Statistical Modeling 2

CS5240 Theoretical Foundations in Multimedia

Leow Wee Kheng

Department of Computer Science
School of Computing
National University of Singapore
Last Time...

We studied

- Expectation, Variance, Covariance
- Maximum likelihood estimation
- Maximum a posteriori estimation
- Difference measures for distributions

Difference measures provide an overall evaluation of estimated models. For more detailed evaluation, need regression methods.
Linear regression studies linear relationships between variables.
Linear regression model is given by

\[ y_i = \sum_{j=1}^{m} a_j x_{ij} + \epsilon_i. \]  

\[ (1) \]

- \( x_{ij} \) are **known non-random** independent variables.
- \( a_j \) are **unobservable (unknown) non-random** parameters.
- Constant terms can be added with \( a_{m+1} \) and \( x_{i,m+1} = 1 \).
- \( \epsilon_i \) are **unknown random** noise, error, or disturbance.
- \( y_i \) are **known random** dependent variables.
Error and Residual

In statistics, error and residual are related but different.

- Error is difference between observed and true, unobservable value.
- Residual is difference between observed and estimated value.

In terms of linear model

\[ y_i = \sum_{j=1}^{m} a_j x_{ij} + \epsilon_i, \quad i = 1, \ldots, n, \quad (2) \]

error \[ \epsilon_i = y_i - \sum_{j=1}^{m} a_j x_{ij}, \quad (3) \]

residual \[ r_i = y_i - \sum_{j=1}^{m} \hat{a}_j x_{ij}, \quad (4) \]

where \( \hat{a}_j \) are the estimated parameter values.
In matrix form, Eq. 1 is

$$y = Xa + \epsilon,$$

(5)

where

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}, \quad a = \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$ 

Linear regression solves for \(\hat{a}\) with

$$y = X\hat{a}.$$  

(6)

The estimated parameter values \(\hat{a}\), from linear least square, is

$$\hat{a} = (X^\top X)^{-1}X^\top y.$$  

(7)
Linear least-square estimate $\hat{a}$ is unbiased and consistent when $\epsilon_i$ are

- zero mean: $E[\epsilon_i] = 0$ for all $i$
- finite variance: $\text{Var}[\epsilon_i] = \sigma^2 < \infty$ for all $i$

First, the expectation of $\hat{a}$ is

$$E[\hat{a}] = E[(X^\top X)^{-1}X^\top(Xa + \epsilon)]$$

$$= E[(X^\top X)^{-1}X^\top Xa + (X^\top X)^{-1}X^\top \epsilon]$$

$$= E[a] + (X^\top X)^{-1}X^\top E[\epsilon]$$

$$= a.$$

So, $\hat{a}$ is unbiased.
Next, the variance of $\hat{a}$ is

\[
\text{Var}[\hat{a}] = \text{Var}[(X^\top X)^{-1}X^\top(Xa + \epsilon)]
\]
\[
= \text{Var}[(X^\top X)^{-1}X^\top Xa + (X^\top X)^{-1}X^\top \epsilon]
\]
\[
= 0 + (X^\top X)^{-1}X^\top \text{Var}[\epsilon][(X^\top X)^{-1}X^\top]^\top
\]
\[
= \sigma^2(X^\top X)^{-1}X^\top X(X^\top X)^{-1}
\]
\[
= \sigma^2(X^\top X)^{-1} \to 0 \text{ as } n \to \infty.
\]

So, $\hat{a}$ is consistent.

So, linear least square is a statistically good estimator.
Simple Linear Regression

Simple linear regression (SLR) has single independent variable

\[ y_i = a x_i + b + \epsilon_i, \quad i = 1, \ldots, n. \tag{8} \]

It minimizes the sum-squared error

\[ E = \sum_{i=1}^{n} (y_i - ax_i - b)^2. \tag{9} \]

The minimum of \( E \) is attained with

\[ \frac{dE}{da} = -2 \sum_{i=1}^{n} (y_i - ax_i - b)x_i = 0, \]

\[ \frac{dE}{db} = -2 \sum_{i=1}^{n} (y_i - ax_i - b) = 0. \tag{10} \]
Solving Eq. 10 yields (Homework)

\[
\hat{a} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\text{Cov}[X,Y]}{\text{Var}[X]}, \quad \hat{b} = \bar{y} - \hat{a}\bar{x},
\]

(11)

where

\[
\bar{x} = \sum_{i=1}^{n} x_i, \quad \bar{y} = \sum_{i=1}^{n} y_i.
\]

(12)

The estimate \( \hat{a} \) is similar to Pearson’s correlation coefficient

\[
\rho(X,Y) = \frac{\text{Cov}[X,Y]}{\sqrt{\text{Var}[X] \text{Var}[Y]}}.
\]

(13)

which is symmetric to \( X \) and \( Y \).
**SLR vs. MLE**

Let $X$ and $Y$ denote two random variables that are related linearly:

$$Y = aX + b + \epsilon.$$  \hspace{1cm} (14)

Suppose error $\epsilon$ follows Gaussian pdf with zero mean and variance $\sigma^2$. The sum-squared error for $n$ pairs of data $x_i$ and $y_i$ is

$$E = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$  \hspace{1cm} (15)

Then, the likelihood $L$ is

$$L(y_1, \ldots, y_n; a, b, \sigma) = \frac{1}{\sqrt{(2\pi)^n\sigma^{2n}}} \exp \left( -\frac{E}{2\sigma^2} \right).$$  \hspace{1cm} (16)

Can show that maximizing log $L$ implies minimizing $E$. (Homework) So, SLR is equivalent to MLE.
Best linear unbiased estimator (BLUE)

- “Best” means estimate \( \hat{a} \) has smallest variance.
- Unbiased estimation means \( E[\hat{a}] = a \).

Gauss-Markov theorem shows that linear least-square estimate

\[
\hat{a} = (X^\top X)^{-1}X^\top y. \tag{17}
\]

is BLUE.

Gauss-Markov assumptions on error \( \epsilon_i \):

- zero mean: \( E[\epsilon_i] = 0 \) for all \( i \)
- finite variance: \( \text{Var}[\epsilon_i] = \sigma^2 < \infty \) for all \( i \)
- uncorrelated: \( \text{Cov}[\epsilon_i, \epsilon_j] = 0 \) for all \( i \neq j \).
Let $\tilde{a} = Cy$ denote another unbiased linear estimator of $a$ where

$$C = (X^\top X)^{-1}X^\top + D.$$  

(18)

We want to show that $\text{Var}[\tilde{a}] \geq \text{Var}[\hat{a}]$.

First, the expectation of $\tilde{a}$ is

$$E[Cy] = E[((X^\top X)^{-1}X^\top + D)(Xa + \epsilon)]$$

$$= ((X^\top X)^{-1}X^\top + D)Xa + ((X^\top X)^{-1}X^\top + D)E[\epsilon]$$

$$= (X^\top X)^{-1}X^\top Xa + DXa$$

$$= (I + DX)a.$$

Since $\tilde{a}$ is unbiased, $DX = 0$. 

Leow Wee Kheng (NUS)  
Statistical Modeling 2
Next, the variance of $\tilde{a}$ is

$$
\text{Var}[C\mathbf{y}] = C\text{Var}[\mathbf{y}]C^\top
$$

$$
= C\text{Var}[\epsilon]C^\top \quad \text{(only $\epsilon$ is random)}
$$

$$
= \sigma^2 CC^\top
$$

$$
= \sigma^2 ((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D})(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{D}^\top)
$$

$$
= \sigma^2 ((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}^\top +
\mathbf{D}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{D}\mathbf{D}^\top))
$$

$$
= \sigma^2 (\mathbf{X}^\top\mathbf{X})^{-1} + 0 + 0 + \sigma^2 \mathbf{D}\mathbf{D}^\top
$$

$$
= \text{Var}[\hat{a}] + \sigma^2 \mathbf{D}\mathbf{D}^\top.
$$

Since $\mathbf{D}\mathbf{D}^\top$ is always positive semi-definite, we have

$$
\text{Var}[\tilde{a}] \geq \text{Var}[\hat{a}]. \quad (19)
$$

So, linear least-square estimate $\hat{a}$ is BLUE.
Nonlinear regression studies nonlinear relationships between variables.
Nonlinear regression model is given by

\[ y_i = f(x_i; a) + \epsilon_i, \quad i = 1, \ldots, n \]  

(20)

where \( f(x_i; a) \) is a nonlinear function.

As with nonlinear fitting, there are two basic approaches:

- If \( f(x_i; a) \) is linear with respect to \( a \), then linear least square can be applied to solve for \( \hat{a} \).
- Otherwise, optimization algorithms are needed, e.g.,
  - gradient descent
  - Gauss-Newton
  - Levenberg-Marquardt
  - quasi-Newton
  - Newton
Mixture Models

In previous lecture...

A nonlinear function $f(x; w)$ can be represented by a linear combination of basis functions $f_j(x), j = 1, \ldots, m$:

$$f(x; w) = \sum_{j=1}^{m} w_j f_j(x)$$  \hspace{1cm} (21)

where weights or mixing proportions $w = [w_1 \cdots w_m]^\top$.

When $f_j(x)$ are pdf, and

$$\sum_{j=1}^{m} w_j = 1,$$  \hspace{1cm} (22)

Eq. 21 is called a mixture model.
A mixture model with Gaussian pdf is a **Gaussian mixture**. Gaussian mixture is good for modeling **multimodal distribution**.
As shown previously, MLE of parameters of single pdf is easy.

But a mixture model has multiple pdf parameters and $w_j$, with a constraint

$$
\sum_{j=1}^{m} w_j = 1. \quad (23)
$$

This is a constrained optimization problem. Need a good constrained optimization method.
Expectation-maximization (EM) is a general approach for MLE.

EM models **complete** vs. **incomplete data**.

- **X** is random vector of **complete data** \( x \) having pdf \( f(x; \phi) \).
- **Y** is random vector of **observed data** \( y \) having pdf \( g(y; \phi) \).
- **Z** is random vector of **unobserved or unobservable data** \( z \).
- \( x = [y^\top \ z^\top]^\top \).
\( x \in \mathcal{X} \) is not directly observable.
\( x \in \mathcal{X}(y) \subset \mathcal{X} \) is observed indirectly through \( y \in \mathcal{Y} \):

\[
y = y(x).
\]

Incomplete-data pdf \( g(y; \phi) \) is related to complete-data pdf \( f(x; \phi) \) by

\[
g(y; \phi) = \int_{\mathcal{X}(y)} f(x; \phi) dx. \tag{24}
\]
The goal is to find the $\phi$ that maximizes complete-data log-likelihood

$$
\log L(\phi; x) = \log f(x; \phi).
$$  \hfill (25)

Since $f(x; \phi)$ is unobservable, we maximize the expectation of $\log L$, given the observed data $y$ and using current parameter $\phi^{(k)}$:

$$
Q(\phi; \phi^{(k)}) = E[\log L(\phi; x) | y],
$$  \hfill (26)

where $k$ is the current iteration number.

The incomplete-data likelihood is measurable because $y$ is observed:

$$
L(\phi; y) = g(y; \phi).
$$  \hfill (27)
**Expectation-Maximization**

**E-Step:** Compute expectation of complete-data log-likelihood:

\[ Q(\phi; \phi^{(k)}) = \mathbb{E}[\log L(\phi; x) | y]. \]

**M-Step:** Find \( \phi^{(k+1)} \) that maximizes \( Q(\phi; \phi^{(k)}) \):

\[ \phi^{(k+1)} = \arg \max Q(\phi; \phi^{(k)}). \]

EM performs alternating optimization.
EM has a surprisingly simple structure.
But, finding \( \phi^{(k+1)} \) that maximizes \( Q(\phi; \phi^{(k)}) \) can be difficult.
Instead, finding \( \phi^{(k+1)} \) that increases \( Q(\phi; \phi^{(k)}) \) is easier.
**Expectation-Maximization**

**E-Step:** Compute expectation of complete-data log-likelihood:

$$Q(\phi; \phi^{(k)}) = E[\log L(\phi; x) | y].$$

**M-Step:** Find any $\phi^{(k+1)}$ that increases $Q(\phi; \phi^{(k)})$:

$$Q(\phi^{(k+1)}; \phi^{(k)}) \geq Q(\phi; \phi^{(k)}).$$

E-step and M-step are repeated until convergence, e.g.,

$$L(\phi^{(k+1)}; y) - L(\phi^{(k)}; y) < \epsilon.$$

What happen to $z$?
It has been proved that EM algorithm always converges [1, 2].
EM for Mixing Proportions

Consider a mixture model in which only the weights $w_j$ are unknown:

$$f(y; w) = \sum_{j=1}^{m} w_j f_j(y).$$

(28)

Then, $\phi = w = [w_1 \cdots w_{m-1}]^\top$ because $w_j$ are normalized:

$$w_m = 1 - \sum_{j=1}^{m-1} w_j.$$  

(29)

Suppose $n$ iid data $y_1, \ldots, y_n$ are observed. Let $y = [y_1^\top \cdots y_n^\top]^\top$.

The likelihood of observing $y$ is

$$L(\phi; y) = \prod_{i=1}^{n} f(y_i; w) = \prod_{i=1}^{n} \sum_{j=1}^{m} w_j f_j(y_i).$$

(30)
The log-likelihood of observing $y$ is

$$
\log L(\phi; y) = \sum_{i=1}^{n} \log \sum_{j=1}^{m} w_j f_j(y_i).
$$

(31)

This equation is difficult to maximize because of the log of summation.

To apply EM, introduce unobserved data $z = [z_{ij}]$.

- $z_{ij}$ indicates which component of the mixture does $y_i$ comes from.

$$
z_{ij} = \begin{cases} 
1 & \text{if } y_i \text{ comes from } j\text{-th component}, \\
0 & \text{otherwise.}
\end{cases}
$$

(32)

- Each $y_i$ comes from only one of the mixture components. For each $i$, $z_{ij} = 1$ for a particular $j$ and 0 otherwise.
Expectation-Maximization
EM for Mixing Proportions

**Example**

<table>
<thead>
<tr>
<th>observed data</th>
<th>component</th>
<th>pdf</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_1$</td>
<td>2</td>
<td>$f_2$</td>
</tr>
<tr>
<td>$y_2$</td>
<td>3</td>
<td>$f_3$</td>
</tr>
<tr>
<td>$y_3$</td>
<td>1</td>
<td>$f_1$</td>
</tr>
<tr>
<td>$y_4$</td>
<td>2</td>
<td>$f_2$</td>
</tr>
</tbody>
</table>

Then, the likelihood of observing $y_1, \ldots, y_4$ is

$$L(\phi; y) = \prod_{i=1}^{4} f(y_i; w)$$

$$= (w_2 f_2(y_1)) (w_3 f_3(y_2)) (w_1 f_1(y_3)) (w_2 f_2(y_4)).$$

Now, combine $y$ and $z$ to form complete data $x = [y^\top \ z^\top]^\top$.

Then, the complete-data likelihood of $x$ can be written as

$$L(\phi; x) = \prod_{i=1}^{4} \prod_{j=1}^{3} (w_j f_j(y_i))^{z_{ij}}.$$
So, by introducing unobserved data $z$ as shown previously, the complete-data likelihood function of $x_i$ becomes

$$L(\phi; x) = \prod_{i=1}^{n} \prod_{j=1}^{m} (w_j f_j(y_i))^{z_{ij}}. \quad (33)$$

Then, the complete-data log-likelihood is

$$\log L(\phi; x) = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} (\log w_j + \log f_j(y_i)). \quad (34)$$

In the E-Step, we want to compute

$$Q(\phi; \phi^{(k)}) = \text{E}[\log L(\phi; x) | y]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \text{E}[z_{ij} | y] (\log w_j + \log f_j(y_i)). \quad (35)$$
The expectation of $z_{ij}$ given $y$ is simply itself:

$$E[z_{ij} \mid y] = z_{ij}. \quad (36)$$

At the same time,

$$E[z_{ij} \mid y] = 1 \cdot P(z_{ij} = 1 \mid y) + 0 \cdot P(z_{ij} = 0 \mid y)$$

$$= P(z_{ij} = 1 \mid y)$$

$$= \frac{P(z_{ij} = 1 \land y)}{P(y)}$$

$$= \frac{w_j f_j(y_i)}{f(y_i; \mathbf{w})}. \quad (37)$$

So,

$$z_{ij}^{(k)} = \frac{w_j^{(k)} f_j(y_i)}{f(y_i; \mathbf{w}^{(k)})}. \quad (38)$$
In the M-Step, find \( w \) that either maximizes or increases \( Q(\phi; \phi^{(k)}) \).

Let’s first try maximization method.

\[
\frac{\partial Q(\phi; \phi^{(k)})}{\partial w_j} = \frac{\partial \log L(\phi; x)}{\partial w_j} = \sum_{i=1}^{n} \frac{z_{ij}}{w_j} \neq 0. \tag{39}
\]

This leads to \( z_{ij} = 0 \) which is not useful.

Next, try “counting” method.

For example, in the previous example,

\[
z = \begin{bmatrix}
0 & 1 & 0 \\
0 & 0 & 1 \\
1 & 0 & 0 \\
0 & 1 & 0
\end{bmatrix}.
\]

Then, MLE of \( w_j \) is just the average of column \( j \):

\[
w_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} z_{ij}^{(k)}. \tag{40}
\]
EM for Mixing Proportions

E-Step: Compute, for $i = 1, \ldots, n$, $j = 1, \ldots, m$,

$$z_{ij}^{(k)} = \frac{w_j^{(k)} f_j(y_i)}{f(y_i; w^{(k)})}. \quad (41)$$

M-Step: Compute, for $j = 1, \ldots, m$,

$$w_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} z_{ij}^{(k)}. \quad (42)$$
Gaussian mixture in $h$-dimensional space is given by

$$G(y; \phi) = \sum_{j=1}^{m} w_j g_j(y; \mu_j, \Sigma_j), \quad (43)$$

where $g_j(y; \mu_j, \Sigma_j)$ is a $h$-D Gaussian

$$g_j(y; \mu_j, \Sigma_j) = \frac{1}{\sqrt{(2\pi)^h |\Sigma_j|}} \exp \left( -\frac{1}{2} (y - \mu_j)^\top \Sigma_j^{-1} (y - \mu_j) \right), \quad (44)$$

$\mu_j$ and $\Sigma_j$ are the mean vectors and covariance matrices, and $\phi$ contains $w_j, \mu_j, \Sigma_j$, for $j = 1, \ldots, m$.

Follow the same reasoning as in the previous example...

Introduce unobserved data $z$ to form the complete data $x = [y^\top \ z^\top]^\top$. 

Leow Wee Kheng (NUS)  Statistical Modeling 2
Complete-data likelihood is

\[
L(\phi; \mathbf{x}) = \prod_{i=1}^{n} \prod_{j=1}^{m} (w_j g_j (y_i; \mu_j, \Sigma_j))^{z_{ij}}. \tag{45}
\]

Then, the complete-data log-likelihood is

\[
\log L(\phi; \mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} (\log w_j + \log g_j (y_i; \mu_j, \Sigma_j))
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \left( \log w_j - \frac{h}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_j| \right) - \frac{1}{2} (y_i - \mu_j)^\top \Sigma_j^{-1} (y_i - \mu_j). \tag{46}
\]
In the **E-Step**, as in the previous example, we get

\[ z_{ij}^{(k)} = \frac{w_j^{(k)} g_j(y_i; \mu_j, \Sigma_j)}{G(y_i; \phi)}. \]  

(47)

In the **M-Step**, as in the previous example, we get

\[ w_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} z_{ij}^{(k)}. \]  

(48)

To get MLE of \( \mu_j \) and \( \Sigma_j \), have to differentiate \( \log L \) wrt \( \mu_j \) and \( \Sigma_j \).
Trace of Matrix

The trace of a matrix $A$ is the sum of its diagonal elements $a_{ii}$:

$$\text{tr}(A) = \sum_{i=1}^{n} a_{ii}. \quad (49)$$

Strictly speaking, scalar $c$ and a $1\times1$ matrix $[c]$ are not the same thing. Nevertheless, since $\text{tr}([c]) = c$, we often write, for simplicity $\text{tr}(c) = c$.

Properties

- $\text{tr}(A) = \sum_{i=1}^{n} \lambda_i$, where $\lambda_i$ are the eigenvalues of $A$.
- $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$
- $\text{tr}(cA) = c \text{tr}(A)$
- $\text{tr}(A) = \text{tr}(A^\top)$
- $\text{tr}(ABCD) = \text{tr}(BCDA) = \text{tr}(CDAB) = \text{tr}(DABC)$
Derivatives of Trace

For variable matrix $X$ and constant matrices $A$, $B$, $C$,

(DT1) \[ \frac{\partial \text{tr}(X)}{\partial X} = I \]

(DT2) \[ \frac{\partial \text{tr}(X^k)}{\partial X} = kX^{k-1} \]

(DT3) \[ \frac{\partial \text{tr}(AX)}{\partial X} = \frac{\partial \text{tr}(XA)}{\partial X} = A \]

(DT4) \[ \frac{\partial \text{tr}(AX^\top)}{\partial X} = \frac{\partial \text{tr}(X^\top A)}{\partial X} = A^\top \]

(DT5) \[ \frac{\partial \text{tr}(X^\top AX)}{\partial X} = X^\top (A + A^\top) \]
\[
\frac{\partial \text{tr}(X^{-1}A)}{\partial X} = -X^{-\top}AX^{-\top}
\]

\[
\frac{\partial \text{tr}(AXB)}{\partial X} = \frac{\partial \text{tr}(BAX)}{\partial X} = BA
\]

\[
\frac{\partial \text{tr}(AXBX^{\top}C)}{\partial X} = BX^{\top}CA + B^{\top}X^{\top}A^{\top}C^{\top}
\]
Derivatives of Determinant

For variable matrix $X$ and constant matrices $A$, $B$,

(DD1) \[ \frac{\partial |X|}{\partial X} = |X|X^{-1} \]

(DD2) \[ \frac{\partial |X^k|}{\partial X} = k|X^k|X^{-1} \]

(DD3) \[ \frac{\partial \log |X|}{\partial X} = X^{-1} \]

(DD4) \[ \frac{\partial |AXB|}{\partial X} = |AXB|X^{-1} \]

(DD5) \[ \frac{\partial |X^\top AX|}{\partial X} = 2|X^\top AX|X^{-1} \]
First, differentiate log $L$ with respect to $\mu_j$ and set to 0.

$$\frac{\partial \log L(\phi; x)}{\partial \mu} = \sum_{i=1}^{n} z_{ij} \frac{1}{2} (y_i - \mu_j)^\top (\Sigma_j^{-1} + \Sigma_j^{-\top}) = 0$$  \hspace{1cm} (C11)

Since $\Sigma_j$ is positive definite, the minimum is attained with

$$\sum_{i=1}^{n} z_{ij} (y_i - \mu) = 0$$

$$\sum_{i=1}^{n} z_{ij} \mu = \sum_{i=1}^{n} z_{ij} y_i.$$  

So,

$$\mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(k)} y_i}{\sum_{i=1}^{n} z_{ij}^{(k)}}.$$  \hspace{1cm} (50)
By applying the properties of trace, \( \log L \) can be written as

\[
\log L(\phi; x) = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \left( \log w_j - \frac{h}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_j| ight)

- \frac{1}{2} \text{tr} \left( \Sigma_j^{-1} (y_i - \mu_j)(y_i - \mu_j)^\top \right).
\] (51)

Next, differentiate \( \log L \) with respect to \( \Sigma_j \) and set to 0.

\[
\frac{\partial \log L(\phi; x)}{\partial \Sigma_j} = \sum_{i=1}^{n} z_{ij} \left( -\frac{1}{2} \Sigma_j^{-1} + \frac{1}{2} \Sigma_j^{-\top} (y_i - \mu_j)(y_i - \mu_j)^\top \Sigma_j^{-\top} \right) = 0 \tag{DT6}
\] (DD3)

Since \( \Sigma_j \) is symmetric, we obtain

\[
\sum_{i=1}^{n} z_{ij} \Sigma_j^{-1} = \Sigma_j^{-1} \sum_{i=1}^{n} z_{ij} (y_i - \mu_j)(y_i - \mu_j)^\top \Sigma_j^{-1} \tag{52}
\]
So,

\[
\Sigma_j^{(k+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(k)} (y_i - \mu_j^{(k+1)}) (y_i - \mu_j^{(k+1)})^T}{\sum_{i=1}^{n} z_{ij}^{(k)}}.
\]  

(53)
EM for Gaussian Mixture

E-Step: Compute, for $i = 1, \ldots, n$, $j = 1, \ldots, m$,

$$
z^{(k)}_{ij} = \frac{w_j^{(k)} g_j(y_i; \mu_j, \Sigma_j)}{G(y_i; \phi)}. \quad (54)
$$

M-Step: Compute, for $j = 1, \ldots, m$,

$$
w^{(k+1)}_j = \frac{1}{n} \sum_{i=1}^{n} z^{(k)}_{ij}. \quad (55)
$$

$$
\mu^{(k+1)}_j = \frac{\sum_{i=1}^{n} z^{(k)}_{ij} y_i}{\sum_{i=1}^{n} z^{(k)}_{ij}}. \quad (56)
$$

$$
\Sigma^{(k+1)}_j = \frac{\sum_{i=1}^{n} z^{(k)}_{ij} (y_i - \mu^{(k+1)}_j)(y_i - \mu^{(k+1)}_j)^\top}{\sum_{i=1}^{n} z^{(k)}_{ij}}. \quad (57)
$$
In using mixture models, how to determine the number $h$ of mixture components required for an application?

In applying EM, how to initialize the variables?

In applying EM, is there a way to check that the update equations of the unobserved variable $z_{ij}$ and parameters $\phi$ make sense?
1. Describe the essence of the following in one sentence each:
   - Linear Regression
   - Nonlinear Regression
   - Best linear unbiased estimator (BLUE)
   - Expectation-maximization

2. Show that, for simple linear regression,

   \[ \hat{a} = \frac{\text{Cov}[X,Y]}{\text{Var}[X]}, \quad \hat{b} = \bar{y} - \hat{a}\bar{x}. \]

3. Show that minimizing \( E \) (Eq. 15) of simple linear regression is equivalent to maximizing its log-likelihood (Eq. 16).
References


