IOSR Journal of Applied Geology and Geophysics (IOSR-JAGG)
e-ISSN: 2321–0990, p-ISSN: 2321–0982.Volume 14, Issue 2 Ser. I (Mar. – Apr. 2026), 58-67
www.iosrjournals.org

Application of GMM in fluid geochemistry of Indian hot springs using PCA factor score data

Amitabha Roy
Ex-Senior Director, Geological Survey of India

Abstract
This study applies Gaussian Mixture Models (GMMs) to the fluid geochemistry of Indian hot springs, utilizing Principal Component Analysis (PCA) factor score data. GMMs, a type of unsupervised machine learning model, represent data as a mixture of multiple Gaussian distributions, allowing for soft clustering and nuanced data interpretation. The model's parameters, including means, covariances, and weights, provide a comprehensive representation of data distributions. By interpreting the multidimensional GMM through scatter plots and confidence ellipses, this research enhances our understanding of geothermal systems and identifies patterns in the complex data. The application of GMMs in this context demonstrates the model's effectiveness in capturing complex data distributions and opens avenues for further research in fluid geochemistry and related fields. The study contributes to the existing body of research on the geothermal geochemistry of Indian hot springs, building on previous work on geostatistics, latent class clustering, and trend surface mapping. The results of this study have implications for the exploration and characterization of geothermal resources, highlighting the potential of GMMs as a valuable tool in the field of geosciences.

Keywords: *Gaussian Mixture Models, fluid geochemistry, Indian hot springs, Principal Component Analysis, geothermal systems, unsupervised machine learning*.

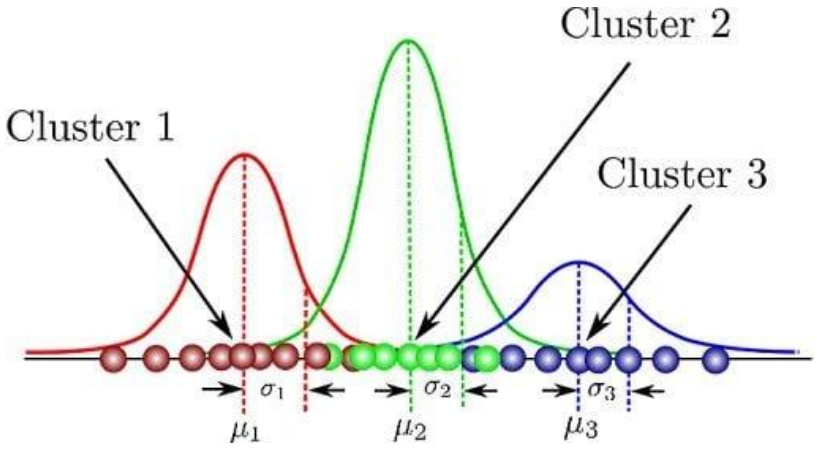
Date of Submission: 14-03-2026 Date of Acceptance: 31-03-2026

I. Introduction

A Gaussian mixture model is a soft clustering technique used in unsupervised learning to determine the probability that a given data point belongs to a cluster. It's composed of several Gaussians, each identified by $k \in \{1,\ldots, K\}$, where K is the number of clusters in a data set.

Each Gaussian *k* in the mixture is comprised of the following parameters, which are graphically illustrated as below:

- A mean μ that defines its center.
- A covariance Σ that defines its width. This would be equivalent to the dimensions of an ellipsoid in a multivariate scenario.
- A mixing probability or coefficients π that defines how big or small the Gaussian function will be.
- σ stands for standard deviation



In the above figure, there are three Gaussian functions or clusters, hence K = 3. Each Gaussian explains the data contained in each of the three clusters available. The mixing coefficients (π or **pi**) are themselves probabilities and must meet this condition.

Key Concepts of GMM

- **Probabilistic Model:** Assumes data points are generated from a mix of different Gaussian distributions.
- **Soft Clustering:** Instead of "hard" assigning a point to exactly one group (like K-Means), GMMs use soft assignment. Each point is given a probability of belonging to every cluster (e.g., 70% Cluster A, 30% Cluster B).
- **Expectation-Maximization (EM) Algorithm:** This is the iterative process used to train the model:
    1. **E-Step (Expectation):** Calculate the probability (responsibility) that each Gaussian generated each data point based on current parameters.
    2. **M-Step (Maximization):** Update the means, covariances, and weights to better fit these assigned probabilities.
    - This repeats until the model's likelihood stabilizes.

K-means clustering, Gaussian Mixture Models (GMM), and Latent Class Clustering are all unsupervised learning techniques used to group data points, but they differ fundamentally in their assumptions, how they assign data to clusters, and the shapes of the clusters they identify.

Comparison Table

Feature	K-Means	Gaussian Mixture (GMM)	Latent Class (LCA)
Type	Centroid-based (Distance)	Model-based (Probability)	Model-based (Probability)
Assignment	Hard (Binary: 0 or 1)	Soft (Probabilistic: 0-100%)	Soft (Probabilistic: 0-100%)
Cluster Shape	Spherical/Globular	Flexible (Elliptical)	Flexible (Categorical focus)
Data Type	Continuous (Numerical)	Continuous (Numerical)	Categorical/Ordinal
Speed	Fast	Slower (Iterative)	Slower
Assumption	Clusters are tight, spherical	Data is a mixture of Gaussians	Data is a mixture of multinomials

Common Uses of Gaussian Mixture Models

GMMs find application in a diverse range of fields:

- Anomaly Detection: Identifying unusual data patterns.
- Image Segmentation: Grouping pixels in images based on color or texture.
- Speech Recognition: Assisting in the recognition of phonemes in audio data.
- Handwriting Recognition: Simulating different handwriting styles.
- Customer Segmentation: Grouping customers with similar behaviors or preferences.
- Data Clustering: Finding natural groups in data.
- Computer Vision: Object detection and background removal.
- Bioinformatics: Analyzing gene expression data.
- Recommendation Systems: Personalizing user experiences.
- Medical Imaging: Tissue classification and abnormality detection.
- Finance: Asset price modeling and risk management.

II. The mathematics of GMM

Model Representation

At its core, a GMM is a combination of several Gaussian components. These components are defined by their mean vectors, covariance matrices, and weights, providing a comprehensive representation of data distributions. The probability density function of a GMM is a sum of its components, each weighted accordingly.
Notation:

- **K**: Number of Gaussian components
- **N**: Number of data points
- **D**: Dimensionality of the data

GMM Parameters:

- **Means (μ):** Center locations of Gaussian components.
- **Covariance Matrices (Σ):** Define the shape and spread of each component.
- **Weights (π):** Probability of selecting each component.

The mathematical formula for the probability density function (PDF) of a Gaussian Mixture Model **(GMM)** is a weighted sum of individual Gaussian component densities:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

The Components:

- $\boldsymbol{x}$: The data point (input vector).
- $\boldsymbol{K}$: The number of clusters or Gaussian components.
- $\boldsymbol{\pi_k}$: The **mixing weight** for the k-th component. All weights must sum to 1 ($\sum \pi_k = 1$) and be non-negative.
- $\boldsymbol{\mathcal{N}(x \mid \mu_k, \Sigma_k)}$: The k-th Gaussian distribution, defined by:

III. Application of GMM in fluid geochemistry using PCA factor score data

Posterior probabilities:

Observation	1	2	3	4	5	Cluster
Obs1	1.000	0.000	0.000	0.000	0.000	1
Obs2	1.000	0.000	0.000	0.000	0.000	1
Obs3	1.000	0.000	0.000	0.000	0.000	1
Obs4	1.000	0.000	0.000	0.000	0.000	1
Obs5	1.000	0.000	0.000	0.000	0.000	1
Obs6	1.000	0.000	0.000	0.000	0.000	1
Obs7	1.000	0.000	0.000	0.000	0.000	1
Obs8	1.000	0.000	0.000	0.000	0.000	1
Obs9	1.000	0.000	0.000	0.000	0.000	1
Obs10	0.000	0.000	0.000	1.000	0.000	4
Obs11	1.000	0.000	0.000	0.000	0.000	1
Obs12	0.999	0.000	0.000	0.001	0.000	1
Obs13	0.003	0.997	0.000	0.000	0.000	2
Obs14	0.000	0.000	0.000	1.000	0.000	4
Obs15	1.000	0.000	0.000	0.000	0.000	1
Obs16	1.000	0.000	0.000	0.000	0.000	1
Obs17	0.000	0.000	0.000	1.000	0.000	4

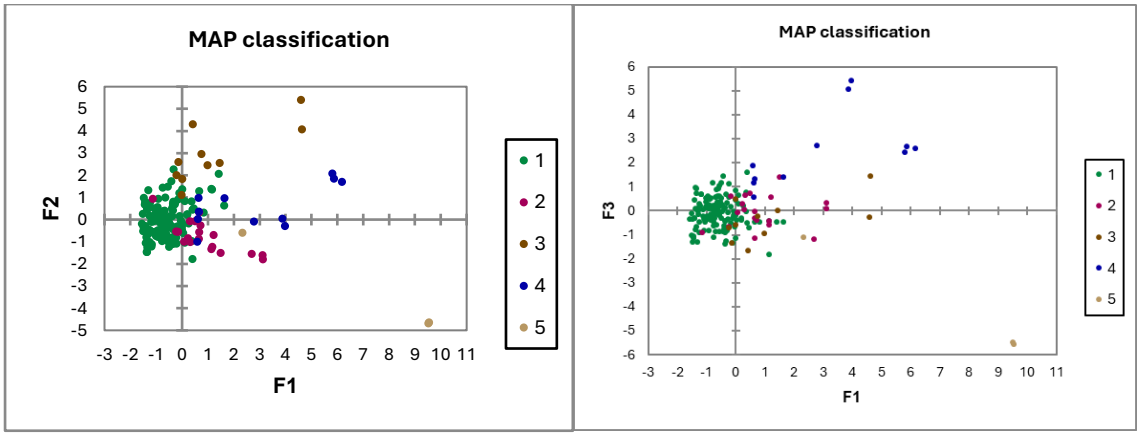
Obs18	1.000	0.000	0.000	0.000	0.000	1
Obs19	0.995	0.005	0.000	0.000	0.000	1
Obs20	0.997	0.003	0.000	0.000	0.000	1
Obs21	1.000	0.000	0.000	0.000	0.000	1
Obs22	0.998	0.002	0.000	0.000	0.000	1
Obs23	0.968	0.032	0.000	0.000	0.000	1
Obs24	0.907	0.093	0.000	0.000	0.000	1
Obs25	0.001	0.999	0.000	0.000	0.000	2
Obs26	0.586	0.414	0.000	0.000	0.000	1
Obs27	1.000	0.000	0.000	0.000	0.000	1
Obs28	1.000	0.000	0.000	0.000	0.000	1
Obs29	0.998	0.002	0.000	0.000	0.000	1
Obs30	1.000	0.000	0.000	0.000	0.000	1
Obs31	0.000	0.000	1.000	0.000	0.000	3
Obs32	0.375	0.000	0.256	0.369	0.000	1
Obs33	0.997	0.003	0.000	0.000	0.000	1
Obs34	1.000	0.000	0.000	0.000	0.000	1
Obs35	1.000	0.000	0.000	0.000	0.000	1
Obs36	0.000	0.000	1.000	0.000	0.000	3
Obs37	0.000	0.000	0.000	1.000	0.000	4
Obs38	0.000	1.000	0.000	0.000	0.000	2
Obs39	1.000	0.000	0.000	0.000	0.000	1
Obs40	1.000	0.000	0.000	0.000	0.000	1
Obs41	0.999	0.001	0.000	0.000	0.000	1
Obs42	1.000	0.000	0.000	0.000	0.000	1
Obs43	1.000	0.000	0.000	0.000	0.000	1
Obs44	1.000	0.000	0.000	0.000	0.000	1
Obs45	1.000	0.000	0.000	0.000	0.000	1
Obs46	0.000	0.000	0.000	0.000	1.000	5
Obs47	0.999	0.000	0.000	0.001	0.000	1
Obs48	0.996	0.000	0.000	0.004	0.000	1
Obs49	0.934	0.066	0.000	0.000	0.000	1
Obs50	0.994	0.006	0.000	0.000	0.000	1
Obs51	0.967	0.033	0.000	0.000	0.000	1
Obs52	0.000	0.000	0.000	0.000	1.000	5
Obs53	1.000	0.000	0.000	0.000	0.000	1
Obs54	0.000	1.000	0.000	0.000	0.000	2
Obs55	0.000	0.000	1.000	0.000	0.000	3
Obs56	0.000	1.000	0.000	0.000	0.000	2
Obs57	1.000	0.000	0.000	0.000	0.000	1
Obs58	0.000	1.000	0.000	0.000	0.000	2
Obs59	0.997	0.003	0.000	0.000	0.000	1
Obs60	0.000	1.000	0.000	0.000	0.000	2
Obs61	0.850	0.150	0.000	0.000	0.000	1
Obs62	1.000	0.000	0.000	0.000	0.000	1
Obs63	1.000	0.000	0.000	0.000	0.000	1
Obs64	0.091	0.909	0.000	0.000	0.000	2
Obs65	1.000	0.000	0.000	0.000	0.000	1

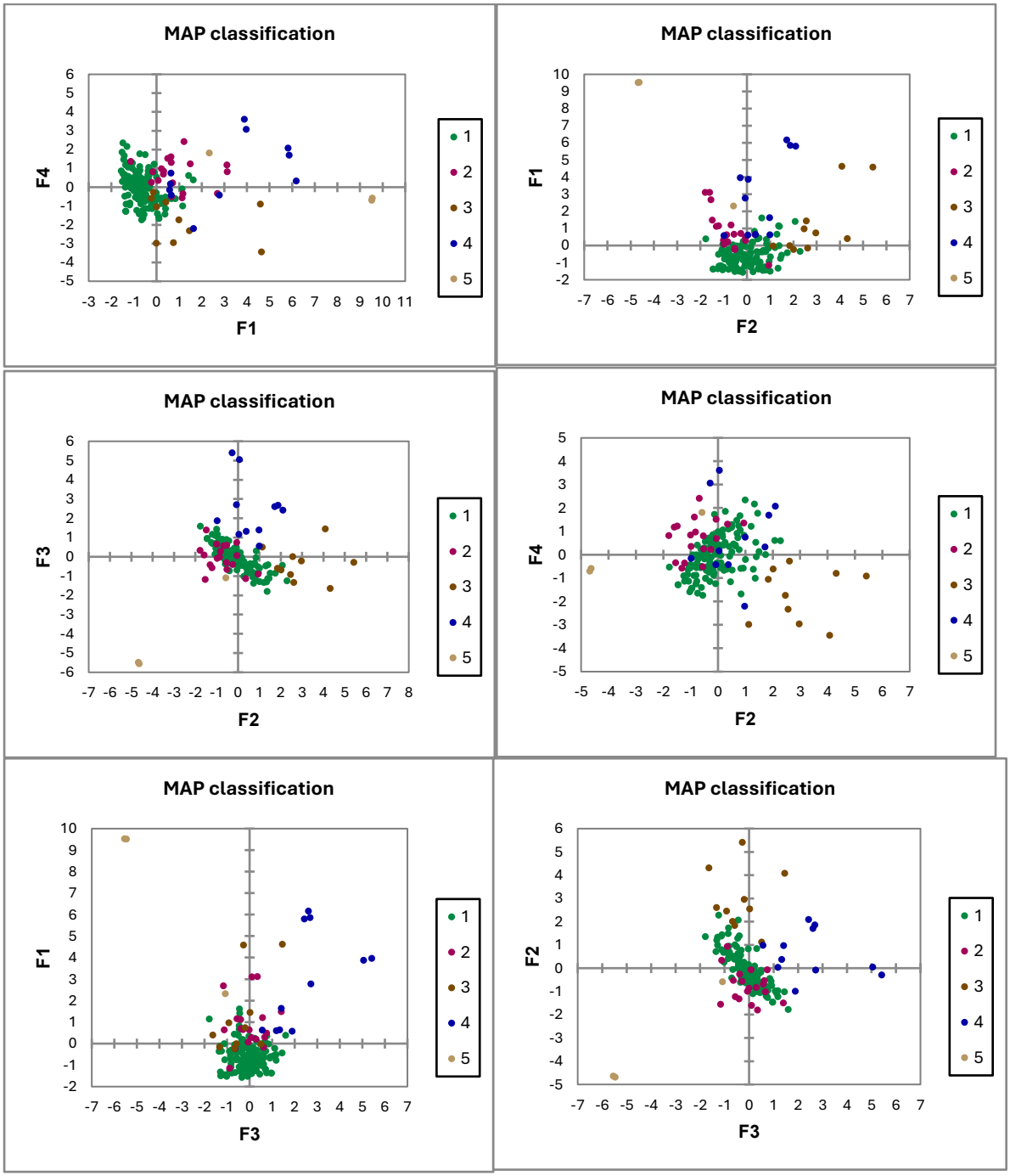
Obs66	0.999	0.001	0.000	0.000	0.000	1
Obs67	0.998	0.002	0.000	0.000	0.000	1
Obs68	0.152	0.848	0.000	0.000	0.000	2
Obs69	0.989	0.011	0.000	0.000	0.000	1
Obs70	1.000	0.000	0.000	0.000	0.000	1
Obs71	0.000	0.000	1.000	0.000	0.000	3
Obs72	0.000	0.000	0.000	1.000	0.000	4
Obs73	1.000	0.000	0.000	0.000	0.000	1
Obs74	0.000	0.000	0.998	0.002	0.000	3
Obs75	1.000	0.000	0.000	0.000	0.000	1
Obs76	0.991	0.000	0.000	0.009	0.000	1
Obs77	0.000	0.000	0.000	1.000	0.000	4
Obs78	0.998	0.000	0.000	0.002	0.000	1
Obs79	1.000	0.000	0.000	0.000	0.000	1
Obs80	1.000	0.000	0.000	0.000	0.000	1
Obs81	1.000	0.000	0.000	0.000	0.000	1
Obs82	0.999	0.000	0.000	0.001	0.000	1
Obs83	0.000	0.000	0.000	1.000	0.000	4
Obs84	1.000	0.000	0.000	0.000	0.000	1
Obs85	1.000	0.000	0.000	0.000	0.000	1
Obs86	1.000	0.000	0.000	0.000	0.000	1
Obs87	0.001	0.000	0.000	0.999	0.000	4
Obs88	0.996	0.004	0.000	0.000	0.000	1
Obs89	1.000	0.000	0.000	0.000	0.000	1
Obs90	0.998	0.002	0.000	0.000	0.000	1
Obs91	0.998	0.002	0.000	0.000	0.000	1
Obs92	0.000	0.000	0.000	1.000	0.000	4
Obs93	0.998	0.002	0.000	0.000	0.000	1
Obs94	1.000	0.000	0.000	0.000	0.000	1
Obs95	0.999	0.000	0.000	0.001	0.000	1
Obs96	0.994	0.000	0.006	0.000	0.000	1
Obs97	0.983	0.000	0.017	0.000	0.000	1
Obs98	0.999	0.001	0.000	0.000	0.000	1
Obs99	0.998	0.000	0.001	0.000	0.000	1
Obs100	0.997	0.000	0.000	0.003	0.000	1
Obs101	0.998	0.002	0.000	0.000	0.000	1
Obs102	0.999	0.000	0.000	0.001	0.000	1
Obs103	0.802	0.198	0.000	0.000	0.000	1
Obs104	0.003	0.997	0.000	0.000	0.000	2
Obs105	0.991	0.008	0.000	0.000	0.000	1
Obs106	1.000	0.000	0.000	0.000	0.000	1
Obs107	0.999	0.000	0.000	0.000	0.000	1
Obs108	0.000	0.000	1.000	0.000	0.000	3
Obs109	0.002	0.000	0.998	0.000	0.000	3
Obs110	1.000	0.000	0.000	0.000	0.000	1
Obs111	0.824	0.175	0.000	0.001	0.000	1
Obs112	0.982	0.000	0.000	0.017	0.000	1
Obs113	0.996	0.000	0.004	0.000	0.000	1

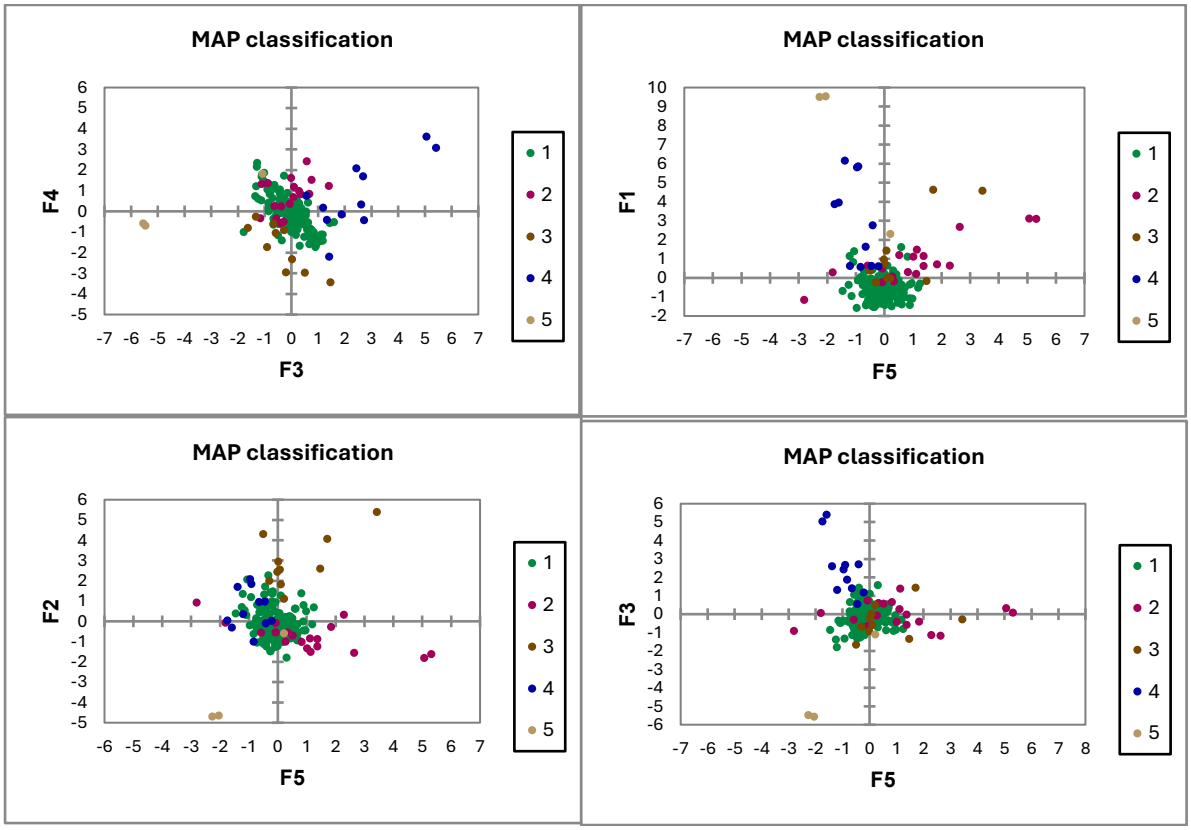
Obs114	0.999	0.000	0.000	0.001	0.000	1
Obs115	1.000	0.000	0.000	0.000	0.000	1
Obs116	1.000	0.000	0.000	0.000	0.000	1
Obs117	0.998	0.002	0.000	0.000	0.000	1
Obs118	1.000	0.000	0.000	0.000	0.000	1
Obs119	0.957	0.043	0.000	0.000	0.000	1
Obs120	0.000	0.000	1.000	0.000	0.000	3
Obs121	0.999	0.001	0.000	0.000	0.000	1
Obs122	0.997	0.003	0.000	0.000	0.000	1
Obs123	1.000	0.000	0.000	0.000	0.000	1
Obs124	0.996	0.000	0.002	0.002	0.000	1
Obs125	0.000	0.000	0.000	1.000	0.000	4
Obs126	0.922	0.000	0.000	0.078	0.000	1
Obs127	0.999	0.000	0.001	0.000	0.000	1
Obs128	0.000	0.000	0.000	1.000	0.000	4
Obs129	1.000	0.000	0.000	0.000	0.000	1
Obs130	1.000	0.000	0.000	0.000	0.000	1
Obs131	1.000	0.000	0.000	0.000	0.000	1
Obs132	1.000	0.000	0.000	0.000	0.000	1
Obs133	1.000	0.000	0.000	0.000	0.000	1
Obs134	1.000	0.000	0.000	0.000	0.000	1
Obs135	1.000	0.000	0.000	0.000	0.000	1
Obs136	1.000	0.000	0.000	0.000	0.000	1
Obs137	0.999	0.001	0.000	0.000	0.000	1
Obs138	1.000	0.000	0.000	0.000	0.000	1
Obs139	1.000	0.000	0.000	0.000	0.000	1
Obs140	0.000	0.000	0.000	0.000	1.000	5
Obs141	0.902	0.098	0.000	0.000	0.000	1
Obs142	0.000	0.000	1.000	0.000	0.000	3
Obs143	1.000	0.000	0.000	0.000	0.000	1
Obs144	1.000	0.000	0.000	0.000	0.000	1
Obs145	0.000	1.000	0.000	0.000	0.000	2
Obs146	1.000	0.000	0.000	0.000	0.000	1
Obs147	0.000	1.000	0.000	0.000	0.000	2
Obs148	0.002	0.998	0.000	0.000	0.000	2
Obs149	0.000	1.000	0.000	0.000	0.000	2
Obs150	0.225	0.775	0.000	0.000	0.000	2
Obs151	0.973	0.027	0.000	0.000	0.000	1
Obs152	0.005	0.995	0.000	0.000	0.000	2
Obs153	1.000	0.000	0.000	0.000	0.000	1
Obs154	0.000	1.000	0.000	0.000	0.000	2
Obs155	1.000	0.000	0.000	0.000	0.000	1
Obs156	0.988	0.000	0.000	0.012	0.000	1
Obs157	1.000	0.000	0.000	0.000	0.000	1
Obs158	0.990	0.010	0.000	0.000	0.000	1
Obs159	0.985	0.015	0.000	0.000	0.000	1
Obs160	0.997	0.003	0.000	0.000	0.000	1
Obs161	0.968	0.032	0.000	0.000	0.000	1

Obs162	0.989	0.011	0.000	0.000	0.000	1
Obs163	0.999	0.000	0.000	0.001	0.000	1
Obs164	1.000	0.000	0.000	0.000	0.000	1
Obs165	0.998	0.002	0.000	0.000	0.000	1
Obs166	0.981	0.019	0.000	0.000	0.000	1
Obs167	1.000	0.000	0.000	0.000	0.000	1
Obs168	1.000	0.000	0.000	0.000	0.000	1
Obs169	0.002	0.998	0.000	0.000	0.000	2
Obs170	0.999	0.000	0.000	0.000	0.000	1
Obs171	1.000	0.000	0.000	0.000	0.000	1
Obs172	1.000	0.000	0.000	0.000	0.000	1
Obs173	0.655	0.345	0.000	0.000	0.000	1
Obs174	0.236	0.764	0.000	0.000	0.000	2
Obs175	0.998	0.002	0.000	0.000	0.000	1
Obs176	0.651	0.349	0.000	0.000	0.000	1
Obs177	0.987	0.000	0.000	0.013	0.000	1
Obs178	1.000	0.000	0.000	0.000	0.000	1
Obs179	1.000	0.000	0.000	0.000	0.000	1
Obs180	1.000	0.000	0.000	0.000	0.000	1
Obs181	0.022	0.000	0.978	0.000	0.000	3

Map Classification




MAP classification
F4
F1
MAP classification
F1
F2
MAP classification
F3
F2
MAP classification
F4
F2
MAP classification
F1
F3
MAP classification
F2
F3
1
2
3
4
5



In a Gaussian Mixture Model (GMM) or any multivariate analysis, for 5-dimensional data, there are 10 unique pairwise scatter plots based on the formula: $\frac{d(d-1)}{2}$
The appearance of 12 plots typically indicates that 2 diagnostic charts (such as BIC and AIC scores) have been added to the visualization to help select the optimal number of clusters.

IV. Interpreting a multidimensional Gaussian Mixture Model (GMM)

Interpreting a multidimensional Gaussian Mixture Model (GMM) from scatter plots involves visualizing the data points alongside the known parameters—means (centroids), covariances (shapes/spread), and mixture weights (component importance)—to understand how the model has structured the data. The most effective approach uses 2D or 3D scatter plots overlaid with confidence ellipses representing each Gaussian component. Here is a step-by-step approach to interpreting GMM scatter plots:

1. Visualizing the Components
To interpret GMMs effectively, scatter plots are often enhanced with specific visual aids:

- **Cluster Centroids**: Plotting the mean vector () of each Gaussian as a distinct marker (e.g., a red 'X') identifies the "center" of each sub-population.
- **Confidence Ellipses**: Drawing elliptical contours around each center represents the covariance matrix (). These ellipses show where the majority of data points for that component are expected to fall (e.g., a 95% confidence interval).
- **Soft Labeling (Color/Hue)**: Using color to represent the "responsibility" or posterior probability of a point belonging to a cluster allows you to see overlapping regions where points have mixed membership.

2. Interpreting Geometric Shapes
The shape and orientation of the clusters in a scatter plot reveal the nature of the data's variance:

- **Spherical Clusters**: If the clusters look like circles, the model likely uses a spherical/isotropic covariance, meaning variance is equal in all directions and features are independent.
- **Elongated/Tilted Ellipses**: These indicate a full covariance matrix, where features are correlated. The orientation of the ellipse's axes shows the direction of maximum variance.
- **Axis-Aligned Ellipses**: If the ellipses are elongated but not tilted, it suggests a diagonal covariance, meaning features have different variances but no correlation.

3. Handling High Dimensionality

For data with more than three features, use these methods to generate interpretable scatter plots:

- **Scatter Plot Matrix (Pair Plots)**: Create a grid of 2D scatter plots for every pair of features. This helps identify which specific dimensions contribute most to cluster separation.
- **Principal Component Analysis (PCA)**: Reduce the data to the first two or three principal components. Plotting these captures the most variance and often makes the Gaussian "blobs" more distinct for visual inspection.

4. Evaluating Model Fit

- **Density Estimation**: Beyond simple points, use Kernel Density Estimation (KDE) or GMM probability contours to see how well the "bumps" of the model overlap with actual data concentrations.
- **Overlap and Ambiguity**: Scatter plots excel at showing "fringe" observations—points in the middle of two clusters—which highlight the GMM's ability to handle uncertainty through soft assignments.

Summary of GMM Interpretation

Visual Feature	GMM Parameter	Interpretation
Center of ellipse	Mean ()	Typical value for the cluster
Spread of ellipse	Covariance ()	Correlation & Variance of features
Color/Group Size	Weight ()	Relative prevalence of the cluster
Overlap of ellipses	Posterior Probability	Uncertainty in classification

V. Conclusion

In conclusion, the application of Gaussian Mixture Models (GMM) in the fluid geochemistry of Indian hot springs demonstrates the model's effectiveness in capturing complex data distributions through soft clustering, thereby enhancing our understanding of geothermal systems. This approach not only facilitates nuanced data interpretation but also opens avenues for further research in various scientific fields.

References

[1] Amitabha Roy, 2023. Geostatistics as applied to fluid geochemistry of Indian hot springs. J. Appl. Geol. & Geophys (ISOR-JAGG), V.11, Issue 4, Ser. II, pp. 01-37

[2] Amitabha Roy, 2024. Geostatistics applied to Fluid Geochemistry of Geothermal Fields in Peninsular and Extra- Peninsular India. White Falcon Publishing, Chandigarh, India, pp. 1- 142.

[3] Amitabha Roy, 2023. Latent class clustering and profile analysis to uncover hidden patterns within the mixed observed distribution of multivariate fluid geothermal geochemistry data in Indian hot springs. J. Appl. Geol. & Geophys (ISOR-JAGG), V.13, Issue 2, Ser. I pp. 21-26

[4] Amitabha Roy[1] , Sk Mafizul Haque[2] , 2025. Trend Surface Mapping of Principal Component Factor Scores to Assess the Geothermal Geochemistry of India's Peninsular and Extra-Peninsular Hot Springs. J. Appl. Geol. & Geophys (ISOR-JAGG), V.13, Issue 3, Ser. I pp. 51-62

[5] Weiguo Lu et. al., 2025. An efficient Gaussian mixture model and its application to neural networks. Knowledge-Based Systems Volume 310