<table>
<thead>
<tr>
<th>著者</th>
<th>WATANABE Kazuho, WATANABE Sumio</th>
</tr>
</thead>
<tbody>
<tr>
<td>編集者</td>
<td>Interdisciplinary Information Sciences</td>
</tr>
<tr>
<td>巻</td>
<td>13</td>
</tr>
<tr>
<td>号</td>
<td>2</td>
</tr>
<tr>
<td>頁范围</td>
<td>151-161</td>
</tr>
<tr>
<td>年</td>
<td>2007</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10097/45537">http://hdl.handle.net/10097/45537</a></td>
</tr>
</tbody>
</table>
Estimating the Data Region Using Minimum and Maximum Values

Kazuho WATANABE¹ and Sumio WATANABE²

¹Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology
²Precision and Intelligence Laboratory, Tokyo Institute of Technology

Received March 6, 2006; final version accepted March 12, 2007

In the field of pattern recognition or outlier detection, it is often necessary to estimate the region where data of a particular class are generated. In other words, it is required to accurately estimate the support of the distribution that generates the data. Considering the 1-dimensional distribution whose support is a finite interval, the data region is estimated effectively 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 show the average loss of the estimator and derive the optimally improved estimators for given loss functions. The method can be extended to estimate the higher dimensional input space.

KEYWORDS: data region, extreme-value theory, asymptotic distribution, novelty detection

1. Introduction

In the field of pattern recognition or outlier detection, it is often necessary to estimate the region where data of a particular class are generated. By setting a boundary that identifies the data region by learning from only examples of the target class, a new input is detected despite whether it is in the target class or it is, in fact, unknown.

Estimating the data region is one of the unsupervised learning tasks and is often used for outlier detection or novelty detection where the detection of uncharacteristic objects from a dataset is required. Unlike in the usual classification problems, in these problems, what is important is setting a boundary, rather than to distinguish two or more classes, to describe a particular kind of object represented by training samples. Ideally, the set boundary should cover all the objects of the target class and should have all other possible objects outside of it. This technique is called data domain description or one-class classification and is carried out by estimating the region where the data of the target class may appear.

The estimated region gives the idea how the target class is described in the space of data. This corresponds to identifying a specific individual in biometrics such as face, iris and fingerprint authentication. This technique can also be used for a classification problem where obtaining samples from one class is easy while that from the other class requires high costs. In a fault detection system, for instance, it is expensive to obtain the (outlier) data that represent the faulty situations of a machine. By estimating the region of only the target data that represent the normal situations of the machine, one can detect the faulty datum that shows up outside the region.

Various methods have been developed to estimate the data region. Most often the probability density of the data is estimated using Parzen density estimation or Gaussian mixture model (Markou and Singh, 2003). In these methods, the region over some probability threshold is estimated. However, setting the thresholds remains a major problem with these techniques. Assuming the data are generated from a probability distribution that has a compact support, the more precise prediction for a new input is realized by more accurately estimating the support of the distribution. For this purpose, some methods inspired by the support vector machines were proposed (Scholkopf et al., 2001; Tax and Duin, 2004).

However, in a parametric model of the distribution whose support is a finite interval and depends on a parameter, the log-likelihood of the model diverges and is not differentiable with respect to its parameter. Hence, the regularity conditions of statistical estimation do not hold. For example, uniform distributions and Beta distributions are included in such non-regular models. Their properties have been discussed in some studies and it has been known that they have quite different properties from those of regular statistical models. In fact, several examples where the maximum likelihood estimators can be asymptotically inadmissible were shown (Akahira and Takeuchi, 1995). In estimating the support of the 1-dimensional distribution, the maximum and minimum values in the samples are used effectively. The properties of these values in the samples taken from a distribution have been studied in the theory of extreme-value statistics (Leadbetter et al., 1983; Gumbel, 1958). Also, estimators of the 1-dimensional data region using several number of extreme values have been devised based on the minimization of their asymptotic variance in some studies (Hall, 1982; Dekkers et al., 1989; Hall and Wang 1999).

* Corresponding author. E-mail: kazuho23@pi.titech.ac.jp
In this research, we focus on a new method to estimate the data region using the maximum and minimum values in the samples. We define the average loss of the estimator and derive the asymptotic behavior of it based on the theory of extreme-value statistics. The optimally improved estimators for given loss functions are derived.

In Section 2, the general theory of extreme-value statistics is introduced. In Section 3, we propose a method to estimate the data region and derive the average loss of the method. This method is derived by assuming some conditions on the distribution that generates the data. Therefore, we need to estimate some parameters of the distribution in order to use this method without any information of these parameters. In Section 4, a method to estimate these additional parameters is given. The efficiency of this method is demonstrated by simulation in Section 5.

The proposed method can also be applied to estimate a multi-dimensional data region by taking a 1-dimensional data set from multi-dimensional data. Using mappings, we obtain sets of 1-dimensional data. Then, a multi-dimensional data region is estimated by applying the proposed method to each 1-dimensional data set. In the latter half of Section 5, we present an experimental study on 2-dimensional data and demonstrate experimentally that the proposed method works effectively to estimate a 2-dimensional data region. Discussion and conclusion are given in Section 6 and Section 7. The proofs of theorems in Section 3 are in Appendix.

2. Extreme-Value Statistics

Suppose that \( X_n = \{X_1, X_2, \ldots, X_n\} \) is a set of \( n \) samples independently and identically taken from a probability distribution with a density function \( f(x) \) and a distribution function \( F(x) \). Putting these samples in the order so that

\[
X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)},
\]

the random variable defined by the \( k \)-th smallest value \( X_{(k)} (1 \leq k \leq n) \) is called the order statistic. Especially, the extremely large or small variables such as \( \max_{1 \leq i \leq n} X_i \) and \( \min_{1 \leq i \leq n} X_i \) are called extreme-value statistics. In this section, we discuss the maximum value \( M_n = \max_{1 \leq i \leq n} X_i \).

The distribution function of \( M_n \) is given by

\[
P[M_n \leq x] = P[X_1 \leq x, \ldots, X_n \leq x] = F(x)^n.
\]

Differentiating this distribution function, we obtain the density function \( f_{\max}(x) \) of \( M_n \) as follows,

\[
f_{\max}(x) = nf(x)F(x)^{n-1}.
\]

The following property of the maximum value \( M_n = \max_{1 \leq i \leq n} X_i \) is well known on its asymptotic distribution function \( F(x)^n \) as \( n \to \infty \) (Leadbetter et al., 1983; Gumbel 1958). If, for some sequences \( a_n (> 0) \) and \( b_n \), the distribution function of

\[
a_n(M_n - b_n)
\]

has an asymptotic distribution function \( G(x) \), then \( G(x) \) have one of the following three forms by changing the location and the scale.

1. \( G(x) = \exp(-e^{-x}) \) \( (-\infty < x < \infty) \),
2. \( G(x) = \begin{cases} 0 & (x \leq 0), \\ \exp(-x^{-\alpha}) & (x > 0), \end{cases} \)
3. \( G(x) = \begin{cases} \exp(-(-x)^{\beta}) & (x \leq 0), \\ 1 & (x > 0), \end{cases} \)

for some \( \alpha > 0 \).

This property of the maximum of samples is equivalent to that of the sum of samples stated in the central limit theorem. The minimum value \( \min_{1 \leq i \leq n} X_i \) also has the same property and corresponding three forms of distribution are known.

3. Estimation of the Data Region

We assume the following conditions on the density function \( f(x) \).

(i) \( \begin{cases} f(x) > 0 & (b < x < a), \\ f(x) = 0 & \text{(otherwise)}. \end{cases} \)

(ii) For \( 0 < \alpha, \beta < \infty \) and \( 0 < A, B < \infty \),

\[
\lim_{x \to a^-} (a - x)^{1-\alpha} f(x) = A,
\]

\[
\lim_{x \to a^+} (x - b)^{1-\beta} f(x) = B.
\]
The condition (i) means that the support of the distribution is a finite interval \([b, a]\). The condition (ii) means the behaviors of \(f(x)\) around the endpoints \(a\) and \(b\) are determined by the parameters \(\alpha, A\) and \(\beta, B\) respectively. We see that in the neighborhood of \(a\), \(f(x)\) diverges for \(0 < \alpha < 1\) and converges to the positive number \(A\) for \(\alpha = 1\) or to 0 for \(\alpha > 1\), and the same holds for the parameter \(\beta\) around \(b\). For example, when \(f(x)\) is the Beta distribution on the interval \([0, 1]\), that is, \(f(x) = \frac{1}{B(\phi_1, \phi_2)} x^{\phi_1-1}(1-x)^{\phi_2-1}\) where \(B(\phi_1, \phi_2) = \int_{0}^{1} u^{\phi_1-1}(1-u)^{\phi_2-1} du\), these conditions are satisfied with \(a = 1, b = 0, \alpha = \phi_1, \beta = \phi_2, A = B = \frac{1}{B(\phi_1, \phi_2)}\).

Under the above conditions, we estimate \(a\) and \(b\), the endpoints of the density \(f(x)\), by the estimators \(\hat{a}, \hat{b}\). We define the estimators \(\hat{a}, \hat{b}\) using the maximum and minimum values in the samples,

\[
\hat{a} = \max_{1 \leq i \leq n} X_i + \frac{c_a}{n^{1/\alpha}},
\]

\[
\hat{b} = \min_{1 \leq i \leq n} X_i - \frac{c_b}{n^{1/\beta}},
\]

where \(c_a\) and \(c_b\) are the coefficients that determine the correction from the maximum and minimum values respectively. The average loss of these estimators is defined later. Then we show the asymptotic behavior of the average loss and derive the optimal \(c_a\) and \(c_b\) that minimize the loss. Before defining the average loss of \(\hat{a}\) and \(\hat{b}\), let us consider the asymptotic distributions of the maximum and minimum values in the samples that determine the properties of the estimator \(\hat{a}\) and \(\hat{b}\). Following Theorem 1 and Theorem 2 are obtained.

**Theorem 1.** For \(M_n = \max_{1 \leq i \leq n} X_i, m_n = \min_{1 \leq i \leq n} X_i\), the asymptotic distributions of \((\frac{4}{9} n)^{1/\alpha}(M_n - a)\) and \((\frac{8}{9} n)^{1/\beta}(m_n - b)\) have following distribution functions \(G_{\max}(x)\) and \(G_{\min}(x)\),

\[
G_{\max}(x) = \begin{cases} 
\exp(-(-x)\alpha) & (x \leq 0), \\
1 & (x > 0), \\
0 & (x < 0), \\
1 - \exp(-x\beta) & (x \geq 0).
\end{cases}
\]

\[
G_{\min}(x) = \begin{cases} 
\exp(-(-x)^\alpha) & (x \leq 0), \\
1 & (x > 0), \\
0 & (x < 0), \\
1 - \exp(-x\beta) & (x \geq 0).
\end{cases}
\]

**Theorem 2.** Denoting by \(G_n(s, t)\ the joint distribution function of \((\frac{4}{9} n)^{1/\alpha}(M_n - a), (\frac{8}{9} n)^{1/\beta}(m_n - b)\),

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

From Theorem 2 it is noted that the maximum \(M_n\) and minimum \(m_n\) are asymptotically independent of each other. Therefore we consider the estimation of each endpoints separately. More specifically, we put the left endpoint \(b = 0\) and consider estimating only the right endpoint \(a > 0\) hereafter. The estimator \(\hat{b}\) of the left endpoint \(b\) and the coefficient \(c_b\) are determined by the estimator \(\hat{a}\) and coefficient \(c_a\) derived from the set of samples \([-X_1, \ldots, -X_n]\).

We define the average loss of the estimator \(\hat{a}\) eq. (1) by

\[
E_{X_1}[U(|a - \hat{a}|)],
\]

where \(E_{X_1}[\cdot]\) denotes the expectation value over all sets of samples and the function \(U(x)\) is an arbitrary analytic function which satisfies \(U(0) = 0, \text{and } U(x) \geq 0\). We first discuss the optimal estimators for the symmetric loss functions given in eq. (3). The optimal estimators for asymmetric loss functions are also discussed in Section 6.

In order to evaluate the average loss eq. (3), we prove the following Theorem 3 and Theorem 4.

**Theorem 3.** Denoting by \(H_n(t)\ the distribution function of \(n^{1/\alpha}(\hat{a} - a)^k\) for any odd number \(k > 0\) and by \(\overline{H}_n(t)\) the distribution function of \(n^{k/\alpha}(|\hat{a} - a|^k)\) for any natural number \(k\),

\[
\lim_{n \to \infty} H_n(t) = \begin{cases} 
1 & (t \geq c_a^k), \\
\exp\left(-\frac{A}{\alpha}(c_a - t^{1/\alpha})^\alpha\right) & (t < c_a^k),
\end{cases}
\]

where \(t^{1/\alpha} = |t|^{1/\alpha}\ for \ t \leq 0\ and

\[
\lim_{n \to \infty} \overline{H}_n(t) = \begin{cases} 
1 - \exp\left(-\frac{A}{\alpha}(c_a + t^{1/\alpha})^\alpha\right) & (t \geq c_a^k), \\
\exp\left(-\frac{A}{\alpha}(c_a - t^{1/\alpha})^\alpha\right) - \exp\left(-\frac{A}{\alpha}(c_a + t^{1/\alpha})^\alpha\right) & (0 < t < c_a^k),
\end{cases}
\]

If \(c_a = 0\, the\ above\ asymptotic\ distribution\ functions\ are\ Weibull\ distribution\ functions.\ By\ differentiating\ the\ above\ distribution\ functions\ eq. (5),\ we\ obtain\ the\ asymptotic\ density\ functions\ of\ \(n^{k/\alpha}(|\hat{a} - a|^k).\)
We illustrate them in Fig. 1 for \( k = 1, C_{11} = 1, 2, 3 \), and in Fig. 2 for \( k = 2, C_{11} = 1, 2, 3 \). In these figures the values of the coefficients \( c_a \) are set to be optimal described below.

Evaluating the expectation values of \( n^k |\bar{a} - a|^k \) and \( n^k |\bar{a} - a|^k \), we obtain the following theorem.

**Theorem 4.** For an arbitrary natural number \( k \),

\[
\lim_{n \to \infty} E_X[n^{k/\alpha}(\bar{a} - a)^k] = c_a^k + \sum_{i=1}^{k} c_a^{k-i} \binom{k}{i} \left( \frac{\alpha}{A} \right)^\frac{i}{\alpha} (-1)^i \Gamma \left( \frac{i}{\alpha} \right),
\]

and if \( k \) is an odd number,

\[
\lim_{n \to \infty} E_X[n^{k/\alpha}|\bar{a} - a|^k] = c_a^k + \sum_{i=1}^{k} c_a^{k-i} \binom{k}{i} \left( \frac{\alpha}{A} \right)^\frac{i}{\alpha} (-1)^i \left\{ 2 \gamma \left( \frac{i}{\alpha}, \frac{A}{\alpha} c_a^2 \right) - \Gamma \left( \frac{i}{\alpha} \right) \right\},
\]

where \( \gamma(x, p) \) and \( \Gamma(x) \) are respectively the incomplete gamma function and the gamma function. The definitions of these functions are

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

From Theorem 4 we obtain the following two corollaries on the average loss eq. (3).

**Corollary 1.** If \( a_1 = \frac{\partial U(0)}{\partial x} \neq 0 \), then

\[
E_X[U(|a - \bar{a}|)] = \frac{a_1}{n^{1/\alpha}} \left[ c_a - \left( \frac{\alpha}{A} \right)^\frac{1}{\alpha} \left\{ 2 \gamma \left( \frac{1}{\alpha}, \frac{A}{\alpha} c_a^2 \right) - \Gamma \left( \frac{1}{\alpha} \right) \right\} \right] + o \left( \frac{1}{n^{1/\alpha}} \right).
\]

Denoting the optimal coefficient \( c_a^* \) that minimizes this average loss by \( c_a^* \),

\[
c_a^* = \left( \frac{\alpha}{A} \log 2 \right)^\frac{1}{\alpha}.
\]
(Proof of Corollary 1). Since

\[ U(\alpha, A) = \alpha \pi - \frac{\beta \pi}{\alpha} \]

put \( k = 1 \) in eq. (6). We obtain \( c_1^* \) by differentiating the average loss with respect to \( c_1 \) using

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

(Q.E.D.)

Corollary 2. If \( a_1 = \frac{dU(0)}{d\alpha} = 0 \) and \( a_2 = \frac{d^2U(0)}{d\alpha^2} \neq 0 \), then

\[
E_X[U(\alpha, A)] = \frac{a_2}{n^{3/\alpha}} \left( c_2^2 - 2 \left( \frac{\alpha}{A} \right)^{\frac{1}{2}} \Gamma \left( \frac{1}{\alpha} \right) c_1 + \left( \frac{\alpha}{A} \right)^{\frac{3}{2}} \Gamma \left( \frac{2}{\alpha} \right) \right) + o \left( \frac{1}{n^{3/\alpha}} \right),
\]

and

\[
c_2^* = \left( \frac{\alpha}{A} \right)^{\frac{1}{2}} \Gamma \left( \frac{1}{\alpha} \right).
\]

(Proof of Corollary 2). Since

\[ U(\alpha, A) = \alpha \pi - \frac{\beta \pi}{\alpha}, \]

put \( k = 2 \) in eq. (7). (Q.E.D.)

4. Estimation of the Other Parameters

In the previous section, the optimal estimators were derived under the conditions (i) and (ii) with parameters \( \alpha, \beta, A, B \). These parameters determine the behavior of the density function \( f(x) \) around the endpoints. If we estimate the data region without any information on these parameters, we need to estimate them using samples. In this section, we consider the case only the right endpoint is estimated and propose a method to estimate the parameters \( \alpha \) and \( A \) of condition (ii) in the previous section.

The asymptotic distribution of the maximum value in the \( k \) samples is uniquely specified if parameters \( \alpha, A \) and endpoint \( a \) are given, and its density function \( p_k(x|\alpha, A, a) \) is given by

\[
p_k(x|\alpha, A, a) = Ak(a - x)^{\alpha-1} \exp \left( -\frac{Ak}{\alpha} (a - x) \right),
\]

from Theorem 1 in the previous section.

Suppose that \( m \) samples \( x_1, x_2, \ldots, x_m \) are independently and identically taken from the above distribution. The log-likelihood function \( L(\alpha, A, a) \) for the \( m \) samples is given by

\[
L(\alpha, A, a) = \sum_{i=1}^{m} \left\{ \log Ak + (\alpha - 1) \log(a - x_i) - \frac{Ak}{\alpha} (a - x_i) \right\}.
\]

By dividing \( n \) samples taken from \( f(x) \) into \( m \) groups and taking the maximum value in each group, the samples of maximum values are obtained. Assuming these are taken from the density eq. (12), we propose to estimate parameters \( \alpha \) and \( A \) by maximizing the log-likelihood eq. (13) in the case \( k = n/m \). Parameters \( \beta, B \) and the left endpoint \( b \) is estimated in the same way using the set of samples \( \{-x_1, \ldots, -x_m\} \). Both \( m \) and \( k \) should be greater as \( n \) grows to ensure the convergence to the asymptotic distribution eq. (12) and acquire sufficient samples for the estimation of the parameters \( \alpha \) and \( A \). This means \( k \) and \( m \) should grow in the order smaller than \( n \).

The log-likelihood eq. (13) can be maximized by an iterative algorithm. However, the maximization with respect to three variables \( \alpha, A \) and \( a \) sometimes makes the log-likelihood to diverge. For this reason, based on the result of the previous section, we propose the more stable and reduced variate method where only two variables \( \alpha \) and \( A \) need to be estimated.

In the previous section, the endpoint \( a \) was estimated by the estimator

\[
\hat{a} = \max_{1 \leq i \leq n} X_i + \frac{c_1^*}{\alpha^{1/\alpha}}
\]

with given \( \alpha \) and \( A \). From Corollary 1 and Corollary 2 in the previous section, \( c_1^* \) that minimizes the average loss is represented by parameters \( \alpha \) and \( A \). Therefore, the log-likelihood function eq. (13) becomes a function of two parameters \( \alpha \) and \( A \) by replacing the parameter \( a \) with that of the form...
\[ a = \max_{1 \leq i \leq n} X_i + \frac{c^*_n}{n^{1/\alpha}}. \]  

(14)

Denoting by \( L(\alpha, A) \) this log-likelihood function with two variables, we propose to estimate parameters \( \alpha \) and \( A \) by maximizing \( L(\alpha, A) \) with an iterative algorithm. The gradient \( \left( \frac{\partial L}{\partial A}, \frac{\partial L}{\partial \alpha} \right) \) of \( L(\alpha, A) \) used in the iterative algorithm is given by

\[
\frac{\partial L}{\partial \alpha} = \sum_{i=1}^{m} \left[ \frac{1}{\alpha} \frac{Ak}{\alpha} (a - x_i)^\alpha \right] \log(a - x_i) \]

\[
+ \frac{Ak}{\alpha^2} (a - x_i)^\alpha + \frac{1}{\alpha - x_i} \log(a - Ak(a - x_i)^\alpha) \frac{\partial a}{\partial \alpha},
\]

\[
\frac{\partial L}{\partial A} = \sum_{i=1}^{m} \left[ \frac{1}{A} \frac{k}{\alpha} (a - x_i)^\alpha + \frac{1}{\alpha - x_i} \log(a - Ak(a - x_i)^\alpha) \frac{\partial a}{\partial A} \right],
\]

where

\[
\frac{\partial a}{\partial \alpha} = \frac{1}{\alpha^2} \left( 1 - \log \frac{\alpha log 2}{An} \right) \left( \frac{\alpha}{An} \log 2 \right)^{1/\alpha},
\]

\[
\frac{\partial a}{\partial A} = -\frac{1}{\alpha A} \left( \frac{\alpha}{An} \log 2 \right)^{1/\alpha},
\]

using \( c^*_n \) of Corollary 1 in eq. (14), or

\[
\frac{\partial a}{\partial \alpha} = \frac{1}{\alpha^2} \left( \left( 1 - \log \frac{\alpha}{An} \right) \Gamma \left( 1 + \frac{1}{\alpha} \right) - \Gamma' \left( 1 + \frac{1}{\alpha} \right) \right) \left( \frac{\alpha}{An} \right)^{1/\alpha},
\]

\[
\frac{\partial a}{\partial A} = -\frac{1}{\alpha A} \left( \frac{\alpha}{An} \right)^{1/\alpha} \Gamma \left( 1 + \frac{1}{\alpha} \right),
\]

using \( c^*_n \) of Corollary 2.

5. Experiments

In this section, we demonstrate the efficiency of the proposed method by experimental results.

5.1 Estimation of the 1-dimensional data region

First, taking \( n \) samples from the uniform distribution

\[ f(x) = \begin{cases} 
1 & (0 < x < 1), \\
0 & \text{(otherwise)},
\end{cases} \]

we estimated the right endpoint \( a = 1.0 \). This is the case when the parameters \( \alpha = A = 1.0 \). If \( \alpha \) and \( A \) are known, the estimator \( \hat{a} \) that minimizes \( E_X[|\hat{a} - a|] \) is given by

\[ \hat{a} = \max_{1 \leq i \leq n} X_i + \frac{\log 2}{n} \]

from Corollary 1 in Section 3. We set \( n = 10000 \) and estimated the endpoint \( a = 1.0 \) using the above estimator \( \hat{a} \).

Then, on the assumption that \( \alpha \) and \( A \) were unknown, we estimated the endpoint \( a \), estimating \( \alpha, A \) simultaneously by using the method described in Section 4. In this case, we divided 10000 samples into 100 groups and obtained 100 samples from the distribution of maximum value eq. (12).

This estimation was simulated 2000 times respectively and the averages of \( n(a - \hat{a}) \) and \( n|a - \hat{a}| \) were calculated. The results are presented in Table 1. The standard deviations are in parentheses.

The other experiment was conducted in the case \( n \) samples were taken from the density

\[ f(x) = \begin{cases} 
2 - 2x & (0 < x < 1), \\
0 & \text{(otherwise)}. 
\end{cases} \]

In this case the endpoint \( a = 1.0 \) and the parameters \( \alpha = A = 2.0 \). If \( \alpha \) and \( A \) are known, the estimator \( \hat{a} \) that minimizes \( E_X[|\hat{a} - a|] \) is given by

\[ \hat{a} = \max_{1 \leq i \leq n} X_i + \sqrt{\frac{\log 2}{n}}. \]
We estimated the endpoint \( \alpha \) again using the above \( \hat{\alpha} \) and the \( \hat{\alpha} \) that was constructed by estimating \( \alpha \) and \( A \) simultaneously. The results of this case are given in Table 2.

We see that in the case where \( \alpha, A \) are known, those results strongly support our theoretical results. If we estimate the parameters \( \alpha, A \) simultaneously, the average loss becomes larger compared with the one of the case \( \alpha, A \) are known. However, in the case when \( \alpha = A = 1.0 \), the results change little by estimating \( \alpha \) and \( A \) simultaneously.

### 5.2 Estimation of the 2-dimensional data region

A multi-dimensional data region can also be estimated by using the proposed method. For given multi-dimensional data generated in some region, we consider a mapping from multi-dimensional space to \( \mathcal{R} \). The proposed method can be applied to 1-dimensional data obtained by applying this mapping to the given data. Using some mappings, some sets of 1-dimensional data are obtained. Then a multi-dimensional data region is estimated by applying the proposed method to each 1-dimensional data set.

In order to investigate the effectiveness of the proposed method when it is applied to estimate a multi-dimensional data region, we conducted an experiment where a 2-dimensional data region was estimated.

Given a set of \( d \)-dimensional data \( \{x_1, \ldots, x_n\} \), we considered following two procedures, (P1) and (P2), in which we made \( k \) mappings \( g_1(x), \ldots, g_k(x) \) from a datum \( x \in \mathcal{R}^d \).

(P1): Generating vectors \( a_1, \ldots, a_k \in \mathcal{R}^d \) randomly, define \( g_i(x) \) by the inner product of the datum \( x \) and the vector \( a_i \), that is, \( g_i(x) = \langle a_i, x \rangle \).

(P2): Selecting \( k \) data \( x_1, \ldots, x_k \) from \( \{x_1, \ldots, x_n\} \), define \( g_i(x) \) by the distance between \( x \) and \( x_i \) in the kernel-induced feature space (Cristianini and Shawe-Taylor, 2000). That is,

\[
g_i(x) = \|\phi(x) - \phi(x_i)\| = \sqrt{K(x,x) - 2K(x,x_i) + K(x_i,x_i)},
\]

where \( \phi(x) \) is a mapping from the input space to the feature space, \( \| \cdot \| \) is the norm in the feature space and \( K(x,y) = \langle \phi(x), \phi(y) \rangle \) is a kernel function.

We fixed the number of samples \( n = 200 \) and generated 2-dimensional data uniformly in the region \( D^* \) defined by

\[
D^* = \{ (x_1, x_2)^T \in \mathcal{R}^2 \mid x_1^2 + 2x_1^2x_2^2 + x_2^2 - 2\sqrt{2}(x_1x_2 + x_1 + x_2) + 4 \leq 0 \}.
\]  

Fig. 3 displays an example of this artificial data set.

We made 40 sets of 1-dimensional data using each procedure described above respectively. More specifically, in the first procedure (P1), vectors \( a_1, \ldots, a_{40} \) were generated uniformly on the unit circle, and in (P2), we used the polynomial kernel with degree 2, and picked out 40 data \( x_1, \ldots, x_{40} \) that are closer to the average vector \( \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \) in the feature space\(^1\). Examples of the boundaries for the above data set estimated by the procedures P1 (dashed line) and P2 (solid line) are also presented in Fig. 3.

To each 1-dimensional data set \( \{g_i(x_1), \ldots, g_i(x_n)\} \), we applied the proposed method and estimated the right endpoint \( a \). The coefficient \( c_n \) was set to be optimal \( c_n^* \) described in Corollary 1. The left endpoint \( b \) was estimated by applying the proposed method to data set \( \{-g_i(x_1), \ldots, -g_i(x_n)\} \) in the same way. For all data sets, 200 samples were divided into 40 groups and we obtained 40 samples from the distribution of the maximum value eq. (12) to estimate parameters \( \alpha \) and \( A \). Then we calculated the error rate on 10000 test samples. The test samples were generated on the lattice-shaped points in the region including the true data region \( D^* \) in it. By doing so, we evaluated the difference between the estimated region and the true region \( D^* \) by the square measure approximately.

\(^1\) However, the explicit expression of the mapping \( \phi(x) \) is not required since the mapping is implicitly carried out and the norm in the feature space is computed by using only the corresponding kernel function \( K(x,y) = (1 + \langle x, y \rangle)^2 \).
The error rate was calculated by the sum of the two types of the error. One is the rate of the test data that are in the estimated region but out of the true region $D^*$, We denote this type of the error rate by $O_+$. The other is the rate of the test data that are out of the estimated region but in $D^*$. We denote this by $T_-$.

In order to investigate the effectiveness of the proposed method where the coefficients $c_a$ and $c_b$ were set to be optimal, we also calculated the error rate in the case when the coefficients $c_a = c_b = 0$, that is, the maximum and minimum values were used directly to estimate the endpoints of the distribution of each 1-dimensional data set. Table 3 shows the average of the error rate over 5 data sets in the two cases when the procedures (P1) and (P2) were used in extracting 1-dimensional data sets. In parentheses, $T_-$ (left) and $O_+$ (right) are given.

We see that by setting $c_a$ and $c_b$ to be optimal, though $O_+$ increases, the accuracy of estimation was improved with respect to the total error rate.

### 6. Discussion

The proposed method estimates the data region by using the maximum and minimum values in the samples. Several other methods to estimate the endpoints $a$ and $b$ were discussed in some statistical literatures (Hall, 1982; Dekkers et al., 1989; Hall and Wang, 1999). These methods use not only the maximum or the minimum but also some number of extreme values. The maximum and the minimum values are the sufficient statistics in estimating $a$ and $b$ when $f(x)$ is a density function of a certain parametric class such as uniform distributions. Hence, it is suggested that the proposed method is more effective than other methods when the parameters $\alpha_1$ and $\alpha_2$ are small. The more exact comparison of the proposed method to other methods with respect to several loss functions is the subject of further research.

As another approaches, Parzen density estimation or Gaussian mixture models are used for estimating the data region (Markou and Singh, 2003). These methods estimate the underlying distribution by the probability density functions whose supports are the entire space. They determine if the probability that a datum is inside the target region is high or not whereas the methods to estimate the support of the distribution determine if there is any non-zero probability or not.

In section 3, the optimal coefficients given in eqs. (9) and (11) were obtained by calculating the average loss of the estimators with their asymptotic distributions. They can also be derived in another way. Consider the fact on a random variable $X$ that the function of $t$, $E_X[X - |t|]$ and $E_X[(X - t)^2]$ are respectively minimized when $t$ is equal to the median of $X$ and the expectation of $X$. Using the fact and calculating the median and the expectation of the random variable $\max_{1 \leq i \leq n} X_i$ under the condition (i) and (ii) in Section 3, we obtain the optimal coefficients eqs. (9) and (11) as well.

In the previous sections, we evaluated $U(\hat{a} - a)$ as loss functions and derived the optimal estimators for the loss functions. These loss functions are symmetric with respect to $\hat{a}$ being greater(overestimation) or smaller(underestimation) than $a$. In some situations, it is more suitable to use loss functions that are not symmetric with respect to that. Let us consider the case that we use such asymmetric loss functions. Let $l_1$ and $l_2$ be positive constants and function $L(x)$ be

![Fig. 3. An example of the 2-dimensional data set and the boundaries estimated by the procedures P1 (dashed line) and P2 (solid line).](image-url)
\[ L(x) = \begin{cases} -l_1 x & (x < 0), \\ l_2 x & (x \geq 0). \end{cases} \]

Redefining the average loss by \( E_{X^n}[L(\hat{a} - a)] \), we also obtain the following theorem in the same way.

**Theorem 5.** Suppose \( X_1, \ldots, X_n \) are independently and identically taken from density \( f(x) \), and if the density function \( f(x) \) satisfies the conditions (i), (ii) in Section 3, then

\[ E_{X^n}[L(\hat{a} - a)] = \frac{l_2}{n^2} \left[ c_a - \left( \frac{\alpha}{A} \right)^{\frac{1}{\alpha}} \left( 1 + \frac{l_1}{l_2} \right) \gamma \left( \frac{1}{\alpha}, \frac{A}{\alpha} c_a^\alpha \right) - \frac{l_1}{l_2} \Gamma \left( \frac{1}{\alpha} \right) \right] + o \left( \frac{1}{n^2} \right), \tag{16} \]

and the optimal coefficient \( c_a^* \) that minimizes this average loss is given by

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

### 7. Conclusion

In this paper, we proposed a method to estimate the data region using the maximum and minimum values in the samples. We calculated the asymptotic distributions of these values and derived the optimal estimators for given loss functions. By extracting 1-dimensional data, the proposed method has abundant practical applications. It can also be applied effectively to estimate a multi-dimensional data region by using some mappings to \( R^1 \). We demonstrated this by the experimental study on 2-dimensional data in the latter half of Section 5. The shape of the estimated region and the accuracy of the estimation depend on the selection and the number of mappings to extract 1-dimensional data sets from the multi-dimensional data. To maximize the effectiveness of the proposed method, questions like the selection of those mappings and the optimization of the number of them have to be addressed.

### Acknowledgments

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research 15500130 and for JSPS Fellows 0404637.

### Appendix

#### Proof of Theorem 1.

From assumptions (i) and (ii), in a neighborhood of \( a \) we have

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

and

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

Putting

\[ u_n = \left( \frac{A}{\alpha n} \right)^{-1/\alpha} x + a, \]

for \( x \leq 0, \)

\[ \Pr \left( \left( \frac{A}{\alpha n} \right)^{1/\alpha} (M_n - a) \leq x \right) = F(u_n)^n \]

\[ = \left[ 1 - (1 - F(u_n))^n \right]^n \]

\[ = \left\{ 1 - \left( -\frac{x}{n} \right)^\alpha + o \left( \frac{1}{n} \right) \right\}^n \]

\[ \to \exp \left( -\left( -\frac{x}{n} \right)^\alpha \right) \quad (n \to \infty). \]

Since \( 1 - F(u_n) = 0 \) for \( x > 0 \), we obtain \( G_{\max}(x) \). By a similar argument, we also obtain \( G_{\min}(x) \). (Q.E.D.)

#### Proof of Theorem 2.

Putting
\[ u_n = \left( \frac{A}{n} \right)^{-\alpha} s + a, \quad v_n = \left( \frac{B}{n} \right)^{-\beta} t + b. \]

we have

\[
\Pr\left\{ \left( \frac{A}{n} \right)^{1/\alpha} (M_n - a) \leq s, \left( \frac{B}{n} \right)^{1/\beta} (m_n - b) \leq t \right\} = F(u_n)^n - (F(u_n) - F(v_n))^n. \tag{17}
\]

It follows from Theorem 1 that

\[ F(u_n)^n \to \exp\left(-(-s)^\alpha\right) \quad (n \to \infty), \]

and that

\[
[F(u_n) - F(v_n)]^n = \left[ 1 - (1 - F(u_n)) - F(v_n) \right]^n
\]

\[ = \left\{ 1 - \frac{(-s)^\alpha}{n} - \frac{t^\beta}{n} + o\left( \frac{1}{n} \right) \right\}^n
\]

\[ \to \exp\left(-(-s)^\alpha - t^\beta\right) \quad (n \to \infty). \]

Then we have

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

(Q.E.D.)

**Proof of Theorem 3.** We define

\[ \tau = a - n^{-\frac{1}{\alpha}}(t^1 + c_a), \]

\[ \tilde{\tau} = a - n^{-\frac{1}{\alpha}}(-t^1 + c_a), \]

then \( n^{k/\alpha}(\tilde{a} - a)^k \leq \tau \) is equivalent to

\[ \max_{1 \leq i \leq n} X_i \leq \tilde{\tau} \]

for any odd number \( k > 0 \), and \( n^{k/\alpha}|\tilde{a} - a|^k \leq \tau \) is equivalent to

\[ \tau \leq \max_{1 \leq i \leq n} X_i \leq \tilde{\tau} \]

for any natural number \( k \).

Hence

\[ H_n(\tau) = \begin{cases} 1 & (\tau \geq c_a^k), \\ F(\tau)^n & (\tau < c_a^k), \end{cases} \]

and

\[ \overline{H}_n(\tilde{\tau}) = \begin{cases} 1 - F(\tilde{\tau})^n & (\tau \geq c_a^k), \\ F(\tilde{\tau})^n - F(\tau)^n & (\tau < c_a^k). \end{cases} \]

Since, in a neighborhood of \( a \), we have

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

hence

\[ F(\tau)^n = \left\{ 1 - \frac{A}{\alpha n} (t^{1/k} + c_a)^\alpha + o\left( \frac{1}{n} \right) \right\}^n \]

\[ \to \exp\left(-\frac{A}{\alpha}(c_a + t^{1/k})^\alpha\right) \quad (n \to \infty), \]

and

\[ F(\tilde{\tau})^n = \left\{ 1 - \frac{A}{\alpha n} (-t^{1/k} + c_a)^\alpha + o\left( \frac{1}{n} \right) \right\}^n \]

\[ \to \exp\left(-\frac{A}{\alpha}(c_a - t^{1/k})^\alpha\right) \quad (n \to \infty). \]

Thus we have eq. (4) and eq. (5). (Q.E.D.)
Proof of Theorem 4. If \( k \) is an odd number,

\[
E_X[n^k/\alpha(\hat{\alpha} - \alpha)^k] \rightarrow c_n^k = \int_{-\infty}^{\infty} \exp\left(-\frac{A}{\alpha}(c_a - t^{1/k})^\alpha\right) dt \quad (n \to \infty)
\]

\[
= c_n^k - \frac{k}{A} \int_{0}^{\infty} \left( c_a - \left(\frac{\alpha}{A} t\right)^{\frac{1}{\alpha}} \right)^{k-1} \left(\frac{\alpha}{A} t\right)^{\frac{1}{\alpha} - 1} e^{-t} dt.
\]

For natural number \( k \)

\[
E_X[n^k/\alpha(\hat{\alpha} - \alpha)^k] \rightarrow c_n^k = \int_{-\infty}^{\infty} \exp\left(-\frac{A}{\alpha}(c_a - t^{1/k})^\alpha\right) dt
\]

\[
+ \int_{0}^{\infty} \exp\left(-\frac{A}{\alpha}(c_a + t^{1/k})^\alpha\right) dt \quad (n \to \infty)
\]

\[
= c_n^k - \frac{k}{A} \int_{0}^{\infty} \left( c_a - \left(\frac{\alpha}{A} t\right)^{\frac{1}{\alpha}} \right)^{k-1} \left(\frac{\alpha}{A} t\right)^{\frac{1}{\alpha} - 1} e^{-t} dt
\]

\[
+ \frac{k}{A} \int_{0}^{\infty} \left(\frac{\alpha}{A} t\right)^{\frac{1}{\alpha}} - c_a \left(\frac{\alpha}{A} t\right)^{\frac{1}{\alpha} - 1} e^{-t} dt.
\]

(18)

(19)

For \( k \) even, this reduces to the right hand side of eq. (18). Hence, expanding \((c_a - (\frac{\alpha}{A} t)^{\frac{1}{\alpha}})^{k-1}\) in eq. (18) and eq. (19) respectively, we obtain eq. (6) for an arbitrary natural number \( k \) and eq. (7) for an odd number \( k \). (Q.E.D.)

REFERENCES


