

# 楽観的なクラスタリングと天文学への応用

副題：一人の統計手法研究者から見た科学応用研究

奥野彰文<sup>1,2</sup>

<sup>1</sup>統計数理研究所 <sup>2</sup>理化学研究所

# 自己紹介



おくの あきふみ  
奥野 彰文  
博士(情報学)



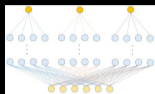
ニューラルネット (深層学習)



伝統的な統計学

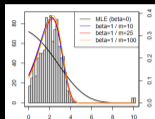
小サンプル領域での機械学習手法研究

## Computation (Machine Learning)

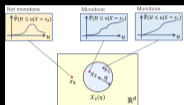


WAIC + NN  
Okuno and Yano (JCGS2023)

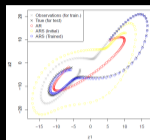
$$\mathbb{E}_\tau \left( \left\| \frac{\partial}{\partial \theta} d_{\beta}(\hat{Q}, P_{\theta(\tau)}) \right\|^2 \right) \rightarrow \text{in prob. } 0, \quad (T \rightarrow \infty).$$



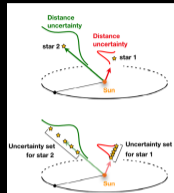
Okuno (2307.05251)



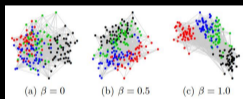
Okuno and Harada (2303.17823)



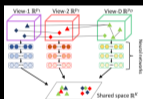
Okuno et al. (2306.16593)



Okuno and Hattori (2204.08205)  
Hattori, Okuno, Roederer (ApJ)



Okuno and Shimodaira (AISTATS2019a)

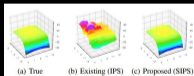


Okuno et al. (ICML2018)



$$\mathcal{V}(\hat{f}_{n,h}(x, x')) = O\left(n^{-\min\{2s, 2\beta+s\}/(s+d)}\right)$$

Okuno and Yano (SPL2023)



Okuno et al. (AISTATS2019b)



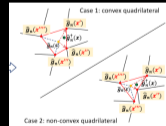
Mizutani, Okuno et al. (2005.00670)

$$\mathcal{E}(\hat{g}_n) = O(n^{-2\beta/(2\beta+d)})$$

Okuno and Shimodaira (NeurIPS2020)  
Cao, Okuno et al. (2112.13951)

$$\inf_{\hat{f}_n} \sup_{f_* \in \mathcal{F}_{INV}^{Lip}} R_{INV}(\hat{f}_n, f_*) \asymp n^{-2/(2+d)}$$

Okuno and Imaizumi (2112.00213)



## Visualization/Integration (Representation Learning)

## Theory (Mathematical Statistics)

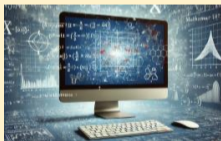
# 最近の個人的興味: 異分野連携



天文学



核融合プラズマ



計算代数/HPC

(推進中)

企業共同研究

▶ まだ駆け出しですが、少しずつ進めています。

## 嬉しいこと

- ▶ 新しい問題設定がたくさんある（嬉しい）
- ▶ 何かの役に立っている感がある（嬉しい）
- ▶ 知り合いがたくさん増える（嬉しい）

## つらいこと

- ▶ お互いに言っていることがよくわからない
- ▶ 論文化しづらい

一般論の前に、私の事例をお話します。

## 研究事例: 楽観的なクラスタリングと天文学への応用

- ▶ Kohei Hattori, Akifumi Okuno, and Ian U. Roederer. “Finding r-II sibling stars in the Milky Way with the Greedy Optimistic Clustering algorithm”. *Astrophysical Journal* (2023).
- ▶ Akifumi Okuno and Kohei Hattori. “A Greedy and Optimistic Approach to Clustering with a Specified Uncertainty of Covariates”. arXiv:2204.08205. Re-submitted

# 事の始まり

大部屋の真ん中で昼食を食べていた, ある日:



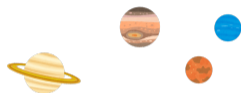
統計処理について  
少し聞きたいんですが...

**服部公平先生**

(国立天文台から統計数理研究所に出向されている, 天文学者)

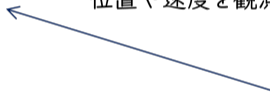
Several hours later...

服部先生の説明により，問題の概形が分かってきた。



現在

位置や速度を観測



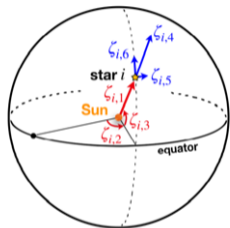
NASAなどが持つ天文台

- ▶ どの天体とどの天体が同じ銀河に由来するのか，知りたい。



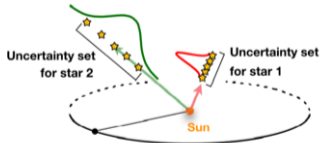
More hours later...

服部先生の説明により，問題の詳細も分かってきた。



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.

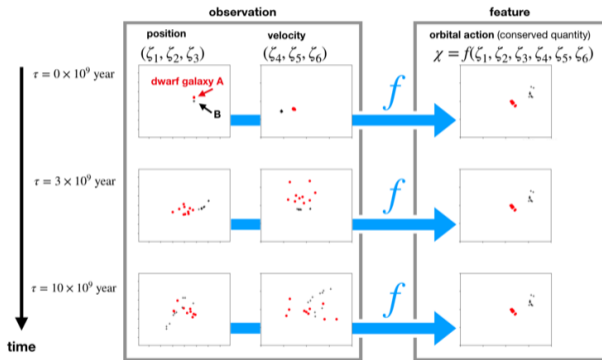


(b) Uncertainty sets for individual stars.

▶ 各天体は観測の精度がバラバラ。（太陽系から遠い天体ほど精度が低め）

# More and more hours later...

服部先生の説明により、問題のさらなる詳細も分かってきた。



- ▶ 観測値を時間不変な特徴量に非線形変換している。
- ▶ 正規性などは完全に破壊される。

## 典型的な統計手法の設定との乖離

各個体を表す確率変数 $X_i$ について、よくある仮定：

$$\text{i.i.d. } X_i \sim N(0, \Sigma) \quad (\text{または, } X_i \sim N(\mu_{k_i}, \Sigma)).$$

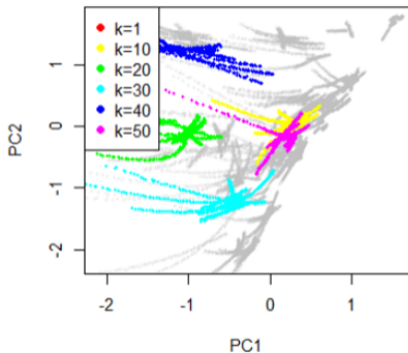
今回の設定：

$$\text{個体ごとに異なる分布 } X_i \sim p_i.$$

- ▶ 計算可能な形で、複雑な設定をどう扱うかが問題だった。

# 不確実性集合

不確実性集合  $\mathcal{X}_i$ : この集合は確率評価  $\mathbb{P}(X_i \in \mathcal{X}_i) \geq 1 - \eta$  を満たす.



(シミュレーション)  $n = 275$ 天体の不確実性集合の2次元可視化

# 不確実性をどう扱うか?

当初, 3つのアプローチを考えていました.

クラスタリングの損失 $L(X_1, X_2, \dots, X_n; c_1, c_2, \dots, c_n)$ について

- (1) 期待値を取る:  $\mathbb{E}_{X_1 \sim \hat{p}_1, \dots, X_n \sim \hat{p}_n} [L(X_1, X_2, \dots, X_n; c_1, c_2, \dots, c_n)]$ .
- (2) 最悪ケースを取る:  $\max_{x_1 \in \hat{\mathcal{X}}_1, \dots, x_n \in \hat{\mathcal{X}}_n} L(x_1, x_2, \dots, x_n; c_1, c_2, \dots, c_n)$ .
- (3) ベストケースを取る:  $\min_{x_1 \in \hat{\mathcal{X}}_1, \dots, x_n \in \hat{\mathcal{X}}_n} L(x_1, x_2, \dots, x_n; c_1, c_2, \dots, c_n)$ .

本研究では, (3)をGreedyに最適化する手法を提案しました. 詳細は省略.

▶ シミュレーションで確かめると, ほぼすべての設定で既存法より良い結果に.

# 実際のデータ解析へ

- ▶ 既存研究 Roederer (2018): 83天体のうち48天体を除いた35天体のクラスタリング
- ▶ 本研究 (Hattori et al., ApJ 2023): 83天体+新規観測=161天体のクラスタリング

$k$	$N_{\text{member},k}$	$(J_r, J_z, J_\phi)$	$(\sigma_{J_r}, \sigma_{J_z}, \sigma_{J_\phi})$	$\sigma_k$	$([\text{Fe}/\text{H}])$	$\sigma_{[\text{Fe}/\text{H}]}$ ( $q_{[\text{Fe}/\text{H}]}$ )	$([\text{Eu}/\text{H}])$	$\sigma_{[\text{Eu}/\text{H}]}$ ( $q_{[\text{Eu}/\text{H}]}$ )	Comment <sup>(a)</sup>
		kpc km s <sup>-1</sup>	kpc km s <sup>-1</sup>	kpc km s <sup>-1</sup>	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 <sup>3/3</sup> (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 <sup>4/4</sup> , F <sup>2/3</sup> (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 <sup>3/4</sup> (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 <sup>2/2</sup> (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 <sup>3/4</sup> (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 <sup>3/3</sup> , F <sup>1/3</sup> , H <sup>2/2</sup> (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 <sup>1/4</sup> (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

## さらにその先へ

- ▶ 数理的側面 (Okuno, arXiv:2407.10418)
  - ▶ 誤差変数モデルの最尤推定として解釈できる.
  - ▶ 外れ値頑健推定と類似の効果が考えられる.
- ▶ 計算的側面 (Hattori, Okuno, and Terada, ongoing)
  - ▶ 40万天体のデータがあるそうです…!
  - ▶ 計算量が単純には620万倍に. 高速化により, これでも計算が可能です…!



天文学者と統計学者の意見交換会@国立天文台

## 異分野連携に関する所感



異分野連携は一日にして成らず.

- ▶ 最初はお互いに言葉が通じない.
- ▶ 信頼関係がなければ研究は進まない.

服部先生とは普段一緒にお昼ご飯を食べていたのも大きかったです.  
共同研究は人と人との営み.

# 戦略的研究プロジェクト（2022年8月～）

統計数理核融合



DATA SCIENCE  
プラズマ物理と相補的なプラズマデータに対する統計数理モデリング  
情報システム研究機構 戦略的研究PJ  
2022-SRP-13  
powered by DALL-E3

プロジェクト概要

「プラズマ物理と相補的なプラズマデータに対する統計数理モデリング」（代表・三分一史和）は 大学共同利用機関法人情報・システム研究機構の戦略的研究プロジェクトに採択された、統計数理研究所と核融合科学研究所を中心とした共同研究プロジェクトです。

核融合におけるプラズマは極めて高温で複雑な非線形環境にあります。プラズマの基本現象の理解は進んでいるものの、異なる現象の相互作用を統合的に理解し制御することが核融合発電などの実現には不可欠です。世界各国が協力して進めるITERプロジェクトなどがその例です。しかし、プラズマの電流が突然消失する「ディスラプション」など未解明の課題も多くあり、本研究ではデータ駆動アプローチとモデル駆動アプローチを併用してプラズマの挙動の予測・制御を目指します。統計数理的手法を活用し、リアルタイムの予測や乱流データの解析を行います。統計数理コミュニティと核融合科学コミュニティの協力により学術界や産業界への貢献と、それに伴う「統計数理核融合学」の創成を目指しています。

ROIS戦略的研究プロジェクト



<https://statplasma.github.io>

- ▶ (文字通り) 何をコメントされているか全くわからない。



日本大学  
佐々木真先生

各構造からのエネルギー輸送は

$$J(\zeta | \alpha, \beta) = \int \phi_\zeta [\phi_\alpha, \nabla_\perp^2 \phi_\beta] d^3x$$

で計算できます。(計算して図に入れてもらえますか?)

- ▶  $\nabla_\perp$ とは?  $d^3x$ とは?
- ▶ そもそもエネルギー輸送って何ですか…?
- ▶ 統計手法を使う使わない以前に、何をやりたいのか理解するところから始まる。

# 異分野連携あるある：難しいと思っているポイントが全く違う

(科学者視点)

- ▶ (学部生レベルの) 電磁気学だから簡単にわかってもらえる (だろう)

(統計学者視点)

- ▶ (学部生レベルの) 単純なクラスタリングだから知られている (だろう)

時間をかけて少しずつ歩み寄るしかない。



- ▶ 手法が古典的なものでも応用側にはインパクトがあったり、
- ▶ 応用としては当然でも手法側の想像しない設定があったり。

逆に言えば、(連携研究では) 必ずしも難しい理屈が必要にはならない。

- ▶ 「特定の手法が使える実応用問題」を探すのは難しいが、
- ▶ 「特定の实応用問題に適した手法」を作るのは比較的容易。

「応用側が何をやりたがっているか？」を適切に切り出すことが重要。

## お互いにメリットがあるとよい

- ▶ 2023年度核融合プラズマ学会: 統計側から2件の講演
- ▶ 2024年度統計関連学会連合大会: 核融合プラズマ側から3件の講演
- ▶ 2024年度核融合プラズマ学会: 統計側から2件+の講演(?)

片側に“だけ”負担が集中すると長続きしない

## いろいろな学会を見て思ったこと

- ▶ どの学会もそこそこ煮詰まっている。
- ▶ 多くの方は現状維持派，または何かアクションを起こすほどバイタリティはない。
- ▶ <10% くらい改革派が居て，突破口を探している。



核融合研究所  
横山雅之先生

核融合プラズマと統計数理をかけあわせた  
統計数理核融合学を作しましょう…!

## Take-home message

結局，少しずつ知り合いを増やして，少しずつ相互理解するしかない

- ▶ (既存のグループに入るのでなければ) 異分野連携は一日にして成らず.
- ▶ 知り合いを作り，何度も集まり，議論してようやく芽が出る.
- ▶ “お互いに歩み寄れる”研究者と知り合ったら，連絡が途絶えないように…!

雑談・ご相談など，お気軽に [okuno@ism.ac.jp](mailto:okuno@ism.ac.jp) まで.