

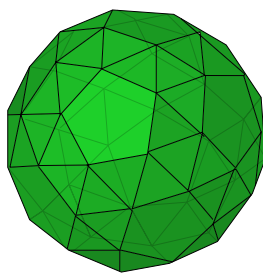
PROJECTS IN MATHEMATICS AND APPLICATIONS

# GAUSSIAN MIXTURE MODEL

Ngày 15 tháng 8 năm 2021

Huỳnh Tiến Dũng <sup>\*</sup>  
Châu Đăng Minh <sup>‡</sup>

<sup>†</sup>Nguyễn Minh Tú  
<sup>§</sup>Nguyễn Đức Triều



---

<sup>\*</sup>Trường THPT Chuyên Bảo Lộc

<sup>†</sup>Trường Phổ Thông Năng Khiếu

<sup>‡</sup>Trường ĐH Bách Khoa

<sup>§</sup>Trường THPT chuyên Hà Nội - Amsterdam

# Lời cảm ơn

Lời đầu tiên, chúng em xin gửi lời cảm ơn chân thành nhất đến founders của trại hè, anh Cấn Trần Thành Trung, anh Trần Hoàng Bảo Linh, anh Lê Việt Hải cùng ban tổ chức PiMA đã xây dựng, duy trì và phát triển chương trình 5 năm liên tục để chúng em có cơ hội trải nghiệm trại hè Toán ứng dụng bổ ích và thú vị.

Bên cạnh đó, chúng em cảm ơn anh, chị Mentors đứng lớp giảng dạy kiến thức về Giải tích, Đại số tuyến tính, Xác suất thống kê và Học máy trong tuần lễ đầu tiên. Sự nhiệt huyết của anh, chị đã tiếp thêm năng lượng cho chúng em cháy hết mình trong những tuần lễ còn lại. Tiếp đến, chúng em muốn gửi lời cảm ơn đặc biệt đến anh, chị Mentors, chị Phạm Thanh Ngọc, anh Nguyễn Hồ Thăng Long và anh Vũ Lê Thế Anh đã theo dõi và giúp đỡ rất tận tình nhóm trong quá trình hoàn thành dự án. Đề tài của chúng em sẽ không thể hoàn thiện nếu thiếu đi sự giúp đỡ của anh, chị.

Trong những ngày hè vừa qua, được đồng hành với các bạn Mentees và các anh, chị Mentors từ khắp nơi trên thế giới, chúng em xin cảm ơn mọi người vì trải nghiệm đầy ý nghĩa và khó quên này. Hi vọng trong những năm tiếp tới, PiMA có thể tiếp tục phát triển và truyền lửa đam mê cho nhiều bạn THPT khác.

Xin cảm ơn.

*Ngày 15 tháng 8 năm 2021.*

*Nhóm 4.*

# Tóm tắt nội dung

Bài báo cáo mở đầu bằng trình bày sơ lược về khái niệm hỗn hợp Gaussian (Mixture of gaussian) và cách dữ liệu được tạo ra dưới góc nhìn của Gaussian mixture model (**GMM**). Tiếp đến sẽ trình bày về **GMM** dưới góc độ toán học và các bước cập nhật tham số của mô hình. Sau đó, ta trình bày về một thuật toán tổng quát để ước lượng tham số của mô hình thống kê khi xuất hiện biến ẩn, thuật toán Expectation-Maximization **EM**, và xây dựng lại các công thức cập nhật tham số trong **GMM**. Ngoài ra, ta mở rộng thuật toán **EM** thành một thuật toán học bán giám sát và áp dụng vào **GMM**. Tiếp đến, ta áp dụng **GMM** trên các tập dữ liệu tự khởi tạo và tập dữ liệu thực với mục đích phân cụm và khởi tạo thêm nhiều điểm dữ liệu mới. Cuối cùng, ta thảo luận về cách chọn số cụm cho **GMM** trong thực nghiệm.

# Mục lục

<b>1</b>	<b>Định nghĩa và tính chất</b>	<b>2</b>
1.1	Định nghĩa . . . . .	2
1.2	Tính chất . . . . .	3
<b>2</b>	<b>Tổng quan về Gaussian Mixture Model (GMM)</b>	<b>4</b>
2.1	Ví Dụ . . . . .	4
2.2	Cách dữ liệu được tạo ra trong GMM . . . . .	5
2.3	Mục tiêu của GMM . . . . .	5
<b>3</b>	<b>Gaussian Mixture Model (GMM)</b>	<b>6</b>
3.1	Mô hình toán học . . . . .	6
3.2	Sử dụng ước lượng hợp lý cực đại (MLE) . . . . .	6
3.3	Sử dụng thuật toán EM trong GMM . . . . .	7
<b>4</b>	<b>Thuật toán Expectation-maximization (EM)</b>	<b>9</b>
4.1	Bất đẳng thức Jensen . . . . .	9
4.2	Mô hình toán học . . . . .	9
4.3	Xây dựng hàm evidence lower bound (ELBO) . . . . .	10
4.4	Hoàn thành thuật toán EM . . . . .	11
4.5	Xây dựng EM trong GMM . . . . .	15
4.6	Kỳ vọng và tối đa trong EM . . . . .	18
<b>5</b>	<b>Thuật toán EM bán giám sát (Semi-supervised EM)</b>	<b>20</b>
5.1	GMM bán giám sát . . . . .	22
<b>6</b>	<b>Áp dụng mô hình - Kết luận đánh giá</b>	<b>23</b>
6.1	Visualize GMM sau các bước thuật toán EM . . . . .	23
6.2	Một số datasets khác sử dụng GMM . . . . .	23
6.3	GMM như một generative model . . . . .	25
6.4	Chọn số cụm cho GMM . . . . .	27
6.5	GMM bán giám sát . . . . .	28

# Danh pháp tiếng Anh

Những thuật ngữ được dùng trong bài báo cáo vào danh pháp tiếng Anh của chúng:

1. Bán giám sát: Semi-supervised	22. Kỳ vọng: Expectation
2. Biến ẩn: Latent Variable	23. Kỳ vọng của Log-Likelihood đầy đủ: Expectation of Complete Log-Likelihood
3. Biến ngẫu nhiên: Random variable	24. Lồi: Convex
4. Chặt: Strict	25. Lõm: Concave
5. Cụm: Cluster	26. Ma trận: Matrix
6. Đạo hàm riêng : Partial Derivative	27. Ma trận hiệp phương sai: Covariance matrix
7. Điểm dữ liệu: Data point	28. Nhân tử Lagrange: Lagrange multiplier
8. Độc lập thống kê: Independent and identically distributed	29. Phân phối: Distribution
9. Đối xứng: Symmetric	30. Phân phối chuẩn: Gaussian distribution
10. Đơn điệu: Monotonic	31. Phân phối chuẩn nhiều chiều: Multivariate gaussian distribution
11. Đường cong chuông: Bell shape	32. Phương sai: Variance
12. Giá trị trung bình: Mean	33. Siêu tham số: Hyperparameter
13. Hàm chỉ thị: Indicator function	34. Tập dữ liệu: Dataset
14. Hàm khối xác suất: Probability mass function	35. Thuật toán lặp: Iterative algorithm
15. Hàm mật độ xác suất: Probability density function	36. Tiên nghiệm: Prior
16. Hàm xác suất: Probability function	37. Ước lượng hợp lý cực đại: Maximum likelihood estimate
17. Hậu nghiệm: Posterior	38. Vector ngẫu nhiên: Random vector
18. Học giám sát: Supervised learning	
19. Học không giám sát: Unsupervised learning	
20. Hỗn hợp Gaussian: Mixture of gaussian	
21. Khả nghịch: Invertible	

# 1 Định nghĩa và tính chất

## 1.1 Định nghĩa

**Định nghĩa 1.1. Phân phối Categorical.** Xét một thí nghiệm ngẫu nhiên có  $k$  kết quả  $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ . Xác suất cho kết quả  $\omega_j$  là  $\phi_j$ , với  $\sum_{j=1}^k \phi_j = 1$ . Gọi  $Z$  là biến ngẫu nhiên thỏa  $Z(\omega_j) = j$ . Ta nói  $Z$  có **phân phối Categorical** khi  $Z$  có hàm khối xác suất:

$$p_Z(z = j; \phi) \triangleq \begin{cases} \phi_j & \text{khi } 1 \leq j \leq k \\ 0 & \text{khi } j \text{ khác} \end{cases}$$

Kí hiệu:  $Z \sim \text{Categorical}(\phi)$ . Trong đó  $\phi = \{\phi_1, \phi_2, \dots, \phi_k\}$

**Định nghĩa 1.2. Vector ngẫu nhiên** là một hàm thực hiện ánh xạ từ không gian mẫu  $\Omega$  đến không gian vector  $\mathbb{R}^d$

$$X : \Omega \rightarrow \mathbb{R}^d$$

Với  $d$  biến ngẫu nhiên  $X_1, X_2, \dots, X_d$ . **Vector ngẫu nhiên**  $X$  được viết:

$$X \triangleq \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{bmatrix}$$

**Định nghĩa 1.3. Hàm xác suất của vector ngẫu nhiên** là hàm xác suất kết hợp của những biến ngẫu nhiên thành phần. Một vector ngẫu nhiên  $X = [X_1, X_2, \dots, X_d]^T \in \mathbb{R}^d$  có hàm xác suất  $f_X : \mathbb{R}^d \rightarrow \mathbb{R}$  được định nghĩa:

$$f_X(x) \triangleq f_{X_1, X_2, \dots, X_d}(x_1, x_2, \dots, x_d)$$

**Định nghĩa 1.4. Phân phối chuẩn nhiều chiều.** Một vector ngẫu nhiên  $X \in \mathbb{R}^d$  có **phân phối chuẩn nhiều chiều** với tham số: vector giá trị trung bình  $\mu \in \mathbb{R}^d$  và ma trận hiệp phương sai  $\Sigma \in \mathbb{R}^{d \times d}$  (Với ma trận  $\Sigma$  đối xứng xác định dương). Khi đó  $X$  có hàm mật độ xác suất:

$$f_X(x; \mu, \Sigma) \triangleq \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

Kí hiệu:  $X \sim \mathcal{N}(\mu, \Sigma)$

**Định nghĩa 1.5. Hàm chỉ thị.** Với  $A$  là một mệnh đề, ta có định nghĩa:

$$\mathbb{1}\{A\} \triangleq \begin{cases} 1 & \text{nếu } A \text{ đúng.} \\ 0 & \text{ngược lại.} \end{cases}$$

## 1.2 Tính chất

Những đạo hàm liên quan đến ma trận  $A \in \mathbb{R}^{d \times d}$  **đối xứng, khả nghịch** và vector  $x, y \in \mathbb{R}^d$  được dùng:

$$\begin{aligned}\frac{\partial}{\partial x} [x^T y] &= y \\ \frac{\partial}{\partial x} [x^T A x] &= 2Ax \\ \frac{\partial}{\partial A} [x^T A x] &= x x^T \\ \frac{\partial}{\partial A} [\log |A|] &= A^{-1}\end{aligned}$$

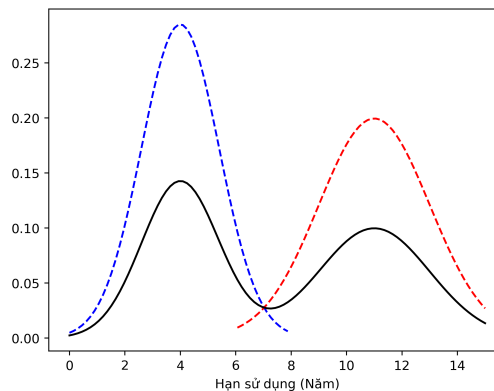
## 2 Tổng quan về Gaussian Mixture Model (GMM)

Gaussian Mixture Model (**GMM**) là mô hình xác suất miêu tả các quần thể con của một tập dữ liệu không gắn nhãn. Mỗi quần thể có thể khác nhau nhưng những điểm dữ liệu trong cùng một quần thể có thể được mô hình bằng một phân phối chuẩn. **GMM** là mô hình học không giám sát.

Cách tốt nhất để hiểu **GMM** là qua một vài ví dụ.

### 2.1 Ví Dụ

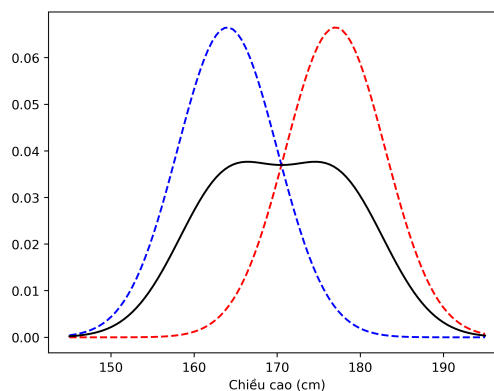
Hạn sử dụng một chiếc tivi loại thường sẽ phân phối theo phân phối chuẩn có giá trị trung bình là 4 năm và phương sai là 1.4 năm. Tương tự, hạn sử dụng một chiếc tivi loại tốt sẽ phân phối theo phân phối chuẩn có giá trị trung bình là 11 năm và phương sai là 2 năm.



Hình 1: Hàm xác suất của hạn sử dụng tivi loại thường (Đường nét đứt màu xanh), tivi loại tốt (Đường nét đứt màu đỏ), hai loại tivi (Đường nét liền màu đen)

Mặc dù được tổng hợp từ hai phân phối chuẩn nhưng phân phối của cả hai loại tivi không phải là một phân phối chuẩn. Một phân phối chuẩn phải có dạng đường cong chuông và tiến dần về không khi đi xa giá trị trung bình.

Một ví dụ khác, chiều cao của một người đàn ông sẽ phân phối theo phân phối chuẩn có giá trị trung bình là 177cm và phương sai là 6cm. Chiều cao của một người phụ nữ sẽ phân phối theo phân phối chuẩn có giá trị trung bình là 164cm và phương sai là 6cm.



Hình 2: Hàm xác suất của chiều cao phụ nữ (Đường nét đứt màu xanh), chiều cao đàn ông (Đường nét đứt màu đỏ), cả hai giới (Đường nét liền màu đen)

Trong trường hợp này, phân phối tổng hợp nhìn giống một phân phối chuẩn hơn. Nhưng đây vẫn không phải là một phân phối chuẩn.

Cả hai ví dụ trên là **hỗn hợp Gaussian (mixture of Gaussian)**: một phân phối được tổng hợp từ những phân phối chuẩn.



## 2.2 Cách dữ liệu được tạo ra trong GMM

Cho **hỗn hợp Gaussian** gồm  $k$  cụm, ứng với mỗi cụm  $j$  là một phân phối chuẩn được tham số bởi vector giá trị trung bình  $\mu_j$  và ma trận hiệp phương sai  $\Sigma_j$ . Một điểm dữ liệu được hình thành từ hai bước:

1. Chọn ngẫu nhiên một trong  $k$  cụm, với xác suất chọn được cụm  $j$  là  $\phi_j$ .
2. Giả sử cụm được chọn là  $j$ , thì điểm dữ liệu lấy từ phân phối chuẩn tham số bởi  $\mu_j, \Sigma_j$ .

## 2.3 Mục tiêu của GMM

Hiểu được một điểm dữ liệu được tạo ra từ hai bước trên. **GMM** sẽ mô hình lại quá trình tạo ra dữ liệu bằng cách tìm các tham số  $\phi_j, \mu_j$  và  $\Sigma_j$ . Sau khi tìm lại được các tham số, ta có thể phân cụm các điểm dữ liệu hoặc tạo thêm nhiều điểm dữ liệu mới từ tham số có được.

## 3 Gaussian Mixture Model (GMM)

### 3.1 Mô hình toán học

Ta có tập dữ liệu  $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ , với mỗi  $x^{(i)} \in \mathbb{R}^d$  là một giá trị xác định của vector ngẫu nhiên  $X^{(i)}$ . Các Vector ngẫu nhiên  $X^{(i)}$  độc lập thống kê với nhau. Đồng thời ta có  $n$  biến ngẫu nhiên ẩn  $Z^{(1)}, Z^{(2)}, \dots, Z^{(n)}$  với  $Z^{(i)} : \Omega \rightarrow \{1, 2, \dots, k\}$ , mỗi giá trị xác định  $z^{(i)}$  của biến ngẫu nhiên  $Z^{(i)}$  thể hiện cho cụm mà điểm dữ liệu  $x^{(i)}$  thuộc vào.

Ta giả sử  $Z^{(i)} \sim \text{Categorical}(\phi)$ , trong đó  $\phi = \{\phi_1, \phi_2, \dots, \phi_k\}$ . Khi đó, hàm khối xác suất của biến ngẫu nhiên  $Z^{(i)}$  là:

$$p_{Z^{(i)}}(z^{(i)} = j; \phi) = \begin{cases} \phi_j & \text{khi } 1 \leq j \leq k \\ 0 & \text{khi } j \text{ khác} \end{cases}$$

Ta có giả sử tiếp theo,  $X^{(i)}|Z^{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j)$ . Khi đó, hàm mật độ xác suất của vector ngẫu nhiên  $X^{(i)}|Z^{(i)} = j$  là:

$$f_{X^{(i)}|Z^{(i)}=j}(x^{(i)}; \mu_j, \Sigma_j) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1}(x^{(i)} - \mu_j)\right)$$

Để ngắn gọn trong việc trình bày bài toán, ta sẽ viết lại các hàm xác suất trên như sau:

$$\begin{aligned} p(z^{(i)} = j; \phi) &\triangleq p_{Z^{(i)}}(z^{(i)} = j; \phi) \\ p(x^{(i)}|z^{(i)} = j; \mu, \Sigma) &\triangleq f_{X^{(i)}|Z^{(i)}=j}(x^{(i)}; \mu_j, \Sigma_j) \end{aligned}$$

Tóm lại, những tham số có trong **GMM** là:

- $\phi_1, \phi_2, \dots, \phi_k$  với  $\phi_i \in \mathbb{R}$
- $\mu_1, \mu_2, \dots, \mu_k$  với  $\mu_i \in \mathbb{R}^d$
- $\Sigma_1, \Sigma_2, \dots, \Sigma_k$  với  $\Sigma_i \in \mathbb{R}^{d \times d}$

Khi đó, một cách toán học, điểm dữ liệu  $x^{(i)}$  sẽ được hình thành từ hai bước:

1.  $z^{(i)}$  được lấy ngẫu nhiên từ phân phối  $Z^{(i)} \sim \text{Categorical}(\phi)$ .
2. Giả sử  $z^{(i)} = j$ , điểm dữ liệu  $x^{(i)}$  được lấy ngẫu nhiên từ phân phối  $\mathcal{N}(\mu_j, \Sigma_j)$ .

### 3.2 Sử dụng ước lượng hợp lý cực đại (MLE)

Mục tiêu của **GMM** là tìm lại các tham số  $\phi, \mu, \Sigma$  ban đầu. Ta thử tiếp cận việc tìm lại các tham số ban đầu bằng **MLE**. Ta có hàm Log-Likelihood:

$$\begin{aligned} l(\phi, \mu, \Sigma) &= \sum_{i=1}^n \log p(x^{(i)}; \phi, \mu, \Sigma) \\ &= \sum_{i=1}^n \log \sum_{z^{(i)}=1}^k p(x^{(i)}|z^{(i)}; \mu, \Sigma) p(z^{(i)}; \phi) \\ &= \sum_{i=1}^n \log \sum_{j=1}^k \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1}(x^{(i)} - \mu_j)\right) \phi_j \end{aligned}$$

Một phân phối chuẩn nhiều biến sẽ có đạo hàm riêng theo  $\mu$  là:

$$\begin{aligned}
\frac{\partial}{\partial \mu} [p(x; \mu, \Sigma)] &= \frac{\partial}{\partial \mu} \left[ \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right) \right] \\
&= \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right) \frac{\partial}{\partial \mu} \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \\
&= \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right) \Sigma^{-1} (x - \mu) \\
&= p(x; \mu, \Sigma) \Sigma^{-1} (x - \mu)
\end{aligned}$$

Khi đó, ta lấy đạo hàm riêng của hàm Log-Likelihood theo biến  $\mu_j$ , ta được:

$$\begin{aligned}
\frac{\partial}{\partial \mu_j} [l(\phi, \mu, \Sigma)] &= \sum_{i=1}^n \frac{1}{\sum_{r=1}^k p(x^{(i)} | z^{(i)} = r; \mu, \Sigma) \phi_r} \frac{\partial}{\partial \mu_j} \left[ \sum_{r=1}^k p(x^{(i)} | z^{(i)} = r; \mu, \Sigma) \phi_r \right] \\
&= \sum_{i=1}^n \frac{1}{\sum_{r=1}^k p(x^{(i)} | z^{(i)} = r; \mu, \Sigma) \phi_r} \frac{\partial}{\partial \mu_j} [p(x^{(i)} | z^{(i)} = j; \mu, \Sigma) \phi_j] \\
&= \sum_{i=1}^n \frac{1}{\sum_{r=1}^k p(x^{(i)} | z^{(i)} = r; \mu, \Sigma) \phi_r} p(x^{(i)} | z^{(i)} = j; \mu, \Sigma) \phi_j \Sigma_j^{-1} (x^{(i)} - \mu_j)
\end{aligned}$$

Ta cho đạo hàm riêng của hàm Log-Likelihood theo biến  $\mu_j$  bằng không, ta được phương trình:

$$\frac{\partial}{\partial \mu_j} [l(\phi, \mu, \Sigma)] = 0$$

Ta nhận xét phương trình trên vừa có phần **tuyến tính**, **hàm mũ** và **hàm mũ dưới mẫu số**. Nên việc giải phương trình trên là không khả thi !!!

### 3.3 Sử dụng thuật toán EM trong GMM

Ta sẽ phát biểu kết quả thu được khi sử dụng thuật toán **EM** trong **GMM** trước, sau đó sẽ xây dựng thuật toán **EM** một cách tổng quát sau.

Thuật toán **EM** là thuật toán lặp (iterative algorithm), gồm 2 bước: bước E và bước M. Cụ thể hơn, trong **GMM**, bước E đang xấp xỉ tốt nhất những giá trị biến ẩn  $z^{(i)}$ , bước M cập nhật các tham số của mô hình dựa trên các xấp xỉ ở bước E. Việc cập nhật các tham số ở bước M sẽ không quá khó vì ta xem như các giá trị xấp xỉ được ở bước E là chính xác, sau đó ta thực hiện bài toán tối đa trên các giá trị xấp xỉ đó.

Thuật toán **EM** trong **GMM** được trình bày như sau:

Khởi tạo các giá trị  $\phi_j, \mu_j, \Sigma_j$  ngẫu nhiên

Thực hiện cho đến khi thuật toán hội tụ: {

(Bước E) Với mỗi giá trị  $i, j$ :

$$w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

(Bước M) Cập nhật các tham số:

$$\begin{aligned}\phi_j &:= \frac{1}{n} \sum_{i=1}^n w_j^{(i)} \\ \mu_j &:= \frac{\sum_{i=1}^n w_j^{(i)} x^{(i)}}{\sum_{i=1}^n w_j^{(i)}} \\ \Sigma_j &:= \frac{\sum_{i=1}^n w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^n w_j^{(i)}}\end{aligned}$$

}

Điều kiện hội tụ:  $|l(\phi^{(t+1)}, \mu^{(t+1)}, \Sigma^{(t+1)}) - l(\phi^{(t)}, \mu^{(t)}, \Sigma^{(t)})| \leq \epsilon$

Ở bước E, ta tính xác suất hậu nghiệm của  $z^{(i)}$  khi biết  $x^{(i)}$  bằng định lý Bayes.

$$p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)} | z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{\sum_{l=1}^k p(x^{(i)} | z^{(i)} = l; \mu, \Sigma) p(z^{(i)} = l; \phi)}$$

Kết quả thu được sau mỗi vòng lặp EM:

- Bước E:  $n \times d$  giá trị  $w$ .
- Bước M:  $k$  tham số  $\phi$ ,  $k$  tham số  $\mu$ ,  $k$  tham số  $\Sigma$ .

Ở đây, giá trị  $w_j^{(i)}$  được hiểu là xác suất của điểm dữ liệu  $x^{(i)}$  thuộc vào phân phối  $\mathcal{N}(\mu_j, \Sigma_j)$ . Trong **GMM** mọi điểm dữ liệu  $x^{(i)}$  đều góp phần vào cập nhật tham số  $\mu_j, \Sigma_j$ . Mức độ đóng góp của các điểm dữ liệu được đánh trọng số bởi giá trị  $w_j^{(i)}$ . Ta có thể hiểu, nếu một điểm dữ liệu có xác suất cao thuộc vào phân phối chuẩn  $j$  thì điểm dữ liệu đó sẽ đóng góp nhiều trong việc xây dựng lại tham số  $\mu_j, \Sigma_j$ ; và ngược lại.

Bên cạnh đó, vì hàm Log-Likelihood không phải là một hàm lồi nên có thể thuật toán **EM** sẽ hội tụ tại điểm cực trị cục bộ. Do đó, người ta thường cho chạy thuật toán **EM** nhiều lần với các giá trị tham số  $\phi_j, \mu_j, \Sigma_j$  khởi tạo ban đầu khác nhau.

Một cách hiểu khác của thuật toán **EM** trong **GMM**, ở bước E, thuật toán cố gắng đoán các giá trị biến ẩn  $z^{(i)}$ . Dựa vào những dự đoán đó, ở bước M thuật toán xây dựng lại các tham số  $\phi_j, \mu_j, \Sigma_j$  ban đầu.

## 4 Thuật toán Expectation-maximization (EM)

Ở phần trước, ta đã phát biểu kết quả thu được khi sử dụng thuật toán **EM** vào **GMM**. Ở phần này, ta sẽ tìm hiểu về thuật toán **EM** một cách tổng quát, thuật toán ước lượng tham số của mô hình thống kê khi xuất hiện biến ẩn. Sau đó, ta sẽ xây dựng lại bước E, tính các giá trị  $w_j^{(i)}$ ; bước M, cập nhật các tham số  $\phi_j, \mu_j, \Sigma_j$  của **GMM**. Trước tiên, ta tìm hiểu về kết quả được dùng trong thuật toán gọi là **bất đẳng thức Jensen**.

### 4.1 Bất đẳng thức Jensen

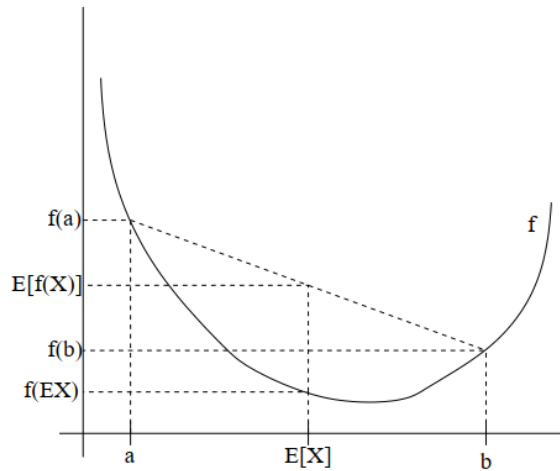
Xét hàm số  $f : \mathbb{R} \rightarrow \mathbb{R}$ , hàm  $f$  được gọi là hàm lồi khi  $f''(x) \geq 0, \forall x \in \mathbb{R}$ ; hàm  $f$  được gọi là hàm lồi **chặt** khi  $f''(x) > 0, \forall x \in \mathbb{R}$ . Bất đẳng thức Jensen được phát biểu như sau:

**Theorem.** Cho  $f$  là hàm lồi,  $X$  là biến ngẫu nhiên, khi đó:

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

Nếu  $f$  là hàm lồi **chặt** thì đẳng thức  $\mathbb{E}[f(X)] = f(\mathbb{E}[X])$  xảy ra khi và chỉ khi  $X = \mathbb{E}[X]$  với xác suất 1 (Nói cách khác, nếu  $X$  là một hằng số).

Ta có thể hiểu rõ hơn về bất đẳng thức qua hình vẽ dưới đây:



Trong hình vẽ, hàm lồi  $f$  được thể hiện bằng đường nét liền, biến ngẫu nhiên  $X$  có xác suất 0.5 nhận giá trị  $a$  và xác suất 0.5 nhận giá trị  $b$ . Do đó, kì vọng của biến ngẫu nhiên  $X$  là trung điểm của  $a$  và  $b$  trên trục  $x$ .

Khi đó,  $f(\mathbb{E}[X])$  là trung điểm của  $f(a), f(b)$  trên trục  $y$ . Từ hình vẽ, ta thấy do  $f$  là hàm lồi nên ta có được kết quả  $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$ .

**Nhận xét.** Hàm  $f$  là hàm lõm (**chặt**) khi và chỉ khi  $-f$  là hàm lồi (**chặt**). Khi đó, bất đẳng thức Jensen đúng cho hàm lõm  $f$  với chiều của bất đẳng thức đổi lại. ( $\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$ )

### 4.2 Mô hình toán học

Ta có bài toán ước lượng tham số của mô hình thống kê trên tập dữ liệu  $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ , với  $n$  điểm dữ liệu được lấy mẫu một cách ngẫu nhiên. Giả sử tham số của mô hình thống kê cần tìm là  $\theta$ . Ta có giá trị biến ẩn đi kèm với  $x^{(i)}$  là  $z^{(i)}$ . Khi đó hàm xác suất của  $x$  được tính bằng cách lấy tổng của hàm xác suất  $x$  và  $z$  ở mọi giá trị  $z$ :

$$p(x; \theta) = \sum_z p(x, z; \theta) \quad (1)$$

Ta cần tìm tham số  $\theta$  tối đa hàm Log-Likelihood của dữ liệu. Hàm Log-Likelihood của bài toán:

$$l(\theta) = \sum_{i=1}^n \log p(x^{(i)}; \theta) \quad (2)$$

$$= \sum_{i=1}^n \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) \quad (3)$$

Ta có thể trực tiếp lấy đạo hàm của hàm Log-Likelihood theo  $\theta$  nhưng ta không thể giải cho đạo hàm đó bằng không, như đã được trình bày trong phần 3.2 trang 6.

Vì lẽ đó, thuật toán **EM** tiếp cận việc tối đa hàm Log-Likelihood theo một cách khác. Ta sẽ tối đa hàm Log-Likelihood bằng cách xây dựng một cận dưới của hàm, ở bước E; và tối đa cận dưới vừa được xây dựng, ở bước M. Bằng cách lặp lại bước E và bước M ta sẽ tối đa được hàm Log-Likelihood.

### 4.3 Xây dựng hàm evidence lower bound (ELBO)

Để thuận tiện trong việc trình bày ở các bước tiếp theo, ta sẽ bỏ qua việc lấy tổng hàm xác suất của mọi điểm dữ liệu  $\sum_{i=1}^n$ . Ta tối đa hàm Log-Likelihood của **một điểm dữ liệu**  $x$ . Sau khi đã xây dựng thuật toán trên một điểm dữ liệu, ta có thể lấy tổng của  $n$  điểm dữ liệu và có một thuật toán hoàn chỉnh. Tóm lại, trước nhất, ta tìm cách tối đa:

$$\log p(x; \theta) = \log \sum_z p(x, z; \theta) \quad (4)$$

Gọi  $Z$  là biến ngẫu nhiên của giá trị biến ẩn  $z$ , ta giả sử  $Q$  là một phân phối **bất kỳ** của biến ngẫu nhiên  $Z$ . Khi đó  $Q$  phải thỏa mãn:  $\sum_z Q(z) = 1$  và  $Q(z) \geq 0$ .

Ta biến đổi hàm Log-Likelihood:

$$\begin{aligned} \log p(x; \theta) &= \log \sum_z p(x, z; \theta) \\ &= \log \sum_z Q(z) \frac{p(x, z; \theta)}{Q(z)} \\ &= \log \mathbb{E}_{z \sim Q} \left[ \frac{p(x, z; \theta)}{Q(z)} \right] \end{aligned} \quad (5)$$

Ta có nhận xét, hàm  $f(x) = \log(x)$  là hàm lồi **chặt** trên  $\mathbb{D} = (0; +\infty)$ , do đó:

$$\log p(x; \theta) \geq \mathbb{E}_{z \sim Q} \left[ \log \frac{p(x, z; \theta)}{Q(z)} \right] \quad (6)$$

$$= \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)} \quad (7)$$

Để thuận tiện, ta sẽ gọi biểu thức (7) là hàm **evidence lower bound** (ELBO), được ký hiệu:

$$\text{ELBO}(x; Q, \theta) \triangleq \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)} \quad (8)$$

Với một giá trị  $x$  và phân phối  $Q$  cố định thì hàm ELBO là hàm số theo biến số  $\theta$ . Ta thấy, **với mọi** phân phối  $Q$ , hàm ELBO cho ta cận dưới của  $\log(x; \theta)$ . Để xấp xỉ tốt nhất hàm Log-Likelihood ta nên chọn phân phối  $Q$  sao cho, khi ta cố định một giá trị  $\theta_0$  thì hàm  $\text{ELBO}(x; Q, \theta_0)$  sẽ đúng bằng  $\log(x; \theta_0)$ .

Vì  $\log(x)$  là hàm lõm **chặt** nên ta có thể dùng điều kiện xảy ra dấu bằng của bất đẳng thức Jensen. Dấu bằng xảy ra khi và chỉ khi:

$$\begin{aligned}\frac{p(x, z; \theta)}{Q(z)} &= c \\ p(x, z; \theta) &= c Q(z) \\ \sum_z p(x, z; \theta) &= c \sum_z Q(z) \\ \sum_z p(x, z; \theta) &= c\end{aligned}$$

Do vậy, để đẳng thức xảy ra thì:

$$\begin{aligned}Q(z) &= \frac{p(x, z; \theta)}{\sum_z p(x, z; \theta)} \\ &= \frac{p(x, z; \theta)}{p(x; \theta)} \\ &= p(z|x; \theta)\end{aligned}\tag{9}$$

Tóm lại, ta sẽ chọn phân phối  $Q$  bằng phân phối hậu nghiệm của  $z$  khi biết được  $x$  tại một giá trị  $\theta_0$  cố định. Hay  $Q(z) = p(z|x; \theta_0)$

Ta thử kiểm tra lại bằng cách thế  $Q(z) = p(z|x; \theta)$  vào hàm ELBO:

$$\begin{aligned}\text{ELBO}(x; Q, \theta) &= \sum_z p(z|x; \theta) \log \frac{p(x, z; \theta)}{p(z|x; \theta)} \\ &= \sum_z p(z|x; \theta) \log \frac{p(z|x; \theta)p(x; \theta)}{p(z|x; \theta)} \\ &= \sum_z p(z|x; \theta) \log p(x; \theta) \\ &= \log p(x; \theta) \sum_z p(z|x; \theta) \\ &= \log p(x; \theta)\end{aligned}$$

Như vậy ta đã xây dựng được hàm ELBO theo như ý muốn:

$$\log p(x; \theta) \geq \text{ELBO}(x; Q, \theta) \quad \forall Q, \theta, x\tag{10}$$

Một cách trực quan, thuật toán **EM** lần lượt cập nhật  $Q$  và  $\theta$  như sau:

1. Cố định  $\theta$ , thay đổi hàm ELBO sao cho  $\text{ELBO}(x; Q, \theta) = \log p(x; \theta)$  tại giá trị  $\theta$  cố định bằng cách  $Q(z) := p(z|x; \theta)$ .
2. Cố định  $Q$ , tối đa hàm ELBO theo biến  $\theta$ .

## 4.4 Hoàn thành thuật toán EM

Ở trên, ta đã tìm được hàm cận dưới phù hợp cho hàm Log-Likelihood của một điểm dữ liệu  $x$ . Tiếp đến, ta sẽ xây dựng cận dưới cho hàm Log-Likelihood của một tập dữ liệu.

Cho tập dữ liệu  $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ . Lưu ý rằng, phân phối  $Q$  ở trên sẽ đặc trưng cho một điểm dữ liệu. Do đó, khi ta có  $n$  điểm dữ liệu thì ta sẽ có  $n$  phân phối  $Q$  tương ứng  $Q_1, Q_2, \dots, Q_n$ . Mỗi điểm dữ liệu  $x^{(i)}$  sẽ có một hàm ELBO là:

$$\log p(x^{(i)}; \theta) \geq \text{ELBO}(x^{(i)}; Q_i, \theta) = \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \quad (11)$$

Lấy tổng của hàm ELBO tại mọi điểm dữ liệu ta sẽ có cận dưới của hàm Log-Likelihood:

$$\begin{aligned} l(\theta) &\geq \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i, \theta) \\ &= \sum_{i=1}^n \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \end{aligned} \quad (12)$$

Với mọi tập phân phối  $Q_1, Q_2, \dots, Q_n$ , biểu thức (12) cho ta cận dưới của hàm Log-Likelihood. Với cùng ý tưởng khi ta xây dựng phân phối  $Q$  cho một điểm dữ liệu thì ta sẽ chọn phân phối  $Q_i$  sao cho:

$$Q_i(z^{(i)}) = p(z^{(i)} | x^{(i)}; \theta)$$

Tóm lại, ta sẽ chọn những phân phối  $Q_i$  bằng phân phối hậu nghiệm của  $z^{(i)}$  khi biết được  $x^{(i)}$  tại một giá trị  $\theta_0$  cố định. Hay  $Q_i(z^{(i)}) = p(z^{(i)} | x^{(i)}; \theta_0)$

Bây giờ, ta có thể trình bày thuật toán **EM** một cách tổng quát:

Khởi tạo giá trị  $\theta^{(0)}$  ngẫu nhiên

Thực hiện cho đến khi thuật toán hội tụ: {

(Bước E) Với mỗi giá trị  $i$ :

$$Q_i^{(t)}(z^{(i)}) := p(z^{(i)} | x^{(i)}; \theta^{(t)})$$

(Bước M) Cập nhật  $\theta$ :

$$\begin{aligned} \theta^{(t+1)} &:= \arg \max_{\theta} \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i^{(t)}, \theta) \\ &= \arg \max_{\theta} \sum_{i=1}^n \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i^{(t)}(z^{(i)})} \end{aligned} \quad (13)$$

}

Điều kiện hội tụ:  $|l(\theta^{(t+1)}) - l(\theta^{(t)})| \leq \epsilon$

Kết quả thu được sau mỗi vòng lặp EM:

- Bước E:  $n$  phân phối  $Q_i$ .
- Bước M: tham số  $\theta$ .



Để chắc chắn thuật toán **EM** hoạt động tốt ta sẽ chứng minh rằng tham số  $\theta$  có được sau mỗi vòng lặp **EM** sẽ **đơn điệu** cải thiện hàm Log-Likelihood. Nói cách khác, tham số  $\theta$  có được sau mỗi vòng lặp **EM**, **luôn luôn** cho được giá trị hàm Log-Likelihood lớn hơn những tham số  $\theta$  trước đó. Chứng minh trên sẽ tương đương với việc chứng minh:

$$l(\theta^{(t+1)}) \geq l(\theta^{(t)}) \quad \forall t \quad (14)$$

Nhắc lại, về cách chọn các phân phối  $Q_i^{(t)}(z^{(i)}) := p(z^{(i)}|x^{(i)}; \theta^{(t)})$  đã giúp ta thỏa mãn điều kiện dấu bằng xảy ra của bất đẳng thức Jensen, hay tổng hàm ELBO của mọi điểm dữ liệu sẽ bằng giá trị của hàm Log-Likelihood tại  $\theta^{(t)}$

$$l(\theta^{(t)}) = \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i^{(t)}, \theta^{(t)}) \quad (15)$$

Vậy ta có thể chứng minh bất đẳng thức (14) như sau:

$$l(\theta^{(t+1)}) \geq \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i^{(t)}, \theta^{(t+1)}) \quad (16)$$

$$\geq \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i^{(t)}, \theta^{(t)}) \quad (17)$$

$$= l(\theta^{(t)}) \quad (18)$$

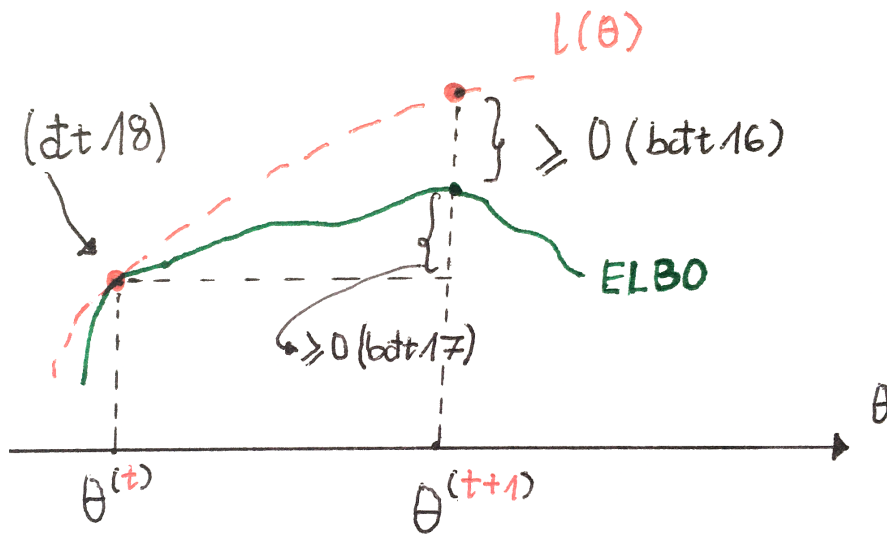
Bất đẳng thức (16) đúng vì đó là cách chúng ta xây dựng hàm ELBO như đã trình bày ở bất đẳng thức (10).

Bất đẳng thức (17) đúng vì ở bước M ta chọn giá trị  $\theta^{(t+1)}$  sao cho tối đa hàm ELBO.

$$\theta^{(t+1)} := \arg \max_{\theta} \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i^{(t)}, \theta)$$

Đẳng thức (18) đúng theo đẳng thức (15).

Ta có thể hiểu rõ hơn chứng minh trên qua hình vẽ dưới đây:

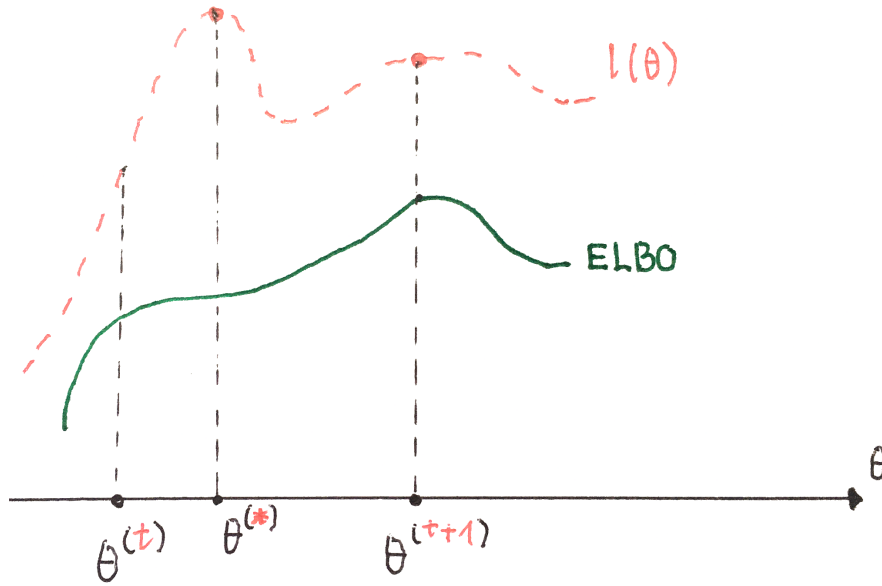


Hình 3: Hàm Log-Likelihood: đường nét đứt màu đỏ. Hàm ELBO: đường nét liền màu xanh.

Như vậy ta đã chứng minh được bất đẳng thức (14).

Sau mỗi vòng lặp **EM** ta chắc chắn rằng tham số  $\theta$  có được sẽ cải thiện hàm Log-Likelihood. Do đó, điều kiện dừng của thuật toán **EM** lúc bấy giờ sẽ là, nếu sau mỗi vòng lặp **EM** giá trị  $\theta$  mới có được không cải thiện hàm Log-Likelihood được nhiều thì ta sẽ dừng thuật toán **EM**. Vậy điều kiện dừng sẽ là:  $|l(\theta^{(t+1)}) - l(\theta^{(t)})| \leq \epsilon$ .

Một quan sát nhỏ, điểm mấu chốt của thuật toán **EM** là ở cách chúng ta xây dựng hàm ELBO sao cho  $l(\theta^{(t)}) = \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i^{(t)}, \theta^{(t)})$ . Nếu dấu bằng của bất đẳng thức Jensen không xảy ra thì ta có thể bỏ qua giá trị  $\theta$  tối ưu cần phải tìm. Ta có thể hiểu rõ hơn tại sao dấu bằng lại quan trọng qua hình vẽ sau:



Hình 4: Hàm Log-Likelihood: đường nét đứt màu đỏ. Hàm ELBO: đường nét liền màu xanh.

Bằng cách tối đa hàm ELBO, ta đã tìm được giá trị tham số  $\theta^{(t+1)}$  là giá trị lớn nhất của hàm ELBO. Nhưng vì hàm ELBO không bằng hàm Log-Likelihood tại  $\theta^{(t)}$  nên ta đã bỏ qua giá trị  $\theta^{(*)}$  tối ưu.

## 4.5 Xây dựng EM trong GMM

Sau khi đã hiểu về thuật toán **EM** tổng quát, ta sẽ xây dựng lại bước E và bước M của **GMM**.

Nhắc lại, vì biến ẩn  $z^{(i)}$  chỉ có thể nhận các giá trị trong tập hợp  $\{1, 2, \dots, k\}$ , với  $k$  là số phân phối chuẩn có trong mô hình, nên với mỗi điểm dữ liệu  $x^{(i)}$  ta chỉ cần quan tâm đến giá trị của phân phối  $Q_i$  tại những điểm  $z^{(i)} \in \{1, 2, \dots, k\}$ . Sử dụng lại kí hiệu của phần 3.1 trang 6 và phần 4.4 trang 11, với mỗi giá trị  $i, j$ ; ta có bước E:

$$w_j^{(i)} = Q_i(z^{(i)} = j) = p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

Khi đó ta có tổng của hàm ELBO tại mọi điểm dữ liệu là:

$$\begin{aligned} \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i; \theta) &= \sum_{i=1}^n \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)}{Q_i(z^{(i)})} \\ &= \sum_{i=1}^n \sum_{j=1}^k Q_i(z^{(i)} = j) \log \frac{p(x^{(i)} | z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{Q_i(z^{(i)} = j)} \\ &= \sum_{i=1}^n \sum_{j=1}^k w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right) \cdot \phi_j}{w_j^{(i)}} \end{aligned}$$

Để xây dựng công thức cập nhật cho  $\mu, \Sigma$ , ta sẽ sử dụng lại những tính chất về đạo hàm của vector và ma trận được trình bày ở phần 1.2 trang 3.

Để tìm ra công thức cập nhật cho  $\mu_l$ , ta sẽ lấy đạo hàm riêng của tổng hàm ELBO theo  $\mu_l$ :

$$\begin{aligned} \frac{\partial}{\partial \mu_l} \left[ \sum_{i=1}^n \sum_{j=1}^k w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right) \cdot \phi_j}{w_j^{(i)}} \right] \\ = -\frac{\partial}{\partial \mu_l} \left[ \sum_{i=1}^n \sum_{j=1}^k w_j^{(i)} \frac{1}{2} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right] \\ = \frac{1}{2} \sum_{i=1}^n w_l^{(i)} \frac{\partial}{\partial \mu_l} [2\mu_l^T \Sigma_l^{-1} x^{(i)} - \mu_l^T \Sigma_l^{-1} \mu_l] \\ = \sum_{i=1}^n w_l^{(i)} (\Sigma_l^{-1} x^{(i)} - \Sigma_l^{-1} \mu_l) \end{aligned}$$

Giải cho đạo hàm bằng không, ta được:

$$\begin{aligned} \sum_{i=1}^n w_l^{(i)} (\Sigma_l^{-1} x^{(i)} - \Sigma_l^{-1} \mu_l) &= 0 \\ \mu_l \sum_{i=1}^n w_l^{(i)} &= \sum_{i=1}^n w_l^{(i)} x^{(i)} \\ \mu_l &= \frac{\sum_{i=1}^n w_l^{(i)} x^{(i)}}{\sum_{i=1}^n w_l^{(i)}} \end{aligned}$$

Vậy ta đã tìm được công thức cập nhật cho  $\mu_l$

Tiếp đến ta sẽ xây dựng công thức cập nhật cho  $\Sigma_l$ , ta sẽ lấy đạo hàm riêng của tổng ELBO theo  $\Sigma_l^{-1}$ :

$$\begin{aligned}
& \frac{\partial}{\partial \Sigma_l^{-1}} \left[ \sum_{i=1}^n \sum_{j=1}^k w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \exp \left( -\frac{1}{2} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right) \cdot \phi_j}{w_j^{(i)}} \right] \\
&= \frac{\partial}{\partial \Sigma_l^{-1}} \left[ \sum_{i=1}^n w_l^{(i)} \log \frac{\frac{1}{(2\pi)^{d/2} |\Sigma_l|^{1/2}} \exp \left( -\frac{1}{2} (x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l) \right) \cdot \phi_l}{w_l^{(i)}} \right] \\
&= \frac{\partial}{\partial \Sigma_l^{-1}} \left[ \sum_{i=1}^n w_l^{(i)} \log \left[ \frac{1}{|\Sigma_l|^{1/2}} \exp \left( -\frac{1}{2} (x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l) \right) \right] \right] \\
&= \frac{\partial}{\partial \Sigma_l^{-1}} \left[ \sum_{i=1}^n \frac{1}{2} w_l^{(i)} \log |\Sigma_l|^{-1} - \frac{1}{2} w_l^{(i)} ((x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l)) \right] \\
&= \frac{1}{2} \frac{\partial}{\partial \Sigma_l^{-1}} \left[ \sum_{i=1}^n w_l^{(i)} \log |\Sigma_l^{-1}| - w_l^{(i)} ((x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l)) \right] \\
&= \frac{1}{2} \left[ \sum_{i=1}^n w_l^{(i)} (\Sigma_l^{-1})^{-1} - w_l^{(i)} (x^{(i)} - \mu_l)(x^{(i)} - \mu_l)^T \right] \\
&= \frac{1}{2} \left[ \sum_{i=1}^n w_l^{(i)} \Sigma_l - w_l^{(i)} (x^{(i)} - \mu_l)(x^{(i)} - \mu_l)^T \right]
\end{aligned}$$

Giải đạo hàm bằng không, ta được

$$\begin{aligned}
\sum_{i=1}^n w_l^{(i)} \Sigma_l - w_l^{(i)} (x^{(i)} - \mu_l)(x^{(i)} - \mu_l)^T &= 0 \\
\Sigma_l \sum_{i=1}^n w_l^{(i)} &= \sum_{i=1}^n w_l^{(i)} (x^{(i)} - \mu_l)(x^{(i)} - \mu_l)^T \\
\Sigma_l &= \frac{\sum_{i=1}^n w_l^{(i)} (x^{(i)} - \mu_l)(x^{(i)} - \mu_l)^T}{\sum_{i=1}^n w_l^{(i)}}
\end{aligned}$$

Vậy ta đã tìm được công thức cập nhật cho  $\Sigma_l$ .

Tiếp đến ta sẽ xây dựng công thức cập nhật cho  $\phi_l$ . Ta có nhận xét, trong tổng hàm ELBO, có nhiều giá trị không phụ thuộc vào biến  $\phi_l$ . Sau khi biến đổi tổng hàm ELBO, ta nhận thấy tối đa tổng hàm ELBO theo biến  $\phi_l$  sẽ tương đương với tối đa hàm số sau:

$$\sum_{i=1}^n \sum_{j=1}^k w_j^{(i)} \log \phi_j$$

Một lưu ý nhỏ, ta tối đa hàm số trên dưới điều kiện  $\sum_{j=1}^k \phi_j = 1$  và  $\phi_j \geq 0$ , nên ta không thể dùng phương pháp lấy đạo hàm riêng thông thường. Thay vào đó, ta sẽ dùng phương pháp nhân tử Lagrange. Ta xây dựng hàm Lagrangian:

$$\mathcal{L}(\phi) = \sum_{i=1}^n \sum_{j=1}^k w_j^{(i)} \log \phi_j + \beta \left( \sum_{j=1}^k \phi_j - 1 \right)$$

Ở đây, ta không thêm vào điều kiện  $\phi_j \geq 0$  vì kết quả  $\phi_j$  thu được sẽ tự thỏa mãn điều kiện trên. Lấy đạo hàm riêng của hàm Lagrangian theo biến  $\phi_l$  ta được:

$$\frac{\partial}{\partial \phi_l} \mathcal{L}(\phi) = \sum_{i=1}^n \frac{w_l^{(i)}}{\phi_l} + \beta$$

Giải cho hàm số trên bằng không, ta được:

$$\phi_l = \frac{\sum_{i=1}^n w_l^{(i)}}{-\beta}$$

Lấy tổng trên mọi giá trị của  $l$  ta sẽ được:

$$\begin{aligned} -\beta \sum_{l=1}^k \phi_l &= \sum_{i=1}^n \sum_{l=1}^k w_l^{(i)} \\ -\beta &= \sum_{i=1}^n 1 \\ -\beta &= n \end{aligned}$$

Vậy công thức cập nhật của  $\phi_l$  là:

$$\phi_l = \frac{1}{n} \sum_{i=1}^n w_l^{(i)}$$

## 4.6 Kỳ vọng và tối đa trong EM

Mục tiêu của phần trình bày tiếp theo là cho thấy được tại sao bước E được gọi là bước kỳ vọng và bước M được gọi là bước tối đa trong **EM**.

Ý tưởng về xây dựng hàm ELBO và những phân phối  $Q_i$  là ý tưởng được phát triển từ nguyên tác (Dempster et al., 1977) [2]. Do vậy phần tiếp theo ta sẽ trình bày thuật toán **EM** giống hơn với nguyên tác.

Nhắc lại kết quả có được trong phần trước:

$$l(\theta) \geq \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i, \theta) = \sum_{i=1}^n \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

Trong bước E, ta xây dựng các phân phối  $Q_i$  sao cho:

$$Q_i^{(t)}(z^{(i)}) := p(z^{(i)} | x^{(i)}; \theta^{(t)})$$

Nghĩa là chúng ta đang xây dựng hàm mục tiêu mới được kí hiệu như sau:

$$\mathcal{L}(\theta; \theta^{(t)}) \triangleq \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i^{(t)}, \theta) := \sum_{i=1}^n \sum_{z^{(i)}} p(z^{(i)} | x^{(i)}; \theta^{(t)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(z^{(i)} | x^{(i)}; \theta^{(t)})}$$

Kí hiệu  $\mathcal{L}(\theta; \theta^{(t)})$  sẽ được hiểu là:  $\mathcal{L}$  là một hàm số theo biến  $\theta$  và được tham số bởi giá trị  $\theta^{(t)}$  ở vòng lặp **EM** trước.

Khi đó, trong bước M, việc tối đa  $\mathcal{L}(\theta; \theta^{(t)})$  theo biến  $\theta$  sẽ được trình bày như sau:

$$\begin{aligned} \theta^{(t+1)} &:= \arg \max_{\theta} \mathcal{L}(\theta; \theta^{(t)}) \\ &= \arg \max_{\theta} \left[ \sum_{i=1}^n \sum_{z^{(i)}} p(z^{(i)} | x^{(i)}; \theta^{(t)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(z^{(i)} | x^{(i)}; \theta^{(t)})} \right] \\ &= \arg \max_{\theta} \left[ \sum_{i=1}^n \sum_{z^{(i)}} p(z^{(i)} | x^{(i)}; \theta^{(t)}) \log p(x^{(i)}, z^{(i)}; \theta) \right] \\ &= \arg \max_{\theta} \left[ \sum_{i=1}^n \mathbb{E}_{z^{(i)} \sim p(z^{(i)} | x^{(i)}; \theta^{(t)})} [\log p(x^{(i)}, z^{(i)}; \theta)] \right] \end{aligned}$$

Hàm số bên trong hàm  $\arg \max$  được gọi là hàm **kỳ vọng của Log-Likelihood đầy đủ (Expectation of Complete Log-Likelihood)** của dữ liệu.

Qua cách biến đổi ở bước M như trên, ta thấy trong hàm  $\mathcal{L}(\theta; \theta^{(t)})$  chỉ có phần **kỳ vọng của Log-Likelihood đầy đủ** là phụ thuộc vào biến  $\theta$ . Do đó ở bước E, ta chỉ cần xây dựng hàm  $\mathcal{L}(\theta; \theta^{(t)})$  là hàm **kỳ vọng của Log-Likelihood đầy đủ** của dữ liệu với giá trị  $\theta^{(t)}$  tìm được ở vòng lặp **EM** trước.

Vậy ta có một cách phát biểu khác cho thuật toán **EM**:

Khởi tạo giá trị  $\theta^{(0)}$  ngẫu nhiên

Thực hiện cho đến khi thuật toán hội tụ: {

(Bước E) Xây dựng hàm kì vọng của Log-Likelihood đầy đủ:

$$\mathcal{L}(\theta; \theta^{(t)}) := \sum_{i=1}^n \mathbb{E}_{z^{(i)} \sim p(z^{(i)} | x^{(i)}; \theta^{(t)})} [\log p(x^{(i)}, z^{(i)}; \theta)]$$

(Bước M) Tối đa hàm số vừa xây dựng bằng cách cập nhật  $\theta$ :

$$\theta^{(t+1)} := \arg \max_{\theta} \mathcal{L}(\theta; \theta^{(t)})$$

}

Điều kiện hội tụ:  $|l(\theta^{(t+1)}) - l(\theta^{(t)})| \leq \epsilon$

Kết quả thu được sau mỗi vòng lặp EM:

- Bước E: hàm số  $\mathcal{L}$ .
- Bước M: tham số  $\theta$ .

Qua cách phát biểu thuật toán **EM** như trên, ta thấy được ở bước E, ta đang xây dựng cận dưới của hàm Log-Likelihood, cận dưới đó là **kỳ vọng của Log-Likelihood đầy đủ** của dữ liệu. Ở bước M, ta sẽ tối đa cận dưới vừa tìm được.

## 5 Thuật toán EM bán giám sát (Semi-supervised EM)

Thuật toán **EM** là thuật toán ước lượng tham số của mô hình thống kê khi xuất hiện biến ẩn. Thuật toán **EM** là thuật toán học không giám sát vì ta không có nhãn cho các điểm dữ liệu. Ở phần này, ta sẽ mở rộng thuật toán **EM** thành một thuật toán học bán giám sát, nghĩa là chúng ta sẽ có những điểm dữ liệu không có nhãn cùng với một vài điểm dữ liệu có nhãn.

Nhắc lại về mô hình toán học, ta có tập dữ liệu  $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ , ứng với mỗi điểm dữ liệu  $x^{(i)}$  ta có một giá trị biến ẩn  $z^{(i)}$ , khi đó tập các biến ẩn là  $\{z^{(1)}, z^{(2)}, \dots, z^{(n)}\}$ . Mô hình thống kê được tham số bởi  $\theta$ . Hàm Log-Likelihood không giám sát là:

$$\begin{aligned} l_{\text{unsup}}(\theta) &= \sum_{i=1}^n \log p(x^{(i)}; \theta) \\ &= \sum_{i=1}^n \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) \\ &\geq \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i, \theta) \end{aligned}$$

Khi này, ta có thêm  $\tilde{n}$  điểm dữ liệu có nhãn  $\{(\tilde{x}^{(1)}, \tilde{z}^{(1)}), (\tilde{x}^{(2)}, \tilde{z}^{(2)}), \dots, (\tilde{x}^{(\tilde{n})}, \tilde{z}^{(\tilde{n})})\}$ . Ta xây dựng hàm Log-Likelihood giám sát cho  $\tilde{n}$  điểm dữ liệu có nhãn:

$$\begin{aligned} l_{\text{sup}}(\theta) &= \sum_{i=1}^{\tilde{n}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta) \\ &\geq \sum_{i=1}^{\tilde{n}} \text{ELBO}(\tilde{x}^{(i)}; \tilde{Q}_i, \theta) \end{aligned}$$

Hàm Log-Likelihood trong trường hợp thuật toán bán giám sát sẽ được định nghĩa là tổng của hàm Log-Likelihood không giám sát và hàm Log-Likelihood giám sát được đánh trọng số bởi siêu tham số  $\alpha$ :

$$l_{\text{semi-sup}}(\theta) \triangleq l_{\text{unsup}}(\theta) + \alpha l_{\text{sup}}(\theta)$$

Siêu tham số  $\alpha$  sẽ được hiểu là mức độ quan trọng của dữ liệu gán nhãn. Nếu  $\alpha = 0$  thì dữ liệu gán nhãn không quan trọng và thuật toán trở thành thuật toán không giám sát thuần túy; và ngược lại.

Ta có cận dưới của hàm Log-Likelihood bán giám sát:

$$l_{\text{semi-sup}}(\theta) \geq \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i, \theta) + \alpha \sum_{i=1}^{\tilde{n}} \text{ELBO}(\tilde{x}^{(i)}; \tilde{Q}_i, \theta)$$



Đối với dữ liệu gán nhãn thì ở bước E, ta xây dựng các phân phối  $\tilde{Q}^{(i)}$  sao cho:

$$\tilde{Q}_i(\tilde{z}^{(i)}) := p(\tilde{z}^{(i)}|\tilde{x}^{(i)}; \theta)$$

Nhưng vì  $\tilde{z}^{(i)}$  là nhãn của  $\tilde{x}^{(i)}$  nên  $p(\tilde{z}^{(i)}|\tilde{x}^{(i)}; \theta) = 1 \forall i, \theta$ . Hay  $\tilde{Q}_i(\tilde{z}^{(i)}) = 1 \forall i$ . Vậy

$$l_{\text{semi-sup}}(\theta) \geq \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i, \theta) + \alpha \sum_{i=1}^{\tilde{n}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta)$$

Bây giờ, ta có thể trình bày thuật toán **EM bán giám sát**:

Khởi tạo giá trị  $\theta^{(0)}$  ngẫu nhiên

Thực hiện cho đến khi thuật toán hội tụ: {

(Bước E) Với mỗi giá trị  $i \in \{1, 2, \dots, n\}$ :

$$Q_i^{(t)}(z^{(i)}) := p(z^{(i)}|x^{(i)}; \theta^{(t)})$$

(Bước M) Cập nhật  $\theta$ :

$$\begin{aligned} \theta^{(t+1)} &:= \arg \max_{\theta} \left[ \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i^{(t)}, \theta) + \alpha \left( \sum_{i=1}^{\tilde{n}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta) \right) \right] \\ &= \arg \max_{\theta} \left[ \sum_{i=1}^n \left( \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i^{(t)}(z^{(i)})} \right) + \alpha \left( \sum_{i=1}^{\tilde{n}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta) \right) \right] \end{aligned}$$

$$\text{Điều kiện hội tụ: } |l_{\text{semi-sup}}(\theta^{(t+1)}) - l_{\text{semi-sup}}(\theta^{(t)})| \leq \epsilon$$

Kết quả thu được sau mỗi vòng lặp EM:

- Bước E:  $n$  phân phối  $Q_i$ .
- Bước M: tham số  $\theta$ .

Để chắc chắn thuật toán **EM bán giám sát** hoạt động tốt ta sẽ chứng minh rằng tham số  $\theta$  có được sau mỗi vòng lặp **EM** sẽ **đơn điệu** cải thiện hàm Log-Likelihood bán giám sát. Nói cách khác, tham số  $\theta$  có được sau mỗi vòng lặp **EM**, **luôn luôn** cho được giá trị hàm Log-Likelihood bán giám sát lớn hơn những tham số  $\theta$  trước đó. Bằng phép biến đổi tương tự như phần trước, ta cũng sẽ chứng minh được:

$$l_{\text{semi-sup}}(\theta^{(t+1)}) \geq l_{\text{semi-sup}}(\theta^{(t)}) \quad \forall t$$

## 5.1 GMM bán giám sát

Ta có tập dữ liệu không gắn nhãn được phát biểu như phần 3.1 trang 6.

Ngoài ra, ta có tập dữ liệu gắn nhãn được trình bày như sau. Tập dữ liệu  $\{\tilde{x}^{(1)}, \tilde{x}^{(2)}, \dots, \tilde{x}^{(\tilde{n})}\}$ , với mỗi  $\tilde{x}^{(i)} \in \mathbb{R}^d$  là một giá trị xác định của vector ngẫu nhiên  $\tilde{X}^{(i)}$ . Các Vector ngẫu nhiên  $\tilde{X}^{(i)}$  độc lập thống kê với nhau. Đồng thời ta có  $\tilde{n}$  biến ngẫu nhiên  $\tilde{Z}^{(1)}, \tilde{Z}^{(2)}, \dots, \tilde{Z}^{(\tilde{n})}$  với  $\tilde{Z}^{(i)} : \Omega \rightarrow \{1, 2, \dots, k\}$ , mỗi giá trị xác định  $\tilde{z}^{(i)}$  của biến ngẫu nhiên  $\tilde{Z}^{(i)}$  là nhãn của điểm dữ liệu  $\tilde{x}^{(i)}$  thể hiện cho cụm mà điểm dữ liệu  $\tilde{x}^{(i)}$  thuộc vào.

Ở đây, giá trị  $\tilde{z}^{(i)}$  là giá trị quan sát được nên ta không cần giả sử về phân phối của  $\tilde{Z}^{(i)}$ . Thay vào đó, ta sẽ có giả sử về phân phối  $\tilde{X}^{(i)} | \tilde{Z}^{(i)}$  như sau:

$$\tilde{X}^{(i)} | \tilde{Z}^{(i)} \sim \mathcal{N}(\mu_{\tilde{z}^{(i)}}, \Sigma_{\tilde{z}^{(i)}})$$

Tóm lại, ta có  $n + \tilde{n}$  điểm dữ liệu, trong đó  $n$  điểm dữ liệu  $x^{(i)}$  không gắn nhãn với giá trị biến ẩn tương ứng là  $z^{(i)}$ ;  $\tilde{n}$  điểm dữ liệu  $\tilde{x}^{(i)}$  được gắn nhãn bởi  $\tilde{z}^{(i)}$ . Mục tiêu của **GMM bán giám sát** là tìm lại các tham số:

- $\phi_1, \phi_2, \dots, \phi_k$  với  $\phi_i \in \mathbb{R}$
- $\mu_1, \mu_2, \dots, \mu_k$  với  $\mu_i \in \mathbb{R}^d$
- $\Sigma_1, \Sigma_2, \dots, \Sigma_k$  với  $\Sigma_i \in \mathbb{R}^{d \times d}$

Bằng cách xây dựng mô hình như phần 4.5 trang 15. Ta có được thuật toán **EM bán giám sát** trong **GMM** như sau:

Khởi tạo các giá trị  $\phi_j, \mu_j, \Sigma_j$  ngẫu nhiên

Thực hiện cho đến khi thuật toán hội tụ: {

(Bước E) Với mỗi giá trị  $i \in \{1, 2, \dots, n\}, j$ :

$$w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

(Bước M) Cập nhật các tham số:

$$\begin{aligned} \phi_j &:= \frac{\sum_{i=1}^n w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = j\}}{n + \alpha \tilde{n}} \\ \mu_j &:= \frac{\sum_{i=1}^n w_j^{(i)} x^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = j\} \tilde{x}^{(i)}}{\sum_{i=1}^n w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = j\}} \\ \Sigma_j &:= \frac{\sum_{i=1}^n w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = j\} (\tilde{x}^{(i)} - \mu_j)(\tilde{x}^{(i)} - \mu_j)^T}{\sum_{i=1}^n w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = j\}} \end{aligned}$$

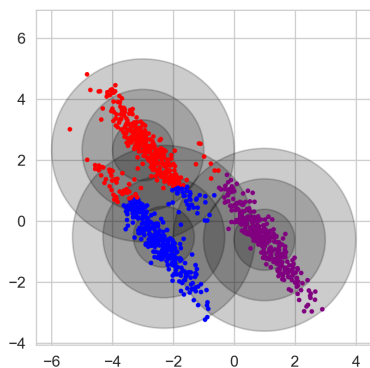
}

Điều kiện hội tụ:  $|l_{\text{semi-sup}}(\phi^{(t+1)}, \mu^{(t+1)}, \Sigma^{(t+1)}) - l_{\text{semi-sup}}(\phi^{(t)}, \mu^{(t)}, \Sigma^{(t)})| \leq \epsilon$

## 6 Áp dụng mô hình - Kết luận đánh giá

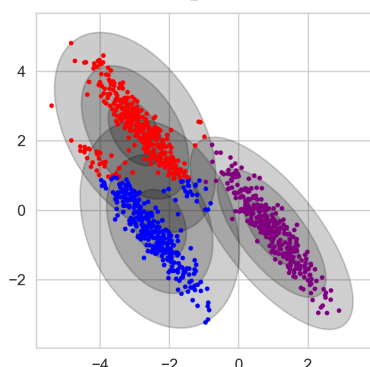
### 6.1 Visualize GMM sau các bước thuật toán EM

Ta sẽ dùng EM để tìm 3 phân phối Gaussian fit với 3 cụm của dataset **blobs-stretched**:



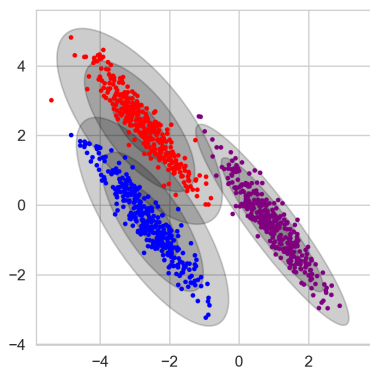
Step 0

Tại bước 0 - bước initialization, ta khởi tạo các  $\mu_i$  bằng các centroids tìm được trong thuật toán K-means, độ ưu tiên  $\pi_i = \frac{1}{3}$ , covariance matrices là các ma trận đơn vị.



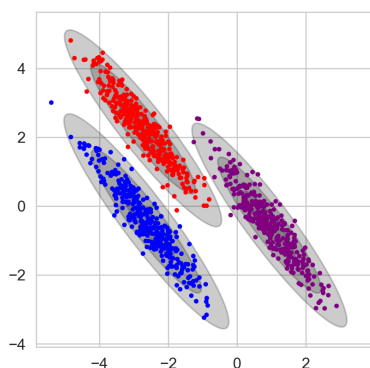
Step 1

Sau 1 bước EM, ta dễ dàng thấy các phân phối tìm được bắt đầu "co lại" để vừa với các cụm



Step 3

Sau 3 bước EM, phân phối Gaussian của cụm màu tím đã vừa vặn, cụm đỏ và xanh cũng phân đúng cho nhau và không còn nhầm lẫn một số điểm ở giữa.

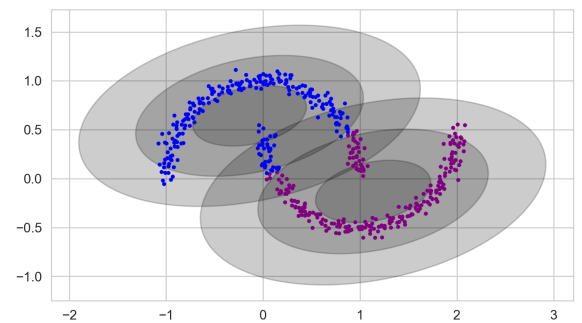
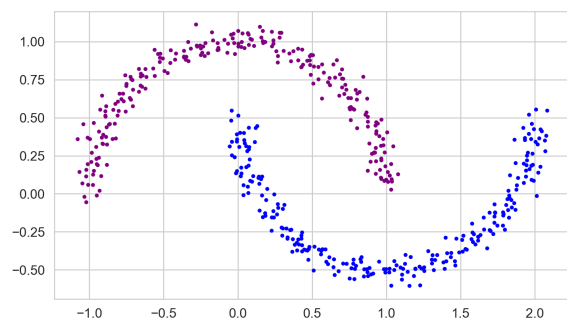
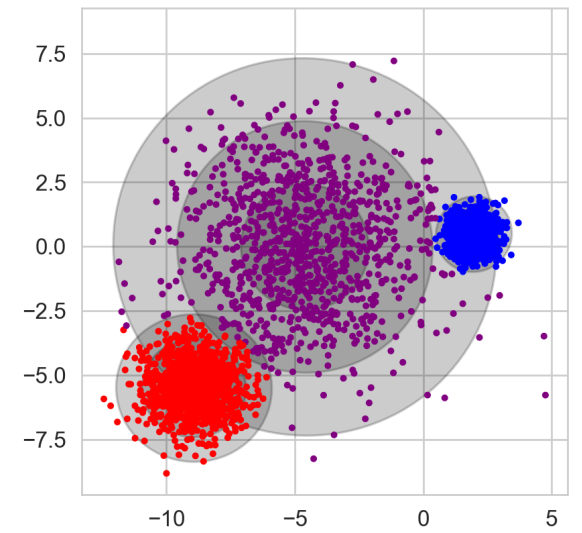
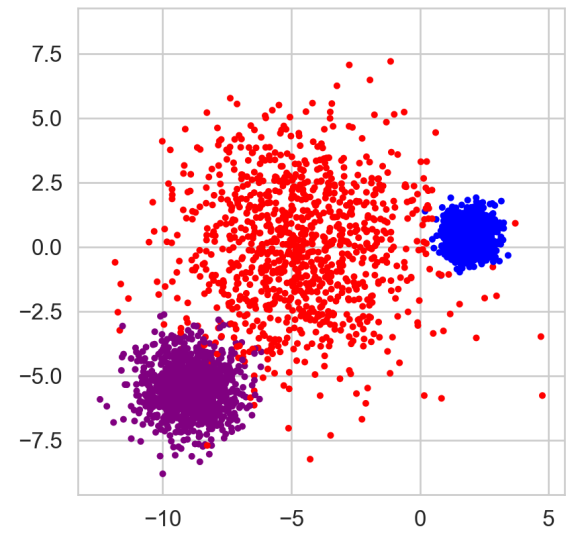
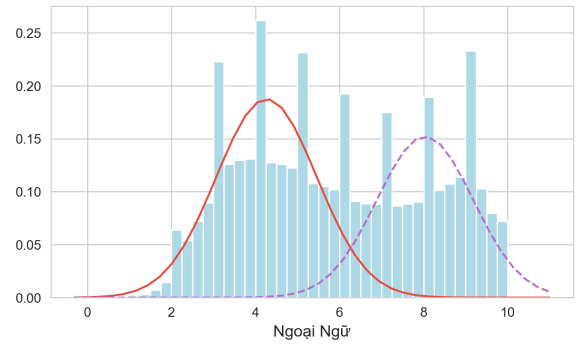
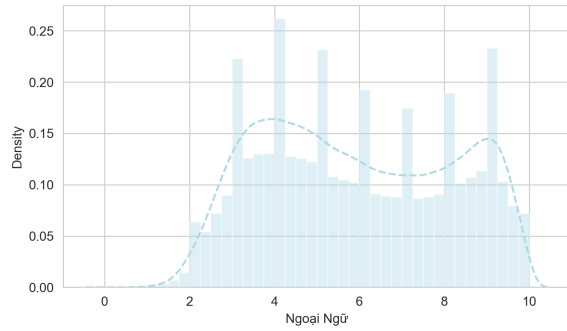
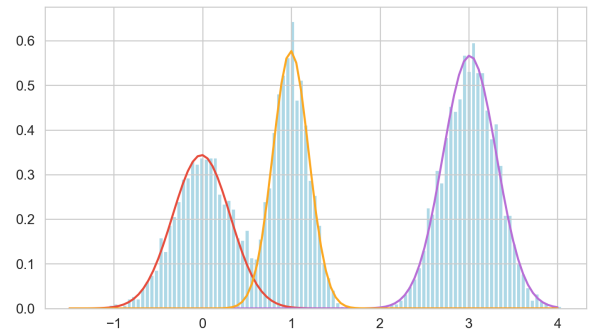
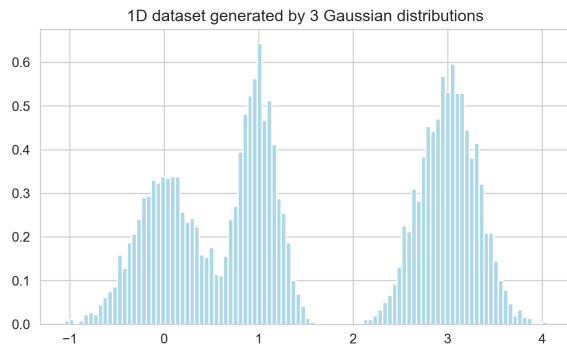


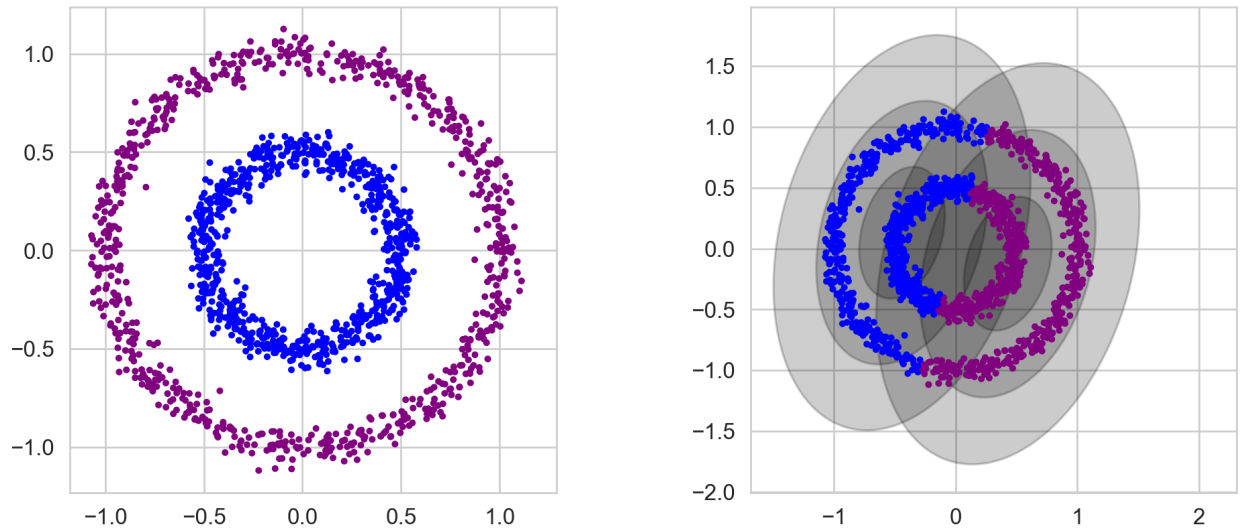
Step 11

Sau 11 bước EM, thuật toán đã tìm được 3 phân phối Gaussian fit với datasets.

### 6.2 Một số datasets khác sử dụng GMM

Các datasets ở bên trái, bên phải là sử dụng GMM để phân các cụm.





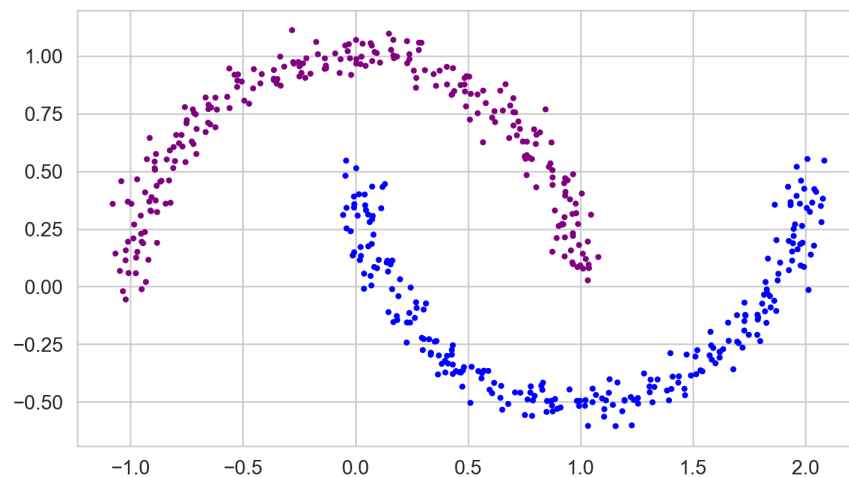
Ta thấy rằng 3 datasets đầu, thuật toán tìm được các cụm rất hợp lý và chính xác.

Trong khi đó, ở 2 datasets cuối, thuật toán không tìm được các cụm hợp lý, điều này là do 2 datasets này hình thù quá đặc biệt, quá khác biệt so với phân phối chuẩn Gaussian.

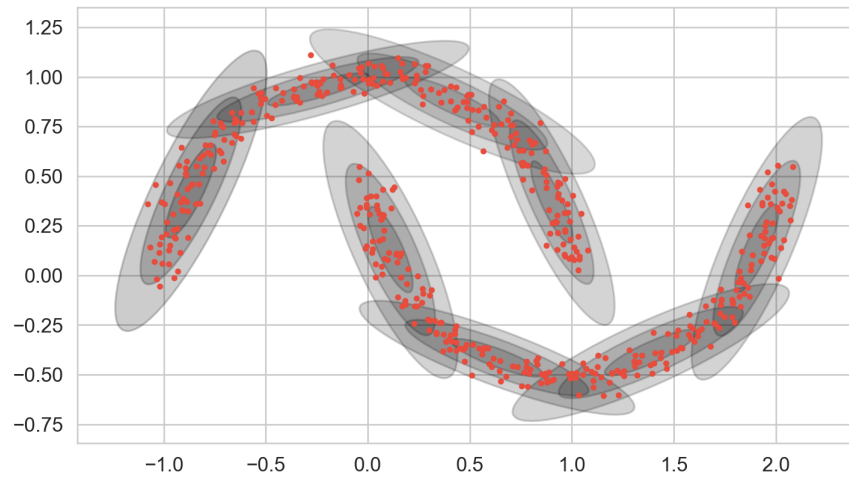
**Nhận xét:** Ta nhận thấy rằng GMM vừa có sự cải tiến hơn so với K-means (như dataset **blobs-stretched**) nhưng cũng gặp khó khăn như K-means với các dataset hình thù đặc biệt. Sự cải tiến này là do có thêm tham số **độ ưu tiên**  $\pi_i$  và **covariance matrices**  $\Sigma_i$ .

### 6.3 GMM như một generative model

Ta hãy cùng quay lại với dataset **2-moon** ở trên:

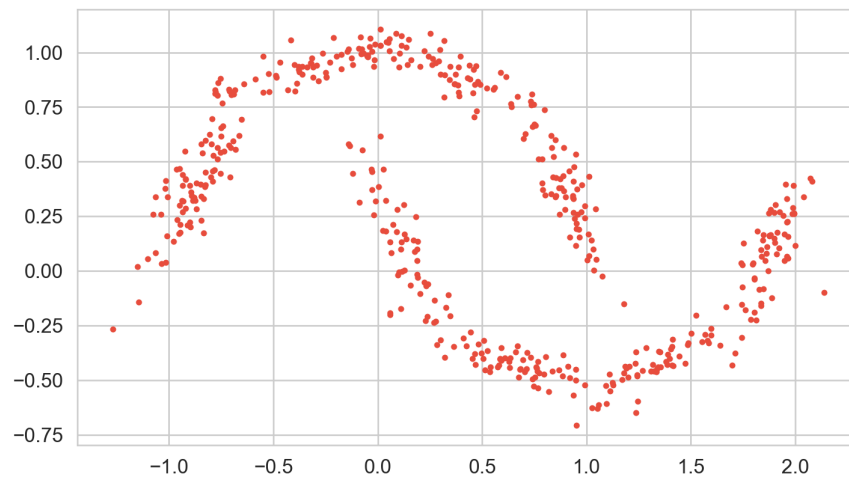


Ta sẽ dùng GMM với số cụm là 8.



Ta đã được 8 phân phối Gaussian mô tả khá hợp lý dataset này. Sau đó, ta sẽ sinh ngẫu nhiên thêm 500 điểm data bằng 8 phân phối đã tìm được.

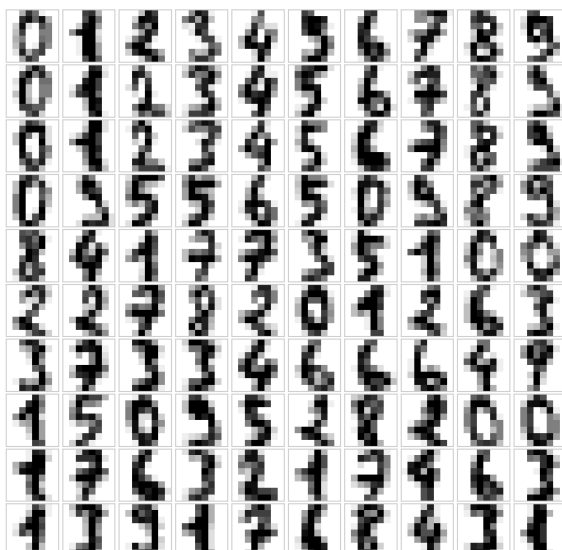
Ta được 500 điểm data mới như sau:



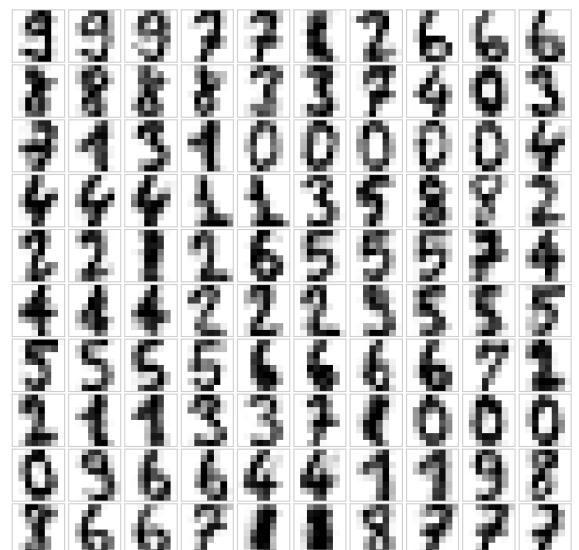
Các điểm data mới sinh ra phân phối rất khớp với datasets ban đầu.

Đây cũng là một công dụng của GMM nhằm giúp sinh thêm nhiều examples mới cho các model học máy.

Ta hãy cùng đến với một ví dụ thực tế hơn, sinh thêm các handwritten digits từ dataset cho trước bằng GMM.



100 real examples from sklearn



100 new generated examples

## 6.4 Chọn số cụm cho GMM

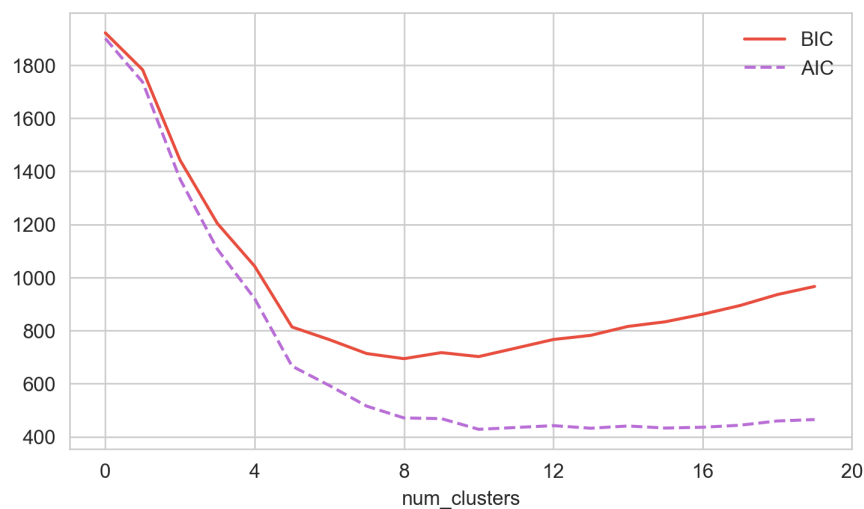
Ta sẽ dễ dàng thấy rằng nếu ta tăng số cụm của GMM lên càng nhiều, thì log-likelihood sẽ càng tăng. Nhưng đôi khi việc phân bằng số cụm nhiều thì model của ta sẽ không đủ "general" để đưa ra được các cách nhóm cụm hữu ích, hoặc nếu số cụm ít quá thì model sẽ không đủ "detailed" để mô tả các đặc điểm nổi bật của data.

Đối với các datasets từ 3D trở xuống, ta có thể chọn số cụm hợp lí bằng cách visualize dataset một cách trực quan. Còn đối với các datasets còn lại, một là giảm chiều của datasets để có thể visualize, hai là sử dụng chuẩn AIC và BIC.

Ta sẽ ví dụ sử dụng chuẩn AIC và BIC để có thể chọn số cụm sao cho phù hợp với data.

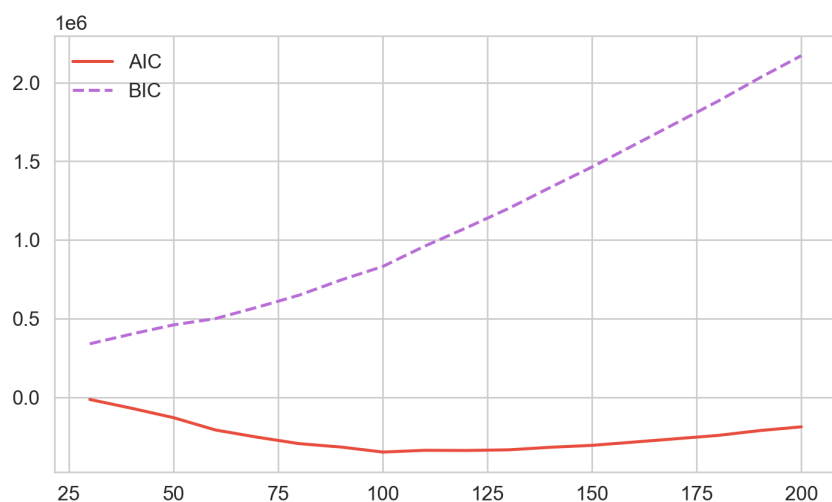
Nói ngắn gọn chuẩn AIC và BIC sẽ cho ta các giá trị đánh giá dựa trên sự trade-up giữa **số cụm** và **khả năng mô tả dataset**, giá trị càng thấp càng tốt. Từ đó tìm ra được số cụm hợp lí nhất sao cho khả năng mô tả dataset đủ tốt.

Ví dụ sử dụng chuẩn AIC và BIC cho dataset **2-moon** ở trên để mô tả tốt dataset.



Ta thấy số cụm nên sử dụng cho GMM để mô tả hiệu quả phân phối **2-moon** là vào khoảng từ 8 đến 12, vùng mà minimize giá trị AIC và BIC.

Ví dụ sử dụng chuẩn AIC và BIC cho dataset **handwritten digits** ở trên để mô tả tốt dataset.

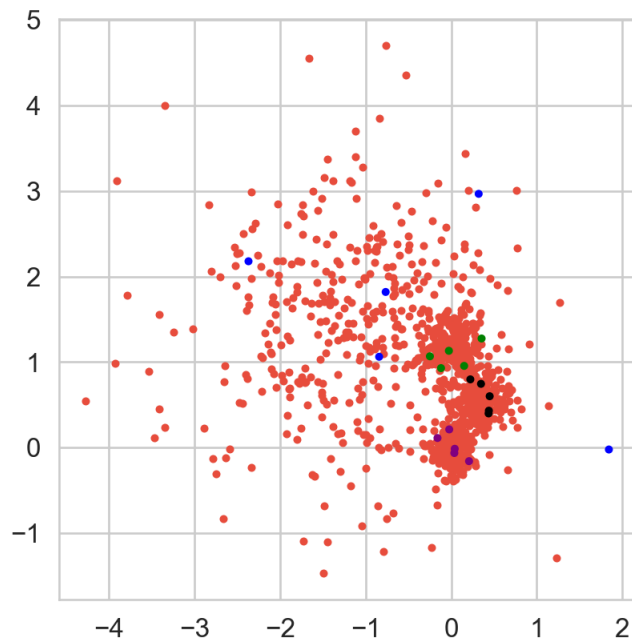


Ta thấy số cụm nên sử dụng cho GMM để mô tả hiệu quả **handwritten digits** là vào tầm khoảng 100.

## 6.5 GMM bán giám sát

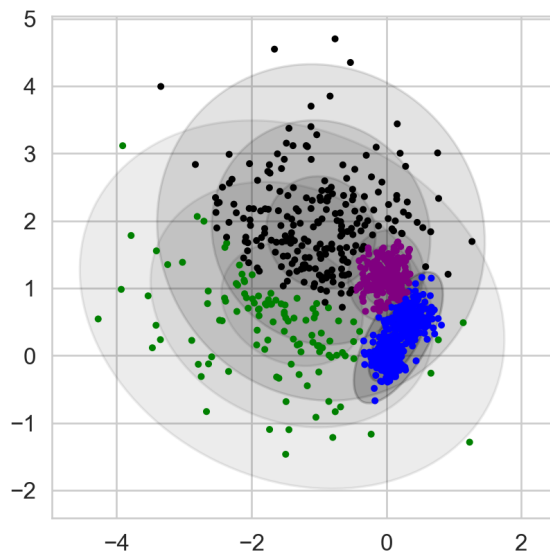
Ta sẽ xây dựng 2 mô hình GMM bán giám sát và GMM truyền thống và so sánh bằng một vài datasets.

Đầu tiên, ta hãy cũng nhìn vào dataset 2D sau đây:

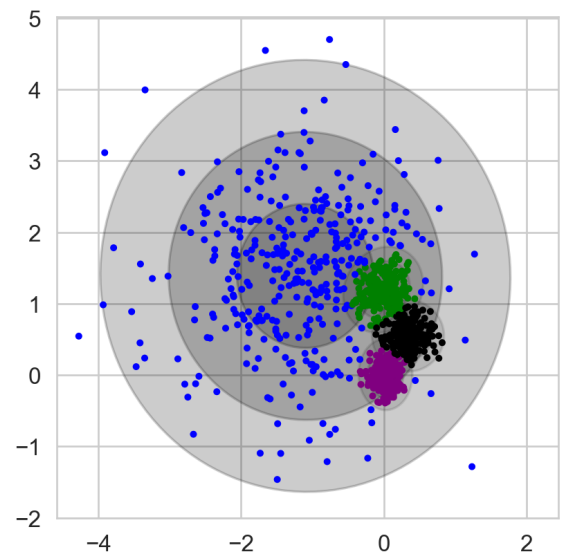


Đây là một dataset được tạo ra bởi 4 phân phối Gaussian gồm 1 phân phối rộng và 3 phân phối nhỏ gần nhau. Như ta thấy, mỗi phân phối có cho trước 5 điểm (khác với màu đỏ) thuộc phân phối đó.

Chạy 2 mô hình GMM, ta nhận được cách phân cụm sau đây:



GMM truyền thống



GMM bán giám sát

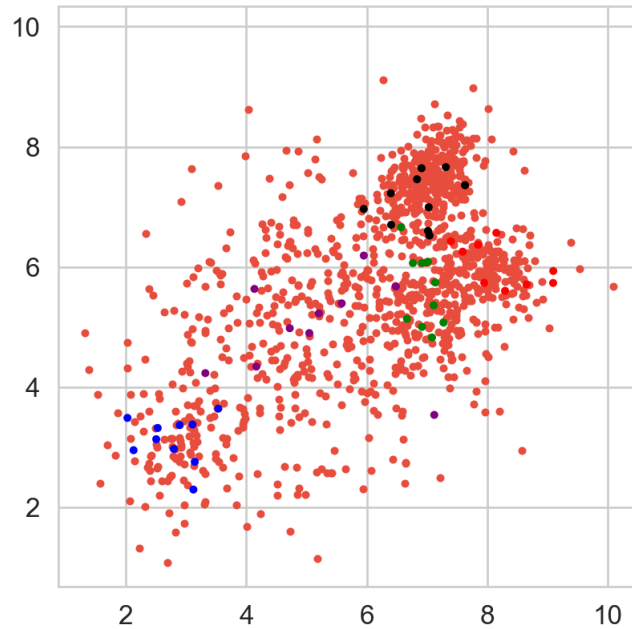
Ta thấy GMM với thuật toán EM truyền thống không phân cụm chính xác mà tách phân phối Gaussian lớn thành 2 phân phối và gộp 2 phân phối Gaussian nhỏ lại thành một.

Trong khi đó, chỉ với 5 điểm được cho biết trước ở mỗi phân phối Gaussian, GMM bán giám sát đã có thể phân cụm rất chính xác 4 cụm đã cho

Không những thế, GMM bán giám sát chỉ chạy trong khoảng 25 EM steps để đạt điều kiện converged, trong khi GMM truyền thống chạy trong 150 EM steps.

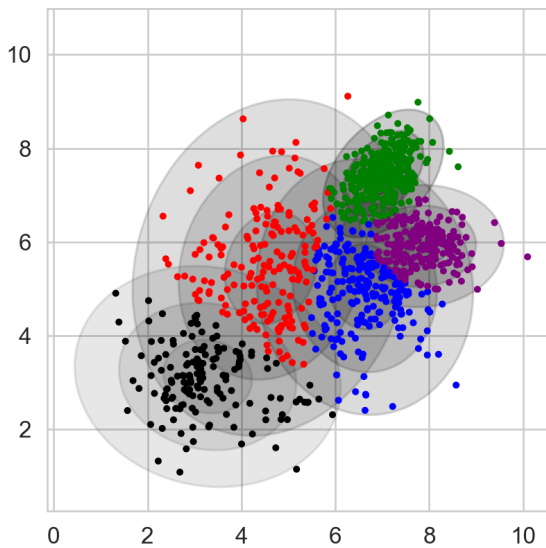
Tương tự như vậy, với dataset sau đây:



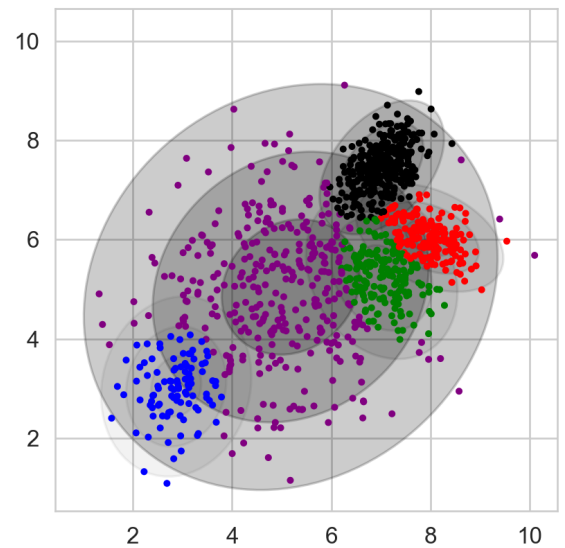


Đây là một dataset được tạo ra bởi 5 phân phối Gaussian. Như ta thấy, mỗi phân phối có cho trước 10 điểm (khác với màu đỏ) thuộc phân phối đó.

Chạy 2 mô hình GMM, ta nhận được cách phân cụm sau đây:



GMM truyền thống



GMM bán giám sát

Nhờ vào 10 điểm cho trước của mỗi phân phối, GMM bán giám sát phân cụm được chính xác và nhanh hơn so với GMM truyền thống.

## Tài liệu

- [1] Sean Borman. The expectation maximization algorithm a short tutorial, July 2004.
- [2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
- [3] Tengyu Ma and Andrew Ng. Cs229 notes 8, May 2019. <http://cs229.stanford.edu/summer2019/cs229-notes8.pdf>.
- [4] Andrew Ng. Cs229 notes 7b. <http://cs229.stanford.edu/summer2019/cs229-notes7b.pdf>.
- [5] Ramesh Sridharan. Gaussian mixture models and the em algorithm. <https://people.csail.mit.edu/rameshvs/content/gmm-em.pdf>.
- [6] Stanford. Cs229 summer 2019, problem set 3.
- [7] C. F. Jeff Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95 – 103, 1983.