

固有値法による射影追跡密度推定

Projection Pursuit Density Estimation Using Eigen Value Method

市橋 秀友
長坂 一徳
三好 哲也
野崎 大輔

大阪府立大学工学部経営工学科

〒593 大阪府堺市学園町1-1

北口 嘉亮

大阪府立産業技術総合研究所

〒590-02 大阪府和泉市あゆみ野2丁目7-1

喜田 義宏

大阪工業大学

〒550 大阪府大阪市旭区大宮5-16-1

H. ICHIHASHI
K. NAGASAKA
T. MIYOSHI
D. NOZAKI

College of Engineering, University of Osaka Prefecture
1-1 Gakuen-cho, Sakai, Osaka 593, Japan

Y. KITAGUCHI

Osaka Prefectural Industrial Technology Research Institute
7-1, 2-chome, Ayumino, Izumi, Osaka 590-02, Japan

Y. KITA

Osaka Institute of Technology
1-53 Enokoshima, Nishi-ku, Osaka 550, Japan

要 約

核関数法 (kernel method) は, ある母集団の確率密度関数をデータから直接推定するノンパラメトリックな方法である. しかし, 多次元空間上に直接推定された密度関数から分布の特徴的な形状を見つけるのは困難である. J.H.Friedman らにより始められた射影追跡回帰の目的は, 高次元空間におけるデータ点を低次元空間に射影することによってデータの下にある特徴的な非線形写像を発見しようとするもので, 線形射影と一変数の近似関数が用いられている.

本研究では核関数法により推定された多次元分布をもとに, 固有値法による射影追跡回帰を用いてデータを 1 次元の直線上に射影する. そして射影された 1 次元データに対して再び核関数法を適用することによってもとの多次元分布の特徴的な形状をできるだけ残した一次元分布を推定する.

キーワード

ファジィクラス, 射影追跡, 固有値法, 核関数法, 確率密度推定

Abstract

This paper focuses on the use of kernel method and projection pursuit regression for non-parametric probability density estimation. Direct application of the kernel method is not able to pick up characteristic features of multidimensional density function. The proposed projection pursuit regression method based on the eigen value problem is able to bypass the “curse of dimensionality” in multidimensional density estimation.

Keywords

Fuzzy Class, Projection Pursuit, Eigen Value Method, Kernel Method, Density Estimation

1. はじめに

核関数法 (kernel method)[1, 2, 3] はポテンシャル関数法とも呼ばれ, ある母集団の確率密度関数をデータから直接推定する方法であり統計的な判別分析に多く用いられている. J.W.Van Ness and C.Simpson[1] や J.W.Van Ness[2] は単峰性, 対称性の核関数であればその形状が異なっても判別分析の結果にあまり影響しないことを報告している. しかし, 核関数法はノンパラメトリックな方法であり, 多次元空間上に直接推定された密度関数から分布の特徴的な形状を見つけるのは困難である.

J.H.Friedman and J.W.Tukey[4] により始められた射影追跡の目的は, 高次元空間におけるデータ点を低次元空間に射影することによってデータの興味深い構造を発見することである[6, 7]. また射影追跡回帰[5] は線形射影を用いてデータの下にある非線形写像を見つけようとするものである. そのために, ある目的関数すなわち射影指標を最大 (または最小) 化する線形射影をコンピュータを用いて自動探索する.

本研究では低次元空間での非線形写像を近似しようとする射影追跡回帰の考え方に基づくノンパラメトリックな射影追跡密度推定法を提案している. まず, 核関数法により推定された多次元分布をもとに, 固有値法による射影追跡回帰[8]を用いてデータを1次元の直線上に射影する. そして射影された1次元データに対して再び核関数法を適用することによってもとの多次元分布の特徴的な形状をできるだけ残した一次元分布を推定する. 多次元空間にあるデータを低次元空間に射影するという考え方は, 多変量解析で一般に用いられているもので, 主成分分析や判別分析などは射影追跡の特別な場合であるとみなされている[6, 7]. 提案手法は固有値法により, 尖度の大きな単峰形関数の特徴的な形状として発見しようとするものである. したがって回帰や密度推定を目的とした通常的手法とは射影指標[6]が異なっているが射影追跡手法の一つである. 応用例として, 小径エンドミルの加工精度の実験計画における水準の設定に

ついて述べる.

2. 射影追跡確率密度推定

2.1 核関数法による確率密度推定の概説

J 個の d 次元データの第 j 番目のデータをベクトル \mathbf{a}_j とし, その第 i 成分を中心 a_{ij} として, 平滑化パラメータ b (smoothing parameter) を持ったカーネル (核関数, パーゼンの窓関数) を考える. 本研究ではカーネルに次のガウス関数を用いる. ここで, x_i を d 次元ベクトル \mathbf{x} の第 i 成分とすると,

$$G_{ij}(x_i) = \exp\left(-\frac{(x_i - a_{ij})^2}{b}\right) \quad (1)$$

$$K_j(\mathbf{x}) = \prod_{i=1}^d G_{ij}(x_i) \quad (2)$$

$d = 1$ のとき

$$K_j(x_1) = \exp\left(-\frac{(x_1 - a_{1j})^2}{b}\right) \quad (3)$$

$d = 2$ のとき

$$\begin{aligned} K_j(x_1, x_2) \\ = \exp\left(-\frac{(x_1 - a_{1j})^2 + (x_2 - a_{2j})^2}{b}\right) \end{aligned} \quad (4)$$

となる. 1つのカーネルの積分は, 以下 $\int_{-\infty}^{\infty}$ を表すとする, と,

$$\begin{aligned} I_d &= \int \cdots \int K_j(\mathbf{x}) d\mathbf{x} \\ &= (\pi b)^{d/2} \end{aligned} \quad (5)$$

となる. たとえば, $d = 1$ のとき

$$\begin{aligned} I_1 &= \int K_j(x_1) dx_1 \\ &= \sqrt{\pi b} \end{aligned} \quad (6)$$

であり, $d = 2$ のとき

$$\begin{aligned} I_2 &= \int \int K_j(x_1, x_2) dx_1 dx_2 \\ &= \pi b \end{aligned} \quad (7)$$

である．推定される確率密度関数 $\hat{f}(\mathbf{x})$ はデータ件数を J として

$$\hat{f}(\mathbf{x}) = \frac{1}{JI_d} \sum_{j=1}^J K_j(\mathbf{x}) \quad (8)$$

となる．この推定量は J を大きくする早さに対して， b を十分ゆっくりゼロに近づけるととき，確率密度関数の一致推定量となることが知られている．たとえば， $d = 1$ のとき

$$\hat{f}(x_1) = \frac{1}{J\sqrt{\pi b}} \sum_{j=1}^J \exp\left(-\frac{(x_1 - a_{1j})^2}{b}\right) \quad (9)$$

であり， $d = 2$ のとき

$$\hat{f}(x_1, x_2) = \frac{1}{J\pi b} \sum_{j=1}^J \exp\left(-\frac{(x_1 - a_{1j})^2 + (x_2 - a_{2j})^2}{b}\right) \quad (10)$$

である．核関数法を適用する際に問題となるのは，平滑化パラメータ b としてどのような値を選択すればよいかということである．このような密度推定の問題では真の分布を $f(\mathbf{x})$ ，推定した分布（モデルが定める密度関数）を $\hat{f}(\mathbf{x})$ とし，K-L情報量

$$\begin{aligned} E\left[\log \frac{f(\mathbf{x})}{\hat{f}(\mathbf{x})}\right] &= \int \cdots \int \left(\log \frac{f(\mathbf{x})}{\hat{f}(\mathbf{x})}\right) f(\mathbf{x}) d\mathbf{x} \\ &= \int \cdots \int (\log f(\mathbf{x})) f(\mathbf{x}) d\mathbf{x} \\ &\quad - \int \cdots \int (\log \hat{f}(\mathbf{x})) f(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (11)$$

によって， $\hat{f}(\mathbf{x})$ を評価することができる．(11)式の右辺の第1項目は定数であるので，第2項目の平均対数尤度

$$\int \cdots \int_{-\infty}^{\infty} (\log \hat{f}(\mathbf{x})) f(\mathbf{x}) d\mathbf{x} \quad (12)$$

で評価すればよい．

J 個の独立な観測値 $\{\mathbf{a}_1, \dots, \mathbf{a}_j, \dots, \mathbf{a}_J\}$ が得られているとき，平均対数尤度は対数尤度の J 分の1

$$\frac{1}{J} \sum_{j=1}^J \log \hat{f}(\mathbf{a}_j) \quad (13)$$

で近似される．符号に注意すると，対数尤度が大きいほどそのモデルは真の分布に近いと考えられる．

以上のことをふまえ次に平滑化パラメータ b を決定する．観測値 \mathbf{a}_j は与えられたものとして固定し， \hat{f} を b の関数と考えるとき，次の関数を尤度(likelihood)と呼び

$$L(b|X) = \prod_{j=1}^J \hat{f}(\mathbf{a}_j, b|X) \quad (14)$$

で表す．ここで \mathbf{a}_j は第 j 番目のデータ点， X はデータの集合 $\{\mathbf{a}_j | j = 1, \dots, J\}$ を表すものとする．

(14)式の両辺の対数をとったものである(13)式の対数尤度を最大にする値を選択すると $b = 0$ のときに最大となり，核関数は各々のデータ点におけるデルタ関数になってしまう．そこで各種の交差確認法(cross-validation method)が用いられている．その中でJ.D.F.Habbema and J.Hermans[9]は次のようなジャックナイフ法を提案している．

文献[9]のジャックナイフ法では(14)式のかわりに

$$L^*(b|X) = \prod_{j=1}^J \hat{f}(\mathbf{a}_j, b|X - \mathbf{a}_j) \quad (15)$$

のように， \mathbf{a}_j の点での尤度の計算に \mathbf{a}_j を除いた残りのサンプル $X - \mathbf{a}_j$ だけを利用する擬尤度関数(pseudo-likelihood function)を採用している．いま，第 j 番目のデータ点を $\mathbf{a}_j = (a_{1j}, a_{2j}, \dots, a_{dj})$ とすると，擬対数尤度(pseudo-logarithm-likelihood, PLL)は

$$\begin{aligned} f_k^*(\mathbf{a}_k) &= \frac{1}{(J-1)I_d} \sum_{j=1, j \neq k}^J \left(\prod_{i=1}^d G_{ij}(a_{ij}) \right) \end{aligned} \quad (16)$$

$$PLL = \frac{1}{J} \sum_{k=1}^J \log f_k^*(\mathbf{a}_k) \quad (17)$$

と表され(17)式のPLLを最大にする b の値を選択する．

2.2 固有値法による多次元確率密度関数の射影

本節では固有値法による射影追跡回帰[8]の密度推定への応用について述べる． $\mathbf{x}_j (j = 1, \dots, J)$ の集合をファジイ集合とし， $m(f_j)$ をメンバシップ関数とする．ファジイ集合（ファジイクラス）を C と D の 2 つとし，それぞれ $\{(\mathbf{x}_j, m(f_j)) ; j = 1, \dots, J\}$ と $\{(\mathbf{x}_j, 1 - m(f_j)) ; j = 1, \dots, J\}$ とする．ただし， $m(f_j)$ は確率密度関数の推定値 $\hat{f}(\mathbf{a}_j) (j = 1, \dots, J)$ の値を区間 $[0, 1]$ に規準化した \mathbf{a}_j における値である．クラス C は確率密度の大きいデータ点のファジイ集合であり，D は大きくないデータ点のファジイ集合である．従ってデータの分布が単峰形の密度関数に従う場合は C は中央で D はその両側である．そして C の分散が小さく D の分散が大きくなるような射影ベクトルを求める．このことは尖度の大きい単峰形の一次元密度関数を求めることになる． \mathbf{y} を単位ベクトル $\mathbf{p} = (p_1, \dots, p_n)^T$ により定まる直線への \mathbf{x} の正射影とすると

$$\mathbf{y} = \mathbf{p}^T \mathbf{x} \quad (18)$$

となる．ただし， T は転置を意味する．ファジイクラス C 内の \mathbf{y} の標本分散は，

$$\begin{aligned} \sigma_C^2 &= \frac{1}{J_C} \sum_{j=1}^J m(f_j) (y_j - \bar{y})^2 \\ &= \mathbf{p}^T \sum_C \mathbf{p} \end{aligned} \quad (19)$$

となる．ただし，

$$J_C = \sum_{j=1}^J m(f_j) \quad (20)$$

$$\bar{y} = \frac{1}{J_C} \sum_{j=1}^J m(f_j) \cdot y_j \quad (21)$$

$$\sum_C = \frac{1}{J_C} \sum_{j=1}^J m(f_j) (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^T \quad (22)$$

$$\bar{\mathbf{x}} = \frac{1}{J_C} \sum_{j=1}^J m(f_j) \cdot \mathbf{x}_j \quad (23)$$

である．一方，ファジイクラス D の \bar{y} （クラス C 内の平均）からの変動を

$$\begin{aligned} \sigma_D^2 &= \frac{1}{J - J_C} \sum_{j=1}^J (1 - m(f_j)) (y_j - \bar{y})^2 \\ &= \mathbf{p}^T \sum_D \mathbf{p} \end{aligned} \quad (24)$$

とする．ただし，

$$\begin{aligned} \sum_D &= \frac{1}{J - J_C} \sum_{j=1}^J (1 - m(f_j)) \cdot \\ &\quad (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^T \end{aligned} \quad (25)$$

である．クラス C 内の分散 (σ_C^2) を小さくし，一方でクラス D の y_j ができるだけ \bar{y} （クラス C 内の平均）から離れるような \mathbf{p} を求める．ただし，射影後の密度関数は大まかには単峰形であると仮定する．射影によってデータの構造を良く表現できる射影ベクトル \mathbf{p} を求めるための評価関数すなわち射影指標を

$$Q(\mathbf{p}) = \frac{\mathbf{p}^T \sum_D \mathbf{p}}{\mathbf{p}^T \sum_C \mathbf{p}} \quad (26)$$

とする．射影ベクトル \mathbf{p} は，この $Q(\mathbf{p})$ を最大化することで求められ，

$$\sum_D \mathbf{p} = \lambda \sum_C \mathbf{p} \quad (27)$$

なる固有値問題に帰着される．

2.3 1次元確率密度関数の推定手順

2.1 節と 2.2 節の方法を用いて 2 次元データ点 (a_1, a_2) から特徴的な 1 次元の分布形を推定する手順を以下に示す．

(step 1)

2 次元データ点を中心として 2 次元のカーネルを置き $\hat{f}(x_1, x_2)$ を (10) 式の核関数法により求める．データ点上に置く 2 次元のカーネルの平滑化パラメータ b の値を (17) 式のジャックナイフ法を用いて決定する． j 番目の点 (a_{1j}, a_{2j}) における確率密度の推定値は $\hat{f}(a_{1j}, a_{2j})$ である．

(step 2)

$f_j = \hat{f}(a_{1j}, a_{2j})$ として固有値法(2.2節)を用いて、射影によってデータの構造を良く表現できるような射影ベクトル \mathbf{p} を求める。

(step 3)

2次元のデータ点を(18)式により1次元の直線に射影する。射影されたデータ点上に置く1次元のカーネルの平滑化パラメータ b の値をジャックナイフ法を用いて決定し、1次元の密度関数を求める。

3. 確率密度推定の数値例

2次元のテストデータ (a_1, a_2) 150個を以下の手順で作成した。

(step 1)

5個ずつの正規乱数 $X_i \sim N(\mu, \sigma^2)$ を(28)式に代入して a_1^* を作成した。ただし、 $\mu = 0$, $\sigma^2 = 1$ である。 a_1^* は自由度 $\phi = 5$ の χ^2 分布に従う。

$$a_1^* = \frac{1}{\sigma^2} \sum_{i=1}^5 (X_i - \mu)^2 \cdot \frac{1}{20} \quad (28)$$

次に a_2^* を一様乱数 ($\in [0, 1]$) により発生させた。

(step 2)

(step 1) で発生させた (a_1^*, a_2^*) を(29)式, (30)式で座標変換し2次元のテストデータ 150個を得た。

$$a_1 = (a_1^*, a_2^*) \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \quad (29)$$

$$a_2 = (a_1^*, a_2^*) \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} \quad (30)$$

2.3節で述べた手順ごとの結果を以下に示す。表1-(a) はstep 1でのジャックナイフ法の結果で $b = 0.010$ が選ばれている。図1は核関数法により求めた2次元の分布形であり、確率密度

表 1: テストデータに対するジャックナイフ法の結果

(a) 2次元		(b) 1次元	
b の値	PLL の値	b の値	PLL の値
0.004	0.3209	0.001	0.4317
0.005	0.3629	0.002	0.4950
0.006	0.3865	0.003	0.5140
0.007	0.4001	0.004	0.5214
0.008	0.4078	0.005	0.5240
0.009	0.4119	0.006	0.5243
0.010	0.4135	0.007	0.5234
0.011	0.4134	0.008	0.5216
0.012	0.4121	0.009	0.5192

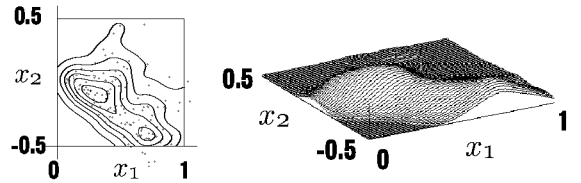


図 1: 核関数法により推定した密度関数の形状 (2次元)

の等高線図と3次元グラフィックスを示している。射影ベクトル $\mathbf{p} = (p_1, p_2)$ は次のようになった。 $p_1 = 0.603$, $p_2 = 0.797$ 。1次元の直線に射影したデータ点に対するジャックナイフ法の結果を表1-(b)に示す。 $b = 0.006$ が選ばれている。核関数法による1次元の分布形を図2に示す。図2に示すように自由度 $\phi = 5$ の χ^2 分布に近い形状が得られていて(28)式の a_1^* の分布の形状をほぼ復元している。

4. 実験計画における水準設定への応用

本章では小径エンドミルの加工精度の実験における実験計画の水準設定[10]に提案の射影追跡確率密度推定法を応用した。小径エンドミル

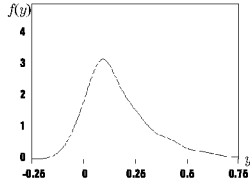


図 2: 核関数法により推定した密度関数の形状
(1 次元)

先端における振れと先端から 3mm における振れを測定した．GC 主軸を手で回転させマイクロメーターの針が最大値を示すところ（刃先）における測定値を A ，そこから 180 度回転させたところ（反対側の刃先）における測定値を B とすると，振れ量は A と B の差の絶対値をとったものとなる．単位は μm である．小径エンドミルの加工において，エンドミル工具の振れ量は加工精度に大きく影響すると考えられ，振れ量の確率分布を推定することは，加工精度に関する知見を得る上で重要である．測定したデータを単位区間 $[0,1]$ に規準化し 3 章と同様の手順で推定を行う．ただし，データはこの場合正であるので 2.1 節の I_d について，以下のような補正を行う必要がある． $d = 1$ のとき

$$I_1 = \sum_{j=1}^J (1 - q_{1j}) \sqrt{\pi b} \quad (31)$$

ただし， q_{1j} は第 j 番目のカーネルの負の部分の面積比率であり，J.D.Williams[11] による正規分布関数の近似式（平均 0，分散 1）を用いてつぎのように表す．

$$q_{1j} = \frac{1}{2} - \sqrt{1 - \exp\left(-\frac{2\mu_{1j}^2}{\pi}\right)} \quad (32)$$

ただし，

$$\mu_{1j} = -\frac{a_{1j}}{\sqrt{b/2}} \quad (33)$$

である． $d = 2$ のとき

$$I_2 = \sum_{j=1}^J ((1 - q_{1j}) + (1 - q_{2j}) + q_{1j}q_{2j}) \pi b \quad (34)$$

表 2: 実験データに対するジャックナイフ法の結果

(a) 2 次元		(b) 1 次元	
b の値	PLL の値	b の値	PLL の値
0.011	-4.8964	0.011	-4.53379
0.012	-4.8918	0.012	-4.53351
0.013	-4.8889	0.013	-4.53339
0.014	-4.8872	0.014	-4.53338
0.015	-4.8865	0.015	-4.53348
0.016	-4.8865	0.016	-4.53366
0.017	-4.8871	0.017	-4.53391
0.018	-4.8882	0.018	-4.53423

ただし， q_{1j} は第 j 番目のカーネルを x_1 軸に平行に切断したときの x_1 方向の負の部分の面積比率，また q_{2j} は第 j 番目のカーネルを x_2 軸に平行に切断したときの x_2 方向の負の部分の面積比率であり，近似式によりつぎのように表される．

$$q_{1j} = \frac{1}{2} - \sqrt{1 - \exp\left(-\frac{2\mu_{1j}^2}{\pi}\right)} \quad (35)$$

$$q_{2j} = \frac{1}{2} - \sqrt{1 - \exp\left(-\frac{2\mu_{2j}^2}{\pi}\right)} \quad (36)$$

ただし，

$$\mu_{1j} = -\frac{a_{1j}}{\sqrt{b/2}} \quad (37)$$

$$\mu_{2j} = -\frac{a_{2j}}{\sqrt{b/2}} \quad (38)$$

である．2.3 節で述べた手順ごとの結果を以下に示す．表 2-(a) は step 1 でのジャックナイフ法の結果で $b = 0.015$ が選ばれている．図 3 は核関数法により求めた 2 次元の分布形であり，確率密度の等高線図と 3 次元グラフィクスを示している．

射影ベクトル $\mathbf{p} = (p_1, p_2)$ は次のように求めた． $p_1 = 0.872$ ， $p_2 = 0.488$ ．1 次元の直線に射影したデータ点に対するジャックナイフ法の結果を表 2-(b) に示す． $b = 0.014$ が選ばれて

図 3: 核関数法により推定した密度関数の形状
(2 次元)

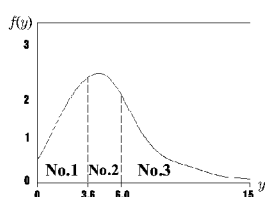


図 4: 核関数法により推定した密度関数の形状
(1 次元)

いる。核関数法による 1 次元の分布形を図 4 に示す。図 3 の 2 次元分布形をベクトル \mathbf{p} で定める直線に射影することにより特徴的な分布形状を求めることができた。図 4 には、面積を 3 等分するように 3 つの水準を設定した例も示している。

5. おわりに

本研究では、核関数法と射影追跡回帰の考え方を採用することにより多次元データの構造を特徴的に示す 1 次元確率分布の推定法を提案した。また小径エンドミルの振れ量の実測を行って実験計画における水準の設定に応用した。カーネルの平滑化パラメータを決定する際ジャックナイフ法を用いたが、他の種々の CV 法との比較、求まる射影ベクトルへの影響等を調べることで、またより実用的な応用を考えることが今後の課題である。

参考文献

- [1] J.W. Van Ness and C. Simpson : On the effects of dimension in discriminant analysis; *Technometrics*, Vol.18, pp.175-187(1976)
- [2] J.W. Van Ness : On the effects of dimension in discriminant analysis for unequal covariance populations; *Technometrics*, Vol.21, pp.119-127(1979)
- [3] P. Smith: Probability density estimation and local basis function networks; *Computational learning theory and natural learning systems*, Vol.2 (S.J. Hanson, T. Petsche, M. Kearns and R.L. Rivest, editors), The MIT Press, pp.233-248(1994)
- [4] J.H. Friedman and J.W. Tukey: A projection pursuit algorithm for exploratory data analysis; *IEEE Transactions on Computers*, Vol.C-23, pp.881-890(1974)
- [5] J.H. Friedman and W. Stuetzle: Projection pursuit regression; *Journal of the American Statistical Association*, Vol.76, No.376, pp.817-823(1981)
- [6] P.J. Huber: Projection pursuit; *The Annals of Statistics*, Vol.13, No.2, pp.435-475(1985)
- [7] 岩崎 : 射影追跡 その考え方と実際 ; 計算機統計学, Vol.4, No.2, pp.41-56(1991)
- [8] 三好, 中尾, 市橋, 長坂 : ニューロ・ファジィ射影追跡による回帰と判別; 第 11 回ファジィシステムシンポジウム講演論文集, pp.811-814(1995)
- [9] J.D.F.Habbema and J.Hermans: A stepwise discriminant analysis program using density estimation; *COMPSTAT*, Physica-Verlag, Vienna, pp.101-110(1974)

- [10] 矢野：加工品質工学，工業調査会(1994)
- [11] J.D.Williams: An approximation to the probability integral; Annals of Mathematical Statistics, Vol.17, pp.363-365(1946)

[問い合わせ先]

〒593

大阪府堺市学園町1-1

大阪府立大学工学部経営工学科

市橋 秀友

TEL : 0722-52-1161

FAX : 0722-59-3340

E-mail: ichi@ie.osakafu-u.ac.jp