

Supplementary Material for
A clustering framework for skewed features with low
true cluster separation

Muntazir Mehdi¹, Haydar Demirhan^{1*}, Sona Taheri¹

¹School of Science, Mathematical and Geospatial Sciences, RMIT
University, Melbourne, Australia.

*Corresponding author(s). E-mail(s): haydar.demirhan@rmit.edu.au;
Contributing authors: muntazir.mehdi22@gmail.com;
sona.taheri@rmit.edu.au;

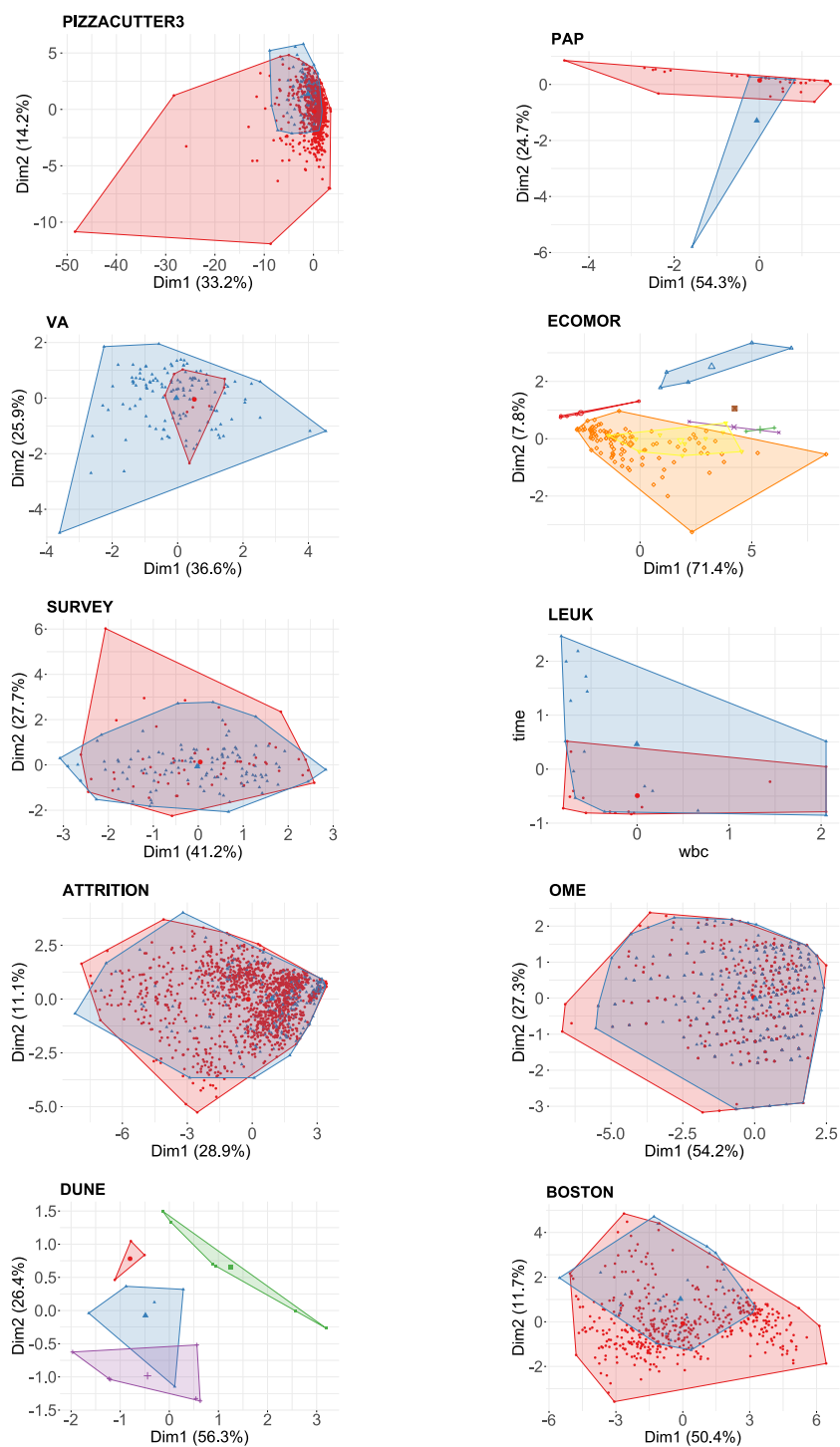


Figure S1. True clustering for benchmarking datasets.

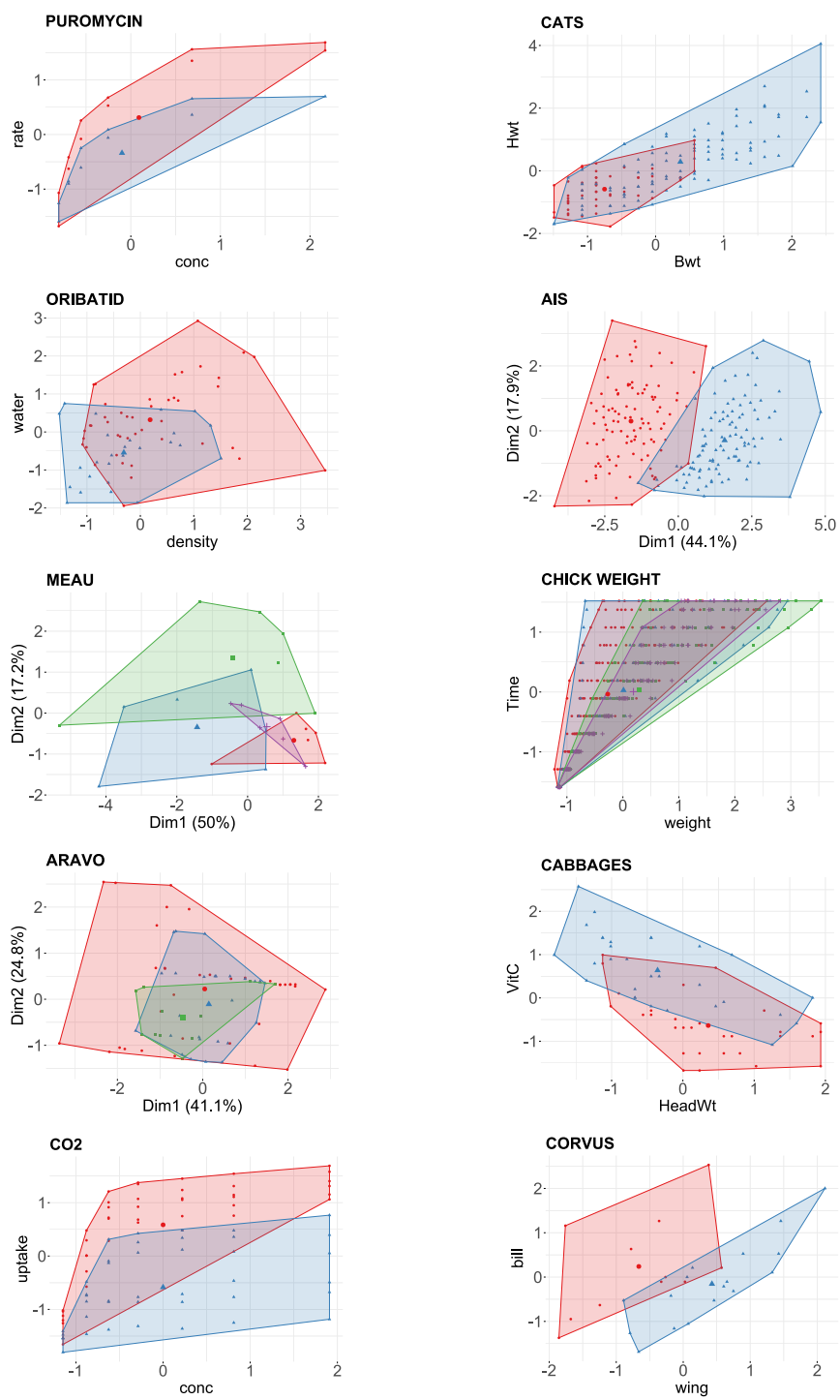


Figure S2. True clustering for benchmarking datasets.

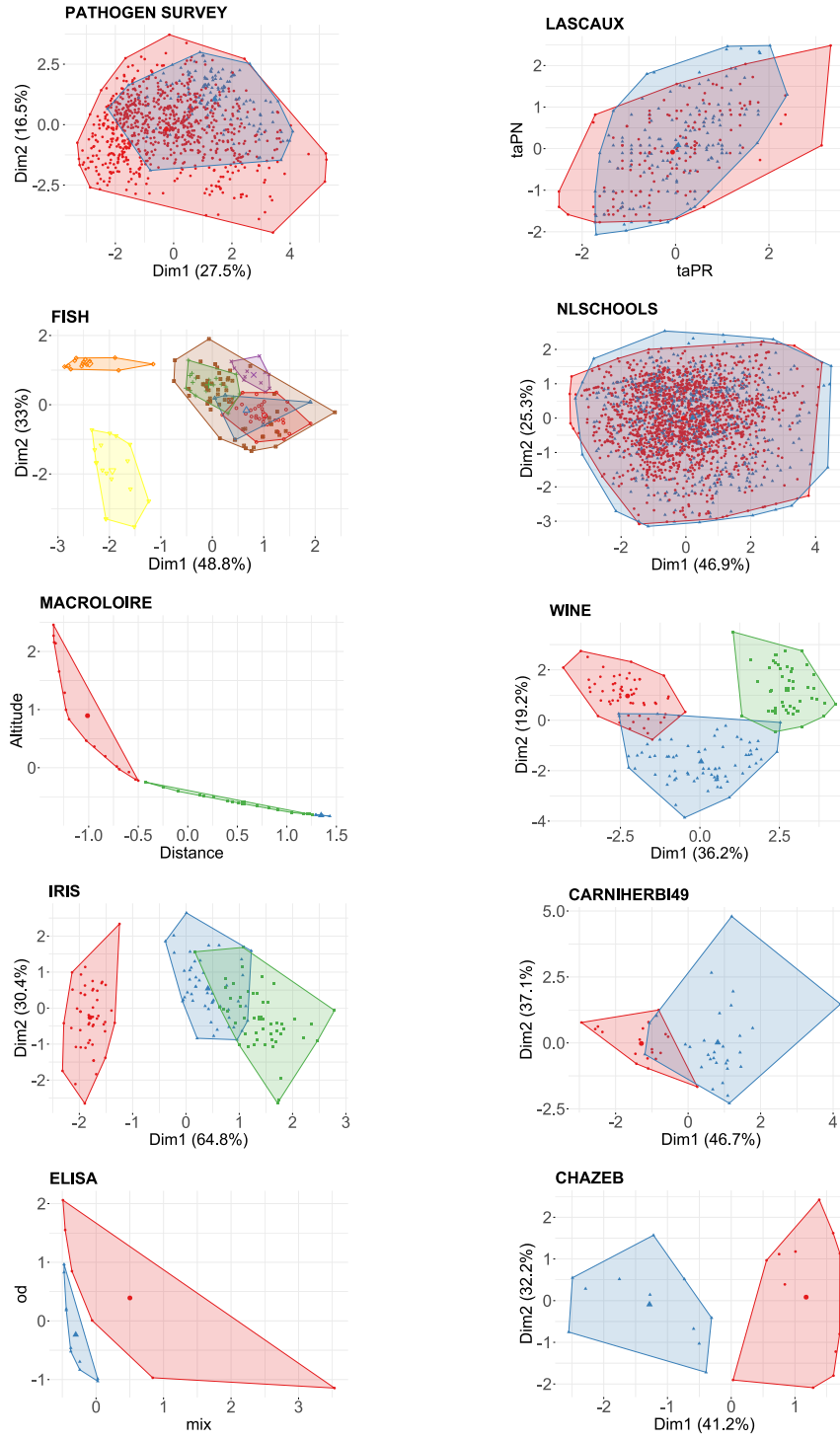


Figure S3. True clustering for benchmarking datasets.

Table S1. Kappa values for benchmarking clustering methods.

Datasets	Classical and Deep Clustering							Bayesian Clustering		
	K-Means	Clara	PAM	HC	Genie	MGMM	DEC	DPMVN	DPMNIG	DPGMVLG
PizzaCutter3	<i>0.203</i>	0.037	0.014	0.0008	0.063	0.118	0.186	0.042	0.147	0.230
Pap	0.035	0.035	0.035	0.035	<i>0.374</i>	0.058	0.037	0.068	0.066	0.530
VA	0.058	0.042	<i>0.081</i>	0.037	0.054	0.015	0.063	0.020	0.070	0.182
Ecomor	0.134	0.223	0.174	0.281	<i>0.420</i>	0.145	0.182	0.375	0.016	0.438
Survey	0.045	0.052	0.052	0.093	<i>0.107</i>	0.024	0.081	0.042	0.009	0.125
Leuk	0.047	<i>0.341</i>	<i>0.341</i>	0.011	0.213	0.151	0.344	0.105	0.334	0.393
Attrition	0.067	0.115	<i>0.116</i>	0.104	0.095	0.088	0.043	0.006	0.097	0.138
OME	0.015	0.020	0.011	0.002	0.002	0.002	0.005	0.025	<i>0.047</i>	0.166
Dune	0.296	0.299	0.299	<i>0.351</i>	0.304	<i>0.351</i>	0.322	0.274	0.306	0.463
Boston	0.002	0.001	0.006	0.003	0.018	<i>0.042</i>	0.019	0.020	0.002	0.242
Puromycin	0.059	<i>0.225</i>	<i>0.225</i>	0.059	0.073	0.059	0.059	0.135	0.059	0.383
Cats	0.399	0.420	0.430	0.408	0.405	0.373	0.375	0.377	<i>0.495</i>	0.567
Oribatid	0.122	0.142	<i>0.228</i>	0.142	0.142	0.081	0.154	0.049	0.064	0.370
Ais	0.921	0.792	0.921	0.911	0.822	0.951	0.723	0.023	<i>0.931</i>	0.901
Meau	0.722	0.722	0.722	0.611	0.722	<i>0.667</i>	0.389	0.189	0.167	0.722
Chick Weight	<i>0.054</i>	0.049	0.039	0.033	0.021	0.018	0.052	0.027	0.008	0.125
Aravo	0.129	<i>0.197</i>	0.140	0.103	0.168	0.077	0.230	0.106	0.054	0.394
Cabbages	<i>0.500</i>	<i>0.500</i>	<i>0.500</i>	0.467	0.400	0.467	0.467	0.049	0.067	0.567
CO2	0.310	0.357	0.310	<i>0.381</i>	0.310	0.071	0.143	0.193	0.167	0.571
Corvus	0.019	0.019	0.019	0.019	<i>0.643</i>	0.410	0.238	0.213	0.150	0.701
Pathogen Survey	0.225	0.092	0.150	<i>0.294</i>	0.189	0.224	0.217	0.0006	0.155	0.332
Lascaux	0.065	<i>0.088</i>	0.065	<i>0.088</i>	0.082	0.038	0.078	0.036	0.082	0.201
Fish	0.566	0.594	0.616	0.621	0.651	<i>0.653</i>	0.573	0.554	0.243	0.666
Nlschools	0.092	0.072	0.071	<i>0.093</i>	0.046	0.034	0.090	0.025	0.021	0.110
Macroloire	0.289	0.259	0.237	0.352	<i>0.570</i>	0.418	0.205	0.386	0.507	0.770
Wine	<i>0.949</i>	0.907	0.865	0.890	0.865	0.966	0.814	0.019	0.008	0.932
Iris	0.750	0.710	0.720	0.670	0.820	<i>0.880</i>	0.720	0.490	0.500	0.900
Carniherbi49	<i>0.831</i>	<i>0.831</i>	<i>0.831</i>	<i>0.831</i>	<i>0.831</i>	0.310	0.752	0.184	0.350	0.875
Elisa	0.200	0.097	0.097	0.200	<i>0.310</i>	0.143	0.200	0.200	0.143	0.385
Chazeb	0.142	0.569	0.569	<i>0.572</i>	<i>0.572</i>	0.324	0.556	0.179	0.215	0.657

Horizontal lines separate the datasets according to the level of low true cluster separation following Table 2. Values in **bold** and *italic* fonts indicate the best and second-best method, respectively.

Algorithm S1. Generation of Symmetric Low-Separability Clusters **Input:** N : Sample size, C : Number of clusters, D : Number of features, τ : Controls how close clusters are, σ_μ^2 : Controls the variation of cluster centers, σ_τ^2 : Controls the impact of τ on the formation of clusters.

Output: Data matrix with true clusters and true cluster centers.

- 1: Distribute N points approximately equally across the C clusters.
- 2: For each cluster,

- draw a D -dimensional vector from a $\text{Uniform}(0, (1 - \tau)\sigma_\tau^2)$ distribution, then
- add a Gaussian perturbation with standard deviation σ_μ^2 in each dimension.

The resulting vectors are the cluster centers; they lie close to each other, yielding low separability controlled by τ .

- 3: Compute

$$\sigma_\Sigma^2 = 1 + \sigma_\tau^2 \cdot \tau,$$

and set the covariance matrix for all clusters to

$$\Sigma = \sigma_\Sigma^2 \cdot \mathbf{I}_D.$$

where $\mathbf{I}_D$ is a D -dimensional identity matrix. This defines spherical Gaussian clusters with symmetric features and increased spread for higher overlap.

- 4: For each cluster, generate the assigned number of points from a multivariate normal distribution with:

$$\mathbf{x}_i \sim \text{Normal}(\boldsymbol{\mu}_i, \Sigma),$$

where $\boldsymbol{\mu}_i$ is the center of cluster i .

- 5: Stack all sampled points into a data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ and record the corresponding cluster labels.
- 6: **return** the data matrix $\mathbf{X}$, the label vector, and the matrix of true cluster centers.

Table S2. ARI values for benchmarking clustering methods.

Datasets	Classical and Deep Clustering							Bayesian Clustering		
	K-Means	Clara	PAM	HC	Genie	MGMM	DEC	DPMVN	DPMNIG	DPGMVLG
PizzaCutter3	<i>0.141</i>	-0.036	0.004	-0.005	0.028	0.086	0.149	0.066	0.008	0.190
Pap	0.015	0.015	0.015	0.015	<i>0.340</i>	0.011	-0.029	0.017	0.112	0.479
VA	0.009	-0.036	<i>0.038</i>	-0.038	0.034	-0.017	0.033	-0.001	-0.019	0.142
Ecomor	0.058	0.133	0.115	0.222	<i>0.431</i>	0.086	0.127	0.381	0.164	0.512
Survey	-0.00002	0.0005	0.0005	0.001	<i>0.052</i>	-0.012	0.036	-0.009	-0.007	0.060
Leuk	-0.021	0.087	0.087	-0.018	0.014	-0.009	0.090	-0.015	<i>0.101</i>	0.128
Attrition	-0.038	<i>0.028</i>	0.019	-0.012	0.044	0.002	-0.055	0.003	-0.001	0.069
OME	-0.0005	0.0001	-0.0006	-0.0004	-0.0004	-0.0004	-0.0005	0.005	<i>0.006</i>	0.027
Dune	<i>0.255</i>	0.157	0.157	0.083	0.007	0.083	0.153	0.068	0.047	0.299
Boston	-0.002	-0.000	0.002	0.001	-0.008	<i>0.020</i>	0.006	0.001	-0.019	0.222
Puromycin	-0.038	<i>0.004</i>	<i>0.004</i>	-0.038	-0.020	-0.038	-0.038	-0.029	-0.038	0.117
Cats	0.109	0.156	0.155	0.168	0.107	0.102	0.077	0.144	<i>0.295</i>	0.363
Oribatid	-0.015	-0.010	<i>0.026</i>	-0.010	-0.046	-0.020	-0.012	0.014	-0.031	0.204
Ais	0.847	0.626	0.847	0.829	0.674	0.903	0.520	-0.0004	<i>0.866</i>	0.811
Meau	0.520	0.520	0.520	0.322	<i>0.486</i>	0.446	0.121	0.011	0.071	0.478
Chick Weight	<i>0.011</i>	0.010	0.004	0.005	0.001	-0.002	0.010	-0.001	-0.003	0.058
Aravo	0.061	<i>0.093</i>	0.065	0.044	0.016	0.008	0.049	-0.014	-0.010	0.187
Cabbages	<i>0.237</i>	<i>0.237</i>	<i>0.237</i>	0.205	0.149	0.204	0.204	-0.014	0.0002	0.310
CO2	0.085	0.117	0.085	<i>0.135</i>	0.086	-0.006	0.009	0.112	0.018	0.318
Corvus	-0.033	-0.033	-0.033	-0.033	<i>0.392</i>	0.218	0.011	0.088	0.032	0.492
Pathogen Survey	0.046	0.017	0.009	0.107	0.004	0.076	<i>0.137</i>	-0.003	0.015	0.169
Lascaux	0.001	0.005	0.001	0.005	<i>0.006</i>	0.001	0.004	0.003	<i>0.006</i>	0.038
Fish	0.484	0.538	0.542	0.559	0.605	0.555	0.490	0.516	<i>0.583</i>	0.605
Nlschools	0.016	0.014	0.009	<i>0.040</i>	0.028	-0.0004	0.031	-0.008	-0.017	0.062
Macroloire	0.310	0.281	0.270	0.390	0.461	<i>0.495</i>	0.281	0.242	0.451	0.710
Wine	<i>0.897</i>	0.814	0.741	0.790	0.741	0.930	0.658	-0.010	-0.002	0.865
Iris	0.611	0.573	0.594	0.544	0.694	<i>0.786</i>	0.586	0.544	0.568	0.813
Carniherbi49	<i>0.694</i>	<i>0.694</i>	<i>0.694</i>	<i>0.694</i>	<i>0.694</i>	0.022	0.561	-0.046	0.103	0.765
Elisa	<i>0.083</i>	-0.050	-0.050	<i>0.083</i>	<i>0.083</i>	0.000	<i>0.083</i>	<i>0.083</i>	0.000	0.200
Chazeb	-0.025	0.289	0.289	0.292	0.292	0.068	<i>0.294</i>	0.019	0.037	0.401

Horizontal lines separate the datasets according to the level of low true cluster separation following Table 2. Values in **bold** and *italic* fonts indicate the best and second-best method, respectively.

Table S3. NMI values for benchmarking clustering methods.

Datasets	Classical and Deep Clustering							Bayesian Clustering		
	K-Means	Clara	PAM	HC	Genie	MGM	DEC	DPMVN	DPMNIG	DPGMVLG
PizzaCutter3	0.038	0.007	0.0002	0.001	0.004	0.015	0.037	0.075	0.043	<i>0.054</i>
Pap	0.001	0.001	0.001	0.001	<i>0.186</i>	0.005	0.005	0.108	0.092	0.253
VA	0.009	0.013	0.011	0.012	0.004	0.002	0.006	0.005	<i>0.023</i>	0.037
Ecomor	0.159	0.253	0.231	0.321	<i>0.385</i>	0.198	0.228	0.491	0.112	0.358
Survey	0.002	0.002	0.002	0.007	0.026	0.0009	0.010	<i>0.032</i>	0.0001	0.026
Leuk	0.002	<i>0.108</i>	0.108	0.0001	0.033	0.016	0.132	0.011	0.086	<i>0.115</i>
Attrition	0.013	0.015	<i>0.017</i>	0.019	0.008	0.011	0.012	0.003	0.015	<i>0.017</i>
OME	0.0002	0.0006	0.0001	0.000005	0.000005	0.000005	0.00004	<i>0.012</i>	0.011	0.020
Dune	0.380	<i>0.432</i>	<i>0.432</i>	0.359	0.288	0.359	0.330	0.337	0.247	0.536
Boston	0.00001	0.00000	0.00006	0.00001	0.00149	0.00282	0.0007	<i>0.0283</i>	0.0034	0.0704
Puromycin	0.003	<i>0.040</i>	<i>0.040</i>	0.003	0.009	0.003	0.003	0.014	0.003	0.118
Cats	0.210	0.170	0.197	0.139	<i>0.236</i>	0.153	0.219	0.128	0.191	0.251
Oribatid	0.015	0.021	0.046	0.021	<i>0.087</i>	0.007	0.026	0.023	0.011	0.155
Ais	0.760	0.598	0.760	0.738	0.572	0.856	0.494	0.040	<i>0.815</i>	0.717
Meau	0.628	0.628	0.628	0.471	<i>0.620</i>	0.566	0.317	0.224	0.128	0.594
Chick Weight	<i>0.013</i>	0.012	0.007	0.009	0.004	0.002	0.012	0.001	0.0003	0.032
Aravo	0.082	<i>0.120</i>	0.113	0.069	0.086	0.031	0.070	0.013	0.012	0.150
Cabbages	<i>0.191</i>	0.189	0.189	0.167	0.162	0.164	0.164	0.017	0.013	0.249
CO2	0.071	0.096	0.072	0.115	0.079	0.004	0.015	<i>0.138</i>	0.026	0.258
Corvus	0.0004	0.0004	0.0004	0.0004	<i>0.349</i>	0.228	0.048	0.106	0.158	0.394
Pathogen Survey	0.082	0.011	0.038	<i>0.124</i>	0.072	0.062	0.041	0.0008	0.060	0.125
Lascaux	0.003	0.006	0.003	0.006	0.007	0.002	0.005	<i>0.018</i>	0.007	0.030
Fish	0.681	<i>0.746</i>	<i>0.746</i>	0.744	0.661	0.747	0.685	0.647	0.575	0.636
Nlschools	0.0071	0.004	0.004	<i>0.0073</i>	0.006	0.001	0.006	0.004	0.001	0.012
Macroloire	0.466	0.451	0.445	0.516	0.542	<i>0.607</i>	0.451	0.336	0.526	0.633
Wine	<i>0.873</i>	0.825	0.782	0.783	0.751	0.910	0.657	0.029	0.005	0.830
Iris	0.637	0.612	0.643	0.593	0.694	<i>0.777</i>	0.622	0.545	0.579	0.793
Carniherbi49	<i>0.591</i>	<i>0.591</i>	<i>0.591</i>	<i>0.591</i>	<i>0.591</i>	0.195	0.485	0.135	0.097	0.711
Elisa	<i>0.098</i>	0.007	0.007	<i>0.098</i>	0.073	0.016	0.098	<i>0.098</i>	0.016	0.209
Chazeb	0.017	0.277	0.277	0.376	<i>0.379</i>	0.188	0.353	0.185	0.070	0.454

Horizontal lines separate the datasets according to the level of low true cluster separation following Table 2. Values in **bold** and *italic* fonts indicate the best and second-best method, respectively.

Table S4. Parameter settings for the benchmark methods.

Setting	K-Means	Clara	PAM
R Function	eclust	eclust	eclust
Distance	Euclidean	Euclidean	Euclidean
Linkage	default	default	default
Cluster selection	Fixed G	Fixed G	Fixed G
$k_{\max}$	10	10	10
Standardize	Yes (pre-scaled)	Yes (pre-scaled)	Yes (pre-scaled)
n_{boot}	100	100	100
Gini threshold	NA	NA	NA
Covariance models	NA	NA	NA
Initialization	NA	NA	NA
Seed	123	123	123
Setting	HC	Genie	MGMM
R Function	hclust	gclust	Mclust
Distance	Euclidean	Euclidean	NA
Linkage	Ward.D2	NA	NA
Cluster selection	Fixed G	Fixed G	Fixed G
$k_{\max}$	NA	NA	NA
Standardize	Yes (pre-scaled)	Yes (pre-scaled)	Yes (pre-scaled)
n_{boot}	NA	NA	NA
Gini threshold	NA	See Table 5	NA
Covariance models	NA	NA	All available
Initialization	NA	NA	Hierarchical
Seed	NA	NA	NA

$k_{\max}$: maximum number of clusters considered.

n_{boot} : number of Monte Carlo bootstrap samples.

NA: not applicable for the corresponding method.