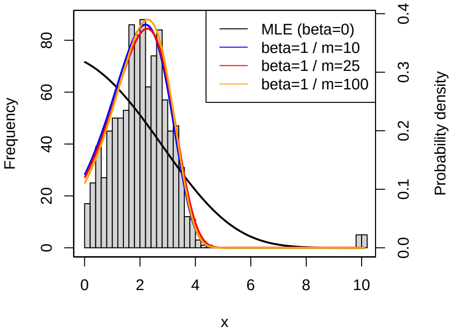
冪密度ダイバージェンスの最小化 およびそのパッケージ化の近況

奥野彰文[1,2]

[1]統計数理研究所, [2]理研AIP

何をしたのか



- 一般の確率モデルp_θでのロバスト推定

$$\min_{\theta}\left\{-\frac{1}{\beta}\frac{1}{n}\sum_{i=1}^{n}p_\theta(x_i)^\beta+\frac{1}{1+\beta}\int p_\theta(x)^{1+\beta}\mathrm{d}x\right\}$$

一般の確率モデルでのロバスト推定 (Okuno, AISM2024)

分布の推定

- 観測値$x_1, x_2, \ldots, x_n$が従う分布Qを知りたい.

経験分布$\hat{Q}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i \leq x)$と確率モデル$P_\theta$のKLダイバージェンス(=乖離度)

$$D(\hat{Q}_n, P_\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i) + \mathrm{Const.}$$

を最小化してP_θが求まる．最尤推定と等価.

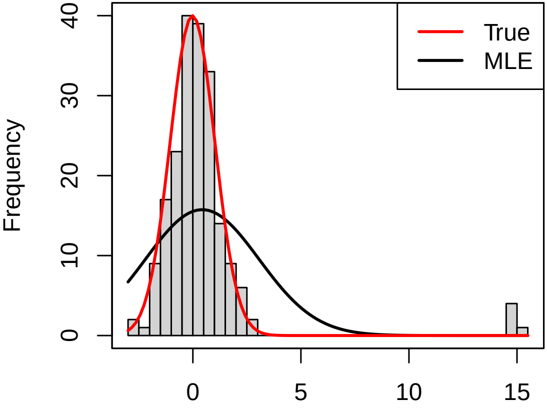


Figure: 最尤推定が外れ値に影響を受けるという典型的な図

ロバストな分布の推定

冪密度ダイバージェンス (Density power divergence, DPD; Basu et al. 1998):

$$D_\beta(\hat{Q}, P_\theta) = -\frac{1}{\beta}\frac{1}{n}\sum_{i=1}^{n} p_\theta(x_i)^\beta + \frac{1}{1+\beta}\int p_\theta(x)^{1+\beta}\mathrm{d}x + \mathrm{Const.}$$

を最小化するとよい.

特に $\beta \searrow 0$ のときKLダイバージェンス($\Leftrightarrow$ 最尤推定)になる.

正規分布の場合，$\xi = 0.1$.

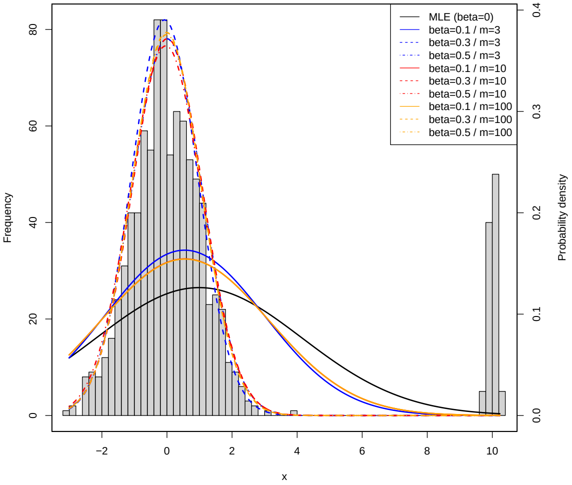


Figure: 外れ値の影響を軽減できた

研究を始める前の Q & A

Q. 外れ値を目視で外せばよいのでは
   A. 私もそう思います．希望があるとすれば高次元*や時系列，自動化の文脈など...

Q. 本当に「ロバスト」な推定なんですか
   A. ロバストって結局何なんでしょうか ⇒ https://arxiv.org/abs/2407.10418

Q. Huberみたいな昔ながらのロバスト推定ではダメですか
   A. 平均等だけでなく，分布そのものを推定できることが強みです．

Q. 計算コストはどうですか
   A. 正規分布など特定の分布以外での最適化は厳しいです．

最適化のボトルネック

$$D_\beta(\hat{Q}, P_\theta) = -\frac{1}{\beta}\frac{1}{n}\sum_{i=1}^{n} p_\theta(x_i)^\beta + \underbrace{\frac{1}{1+\beta}\int p_\theta(x)^{1+\beta}\mathrm{d}x}_{=:r_\theta^{(\beta)}} + \mathrm{Const.}$$

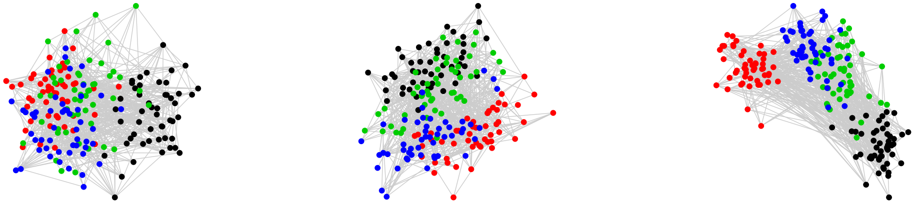
- $P_\theta = N(\mu, \sigma^2)$とすると，$r_\theta^{(\beta)} = (2\pi\sigma^2)^{-\beta/2}(1+\beta)^{-3/2}$.
- 計算できる：正規，一般化指数，一般化パレート，ワイブルモデル.
- それ以外で計算できない．例えば一般の指数型分布モデルで計算できない
  - 混合正規モデルでの近似最適化 (Fujisawa and Eguchi, 2006),
  - ポアソンモデルでの近似最適化 (Kawashima and Fujisawa, 2019),
  - 歪正規モデルでの近似最適化 (Nandy et al., 2022).

分布推定できることが強みなのに，最適化できる確率モデルが少ない矛盾

雑談：かつて1本だけ書いたロバスト推定論文

- ポアソン回帰のような設定 (Okuno and Shimodaira, AISTATS2019):
  確率分布 $P(y \mid \mathbf{x})$ ではなく，平均関数 $\mu(\mathbf{x}) = \mathbb{E}(y \mid \mathbf{x})$ を推定する.

$$L_{\beta,n}(\boldsymbol{\theta}) := \sum_{(i,j)\in\mathcal{I}_n} \left\{ -w_{ij}\frac{\mu_{\boldsymbol{\theta}}(\mathbf{x}_i,\mathbf{x}_j)^{\beta}-1}{\beta} + \frac{\mu_{\boldsymbol{\theta}}(\mathbf{x}_i,\mathbf{x}_j)^{1+\beta}}{1+\beta} \right\}.$$



$\beta = 0$ $\beta = 0.5$ $\beta = 1$

- ⋯という逃げ方もあるが，解く問題が変わっている.

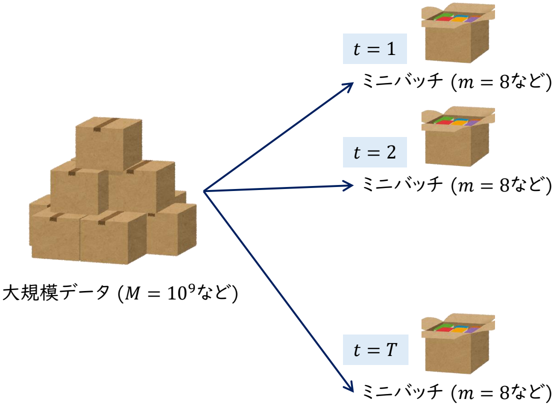
数値積分ではダメなのか

- 勾配法+数値積分: $y_1^{(t)}, y_2^{(t)}, \cdots \sim p_{\theta^{(t)}}$について

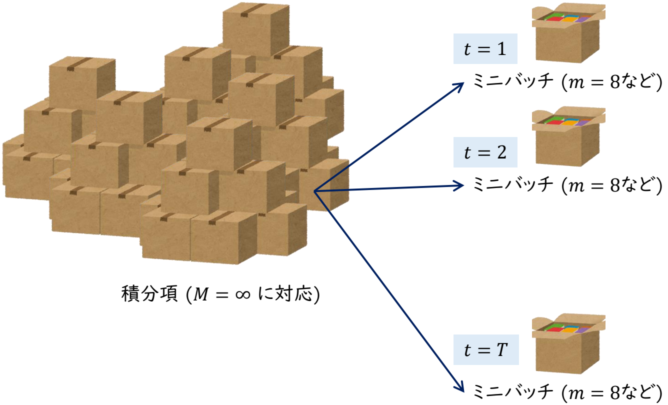
$$
\begin{aligned}
\theta^{(t+1)} &\leftarrow \theta^{(t)} - \gamma \nabla D_\beta(\hat{Q}, P_{\theta^{(t)}}) \\
&= \theta^{(t)} - \gamma \left\{ -\frac{1}{\beta}\frac{1}{n}\sum_{i=1}^{n} \nabla\{p_{\theta^{(t)}}(x_i)^\beta\} + \frac{1}{1+\beta} \operatorname*{p\,lim}_{M\to\infty} \frac{1}{M}\sum_{j=1}^{M} \nabla\{p_{\theta^{(t)}}(y_j^{(t)})^\beta\} \right\}
\end{aligned}
$$

- 各反復で数値積分が必要：次元dが増えると必要なMは大きくなる $(M \sim \varepsilon^{1/d})$
- 学習率 γ は固定，数値積分のサンプルサイズ M は可能な限り大きくする.
- 計算量$O(TM)$, 近似誤差は$T \to \infty$でも$O_p(1/\sqrt{M}) > 0$.

ところで，深層学習で使われるミニバッチ最適化 (M =巨大)



積分 $\Leftrightarrow M = \infty$



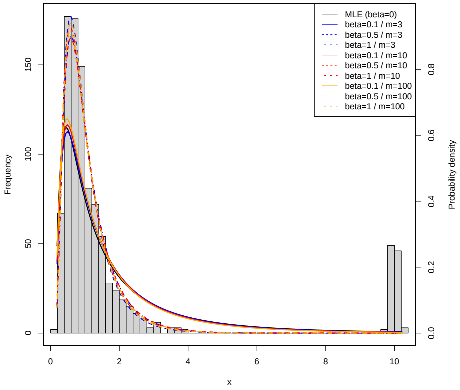
提案法

$y_1^{(t)}, y_2^{(t)}, \cdots \sim P_{\theta^{(t)}}$について

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)} \left\{ -\frac{1}{\beta}\frac{1}{n}\sum_{i=1}^{n} \nabla \{p_{\theta^{(t)}}(x_i)^{\beta}\} + \frac{1}{1+\beta}\frac{1}{m}\sum_{j=1}^{m} \nabla \{p_{\theta^{(t)}}(y_j^{(t)})^{\beta}\} \right\}$$

- 学習率を下げるのがポイント$\gamma^{(t)} \to 0\ (t \to \infty)$.
- mは何でもよい．例えば$m = 8$．$m = 1$でもよい!
- 近似誤差 $O_p(1/\sqrt{M}) > 0$ のある数値積分と違い，真値に収束する! $\hat{\theta}^{(t)} \to \theta_*$.
- 最適化分野ではよく知られている: 確率的最適化 (Robbins and Monro, 1951).
- Contrastive divergence (Hinton et al., 2002).

逆ガウス分布の場合，$\xi = 0.1$.



$$p_\theta(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right)$$

(*逆ガウス分布は積分項が明示的に計算できない)

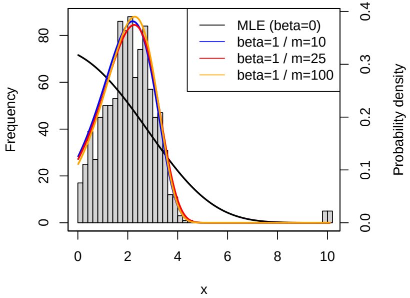


Figure: Gompertz分布のフィッティング

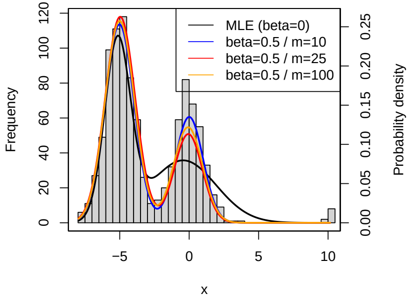
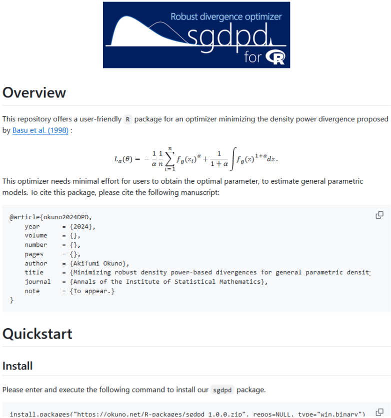


Figure: 混合ガウス分布のフィッティング

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

(*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)

sgdpdパッケージ


Robust divergence optimizer sgdpd for R

Overview

This repository offers a user-friendly `R` package for an optimizer minimizing the density power divergence proposed by Basu et al. (1998) :

$$L_\alpha(\theta) = -\frac{1}{\alpha}\frac{1}{n}\sum_{i=1}^{n} f_\theta(z_i)^\alpha + \frac{1}{1+\alpha}\int f_\theta(z)^{1+\alpha} dz.$$

This optimizer needs minimal effort for users to obtain the optimal parameter, to estimate general parametric models. To cite this package, please cite the following manuscript:

```
@article{okuno2024DPD,
    year      = {2024},
    volume    = {},
    number    = {},
    pages     = {},
    author    = {Akifumi Okuno},
    title     = {Minimizing robust density power-based divergences for general parametric densit
    journal   = {Annals of the Institute of Statistical Mathematics},
    note      = {To appear.}
}
```

Quickstart

Install

Please enter and execute the following command to install our `sgdpd` package.

```
install.packages("https://okuno.net/R-packages/sgdpd_1.0.0.zip", repos=NULL, type="win.binary")
```

Figure: `https://github.com/oknakfm/sgdpd`

- "sgdpd" をyoutubeで検索！

実は課題が山積み

- 提案分布$p_{\theta(t)}$から乱数を生成できると限らない.
  - 提案分布$\tilde{p}_t$に分離した場合の一般の収束条件を求めていない.
  - p_θが混合ガウスなら$\mathbb{E}(\{p_{\theta(t)}(y)/\tilde{p}_t(y)\}^4) < \infty$で収束.
- sgdpdパッケージでは，標本平均/標本分散を用いた正規分布が提案分布$\tilde{p}_t$.
  - 設定によっては著しく収束が悪い．特に回帰などは提案分布が退化して絶望的.
  - 自己回帰での共同研究が進行中 (with Nandy, Pyne, and Basu from ISI)

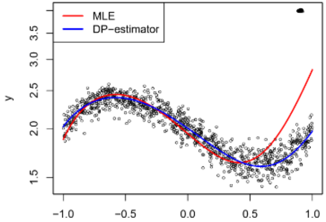


Figure: 多項式回帰の例．苦肉の策として “conditions=c(1)” という命令を実装.

まとめ

- Akifumi Okuno. Minimizing robust density power-based divergences for general parametric density models. AISM. 2024.
(一般の確率モデルでのロバストダイバージェンス最小化)
- sgdpdパッケージ: `https://github.com/oknakfm/sgdpd`
- 日本語での解説: Jxiv.642

質問やコメントは `okuno@ism.ac.jp` まで.