論文解説：一般の確率モデルでの冪密度ダイバージェンス最小化

奥野彰文 *[1,2]

[1] 統計数理研究所 数理・推論研究系　[2] 理化学研究所 AIP センター

要旨

本稿は雑誌 Annals of the Institute of Statistical Mathematics に採択された，冪密度ダイバージェンスの最小化に関する我々の論文: Okuno (2024) の解説です．解説の平易さを優先するため，より厳密な記述については原著論文をご参照ください．

キーワード: 冪密度ダイバージェンス，β-ダイバージェンス，確率的最適化

1　研究背景

1.1　最尤推定と KL ダイバージェンス

観測値 $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ が従う確率分布を推定する問題を考えます．ここでは特に確率モデル P_θ のパラメータ θ の推定を考えましょう[a]．p_θ を P_θ の密度関数とすると，一般には最尤推定

$$\arg\max_\theta \prod_{i=1}^n p_\theta(x_i)$$

により確率モデル P_θ のパラメータ θ が推定できることが知られています．

ここで，最尤推定の別の解釈も考えてみましょう．経験分布 $\hat{Q}_n(x) = n^{-1}\sum_{i=1}^n \mathbb{1}(x_i \le x)$ と確率モデル P_θ の Kullback-Leibler (KL) ダイバージェンスは $D(\hat{Q}_n, P_\theta) = -n^{-1}\sum_{i=1}^n \log p_\theta(x_i) + C$ (C は θ によらない定数) なので，その最小元

$$\hat{\theta}_n = \arg\min_\theta D(\hat{Q}_n, P_\theta) \tag{1}$$

は実は最尤推定量と同値であることが示せます．ダイバージェンスを確率分布の距離のようなものだと思えば，最尤推定とは確率モデル P_θ が観測値から構成される経験分布 $\hat{Q}_n$ に最も近くなるようパラメータ θ を推定しているとみなせて，とても直感的に理解することができます．同様に，例えば回帰や分類・次元削減など統計の様々な問題は経験分布と確率モデルのダイバージェンス最小化を介して定式化でき，理論的な扱いの容易さや枠組みの一般性などからこれまでも様々なダイバージェンスが提案され，様々な問題に利用されてきました．

1.2　最尤推定量の性質

最尤推定は非常に単純な推定量でありながら，理論的に良い性質を持ちます．例えば $x_1, x_2, \ldots, x_n$ がパラメータ θ_* の確率分布 P_{θ_*} から生成されると仮定すれば，最尤推定量，つまり KL ダイバージェンスを用いた推定量 (1) はいくつかの緩やかな条件の下で一致性

$$\hat{\theta}_n \to \theta_* \quad (\text{in probability}, n \to \infty) \tag{2}$$

を持ちます．つまり，確率モデル P_θ が真の分布を含み，かつ観測が無限に得られるのであれば，推定された確率モデルは真の分布を復元します．一致性に加えて，最尤推定では推定量の漸近分散が理論的な下限を達成することも知られています．したがって，いくつかの仮定の下で漸近的な観点では最尤推定が最も効率的にパラメータを推定できるといえます (漸近有効性)．

1.3　外れ値の影響

最尤推定は漸近的に良い性質を持つ一方で，すべての意味でベストな推定法であるとは限りません．さて，ここでは一例として，ある集団の年収期待値の推定を考えましょう．年収期待値の真値が 500 万円であるとし，そこから 99 人の年収を観測したとします．正規モデルを仮定すれば，この 99 人の年収の最尤推定量[b]はおよそ 500 万円となるはずですが，ここに年収 1 兆円の大富豪が 1 人紛れ込むと，年収の最尤推定量はおよそ 100 億円まで跳ね上がります．このように極端に大きい (または極端に小さい) 値を特に外れ値と呼び，最尤推定量は外れ値に強い影響を受けることが分かります．上記以外でも，例えばデータ入力時に誤った観測値を入力したり，測定のためのセンサーが故障したりなど，実際のデータ解析において外れ値や異常値が紛れ込んでしまう状況は多々あります．これら外れ値はデータの重要な性質を表す場合もありますが，データ全体の傾向を把握したい多くの場合，例えば上記の例のように外れ値が要約統計量を極端に歪めてしまい，解析に悪影響を及ぼします．

* 責任著者，okuno@ism.ac.jp

[a] 確率モデル P_θ として，例えば $\theta = (\mu, \sigma^2)$ をパラメータとする正規分布 $N(\mu, \sigma^2)$ などを利用するとよいでしょう．

[b] 正規分布の期待値パラメータの最尤推定量は観測値の平均です．

1.4 冪密度ダイバージェンス

既に説明した通り，最尤推定量は一致性や漸近有効性など良い性質を持つ (1.2 節) 一方で，外れ値の混入に強い影響を受けます (1.3 節).

最尤推定量は KL ダイバージェンスの最小化解 (1) としても理解することができるので，これらの性質は利用した KL ダイバージェンス (つまり確率分布の距離) に依存したものと考えられます．そこで，KL ダイバージェンスを別のダイバージェンスに置き換えることで，様々な性質をもった推定量を定義することができます．特に外れ値の問題に対処するため，Basu et al. (1998) は冪密度ダイバージェンス:

$$D_\beta(\hat{Q}_n, P_\theta) := -\frac{1}{\beta n}\sum_{i=1}^{n} p_\theta(x_i)^\beta + \underbrace{\frac{1}{1+\beta}\int p_\theta(x)^{1+\beta}\mathrm{d}x}_{=:r_\theta^{(\beta)}} + C_\beta \quad (3)$$

を提案しました．C_β はパラメータ θ に依らない定数で，$\beta > 0$ はユーザの指定するハイパーパラメータです[c]．典型的には $\beta = 0.1, 0.5, 1$ などを利用します．最尤推定量と同値な最小元 (1) の KL ダイバージェンスを冪密度ダイバージェンスに置き換えると，推定量

$$\hat{\theta}_{\beta,n} := \arg\min_\theta D_\beta(\hat{Q}_n, p_\theta) \quad (4)$$

が定義できます．簡単のため，以降はこの推定量 (4) を β-推定量と呼びましょう．β-推定量は $\beta > 0$ のとき外れ値の影響を受けにくいことが知られており，この性質を特に (外れ値への) ロバスト性と呼びます.

β-推定量は $\beta \searrow 0$ の極限で最尤推定量に収束するので，最尤推定量のロバストな拡張とみなすことができます．一方で，β-推定量 ($\beta > 0$) は一般に漸近有効でなく，外れ値の影響力を問題視するのか，推定量の有効性を求めるかなどの必要に応じて最尤推定量と β-推定量を適切に使い分けることが重要です[de].

1.5 問題点: 積分項に起因する最適化の難しさ

さて，β-推定量の理論的な性質に関する研究はたくさんありますが，実際に観測値から β-推定量 (4) を計算するにはいくつかの困難があります．β-推定量 (4) を計算するには冪密度ダイバージェンスを最小化する必要があり，通常このような最小化問題では学習率 $\omega > 0$ と $\nabla = \partial/\partial\theta$ を用いた勾配法

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \omega\nabla D_\beta(\hat{Q}_n, P_{\theta^{(t)}}) \quad (5)$$

により反復的にパラメータを更新する戦略がとられます．勾配法では冪密度ダイバージェンス (3) の勾配を計算する必要があり，つまり (3) に含まれる積分項 $r_\theta^{(\beta)} = \frac{1}{1+\beta}\int p_\theta(x)^{1+\beta}\mathrm{d}x$ の勾配を明示的に計算できる必要があります.

多くの既存研究では確率モデル P_θ が正規分布であることを仮定することで，積分項を解析的に展開し，その勾配を計算しています．一方で，積分項 $r_\theta^{(\beta)}$ を解析的に展開できる確率モデルは限られていて，既存研究が存在するものでは (一般化) 指数分布，一般化パレート分布，ワイブル分布のみでこの積分項が解析的に展開可能であると示されているものの，例えば一般の指数型分布族などに対してさえ積分項を展開することができません.

解析的な計算が難しいので，ポアソン分布および (一次元) 歪正規分布に対して積分項の勾配 $\nabla r_\theta^{(\beta)}$ を有限近似することで冪密度ダイバージェンスを最小化する既存研究などもありますが，特に観測値の次元 d が 3 以上の場合など，積分の近似精度を十分良くするには勾配法の各反復で巨大な計算量が必要となります.

2 本研究の貢献

2.1 提案法

本研究では Robbins and Monro (1951) に起源をもつ確率的な最適化法を利用し，明示的な積分計算を回避する効率的な冪密度ダイバージェンス最小化法を提案しています．より具体的には，(5) の勾配 $\nabla D_\beta(\hat{Q}_n, P_\theta)$ を確率的な勾配 $g_t(\theta^{(t)})$ に置き換えた確率的勾配法

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \omega_t g_t(\theta^{(t)}) \quad (6)$$

によりパラメータを更新します．ここで $\omega_t > 0$ は反復に応じて減少する学習率であり，ユーザが指定する任意のサンプラー $\tilde{p}_t$ からランダムに生成された[f]乱数

[c] 冪密度ダイバージェンスは，後続論文でハイパーパラメータが β と表記されていたことから β-ダイバージェンスなどとも呼ばれますが，原著論文 Basu et al. (1998) では α が使われています．また Basu 先生らのグループは現在に至るまで一貫してハイパーパラメータ α を用いていますが，α-ダイバージェンスと呼ぶと別のダイバージェンスが連想されてしまうので，混同しないよう注意が必要です.

[d] 冪密度などロバストなダイバージェンスについて日本語で学べる書籍として，「ロバスト統計 外れ値への対処の仕方」(藤澤洋徳・著) があります．漸近効率やロバスト性の指標などについても議論されています.

[e] 1.3 節で紹介した例ですと，例えば中央値を用いることで外れ値の影響を無視することができます．このようなロバスト推定手法は多数提案されていますが，ロバストダイバージェンスを利用する強みは主に，期待値パラメータなど個別の統計量だけでなく分布そのものを推定できること，および推定を数理的に統一的に解釈できて見通しがよくなることだと思います.

[f] $\tilde{p}_t$ が $p_{\theta^{(t)}}$ と大きく異なると確率的最適化の収束条件が満たされなくなる場合があります．したがって，理想的には $\tilde{p}_t = p_{\theta^{(t)}}$

$y_1^{(t)}, \dots, y_m^{(t)} \in \mathbb{R}^d$ と $t_\theta(x) = \nabla \log p_\theta(x)$ を用いて

$$g_t(\theta^{(t)}) := -\frac{1}{\beta n}\sum_{i=1}^{n} p_\theta(x_i)^\beta t_\theta(x_i) + \frac{1}{1+\beta}\frac{1}{m}\sum_{j=1}^{m}\frac{p_\theta(y_j^{(t)})^\beta}{\tilde{p}_t(y_j^{(t)})} p_\theta(y_j^{(t)})^\beta t_\theta(y_j^{(t)}) \quad (7)$$

とすれば，確率勾配が $\{y_j^{(t)}\}_j$ に関する不偏性

$$\mathbb{E}(g_t(\theta^{(t)})) = \nabla D_\beta(\hat{Q}_n, P_{\theta^{(t)}})$$

を満たしますから，いくつかの条件の下で収束

$$\theta^{(t)} \to \hat{\theta}_{\beta,n} \quad (\text{in probability}, t \to \infty)$$

が示せます．Robbins and Monro 型の確率的最適化法はかなり広いクラスの損失関数を最小化できることが示されていて，例えば Lan (2020) などに一般の損失関数での確率的最適化の収束理論がまとめられています．

2.2 提案法のメリット

既存法に比べ，提案法は以下のメリットを持ちます．

(1) **一般の確率モデルで計算できる．** 既存法で計算が難しい，一般の確率モデルに対する冪密度ダイバージェンスを最小化できます．例えば，通常の勾配法では計算が困難であった，混合正規分布のパラメータ推定等も容易に実行できます．

(2) **効率的に計算できる．** 提案法では，確率勾配の計算のためランダムに $m \in \mathbb{N}$ 個の $y_1^{(t)}, y_2^{(t)}, \dots, y_m^{(t)}$ を生成しますが，$m \in \mathbb{N}$ がどのような値の場合でもパラメータの収束が証明できます．例えば $m = 1$ で確率勾配を計算した場合であっても確率的勾配法は最適解に収束することが期待できます[g]．数値積分を用いる有限近似のアプローチではたくさんのサンプルを生成する必要がある一方で，提案法は各反復での計算量がほぼ定数なので，効率的にパラメータ推定ができます．

(3) **非凸最適化への親和性が高い．** 冪密度などロバストダイバージェンスは一般に (パラメータに関して) 非凸な関数であることが知られています．通常の勾配法 (5) は局所解に陥りがちですが，確率的勾配法 (6) は確率的な変動によりこのような局所解から抜け出せる場合があることが知られています．これらは近年主に深層学習の枠組みで研究が進められていますが，確率的勾配法とロバストダイバージェンスの相性の良さが分かります．

実験を交えた，既存法とのさらに詳細な比較については Okuno (2024) をご参照ください．

3 提案法の実装

https://github.com/oknakfm/sgdpd にて，提案法に関する R パッケージ sgdpd を提供しています．パッケージのインストール方法，および詳細な利用法については readme およびマニュアルをご覧ください．

3.1 数値実験

一例として，ここでは位置パラメータ μ, スケールパラメータ σ, 形状パラメータ ξ を持つ歪正規分布[h]を考えます．真値を $\theta_* = (\mu_*, \sigma_*, \xi_*) = (0, 1, 4)$ とし，歪正規分布から $n = 1000$ 個の観測値を生成する際に $N(10, 0.01^2)$ から生成した外れ値を 5% 混入させます．外れ値が混入した観測値から，提案法により歪正規分布モデルをフィッティングした結果が図 1 です．最尤推定量 (MLE) が外れ値に影響を受けているのに対して，提案法により計算した β-推定量が外れ値の影響を取り除けていることが分かります．

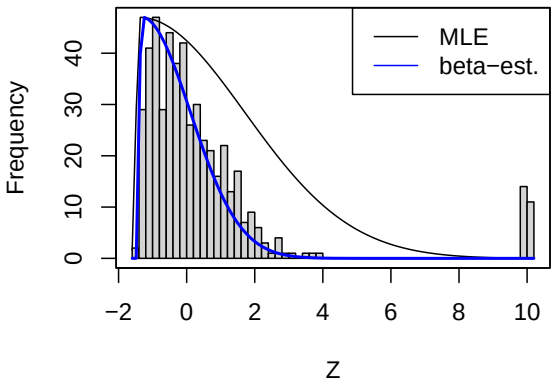

図 1: 歪正規分布の推定

4 まとめ

本研究では外れ値に影響を受けにくい統計的推定を行うため，冪密度ダイバージェンスを利用した推定量を考えました．冪密度ダイバージェンスは積分項を含むので，正規分布モデルなど一部の確率モデルを除き通常の勾配法を計算することは難しいのですが，うまくデザインした確率勾配を利用した確率的勾配法により冪密度ダイバージェンスを最小化する方法を提案しました．https://github.com/oknakfm/sgdpd に提案法の実装を R パッケージとして公開しています．

が好ましいものの，乱数が生成できない確率モデルを利用する場合などは熟慮が必要です．

[g] $m = 1$ の場合，確率勾配の分散が大きくなるので収束レートは悪くなりますが，収束はします．

[h] 歪正規分布について，積分項 $r_\theta^{(\beta)}$ を解析的に展開することは (少なくとも私には) 不可能ですし，実際に数値積分で解析的な展開を回避した既存研究がありますが，提案法により極めて簡単に最適化を実行できます．

参考文献

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559.

Lan, G. (2020). *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Series in the Data Sciences. Springer International Publishing.

Okuno, A. (2024). Minimizing robust density power-based divergences for general parametric density models. *Annals of the Institute of Statistical Mathematics*. to appear.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407.