, we add a summation operator over variable n

$$\sum_{n=1}^{N} k(x, x_n) = \sum_{n=1}^{N} \Big(\frac{exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big)}{\sum_{m=1}^{N} exp\Big(-\frac{(x-x_m)^2}{2\sigma^2}\Big)}\Big) \tag{133}$$

Since the denominator is one term that has executed summation over variable n, it is obvious that the denominator $\sum_{m=1}^{N} exp\Big(-\frac{(x-x_m)^2}{2\sigma^2}\Big)$ is independent on n. Thus, it is valid to pull the denominator out of the newly added summation operator as follows.

$$\sum_{n=1}^{N} k(x, x_n) = \frac{1}{\sum_{m=1}^{N} exp\Big(-\frac{(x-x_m)^2}{2\sigma^2}\Big)} \cdot \sum_{n=1}^{N} exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big) \tag{134}$$

which can be written as

$$\sum_{n=1}^{N} k(x, x_n) = \frac{\sum_{n=1}^{N} exp\Big(-\frac{(x-x_n)^2}{2\sigma^2}\Big)}{\sum_{m=1}^{N} exp\Big(-\frac{(x-x_m)^2}{2\sigma^2}\Big)} \tag{135}$$

It is obvious that the numerator and denominator are of the same value. Thus

$$\sum_{n=1}^{N} k(x, x_n) = 1 \tag{136}$$

4 Maximum Margin Hyperplane

4.1 Two points are sufficient to determine the hyperplane

We start from the optimisation objective if the maximum margin hyperplane is desired.

$$\arg\max \ \mathit{margin} \tag{137}$$

As widely known, margin is defined to be the minimal distance a set of data objects of the same class to the decision surface (hyperplane). Therefore, we have

$$\arg\max \ \min \ \mathit{distance} \tag{138}$$

Now introduce the expression for the distance of each data objects to one hyperplane from lecture notes.

$$\mathit{distance} = \frac{t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b)}{||\boldsymbol{w}||} \tag{139}$$

As derivation provided above, we have the following complete optmisation objective for the problem of maximum margin hyperplane.

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \ \min\big(\frac{t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b)}{||\boldsymbol{w}||}\big) \tag{140}$$

Then we need to tranform the acquired optimisation objective. Since the $||\boldsymbol{w}||$ is independent on n, which is the index of data objects, we can extract the norm $||\boldsymbol{w}||$ outward, and the resulted optmisation objective is to be

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \frac{1}{||\boldsymbol{w}||} \min\big(t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b)\big) \tag{141}$$

Then we simplify the problem by extracting a common factor from $\boldsymbol{w}$ and b to make the following happen,

$$\forall n, \ t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) \geq 1 \tag{142}$$

Note that the above trick is available since the case here only involves 2 points and thus the space is linearly separable for given data objects. And we can further extend the consequence of the trick as follows,

$$\min\big(t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b)\big) = 1 \tag{143}$$

In this manner, the optimisation objective is converted to be

$$\begin{aligned} \boldsymbol{w}^* &= \arg\max_{w} \frac{1}{||\boldsymbol{w}||} \\ &\text{s.t. } \forall n, \ t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) \geq 1 \end{aligned} \tag{144}$$

Further simplify the optimisation objective, we have

$$\begin{aligned} \boldsymbol{w}^* &= \arg\min_{w} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \\ &\text{s.t. } \forall n, \ t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) \geq 1 \end{aligned} \tag{145}$$

Construct Lagrangian as follows,

$$\mathcal{L} = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{n=1}^{N} a_n\big(t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) - 1\big) \tag{146}$$

$$\begin{aligned} s.t. \quad & a_n \geq 0 \\ & t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) - 1 \geq 0 \\ & a_n\big(1 - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b)\big) = 0 \end{aligned} \tag{147}$$

Given two points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, which are of different class, it is necessary to treat those two data objects to have minimal distance to the hyperplane (margin) for each class. Besides, two-point-case for SVM must be linearly separable (two points cannot be equal to each other).

$$t_1(\boldsymbol{w}^T\boldsymbol{x}_1 + b) = 1 \tag{148}$$
$$t_2(\boldsymbol{w}^T\boldsymbol{x}_2 + b) = 1 \tag{149}$$

And the corresponding Lagrangian would be

$$\mathcal{L} = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + a_1\big(1 - t_1(\boldsymbol{w}^T\boldsymbol{x}_1 + b)\big) + a_2\big(1 - t_2(\boldsymbol{w}^T\boldsymbol{x}_2 + b)\big) \tag{150}$$
$$= \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + (a_1 + a_2) - a_1t_1(\boldsymbol{w}^T\boldsymbol{x}_1 + b) - a_2t_2(\boldsymbol{w}^T\boldsymbol{x}_2 + b) \tag{151}$$

To simplify the problem, we specify the t_1 and t_2 to be +1, -1 respectively, that is

$$t_1 = 1 \tag{152}$$
$$t_2 = -1 \tag{153}$$

Therefore, the derived formula can be represented as

$$\boldsymbol{w}^T\boldsymbol{x}_1 + b = 1 \tag{154}$$
$$\boldsymbol{w}^T\boldsymbol{x}_2 + b = -1 \tag{155}$$

And the lagrangian can be simplified to be

$$\mathcal{L} = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + (a_1 + a_2) - a_1(\boldsymbol{w}^T\boldsymbol{x}_1 + b) + a_2(\boldsymbol{w}^T\boldsymbol{x}_2 + b) \tag{156}$$

Take derivative with regard to $\boldsymbol{w}^T$, and b, we have

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{w}} = \boldsymbol{w} + (a_2\boldsymbol{x}_2 - a_1\boldsymbol{x}_1) \tag{157}$$
$$\frac{\partial\mathcal{L}}{\partial b} = a_2 - a_1 \tag{158}$$
$$\frac{\partial\mathcal{L}}{\partial a_1} = 1 - \boldsymbol{w}^T x_1 - b \tag{159}$$
$$\frac{\partial\mathcal{L}}{\partial a_2} = 1 + \boldsymbol{w}^T x_1 + b \tag{160}$$

Set partial derivative of lagrangian term with regard to $\boldsymbol{w}$, b, a_1 and a_2, we have

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{w}} = 0,\ \frac{\partial\mathcal{L}}{\partial b} = 0,\ \frac{\partial\mathcal{L}}{\partial a_1} = 0,\ \frac{\partial\mathcal{L}}{\partial a_2} = 0 \tag{161}$$

Solve the set of equation, we have

$$a_1 = a_2 = \frac{2}{(\boldsymbol{x}_2 - \boldsymbol{x}_1)^T(\boldsymbol{x}_2 - \boldsymbol{x}_1)} = \frac{2}{||\boldsymbol{x}_2 - \boldsymbol{x}_1||^2} \tag{162}$$
$$\boldsymbol{w} = \frac{2(\boldsymbol{x}_1 - \boldsymbol{x}_2)}{(\boldsymbol{x}_2 - \boldsymbol{x}_1)^T(\boldsymbol{x}_2 - \boldsymbol{x}_1)} = \frac{2(\boldsymbol{x}_1 - \boldsymbol{x}_2)}{||\boldsymbol{x}_2 - \boldsymbol{x}_1||^2} \tag{163}$$
$$b = \frac{(\boldsymbol{x}_2 - \boldsymbol{x}_1)^T(\boldsymbol{x}_1 + \boldsymbol{x}_2)}{(\boldsymbol{x}_2 - \boldsymbol{x}_1)^T(\boldsymbol{x}_2 - \boldsymbol{x}_1)} = \frac{(\boldsymbol{x}_2 - \boldsymbol{x}_1)^T(\boldsymbol{x}_1 + \boldsymbol{x}_2)}{||\boldsymbol{x}_2 - \boldsymbol{x}_1||^2} \tag{164}$$

Note that $||\cdot||$ is l2-norm.

As we can see, whatever the dimension of given data objects $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is, two points are sufficient to determine the hyperplane (the $\boldsymbol{w}$ and b can be figured out).

4.2 Dimension of this maximum margin hyperplane

Obviously, the dimension of this maximum margin hyperplane is $D-1$, where the D is dimensionality of two given points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

Formally, this is because the hyperplane we found is perpendicular to the $\boldsymbol{x}_1 - \boldsymbol{x}_2$ (as shown in the above eauation, direction vector $\boldsymbol{w}$ of the hyperplane is the same as $\boldsymbol{x}_1 - \boldsymbol{x}_2$), and thus all points in this hyperplane lose the freedom in the direction $\boldsymbol{x}_1 - \boldsymbol{x}_2$. And since the given point $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is of dimension D, the dimension of hyperplane should be $D-1$.

4.3 Discriminant function $f(\boldsymbol{x})$

The discriminant function $f(\boldsymbol{x})$ can be

$$f(x) = \boldsymbol{w}^T \cdot x + b \tag{165}$$

$$= \frac{2(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T}{||\boldsymbol{x}_2 - \boldsymbol{x}_1||^2} \cdot \boldsymbol{x} + \frac{(\boldsymbol{x}_2 - \boldsymbol{x}_1)^T(\boldsymbol{x}_1 + \boldsymbol{x}_2)}{||\boldsymbol{x}_2 - \boldsymbol{x}_1||^2} \tag{166}$$

Since we only care about the sign of $f(x)$, the $f(x)$ can be further simplified by varying the scale.

$$f(x) = 2(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \cdot \boldsymbol{x} + (\boldsymbol{x}_2 - \boldsymbol{x}_1)^T(\boldsymbol{x}_1 + \boldsymbol{x}_2) \tag{167}$$

$$= (\boldsymbol{x}_2 - \boldsymbol{x}_1)^T\big(\boldsymbol{x}_1 + \boldsymbol{x}_2 - 2\boldsymbol{x}\big) \tag{168}$$

Classfication result:

$$y(\boldsymbol{x}) = \begin{cases} 1 & \text{if } (\boldsymbol{x}_2 - \boldsymbol{x}_1)^T\big(\boldsymbol{x}_1 + \boldsymbol{x}_2 - 2\boldsymbol{x}\big) > 0 \\ reject & \text{if } (\boldsymbol{x}_2 - \boldsymbol{x}_1)^T\big(\boldsymbol{x}_1 + \boldsymbol{x}_2 - 2\boldsymbol{x}\big) = 0 \\ -1 & \text{if } (\boldsymbol{x}_2 - \boldsymbol{x}_1)^T\big(\boldsymbol{x}_1 + \boldsymbol{x}_2 - 2\boldsymbol{x}\big) < 0 \end{cases} \tag{169}$$

Note that when $f(\boldsymbol{x}) = 0$, this maximum margin classifier cannot determine which class the given point $\boldsymbol{x}$ is in. Therefore, this classifier reject the decision of class of $\boldsymbol{x}$.

4.4 Discuss the values of the Lagrange parameters

From (146), we have general version of lagrangian $\mathcal{L}$

$$\begin{aligned} \mathcal{L} &= \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{n=1}^{N} a_n\big(1 - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b)\big) \\ s.t. \quad & a_n \geq 0 \\ & 1 - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) \leq 0 \\ & a_n\big(1 - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b)\big) = 0 \end{aligned} \tag{170}$$

Based on the KKT condition, we have

$$\begin{cases} a_n = 0 & \text{if } 1 - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) < 0 \\ a_n > 0 & \text{if } 1 - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) = 0 \end{cases} \tag{171}$$

Since we have known that when condition $1 - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) = 0$, the point $\boldsymbol{x}_n$ has the margin (the ones with smallest distance to hyperplane in separate space). That is, $\boldsymbol{x}_n$ is the smallest margin that affects the position and orientation of the hyperplane, a_n can be of any value that is bigger than zero. Conversely, if $\boldsymbol{x}_n$ is not the margin , a_n must be zero to ignore the value of the condition term, since the optimisation objective of our model is to find a hyperplane so that the margin is maximum (non-margin objects are not taken into consideration).

Therefore, to some extent, the functionality of lagrangian parameter a_n is to ignore points that are unnecessary to be considered.

In our problem, since there is only two points and they must be respective margin in its located space, both of these two points satisfy $1 - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) = 0$. And hence, these two points participate in determining the position and orientation of hyperplane.

5 Kernels and Feature Maps

5.1 Prove $k(\boldsymbol{x}, \boldsymbol{z})$ to be Valid Kernel in 2 Dimension

To prove $k(\boldsymbol{x}, \boldsymbol{z})$ is a valid kernel in 2 dimension, we should figure out the specific feature mapping that corresponds to the $k(\boldsymbol{x}, \boldsymbol{z})$. But first we declare

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad \boldsymbol{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \tag{172}$$

Then we start to figure out the feature map,

$$\begin{aligned}
k(\boldsymbol{x}, \boldsymbol{z}) &= (\boldsymbol{x}^T \boldsymbol{z} + 1)^3 && (173)\\
&= (x_1 z_1 + x_2 z_2 + 1)^3 && (174)\\
&= (x_1 z_1 + x_2 z_2)^3 + 3(x_1 z_1 + x_2 z_2)^2 + 3(x_1 z_1 + x_2 z_2) + 1 && (175)\\
&= x_1^3 z_1^3 + 3x_1^2 z_1^2 x_2^1 z_2^1 + 3x_1^1 z_1^1 x_2^2 z_2^2 + x_2^3 z_2^3 \\
&\quad + 3x_1^2 z_1^2 + 6x_1^1 z_1^1 x_2^1 z_2^1 + 3x_2^2 z_2^2 + 3x_1^1 z_1^1 + 3x_2^1 z_2^1 + 1 && (176)\\
&= x_1^3 \cdot z_1^3 + \sqrt{3} x_1^2 x_2 \cdot \sqrt{3} z_1^2 z_2 + \sqrt{3} x_1 x_2^2 \cdot \sqrt{3} z_1 z_2^2 + x_2^3 \cdot z_2^3 \\
&\quad + \sqrt{3} x_1^2 \cdot \sqrt{3} z_1^2 + \sqrt{6} x_1 x_2 \cdot \sqrt{6} z_1 z_2 + \sqrt{3} x_2^2 \cdot \sqrt{3} z_2^2 \\
&\quad + \sqrt{3} x_1 \cdot \sqrt{3} z_1 + \sqrt{3} x_2 \cdot \sqrt{3} z_2 + 1 \cdot 1 && (177)
\end{aligned}$$

Next, define feature map $\Phi(\boldsymbol{x})$ to be,

$$\Phi(\boldsymbol{x}) = \begin{pmatrix} x_1^3 \\ \sqrt{3} x_1^2 x_2 \\ \sqrt{3} x_1 x_2^2 \\ x_2^3 \\ \sqrt{3} x_1^2 \\ \sqrt{6} x_1 x_2 \\ \sqrt{3} x_2^2 \\ \sqrt{3} x_1 \\ \sqrt{3} x_2 \\ 1 \end{pmatrix} \tag{178}$$

Obviously, $k(\boldsymbol{x}, \boldsymbol{z})$ can be decomposed as inner product of $\Phi(\boldsymbol{x})$ and $\Phi(\boldsymbol{z})$

$$k(\boldsymbol{x}, \boldsymbol{z}) = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{z}) \rangle \tag{179}$$

where the inner product is defined to be,

$$\langle A, B \rangle = tr(A^T B) \tag{180}$$

Therefore, we can conclude that

$k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T \boldsymbol{z} + 1)^3$ is a valid kernel when input space has only two dimensions.

5.2 Generalize Input Space to Higher Dimension

Yes! We can use the same method in section 5.1 to validate the extrapolation on dimension n. The specific technique is to utilise the Mathematical Induction.

Proof Objective:

$$\forall n \geq 2,\ dim(\boldsymbol{x}) = dim(\boldsymbol{z}) = n,\ k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T\boldsymbol{z} + 1)^3 \ is\ valid\ kernel. \tag{181}$$

where the $dim()$ operator returns the dimensionality of its argument.

Base case:
the statement that kernel $k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T\boldsymbol{z} + 1)^3$ is valid for $n = 2$ has been proven in (179).

Inductive cases: Assume that the

$$k_M(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T\boldsymbol{z} + 1)^3,\ dim(\boldsymbol{x}) = dim(\boldsymbol{z}) = M\ is\ valid\ kernel\ function. \tag{182}$$

Prove that

$$k_{M+1}(\boldsymbol{x}^*, \boldsymbol{z}^*) = ((\boldsymbol{x}^*)^T\boldsymbol{z}^* + 1)^3,\ is\ also\ valid\ kernel\ function. \tag{183}$$

where $\boldsymbol{x}^* = (\boldsymbol{x}, x_{M+1}),\ \boldsymbol{z}^* = (\boldsymbol{z}, z_{M+1})$.

Proof for Inductive Cases:
First of all, extend $k_M(\boldsymbol{x}, \boldsymbol{z})$,

$$k_M(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T\boldsymbol{z} + 1)^3 \tag{184}$$

$$= \big((\sum_{i=1}^{M} x_i z_i) + 1\big)^3 \tag{185}$$

$$= (\sum_{i=1}^{M} x_i z_i)^3 + 3(\sum_{i=1}^{M} x_i z_i)^2 + 3(\sum_{i=1}^{M} x_i z_i)^1 + 1 \tag{186}$$

And from the assumption of inductive cases (182), we have,

$$\exists \Phi(\cdot),\ such\ that\ \Phi_M^T(\boldsymbol{x})\Phi_M(\boldsymbol{z}) = k_M(\boldsymbol{x}, \boldsymbol{z}) \tag{187}$$

Then use $\Phi_M(\cdot)$ to denote the $\Phi(\cdot)$ above in the extension of $k_M(\boldsymbol{x}, \boldsymbol{z})$

$$k_M(\boldsymbol{x}, \boldsymbol{z}) = (\sum_{i=1}^{M} x_i z_i)^3 + 3(\sum_{i=1}^{M} x_i z_i)^2 + 3(\sum_{i=1}^{M} x_i z_i)^1 + 1 = \Phi_M^T(\boldsymbol{x})\Phi_M(\boldsymbol{z}) \tag{188}$$

Now, start to manipulate $k_{M+1}(\boldsymbol{x}^*, \boldsymbol{z}^*)$ as follows,

$$k_{M+1}(\boldsymbol{x}^*, \boldsymbol{z}^*) = \big((\sum_{i=1}^{M+1} x_i z_i) + 1\big)^3 \tag{189}$$

$$= \big((\sum_{i=1}^{M} x_i z_i) + (x_{M+1}z_{M+1} + 1)\big)^3 \tag{190}$$

$$= (\sum_{i=1}^{M} x_i z_i)^3 + 3(\sum_{i=1}^{M} x_i z_i)^2(x_{M+1}z_{M+1} + 1) + 3(\sum_{i=1}^{M} x_i z_i)(x_{M+1}z_{M+1} + 1)^2 + (x_{M+1}z_{M+1} + 1)^3 \tag{191}$$

$$= \boxed{(\textstyle\sum_{i=1}^{M} x_i z_i)^3} + 3(\sum_{i=1}^{M} x_i z_i)^2 x_{M+1}z_{M+1} + \boxed{3(\textstyle\sum_{i=1}^{M} x_i z_i)^2} + 3(\sum_{i=1}^{M} x_i z_i)(x_{M+1}z_{M+1})^2 + 6(\sum_{i=1}^{M} x_i z_i)(x_{M+1}z_{M+1}) + \boxed{3(\textstyle\sum_{i=1}^{M} x_i z_i)} + (x_{M+1}z_{M+1})^3 + 3(x_{M+1}z_{M+1})^2 + 3x_{M+1}z_{M+1} + \boxed{1} \tag{192}$$

Next, replace the sum of boxed terms in (192) by using (188), we have the followings,

$$\begin{aligned} k_{M+1}(\boldsymbol{x}^*, \boldsymbol{z}^*) = & \overbrace{\boxed{\Phi_M^T(\boldsymbol{x})\Phi_M(\boldsymbol{z})}}^{\Phi_M(\boldsymbol{x})\cdot\Phi_M(\boldsymbol{z})} + \overbrace{\boxed{3(\textstyle\sum_{i=1}^M x_i z_i)^2 x_{M+1}z_{M+1}}}^{\sqrt{3}\Phi^*_{(M+1)}(\boldsymbol{x}^*)\cdot\sqrt{3}\Phi^*_{(M+1)}(\boldsymbol{z}^*)} + \overbrace{\boxed{3(\textstyle\sum_{i=1}^M x_i z_i)(x_{M+1}z_{M+1})^2}}^{\sqrt{3}\Phi^{**}_{(M+1)}(\boldsymbol{x}^*)\cdot\sqrt{3}\Phi^{**}_{(M+1)}(\boldsymbol{z}^*)} \\ & + \underbrace{\boxed{6(\textstyle\sum_{i=1}^M x_i z_i)(x_{M+1}z_{M+1})}}_{\sqrt{6}\Phi^{***}_{(M+1)}(\boldsymbol{x}^*)\cdot\sqrt{6}\Phi^{***}_{(M+1)}(\boldsymbol{z}^*)} + \underbrace{\boxed{(x_{M+1}z_{M+1})^3}}_{x^3_{M+1}\cdot z^3_{M+1}} + \underbrace{\boxed{3(x_{M+1}z_{M+1})^2}}_{\sqrt{3}x^2_{M+1}\cdot\sqrt{3}z^2_{M+1}} + \underbrace{\boxed{3x_{M+1}z_{M+1}}}_{\sqrt{3}x_{M+1}\cdot\sqrt{3}z_{M+1}} \end{aligned} \tag{193}$$

In above formula, the decomposition of each term is presented. For safety, we also write down the formal representation as follows,

$$\Phi_M^T(\boldsymbol{x})\Phi_M(\boldsymbol{z}) = \Phi_M(\boldsymbol{x})\cdot\Phi_M(\boldsymbol{z}) \tag{194}$$

$$3(\sum_{i=1}^M x_i z_i)^2 x_{M+1}z_{M+1} = \sqrt{3}\Phi^*_{(M+1)}(\boldsymbol{x}^*)\cdot\sqrt{3}\Phi^*_{(M+1)}(\boldsymbol{z}^*) \tag{195}$$

$$3(\sum_{i=1}^M x_i z_i)(x_{M+1}z_{M+1})^2 = \sqrt{3}\Phi^{**}_{(M+1)}(\boldsymbol{x}^*)\cdot\sqrt{3}\Phi^{**}_{(M+1)}(\boldsymbol{z}^*) \tag{196}$$

$$6(\sum_{i=1}^M x_i z_i)(x_{M+1}z_{M+1}) = \sqrt{6}\Phi^{***}_{(M+1)}(\boldsymbol{x}^*)\cdot\sqrt{6}\Phi^{***}_{(M+1)}(\boldsymbol{z}^*) \tag{197}$$

$$(x_{M+1}z_{M+1})^3 = x^3_{M+1}\cdot z^3_{M+1} \tag{198}$$

$$3(x_{M+1}z_{M+1})^2 = \sqrt{3}x^2_{M+1}\cdot\sqrt{3}z^2_{M+1} \tag{199}$$

$$3x_{M+1}z_{M+1} = \sqrt{3}x_{M+1}\cdot\sqrt{3}z_{M+1} \tag{200}$$

where $\Phi^*_{(M+1)}$, $\Phi^{**}_{(M+1)}$, $\Phi^{***}_{(M+1)}$ are defined as follows,

$$\Phi^*_{(M+1)}(\boldsymbol{x}^*) = \begin{pmatrix} x_1x_1x_{M+1} \\ x_1x_2x_{M+1} \\ \vdots \\ x_1x_jx_{M+1} \\ x_1x_{j+1}x_{M+1} \\ \vdots \\ x_1x_Mx_{M+1} \\ x_2x_1x_{M+1} \\ \vdots \\ x_ix_jx_{M+1} \\ x_ix_{j+1}x_{M+1} \\ \vdots \\ x_ix_Mx_{M+1} \\ x_{i+1}x_1x_{M+1} \\ \vdots \\ x_Mx_Mx_{M+1} \end{pmatrix} \tag{201}$$

$$\Phi^{**}_{(M+1)}(\boldsymbol{x}^*) = \begin{pmatrix} x_1x^2_{M+1} \\ x_2x^2_{M+1} \\ \vdots \\ x_ix^2_{M+1} \\ x_{i+1}x^2_{M+1} \\ \vdots \\ x_Mx^2_{M+1} \end{pmatrix} \tag{202}$$

$$\Phi^{***}_{(M+1)}(\boldsymbol{x}^*) = \begin{pmatrix} x_1x_{M+1} \\ x_2x_{M+1} \\ \vdots \\ x_ix_{M+1} \\ x_{i+1}x_{M+1} \\ \vdots \\ x_Mx_{M+1} \end{pmatrix} \tag{203}$$

Therefore, we can figure out the feature mapping Φ_{M+1}, such that the following is satisfied,

$$k_{M+1}(\boldsymbol{x}^*, \boldsymbol{z}^*) = \Phi^T_{M+1}(\boldsymbol{x}^*)\Phi_{M+1}(\boldsymbol{z}^*) \tag{204}$$

where the specific feature mapping is of the form as follows,

$$\Phi_{M+1}(\boldsymbol{x}^*) = \begin{pmatrix} \Phi_M(\boldsymbol{x}) \\ \sqrt{3}\Phi^*_{(M+1)}(\boldsymbol{x}^*) \\ \sqrt{3}\Phi^{**}_{(M+1)}(\boldsymbol{x}^*) \\ \sqrt{6}\Phi^{***}_{(M+1)}(\boldsymbol{x}^*) \\ x^3_{M+1} \\ \sqrt{3}x^2_{M+1} \\ \sqrt{3}x_{M+1} \end{pmatrix} \tag{205}$$

Hence, $k_{M+1}(\boldsymbol{x}^*, \boldsymbol{z}^*)$ is also a valid function since we have proven that

$$\exists \Phi_{M+1}(\boldsymbol{x}^*),\ \textit{such that}\ k_{M+1}(\boldsymbol{x}^*, \boldsymbol{z}^*) = \langle \Phi_{M+1}(\boldsymbol{x}^*),\ \Phi_{M+1}(\boldsymbol{z}^*) \rangle \tag{206}$$

where the definition of inner product follows (180).

Hence, we successfully prove the inductive part (183). Based on the theory of Mathematical Induction, it is concluded that

$$\forall n \geq 2,\ dim(\boldsymbol{x}) = dim(\boldsymbol{z}) = n,\ k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T\boldsymbol{z} + 1)^3\ \textit{is valid kernel.} \tag{207}$$

5.3 Generalize Kernel to Higher Power

Still Yes. We can still prove the validity of generalized kernel by finding feature map. The technique for proving generalization on p is rather similiar to section 5.2.

Proof Objective:

$$\forall p \geq 3, k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T \boldsymbol{z} + 1)^p \ \textit{is a valid kernel.} \tag{208}$$

Base Case: $p = 3$
For the case of dimension $n = 1$, it is obvious that the $k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}\boldsymbol{z} + 1)^3$ is a valid kernel. We provide a quick proof in the followings.

$$k(x, z) = (xz + 1)^3 = x^3 z^3 + 3x^2 z^2 + 3xz + 1 \tag{209}$$

We can define feature mapping $\boldsymbol{\Phi}(x)$ to be

$$\boldsymbol{\Phi}(x) = \begin{pmatrix} x^3 \\ \sqrt{3}x^2 \\ \sqrt{3}x \\ 1 \end{pmatrix} \tag{210}$$

In this manner, we can represent $k(x, z)$ as inner product of two feature mapping.

$$k(x, z) = \langle \boldsymbol{\Phi}(x),\ \boldsymbol{\Phi}(z) \rangle \tag{211}$$

where the definition of inner product follows (180).
Thus, the statement that $k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}\boldsymbol{z} + 1)^3$ is a valid kernel for dimension $n = 1$ holds.

And the statement that $k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T \boldsymbol{z} + 1)^3$ is valid kernel for dimensions $n \geq 2$ has been proven in last section 5.2.
Therefore, it can be conclude that for any positive integer dimension of $\boldsymbol{x}$ and $\boldsymbol{z}$, where $dim(\boldsymbol{x}) = dim(\boldsymbol{z}) = n \geq 0$, we have $k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T \boldsymbol{z} + 1)^3$ is valid kernel.

Inductive Cases:
Assume that

$$k_Q(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T \boldsymbol{z} + 1)^Q \ \textit{is valid kernel function} \tag{212}$$

Prove that

$$k_{Q+1}(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T \boldsymbol{z} + 1)^{Q+1} \ \textit{is also valid kernel function} \tag{213}$$

Proof for Inductive Cases:
Start from the assumption (212) and the definition of valid kernel,

$$\exists \Phi(\boldsymbol{x}),\ \textit{such that } k_Q(\boldsymbol{x}, \boldsymbol{z}) = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{z}) \rangle = \Phi^T(\boldsymbol{x})\Phi(\boldsymbol{z}) \tag{214}$$

Since the $k_M(\boldsymbol{x}, \boldsymbol{z})$ we talk about has the form,

$$k_Q(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T \boldsymbol{z} + 1)^Q = (\sum_{i=1}^{n} x_i z_i + 1)^Q \tag{215}$$

where n is the dimensionality of vector $\boldsymbol{x}$ and $\boldsymbol{z}$.
We can derive the feature map for $k_Q(\boldsymbol{x}, \boldsymbol{z})$, denoted by Φ_Q as follows,

$$(\Phi_Q(\boldsymbol{x}))^T \Phi_Q(\boldsymbol{z}) = (\sum_{i=1}^{n} x_i z_i + 1)^Q \tag{216}$$

Then start to manipulate the $k_{Q+1}(\boldsymbol{x}, \boldsymbol{z})$,

$$k_{Q+1}(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T \boldsymbol{z} + 1)^{Q+1} \tag{217}$$

$$= (\boldsymbol{x}^T \boldsymbol{z} + 1)^{Q} \ (\boldsymbol{x}^T \boldsymbol{z} + 1) \tag{218}$$

$$= \underbrace{(\Phi_Q(\boldsymbol{x}))^T \Phi_Q(\boldsymbol{z}) \left(\sum_{i=1}^{n} x_i z_i \right)}_{x_i \Phi_Q(\boldsymbol{x}) \cdot z_i \Phi_Q(\boldsymbol{z})} + \underbrace{(\Phi_Q(\boldsymbol{x}))^T \Phi_Q(\boldsymbol{z})}_{\Phi_Q(\boldsymbol{x}) \cdot \Phi_Q(\boldsymbol{z})} \tag{219}$$

Note that the derivation from (218) to (219) make use of what we have already derived formula (216).

Obviously, we can derive the feature map $\Phi_{Q+1}(\boldsymbol{x})$ for $k_{Q+1}(\boldsymbol{x}, \boldsymbol{z})$, that is,

$$\Phi_{Q+1}(\boldsymbol{x}) = \begin{pmatrix} x_1 \Phi_Q(\boldsymbol{x}) \\ \vdots \\ x_i \Phi_Q(\boldsymbol{x}) \\ \vdots \\ x_n \Phi_Q(\boldsymbol{x}) \\ \Phi_Q(\boldsymbol{x}) \end{pmatrix} \tag{220}$$

Such that,

$$k_{Q+1}(\boldsymbol{x}, \boldsymbol{z}) = (\Phi_{Q+1}(\boldsymbol{x}))^T \Phi_{Q+1}(\boldsymbol{z}) = \langle \Phi_{Q+1}(\boldsymbol{x}), \Phi_{Q+1}(\boldsymbol{z}) \rangle \Rightarrow \ \textit{valid kernel} \tag{221}$$

where the definition of inner product follows (180).

Note that each term $x_i \Phi_Q(\boldsymbol{x})$ $(i = 1..n)$ in (220) represents one vector whose element is multiplication of x_i and one corresponding element of original feature mapping $\Phi_Q(\boldsymbol{x})$. And the resulted feature mapping $\Phi_{Q+1}(\boldsymbol{x})$ can be regarded as a huge vector formed by stacking the shown set of vectors in (220).

Therefore, we successfully prove the inductive cases (213). Based on the theory of Mathematical Induction, since base case and inductive cases are proven, it is concluded that

$$\forall p \geq 3, k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T \boldsymbol{z} + 1)^p \ \textit{is a valid kernel.} \tag{222}$$

6 EM for Trees

6.1 Define latent variables and a joint generative model

In this case, the definition of latent variable $\boldsymbol{z} = \{0,1\}^k, k = \{1,2,..,K\}$, represents the real tree from which one measurement $\boldsymbol{x}_n$ was generated (put it another way, z_k means the measurement corresponds to the tree whose index is k). We model the tree one laser comes from as latent variable because this is not directly observed but are rather inferred.

Note that here we use the 1-of-k coding scheme for $\boldsymbol{z}$, that is

$$z_k = 1 \ or \ 0 \tag{223}$$

$$\forall \boldsymbol{z}, \ \sum_{k=1}^{K} z_k = 1 \tag{224}$$

The joint generative model of the observations and latent variables given the parameters $p(X, Z)$ means under current distribution of trees in this forest, the likelihood of collecting set of measurements X and such measurements to one certain tree.

Then we provide the derivation of joint distribution $p(\boldsymbol{x}_n, \boldsymbol{z}_n)$.

Based on the probability distribution of measurement $\boldsymbol{x}_n$ conditioned on z_{nk},

$$p(\boldsymbol{x}_n | z_{nk} = 1) = \mathcal{N}(\boldsymbol{\mu}_k | \sigma^2 \boldsymbol{I}) \tag{225}$$

Since we use 1-of-k coding scheme, at all states of $\boldsymbol{z}_n$, only one element z_{nk} can be one while others are zeros. That is, no measurement is from two trees, which conforms to our assumption in the question that the trees have a small diameter and thus the offset of measurement from the true center is negligible. Therefore, we can repreent $p(\boldsymbol{x}_n | \boldsymbol{z}_n)$ in the way as follows,

$$p(\boldsymbol{x}_n | \boldsymbol{z}_n) = \prod_{k=1}^{K} \left(\mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}) \right)^{z_{nk}} \tag{226}$$

Similarly according to the 1-of-k coding scheme, set the $p(z_k = 1) = \pi_k$, in order to derive $p(\boldsymbol{z})$, we have

$$p(\boldsymbol{z}_n) = \prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{227}$$

Consider the joint distribution of laser-range measurement $p(\boldsymbol{x}_n, \boldsymbol{z}_n)$

$$p(\boldsymbol{x}_n, \boldsymbol{z}_n) = \ p(\boldsymbol{z}_n) \ p(\boldsymbol{x}_n | \boldsymbol{z}_n) \qquad \text{product rule} \tag{228}$$

$$= \prod_{k=1}^{K} \pi_k^{z_{nk}} \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})^{z_{nk}} \qquad \text{cite (226) and (227)} \tag{229}$$

$$= \prod_{k=1}^{K} \left(\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}) \right)^{z_{nk}} \qquad \text{abstract the multiplication} \tag{230}$$

Again, based on the 1-of-k coding scheme, for each measurement $\boldsymbol{x}_n$, the corresponding latent variable has only one element $z_{nk} = 1$ and others $z_{nk} = 0$. Therefore, all terms except one that $z_{nk} = 1$ is one because the power is zero. We can easily work out the multiplication over k and derive the following representation of $p(\boldsymbol{x}_n, \boldsymbol{z}_n)$.

$$p(\boldsymbol{x}_n, \boldsymbol{z}_n) = 1 \times 1 \cdots \times \left(\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}) \right) \times \cdots \times 1 \tag{231}$$

$$= \pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}) \tag{232}$$

where we use k to denote always the non-zero element index of $\boldsymbol{z}$.

Based on the sum rule, we derive the marginal probability of $\boldsymbol{x}_n$,

$$p(\boldsymbol{x}_n) = \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}) \tag{233}$$

6.2 Derive the expected log likelihood of the joint model

To derive the expected log likelihood, we start from the log likelihood of the whole collected data sets $\boldsymbol{X}$, each row of whom is $\boldsymbol{x}_n$.

$$ln\ p(\boldsymbol{X}|\theta) = ln \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta) \qquad \text{by i.i.d} \tag{234}$$

$$= \sum_{n=1}^{N} ln\ p(\boldsymbol{x}_n|\theta) \qquad \text{put log function in} \tag{235}$$

Next, to further extend the derivation shown above, we have to apply the following series of tricks.

$$p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta) = p(\boldsymbol{x}_n|\theta) \cdot p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta) \tag{236}$$

$$ln\ p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta) = ln\ p(\boldsymbol{x}_n|\theta) + ln\ p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta) \tag{237}$$

$$ln\ p(\boldsymbol{x}_n|\theta) = ln\ p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta) - ln\ p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta) \tag{238}$$

Multiply for both side a distribution $q(\boldsymbol{z}_n)$, for which $\sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n) = 1$.

$$q(\boldsymbol{z}_n) \cdot ln\ p(\boldsymbol{x}_n|\theta) = q(\boldsymbol{z}_n) \cdot ln\ p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta) - q(\boldsymbol{z}_n) \cdot ln\ p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta) \tag{239}$$

And summing over $\boldsymbol{z}_n$ for both side.

$$\sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ p(\boldsymbol{x}_n|\theta)\Big) = \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta)\Big) - \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta)\Big) \tag{240}$$

Since $ln\ p(\boldsymbol{x}_n|\theta)$ is independent on $\boldsymbol{z}_n$, the left-hand-side term in above equation can be simplified as follows,

$$\sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ p(\boldsymbol{x}_n|\theta)\Big) = \Big(\sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n)\Big) \cdot ln\ p(\boldsymbol{x}_n|\theta) = ln\ p(\boldsymbol{x}_n|\theta) \tag{241}$$

Note that in above derivation, we use the normalisation property of $\sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n) = 1$.
Take the simplified term back to equation (240), we have

$$ln\ p(\boldsymbol{x}_n|\theta) = \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta)\Big) - \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta)\Big) \tag{242}$$

$$= \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta)\Big) - \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta)\Big) + \sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n) ln\ q(\boldsymbol{z}_n) - \sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n) ln\ q(\boldsymbol{z}_n) \tag{243}$$

$$= \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta) - q(\boldsymbol{z}_n) ln\ q(\boldsymbol{z}_n)\Big) - \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta) - q(\boldsymbol{z}_n) ln\ q(\boldsymbol{z}_n)\Big) \tag{244}$$

$$= \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta)}{q(\boldsymbol{z}_n)}\Big) - \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ \frac{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta)}{q(\boldsymbol{z}_n)}\Big) \tag{245}$$

Then put the subtraction into the term within the summation, we have

$$ln\ p(\boldsymbol{x}_n|\theta) = \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta)}{q(\boldsymbol{z}_n)}\Big) + \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ \frac{q(\boldsymbol{z}_n)}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta)}\Big) \tag{246}$$

Take above equation (246) into what we have previously derived (235),

$$ln\ p(\boldsymbol{X}|\theta) = \sum_{n=1}^{N} \left(\sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta)}{q(\boldsymbol{z}_n)}\Big) + \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ \frac{q(\boldsymbol{z}_n)}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta)}\Big) \right) \tag{247}$$

$$= \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta)}{q(\boldsymbol{z}_n)}\Big) + \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\ \frac{q(\boldsymbol{z}_n)}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta)}\Big) \tag{248}$$

Obviously, the second term is KL divergence between $q(\boldsymbol{z}_n)$ and $p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta)$. As we all know, KL divergence is always greater than zero, that is

$$KL\Big(q(\boldsymbol{z}_n) \;||\; p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta)\Big) = \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\; \frac{q(\boldsymbol{z}_n)}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta)}\Big) \geq 0 \tag{249}$$

That is to say, the first term is always lower bound of $ln\; p(\boldsymbol{X}|\theta)$. Formally, we have

$$ln\; p(\boldsymbol{X}|\theta) \geq \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n) \cdot ln\; \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta)}{q(\boldsymbol{z}_n)}\Big) \tag{250}$$

Now we apply the Expectation-Maximisation algorithm to derive the expected complete log likelihood. We first fix θ and find smallest KL divergence in E-step to maximise the lower bound, that is, based on the property of KL divergence, to find a distribution $q(\boldsymbol{z}_n)$ equaling to distribution $p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta)$, in which case the KL divergence is zero. And then in the M-step, we fix found $q(\boldsymbol{z}_n)$ and maximise the lower bound with regard to parameter θ.

That is to say, in each M-step, we have

$$q(\boldsymbol{z}_n) = p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta^{old}) \tag{251}$$

Now we turn to the optimisation objective in M-step. The optimisation of lower bound with regard to θ can be written as,

$$\arg\max_{\theta} \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n)) \cdot ln\; \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta)}{q(\boldsymbol{z}_n)}\Big) \tag{252}$$

That is to optimise the following objective

$$\arg\max_{\theta} \left(\sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n)) \cdot ln\; p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta)\Big) - \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n)) \cdot ln\; q(\boldsymbol{z}_n)\Big)\right) \tag{253}$$

Since $q(\boldsymbol{z}_n)$ is not dependent on θ, the optimisation objective above can be rewritten as

$$\arg\max_{\theta} \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} \Big(q(\boldsymbol{z}_n)) \cdot ln\; p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta)\Big) \tag{254}$$

Based on equation (251), we derived optimisation objective for M-step as follows.

$$\arg\max_{\theta} \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} \Big(p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta^{old}) \cdot ln\; p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta)\Big) \tag{255}$$

As we can see, based on the definition of expectation, the term to be optimised above is exactly the expected log likelihood of this joint model.

$$Q(\theta, \theta^{old}) = \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} \Big(p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta^{old}) \cdot ln\; p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta)\Big) \tag{256}$$

6.3 Derive the E-step update and show how to compute expectations

First from the (248), we have

$$ln\ p(\boldsymbol{X}|\theta) = \sum_{n=1}^{N}\sum_{\boldsymbol{z}_n}\Big(q(\boldsymbol{z}_n)\cdot ln\ \frac{p(\boldsymbol{x}_n,\boldsymbol{z}_n|\theta)}{q(\boldsymbol{z}_n)}\Big) + \sum_{n=1}^{N}\sum_{\boldsymbol{z}_n}\Big(q(\boldsymbol{z}_n)\cdot ln\ \frac{q(\boldsymbol{z}_n)}{p(\boldsymbol{z}_n|\boldsymbol{x}_n,\theta)}\Big) \tag{257}$$

In the Expectation-step, we are gonna to minimise the KL divergence as follows. As widely known, the lowest value of KL divergence is zero if and only if the $q(\boldsymbol{z}_n) = p(\boldsymbol{z}_n|\boldsymbol{x}_n)$.

$$KL\Big(q(\boldsymbol{z}_n)\ ||\ p(\boldsymbol{z}_n|\boldsymbol{x}_n,\theta)\Big) = \sum_{\boldsymbol{z}_n}\Big(q(\boldsymbol{z}_n)\cdot ln\ \frac{q(\boldsymbol{z}_n)}{p(\boldsymbol{z}_n|\boldsymbol{x}_n,\theta)}\Big) \geq 0 \tag{258}$$

Next we are going to derive the $q(\boldsymbol{z}_n)$ by working out the form of $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$. But first we should derive the expression of marginal distribution $p(\boldsymbol{x}_n)$ and

To derive the Expectation step, then we manipulate the $p(\boldsymbol{z}|\boldsymbol{x})$, which is the expected probability of one data object $\boldsymbol{x}$ being assigned to one cluster (tree in this case), given the observation pattern (laser measurement in this case).

$$p(\boldsymbol{z}_n|\boldsymbol{x}_n) = \frac{p(\boldsymbol{z}_n,\boldsymbol{x}_n)}{p(\boldsymbol{x}_n)} \qquad \text{by product rule} \tag{259}$$

$$= \frac{\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\sigma^2\boldsymbol{I})}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\sigma^2\boldsymbol{I})} \qquad \text{cite (232) and (233)} \tag{260}$$

where the π_k, which is $p(z_{nk}=1)$, can be treated as the coefficient (significance) of k-th mixture of gaussian component. And the specific form of $\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\sigma^2\boldsymbol{I})$ is as follows,

$$\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\sigma^2\boldsymbol{I}) = \frac{1}{(2\pi)^{\frac{D}{2}}\sigma^D}exp\Big(-\frac{(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T(\boldsymbol{x}_n-\boldsymbol{\mu}_k)}{2\sigma^2}\Big) \tag{261}$$

Now we figure out the form of $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ and thus the $q(\boldsymbol{z}_n)$ that maximise the lower bound for each data object n can be computed.

By convention, we utilise the notation γ_{nk} to denote the probability of one unlabel measurement $\boldsymbol{x}_n$ being attributed to certain tree $\boldsymbol{z}_n$ given the value of measurement $\boldsymbol{x}_n$, that is

$$\gamma_{nk} \overset{def}{=} p(\boldsymbol{z}_n|\boldsymbol{x}_n) \tag{262}$$

By (260) to replace the $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$, we have

$$\gamma_{nk} = \frac{\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\sigma^2\boldsymbol{I})}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\sigma^2\boldsymbol{I})} \tag{263}$$

The procedure of computing expectation of responsibilities.

- For start-up., if no given knowledge of approximate location of each tree (no preknowledge of parameters of gaussian components), randomly initialise the parameters of all k gaussian components presented above.
- E-step: Then tor each unlabelled measurement $\boldsymbol{x}_n$, we should figure out its probability of coming from all K tree components (all possible state of $\boldsymbol{z}$), that is, the responsibility of each component to explain currently treated measurement $\boldsymbol{x}_n$. In this step, we ought to have $N \times K$ times of computation for this responsibility.

6.4 Derive the M-step for EM

In the solution to 6.2, we have figured out in advance the optimisation objective of the M-step, which is to maximise the following term

$$Q(\theta, \theta^{old}) = \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} \Big(p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta^{old}) \cdot ln\ p(\boldsymbol{x}_n, \boldsymbol{z}_n|\theta) \Big) \tag{264}$$

This can be rewritten as the follows based on (232)

$$Q(\theta, \theta^{old}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \Big(\gamma_{nk} \cdot ln\ (\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})) \Big) \tag{265}$$

This is valid since we have figured out the expression of $p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta^{old})$ in 6.3 and that probability term has be computed in E-step, the $p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta^{old})$ is a constant in M-step. Hence, we just use γ_{nk} to represent $p(\boldsymbol{z}_n|\boldsymbol{x}_n, \theta^{old})$.

Manipulate the partial derivative of Q with regard to $\boldsymbol{\mu}_k$

$$\frac{\partial \sum_{n=1}^{N} \sum_{k=1}^{K} \Big(\gamma_{nk} \cdot ln\ (\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})) \Big)}{\partial \boldsymbol{\mu}_k} \tag{266}$$

$$= \sum_{n=1}^{N} \frac{\partial \sum_{k=1}^{K} \Big(\gamma_{nk} \cdot ln\ (\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})) \Big)}{\partial \boldsymbol{\mu}_k} \tag{267}$$

$$= \sum_{n=1}^{N} \frac{\partial \Big(\gamma_{nk} \cdot ln\ (\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})) \Big)}{\partial \boldsymbol{\mu}_k} \tag{268}$$

$$= \sum_{n=1}^{N} \gamma_{nk} \cdot \frac{\partial ln\ (\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}))}{\partial \boldsymbol{\mu}_k} \tag{269}$$

$$= \sum_{n=1}^{N} \gamma_{nk} \cdot \frac{1}{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})} \cdot \frac{\partial (\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}))}{\partial \boldsymbol{\mu}_k} \tag{270}$$

$$= \sum_{n=1}^{N} \gamma_{nk} \cdot \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})}{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})} \cdot \frac{\partial \Big(- \frac{(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_n - \boldsymbol{\mu}_k)}{2\sigma^2} \Big)}{\partial \boldsymbol{\mu}_k} \tag{271}$$

$$= \sum_{n=1}^{N} \gamma_{nk} \cdot \frac{\partial \Big(- \frac{(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_n - \boldsymbol{\mu}_k)}{2\sigma^2} \Big)}{\partial \boldsymbol{\mu}_k} \tag{272}$$

$$= \sum_{n=1}^{N} \gamma_{nk} \cdot \frac{(\boldsymbol{x}_n - \boldsymbol{\mu}_k)}{\sigma^2} \tag{273}$$

Since the σ is independent on the n, the final result of derivative of likelihood w.r.t $\boldsymbol{\mu}_k$ can be represented as

$$\frac{\partial Q}{\partial \boldsymbol{\mu}_k} = \frac{1}{\sigma^2} \sum_{n=1}^{N} \gamma_{nk} \cdot (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \tag{274}$$

Take derivative to the likelihood w.r.t $\boldsymbol{\mu}_k$ to get the extrenum,

$$\frac{\partial Q}{\partial \boldsymbol{\mu}_k} = 0 \tag{275}$$

Solve the equation above to acquire the extrenum,

$$\sum_{n=1}^{N} \gamma_{nk} \cdot \frac{(\boldsymbol{x}_n - \boldsymbol{\mu}_k)}{\sigma^2} = 0 \tag{276}$$

$$\frac{1}{\sigma^2} \sum_{n=1}^{N} \gamma_{nk} \cdot (\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \tag{277}$$

$$\sum_{n=1}^{N} \gamma_{nk} \cdot (\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \tag{278}$$

$$\Big(\sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n - \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{\mu}_k \Big) = 0 \tag{279}$$

$$\sum_{n=1}^{N} \gamma_{nk} \boldsymbol{\mu}_k = \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n \tag{280}$$

$$\boldsymbol{\mu}_k \sum_{n=1}^{N} \gamma_{nk} = \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n \tag{281}$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma_{nk}} \tag{282}$$

Note that from (277) to (278), we assume that $\sigma^2 \neq 0$. The derivation from (280) to (281) is valid since $\boldsymbol{\mu}_k$ is independent on summation variable n.

Then we define N_k to be effective number of measurements $\boldsymbol{x}_n$ being explained by k-th tree component. That is

$$N_k = \sum_{n=1}^{N} \gamma_{nk} \tag{283}$$

Therefore, we have

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n \tag{284}$$

Similarly, take derivative to Q w.r.t σ,

$$\frac{\partial \sum_{n=1}^{N} \sum_{k=1}^{K} \Big(\gamma_{nk} \cdot ln\ (\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})) \Big)}{\partial \sigma} \tag{285}$$

$$= \sum_{n=1}^{N} \frac{\partial \sum_{k=1}^{K} \Big(\gamma_{nk} \cdot ln\ (\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})) \Big)}{\partial \sigma} \tag{286}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\partial \Big(\gamma_{nk} \cdot ln\ (\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})) \Big)}{\partial \sigma} \tag{287}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \cdot \frac{\partial \Big(ln\ (\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})) \Big)}{\partial \sigma} \tag{288}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \cdot \frac{1}{\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})} \cdot \frac{\partial \Big(\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}) \Big)}{\partial \sigma} \tag{289}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \cdot \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})}{\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})} \cdot \Big(-\frac{2}{\sigma} + \frac{(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_n - \boldsymbol{\mu}_k)}{\sigma^3} \Big) \tag{290}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \cdot \Big(-\frac{2}{\sigma} + \frac{(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_n - \boldsymbol{\mu}_k)}{\sigma^3} \Big) \tag{291}$$

Set the derivative to zero, we will obtain the extrenum for σ

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\cdot\Big(-\frac{2}{\sigma}+\frac{(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T(\boldsymbol{x}_n-\boldsymbol{\mu}_k)}{\sigma^3}\Big)=0 \tag{292}$$

Since σ is independent on n and k, we have

$$\sigma^3\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\cdot\Big(-2\sigma^2+(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\Big)=0 \tag{293}$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\cdot\Big(-2\sigma^2+(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\Big)=0 \tag{294}$$

Further manipulation as follows,

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}2\sigma^2=\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T(\boldsymbol{x}_n-\boldsymbol{\mu}_k) \tag{295}$$

Since both σ is independent on n and k, we have

$$2\sigma^2\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}=\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T(\boldsymbol{x}_n-\boldsymbol{\mu}_k) \tag{296}$$

It is evident that

$$\sigma^2=\frac{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T(\boldsymbol{x}_n-\boldsymbol{\mu}_k)}{2\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}} \tag{297}$$

Based on the definition of $\gamma_{nk}=\frac{\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\sigma^2\boldsymbol{I})}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\sigma^2\boldsymbol{I})}$ in (263), we have

$$\sum_{k=1}^{K}\gamma_{nk}=\sum_{k=1}^{K}\frac{\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\sigma^2\boldsymbol{I})}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\sigma^2\boldsymbol{I})} \tag{298}$$

$$=\frac{\sum_{k=1}^{K}\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\sigma^2\boldsymbol{I})}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\sigma^2\boldsymbol{I})} \tag{299}$$

$$=1 \tag{300}$$

Note that the derivation from (298) to (299) is valid because the whole term denominator $\sum_{k=1}^{K}\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\sigma^2\boldsymbol{I})$ is independent on k .

Hence, the extrenum of σ^2 can be simplified as follows.

$$\sigma^2=\frac{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T(\boldsymbol{x}_n-\boldsymbol{\mu}_k)}{2\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}} \tag{301}$$

$$=\frac{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T(\boldsymbol{x}_n-\boldsymbol{\mu}_k)}{2\sum_{n=1}^{N}1} \tag{302}$$

$$=\frac{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T(\boldsymbol{x}_n-\boldsymbol{\mu}_k)}{2N} \tag{303}$$

$$=\frac{1}{2N}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T(\boldsymbol{x}_n-\boldsymbol{\mu}_k) \tag{304}$$

$$=\frac{1}{2N}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T(\boldsymbol{x}_n-\boldsymbol{\mu}_k) \tag{305}$$

Finally, let's derive the extrenum of Q with regard to π_k, but since π_k is constrained by condition $\sum_{k=1}^{K} \pi_k = 1$, the derivation would be a bit more complicated.

We construct Lagrangian $\mathcal{L}'(\boldsymbol{\pi}, \lambda)$ with lagrangian multiplier λ

$$\mathcal{L}'(\boldsymbol{\pi}, \lambda) = \sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\gamma_{nk} \cdot ln\ (\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2\boldsymbol{I}))\Big) + \lambda(1 - \sum_{k=1}^{K}\pi_k) \tag{306}$$

Then take derivative with regard to π_k, we have

$$\frac{\partial \mathcal{L}'(\boldsymbol{\pi}, \lambda)}{\partial \pi_k} = \sum_{n=1}^{N}\Big(\gamma_{nk} \cdot \frac{\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2\boldsymbol{I})}{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2\boldsymbol{I})}\Big) - \lambda \tag{307}$$

$$= \sum_{n=1}^{N}\Big(\gamma_{nk} \cdot \frac{1}{\pi_k}\Big) - \lambda \tag{308}$$

$$= \Big(\frac{1}{\pi_k}\sum_{n=1}^{N}\gamma_{nk}\Big) - \lambda \tag{309}$$

$$= \frac{N_k}{\pi_k} - \lambda \tag{310}$$

Set the derivative towards langarangian to zero, we have

$$\frac{N_k}{\pi_k} - \lambda = 0 \tag{311}$$

$$\pi_k = \frac{N_k}{\lambda} \tag{312}$$

Take the expression π_k back to $\mathcal{L}'(\boldsymbol{\pi}, \lambda)$, we have

$$\mathcal{L}(\lambda) = \sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\gamma_{nk} \cdot ln\ (\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2\boldsymbol{I}))\Big) + \lambda(1 - \sum_{k=1}^{K}\pi_k) \tag{313}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\gamma_{nk} \cdot ln\ (\frac{N_k}{\lambda}\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2\boldsymbol{I}))\Big) + \lambda - \lambda \cdot \sum_{k=1}^{K}\frac{N_k}{\lambda} \tag{314}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\gamma_{nk} \cdot ln\ (N_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2\boldsymbol{I})) - \gamma_{nk} \cdot ln\ \lambda\Big) + \lambda - \sum_{k=1}^{K}N_k \tag{315}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\gamma_{nk} \cdot ln\ (N_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \sigma^2\boldsymbol{I})) - \gamma_{nk} \cdot ln\ \lambda\Big) + \lambda - N \tag{316}$$

Take the derivative to $\mathcal{L}(\lambda)$ with regard to λ ,

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} = \sum_{n=1}^{N}\sum_{k=1}^{K}\Big(-\frac{\gamma_{nk}}{\lambda}\Big) + 1 \qquad \text{eliminate irrelavent terms} \tag{317}$$

$$= -\frac{1}{\lambda}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\gamma_{nk}\Big) + 1 \qquad \lambda \text{ independent on } n \text{ and } k \tag{318}$$

$$= -\frac{1}{\lambda}(\sum_{n=1}^{N}1) + 1 \qquad \text{cite (300)} \tag{319}$$

$$= -\frac{N}{\lambda} + 1 \tag{320}$$

Set derivative to zero to get minima of $\mathcal{L}(\lambda)$

$$-\frac{N}{\lambda} + 1 = 0 \tag{321}$$

$$\lambda = N \tag{322}$$

Take the final result back to what we have previously derived in (312).

$$\pi_k = \frac{N_k}{N} \tag{323}$$

which is the update formual for π_k in M-step.

In summary, we acquire all the M-step update as follows.

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n \tag{324}$$

$$\sigma^2 = \frac{1}{2N} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \tag{325}$$

$$\pi_k = \frac{N_k}{N} \tag{326}$$

where N_k is defined to be effective number of measurements $\boldsymbol{x}_n$ being explained by k-th tree component.

7 (Semi-/Un-)Supervised Learning and EM

7.1 Supervised Learning

We apply the supervised learning on Iris Data and use 10-fold cross validation to figure out the error rate by using the classficiation approach above.

Data Presentation. Specific values of mean vector and covariance matrix of each gaussian distribution in each cross validation training set are shown in the appendix Cross Validation for Supervised Learning. But here we just present the overall error rate in the required 10-fold cross validation and specific error rates of each validation set.

Average ErrorRate: 2.6667%

Specific Error in each set:

Set 1	Set 2	Set 3	Set 4	Set 5
0.000%	6.667%	0.000%	0.000%	0.000%
Set 6	Set 7	Set 8	Set 9	Set 10
6.667%	0.000%	6.667%	6.667%	0.000%

Note that my implementation for obtaining cross validation set in the above 10-fold cross validation is called "repeatedly random subsamppling validation", which is widely used. (Refer to http://en.wikipedia.org/wiki/Cross-validation_(statistics))

To run my program, please type the following in the command.

```
python ./Supervised.py
```

7.2 Unsupervised Learning

We apply Expectation-Maximisation technique to implement such a unsupervised learning task. For specific results of this task, please see the appendix Expectation Maximisation for Unsupervised Learning.

item	Case 1	Case 2	Case 3	Case 4	Case 5
Converged log likelihood	-185.276241	-196.953446	-180.185510	-180.185529	-180.185537

As illustrated in the above table, which indicates the log likelihood while algorithm converges (likelihood has little change in adjacent iteration, threshold all set to $10e^{-6}$.) It is obvious that each time we run EM algorithm to get three 4D gaussian distributions, the results are different from each others. There are many possible reasons as follows.

- MultiModial Distribution
- Sensitivity to Initilisation

Note that here we count a pair of E-step and M-step as one iteration.

To run my program, please type the following in the command.

```
python ./Unsupervised.py
```

Note that all the parameters have default values, for details of command line argument, please see help documentation by typing

```
python ./Unsupervised.py -h
```

7.3 Semi-supervised Learning

Approach Description: Use 5% of labeled data to construct initial gaussian components, that is, to fit the parameters of initial gaussian components by using MLE, and then utilise 95% of unlabeled data to adjust (perhaps with less weight comparing to the labeled data) the parameters of components by Expectation-Maximisation (EM) algorithm.

One strength of this approach lies in unnecessity of subjectively determining number of mixture components k for unsuperivsed learning. It is quite tough for human to determine one "good" k when human has no preknowledge of specific problem. Even though sometime large k may lead to small error or large likelilhood, which is desired mathematically, the result may not be interpretable in practice, which means the k make no sense in reality. If we apply the model above, we just need to label the data that has been observed as objects of different classes, and thus decrease the risk of bad k (number of components).

On top of that, since only 5% of data objects are provided for supervised learning in the first stage, the fitted parameters may be good qualitatively (it captures the approximate location of one class of objects) but not precise quantitatively (fitted parameters in a measure is far from the real one because the 5% data happens to be located in the brisk of the whole class's distribution). But if we after that apply the EM algorithm to adjust the acquired parameters with a relatively enormous size of dataset, the concerns about precision of parameters fitted by small size of dataset (5% of the whole) can be removed.

Another significant reason of doing so comes from the weakness of random initialisation for startup in EM algorithm. If bad random initialisation is encountered, the parameters of model such as $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and π in mixture of gaussians would be likely to converge to bad local optima rather than global optima or relatively satisfying optima. Therefore, training task has to be run for several times to derive a number of results, then compare the acquired results and pick up the one with smallest error or largest likelihood , which obviously is unexpectedly time-consuming. On the contrast, if a set of labeled data is introduced, even though in a limited quantity, we can make use of those data to fit initial gaussian model (perhaps maixmum likelihood estimation) to avoid running one machine learning task repeatedly (Sometimes tremendous ammount data is involved in a task, we may not wish to run them for several times because they would be either impossible or too costly.).

A Data Presentation

A.1 Cross Validation for Supervised Learning

Cross Validation Set 1:

Error of this Testing Set: 0.00%

For class Iris-setosa:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 5.01777778 & 3.44666667 & 1.46444444 & 0.25111111 \end{pmatrix} \tag{327}$$

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 0.12104040 & 0.10028788 & 0.01951010 & 0.01088889 \\ 0.10028788 & 0.15163636 & 0.01351515 & 0.00937879 \\ 0.01951010 & 0.01351515 & 0.03143434 & 0.00663131 \\ 0.01088889 & 0.00937879 & 0.00663131 & 0.01210101 \end{pmatrix} \tag{328}$$

For class Iris-versicolor:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.92444444 & 2.76666667 & 4.26000000 & 1.33333333 \end{pmatrix} \tag{329}$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.27416162 & 0.08174242 & 0.18645455 & 0.05553030 \\ 0.08174242 & 0.09909091 & 0.08000000 & 0.03977273 \\ 0.18645455 & 0.08000000 & 0.22927273 & 0.07409091 \\ 0.05553030 & 0.03977273 & 0.07409091 & 0.03909091 \end{pmatrix} \tag{330}$$

For class Iris-virginica:

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.63777778 & 2.96444444 & 5.58666667 & 2.01555556 \end{pmatrix} \tag{331}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.41967677 & 0.10637374 & 0.31574242 & 0.06098990 \\ 0.10637374 & 0.10961616 & 0.08133333 & 0.05011111 \\ 0.31574242 & 0.08133333 & 0.31845455 & 0.05748485 \\ 0.06098990 & 0.05011111 & 0.05748485 & 0.07497980 \end{pmatrix} \tag{332}$$

Cross Validation Set 2:

Error of this Testing Set: 6.67%

For class Iris-setosa:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 5.01333333 & 3.43555556 & 1.47555556 & 0.24666667 \end{pmatrix} \tag{333}$$

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 0.12345455 & 0.10451515 & 0.01737879 & 0.01163636 \\ 0.10451515 & 0.15370707 & 0.01225253 & 0.01080303 \\ 0.01737879 & 0.01225253 & 0.03143434 & 0.00662121 \\ 0.01163636 & 0.01080303 & 0.00662121 & 0.01209091 \end{pmatrix} \tag{334}$$

For class Iris-versicolor:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.97777778 & 2.76888889 & 4.27111111 & 1.32222222 \end{pmatrix} \tag{335}$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.26494949 & 0.08724747 & 0.18161616 & 0.05573232 \\ 0.08724747 & 0.10037374 & 0.07635354 & 0.03707071 \\ 0.18161616 & 0.07635354 & 0.21528283 & 0.06679293 \\ 0.05573232 & 0.03707071 & 0.06679293 & 0.03494949 \end{pmatrix} \tag{336}$$

For class Iris-virginica:

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.60888889 & 2.96444444 & 5.55333333 & 2.01111111 \end{pmatrix} \tag{337}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.41946465 & 0.10305051 & 0.32156061 & 0.05694444 \\ 0.10305051 & 0.11143434 & 0.07557576 & 0.04926768 \\ 0.32156061 & 0.07557576 & 0.31618182 & 0.04734848 \\ 0.05694444 & 0.04926768 & 0.04734848 & 0.07782828 \end{pmatrix} \tag{338}$$

Cross Validation Set 3:
Error of this Testing Set: 0.00%
For class Iris-setosa:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 5.01041667 & 3.41875000 & 1.46250000 & 0.25000000 \end{pmatrix} \tag{339}$$

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 0.12520833 & 0.09873670 & 0.01635638 & 0.01095745 \\ 0.09873670 & 0.13900266 & 0.01135638 & 0.01159574 \\ 0.01635638 & 0.01135638 & 0.03132979 & 0.00638298 \\ 0.01095745 & 0.01159574 & 0.00638298 & 0.01106383 \end{pmatrix} \tag{340}$$

For class Iris-versicolor:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.91428571 & 2.75952381 & 4.25238095 & 1.32857143 \end{pmatrix} \tag{341}$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.29344948 & 0.08937282 & 0.19679443 & 0.05519164 \\ 0.08937282 & 0.10783391 & 0.08851336 & 0.04191638 \\ 0.19679443 & 0.08851336 & 0.23718931 & 0.07700348 \\ 0.05519164 & 0.04191638 & 0.07700348 & 0.03721254 \end{pmatrix} \tag{342}$$

For class Iris-virginica:

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.60666667 & 2.98000000 & 5.57333333 & 2.03777778 \end{pmatrix} \tag{343}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.39972727 & 0.09786364 & 0.29290909 & 0.04542424 \\ 0.09786364 & 0.11118182 & 0.07445455 & 0.05100000 \\ 0.29290909 & 0.07445455 & 0.29018182 & 0.04512121 \\ 0.04542424 & 0.05100000 & 0.04512121 & 0.08104040 \end{pmatrix} \tag{344}$$

Cross Validation Set 4:
Error of this Testing Set: 0.00%
For class Iris-setosa:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 5.04047619 & 3.47619048 & 1.45952381 & 0.24761905 \end{pmatrix} \tag{345}$$

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 0.12929733 & 0.09586527 & 0.01704413 & 0.01168409 \\ 0.09586527 & 0.12771196 & 0.00828107 & 0.01042973 \\ 0.01704413 & 0.00828107 & 0.02978513 & 0.00782811 \\ 0.01168409 & 0.01042973 & 0.00782811 & 0.01182346 \end{pmatrix} \tag{346}$$

For class Iris-versicolor:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.92978723 & 2.76170213 & 4.25744681 & 1.32978723 \end{pmatrix} \tag{347}$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.27909343 & 0.08551341 & 0.18586031 & 0.05604995 \\ 0.08551341 & 0.09806660 & 0.07985661 & 0.03964385 \\ 0.18586031 & 0.07985661 & 0.21423682 & 0.07107771 \\ 0.05604995 & 0.03964385 & 0.07107771 & 0.03735430 \end{pmatrix} \tag{348}$$

For class Iris-virginica:

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.59565217 & 2.95000000 & 5.56086957 & 2.01739130 \end{pmatrix} \tag{349}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.37953623 & 0.07911111 & 0.28915942 & 0.06074396 \\ 0.07911111 & 0.09277778 & 0.06244444 & 0.05111111 \\ 0.28915942 & 0.06244444 & 0.30110145 & 0.05514010 \\ 0.06074396 & 0.05111111 & 0.05514010 & 0.07613527 \end{pmatrix} \tag{350}$$

Cross Validation Set 5:
Error of this Testing Set: 0.00%
For class Iris-setosa:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 4.99565217 & 3.43043478 & 1.45652174 & 0.25000000 \end{pmatrix} \tag{351}$$

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 0.11642512 & 0.09502415 & 0.01380676 & 0.00977778 \\ 0.09502415 & 0.14527536 & 0.01024155 & 0.00822222 \\ 0.01380676 & 0.01024155 & 0.03140097 & 0.00622222 \\ 0.00977778 & 0.00822222 & 0.00622222 & 0.01144444 \end{pmatrix} \tag{352}$$

For class Iris-versicolor:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.91739130 & 2.76956522 & 4.24565217 & 1.31521739 \end{pmatrix} \tag{353}$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.27257971 & 0.09298551 & 0.18429952 & 0.05439614 \\ 0.09298551 & 0.09994203 & 0.08808696 & 0.04180676 \\ 0.18429952 & 0.08808696 & 0.22698068 & 0.07351208 \\ 0.05439614 & 0.04180676 & 0.07351208 & 0.03865217 \end{pmatrix} \tag{354}$$

For class Iris-virginica:

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.61162791 & 2.96744186 & 5.56511628 & 2.01627907 \end{pmatrix} \tag{355}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.41629014 & 0.10205426 & 0.32232004 & 0.05933001 \\ 0.10205426 & 0.11320044 & 0.07598007 & 0.04911406 \\ 0.32232004 & 0.07598007 & 0.31375415 & 0.05677187 \\ 0.05933001 & 0.04911406 & 0.05677187 & 0.07901440 \end{pmatrix} \tag{356}$$

Cross Validation Set 6:
Error of this Testing Set: 6.67%
For class Iris-setosa:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 5.03043478 & 3.42826087 & 1.46956522 & 0.25217391 \end{pmatrix} \tag{357}$$

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 0.12616425 & 0.10689855 & 0.01428019 & 0.00971014 \\ 0.10689855 & 0.15273913 & 0.01510145 & 0.01049275 \\ 0.01428019 & 0.01510145 & 0.02749758 & 0.00606763 \\ 0.00971014 & 0.01049275 & 0.00606763 & 0.01143961 \end{pmatrix} \tag{358}$$

For class Iris-versicolor:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.94000000 & 2.76666667 & 4.25777778 & 1.32222222 \end{pmatrix} \tag{359}$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.26972727 & 0.09227273 & 0.19331818 & 0.06113636 \\ 0.09227273 & 0.10909091 & 0.09106061 & 0.04598485 \\ 0.19331818 & 0.09106061 & 0.23840404 & 0.08073232 \\ 0.06113636 & 0.04598485 & 0.08073232 & 0.04267677 \end{pmatrix} \tag{360}$$

For class Iris-virginica:

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.57045455 & 2.95000000 & 5.53863636 & 2.05227273 \end{pmatrix} \tag{361}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.33561839 & 0.05872093 & 0.26767970 & 0.05251057 \\ 0.05872093 & 0.09279070 & 0.04965116 & 0.05337209 \\ 0.26767970 & 0.04965116 & 0.29777484 & 0.05281712 \\ 0.05251057 & 0.05337209 & 0.05281712 & 0.07604123 \end{pmatrix} \tag{362}$$

Cross Validation Set 7:
Error of this Testing Set: 0.00%
For class Iris-setosa:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 5.03777778 & 3.45111111 & 1.46888889 & 0.25111111 \end{pmatrix} \tag{363}$$

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 0.12194949 & 0.09393434 & 0.01438384 & 0.01075253 \\ 0.09393434 & 0.12801010 & 0.00844444 & 0.01210101 \\ 0.01438384 & 0.00844444 & 0.03173737 & 0.00662626 \\ 0.01075253 & 0.01210101 & 0.00662626 & 0.01164646 \end{pmatrix} \tag{364}$$

For class Iris-versicolor:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.94666667 & 2.77777778 & 4.26888889 & 1.32666667 \end{pmatrix} \tag{365}$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.26072727 & 0.07310606 & 0.17921212 & 0.05077273 \\ 0.07310606 & 0.09176768 & 0.07611111 & 0.03833333 \\ 0.17921212 & 0.07611111 & 0.22810101 & 0.07334848 \\ 0.05077273 & 0.03833333 & 0.07334848 & 0.03972727 \end{pmatrix} \tag{366}$$

For class Iris-virginica:

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.58666667 & 2.95777778 & 5.55777778 & 2.02888889 \end{pmatrix} \tag{367}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.40709091 & 0.08328788 & 0.29192424 & 0.04471212 \\ 0.08328788 & 0.09931313 & 0.05908586 & 0.05033838 \\ 0.29192424 & 0.05908586 & 0.28658586 & 0.04261111 \\ 0.04471212 & 0.05033838 & 0.04261111 & 0.08073737 \end{pmatrix} \tag{368}$$

Cross Validation Set 8:
Error of this Testing Set: 6.67%
For class Iris-setosa:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 4.99782609 & 3.42173913 & 1.46086957 & 0.24565217 \end{pmatrix} \tag{369}$$

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 0.13266184 & 0.10471498 & 0.01769082 & 0.01121256 \\ 0.10471498 & 0.15062802 & 0.01264734 & 0.00920773 \\ 0.01769082 & 0.01264734 & 0.03265700 & 0.00649275 \\ 0.01121256 & 0.00920773 & 0.00649275 & 0.01142512 \end{pmatrix} \tag{370}$$

For class Iris-versicolor:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.94090909 & 2.75454545 & 4.23409091 & 1.31136364 \end{pmatrix} \tag{371}$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.28154334 & 0.08934461 & 0.20089852 & 0.05905920 \\ 0.08934461 & 0.10393235 & 0.08484144 & 0.03959831 \\ 0.20089852 & 0.08484144 & 0.23718288 & 0.07448732 \\ 0.05905920 & 0.03959831 & 0.07448732 & 0.03730973 \end{pmatrix} \tag{372}$$

For class Iris-virginica:

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.58000000 & 2.96000000 & 5.57111111 & 2.00888889 \end{pmatrix} \tag{373}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.43254545 & 0.10100000 & 0.33395455 & 0.04563636 \\ 0.10100000 & 0.11336364 & 0.08086364 & 0.04945455 \\ 0.33395455 & 0.08086364 & 0.32846465 & 0.05412626 \\ 0.04563636 & 0.04945455 & 0.05412626 & 0.07582828 \end{pmatrix} \tag{374}$$

Cross Validation Set 9:
Error of this Testing Set: 6.67%
For class Iris-setosa:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 5.00444444 & 3.43111111 & 1.44666667 & 0.23777778 \end{pmatrix} \tag{375}$$

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 0.13816162 & 0.10963131 & 0.01728788 & 0.01119192 \\ 0.10963131 & 0.15219192 & 0.01033333 & 0.00788889 \\ 0.01728788 & 0.01033333 & 0.02800000 & 0.00410606 \\ 0.01119192 & 0.00788889 & 0.00410606 & 0.00876768 \end{pmatrix} \tag{376}$$

For class Iris-versicolor:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.94772727 & 2.78409091 & 4.26363636 & 1.33636364 \end{pmatrix} \tag{377}$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.25371564 & 0.08403277 & 0.17014799 & 0.05078224 \\ 0.08403277 & 0.10695032 & 0.08429175 & 0.04268499 \\ 0.17014799 & 0.08429175 & 0.19306554 & 0.06437632 \\ 0.05078224 & 0.04268499 & 0.06437632 & 0.03725159 \end{pmatrix} \tag{378}$$

For class Iris-virginica:

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.56956522 & 2.97826087 & 5.54347826 & 2.03043478 \end{pmatrix} \tag{379}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.42216425 & 0.10376812 & 0.32135266 & 0.05516908 \\ 0.10376812 & 0.11151691 & 0.08052174 & 0.05156522 \\ 0.32135266 & 0.08052174 & 0.32428986 & 0.05464734 \\ 0.05516908 & 0.05156522 & 0.05464734 & 0.08038647 \end{pmatrix} \tag{380}$$

Cross Validation Set 10:
Error of this Testing Set: 0.00%
For class Iris-setosa:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 4.98888889 & 3.39777778 & 1.46000000 & 0.24222222 \end{pmatrix} \tag{381}$$

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 0.12146465 & 0.09747475 & 0.01568182 & 0.00934343 \\ 0.09747475 & 0.13976768 & 0.01172727 & 0.01009596 \\ 0.01568182 & 0.01172727 & 0.03200000 & 0.00559091 \\ 0.00934343 & 0.01009596 & 0.00559091 & 0.01067677 \end{pmatrix} \tag{382}$$

For class Iris-versicolor:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.89534884 & 2.76279070 & 4.22790698 & 1.30465116 \end{pmatrix} \tag{383}$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.28188261 & 0.09410853 & 0.19156146 & 0.05430786 \\ 0.09410853 & 0.10239203 & 0.09225360 & 0.04589147 \\ 0.19156146 & 0.09225360 & 0.23253599 & 0.07415282 \\ 0.05430786 & 0.04589147 & 0.07415282 & 0.03902547 \end{pmatrix} \tag{384}$$

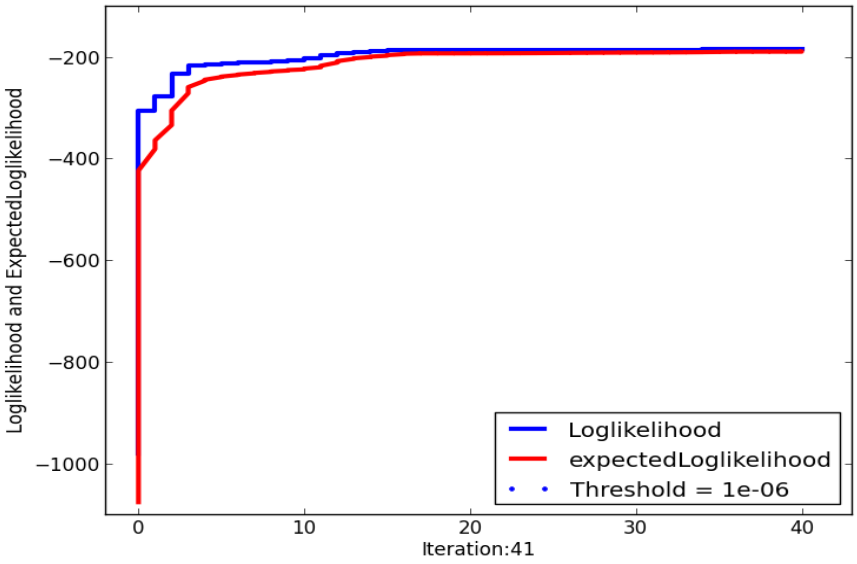
For class Iris-virginica:

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.59574468 & 2.96595745 & 5.57021277 & 2.03191489 \end{pmatrix} \tag{385}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.42954672 & 0.10093895 & 0.32117484 & 0.05122572 \\ 0.10093895 & 0.10925069 & 0.07874653 & 0.05154487 \\ 0.32117484 & 0.07874653 & 0.31691952 & 0.05118871 \\ 0.05122572 & 0.05154487 & 0.05118871 & 0.07917669 \end{pmatrix} \tag{386}$$

A.2 Expectation Maximisation for Unsupervised Learning

Specific data in first random initialisation



Initilisation of parameters:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 6.94825182 & 3.43028943 & 5.13165042 & 1.73043353 \end{pmatrix} \tag{387}$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 6.03430983 & 2.44263481 & 4.19416609 & 1.34434648 \end{pmatrix} \tag{388}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 7.41659032 & 2.74307646 & 3.54490011 & 0.37515563 \end{pmatrix} \tag{389}$$

$$\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I} \tag{390}$$

$$\boldsymbol{\pi} = \begin{pmatrix} 0.63777939 & 0.17825037 & 0.18397024 \end{pmatrix} \tag{391}$$

Iteration Quantity: 41
Parameters of resulted Guassians:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 6.30229683 & 2.90022864 & 5.06049687 & 1.76245416 \end{pmatrix} \tag{392}$$

$$\boldsymbol{\Sigma}_0 \begin{pmatrix} 0.42223868 & 0.09703482 & 0.43821003 & 0.15608046 \\ 0.09703482 & 0.10551541 & 0.11569370 & 0.07015344 \\ 0.43821003 & 0.11569370 & 0.62024483 & 0.24770913 \\ 0.15608046 & 0.07015344 & 0.24770913 & 0.16391035 \end{pmatrix} \tag{393}$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 6.06077537 & 2.73103951 & 4.13451491 & 1.24428925 \end{pmatrix} \tag{394}$$

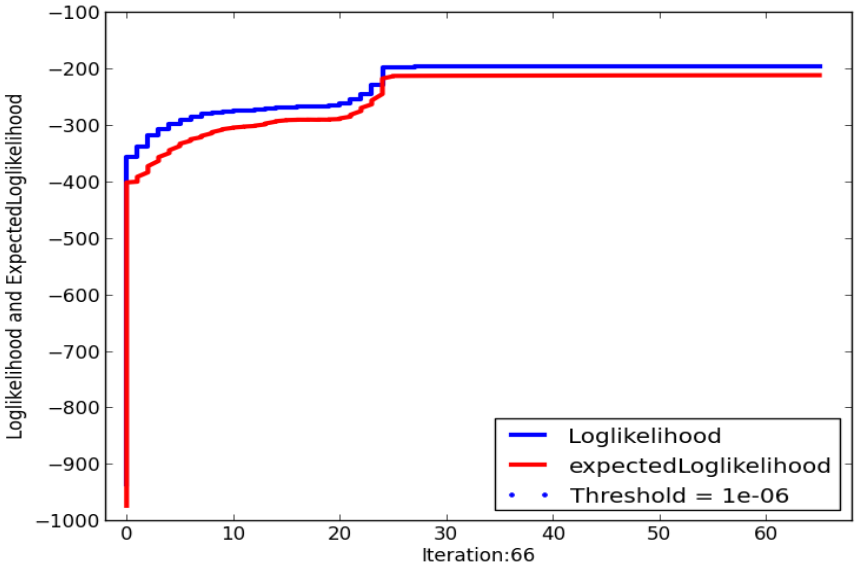
$$\boldsymbol{\Sigma}_1 \begin{pmatrix} 0.44986364 & 0.20624275 & 0.31552694 & 0.10820035 \\ 0.20624275 & 0.10624364 & 0.13904845 & 0.05150205 \\ 0.31552694 & 0.13904845 & 0.23264190 & 0.07652083 \\ 0.10820035 & 0.05150205 & 0.07652083 & 0.02840178 \end{pmatrix} \tag{395}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 5.00600023 & 3.42800051 & 1.46200007 & 0.24599998 \end{pmatrix} \tag{396}$$

$$\boldsymbol{\Sigma}_2 \begin{pmatrix} 0.12176394 & 0.09723179 & 0.01602797 & 0.01012402 \\ 0.09723179 & 0.14081549 & 0.01146392 & 0.00911203 \\ 0.01602797 & 0.01146392 & 0.02955600 & 0.00594801 \\ 0.01012402 & 0.00911203 & 0.00594801 & 0.01088400 \end{pmatrix} \tag{397}$$

$$\boldsymbol{\pi} = \begin{pmatrix} 0.55543518 & 0.11123164 & 0.33333318 \end{pmatrix} \tag{398}$$

Specific data in second random initialisation



Initilisation of parameters:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 5.47252304 & 3.04349496 & 3.20982127 & 1.74800758 \end{pmatrix} \tag{399}$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.37077336 & 3.19556116 & 5.46395648 & -0.03621684 \end{pmatrix} \tag{400}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.08856967 & 3.14829166 & 5.74406265 & 1.40878059 \end{pmatrix} \tag{401}$$

$$\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I} \tag{402}$$

$$\boldsymbol{\pi} = \begin{pmatrix} 0.92942981 & 0.06546349 & 0.0051067 \end{pmatrix} \tag{403}$$

Iteration Quantity: 66
Parameters of resulted Guassians:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 5.00600011 & 3.42800025 & 1.46200004 & 0.24599999 \end{pmatrix} \tag{404}$$

$$\boldsymbol{\Sigma}_0 \begin{pmatrix} 0.12176397 & 0.09723189 & 0.01602799 & 0.01012401 \\ 0.09723189 & 0.14081575 & 0.01146396 & 0.00911202 \\ 0.01602799 & 0.01146396 & 0.02955600 & 0.00594800 \\ 0.01012401 & 0.00911202 & 0.00594800 & 0.01088400 \end{pmatrix} \tag{405}$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 6.11764829 & 2.84264967 & 4.67930452 & 1.59705880 \end{pmatrix} \tag{406}$$

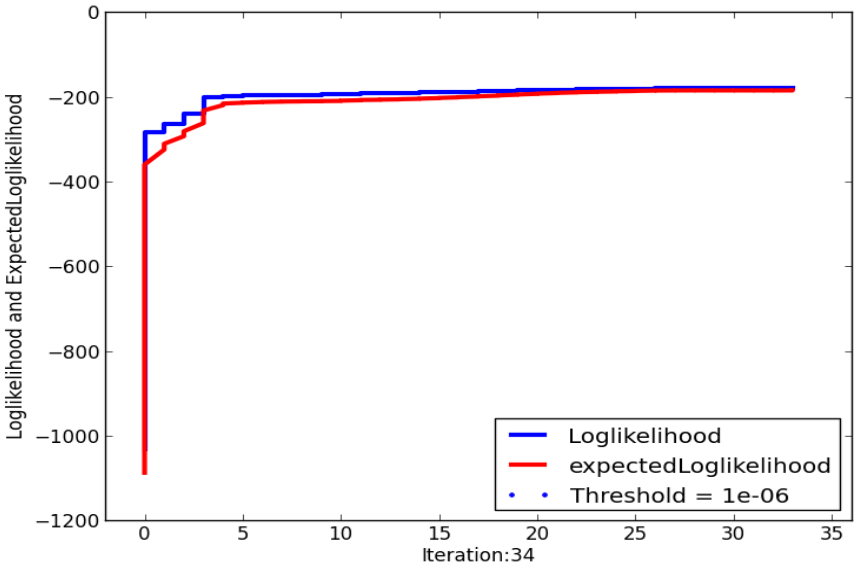
$$\boldsymbol{\Sigma}_1 \begin{pmatrix} 0.31117193 & 0.10177389 & 0.28584863 & 0.14245983 \\ 0.10177389 & 0.09664535 & 0.12171721 & 0.07841352 \\ 0.28584863 & 0.12171721 & 0.47102262 & 0.24838159 \\ 0.14245983 & 0.07841352 & 0.24838159 & 0.17198206 \end{pmatrix} \tag{407}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.90609957 & 3.00296152 & 5.91751825 & 2.02823662 \end{pmatrix} \tag{408}$$

$$\boldsymbol{\Sigma}_2 \begin{pmatrix} 0.47944373 & 0.10318151 & 0.37851617 & -0.00947992 \\ 0.10318151 & 0.14649679 & 0.06689303 & 0.02639494 \\ 0.37851617 & 0.06689303 & 0.33139405 & 0.01686062 \\ -0.00947992 & 0.02639494 & 0.01686062 & 0.05638487 \end{pmatrix} \tag{409}$$

$$\boldsymbol{\pi} = \begin{pmatrix} 0.33333326 & 0.54461183 & 0.12205491 \end{pmatrix} \tag{410}$$

Specific data in third random initialisation



Initilisation of parameters:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 5.58180881 & 4.84821904 & 4.19996763 & 1.66923484 \end{pmatrix} \tag{411}$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.98138886 & 3.26599148 & 2.48486145 & -0.38105446 \end{pmatrix} \tag{412}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 5.81470396 & 4.52844168 & 4.91642107 & 3.99937301 \end{pmatrix} \tag{413}$$

$$\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I} \tag{414}$$

$$\boldsymbol{\pi} = \begin{pmatrix} 0.5694223 & 0.22911455 & 0.20146316 \end{pmatrix} \tag{415}$$

Iteration Quantity: 34
Parameters of resulted Guassians:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 5.91506499 & 2.77785242 & 4.20175063 & 1.29704350 \end{pmatrix} \tag{416}$$

$$\boldsymbol{\Sigma}_0 \begin{pmatrix} 0.27532206 & 0.09691983 & 0.18469154 & 0.05440596 \\ 0.09691983 & 0.09264056 & 0.09113461 & 0.04299639 \\ 0.18469154 & 0.09113461 & 0.20070652 & 0.06101118 \\ 0.05440596 & 0.04299639 & 0.06101118 & 0.03201270 \end{pmatrix} \tag{417}$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.00600000 & 3.42800000 & 1.46200000 & 0.24600000 \end{pmatrix} \tag{418}$$

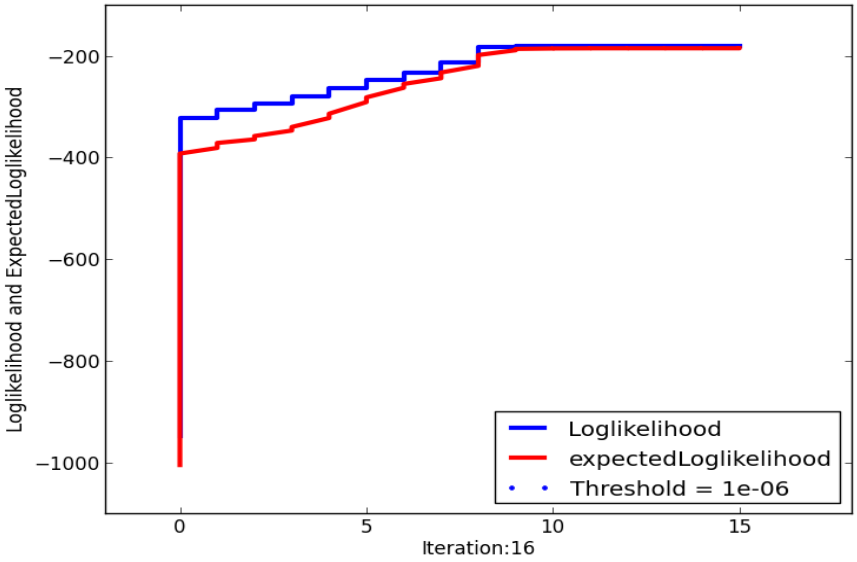
$$\boldsymbol{\Sigma}_1 \begin{pmatrix} 0.12176400 & 0.09723200 & 0.01602800 & 0.01012400 \\ 0.09723200 & 0.14081600 & 0.01146400 & 0.00911200 \\ 0.01602800 & 0.01146400 & 0.02955600 & 0.00594800 \\ 0.01012400 & 0.00911200 & 0.00594800 & 0.01088400 \end{pmatrix} \tag{419}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.54467220 & 2.94870862 & 5.47980119 & 1.98476235 \end{pmatrix} \tag{420}$$

$$\boldsymbol{\Sigma}_2 \begin{pmatrix} 0.38704847 & 0.09220736 & 0.30276907 & 0.06159530 \\ 0.09220736 & 0.11034094 & 0.08425917 & 0.05599446 \\ 0.30276907 & 0.08425917 & 0.32766484 & 0.07441712 \\ 0.06159530 & 0.05599446 & 0.07441712 & 0.08573678 \end{pmatrix} \tag{421}$$

$$\boldsymbol{\pi} = \begin{pmatrix} 0.29931064 & 0.33333333 & 0.36735603 \end{pmatrix} \tag{422}$$

Specific data in fourth random initialisation



Initilisation of parameters:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 4.93395858 & 3.72289358 & 2.88809351 & 1.81106115 \end{pmatrix} \tag{423}$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.74870051 & 1.17972676 & 3.84560644 & 0.78416837 \end{pmatrix} \tag{424}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 7.63689574 & 2.95945543 & 5.19796747 & 1.71749004 \end{pmatrix} \tag{425}$$

$$\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I} \tag{426}$$

$$\boldsymbol{\pi} = \begin{pmatrix} 0.86693085 & 0.06884158 & 0.06422758 \end{pmatrix} \tag{427}$$

Iteration Quantity: 16
Parameters of resulted Guassians:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 5.00600000 & 3.42800000 & 1.46200000 & 0.24600000 \end{pmatrix} \tag{428}$$

$$\boldsymbol{\Sigma}_0 \begin{pmatrix} 0.12176400 & 0.09723200 & 0.01602800 & 0.01012400 \\ 0.09723200 & 0.14081600 & 0.01146400 & 0.00911200 \\ 0.01602800 & 0.01146400 & 0.02955600 & 0.00594800 \\ 0.01012400 & 0.00911200 & 0.00594800 & 0.01088400 \end{pmatrix} \tag{429}$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.91510204 & 2.77785066 & 4.20180128 & 1.29706329 \end{pmatrix} \tag{430}$$

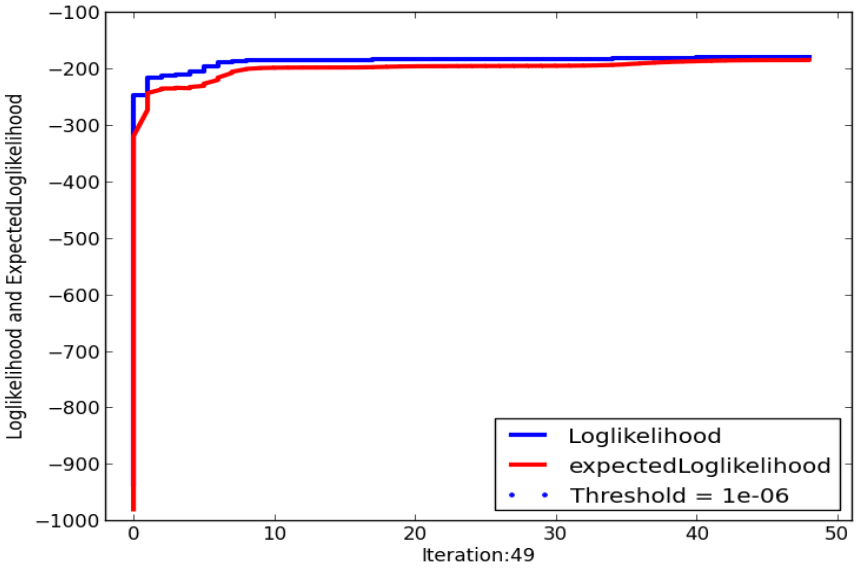
$$\boldsymbol{\Sigma}_1 \begin{pmatrix} 0.27532744 & 0.09691654 & 0.18470680 & 0.05441350 \\ 0.09691654 & 0.09264040 & 0.09113279 & 0.04299613 \\ 0.18470680 & 0.09113279 & 0.20073024 & 0.06102147 \\ 0.05441350 & 0.04299613 & 0.06102147 & 0.03201724 \end{pmatrix} \tag{431}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.54468870 & 2.94872272 & 5.47985468 & 1.98479721 \end{pmatrix} \tag{432}$$

$$\boldsymbol{\Sigma}_2 \begin{pmatrix} 0.38706100 & 0.09220597 & 0.30277029 & 0.06158850 \\ 0.09220597 & 0.11033973 & 0.08424916 & 0.05598870 \\ 0.30277029 & 0.08424916 & 0.32763930 & 0.07439355 \\ 0.06158850 & 0.05598870 & 0.07439355 & 0.08572418 \end{pmatrix} \tag{433}$$

$$\boldsymbol{\pi} = \begin{pmatrix} 0.33333333 & 0.29933787 & 0.36732879 \end{pmatrix} \tag{434}$$

Specific data in fifth random initialisation



Initilisation of parameters:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 4.74122040 & 3.15279417 & 1.77909908 & 0.80293032 \end{pmatrix} \tag{435}$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 6.36558270 & 4.20423977 & 4.12724521 & 1.24996769 \end{pmatrix} \tag{436}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 7.34388669 & 3.54991993 & 4.39371314 & 2.05949636 \end{pmatrix} \tag{437}$$

$$\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I} \tag{438}$$

$$\boldsymbol{\pi} = \begin{pmatrix} 0.02053414 & 0.0965958 & 0.88287006 \end{pmatrix} \tag{439}$$

Iteration Quantity: 49
Parameters of resulted Guassians:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 5.00600000 & 3.42800000 & 1.46200000 & 0.24600000 \end{pmatrix} \tag{440}$$

$$\boldsymbol{\Sigma}_0 \begin{pmatrix} 0.12176400 & 0.09723200 & 0.01602800 & 0.01012400 \\ 0.09723200 & 0.14081600 & 0.01146400 & 0.00911200 \\ 0.01602800 & 0.01146400 & 0.02955600 & 0.00594800 \\ 0.01012400 & 0.00911200 & 0.00594800 & 0.01088400 \end{pmatrix} \tag{441}$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5.91484076 & 2.77783583 & 4.20129188 & 1.29686598 \end{pmatrix} \tag{442}$$

$$\boldsymbol{\Sigma}_1 \begin{pmatrix} 0.27531479 & 0.09697069 & 0.18462390 & 0.05437052 \\ 0.09697069 & 0.09265277 & 0.09115580 & 0.04299921 \\ 0.18462390 & 0.09115580 & 0.20053150 & 0.06093600 \\ 0.05437052 & 0.04299921 & 0.06093600 & 0.03197655 \end{pmatrix} \tag{443}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 6.54438345 & 2.94859425 & 5.47921795 & 1.98439211 \end{pmatrix} \tag{444}$$

$$\boldsymbol{\Sigma}_2 \begin{pmatrix} 0.38703762 & 0.09221031 & 0.30286958 & 0.06172628 \\ 0.09221031 & 0.11033498 & 0.08432941 & 0.05603638 \\ 0.30286958 & 0.08432941 & 0.32798018 & 0.07468452 \\ 0.06172628 & 0.05603638 & 0.07468452 & 0.08588114 \end{pmatrix} \tag{445}$$

$$\boldsymbol{\pi} = \begin{pmatrix} 0.33333333 & 0.29903553 & 0.36763114 \end{pmatrix} \tag{446}$$