数理解析研究所講究録 第2091巻 2018年 38-54

クロスバリデーションによる誤判別確率の推定に対するバイアス補正法

中川　智之
東京理科大学 理工学部 情報科学科 *
Tomoyuki Nakagawa
Department of Information Sciences, Faculty of Science & Technology
Tokyo University of Science

1　導入

判別分析では古くから多くの判別手法 (例えば, Fisher の線形判別 [5] や二次判別など) が提案されており, 近年でもサポートベクターマシン (SVM) やスパース判別分析 (SDA) [1] などの高次元データや大規模データに対応した判別手法が提案されている. 数多くある判別手法を比較する際に最もよく用いられる指標が誤判別確率である. 誤判別確率は“間違った母集団に判別してしまう確率”で, 2 つの母集団 Π_1, Π_2 の場合には,

$$P(2|1) = \Pr\{x \in \Pi_1 \text{ を誤って } \Pi_2 \text{ に判別する }\},$$
$$P(1|2) = \Pr\{x \in \Pi_2 \text{ を誤って } \Pi_1 \text{ に判別する }\},$$

と表せ, 誤判別確率をより小さくする判別手法が良いことがわかる. しかしながら, 誤判別確率を正確に知ることは困難であり, 何らかの手法で推定する必要がある. 誤判別確率の推定方法には大きく分けて“パラメトリックな手法”と“ノンパラメトリックな手法”の 2 種類ある. パラメトリックな手法は [10] や [6] などの漸近理論を用いて近似式を導出する方法がよく知られており, 近年でも [12] などで行われている. ノンパラメトリックな手法はクロスバリデーション ([7], [11]) やブートストラップ ([3], [4]) などに代表される分

* 〒278-8510 千葉県野田市山崎 2641

布などの仮定しない方法である．パラメトリックな手法は理論的な妥当性があり，高次のオーダーまで近似精度を上げることができる．しかしながら，分布や判別手法ごとに導出する必要があり，適用範囲がとても狭いことが問題である．一方，ノンパラメトリックな手法は分布や判別手法などの仮定が必要ないため，適用範囲が広い．実際に適用しやすく多くの場面で用いられているが，理論的な妥当性が乏しく，[8] や [3] などでは，クロスバリデーションによる推定は漸近不偏性を持つと述べられているが，これは標本数が十分大きい場合に限ってである．特に高次元データに対しては妥当性はほとんど分かっていない．

高次元データにも対応できる推定方法としては，[2] や [6] で Fisher の線形判別などに関して高次元大標本漸近理論に基づいて漸近近似式を与えている．さらに，[?] では高次のオーダーまでバイアス補正をした推定方法を提案している．一方で，[9] ではクロスバリデーションによる推定について高次元大標本漸近理論に基づいて，漸近不偏性や一致性をある条件の下で示しており，クロスバリデーションの理論的な妥当性を導出している．さらに，[9] ではクロスバリデーションによる推定に対してバイアス補正法を提案している．

本稿では，[9] で提案されているバイアス補正法について，仮定を満たさない場合での振る舞いを数値実験を用いて検証を行う．第 2 節では，クロスバリデーションの漸近性質について紹介する．第 3 節では，バイアス補正法について紹介する．第 4 節では正規分布と t 分布について数値実験を行い，推定量の比較を行う．

2　高次元大標本におけるクロスバリデーションの漸近性質

本節では次の 2 つの p 次元の正規母集団を考える．

$$\Pi_1 : N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}),\ \Pi_2 : N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

さらに 2 つの母集団に対する判別ルールを判別関数 $d(\cdot)$ を用いて，次のように定義する．

$$d(\boldsymbol{x}) > c \Rightarrow \boldsymbol{x} \in \Pi_1,$$
$$d(\boldsymbol{x}) \leq c \Rightarrow \boldsymbol{x} \in \Pi_2.$$

ここで，c はカットオフポイントである．このとき，誤判別確率は，

$$P(2|1) = \Pr(d(\boldsymbol{x}) \leq c \mid \boldsymbol{x} \in \Pi_1),$$
$$P(2|1) = \Pr(d(\boldsymbol{x}) > c \mid \boldsymbol{x} \in \Pi_2),$$

と表すことができる．判別関数は Π_k からのテストデータ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_k} (k = 1, 2)$ を用いて構成される．例えば，Fisher の線形判別は

$$d_F(\boldsymbol{x}) = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}^{-1} \left\{ \boldsymbol{x} - \frac{1}{2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \right\},$$

と表せる．ここで，$\bar{\boldsymbol{x}}_k$ は Π_k の標本平均 $(k = 1, 2)$ で，$\boldsymbol{S}$ はプールされた分散共分散行列である．さらに，本稿で扱う高次元大標本漸近理論を

$$p, N_1, N_2 \to \infty, \ \frac{N}{N_k} = O(1) \ \ (k = 1, 2), \ \frac{p}{N} \to c_0 \in (0, 1), \ (N - p + 2 > 0),$$

と定義する．ここで，$N = N_1 + N_2$ である．このとき，クロスバリデーションによる推定は以下のように表せる．

$$\hat{P}_{CV} = N_1^{-1} \sum_{i=1}^{N_1} 1(d^{(-i)}(\boldsymbol{x}_{1i}) \leq c),$$

ここで，$1(\cdot)$ は定義関数で，$d^{(-i)}$ は $\boldsymbol{x}_{1i}$ を除いた判別関数である．そして，次の定理が成り立つ．

定理 2.1. 誤判別確率 $P(2|1)$ に対して，

$$P(2|1) = Q_0\left(\frac{p}{N_1}, \frac{p}{N_2}\right) + \frac{1}{N} Q_1\left(\frac{p}{N_1}, \frac{p}{N_2}\right) + O_2, \tag{1}$$

となる展開が与えられてると仮定する．ここで $Q_0(x_1, x_2)$ と $Q_1(x_1, x_2)$ は $(p/N_1, p/N_2)$ 周りで C^1 級関数ある．このとき，バイアスは

$$\mathrm{E}[\hat{P}_{CV}(2|1)] - P(2|1) = O_1,$$

となる．つまり，$\hat{P}_{CV}(2|1)$ は漸近的に不偏推定量である．ここで，O_k は高次元大標本漸近理論の枠組みで $(p^{-1}, N_1^{-1}, N_2^{-1}, N^{-1})$ に関するオーダーとする．

次に平均二乗誤差 (MSE) の評価をみる．$\hat{P}(2|1)$ の MSE は次のように計算できる．

$$\begin{aligned}
&\mathrm{MSE}\left(\hat{P}_{CV}(2|1)\right) = \mathrm{Bias}\left(\hat{P}_{CV}(2|1)\right)^2 + \mathrm{Var}\left(\hat{P}_{CV}(2|1)\right), \\
&\mathrm{Var}\left(\hat{P}_{CV}(2|1)\right) \\
&\quad = \Pr\left(d^{(-1)}(\boldsymbol{x}_{11}) \leq c, d^{(-2)}(\boldsymbol{x}_{12}) \leq c\right) - \Pr\left(d^{(-1)}(\boldsymbol{x}_{11}) \leq c\right)^2 \\
&\qquad + \frac{1}{N_1}\left[\Pr\left(d^{(-1)}(\boldsymbol{x}_{11}) \leq c\right) - \Pr\left(d^{(-1)}(\boldsymbol{x}_{11}) \leq c, d^{(-2)}(\boldsymbol{x}_{12}) \leq c\right)\right].
\end{aligned}$$

この計算から $d^{(-1)}(\boldsymbol{x}_{11})$ と $d^{(-2)}(\boldsymbol{x}_{12})$ が漸近的に独立であれば, $\hat{P}_{CV}(2|1)$ が一致性を持つことがわかる. つまり,

$$\Pr\left(d^{(-1)}(\boldsymbol{x}_{11}) \leq c, d^{(-2)}(\boldsymbol{x}_{12}) \leq c\right) - \Pr\left(d^{(-1)}(\boldsymbol{x}_{11}) \leq c\right)^2 \to 0, \tag{2}$$

が成り立てば一致性を持つ. [9] では, Fisher の線形判別を含むクラスの判別手法で (2) が成り立つことを示している.

3 バイアス補正法

クロスバリデーションは漸近的に不偏推定量ではあるが, 標本数が十分に大きくない場合などはバイアスが大きくなるためバイアス補正をする必要がある. 本節では [9] で提案されている“2 つ抜きの CV を用いる方法”, “少しだけ残す CV の方法”, “カットオフポイントをずらす方法”の 3 つの方法について紹介する.

3.1 Method I : 2 つ抜きの CV を用いる方法

Method I はノンパラメトリックなバイアス補正法であり, [13] で情報量規準のバイアス補正法として提案されている. 本稿ではこの方法を誤判別確率に対するクロスバリデーションに応用する. さらに, 高次元データにも対応できるようなバイアス補正法に改良する. 2 つ抜きの CV は次のように定義できる.

$$\hat{P}_{CV_2}(2|1) = \frac{1}{{}_{N_1}C_2} \sum_{i<j}^{N_1} \frac{1}{2} \sum_{k\in\{i,j\}} 1\left(d^{(-i,-j)}(\boldsymbol{x}_{1k}) \leq c\right).$$

ここで $N_j^{(-\ell)} = N_j - \ell$, $N^{(-\ell)} = N - \ell$ で, $\boldsymbol{x}_{1i}$ と $\boldsymbol{x}_{1j}$ を除いたときの判別関数を $d^{(-i,-j)}$ とする. このとき, 推定量を

$$\hat{P}_{\mathrm{I}}(2|1) = \left\{\hat{P}_{CV}(2|1) - \frac{N_1^{(-2)}}{N_1}\left(\hat{P}_{CV_2}(2|1) - \hat{P}_{CV}(2|1)\right)\right\},$$

このように与えると, バイアスは

$$\mathrm{E}\left[\hat{P}_{\mathrm{I}}(2|1)\right] - P(2|1) = O_2,$$

となる. また, 3 つ抜きの CV を用いるとさらにバイアス補正が可能になる.

3.2 Method II : 少しだけ抜く CV

本節ではクロスバリデーションを用いる際に 1 つ抜くのではなく, 少しだけ残す方法を考える. この方法は [14] と [15] で情報量規準のバイアス補正法として提案されており, 本稿ではこの方法を誤判別確率に対するクロスバリデーションに応用する. さらに, 高次元データにも対応できるようなバイアス補正法に改良する. $F_{N-1}^{(-i)}$ と F_i をそれぞれ $\boldsymbol{x}_{11},\ldots,\boldsymbol{x}_{1i-1},\boldsymbol{x}_{1i+1}\ldots,\boldsymbol{x}_{1N_1}$ と $\boldsymbol{x}_{1i}$ の経験分布とする. このとき, 判別関数 $\hat{d}^{(-i;\lambda)}$ を $(1-u_\lambda)F_{N-1}^{(-i)}+u_\lambda F_i$ を用いて構成する. ここで $u_\lambda=(1-\lambda)/(N_1-\lambda)$ とする. 例えば, 判別関数 d_θ が θ によってパラメータ付されているとすると, θ の推定量を

$$\hat{\theta}^{(-i;\lambda)}=\arg\max_{\theta\in\Theta}\left\{\frac{1}{N_1-\lambda}\sum_{k\neq i}^{N_1}\log f(\boldsymbol{x}_{1k};\theta)+\frac{1-\lambda}{N_1-\lambda}\log f(\boldsymbol{x}_{1i};\theta)\right\},$$

で与える. ここで f は $\boldsymbol{x}_{1i}$ の確率密度関数である. このとき $\hat{d}^{(-i;\lambda)}=d_{\hat{\theta}^{(-i;\lambda)}}$ である. 正規分布の場合は, 平均 $\bar{\boldsymbol{x}}_1^{(-i;\lambda)}$ と分散共分散行列 $\boldsymbol{S}^{(-i;\lambda)}$ は

$$\bar{\boldsymbol{x}}_1^{(-i;\lambda)}=\frac{N_1-1}{N_1-\lambda}\bar{\boldsymbol{x}}_1^{(-i)}+\frac{1-\lambda}{N_1-\lambda}\boldsymbol{x}_{1i}$$

$$\boldsymbol{S}^{(-i;\lambda)}=\frac{1}{N^{(-\lambda)}}\left\{\left(N^{(-3)}\right)\boldsymbol{S}^{(-i)}+\frac{N_1^{(-1)}}{N_1^{(-\lambda)}}(1-\lambda)\left(\boldsymbol{x}_{1i}-\bar{\boldsymbol{x}}_1^{(-i)}\right)\left(\boldsymbol{x}_{1i}-\bar{\boldsymbol{x}}_1^{(-i)}\right)^\top\right\}$$

で与えられる. このとき, $\hat{d}^{(i;\lambda)}$ を用いて

$$\hat{P}_{CV_\lambda}(2|1)=\frac{1}{N_1}\sum_{i=1}^{N_1}1(\hat{d}^{(-i;\lambda)}(\boldsymbol{x}_{1i})\leq c),$$

とすると, λ をうまく選ぶことができればバイアスを補正することが可能になる. [9] では λ を正規分布で Fisher の線形判別に対して次のように与えている.

$$\lambda=1-\kappa(\Delta)/N,$$

$$\kappa(\Delta)=\frac{N}{4N_1}\left\{2-\left(\Delta^2+\frac{p}{N_1}+\frac{p}{N_2}\right)^{-1}\left(\Delta^2+\frac{p}{N_2}-\frac{p}{N_1}\right)\right\}.$$

また, この方法は $\lambda=1$ の場合は通常のクロスバリデーションになる.

3.3 Method III : カットオフポイントをずらす方法

判別分析ではカットオフポイント c を変えることで誤判別確率が変化するため, 本節ではクロスバリデーションを行う際に

$$\hat{P}_{CV_{c+c_1N^{-1/2}}}(2|1) = \sum_{i=1}^{N_1} 1(d^{(-i)}(x_{i1}) \leq c + c_1 N^{-1/2}),$$

のように c をずらすことでバイアス補正を可能にする. Method II と同様に c_1 を選択することでバイアス補正ができる. 正規分布で Fisher の線形判別の場合は c_1 は

$$\begin{aligned}
\eta^{(-1)} &= \frac{n-1}{N-p-1}\left(\Delta^2 + \frac{p}{N_2} - \frac{bp}{n_1} + p(1-b)\right) = \eta + \eta_1 + O_2 \\
(s^{(-1)})^2 &= 4\frac{(n-1)^2(N-1)}{(N-p-1)^3}\left(\Delta^2 + \frac{pb^2}{n_1} + \frac{p}{N_2}\right) \\
&= s^2 + s_1 + O_2 \\
c_1(\Delta) &= \frac{N}{s}\left\{\frac{s_1}{2}(c-\eta) - s\eta_1\right\}.
\end{aligned}$$

で与えられる. ここで

$$\begin{aligned}
\eta &= \frac{p}{N-p}\mathrm{tr}(\boldsymbol{A\Omega}^*) = \frac{n}{N-p}\left(\Delta^2 + \frac{p}{N_2} - \frac{bp}{N_1} + p(1-b)\right), \\
s^2 &= 4\frac{n^2N}{(N-p)^3}\left(\Delta^2 + \frac{pb^2}{N_1} + \frac{p}{N_2}\right), \\
\eta_1 &= \left(\frac{1}{N-p} + \frac{n}{(N-p)^2}\right)\left(\Delta^2 + \frac{p}{N_2} - \frac{bp}{N_1} + p(1-b)\right) - \frac{bnp}{(N-p)N_1^2}, \\
s_1 &= 4\frac{Nn^2}{(N-p)^3}\left(\frac{3}{N-p} - \frac{2}{n} - \frac{1}{N}\right)\left(\Delta^2 + \frac{pb^2}{N_1} + \frac{p}{N_2}\right) + 4\frac{pb^2Nn^2}{N_1^2(N-p)^3}.
\end{aligned}$$

である.

4 数値実験

本節では, 前節で紹介したバイアス補正法と [12] で提案されたパラメトリックな推定量 Q_TNW の比較を行う. 分布は正規分布と t 分布 (自由度 3) を用い, 判別手法は Fisher の線形判別を用いる. 一般性を失うことなくパラメータを $\boldsymbol{\mu}_1 = \Delta(1,...,1)'/2\sqrt{p}$, $\boldsymbol{\mu}_2 = -\Delta(1,...,1)'/2\sqrt{p}$, $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ と仮定できる. CV, I, II, III, TNW をそれぞれクロスバリ

デーション, Method I, II, II, Q_{TNW} とする. N_1, N_2, p, Δ を $N_1, N_2 = 15, 20, 25, 30, 35$, $p/N = 1/5, 3/5$, $\Delta = 1.05, 1.68, 2.56, 3.29$ の場合でモンテカルロシミュレーションで比較する. Δ は $\Phi(-\Delta/2)$ がそれぞれ 0.30, 0.20, 0.10, 0.05 となるようになる値である. Methods II, III には Δ の推定量が必要であるが, 本稿では $\hat{\Delta}^2$ を

$$\hat{\Delta}^2 = \frac{n-p-3}{n} D^2 - \frac{pN}{N_1 N_2}.$$

で与える. ここで $D^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ である. $\hat{\Delta}^2$ は Δ^2 の正規分布の場合に不偏推定量で一致推定量である [12].

4.1 正規分布

正規分布の場合は, 図 1, 2 を見てもわかるように, 全ての推定量が 0 に収束していることがわかる. さらに, 第 3 節で提案されたバイアス補正法はどの場合おいても有効であることがわかる. さらに, Q_{TNW} よりもクロスバリデーションを用いた手法の方がバイアス補正していることがわかる. しかしながら, 図 3, 4 を見ると Q_{TNW} が MSE を小さくしている. この結果からクロスバリデーションの推定量は分散が大きくなることがわかる. さらに, Method I は分散をクロスバリデーションより大きくすることがわかる.

4.2 t 分布

[9] では正規分布の場合のみに λ と c_1 の値を決定しているが, 非正規の場合がわかっていない. 本節では非正規の場合の精度を確かめるため t-分布で比較を行う. 図 5, 6 を見ると正規分布の場合で提案されている Q_{TNW} はバイアスが大きくなる場合がある. クロスバリデーションは漸近的に 0 に近づいていることがわかる. また, Method I はバイアス補正ができている. Method II, III は λ と c_1 が十分ではないため, バイアス補正はできているが, Method I ほどではない. しかし, MSE は正規性が崩れていたしても Q_{TNW} が小さいことがわかる.

5 まとめ

本稿はバイアス補正法を紹介し, 正規分布と t-分布 (自由度 3) の場合で数値比較を行った. バイアスだけ見れば, Method I が最も良い推定量であることがわかった. Method I はノンパラメトリックな手法であるため, 分布に関係なく推定できるため, 分散を気に

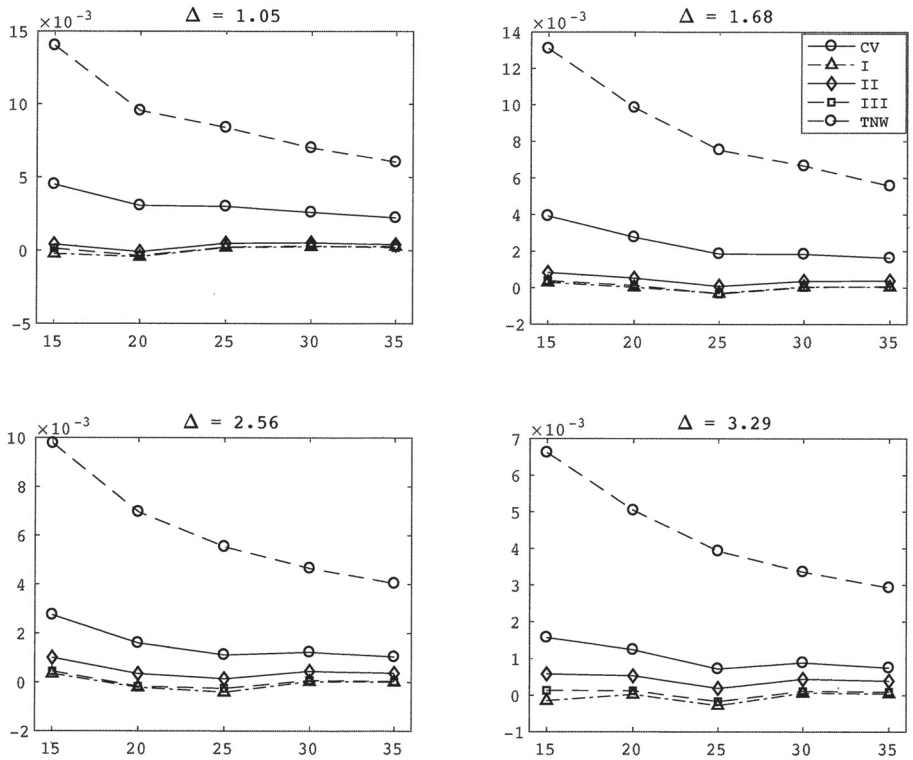

図 1　Bias in the case of $p/N = 1/5$(正規分布)

しなければ推定量としては十分である．しかし, MSE を見ると $Q_{TNW}g$ が良い推定量であった．これは, クロスバリデーションによる推定が分散が大きいことに原因がある．

今後の課題として, クロスバリデーションに変わる分散が小さいノンパラメトリックな手法を考える必要がある．また, クロスバリデーションを用いるには計算負荷も大きいため, 計算負荷を軽減できる手法が必要である．

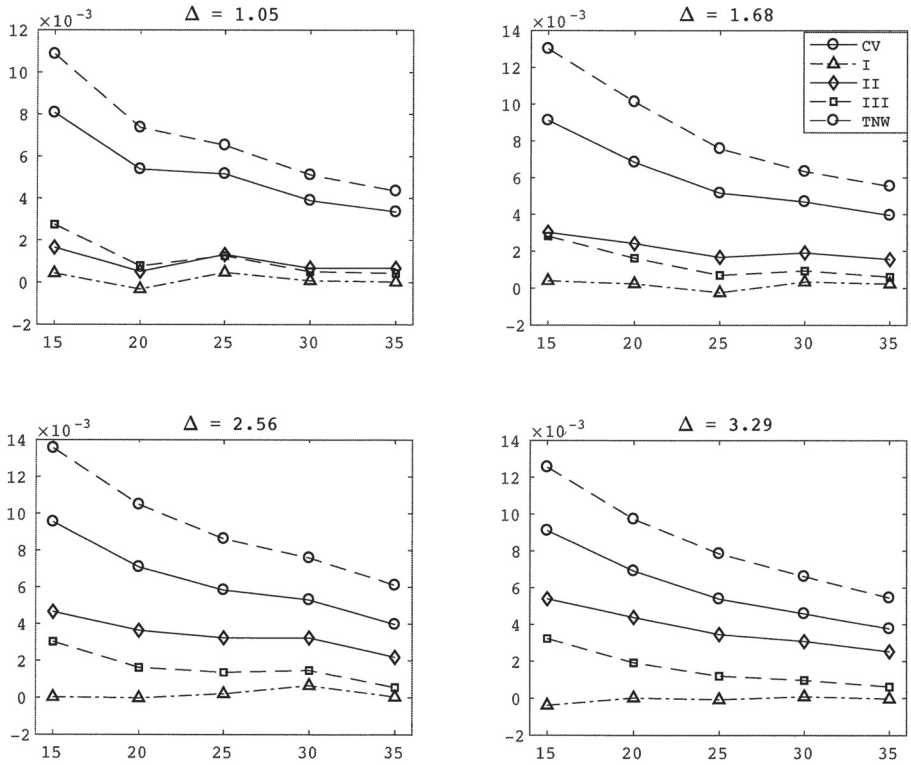

図 2　Bias in the case of $p/N = 3/5$(正規分布)

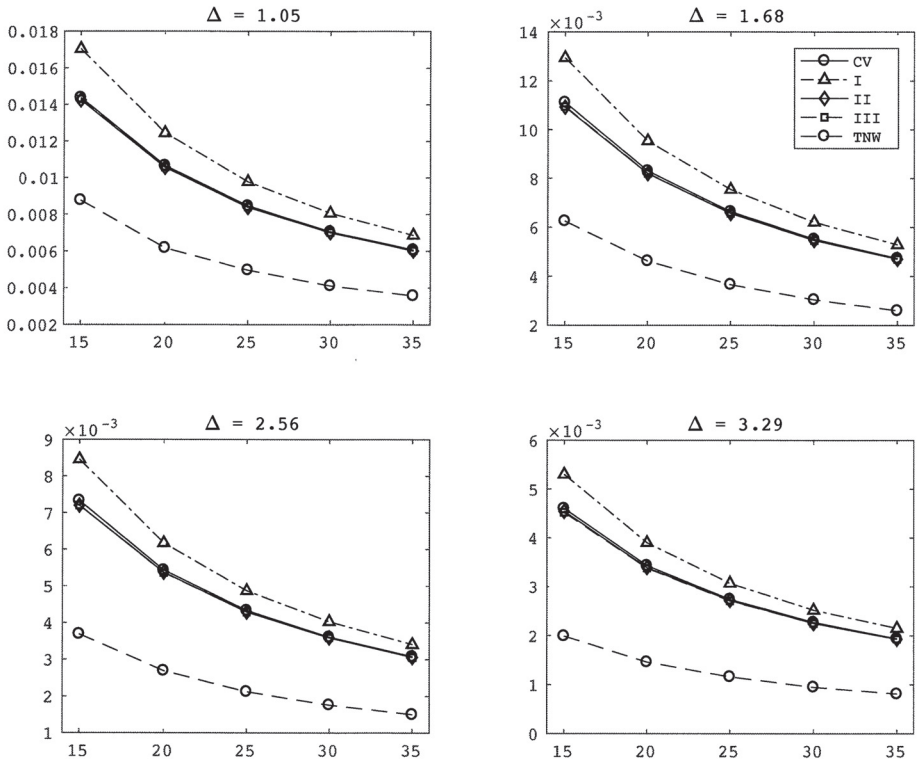

図 3　MSE in the case of $p/N = 1/5$(正規分布)

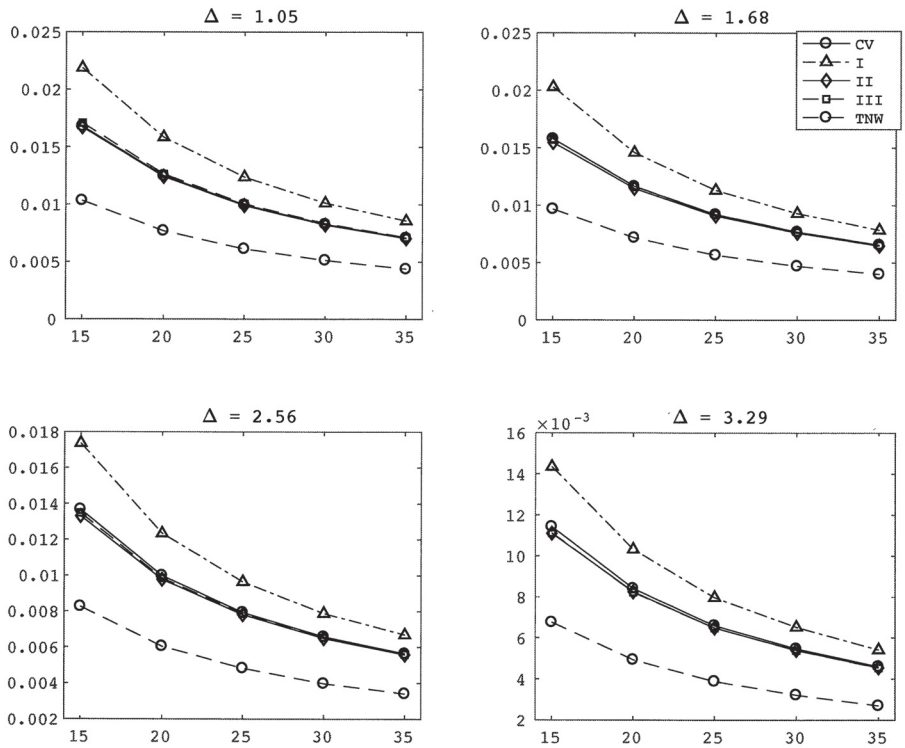

図 4　MSE in the case of $p/N = 3/5$(正規分布)

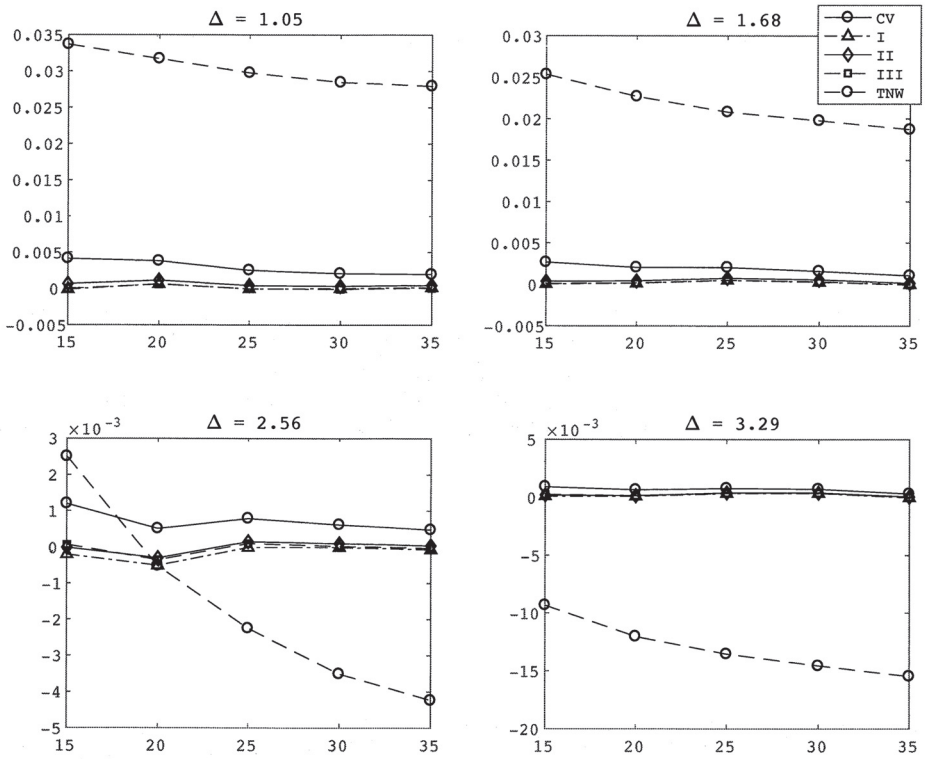

図 5　Bias in the case of $p/N = 1/5$(t 分布)

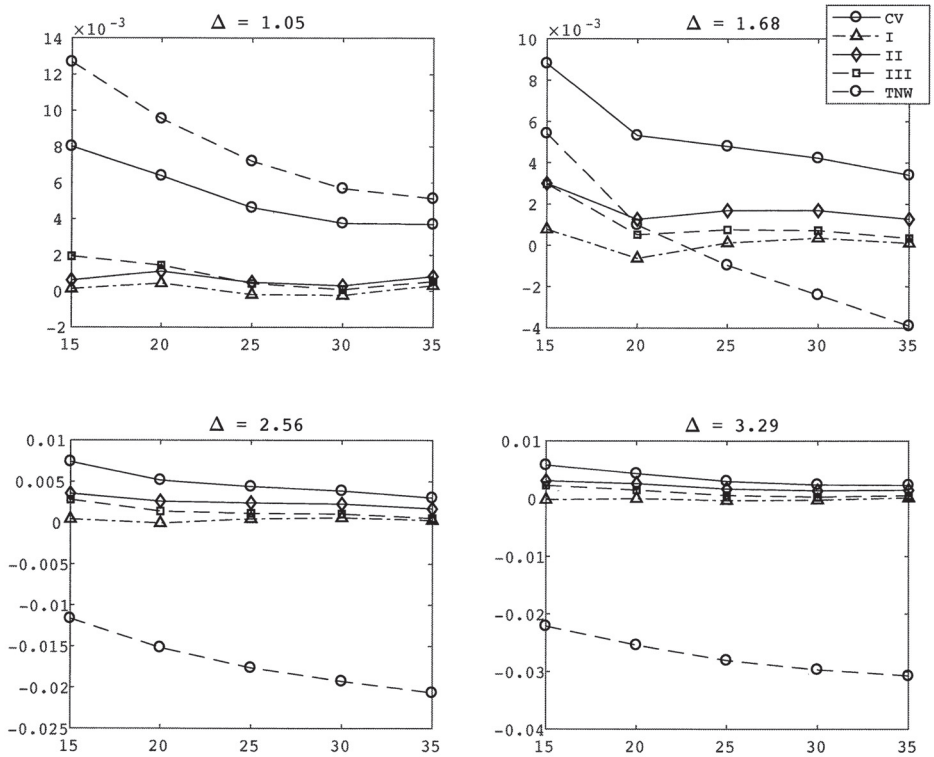

図 6　Bias in the case of $p/N = 3/5$(t 分布)

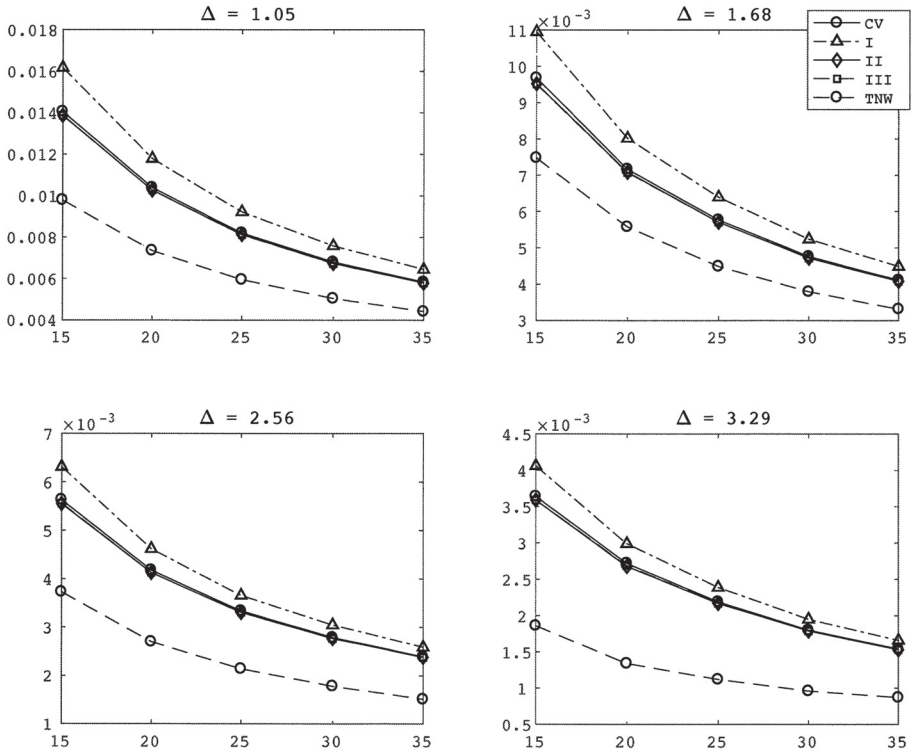

図 7　MSE in the case of $p/N = 1/5$(t 分布)

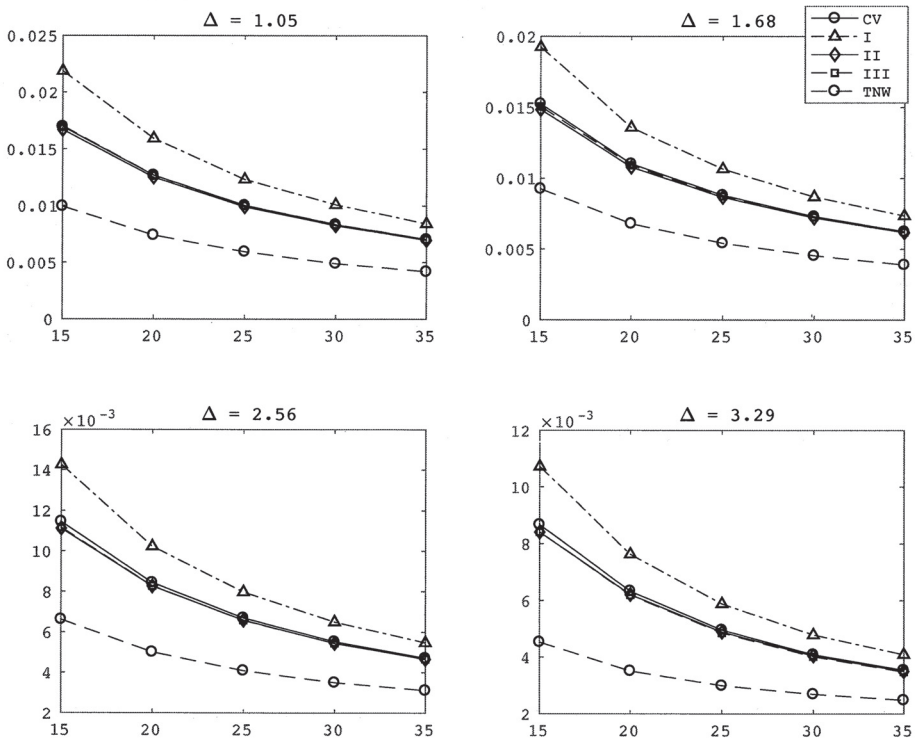

図 8　MSE in the case of $p/N = 3/5$(t 分布)

参考文献

[1] L. Clemmensen, D. Witten, T. Hastie and B. Ersbøll, Sparse discriminant analysis. *Technometrics*, Vol. **53**, No.4, (2011), 406–413.

[2] A. D. Deev, Representation of statistics of discriminant analysis and asymptotic expansions when space dimensions are comparable with sample size. *Soviet Math. Dokl.*, **11** (1970), 1547–1550.

[3] B. Efron, Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Am. Statist. Assoc.*, **78** (1983), 316-331.

[4] B. Efron and R. Tibshirani, Improvement on cross-validation: The .632+ Bootstrap method. *J. Amer. Statist. Assoc.*, **92** (1997), 548–560.

[5] R. A. Fisher, The use of multiple measurements in taxonomic problems. *The Annals of Human Genetics*, **7** (1936), 111–132.

[6] Y. Fujikoshi and T. Seo, Asymptotic aproximations of EPMC's of the linear and the quadratic discriminant functions when the sample sizes and the dimension are large, *Random Oper. Stochastic Equations*, **6** (1998), 269–280.

[7] P. A. Lachenbruch and M. R. Mickey, Estimation of error rates in discriminant analysis. *Technometrics*, **10** (1968), 1–11.

[8] G. J. McLachlan, An asymptotic unbiased technique for estimating the error rates in discriminant analysis. *Biometrics*, **30** (1974), 230–249.

[9] T. Nakagawa, Estimating the probabilities of misclassification using CV when the dimension and the sample sizes are large. *Hiroshima Math. J.*, (2018), in press.

[10] M. Okamoto, An asymptotic expansion for the distribution of the linear discriminant function. *Ann. Math. Statist.*, **34** (1963), 1286–301.

[11] M. Stone, Cross-validatory choice and assessment of statistical predictions. *J. R. Statist. Soc.*, **B36** (1974), 111–147.

[12] T. Tonda, T. Nakagawa and H. Wakaki, EPMC estimation in discriminant analysis when the dimension and sample sizes are large. *Hiroshima Math. J.*, Vol.**47** (2017), No.1, 43–62.

[13] H. Yanagihara and H. Fujisawa, Iterative bias correction of the cross-validation criterion. *Scand. J. Stat.*, Vol.**39** (2012), 116–130.

[14] H. Yanagihara, T. Tonda and C. Matsumoto, Bias correction of cross-validation criterion based on Kullback-Laeibler information under a general condition. *J. Multivariate Anal.*, **97** (2006), 1965–1975.

[15] H. Yanagihara, K.-H. Yuan, H. Fujisawa and K. Hayashi, A class of cross-validatory model selection criteria. *Hiroshima Math. J.*, **43** (2013), 149–177.