Supplementary material for:

A game-inspired algorithm for marginal and global clustering

Miguel de Carvalho,[a,b,*] Gabriel Martos[b], and Andrej Svetlošák[a]

[a] School of Mathematics, University of Edinburgh, United Kingdom

[b] Department of Mathematics, CIDMA, Universidade de Aveiro, Portugal

[c] Universidad Torcuato Di Tella, Buenos Aires, Argentina

1 Background on Voronoi diagrams

Definitions

In this section we provide background on Voronoi tesselations and Delaunay triangulations. Let W be a finite subset of $\mathbb{R}^d$. The elements of W are called sites. For a specific site $w \in W$ a Voronoi cell $V(w)$ is defined as the set of points in $\mathbb{R}^d$ that are strictly closer to w than to any other site in W. Closeness is measured by a given norm, usually the Euclidean norm or the Mahalanobis norm. More generally, the Voronoi cell can be defined for a nonempty set of sites $U \subseteq W$; formally, $V(U)$ is defined as the set of points in $\mathbb{R}^d$ that are equidistant from all members of U, and closer to any member of U than to any site in $W \backslash U$. Together, all Voronoi cells fully partition the space $\mathbb{R}^d$; this partition is known as the Voronoi diagram and it is given by the family $\{V(U) : U \subseteq W\}$.

Related to Voronoi diagrams are Delaunay triangulations. For a set of sites $U \subseteq W$ a Delaunay face $D(U) \subset \mathbb{R}^d$ is the set of points in a sphere through all the sites of U. For a specific Delaunay face $D(U)$ all other sites in W are on the exterior of $D(U)$. Therefore, $D(U)$ is the interior of the convex hull of U. The collection of all Delaunay faces is called a Delaunay triangulation. This triangulation is a dual graph of its corresponding Voronoi diagram. Figure 1 provides an example of a Voronoi diagram and its corresponding Delaunay triangulation.

Notes & Comments

Voronoi tesselations and Delaunay triangulations are well known. The definitions above are adapted from [1, 2]. In the context of cluster analysis, Voronoi diagrams are well-known for representing the

regions associated with clusters formed by $\mathcal{K}$-means and $\mathcal{K}$-medoids algorithms, where each partition corresponds to the subset of the sample space nearest to a particular cluster center. The concept of Voronoi tesselations was introduced and studied by [3], Voronoi [4], and Thiessen [5]. Since Voronoi tessellations have been introduced by multiple authors, they are also known as Dirichlet tesselations and Thiessen diagrams. Delaunay triangulations (or tesselations) were proposed in [6]. Further technical details on Voronoi diagrams and Delaunay triangulations can be found in [7].

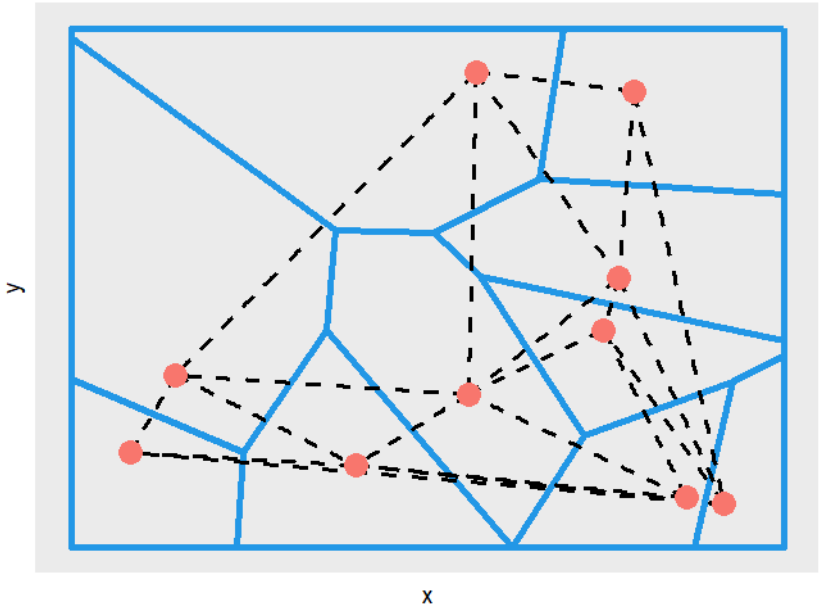

Figure 1: Example of a Voronoi diagram and its corresponding Delaunay triangulation. The red dots are sites, Voronoi faces are delimited by full blue lines, and the Delaunay triangulation is represented by the dotted lines.

2 Further numerical results

2.1 Outliers, skewness, and imbalanced mixing proportions

In this section, we report a Monte Carlo simulation study based on the scenarios introduced below.

Scenario A: Outliers

This scenario is a variation of Scenario 1 from the main paper, now with data contaminated by a uniform distribution in $[-10, 10]^2$. We consider a contamination level of 4%. In Figure 2 we depict a one shot example of the RC algorithm for a sample of size $n = 500$ and $u = 0.1$ in Scenario A. The resulting partition resembles the one obtained in the one shot example of Scenario 1. Moreover, outliers are assigned to clusters whose centers are closest in terms of Euclidean distance. Such pos-

itive performance extends beyond this one shot experiment, as demonstrated by the Monte Carlo evidence presented in Figure 5.

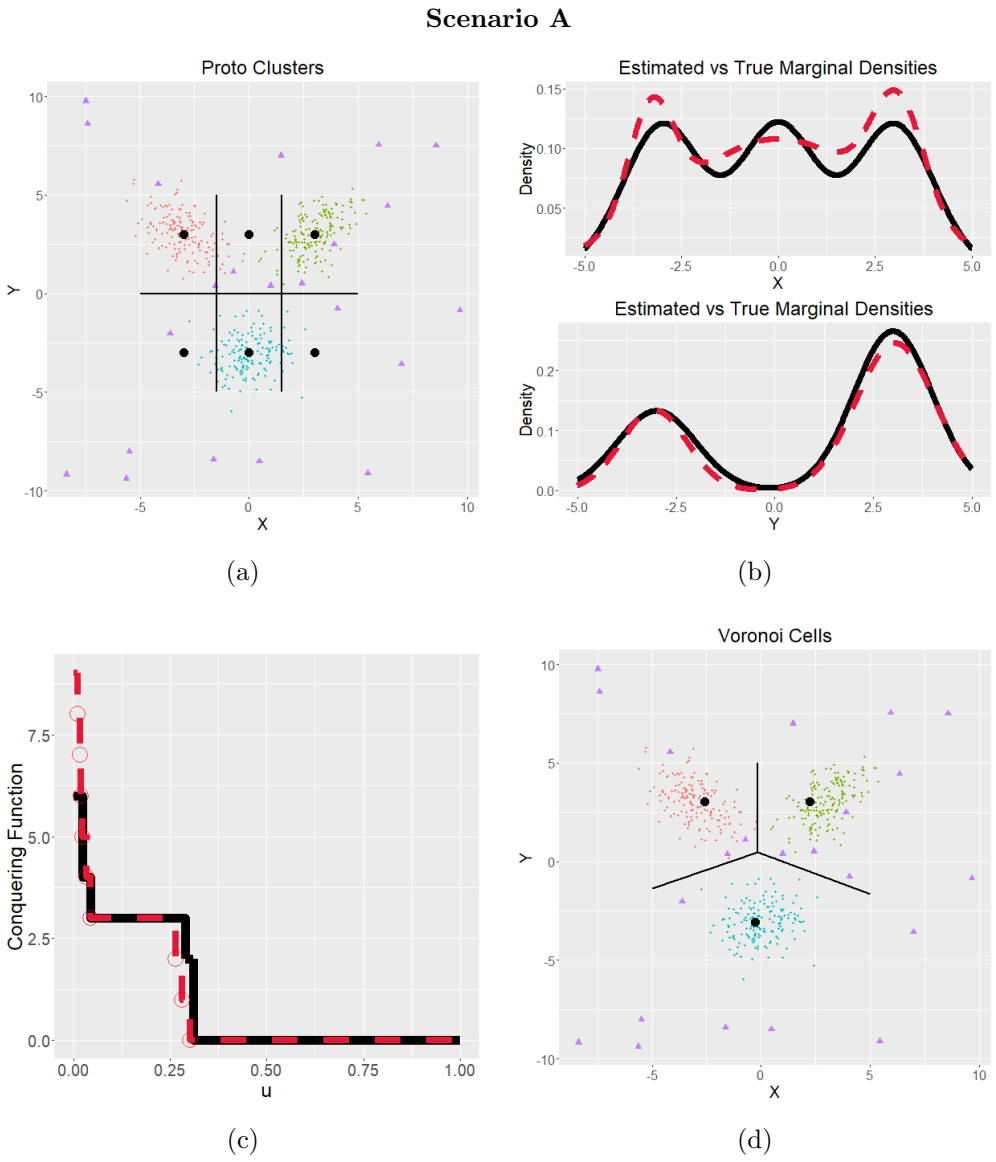

Figure 2: One shot experiments for Scenario A. (a) Simulated data and protoclusters (outliers represented using ▲). (b) Estimated (dashed) vs true (solid) marginal densities. (c) Estimated (dashed) vs true (solid) conquering functions. (d) Voronoi cells of the conquerors for $u = 0.1$.

Scenario B: Skewness

This scenario involves a mixture of $\mathcal{K} = 3$ skewed bivariate normal distributions. Data is generated using the `R` package `sn` [8] with the following parametrization: $\boldsymbol{\mu}_1 = (-3, 3)^{\mathrm{T}}$, $\boldsymbol{\mu}_2 = (3, 3)^{\mathrm{T}}$, $\boldsymbol{\mu}_3 = (0, -3)^{\mathrm{T}}$, $\boldsymbol{\alpha}_1 = (10, 4)^{\mathrm{T}}$, $\boldsymbol{\alpha}_2 = (-10, -4)^{\mathrm{T}}$, $\boldsymbol{\alpha}_3 = (0, 5)^{\mathrm{T}}$ and $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2 = \boldsymbol{\Omega}_3 = 2\mathbf{I}_2$; where $\boldsymbol{\mu}_i$,

$\boldsymbol{\alpha}_i$ and $\boldsymbol{\Omega}_i$ are the location, slant, and scale parameters corresponding to data in clusters $i = 1, 2, 3$; see Azzalini [9, Ch. 5] for further details on this parametrization. In Figure 3 we depict a one shot example of the RC algorithm corresponding to a sample of size $n = 500$ and $u = 0.1$ in Scenario B. As can be seen in panel (d), for this particular data, RC identifies 4 clusters.

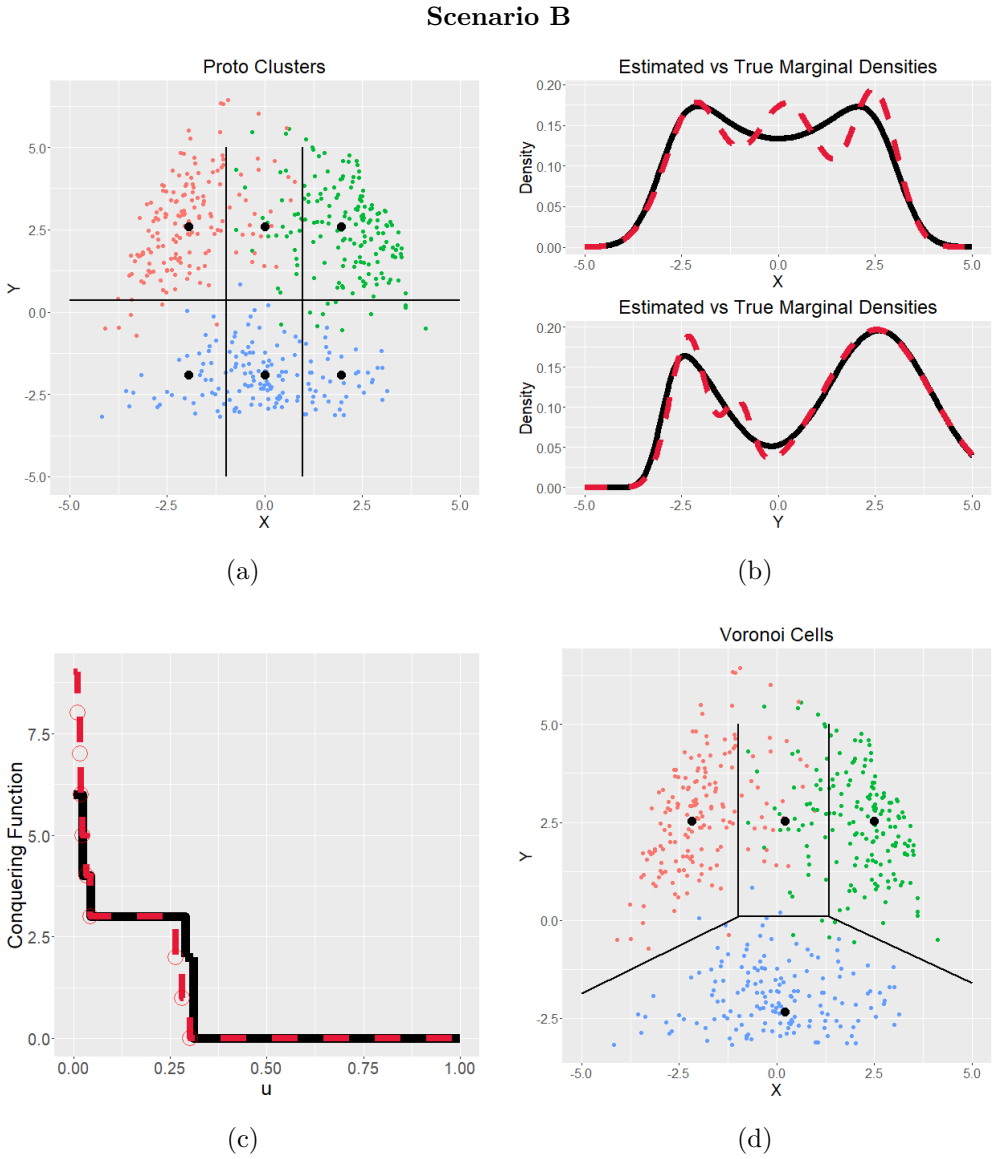

Figure 3: One shot experiments for Scenario B. (a) Simulated data and protoclusters. (b) Estimated (dashed) vs true (solid) marginal densities. (c) Estimated (dashed) vs true (solid) conquering functions. (d) Voronoi cells of the conquerors for $u = 0.1$.

Scenario C: Imbalanced mixing proportions

This scenario resembles Scenario 2 in the main paper but considers imbalanced mixing propor-

tions. In Figure 4 we depict a one shot example of the RC algorithm corresponding to a sample of size $n = 500$ and $u = 0.1$; mixing proportions for cluster 1 (●), 2 (●) and 3 (●) are 0.5, 0.3 and 0.2 respectively. Panel (d) resembles the solution obtained for the one shot experiment in Scenario 2.

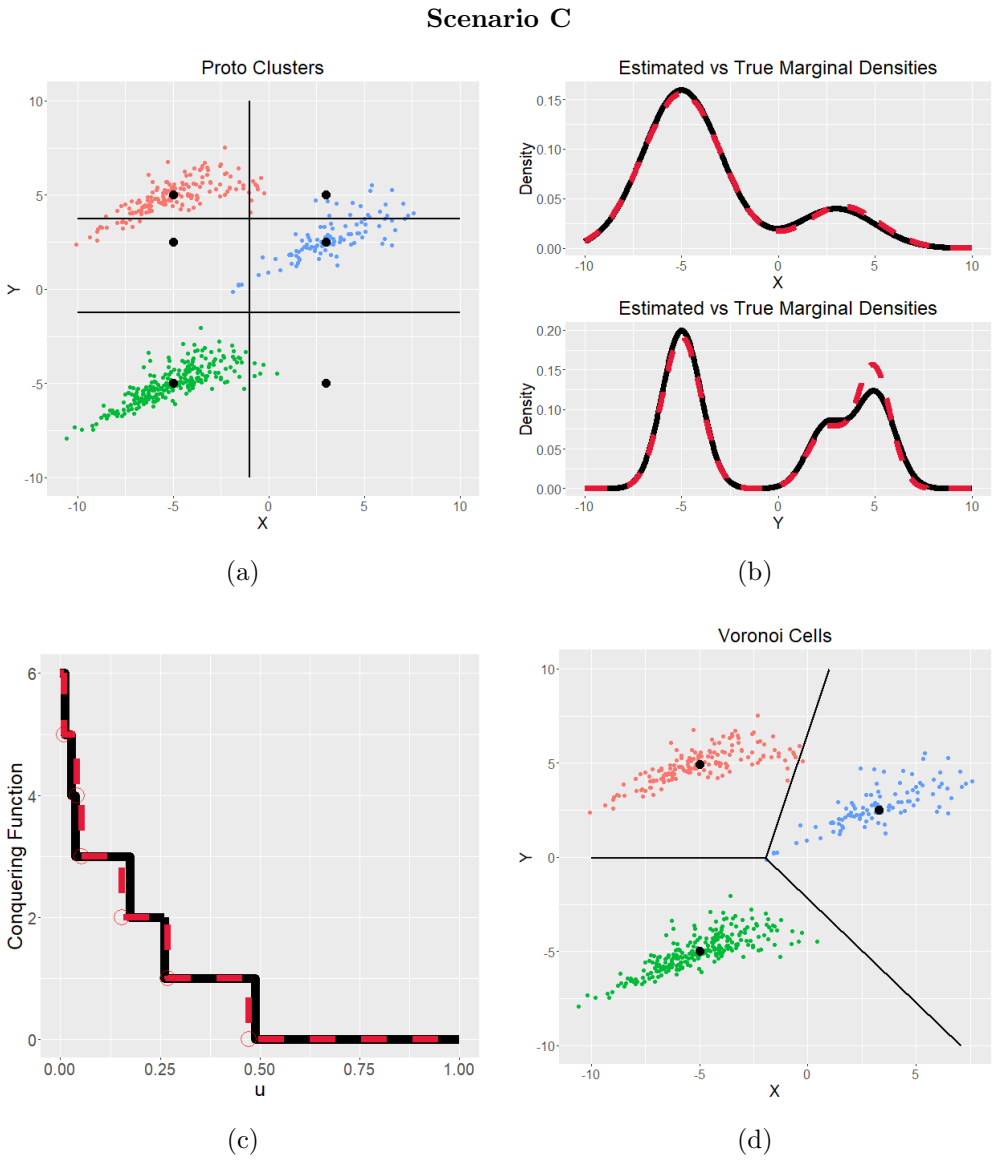

Figure 4: One shot experiments for Scenario C. (a) Simulated data and protoclusters. (b) Estimated (dashed) vs true (solid) marginal densities. (c) Estimated (dashed) vs true (solid) conquering functions. (d) Voronoi cells of the conquerors for $u = 0.1$.

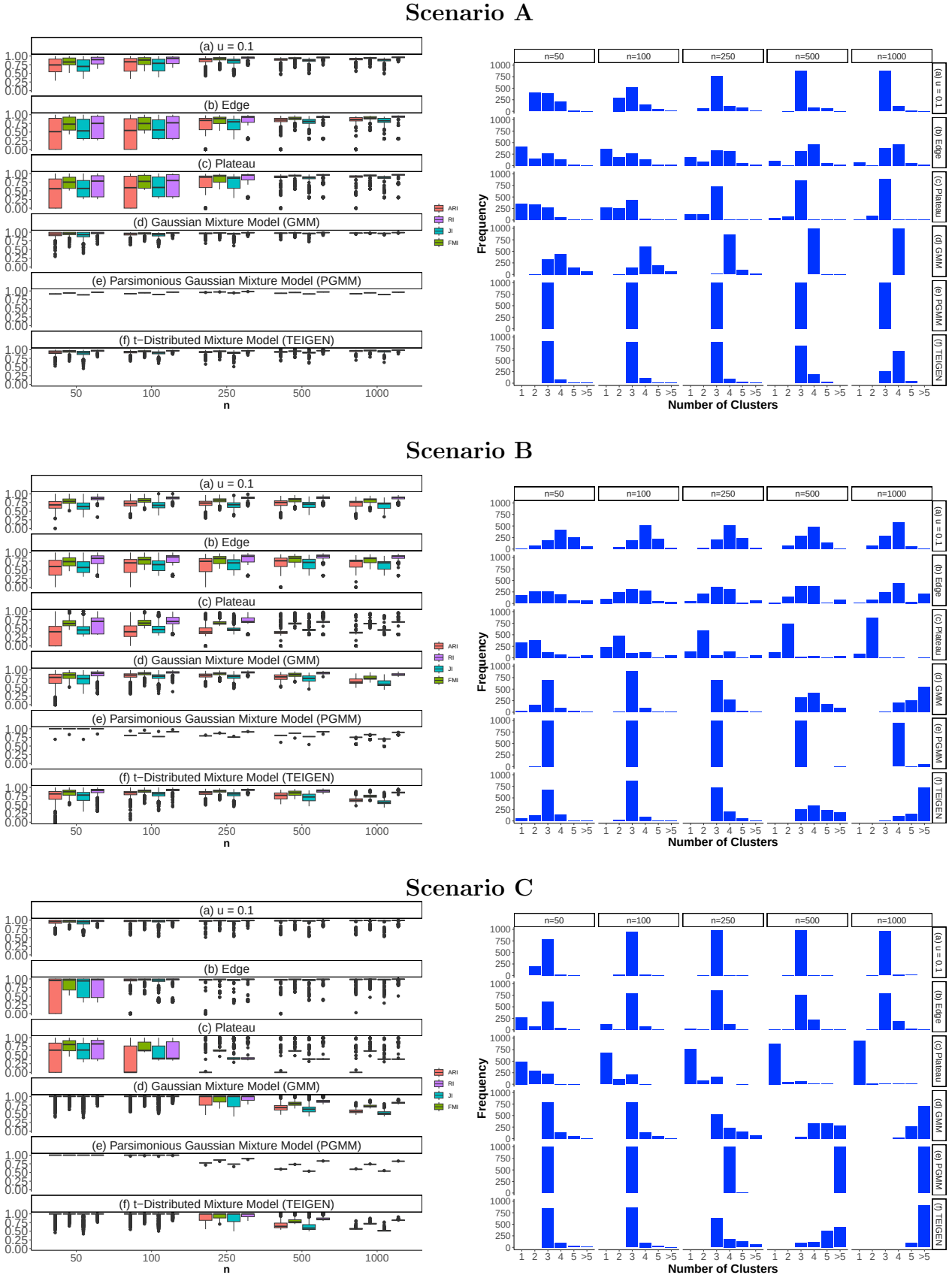

Figure 5: Monte Carlo simulation study for additional simulation scenarios with outliers, skewness, and imbalanced mixing proportions: (Left) Performance metrics (ARI, RI, JI, FMI). (Right) Empirical distribution on the number of detected clusters.

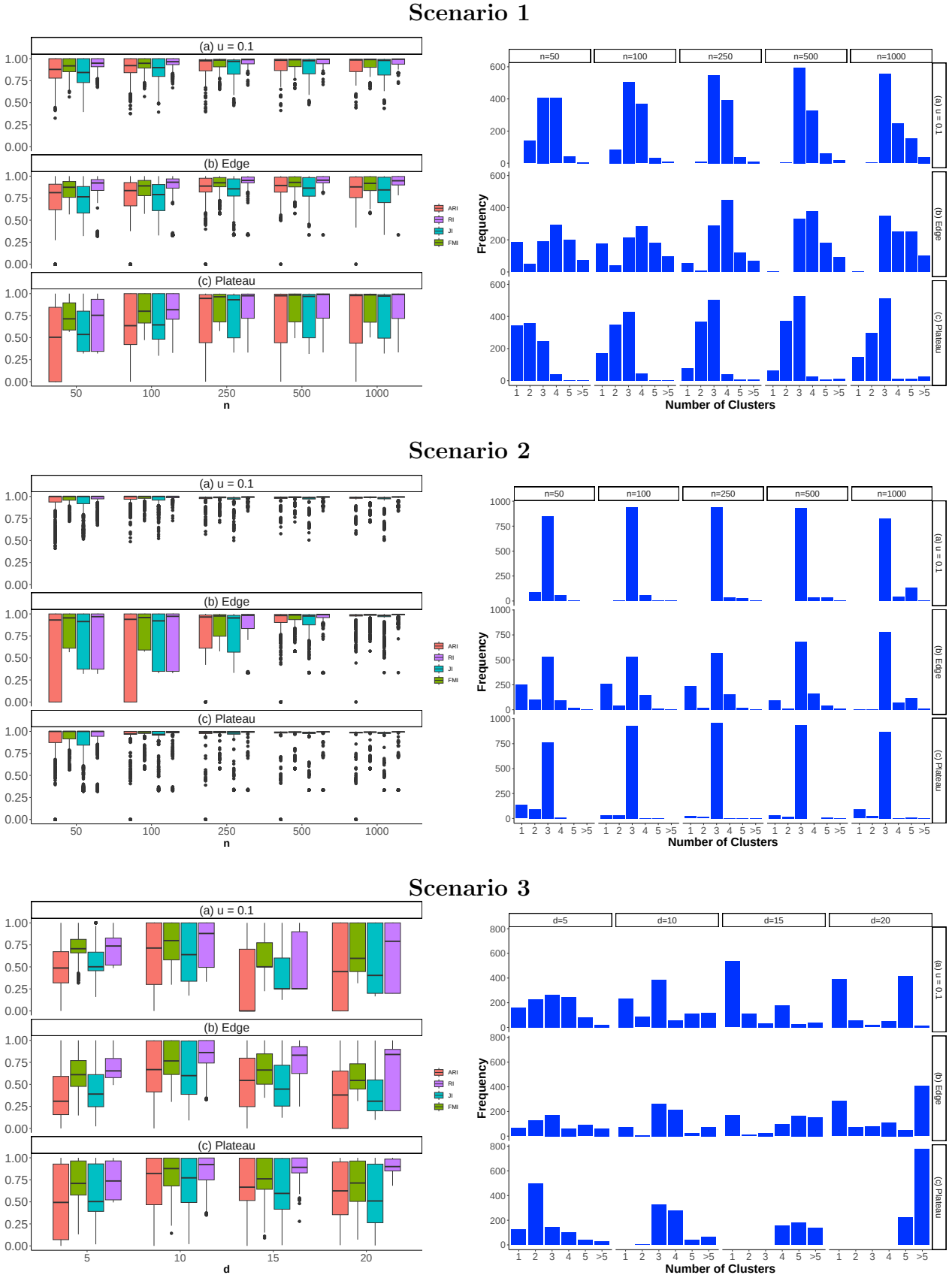

Figure 6: Monte Carlo simulation study for Mahalanobis norm-based version of the RC algorithm: (Left) Performance metrics (ARI, RI, JI, FMI). (Right) Empirical distribution on the number of detected clusters.

Monte Carlo evidence—taking stock

To assess the performance of the RC clustering algorithm in Scenarios A, B, and C we run a Monte Carlo simulation study considering sample sizes $n \in \{50, 100, 250, 500, 1000\}$. Some remarks on the simulation results are in order:

- Figure 5 (Scenario A) shows that the results for RC, PGMM, and TEIGEN are tantamount to the ones of Scenario 1 in the paper. In this regard, the RC clustering algorithm appears to be robust to a moderate amount of uniformly distributed outliers.

- Figure 5 (Scenario B) indicates that in the case of skewed data, RC and GMM frequently tend to under or over identify clusters in the data. TEIGEN and PGMM, on the other hand, appear to produce sensible cluster solutions even for relatively small sample sizes in the presence of moderately skewed data.

- Figure 5 (Scenario C) suggests that in the presence of imbalanced mixing proportions, RC (edge) behaves better than RC (plateau). Similarly to the Scenario 2 in the main paper, RC (edge) overperforms GMM, PGMM, and TEIGEN. Such overperformance stems once more from RC flexibility in adjusting to a different number of clusters per margin.

All in all, the evidence above suggests, that the RC algorithm demonstrates reasonable resilience when faced with a moderate number of outliers in the data or in scenarios with slightly imbalanced mixing proportions. Scenario B suggests however that in the presence of skewed data, it may underperform in comparison to TEIGEN and PGMM.

2.2 Alternative metric for conquering

After considering the comments provided by an anonymous reviewer, we reran the Monte Carlo simulations for Scenarios 1, 2, and 3 as described in the paper, this time using the Mahalanobis norm to allocate observations to protoclusters and clusters—instead of the Euclidean norm as in the main paper. Thus, the lth data point is encoded into the protocluster (Step 2) or cluster (Step 3) that minimize the following Mahalanobis norm:

$$\text{Enc}(l) = \arg\min_{\mathbf{i}} \ (\mathbf{x}_l - \boldsymbol{\mu}_\mathbf{i})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_l - \boldsymbol{\mu}_\mathbf{i}),$$

where $\boldsymbol{\mu}_{\mathbf{i}} = (\mu_{X_1}^{(i_1)}, \ldots, \mu_{X_d}^{(i_d)})^{\mathrm{T}}$ is the sample mean corresponding to protocluster or cluster $\mathbf{i} \in I$, and $\boldsymbol{\Sigma}$ is the sample covariance matrix. As can be seen in Figure 6, the simulation results considering this alternative encoding strategy are equivalents in Scenarios 1 and 2 ($d = 2$). In the case of Scenario 3 (moderately high–dimensional data), the performance of RC via Mahalanobis norm decreases. For moderately large d and relatively small sample size n, the computation of Mahalanobis distance entails numerical issues (i.e., $\boldsymbol{\Sigma}$ is frequently an ill–conditioned matrix that poorly estimates the true covariance matrix). Versions of the Mahalanobis norm for addressing clustering problems in the context of high-dimensional data have been proposed [e.g., 10]; incorporating these into the RC algorithm remains a task for future research.

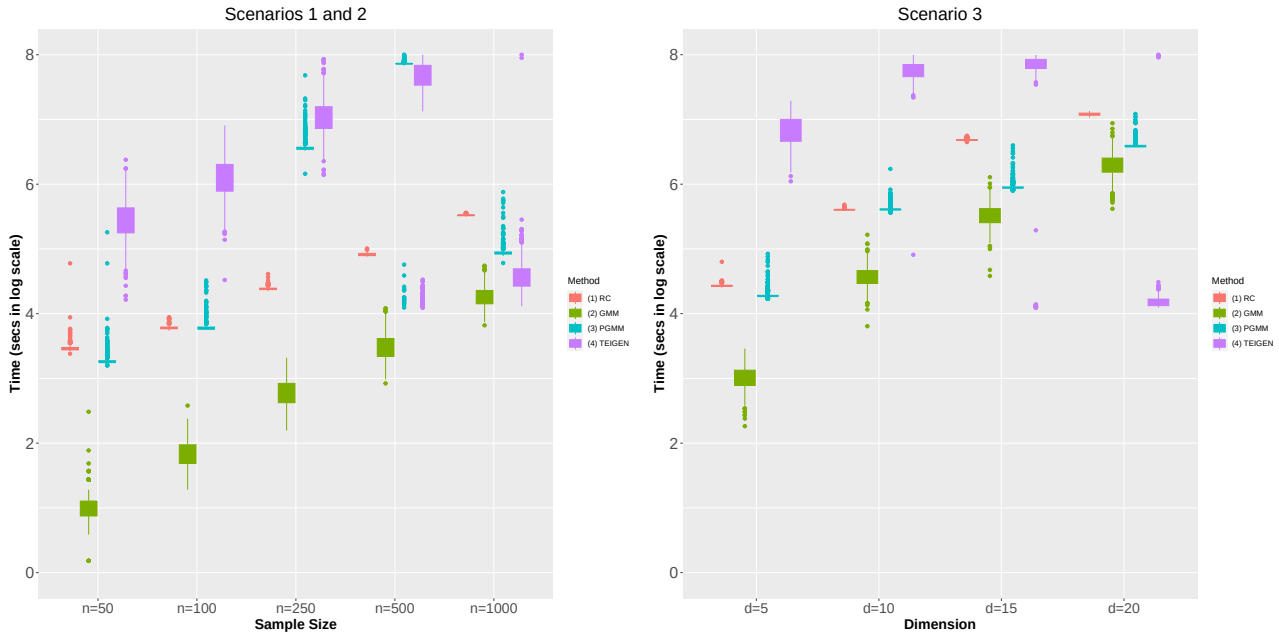

Figure 7: Side-by-side boxplots of execution times.

2.3 Analysis of computational times

In Figure 7, we compare the execution time of RC with other clustering methods under the Monte Carlo simulations in Scenarios 1, 2, and 3. The observed differences are partially explained by different degrees of efficiency in the implementation of each method; for example, GMM fits are obtained with the `mclust` package, whose key routines are written in Fortran. The version of the RC clustering method implemented here relies on the `mombf` package [11] for Step 1, and exhibits computational efficiency comparable to GMM.

3 Further empirical results

In Table 1 we report external agreement metrics for all data sets in the paper where we have access to the true labels. It can be seen that the global performance of RC is in line with that of other mainstream alternative clustering methods.

Method	**Metric**	**Banknotes**	**Wine**	**Rice**	**Iris**
RC	ACC	0.995	0.758	0.921	0.702
	NMI	0.665	0.417	0.405	0.593
	RI	0.990	0.737	0.766	0.793
	ARI	0.979	0.414	0.633	0.584
	JI	0.980	0.441	0.689	0.619
	FMI	0.999	0.613	0.799	0.797
GMM	ACC	0.730	0.949	0.329	0.666
	NMI	0.674	0.888	0.303	0.636
	RI	0.840	0.932	0.578	0.776
	ARI	0.680	0.848	0.169	0.568
	JI	0.679	0.817	0.218	0.595
	FMI	0.824	0.899	0.431	0.774
PGMM	ACC	0.740	0.825	0.606	0.980
	NMI	0.674	1.002	0.378	1.021
	RI	0.793	0.906	0.668	0.973
	ARI	0.592	0.779	0.342	0.941
	JI	0.591	0.731	0.413	0.927
	FMI	0.768	0.851	0.609	0.961
TEIGEN	ACC	0.800	0.511	0.424	0.666
	NMI	0.674	0.881	0.365	0.636
	RI	0.839	0.783	0.613	0.776
	ARI	0.679	0.444	0.234	0.568
	JI	0.678	0.394	0.298	0.595
	FMI	0.823	0.605	0.507	0.771

Table 1: External agreement metrics for different clustering methods and real data sets with labels.

References

[1] Y. Ito. Voronoi Tessellation. In Encyclopedia of Applied and Computational Mathematics, ed. Engquist, B., Springer, New York, pp. 1509–1547, 2015.

[2] S. Fortune. Voronoi diagrams and Delaunay triangulations. In Handbook of Discrete and Computational Geometry, Chapman & Hall/CRC, Boca Raton FL, pp. 705–721, 2017.

[3] L. G. Dirichlet. Über die reduction der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen. Journal für die Reine und Angewandte Mathematik (Crelles Journal) (1850) 209–227.

[4] G. Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs. Journal für die Reine und Angewandte Mathematik (Crelles Journal) (1908) 198–287.

[5] A.H. Thiessen. Precipitation averages for large areas. Monthly Weather Rev. 39 (1911) 1082–1089.

[6] B. Delaunay. Sur la sphere vide. Izv. Akad. Nauk SSSR 7 (1934) 1–2.

[7] A. Okabe, B. Boots, K. Sugihara, S.N. Chiu. Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. Wiley, New York, 2009.

[8] A. Azzalini. The R package `sn`: The skew-normal and related distributions such as the skew-t and the SUN (version 2.1.1), Università degli Studi di Padova, http://azzalini.stat.unipd.it/SN/ (2023).

[9] A. Azzalini. The Skew-Normal and Related Families, Cambridge University Press, Cambridge MA, 2013.

[10] M. Fauvel, J. Chanussot, J.A. Benediktsson, A. Villa. Parsimonious Mahalanobis kernel for the classification of high dimensional data. Pattern Recognit. 46 (2013) 845–854.

[11] D. Rossell. Bayesian model selection and averaging with `mombf`. Unpublished manuscript (2018).

[12] I. Cinar, M. Koklu. Classification of rice varieties using Artificial Intelligence methods. Int. J. Int. Sys. Appl. Eng. 7 (2019) 188–194.

[13] M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs. `NbClust`: an R package for determining the relevant number of clusters in a data set. J. Statist. Soft. 61 (2014) 1–36.