genMPGMM: A synthetic data generator with controlled feature partitions and cluster similarity for evaluating pattern discovery methods

Karolina Widzisz[1], Joanna Zyla[2], Mateusz Kania[3], Andrzej Polanski[1]

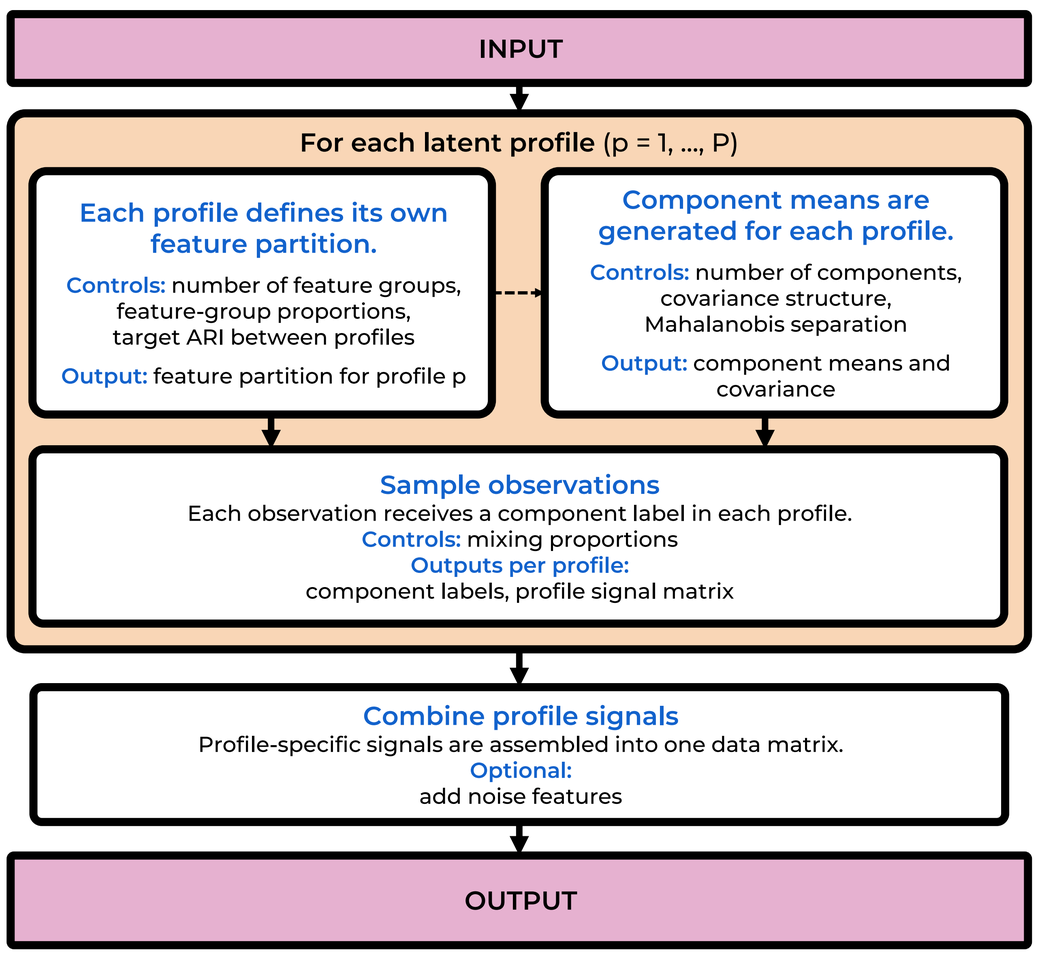
[1] Department of Computer Graphics, Vision and Digital Systems, Silesian University of Technology Gliwice, Poland
[2] Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland
[3] Department of Applied Informatics, Silesian University of Technology, Gliwice, Poland.

Motivation

High-dimensional data may contain several overlapping latent structures, such as profile-specific feature groups and observation clusters. Since real datasets rarely provide complete ground truth, evaluating clustering, biclustering and multi-view methods remains challenging. genMPGMM provides controlled synthetic Gaussian mixture data with known feature partitions, mixture components, predefined ARI similarity and adjustable Mahalanobis separation, enabling reproducible benchmarking of pattern discovery methods.

genMPGMM(...)

Parameter	Controls
`n_feature_patterns`	No. of feature groups per profile
`n_components`	No. of mixture components per profile
`feature_group_proportions`	Sizes of feature groups
`mixing_proportions`	Component proportions
`target_ari`	Similarity between feature partitions
`dist_mahalanobis`	Component separation
`noise_feature_fraction`	Proportion of noise features



Output	Description
`X`	Synthetic data matrix with features as rows and observations as columns
`z`	Component assignments for observations in each profile
`s`	Feature-group assignments for each profile
`Mu`	Component mean vectors for each profile
`Sigma`	Component covariance matrices for each profile
`achieved_ari_features`	Achieved similarity between feature partitions
`achieved_mahalanobis`	Achieved separation between mixture components
`achieved_proportions`	Achieved feature-group and component proportions

The generated **`MPObject`** can be inspected with **`summary()`** for numerical diagnostics and **`plot()`** for visual diagnostics of the simulated latent structure. For reproducible documentation, **`report_MPObject()`** exports a report combining the key settings, summary tables and diagnostic plots.

Silesian University of Technology

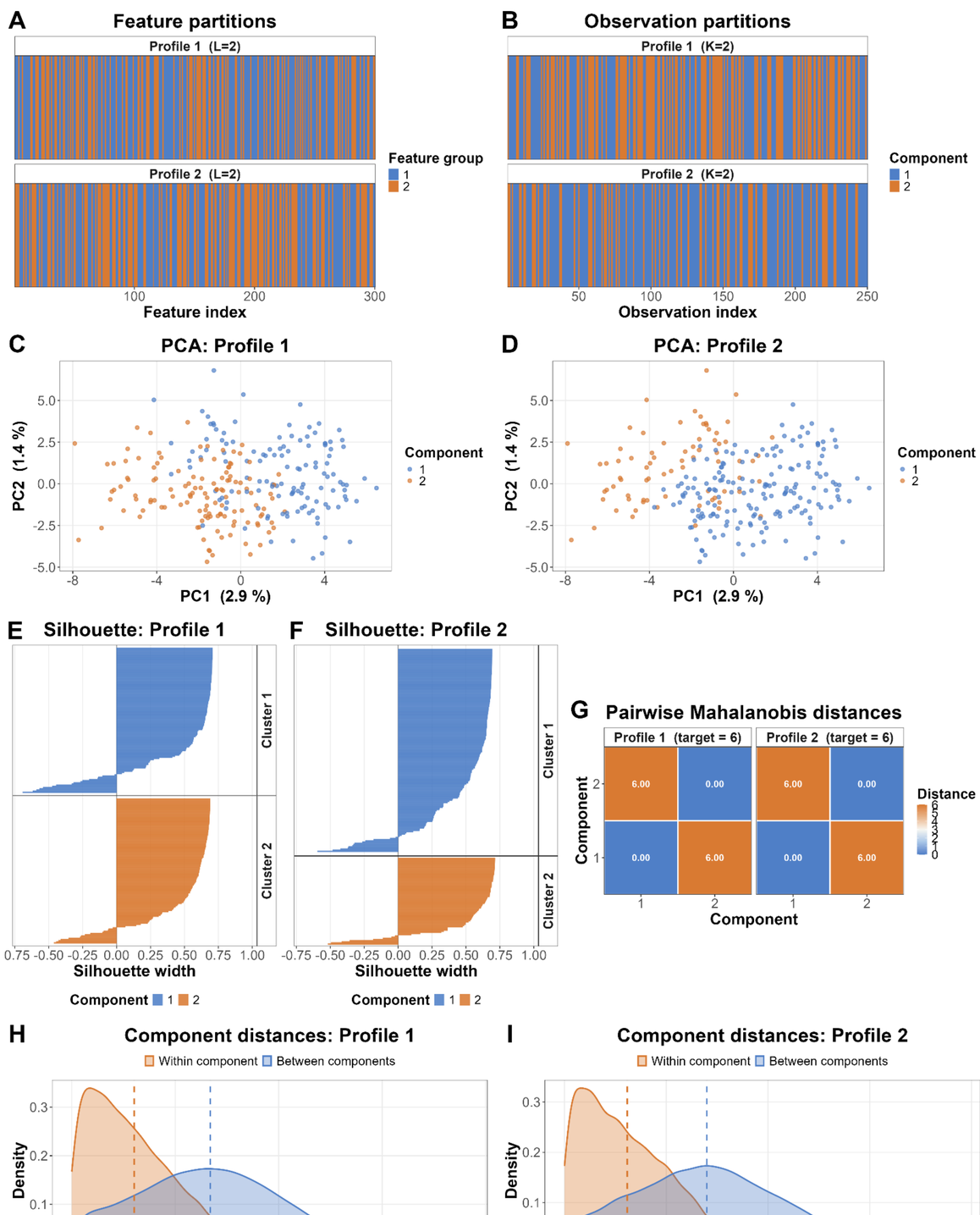
RESEARCH UNIVERSITY EXCELLENCE INITIATIVE
Ministry of Science and Higher Education

Acknowledgements

This study was supported by Silesian University of Technology grant for maintaining and developing research potential 02/090/BKM26/0069

Results

The example dataset contains M = 300 features and N = 250 observations from two latent profiles. Both profiles include two feature groups and two Gaussian mixture components. Feature-group proportions are 0.6/0.4 and 0.5/0.5, while component mixing proportions are 0.5/0.5 and 0.7/0.3 for Profiles 1 and 2, respectively. Feature-partition similarity is controlled with target ARI = 0.40, and component separation of Mahalanobis distance = 6. Diagonal covariances, no noise features and seed = 42 were used.



Conclusion

genMPGMM provides a flexible framework for generating synthetic multi-profile datasets with known ground truth. The package allows users to control key aspects of the simulation, including feature partitions, mixture components, component proportions, Mahalanobis separation, and noise level. Together with numerical summaries, diagnostic plots, and reproducible reports, it supports systematic benchmarking of clustering, biclustering, and partition-comparison methods under well-defined simulation settings.

Scan to access the package repository

GitHub