Investigating the Clustering Availability of Mixtures of Low-Rank Gaussian (MoLRG) Distributions

Ruixuan Deng
Department of Statistics
University of Michigan, Ann Arbor

1 Introduction

In recent years, Gaussian mixture models (GMMs) have emerged as fundamental data analysis tools for clustering and density estimation. These models naturally capture the heterogeneous nature of complex data through their multi-modal structure. However, as data dimensionality increases, GMMs face significant challenges in parameter estimation, particularly in modeling the covariance structure. A standard GMM with full covariance matrices requires $O(d^2)$ parameters per component for d-dimensional data, leading to computational intractability and potential overfitting in high dimensions.

Low-rank Gaussian distributions offer an elegant solution to this challenge. By constraining the covariance matrices to have low-rank structure, these models can efficiently capture high-dimensional dependencies while dramatically reducing the number of required parameters. This approach aligns with the empirical observation that real-world high-dimensional data often exhibits intrinsic low-dimensional structure, making low-rank Gaussian distributions particularly suitable for modeling such data.

This study investigates a fundamental question in the application of low-rank Gaussian mixtures: **How does the complexity of the covariance structure—characterized by the number of components, rank constraints, and noise levels—affect clustering performance?** Understanding these relationships is crucial for developing robust algorithms and providing practical guidance for model selection in applications ranging from computer vision to genomics.

2 Methodology

2.1 Model Formulation

We say that a d-dimensional random vector $X \in \mathbb{R}^d$ following a mixture of low-rank Gaussian distributions (MoLRG) with K components if:

$$X \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{U}_k\mathbf{U}_k^T + \sigma^2\mathbf{I}) \tag{1}$$

where $\mathbf{U}_k \in \mathbb{R}^{d\times r}, r \ll d$ denotes the orthonormal basis of the k-th component and $\pi_k \geq 0$ is the mixing proportion of the k-th mixture component. Term $\sigma^2\mathbf{I}$ accounts for isotropic noise.

2.2 Experimental Setting

For exact control over the complexity of data distributions, we generate synthetic data sets in various settings:

Preprint. Under review.

- Number of Components K: {5, 10, 15}
- Rank Structure r:{low(all 5), high(all 20), Mixed}
- Noise Scale σ: {0, 0.1, 0.3}

We then implement and compare four distinct clustering approaches (K-Means, hierarchical clustering, GMM and spectral clustering) based on Adjusted Random Index (ARI) over the synthetic datasets. Each dataset contains 1000 data samples and each clustering experiment is repeated 30 times to ensure the robustness of the results.

3 Results and Discussion

Our experimental analysis reveals several key findings about the performance of different clustering methods under varying conditions, as measured by the Adjusted Rand Index (ARI). Hierarchical clustering and K-means demonstrate robust performance under various circumstances, while GMM and spectral clustering generally underperform. Notably, GMM shows superior performance only in the mixed rank configuration, achieving an ARI of approximately 0.25 and substantially outperforming other methods.

The robust performance of hierarchical clustering and K-means can be attributed to several factors. First, these methods rely on distance-based metrics rather than probabilistic assumptions about the underlying data distribution. In high-dimensional settings, this simplicity becomes an advantage as these methods are less sensitive to the curse of dimensionality. Second, both methods make minimal assumptions about cluster shape and structure, allowing them to adapt to various data configurations. K-means' centroid-based approach and hierarchical clustering's progressive merging strategy provide natural safeguards against the instabilities that can arise in high-dimensional spaces.

The generally poor performance of GMM, except in mixed rank scenarios, reveals interesting insights about model complexity and estimation challenges. GMM's superior performance with mixed rank configurations suggests that it can capture imbalanced covariance structures when they exist. However, its poor performance in other scenarios likely stems from difficulties in estimating high-dimensional covariance matrices and the challenge of maintaining orthogonality constraints. While we considered implementing the EM algorithm for parameter estimation, the convergence challenges in high-dimensional settings made this impractical.

Spectral clustering's consistent underperformance is also noteworthy. Despite its theoretical advantages in capturing non-linear cluster structures, the method struggles in our high-dimensional, low-rank setting. This may be due to the degradation of pairwise distances in high dimensions, which affects the quality of the constructed similarity matrix. Additionally, the eigendecomposition step becomes less reliable as dimensionality increases, potentially obscuring the true cluster structure.

Our findings suggest that in high-dimensional settings with low-rank structure, simpler, distance-based clustering methods may be more reliable than their more sophisticated counterparts. The success of hierarchical clustering and K-means points to the importance of algorithmic stability over model complexity, particularly when dealing with high-dimensional data that exhibits low-rank characteristics.

A Appendix

A.1 Code Repository

The complete implementation of this analysis is available at `https://github.com/luckyeric320/STATS_506`.

A.2 Additional Figures

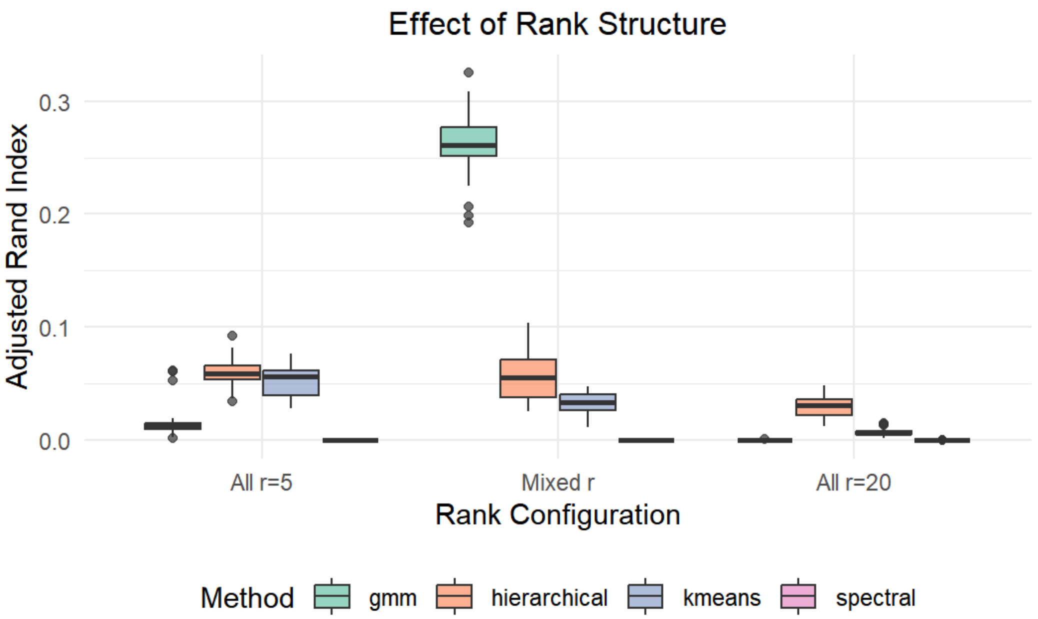

Figure 1: Effect of Rank Structure on Clustering Performance. The plot shows the Adjusted Rand Index for different clustering methods across various rank configurations (All r=5, Mixed r, All r=20).

A.3 Attribution of Sources

In accordance with the course requirements, I acknowledge the following sources for this project:

- Statistical computing software: R version 4.3.0
- Clustering implementations from standard R packages
- Claude Sonnet for polishing my writing and assisting me on writing code for experiment.

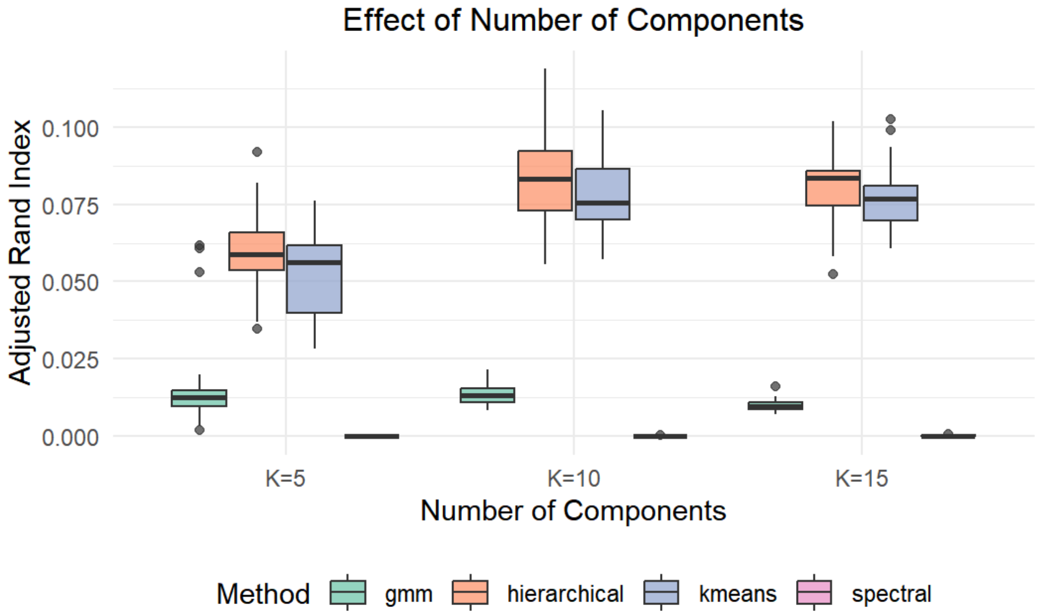

Figure 2: Effect of Number of Components on Clustering Performance. The plot demonstrates how different clustering methods perform as the number of components (K) varies from 5 to 15.

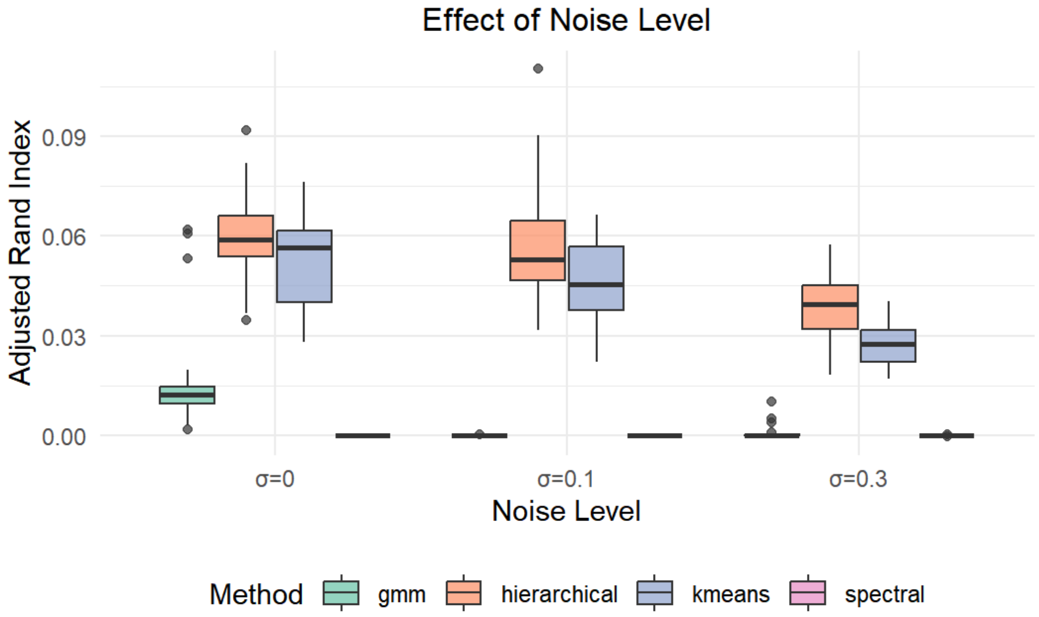

Figure 3: Effect of Noise Level on Clustering Performance. The plot illustrates the impact of different noise levels ($\sigma = 0, 0.1, 0.3$) on clustering performance across methods.

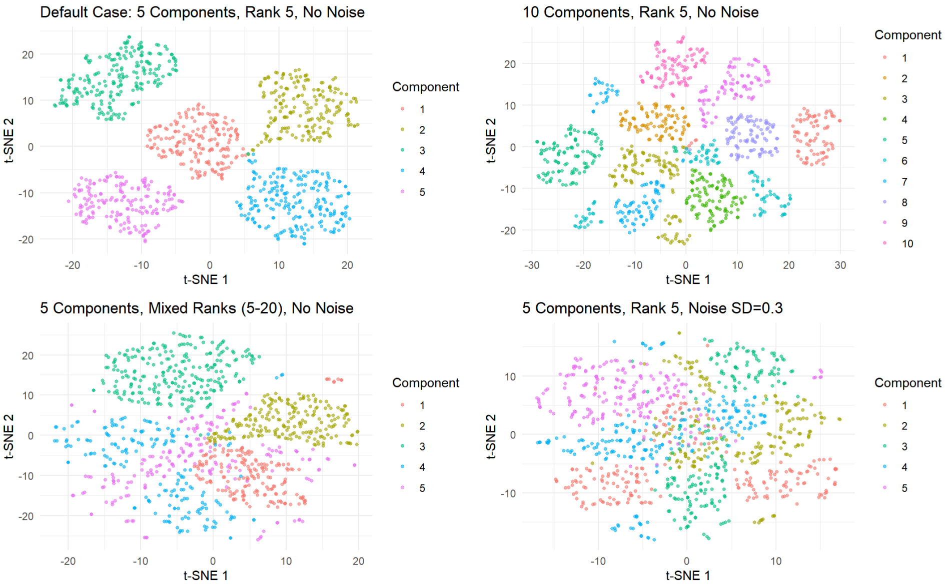

Figure 4: t-SNE of Synthetic Datasets. The plot illustrates the impact of different parameter settings over the dataset.