

特定された不確実性を利用する 貪欲かつ楽観的なクラスタリングと天文学への応用

奥野 彰文 (統数研/総研大/理研AIP)

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

<https://arxiv.org/abs/2204.08205> (methodology; under review)

<https://arxiv.org/abs/2207.04110> (application; accepted to **ApJ2023**)

自己紹介

- 阪大基礎工 情報科 数理 (学部)
- 阪大基礎工 数理 (修士)
- 京大情報学 数理 (博士)



https://twitter.com/public_aokn

専門：数理統計と統計的機械学習（？）

指導教員：下平英寿先生



<https://twitter.com/hshimodaira>

数理寄りでしたが、最近はもっと応用/計算に興味あり。

私の論文で今まで使った研究キーワード

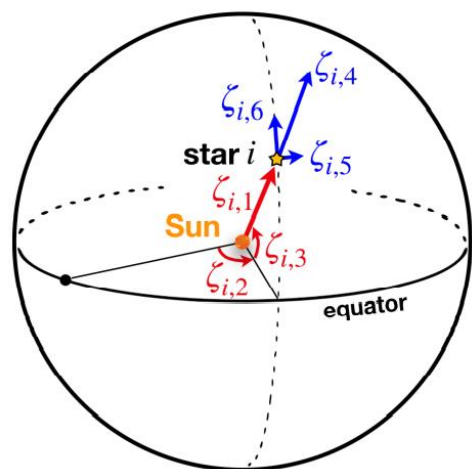
- グラフの統計解析 (相関分析)
- ニューラルネット
- 特徴量学習 (表現学習, 埋め込み)
- モデル選択/汎化の評価 (特にWAIC, 特異/過剰パラメータモデル)
- ノンパラ推定/ミニマックス最適性/収束レート
- 形状制約付きの関数推定 (可逆関数推定, (部分的)単調関数推定)
- ロバスト推定/ロバスト最適化

今日の話は (広義で) ここに含まれている

Research Background

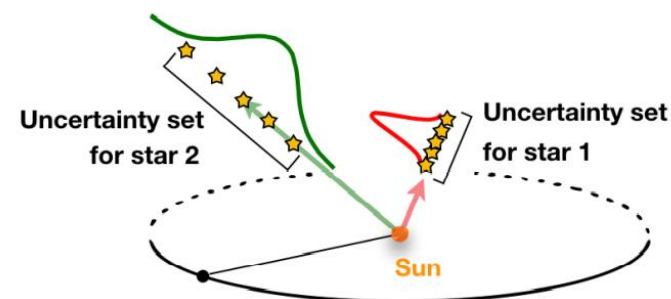
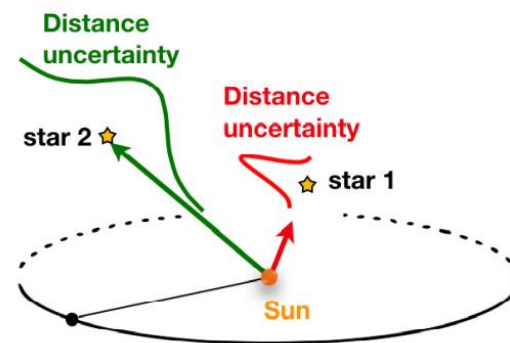
Setting (1/3)

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 = $1/\text{Distance}$	$\zeta_{i,4} = d(1/\zeta_{i,1})/dt$
$\zeta_{i,2}$ Azimuthal angle	$\zeta_{i,5} = d\zeta_{i,2}/dt$
$\zeta_{i,3}$ Polar angle	$\zeta_{i,6} = d\zeta_{i,3}/dt$

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

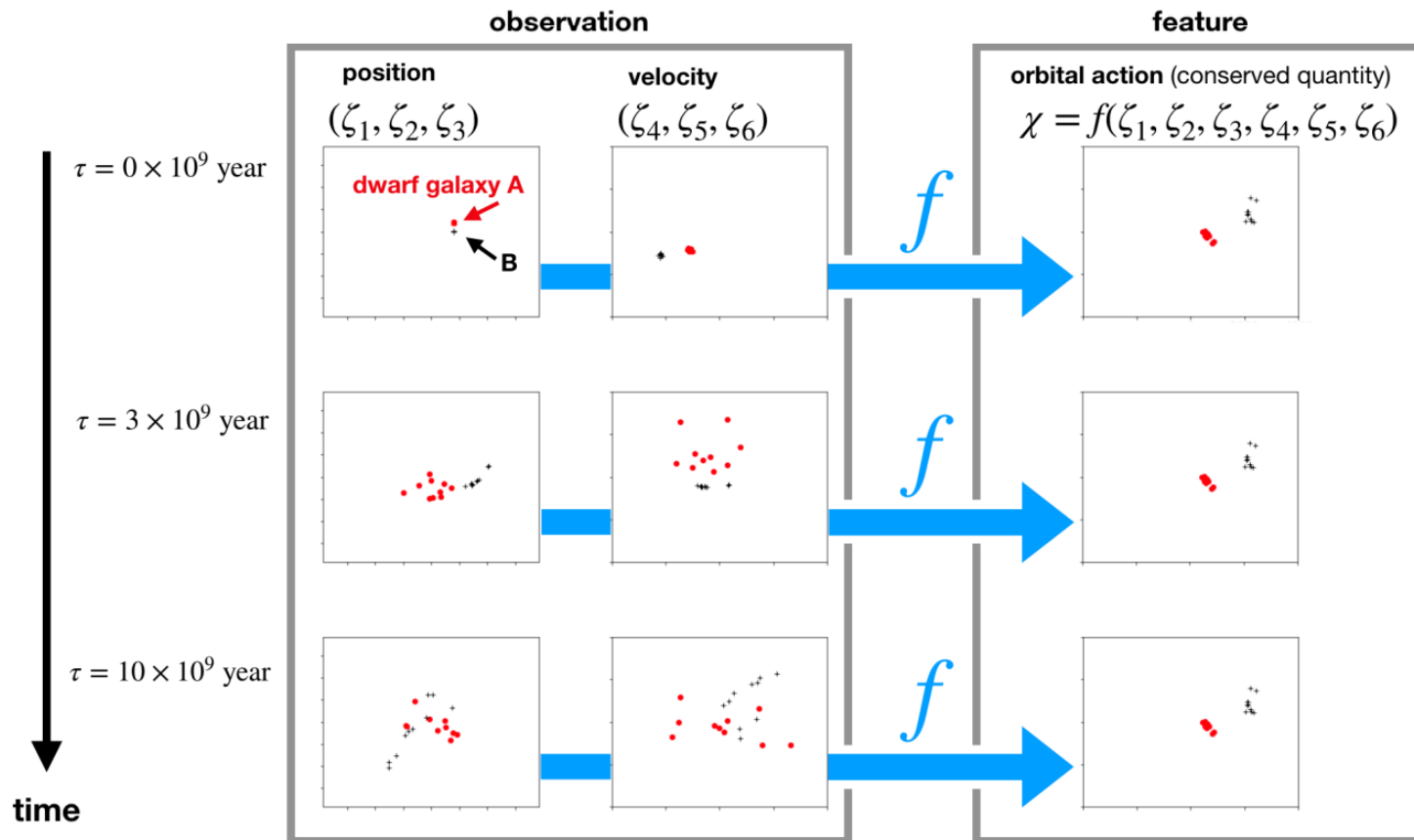


(b) Uncertainty sets for individual stars.

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

Setting (2/3)

Nonlinear pre-processing is important:
from observation to conserved feature



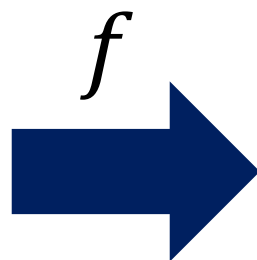
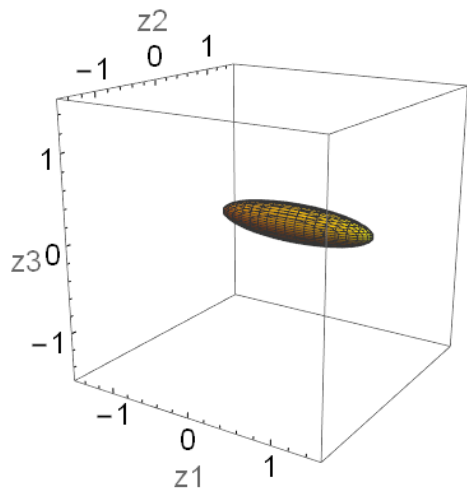
Setting (3/3)

Non-linear preprocessing twists the uncertainty sets!

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

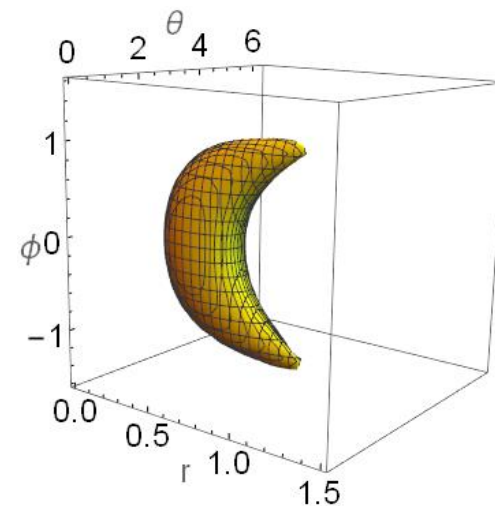
“Uncertainty set”

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



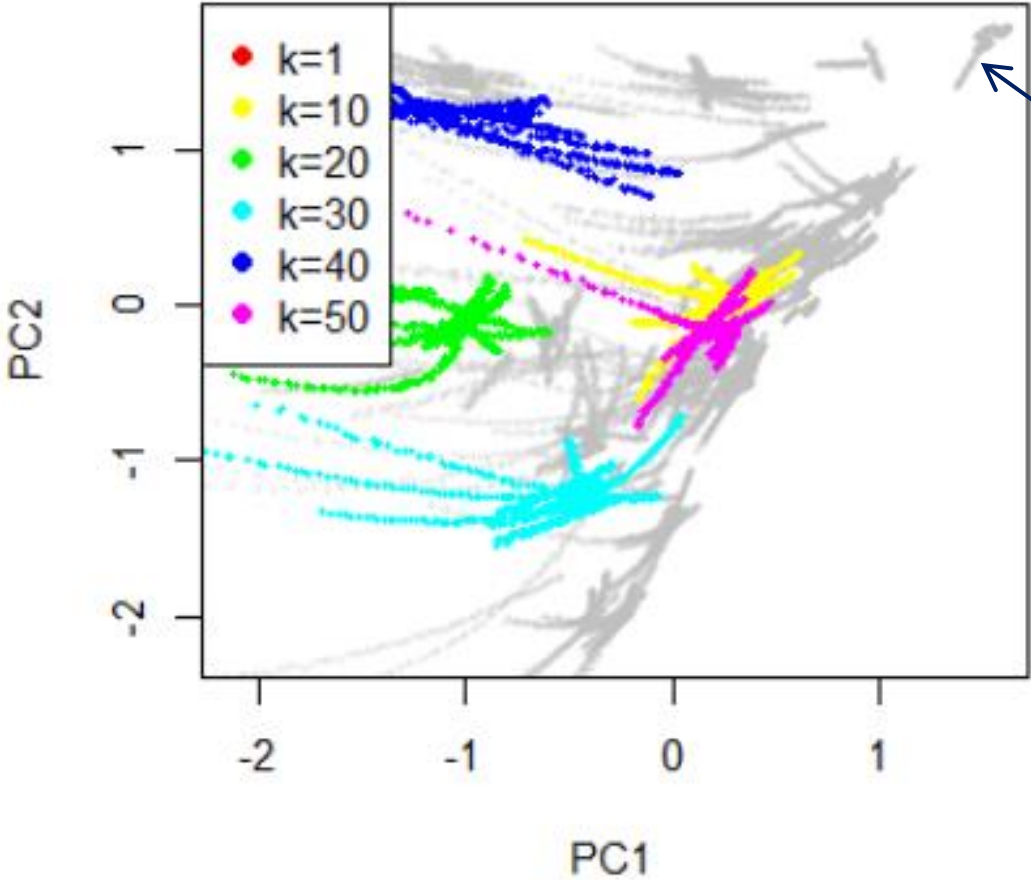
$X_i \in \mathbb{R}^q$: feature
(conserved quantity)

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



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

Example: feature uncertainty sets of stars



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

Previous Studies

In Astronomy

- 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 (<https://arxiv.org/abs/2207.04110>):

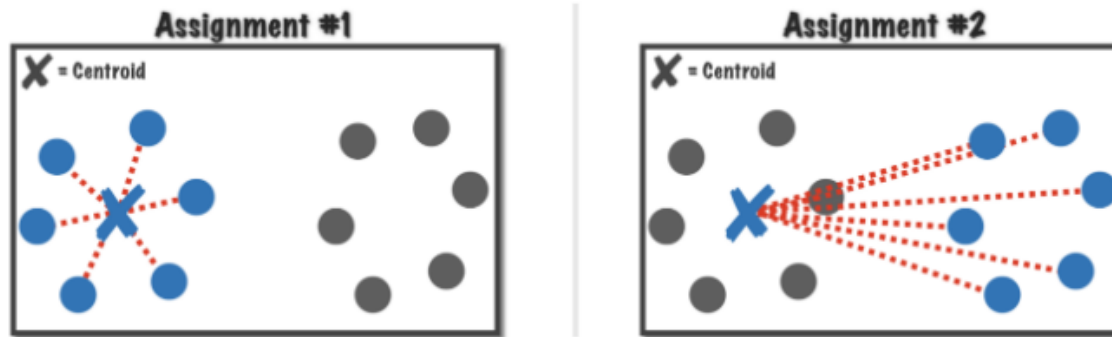
We use observations of $35 + 48 + 78 = 161$ stars!
Used in R(2018) New observations
Ignored in R(2018)

In Statistics

Most of classical approaches **does not consider the uncertainty.**

Ex) K-means (MacQueen et al., 1967)

$$\min_{\mu_1, \mu_2, \dots, \mu_k} \sum_{k=1}^K \sum_{i=1}^n \underbrace{S_{ik}}_{\text{Assignment}} \underbrace{\|x_i - \mu_k\|_2^2}_{\text{Cluster center}}$$



Which data points should be assigned to this centroid?

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

1) Clustering with the uncertainty

1. UK-means (Chau et al. KDD2006)

x_i follows a normal distribution p_i :

$$\min_{\mu_1, \mu_2, \dots, \mu_k} \sum_{k=1}^K \sum_{i=1}^n s_{ik} \underbrace{E_{p_i}(\|x_i - \mu_k\|_2^2)}$$

Minimizing the **expected** distance to cluster center μ_k

➤ Ngai et al. (ICDM2006)

Fast approximation of UK-means

➤ Lee et al. (ICDM2007 workshop)

Proves that UK-means = K-means applied to $x'_i := E_{p_i}(x_i)$

~~Proves meaninglessness of K-means...~~

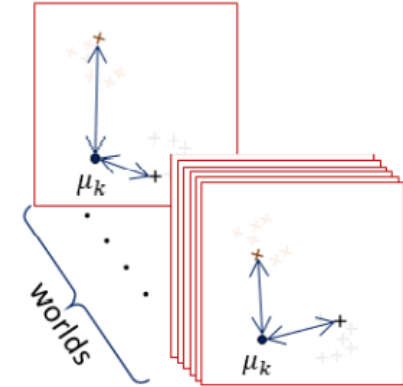
➤ Cormode and McGregor. (2008)

Another fast approximation of UK-means, UK-median

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

Parallel computation for **all the possible combinations**
→ aggregation.

Large computational complexity



3. Hierarchical clustering + Distribution distance (Kriegel and Pfeifle 2005; Jiang et al. 2013)

Compute the distance between distributions p_i, p_j of covariates X_i, X_j
→ distance-based clustering

All the approaches take **expectation**

2) Robust optimization

Optimizing the **worst case** (=pessimistic attitude)

Application to K-means clustering (Vo et al. 2016)

$$\max_{\tilde{x}_1 \in \mathcal{X}_1, \dots, \tilde{x}_n \in \mathcal{X}_n} \min_{\mu_1, \mu_2, \dots, \mu_k} \sum_{k=1}^K \sum_{i=1}^n s_{ik} \|\tilde{x}_i - \mu_k\|_2^2$$

Worst-case loss

A major idea in the field of optimization.

Also known as **adversarial attack**; now gathers attention from ML community.

Summary of previous studies

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



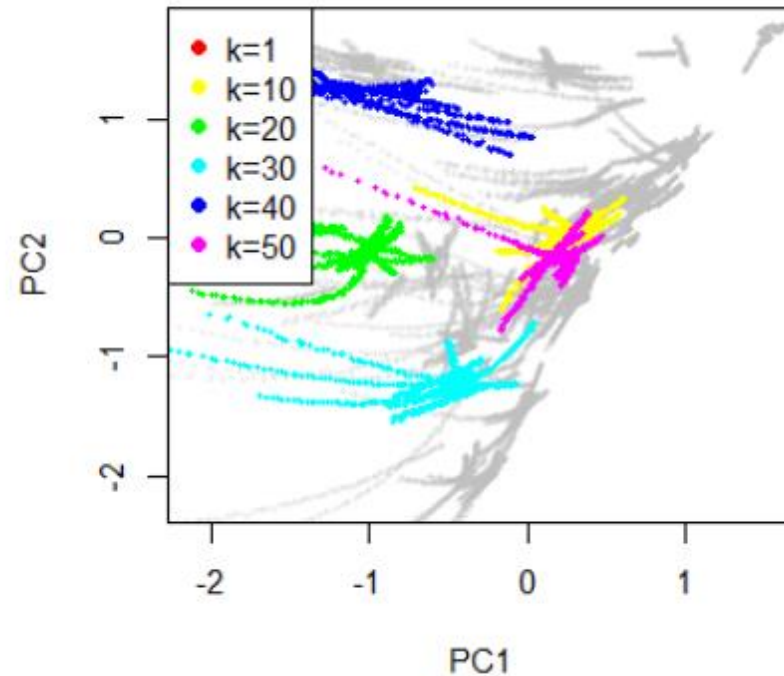
Taking expectation w.r.t. covariate uncertainty

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



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

Uncertainty set in our setting...



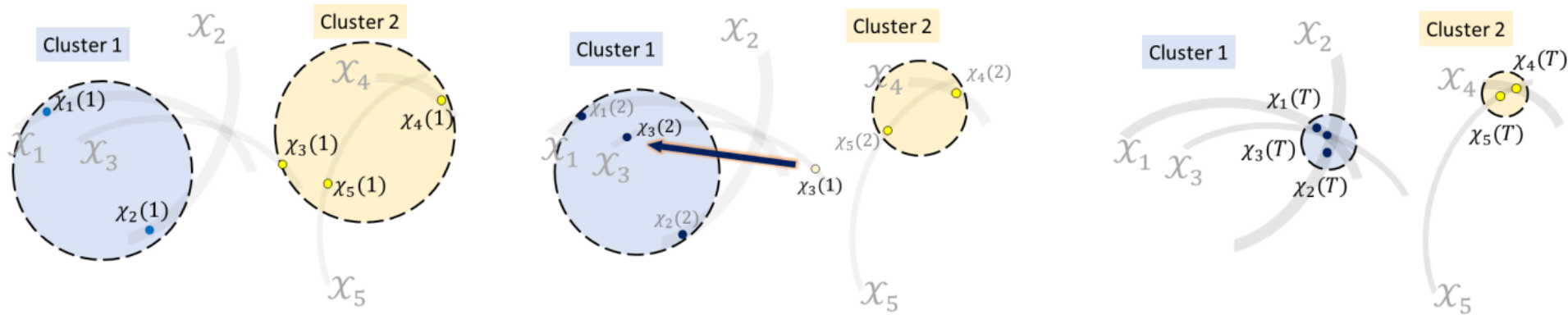
- No meaning to take expectation
- Worst case analysis results in devastating output
(only 1 big cluster is output then)

Proposal

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

Greedy, and Optimistic Clustering (GOC) Algorithm

Minimizing the cluster 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:

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

and define an empirical feature uncertainty set as follows:

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

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

Step 2) repeat the following steps:

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

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

$$\hat{\mathbf{c}}_i(t) \leftarrow \arg \min_{k \in [K(t)]} \|\chi_i(t) - \hat{\mu}_k(\Xi(t-1), \hat{\mathbf{c}}^\dagger(t))\|_2$$

In case of K-means...

$(\chi_1, \chi_2, \dots, \chi_n)$: covariates, (s_{ik}) : cluster assignments

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

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

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

Taking max = robust optimization

Experiments

Experiments : Realistic Dataset

- [oknakfm/GOC \(github.com\)](https://github.com/oknakfm/GOC)
- $N=275$, $q = 3$ 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:

$$\bar{\chi}_i := \frac{1}{m_i} \sum_{\chi \in \tilde{\mathcal{X}}_i^{(m_i)}} \chi \in \mathbb{R}^3 \quad (i \in [n]), \iff \text{Taking expectation w.r.t. covariates}$$

GOC with K-means: scores are improved

(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		
<i>F</i> -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

Proposal

Large penalty
 \approx K-means + representatives

Optimistic+K-means

> Representatives+K-means

\geq UK-means

>> Pessimistic+K-means (robust optimization)

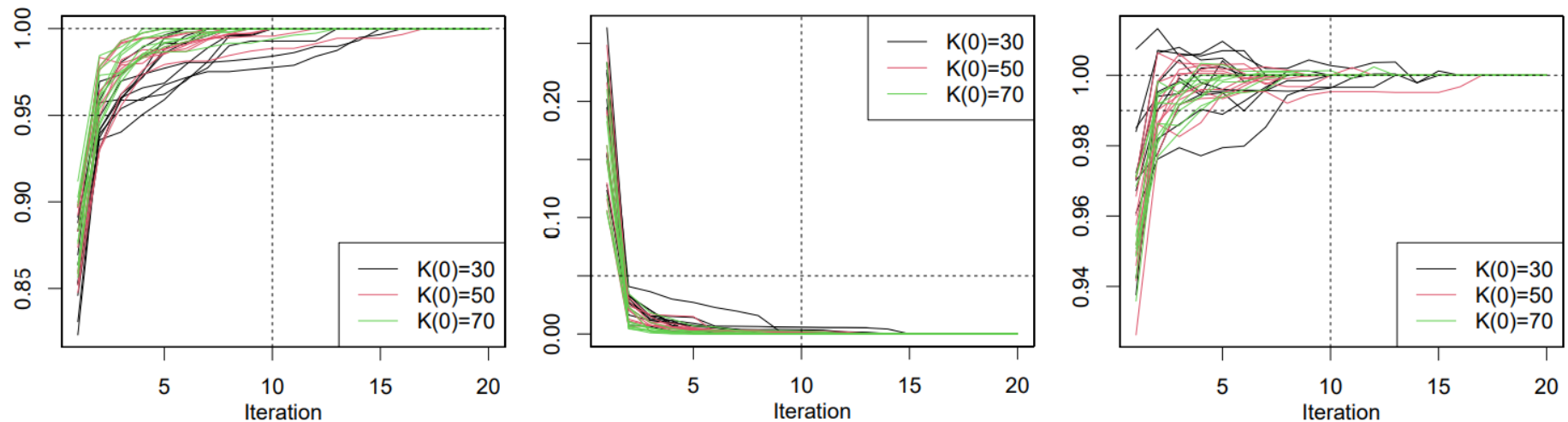
With different number of initial clusters

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: $\eta_1(t) = \text{NMI}(\hat{\mathbf{c}}(t), \hat{\mathbf{c}}(\infty))$.
 (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: $\eta_3(t) = \text{NMI}(\hat{\mathbf{c}}(t), \mathbf{c}^*) / \text{NMI}(\hat{\mathbf{c}}(\infty), \mathbf{c}^*)$.

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

Overlap to Roederer (2018)

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

k	$N_{\text{member},k}$	(J_r, J_z, J_ϕ) kpc km s ⁻¹	$(\sigma_{J_r}, \sigma_{J_z}, \sigma_{J_\phi})$ kpc km s ⁻¹	σ_k kpc km s ⁻¹	$\langle[\text{Fe}/\text{H}]\rangle$ dex	$\sigma_{[\text{Fe}/\text{H}]}$ ($q_{[\text{Fe}/\text{H}]}$) dex (percentile)	$\langle[\text{Eu}/\text{H}]\rangle$ dex	$\sigma_{[\text{Eu}/\text{H}]}$ ($q_{[\text{Eu}/\text{H}]}$) dex (percentile)	Comment ^(a)
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

Discussion

Similar approach can be found in regression problem:

- Errors-in-variables (EiV) regression

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

Covariate X has an error D

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

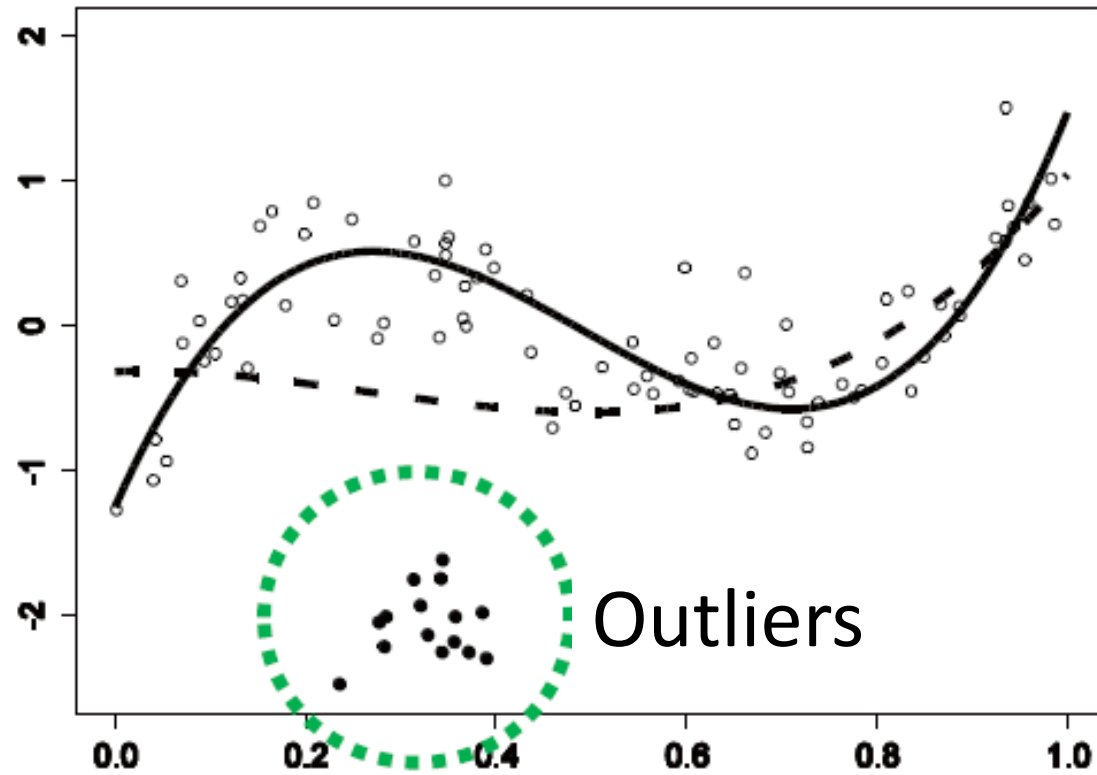
$$E \stackrel{\text{i.i.d.}}{\sim} \mathbb{Q}_E, \quad \mathbb{E}(E) = 0, \quad \mathbb{V}(E) = \sigma_E^2,$$

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

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

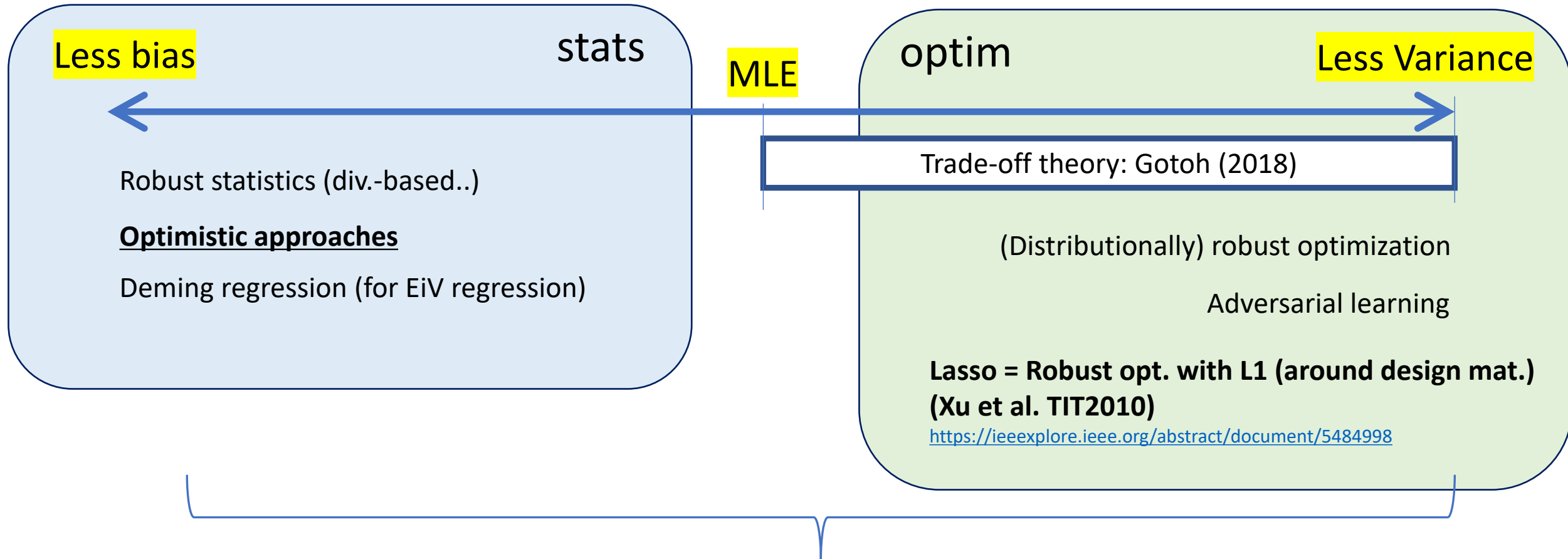
Deming regression selects X (as well as GOC)

Robust statistics is optimistic



Robust statistics aims to ignore outliers
(by considering the **optimistic case**)

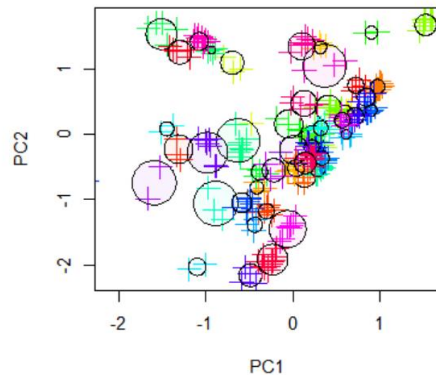
Bias-variance tradeoff under misspecification



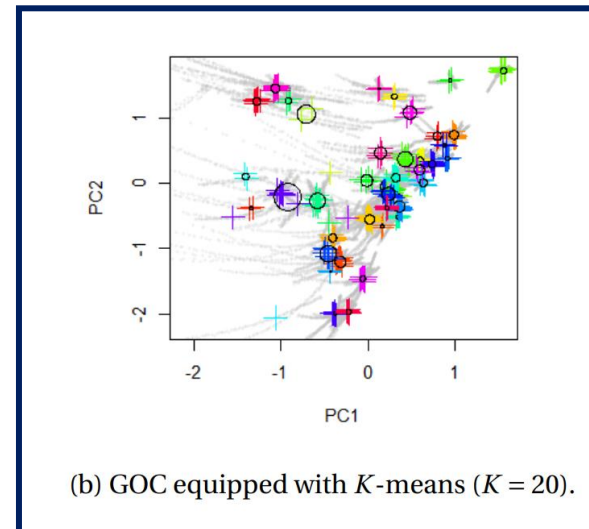
Can we integrate these two?
= Robust statistics applied to Lasso (?)

Conclusion

- We proposed GOC, that considered the uncertainty for each individual.
- GOC was applied to realistic dataset, and the scores are improved.
- GOC was also applied to the real dataset (in H., O., and R., ApJ2023).



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



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

<https://arxiv.org/abs/2204.08205> (methodology; under review)

<https://arxiv.org/abs/2207.04110> (application; accepted to **ApJ2023**)