Supplementary material to:
"**A greedy and optimistic clustering for leveraging individual covariate uncertainty**"

Akifumi Okuno[a] and Kohei Hattori[b]

[a] ISM and RIKEN AIP, okuno@ism.ac.jp
[b] NAOJ and ISM and U. Michigan, khattori@ism.ac.jp

A Additional example

Example 3 *For the general clustering problem of individuals* $[n]$ *equipped with observed covariate instances* $z_i \in \mathbb{R}^d$ *(*$i \in [n]$*), we can assume that* z_i *is an instance of the covariate* $Z_i \in \mathbb{R}^d$ *following the distribution* $\mathbb{P}_{Z_i}$*, which is typically a normal distribution* $\mathbb{P}_{Z_i} = \mathcal{N}(z_i, \Sigma_i)$*. The uncertainty set,* $\mathcal{Z}_i$*, can be specified by* $\mathcal{Z}_i := \{Z \in \mathbb{R}^d \mid p_i(Z) > \varepsilon\}$*, where* p_i *denotes the probability density of* $\mathbb{P}_{Z_i}$*. To cluster individuals* $i \in [n]$*, a nonlinear dimensionality reduction, including a kernel principal component analysis (Schölkopf et al. 1998), can be applied beforehand to the observed covariate instances* $\{z_i\}_{i=1}^n$ *to obtain* $x_i = f(z_i)$ *and* $\mathcal{X}_i = f(\mathcal{Z}_i)$*.*

B Related works

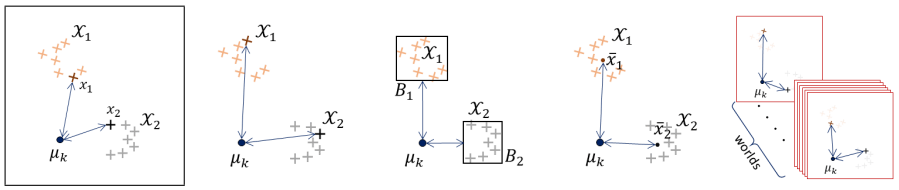
This section provides a fully detailed version of the related works presented in Section 1.2 of the main body. Although it includes more detailed explanations, the cited references remain the same as those in Section 1.2.

Robust optimization. In contrast to the optimistic approach (i.e., optimizing the best-case loss functions over the uncertainty sets) considered in this paper, *robust optimization (RO)* minimizes the worst-case (pessimistic) loss functions (Ben-Tal and Nemirovski 2002; Bertsimas et al. 2011), with application to statistical problems including classification (Xu et al. 2009; Takeda et al. 2013). Vo et al. (2016) applies RO to clustering: see Figure 6(b) for illustration. However, RO applied to clustering problem demonstrates extremely low scores in our setting, as discussed in Section 3.3. While RO typically employ small balls equipped with the p-norm ($p = 1, 2$) centered at instances of observed covariates as convex uncertainty sets (Ben-Tal and Nemirovski 2002; Vo et al. 2016), this study considers non-convex uncertainty sets due to the non-linear pre-processing.

Clustering of the uncertain data. A similar approach to ours can be found in Ngai et al. (2006), which assumes that the uncertainty set $\mathcal{X}_i$ consists of finite points and considers the minimum box B_i that contains $\mathcal{X}_i$; (a bound of) the Hausdorff distance between the boundary ∂B_i and cluster center is used for K-means clustering, instead of the Euclidean distance. See Figure 6(c). However, Ngai et al. (2006) is not compatible with our setting, as it implicitly assumes the convexity of the set $\mathcal{X}_i$ (also see Supplement E for discussion; it is not suitable for our non-convex setting).

Other directions for clustering uncertain data are classified into the following four types.

(i) The first type employs the probability density function p_i of the feature X_i ($i = 1, 2, \ldots, n$), and takes expectation of the clustering loss function. While the simple K-means (MacQueen 1967) minimizes the squared Euclidean distance between the feature instance and the cluster centers, UK-means (Chau et al. 2006) considers the expectation of the distance between feature and the cluster centers (with respect to $(p_1, p_2, \ldots, p_n)$). As pointed out in Lee et al. (2007) and Cormode and McGregor (2008), UK-means is equivalent to K-means applied to the expectation of features $\bar{x}_i = \mathbb{E}(X_i)$ ($i = 1, 2, \ldots, n$). See Figure 6(d). Cormode and McGregor (2008) also provides approximation algorithms for the variants of UK-means (e.g., UK-median).



(a) GOC (b) Pessimistic (c) Dist. to Box (d) UK-means (e) PW model

Fig. 6: Comparison of how to measure the distance from the cluster center μ_k (to the uncertainty set $\mathcal{X}_i$). (a) Proposed GOC considers the distance to the nearest instance, (b) robust optimization applied to clustering (a slight modification of Vo et al. (2016)) considers the distance to the farthest instance, (c) Ngai et al. (2006) computes the distance to the boundary of the minimum box $B_i(\supset \mathcal{X}_i)$, (d) UK-means is equivalent to applying K-means to $\bar{x}_i = \mathbb{E}(X_i)$, and (e) PW model considers all the possible worlds $\Xi_1, \Xi_2, \cdots \in \mathcal{A}_n$ and aggregates their clustering results.

Although UK-means consequently considers only the expectation $\bar{x}_i$, it ignores the variances of each density p_i: Gullo et al. (2013) incorporates the variance penalty of each object (to each cluster loss) to UK-means. From the same perspective, Gullo and Tagarelli (2012) first computes the average $\bar{q}_k$ of p.d.f.s of cluster members for each cluster k, and minimizes the sum of variance of $\bar{q}_k$ $(k = 1, 2, \ldots, K)$.

(ii) The second type also employs the density function p_i, and measures the distance between the pair (p_i, p_j) (e.g., KL-divergence) instead of the distance between single observed instances X_i, X_j. Jiang et al. (2013) applies similarity-based clustering methods to the distances (also see Experiment 5 in Section 4 for the related approach).

(iii) The third type considers the distribution of the distance, between the uncertain features X_i, X_j. See Kriegel and Pfeifle (2005a), Kriegel and Pfeifle (2005b), and Zhang et al. (2017).

(iv) The fourth type called possible-world (PW) model, considers all the possible combinations of the feature candidates. These combinations are called "worlds" $\Xi_1, \Xi_2, \cdots \in \mathcal{A}_n := \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$, and PW approaches apply a clustering algorithm to each world $\Xi_1, \Xi_2, \ldots$ (in parallel), and aggregates all the clustering results (Volk et al. 2009; Züfle et al. 2014; Liu et al. 2021). See Figure 6(e) for illustration.

However, the above approaches (i)-(iii) assume the tractability of the probability densities; they are not compatible with our setting, which considers the non-linearly transformed features whose distribution is generally intractable. Although studies in line with (iii) are practically implemented by using empirical distance computed via sampling features from the distributions, computing the distance between all the objects require high computational complexity. Furthermore, in contrast to our approach finding only the optimistic candidates, PW models (iv) overall require unrealistic computational cost in our setting, as the number of (even a subset of) possible worlds $\mathcal{A}_n$ is huge. We note that, clustering uncertain data is distinct from fuzzy clustering (see, e.g., Bezdek (1981)), which outputs multiple assignments of clusters with deterministic input. Although the approach by Kumar and Patel (2007) for clustering data with measurement errors may seem to share a similar problem setting with ours, they assume that each covariate x_i is generated from a normal distribution with fixed cluster centers. In contrast, our method allows the distribution center of the covariate to vary.

C Detailed Descriptions of Synthetic Dataset

C.1 General Description of the Simulation

In galactic astronomy, it is believed that the Milky Way was formed through the merging of smaller systems, such as dwarf galaxies. In the numerical experiments described in Section 4.1, we generated mock data by simulating the formation process of the Milky Way. To simplify this case, we assume that $K_* = 50$ dwarf galaxies merge with the Milky Way and are instantaneously disrupted at time $\tau = 0$. Each dwarf galaxy contains 30,000 sibling stars. When a dwarf galaxy is disrupted, sibling stars begin moving independently. After $\tau = 0$, and until the current epoch ($\tau = 10 \times 10^9$ years), the motions of these sibling stars are treated as test particles (i.e., particles with zero mass) moving within the gravitational potential of the Milky Way. For each dwarf galaxy, the positions and velocities of the sibling stars at $\tau = 0$ (i.e., the initial conditions) slightly differ from each other. The small difference in the initial conditions evolves over cosmic time, and the positions and velocities of the sibling stars are completely different from each other in the current epoch, although they originate from the same dwarf galaxy.

C.2 Visualization of the Simulation

To provide an intuitive understanding of the simulation, Figure 7 shows a subset of sibling stars in two dwarf galaxies A and B that merge with the Milky Way at $\tau = 0$. At $\tau = 0$, the sibling stars in each dwarf galaxy have identical positions and slightly different velocities, making these two groups clearly distinguishable in terms of their positions and velocities. At $\tau = 3 \times 10^9$ years, the positions and velocities of the sibling stars exhibit a wider distribution, and the two groups of stars are marginally distinguishable in terms of their positions and velocities. At the current epoch, $\tau = 10 \times 10^9$ years, the positions and velocities of the sibling stars show a mixed distribution. At this point, it is difficult to separate two groups of stars from each other in terms of their positions and velocities.

From these three snapshots, it is evident that finding sibling stars within a six-dimensional position and velocity space becomes more difficult as the system evolves over time. Importantly, this difficulty is unrelated to the accuracy of the data. Even if we have a perfect measurement of the positions and velocities of the stars, finding sibling stars is a difficult task if we use the raw data of position and velocity.

However, the case appears to be simpler if we look at the system in a three-dimensional phase space spanned by the orbital action. (As a reminder of the readers, the orbital action is a three-dimensional conserved quantity, which is a function of position and velocity; and it describes the stellar orbital properties.) Because the orbital action is conserved for each star, the distribution of sibling stars in the orbital action space is also conserved over time, as shown in the rightmost panels in Figure 7. Therefore, using the action distribution instead of the position and velocity distributions is an indispensable strategy for identifying sibling stars. As shown in Example 1, in the presence of observational uncertainties in the stellar positions and velocities, it is difficult to find sibling stars in the action space, which motivated us to introduce the GOC algorithm as a new type of clustering approach.

C.3 Detailed Implementations of the Simulation

To generate mock data, we first randomly generated $K_* = 50$ centroids with positions $\vec{x}_k^{\mathrm{centroid}}$ and velocities $\vec{v}_k^{\mathrm{centroid}}$ ($k = 1, \cdots, K_*$) using a realistic distribution function model of the Milky Way, similar to that described in Hattori et al. (2021). The kth centroid corresponds to the position and velocity of the kth dwarf galaxy at $\tau = 0$. For the kth dwarf galaxy, we generated $N_{\mathrm{sibling}} = 30000$ positions and velocities, $\vec{x}_{ks}(\tau = 0)$ and $\vec{v}_{ks}(\tau = 0)$ ($s = 1, \cdots, N_{\mathrm{sibling}}$), such that $\vec{x}_{ks}(\tau = 0) = \vec{x}_k^{\mathrm{centroid}}$ and $\vec{v}_{ks}(\tau = 0) \sim \mathcal{N}(\vec{v}_k^{\mathrm{centroid}}, \Sigma)$

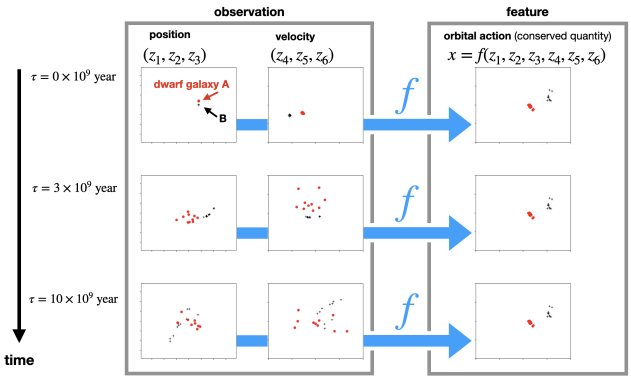


Fig. 7: Disruption of two dwarf galaxies in the Milky Way. Even in the absence of observational uncertainty, finding sibling stars from a six-dimensional position and the velocity space is difficult at the current epoch ($\tau = 10 \times 10^9$ years). By contrast, finding sibling stars in a three-dimensional orbital action space is easier because the orbital actions are conserved over time. Note that, for clarity, this figure only shows two-dimensional projections of the three-dimensional position, velocity, and orbital action.

with $\Sigma = (5 \text{ km s}^{-1})^2 I$. Here, $I \in \mathbb{R}^{3\times 3}$ denotes the identity matrix. These positions and velocities correspond to the initial conditions of the sibling stars at $\tau = 0$. From these initial conditions, we integrated the orbits of $K_* N_{\text{sibling}}$ stars for 10×10^9 years (which is approximately the age of the universe) under a widely used gravitational potential model of the Milky Way described in McMillan (2017) and derived the current-day positions and velocities. At this point, N_{sibling} stars originating from the same dwarf galaxy are no longer located close to each other (see the bottom-left panel in Figure 7). To mimic the observations, for the kth group, we randomly select n_k stars that are close to the current position of the Sun. (Note that stars that are too far away from the Sun are too faint to be observed.) The assumed position and velocity of the Sun is the same as those in Doke and Hattori (2022). We chose $n_k = 1 + \{(k-1) \bmod 10\}$, where "$x$ mod a" denotes the residual of the division (x divided by a). With this, we have $n = 275$ stars in total, such that we have 1 member star for $k = 1, 11, 21, 31, 41$; we have 2 member stars for $k = 2, 12, 22, 32, 42$; and so on.

For completeness, in the following, we briefly mention how we converted the simulated data into the uncertainty set used by the GOC algorithm. (See Section 4.1 for a full description.) First, we converted the simulated stellar positions and velocities of $n = 275$ stars into observable quantities, as illustrated schematically in Figure 2(a). Note that the stellar positions and velocities in the simulation are *true* quantities that are unavailable in reality. To mimic the actual observation, we add a random error to the observable quantities, which are then used to construct the uncertainty set.

By following the same procedure, we run 10 independent simulations. Each simulation is used to construct a dataset.

D Convergence of GOC Using K-Medoids and GMM

Regarding the comparison of the convergence of the K-means clustering shown through Experiment 3 described in Section 4.3, Figures 8–10 show the convergence of K-medoids, GMM (`ClusterR`), and GMM (`Mclust`+BIC), respectively, all of which demonstrated the same tendencies.

E Discussion: Convex Uncertainty Sets

While we employ the empirical uncertainty set $\tilde{\mathcal{X}}_i^{(m_i)}$ which is not restricted to be convex, we may consider an alternative convex set $\tilde{T}_i^{(m_i)}$ containing $\tilde{\mathcal{X}}_i^{(m_i)}$: finding possible feature candidates over the set

$$\mathcal{A}_n^\dagger := \tilde{T}_1^{(m_1)} \times \tilde{T}_2^{(m_2)} \times \cdots \times \tilde{T}_n^{(m_n)}$$

instead of $\mathcal{A}_n$, is expected to be more efficiently computed by the existing optimization techniques related to convex sets. For computational reasons, Ngai et al. (2006) considers a minimum box B_i containing $\tilde{\mathcal{X}}_i^{(m_i)}$ as the convex set $\tilde{T}_i^{(m_i)}$, and Vo et al. (2016) assumes that the uncertainty set is box-shaped (i.e., convex).

Referring to Vo et al. (2016), we may employ difference-of-convex algorithm (DCA; see, e.g., Le Thi and Tao 2005) to solve a specific form of GOC (particularly, GOC equipped with K-means) more efficiently. We think that this convex modification of GOC would be a future research worth considering, while we do not employ this convex set $\tilde{T}_i^{(m_i)}$ in this study by the following reasons: (i) to exploit the convex techniques, we need to heavily restrict the types of clustering oracle $\mathfrak{C}$ (whereby the applicability of GOC would be much degraded, and the implementation would be mathematically difficult for users), and (ii) the convex set $\tilde{T}_i^{(m_i)}$ may contain large unnecessary regions in some situations (see Figure 11).

F Evaluation metrics

We define the following scores, using the estimated clusters $\hat{\boldsymbol{c}} = (\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_n) \in [K]^n$ as well as the true clusters $\boldsymbol{c}^* = (c_1^*, c_2^*, \ldots, c_n^*) \in [K_*]^n$, $n_{kl} := \sum_{s=1}^n \mathbb{1}(\hat{c}_s = k)\mathbb{1}(c_s^* = l)$, $n_{k\cdot} := \sum_{l=1}^{K_*} n_{kl}$, and $n_{\cdot l} := \sum_{k=1}^{K} n_{kl}$.

1. Normalized mutual information $\mathrm{NMI}(\hat{\boldsymbol{c}}, \boldsymbol{c}^*)$ is defined by $2\mathcal{I}/(\mathcal{H}^{(1)} + \mathcal{H}^{(2)})$, using the mutual information $\mathcal{I}$ and entropy $\mathcal{H}$:

$$\mathcal{I} := \sum_{k=1}^{K}\sum_{l=1}^{K_*} \frac{n_{kl}}{n} \log \frac{n \cdot n_{kl}}{n_{\cdot l} \cdot n_{k\cdot}},$$

$$\mathcal{H}^{(1)} := -\sum_{k=1}^{K} \frac{n_{k\cdot}}{n} \log \frac{n_{k\cdot}}{n}, \quad \mathcal{H}^{(2)} := -\sum_{l=1}^{K_*} \frac{n_{\cdot l}}{n} \log \frac{n_{\cdot l}}{n}.$$

2. The F**-measure** $F(\hat{\boldsymbol{c}}, \boldsymbol{c}_*)$ is defined as

$$\sum_{l=1}^{K_*} \frac{n_{\cdot l}}{n} \max_{k \in [K]} \left\{ \left(\frac{n_{kl}}{n_{k\cdot}} + \frac{n_{kl}}{n_{\cdot l}} \right)^{-1} \left(\frac{2n_{kl}^2}{n_{k\cdot} n_{\cdot l}} \right) \right\}.$$

Both the NMI and F-measure take values within $[0, 1]$, and attain a value of 1 if and only if the estimated clusters $\hat{\boldsymbol{c}}$ perfectly match the true clusters $\boldsymbol{c}^*$ (up to the permutation of the cluster labels).

G Applicability of GOC algorithm to incomplete data

In Example 1, we assume that 6-dimensional position and velocity data are available for all the stars under consideration. However, this assumption does not always hold in practice. For instance, if we are interested in very faint stars, the distance—corresponding to $z_{i,1}$ in Fig. 2(a)—cannot be reliably measured. Likewise, if the stellar spectrum is unavailable, the line-of-sight velocity—that is, the velocity component directed toward or away from us,

represented by $z_{i,4}$ in Fig. 2(a)—is also unmeasurable. In such cases, only 5-dimensional data are available, unlike the scenario in Example 1.

Nevertheless, the GOC algorithm may still be applicable by leveraging appropriate domain knowledge. For example, even if the stellar distance is unknown, we can reasonably assume it lies between 0 and the size of the Milky Way, as most observable stars reside within the Milky Way. Similarly, if the line-of-sight velocity is missing, we can assume it falls within the range $-v_{\mathrm{esc}}$ to v_{esc}, where $v_{\mathrm{esc}} \simeq 600\ \mathrm{km\ s^{-1}}$ is the escape velocity from the Milky Way, since most visible stars are gravitationally bound to the Milky Way.

Thus, with appropriate domain-based constraints, we can construct uncertainty set for each star, making it possible to apply the GOC algorithm even with incomplete data.

References

Ben-Tal, A. and Nemirovski, A. (2002). Robust optimization–methodology and applications. *Mathematical Programming*, 92(3):453–480.

Bertsimas, D., Brown, D. B., and Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM review*, 53(3):464–501.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.

Chau, M., Cheng, R., Kao, B., and Ng, J. (2006). Uncertain data mining: An example in clustering location data. In *Proceedings of the 10th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, page 199–204.

Cormode, G. and McGregor, A. (2008). Approximation algorithms for clustering uncertain data. In *Proceedings of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 191–200.

Doke, Y. and Hattori, K. (2022). Probability of forming gaps in the gd-1 stream by close encounters of globular clusters. *The Astrophysical Journal*, 941(2):129.

Gullo, F., Ponti, G., and Tagarelli, A. (2013). Minimizing the variance of cluster mixture models for clustering uncertain objects. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(2):116–135.

Gullo, F. and Tagarelli, A. (2012). Uncertain centroid based partitional clustering of uncertain data. *Proceedings of the VLDB Endowment*, 5(7):610–621.

Hattori, K., Valluri, M., and Vasiliev, E. (2021). Action-based distribution function modelling for constraining the shape of the Galactic dark matter halo. *Monthly Notices of the Royal Astronomical Society*, 508(4):5468–5492.

Jiang, B., Pei, J., Tao, Y., and Lin, X. (2013). Clustering uncertain data based on probability distribution similarity. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):751–763.

Kriegel, H.-P. and Pfeifle, M. (2005a). Density-based clustering of uncertain data. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 672–677.

Kriegel, H.-P. and Pfeifle, M. (2005b). Hierarchical density-based clustering of uncertain data. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 689–692. IEEE.

Kumar, M. and Patel, N. R. (2007). Clustering data with measurement errors. *Computational Statistics & Data Analysis*, 51(12):6084–6101.

Le Thi, H. A. and Tao, P. (2005). The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133:23–46.

Lee, S. D., Kao, B., and Cheng, R. (2007). Reducing UK-means to K-means. In *Seventh IEEE International Conference on Data Mining Workshops*, pages 483–488.

Liu, H., Zhang, X., Zhang, X., Li, Q., and Wu, X.-M. (2021). RPC: Representative possible world based consistent clustering algorithm for uncertain data. *Computer Communications*, 176:128–137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Oakland, CA, USA.

McMillan, P. J. (2017). The mass distribution and gravitational potential of the Milky Way. *Monthly Notices of the Royal Astronomical Society*, 465(1):76–94.

Ngai, W. K., Kao, B., Chui, C. K., Cheng, R., Chau, M., and Yip, K. Y. (2006). Efficient clustering of uncertain data. In *Proceedings of the Sixth International Conference on Data Mining*, pages 436–445.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.

Takeda, A., Mitsugi, H., and Kanamori, T. (2013). A unified classification model based on robust optimization. *Neural Computation*, 25(3):759–804.

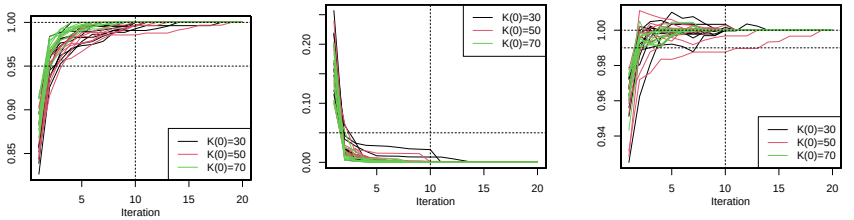
Vo, X. T., Le Thi, H. A., and Pham Dinh, T. (2016). Robust optimization for clustering. In *Intelligent Information and Database Systems*, pages 671–680. Springer Berlin Heidelberg.

Volk, P. B., Rosenthal, F., Hahmann, M., Habich, D., and Lehner, W. (2009). Clustering uncertain data with possible worlds. In *IEEE 25th International Conference on Data Engineering*, pages 1625–1632.

Xu, H., Caramanis, C., and Mannor, S. (2009). Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(51):1485–1510.

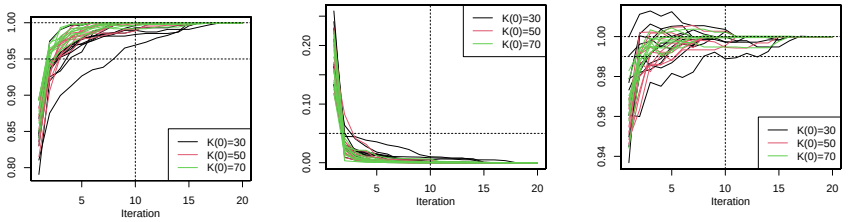
Zhang, X., Han, L., and Xiaotong, Z. (2017). Novel density-based and hierarchical density-based clustering algorithms for uncertain data. *Neural Networks*, 93:240–255.

Züfle, A., Emrich, T., Schmid, K. A., Mamoulis, N., Zimek, A., and Renz, M. (2014). Representative clustering of uncertain data. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 243–252.



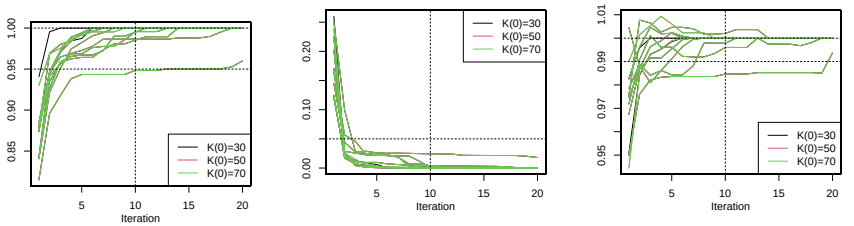
(a) $\mathrm{NMI}(\hat{\boldsymbol{c}}(t), \hat{\boldsymbol{c}}(\infty))$. (b) $n^{-1}\sum_{i=1}^{n}\|x_i(t) - x_i(\infty))\|_2^2$. (c) $\mathrm{NMI}(\hat{\boldsymbol{c}}(t), \boldsymbol{c}^*)/\mathrm{NMI}(\hat{\boldsymbol{c}}(\infty), \boldsymbol{c}^*)$.

Fig. 8: Convergence of K-medoids. We observe a similar tendency to K-means (see Figure 5).



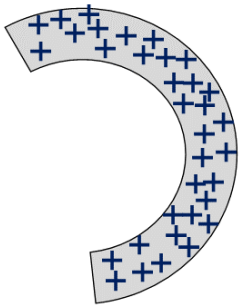
(a) $\mathrm{NMI}(\hat{\boldsymbol{c}}(t), \hat{\boldsymbol{c}}(\infty))$. (b) $n^{-1}\sum_{i=1}^{n}\|x_i(t) - x_i(\infty))\|_2^2$. (c) $\mathrm{NMI}(\hat{\boldsymbol{c}}(t), \boldsymbol{c}^*)/\mathrm{NMI}(\hat{\boldsymbol{c}}(\infty), \boldsymbol{c}^*)$.

Fig. 9: Convergence of GMM (`ClusterR`). We observe a similar tendency to K-means (see Figure 5).

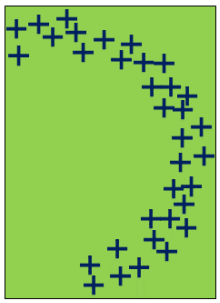


(a) $\mathrm{NMI}(\hat{\boldsymbol{c}}(t), \hat{\boldsymbol{c}}(\infty))$. (b) $n^{-1}\sum_{i=1}^{n}\|x_i(t) - x_i(\infty))\|_2^2$. (c) $\mathrm{NMI}(\hat{\boldsymbol{c}}(t), \boldsymbol{c}^*)/\mathrm{NMI}(\hat{\boldsymbol{c}}(\infty), \boldsymbol{c}^*)$.

Fig. 10: Convergence of GMM (`Mclust`+BIC). We observe a trend very similar to that of K-means (see Figure 5). However, regardless of the initial number of clusters $K(0)$, all paths converge to nearly identical outcomes. This is likely because, in the first iteration, the BIC-based model selection yields almost the same cluster count $K(1)$ irrespective of the value of $K(0)$, leading to identical computations thereafter.



(a) Underlying uncertianty set $\mathcal{X}_i$ (colored in grey)



(b) Smallest box T_i containing $\tilde{\mathcal{X}}_i^{(m_i)}$

Fig. 11: Compared to the underlying uncertainty set $\mathcal{X}_i$ (where the empirical uncertainty set $\tilde{\mathcal{X}}_i^{(m_i)}$ is shown by the "+" symbols), the smallest box T_i containing $\tilde{\mathcal{X}}_i^{(m_i)}$ includes a large unnecessary regions.