

正規性を仮定しない分布推定

奥野彰文^{1,2,3}

¹統計数理研究所, ²総合研究大学院大学, ³理化学研究所

自己紹介

- ▶ <https://okuno.net>
- ▶ 統計学と機械学習の手法研究をしています。
- ▶ 対象は多岐にわたる：だいたいなんでもやります。理論+新規手法開発+応用。

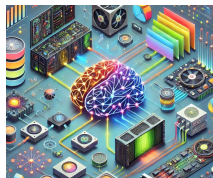


統計学

小サンプル
単純モデル
精密な理論保証

線形回帰モデル,
信頼度評価, 漸近正規性, ...

実は溝がある



機械学習

ビッグデータ
巨大モデル
ラフな理論保証

大規模言語モデル (LLM),
深層学習 (Deep Learning), ...

- ▶ 長尾先生（地震研）にお誘いいただき、板野先生（秋田大）の計画に参加。

なぜ手法研究が必要なのか？

データ解析は
車の運転のようなもの



解析できる人は
増えてきた

手法研究者は
車を調整するエンジニア



ほとんどいない

- ▶ 普段は統計手法をチューニングしたり，用途に合わせて改造しています。

応用研究とのかかわり



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

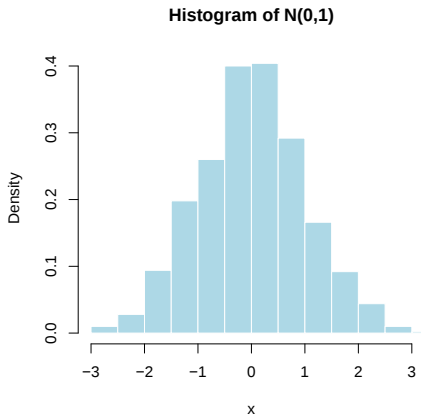
- ▶ プラ核学会 (2023年) : 非線形モデルの学習での高次変動正則化
- ▶ プラ核学会 (2024年) : 簡約流体モデルに基づく数値乱流場の構造クラスタリング

次回はJpGUでも応用の話ができると嬉しいです.

正規性を仮定しない分布推定

分布推定とは？

- ▶ データの“形状”を記述する方法.



このデータ（のヒストグラム）は山に似ている。

最尤推定法

- ▶ サンプルサイズ n が大きく、いくつかの理論的条件を満たすとき、
(ある意味で) 数理的に最も性質が良い方法.

データ $x_1, x_2, \dots, x_n$ があるとき、正規分布の確率密度を $p_\theta(x)$ とすると、

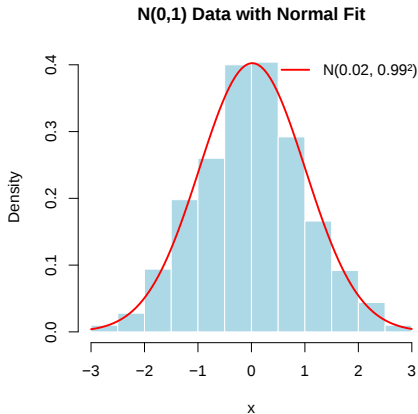
$$\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^n \log p_\theta(x_i).$$

正規分布の場合、 $\hat{\mu} = n^{-1} \sum_{i=1}^n x_i$, $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$.

という話が始めるとややこしくなりますが…

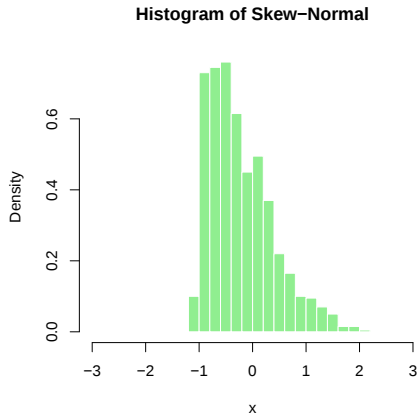
最尤推定法：要するに

- ▶ ヒストグラムにフィットする赤い曲線が得られます。
- ▶ 真値 $\theta_* = (0, 1)$, 推定値 $\hat{\theta} = (0.02, 0.99)$.



ただしいろいろな問題がある

- ▶ 正規分布を使った分布推定（フィッティング）はよく知られている。
- ▶ が、正規分布は対称な関数しか表現できない。



こういうデータはきびしい…?

無理やり正規分布でフィッティングしてみると

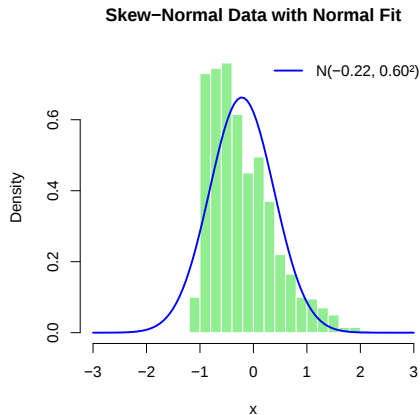
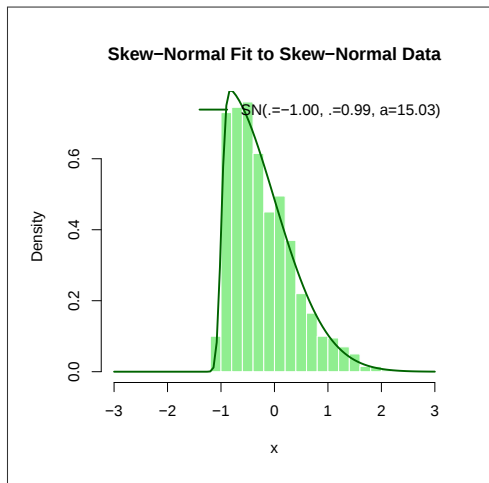
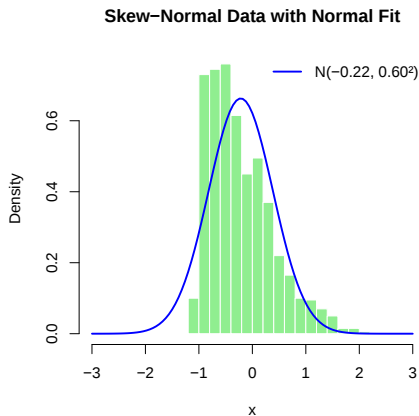


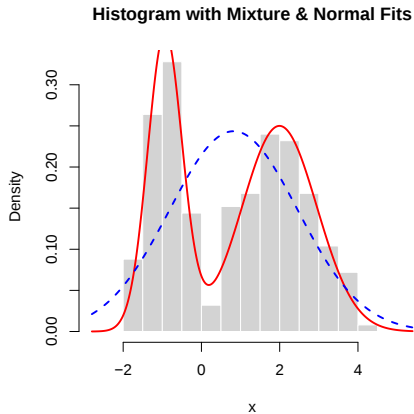
Figure: 大丈夫といえば，大丈夫…？

最尤法の適用可能範囲はとても広い

- ▶ 歪んだ分布を扱えるモデルを使えばよい (歪正規分布)



一次元で歪んでいるだけならまだよいのですが



(a) 2つ以上の山がある場合



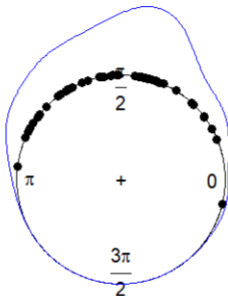
(b) 高次元で人間が認識できない場合

► (適切なモデルを選べば) こういう問題にも最尤法が使えます.

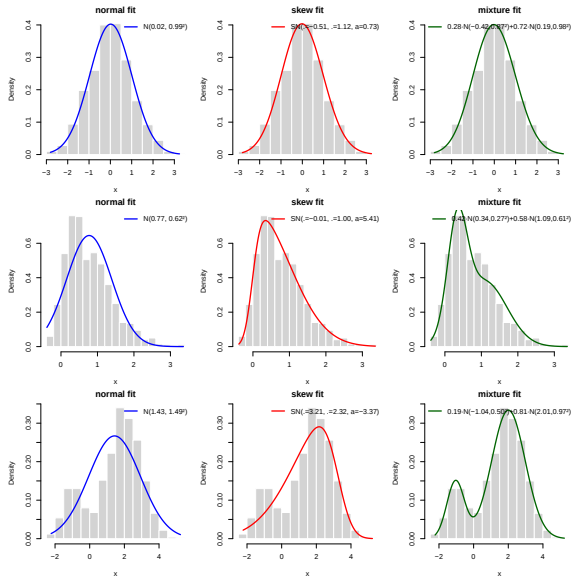
形式の異なるデータ？

- ▶ 角度データは $\theta = 1.99\pi[\text{rad}]$ から少し増えると $0[\text{rad}]$ に戻る：
⇒ 普通のデータとは扱いが異なる.

Circular Histogram with von Mises Fit



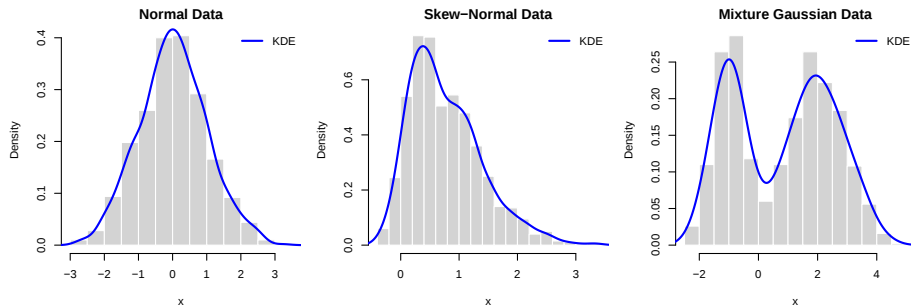
- ▶ こういうものも扱える. 様々なノウハウがある.



どのデータの時、どのモデルを使うべきかという問題もある…

カーネル密度推定 (KDE)

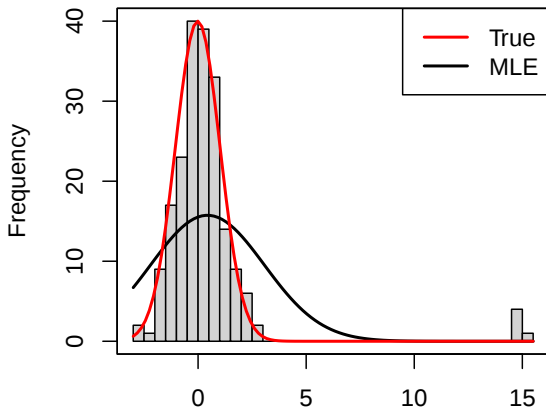
$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i)$$



モデルを事前に指定しなくても, n が大きければ綺麗にフィットする.
(今日はこの話はしません)

正規性を仮定しないロバスト推定 (Okuno, AISM2024)

外れ値が混入する問題



最尤推定は外れ値（右端）に強い影響を受ける

- ▶ という問題を扱うのが「ロバスト統計」と呼ばれる分野です。

ロバスト統計理論が発展している

- ▶ 最尤推定 :

$$\hat{\theta} = \arg \min_{\theta} \sum_{i=1}^n \log p_{\theta}(x_i).$$

- ▶ ロバスト推定 (β -推定; Basu et al. 1998) :

$$\hat{\theta}_{\beta} = \arg \min_{\theta} \underbrace{\left\{ -\frac{1}{\beta} \frac{1}{n} \sum_{i=1}^n p_{\theta}(x_i)^{\beta} + \frac{1}{1+\beta} \int p_{\theta}(x)^{1+\beta} dx \right\}}_{=: d_{\beta}(\hat{Q}, P_{\theta})}$$

$\beta > 0$ でロバストな推定量になる, とされている. 積分項が鬼門で, 正規分布などでのみ展開できる: $(2\pi\sigma^2)^{-\beta/2}(1+\beta)^{-1/2}$. それ以外の分布ではそもそも計算できない...

- ▶ 最適化時の反復のたび, 数値積分をするのは現実的でない.

効率的な最適化法を作りました

- ▶ 確率的な方法により、積分の明示的な計算を回避する。
- ▶ 理論的に収束が言えて、効率的。次元が上がると差が顕著。



Figure: <https://github.com/oknakfm/sgdpd>

収束定理

Theorem (Okuno (AISM2024) Proposition 2)

Assume that the learning rate $\{\eta_t\}_{t=1}^T$ satisfies $\eta_t < 2/L$, $\sum_{t=1}^T \eta_t \rightarrow \infty$ and $\{\sum_{t=1}^T \eta_t\}^{-1} \sum_{t=1}^T \eta_t^2 \rightarrow 0$ as $T \rightarrow \infty$. Under some assumptions, the sequence $\{\theta^{(t)}\}$ obtained by our sgdpd algorithm satisfies

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

$\mathbb{E}_\tau$ represents the expectation with respect to the number of iterations $\tau \in \{1, 2, \dots, T\}$ randomly chosen with the probability $\mathbb{P}(\tau = k \mid T) = \{2\eta_k - L\eta_k^2\} / \sum_{k=1}^T \{2\eta_k - L\eta_k^2\}$ ($k = 1, 2, \dots, T$).

- ▶ 要するに、提案法によって損失関数の勾配が0に収束する。

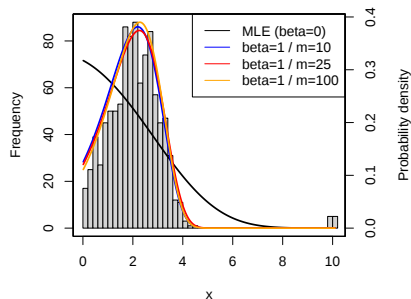


Figure: Gompertz分布のフィッティング

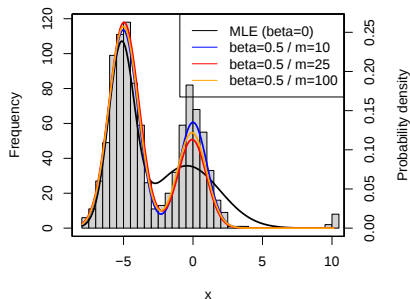
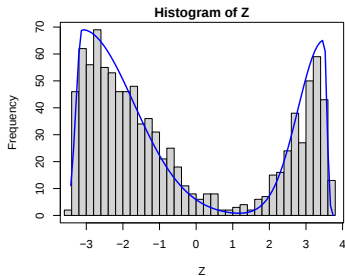


Figure: 混合ガウス分布のフィッティング

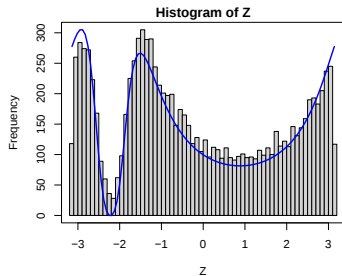
$$\text{Gompertz: } p_{\theta}(x) = \lambda \exp \left(\omega x + \frac{\lambda}{\omega} \{1 - \exp(\omega x)\} \right), (x \geq 0).$$

(*Gompertz分布も混合ガウス分布も積分項が明示的に計算できない)

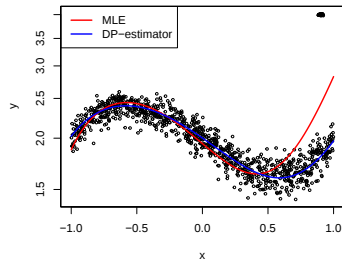
ほかにもいろいろできる.



(a) 混合“歪”正規分布



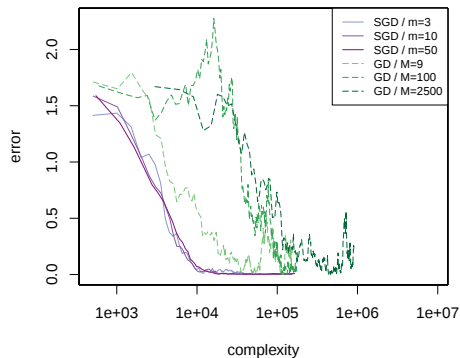
(b) Kato-Jones分布



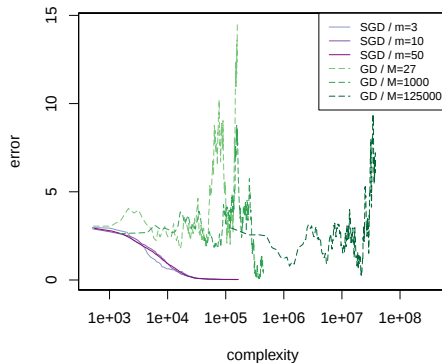
(c) 多項式回帰

▶ 正規分布だけだと, “歪み”などいろいろな情報を捨ててしまう.

計算速度



(a) $d = 2$



(b) $d = 3$

▶ 提案法（実線）は数値積分を介した方法（破線）に比して高速に収束する。

まとめ

- ▶ 分布の推定法が色々あります.
- ▶ 外れ値にロバストな分布推定を効率化しました (Okuno, AISM2024).
- ▶ 分布推定だけではなく, なんでもやっています.
(学生時代は次元削減とか特徴学習をやっていました)
- ▶ 何でもお気軽にご相談いただければと思います.

連絡先 : okuno@ism.ac.jp