Confidence interval of log odds ratio for the posterior probabilities under heteroscedastic multivariate normal groups for large dimension

Takayuki Yamada*
General Studies, College of Engineering,
Nihon University,
1 Nakagawara, Tokusada, Tamuramachi, Koriyama, Fukushima 963-8642, Japan
E-mail address: yma801228@gmail.com

Abstract

This paper is concerned with interval estimation of the log-odds of the posterior probabilities under heteroscedastic multivariate normal groups when the prior probabilities are equal. We treat the case that the dimension is large, but does not exceed sample sizes. We give the limiting distribution of the unbiased estimator for the log-odds as the sample sizes and the dimension tends to infinity together, and give the approximated confidence interval based on the asymptotic distribution. Small scale simulation is performed to check the precision.

Keywords: Log-odds of the posterior probabilities; heteroscedastic multivariate normal groups; (n,p)-asymptotic.

1. Introduction

Posterior probability that $\boldsymbol{x}$ belongs to G_i $(i=1,2)$ is $\tau_i(\boldsymbol{x})=\pi_i f_i(\boldsymbol{x})/f(\boldsymbol{x})$, where $f(\boldsymbol{x})=\sum_{i=1}^{g}\pi_i f_i(\boldsymbol{x})$, $f_i(\boldsymbol{x})$ is the probability density function when $\boldsymbol{x}$ belongs to G_i, and π_i is the prior probability. The log-odds is given by $\eta(\boldsymbol{x})=\log(\tau_1(\boldsymbol{x})/\tau_2(\boldsymbol{x}))$. Consider the case that G_i has multivariate normal distribution $N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$. Let $\Psi_{\mathrm{U}}=\{\pi_1,\pi_2,\Theta_{\mathrm{U}}\}$, $\Theta_{\mathrm{U}}=\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2\}$. The log odds $\eta(\boldsymbol{x})=\eta(\boldsymbol{x};\Psi_{\mathrm{U}})$ can be expressed as $\eta(\boldsymbol{x})=\log(\pi_1/\pi_2)+\xi(\boldsymbol{x})$, where

$$\xi(\boldsymbol{x})=\xi(\boldsymbol{x};\Theta_{\mathrm{U}})=-\frac{1}{2}\{\delta_{1\mathrm{U}}(\boldsymbol{x})-\delta_{2\mathrm{U}}(\boldsymbol{x})\}$$

with

$$\begin{aligned}\delta_{i\mathrm{U}}(\boldsymbol{x})&=\delta(\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)+\log|\boldsymbol{\Sigma}_i|\\&=(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)+\log|\boldsymbol{\Sigma}_i|\quad(i=1,2).\end{aligned}$$

For ease, we treat the case that $\pi_1=\pi_2$. The uniform minimum variance unbiased estimator of $\xi(\boldsymbol{x};\Theta_{\mathrm{U}})$ is given by

$$\widehat{\xi(\boldsymbol{x};\Theta_{\mathrm{U}})}=\xi(\boldsymbol{x};\widehat{\Theta_{\mathrm{U}}})=-\frac{1}{2}\{\widehat{\delta_{1\mathrm{U}}(\boldsymbol{x})}-\widehat{\delta_{2\mathrm{U}}(\boldsymbol{x})}\},$$

where $\widehat{\Theta_{\mathrm{U}}}=\{\bar{\boldsymbol{x}}_1,\bar{\boldsymbol{x}}_2,\boldsymbol{S}_1,\boldsymbol{S}_2\}$,

$$\widehat{\delta_{i\mathrm{U}}(\boldsymbol{x})}=\frac{\delta(\boldsymbol{x};\bar{\boldsymbol{x}}_i,\boldsymbol{S}_i)}{c_1(n_i)}-\frac{p}{N_i}+\log|\boldsymbol{S}_i|+p\log n_i-c_2(n_i),$$

$c_1(n_i)=n_i/(n_i-p-1)$, $c_2(n_i)=p\log 2+\sum_{j=1}^{p}\psi((n_i-p+j)/2)$, $\psi=(d/dy)\log\Gamma(y)$,

$$\bar{\boldsymbol{x}}_i=\frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{x}_{ij},\quad \boldsymbol{S}_i=\frac{1}{n_i}\sum_{j=1}^{N_i}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)',\quad n_i=N_i-1\quad(i=1,2),$$

$$\boldsymbol{x}_{ij}\overset{\text{i.i.d.}}{\sim}N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\quad(j=1,\ldots,N_i,i=1,2).$$

From Lemma 1,

$$\frac{1}{\sqrt{2}\boldsymbol{b}_i'\boldsymbol{b}_i}\frac{\sqrt{m_i}}{p}\left[\frac{\delta_{i\mathrm{U}}(\boldsymbol{x};\bar{\boldsymbol{x}}_i,\boldsymbol{S}_i)}{c_1(n_i)}-\frac{p}{N_i}-p\boldsymbol{b}_i'\boldsymbol{b}_i\right]\xrightarrow{\mathcal{D}}N(0,1).$$

On the other hand, from Lemma 2,

$$\frac{\sqrt{m_i}}{p}(\log|\boldsymbol{S}_i|-\log|\boldsymbol{\Sigma}_i|+p\log n_i-c_2(n_i))\xrightarrow{p}0\quad(i=1,2)$$

under the asymptotic framework A. Using Slutsky's theorem,

$$\frac{\sqrt{m_i}}{p}\left[\widehat{\delta_{i\mathrm{U}}(\boldsymbol{x})}-\delta_{i\mathrm{U}}(\boldsymbol{x})\right]\xrightarrow{\mathcal{D}}N(0,2\lim_{\mathrm{A}}\{\delta(\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)/p\}^2).$$

Proposition 1. *Under the asymptotic framework* A*,*

$$\frac{\sqrt{m}}{p}\left[\xi(\widehat{\boldsymbol{x};\Theta_{\mathrm{U}}})-\xi(\boldsymbol{x};\Theta_{\mathrm{U}})\right]\xrightarrow{\mathcal{D}}N(0,(1/2)\lim_{\mathrm{A}}\delta^*(\boldsymbol{x};\widehat{\Theta_{\mathrm{U}}})/p^2),$$

where $m=m_1+m_2$*,*

$$\delta^*(\boldsymbol{x};\Theta_{\mathrm{U}})=\sum_{i=1}^{2}\frac{m}{m_i}\{\delta(\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\}^2.$$

For actual use, it is needed to estimate the asymptotic variance. It is natural to estimate $\delta(\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ by $\delta(\boldsymbol{x};\bar{\boldsymbol{x}}_i,\boldsymbol{S}_i)$. Under the asymptotic framework A, consistency of $\delta(\boldsymbol{x};\bar{\boldsymbol{x}}_i,\boldsymbol{S}_i)/p$ holds. Thus, we propose confidence interval of $\xi(\boldsymbol{x};\Theta_{\mathrm{U}})$ with confidence level $(1-\alpha)\%$ as

$$\mathrm{CI_p}:\xi(\widehat{\boldsymbol{x};\Theta_{\mathrm{U}}})\pm\sqrt{\frac{1}{2}\sum_{i=1}^{2}\frac{1}{m_i}\{\delta(\boldsymbol{x};\bar{\boldsymbol{x}}_i,\boldsymbol{S}_i)\}^2\cdot z_{1-\alpha/2}}$$

3. Simulation

In order to see the performance of the proposed confidence interval, we do small scale simulation for the actual confidence level based on 10,000 repetitions. Let $\boldsymbol{P}_i$ be an orthogonal matrix whose first column is proportional to $\boldsymbol{\Sigma}_i^{-1/2}(\boldsymbol{\mu}_i-\boldsymbol{x})$. Transforming $\boldsymbol{t}\to\boldsymbol{t}^*=\boldsymbol{P}_i(\boldsymbol{t}-\boldsymbol{x})$,

$$\begin{aligned}&\boldsymbol{x}_{ij}\to\boldsymbol{x}_{ij}^*\sim N_p(\delta(\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\cdot\boldsymbol{e}_1,\boldsymbol{I}_p)\quad(j=1,\ldots,N_i,i=1.2),\\&\boldsymbol{x}\to\boldsymbol{x}^*=\boldsymbol{0}.\end{aligned}$$

We did simulation when $N_1=N_2=50$, $p=10,20,30,40$, $\delta(\boldsymbol{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)=\delta(\boldsymbol{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)=1.68$. In the following table, we gave the actual confidence levels for the nominal level $1-\alpha=0.95$.

Table. Actual confidence levels based on 10,000 repetitions for nominal level 0.95 when $N_1=N_2=50$.

	p			
	10	20	30	40
$\mathrm{CI_{CFR}}$	0.96	1.00	1.00	1.00
$\mathrm{CI_p}$	0.88	0.95	0.97	0.99

We can see from Table that our proposed confidence interval has a good approximation when $p=20,30$. When $p=10$, our proposal underestimates the confidence level, and when $p=40$, our proposal overestimates. Critchley et al. (1987)'s confidence interval overestimates the confidence level for all cases.

4. Proof of lemmas

Proof of Lemma 1. It holds that

$$\frac{m-1}{y} = 1 - \left(\frac{y}{m+1} - 1\right) + \left(\frac{y}{m+1} + \frac{m-1}{y} - 2\right).$$

Consider the probability convergence of

$$\left(\frac{1}{\sqrt{N}}\boldsymbol{z} - \sqrt{p}\boldsymbol{b}\right)' \left(\frac{1}{\sqrt{N}}\boldsymbol{z} - \sqrt{p}\boldsymbol{b}\right) \cdot \left(\frac{y}{m+1} - 1\right) = \left(\frac{1}{N}\boldsymbol{z}'\boldsymbol{z} - 2\sqrt{\frac{p}{N}}\boldsymbol{b}'\boldsymbol{z} + p\boldsymbol{b}'\boldsymbol{b}\right) \cdot \left(\frac{y}{m+1} - 1\right).$$

It can be expressed that

$$E\left[\frac{1}{N}\boldsymbol{z}'\boldsymbol{z}\left(\frac{y}{m+1} - 1\right)\right] = 0, \quad \mathrm{Var}\left[\frac{1}{N}\boldsymbol{z}'\boldsymbol{z}\left(\frac{y}{m+1} - 1\right)\right] = \frac{p(p+2)}{N^2}\frac{2}{m+1},$$
$$E\left[\boldsymbol{b}'\boldsymbol{z}\left(\frac{y}{m+1} - 1\right)\right] = 0, \quad \mathrm{Var}\left[\boldsymbol{b}'\boldsymbol{z}\left(\frac{y}{m+1} - 1\right)\right] = \frac{2}{m+1}\boldsymbol{b}'\boldsymbol{b}.$$

Thus under the asymptotic framework A,

$$\left(\frac{1}{N}\boldsymbol{z}'\boldsymbol{z} - 2\sqrt{\frac{p}{N}}\boldsymbol{b}'\boldsymbol{z} + p\boldsymbol{b}'\boldsymbol{b}\right) \cdot \left(\frac{y}{m+1} - 1\right) - p\boldsymbol{b}'\boldsymbol{b}\left(\frac{y}{m+1} - 1\right) \xrightarrow{p} 0.$$

Next, we focus on the probability convergence of

$$\left(\frac{1}{\sqrt{N}}\boldsymbol{z} - \sqrt{p}\boldsymbol{b}\right)' \left(\frac{1}{\sqrt{N}}\boldsymbol{z} - \sqrt{p}\boldsymbol{b}\right) \cdot \left(\frac{y}{m+1} + \frac{m-1}{y} - 2\right)$$
$$= \left(\frac{1}{N}\boldsymbol{z}'\boldsymbol{z} - 2\sqrt{\frac{p}{N}}\boldsymbol{b}'\boldsymbol{z} + p\boldsymbol{b}'\boldsymbol{b}\right) \cdot \left(\frac{y}{m+1} + \frac{m-1}{y} - 2\right).$$

It holds that

$$E\left[\frac{1}{N}\boldsymbol{z}'\boldsymbol{z}\left(\frac{y}{m+1} + \frac{m-1}{y} - 2\right)\right] = 0, \mathrm{Var}\left[\frac{1}{N}\boldsymbol{z}'\boldsymbol{z}\left(\frac{y}{m+1} + \frac{m-1}{y} - 2\right)\right] = \frac{8p(p+2)}{N^2(m+1)(m-3)},$$
$$E\left[\boldsymbol{b}'\boldsymbol{z}\left(\frac{y}{m+1} + \frac{m-1}{y} - 2\right)\right] = 0, \mathrm{Var}\left[\boldsymbol{b}'\boldsymbol{z}\left(\frac{y}{m+1} + \frac{m-1}{y} - 2\right)\right] = \frac{8}{(m+1)(m-3)}\boldsymbol{b}'\boldsymbol{b}.$$

Thus under the asymptotic framework A,

$$\left(\frac{1}{N}\boldsymbol{z}'\boldsymbol{z} - 2\sqrt{\frac{p}{N}}\boldsymbol{b}'\boldsymbol{z} + p\boldsymbol{b}'\boldsymbol{b}\right) \cdot \left(\frac{y}{m+1} + \frac{m-1}{y} - 2\right) - p\boldsymbol{b}'\boldsymbol{b}\left(\frac{y}{m+1} + \frac{m-1}{y} - 2\right) \xrightarrow{p} 0.$$

Combining these probability convergences, and taking consideration of the equality:

$$-\left(\frac{y}{m+1} - 1\right) + \left(\frac{y}{m+1} + \frac{m-1}{y} - 2\right) = \frac{m-1}{y} - 1,$$

we have that

$$\frac{\left(\frac{1}{\sqrt{N}}\boldsymbol{z} - \sqrt{p}\boldsymbol{b}\right)'\left(\frac{1}{\sqrt{N}}\boldsymbol{z} - \sqrt{p}\boldsymbol{b}\right)}{y/(m-1)} - \left(\frac{1}{\sqrt{N}}\boldsymbol{z} - \sqrt{p}\boldsymbol{b}\right)' \left(\frac{1}{\sqrt{N}}\boldsymbol{z} - \sqrt{p}\boldsymbol{b}\right) - \left(\frac{m-1}{y} - 1\right) \cdot p\boldsymbol{b}'\boldsymbol{b} \xrightarrow{p} 0.$$

Furthermore, it holds that $E[|\boldsymbol{z}'\boldsymbol{z}/(\sqrt{p}N)] = \sqrt{p}/N$, which converges to 0 under the asymptotic framework A. From Markov inequality, $\{1/(\sqrt{p}N)\}\boldsymbol{z}'\boldsymbol{z} \xrightarrow{p} 0$. In addition, since $E[p^{-1/2}\boldsymbol{b}'\boldsymbol{z}] = 0$ and $\mathrm{Var}(p^{-1/2}\boldsymbol{b}'\boldsymbol{z}) = p^{-1}\boldsymbol{b}'\boldsymbol{b}$, Chebyshev inequality leads that $(1/\sqrt{p})\boldsymbol{b}'\boldsymbol{z} \xrightarrow{p} 0$ under the asymptotic framework A. Using Slutsky's theorem,

$$\frac{1}{\sqrt{p}}\left[\left(\frac{1}{\sqrt{N}}\boldsymbol{z} - \sqrt{p}\boldsymbol{b}\right)' \left(\frac{1}{\sqrt{N}}\boldsymbol{z} - \sqrt{p}\boldsymbol{b}\right) - \frac{p}{N} - p\boldsymbol{b}'\boldsymbol{b}\right] \xrightarrow{p} 0.$$

On the other hand, since it holds that $E[|\sqrt{m}/y|] = \sqrt{m}/(m-1) \to 0$ under the asymptotic framework A, from Markov inequality, $\sqrt{m}/y \xrightarrow{p} 0$. In addition, $\sqrt{m}(y/m-1) \xrightarrow{\mathcal{D}} N(0,2)$. From delta methods and Slutsly theorem, $\sqrt{m}\{(m-1)/y-1\} \xrightarrow{p} N(0,2)$. Combining these results,

$$\frac{1}{\sqrt{2}\boldsymbol{b}'\boldsymbol{b}}\frac{\sqrt{m}}{p}\left[\frac{\left(\frac{1}{\sqrt{N}}\boldsymbol{z}-\sqrt{p}\boldsymbol{b}\right)'\left(\frac{1}{\sqrt{N}}\boldsymbol{z}-\sqrt{p}\boldsymbol{b}\right)}{y/(m-1)}-\frac{p}{N}-p\boldsymbol{b}'\boldsymbol{b}\right] \xrightarrow{\mathcal{D}} N(0,1).$$

□

Before proving Lemma 2, we give a result concerning the boundaries of series.

Lemma 3. *Let $f(x)$ be a non-negative decreasing function. Then, for positive constant $a > 1$,*

$$\lim_{n\to\infty}\int_1^{n+1} f(x+a)dx < \sum_{k=1}^{\infty} f(k+a) < \lim_{n\to\infty}\int_1^{n+1} f(x+a-1)dx$$

Proof. Since $f(x)$ is a decreasing function, for positive integer ℓ, it holds that

$$\int_\ell^{\ell+1} f(x)dx > \{(\ell+1)-\ell\}\cdot f(\ell+1) = f(\ell+1),$$
$$\int_\ell^{\ell+1} f(x)dx < \{(\ell+1)-\ell\}\cdot f(\ell) = f(\ell),$$

and so

$$\sum_{k=1}^{n} f(k+a) < \int_1^{n+1} f(x+a-1)dx < \sum_{k=1}^{n} f(k+a-1).$$

□

Proof of Lemma 2. Characteristic function of $V = \log|\boldsymbol{W}|$ is given as

$$C(t) = \frac{\Gamma_p(\frac{n}{2}+it)}{\Gamma_p(\frac{n_1}{2})},$$

where $\Gamma_p(a/2) = \pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma((a-i+1)/2)$. The cumulant generating function $K(t) = \log C(t)$ can be expressed as

$$K(t) = \sum_{j=1}^{p}\left\{\log\Gamma\left(\frac{n-p+j}{2}+it\right)-\log\Gamma\left(\frac{n-p+j}{2}\right)\right\}.$$

Using Taylor expansion of $\log\Gamma((n-p+j)/2+x)$ at $x=0$, it formally can be expanded as

$$K(t) = \sum_{s=1}^{\infty}\frac{\kappa^{(s)}}{s!}$$

with s-th cumulant $\kappa^{(s)}$, which is given as follows.

$$\kappa^{(s)} = \sum_{j=1}^{p}\psi^{(s-1)}\left(\frac{n-p+j}{2}\right),$$

where

$$\psi^{(s)}(a) = \begin{cases} -C+\displaystyle\sum_{k=0}^{\infty}\left(\frac{1}{1+k}-\frac{1}{k+a}\right) & (s=0), \\ \displaystyle\sum_{k=0}^{\infty}\frac{(-1)^{s+1}s!}{(k+a)^{s+1}} & (s\geq 1). \end{cases}$$

Here, C denotes Euler's constant. For $\mathrm{Var}(V) = \kappa^{(2)}$, it is found from Lemma 3 that

$$\sum_{j=1}^{p} \lim_{k\to\infty} \int_1^{k+1} \frac{1}{(x+a_j-1)^2} dx < \mathrm{Var}(V) < \sum_{j=1}^{p} \lim_{k\to\infty} \int_1^{k+1} \frac{1}{(x+a_j-2)^2} dx,$$

where $a_j = (n-p+j)/2$, and so

$$\sum_{j=1}^{p} \frac{2}{n-p+j} < \mathrm{Var}(V) < \sum_{j=1}^{p} \frac{2}{n-p+j-2}. \tag{1}$$

For the left-hand side of the inequality, using Lemma 3 again,

$$\sum_{j=1}^{p} \frac{2}{n-p+j} > 2\int_1^{p+1} \frac{1}{n-p+x} dx = 2\log\left(1+\frac{p}{n-p+1}\right).$$

On the other hand, the right-hand side of the inequality (1) is bounded by

$$\sum_{j=1}^{p} \frac{2}{n-p+1-2} = \frac{2p}{n-p-1}.$$

Thus $\mathrm{Var}(V)$ converges a positive constant as $n, p \to \infty$, $p/n \to c \in (0,1)$. The lemma follows from Chebyshev inequality. □

References

Critchley, F., Ford, I. and Rijal, O. (1987). Uncertainly in discrimination. *Proc. Conf. DIANA II*. Prague: Math. Inst. of the Czechoslovak Academy of Sciences, pp. 83–106.

Fujikoshi, Y., Ulyanov, V.V. and Shimizu, R. (2010). *Multivariate Statistics High-Dimensional and Large-Sample Approximations*, Wiley, Hoboken, NJ.