天文学データと統計数理手法の交差点

服部　公平　　統計思考院　助教　**khattori [AT] ism.ac.jp**

2023年5月26日　統計数理研究所　オープンハウス

0. 余談

私は2020年9月に国立天文台から統計思考院に出向し、３年近くが経ちます。「給料は国立天文台が出すので、統計学者に囲まれてたくましくなって下さい」というポストなので、さまざまな統計学者・数学者の方々と知り合いたいところですが、コロナ禍でお会いしたことのない所員の方も多いというのが現状です。先日、統計数理セミナーの後に卓球の会があると聞いて飛び入り参加しましたところ、「統計数理セミナーの司会の人」程度には認識されているようですが、これを機にお話する機会があれば良いと思っております。

1. 概要

天文学データは、統計数理手法の応用先として興味深い対象である。
データは無料公開が原則であり、機密情報もないため、アカデミック志向の統計学者とは相性が良い。
ここでは、天文学データがどのような分野と関連するかについて、私の周辺の天文学の話題をいくつか紹介する。
（もし興味があれば、上記の連絡先までご一報いただければ幸いです。）

2. 具体例

「テストデータの予測ラベル全体の分布」をチューニングしたい

$$y = f(w, x)$$

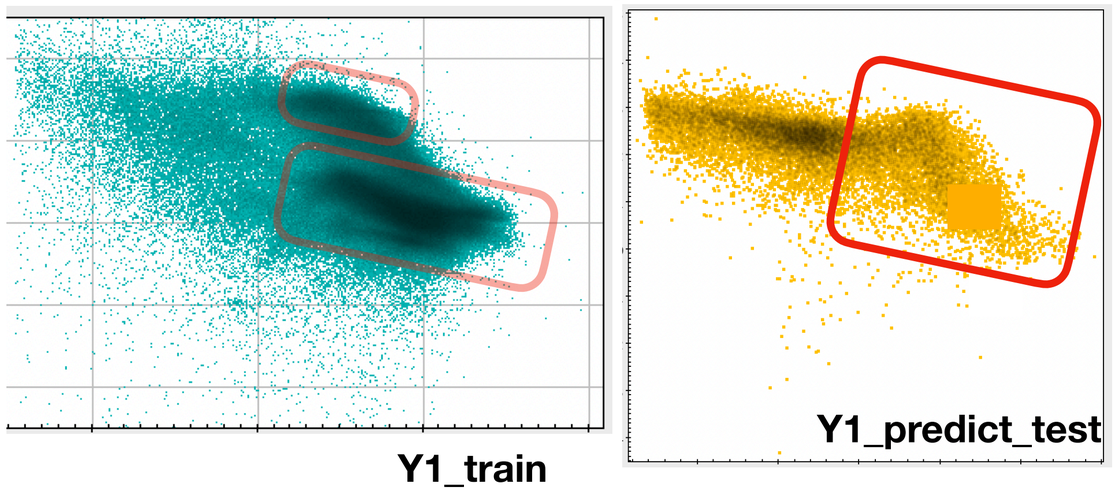
データ x
(110次元実数)　→　□　→　実数ラベル
y = (y1, y2)
(2次元実数)

モデルパラメータ {w}
教師データ {X_train, Y_train}
テストデータ {X_test}
のとき以下の要請を満たすような
モデル f またはパラメータwを決定したい。

・要請A (教師データの二乗誤差要請)
Y1_train, Y2_train = f (w, X_train)

・要請B「テストデータの予測ラベルの分布」
「教師データのラベル分布」の類似性の要請
Y1_predict_test, Y2_predict_test = f (w, X_test)
とおくとき、以下の２つが似た分布を示して欲しい。
P(Y2_predict_test | Y1_predict_test)
P(Y2_train | Y1_train)

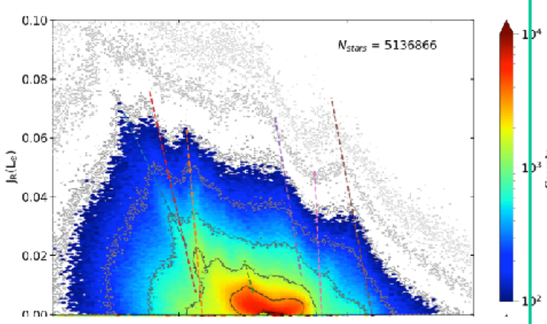
Y2_train　　Y2_predict_test



現在直面している課題：
（左）教師データには赤で囲った２つの山をもつ
（右）Random Forestでナイーブに回帰すると同じ領域が単峰となる。

密度分布の中の特徴的な構造を検出したい。

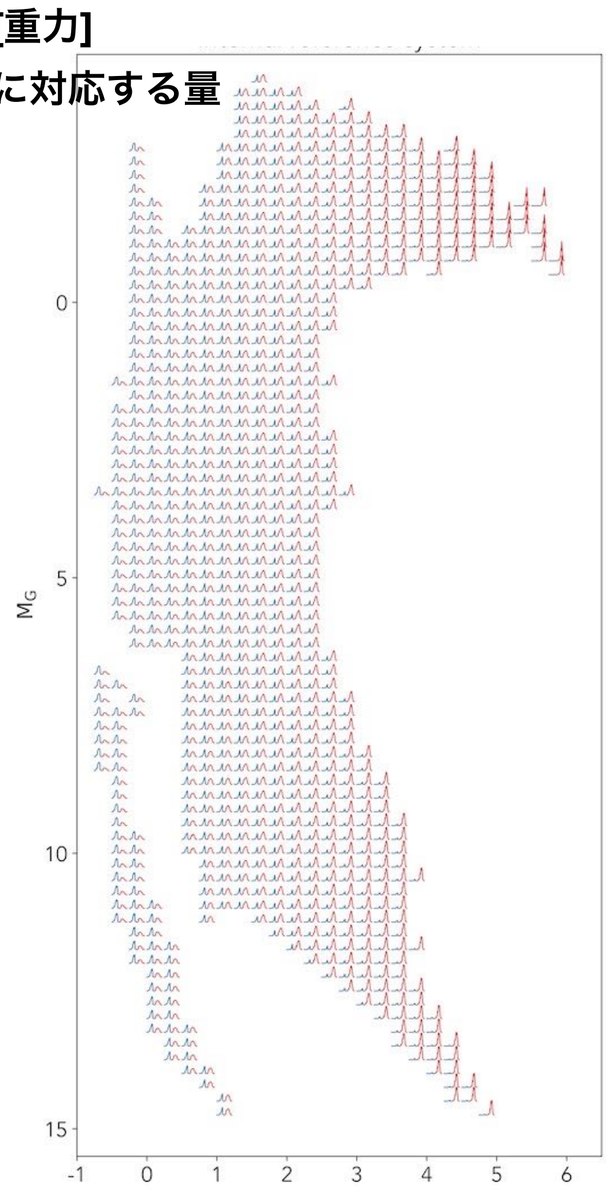
観測データ {X_i | i = 1, ..., N}　(N ~ 1e6)
観測誤差 {e_i | i = 1, ..., N}
物理量 Y = f(X) [関数 f は既知]



Yの密度分布に現れるエッジ構造を検出したい。
観測誤差がなければ、(i) Y=f(X) を計算 (ii) 密度 rho(Y) を計算
(iii) rho(Y) のエッジを探す、とすれば良い。
観測誤差がある場合に同じことをすると、弱いエッジ構造が
観測誤差によって消される。密度分布全体は多少間違っても良いので、
密度分布のエッジ(輪郭)を検出する方法はあるだろうか？

類似した多次元ベクトルを探索したい

[重力]
に対応する量



データ：110次元ベクトル
（２億天体）

高精度のデータもあれば、
低精度のデータもある。
このデータは主に２パラメータ
で記述される。
（星の[温度]と[重力]）

ある意味で110次元ベクトルは
２次元の潜在変数で記述できる。

低精度データの潜在変数を
推定するために、
よく似た高精度データと
対応づけしたい。

多次元データでこのような
クラスタリングを効率的に
行う方法はあるだろうか？

3. その他にどんな数理的な問題が存在するか

・大規模データ、多次元データの宝庫である。
→「希少な現象の発見」「異常検知」
→「極値統計学」
→「多次元データのクラスタリング・分類」
→「次元削減」「可視化」
→「アルゴリズムの高速化」「メモリの最適化」
→「即応分類（観測の最中に分類する）」

・（理論天文学）計算コストが高い
→「機械学習」
→「Simulation-based inference」
「Approximate Bayesian Computation」
（例：シミュレーション結果と観測結果の類似度）
→「シミュレーション設定の最適化」
（例：パラメータを変えた計算はN回しかできない。
どんなN組のパラメータを選ぶべきか）

・（観測天文学）時系列データのケイデンスは一般に不規則
（観夜のみ観測可能、悪天候時は観測不可能）
→「不規則ケイデンスの時系列データの分類」
→「観測スケジュールの最適化」

・（観測天文学）観測誤差が大きい
→「欠損データの統計学」
（欠損データは誤差＝無限大に対応）
→「誤差が大きいデータのクラスタリング」
→「Denoise (誤差を消す) を行う手法」
→「Non-Gaussianの誤差のモデル化」