楽観的なクラスタリングと その背後にある数理についての検討

奥野彰文[1,2]

[1]統計数理研究所, [2]理研AIP

共同研究者：服部公平（統数研/天文台）， Ian Roederer （ミシガン大）

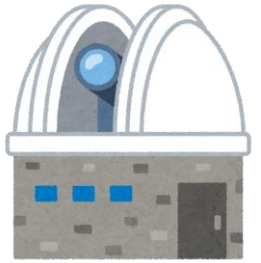
手法提案：Okuno and Hattori (arXiv:2204.08205)
実データへの適用：Hattori, Okuno and Roederer (ApJ2023)
関係する理論っぽい話： Okuno (arXiv:2407.10418)

Research Background

やりたいこと（ざっくり）

現在

位置や速度を観測



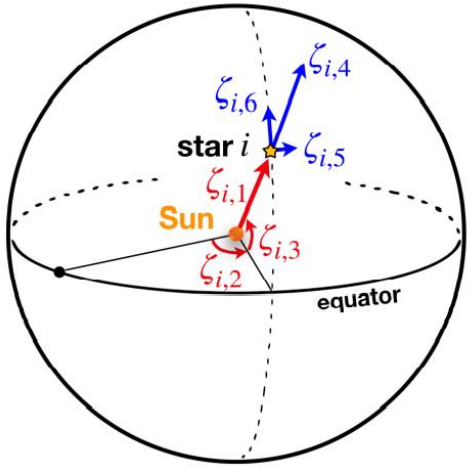
NASAなどが持つ天文台

どの天体とどの天体が**同じ銀河からやってきた**のか，知りたい．

Setting (1/3)

Position and velocity of stars:

their **observation uncertainty are different for each star**.



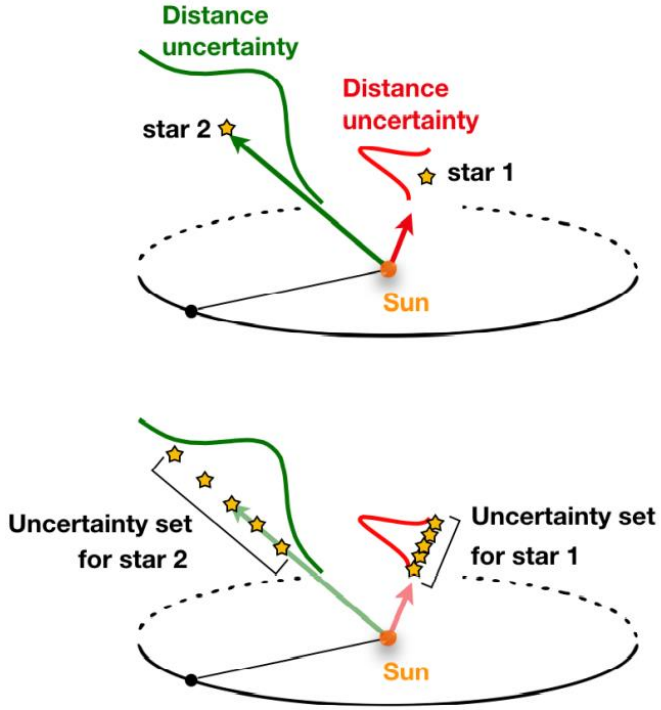
Position (relative to the Sun) | Velocity (relative to the Sun)

$\zeta_{i,1}$ Parallax = 1/Distance | $\zeta_{i,4} = \mathrm{d}(1/\zeta_{i,1})/\mathrm{dt}$

$\zeta_{i,2}$ Azimuthal angle | $\zeta_{i,5} = \mathrm{d}\zeta_{i,2}/\mathrm{dt}$

$\zeta_{i,3}$ Polar angle | $\zeta_{i,6} = \mathrm{d}\zeta_{i,3}/\mathrm{dt}$

(a) Stellar position and velocity observed from the Sun.



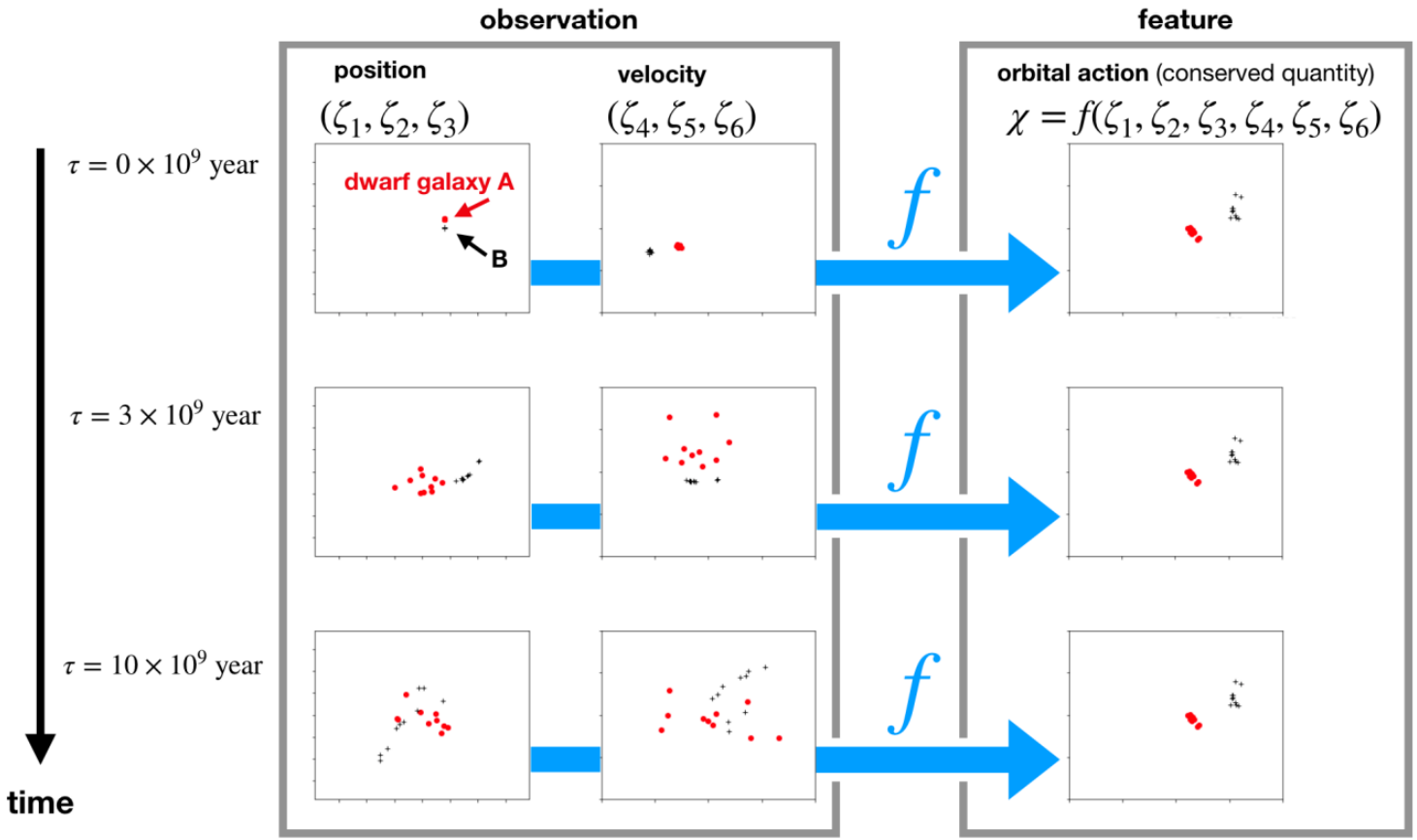
(b) Uncertainty sets for individual stars.

”Uncertainty set” for the observation: $\mathcal{Z}_i := \{Z \in \mathbb{R}^d \mid p_i(Z) > \varepsilon\},$

Setting (2/3)

Nonlinear pre-processing is important:

from observation to conserved feature



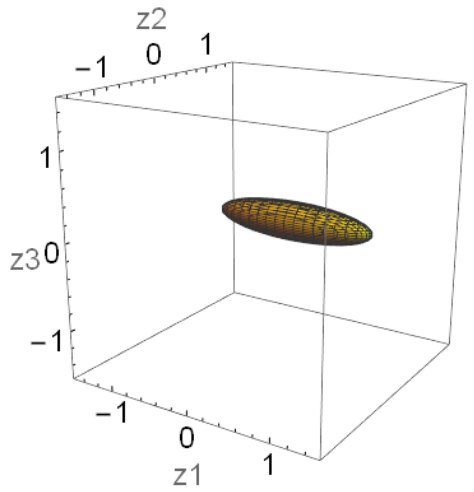
Setting (3/3)

Non-linear preprocessing twists the uncertainty sets!

$Z_i \in \mathbb{R}^d$: covariate
(observed position/velocity)

"Uncertainty set"

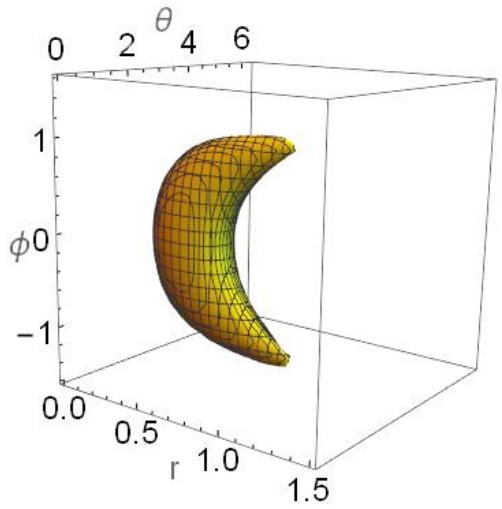
$$\mathcal{Z}_i := \{Z \in \mathbb{R}^d \mid p_i(Z) > \varepsilon\},$$



f

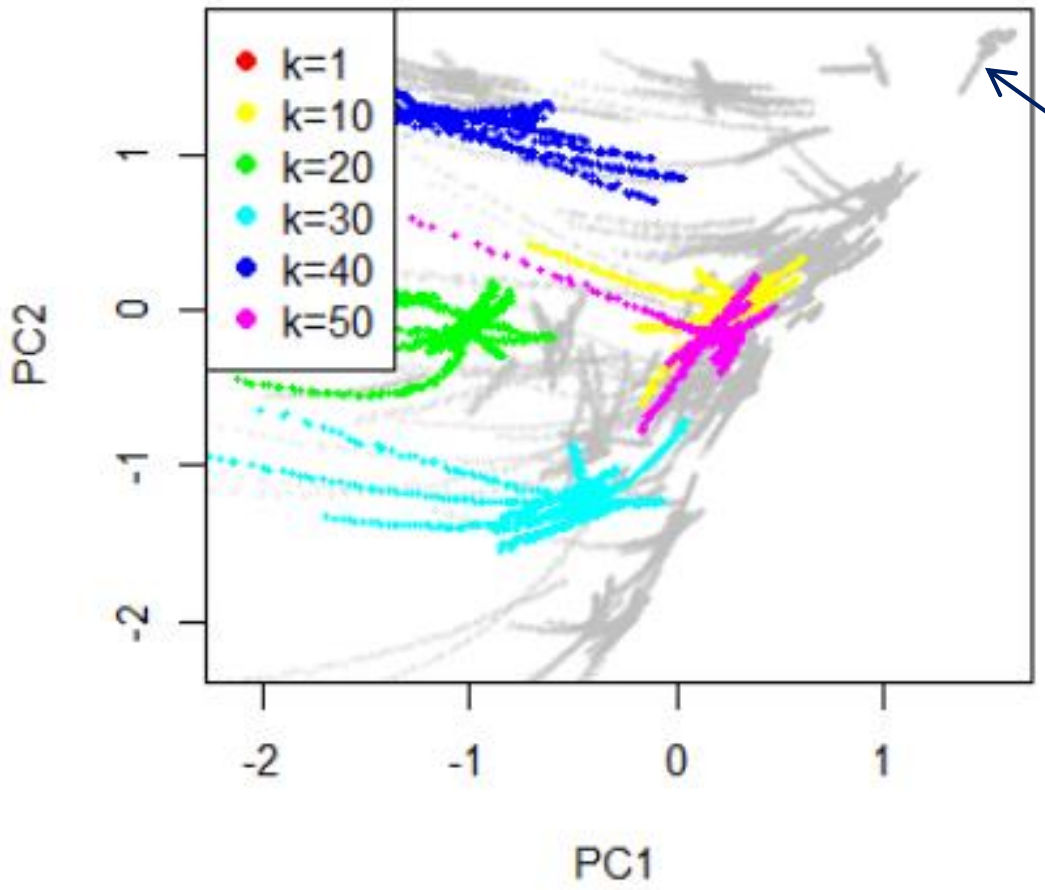
$X_i \in \mathbb{R}^q$: feature
(conserved quantity)

$$\mathcal{X}_i = f(\mathcal{Z}_i) := \{f(Z) \mid Z \in \mathcal{Z}_i\} \subset \mathbb{R}^d,$$



$$\mathbb{P}(X_i \in \mathcal{X}_i) \geq \mathbb{P}(Z_i \in \mathcal{Z}_i) \overset{(1)}{\geq} 1 - \eta.$$

Example: feature uncertainty sets of stars



Each curve represents the
feature uncertainty set of each star.
(Obviously, non-convex)

Progress from the previous 2018 study

- Roederer (2018):
  48 stars having large uncertainty are removed from observed 83 stars.
  (Namely, only **35 stars** are considered)

4 years later
Additional observations are obtained.

- Okuno and Hattori (https://arxiv.org/abs/2204.08205, This study)
- Hattori, Okuno, and Roederer (ApJ2023):

  We use observations of **161 stars**!

Summary of previous studies

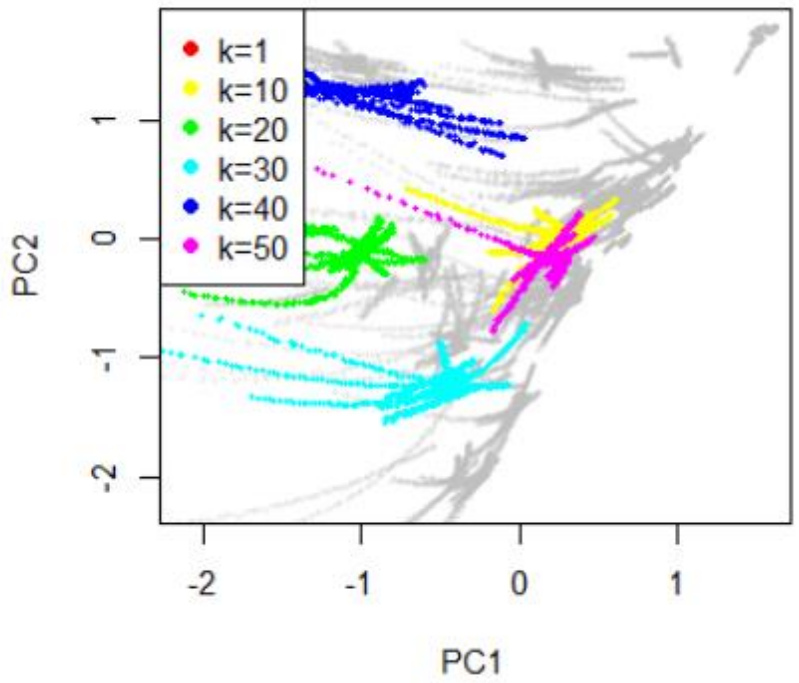
- UK-means (Chau et al. 2006; Ngai et al. 2006; Lee et al. 2007;…)
- Possible-world models (Volk et al. 2009; Zufle et al. 2014)
- Distribution distance + Hier. clust. (Kriegel and Pfeife 2005; …)

Taking average w.r.t. covariate uncertainty

- Robust optimization + clustering (Vo et al. 2016)

Taking worst case w.r.t. covariate uncertainty.
(Pessimistic attitude)

Uncertainty set in our setting…



- No meaning to take average
- Worst case analysis results in devastating output

(only 1 big cluster is output then)

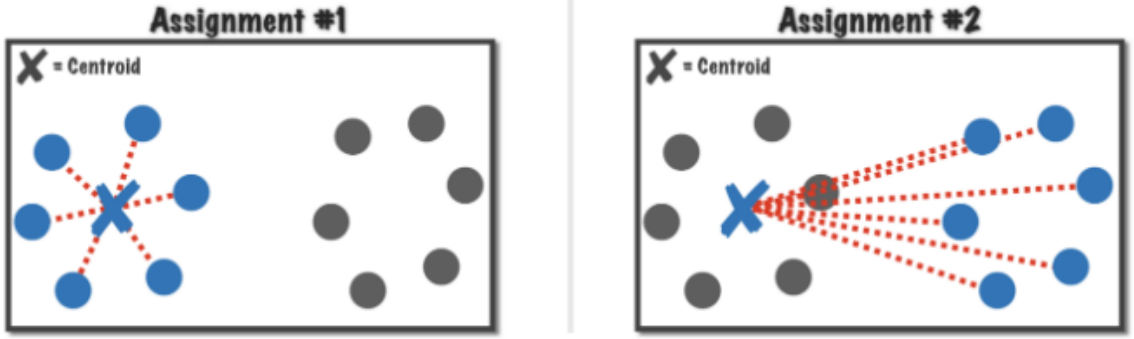
既存研究

不確実性を考慮しないクラスタリング

例えば K-means (MacQueen et al., 1967)

$$\min_{\mu_1,\mu_2,\ldots,\mu_k} \sum_{k=1}^{K} \sum_{i=1}^{n} s_{ik} \|x_i - \mu_k\|_2^2$$

s_{ik}: クラスタ割当　μ_k: クラスタ中心　x_i: 各個体の特徴量



(https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c より)

1999; Nugent and Meila, 2010), among other areas. Owing to its versatility and numerous applications, many types of clustering algorithms have been developed, including K-means (MacQueen et al., 1967), mean-shift (Fukunaga and Hostetler, 1975; Cheng, 1995), spectral clustering (Chung, 1997; Von Luxburg, 2007), convex clustering (Pelckmans et al., 2005), and likelihood-based approaches using stochastic block models (Holland et al., 1983) and Gaussian mixture models (McLachlan and Peel, 2000). For comprehensive surveys of various clustering algorithms, see Jain and Dubes (1988), Everitt (1993), and Xu and Wunsch (2005).

不確実なデータのクラスタリング

UK-means (Chau et al. KDD2006)

各特徴量x_iが正規分布p_iに従うとし，　$\min_{\mu_1,\mu_2,\ldots,\mu_k} \sum_{k=1}^{K} \sum_{i=1}^{n} s_{ik} E_{p_i}(\|x_i - \mu_k\|_2^2)$

クラスタ中心との距離の期待値を最小化する！

- Ngai et al. (ICDM2006)

  UK-meansの高速な近似

- Lee et al. (ICDM2007 workshop)

  UK-meansは実は特徴量の期待値 $x_i' := E_{p_i}(x_i)$にK-meansをかけるのと等価(※自明)

  ~~観測値x_iを使って$p_i = \phi((x - x_i)/\sigma)$とかしてたらUK-meansは意味ないのでは・・・~~

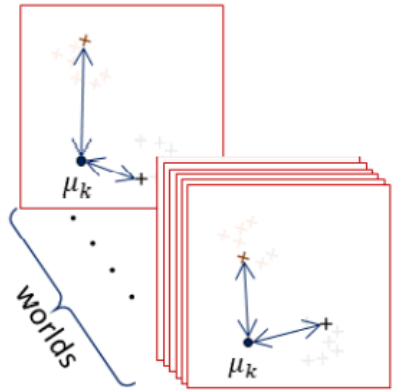
- Cormode and McGregor. (2008)

  UK-means，　UK-medianなどの高速なapproximation algorithm

Possible-world models (Volk et al. 2009; Zufle et al. 2014)

可能な組み合わせ全てでクラスタリングを並列計算
最後にaggregation.

CPUコア数が無限ならすぐ終わるが総計算量はとても大きい…



分布間距離を用いる (Kriegel and Pfeifle 2005; Jiang et al. 2013)

特徴量x_i, x_jの確率分布p_i, p_jの距離をKLなどで測って
距離ベースのクラスタリング法を適用

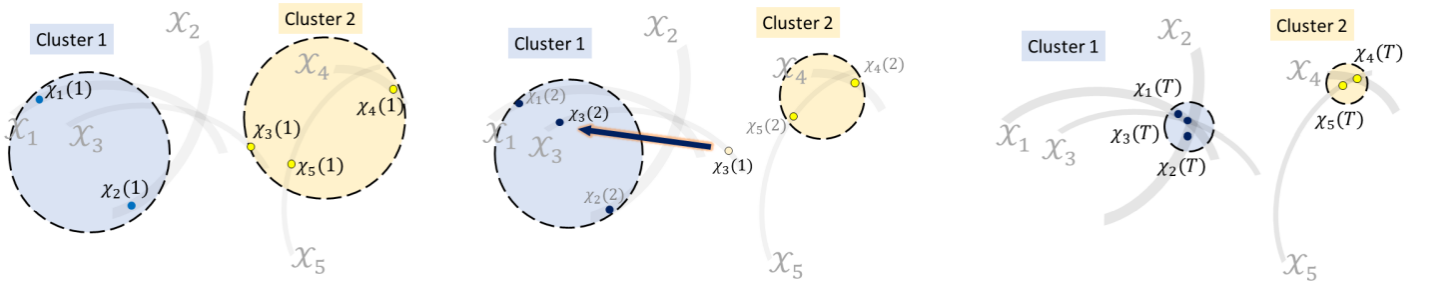
どれも”不確実さを平均化する”という発想

Proposal

Our idea：optimizing the best case (=**optimistic** attitude)

Greedy, and Optimistic Clustering (GOC) Algorithm

Minimizing the clustering radius



(a) Step (II): Temporal cluster assignments are obtained as $\hat{c}_1^\dagger(2) = \hat{c}_2^\dagger(2) = 1, \hat{c}_3^\dagger(2) = \hat{c}_4^\dagger(2) = \hat{c}_5^\dagger(2) = 2$.

(b) Step (III): Feature candidates and cluster assignments are updated ($\hat{c}_3^\dagger(2) = 2$ is reassigned to $\hat{c}_3(2) = 1$).

(c) After a sufficiently large number of iterations $T \in \mathbb{N}$, we expect to obtain the condensed clusters.

We may employ arbitrary clustering method as the clustering oracle.
(e.g. k-means, GMM, mean-shift, …)

Bit detailed..

Step 1) making a discrete set approximating the uncertainty set:

$$\zeta_i^{(1)}, \zeta_i^{(2)}, \ldots, \zeta_i^{(m_i)} \text{ are instances i.i.d. drawn from Unif.}(\mathcal{Z}_i), \tag{4}$$

and define an empirical feature uncertainty set as follows:

$$\tilde{\mathcal{X}}_i^{(m_i)} := \{\chi_i^{(j)}\}_{j=1}^{m_i} \subset \mathbb{R}^q, \ \chi_i^{(j)} = f(\zeta_i^{(j)}), \quad (j = 1, 2, \ldots, m_i). \tag{5}$$

$\{m_i\}_{i=1}^n \subset \mathbb{N}$ are hyperparameters, typically, $m_i = 100$. By assuming the following non-degenerate condition

Step 2) repeat the following steps:

1. Clustering the current representatives $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2, \ldots$ (e.g., K-means)
2. Compute the clustering centers
3. Assign new representatives to minimize the distance to cluster centers

$$\chi_i(t) \leftarrow \underset{\chi \in \tilde{\mathcal{X}}_i^{(m_i)}}{\operatorname{argmin}} \min_{k \in [K(t)]} \left\{ \| \chi - \hat{\mu}_k(\Xi(t-1), \hat{\boldsymbol{c}}^\dagger(t)) \|_2^2 + \lambda \mathrm{Pen}_i(\chi) \right\},$$

$$\hat{c}_i(t) \leftarrow \underset{k \in [K(t)]}{\operatorname{argmin}} \| \chi_i(t) - \hat{\mu}_k(\Xi(t-1), \hat{\boldsymbol{c}}^\dagger(t)) \|_2$$

Experiments

Experiments：Realistic Dataset

- oknakfm/GOC (github.com)
- N=275, $q = 3$ dimensional
- Each star i has discrete uncertainty set with $m_i = 101$ points
- 10 instances are generated with 10 different random seeds
- Same birthplace = same class

- Baseline: clustering with some representatives:

$$\bar{\chi}_i := \frac{1}{m_i} \sum_{\chi \in \tilde{\mathcal{X}}_i^{(m_i)}} \chi \in \mathbb{R}^3 \quad (i \in [n]),$$

GOC with K-means: scores are improved

Proposal

Large penalty
≈ K-means + representatives

(b) $K(0) = 50$

		$\lambda = 0$	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1$
NMI	GOC	0.880 ± 0.024	0.879 ± 0.027	0.871 ± 0.024	0.846 ± 0.026
	Baseline	0.839 ± 0.026			
F-measure	GOC	0.750 ± 0.039	0.752 ± 0.046	0.736 ± 0.041	0.694 ± 0.045
	Baseline	0.685 ± 0.048			
#clusters	GOC	46.1 ± 1.10	46.4 ± 0.97	47.2 ± 1.23	48.5 ± 0.85
#iterations	GOC	15.8 ± 3.23	15 ± 3.33	12.8 ± 3.74	7.4 ± 1.84

Optimistic+K-means

> Representatives+K-means

≧ UK-means

>> Pessimistic+K-means

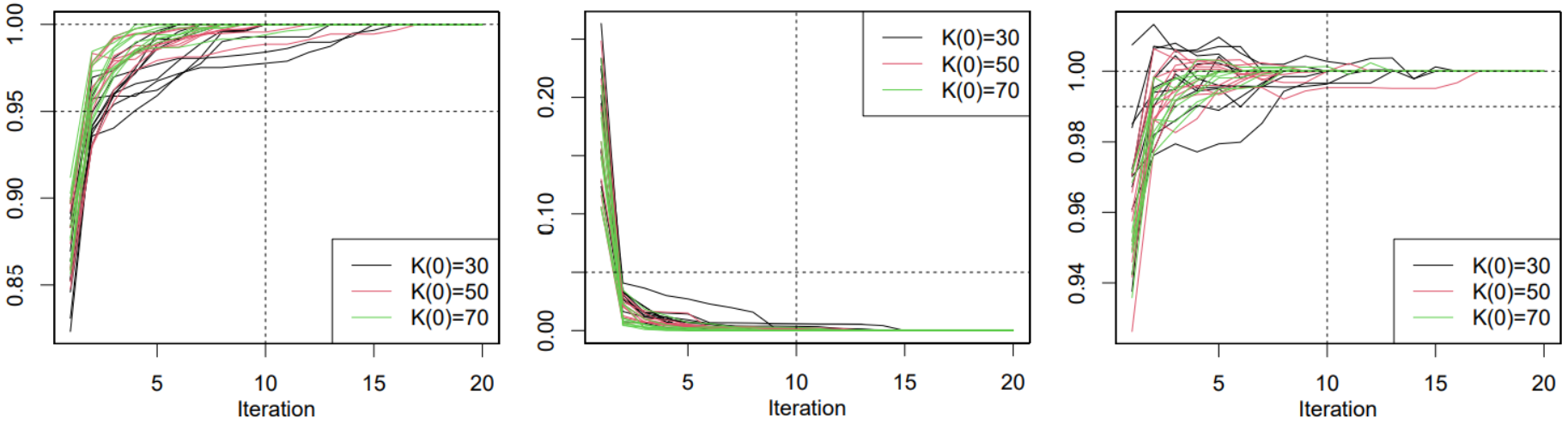
With different number of initial clusters

Table 2: K-means with fixed $\lambda = 0.01$ and increasing $K(0)$.

		$K(0) = 30$	$K(0) = 40$	$K(0) = 50$	$K(0) = 60$	$K(0) = 70$
NMI	GOC	0.834 ± 0.023	0.868 ± 0.023	0.879 ± 0.027	0.874 ± 0.027	0.878 ± 0.022
	Baseline	0.808 ± 0.020	0.832 ± 0.026	0.839 ± 0.026	0.839 ± 0.027	0.837 ± 0.021
F-measure	GOC	0.641 ± 0.026	0.719 ± 0.037	0.752 ± 0.046	0.741 ± 0.044	0.747 ± 0.036
	Baseline	0.604 ± 0.024	0.671 ± 0.039	0.685 ± 0.048	0.682 ± 0.047	0.673 ± 0.035
#clusters	GOC	28.4 ± 0.84	37.6 ± 0.84	46.4 ± 0.97	55.9 ± 1.52	64.2 ± 1.55
#iterations	GOC	15.8 ± 3.74	15.5 ± 5.32	15.0 ± 3.33	13.5 ± 1.96	12.2 ± 2.10

In any case, the proposed GOC improves the clustering scores

While 15 iterations is needed for exact convergence,
it almost converges in an early iteration.



(a) Convergence of cluster assignments: $\eta_1(t) = \mathrm{NMI}(\hat{\boldsymbol{c}}(t), \hat{\boldsymbol{c}}(\infty))$.

(b) Convergence of feature candidates: $\eta_2(t) = n^{-1}\sum_{i=1}^{n} \|\chi_i(t) - \chi_i(\infty)\|_2^2$.

(c) Convergence of NMI: $\eta_3(t) = \mathrm{NMI}(\hat{\boldsymbol{c}}(t), \boldsymbol{c}^*)/\mathrm{NMI}(\hat{\boldsymbol{c}}(\infty), \boldsymbol{c}^*)$.

Only 2~3 iterations is needed to obtain better results (than existing methods)

With different clustering oracles

(b) $K(0) = 50$

		K-means	K-medoids	GMM (`ClusterR`)	GMM (`Mclust+BIC`)
NMI	GOC	0.879 ± 0.027	0.879 ± 0.024	0.864 ± 0.022	0.807 ± 0.035
	Baseline	0.839 ± 0.026	0.841 ± 0.026	0.828 ± 0.023	0.786 ± 0.032
F-measure	GOC	0.752 ± 0.046	0.753 ± 0.031	0.718 ± 0.039	0.571 ± 0.059
	Baseline	0.685 ± 0.048	0.698 ± 0.042	0.654 ± 0.04	0.539 ± 0.057
#clusters	GOC	46.4 ± 0.97	46.3 ± 1.83	48 ± 1.05	24 ± 3.89
#iterations	GOC	15 ± 3.33	16.2 ± 2.94	16.7 ± 5.40	18.3 ± 9.11

Real data analysis

(in Hattori, Okuno, and Roederer, https://arxiv.org/abs/2207.04110)

Overlap to Roederer (2018)

k	$N_{\mathrm{member},k}$	(J_r, J_z, J_ϕ)	$(\sigma_{Jr}, \sigma_{Jz}, \sigma_{J\phi})$	σ_k	$\langle$[Fe/H]$\rangle$	$\sigma_{\mathrm{[Fe/H]}}$ ($q_{\mathrm{[Fe/H]}}$)	$\langle$[Eu/H]$\rangle$	$\sigma_{\mathrm{[Eu/H]}}$ ($q_{\mathrm{[Eu/H]}}$)	Comment[a]
		kpc km s^{-1}	kpc km s^{-1}	kpc km s^{-1}	dex	dex (percentile)	dex	dex (percentile)	
1	9	(129, 265, 1209)	(107, 140, 112)	121	−2.78	0.22 (0.56)	−1.64	0.32 (5.06)	Tier-1 – New
2	9	(942, 52, 102)	(113, 49, 104)	93	−1.65	0.25 (1.08)	−0.62	0.22 (0.78)	Tier-1 – $D^{3/3}$(R18), DTG10(Y20)
3	18	(464, 118, −711)	(123, 113, 117)	118	−2.37	0.37 (1.86)	−1.45	0.35 (1.20)	Tier-1 – $A^{4/4}$,$F^{2/3}$(R18), DTG38(Y20)
4	12	(115, 195, −889)	(63, 56, 67)	63	−2.42	0.33 (2.76)	−1.48	0.33 (2.78)	Tier-1 – $C^{3/4}$(R18)
5	5	(954, 354, 773)	(70, 114, 15)	78	−2.62	0.21 (4.92)	−1.60	0.20 (4.30)	Tier-1 – $G^{2/2}$(R18)
6	2	(67, 67, −2504)	(35, 65, 7)	43	−2.55	0.05 (10.11)	−1.36	0.36 (63.38)	Tier-3 – New
7	2	(971, 206, −2749)	(57, 30, 86)	62	−2.83	0.05 (10.83)	−1.82	0.11 (21.37)	Tier-2 – New
8	2	(3519, 3390, 163)	(8, 71, 110)	76	−1.66	0.06 (11.65)	−0.39	0.23 (43.04)	Tier-3 – New
9	6	(112, 873, 829)	(60, 193, 79)	125	−2.87	0.31 (12.58)	−1.65	0.33 (13.92)	Tier-1 – $B^{3/4}$(R18)
10	4	(878, 1190, 1208)	(51, 110, 39)	73	−2.27	0.26 (15.28)	−1.34	0.26 (16.82)	Tier-2 – New
11	2	(936, 388, 1757)	(101, 32, 0)	61	−1.39	0.08 (17.42)	−0.58	0.03 (7.01)	Tier-2 – New
12	2	(256, 1282, −1180)	(13, 54, 180)	108	−2.48	0.09 (18.15)	−1.40	0.07 (14.35)	Tier-2 – New
13	6	(450, 807, 47)	(77, 103, 106)	96	−2.51	0.35 (18.28)	−1.43	0.26 (6.44)	Tier-2 – New
14	2	(969, 254, −1940)	(47, 111, 167)	119	−2.90	0.11 (22.26)	−1.43	0.86 (97.43)	Tier-3 – New
15	18	(518, 153, −177)	(105, 109, 105)	106	−2.43	0.50 (30.62)	−1.51	0.49 (30.46)	Tier-4 – $E^{3/3}$,$F^{1/3}$,$H^{2/2}$(R18), DTG38(Y20)
16	13	(340, 214, 521)	(131, 120, 64)	109	−2.21	0.50 (34.18)	−1.25	0.54 (52.98)	$B^{1/4}$(R18)
17	4	(129, 1145, −419)	(153, 65, 88)	108	−1.94	0.42 (43.80)	−0.60	0.39 (38.32)	Tier-4 – New
18	3	(370, 243, −1889)	(32, 4, 38)	29	−2.64	0.41 (51.06)	−1.42	0.41 (51.54)	–
19	7	(107, 292, −287)	(88, 130, 86)	103	−2.12	0.53 (54.58)	−1.11	0.56 (67.74)	–
20	4	(1308, 342, −1017)	(18, 98, 49)	64	−2.20	0.49 (58.04)	−1.32	0.51 (63.82)	–
21	2	(2163, 215, −1052)	(12, 11, 39)	24	−2.36	0.43 (70.65)	−1.17	0.01 (2.44)	Tier-3 – New

先ほどの天体の分類では，

161天体をクラスタリングした．

(組み合わせ数： 12880通り)

～約半年前～

私　　服部さん

実は**40万天体**のデータがあります

(組み合わせ数： 約**800億通り**)

約621万倍: これまでは1秒で終わった計算も，72日かかります

背後にある数理(?)

arXiv:2407.10418

普通のk-meansクラスタリング

In this section, we interpret the results of the proposed GOC algorithm using K-means clustering (MacQueen et al., 1967) with a constant number of clusters $K(t) = K$. Let

$$\mathscr{S}_{K,n} := \left\{ \{s_{ik}\}_{i\in[n],k\in[K]} \subset \{0,1\} \mid \sum_{k=1}^{K} s_{ik} = 1,\ \forall i \in [n] \right\}$$

be a set of cluster-assignment indicators $\boldsymbol{s} = \{s_{ik}\}$, where $s_{ik} = 1$ denotes that individual i is assigned to cluster k (and 0 otherwise). Using a loss function

$$\ell(\Xi; \boldsymbol{s}) := \min_{\mu_1,\mu_2,\ldots,\mu_K\in\mathbb{R}^q} \sum_{k=1}^{K}\sum_{i=1}^{n} s_{ik}\|\chi_i - \mu_k\|_2^2,$$

conventional K-means clustering applied to a fixed instance $\Xi = (\chi_1, \chi_2, \ldots, \chi_n) \in \mathscr{A}_n$ computes the cluster assignments $\hat{\boldsymbol{c}} = (\hat{c}_1, \ldots, \hat{c}_n) \in [K]^n$ by solving the following problem:

$$\text{(Conventional)} \qquad \hat{c}_i := \arg\max_{k\in[K]} s_{ik}, \quad \hat{\boldsymbol{s}} := \arg\min_{\boldsymbol{s}\in\mathscr{S}_{K,n}} \{\ell(\Xi; \boldsymbol{s})\}.$$

Optimism vs Pessimism.

Further, the GOC algorithm equipped with K-means clustering is expected to solve the following minimization problem in a greedily manner:

$$\text{(Optimistic)} \qquad \hat{c}_i^{(\mathrm{GOC})} = \arg\max_{k\in[K]} \hat{s}_{ik}^{(\mathrm{GOC})}, \quad \hat{\boldsymbol{s}}^{(\mathrm{GOC})} := \arg\min_{\boldsymbol{s}\in\mathscr{S}_{K,n}} \min_{\tilde{\Xi}\in\mathscr{A}_n} \left\{ \ell(\tilde{\Xi};\boldsymbol{s}) + \lambda \sum_{i=1}^{n} \mathrm{Pen}_i(\tilde{\chi}_i) \right\}$$

for $\tilde{\Xi} = (\tilde{\chi}_1, \tilde{\chi}_2, \ldots, \tilde{\chi}_n)$. By contrast, we may consider a pessimistic variant of the GOC algorithm, i.e., the greedy and pessimistic clustering (GPC), when solving the following problem:

$$\text{(Pessimistic)} \qquad \hat{c}_i^{(\mathrm{GPC})} = \arg\max_{k\in[K]} \hat{s}_{ik}^{(\mathrm{GPC})}, \quad \hat{\boldsymbol{s}}^{(\mathrm{GPC})} := \arg\min_{\boldsymbol{s}\in\mathscr{S}_{K,n}} \max_{\tilde{\Xi}\in\mathscr{A}_n} \left\{ \ell(\tilde{\Xi};\boldsymbol{s}) + \lambda \sum_{i=1}^{n} \mathrm{Pen}_i(\tilde{\chi}_i) \right\},$$

Min-min型の回帰はかなり昔からある

- Errors-in-variables (EiV) regression

Covariate X has an error D

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} Z + D \\ \alpha_* + Z\beta_* + E \end{pmatrix} \in \mathbb{R}^2, \quad Z \overset{\text{i.i.d.}}{\sim} N(\mu_Z, \sigma_Z^2),$$

$$D \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_D, \quad \mathbb{E}(D) = 0, \quad \mathbb{V}(D) = \sigma_D^2,$$

$$E \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_E, \quad \mathbb{E}(E) = 0, \quad \mathbb{V}(E) = \sigma_E^2,$$

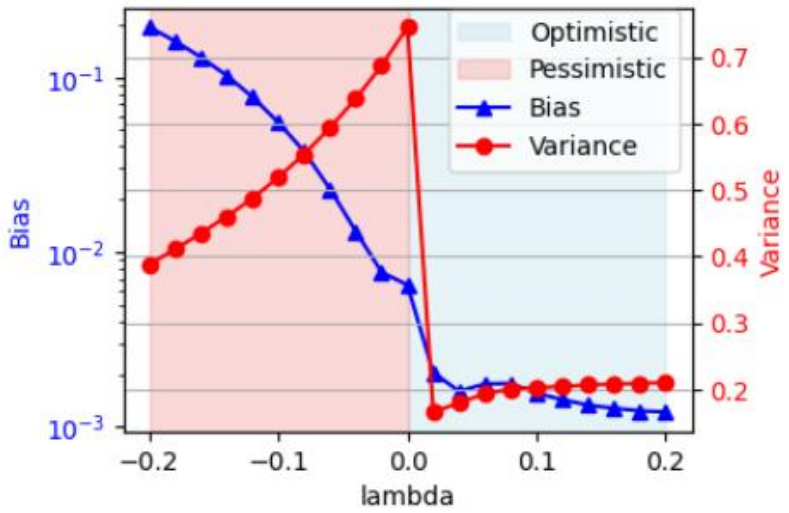
Deming regression is known to reduce the bias (Adcock 1878, Kummell 1879, Deming 1943, …)

$$\hat{\gamma}_n^{\mathrm{D}}(\lambda) := \underset{\gamma=(\alpha,\beta)\in\mathbb{R}^2}{\operatorname{argmin}} \min_{\tilde{\boldsymbol{x}}_n \in \mathbb{R}^n} \left\{ \|\boldsymbol{y}_n - \alpha \mathbf{1}_n - \beta \tilde{\boldsymbol{x}}_n\|_2^2 + \lambda \|\tilde{\boldsymbol{x}}_n - \boldsymbol{x}_n\|_2^2 \right\}$$

Deming regression selects X (as well as GOC)

Min-min推定 ⇔ (外れ値)ロバスト推定

- Gannaz (SC2007)による結果：
  応答変数のmin-min ⇔ 外れ値ロバスト推定 (Huber)

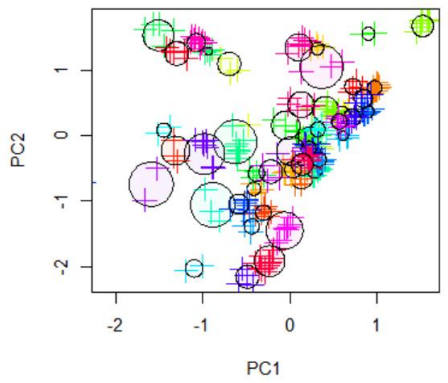


モデルが「本来的に正しく特定」されていて,
かつ「外れ値によって悪影響を受けている」なら
その影響を軽減できる.

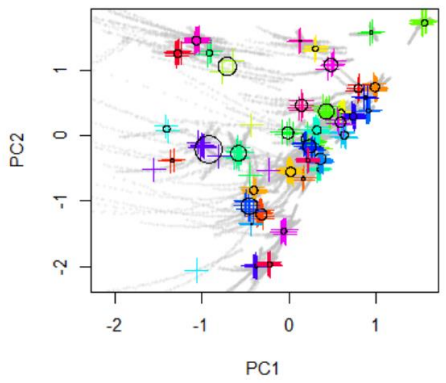
提案したGOCにもおそらく似たような効果が出ている.

Conclusion

- We proposed GOC, that considers the uncertainty for each individual.
- GOC is applied to realistic dataset, and the scores are improved.
- GOC is also applied to the real dataset (in H., O., and R., 2022).



(a) Conventional K-means ($K = 20$).



(b) GOC equipped with K-means ($K = 20$).