論　文

パターン認識における予測分布の偏りに関する一考察

岩村　雅一[†]　　大町真一郎[†]　　阿曽　弘具[†]

On the Bias of Predictive Distribution in Pattern Recognition

Masakazu IWAMURA[†], Shinichiro OMACHI[†], and Hirotomo ASO[†]

あらまし　学習用サンプルから分布を推定する際，未知パラメータを定数として推定する最尤推定に対し，未知パラメータを確率変数とするベイズ推定がある．ベイズ推定を用いたパターン認識は，クラス間でサンプル数が等しい場合は有効であるが，等しくない場合には尤度の偏りのために認識性能が向上しないことが指摘されている．クラス間でサンプル数が異なる場合のパターン認識では尤度の偏りに差が生じ，この偏りの差が認識性能を悪化させていると考えられる．本論文では，尤度の偏りを補正することで認識性能が改善できると考え，予測分布の偏りを理論的に導き，厳密解を与える．予測分布の偏りの理論式が得られたことにより，尤度の補正が可能となる．実際に尤度の補正法を提案し，これにより認識性能が改善することを検証する．さらに本論文で導く理論式から，これまで経験的に得られた知見の理論的根拠を明らかにする．

キーワード　パターン認識，ベイズ推定，予測分布，確率の偏り，Geisser の予測分布

1. まえがき

学習用サンプルから分布を推定する際，未知パラメータを定数として推定する最尤推定に対し，未知パラメータを確率変数と考えて，その分布を推定し，パラメータの分布をもとに次のサンプルの予測分布を求めるベイズ推定がある [1], [2]．ベイズ推定は最尤推定よりも優れた推定が行えるとされる [2] ~ [4]．

ベイズ推定では，サンプルが与えられる前に持っている真のパラメータに関する知識を「事前分布」という形で表現し，サンプルから得られた情報と統合した「事後分布」を求め，次のサンプルの分布を「予測分布」として推定する．事前分布は一般に設計者の主観で決めるパラメータであるが，客観的な立場から，事前知識が無いことを表す無情報事前分布を用いる場合がある．

共分散行列が未知のときに予測分布を求める方法としては，真の共分散行列が特定の行列をパラメータとする逆 Wishart 分布に従うとする Keehn の方法 [5] や，パラメータが一様に分布すると仮定した無情報事前分布を用いる Geisser の方法 [6] などがある．ベイズ推定をパターン認識に適用したとき，各クラスの学習サンプル数が等しい場合は認識性能が向上することが知られているが，文献 [7] において学習サンプル数がクラス間で異なる場合は Keehn の方法，Geisser の方法ともに有効でないことが指摘されている．そのため，韓ら [8] は，Keehn の方法でパラメータとして用いられる信頼度定数の認識に有効な定め方を実験的に示し，パターン認識にベイズ推定を用いることの有効性を示した．しかし，この方法は Keehn の方法を対象としたもので一般的なものではない．また，認識の際の有効性に主眼が置かれているため，認識性能が低下する原因や予測分布の偏り方については考察されていない．

サンプル数が大きい場合にベイズ推定を行うと，未知入力ベクトルの予測分布が漸近的に多次元正規分布に一致することが知られている．一方，サンプル数が十分大きくない場合については，予測分布の期待値が多次元正規分布と比べて小さく偏ると考えられる．この偏りはサンプル数によって異なり，一様でない．そのため，各クラスのサンプル数が等しい場合には，偏りに（あまり）差が生じず，認識性能に影響を及ぼさないが，サンプル数が異なる場合には偏りに差が生じて，認識性能が低下するものと考えられる．

学習サンプル数がクラス間で異なる場合に認識性

[†] 東北大学大学院工学研究科，仙台市
Graduate School of Engineering, Tohoku University, Sendai-shi, 980–8579 Japan

能が低下する原因が，サンプル数を n としたとき，$n \to \infty$ の近似が成り立たないときに生じる予測分布の偏りの差にあるならば，各クラスの偏りを補正することで認識性能が改善するはずである．

本論文では，Geisser の方法について，予測分布の偏りの理論式（厳密解）を導出し，サンプル数が小さい場合の予測分布の偏りの程度を定量的に評価し，誤認識が起こりやすい状況についての根拠を明らかにするとともに，解決法を与えることを目的とする．具体的にはまず，サンプル数が少ない場合に尤度の期待値が偏ることをサンプルから求めた平均値を用いて確認する．次に，予測分布の偏りの理論式を導出し，サンプル数がクラス間で異なる場合には，導いた理論式で尤度を補正すれば認識精度が向上することを実験によって示す．これは，ベイズ推定を用いたパターン認識では予測分布の偏りが認識性能を悪化させていることを検証するものである．さらに，本論文で導く理論式を用いて，文献 [8] で指摘されている「分散の小さなクラスの学習サンプルが少ないほど認識率が低下する」という経験的に得られた知見に理論的根拠を与える．

なお，本論文で示す予測分布の偏りの理論式の導出法は，別の事前分布を仮定した場合（例えば，Keehn の方法）にも容易に適用できる高い一般性を持つ．

2. 数学的準備

本論文で用いる数学的な事柄について最初に紹介する．本節の内容については文献 [9] ~ [11] が詳しい．

2.1 Wishart 分布

正規分布 $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ に従う各々独立な n 個の d 次元確率ベクトル $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ の平方和積和行列 $\boldsymbol{W}$ を

$$\boldsymbol{W} = \sum_{t=1}^{n} \boldsymbol{X}_t \boldsymbol{X}_t^{\mathrm{T}} \tag{1}$$

とおくと，$\boldsymbol{\Sigma}$ が与えられたときの $\boldsymbol{W}$ の分布 $\mathrm{W}_d(n, \boldsymbol{\Sigma})$ は Wishart 分布として知られ，その確率密度は次式で与えられる．

$$P(\boldsymbol{W}) = \frac{|\boldsymbol{\Sigma}|^{-\frac{n}{2}} |\boldsymbol{W}|^{\frac{1}{2}(n-d-1)}}{2^{\frac{1}{2}nd} \boldsymbol{\Gamma}_d\left(\frac{n}{2}\right)} \exp\left(-\frac{1}{2} \operatorname{tr} \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\right) \tag{2}$$

ただし，$\boldsymbol{\Gamma}_d(a)$ は次式で与えられる多変量ガンマ関数である．

$$\boldsymbol{\Gamma}_d(a) = \pi^{\frac{1}{4}d(d-1)} \prod_{j=1}^{d} \Gamma\left[a - \frac{1}{2}(j-1)\right] \tag{3}$$

2.2 標本共分散行列の確率密度

標本共分散行列 $\hat{\boldsymbol{\Sigma}}$，標本平均ベクトル $\hat{\boldsymbol{\mu}}$ を

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \sum_{t=1}^{n} (\boldsymbol{X}_t - \hat{\boldsymbol{\mu}}) (\boldsymbol{X}_t - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \tag{4}$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{X}_t \tag{5}$$

とおくと，$\hat{\boldsymbol{\Sigma}}$ の分布は $\mathrm{W}_d(n-1, \frac{1}{n-1}\boldsymbol{\Sigma})$ で与えられ，$\hat{\boldsymbol{\Sigma}}$ の確率密度は次式で与えられる．

$$P(\hat{\boldsymbol{\Sigma}}) = \frac{\left(\frac{n-1}{2}\right)^{\frac{1}{2}(n-1)d} |\boldsymbol{\Sigma}|^{-\frac{1}{2}(n-1)} |\hat{\boldsymbol{\Sigma}}|^{\frac{1}{2}(n-d-2)}}{\boldsymbol{\Gamma}_d\left(\frac{n-1}{2}\right)} \cdot \exp\left(-\frac{n-1}{2} \operatorname{tr} \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\Sigma}}\right) \tag{6}$$

2.3 逆 Wishart 分布

$\mathrm{W}_d(n, \boldsymbol{\Sigma})$ に従う $\boldsymbol{W}$ を用いて $\boldsymbol{V} = \boldsymbol{W}^{-1}$ とおくとき，$\boldsymbol{V}$ の分布を逆 Wishart 分布 $\mathrm{W}_d^{-1}(n, \boldsymbol{\Sigma}^{-1})$ という．$\boldsymbol{V}$ が正定値のとき，$\boldsymbol{W} \to \boldsymbol{W}^{-1}$ の変換のヤコビアンは

$$J(\boldsymbol{W}) = |\boldsymbol{V}|^{-d-1} = |\boldsymbol{W}|^{d+1} \tag{7}$$

となるので，$\boldsymbol{V}$ の確率密度は

$$P(\boldsymbol{V}) = \frac{|\boldsymbol{V}|^{-\frac{1}{2}(n+d+1)}}{2^{\frac{1}{2}nd} |\boldsymbol{\Sigma}|^{\frac{n}{2}} \boldsymbol{\Gamma}_d\left(\frac{n}{2}\right)} \exp\left(-\frac{1}{2} \operatorname{tr} \boldsymbol{\Sigma}^{-1} \boldsymbol{V}^{-1}\right) \tag{8}$$

となる．

2.4 $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ から $\boldsymbol{W}$ への変換比率

$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ の同時密度関数は

$$\begin{aligned} P(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) &= \frac{1}{(2\pi)^{\frac{nd}{2}} |\boldsymbol{\Sigma}|^{\frac{n}{2}}} \exp\left(-\frac{1}{2} \sum_{j=1}^{n} \boldsymbol{X}_j^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_j\right) \\ &= \frac{|\boldsymbol{\Sigma}|^{-\frac{n}{2}}}{(2\pi)^{\frac{nd}{2}}} \exp\left(-\frac{1}{2} \operatorname{tr} \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\right) \end{aligned} \tag{9}$$

である．式 (2) と式 (9) の比をとることで，$\boldsymbol{X}$ の空間から $\boldsymbol{W}$ の空間への変換比率 $\mathrm{H}(\boldsymbol{W}, n)$ が得られる．

$$\mathrm{H}(\boldsymbol{W},n)=\frac{P(\boldsymbol{W})}{P(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)}=\frac{|\boldsymbol{W}|^{\frac{1}{2}(n-d-1)}}{\pi^{\frac{1}{4}d(d-1)-\frac{1}{2}nd}\prod_{j=1}^{d}\Gamma\left[\frac{1}{2}(n+1-j)\right]} \tag{10}$$

2.5 多変量ベータ分布

1 変量のベータ分布を多変量に拡張，一般化したものを多変量ベータ分布といい，$\boldsymbol{U}$ を $d\times d$ の確率変数行列とすると次式で与えられる．

$$P(\boldsymbol{U})=\frac{1}{\boldsymbol{B}_d(a,b)}|\boldsymbol{U}|^{\frac{a-d-1}{2}}\,|\boldsymbol{I}_d-\boldsymbol{U}|^{\frac{b-d-1}{2}} \tag{11}$$

ただし，$\boldsymbol{I}_d$ は d 次元の単位行列である．また，$\boldsymbol{B}_d(\cdot,\cdot)$ は多変量ベータ関数で，1 変量同様，$\boldsymbol{\Gamma}_d(\cdot)$ との間には

$$\boldsymbol{B}_d(a,b)=\frac{\boldsymbol{\Gamma}_d(a)\,\boldsymbol{\Gamma}_d(b)}{\boldsymbol{\Gamma}_d(a+b)} \tag{12}$$

の関係がある．

3. Geisser の予測分布 [6]

3.1 予測分布の導出

真の分布に関する情報を何も持っていないことを表す無情報事前分布を事前分布として採用した場合に，既に得られた N 個のサンプルから $N+1$ 番目のサンプルを予測する予測分布を導く．真の平均ベクトルは既知，真の共分散行列は未知であるとする．

初めに，N 個のサンプルから求まる標本共分散行列 $\boldsymbol{S}$ から確率変数である真の共分散行列 $\acute{\boldsymbol{\Sigma}}$ の分布 $P(\acute{\boldsymbol{\Sigma}}^{-1}|\boldsymbol{S})$ を求める．まず，標本共分散行列

$$\boldsymbol{S}=\frac{1}{N}\sum_{t=1}^{N}(\boldsymbol{X}_t-\boldsymbol{\mu})(\boldsymbol{X}_t-\boldsymbol{\mu})^{\mathrm{T}} \tag{13}$$

の確率密度関数は，$\acute{\boldsymbol{\Sigma}}$ を用いた Wishart 分布 $\mathrm{W}_d(N,\frac{1}{N}\acute{\boldsymbol{\Sigma}})$ で与えられ，次式が得られる．

$$P(\boldsymbol{S}|\acute{\boldsymbol{\Sigma}}^{-1})=\frac{\left(\frac{N}{2}\right)^{\frac{1}{2}Nd}|\acute{\boldsymbol{\Sigma}}|^{-\frac{N}{2}}|\boldsymbol{S}|^{\frac{1}{2}(N-d-1)}}{\boldsymbol{\Gamma}_d\left(\frac{N}{2}\right)}\cdot\exp\left(-\frac{N}{2}\operatorname{tr}\acute{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}\right) \tag{14}$$

ここで $\acute{\boldsymbol{\Sigma}}$ の代わりに $\acute{\boldsymbol{\Sigma}}^{-1}$ を用いているのは，後で $\acute{\boldsymbol{\Sigma}}^{-1}$ について積分するためである．

無情報事前分布を仮定した真の共分散行列の事前分布 g は，$\acute{\boldsymbol{\Sigma}}^{-1}$ の空間で次のように分布しているとする．

$$g\left(\acute{\boldsymbol{\Sigma}}^{-1}\right)d\acute{\boldsymbol{\Sigma}}^{-1}\propto\left|\acute{\boldsymbol{\Sigma}}\right|^{\frac{1}{2}v}d\acute{\boldsymbol{\Sigma}}^{-1} \tag{15}$$

ただし，

$$d\acute{\boldsymbol{\Sigma}}^{-1}=\prod_{i\geqq j}d\acute{\sigma}_{ij} \tag{16}$$

である．式 (15) で $\propto$ を用いるのは，右辺を全空間で積分すると無限大になり，確率の公理を満たさないためである．v は $v\leqq N$ を満たす整数で，本論文では文献 [6] で示されている $v=d+1$ を用いる．

$\acute{\boldsymbol{\Sigma}}^{-1}$ と $\boldsymbol{S}$ の同時分布は式 (14) と式 (15) の積により，

$$P(\acute{\boldsymbol{\Sigma}}^{-1},\boldsymbol{S})\propto\frac{\left(\frac{N}{2}\right)^{\frac{1}{2}Nd}|\acute{\boldsymbol{\Sigma}}|^{-\frac{N-v}{2}}|\boldsymbol{S}|^{\frac{1}{2}(N-d-1)}}{\boldsymbol{\Gamma}_d\left(\frac{N}{2}\right)}\cdot\exp\left(-\frac{N}{2}\operatorname{tr}\acute{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}\right) \tag{17}$$

となる．$P(\acute{\boldsymbol{\Sigma}}^{-1}|\boldsymbol{S})$ を得るために，$P(\boldsymbol{S})=\int P(\acute{\boldsymbol{\Sigma}}^{-1},\boldsymbol{S})d\acute{\boldsymbol{\Sigma}}^{-1}$ を求める．ここで次式の関係 [12]

$$\int_{\boldsymbol{P}\text{が正定値}}|\boldsymbol{P}|^{\frac{a-d-2}{2}}\exp\left(-\operatorname{tr}\boldsymbol{A}\boldsymbol{P}\right)d\boldsymbol{P}=\frac{\boldsymbol{\Gamma}_d\left(\frac{a-1}{2}\right)}{|\boldsymbol{A}|^{\frac{a-1}{2}}} \tag{18}$$

を用いることで

$$P(\boldsymbol{S})\propto\frac{\left(\frac{N}{2}\right)^{\frac{1}{2}Nd}|\boldsymbol{S}|^{\frac{1}{2}(N-d-1)}}{\boldsymbol{\Gamma}_d\left(\frac{N}{2}\right)}\frac{\boldsymbol{\Gamma}_d\left(\frac{N-v+d+1}{2}\right)}{|\frac{N}{2}\boldsymbol{S}|^{\frac{N-v+d+1}{2}}} \tag{19}$$

が得られる．式 (17)，式 (19) から次式のように $P(\acute{\boldsymbol{\Sigma}}^{-1}|\boldsymbol{S})$ が得られる．

$$P(\acute{\boldsymbol{\Sigma}}^{-1}|\boldsymbol{S})=\frac{P(\acute{\boldsymbol{\Sigma}}^{-1},\boldsymbol{S})}{P(\boldsymbol{S})}=\frac{|\acute{\boldsymbol{\Sigma}}|^{-\frac{N-v}{2}}|\frac{N}{2}\boldsymbol{S}|^{\frac{N-v+d+1}{2}}}{\boldsymbol{\Gamma}_d\left(\frac{N-v+d+1}{2}\right)}\cdot\exp\left(-\frac{N}{2}\operatorname{tr}\acute{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}\right) \tag{20}$$

式 (20) の密度関数が等号で表されるのは，式 (17) と

式 (19) を全空間で積分したときの比が等しいためである．

次に，$P(\acute{\boldsymbol{\Sigma}}^{-1}|\boldsymbol{S})$ を用いて $N+1$ 番目のサンプルの予測分布 $P(\boldsymbol{x}|\boldsymbol{S})$ を求める．正規分布 $\mathrm{N}(\boldsymbol{0}, \acute{\boldsymbol{\Sigma}})$ に従う $\boldsymbol{x}$ の分布は

$$P(\boldsymbol{x}|\acute{\boldsymbol{\Sigma}}^{-1}) = (2\pi)^{-\frac{d}{2}} \left|\acute{\boldsymbol{\Sigma}}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\acute{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x}\right) \tag{21}$$

であるので，予測分布 $P(\boldsymbol{x}|\boldsymbol{S})$ は，

$$\begin{aligned} P(\boldsymbol{x}|\boldsymbol{S}) &= \int P(\boldsymbol{x}|\acute{\boldsymbol{\Sigma}}^{-1}) P(\acute{\boldsymbol{\Sigma}}^{-1}|\boldsymbol{S}) d\acute{\boldsymbol{\Sigma}}^{-1} \\ &= \int (2\pi)^{-\frac{d}{2}} \frac{|\acute{\boldsymbol{\Sigma}}|^{-\frac{N-v+1}{2}} |\frac{N}{2}\boldsymbol{S}|^{\frac{N-v+d+1}{2}}}{\boldsymbol{\Gamma}_d\left(\frac{N-v+d+1}{2}\right)} \\ &\qquad \cdot \exp\left(-\frac{1}{2}\operatorname{tr}\acute{\boldsymbol{\Sigma}}^{-1}\left(N\boldsymbol{S}+\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right)\right) d\acute{\boldsymbol{\Sigma}}^{-1} \\ &= \frac{(2\pi)^{-\frac{d}{2}} |\frac{N}{2}\boldsymbol{S}|^{\frac{N-v+d+1}{2}}}{\left|\frac{1}{2}\left(N\boldsymbol{S}+\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right)\right|^{\frac{N-v+d+2}{2}}} \frac{\boldsymbol{\Gamma}_d\left(\frac{N-v+d+2}{2}\right)}{\boldsymbol{\Gamma}_d\left(\frac{N-v+d+1}{2}\right)} \\ &= \frac{(2\pi)^{-\frac{d}{2}} |\frac{N}{2}\boldsymbol{S}|^{-\frac{1}{2}}}{\left|\boldsymbol{I}+\frac{1}{N}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{S}^{-1}\right|^{\frac{N-v+d+2}{2}}} \frac{\Gamma\left(\frac{N-v+d+2}{2}\right)}{\Gamma\left(\frac{N-v+2}{2}\right)} \end{aligned} \tag{22}$$

となる．途中，式 (18) の関係を再度用い，関係 $\boldsymbol{x}^{\mathrm{T}}\acute{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x} = \operatorname{tr}\left(\acute{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right)$ 及び $\boldsymbol{\Gamma}_d$ の定義（式 (3)）を用いた．さらに，

$$\left|\boldsymbol{I}+\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{C}^{-1}\right| = 1+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{x} \tag{23}$$

が成立するので（文献 [10] の p.594），式 (22) は

$$\begin{aligned} P(\boldsymbol{x}|\boldsymbol{S}) &= (2\pi)^{-\frac{d}{2}} |\boldsymbol{S}|^{-\frac{1}{2}} G_1(N,d;v) \\ &\qquad \cdot \left(1+\frac{1}{N}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{x}\right)^{-\frac{N-v+d+2}{2}} \end{aligned} \tag{24}$$

となる．ただし，

$$G_1(N,d;v) = \frac{\Gamma\left(\frac{N-v+d+2}{2}\right)}{\Gamma\left(\frac{N-v+2}{2}\right)\left(\frac{N}{2}\right)^{\frac{d}{2}}} \tag{25}$$

である．

これまで平均ベクトルが既知の場合を考えていたが，平均ベクトルもサンプルから推定する場合について考える．実際の学習用のサンプル数を n とすると，平均ベクトルを推定したときの共分散行列は自由度 $n-1$ の Wishart 分布 $\mathrm{W}_d(n-1, \frac{1}{n-1}\acute{\boldsymbol{\Sigma}})$ に従うので，N を $n-1$ で置き換え，さらに $v=d+1$ を代入すると，

$$\begin{aligned} P(\boldsymbol{x}|\boldsymbol{S}) &= (2\pi)^{-\frac{d}{2}} |\boldsymbol{S}|^{-\frac{1}{2}} G_2(n,d) \\ &\qquad \cdot \left(1+\frac{1}{n-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{x}\right)^{-\frac{n}{2}} \end{aligned} \tag{26}$$

$$G_2(n,d) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-d}{2}\right)\left(\frac{n-1}{2}\right)^{\frac{d}{2}}} \tag{27}$$

が得られる．

ベイズ推定を用いたパターン認識では，予測分布 $P(\boldsymbol{x}|\boldsymbol{S})$ を 2 次識別関数の代わりに識別関数として用いる．

3.2 予測分布の漸近的性質

$\boldsymbol{\Sigma}$ を真の共分散行列とする．n が大きいとき，式 (26) は多次元正規分布

$$P(\boldsymbol{x}|\boldsymbol{\Sigma}) = (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right) \tag{28}$$

に近づく．これは以下の理由による。まず、最尤推定量である標本共分散行列 $\boldsymbol{S}$ は漸近的に $\boldsymbol{\Sigma}$ と一致する。そして、$n\to\infty$ のとき，Stirling の公式

$$\Gamma(n+1) \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \tag{29}$$

から，

$$G_2(n,d) \to 1 \tag{30}$$

となる．さらに，自然対数の底 e の定義から $n\to\infty$ のとき，任意の実数 A に関して

$$\begin{aligned} \left(1+\frac{1}{n-1}A\right)^{-\frac{n}{2}} &= \left\{(1+\frac{1}{n-1}A)^{\frac{n-1}{A}}\right\}^{-\frac{nA}{2(n-1)}} \\ &\to e^{-\frac{1}{2}A} \end{aligned} \tag{31}$$

が成り立ち、ここで

$$A = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{x} \tag{32}$$

とおくと，式 (26) の $\left(1+\frac{1}{n-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{x}\right)^{-\frac{n}{2}}$ は $n\to\infty$ で $\exp\left(-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)$ となる．したがって，$n\to\infty$ のとき，式 (26) は漸近的に式 (28) と一致する．

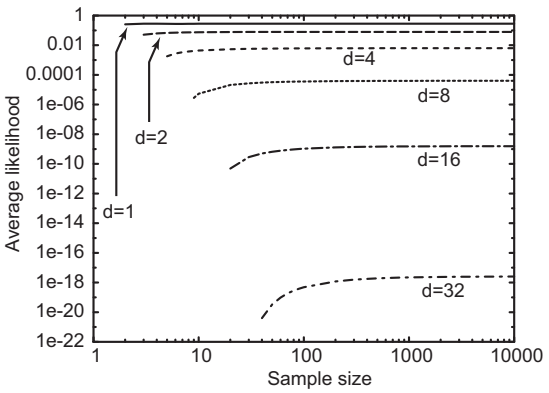


図 1　尤度の期待値 $B_d(n)$ (分散 1 の場合)
Fig. 1　Average likelihood $B_d(n)$ in the case of variance 1.

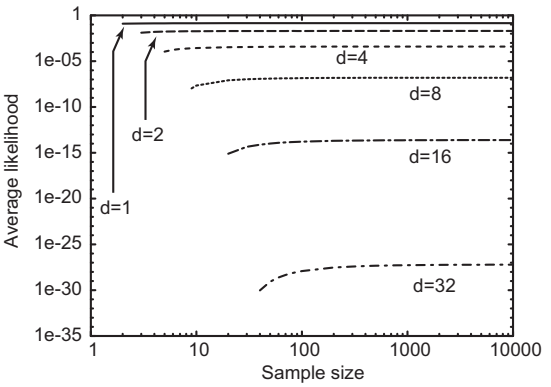


図 2　尤度の期待値 $B_d(n)$ (分散 4 の場合)
Fig. 2　Average likelihood $B_d(n)$ in the case of variance 4.

4. 予測分布の偏りの補正

4.1 予測分布の期待値

Geisser の予測分布 (式 (26)) の正規分布 (式 (28)) からの偏りはサンプル数 n に応じて変わる．この偏りの n に応じた変化を調べるために，分散 1 と 4 の等方性正規分布に従う n 個の人工サンプルを実際に作成し，式 (26) を用いて計算される尤度の期待値 $B_d(n)$ を求める．$B_d(n)$ を式で表すと，

$$B_d(n) = \frac{1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n} P(\boldsymbol{x}_{ti}|\boldsymbol{S}_t(n)) \tag{33}$$

である．ただし，$\boldsymbol{x}_{t1}, \boldsymbol{x}_{t2}, \ldots, \boldsymbol{x}_{tn}$ は t 回目の実験で用いた n 個のサンプル，$\boldsymbol{S}_t(n)$ は t 回目の実験で n 個のサンプルから求めた標本共分散行列である．T は実験した回数で，ここでは 1,000 である．

次元数 1，2，4，8，16，32 のそれぞれについて，分散が 1，4 の等方性正規分布のときの $B_d(n)$ を図 1，2 に示す．図から，$B_d(n)$ はサンプル数 n が小さくなるほど，偏りの程度が大きくなることがわかる．

サンプル数の違いで尤度の偏り方が異なるということは，予測分布を認識に用いる場合，サンプルが少ないクラスの尤度を過小評価してしまい，不公平な比較になるということを意味する．尤度の偏りをあらかじめ知ることができれば，偏りを補正して公平に比較できると考えられる．

4.2 $\boldsymbol{B_d(n)}$ の理論式

サンプルが正規分布に従うときの尤度の期待値 $B_d(n)$ を解析的に導く．サンプルが正規分布 $\mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ に従うときの Geisser の予測分布の期待値 $B_d(n)$ は式 (21)，式 (26)，式 (6) を用いて次式で与えられる．

$$\begin{aligned} B_d(n) &= \iint P(\boldsymbol{x}|\boldsymbol{\Sigma}^{-1})P(\boldsymbol{x}|\boldsymbol{S})P(\boldsymbol{S})d\boldsymbol{S}d\boldsymbol{x} \\ &= \frac{(2\pi)^{-d}\left(\frac{n-1}{2}\right)^{\frac{1}{2}(n-1)d} G_2(n,d)}{|\boldsymbol{\Sigma}|^{\frac{n}{2}}\boldsymbol{\Gamma}_d\left(\frac{n-1}{2}\right)} \\ &\quad \cdot \iint |\boldsymbol{S}|^{\frac{1}{2}(n-d-3)} \exp\left(-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right) \\ &\quad \cdot \left(1+\frac{1}{n-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{x}\right)^{-\frac{n}{2}} \\ &\quad \cdot \exp\left(-\frac{n-1}{2}\,\mathrm{tr}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right) d\boldsymbol{S}d\boldsymbol{x} \end{aligned} \tag{34}$$

以後，式 (34) の積分部分 (2 行目以降) についてのみ考え，最後に定数部分 (1 行目) と統合する．まず，$(n-1)\boldsymbol{S} = \boldsymbol{W}$ と変換する．$\boldsymbol{S}$，$\boldsymbol{W}$ のパラメータ数が $\frac{d(d+1)}{2}$ であることに注意すれば，このときのヤコビアンは次のようになる．

$$\begin{aligned} J(\boldsymbol{S}\to\boldsymbol{W}) &= \left|\frac{1}{n-1}\boldsymbol{I}_{\frac{d(d+1)}{2}}\right| \\ &= (n-1)^{-\frac{d(d+1)}{2}} \end{aligned} \tag{35}$$

ただし，$\boldsymbol{I}_{\frac{d(d+1)}{2}}$ は $\frac{d(d+1)}{2}$ 次元の単位行列を表す．この結果を用いると，式 (34) の積分部分は，

$$(n-1)^{-\frac{d(n-2)}{2}} \iint |\boldsymbol{W}|^{\frac{1}{2}(2n-d-3)} \left|\boldsymbol{W}+\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right|^{-\frac{n}{2}} \cdot \exp\left(-\frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}+\boldsymbol{W}\right)\right) d\boldsymbol{W}d\boldsymbol{x} \quad (36)$$

となる．途中，式 (23) の関係を用いた．

次に，$\boldsymbol{X}=\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$ と変換する．この変換の比率は式 (10) で与えられ，

$$\mathrm{H}(\boldsymbol{X},1)=\frac{\pi^{\frac{d}{2}}|\boldsymbol{X}|^{-\frac{d}{2}}}{\boldsymbol{\Gamma}_d\left(\frac{1}{2}\right)} \quad (37)$$

となるので，

$$(n-1)^{-\frac{d(n-2)}{2}}\frac{\pi^{\frac{d}{2}}}{\boldsymbol{\Gamma}_d\left(\frac{1}{2}\right)} \cdot \iint |\boldsymbol{W}|^{\frac{1}{2}(2n-d-3)}|\boldsymbol{W}+\boldsymbol{X}|^{-\frac{n}{2}}|\boldsymbol{X}|^{-\frac{d}{2}} \cdot \exp\left(-\frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}+\boldsymbol{W})\right) d\boldsymbol{W}d\boldsymbol{X} \quad (38)$$

が得られる．

そして，$\boldsymbol{C}=\boldsymbol{W}+\boldsymbol{X}$ とおき，積分変数を $\boldsymbol{X}$ から $\boldsymbol{C}$ に変える．この変換のヤコビアン J は $J=1$ である．$\boldsymbol{X}=\boldsymbol{C}-\boldsymbol{W}$ を式 (38) に代入し，整理すると，

$$(n-1)^{-\frac{d(n-2)}{2}}\frac{\pi^{\frac{d}{2}}}{\boldsymbol{\Gamma}_d\left(\frac{1}{2}\right)} \iint |\boldsymbol{W}|^{\frac{1}{2}(2n-d-3)}|\boldsymbol{C}|^{-\frac{n+d}{2}} \cdot |\boldsymbol{I}-\boldsymbol{W}\boldsymbol{C}^{-1}|^{-\frac{d}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\right) d\boldsymbol{W}d\boldsymbol{C} \quad (39)$$

となる．

さらに，

$$\boldsymbol{C}=\boldsymbol{T}^{\mathrm{T}}\boldsymbol{T} \quad (40)$$

$$\boldsymbol{W}=\boldsymbol{T}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{T} \quad (41)$$

とおく．$\boldsymbol{T}$ は上三角行列である．積分変数を $\boldsymbol{W}$ から $\boldsymbol{U}$ に変えれば，変換のヤコビアンは次式で与えられる（文献 [11] の p.109）．

$$J(\boldsymbol{W}\to\boldsymbol{U})=|\boldsymbol{C}|^{\frac{d+1}{2}} \quad (42)$$

したがって，

$$|\boldsymbol{I}-\boldsymbol{W}\boldsymbol{C}^{-1}|=\left|\boldsymbol{T}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{U})\boldsymbol{T}\left(\boldsymbol{T}^{\mathrm{T}}\boldsymbol{T}\right)^{-1}\right| =|\boldsymbol{I}-\boldsymbol{U}| \quad (43)$$

の関係に注意すれば，式 (39) は

$$(n-1)^{-\frac{d(n-2)}{2}}\frac{\pi^{\frac{d}{2}}}{\boldsymbol{\Gamma}_d\left(\frac{1}{2}\right)} \cdot \int |\boldsymbol{U}|^{\frac{1}{2}(2n-d-3)}|\boldsymbol{I}-\boldsymbol{U}|^{-\frac{d}{2}}d\boldsymbol{U} \cdot \int |\boldsymbol{C}|^{\frac{n-d-2}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\right)d\boldsymbol{C} \quad (44)$$

となる．ここで，$\boldsymbol{U}$ に関する積分が多変量ベータ分布（式 (11)），$\boldsymbol{C}$ に関する積分が Wishart 分布（式 (2)）の一部になっていることから，式 (44) は，

$$(n-1)^{-\frac{d(n-2)}{2}}\pi^{\frac{d}{2}}\frac{\boldsymbol{\Gamma}_d\left(\frac{2n-2}{2}\right)}{\boldsymbol{\Gamma}_d\left(\frac{2n-1}{2}\right)}\frac{2^{\frac{d(n-1)}{2}}\boldsymbol{\Gamma}_d\left(\frac{n-1}{2}\right)}{|\boldsymbol{\Sigma}|^{-\frac{n-1}{2}}} \quad (45)$$

となる．

式 (45) に式 (34) の定数部分を掛け，整理すると，

$$B_d(n)=(2\pi)^{-\frac{d}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\frac{\Gamma\left(\frac{2n-d-1}{2}\right)}{\Gamma\left(\frac{2n-1}{2}\right)}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-d}{2}\right)} \quad (46)$$

が得られる．式 (46) から，$B_d(n)$ が特徴量の次元数，サンプル数，真の共分散行列で定まる関数であることがわかる．

4.3 予測分布の偏りの補正法

予測分布を用いた認識を行う場合に，尤度の偏りを補正する方法を述べる．

共分散行列の推定に用いたサンプル数を n，適当な大きな数を n' とする．n と n' のそれぞれについて，式 (46) から確率の偏りの理論値 $B_d(n)$ と $B_d(n')$ を求めると，$B_d(n)$ と $B_d(n')$ の差が偏りを与える．したがって，学習サンプル数が n 個のクラスの尤度の偏りは $B_d(n')/B_d(n)$ を掛けることによって補正できる．このとき $\boldsymbol{\Sigma}$ に関する項が消えるため，補正量は特徴量の次元数，サンプル数のみに依存する．

5. 実　　験

Geisser の予測分布（式 (26)）のクラス毎の偏りを 4.3 の方法で補正した場合に認識性能が改善することを 2 クラスの文字認識実験で確認する．

実験には正規分布に従う人工サンプルを用い，次元数 1，2，4，8，16，32 のときのそれぞれについて，次の等方性正規分布を用いる．クラス 1 の分布は平均

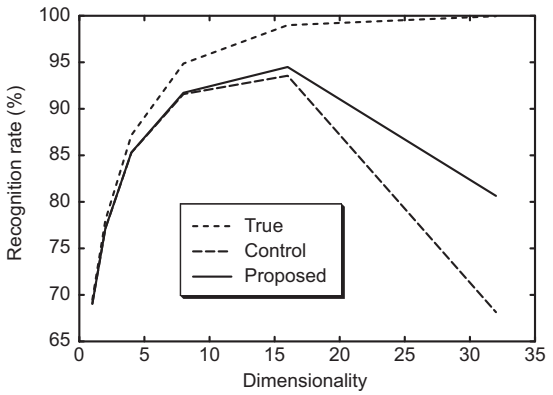


図 3　認識率(サンプル数比固定 , サンプル数比 : 条件 S1)
Fig. 3　The recognition rates of the experiments (fixed sample size, ratio of sample sizes is condition S1).

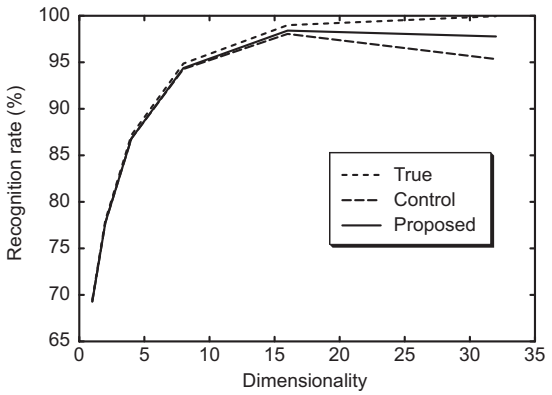


図 4　認識率(サンプル数比固定 , サンプル数比 : 条件 S2)
Fig. 4　The recognition rates of the experiments (fixed sample size, ratio of sample sizes is condition S2).

$\boldsymbol{\mu}_1 = (0, 0, \cdots, 0)$, 分散 1 の等方性正規分布とし , クラス 2 の分布は平均 $\boldsymbol{\mu}_2 = (1, 1, \cdots, 1)$, 分散 4 の等方性正規分布とする .

実験はサンプル数の比を固定して次元数を変える場合と , 次元数を固定してサンプル数の比を変える場合の 2 種類行う . サンプル数を固定する場合は、クラス 1 のサンプル数を 40 , クラス 2 のサンプル数を 10,000 とした場合 (条件 S1) と , クラス 1 のサンプル数を 10,000 , クラス 2 のサンプル数を 40 とした場合 (条件 S2) の 2 通りとする . 次元数を固定する場合は次元数を 16 に固定し , クラス 2 のサンプル数を 10,000 に固定してクラス 1 のサンプル数を変える場合 (条件 D1) と , クラス 1 のサンプル数を 10,000 に固定してクラス 2 のサンプル数を変える場合 (条件 D2) の 2 通りとする .

認識実験を各 1,000 回行い , 得られた認識率を平均する . テストサンプルには学習用とは異なるサンプルを 1,000 個ずつ用いる . 尤度の偏りの補正で用いる n' を 10,000,000 とする .

サンプル数比を固定した場合の認識結果を図 3 , 図 4 , 次元数を固定した場合の認識結果を図 5 , 図 6 に , 提案手法の補正効果を示す認識率の差分を図 7 ~ 10 に示す . 図中の “Control” は尤度の偏りを補正しない場合 , “Proposed” は尤度の偏りを補正する場合 , “True” は真の分布を辞書にして “Control” や “Proposed” で用いるテストサンプル (計 1,000,000 個) を 2 次識別関数で認識した場合の認識率を表す .

サンプル数比を固定した場合については , “Control , ” “Proposed” ともに 16 次元の認識率が最も高かった . これは , 次元数が大きくなると , “True” の認識率が単調増加していることからわかるようにベイズエラーが減少することと , 次元数に対して学習サンプル数が不足することによる誤差の増大との効果によるものと考えられる . 図 3 と図 4 , 図 7 と図 8 の比較から , 文献 [8] で指摘されているように , 分散の小さなクラスの学習サンプルが少ないほど認識率が低下する傾向がみられ , より大きな補正効果が得られた . 逆に , 補正効果は分散の大きなクラスの学習サンプルが少ないほど低い傾向がみられ , 図 8 の次元数が 1 の場合には補正効果が得られなかった . これらのことについては , 6. で考察する .

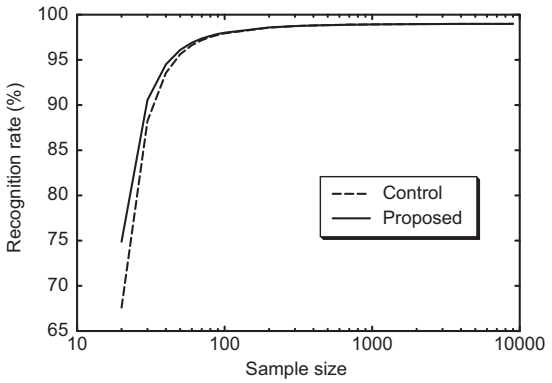


図 5　認識率 (次元数固定 , サンプル数比 : 条件 D1)
Fig. 5　The recognition rates of the experiments (fixed dimensionality, ratio of sample sizes is condition D1).

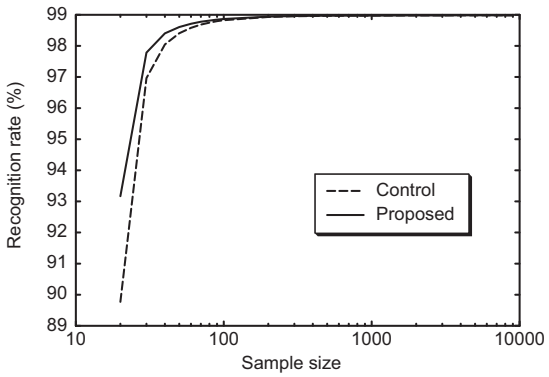


図 6　認識率 (次元数固定，サンプル数比：条件 D2)
Fig. 6　The recognition rates of the experiments (fixed dimensionality, ratio of sample sizes is condition D2).

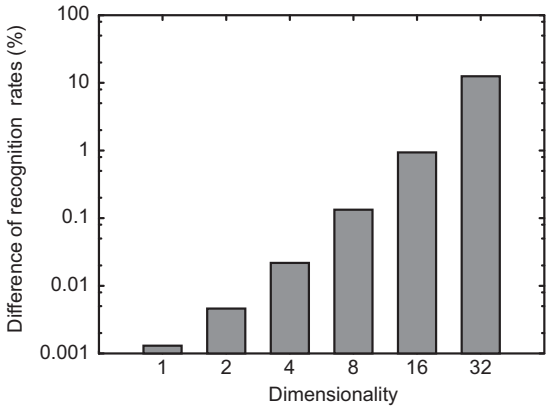


図 7　提案手法の補正効果 (サンプル数比固定，サンプル数比：条件 S1)
Fig. 7　The effect of the proposed method (fixed sample size, ratio of sample sizes is condition S1).

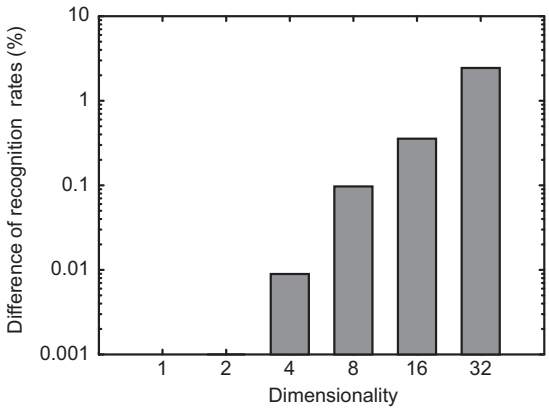


図 8　提案手法の補正効果 (サンプル数比固定，サンプル数比：条件 S2)
Fig. 8　The effect of the proposed method (fixed sample size, ratio of sample sizes is condition S2).

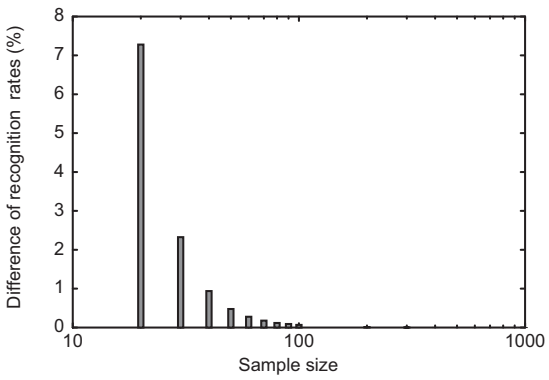


図 9　提案手法の補正効果 (次元数固定，サンプル数比：条件 D1)
Fig. 9　The effect of the proposed method (fixed dimensionality, ratio of sample sizes is condition D1).

次元数を固定した場合については，認識率はサンプル数に対して単調に増加した．原因として，学習サンプル数の増加とともに予測分布の偏りが減少したこと，学習サンプル数が増加したことによって推定誤差が減少したことが考えられる．図 9，図 10 から，次元数が小さいほど提案手法の補正効果が大きいことがわかる．また，サンプル数比を固定した場合と同様，分散の小さなクラスの学習サンプルが少ないほど認識率が低下する傾向がみられた．

尤度の偏りを提案手法で補正することによって認識性能に一定の改善がみられたことから，ベイズ推定を用いたパターン認識では予測分布の偏りが認識性能を悪化させていることが確認できる．

6. 考　　察

文献 [8] では，分散の小さなクラスの学習サンプルが少ないほど認識率が低下する傾向があることが指摘されている．本論文の認識実験においても同様の結果がみられた．この理由は，4. 2 で導いた理論式を用いて簡潔に説明できる．式 (46) によると、サンプルが少ないときの予測分布の偏りは真の分散が小さいほど大きい．このことは，分散が大きいクラスの尤度が相対的に高くなり，分散が小さいクラスに属すべきサンプルは分散が大きいクラスに誤認識されやすくなることを意味する．したがって，分散が小さいクラスのサ

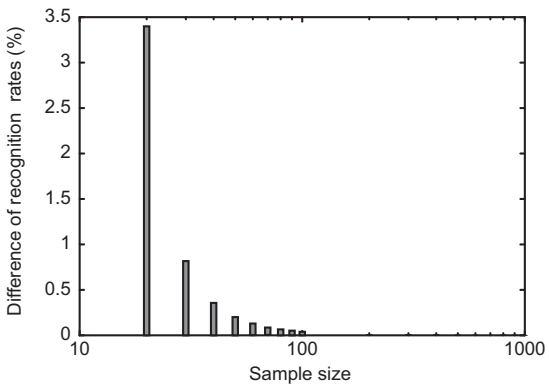


図 10　提案手法の補正効果 (次元数固定，サンプル数比 : 条件 D2)

Fig. 10　The effect of the proposed method (fixed dimensionality, ratio of sample sizes is condition D2).

ンプルが少ないという，予測分布により尤度に大きな偏りが生じる条件が揃ったときに認識性能が大きく低下するといえる．逆に，分散が大きいクラスのサンプルが少ない場合には，尤度に生じる偏りのバランスが拮抗し，分散が小さいクラスのサンプルが少ない場合よりも誤認識が起こりにくくなるといえる．

7. むすび

ベイズ推定を用いたパターン認識では，クラス間でサンプル数が等しい場合は有効であるが，クラス間でサンプル数が異なる場合には認識性能が向上しないことが指摘されている．サンプル数が十分大きくない場合には，予測分布の期待値が多次元正規分布と比べて小さく偏り，偏り方はサンプル数によって異なることから，サンプル数が異なる場合には予測分布の偏りに差が生じて，認識性能が低下するものと考えられる．

本論文では，尤度の偏りを補正することで認識性能が改善できることを示した．すなわち，予測分布の偏りの理論式 (厳密解) を導出し，サンプル数がクラス間で異なる場合は，導いた理論式で尤度を補正することで認識精度が向上することを確認した．また，理論式から，文献 [8] で指摘されている「分散の小さなクラスの学習サンプルが少ないほど認識率が低下する」という経験的に得られた知見の根拠を明らかにした．

本論文で示した予測分布の偏りの理論式の導出法は，別の事前分布を仮定した場合 (例えば，Keehn の方法) にも容易に適用できる高い一般性を持つ．

文　献

[1] 竹村彰通，現代数理統計学，創文社，東京，1991．
[2] 上田修功，“ベイズ学習 [II]—ベイズ学習の基礎—，” 電子情報通信学会誌，vol.85, no.6，pp.421–426，June，2002．
[3] F. Komaki, “On asymptotic properties of predictive distributions,” Biometrika, vol.83, no.2, pp.299–313, June, 1996.
[4] 駒木文保，“予測分布の理論について，” 信学論 (A)，vol.J83-A, no.6，pp.612–619，June，2000．
[5] D. G. Keehn, “A note on learning for gaussian properties,” IEEE Trans. on Inf. Theory, vol.IT-11, pp.126–132, Jan., 1965.
[6] S. Geisser, “Posterior odds for multivariate normal parameters,” Journal of the Royal Statistical Society. Series B, Methodological, vol.26, no.1, pp.69–76, 1964.
[7] S. J. Raudys and A. K. Jain, “Small sample size effects in statistical pattern recognition: Recommendations for practitioners,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol.13, no.3, pp.252–264, Mar., 1991.
[8] 韓雪仙，若林哲史，木村文隆，三宅康二，“ベイズアプローチによる最適識別系の有限標本効果に関する考察—学習標本の大きさがクラス間で異なる場合—，” 信学論 (D-II)，vol.J82-D-II, no.4，pp.621–630，Apr.，1999．
[9] 竹村彰通，多変量推測統計の基礎，共立出版，東京，1991．
[10] T. W. Anderson, An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, Inc., 2nd edition, 1984.
[11] R. J. Muirhead, Aspects of Multivariate Statistical Theory, John Wiley & Sons, Inc., New York, 1982.
[12] H. Cramer, Mathematical methods of statistics, pp.391–394, Princeton University Press, 1946.

(平成 xx 年 xx 月 xx 日受付)

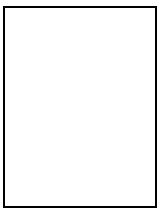
岩村　雅一　(正員)

平 10 東北大・工・通信卒．平 15 同大大学院博士課程了．同年同大学院工学研究科助手．現在に至る．博士 (工学)．パターン認識に関する研究に従事．情報処理学会会員．

大町真一郎　(正員)

昭 63 東北大・工・情報卒．平 5 同大大学院博士課程了．同年同大情報処理教育センター助手．平 8 同大工学部助手．平 11 同大大学院工学研究科助教授，現在に至る．博士 (工学)．その間，平 12 ~ 13 米国ブラウン大学客員助教授．パターン認識，コ

ンピュータビジョン，並列処理，文字認識システムの開発などの研究に従事．IEEE，情報処理学会，人工知能学会，Pattern Recognition Society 各会員．



阿曽　弘具　（正員：フェロー）

昭 43 東北大・工・電気卒．昭 49 同大大学院博士課程了．昭 48 東北大・工・助手，昭 54 名大・工・講師．昭 57 同助教授，昭 61 東北大・工・助教授を経て，平 3 同教授．現在，同大大学院工学研究科教授．工博．その間，学習オートマトン，セル構造オートマトン，並行処理理論，シストリックアルゴリズム設計論，文字認識，音声認識，ニューラルネットワークなどの研究に従事．平 3 年度本会業績賞受賞．IEEE, ACM, EATCS，情報処理学会，人工知能学会，日本認知科学会，LA 各会員．