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

奥野彰文 (統数研, 理研AIP)
https://okuno.net/

1. 研究背景

確率モデルP_θの最尤推定は，経験分布$\hat{Q}$との乖離を表すカルバックライブラー(KL)ダイバージェンス$D(\hat{Q}, P_\theta)$の最小化と等しい．統計の様々な問題は経験分布と確率モデルの間のダイバージェンスの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，様々な問題に適用されてきた．

KLダイバージェンスはデータ中に含まれる異常値に強い影響を受けることが知られており，そのロバストな拡張として冪密度ダイバージェンス(Basu et al. 1998)が提案された．冪密度ダイバージェンス最小化は，その(経験的)クロスエントロピー

$$d_\beta(\hat{Q}, P_\theta) \coloneqq -\frac{1}{\beta} n^{-1} \sum p_\theta(x_i)^\beta + \underbrace{\frac{1}{1+\beta} \int p_\theta(x)^{1+\beta} \mathrm{d}x}_{=: r_\theta^{(\beta)}}$$

の最小化と等しい．一方で，バイアス補正に必要な後半の積分項$r_\theta^{(\beta)}$を解析的に導出することは難しい．積分項が明示的に求まらなければ，最適化問題は難しくなる．多くの場合は正規分布が利用され，そのとき$r_\theta^{(\beta)} = (2\pi\sigma^2)^{-\beta/2}(1+\beta)^{-3/2}$である．

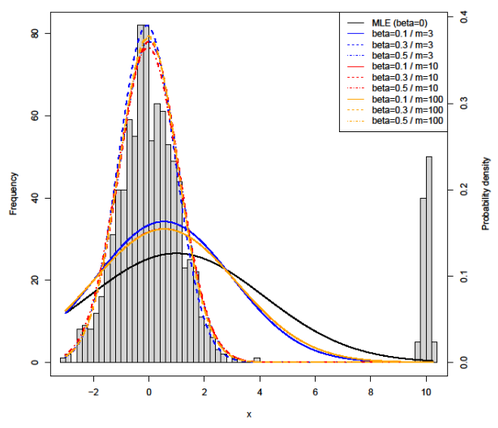

図1: 典型的な外れ値+正規分布でのフィッティングの例．
最尤推定(黒線)より
ロバスト推定(カラー)のほうが
外れ値の影響を受けにくい．

ほかに指数分布 (Jones et al. 2001), 一般化パレート分布 (Juarez and Schucany, 2004), ワイブル分布 (Basu et al. 2016)で積分が計算可能であるが，より一般の分布，例えば混合正規分布などにさえ積分を明示的に展開することができない．数種類の分布以外で積分が明示的にわからない，つまり最適化できないのだから，外れ値の悪影響以前に，データの従う分布によっては確率モデルそれ自体が強制的な誤特定の影響を強く受けてしまう．

同様の問題意識を持った研究が（少ないながらも）ある．例えばFujisawa and Eguchi (2006)では混合正規分布での冪密度ダイバージェンスの上界の最小化を提案したり，Kawashima and Fujisawa (2019)ではポアソン分布での数値積分を介した有限近似ダイバージェンス最小化を提案しているなど，近似によるアプローチがとられてきた．

2. アイデア

(非確率的な)勾配法を利用すると積分部分の厳密評価が必要になる一方で，Robbins and Monro (1951)に端を発する確率的勾配法(Stochastic gradient descent, SGD)では厳密な勾配を計算する必要がなく，厳密な積分計算を回避できる．

昨今の深層学習技術の根幹をなしているSGDは，観測された全データから一部のデータを(確率的に)サブサンプリングし，一部データのみで勾配を計算する方法である．これをより一般に書き直すと，t回目の反復での確率的な勾配$g_t(\theta^{(t)})$が不偏性

$$\mathbb{E}\left[g_t(\theta^{(t)})\right] = \frac{d}{d\theta} d_\beta(\hat{Q}, P_\theta)$$

を満たせば，適当に減少する学習率$\eta_t \searrow 0$を用いた確率的勾配法

$$\theta^{(t+1)} = \theta^{(t)} - \eta_t g_t(\theta^{(t)})$$

は$d_\beta(\hat{Q}, P_\theta)$の勾配を0に収束させることが知られている．したがって，不偏な確率的勾配さえ構成できればよい．

3. 提案 (不偏な確率的勾配)

$s_\theta(x) = \mathrm{d}\log p_\theta(x)/\mathrm{d}\theta$とし，$y_j^{(t)} (j = 1, \ldots, m)$を$p_{\theta^{(t)}}$からのサンプルとする．このとき確率的勾配を

$$g_t(\theta^{(t)}) = -n^{-1} \sum p_{\theta^{(t)}}(x_i)^\beta s_{\theta^{(t)}}(x_i) + m^{-1} \sum p_{\theta^{(t)}}\left(y_j^{(t)}\right)^\beta s_{\theta^{(t)}}\left(y_j^{(t)}\right)$$

とすると，$m \in \mathbb{N}$によらず不偏であり，この確率的勾配を用いた確率的勾配法は冪密度ダイバージェンスを最小化する．

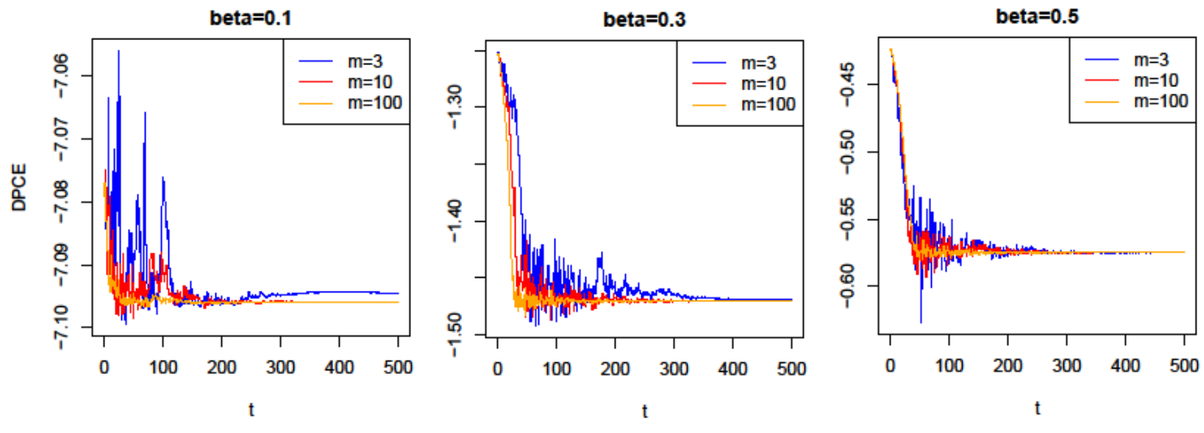

図2: 正規分布では冪密度ダイバージェンスが明示的に計算でき，確率的最適化での遷移を記録したもの．
なおフィッティングの結果は図1.

※テクニカルにはRobbins and Monro (1951)そのものだし，Intractable likelihood estimationやContrastive divergenceなど，類似の話は既にあることに注意．

4. この研究で出来るようになったこと

① 任意の確率モデルで冪密度ダイバージェンスが最小化できるようになった（例えば右のゴンペルツ/混合正規分布）．
② Kanamori and Fujisawa (2015)を援用するとガンマダイバージェンス (Fujisawa and Eguchi, 2008)も最小化できる．
③ 回帰や判別など別の設定でも同様の最適化可能．
④ もっと一般のダイバージェンスでも最適化可能．

詳細はプレプリント https://arxiv.org/abs/2307.05251 を参照．

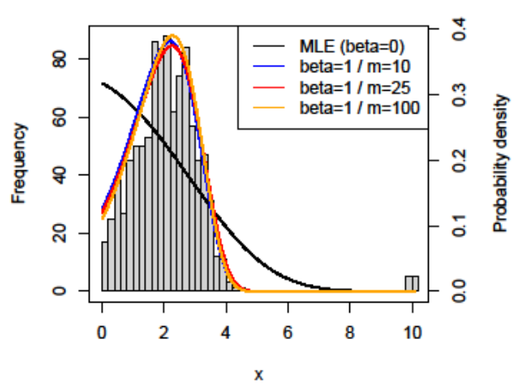

図3: ゴンペルツ分布
$p_\theta(x) = \lambda \exp(\omega x + \lambda\omega^{-1}(1 - \exp(\omega x)))$

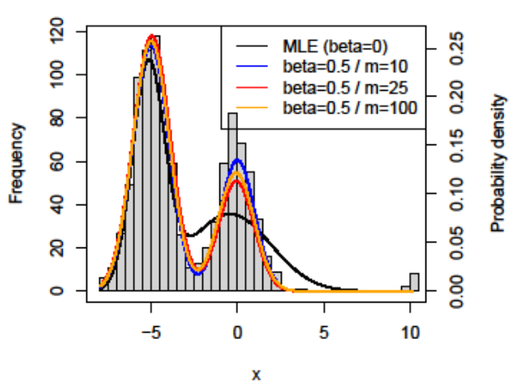

図4: 混合正規分布
$p_\theta(x) = \alpha\phi(x; \mu_1, \sigma_1^2) + (1-\alpha)\phi(x; \mu_2, \sigma_2^2)$