社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

極値統計量を用いたデータ発生領域の学習法

渡辺　一帆†　　渡辺　澄夫††

† 東京工業大学大学院総合理工学研究科電子機能システム専攻　〒 226–8503 横浜市緑区長津田 4259
†† 東京工業大学精密工学研究所　〒 226–8503 横浜市緑区長津田 4259
E-mail: †{kazuho23,swatanab}@pi.titech.ac.jp

あらまし　パターン認識，異常値の検出といった問題においては，学習によって，ある特定のパターンのデータが存在する領域を精度よく推定することで，入力データの識別が行なわれる．本研究は，データが発生する範囲を推定するために，1 次元の確率分布の端点を，サンプルの最大値および最小値を用いて推定する方法を考察する． サンプルの最大値，最小値の漸近分布は極値統計学において研究されてきた．そこで本研究では，サンプルの最大値，最小値に補正を加えることで，その端点を推定するときの推定誤差を導出し，その推定誤差を最小にする推定法を与える．
キーワード　データ発生領域，極値統計，極限分布

Learning the Data Region using Extreme-value Statistics

Kazuho WATANABE† and Sumio WATANABE††

† Department of Advanced Applied Electronics, Tokyo Institute of Technology　4259 Nagatsuta, Midori-ku, Yokohama, 226–8503 Japan
†† P&I Lab, Tokyo Institute of Technology　4259 Nagatsuta, Midori-ku, Yokohama, 226–8503 Japan
E-mail: †{kazuho23,swatanab}@pi.titech.ac.jp

Abstract　In the field of pattern recognition or outlier detection, it is desirable to estimate the region where the data of a particular class are generated. In other words, precise prediction is realized by accurately estimating the support of the distribution that generates the data. Considering the 1-dimensional distribution whose support is a finite interval, the data region is characterized by the maximum value and the minimum value in the samples. Limiting distributions of these values have been studied in the extreme-value theory in statistics. In this research, we propose a method to estimate the data region using the maximum value and the minimum value in the samples. We calculate the average loss of the estimator, and derive the optimal estimators for given loss functions.
Key words　data region, limiting distribution, extreme value statistics

1. ま え が き

パターン認識や異常値の検出などの一つの方法として，ある特定のクラスのデータや正常なデータの具体例を用いて，それらのデータの発生する領域を推定することが行なわれる．データの存在する領域と存在しない領域との境界を得ることで，入力データの識別がおこなわれる．データがある確率分布にしたがって発生していると仮定すれば，その確率分布が正の値をとる範囲を正確に推測することによって，高精度な予測が実現される．

しかし，データを発生している確率分布がある有限の範囲でのみ正の値をとる場合，そのモデルの対数尤度は発散してしまうため，統計的推測の漸近理論における正則条件は成立しない．

この場合の統計的推測についても様々な研究がなされており，統計的正則モデルとはかなり異なった性質をもつことが明らかにされてきた [2]．

他方では，ある確率分布から得られるサンプルの最大値，最小値等の性質は極値統計の理論で議論されてきた [1] [3]．

本研究では，一次元のデータ，もしくは，高次元データの一次元の指標といった確率変数の確率分布の端点を推定することで，データの発生する領域を推定する方法を提案する． 得られたサンプルの最大値，最小値に補正を加えて端点を推定する場

合，極値統計に基づいて最大値，最小値の漸近分布を考えることで，予測損失を最小とするような補正を与えることができることを示す．

まず第２章で極値統計の一般論として，最大値の漸近分布について述べ，第３章では，それに基づきデータの発生領域の推定法と，そのときの予測損失を導出する．得られる結果は１次元の確率分布についてのものであるが，高次元のデータについても様々な１次元の指標を考えると，この方法を適用することが考えられる．これについては第４章で考察を行ない，高次元データへ適用する場合に必要となる今後の課題について述べる．

2. 極値統計量と漸近分布

互いに独立に同じ分布 (密度関数を $f(x)$, 累積分布関数を $F(x)$ とする) に従う n 個のサンプル $X_1, X_2, \cdots, X_n$ を，大きさの順に並べ換え，

$$X_{(1)} \leqq X_{(2)} \leqq \cdots \leqq X_{(n)}$$

としたときの，k 番目の値で決まる確率変数 $X_{(k)} (1 \leqq k \leqq n)$ を順序統計量という．最小値 $\min_{1 \leqq i \leqq n} X_i$，最大値 $\max_{1 \leqq i \leqq n} X_i$ は特に極値統計量と呼ばれる．ここでは特に最大値 $M_n = \max_{1 \leqq i \leqq n} X_i$ のみを扱う．最小値 $\min X_i$ についても同様なことが成り立つ．

M_n の分布関数は，

$$\begin{aligned} P\{M_n \leqq x\} &= P\{X_1 \leqq x, \cdots, X_n \leqq x\} \\ &= F(x)^n \end{aligned}$$

で与えられる．M_n の密度関数 $f_{max}(x)$ は，この分布関数を微分して，

$$f_{max}(x) = n f(x) F(x)^{n-1}$$

となる．

$M_n = \max_{1 \leqq i \leqq n} X_i$ の分布関数 $F(x)^n$ の $n \to \infty$ での漸近分布について以下のことが知られている．

数列 $a_n (> 0), b_n$ を用いて，

$$a_n (M_n - b_n)$$

の分布関数がある分布関数 $G(x)$ に収束する場合，$G(x)$ は，次の３種類の形に限られることが知られている．

（1） $$G(x) = \exp(-e^{-x}) \quad -\infty < x < \infty$$

（2） $$G(x) = \begin{cases} 0 & (x \leqq 0) \\ \exp(-x^{-\alpha}) & (x > 0) \end{cases}$$

（3） $$G(x) = \begin{cases} \exp(-(-x)^{\alpha}) & (x \leqq 0) \\ 1 & (x > 0) \end{cases}$$

(ここで α はある正の実数)

最小値 $\min_{1 \leqq i \leqq n} X_i$ についても同様なことが言えて，対応する３種類の分布関数が知られている．

3. データ発生領域の推定法

n 個のデータ $X_1, X_2, \cdots, X_n$ が確率密度関数 $f(x)$ にしたがって得られたとする．$f(x)$ の累積分布関数を $F(x)$ とする．確率密度関数 $f(x)$ は以下の条件を満たすとする．

（i） $$\begin{cases} f(x) > 0 & (b < x < a) \\ f(x) = 0 & (\text{その他}) \end{cases}$$

（ii） $0 < \alpha, \beta < \infty$，$0 < A, B < \infty$ に対し，

$$\lim_{x \to a-0} (a-x)^{1-\alpha} f(x) = A$$

$$\lim_{x \to b+0} (x-b)^{1-\beta} f(x) = B$$

このとき，分布の端点である a, b を，

$$\hat{a} = \max_{1 \leqq i \leqq n} X_i + \frac{c_a}{n^{1/\alpha}}$$

$$\hat{b} = \min_{1 \leqq i \leqq n} X_i - \frac{c_b}{n^{1/\beta}}$$

を用いて推定する．

このときの予測損失を評価することで，予測損失が最小になるように c_a と c_b を決めることを考える．推定量 $\hat{a}$，$\hat{b}$ の性質はサンプルの最大値および最小値の性質によって決まるので，まずこれらの値の漸近分布について考察する．

[定理 1]　$M_n = \max_{1 \leqq i \leqq n} X_i$，$m_n = \min_{1 \leqq i \leqq n} X_i$ とするとき，$(\frac{A}{\alpha} n)^{1/\alpha} (M_n - a)$，$(\frac{B}{\beta} n)^{1/\beta} (m_n - b)$ の漸近分布は，それぞれ

$$G_{\max}(x) = \begin{cases} \exp(-(-x)^{\alpha}) & (x \leqq 0) \\ 1 & (x > 0) \end{cases}$$

$$G_{\min}(x) = \begin{cases} 0 & (x < 0) \\ 1 - \exp(-x^{\beta}) & (x \geqq 0) \end{cases}$$

となる．

(定理１の証明)

$f(x)$ についての条件から $x = a$ の近傍で，

$$f(x) = A(a-x)^{\alpha-1} + o((a-x)^{\alpha-1})$$

$$1 - F(x) = \frac{A}{\alpha}(a-x)^{\alpha} + o((a-x)^{\alpha})$$

と表せる．

$$u_n = \Big(\frac{A}{\alpha}n\Big)^{-1/\alpha}x + a$$

とおくと，$x \leqq 0$ では，

$$\begin{aligned}
& \Pr\{(\frac{A}{\alpha}n)^{1/\alpha}(M_n - a) \leqq x\} \\
=\ & F(u_n)^n \\
=\ & \{1 - (1 - F(u_n))\}^n \\
=\ & \Big\{1 - \frac{(-x)^\alpha}{n} + o(\frac{1}{n})\Big\}^n \\
\to\ & \exp(-(-x)^\alpha) \quad (n \to \infty)
\end{aligned}$$

$x > 0$ では $1 - F(u_n) = 0$ となるので，$G_{\max}(x)$ が得られた．また，$G_{\min}(x)$ も同様にして得られる．

(定理 1 の証明終)

［定理 2］ $\big((\frac{A}{\alpha}n)^{1/\alpha}(M_n - a),\ \ (\frac{B}{\beta}n)^{1/\beta}(m_n - b)\big)$ の分布関数を $G_n(s,t)$ とおくと，

$$\lim_{n\to\infty} G_n(s,t) = G_{\max}(s)G_{\min}(t)$$

となる．

(定理 2 の証明)

$$u_n = \Big(\frac{A}{\alpha}n\Big)^{-1/\alpha}s + a,\ \ v_n = \Big(\frac{B}{\beta}n\Big)^{-1/\beta}t + b$$

とおくと，

$$\begin{aligned}
& \Pr\{\big((\frac{A}{\alpha}n)^{1/\alpha}(M_n - a) \leqq s,\ (\frac{B}{\beta}n)^{1/\beta}(m_n - b)\big) \leqq t\} \\
=\ & F(u_n)^n - \{F(u_n) - F(v_n)\}^n \qquad (1)
\end{aligned}$$

定理 1 から

$$F(u_n)^n \to \exp(-(-s)^\alpha) \quad (n \to \infty)$$

であり，

$$\begin{aligned}
& \{F(u_n) - F(v_n)\}^n \\
=\ & \{1 - (1 - F(u_n)) - F(v_n)\}^n \\
=\ & \Big\{1 - \frac{(-s)^\alpha}{n} - \frac{t^\beta}{n} + o(\frac{1}{n})\Big\}^n \\
\to\ & \exp(-(-s)^\alpha - t^\beta) \quad (n \to \infty)
\end{aligned}$$

よって，

$$G_n(s,t) \to \exp(-(-s)^\alpha)(1 - \exp(-t^\beta)) \quad (n \to \infty)$$

(定理 2 の証明終)

定理 2 から，最大値 M_n と最小値 m_n は漸近的に独立になるので，それぞれの端点の推定は互いに独立して扱えばよい．そこで，以降は端点 $b = 0$ として，端点 a のみを推定するときの予測損失を考える．予測損失として，

$$E_{X^n}\Big[U(|a - \hat{a}|)\Big] \qquad (2)$$

を用いる．ここで，$E_{X^n}[\cdot]$ はサンプルの出方についての平均を表し，関数 $U(x)$ は，$U(0) = 0$ を満たす任意の解析関数とする．

［定理 3］ k を自然数とするとき，$n^{k/\alpha}|a - \hat{a}|^k$ の累積分布関数を $H_n(t)$ とおくと，

$$\lim_{n\to\infty} H_n(t) = \begin{cases} 1 - \exp(-\frac{A}{\alpha}(c_a + t^{1/k})^\alpha) \\ (t \geqq c_a^k) \\ \exp(-\frac{A}{\alpha}(c_a - t^{1/k})^\alpha) - \exp(-\frac{A}{\alpha}(c_a + t^{1/k})^\alpha) \\ (0 < t < c_a^k) \end{cases}$$

(定理 3 の証明)

$$\underline{t} = a - n^{-\frac{1}{\alpha}}(t^{\frac{1}{k}} + c_a)$$

$$\overline{t} = a - n^{-\frac{1}{\alpha}}(-t^{\frac{1}{k}} + c_a)$$

とおくと，$n^{k/\alpha}|a - \hat{a}|^k \leqq t$ は，

$$\underline{t} \leqq \max_{1 \leqq i \leqq n} X_i \leqq \overline{t}$$

と同値なので，

$$H_n(t) = \begin{cases} 1 - F(\underline{t})^n & (t \geqq c_a^k) \\ F(\overline{t})^n - F(\underline{t})^n & (t < c_a^k) \end{cases}$$

である．

$x = a$ の近傍では，

$$1 - F(x) = \frac{A}{\alpha}(a - x)^\alpha + o((a - x)^\alpha)$$

と表せるので，

$$\begin{aligned}
F(\underline{t})^n &= \Big\{1 - \frac{1}{n}\frac{A}{\alpha}(t^{1/k} + c_a)^\alpha + o\Big(\frac{1}{n}\Big)\Big\}^n \\
&\to \exp(-\frac{A}{\alpha}(c_a + t^{1/k})^\alpha) \quad (n \to \infty)
\end{aligned}$$

$$\begin{aligned}
F(\overline{t})^n &= \Big\{1 - \frac{1}{n}\frac{A}{\alpha}(-t^{1/k} + c_a)^\alpha + o\Big(\frac{1}{n}\Big)\Big\}^n \\
&\to \exp(-\frac{A}{\alpha}(c_a - t^{1/k})^\alpha) \quad (n \to \infty)
\end{aligned}$$

より定理を得る．

(定理 3 の証明終)

$c_a = 0$ のとき，定理 3 の漸近分布はワイブル分布関数になる．

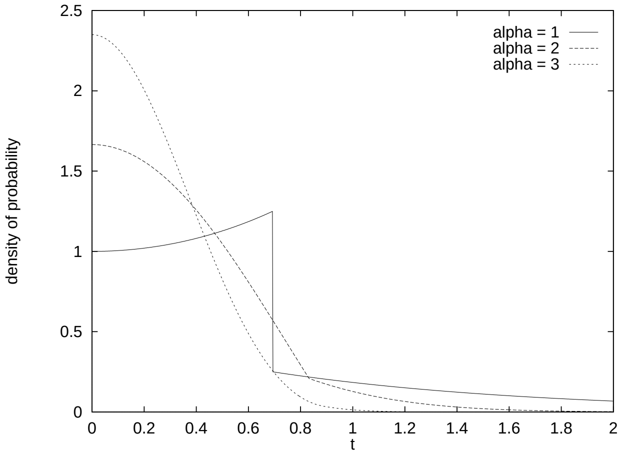

図 1　$n^{\frac{1}{\alpha}}|a-\hat{a}|$ の確率密度 $(\alpha=1,2,3)$

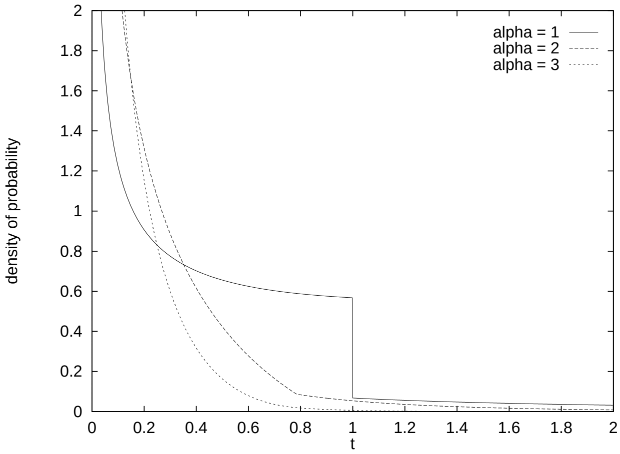

図 2　$n^{\frac{2}{\alpha}}|a-\hat{a}|^2$ の確率密度 $(\alpha=1,2,3)$

定理３の漸近分布を微分することで，$n^{k/\alpha}|a-\hat{a}|^k$ の確率密度関数 $h_n(t)$ が得られる．$k=1$ としたときの $h_n(t)$ を $\alpha=1,2,3$ の場合について図１に示し，$k=2$ としたときの $h_n(t)$ を図２に示す．ここで，最大値からの補正の係数である c_a の値は，後で述べるそれぞれの分布の平均を最小にするような c_a の値を用いた．

定理３で得られた漸近分布の平均を求めることで，以下の定理を得る．

［定理 4］　k を自然数とするとき，

$$\lim_{n\to\infty} E_{X^n}\left[n^{k/\alpha}|a-\hat{a}|^k\right]$$
$$= \begin{cases} c_a^k + \sum_{i=1}^k c_a^{k-i}\binom{k}{i}\left(\frac{\alpha}{A}\right)^{\frac{i}{\alpha}}\frac{i}{\alpha}(-1)^i\Gamma\left(\frac{i}{\alpha}\right) \\ (k\text{ が偶数}) \\ c_a^k + \sum_{i=1}^k c_a^{k-i}\binom{k}{i}\left(\frac{\alpha}{A}\right)^{\frac{i}{\alpha}}\frac{i}{\alpha}(-1)^i\{2\gamma\left(\frac{i}{\alpha},\frac{A}{\alpha}c_a^\alpha\right)-\Gamma\left(\frac{i}{\alpha}\right)\} \\ (k\text{ が奇数}) \end{cases}$$

ここで，$\gamma(x,p)$ と，$\Gamma(x)$ は不完全ガンマ関数とガンマ関数で，その定義式は，

$$\gamma(x,p)=\int_0^p t^{x-1}e^{-t}dt,\quad \Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$$

である．

(定理４の証明)

$$\begin{aligned} & E_{X^n}\left[n^{k/\alpha}|a-\hat{a}|^k\right] \\ \to\ & c_a^k - \int_0^{c_a^k}\exp\left(-\frac{A}{\alpha}(c_a-t^{1/k})^\alpha\right)dt \\ & \quad + \int_0^\infty \exp\left(-\frac{A}{\alpha}(c_a+t^{1/k})^\alpha\right)dt \quad (n\to\infty) \\ =\ & c_a^k - \frac{k}{A}\int_0^{\frac{A}{\alpha}c_a^\alpha}\left\{c_a-\left(\frac{\alpha}{A}t\right)^{\frac{1}{\alpha}}\right\}^{k-1}\left(\frac{\alpha}{A}t\right)^{\frac{1}{\alpha}-1}e^{-t}dt \\ & \quad + \frac{k}{A}\int_{\frac{A}{\alpha}c_a^\alpha}^\infty\left\{\left(\frac{\alpha}{A}t\right)^{\frac{1}{\alpha}}-c_a\right\}^{k-1}\left(\frac{\alpha}{A}t\right)^{\frac{1}{\alpha}-1}e^{-t}dt \end{aligned}$$

ここで $\left(c_a-\left(\frac{\alpha}{A}t\right)^{\frac{1}{\alpha}}\right)^{k-1}$ を２項展開して得られる．

(定理４の証明終)

定理４から (2) 式の予測損失について，次の２つの系が得られる．

［系 1］　$a_1=\frac{\partial U(0)}{\partial x}\neq 0$ のとき

$$\begin{aligned} & E_{X^n}\left[U(|a-\hat{a}|)\right] \\ & = \frac{a_1}{n^{1/\alpha}}\left[c_a-\left(\frac{\alpha}{A}\right)^{\frac{1}{\alpha}}\frac{1}{\alpha}\left\{2\gamma\left(\frac{1}{\alpha},\frac{A}{\alpha}c_a^\alpha\right)-\Gamma\left(\frac{1}{\alpha}\right)\right\}\right]+o\left(\frac{1}{n^{1/\alpha}}\right) \end{aligned}$$

また，これを最小にする c_a を c_a^* と表すと，

$$c_a^*=\left(\frac{\alpha}{A}\log 2\right)^{\frac{1}{\alpha}}$$

である．

(系１の証明)

$$U(|a-\hat{a}|)=a_1|a-\hat{a}|+O(|a-\hat{a}|^2)$$

より，定理３で $k=1$ とすればよい．また，c_a^* は，

$$\frac{\partial\gamma(x,p)}{\partial p}=p^{x-1}e^{-p}$$

であることを用いて，予測損失を c_a について微分して得られる．

［系 2］　$a_1=\frac{\partial U(0)}{\partial x}=0,\ a_2=\frac{1}{2}\frac{\partial^2 U(0)}{\partial x^2}\neq 0$ のとき

$$\begin{aligned} & E_{X^n}\left[U(|a-\hat{a}|)\right] \\ & = \frac{a_2}{n^{2/\alpha}}\left\{c_a^2-2\left(\frac{\alpha}{A}\right)^{\frac{1}{\alpha}}\frac{1}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)c_a+\left(\frac{\alpha}{A}\right)^{\frac{2}{\alpha}}\frac{2}{\alpha}\Gamma\left(\frac{2}{\alpha}\right)\right\}+o\left(\frac{1}{n^{2/\alpha}}\right) \end{aligned}$$

また，

$$c_a^*=\left(\frac{\alpha}{A}\right)^{\frac{1}{\alpha}}\frac{1}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)$$

である．

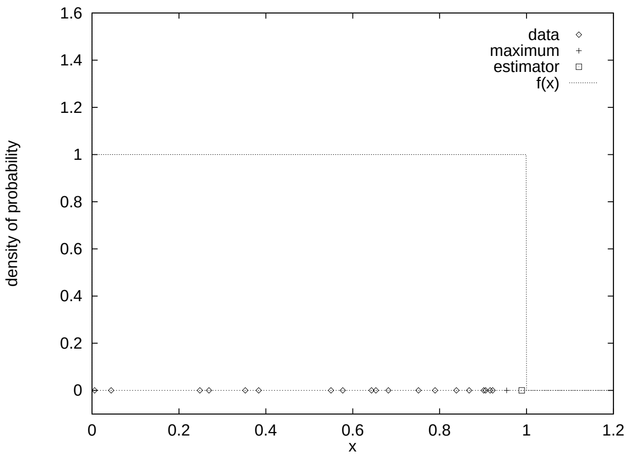

図 3　端点の推定の様子 ($\alpha = 1$ のとき)

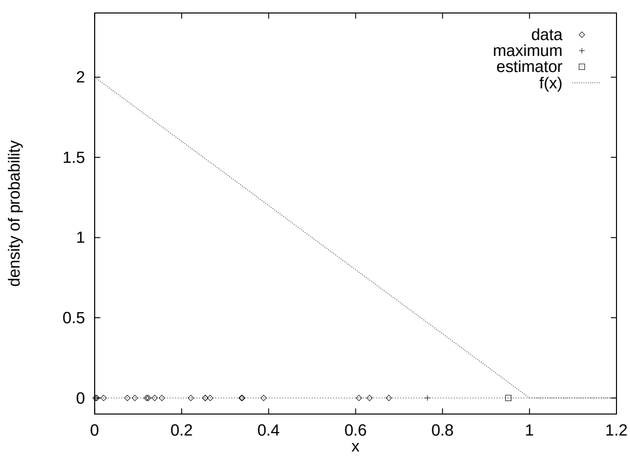

図 4　端点の推定の様子 ($\alpha = 2$ のとき)

(系 2 の証明)

$$U(|a-\hat{a}|) = a_2|a-\hat{a}|^2 + O(|a-\hat{a}|^3)$$

より，定理 4 で $k=2$ とすればよい．

4. 適　用　例

上記の推定法の適用例として，

$$f(x) = \begin{cases} 1 & (0 < x < 1) \\ 0 & (\text{その他}) \end{cases}$$

のとき，確率密度関数 $f(x)$ に従う n 個のサンプルから，分布の端点 $a = 1.0$ を推定する様子を図 3 に示す．この場合は $\alpha = A = 1.0$ であり，$E_{X^n}\left[|\hat{a}-a|\right]$ を最小にするような推定量 $\hat{a}$ は，

$$\hat{a} = \max_{1 \leqq i \leqq n} X_i + \frac{\log 2}{n}$$

で与えられる．図には $n = 20$ とした場合を示した．

次に，

$$f(x) = \begin{cases} 2-2x & (0 < x < 1) \\ 0 & (\text{その他}) \end{cases}$$

の場合に，$a = 1.0$ を推定する様子を図 4 に示す．このとき，$\alpha = A = 2.0$ であり，$E_{X^n}\left[|\hat{a}-a|\right]$ を最小にするような推定量 $\hat{a}$ は，

$$\hat{a} = \max_{1 \leqq i \leqq n} X_i + \sqrt{\frac{\log 2}{n}}$$

で与えられる．

5. 考　　察

データが発生する領域が有限の場合に，その領域の推定法として，一次元の確率分布について，その端点を推定する方法を与えた．一次元の確率分布の端点の推定においては，得られたデータの最大値や最小値を用いた推定法を考えると，極値統計学に基づいて，それらの漸近分布を考察することで，予測損失を最小にするような補正を与えることができることを示した．ここでは一次元の確率分布について，その端点を推定することでデータの発生する領域を推定する方法を与えたが，一般の高次元のデータについても何らかの一次元の指標についてはこの方法を適用することができる．いくつかの指標を用いることで高次元のデータの発生する領域の推定を行なうことも可能である．

また，これまでの議論では，損失関数として $U(|\hat{a}-a|)$ を評価し，この損失に対し最適な推定法を導出したが，端点を過大に推定するのと，過小に推定する場合で損失が異なるような場合にも，同様な議論により損失を最小化する推定法を導くことができる．l_1, l_2 を正の定数とし，関数 $L(x)$ を，

$$L(x) = \begin{cases} -l_1 x & (x < 0) \\ l_2 x & (x \geqq 0) \end{cases}$$

とすると，予測損失として，$E_{X^n}\left[L(\hat{a}-a)\right]$ を用いた場合に，同様の議論から次の定理が成り立つ．

［定理 5］　$X_1, \cdots, X_n$ は，確率分布 $f(x)$ から独立に得られたとする．また密度関数 $f(x)$ は，前章の条件 (i),(ii) を満たすとする．このとき，

$$\begin{aligned} & E_{X^n}\left[L(\hat{a}-a)\right] \\ = & \frac{l_2}{n^{\frac{1}{\alpha}}}\left[c_a - \left(\frac{\alpha}{A}\right)^{\frac{1}{\alpha}}\frac{1}{\alpha}\left\{\left(1+\frac{l_1}{l_2}\right)\gamma\left(\frac{1}{\alpha}, \frac{A}{\alpha}c_a^{\alpha}\right) - \Gamma\left(\frac{1}{\alpha}\right)\right\}\right] + o\left(\frac{1}{n^{\frac{1}{\alpha}}}\right) \end{aligned}$$

であり，これを最小にする c_a を c_a^* とすると，

$$c_a^* = \left(\frac{\alpha}{A}\log\left(1+\frac{l_1}{l_2}\right)\right)^{\frac{1}{\alpha}}$$

である．

これまで述べたデータの発生領域の推定法においては，データを発生している確率密度関数 $f(x)$ について，定数 α, β，A, B は与えられたものとして最適な推定法を導出した．これらの定数は確率密度関数 $f(x)$ の端点付近の変化によって決まるもので，実際にデータのみから領域を推定する場合は，これらの定数も推定する必要がある．これらのパラメータの推定法は今後の課題である．

本研究では１次元の確率分布の端点の推定に，得られたデータの最大値および最小値のみを用いたが，端点を推定する方法は他にも様々なものが考えられる．ここで提案した推定法を他の推定法と比較することは今後の課題である．

6．結　　論

データの発生する領域の推定法として，１次元の確率分布から得られたデータの最大値，最小値を用いた推定法を提案した．データの最大値，最小値の漸近分布を導出し，予測損失を最小にするような推定法を与えた．

文　　献

[1] M.R.Leadbetter,G.Lindgren,H.Rootzen "Extremes and Related Properties of Random Sequences and Processes", Springer-Verlag,New York Heidelberg Berlin,1983.

[2] M.Akahira,K.Takeuchi "Non-Regular Statistical Estimation", Springer-Verlag,New York,1995.

[3] E.J.Gumbel(河田竜夫，岩井重久，加瀬滋男監訳) "極値統計学", 生産技術センター社,1962.