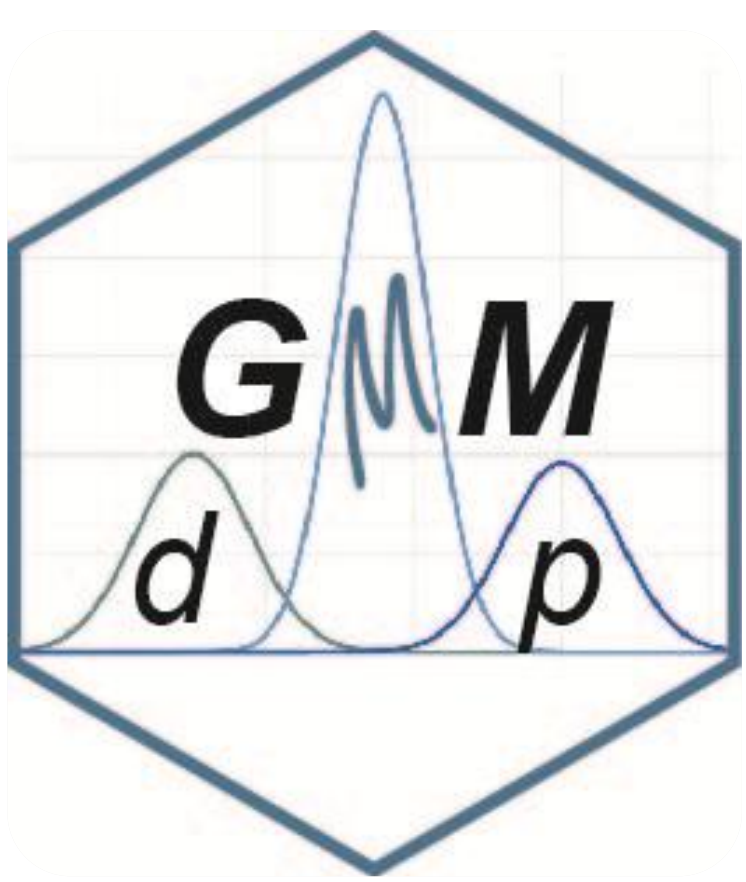




The *dpGMM* package for stable initialization in Gaussian Mixture Modeling

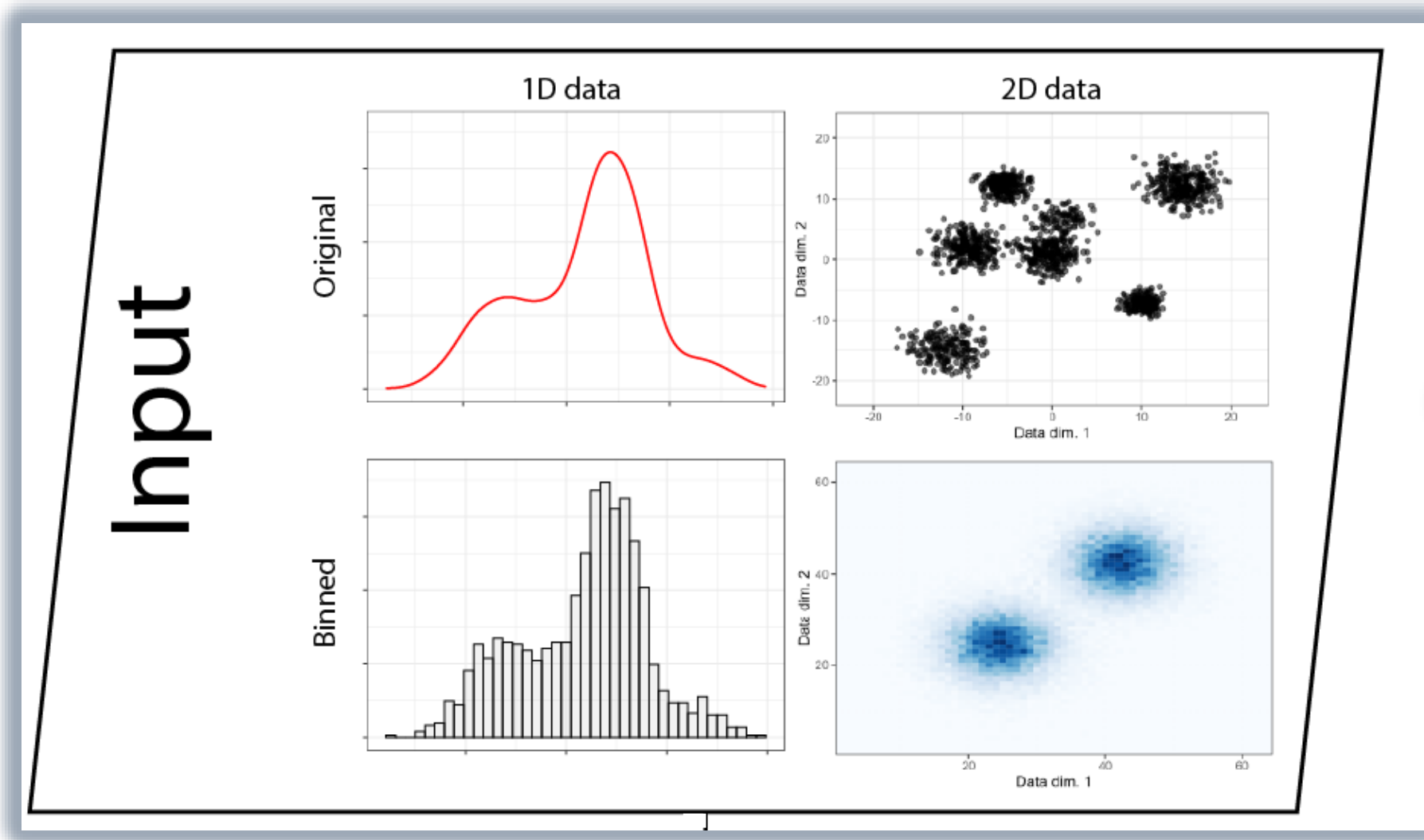
JOANNA ZYLA¹, KAMILA SZUMALA², MICHAŁ MARCZYK¹

¹ Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland

² Department of Systems Biology and Engineering, Silesian University of Technology, Gliwice, Poland

INTRODUCTION

Gaussian Mixture Modeling (GMM) is a one of unsupervised techniques used in many fields of data analysis, such as bioinformatics, pattern recognition, and network traffic analysis. Yet, existing R implementations often lack support for binned data (commonly observed in image analysis) and suffer from initialization instability or massive memory usage. To address these limitations, the novel R package *dpGMM* was introduced. It is specifically designed for robust decomposition of 1D and 2D data, supporting both continuous and binned encodings. The core implementation utilizes the Expectation-Maximization (EM) algorithm, accompanied by a deterministic dynamic programming (DP) initialization strategy based on the Bellman recurrence. For 1D data, the DP algorithm partitions the histogram to accurately estimate initial weights, means, and variances.

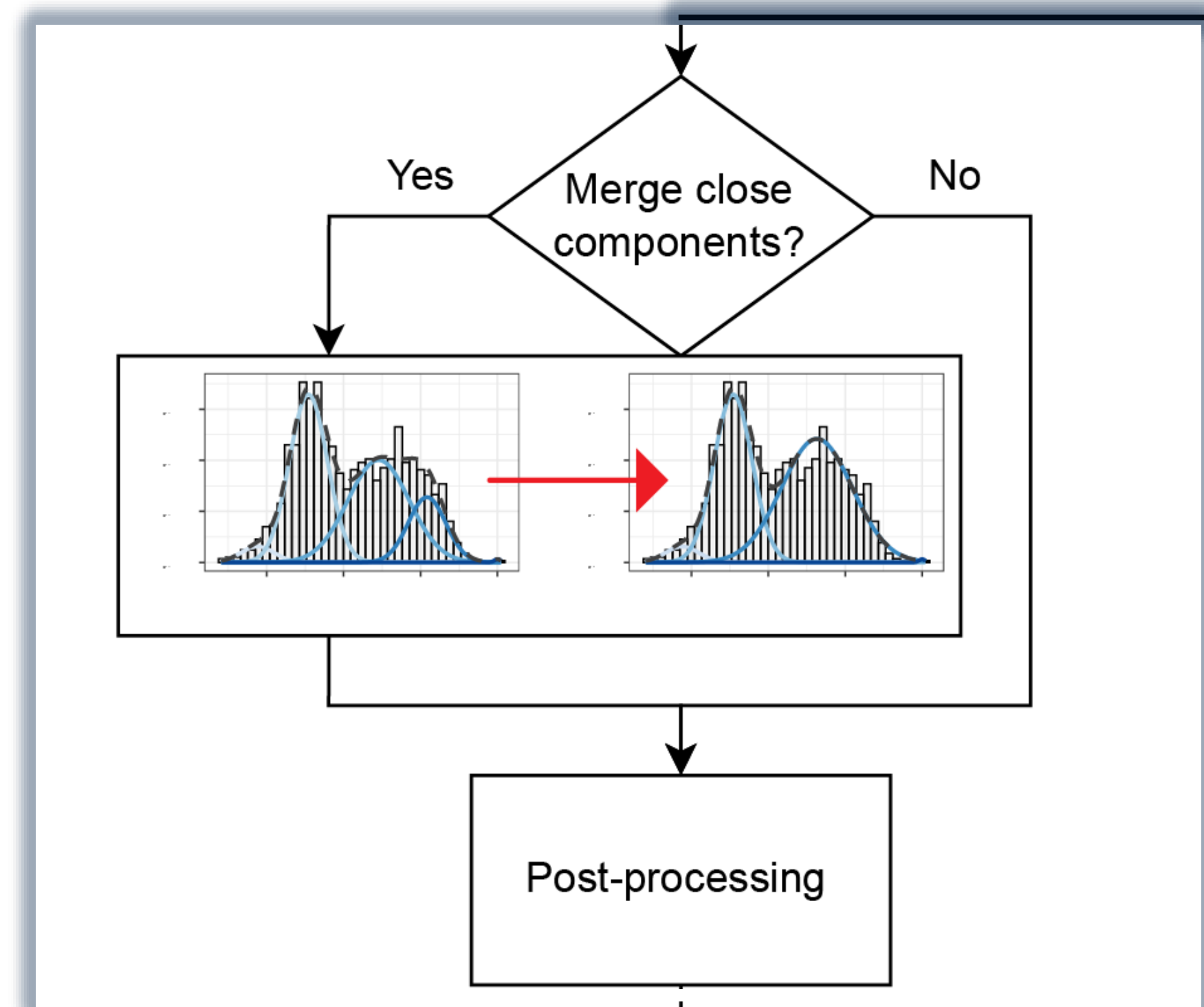
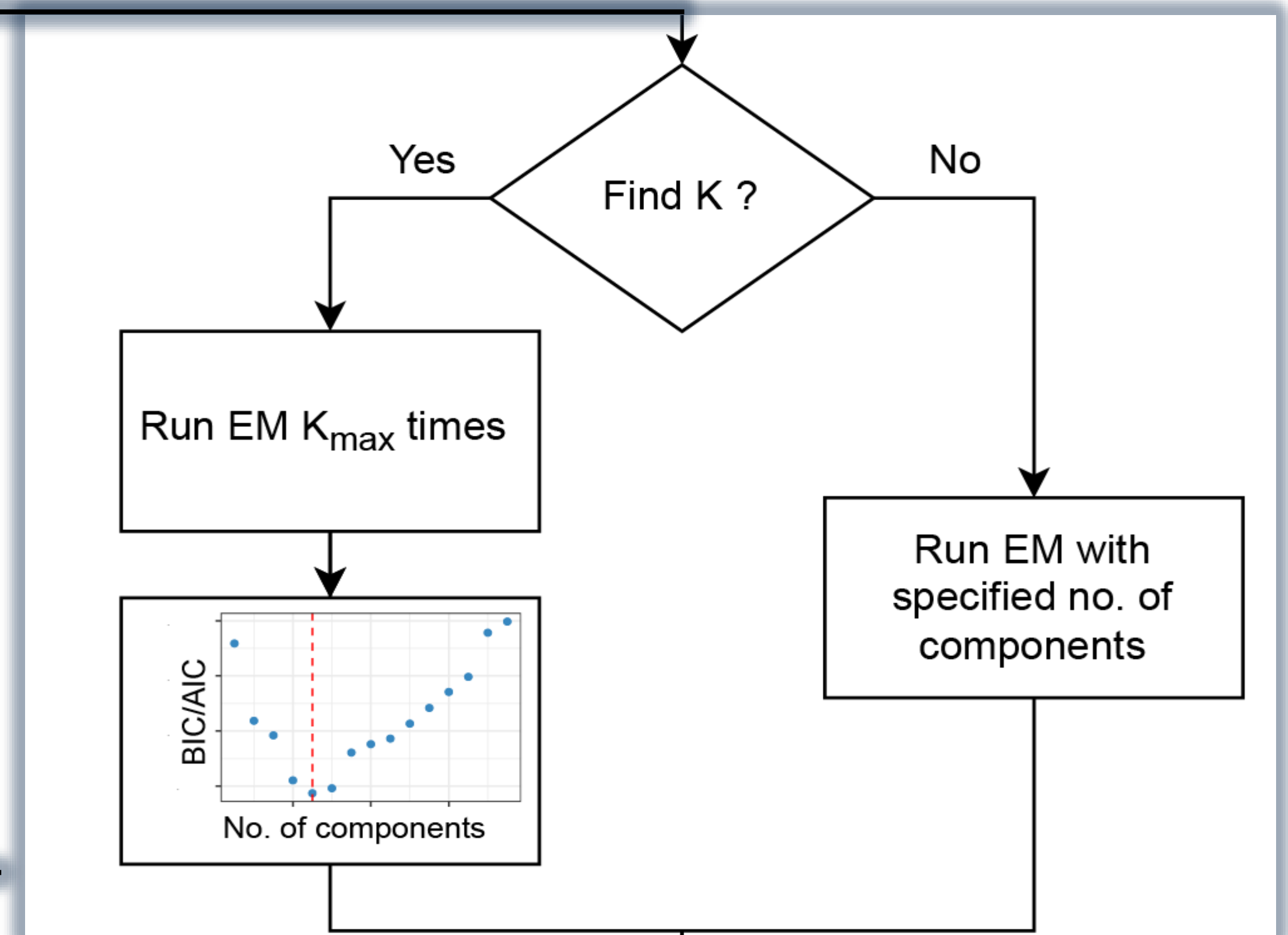


- ✓ 1D and 2D continuous data
- ✓ 1D and 2D binned data, such as histograms useful for analyzing spectra in mass spectrometry or medical image processing.

HEART OF ALGORITHM

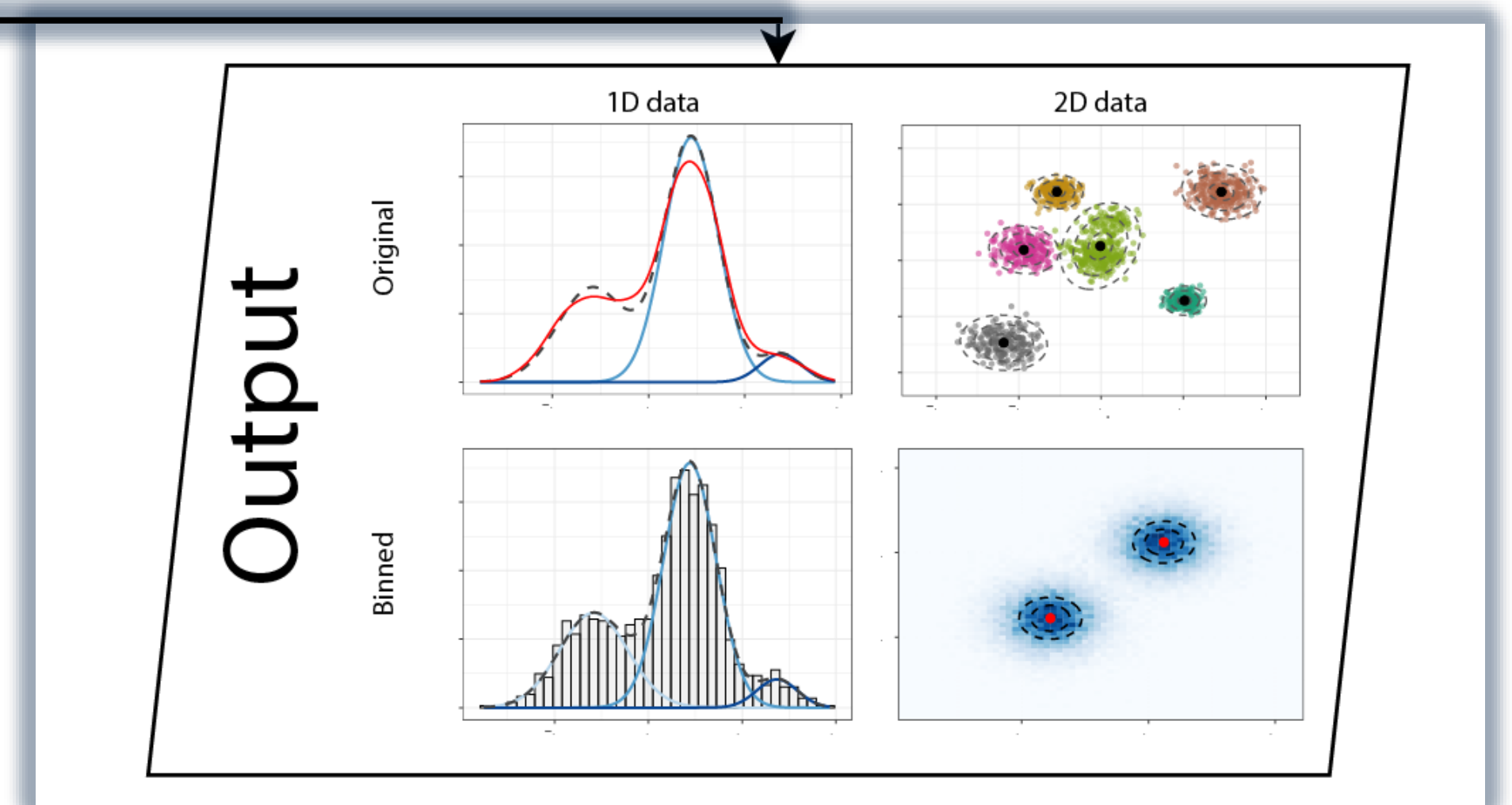


- ✓ Wrapper implementation with **K- search** by 5 various metrics e.g. AIC or BIC.
- ✓ Quick stop option (LR test)
- ✓ Mechanism for **reduction of „pin” component**
- ✓ **EM** algorithm with stable initialization based on **histogram partition** that speed up convergence in the **Bellman recurrence** (Dynamic programming part)
- ✓ **Fixed no. of components** option



- ✓ Merging close components by σ distance
- ✓ Threshold estimation and cluster assignment by Bayes-optimal decision

- ✓ Decomposition visualization by ggplot + return of figure as parameter
- ✓ QQ-plot for decomposition control



EVALUATION

The proposed *dpGMM* package was evaluated against existing tools like ClusterR, mclust, and mixtools using various synthetic datasets. The evaluation included two main scenarios featuring mixtures with both strictly controlled (OV: 0.1, 0.25 and 0.4) and randomly assigned overlap coefficients. The clustering performance for these datasets with known labels was measured using the Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Maximum-Match Measure (MMM), and Cluster Number Deviation (CND). In Figure 1, the comprehensive ranking made on median metric values is presented.

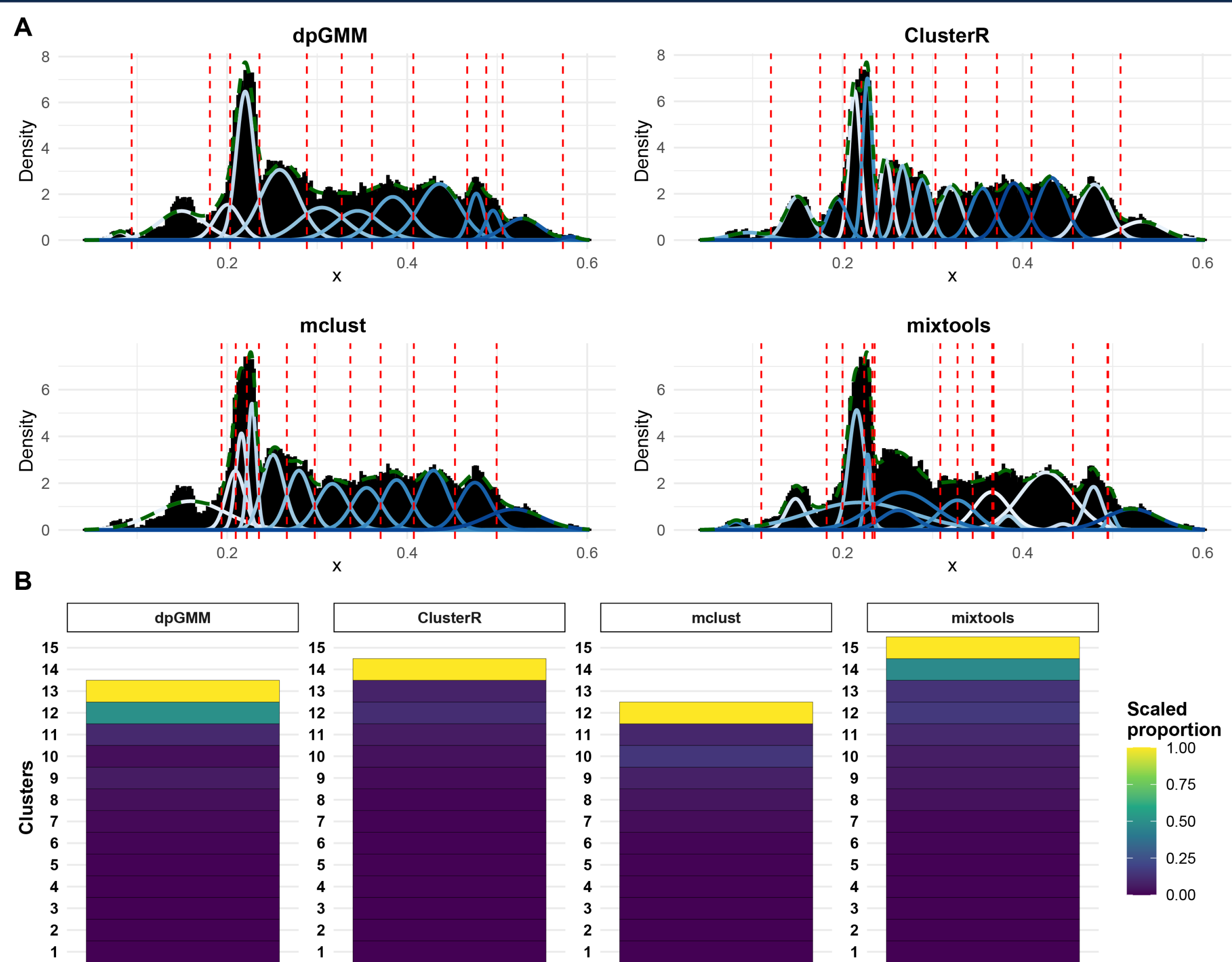


Fig. 2. Results of testing scRNA-Seq dataset. Panel A shows GMM decompositions with threshold values from each tested packages. Panel B shows the proportion of GAM cells in each cluster given where the more yellowish color indicates a higher number of GAM cells. It is expected that in the last clusters this number will be higher.

Ranking of methods

1	1	3	1	2.5	2.5	3	1.5	2	2	3	2	1	1	3	2	31.5	dpGMM	
2	2	1	3	2.5	2.5	3	3.5	2	2	4	3	2	3	2	3	40.5	mclust	
3	3	2	4	2.5	2.5	3	1.5	2	2	2	1	3	2	4	4	41.5	mixtools	
4	4	4	2	2.5	2.5	1	3.5	4	4	1	4	4	4	1	1	46.5	ClusterR	
Total																		
ARI 0.1	ARI 0.25	ARI 0.4	ARI Uncon.	[CND] 0.1	[CND] 0.25	[CND] 0.4	[CND] Uncon	MMM 0.1	MMM 0.25	MMM 0.4	MMM Uncon	NMI 0.1	NMI 0.25	NMI 0.4	NMI Uncon.			

Fig. 1. Heatmap of ranking values for clustering performance metrics calculated based on median values. The lower the value, the better the outcome.

scRNA-SEQ DATA

Next, figure 2 visualizes the performance of the tested GMM packages on a large scRNA-Seq dataset comprising over 80,000 cells. Panel A demonstrates that mclust and mixtools produce suboptimal, overlapping, or poorly fitted decompositions, whereas ClusterR and *dpGMM* appear much more coherent. However, Panel B proves the superior performance of *dpGMM*, as it uniquely and successfully identifies two distinct clusters with a high concentration of GAM cells in the right tail of the distribution.

`install.packages("dpGMM")`



Silesian University
of Technology



KIAED
KATEDRA INŻYNIERII I ANALIZY EKSPLORACYJNEJ DANYCH

CRAN
The Comprehensive R
Archive Network



GitHub

