Laplace 分布の再評価：ベイズ法と活性化関数から

統計数理研究所　　柳本 武美
Takemi Yanagimoto
Institute of Statistical Mathematics

§1. 序：確率分布の有用性

Laplace 分布 $\mathrm{LA}(\mu, \tau)$ の密度関数は、$\mu \in R$, $\tau \in R^+$ に対して

$$p(x|\,\mu, \tau) = \frac{\tau}{2}\exp\{-\tau|x - \mu|\} \tag{1.1}$$

で与えられる。平均 μ は位置母数でもあり、標準偏差 $1/\tau$ は尺度母数でもある。数学者 Laplace (1749 - 1827) の名が冠せられるように古くから知られている。この分布の詳細なレビューは Kotz ら (2001) にある。

一方、logistic 分布 $\mathrm{LG}(\mu, \tau)$ の密度関数は、$\mu \in R$, $\tau \in R^+$ に対して

$$p(x|\,\mu, \tau) = \frac{\tau \exp\{\tau(x - \mu)\}}{\{1 + \exp\{\tau(x - \mu)\}\}^2} \tag{1.2}$$

で与えられる。共に 平均値 μ に関して対称で、x が大きいときに指数関数のオーダーで減少する。また、分布関数も初等関数で容易に表現できる。この性質は順序統計量を扱う場合に便利である。一方、これら二つ分布は指数分布族に属さないし、二次元の十分統計量もない。このように二つの分布の役割は、大まかに見れば極めて良く似ている。

統計学においては確率分布はデータのばらつきを表現するために使われる。確率分布は無数にあるが、実際に使われる分布は限られている。改

めて確率分布の有用性を考えると、独自性と解析的な利便性の二面がある。データのばらつき方を表現する規則性はデータによって異なるから、それに応じた特性が求められる。一方で大まかには似た確率分布であれば、母数の推定などデータの効率的な解析に必要な解析的な性質を満たすことが望ましい。解析的な利便性は、数学的な性質であって確率分布の有用性とは関係がないように感じるかも知れない。しかし、正規分布を初めよく使われる分布は便利な性質を満たしている。

本稿では、Laplace 分布と logistic 分布を例にして、分布の有用性を議論する。今日では、どちらかと言うと logistic 分布の方が馴染まれてはいるが、細かく調べてみても logistic 分布には特に長所が見当たらないことを指摘する。そして深層学習のような分野、つまり自然現象の記述と言うより人の認知過程に関する分野、では Laplace 分布の欠点と考えられていた性質がなくなることを指摘する。

結論として、Laplace 分布の再評価が望まれることを強調する。

§2. Laplace 分布の特性

確率密度 (1.1) から分かるように、平均値 μ で二つの指数密度関数を貼り合わせた密度関数になっている。その意味で double exponential 分布とも呼ばれる。正値上の分布としての指数分布の重要性は論を俟たない。しかし、その事実は Laplace 分布の重要性と直接の関係はない。

当初に注目された点は、標本中央値と μ の MLE が一致することにあった。Laplace (1774), Kaynes (1911) らの関心を惹いた。しかし、推定量としての性能は良くない。実際、標本中央値が記述統計学で重要視される理由は、母集団の分布の非対称性が強く強度に歪んでいる場合である。そうしたデータでは中央値によって集団を代表させる推定量としての統計量も同様の働きが見込まれる。対称な分布の中央値、同時に平均値でもある、の推定量として良い性質を保つかどうかについては何の保証もない上に懐疑的でもある。また、MLE について標本サイズが偶数 $n = 2m$ の場合の μ の推定量を考える。標本 $x_1, \cdots, x_n$ の順序統計量を $x_{(1)}, \cdots, x_{(n)}$ と書

くと、尤度は $(x_{(m)}, x_{(m+1)})$ の区間で一定であり最大値を達成する。だから MLE は一意には定まらない。そこで、$(x_{(m)} + x_{(m+1)})/2$ とする。しかし、後に議論する事後分布から見ると、この操作には何らの妥当性もない。だから、正当性がない二つの推定値が偶々一致しただけで、何らの格別な意味もない。実際、MLE を改良することは容易であり、Govindarajulu (1966), Akahira and Takeuchi (1990) などの研究がある。

注 正規分布では標本平均と μ の MLE が一致することに注目された。この推定量 $\bar{x}$ の性能は良いことが知られている。逆に、この性質は正規分布を特徴付ける。二つの推定量が一致することでは見かけ上同じであるが、その意味するところは全くの逆である。正規分布では MLE が望ましいことを含意するが、Laplace 分布では望ましいとは含意しない。

こうした欠点以上に従来 Laplace 分布が避けられてきた大きい理由は、密度関数が中央値 μ で折れている点である。素朴な自然観では、自然現象とか生理反応は滑らかに変化すると考えられている。だから、多くの応用分野でのデータの変動の記述には不適切と見なされている。もし分布の中央部の細かい変動が重要であれば、折れ点の存在は大きな欠点である。そうではなくて、分布の全体的変動に関心があるのなら、折れ点は重要ではない。

欠点が多いようであるが、様々な長所がある。本稿の後節では、母数のベイズ推定と深層学習における活性化関数と対数尤度比との関係を議論する。

その前に、Subbtin (1923) が導入した exponential power 分布族から Laplace 分布を俯瞰する。係数 $r > 0$ を任意に固定したとき、この分布族の密度関数は母数 (μ, τ) に対して次のように表される。

$$p_r(x|\,\mu, \tau) = \frac{\tau^{1/r}}{2\Gamma((r+1)/r)} \exp\{-\tau|x - \mu|^r\}. \tag{2.1}$$

係数 r を 1, 2 と置くと、各々 Laplace 分布と正規分布になる。また、$\xi = \tau^r$ と置いて、$r \to 0$ の極限分布として一様分布が得られる。この密

度を μ の関数と見ると同じ分布に従うと共に、τ の関数と見ると gamma 分布の密度関数に比例する。係数が $r > 1$ の場合には密度関数は滑らかで折れ点がない。従って、Fisher の情報行列は $r \downarrow 1$ の極限値として導かれる。近年注目されている Lasso は正規標本分布の下で Laplace 事前分布を仮定したベイズモデルと見なされる (Park and Casella, 2008)。このような定式化が出来るのは、両方の分布が同じ exponential power 分布族に属して、尺度母数 τ を見かけ上同一視出来るからである。

一方 logistic 分布の特徴を調べる。既に述べたように、密度関数 (1.2) は滑らかである。また、分布関数も Laplace 分布より表現が簡潔である。従って、打ち切りのある順序統計量の扱いではより容易である。分布関数が成長曲線のモデルと関連する。Logistic の名称の由来は成長曲線との関連である。更に特筆すべき点は、二項分布 $\mathrm{Bi}(p)$ を指数分布族と見なしたときの自然母数 θ との関連である。自然母数は

$$\theta = \log \frac{p}{1-p}$$

で表される。この関数は logistic 分布関数で $\tau = 1$, $\mu = 0$ と置いて、$p = x$ の関数と見た場合の逆関数である。この関係は、二項分布を説明する潜在分布と見なされる。広く用いられている logit モデルの基礎を与えている。

しかし、logit モデルは回帰モデルであって、逆関数は回帰関数として用いられている。通常のデータの変動を表現する訳ではない。また、潜在分布としてもっともらしいとする正当化もない。Logistic 分布の良さを表していると言うより、二項分布の特性であると見る方が自然である。実際 logit モデルが歓迎されたのは、それ以前に毒性学・生物検定で用いられていた probit モデルに無理があったと見なされている。多くの分野では、潜在分布として正規分布を想定するのは無理があるからである。

指数分布族には属さないし、MLE は陽には表現できない。尺度母数 τ にべき乗とか gamma 密度関数との関連もない。また、Kullback-Leibler

分離度は、 $g(\boldsymbol{y})$ の密度 $p(\boldsymbol{y})$ の下での期待値を $\mathrm{E}\{g(\boldsymbol{y});\, p(\boldsymbol{y})\}$ と書いて

$$\mathrm{D}\big(\hat{\theta}, \theta\big) \;=\; \mathrm{E}\{\log\{p(\boldsymbol{y}|\,\hat{\theta})/p(\boldsymbol{y}|\,\theta);\, p(\boldsymbol{y}|\,\hat{\theta})\} \tag{2.2}$$

と表されるが、logistic 分布の間では陽に表現できない。更には、正規分布など他の有用な分布を共に含む一般分布族もない。その結果、標本分布を logistic 分布としたベイズモデルでは事前分布の選択が限られる。逆に事前分布として logistic 分布を仮定する場合の利便性も見あたらない。

改めて logistic 分布を評価すると、密度関数が折れ点のない滑らかな関数であることを除いて格別な利便性は見当たらない。

§3. ベイズ推定における estimand

Laplace 分布 $\mathrm{LA}(\mu, \tau)$ の母数推定を考える。事前密度として無情報事前関数として reference 事前関数 (Garvan and Ghosh, 1997)

$$\pi_R(\mu, \tau) \;\propto\; \frac{1}{\tau} \tag{3.1}$$

を仮定する。他にも Jeffreys' prior, $\pi_J(\mu, \tau) \;\propto\; 1/\sqrt{\tau}$ が知られるが、reference prior の方が理論的な裏付けがある上に多くの例で性能の良い推定量が導出されることが知られている。

もし μ に関しての情報が存在するときには informative 事前密度も次のように定義できる。

$$\pi(\mu, \tau;\, \delta_0, m) \;\propto\; \frac{\delta_0 \tau}{2} \exp\{-\delta_0 \tau |\mu \,-\, m|\} \cdot \frac{1}{\tau}.$$

主要項は Laplace 密度であり、台としての測度が reference 事前関数である。このような簡明な informative 事前分布が定義できて、演算が容易になるのは τ がべき乗係数と扱えるからである。この仮定下で事後密度は簡潔になる。更に、τ に関して事前情報が仮定できる場合にも、gamma 分布を用いた簡単な仮定が可能である。

ベイズ推定では母数の推定は事後平均で推定することが原則である。問題は事後平均を求める母数の選択である。ここでは選択した母数を estimand と呼ぶ。通常は、応用上重要であるとかその母数に関心があるとか

で選ぶと考えられている。説明が容易であるとの意味では、平均 μ と標準偏差 $1/\tau$ が候補に上がる。例えば、初期の研究である Govindarajulu (1966) ではこの母数を選択した上で線形最小不偏推定量を求めている。ところが、平均はともかく標準偏差の事後平均は直感的には違和感がある。何故なら標準偏差より分散の方が事後平均との相性が良さそうだからである。更には上でも述べたように、τ の informative 事前分布を仮定する際には τ の方が gamma 分布との相性が良い。結局直感的に妥当と見なされる estimand の候補として $(\mu, 1/\tau)$, $(\mu, 1/\tau^2)$, (μ, τ) が挙げられる。これらの中から一つを選択することは難しい問題である。選択の判断規準として、応用上とか解析結果の解釈の容易さなどが議論されるが、実際的ではない。むしろ、損失関数の選択とか推定の効率を上げる技術上の都合を原則にする方がより建設的である。

理論統計学では、推定法はリスクを小さくするように選ぶことが原則である。この原則では、予め損失 $\mathrm{L}(\hat{\theta}, \theta)$ を定めておく必要がある。ベイズ法の入門書では、損失

$$\mathrm{L}((\hat{\mu}, \widehat{1/\tau})), (\mu, 1/\tau) \;=\; (\hat{\mu} - \mu)^2 \,+\, (\widehat{1/\tau} - 1/\tau)^2$$

の下で $(\mu, 1/\tau)$ の事後平均がベイズリスクを最小にすること強調される。しかし、この理論は estimand の選択に応じて損失関数を決めているので、どんな estimand の選択も同等に正当化される。だから、適切なベイズ法での estimand の選択には役立たない。

現在多くの研究者が認めている損失に Kullback-Leibler 分離度 (2.2) がある。この分離度は、母数間の距離 (分離度) を予測分布間の距離として定義している。従って、同じ予測子を導く同等な母数は同一視するので、我々の議論にとって都合が良い。二つの Laplace 分布間の距離は次のように表される。

$$\begin{aligned} \mathrm{D}((\hat{\mu}, \hat{\tau}); (\mu, \tau)) \;&=\; n\frac{\tau}{\hat{\tau}} \left\{ \exp(-\hat{\tau}|\hat{\mu} - \mu|) + \hat{\tau}|\hat{\mu} - \mu| - 1 \right\} \\ &\quad + n \left\{ \frac{\tau}{\hat{\tau}} - \log \frac{\tau}{\hat{\tau}} - 1 \right\}. \end{aligned} \tag{3.2}$$

この損失は $\hat{\mu} \doteq \mu$ のとき

$$\frac{1}{2}\tau\hat{\tau}(\hat{\mu} - \mu)^2 + \frac{\tau}{\hat{\tau}} - \log\frac{\tau}{\hat{\tau}} - 1 \tag{3.3}$$

で近似できる。これから $(\tau\mu, \mu)$ の事後平均が上の近似損失を最小にすることが、細々した計算から、得られる。

Estimand の選択についてのこの結果は、上での直感的な議論とは異なる。しかし、奇妙な結果ではない。指数分布族では、Kullback-Leibler 分離度を損失にしたとき、ベイズリスクを最小にする estimand は自然母数である。例えば、正規分布 $\mathrm{N}(\mu, \sigma^2)$ では $(\mu/\sigma^2, 1/\sigma^2)$ となる。事前密度として reference prior を選ぶと、事後平均は $(\hat{\mu}, \hat{\sigma}^2) = (\bar{x}, s^2)$ 但し $s^2 = \sum(x_i - \bar{x})^2/(n-1)$ と同等になる。導かれる推定量は頻度論の立場でもごく自然である。しかも、上での議論したような直感的な estimand である (μ, σ^2) は現れない。

推定量は次のように一次元の数値積分を用いて計算でいる表現に帰着される。

$$\hat{\tau} = \frac{n\int\{\sum|x_i - \mu|\}^{-(n+1)}d\mu}{\int\{\sum|x_i - \mu|\}^{-n}d\mu} \tag{3.4}$$

$$\hat{\mu}\left(= \widehat{\frac{\tau\mu}{\hat{\tau}}}\right) = \frac{\int\mu\{\sum|x_i - \mu|\}^{-(n+1)}d\mu}{\int\{\sum|x_i - \mu|\}^{-(n+1)}d\mu} \tag{3.5}$$

一次元の数値計算に帰着されるのは、τ に関しての積分は gamma 関数の積分表現が適用できるからである。上式の被積分関数は区分的に $\mu - m$ のべき乗の形であり、陽に積分できる。だから n が小さいときは計算量は小さい。大きいときは近似式により計算できるとされている (Box and Tiao, 1962)。

上の議論にも拘わらず estimand は (μ, τ) とする方が魅力的である考える研究者もいるかも知れない。理論的な裏付けはなくても、$\tau\mu$ より μ

の方が意味も取りやすいからである。Reference 事前関数を仮定すると、τ の推定量は提案法と同じになるが、μ の推定量は

$$\tilde{\mu} = \frac{\int \mu\{\sum|x_i - \mu|\}^{-n} d\mu}{\int \{\sum|x_i - \mu|\}^{-n} d\mu}$$

と表される。推定量 (3.5) との違いはべき乗の係数だけに現れる。二つの推定量は大きくは変わらない。しかし、$n = 2$ の場合を比較すると、この推定量の分子の積分は存在しない。この事実は明確な欠点である。一方で、estimand を (μ, τ) と選択することによる長所は何らか見出しがたい。

ここで、改めて Laplace 密度の exponent は $-|\tau x - \tau\mu|$ と書き換えられることに注意する。即ち、密度関数が μ に依存するのは $\tau\mu$ に通してのみである。だから、分布間の距離に基づいてい議論する限りには μ が孤立して現れることはない。

表 1. 提案推定量の平均：$1/\tau$ の場合、繰り返し 10,000

標本サイズ:	3	4	5	6	7
平均:	1.0015	1.0046	0.9989	1.0064	0.9935

幾らかの数値比較を行った範囲では、上の推定量の性能は良好である。その過程で得られた一つの興味深い観察結果は、表 1 に与えたように、$1/\hat{\tau}$ が近似的には不偏である事実である。表 1 を一見すると、厳密にも不偏性が成り立つように見えるが、推定量の表現式 (3.4) を見ても厳密な不偏性は成り立ちそうもない。一方で全くの偶然でもなさそうである。実際同じ exponential power 分布族に属する正規分布では $\tau = 1/\sigma^2$ の事後平均が s^2 となって不偏である。しかも、Laplace 分布では必要な近似 (3.3) も不要である。

この観察結果は、直感的な estimand の選択が危ういことを示唆している。むしろ損失の選択から estimand を導く方が理論的な扱いが可能になり、推定量の柔軟な選択を可能にさせる。また、Laplace 分布ではベイ

ズ母数推定の理論的な展開を可能にする性質が満たされることが観察される。

上で観察された母数推定に対応する推定法は、logistic 分布の場合には困難である。実際、Laplace 分布の場合に比べて推定法を構築するための基礎的な性質が見当たらない。対応する性質を調べる。

1) 無情報事前分布では Jeffreys 事前関数は得られるが、より魅力的な reference prior は文献上見当たらない

2) Informative 事前分布の定義が難しい。Logistic 分布の範囲内で定義することは困難である。

3) Kullback-Leibler 分離度は陽に表現されない。

4) 推定量は二次元の数値積分が必要になる。

5) Estimand の選択についての理論的な展開が出来ない。

これらの事実を要するに、ベイズ推定では logistic 分布には有用な性質は見当たらない。

§4. 対数尤度比としての活性化関数

活性化関数は深層学習で用いられる非線形関数である。深層学習では、高次元のデータを隣り合う層の間で繰り返して逐次的に次元を縮小させる線形変換と、活性化関数によるデータの変換を通じて目的の情報に集約させる。深層学習は、人間の知的営みである認知機能を機械的操作で代用させる試みである人工知能の最大の成功例である。

当初のモデルは生体の反応でしばしば観察される閾値モデルであった。筋繊維に対する電気的刺激に代表される生体の反応は、連続値から二値 $R \to \{0,1\}$ への step 関数である。生体反応の典型として、より複雑な認知機能に際しても基本的な変換であると考えられた。しかし、そのような単純な変換に基づいて機械的に認知機能を表現することが不可能であるので、より一般的な $R \to [0,1]$ が用いられるようになった。この関数に非減少性の制約を入れると確率分布の分布関数に限られる。加えて、潜在的な閾値の分布を表現しているとの説明もできる。この目的のため

に用いられた分布関数が logistic 分布関数である。伝統的にこの分野では sigmoid 関数と呼ばれている。この関数は出力層での softmax 関数としても現れる二項分布の自然母数と平均母数の関係式、自然連結関数、である。しかし、これは logit モデルの場合と同じで回帰関数としての利用であって logistic 分布の関連は直接的ではない。

分類の精度を上げるために、中間層を多くする試みがなされて深層と名付けられた。中間層を増やすと閾値分布の拡張したモデルでは、勾配の喪失など扱いが困難になった。そのため広く用いられる活性関数として ReLU 関数

$$a_{RL}(x) = \mathrm{Max}\{0, x\} \tag{4.1}$$

が提案された。この関数は分布関数ではなく、且つ原点 $x = 0$ で折れている。しかし、従来の統計モデルでは折れ点に格別な意味をもたせることが多いが、深層学習では特別な意味はもたせない。関数の全体的な傾向に関心が払われる。

回帰関数を ReLU 関数の一般形である $\mathrm{Max}\{c, ax + b\}$ とした回帰モデルは hockey-stick モデルとして環境分野で利用された (Hasselblad ら, 1973)。この場合には折れ点が健康影響を及ぼさない閾値と理解できることが、このモデルが用いられた動機である。しかし、そのような利用には十分な配慮が必要である (Yanagimoto and Yamamoto, 1979)。一方で深層学習では ReLU 関数の折れ点に閾値の関連した解釈は不要である。

ReLU 関数から見れば、定数関数 $a_C(x) = 0$ とは大きく離れている。また、step 関数 $a_{St}(x) = 0 \text{ for } x < 0, = 1 \text{ for } x \geq 0$ とも異なる。むしろ、恒等関数 $a_I(x) = x$ の変形であり、その活性化関数と見なされる。恒等関数、より一般的には線形関数、は変換しても活性化関数としては無意味である。そもそも層を逐次進める際に、多次元の線形変換 (affine 変換) が成されているのからである。言い換えると、活性化関数を線形関数にすると多層にする意味はなく、一層のみでの変数変換に帰着する。このことは、深層学習を線形モデルとして扱うことに帰着する。従来の正規

性の仮定の下での線形モデルである。

そこで、活性化関数を対数尤度比の視点から見直す。対数尤度比は二つの密度関数 $p_1(x)$ と $p_0(x)$ に対して

$$\mathrm{LLR}(x) = \log\{p_1(x)/p_0(x)\} \tag{4.2}$$

で定義される。対数尤度比は、Neyman-Pearson の基本補題に見られるように、二つの密度間の相対測度を表す基本量である。検定理論を含めて、理論統計学における基本量である。なおここでの利用は尤度としてより密度比を扱うが、慣例により尤度比の用語を使う。

例 4.1. (恒等関数) 密度関数 $\varphi(x)$ を標準正規分布, $\mathrm{N}(0, 1)$, の密度関数として、 $p_1(x) = \varphi(x - 1/2)$ and $p_0(x) = \varphi(x + 1/2)$. 対数尤度比は

$$\mathrm{LLR}(x) = x \tag{4.3}$$

と恒等関数になる。

恒等関数が平均を移動させた正規密度の比から導かれる。上の例では、代表的な対の分布を選ぶためにある分布族の中で二つの異なる平均をもつ分布を選んだ。この選択は分布の対見つけるためには最も基本的である。更に興味深い点は、分布族として正規密度を用いて活性化関数を導出している事実である。正規分布の仮定とデータの線形変換とは相性が良い。だから多変量解析では多変量正規分布が仮定されて、データが線形変換されてきた。多変量解析を革新する深層学習では、線形変換を改良する活性化関数が望まれる。有用な活性化関数は、定数関数を改良するより恒等関数を改良することにより得られると見込まれる。

例 4.2. (定数関数) 二つの密度関数 $p_1(x)$ と $p_0(x)$ を R^1 を台に持つ同じ密度関数 $p(x)$ とする。対数尤度比は $\mathrm{LLR}(x) = 0$ となり定数関数になる。

対数尤度比から見ると、定数関数は自明な例である。一方で、定数関

数と対をなす step 関数 $a_{St}(x)$ を対数尤度比として導くのは困難である。

§5. Laplace 分布が誘導する活性化関数

活性化関数 ReLU を対数尤度比として導くような密度関数の対を、Laplace 密度関数とこれに関連した密度関数によって見つけることが出来る。記号を簡単にするために、正値区間 $(0,\infty)$ の定義関数を $\mathrm{I}_+(x)$ と書き、非正値区間の定義関数を $\mathrm{I}_-(x)$ と書く。

例 5.1. (ReLU 関数) 密度関数 $p_1(x)$ を Laplace 分布 $\mathrm{LA}(0,1)$ とする。別の密度関数を

$$p_0(x) = (2/3)\exp x\,\mathrm{I}_-(x) + (2/3)\exp -2x\,\mathrm{I}_+(x).$$

とする。その対数尤度比は

$$\mathrm{LLR}(x) = \log\frac{3}{4}\,\mathrm{I}_-(x) + \left(x + \log\frac{3}{4}\right)\,\mathrm{I}_+(x)$$

と表される。従って $a_{RL}(x) = \mathrm{LLR}(x) + \log(4/3)$. となるから ReLU 関数である。

密度関数 $p_0(x)$ は非対称 Laplace 密度関数と呼ばれる。異なる平均の指数分布を原点で左右に貼り合わせた密度関数で、Laplace 密度の直接的な拡張である。

例 5.2. (Ramp 関数) Laplace 分布 $\mathrm{LA}(0,1/2)$ の密度関数を $p(x)$ とする。ここで $p_1(x) = p(x-1/2)$, $p_0(x) = p(x+1/2)$ と置くと.

$$\begin{aligned} \mathrm{LLR}(x) &= -1/2 && \text{for } -1/2 \geq x \\ &= x && \text{for } 1/2 \geq x > -1/2 \\ &= 1/2 && \text{for } x > 1/2. \end{aligned} \tag{5.1}$$

右辺は $\mathrm{Max}\{-1/2, \mathrm{Min}\{x, 1/2\}\}$ と簡潔に書くことも出来る。.

この関数は $x = -1/2$ and $1/2$ の二点で折れ点をもつ。この形状から ramp 関数と呼ばれる。深層学習では $\mathrm{LLR}(x) + 1/2$ を hard-sigmoid 関

数と呼ばれて、一様分布の分布関数として扱われていた。また、回帰関数 (Mudelsee, 2000) あるいは損失 (Collobert ら, 2006) としても用いられる。

上の結果較べて、logistic 密度を用いた対数尤度比から ReLU 関数を導くことは折れ点が表現できないので不可能である。逆に活性化関数を分布関数と見た場合、logistic 分布関数と Laplace 分布関数は特に裾部分で似ている。密度関数での折れ点はなくなり C^1 級の関数になる。対数尤度比の視点から見ると、Laplace 密度と logsitic を入れ替えても大きな差はない。計算量が少し増加する上に、ReLU 関数のような既存の活性関数との関連が弱くなってしまう。

4, 5 節の内容の一部は大草孝介博士 (横浜市立大) との共同研究の結果であることを記して、改めて感謝します。

文献

Akahira, M. and Takeuchi, K. (1990). Loss of information associated with the order statistics and related estimators in the double exponential distribution case. *Australian J. Statist.*, **32**, 281-291.

Bcx, G. E. P. and Tiao, G. C. (1962). A further look at robustness via Bayes's theorem. *Biometrika*, **49**, 419-432.

Collobert, R., Sinz, F., Weston, J. and Bottou. L. (2006). Large scale transductive SVMs. *J. Machine Learn. Res.*, **7**,1687-1712.

Garvan, C.W. and Ghosh, M. (1997). Noninformative priors for dispersion models. *Biometrika*, **84**, 976-982.

Hasselblad, V., Creason, J.P., Nelson, W.C. (1973). Regression using "hockey stick" functions. Read at 101st Ann. Meet. Amer. Pub. Health Assoc.

Govindarajulu, Z. (1966). Best linear estimates under symmetric cen-

soring of the parameters of a double exponential population. *J. Am. Statist. Assoc.*, **61**, 248-258.

Keynes, J.M. (1911) The principal averages and the laws of error which lead to them. *J. Roy. Statist. Soc. A*, **74**, 322-331.

Kotz, S., Kozubowski, T. and Podgorski, K. (2001). *The Laplace Distribution and Generalizations: a Revisit with Applications to Communications, Economics, Engineering, and Finance* . Springer Science+Buisiness Media, New York.

Laplace, P.S. (1774). Memoire sur les suites recurro-recurrentes et sur leurs usages dans la theorie des hasards. *Mem. Acad. Roy. Sci. Paris* **6**, 621-656.

Mudelsee, M. (2000). Ramp function regression: a tool for quantifying climate transitions. *Comput. & Geosci.*, **26**, 293-307.

Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Am. Statist. Assoc.*, **103**, 681-686.

Subbotin, M.T. (1923). On the law of frequency of error. *Mat. Sb.*, **31**, 296-301.

Yanagimoto, T. and Yamamoto, E. (1979). Estimation of safe doses: Critical review of the hockey stick regression method. *Env. Health Persp.*, 32, 193-199.

Yanagimoto, T. and Ohnishi, T. (2009). Bayesian prediction of a density function in terms of e-mixture. *J. Statist. Plann. Inf.*, **139**, 3064-3075.