ĐẠI HỌC HUẾ
KHOA KỸ THUẬT VÀ CÔNG NGHỆ
BỘ MÔN KHOA HỌC DỮ LIỆU VÀ TRÍ TUỆ NHÂN TẠO


HOÀNG THỊ HƯƠNG GIANG

# PHÂN CỤM MỀM

## ĐỒ ÁN
## PHÂN CỤM MỀM

THÀNH PHỐ HUẾ, NĂM 2025

ĐẠI HỌC HUẾ
KHOA KỸ THUẬT VÀ CÔNG NGHỆ
BỘ MÔN KHOA HỌC DỮ LIỆU VÀ TRÍ TUỆ NHÂN TẠO


HOÀNG THỊ HƯƠNG GIANG - 23E1020001

# PHÂN CỤM MỀM

## ĐỒ ÁN
## HỌC MÁY II

**Giảng viên hướng dẫn:**
**TS. Nguyễn Đăng Trị**

**THÀNH PHỐ HUẾ, NĂM 2025**

# LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của TS. Nguyễn Đăng Trị. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Khóa luận/Đồ án tốt nghiệp còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Khóa luận/Đồ án tốt nghiệp của mình. Khoa Kỹ thuật và Công nghệ - Đại học Huế không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

**Tp Huế, ngày ... tháng .... năm ...**

**Hoàng Thị Hương Giang**

# Lời cảm ơn

Em xin chân thành cảm ơn đến Ban giám hiệu Trường Đại học Huế, Ban chủ nhiệm khoa Kỹ Thuật và Công Nghệ đã giúp đỡ sinh viên chúng em trong suốt thời gian học tại trường.
Em xin chân thành cảm ơn thầy Nguyễn Đăng Trị - giảng viên bộ môn Học máy II nhiệt tình giúp em hoàn thành đồ án này.
Do kiến thức bản thân còn hạn chế, trong quá trình làm đồ án của em không tránh khỏi những thiếu sót. Em mong nhận được sự góp ý của thầy để đồ án hoàn chỉnh hơn.
Em xin chân thành cảm ơn !

Hoàng Thị Hương Giang

April, 2025

# Tóm tắt

Bài tiểu luận này tập trung vào phân cụm mềm (soft clustering), một kỹ thuật quan trọng trong lĩnh vực học máy. Khác với phân cụm cứng (hard clustering) - nơi mỗi điểm dữ liệu chỉ thuộc về một cụm duy nhất, phân cụm mềm cho phép một điểm dữ liệu thuộc về nhiều cụm với các mức độ thành viên khác nhau.

Trong số các thuật toán phân cụm mềm, Gaussian Mixture Model (GMM) nổi bật nhờ khả năng mô hình hóa dữ liệu phức tạp bằng cách giả định rằng dữ liệu được tạo ra từ một hỗn hợp các phân phối Gaussian. Bài tiểu luận này sẽ đi sâu vào lý thuyết và ứng dụng của GMM trên cơ sở lý thuyết giải thích các khái niệm cơ bản về GMM, bao gồm hàm mật độ xác suất Gaussian đa biến, ước lượng tham số bằng thuật toán Expectation-Maximization (EM). Có ứng dụng trong các bài toán thực tế như phân tích hình ảnh, phân cụm văn bản và nhận dạng giọng nói. Mục tiêu của bài tiểu luận là cung cấp một cái nhìn tổng quan và chi tiết về GMM, giúp người đọc hiểu rõ về khả năng và ứng dụng của thuật toán này trong lĩnh vực phân cụm dữ liệu.

# MỤC LỤC

# Danh mục hình ảnh

# Danh mục viết tắt

| | |
|---|---|
| GMM | Gaussian Mixture Model |
| AIC | Akaike Information Criterion |
| BIC | Bayesian Information Criterion |
| PCA | Principal Component Analysis |
| GFF | Gaussian Filter Function |
| AIC | Akaike information criterion |
| BIC | Bayesian information criterion |
| PCA | Principal Component Analysis |
| Var | variance (phương sai) |
| Cov | covariance (hiệp phương sai) |
| exp | Exponential Function (Hàm mũ) |

# Phần I

# PHẦN MỞ ĐẦU

# Chương 1

# Mở đầu

## 1.1 Giới thiệu

Phân cụm là một thuật toán học máy không giám sát nhằm mục đích chia một tập dữ liệu thành các nhóm hoặc cụm, trong đó các điểm trong cùng một cụm có các đặc điểm tương tự và khác biệt với các điểm trong các cụm khác. Về cơ bản, các thuật toán phân cụm được phân loại thành hai loại chính: phân cụm mềm và phân cụm cứng. Nhưng ở đây ta chỉ làm việc với phân cụm mềm. [1]

Phân cụm mềm: Thay vì gán mỗi điểm vào một cụm, thuật toán sẽ tính toán xác suất của mỗi điểm dữ liệu thuộc về các cụm khác nhau. Điều này có nghĩa là một điểm có thể thuộc về nhiều cụm với các xác suất khác nhau. Gaussian Mixture Model (GMM) [1] là thuật toán được giới thiệu trong bài viết này.

# Phần II

# NỘI DUNG

# Chương 2

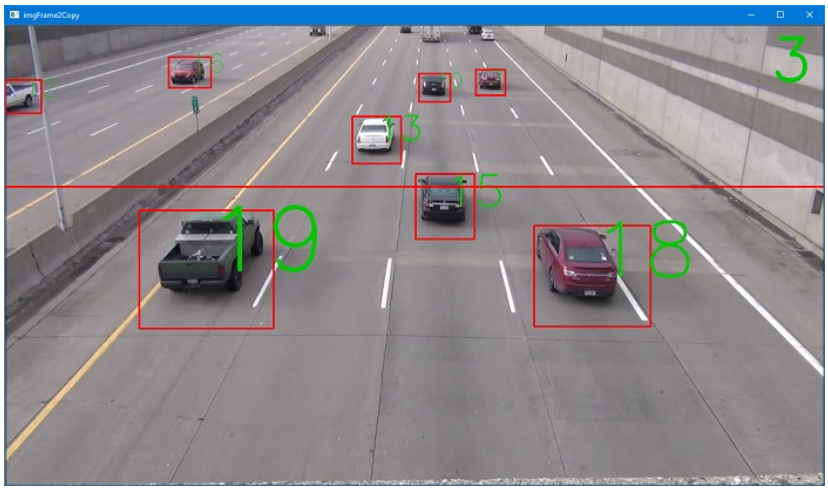
# Tổng quan về tình hình nghiên cứu

## 2.1 Tình hình nghiên cứu trong nước

Tại Việt Nam, nghiên cứu về GMM chủ yếu tập trung vào ứng dụng thực tiễn. Mặc dù có nhiều công trình ứng dụng GMM trong các lĩnh vực khác nhau, nhưng thông tin về các nhà nghiên cứu cụ thể và các nghiên cứu lý thuyết về GMM tại Việt Nam còn hạn chế.

Hệ thống khuyến nghị sản phẩm: Nguyễn Văn Đạt và Tạ Minh Thanh đã đề xuất một thuật toán khuyến nghị dựa trên phân phối sử dụng GMM để nhóm các sản phẩm thành các cụm khác nhau, sau đó áp dụng hàm Gaussian Filter Function (GFF) để tính độ tương đồng và sắp xếp kết quả. Kết quả cho thấy độ chính xác cao và thời gian xử lý nhanh, phù hợp với các ứng dụng thực tế. [2]

Giám sát giao thông: Ngô Quốc Tạo, Nguyễn Văn Căn và Huỳnh Văn Huy đã nghiên cứu phương pháp đếm xe ô tô sử dụng GMM kết hợp với luồng quang học. Phương pháp này cải tiến mô hình nền GMM truyền thống, thích ứng tốt với sự thay đổi ánh sáng và nền động, đạt kết quả chính xác trong điều kiện mật độ xe thấp trên các đường cao tốc. [3]




Hình 2.1: Phương pháp đếm xe ô tô sử dụng GMM

## 2.2 Tình hình nghiên cứu ngoài nước

Mô hình hỗn hợp Gaussian (GMM) là kết quả của quá trình phát triển lâu dài với sự đóng góp của nhiều nhà khoa học. Carl Friedrich Gauss (30 tháng 4, 1777), Gauss đã nghiên cứu và phát triển lý thuyết về phân phối Gaussian vào đầu thế kỷ 19. Các công trình của ông về phương pháp bình phương tối thiểu, một kỹ thuật thống kê quan trọng, đã dẫn đến sự phát hiện và ứng dụng rộng rãi của phân phối Gaussian. Gauss đã đưa ra những công thức toán học và các tính chất quan trọng của phân phối Gaussian, cho phép các nhà khoa học và thống kê viên sử dụng nó để mô hình hóa và phân tích dữ liệu. Phân phối Gaussian đã trở thành một công cụ cơ bản trong thống kê và xác suất, được ứng dụng rộng rãi trong nhiều lĩnh vực khoa học và kỹ thuật., đã đặt nền tảng lý thuyết cơ bản [4] . Tiếp theo, trong thế kỷ 20, nhiều nhà thống kê và nhà nghiên cứu học máy đã xây dựng và phát triển các thuật toán ước lượng tham số cho GMM, đặc biệt là sự ra đời của thuật toán Expectation-Maximization (EM) đã mở ra khả năng ứng dụng rộng rãi cho mô hình này. Christopher M. Bishop, thông qua cuốn sách "Pattern Recognition and Machine Learning", đã góp phần hệ thống hóa và phổ biến kiến thức về GMM. Gần đây, với sự phát triển của học sâu, nhiều nhà nghiên cứu đã tìm cách kết hợp GMM với các mô hình mạng nơ-ron, tạo ra những phương pháp mạnh mẽ hơn. Vì vậy, GMM là một thành tựu chung của cộng đồng khoa học, được xây dựng và hoàn thiện qua nhiều thế hệ. [5]

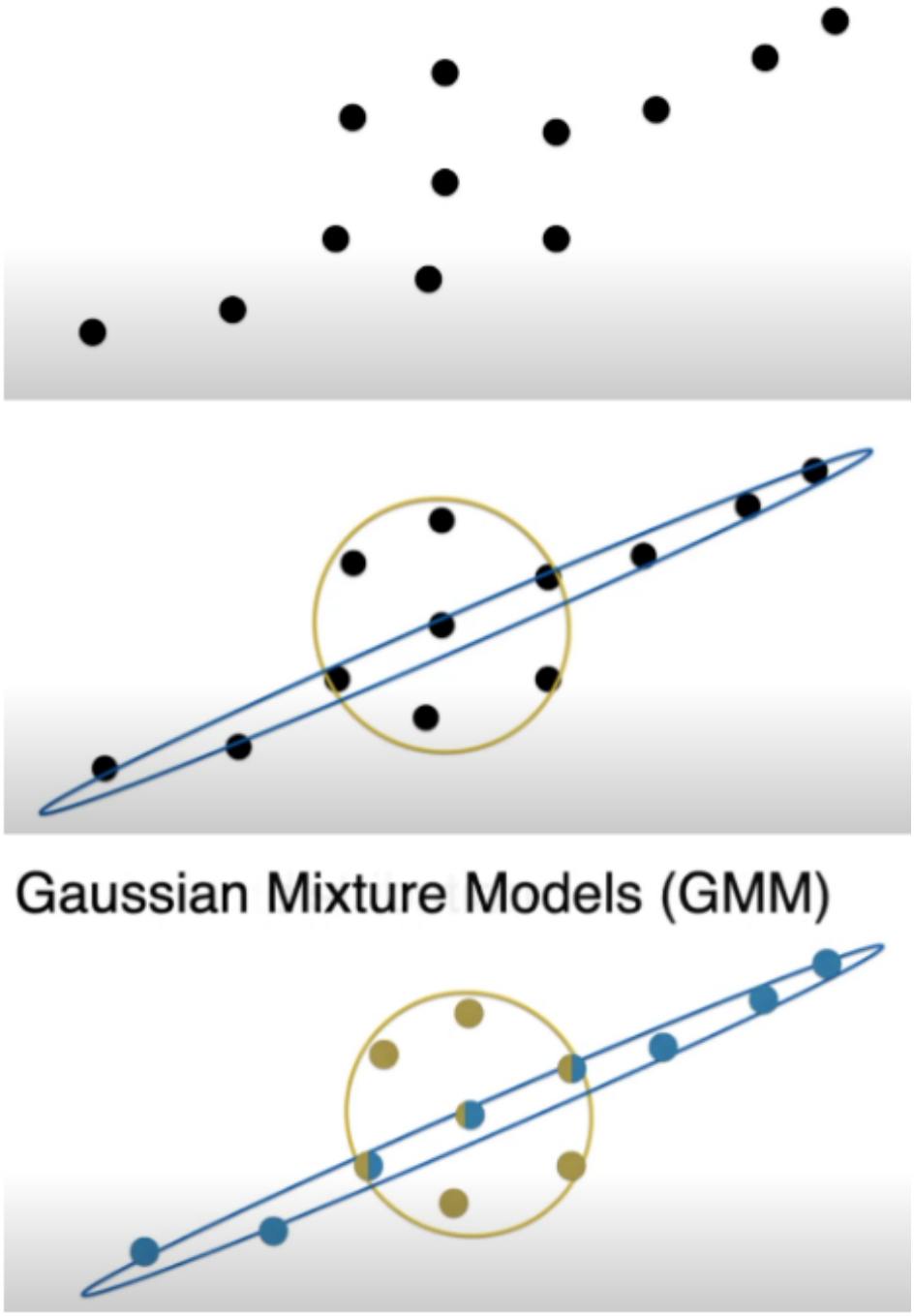
Hình 2.2: Johann Carl Friedrich Gauß (30/4/1777 - 23/2/1855).

# Chương 3

# Đặt vấn đề và mô tả bài toán

## 3.1 Đặt vấn đề

Trong nhiều bài toán phân loại dữ liệu, các phương pháp phân loại cứng như K-means thường gặp khó khăn trong việc xử lý các điểm dữ liệu có sự chồng chéo hoặc thuộc tính không rõ ràng. Thay vào đó, phân loại mềm, đặc biệt là Mô hình Hỗn hợp Gaussian (Gaussian Mixture Model - GMM), trở nên hữu ích hơn nhiều. GMM cho phép mỗi điểm dữ liệu thuộc về nhiều cụm với các mức độ thành viên khác nhau, thể hiện sự không chắc chắn và tính linh hoạt trong việc phân loại. Điều này giúp mô hình hóa dữ liệu phức tạp hiệu quả hơn và đưa ra kết quả phân loại chính xác hơn, khiến quá trình phân tích và hiểu dữ liệu trở nên dễ dàng và trực quan hơn.




Hình 3.1: Minh họa vấn đề

## 3.2 Mô tả bài toán

Trong quá trình điều chỉnh một phân phối Gaussian cho dữ liệu, bao gồm việc tính toán trung bình (mean), ma trận hiệp phương sai (covariance matrix), và hàm mật độ xác suất của phân phối Gaussian đa biến. Sau đây sẽ trình bày, giải thích ý nghĩa, cách sử dụng, và vai trò của chúng trong việc mô hình hóa dữ liệu.

**Công thức tính trung bình (mean) $\mu$:**

$$\mu = \text{Average}$$

Trong ngữ cảnh của phân phối Gaussian, $\mu$ là vector trung bình, biểu thị "trung tâm khối lượng"(center of mass) của tập dữ liệu. Trong không gian đa chiều, $\mu$ là một vector chứa giá trị trung bình của từng biến số (ở đây là $x$ và $y$).

Để tính $\mu$, ta lấy trung bình cộng của tất cả các điểm dữ liệu theo từng chiều. Nếu tập dữ liệu có $n$ điểm $(x_i, y_i)$, thì:

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i$$

[6]

Vector $\mu$ sẽ là $[\mu_x, \mu_y]$. Nhằm xác định vị trí trung tâm của phân phối Gaussian, là điểm mà các giá trị dữ liệu tập trung xung quanh.

**Công thức tính ma trận hiệp phương sai $\Sigma$:**

$$\Sigma = \begin{pmatrix} \text{Var}(x) & \text{Cov}(x,y) \\ \text{Cov}(x,y) & \text{Var}(y) \end{pmatrix}$$

[7]

- **Var(x):** Phương sai của biến $x$, đo lường mức độ phân tán của $x$ xung quanh giá trị trung bình $\mu_x$. Công thức tính phương sai là:

$$\text{Var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2$$

[8]

- **Var(y):** Tương tự, đây là phương sai của biến $y$:

$$\text{Var}(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_y)^2$$

- **Cov(x, y):** Hiệp phương sai giữa $x$ và $y$, đo lường mức độ mà hai biến này thay đổi cùng nhau. Công thức là:

$$\text{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

[9]

Ma trận hiệp phương sai $\Sigma$ mô tả hình dạng và hướng của phân phối Gaussian. Các phần tử trên đường chéo (Var($x$) và Var($y$)) biểu thị độ phân tán của dữ liệu theo các trục $x$ và $y$, trong khi hiệp phương sai (Cov($x$, $y$)) thể hiện mối quan hệ tuyến tính giữa $x$ và $y$. Nếu Cov($x$, $y$) > 0, $x$ và $y$ có xu hướng tăng hoặc giảm cùng nhau; nếu Cov($x$, $y$) < 0, chúng có xu hướng ngược nhau.

**Hàm mật độ xác suất của phân phối Gaussian đa biến $f(x)$:**

Công thức được trình bày trong hai dạng:

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

[10]

và một dạng đơn giản hơn (có thể cho trường hợp một chiều):

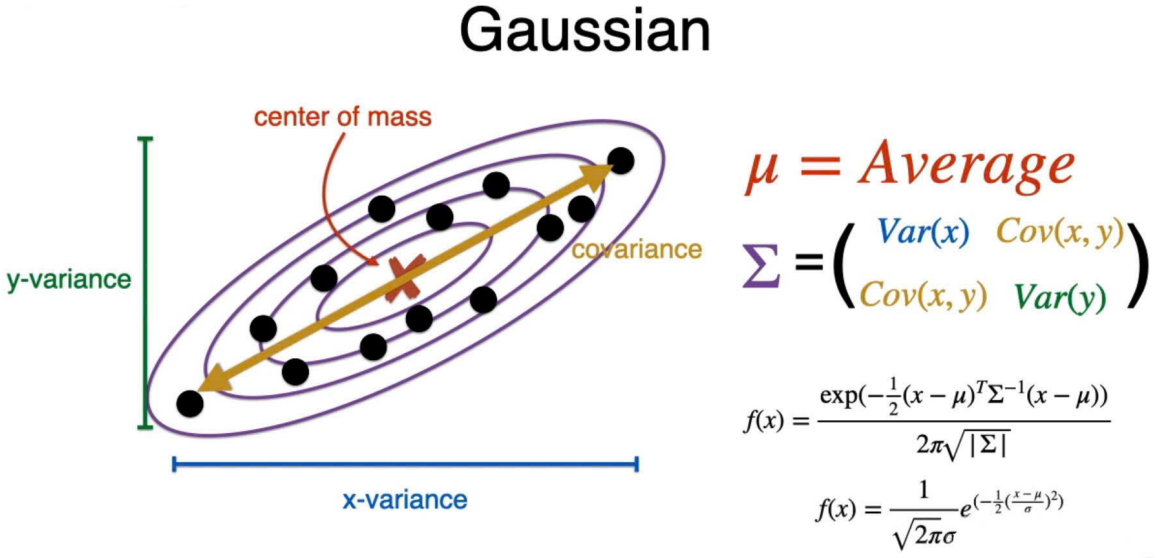
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

[11]

- $x$: Đây là vector dữ liệu, trong trường hợp hai chiều là $[x, y]$.
- $\mu$: Vector trung bình, như đã giải thích ở trên.
- $\Sigma$: Ma trận hiệp phương sai.
- $\Sigma^{-1}$: Ma trận nghịch đảo của $\Sigma$.

- $|\Sigma|$: Định thức của ma trận $\Sigma$, ảnh hưởng đến hằng số chuẩn hóa.
- $(x-\mu)^T\Sigma^{-1}(x-\mu)$: Đây là dạng bậc hai, đo lường khoảng cách Mahalanobis từ $x$ đến $\mu$, có tính đến hiệp phương sai giữa các biến.
- Hằng số chuẩn hóa $2\pi\sqrt{|\Sigma|}$: Đảm bảo tổng xác suất của phân phối bằng 1.

Hàm $f(x)$ cho biết xác suất của một điểm dữ liệu $x$ trong phân phối Gaussian. Phân phối này có dạng hình chuông (bell-shaped) trong không gian một chiều hoặc hình elip trong không gian hai chiều (như minh họa trong hình).




Hình 3.2: Hình ảnh minh họa

# Chương 4

# Đề xuất phương án giải quyết bài toán

## 4.1 Thách thức của bài toán

Trong quá trình huấn luyện mô hình Gaussian Mixture Model (GMM), việc xác định số cụm $K$ tối ưu là một thách thức quan trọng. Nếu lựa chọn $K$ quá nhỏ, mô hình có thể không đủ linh hoạt để mô tả đầy đủ cấu trúc của dữ liệu, dẫn đến việc phân cụm không chính xác. Ngược lại, nếu $K$ quá lớn, mô hình có nguy cơ bị overfitting [12], khiến khả năng tổng quát hóa đối với dữ liệu mới bị suy giảm. Do đó, cần có các phương pháp đánh giá khách quan như tiêu chí Akaike (AIC), tiêu chí Bayesian (BIC) hoặc phương pháp Elbow để lựa chọn $K$ tối ưu, đảm bảo sự cân bằng giữa độ chính xác và tính đơn giản của mô hình. Bên cnh ó, mt thách thc khác ca GMM là kh năng din gii kt qu khi d liu có s chiu ln. Trong không gian nhiều chiều, việc trực quan hóa các cụm trở nên khó khăn, gây cản trở quá trình đánh giá và phân tích mô hình. Ngoài ra, các ma trận hiệp phương sai trong GMM có thể trở nên phức tạp, khiến việc so sánh và hiểu rõ mối quan hệ giữa các cụm trở nên khó khăn hơn. Để khắc phục vấn đề này, các phương pháp giảm chiều như Phân tích Thành phần Chính (PCA) hoặc t-SNE có thể được áp dụng nhằm giúp trực quan hóa dữ liệu tốt hơn, đồng thời giữ lại thông tin quan trọng của các cụm.

## 4.2 Phương án giải quyết

*** Chọn số cụm $K$ tối ưu bằng ( Akaike Information Criterion) AIC và ( Bayesian Information Criterion) BIC**

Một trong những thách thức chính của Gaussian Mixture Model (GMM) là phải chọn trước số cụm $K$. Nếu chọn $K$ quá nhỏ, mô hình sẽ không đủ linh hoạt để mô tả dữ liệu. Nếu chọn $K$ quá lớn, mô hình có thể bị overfitting.

**Tiêu chí Akaike Information Criterion (AIC)**

AIC đánh giá mức độ phù hợp của mô hình và có công thức như sau:

$$\text{AIC} = 2p - 2\log L$$

[13]

**Trong đó:**

- $p$ là số tham số của mô hình (càng nhiều cụm $K$, $p$ càng lớn).
- $L$ là log-likelihood của mô hình (càng lớn, mô hình càng khớp dữ liệu tốt).

**Nguyên tắc chọn $K$:** Chọn $K$ sao cho AIC nhỏ nhất.

**Tiêu chí Bayesian Information Criterion (BIC)**

BIC phạt mô hình phức tạp mạnh hơn AIC để tránh overfitting:

$$\text{BIC} = p\log N - 2\log L$$

[13]

**Sự khác biệt với AIC:**

- Thay vì chỉ phạt $p$, BIC nhân thêm hệ số $\log N$, trong đó $N$ là số điểm dữ liệu.
- Khi $N$ lớn, BIC ưu tiên mô hình đơn giản hơn AIC.

**Nguyên tắc chọn $K$:** Chọn $K$ sao cho BIC nhỏ nhất.

**So sánh AIC và BIC**

| Tiêu chí | Công thức | Khi nào dùng? |
|---|---|---|
| AIC | $2p - 2\log L$ | Khi muốn cân bằng giữa độ chính xác và độ phức tạp |
| BIC | $p\log N - 2\log L$ | Khi ưu tiên mô hình đơn giản và tránh overfitting |

**Thực tế:**

- BIC thường chọn số cụm nhỏ hơn AIC do mức phạt cao hơn.
- Nếu hai tiêu chí cho kết quả giống nhau, đó có thể là số cụm tối ưu.
- Nếu kết quả khác nhau, có thể kiểm tra thêm bằng phương pháp Elbow hoặc Silhouette Score.

*** Overfitting khi số chiều cao - Giải pháp ( Principal Component Analysis ) PCA**

Một vấn đề lớn khi dùng GMM là số chiều của dữ liệu quá cao (High-dimensional Data). Khi số chiều quá cao, GMM có thể gặp phải hai vấn đề chính:

**Overfitting:**

- Khi số chiều lớn nhưng số điểm dữ liệu ít, mô hình dễ bị overfitting do có quá nhiều tham số.
- Ma trận hiệp phương sai $\Sigma_k$ có thể trở nên suy biến (singular), làm mất ổn định tính toán.

**Tốc độ chậm:**

- GMM có độ phức tạp $O(NKd^2)$, trong đó $d$ là số chiều.
- Nếu $d$ lớn, việc tính toán ma trận hiệp phương sai trở nên rất chậm.

**Giải pháp: Giảm chiều bằng Principal Component Analysis (PCA)**

**Nguyên tắc của PCA**

PCA giúp giảm chiều dữ liệu bằng cách tìm các trục quan trọng nhất trong dữ liệu và chỉ giữ lại các thành phần chính (principal components) thay vì giữ toàn bộ các chiều gốc.

$$X' = XW$$

**Trong đó:**

- $X$ là dữ liệu ban đầu ($N$ điểm, $d$ chiều).
- $W$ là ma trận thành phần chính (chọn ra $m$ chiều quan trọng nhất).
- $X'$ là dữ liệu mới có số chiều thấp hơn ($m < d$).

**Mục tiêu:** Chọn $m$ sao cho vẫn giữ được nhiều thông tin nhất có thể.

**Cách áp dụng PCA vào GMM** [14]

**- Chuẩn hóa dữ liệu (Standardization):**

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

Giúp dữ liệu có cùng tỉ lệ, tránh PCA bị chi phối bởi đơn vị đo lường.

**- Chạy PCA để chọn số chiều tối ưu:**

- Giữ lại 95% phương sai để đảm bảo không mất quá nhiều thông tin.
- Dùng Scree Plot hoặc Cumulative Explained Variance để chọn $m$.

**- Chạy GMM trên dữ liệu đã giảm chiều:**

- Huấn luyện GMM với dữ liệu mới có $m$ chiều.
- Chọn số cụm $K$ tối ưu bằng AIC/BIC.

# Chương 5

# Kết quả mô phỏng và đánh giá

## 5.1 Quy Trình thực hiên mô phỏng

Trong bài viết này, thực hiện mô phỏng trên tập dữ liệu bệnh nhân ung thư bằng cách sử dụng các phương pháp tiền xử lý dữ liệu, phân cụm và đánh giá mô hình. [15] [16]

| | Age | Gender | Air Pollution | Alcohol use | Dust Allergy | OccuPational Hazards | Genetic Risk | chronic Lung Disease | Balanced Diet | Obesity | ... | Fatigue | Weight Loss | Shortness of Breath | Wheezing | Swallowing Difficulty | Clubbing of Finger Nails | Frequent Cold | Dry Cough |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 33 | 1 | 2 | 4 | 5 | 4 | 3 | 2 | 2 | 4 | ... | 3 | 4 | 2 | 2 | 3 | 1 | 2 | 3 |
| **1** | 17 | 1 | 3 | 1 | 5 | 3 | 4 | 2 | 2 | 2 | ... | 1 | 3 | 7 | 8 | 6 | 2 | 1 | 7 |
| **2** | 35 | 1 | 4 | 5 | 6 | 5 | 5 | 4 | 6 | 7 | ... | 8 | 7 | 9 | 2 | 1 | 4 | 6 | 7 |
| **3** | 37 | 1 | 7 | 7 | 7 | 7 | 6 | 7 | 7 | 7 | ... | 4 | 2 | 3 | 1 | 4 | 5 | 6 | 7 |
| **4** | 46 | 1 | 6 | 8 | 7 | 7 | 7 | 6 | 7 | 7 | ... | 3 | 2 | 4 | 1 | 4 | 2 | 4 | 2 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **995** | 44 | 1 | 6 | 7 | 7 | 7 | 7 | 6 | 7 | 7 | ... | 5 | 3 | 2 | 7 | 8 | 2 | 4 | 5 |
| **996** | 37 | 2 | 6 | 8 | 7 | 7 | 7 | 6 | 7 | 7 | ... | 9 | 6 | 5 | 7 | 2 | 4 | 3 | 1 |
| **997** | 25 | 2 | 4 | 5 | 6 | 5 | 5 | 4 | 6 | 7 | ... | 8 | 7 | 9 | 2 | 1 | 4 | 6 | 7 |
| **998** | 18 | 2 | 6 | 8 | 7 | 7 | 7 | 6 | 7 | 7 | ... | 3 | 2 | 4 | 1 | 4 | 2 | 4 | 2 |
| **999** | 47 | 1 | 6 | 5 | 6 | 5 | 5 | 4 | 6 | 7 | ... | 8 | 7 | 9 | 2 | 1 | 4 | 6 | 7 |

Hình 5.1: Mô tả dữ liệu

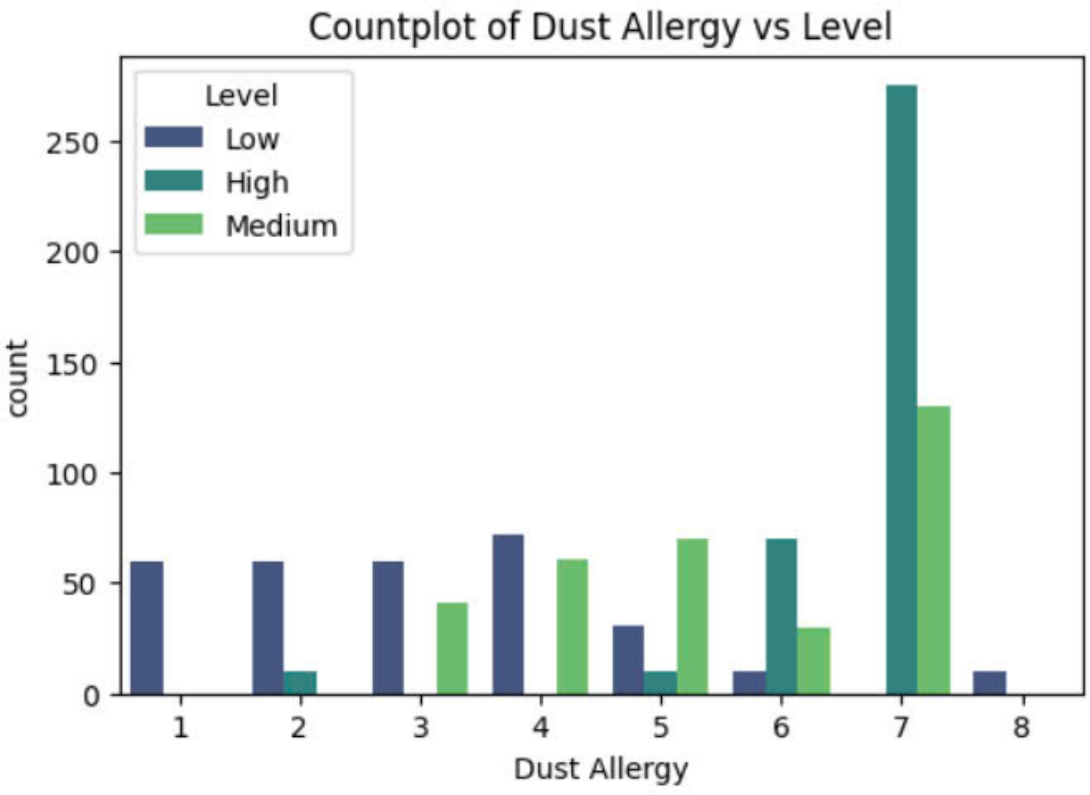
**Bước 1: Tiền xử lý dữ liệu**

```
array([[-0.34784816, -0.81990292, -0.90667901, ..., -0.41855027,
         0.72865507,  0.03955825],
       [-1.68123833, -0.81990292, -0.41391868, ...,  1.54417079,
        -0.6282445 ,  1.23829315],
       [-0.18117439, -0.81990292,  0.07884165, ...,  1.54417079,
        -0.6282445 , -1.15917665],
       ...,
       [-1.01454325,  1.21965659,  0.07884165, ...,  1.54417079,
        -0.6282445 , -1.15917665],
       [-1.59790145,  1.21965659,  1.06436231, ..., -0.90923053,
         0.05020528, -1.15917665],
       [ 0.81886824, -0.81990292,  1.06436231, ...,  1.54417079,
        -0.6282445 , -1.15917665]])
```

Hình 5.2: Dữ liệu sau tiền xử lý

- **Đọc dữ liệu**: Tập dữ liệu được tải vào và xử lý bằng thư viện Pandas.
- **Xóa cột không cần thiết**: Các cột 'index' và 'Patient Id' được loại bỏ vì không có giá trị phân tích.
- **Kiểm tra giá trị thiếu**: Kiểm tra và xác định các giá trị bị thiếu trong tập dữ liệu.
- **Mã hóa biến phân loại**: Các giá trị thuộc tính 'Gender' và 'Level' được chuyển đổi thành dạng số bằng LabelEncoder.
- **Chuẩn hóa dữ liệu**: Sử dụng StandardScaler để chuẩn hóa các thuộc tính liên tục.

**Bước 2: Phân tích dữ liệu bằng trực quan hóa**




Hình 5.3: Trực quan hóa dữ liệu

- **Phân nhóm dữ liệu liên tục**: Các thuộc tính số được chia thành các nhóm (binned) để phân tích mức độ ảnh hưởng đến 'Level'.

- **Vẽ biểu đồ tần suất**: Biểu đồ phân bố của các thuộc tính liên tục được vẽ để quan sát sự khác biệt giữa các mức độ bệnh.

**Bước 3: Áp dụng PCA để giảm chiều dữ liệu**

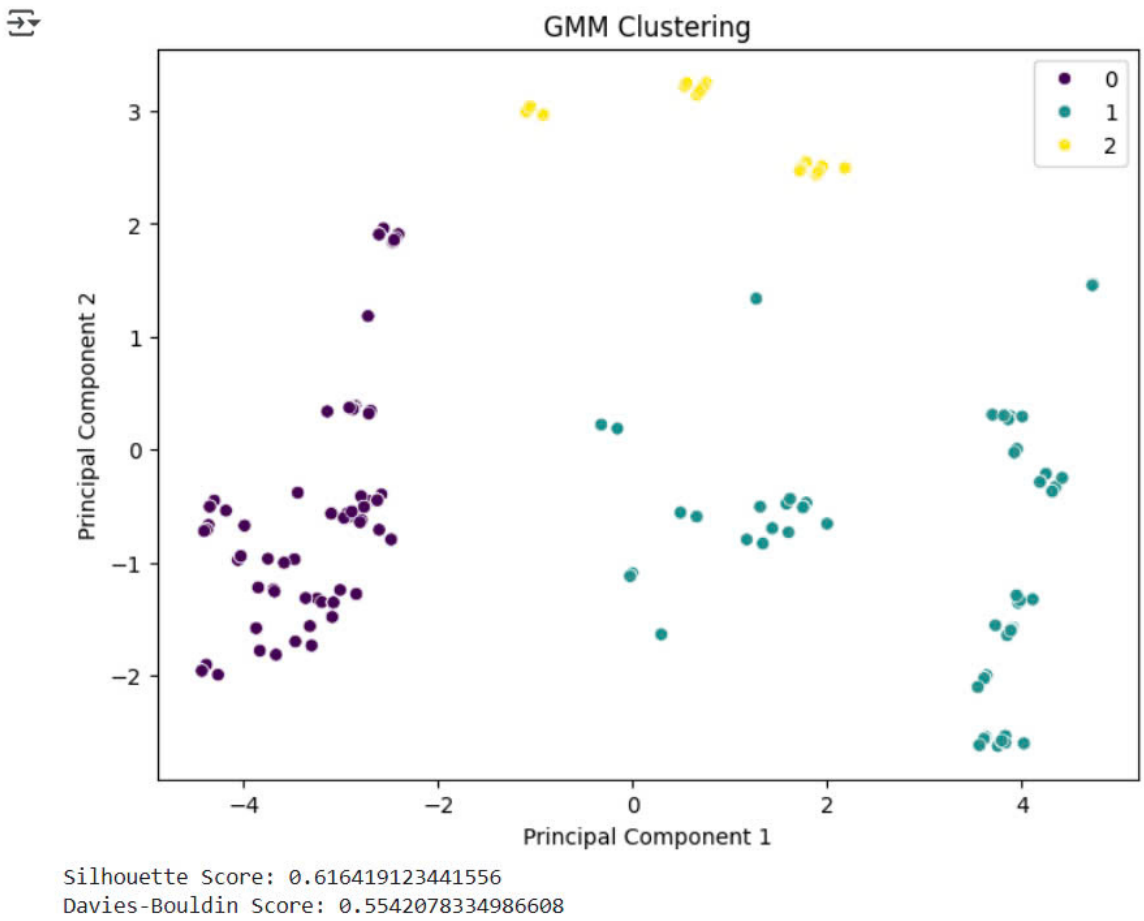
```
array([[-2.5322661 , -0.68560809],
       [-2.49923463,  1.04610398],
       [ 1.71089577,  2.8775569 ],
       ...,
       [ 1.51254993,  2.92877021],
       [ 3.35588653, -2.56456813],
       [ 1.99065749,  2.89378614]])
```

Hình 5.4: Dữ liệu được giảm chiều

- **Sử dụng PCA**: Phân tích thành phần chính (PCA) giúp giảm chiều dữ liệu từ nhiều biến số xuống hai biến số chính.
- **Biểu diễn dữ liệu trên không gian 2D**: Biểu đồ phân tán của các thành phần chính được vẽ để kiểm tra tính phân tách của dữ liệu.

**Bước 4: Áp dụng mô hình Gaussian Mixture Model (GMM) để phân cụm**

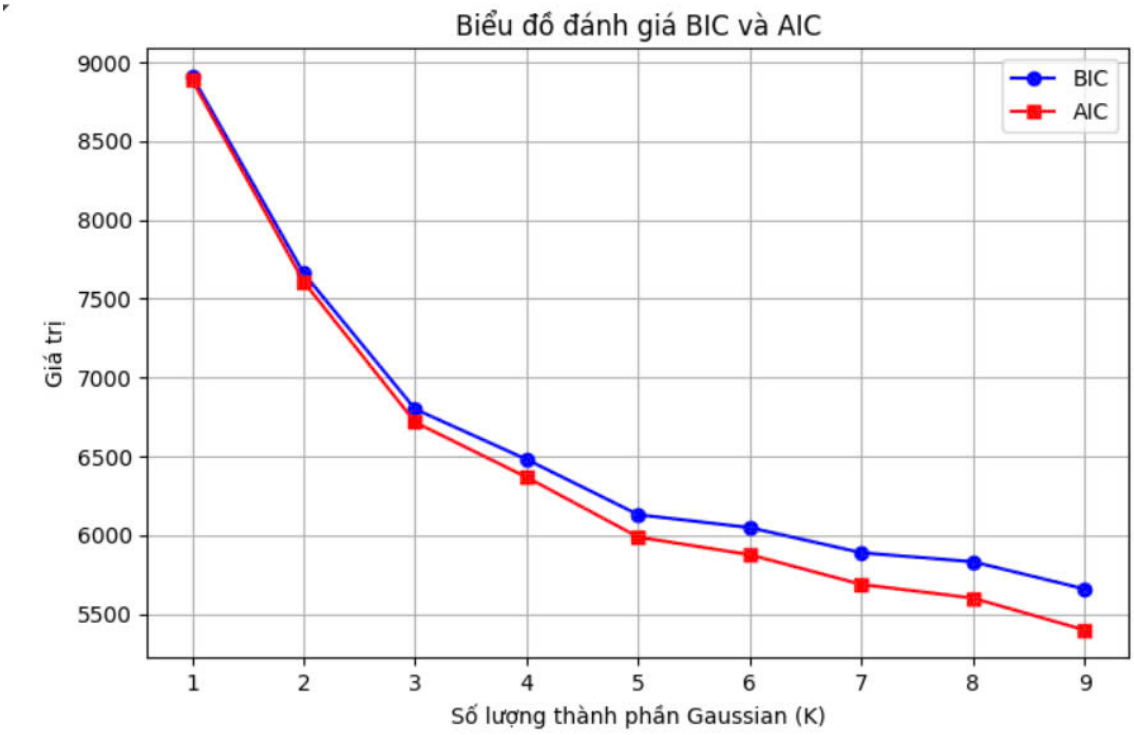
- **Huấn luyện mô hình GMM**: GMM được áp dụng để phân cụm dữ liệu thành các nhóm dựa trên sự phân bố xác suất.
- **Xác định số cụm tối ưu**: Dùng tiêu chí BIC và AIC để tìm số lượng cụm tối ưu.
- **Đánh giá mô hình**: Chỉ số Silhouette và Davies-Bouldin được sử dụng để đánh giá độ chính xác của phân cụm.




Hình 5.5: Dữ liệu được phân cụm bằng GMM

- **Silhouette Score = 0.6164:** Khá tốt. Silhouette Score dao động từ -1 đến 1, càng gần 1 thì cụm phân biệt rõ và chặt chẽ hơn. Trên 0.6 được xem là khá ổn, cho thấy các điểm dữ liệu trong cùng một cụm khá gần nhau và khác biệt rõ với các cụm khác.
- **Davies-Bouldin Index = 0.5542:** Thấp, điều này tốt. Davies-Bouldin Index càng nhỏ càng chứng tỏ cụm càng tách biệt và càng chặt chẽ. Giá trị dưới 1 thường cho thấy kết quả phân cụm khá chất lượng.

  **Tóm lại:** Với 2 chỉ số này, mô hình phân cụm của bạn rất ổn áp. Dữ liệu đang được phân cụm rõ ràng và hợp lý.




Hình 5.6: Biểu đồ đánh giá BIC và AIC

```
Silhouette Score: 0.616419123441556
Davies-Bouldin Index: 0.5542078334986608
```

Hình 5.7: Kết quả chỉ số đánh giá

## 5.2 Đánh giá Kết quả

**Kết quả phân cụm**

Sau khi áp dụng GMM với số cụm tối ưu, dữ liệu được phân chia thành các nhóm khác nhau. Biểu đồ phân tán cho thấy các cụm được hình thành một cách rõ ràng.

**Đánh giá mô hình**

- **Silhouette Score**: Chỉ số này đo lường độ tách biệt giữa các cụm. Giá trị càng cao, cụm càng tách biệt rõ ràng.
- **Davies-Bouldin Score**: Chỉ số này đánh giá mức độ tương tự giữa các cụm, giá trị càng thấp, cụm càng có sự khác biệt tốt.

## 5.3 Thảo luận các vấn đề liên quan đến kết quả mô phỏng

**Hạn chế của mô hình**

- **Phân cụm có thể bị ảnh hưởng bởi số lượng thành phần PCA**: Nếu số lượng thành phần chính được chọn không phù hợp, có thể ảnh hưởng đến chất lượng phân cụm.
- **GMM giả định dữ liệu có phân bố Gaussian**: Nếu dữ liệu không tuân theo phân bố Gaussian, mô hình có thể hoạt động kém hiệu quả.
- **Ảnh hưởng của tiền xử lý dữ liệu**: Các bước như chuẩn hóa dữ liệu hoặc xử lý giá trị thiếu có thể ảnh hưởng đến kết quả cuối cùng.

**Khả năng mở rộng và cải tiến mô hình**

- **Tích hợp thêm các phương pháp phân cụm khác**: So sánh GMM với K-Means hoặc Hierarchical Clustering để kiểm tra độ ổn định của phân cụm.
- **Tăng cường xử lý dữ liệu**: Áp dụng các phương pháp làm sạch dữ liệu mạnh mẽ hơn để giảm nhiễu.
- **Sử dụng kỹ thuật học sâu (Deep Learning)**: Thử nghiệm với Autoencoder hoặc các mô hình mạng nơ-ron để tìm kiếm biểu diễn tốt hơn của dữ liệu.

# Phần III

# KẾT LUẬN VÀ KIẾN NGHỊ

Mô hình phân cụm GMM đã giúp chia dữ liệu bệnh nhân ung thư thành các nhóm có ý nghĩa, giúp các nhà nghiên cứu và bác sĩ có cái nhìn sâu hơn về các đặc điểm của bệnh nhân. Các bước tiền xử lý dữ liệu và đánh giá mô hình đã giúp cải thiện độ chính xác và tính thực tiễn của kết quả.

Bằng cách áp dụng các phương pháp phân tích dữ liệu hiện đại, chúng tôi đã có thể tạo ra mô hình phân cụm hiệu quả, giúp hỗ trợ trong việc chẩn đoán và phân loại bệnh nhân ung thư một cách chính xác hơn

# Tài liệu tham khảo


[1] N. T. Luan, "Introduction to fuzzy clustering – soft clustering," *github*, p. 1, 2015.

[2] T. M. T. Nguyễn Văn Đạt, "Đề xuất thuật toán khuyến nghị theo phân phối dựa trên mô hình hỗn hợp gaussian," *Internet*, pp. 1–12, 2020.

[3] H. V. H. Nguyễn Văn Can, Ngô Quốc Tạo, "Nguyên cứu phương pháp đếm xe ô tô sử dụng mô hình hỗn hợp gaussian và luồng quang học." *Internet*, pp. 1–9, 2021.

[4] C. F. Gauss, "Disquisitiones arithmeticae," *Book "Disquisitiones Arithmeticae"*, 1798.

[5] //, "Lịch sử phát triển gaussian mixture model," *https://en.wikipedia.org/wiki/Mixture*$_m odel History, 2000.$

[6] I. Jairi, "Ordinary least squares and normal equations to estimate linear regression coefficients/parameters," Dec 20, 2021.

[7] @pacmannai, "Reduksi data tanpa ngilangin informasi pentingnya, emang bisa?" *https://x.com/pacmannai/status/1373967708669276*, 22 Mar, 36 tweets, 7 min read.

[8] *Internet/https://vi.wikipedia.org/wiki/Phuongsai*.

[9] F. R. Arman Melkumyan, "A sparse covariance function for exact gaussian process inference in large datasets," *https://www.ijcai.org/Proceedings/09/Papers/320.pdf*, pp. 1–7, 2016.

[10] P. Ahrendt, "The multivariate gaussian probability distribution." *Technical University of Denmark, Tech.Rep, 203.*, pp. 1–14, 2005.

[11] H. Tanaka, “"an inequality for a functional of probability distributions and its application to kac’s one-dimensional model of a maxwellian gas."stochastic processes: Selected papers of hiroshi tanaka. 2002. 83-88.” *Internet*, 2002.

[12] W. L. X. L. Y. J. Y. M. Chunhua Feng, Zhuang Liu, “Improved gaussian mixture model and gaussian mixture regression for learning from demonstration based on gaussian noise scattering,” *journal homepage: www.elsevier.com/locate/aei*, p. 1, 2025.

[13] I. M. S. B. S. Triyani Hendrawati, Aji Hamim Wigena, “Performance evaluation of aic and bic in time series clustering with piccolo method,” *The 1st International Conference on Statistics and Analytics. 2020*, pp. 1–7, 2-8-2019.

[14] N. Lawrence and A. Hyvärinen., “"probabilistic non-linear principal component analysis with gaussian process latent variable models.",” *Journal of machine learning research 6.11 (2005).*, 2005.

[15] “Source data,” *https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/data*.

[16] “Source data,”

*https://colab.research.google.com/drive/1S7oHnJOINqrSQZBqI*$_D 0Wa7mwvri4FuN?usp = sharing.$