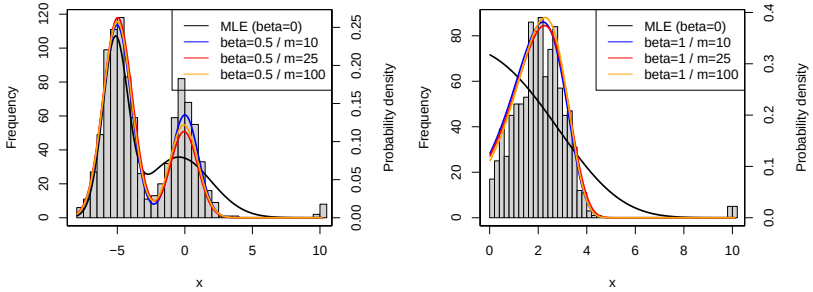
IBIS2024企画セッション1：サイエンスと機械学習
「一人の統計手法研究者から見た科学応用研究」

奥野彰文[1,2]

[1]統計数理研究所　[2]理化学研究所

- フラフラと色々やっている統計学/機械学習手法の研究者です.



(a) 混合正規分布　　(b) ゴンペルツ分布

Figure: Okuno (AISM2024) Fig.1 より転載. 昨年のIBISでの発表：
「一般の確率モデルに対するロバストダイバージェンスの最小化」(優秀賞)

- マニアック手法研究が好き.

手法研究者 meets 異分野連携研究 (科学研究)

天文学

核融合プラズマ

計算代数など

（推進中）

企業共同研究

- 素人が少しずつ何かをやろうとしているところ.
- まだ始めたばかりですが，少しずつ進めています.

今日の話

科学系の知識がない(素人の)手法研究者が科学者と対峙すると，どうなったか?

同じく新規参入したい人，特に手法研究者・科学者両方に向けて
素人視点での経験談が共有できればと思います.

理想

手法研究

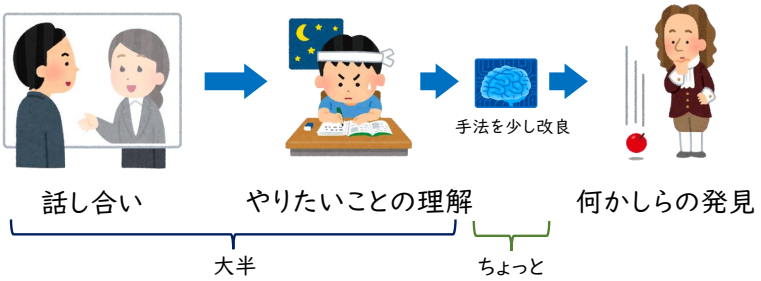
すごい手法の提案

すごい発見

- …という感じでトントン拍子にはいかない.
- 「手法としての差分」が役立つ問題は，そんなに簡単に見つからない.

現実

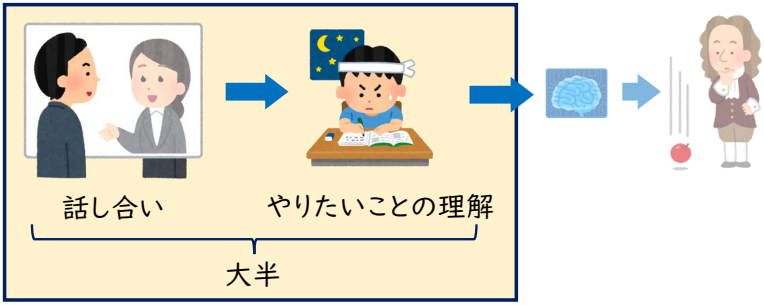
- 「相手の言っていることを理解する」までで大半の時間が消費される.



- AI for Science?
- MA[1] for Science...

[1] Multivariate Analysis

一番重要なこと



- 「いかにして前半部分の時間を用意するか」が一番重要.
- ひとたび（手法の言葉で）目的が明確化されれば，誰がやってもだいたい同じ.
- 翻訳して問題に落とし込むまでが腕の見せ所.

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

服部先生の説明により，問題の概形が分かってきた.

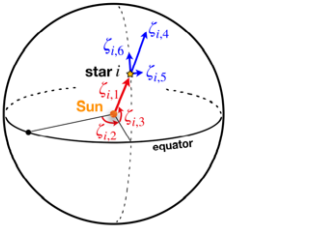
現在

位置や速度を観測

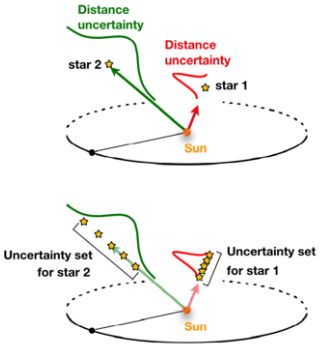
NASAなどが持つ天文台

- 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.

服部先生の説明により，問題の詳細も分かってきた.



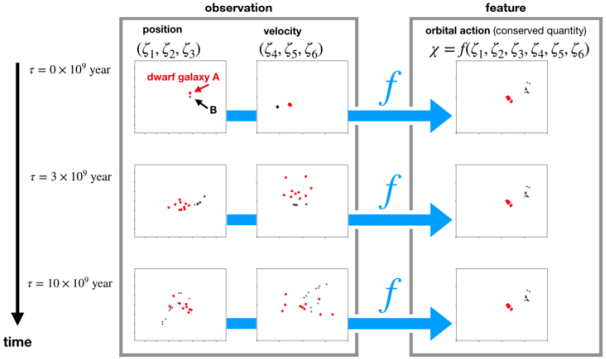
(a) Stellar position and velocity observed from the Sun.



(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）

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：

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

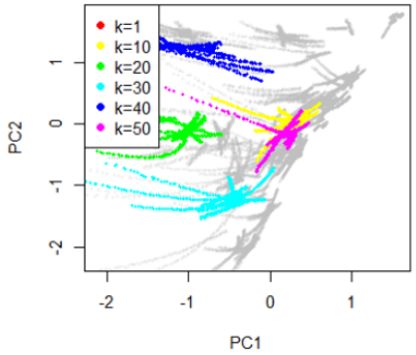
今回の設定：

$$\text{個体ごとに異なる分布 } X_i \sim p_i.$$

- 計算可能な形で，複雑な設定をどう扱うかが問題だった.
- 今までこんな設定を考えたことがなかった ($\Leftarrow$重要!).

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 $L(X_1, X_2, \ldots, X_n; c_1, c_2, \ldots, c_n)$ について

(1) 期待値を取る: $\mathbb{E}_{X_1 \sim \hat{p}_1, \ldots, X_n \sim \hat{p}_n}[L(X_1, X_2, \ldots, X_n; c_1, c_2, \ldots, c_n)]$.

(2) 最悪ケースを取る: $\max_{x_1 \in \hat{\mathcal{X}}_1, \ldots, x_n \in \hat{\mathcal{X}}_n} L(x_1, x_2, \ldots, x_n; c_1, c_2, \ldots, c_n)$.

(3) ベストケースを取る: $\min_{x_1 \in \hat{\mathcal{X}}_1, \ldots, x_n \in \hat{\mathcal{X}}_n} L(x_1, x_2, \ldots, x_n; c_1, c_2, \ldots, c_n)$.

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	$N_{\mathrm{member},k}$	(J_r, J_z, J_ϕ)	$(\sigma_{J_r}, \sigma_{J_z}, \sigma_{J_\phi})$	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

さらにその先へ

- 数理的側面 (Okuno, arXiv:2407.10418, 今日の午後のポスター [1-R-002])
  - 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

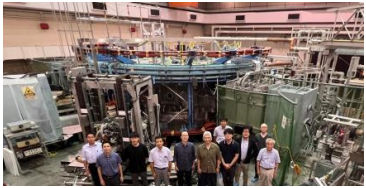
あらためて思うこと

（科学系に限らず）
異分野連携はコミュニケーションコストが高い.

- 最初はお互いに言葉が通じない.
- 9割型，緩やかにフェードアウト.
- （手法的には）すごく簡単なところで解ける問題があったりする.

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.

ROIS戦略的研究プロジェクト（2022年8月〜）



https://statplasma.github.io

- 定期的な集会の重要性を感じています．

- 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，
  「特定の実応用問題に適した手法」を作るのは比較的容易．
- 解析手法のすごさ $\not\Rightarrow$ 応用のニーズ

「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とにかく重要．

現状の問題: コミュニケーションコスト $\gg$ 異分野連携の短期的メリット

基礎研究者か? 応用研究者か?

- 統計学者は基礎/応用両方の立場から共同研究がありうる.

基礎側

応用側

- 両方を体験して思うところはたくさんあります.
- どれくらい個別案件のドメイン知識をつけるべきかなど，いろいろ模索中…

宣伝


大学共同利用機関法人 情報・システム研究機構
統計数理研究所
The Institute of Statistical Mathematics

お問合せ サイトマップ アクセス ポリシー English 所内専用

検索

ホーム 研究所について 研究活動 共同利用 刊行物案内 産学連携 プロジェクト 大学院教育

プレスリリース一覧

index

データ駆動科学における共創型研究拠点形成事業「バーチャルラボ」を始動
～第一弾として四つのバーチャルラボを設立～

ISM2024-03
2024年11月1日

2024年11月、統計数理研究所（所在地：東京都立川市、所長：椿 広計）は、産学官および国内外の研究者らが分野・組織・国境の垣根を越え、データ駆動科学の新学術創成やオープンイノベーションを推進する「バーチャルラボ」共創拠点形成事業を始動しました。Society 5.0を牽引するディープテクノロジーである人工知能（以下「AI」という。）は、科学・産業・国民生活のあらゆる領域を変革する可能性を秘めています。データ科学はその学術的基盤を担います。AIやデータ科学を取り巻く研究環境は大きな変換を遂げつつありま す。欧米やアジア諸国は、巨大IT企業を巻き込みながら膨大な研究費を投じて、AI関連の諸分野において熾烈

- 若手研究者主導で学際融合を図る「諸科学統計数理ラボ」（長：矢野恵佑准教授）
- 統計数理を介して科学の発展に資する創発拠点へ.
- 色々な分野からご参加いただき，今月1日に始動しました！

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) 研究開始まで時間がかかる.
- 知り合いを作りに行き，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"研究者はすごく貴重.
- いい感じの仕組みが作りたいなぁと思っています.

連絡先：okuno@ism.ac.jp