Behaviormetrika
Vol.41, No.2, 2014, 195–223

OPTIMIZATION OF THE GAUSSIAN AND JEFFREYS POWER PRIORS WITH EMPHASIS ON THE CANONICAL PARAMETERS IN THE EXPONENTIAL FAMILY

Haruhiko Ogasawara*

Optimal powers of the Gaussian and Jeffreys priors are obtained so that they minimize the asymptotic mean square error of the linear predictor and the sum of the asymptotic mean square errors of associated parameter estimators. Conditions that the summarized mean square errors using powers of the priors are smaller than those by maximum likelihood are given. In the case of a scalar canonical parameter in the exponential family, a matching prior for the Jeffreys power prior is found, where the Wald confidence interval has second-order accurate coverage. The results are numerically illustrated using the categorical distribution and logistic regression.

1. Introduction

The Jeffreys (1946; 1961, Section 3.10) prior has been used as a noninformative prior in statistical inference with reasonable results in many cases. Its properties have been provided by Kass (1989), Kass and Wasserman (1996), Kass and Vos (1997, Section 7.2.3), and Amari and Nagaoka (2000, p.44) from a geometric viewpoint. It is known that the Jeffreys prior can yield improper priors. Ibrahim and Laud (1991) gave conditions for the improper propriety of the posterior and prior distributions in generalized linear models. Chen, Ibrahim and Kim (2008) investigated the tail property of the Jeffreys prior in typical binomial models.

Firth (1993) gave a weighted score (penalized likelihood) method that gives an estimator asymptotically unbiased up to order $O(n^{-1})$, where n is the sample size. In the case of a canonical (vector) parameter in the exponential family, the estimator by Firth's method becomes algebraically equal to the Bayes modal estimator (posterior mode) using the Jeffreys prior. In other words, the Jeffreys prior gives an asymptotically unbiased estimator in this case. Further, fortunately in many cases of interest, the Jeffreys prior reduces the higher-order asymptotic variance up to order $O(n^{-2})$ (Ogasawara, 2014). That is, the Jeffreys prior in many cases gives asymptotically smaller variance and squared bias than the maximum likelihood estimator (MLE). Ogasawara (2013, Result 1, Table 1) gave a condition for the relative size of the higher-order variances up to order $O(n^{-2})$ of the Jeffreys modal estimator and the

Key Words and Phrases: Gaussian prior; Jeffreys prior; weighted score; penalized likelihood; mean square error; asymptotic cumulants; Cornish-Fisher expansion.

* Otaru University of Commerce
Mail Address: hogasa@res.otaru-uc.ac.jp

This work was partially supported by a Grant-in-Aid for Scientific Research from the Japanese Ministry of Education, Culture, Sports, Science and Technology, No.26330031. Author's address: Department of Information and Management Science, Otaru University of Commerce, 3–5–21, Midori, Otaru 047–8501, Japan.

MLE, and a similar condition for the asymptotic mean square errors (AMSEs) in the case of a scalar canonical parameter.

Ogasawara (2014) derived a general method of bias adjustment minimizing the MSE up to order $O(n^{-2})$ of a parameter estimator by multiplying the weight in the weighted score method by an optimal constant. Note that the weight is a log prior derivative in the case of Bayesian estimation. That is, the constant is an optimal power of the prior.

Powers of priors have been used in different ways. A power of the Gaussian prior corresponds to the precision parameter or the reciprocal of variance, which is also known as a ridge parameter in, for example, logistic regression (le Cessie & van Houwelingen, 1992, Section 2). An optimal value of the ridge parameter has been typically obtained by cross-validation and Akaike's (1973) information criterion (AIC). Zellner's (1986) g in his g-prior for the multiple linear regression model can be seen as a Gaussian power in a conceptual sample of the same size as a real sample. An optimal value of g can be found as the ratio of the sum of the variances to that of the squared biases for parameter estimators as in Ogasawara (2014). Though Zellner (1986) did not give the optimal value, the value with the smallest MSE is consistent with Zellner's (1986, Table 15.1) numerical illustration.

Special cases of powers of the Jeffreys prior, hereafter called as the Jeffreys power priors, are available. Rubin and Schenker (1987) used pseudocounts, in logistic regression, whose sizes in cells are proportional to the number of covariates and inversely proportional to the number of distinct values in covariates. The latter value becomes n when all observations have different sets of values for covariates. Note that pseudocounts in binomial regression are given by the Jeffreys prior (see e.g., Ogasawara, 2014).

Poirier's (1994) neutral prior is the Dirichlet prior in a multinomial logit with J categories using "a fictitious prior sample of size $\underline{r}N$" (p.330) as in Zellner (1986), where N is the usual sample size. When $\underline{r} = J/(2N)$, the prior becomes the Jeffreys prior. The so-called power prior in Ibrahim and Chen (2000, Equation (2.1)) is a power of the likelihood of historical data, where the power takes values between 0 and 1. The prior distribution also has the "initial" or usual prior which is multiplied by the power prior. The power in the power prior can be stochastic with a hierarchical structure (Ibrahim and Chen, 2000, Equation (2.2)), where a conjugate second-stage prior is beta. Ibrahim, Chen and Sinha (2003, p.208) explained the power as that corresponding to the precision parameter in the Gaussian prior.

A purpose of this paper is to derive optimal values of powers of the Gaussian and Jeffreys priors in the case of multiple parameters with appropriate definitions of the summarized AMSE as an extension of Ogasawara's (2014) results for the AMSE of an estimator. Applications are given for the canonical parameters in the categorical distribution and penalized logistic regression. Another purpose is to derive an optimal value for the power in interval estimation of a scalar canonical parameter, where the matching Jeffreys power prior gives a Wald confidence interval with second-order accurate coverage.

2. Asymptotic cumulants of the estimators by the weighted score

Let $\boldsymbol{\theta}$ be a $q \times 1$ parameter vector in a model with $\boldsymbol{\theta}_0$ being a population value. Define l as a log likelihood of $\boldsymbol{\theta}$ based on n independent observations denoted by $r \times 1$ vectors $\mathbf{u}_i$ $(i = 1, \ldots, n)$ with $\mathbf{U} \equiv (\mathbf{u}_1, \ldots, \mathbf{u}_n)'$ that are given by observable variables $\mathbf{U}_i$ $(i = 1, \ldots, n)$. Let $\bar{l} = n^{-1}l$. When there are p fixed covariates, they are denoted by $p \times 1$ vectors $\mathbf{x}_i$ $(i = 1, \ldots, n)$ with $\mathbf{X} \equiv (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$. Define $\mathbf{q}^* = \mathbf{q}^*(\boldsymbol{\theta})$ as a $q \times 1$ vector of the weight in the weighted score or penalized likelihood. In the case of Bayesian estimation, $\mathbf{q}^*$ is a vector of log prior derivatives. The estimator $\hat{\boldsymbol{\theta}}_{\mathrm{W}}$ by the weighted score is given by the solution of

$$\left.\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\mathrm{W}}} + n^{-1}\mathbf{q}^*(\hat{\boldsymbol{\theta}}_{\mathrm{W}}) \equiv \frac{\partial \bar{l}}{\partial \hat{\boldsymbol{\theta}}_{\mathrm{W}}} + n^{-1}\hat{\boldsymbol{q}}^*_{\mathrm{W}} = 0. \tag{2.1}$$

Let $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ be the MLE satisfying $\partial \bar{l}/\partial \hat{\boldsymbol{\theta}}_{\mathrm{ML}} = 0$. Then, under regularity conditions $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is symbolically expanded as

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_0 = \sum_{i=1}^{3} \boldsymbol{\Lambda}^{(i)}\mathbf{l}_0^{(i)} + O_p(n^{-2}), \tag{2.2}$$

where $\boldsymbol{\Lambda}^{(i)} = O(1)$ is given by $\boldsymbol{\theta}_0$, $\mathbf{l}_0^{(i)} = O_p(n^{-i/2})$ $(i = 1, 2, 3)$, and $\mathbf{l}_0^{(i)}$ includes $\boldsymbol{\theta}_0$ and $\mathbf{U}$. The actual expressions of $\boldsymbol{\Lambda}^{(i)}$ and $\mathbf{l}_0^{(i)}$, are given as

$$\begin{aligned}
\boldsymbol{\Lambda}^{(1)}\mathbf{l}_0^{(1)} &= -\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}, \\
\boldsymbol{\Lambda}^{(2)}\mathbf{l}_0^{(2)} &= \boldsymbol{\Lambda}^{-1}\mathbf{M}\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0} - \frac{1}{2}\boldsymbol{\Lambda}^{-1}\mathrm{E}_{\mathrm{T}}(\mathbf{J}_0^{(3)})\left(\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}\right)^{<2>} \\
\boldsymbol{\Lambda}^{(3)}\mathbf{l}_0^{(3)} &= -\boldsymbol{\Lambda}^{-1}\mathbf{M}\boldsymbol{\Lambda}^{-1}\mathbf{M}\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0} + \frac{1}{2}\boldsymbol{\Lambda}^{-1}\mathbf{M}\boldsymbol{\Lambda}^{-1}\mathrm{E}_{\mathrm{T}}(\mathbf{J}_0^{(3)})\left(\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}\right)^{<2>} \\
&\quad + \boldsymbol{\Lambda}^{-1}\mathrm{E}_{\mathrm{T}}(\mathbf{J}_0^{(3)})\left\{\left(\boldsymbol{\Lambda}^{-1}\mathbf{M}\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}\right) \otimes \left(\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}\right)\right\} \\
&\quad - \frac{1}{2}\boldsymbol{\Lambda}^{-1}\{\mathbf{J}_0^{(3)} - \mathrm{E}_{\mathrm{T}}(\mathbf{J}_0^{(3)})\}\left(\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}\right)^{<2>} \\
&\quad - \frac{1}{2}\boldsymbol{\Lambda}^{-1}\mathrm{E}_{\mathrm{T}}(\mathbf{J}_0^{(3)})\left[\left(\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}\right) \otimes \left\{\boldsymbol{\Lambda}^{-1}\mathrm{E}_{\mathrm{T}}(\mathbf{J}_0^{(3)})\left(\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}\right)^{<2>}\right\}\right] \\
&\quad + \frac{1}{6}\boldsymbol{\Lambda}^{-1}\mathrm{E}_{\mathrm{T}}(\mathbf{J}_0^{(4)})\left(\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}\right)^{<3>},
\end{aligned} \tag{2.3}$$

which were given by Ogasawara (2010, Equation (2.4)). In (2.3), $\boldsymbol{\Lambda} = \mathrm{E}_{\mathrm{T}}\left(\left.\frac{\partial^2 \bar{l}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) \equiv \mathrm{E}_{\mathrm{T}}\left(\frac{\partial^2 \bar{l}}{\partial\boldsymbol{\theta}_0\partial\boldsymbol{\theta}_0'}\right)$ and $\mathrm{E}_{\mathrm{T}}(\cdot)$ indicates a true expectation in terms of the variables $\mathbf{U}_i$ $(i = 1, \ldots, n)$ under possible model misspecification; under

a correct model, $\mathrm{E_T}(\cdot)$ is denoted by $\mathrm{E}_\theta(\cdot)$ and $\mathbf{\Lambda} = -\mathbf{I}_0 = \mathrm{E}_\theta\left(\frac{\partial^2 \bar{l}}{\partial\boldsymbol{\theta}_0\partial\boldsymbol{\theta}_0'}\right)$, where $\mathbf{I}_0$ is the information matrix $\mathbf{I}$ for $\boldsymbol{\theta}$ per observation evaluated at $\boldsymbol{\theta}_0$;
$\mathbf{M} \equiv \mathbf{L}_0 - \mathbf{\Lambda} \equiv \mathbf{L}(\boldsymbol{\theta}_0, \mathbf{U}) - \mathbf{\Lambda} = O_p(n^{-1/2})$; $(\cdot)_{O_p(n^{-i/2})}$ indicates that the value is of order $O_p(n^{-i/2})$; $\otimes$ denotes the Kronecker product; $\mathbf{x}^{<k>} = \mathbf{x}\otimes\cdots\otimes\mathbf{x}$ (k times of $\mathbf{x}$) is the k-fold Kronecker product of $\mathbf{x}$; and $\mathbf{J}_0^{(i)} \equiv \frac{\partial^i \bar{l}}{\partial\boldsymbol{\theta}_0(\partial\boldsymbol{\theta}_0')^{<i-1>}}(i = 3, 4)$.

Using (2.2) and (2.3), it is known that $\hat{\boldsymbol{\theta}}_\mathrm{W}$ is expanded as

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_\mathrm{W} - \boldsymbol{\theta}_0 &= -n^{-1}\mathbf{\Lambda}^{-1}\mathbf{q}_0^* + \sum_{i=1}^{3}\mathbf{\Lambda}^{(i)}\mathbf{l}_0^{(i)} - n^{-1}(\hat{\mathbf{L}}^{-1}\hat{\mathbf{q}}_\mathrm{W}^* - \mathbf{\Lambda}^{-1}\mathbf{q}_0^*)_{O_p(n^{-1/2})} + O_p(n^{-2}) \\
&= -n^{-1}\mathbf{\Lambda}^{-1}\mathbf{q}_0^* + \sum_{i=1}^{3}\mathbf{\Lambda}^{(i)}\mathbf{l}_0^{(i)} + n^{-1}\left[\mathbf{\Lambda}^{-1}\mathbf{M}\mathbf{\Lambda}^{-1}\mathbf{q}_0^* - \mathbf{\Lambda}^{-1}\frac{\partial\mathbf{q}_0^*}{\partial\boldsymbol{\theta}_0'}\mathbf{\Lambda}^{(1)}\mathbf{l}_0^{(1)}\right. \\
&\quad \left. - \mathbf{\Lambda}^{-1}\,\mathrm{E_T}\left(\frac{\partial^3\bar{l}}{\partial\boldsymbol{\theta}_0(\partial\boldsymbol{\theta}_0')^{<2>}}\right)\{(\mathbf{\Lambda}^{-1}\mathbf{q}_0^*)\otimes\mathbf{\Lambda}^{-1}\}\mathbf{l}_0^{(1)}\right] + O_p(n^{-2}) \\
&\equiv -n^{-1}\mathbf{\Lambda}^{-1}\mathbf{q}_0^* + \sum_{i=1}^{3}\mathbf{\Lambda}^{(i)}\mathbf{l}_0^{(i)} + n^{-1}\mathbf{L}^{(\mathrm{W})}_{O_p(n^{-1/2})} + O_p(n^{-2}). \qquad (2.4)
\end{aligned}$$

(Ogasawara, 2014, Equation (5.8)), where $\mathbf{q}_0^* = \mathbf{q}^*(\boldsymbol{\theta}_0)$ and $\hat{\mathbf{L}} \equiv \mathbf{L}(\hat{\boldsymbol{\theta}}_\mathrm{W}, \mathbf{U}) \equiv \frac{\partial^2\bar{l}}{\partial\hat{\boldsymbol{\theta}}_\mathrm{W}\partial\hat{\boldsymbol{\theta}}_\mathrm{W}'} = O_p(1)$.

It is known that in the case of the exponential family with canonical parametrization, $\hat{\mathbf{L}}$ does not include $\mathbf{U}$, and consequently written as $\mathbf{L}(\hat{\boldsymbol{\theta}}_\mathrm{W})$ rather than $\mathbf{L}(\hat{\boldsymbol{\theta}}_\mathrm{W}, \mathbf{U})$, and that $\mathbf{M}$ vanishes with $\mathbf{L}_0 = \mathbf{L}(\boldsymbol{\theta}_0) = \mathbf{\Lambda}$ irrespective of model misspecification. A population parameter vector $\boldsymbol{\theta}_0$ under model misspecification is defined as a solution of $\mathrm{E_T}(\partial\bar{l}/\partial\boldsymbol{\theta}_0) = 0$, which yields $\mathrm{E_T}(\mathbf{l}_0^{(1)}) = 0$ as $\mathrm{E}_\theta(\mathbf{l}_0^{(1)}) = 0$.

Denote the i-th cumulant of a variable by $\kappa_i(\cdot)$ under possible model misspecification unless otherwise stated. Then, from (2.4), we have

$$\begin{aligned}
\kappa_1(\hat{\boldsymbol{\theta}}_\mathrm{W} - \boldsymbol{\theta}_0) &= n^{-1}(\boldsymbol{\alpha}_{\mathrm{ML}1} - \mathbf{\Lambda}^{-1}\mathbf{q}_0^*) + O(n^{-2}) \equiv n^{-1}\boldsymbol{\alpha}_{\mathrm{W}1} + O(n^{-2}), \\
\kappa_2(\hat{\boldsymbol{\theta}}_\mathrm{W}) &= n^{-1}\mathbf{A}_{\mathrm{ML}2} + n^{-2}\left\{\mathbf{A}_{\mathrm{ML}\Delta 2} + n\mathrm{E_T}\left(\mathbf{L}^{(\mathrm{W})}\mathbf{l}_0^{(1)'}\mathbf{\Lambda}^{(1)} + \mathbf{\Lambda}^{(1)}\mathbf{l}_0^{(1)}\mathbf{L}^{(\mathrm{W})'}\right)\right\} \\
&\quad + O(n^{-3}) \\
&\equiv n^{-1}\mathbf{A}_{\mathrm{ML}2} + n^{-2}(\mathbf{A}_{\mathrm{ML}\Delta 2} + \mathbf{A}_{\mathrm{W}\,|\,\mathrm{ML}\Delta 2}) + O(n^{-3}) \\
&\equiv n^{-1}\mathbf{A}_{\mathrm{ML}2} + n^{-2}\mathbf{A}_{\mathrm{W}\Delta 2} + O(n^{-3}) \ (\mathbf{A}_{\mathrm{W}2} = \mathbf{A}_{\mathrm{ML}2}), \qquad (2.5)
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{A}_{\mathrm{ML}2} &= n\mathrm{E_T}(\mathbf{\Lambda}^{(1)}\mathbf{l}_0^{(1)}\mathbf{l}_0^{(1)'}\mathbf{\Lambda}^{(1)}) = \mathbf{\Lambda}^{-1}n\mathrm{E_T}(\mathbf{l}_0^{(1)}\mathbf{l}_0^{(1)'})\mathbf{\Lambda}^{-1}, \\
\mathbf{A}_{\mathrm{ML}\Delta 2} &= n^2\mathrm{E_T}(\mathbf{\Lambda}^{(1)}\mathbf{l}_0^{(1)}\mathbf{l}_0^{(2)'}\mathbf{\Lambda}^{(2)'} + \mathbf{\Lambda}^{(2)}\mathbf{l}_0^{(2)}\mathbf{l}_0^{(1)'}\mathbf{\Lambda}^{(1)} + \mathbf{\Lambda}^{(2)}\mathbf{l}_0^{(2)}\mathbf{l}_0^{(2)'}\mathbf{\Lambda}^{(2)'} \\
&\quad + \mathbf{\Lambda}^{(1)}\mathbf{l}_0^{(1)}\mathbf{l}_0^{(3)'}\mathbf{\Lambda}^{(3)'} + \mathbf{\Lambda}^{(3)}\mathbf{l}_0^{(3)}\mathbf{l}_0^{(1)'}\mathbf{\Lambda}^{(1)}) - \boldsymbol{\alpha}_{\mathrm{ML}1}\boldsymbol{\alpha}_{\mathrm{ML}1}', \\
n\mathrm{E_T}(\mathbf{L}^{(\mathrm{W})}\mathbf{l}_0^{(1)'}\mathbf{\Lambda}^{(1)}) &= -\mathbf{\Lambda}^{-1}n\mathrm{E_T}(\mathbf{M}\mathbf{\Lambda}^{-1}\mathbf{q}_0^*\mathbf{l}_0^{(1)'})\mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1}\frac{\partial\mathbf{q}_0^*}{\partial\boldsymbol{\theta}_0'}\mathbf{A}_{\mathrm{ML}2} \\
&\quad + \mathbf{\Lambda}^{-1}\mathrm{E_T}(\mathbf{J}_0^{(3)})\{(\mathbf{\Lambda}^{-1}\mathbf{q}_0^*)\otimes\mathbf{A}_{\mathrm{ML}2}\},
\end{aligned}$$

$$nE_T(\mathbf{M}\boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*\mathbf{l}_0^{(1)'})_{ij} = \sum_{k^*=1}^{q} nE_T\{m_{ik^*}(\mathbf{l}_0^{(1)})_j\}(\boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*)_{k^*} \ (i,j=1,\ldots,q),$$

$n^{-1}\boldsymbol{\alpha}_{\mathrm{ML}1}$ is a $q\times 1$ vector of the asymptotic bias of order $O(n^{-1})$ for $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$, $n^{-1}\mathbf{A}_{\mathrm{ML}2}$ is the usual asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$, and $(\cdot)_{ij}$ and $(\cdot)_j$ denote the (i,j)th element of a matrix and the j-th element of a vector, respectively.

Let θ be an element of $\boldsymbol{\theta}$ with $\hat{\theta}_{\mathrm{ML}}$, $\hat{\theta}_{\mathrm{W}}$ and θ_0 defined similarly. Then, it is known that

$$\begin{aligned}\kappa_3(\hat{\theta}_{\mathrm{W}}) &= \kappa_3(\hat{\theta}_{\mathrm{ML}}) + O(n^{-3}) \equiv n^{-2}\alpha_{\mathrm{ML}3} + O(n^{-3}) \ (\alpha_{\mathrm{W}3} = \alpha_{\mathrm{ML}3}),\\ \kappa_4(\hat{\theta}_{\mathrm{W}}) &= \kappa_4(\hat{\theta}_{\mathrm{ML}}) + O(n^{-4}) \equiv n^{-3}\alpha_{\mathrm{ML}4} + O(n^{-4}) \ (\alpha_{\mathrm{W}4} = \alpha_{\mathrm{ML}4})\end{aligned} \tag{2.6}$$

(see Ogasawara, 2014, Equation (4.1), Theorem 4). Actual expressions of $\boldsymbol{\alpha}_{\mathrm{ML}1}$, $\mathbf{A}_{\mathrm{ML}2}$, $\mathbf{A}_{\mathrm{ML}\Delta 2}$, $\alpha_{\mathrm{ML}3}$ and $\alpha_{\mathrm{ML}4}$ under possible model misspecification for a canonical vector parameter in the exponential family are given in Subsection A.1 of the appendix (for the results under a correct model, see Ogasawara, 2013, Section 3).

Under the same condition, $\mathbf{A}_{\mathrm{W}\Delta 2}$ in $\kappa_2(\hat{\boldsymbol{\theta}}_{\mathrm{W}})$ of (2.5) becomes

$$\begin{aligned}\mathbf{A}_{\mathrm{W}\Delta 2} &= \mathbf{A}_{\mathrm{ML}\Delta 2} + n\,\mathrm{acov}_T(\hat{\mathbf{I}}_{\mathrm{W}}^{-1}\hat{\mathbf{q}}_{\mathrm{W}}^*, \hat{\boldsymbol{\theta}}_{\mathrm{W}}') + n\,\mathrm{acov}_T(\hat{\boldsymbol{\theta}}_{\mathrm{W}}, \hat{\mathbf{q}}_{\mathrm{W}}^{*'}\hat{\mathbf{I}}_{\mathrm{W}}^{-1})\\ &= \mathbf{A}_{\mathrm{ML}\Delta 2} + n\,\mathrm{acov}_T(\hat{\mathbf{I}}_{\mathrm{ML}}^{-1}\hat{\mathbf{q}}_{\mathrm{ML}}^*, \hat{\boldsymbol{\theta}}_{\mathrm{ML}}') + n\,\mathrm{acov}_T(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}, \hat{\mathbf{q}}_{\mathrm{ML}}^{*'}\hat{\mathbf{I}}_{\mathrm{ML}}^{-1})\\ &= \mathbf{A}_{\mathrm{ML}\Delta 2} + \frac{\partial \mathbf{I}_0^{-1}\mathbf{q}_0^*}{\partial \boldsymbol{\theta}_0'}\mathbf{A}_{\mathrm{ML}2} + \mathbf{A}_{\mathrm{ML}2}\frac{\partial \mathbf{q}_0^{*'}\mathbf{I}_0^{-1}}{\partial \boldsymbol{\theta}_0},\end{aligned} \tag{2.7}$$

where $\hat{\mathbf{I}}_{\mathrm{W}}$ and $\hat{\mathbf{I}}_{\mathrm{ML}}$ are $\mathbf{I}(\hat{\boldsymbol{\theta}}_{\mathrm{W}})$ and $\mathbf{I}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})$, respectively, and $\mathrm{acov}_T(\cdot)$ denotes an asymptotic covariance of two variables up to order $O(n^{-1})$ under possible model misspecification.

3. Optimal powers in the power priors minimizing the AMSE

Consider $k\mathbf{q}^*$, where k is a fixed constant. Define $\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)}$ as $\hat{\boldsymbol{\theta}}_{\mathrm{W}}$ when $k\mathbf{q}^*$ is used in place of $\mathbf{q}^*$. Then, $\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)}$ is the Bayes modal estimator using the k-th power of a prior when Bayesian estimation is used. It follows from (2.5) and (2.6) that

$$\begin{aligned}\kappa_1(\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)} - \boldsymbol{\theta}_0) &= n^{-1}(\boldsymbol{\alpha}_{\mathrm{ML}1} - k\boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*) + O(n^{-2}) \equiv n^{-1}\boldsymbol{\alpha}_{\mathrm{W}(k)1} + O(n^{-2}),\\ \kappa_2(\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)}) &= n^{-1}\mathbf{A}_{\mathrm{ML}2} + n^{-2}(\mathbf{A}_{\mathrm{ML}\Delta 2} + k\mathbf{A}_{\mathrm{W}\,|\,\mathrm{ML}\Delta 2}) + O(n^{-3})\\ &\equiv n^{-1}\mathbf{A}_{\mathrm{ML}2} + n^{-2}\mathbf{A}_{\mathrm{W}(k)\Delta 2} + O(n^{-3}) (\mathbf{A}_{\mathrm{W}(k)2} = \mathbf{A}_{\mathrm{ML}2}),\\ \kappa_3(\hat{\theta}_{\mathrm{W}(k)}) &= n^{-2}\alpha_{\mathrm{ML}3} + O(n^{-3}) \ (\alpha_{\mathrm{W}(k)3} = \alpha_{\mathrm{ML}3}),\\ \kappa_4(\hat{\theta}_{\mathrm{W}(k)}) &= n^{-3}\alpha_{\mathrm{ML}4} + O(n^{-4}) \ (\alpha_{\mathrm{W}(k)4} = \alpha_{\mathrm{ML}4}).\end{aligned} \tag{3.1}$$

When there are more than one parameter with $q > 1$, a summarized MSE of the vector estimator is defined in various ways. Define a linear predictor $\eta = \mathbf{p}^{*'}\boldsymbol{\theta}$, where $\mathbf{p}^*$ is a $q\times 1$ vector of fixed constants with $\hat{\eta}_{\mathrm{W}(k)} \equiv \mathbf{p}^{*'}\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)}$ and $\eta_0 \equiv \mathbf{p}^{*'}\boldsymbol{\theta}_0$. When there are p covariates possibly including that for an intercept in regression models with $p = q$, a choice of $\mathbf{p}^*$ is $\bar{\mathbf{x}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$. However, $\bar{\mathbf{x}}$ may frequently be close to

zero except the invariant element for the intercept, and becomes meaningless in such cases. A weight vector to yield principal components using the covariance matrix of the p covariates may be reasonable. However, the weight for the covariate for the intercept becomes zero. It is difficult to define a single reasonable $\mathbf{p}^*$ especially when the covariates are uncorrelated. In the following, assume that $\mathbf{p}^*$ is one of reasonable weights.

The MSE of $\hat{\eta}_{\mathrm{W}(k)}$ is given by

$$\begin{aligned}\mathrm{MSE}(\hat{\eta}_{\mathrm{W}(k)}) &= n^{-1}\mathbf{p}^{*'}\mathbf{A}_{\mathrm{ML}2}\mathbf{p}^* + n^{-2}\mathbf{p}^{*'}(\mathbf{A}_{\mathrm{W}(k)\Delta 2} + \boldsymbol{\alpha}_{\mathrm{W}(k)1}\boldsymbol{\alpha}'_{\mathrm{W}(k)1})\mathbf{p}^* + O(n^{-3}) \\ &= n^{-1}\mathbf{p}^{*'}\mathbf{A}_{\mathrm{ML}2}\mathbf{p}^* \\ &\quad + n^{-2}\mathbf{p}^{*'}\{\mathbf{A}_{\mathrm{ML}\Delta 2} + k\mathbf{A}_{\mathrm{W}\,|\,\mathrm{ML}\Delta 2} + (\boldsymbol{\alpha}_{\mathrm{ML}1} - k\boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*)(\boldsymbol{\alpha}_{\mathrm{ML}1} - k\boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*)'\}\mathbf{p}^* \\ &\quad + O(n^{-3}) \\ &= n^{-1}\mathbf{p}^{*'}\mathbf{A}_{\mathrm{ML}2}\mathbf{p}^* + n^{-2}\mathbf{p}^{*'}(\mathbf{A}_{\mathrm{ML}\Delta 2} + \boldsymbol{\alpha}_{\mathrm{ML}1}\boldsymbol{\alpha}'_{\mathrm{ML}1})\mathbf{p}^* \\ &\quad + n^{-2}k\,\mathbf{p}^{*'}(\mathbf{A}_{\mathrm{W}\,|\,\mathrm{ML}\Delta 2} - \boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*\boldsymbol{\alpha}'_{\mathrm{ML}1} - \boldsymbol{\alpha}_{\mathrm{ML}1}\mathbf{q}_0^{*'}\boldsymbol{\Lambda}^{-1})\mathbf{p}^* \\ &\quad + n^{-2}k^2(\mathbf{p}^{*'}\boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*)^2 + O(n^{-3}), \end{aligned} \tag{3.2}$$

where $n^{-1}\mathbf{A}_{\mathrm{ML}2} + n^{-2}(\mathbf{A}_{\mathrm{W}(k)\Delta 2} + \boldsymbol{\alpha}_{\mathrm{W}(k)1}\boldsymbol{\alpha}'_{\mathrm{W}(k)1})$ is a matrix mean square error (see e.g., Giles & Rayner, 1979, p.320) up to order $O(n^{-2})$.

Then, from (3.2), we have

Result 1. *The MSE of $\hat{\eta}_{\mathrm{W}(k)}$ using the k-th power prior up to order $O(n^{-2})$ under possible model misspecification is minimized when k is*

$$\begin{aligned} k_{\min} &= \frac{1}{2}(\mathbf{p}^{*'}\boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*)^{-2}\mathbf{p}^{*'}(-\mathbf{A}_{\mathrm{W}\,|\,\mathrm{ML}\Delta 2} + \boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*\boldsymbol{\alpha}'_{\mathrm{ML}1} + \boldsymbol{\alpha}_{\mathrm{ML}1}\mathbf{q}_0^{*'}\boldsymbol{\Lambda}^{-1})\mathbf{p}^* \\ &= -\frac{\mathbf{p}^{*'}\mathbf{A}_{\mathrm{W}\,|\,\mathrm{ML}\Delta 2}\mathbf{p}^*}{2(\mathbf{p}^{*'}\boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*)^2} + \frac{\mathbf{p}^{*'}\boldsymbol{\alpha}_{\mathrm{ML}1}}{\mathbf{p}^{*'}\boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*}, \end{aligned} \tag{3.3}$$

where the minimized value is

$$\begin{aligned}\mathrm{MSE}_{O(n^{-2})}(\hat{\eta}_{\mathrm{W}(k)}) &= n^{-1}\mathbf{p}^{*'}\mathbf{A}_{\mathrm{ML}2}\mathbf{p}^* + n^{-2}\mathbf{p}^{*'}(\mathbf{A}_{\mathrm{ML}\Delta 2} + \boldsymbol{\alpha}_{\mathrm{ML}1}\boldsymbol{\alpha}'_{\mathrm{ML}1})\mathbf{p}^* \\ &\quad - \frac{n^{-2}}{4}(\mathbf{p}^{*'}\boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*)^{-2}(-\mathbf{p}^{*'}\mathbf{A}_{\mathrm{W}\,|\,\mathrm{ML}\Delta 2}\mathbf{p}^* + 2\mathbf{p}^{*'}\boldsymbol{\alpha}_{\mathrm{ML}1}\mathbf{p}^{*'}\boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*)^2, \end{aligned} \tag{3.4}$$

and $\mathrm{MSE}_{O(n^{-2})}(\cdot)$ *indicates the MSE up to order*$O(n^{-2})$.

The second definition of the summarized MSE for $\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)}$ is the sum of the MSEs for $(\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)})_i$ $(i = 1, \ldots, q)$, which is well defined and has been used by, for instance, Zellner (1986). Gruber (1998, p.117) called it the total MSE, which is employed in this paper and is denoted by TMSE. Then,

$$\mathrm{TMSE}(\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)}) = \sum_{i=1}^{n}\mathrm{MSE}\{(\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)})_i\}$$

$$= n^{-1}\mathrm{tr}(\mathbf{A}_{\mathrm{ML}2}) + n^{-2}\mathrm{tr}\{\mathbf{A}_{\mathrm{ML}\Delta 2} + k\mathbf{A}_{\mathrm{W}\,|\,\mathrm{ML}\Delta 2} + (\boldsymbol{\alpha}_{\mathrm{ML}1} - k\boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*)(\boldsymbol{\alpha}_{\mathrm{ML}1} - k\boldsymbol{\Lambda}^{-1}\mathbf{q}_0^*)'\} + O(n^{-3}). \quad (3.5)$$

From (3.5), we have

Result 2. *The TMSE of $\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)}$ using the k-th power prior up to order $O(n^{-2})$ under possible model misspecification is minimized when k is*

$$k_{\min} = \frac{1}{2}(\mathbf{q}_0^{*'}\boldsymbol{\Lambda}^{-2}\mathbf{q}_0^*)^{-1}\{-\mathrm{tr}(\mathbf{A}_{\mathrm{W}\,|\,\mathrm{ML}\Delta 2}) + 2\mathbf{q}_0^{*'}\boldsymbol{\Lambda}^{-1}\boldsymbol{\alpha}_{\mathrm{ML}1}\}, \quad (3.6)$$

where the minimized value is

$$\mathrm{TMSE}_{O(n^{-2})}(\hat{\eta}_{\mathrm{W}(k_{\min})}) = n^{-1}\mathrm{tr}(\mathbf{A}_{\mathrm{ML}2}) + n^{-2}\{\mathrm{tr}(\mathbf{A}_{\mathrm{ML}\Delta 2}) + \boldsymbol{\alpha}_{\mathrm{ML}1}'\boldsymbol{\alpha}_{\mathrm{ML}1}\} - \frac{n^{-2}}{4}(\mathbf{q}_0^{*'}\boldsymbol{\Lambda}^{-2}\mathbf{q}_0^*)^{-1}\{-\mathrm{tr}(\mathbf{A}_{\mathrm{W}\,|\,\mathrm{ML}\Delta 2}) + 2\mathbf{q}_0^{*'}\boldsymbol{\Lambda}^{-1}\boldsymbol{\alpha}_{\mathrm{ML}1}\}^2. \quad (3.7)$$

4. Optimal Gaussian and Jeffreys power priors for the canonical parameter in the exponential family

Let a Gaussian prior be proportional to $\exp(-\boldsymbol{\theta}'\boldsymbol{\theta}/2)$. Then, $k\,\mathbf{q}^* = -k\,\boldsymbol{\theta}$. In this section, consider a correct model with a canonical vector parameter in the exponential family. Then, Result 1 gives

Result 3. *The MSE of $\hat{\eta}_{\mathrm{G}(k)}$, which is $\hat{\eta}_{\mathrm{W}(k)}$ when $\mathbf{q}^* = -\boldsymbol{\theta}$, using the Gaussian power prior up to order $O(n^{-2})$ under a correct model with the canonical parameter in the exponential family is minimized when k is*

$$k_{\min} = -\frac{\mathbf{p}^{*'}\mathbf{A}_{\mathrm{G}\,|\,\mathrm{ML}\Delta 2}\mathbf{p}^*}{2(\mathbf{p}^{*'}\mathbf{I}_0^{-1}\boldsymbol{\theta}_0)^2} + \frac{\mathbf{p}^{*'}\boldsymbol{\alpha}_{\mathrm{ML}1}}{\mathbf{p}^{*'}\mathbf{I}_0^{-1}\boldsymbol{\theta}_0}, \quad (4.1)$$

where $\mathbf{A}_{\mathrm{G}\,|\,\mathrm{ML}\Delta 2} = \mathbf{A}_{\mathrm{W}\,|\,\mathrm{ML}\Delta 2}$ when $\mathbf{q}^* = -\boldsymbol{\theta}$ and

$$\mathbf{p}^{*'}\mathbf{A}_{\mathrm{G}\,|\,\mathrm{ML}\Delta 2}\mathbf{p}^* = 2\mathbf{p}^{*'}\frac{\partial \mathbf{I}_0^{-1}(-\boldsymbol{\theta}_0)}{\partial\,\boldsymbol{\theta}_0'}\mathbf{A}_{\mathrm{ML}2}\mathbf{p}^* = 2\mathbf{p}^{*'}[\mathbf{I}_0^{-1}\mathbf{I}_0^{(\mathrm{D}1)}\{(\mathbf{I}_0^{-1}\boldsymbol{\theta}_0)\otimes\mathbf{A}_{\mathrm{ML}2}\} - \mathbf{I}_0^{-1}\mathbf{A}_{\mathrm{ML}2}]\mathbf{p}^* \quad (4.2)$$

with $\mathbf{I}_0^{(\mathrm{D}1)} \equiv -\frac{\partial^3 \bar{l}}{\partial\boldsymbol{\theta}_0(\partial\boldsymbol{\theta}_0')^{\langle 2\rangle}}$. The minimized value is given from (3.4) as

$$\mathrm{MSE}_{O(n^{-2})}(\hat{\eta}_{\mathrm{G}(k_{\min})}) = \mathrm{MSE}_{O(n^{-2})}(\hat{\eta}_{\mathrm{ML}}) - \frac{n^{-2}}{4}(\mathbf{p}^{*'}\mathbf{I}_0^{-1}\boldsymbol{\theta}_0)^{-2}(-\mathbf{p}^{*'}\mathbf{A}_{\mathrm{G}\,|\,\mathrm{ML}\Delta 2}\mathbf{p}^* + 2\mathbf{p}^{*'}\boldsymbol{\alpha}_{\mathrm{ML}1}\mathbf{p}^{*'}\mathbf{I}_0^{-1}\boldsymbol{\theta}_0)^2, \quad (4.3)$$

where $\hat{\eta}_{\mathrm{ML}}$ is defined as $\hat{\eta}_{\mathrm{G}(k)}$ when $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is used in place of $\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)}$.

Similarly, Result 2 gives

Result 4. *The TMSE of* $\hat{\boldsymbol{\theta}}_{\mathrm{G}(k)}$*, which is* $\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)}$ *when* $\mathbf{q}^* = -\boldsymbol{\theta}$*, under the same conditions as in Result 3 is minimized when* k *is*

$$k_{\min} = \frac{1}{2}(\boldsymbol{\theta}_0'\mathbf{I}_0^{-2}\boldsymbol{\theta}_0)^{-1}\{-\mathrm{tr}(\mathbf{A}_{\mathrm{G}\,|\,\mathrm{ML}\Delta 2}) + 2\boldsymbol{\theta}_0'\mathbf{I}_0^{-1}\boldsymbol{\alpha}_{\mathrm{ML}1}\}. \tag{4.4}$$

The minimized value is given from (3.7) as

$$\begin{aligned}\mathrm{TMSE}_{O(n^{-2})}(\hat{\boldsymbol{\theta}}_{\mathrm{G}(k_{\min})}) &= \mathrm{TMSE}_{O(n^{-2})}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) \\ &\quad - \frac{n^{-2}}{4}(\boldsymbol{\theta}_0'\mathbf{I}_0^{-2}\boldsymbol{\theta}_0)^{-1}\{-\mathrm{tr}(\mathbf{A}_{\mathrm{G}\,|\,\mathrm{ML}\Delta 2}) + 2\boldsymbol{\theta}_0'\mathbf{I}_0^{-1}\boldsymbol{\alpha}_{\mathrm{ML}1}\}^2. \end{aligned}\tag{4.5}$$

Consider the Jeffreys power prior under a correct model as in Results 3 and 4 with the canonical parameter in the exponential family. Then, noting that $\mathbf{q}^* = -\mathbf{I}\boldsymbol{\alpha}_{\mathrm{ML}1}$ for the Jeffreys prior in this case (see Ogasawara, 2014, Section 5), we have

Result 5. *The MSE of* $\hat{\eta}_{\mathrm{J}(k)}$*, defined similarly to* $\hat{\eta}_{\mathrm{G}(k)}$*, using the Jeffreys power prior up to order* $O(n^{-2})$ *under a correct model with the canonical parameter in the exponential family is minimized when* k *is*

$$k_{\min} = (\mathbf{p}^{*'}\boldsymbol{\alpha}_{\mathrm{ML}1})^{-2}\mathbf{p}^{*'}n\,\mathrm{acov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}_{\mathrm{ML}1}, \hat{\boldsymbol{\theta}}_{\mathrm{ML}}')\mathbf{p}^* + 1, \tag{4.6}$$

where $\mathrm{acov}_{\boldsymbol{\theta}}(\cdot)$ *is* $\mathrm{acov}_{\mathrm{T}}(\cdot)$ *under a correct model. The minimized value is given from (3.4) as*

$$\begin{aligned}\mathrm{MSE}_{O(n^{-2})}(\hat{\eta}_{\mathrm{J}(k_{\min})}) &= \mathrm{MSE}_{O(n^{-2})}(\hat{\eta}_{\mathrm{ML}}) \\ &\quad - n^{-2}(\mathbf{p}^{*'}\boldsymbol{\alpha}_{\mathrm{ML}1})^{-2}[\mathbf{p}^{*'}\{n\,\mathrm{acov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}_{\mathrm{ML}1}, \hat{\boldsymbol{\theta}}_{\mathrm{ML}}') + \boldsymbol{\alpha}_{\mathrm{ML}1}\boldsymbol{\alpha}_{\mathrm{ML}1}'\}\mathbf{p}^*]^2 \\ &= n^{-1}\mathbf{p}^{*'}\mathbf{I}_0^{-1}\mathbf{p}^* + n^{-2}[\mathbf{p}^{*'}\mathbf{A}_{\mathrm{ML}\Delta 2}\mathbf{p}^* - 2\mathbf{p}^{*'}n\,\mathrm{acov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}_{\mathrm{ML}1}, \hat{\boldsymbol{\theta}}_{\mathrm{ML}}')\mathbf{p}^* \\ &\quad - (\mathbf{p}^{*'}\boldsymbol{\alpha}_{\mathrm{ML}1})^{-2}\{\mathbf{p}^{*'}n\,\mathrm{acov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}_{\mathrm{ML}1}, \hat{\boldsymbol{\theta}}_{\mathrm{ML}}')\mathbf{p}^*\}^2]. \end{aligned}\tag{4.7}$$

Result 6. *The TMSE of* $\hat{\boldsymbol{\theta}}_{\mathrm{J}(k)}$ *under the same conditions as in Results 3 to 5 is minimized when* k *is*

$$k_{\min} = (\boldsymbol{\alpha}_{\mathrm{ML}1}'\boldsymbol{\alpha}_{\mathrm{ML}1})^{-1}\mathrm{tr}\{n\,\mathrm{acov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}_{\mathrm{ML}1}, \hat{\boldsymbol{\theta}}_{\mathrm{ML}}')\} + 1 \tag{4.8}$$

(see (3.6) with (2.7)). The minimized value is given from (3.7) as

$$\begin{aligned}\mathrm{TMSE}_{O(n^{-2})}(\hat{\eta}_{\mathrm{J}(k_{\min})}) &= \mathrm{TMSE}_{O(n^{-2})}(\hat{\eta}_{\mathrm{ML}}) \\ &\quad - n^{-2}(\boldsymbol{\alpha}_{\mathrm{ML}1}'\boldsymbol{\alpha}_{\mathrm{ML}1})^{-1}[\mathrm{tr}\{n\,\mathrm{acov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}_{\mathrm{ML}1}, \hat{\boldsymbol{\theta}}_{\mathrm{ML}}')\} + \boldsymbol{\alpha}_{\mathrm{ML}1}'\boldsymbol{\alpha}_{\mathrm{ML}1}]^2 \\ &= n^{-1}\mathrm{tr}(\mathbf{I}_0^{-1}) + n^{-2}[\mathrm{tr}(\mathbf{A}_{\mathrm{ML}\Delta 2}) - 2\mathrm{tr}\{n\,\mathrm{acov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}_{\mathrm{ML}1}, \hat{\boldsymbol{\theta}}_{\mathrm{ML}}')\} \\ &\quad - (\boldsymbol{\alpha}_{\mathrm{ML}1}'\boldsymbol{\alpha}_{\mathrm{ML}1})^{-1}\{\mathrm{tr}(n\,\mathrm{acov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}_{\mathrm{ML}1}, \hat{\boldsymbol{\theta}}_{\mathrm{ML}}'))\}^2]. \end{aligned}\tag{4.9}$$

Recall that $\mathbf{I} = O(1)$ was defined as an information matrix per observation. When covariates exist, the distributions of $\mathbf{U}_i (i = 1, \ldots, n)$ are generally unidentical as in regression models. Even in such a case, $\mathbf{I}$ is well defined as that averaged over observations. Define $\eta^{(i)} = \mathbf{x}_i'\boldsymbol{\theta}$ as a linear predictor for the i-th observation. Then, we can

consider priors for $\eta^{(i)}(i = 1, \ldots, n)$. Take the geometric mean $\bar{f}_\eta$ of these priors:

$$\bar{f}_\eta = \left\{ \prod_{i=1}^{n} f(\eta^{(i)}) \right\}^{1/n}. \tag{4.10}$$

In the case of logistic regression, whose numerical results will be given in Subsection 6.2:

$$\pi_i \equiv \frac{1}{1 + \exp(-\mathbf{x}_i'\boldsymbol{\beta})} = \frac{1}{1 + \exp(-\eta^{(i)})} \quad (i = 1, \ldots, n), \tag{4.11}$$

where $\boldsymbol{\theta} = \boldsymbol{\beta}$ ($\boldsymbol{\theta}_0 = \boldsymbol{\beta}_0$) and the Fisher information for $\eta^{(i)}$ is $\mathbf{I}^{(i)} \equiv \pi_i(1 - \pi_i)$, which gives the Jeffreys power prior $\{f(\eta^{(i)})\}^k = \{\mathbf{I}^{(i)}\}^{k/2}$ for $\eta^{(i)}$ yielding $\bar{f}_\eta^k = \left\{ \prod_{i=1}^{n} \mathbf{I}^{(i)} \right\}^{k/(2n)}$. Then, the posterior density or the weighted likelihood of $\boldsymbol{\beta}$ becomes

$$\begin{aligned} \exp(l)\bar{f}_\eta^k &= \left\{ \prod_{i=1}^{n} \pi_i^{u_i}(1 - \pi_i)^{1-u_i} \right\} \prod_{i=1}^{n} \{\pi_i(1 - \pi_i)\}^{k/(2n)} \\ &= \prod_{i=1}^{n} \pi_i^{u_i + n^{-1}0.5k}(1 - \pi_i)^{1-u_i+n^{-1}0.5k}. \end{aligned} \tag{4.12}$$

That is, the weighted likelihood is given by the pseudocount $n^{-1}0.5k$ each for two cells for the i-th observation. In this case

$$k\,\mathbf{q}^* = \frac{\partial \log \bar{f}_\eta^k}{\partial \boldsymbol{\beta}} = n^{-1}0.5k \sum_{i=1}^{n} (1 - 2\pi_i)\mathbf{x}_i. \tag{4.13}$$

Assume a correct model as before. Define $\hat{\boldsymbol{\beta}}_{\mathrm{P}}$ as that maximizes (4.12), $\hat{\mathbf{I}}_{\mathrm{P}} \equiv \mathbf{I}(\hat{\boldsymbol{\beta}}_{\mathrm{P}})$ giving $\hat{\mathbf{L}} = -\hat{\mathbf{I}}_{\mathrm{P}}$, and $\hat{\mathbf{q}}_{\mathrm{P}}^* \equiv \mathbf{q}^*(\hat{\boldsymbol{\beta}}_{\mathrm{P}})$. Then, $\mathbf{L}^{(\mathrm{W})}_{O_p(n^{-1/2})}$ in (2.4) and (2.5) becomes

$$\mathbf{L}^{(\mathrm{P})}_{O_p(n^{-1/2})} = -(-\hat{\mathbf{I}}_{\mathrm{p}}^{-1}\hat{\mathbf{q}}_{\mathrm{P}}^* + \mathbf{I}_0^{-1}\mathbf{q}_0^*) = \hat{\mathbf{I}}_{\mathrm{p}}^{-1}\hat{\mathbf{q}}_{\mathrm{P}}^* - \mathbf{I}_0^{-1}\mathbf{q}_0^* \tag{4.14}$$

with $\mathrm{E}_\theta(\mathbf{L}^{(\mathrm{P})}\mathbf{l}_0^{(1)'}\boldsymbol{\Lambda}^{(1)}) = \frac{\partial \mathbf{I}_0^{-1}\mathbf{q}_0^*}{\partial \boldsymbol{\beta}_0'}\mathbf{I}_0^{-1}$, where

$$\frac{\partial \mathbf{I}_0^{-1}\mathbf{q}_0^*}{\partial(\boldsymbol{\beta}_0)_j} = -\mathbf{I}_0^{-1}n^{-1}\sum_{i=1}^{n}(1 - 2\pi_i)\pi_i(1 - \pi_i)\mathbf{x}_i\mathbf{x}_i'x_{ij}\mathbf{I}_0^{-1}\mathbf{q}_0^* - \mathbf{I}_0^{-1}n^{-1}\sum_{i=1}^{n}\pi_i(1 - \pi_i)\mathbf{x}_i x_{ij} \tag{4.15}$$

and $x_{ij} \equiv (\mathbf{x}_i)_j$ $(i = 1, \ldots, n; j = 1, \ldots, q)$. The values of $k_{\min}$ are given from Results 1 and 2, (4.14) and (4.15).

In Results 1 to 6 and the result for $\hat{\boldsymbol{\beta}}_{\mathrm{P}}$, the values of $k_{\min}$ generally depend on unknown $\boldsymbol{\theta}_0$. However, under correct models with a scalar parameter, several $k_{\min}$ values independent of $\boldsymbol{\theta}_0$ are available. For non-canonical parameters, when pseudo power priors with $k\mathbf{q}^* = -k\mathbf{I}\boldsymbol{\alpha}_{\mathrm{ML1}}$ for bias adjustment are used rather than the Jeffreys power prior, we have $k_{\min} = 3$ for the transformed parameter representing the

inter-event time in the Poisson distribution and $k_{\min} = -0.5$ for the mean in the normal distribution with known coefficient of variation (Ogasawara, 2014). In canonical parameters, $k\mathbf{q}^* = -k\mathbf{I}\boldsymbol{\alpha}_{\mathrm{ML1}}$ becomes the Jeffreys power prior, which gives $k_{\min} = 3$ for the logarithm of the Poisson source parameter, $k_{\min} = 2$ for the negative rate parameter in the gamma distribution with known shape parameter, and $k_{\min} = 2$ for the precision parameter in the normal distribution with known mean (Ogasawara, 2013). Though in the canonical parameter, logit, in the Bernoulli distribution $k_{\min}$ depends on the population parameter, it is known that $k_{\min} \geq 3$ (Ogasawara, 2014).

Though $k_{\min}$ values in Results 1 to 6 and the result for $\hat{\boldsymbol{\beta}}_{\mathrm{P}}$ depend on $\boldsymbol{\theta}_0$, the results may be useful to have reasonable values of k after accumulation of information about $k_{\min}$ values by numerical experiments with known $\boldsymbol{\theta}_0$ as will be illustrated later and by $k_{\min}$ obtained from $\hat{\boldsymbol{\theta}}_{\mathrm{W}}$. For a similar purpose, consider the Jeffreys power prior whose power is k. Write

$$\begin{aligned}\mathrm{MSE}_{O(n^{-2})}(\hat{\eta}_{\mathrm{J}(k)}) = n^{-1}\mathbf{p}^{*'}\mathbf{A}_{\mathrm{ML2}}\mathbf{p}^* + n^{-2}\{&\mathbf{p}^{*'}\mathbf{A}_{\mathrm{ML}\Delta 2}\mathbf{p}^* \\ &- 2k\,\mathbf{p}^{*'}n\,\mathrm{acov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}_{\mathrm{ML1}}, \hat{\boldsymbol{\theta}}'_{\mathrm{ML}})\mathbf{p}^* + (1-k)^2(\mathbf{p}^*\boldsymbol{\alpha}_{\mathrm{ML1}})^2\}, \quad (4.16)\end{aligned}$$

which gives

Result 7. *Under the same conditions as in Results 3 to 6, when* $\mathrm{acov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}_{\mathrm{ML1}}, \hat{\boldsymbol{\theta}}'_{\mathrm{ML}})$ *is positive definite*

$$\mathrm{MSE}_{O(n^{-2})}(\hat{\eta}_{\mathrm{J}(k)}) < \mathrm{MSE}_{O(n^{-2})}(\hat{\eta}_{\mathrm{ML}}) \quad \text{for} \quad 0 < k \leq 2. \quad (4.17)$$

Similarly, since

$$\begin{aligned}\mathrm{TMSE}_{O(n^{-2})}(\hat{\boldsymbol{\theta}}_{\mathrm{J}(k)}) &= n^{-1}\mathrm{tr}(\mathbf{A}_{\mathrm{ML2}}) \\ &+ n^{-2}[\mathrm{tr}(\mathbf{A}_{\mathrm{ML}\Delta 2}) - 2k\,\mathrm{tr}\{n\,\mathrm{acov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}_{\mathrm{ML1}}, \hat{\boldsymbol{\theta}}'_{\mathrm{ML}})\} + (1-k)^2\boldsymbol{\alpha}'_{\mathrm{ML1}}\boldsymbol{\alpha}_{\mathrm{ML1}}], \quad (4.18)\end{aligned}$$

we have

Result 8. *Under the same conditions as in Results 3 to 7, when* $\mathrm{tr}\{\mathrm{acov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}_{\mathrm{ML1}}, \hat{\boldsymbol{\theta}}'_{\mathrm{ML}})\} > 0$,

$$\mathrm{TMSE}_{O(n^{-2})}(\hat{\eta}_{\mathrm{J}(k)}) < \mathrm{TMSE}_{O(n^{-2})}(\hat{\eta}_{\mathrm{ML}}) \quad \text{for} \quad 0 < k \leq 2. \quad (4.19)$$

5. Interval estimation of a scalar canonical parameter in the exponential family

Though most of $\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)}$ were introduced earlier to improve point estimation of $\boldsymbol{\theta}_0$, they can also be used for interval estimation using estimated asymptotic cumulants of the studentized $\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)}$. The asymptotic cumulants of the studentized $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{W}(k)}$ under possible model misspecification for a vector canonical parameter are given in Subsection A.2 of the appendix (for the results for a scalar canonical parameter under a correct model, see Ogasawara, 2013, Sections 4 and 5).

In this section, interval estimation of a scalar parameter θ_0 with canonical parametrization in the exponential family under a correct model is dealt with. Under such a condition, define

$$\begin{aligned} t_{\mathrm{W}(k)} &\equiv n^{1/2}\hat{\bar{i}}_{\mathrm{W}(k)}^{1/2}(\hat{\theta}_{\mathrm{W}(k)}-\theta_0) \\ &= n^{1/2}\{\hat{\bar{i}}_{\mathrm{ML}}^{1/2}+(n^{-1}/2)k\hat{\bar{i}}_{\mathrm{ML}}^{(\mathrm{D1})}\hat{\bar{i}}_{\mathrm{ML}}^{-3/2}\hat{q}_{\mathrm{ML}}^{*}+O_p(n^{-2})\} \\ &\quad\times\{\hat{\theta}_{\mathrm{ML}}-\theta_0+n^{-1}k\hat{\bar{i}}_{\mathrm{ML}}^{-1}\hat{q}_{\mathrm{ML}}^{*}+O_p(n^{-2})\} \\ &= t_{\mathrm{ML}}+n^{-1/2}\{\hat{\bar{i}}_{\mathrm{ML}}^{-1/2}+(1/2)(\hat{\theta}_{\mathrm{ML}}-\theta_0)\hat{\bar{i}}_{\mathrm{ML}}^{(\mathrm{D1})}\hat{\bar{i}}_{\mathrm{ML}}^{-3/2}\}k\hat{q}_{\mathrm{ML}}^{*}+O_p(n^{-3/2}), \end{aligned} \tag{5.1}$$

where $\hat{\bar{i}}_{\mathrm{W}(k)}$ is the fisher information averaged over observations $\bar{i}_0$ with θ_0 replaced by $\hat{\theta}_{\mathrm{W}(k)}$; $t_{\mathrm{ML}}\equiv n^{1/2}\hat{\bar{i}}_{\mathrm{ML}}^{1/2}(\hat{\theta}_{\mathrm{ML}}-\theta_0)$; $\hat{\bar{i}}_{\mathrm{ML}}$ is the Fisher information estimated by $\hat{\theta}_{\mathrm{ML}}$; $\hat{\bar{i}}_{\mathrm{ML}}^{(\mathrm{D1})}=\partial\bar{i}/\partial\theta|_{\theta=\hat{\theta}_{\mathrm{ML}}}$; and $\hat{q}_{\mathrm{ML}}^{*}$ is the estimated q^{*} using $\hat{\theta}_{\mathrm{ML}}$. Denote the asymptotic cumulants, independent of n, for t_{ML} and $t_{\mathrm{W}(k)}$ by $\alpha_{\mathrm{ML}j}^{(t)}$ and $\alpha_{\mathrm{W}(k)j}^{(t)}(j=1,2,\Delta2,3,4)$, respectively. Then, from (5.1) it can be shown that

$$\begin{aligned} \kappa_1(t_{\mathrm{W}(k)}) &= n^{-1/2}(\alpha_{\mathrm{ML}1}^{(t)}+k\bar{i}_0^{-1/2}q_0^{*})+O(n^{-3/2}) \\ &\equiv n^{-1/2}\alpha_{\mathrm{W}(k)1}^{(t)}+O(n^{-3/2}), \\ \kappa_2(t_{\mathrm{W}(k)}) &= 1+n^{-1}[\alpha_{\mathrm{ML}\Delta2}^{(t)}+2k\bar{i}_0^{1/2}n\,\mathrm{acov}_\theta(\hat{\theta}_{\mathrm{ML}},\hat{\bar{i}}_{\mathrm{ML}}^{-1/2}\hat{q}_{\mathrm{ML}}^{*}) \\ &\qquad +k\bar{i}_0^{1/2}n\,\mathrm{avar}_\theta(\hat{\theta}_{\mathrm{ML}})\bar{i}_0^{(\mathrm{D1})}\bar{i}_0^{-3/2}q_0^{*}]+O(n^{-2}) \\ &= 1+n^{-1}\{\alpha_{\mathrm{ML}\Delta2}^{(t)}+2k\,n\,\mathrm{acov}_\theta(\hat{\theta}_{\mathrm{ML}},\hat{q}_{\mathrm{ML}}^{*})\}+O(n^{-2}) \\ &\equiv 1+n^{-1}\alpha_{\mathrm{W}(k)\Delta2}^{(t)}+O(n^{-2})\ (\alpha_{\mathrm{W}(k)2}^{(t)}=\alpha_{\mathrm{ML}2}^{(t)}=1), \\ \kappa_3(t_{\mathrm{W}(k)}) &= n^{-1/2}\alpha_{\mathrm{ML}3}^{(t)}+O(n^{-3/2})\ (\alpha_{\mathrm{W}(k)3}^{(t)}=\alpha_{\mathrm{ML}3}^{(t)}), \\ \kappa_4(t_{\mathrm{W}(k)}) &= n^{-1}\alpha_{\mathrm{ML}4}^{(t)}+O(n^{-2})\ (\alpha_{\mathrm{W}(k)4}^{(t)}=\alpha_{\mathrm{ML}4}^{(t)}) \end{aligned} \tag{5.2}$$

(see Ogasawara, 2014, (4.4)), where $\bar{i}_0^{(\mathrm{D1})}=\partial\bar{i}_0/\partial\theta_0$.

Note that (5.2) holds also for non-canonical parameters. Under canonical parametrization with $q^{*}=-\bar{i}\alpha_{\mathrm{ML}1}$, define $\bar{s}=O_p(1)$ as a mean sufficient statistic with $\bar{s}-\mathrm{E}_\theta(\bar{s})=l_0^{(1)}=\mathrm{l}_0^{(1)}$ in (2.2). Denote the $n^{1/2}$ times skewness and n times excess kurtosis of $\bar{s}$ by $\mathrm{sk}(s)=O(1)$ and $\mathrm{kt}(s)=O(1)$, respectively. Then, from (5.2), Subsection A.1 of the appendix and Ogasawara (2013, Theorem 2),

$$\begin{aligned} \alpha_{\mathrm{J}(k)1}^{(t)} &= -k\bar{i}_0^{1/2}\alpha_{\mathrm{ML}1}=\frac{k}{2}\bar{i}_0^{-3/2}\bar{i}_0^{(\mathrm{D1})}=\frac{k}{2}\mathrm{sk}(s)\ (\alpha_{\mathrm{ML}1}^{(t)}=0), \\ \alpha_{\mathrm{J}(k)\Delta2}^{(t)} &= \alpha_{\mathrm{ML}\Delta2}^{(t)}+k\{-\bar{i}_0^{-2}\bar{i}_0^{(\mathrm{D1})}n\,\mathrm{acov}_\theta(\hat{\theta}_{\mathrm{ML}},\hat{\bar{i}}_{\mathrm{ML}})+\bar{i}_0^{-1}\,n\,\mathrm{acov}_\theta(\hat{\theta}_{\mathrm{ML}},\hat{\bar{i}}^{(\mathrm{D1})})\} \\ &= \alpha_{\mathrm{ML}\Delta2}^{(t)}+k\{-\bar{i}_0^{-2}\bar{i}_0^{(\mathrm{D1})}(\bar{i}_0^{(\mathrm{D1})}\bar{i}_0^{-1})+\bar{i}_0^{-1}\bar{i}_0^{(\mathrm{D2})}\bar{i}_0^{-1}\} \\ &= -\frac{3}{4}\{\mathrm{sk}(s)\}^2+\frac{1}{2}\mathrm{kt}(s)+k[-\{\mathrm{sk}(s)\}^2+\mathrm{kt}(s)], \\ \alpha_{\mathrm{J}(k)3}^{(t)} &= \alpha_{\mathrm{ML}3}^{(t)}=\mathrm{sk}(s)\,, \end{aligned} \tag{5.3}$$

$$\alpha^{(t)}_{\mathrm{J}(k)4} = \alpha^{(t)}_{\mathrm{ML}\,4} = -3\{\mathrm{sk}(s)\}^2 + 3\mathrm{kt}(s),$$

where $\bar{i}^{(\mathrm{D}2)}_0 = \partial^2 \bar{i}_0 / \partial \theta_0^2$.

A lower endpoint of the one-sided confidence interval (CI) with second-order accurate coverage and asymptotic confidence level $\tilde{\alpha}$ $(0 < \tilde{\alpha} < 1)$ using $\hat{\theta}_{\mathrm{J}(k)}$ is given by the Cornish-Fisher expansion as

$$\begin{aligned} L^*_{\mathrm{J}}(k, \tilde{\alpha}, n^{-1}) &= \hat{\theta}_{\mathrm{J}(k)} - n^{-1/2} \hat{\bar{i}}^{-1/2}_{\mathrm{J}(k)} z_{\tilde{\alpha}} - n^{-1} \hat{\bar{i}}^{-1/2}_{\mathrm{J}(k)} \left\{ \hat{\alpha}^{(t)}_{\mathrm{J}(k)1} + \frac{\hat{\alpha}^{(t)}_{\mathrm{ML}3}}{6} (z^2_{\tilde{\alpha}} - 1) \right\} \\ &= \hat{\theta}_{\mathrm{J}(k)} - n^{-1/2} \hat{\bar{i}}^{-1/2}_{\mathrm{J}(k)} z_{\tilde{\alpha}} - n^{-1} \hat{\bar{i}}^{-1/2}_{\mathrm{J}(k)} \left(\frac{k}{2} + \frac{z^2_{\tilde{\alpha}} - 1}{6} \right) \widehat{\mathrm{sk}}(s) \end{aligned} \tag{5.4}$$

with $\Pr(\theta_0 < L^*_{\mathrm{J}}(k, \tilde{\alpha}, n^{-1})) = 1 - \tilde{\alpha} + O(n^{-1})$, where $\int_{-\infty}^{z_{\tilde{\alpha}}} \exp(-x^2/2)(1/\sqrt{2})\, dx = \tilde{\alpha}$; $\hat{\alpha}^{(t)}_{\mathrm{J}(k)1}$, $\hat{\alpha}^{(t)}_{\mathrm{ML}3}$, and $\widehat{\mathrm{sk}}(s)$ are sample versions of $\alpha^{(t)}_{\mathrm{J}(k)1}$, $\alpha^{(t)}_{\mathrm{ML}3}$ and $\mathrm{sk}(s)$, respectively (see Ogasawara, 2012). A corresponding upper endpoint $U^*_{\mathrm{J}}(k, \tilde{\alpha}, n^{-1})$ is given by (5.4) replacing $z_{\tilde{\alpha}}$ with $z_{1-\tilde{\alpha}} = -z_{\tilde{\alpha}}$.

Theorem 1. *For a scalar canonical parameter in the exponential family under a correct model and regularity conditions including Cramér's condition (see e.g., Hall, 1992, p.45, Chapter 2), the one-sided upper and lower Wald CIs using $\hat{\theta}_{\mathrm{J}(k)}$ with the asymptotic confidence level $\tilde{\alpha}$ $(0 < \tilde{\alpha} < 1)$have second-order accurate coverage when k is $k_z \equiv -(z^2_{\tilde{\alpha}} - 1)/3$.*

Proof. In (5.4), the term $-n^{-1} \hat{\bar{i}}^{-1/2}_{\mathrm{J}(k)} \left(\frac{k_z}{2} + \frac{z^2_{\tilde{\alpha}} - 1}{6} \right) \widehat{\mathrm{sk}}(s)$ becomes 0 both for upper and lower CIs giving $L^*_{\mathrm{J}}(k_z, \tilde{\alpha}, n^{-1}) = \hat{\theta}_{\mathrm{J}(k_z)} - n^{-1/2} \hat{\bar{i}}^{-1/2}_{\mathrm{J}(k_z)} z_{\tilde{\alpha}}$, which is a lower endpoint of the Wald CI with second-order accurate coverage. Similarly, $U^*_{\mathrm{J}}(k_z, \tilde{\alpha}, n^{-1}) = \hat{\theta}_{\mathrm{J}(k_z)} + n^{-1/2} \hat{\bar{i}}^{-1/2}_{\mathrm{J}(k_z)} z_{\tilde{\alpha}}$ with the same accuracy order. Q.E.D.

A lower endpoint of the one-sided CI with third-order accurate coverage and the asymptotic confidence level $\tilde{\alpha}$ using $\hat{\theta}_{\mathrm{W}(k)}$ is

$$\begin{aligned} L^*_{\mathrm{W}}(k, \tilde{\alpha}, n^{-3/2}) &= \hat{\theta}_{\mathrm{W}(k)} - n^{-1/2} \hat{\bar{i}}^{-1/2}_{\mathrm{W}(k)} z_{\tilde{\alpha}} - n^{-1} \hat{\bar{i}}^{-1/2}_{\mathrm{W}(k)} \left\{ \hat{\alpha}^{(t)}_{\mathrm{W}(k)1} + \frac{\hat{\alpha}^{(t)}_{\mathrm{ML}3}}{6} (z^2_{\tilde{\alpha}} - 1) \right\} \\ &\quad - n^{-3/2} \hat{\bar{i}}^{-1/2}_{\mathrm{W}(k)} \left[\frac{1}{2} \left\{ \hat{\alpha}^{(t)}_{\mathrm{W}(k)\Delta 2} - 2 \hat{\bar{i}}^{3/2}_{\mathrm{W}(k)} n\, \widehat{\mathrm{acov}}_{\theta} \left(\hat{\theta}_{\mathrm{ML}}, \hat{\alpha}^{(t)}_{\mathrm{W}(k)1} + \frac{\hat{\alpha}^{(t)}_{\mathrm{ML}3}}{6} (z^2_{\tilde{\alpha}} - 1) \right) \right\} z_{\tilde{\alpha}} \right. \\ &\quad \left. + \{\hat{\alpha}^{(t)}_{\mathrm{ML}3}\}^2 \left(-\frac{z^3_{\tilde{\alpha}}}{18} + \frac{5}{36} z_{\tilde{\alpha}} \right) + \hat{\alpha}^{(t)}_{\mathrm{ML}4} \left(\frac{z^3_{\tilde{\alpha}}}{24} - \frac{z_{\tilde{\alpha}}}{8} \right) \right] \end{aligned} \tag{5.5}$$

with $\Pr(\theta_0 < L^*_{\mathrm{W}}(k, \tilde{\alpha}, n^{-3/2})) = 1 - \tilde{\alpha} + O(n^{-3/2})$, where $\widehat{\mathrm{acov}}_{\theta}(\cdot)$ is a sample version of $\mathrm{acov}_{\theta}(\cdot)$ (Ogasawara, 2012).

When $\hat{\theta}_{\mathrm{J}(k_z)}$ is used as in Theorem 1, (5.5) becomes

$$L^*_{\mathrm{J}}(k_z, \tilde{\alpha}, n^{-3/2}) = \hat{\theta}_{\mathrm{J}(k_z)} - n^{-1/2} \hat{\bar{i}}^{-1/2}_{\mathrm{J}(k_z)} z_{\tilde{\alpha}}$$

Table 1: The powers minimizing the AMSEs of the estimators of the multinomial logits for the categorical distribution under correct model specification

K = the number of categories, $\mathbf{p}_{(K)'} = (p_1, \ldots, p_K)$ [Case ID]					Gaussian prior minL	Gaussian prior minT	Jeffreys prior minL	Jeffreys prior minT
$K = 2$	$\mathbf{p}_{(2)'}$	SD(p)	θ_0	α_{ML1}				
[2.1]	(.1, .9)	.4	−2.20	−4.44	.75	.75	3.56	3.56
[2.2]	(.9, .1)	.4	2.20	4.44	.75	.75	3.56	3.56
[2.3]	(.3, .7)	.2	−.85	−.95	2.10	2.10	8.25	8.25
$K = 3$	$\mathbf{p}_{(3)'}$	SD(p)	$\boldsymbol{\theta}_0'$	$\boldsymbol{\alpha}_{\mathrm{ML1}}'$				
[3.1]	(.1, .2, .7)	.26	(−1.95, −1.25)	(−4.29, −1.79)	.53	.74	2.81	3.99
[3.2]	(.2, .7, .1)	.26	(.69, 1.95)	(2.50, 4.29)	.72	.72	5.64	5.61
[3.3]	(.3, .3, .4)	.05	(−.29, −.29)	(−.42, −.42)	5.87	7.11	35.0	51.0
$K = 6$								
[6.1]	$\mathbf{p}_{(6)'}$= (.05, .05, .05, .05, .05, .75), SD(p) = .26 $\boldsymbol{\theta}_0'$ = (−2.71 (5)), $\boldsymbol{\alpha}_{\mathrm{ML1}}'$ = (−9.33 (5))				.19	.42	1.47	3.31
[6.2]	$\mathbf{p}_{(6)'}$ = (.05, .05, .05, .05, .75, .05), SD(p) = .26 $\boldsymbol{\theta}_0'$ = (0, 0, 0, 0, 2.71), $\boldsymbol{\alpha}_{\mathrm{ML1}}'$ = (0, 0, 0, 0, 9.33)				1.29	1.38	67.6	21.7
[6.3]	$\mathbf{p}_{(6)'}$ = (.1, .1, .1, .1, .5, .1), SD(p) = .15 $\boldsymbol{\theta}_0'$ = (0, 0, 0, 0, 1.61), $\boldsymbol{\alpha}_{\mathrm{ML1}}'$ = (0, 0, 0, 0, 4.00)				3.06	3.28	91.8	29.3

Note. SD(p) = the standard deviation of $p_1, \ldots, p_K$, minL = the power minimizing the AMSE of the linear predictor, minT = the power minimizing the total AMSE. The notation a (b) indicates that a is repeated b times.

Table 2.1: The powers minimizing the AMSEs assuming a correct model under correct or incorrect models in logistic regression

Methods	Correct model				Incorrect model	
	$n = 50$		$n = 100$		$n = 50$	
	minL	minT	minL	minT	minL	minT
Gaussian prior	2.16	3.08	1.64	1.84	3.68	6.91
Jeffreys prior	2.32	3.29	2.19	2.83	2.83	4.80
Pseudocounts	7.74	10.71	9.75	11.37	11.34	19.45

Note. minL = the power minimizing the AMSE of the linear predictor, minT = the power minimizing the total AMSE.

$$
\begin{aligned}
&- n^{-3/2}\hat{i}_{\mathrm{J}(k_z)}^{-1/2}\left[\frac{1}{2}\hat{\alpha}_{\mathrm{J}(k_z)\Delta 2}^{(t)} z_{\tilde{\alpha}} + \{\hat{\alpha}_{\mathrm{ML3}}^{(t)}\}^2\left(-\frac{z_{\tilde{\alpha}}^3}{18} + \frac{5}{36}z_{\tilde{\alpha}}\right) + \hat{\alpha}_{\mathrm{ML4}}^{(t)}\left(\frac{z_{\tilde{\alpha}}^3}{24} - \frac{z_{\tilde{\alpha}}}{8}\right)\right] \qquad (5.6)\\
&= \hat{\theta}_{\mathrm{J}(k_z)} - n^{-1/2}\hat{i}_{\mathrm{J}(k_z)}^{-1/2} z_{\tilde{\alpha}}\\
&\quad - n^{-3/2}\hat{i}_{\mathrm{J}(k_z)}^{-1/2}\left[\frac{1}{2}\left\{\left(-\frac{3}{4}\{\widehat{\mathrm{sk}}(s)\}^2 + \frac{1}{2}\widehat{\mathrm{kt}}(s)\right) z_{\tilde{\alpha}} - \frac{z_{\tilde{\alpha}}^3 - z_{\tilde{\alpha}}}{3}\left(-\{\widehat{\mathrm{sk}}(s)\}^2 + \widehat{\mathrm{kt}}(s)\right)\right\}\right.\\
&\quad \left. + \{\widehat{\mathrm{sk}}(s)\}^2\left(-\frac{z_{\tilde{\alpha}}^3}{18} + \frac{5}{36}z_{\tilde{\alpha}}\right) + [-3\{\widehat{\mathrm{sk}}(s)\}^2 + 3\widehat{\mathrm{kt}}(s)]\left(\frac{z_{\tilde{\alpha}}^3}{24} - \frac{z_{\tilde{\alpha}}}{8}\right)\right]\\
&= \hat{\theta}_{\mathrm{J}(k_z)} - n^{-1/2}\hat{i}_{\mathrm{J}(k_z)}^{-1/2} z_{\tilde{\alpha}} - n^{-3/2}\hat{i}_{\mathrm{J}(k_z)}^{-1/2}\left\{-\left(\frac{z_{\tilde{\alpha}}^3}{72} + \frac{z_{\tilde{\alpha}}}{36}\right)\{\widehat{\mathrm{sk}}(s)\}^2 + \left(-\frac{z_{\tilde{\alpha}}^3}{24} + \frac{z_{\tilde{\alpha}}}{24}\right)\widehat{\mathrm{kt}}(s)\right\}.
\end{aligned}
$$

A corresponding upper endpoint $U_{\mathrm{J}}^*(k_z, \tilde{\alpha}, n^{-3/2})$ is similarly obtained.

Table 2.2: Simulated and asymptotic standard errors under correct models

		$n = 50$						$n = 100$					
($n^{1/2}$ASE		$\beta_1 = 1.2$		$\beta_2 = .7$		$\beta_* = .1$		$\beta_1 = 1.2$		$\beta_2 = .7$		$\beta_* = .1$	
of $\hat{\beta}_i$ (not z		(2.94)		(2.92)		(2.44)		(3.31)		(2.41)		(2.34)	
or t) $= \alpha_{\mathrm{ML2}}^{1/2}$)	ASE	SD	HASE	SD	HASE	SD	HASE	SD	HASE	SD	HASE	SD	HASE
z													
ML	1	1.30	1.13	1.19	1.12	1.15	1.10	1.09	1.06	1.08	1.06	1.05	1.04
G1	1	.80	.75	.81	.79	.88	.89	.84	.79	.92	.90	.94	.93
GminL	1	.62	*	.64	*	.73	.54	.75	.55	.85	.78	.88	.85
GminT	1	.54	*	.56	*	.66	*	.73	.45	.83	.74	.87	.83
J1	1	1.07	1.01	1.04	1.02	1.02	1.02	1.02	1.01	1.01	1.01	1.01	1.00
JminL	1	.86	.81	.87	.87	.89	.89	.95	.94	.95	.94	.96	.96
JminT	1	.74	.63	.78	.74	.81	.78	.92	.91	.91	.90	.94	.93
P1	1	1.14	1.08	1.09	1.07	1.08	1.06	1.05	1.03	1.05	1.04	1.03	1.02
P3	1	.97	.96	.95	.96	.97	.98	.99	.98	1.00	.99	.99	.99
PminL	1	.74	.57	.76	.63	.81	.76	.82	.75	.86	.83	.89	.87
PminT	1	.66	*	.68	.27	.74	.57	.79	.69	.83	.78	.87	.83
t													
ML	1	.97	.98	.98	1.00	.99	1.00	1.00	1.00	.99	1.00	1.00	1.00
G1	1	.84	.78	.82	.76	.85	.82	.89	.87	.92	.91	.92	.91
GminL	1	.74	.42	.71	.28	.75	.53	.85	.77	.88	.84	.89	.85
GminT	1	.68	*	.64	*	.69	*	.83	.74	.87	.82	.88	.83
J1	1	.95	.96	.94	.96	.94	.94	.99	1.00	.97	.97	.97	.97
JminL	1	.90	.92	.89	.90	.87	.84	.99	1.01	.94	.95	.94	.94
JminT	1	.86	.90	.84	.86	.82	.77	.98	1.01	.93	.93	.93	.92
P1	1	.96	.96	.96	.97	.96	.97	.99	.99	.98	.98	.98	.98
P3	1	.92	.91	.91	.90	.91	.90	.96	.97	.96	.96	.96	.96
PminL	1	.84	.79	.81	.74	.81	.72	.90	.89	.89	.87	.89	.86
PminT	1	.79	.69	.76	.61	.76	.59	.89	.87	.87	.85	.87	.83

Note. β_* = intercept, ASE of z and $t = (\alpha_{\mathrm{ML2}}^{(v)})^{1/2} = 1$ $(v = z, t)$, SD = the standard deviations from simulations, HASE $= (1 + n^{-1}\alpha_{\mathrm{ML}\Delta 2}^{(v)})^{1/2}$ $(v = z, t)$, G = Gaussian, J = Jeffreys, P = pseudocounts. The asterisks indicate that the values are imaginary. See also the footnote of Table 2.1.

6. Numerical illustration

In this section, two numerical examples are shown using the categorical distribution and logistic regression in Subsections 6.1 and 6.2, respectively. The first example under correct model specification has relatively simple algebraic results while the second example is one of frequently used methods in practice and is dealt with under correct and incorrect model specification. Note that the parameters to be estimated in these examples are canonical parameters in the exponential family.

6.1 The categorical distribution

The categorical distribution with $K(\geq 3)$ categories is the multinomial or generalized version of the Bernoulli distribution whose relationship is similar to that of the multinomial and binomial distributions. The probability functions of the discrete

Table 2.3: Simulated and asymptotic biases under correct models

Th. $= \alpha_{\mathrm{W1}}^{(v)}$ $(v = z, t)$	$n = 50$						$n = 100$					
	β_1		β_2		β_*		β_1		β_2		β_*	
	Sim.	Th.	Sim.	Th.	Sim.	Th.	Sim.	Th.	Sim.	Th.	Sim.	Th.
z												
ML	2.80	2.20	1.59	1.32	.41	.28	2.21	2.04	1.67	1.56	.38	.39
G1	−1.65	−1.34	−.93	−.71	−.13	−.08	−2.60	−2.61	−1.66	−1.61	−.78	−.80
GminT	−5.76	−8.70	−3.20	−4.94	−.50	−.83	−5.45	−6.53	−3.65	−4.29	−1.42	−1.80
J1	.15	0	.01	0	.05	0	.05	0	.03	0	−.02	0
JminT	−4.42	−5.04	−2.87	−3.02	−.50	−.64	−3.63	−3.74	−2.75	−2.85	−.76	−.71
P1	1.61	1.42	.81	.78	.22	.17	1.44	1.38	1.07	1.04	.24	.26
P3	−.28	−.14	−.44	−.30	−.06	−.04	−.01	.05	−.06	.01	−.02	.02
PminT	−4.80	−6.16	−3.35	−4.47	−.61	−.86	−4.90	−5.53	−3.87	−4.33	−.89	−1.02
t												
ML	1.03	1.04	.48	.50	.22	.21	.77	.75	.69	.70	.16	.19
G1	−2.39	−2.50	−1.43	−1.53	−.16	−.15	−3.61	−3.91	−2.32	−2.47	−.91	−1.00
GminT	−6.88	−9.87	−3.79	−5.76	−.53	−.90	−6.58	−7.83	−4.30	−5.15	−1.54	−2.00
J1	−1.03	−1.16	−.79	−.82	−.05	−.07	−1.24	−1.30	−.81	−.86	−.22	−.20
JminT	−5.49	−6.21	−3.56	−3.84	−.53	−.71	−4.94	−5.04	−3.51	−3.71	−.93	−.91
P1	.23	.26	−.07	−.04	.10	.10	.10	.08	.17	.18	.04	.07
P3	−1.26	−1.31	−1.08	−1.12	−.11	−.11	−1.21	−1.25	−.84	−.85	−.20	−.18
PminT	−5.82	−7.32	−3.99	−5.29	−.65	−.93	−6.09	−6.83	−4.55	−5.19	−1.02	−1.22

Note. Sim. = simulated values, Th. = theoretical values. See also the footnotes given earlier.

Table 2.4: Simulated and asymptotic skewness and kurtoses under correct models

	Sim. skewness (Th. skewness)						Sim. kurtosis (Th. kurtosis)					
	$n = 50$			$n = 100$			$n = 50$			$n = 100$		
	β_1	β_2	β_*	β_1	β_2	β_*	β_1	β_2	β_*	β_1	β_2	β_*
z												
(Th.)	(4.7	3.3	.3)	(5.2	3.4	.8)	(58	30	11)	(60	42	11)
ML	28.8	11.6	2.7	7.7	5.1	1.2	2491	517	159	124	75	26
G1	.9	.7	.0	1.8	1.7	.3	1	2	3	10	13	7
GminT	−.0	.0	−.0	.7	.8	.1	−0	−0	−0	2	47	3
J1	9.1	5.0	.8	6.2	4.1	1.0	248	102	43	89	54	21
JminT	1.4	1.2	−.0	4.2	2.7	.7	9	10	5	50	30	13
P1	10.0	5.5	.9	6.6	4.4	1.1	216	100	46	97	60	22
P3	3.8	2.4	.2	4.9	3.3	.8	38	25	15	60	39	17
PminT	.4	.3	−.0	1.7	1.2	.3	1	1	1	11	9	5
t												
(Th.)	(−2.3	−1.6	−.1)	(−2.6	−1.7	−.4)	(−5.2	−13.1	−17.0)	(10.2	−4.8	−13.4)
ML	−2.5	−1.7	−.3	−2.6	−1.7	−.3	−4.9	−11.7	−16.2	1.2	−10.9	−12.1
G1	−1.6	−1.0	−.1	−2.2	−1.4	−.3	−.9	−3.8	−6.3	3.0	−7.1	−8.0
GminT	−.8	−.4	−.0	−1.8	−1.2	−.2	−.2	−.9	−1.8	2.9	−5.6	−5.9
J1	−1.7	−1.1	−.1	−2.2	−1.3	−.3	−7.0	−10.1	−10.6	−2.7	−10.3	−10.1
JminT	−.5	−.2	−.0	−1.5	−.7	−.1	−4.6	−5.3	−4.2	−7.8	−8.5	−7.2
P1	−2.2	−1.4	−.2	−2.5	−1.6	−.3	−4.6	−9.9	−13.1	1.1	−10.3	−11.2
P3	−1.8	−1.1	−.1	−2.3	−1.4	−.3	−4.0	−7.3	−9.0	.9	−9.2	−9.6
PminT	−.8	−.5	−.0	−1.6	−.9	−.2	−2.2	−2.8	−2.8	.2	−6.0	−5.2

Note. Th. skewness $= \alpha_{\mathrm{ML3}}^{(v)}$, Th. kurtosis $= \alpha_{\mathrm{ML4}}^{(v)}$ $(v = z, t)$. See also the footnotes given earlier.

Table 2.5: Simulated and asymptotic root mean square errors (RMSEs) under correct models

	$n=50$						$n=100$					
	β_1		β_*		L.P.		β_1		β_*		L.P.	
	Sim.	Th.	Sim.	Th.	Sim.	Th.	Sim.	Th.	Sim.	Th.	Sim.	Th.
RMSE												
ML	.564	.488	.399	.381	.943	.779	.367	.357	.245	.244	.623	.592
G1	.345	.323	.303	.306	.554	.556	.291	.276	.220	.218	.494	.519
GminT	.405	.321	.229	*	.653	.244	.300	.263	.206	.197	.506	.498
J1	.444	.419	.353	.351	.736	.692	.339	.334	.235	.235	.579	.568
JminT	.403	.396	.283	.272	.669	.599	.326	.325	.220	.219	.561	.550
P1	.484	.455	.372	.368	.793	.737	.351	.345	.240	.240	.596	.580
P3	.402	.397	.335	.339	.639	.659	.326	.323	.232	.232	.554	.559
PminT	.393	.352	.256	.201	.648	.498	.307	.292	.204	.196	.520	.516

Note. The asymptotic RMSE (Th. RMSE) $= \{n^{-1}\alpha_{\mathrm{ML}2} + n^{-2}(\alpha_{\mathrm{ML}\Delta 2} + \alpha^2_{\mathrm{ML}1})\}^{1/2}$,
L.P. = linear predictor. The asterisk indicates that the value is imaginary.
See also the footnotes given earlier.

Table 2.5: (continued)

	$n=50$						$n=100$					
	β_1		β_*		L.P.		β_1		β_*		L.P.	
	Sim.	Th.	Sim.	Th.	Sim.	Th.	Sim.	Th.	Sim.	Th.	Sim.	Th.
$\alpha_{\mathrm{W}\Delta 2}$												
ML	297	123	97	64	869	413	197	138	54	46	545	435
G1	−156	−185	−70	−64	−459	−490	−319	−408	−67	−74	−956	−1171
GminT	−307	−827	−169	−330	−852	−2369	−517	−867	−132	−174	−1533	−2523
J1	62	7	13	9	191	94	53	22	7	4	160	124
JminT	−193	−258	−101	−116	−555	−639	−175	−192	−67	−72	−464	−444
P1	133	70	47	39	350	240	114	75	31	27	301	248
P3	−28	−37	−18	−11	−146	−107	−29	−51	−10	−10	−118	−127
PminT	−243	−447	−137	−202	−758	−1442	−414	−577	−136	−167	−1245	−1695
$\alpha^2_{\mathrm{W}1}$												
ML	68	42	1	0.5	193	121	54	46	1	1	150	131
G1	24	15	0	0.0	62	39	74	75	3	3	207	207
GminT	286	653	1	4	755	1766	325	466	11	18	907	1304
J1	0	0	0	0	0	0	0	0	0	0	0	0
JminT	169	219	2	2	510	634	144	153	3	3	416	436
P1	22	17	0	0.2	58	47	23	21	0	0.4	62	59
P3	1	0.2	0	0.0	5	2	0	0.0	0	0.0	0	0.0
PminT	199	327	2	4	644	1104	262	334	4	6	760	967

independent random vectors $\mathbf{Y}_i (i = 1, \ldots, n)$ are given by

$$\Pr(\mathbf{Y}_i = \mathbf{y}_i) = \prod_{j=1}^{K} p_j^{y_{ij}} (i = 1, \ldots, n), \tag{6.1}$$

where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iK})'$, $\mathbf{y}_i = (y_{i1}, \ldots, y_{iK})'$, y_{ij} =0 or 1, $\sum_{j=1}^{K} y_{ij} = 1, 0 \leq p_j \leq 1$ and $\sum_{j=1}^{K} p_j = 1$. One of the standard parametrizations satisfying the model identi-

Table 2.6: Simulated and asymptotic RMSEs under an incorrect model ($n = 50$)

	$\beta_1 \doteq .801$		$\beta_2 \doteq .479$		$\beta_* \doteq .117$		L.P.	
	Sim.	Th.	Sim.	Th.	Sim.	Th.	Sim.	Th.
ML	.283	.273	.245	.240	.222	.220	.440	.412
G1	.232	.165	.200	.087	.192	.143	.355	.211
GminT	.338	.293	.196	*	.124	*	.554	*
J1	.251	.219	.221	.183	.205	.183	.388	.323
J3	.259	.178	.210	*	.179	.064	.419	.158
JminT	.301	.242	.225	*	.161	*	.507	.114
P1	.267	.253	.233	.218	.215	.205	.410	.378
P3	.247	.214	.216	.169	.202	.173	.374	.307
PminT	.307	.236	.221	*	.144	*	.525	*

Note. The asterisks indicate that the values are imaginary.
See also the footnotes given earlier.

Table 2.7: Simulated proportions of a population value below the endpoints of the confidence intervals under a correct model

	Accuracy order	Nominal values .0050	.0250	.1000	.5000	.9000	.9750	.9950
$n = 50$, Wald								
ML	1	.0000	.0095	.0872	.5182	.8958	.9690	.9922
G1	1	.0000	.0001	.0142	.3472	.8395	.9525	.9883
J1	1	.0000	.0048	.0592	.4534	.8703	.9610	.9901
P1	1	.0000	.0050	.0645	.4767	.8819	.9650	.9913
JZ	2	.0000	.0163	.0939	.4972	.9004	.9750	.9949
$n = 50$, C-F								
ML	2	.0070	.0257	.0967	.4969	.8998	.9740	.9944
G1	2	.0145	.0390	.1131	.4912	.8933	.9720	.9938
J1	2	.0107	.0328	.1066	.4966	.8966	.9729	.9941
P1	2	.0103	.0317	.1048	.4971	.8981	.9735	.9942
$n = 100$, Wald								
ML	1	.0011	.0156	.0935	.5177	.8962	.9698	.9922
G1	1	.0000	.0017	.0287	.3751	.8479	.9543	.9887
J1	1	.0005	.0105	.0721	.4677	.8758	.9623	.9905
P1	1	.0007	.0117	.0795	.4924	.8878	.9668	.9917
JZ	2	.0028	.0221	.0982	.5012	.9000	.9754	.9951
$n = 100$, C-F								
ML	2	.0055	.0253	.0995	.5010	.8996	.9745	.9947
G1	2	.0097	.0334	.1103	.4972	.8944	.9727	.9942
J1	2	.0068	.0282	.1041	.5008	.8974	.9736	.9944
P1	2	.0064	.0273	.1024	.5012	.8988	.9741	.9945

Note. 1(2) = 1st (2nd)-order accurate, JZ = Jeffreys modal with the matching power prior. C-F = Cornish-Fisher. See also the footnotes given earlier.

fication is given by

$$p_j = e^{\theta_j} / \left(1 + \sum_{a=1}^{K-1} e^{\theta_a}\right) \quad (j = 1, \ldots, K-1) \quad \text{and} \quad p_K = 1 / \left(1 + \sum_{a=1}^{K-1} e^{\theta_a}\right). \tag{6.2}$$

That is, $\theta_j = \log(p_j/p_K)$ $(j = 1, \ldots, K-1)$ is the generalized logit, where the last

category is used as a reference category.

It can be shown that n times the asymptotic biases of $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \{(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})_1, \ldots, (\hat{\boldsymbol{\theta}}_{\mathrm{ML}})_{K-1}\}'$ are

$$(\boldsymbol{\alpha}_{\mathrm{ML1}})_j = \frac{1}{2}\left(-\frac{1}{p_j} + \frac{1}{p_K}\right) \tag{6.3}$$

and $$n\,\mathrm{acov}\{(\hat{\boldsymbol{\alpha}}_{\mathrm{ML1}})_j, \hat{\boldsymbol{\theta}}'_{\mathrm{ML}}\} = \frac{1}{2}(p_j^{-2}\mathbf{e}_{(j)} + p_K^{-2}\mathbf{1}_{(K-1)})' > 0 \ (j = 1, \ldots, K-1) \tag{6.4}$$

(see Ogasawara, 2015, Lemma 2 and Equation (S.20)), where $\mathbf{e}_{(j)}$ is the $(K-1) \times 1$ vector, whose j-th element is 1 with the remaining ones being 0 and $\mathbf{1}_{(K-1)}$ is the $(K-1) \times 1$ vector of 1's.

The Jeffreys prior evaluated at the population value is given by constant$\times |\mathbf{I}_0|^{1/2}$ with $|\mathbf{I}_0| = p_1 p_2 \cdots p_K = p_1 p_2 \cdots p_{K-1}(1 - p_1 - \cdots - p_{K-1})$ (see Ogasawara, 2015, Lemma 1). The Gaussian prior is defined as constant$\times \exp(-\boldsymbol{\theta}'\boldsymbol{\theta}/2)$ with $\mathbf{q}_0^* = \boldsymbol{\theta}_0$ (recall (2.1)). The optimal powers of these priors for the linear predictors and the TMSEs are obtained by Results 3 to 6. From (6.4), we find that the elements of $n\,\mathrm{acov}(\hat{\boldsymbol{\alpha}}_{\mathrm{ML1}}, \hat{\boldsymbol{\theta}}'_{\mathrm{ML}})$ are all positive and that the matrix is positive definite. Consequently, this case satisfies the conditions of Results 7 and 8 for nice properties of the estimators with the associated power priors.

Using Results 6, (6.3) and (6.4), $k_{\min}$ for the TMSE of the Jeffreys prior is given by

$$\begin{aligned} k_{\min} &= 2\left\{\sum_{j=1}^{K-1}\left(-\frac{1}{p_j} + \frac{1}{p_K}\right)^2\right\}^{-1}\left(\sum_{j=1}^{K-1}\frac{1}{p_j^2} + \frac{K-1}{p_K^2}\right) + 1 \\ &= 2\left\{\sum_{j=1}^{K-1}\frac{1}{p_j^2} + \frac{K-1}{p_K^2} - \frac{2}{p_K}\sum_{j=1}^{K-1}\frac{1}{p_j}\right\}^{-1}\left(\sum_{j=1}^{K-1}\frac{1}{p_j^2} + \frac{K-1}{p_K^2}\right) + 1. \end{aligned} \tag{6.5}$$

Since the factor $\{\cdot\}$ on the right-hand side of (6.5) is smaller than the factor $(\cdot)$, we have

Result 9. *Under the categorical distribution of (6.1) with canonical parametrization of (6.2), the lower bound of $k_{\min}$ for the TMSE up to order $O(n^{-2})$ given by the Jeffreys prior is 3.*

When $K = 2$, which reduces to the Bernoulli distribution, (6.5) becomes $2 + \{1/(2p_1 - 1)^2\} > 3$ when $p_1 \neq 0, 1/2, 1$, which is known as was mentioned earlier. The lower bound 3 in (6.5) is obtained as the limiting value as $p_K \to 1$. Note that the case $p_K = 1$ gives the largest variance of $p_1, \ldots, p_K$(see Ogasawara, 2015, Section 4). In the case of $K = 2$, the largest variance is obtained when $p_1 = 0$ or 1

It is of interest to see that the lower bound 3 does not depend on the number of categories. The largest value of $k_{\min} = +\infty$ is also given as the limiting value of (6.5) when $p_j \to K^{-1}(j = 1, \ldots, K)$. This result is reasonable since the k-th power of the Jeffreys prior distributes equal $k/2$ pseudocounts to each cell of K categories

suggesting equal probabilities for the categories. Note that when k is infinite, the estimators give the population values.

Table 1 shows numerical values of the optimal powers in the associated priors. The values of $k_{\min}$ denoted by minL and minT minimize the AMSE of the linear predictor and the TMSE up to order $O(n^{-2})$, respectively, where $\mathbf{p}^* = \mathbf{1}_{(K-1)}$ is used for the linear predictor for illustration. The three cases of $K = 2$ give the results of the usual binomial logit and are included for comparison. Cases 2.1 and 2.2 are symmetric. The SD(p) indicates the standard deviation of $p_1, \ldots, p_K$. Case 2.3 is less variant than Cases 2.1 and 2.2 and has smaller absolute values of the elements of $\boldsymbol{\theta}_0$ and $\boldsymbol{\alpha}_{\mathrm{ML1}}$, and larger $k_{\min}$'s than those in Cases 2.1 and 2.2.

The multinomial cases of $K = 3$ and $K = 6$ have tendencies similar to those with $K = 2$. However, we find that when the reference category is changed, the results changes (see Cases 3.1 vs. 3.2 and Cases 6.1 vs. 6.2). Cases 6.2 and 6.3 include categories whose proportions are equal to that of the reference category. In these cases, the $k_{\min}$'s have relatively large values in line with theory. The $k_{\min}$'s for the Gaussian prior are smaller than those for the Jeffreys prior in these cases.

6.2 Logistic regression

In this subsection, logistic regression with fixed covariates is used for numerical illustration under correct and incorrect model specification, where the regression coefficients are canonical parameters. The first half of this subsection is for illustration of $k_{\min}$ values, asymptotic cumulants and AMSEs of parameter estimators while the second half is for interval estimation. In the first half, two covariates and the unit covariate for an intercept are used, that is $\mathbf{x}_i = (x_{i1}, x_{i2}, 1)'$. The first two covariates are randomly and independently generated using the standard normal. The parameter vector is $\boldsymbol{\theta} = \boldsymbol{\beta} = (\beta_1, \beta_2, \beta_*)'$, where the last element β_* is an intercept parameter. The population value $\boldsymbol{\beta}_0 = (1.2, 0.7, 0.1)'$ is used. Under the correct model, $\Pr(U_i = 1) = 1/\{1 + \exp(-\mathbf{x}_i'\boldsymbol{\beta})\} \equiv \pi_i (i = 1, \ldots, n)$ while under an incorrect model, the term $-\mathbf{x}_i'\boldsymbol{\beta}$ is perturbed such that $-\mathbf{x}_i'\boldsymbol{\beta} + e_i$, where e_i is independently normally distributed with mean 0 and variance 10 giving $\pi_{\mathrm{T}i} \equiv 1/\{1 + \exp(-\mathbf{x}_i'\boldsymbol{\beta} + e_i)\}$. Two sample sizes $n = 50$ and 100 are used. The degree of model misspecification is gross in that the correlation of π_i and $\pi_{\mathrm{T}i}$ over observations are as low as 0.558. The population value $\boldsymbol{\beta}_0$ under the incorrect model is obtained by using $\pi_{\mathrm{T}i}$ in place of u_i in ML estimation, which gives $\boldsymbol{\beta}_0 \doteq (0.801, 0.479, 0.117)'$ up to the third decimal place. The reduction of the population values of the regression coefficients except the intercept due to a regression effect is observed.

Table 2.1 gives the $k_{\min}$ values which minimize the AMSE of the linear predictor and the TMSE up to order $O(n^{-2})$, where $\mathbf{p}^* = (1, 1, 1)'$ is used for the linear predictor for simplicity. The values for the Gaussian and Jeffreys power priors are similar but the latter is somewhat larger than the former under the correct model. The values by pseudocounts using the geometric mean of the Jeffreys power priors are much larger than the other values. Under the incorrect model, the value by Gaussian becomes

larger than that by Jeffreys. It is found that a lower bound 3 of $k_{\min}$ by Jeffreys for logit in a binomial/multinomial proportion with no covariates does not hold in this case with covariates. However, all the values by Jeffreys in Table 2.1 are greater than 2. The relatively large values by pseudocounts can be partially explained by the fact that π_i and $\pi_{\mathrm{T}i}$ are on average close to 0.5 (the mean of π_i over observations is 0.535 when $n = 50$, and is 0.496 when $n = 100$; the mean of $\pi_{\mathrm{T}i}$ is 0.539). Note that when $\pi_i = 0.5$ for all observations, optimal pseudocounts for two cells are infinite.

Tables 2.2, 2.3 and 2.4 show selected results of simulated and asymptotic cumulants of the estimators using various power priors and ML for comparison. Simulated cumulants are given by simulations with 10^5 replications, where k-statistics (unbiased sample cumulants) are used, which are multiplied by appropriate powers of n for ease of comparison to the asymptotic cumulants independent of n. Three generated observations were discarded, when $n = 50$, due to non-convergence of estimation until 10^5 regular observations were obtained. No generated observations were discarded when $n = 100$.

In Table 2.2, for the Gaussian and Jeffreys power priors, the results with $k = 1$ (G1 and J1), minL (GminL and JminL) and minT (GminT and JminT) are shown while for the pseudocount method, the results with $k = 1$ (P1), 3 (P3), minL (PminL) and minT (PminT) are presented, where P3 is for intermediate results between P1 and PminL. The results are for z and t, where $z = n^{1/2}(\mathbf{I}_0^{-1})_{\theta\theta}^{-1/2}(\hat{\theta}_{\mathrm{W}(k)} - \theta_0)$ with $(\cdot)_{\theta\theta}$ being the diagonal element corresponding to θ_0. That is, z is the standardized parameter estimator using the population asymptotic standard error for ease of comparison to the results by t. The asymptotic cumulants of z are denoted by e.g., $\alpha_{\mathrm{ML1}}^{(z)}$ similarly to $\alpha_{\mathrm{ML1}}^{(t)}$ for t. The SD is a simulated standard error, which is given by the standard deviation from 10^5 estimates in simulations. HASE $= (1 + n^{-1}\alpha_{\mathrm{W}\Delta 2}^{(v)})$ $(v = z, t)$.

From the results of z in Table 2.2, it is seen that the MLE tends to give large SE values while other values are smaller, to various degrees, than those by ML, which is expected as a general property of shrinkage by Bayesian estimation. Among the Bayes estimators, GminL, GminT and PminT show large shrinkage. J1 and P3 give values close to the unit ASE. The simulated SEs different from the unit ASE are well approximated by the HASEs in the table. On the other hand, the results of t are much more stable than those of z in that the SEs are closer to the unit ASE, which is interpreted as a compensatory effect due to the estimated standard error.

In Table 2.2, some of HASEs are imaginary, which indicates negative higher-order asymptotic variances. This can happen in finite samples especially when sample sizes are relatively small since the added higher-order term can be negative. The negative higher-order asymptotic variances are due to the overcorrection by the added term. This can possibly be corrected by considering more higher-order asymptotic variances e.g., that up to order $O(n^{-3})$. In Table 2.2, the imaginary cases come from the relatively smaller sample size $n = 50$ in the table. Note that the simulated standard error of z denoted by SD for GminT, $\beta_1 = 1.2$ and $n = 50$ is as small as .54, whose corresponding usual ASE is 1. In this case, the over correction was due to the over

reduction of the poor unit ASE.

Table 2.3 shows simulated and asymptotic biases of the estimators, where results by power priors showing intermediate values are omitted to save space. The zero asymptotic values by J1 are by construction with small simulated values in line with theory. It is of interest to see that the absolute biases of P3 are approximately as small as J1 for z and those of P1 are the smallest in t.

Table 2.4 gives simulated and asymptotic skewness and kurtosis, where the simulated skewness and kurtosis are given by the third and fourth simulated cumulants multiplied by $n^{1/2}$ and n, respectively. The asymptotic values are common to the MLE and Bayes modal estimators. Some variations of simulated skewness are found especially when $n = 50$ for z. These positive values become negative after studentization. Similar tendency of sign reversal by studentization is also observed for kurtosis.

Table 2.5 shows simulated and asymptotic root MSE (RMSE) and associated values, where the asymptotic RMSE is given by $\{n^{-1}\alpha_{\mathrm{ML2}} + n^{-2}(\alpha_{\mathrm{W}(k)\Delta 2} + \alpha^2_{\mathrm{W}(k)1})\}^{1/2}$. The results for β_2 are omitted to save space while those for the linear predictor are included. Note that the minT's are used even for the linear predictor for illustration. The simulated value corresponding to $\alpha_{\mathrm{W}(k)\Delta 2}$ is given by the simulated variance (SD^2) minus the asymptotic variance of order $O(n^{-1})$ followed by multiplication of n^2. The simulated value corresponding to $\alpha^2_{\mathrm{W}(k)1}$ is n^2 times the square of a simulated bias. It is found that $\alpha^2_{\mathrm{W}(k)1}$ for G1, P1, P3 with similar simulated values are smaller than those by ML, while the rounded zero simulated values by J1 are expected ones.

Table 2.6 gives results under the incorrect model when $n = 50$ corresponding to those in Table 2.5. The results of β_2 and J3 are also added. Note that minT is obtained as if the model with $\pi_i (i = 1, \ldots, n)$ holds. On average, the values have decreased from those in Table 2.5. Some of the asymptotic values are imaginary. The values for β_1 by minT are larger than that by ML. The small values are given by G1, J1, J3 P1 and P3.

In the remaining part of this subsection, numerical illustration is shown for interval estimation of a scalar canonical parameter, where the model with a single covariate with an intercept $\pi_i = 1/\{1 + \exp(-x_i\beta_1 - \beta_*)\}$ $(i = 1, \ldots, n)$ is employed, where β_* is assumed to be known. This is a situation, when common $\Pr(U_i = 1)$ $(i = 1, \ldots, n)$ under the condition without covariates is known from previous research. Then, $\boldsymbol{\theta}$ becomes $\theta = \beta_1$, where the population value $\beta_1 = 1.2$ is used with known $\beta_* = 0.1$, and $\mathbf{x}_i = x_i$ $(i = 1, \ldots, n)$ are randomly generated using the standard normal with two sample sizes $n = 50$ and 100.

Table 2.7 gives simulated proportions of the population value below the endpoint of the CIs under the correct model. No observation was discarded until regular 10^5 cases were obtained. The methods ML, G1, J1, P1 and JZ are used, where JZ is the Jeffreys power prior with $k = k_z$ defined in Section 5. The Wald and second-order accurate Cornish-Fisher CIs are used, where the Wald CI by JZ is second-order accurate as shown earlier. Note that G1, J1 and P1 were not introduced to improve interval estimation but shown here for comparison. The Wald CIs by G1, J1 and P1 show

poorer results than those by ML while JZ gives proportions closer to the nominal values than those of the Wald CI by ML. The last result is due to the different orders of accuracy with an advantage for JZ over ML. When the CIs by the Cornish-Fisher expansion are used, all the proportions have improved.

7. Remarks

A natural question arises. Which power prior is best? In the numerical illustration, the population parameters are used to find minL and minT. In practice, the population values are unknown. Further, models are more or less misspecified. For the Gaussian power priors, considering that minL = 1.64 and minT = 1.84 in Table 2.1 and that G1 gives results better than or similar to those by GminT in Table 2.6, the standard normal prior or $k = 1.5$ may be reasonable when data are similar to those in this study. Similarly, for the Jeffreys power priors $k = 1, 1.5$ or 2 may be reasonable. For the pseudocount methods, $k = 3$ (P3) seems to be better than $k = 1$ (P1) since P3 gives smaller variance and squared bias than those by P1 in Table 2.5 and shows smaller RMSEs under the incorrect model in Table 2.6.

When information about appropriate powers of priors are vague, a hierarchical model with a stochastic power can be used as is used for the power prior based on historical data as mentioned earlier. For example, consider the Jeffreys power prior with stochastic k. Define $i_{\mathrm{J}} \equiv (1/2)\log|\mathbf{I}|$, then $|\mathbf{I}|^{k/2} = e^{ki_{\mathrm{J}}}$. Suppose that $k \sim \mathrm{N}(\mu_{\mathrm{J}}, \sigma_{\mathrm{J}}^2)$, and integrate out k as follows:

$$\int_{-\infty}^{\infty} e^{ki_{\mathrm{J}}} \frac{1}{\sqrt{2}} \exp\left\{-\frac{(k-\mu_{\mathrm{J}})^2}{2\sigma_{\mathrm{J}}^2}\right\} dk = \exp\left(\mu_{\mathrm{J}} i_{\mathrm{J}} + \frac{\sigma_{\mathrm{J}}^2 i_{\mathrm{J}}^2}{2}\right)$$
$$= \exp\left\{\left(\mu_{\mathrm{J}} + \frac{\sigma_{\mathrm{J}}^2 i_{\mathrm{J}}}{2}\right) i_{\mathrm{J}}\right\} \equiv \exp(k_{\mathrm{N}} i_{\mathrm{J}}), \quad (7.1)$$

which is the moment generating function in terms of i_{J} in the second-stage Gaussian prior for k. The vector $\mathbf{q}^*$ becomes

$$\mathbf{q}^* = \frac{\partial k_{\mathrm{N}} i_{\mathrm{J}}}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}}\left(\mu_{\mathrm{J}} i_{\mathrm{J}} + \frac{\sigma_{\mathrm{J}}^2 i_{\mathrm{J}}^2}{2}\right) = \left(\mu_{\mathrm{J}} + \sigma_{\mathrm{J}}^2 i_{\mathrm{J}}\right) \frac{\partial i_{\mathrm{J}}}{\partial \boldsymbol{\theta}}$$
$$= \left(\mu_{\mathrm{J}} + \frac{\sigma_{\mathrm{J}}^2 \log|\mathbf{I}|}{2}\right) \frac{1}{2} \frac{\partial \log|\mathbf{I}|}{\partial \boldsymbol{\theta}}. \quad (7.2)$$

In (7.2), $\frac{\sigma_{\mathrm{J}}^2 \log|\mathbf{I}|}{4} \frac{\partial \log|\mathbf{I}|}{\partial \boldsymbol{\theta}}$ is added to $\mathbf{q}^*$ with non-stochastic $k = \mu_{\mathrm{J}}$.

Appendix

A.1 Asymptotic cumulants of the MLEs for the canonical parameters in the exponential family under possible model misspecification

Recall $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_0 = \sum_{i=1}^{3} \boldsymbol{\Lambda}^{(i)} \mathbf{l}_0^{(i)} + O_p(n^{-2})$.

Define $\mathbf{I}_0^{(\mathrm{D1})} \equiv -\frac{\partial^3 \bar{l}}{\partial \boldsymbol{\theta}_0 (\partial \boldsymbol{\theta}_0')^{<2>}}$ and $\mathbf{I}_0^{(\mathrm{D2})} \equiv -\frac{\partial^4 \bar{l}}{\partial \boldsymbol{\theta}_0 (\partial \boldsymbol{\theta}_0')^{<3>}}$. Then,

$$\begin{aligned}
\boldsymbol{\Lambda}^{(1)} \mathbf{l}_0^{(1)} &= -\boldsymbol{\Lambda}^{-1} \frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0} = \mathbf{I}_0^{-1} \frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0} = \mathbf{I}_0^{-1} \{\bar{\mathbf{s}} - \mathrm{E}_{\mathrm{T}}(\bar{\mathbf{s}})\}, \\
\boldsymbol{\Lambda}^{(2)} \mathbf{l}_0^{(2)} &= -\frac{1}{2} \mathbf{I}_0^{-1} \mathbf{I}_0^{(\mathrm{D1})} (\mathbf{I}_0^{-1} \mathbf{l}_0^{(1)})^{<2>} \quad (\mathbf{l}_0^{(2)} = (\mathbf{l}_0^{(1)})^{<2>}), \\
\boldsymbol{\Lambda}^{(3)} \mathbf{l}_0^{(3)} &= \frac{1}{2} \mathbf{I}_0^{-1} \mathbf{I}_0^{(\mathrm{D1})} [(\mathbf{I}_0^{-1} \mathbf{l}_0^{(1)}) \otimes \{\mathbf{I}_0^{-1} \mathbf{I}_0^{(\mathrm{D1})} (\mathbf{I}_0^{-1} \mathbf{l}_0^{(1)})^{<2>}\}] \\
&\quad - \frac{1}{6} \mathbf{I}_0^{-1} \mathbf{I}_0^{(\mathrm{D2})} (\mathbf{I}_0^{-1} \mathbf{l}_0^{(1)})^{<3>} \quad (\mathbf{l}_0^{(3)} = (\mathbf{l}_0^{(1)})^{<3>}),
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\alpha}_{\mathrm{ML1}} &= n\mathrm{E}_{\mathrm{T}}(\boldsymbol{\Lambda}^{(2)} \mathbf{l}_0^{(2)}) = -\frac{1}{2} \mathbf{I}_0^{-1} \mathbf{I}_0^{(\mathrm{D1})} \mathrm{vec}\{ n\mathrm{E}_{\mathrm{T}}(\mathbf{I}_0^{-1} \mathbf{l}_0^{(1)} \mathbf{l}_0^{(1)\prime} \mathbf{I}_0^{-1})\} \\
&\equiv -\frac{1}{2} \mathbf{I}_0^{-1} \mathbf{I}_0^{(\mathrm{D1})} \mathrm{vec}(\mathbf{A}_{\mathrm{ML2}}),
\end{aligned}$$

where $\bar{\mathbf{s}}$ is the $q \times 1$ vector of sufficient statistics averaged over observations and $\mathrm{vec}(\cdot)$ is a vectorizing operator stacking columns of a matrix. Recall (2.5):

$$\begin{aligned}
\mathbf{A}_{\mathrm{ML}\Delta 2} = n^2 \mathrm{E}_{\mathrm{T}} (&\boldsymbol{\Lambda}^{(1)} \mathbf{l}_0^{(1)} \mathbf{l}_0^{(2)\prime} \boldsymbol{\Lambda}^{(2)\prime} + \boldsymbol{\Lambda}^{(2)} \mathbf{l}_0^{(2)} \mathbf{l}_0^{(1)\prime} \boldsymbol{\Lambda}^{(1)} + \boldsymbol{\Lambda}^{(2)} \mathbf{l}_0^{(2)} \mathbf{l}_0^{(2)\prime} \boldsymbol{\Lambda}^{(2)\prime} \\
&+ \boldsymbol{\Lambda}^{(1)} \mathbf{l}_0^{(1)} \mathbf{l}_0^{(3)\prime} \boldsymbol{\Lambda}^{(3)\prime} + \boldsymbol{\Lambda}^{(3)} \mathbf{l}_0^{(3)} \mathbf{l}_0^{(1)\prime} \boldsymbol{\Lambda}^{(1)}) - \boldsymbol{\alpha}_{\mathrm{ML1}} \boldsymbol{\alpha}_{\mathrm{ML1}}',
\end{aligned}$$

where

$$\begin{aligned}
&n^2 \mathrm{E}_{\mathrm{T}}(\boldsymbol{\Lambda}^{(2)} \mathbf{l}_0^{(2)} \mathbf{l}_0^{(1)\prime} \boldsymbol{\Lambda}^{(1)}) = \boldsymbol{\Lambda}^{(2)} n^2 \mathrm{E}_{\mathrm{T}}(\mathbf{l}_0^{(1)^{<2>}} \mathbf{l}_0^{(1)\prime}) \boldsymbol{\Lambda}^{(1)} \\
&\quad = \boldsymbol{\Lambda}^{(2)} n^2 \kappa_3(\bar{\mathbf{s}}^{<2>}, \bar{\mathbf{s}}') \mathbf{I}_0^{-1} = -\frac{1}{2} \mathbf{I}_0^{-1} \mathbf{I}_0^{(\mathrm{D1})} \mathbf{I}_0^{-1<2>} n^2 \kappa_3(\bar{\mathbf{s}}^{<2>}, \bar{\mathbf{s}}') \mathbf{I}_0^{-1}, \\
&n^2 \mathrm{E}_{\mathrm{T}}(\boldsymbol{\Lambda}^{(2)} \mathbf{l}_0^{(2)} \mathbf{l}_0^{(2)\prime} \boldsymbol{\Lambda}^{(2)\prime}) - \boldsymbol{\alpha}_{\mathrm{ML1}} \boldsymbol{\alpha}_{\mathrm{ML1}}' \\
&\quad = \frac{1}{4} \mathbf{I}_0^{-1} \mathbf{I}_0^{(\mathrm{D1})} n^2 \mathrm{E}_{\mathrm{T}} \{(\mathbf{I}_0^{-1} \mathbf{l}_0^{(1)})^{<2>} (\mathbf{l}_0^{(1)\prime} \mathbf{I}_0^{-1})^{<2>}\} \mathbf{I}_0^{(\mathrm{D1})\prime} \mathbf{I}_0^{-1} - \boldsymbol{\alpha}_{\mathrm{ML1}} \boldsymbol{\alpha}_{\mathrm{ML1}}' \\
&\quad = \frac{1}{2} \mathbf{I}_0^{-1} \mathbf{I}_0^{(\mathrm{D1})} \mathbf{A}_{\mathrm{ML2}}^{<2>} \mathbf{I}_0^{(\mathrm{D1})\prime} \mathbf{I}_0^{-1} + O(n^{-1}), \\
&n^2 \mathrm{E}_{\mathrm{T}}(\boldsymbol{\Lambda}^{(3)} \mathbf{l}_0^{(3)} \mathbf{l}_0^{(1)\prime} \boldsymbol{\Lambda}^{(1)}) = \boldsymbol{\Lambda}^{(3)} n^2 \mathrm{E}_{\mathrm{T}}(\mathbf{l}_0^{(1)^{<3>}} \mathbf{l}_0^{(1)\prime}) \mathbf{I}_0^{-1} \\
&\quad = \sum_{a,b,c,d=1}^{q} (\boldsymbol{\Lambda}^{(3)})_{\cdot abc} \{(\mathbf{A}_{\mathrm{ML2}})_{ab} (\mathbf{A}_{\mathrm{ML2}})_{cd} + (\mathbf{A}_{\mathrm{ML2}})_{ac} (\mathbf{A}_{\mathrm{ML2}})_{bd} \\
&\qquad + (\mathbf{A}_{\mathrm{ML2}})_{bc} (\mathbf{A}_{\mathrm{ML2}})_{ad}\} (\mathbf{I}_0^{-1})_{d\cdot} + O(n^{-1}) \\
&\quad = \frac{1}{2} \mathbf{I}_0^{-1} \mathbf{I}_0^{(\mathrm{D1})} n^2 \mathrm{E}_{\mathrm{T}} \left([(\mathbf{I}_0^{-1} \mathbf{l}_0^{(1)}) \otimes \{\mathbf{I}_0^{-1} \mathbf{I}_0^{(\mathrm{D1})} (\mathbf{I}_0^{-1} \mathbf{l}_0^{(1)})^{<2>}\}] \mathbf{l}_0^{(1)\prime} \mathbf{I}_0^{-1} \right)
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{6}\mathbf{I}_0^{-1}\mathbf{I}_0^{(\mathrm{D2})} n^2 \mathrm{E}_{\mathrm{T}}\{(\mathbf{I}_0^{-1}\mathbf{l}_0^{(1))^{<3>}}\mathbf{l}_0^{(1)'}\mathbf{I}_0^{-1}\} \\
&= \frac{1}{2}\mathbf{I}_0^{-1}\mathbf{I}_0^{(\mathrm{D1})}[\mathbf{A}_{\mathrm{ML2}} \otimes \{\mathbf{I}_0^{-1}\mathbf{I}_0^{(\mathrm{D1})}\mathrm{vec}(\mathbf{A}_{\mathrm{ML2}})\}] \\
&\quad + \mathbf{I}_0^{-1}\mathbf{I}_0^{(\mathrm{D1})} \sum_{a,b=1}^{q} \{(\mathbf{A}_{\mathrm{ML2}})_{\cdot a} \otimes (\mathbf{I}_0^{-1}\mathbf{I}_0^{(\mathrm{D1})})_{\cdot ab}\}(\mathbf{A}_{\mathrm{ML2}})_{b\cdot} \\
&\quad - \frac{1}{2}\mathbf{I}_0^{-1}\mathbf{I}_0^{(\mathrm{D2})}\{\mathbf{A}_{\mathrm{ML2}} \otimes \mathrm{vec}(\mathbf{A}_{\mathrm{ML2}})\} + O(n^{-1}) \\
&= \mathbf{I}_0^{-1}\mathbf{I}_0^{(\mathrm{D1})}\left[\frac{1}{2}\mathbf{A}_{\mathrm{ML2}} \otimes \{\mathbf{I}_0^{-1}\mathbf{I}_0^{(\mathrm{D1})}\mathrm{vec}(\mathbf{A}_{\mathrm{ML2}})\}\right. \\
&\qquad\qquad \left. + \{\mathbf{I}_{(q)}^{*} \otimes (\mathbf{I}_0^{-1}\mathbf{I}_0^{(\mathrm{D1})})\}\{\mathrm{vec}(\mathbf{A}_{\mathrm{ML2}}) \otimes \mathbf{A}_{\mathrm{ML2}}\}\right] \\
&\quad - \frac{1}{2}\mathbf{I}_0^{-1}\mathbf{I}_0^{(\mathrm{D2})}\{\mathbf{A}_{\mathrm{ML2}} \otimes \mathrm{vec}(\mathbf{A}_{\mathrm{ML2}})\} + O(n^{-1}),
\end{aligned}
$$

where $(\cdot)_{d\cdot}$ and $(\cdot)_{\cdot a}$ are the d-th row and the a-th column of a matrix, respectively; $(\cdot)_{\cdot ab}$ and $(\cdot)_{\cdot abc}$ are defined similarly; and $\mathbf{I}_{(q)}^{*}$ is the $q \times q$ identity matrix.

$$
\begin{aligned}
\alpha_{\mathrm{ML3}} &= n^2\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(1)'}\mathbf{l}_0^{(1)})^3\} + 3n^2\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(1)'}\mathbf{l}_0^{(1)})^2\boldsymbol{\lambda}^{(2)'}\mathbf{l}_0^{(2)}\} - 3\alpha_{\mathrm{ML1}}\alpha_{\mathrm{ML2}} + O(n^{-1}) \\
&= (\mathbf{I}_0^{-1})_{\theta\cdot}^{<3>} n^2 \mathrm{E}_{\mathrm{T}}(\mathbf{l}_0^{(1)^{<3>}}) - 3(\mathbf{A}_{\mathrm{ML2}})_{\theta\cdot}^{<2>}\mathbf{I}_0^{(\mathrm{D1})'}(\mathbf{I}_0^{-1})_{\cdot\theta} \\
&= (\mathbf{I}_0^{-1})_{\theta\cdot}^{<3>} n^2 \kappa_3(\bar{\mathbf{s}}) - 3(\mathbf{A}_{\mathrm{ML2}})_{\theta\cdot}^{<2>}\mathbf{I}_0^{(\mathrm{D1})'}(\mathbf{I}_0^{-1})_{\cdot\theta}
\end{aligned}
$$

where $\boldsymbol{\lambda}^{(i)'}$ is the row of $\boldsymbol{\Lambda}^{(i)} (i = 1, 2)$ corresponding to θ_0; and $(\cdot)_{\theta\cdot}$ and $(\cdot)_{\cdot\theta}$ are the row and the column of a matrix corresponding to θ_0, respectively.

$$
\begin{aligned}
\alpha_{\mathrm{ML4}} &= n^3\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(1)'}\mathbf{l}_0^{(1)})^4\} + 4n^3\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(1)'}\mathbf{l}_0^{(1)})^3\boldsymbol{\lambda}^{(2)'}\mathbf{l}_0^{(2)}\} \\
&\quad + 6n^3\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(1)'}\mathbf{l}_0^{(1)})^2(\boldsymbol{\lambda}^{(2)'}\mathbf{l}_0^{(2)})^2\} + 4n^3\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(1)'}\mathbf{l}_0^{(1)})^3\boldsymbol{\lambda}^{(3)'}\mathbf{l}_0^{(3)}\} \\
&\quad - 3n\alpha_{\mathrm{ML2}}^2 - 4\alpha_{\mathrm{ML1}}\alpha_{\mathrm{ML3}} - 6\alpha_{\mathrm{ML2}}\alpha_{\mathrm{ML}\Delta 2} - 6\alpha_{\mathrm{ML1}}^2\alpha_{\mathrm{ML2}} + O(n^{-1}),
\end{aligned}
$$

where $\boldsymbol{\lambda}^{(3)}$ is defined similarly to $\boldsymbol{\lambda}^{(i)} (i = 1, 2)$,

$$
\begin{aligned}
&n^3\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(1)'}\mathbf{l}_0^{(1)})^4\} - 3n\alpha_{\mathrm{ML2}}^2 = (\mathbf{I}_0^{-1})_{\theta\cdot}^{<4>} n^3\kappa_4(\bar{\mathbf{s}}), \\
&n^3\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(1)'}\mathbf{l}_0^{(1)})^3\boldsymbol{\lambda}^{(2)'}\mathbf{l}_0^{(2)}\} \\
&\quad = \sum_{a,b,c,d,e=1}^{q} (\boldsymbol{\lambda}^{(1)})_a(\boldsymbol{\lambda}^{(1)})_b(\boldsymbol{\lambda}^{(1)})_c(\boldsymbol{\lambda}^{(2)})_{de} n^3\mathrm{E}_{\mathrm{T}}\{(\mathbf{l}_0^{(1)})_a(\mathbf{l}_0^{(1)})_b(\mathbf{l}_0^{(1)})_c(\mathbf{l}_0^{(1)})_d(\mathbf{l}_0^{(1)})_e\} \\
&\quad = \left\{\sum_{a,b,c,d,e=1}^{q} (\boldsymbol{\lambda}^{(1)})_a(\boldsymbol{\lambda}^{(1)})_b(\boldsymbol{\lambda}^{(1)})_c(\boldsymbol{\lambda}^{(2)})_{de}\right\} \\
&\qquad \times \sum_{(a,b,c,d,e)}^{10} n\kappa_2\{(\bar{\mathbf{s}})_a, (\bar{\mathbf{s}})_b\} n^2\kappa_3\{(\bar{\mathbf{s}})_c, (\bar{\mathbf{s}})_d, (\bar{\mathbf{s}})_e\} + O(n^{-1}) \\
&\quad = (\mathbf{I}_0^{-1})_{\theta\cdot}^{<3>} n^2\kappa_3(\bar{\boldsymbol{s}})\alpha_{\mathrm{ML1}} \ (\#1)
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{3}{2}\alpha_{\mathrm{ML2}}(\mathbf{I}_0^{-1})_{\theta\cdot}n^2\kappa_3(\bar{\mathbf{s}},\bar{\mathbf{s}}'^{<2>})\mathbf{I}_0^{-1<2>}\mathbf{I}_0^{(\mathrm{D1})'}(\mathbf{I}_0^{-1})_{\cdot\theta} \ (\#3)\\
&- 3\{(\mathbf{A}_{\mathrm{ML2}})_{\theta\cdot}\otimes((\mathbf{I}_0^{-1})_{\theta\cdot}^{<2>}n^2\kappa_3(\bar{\mathbf{s}}^{<2>},\bar{\mathbf{s}}')\mathbf{I}_0^{-1})\}\mathbf{I}_0^{(\mathrm{D1})'}(\mathbf{I}_0^{-1})_{\cdot\theta}(\#6)+O(n^{-1}),
\end{aligned}
$$

where $\sum_{(a,b,c,d,e)}^{10}$ denotes that the sum of 10 terms considering combinations for a, b, c, d and e; and $(\#\ \cdot)$ indicates the number of terms before summation,

$$
\begin{aligned}
&n^3\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(1)'}\mathbf{l}_0^{(1)})^2(\boldsymbol{\lambda}^{(2)'}\mathbf{l}_0^{(2)})^2\}\\
&\quad=\sum_{a,b,c,d,e,f=1}^{q}(\boldsymbol{\lambda}^{(1)})_a(\boldsymbol{\lambda}^{(1)})_b(\boldsymbol{\lambda}^{(2)})_{cd}(\boldsymbol{\lambda}^{(2)})_{ef}\\
&\qquad\times n^3\mathrm{E}_{\mathrm{T}}\{(\mathbf{l}_0^{(1)})_a(\mathbf{l}_0^{(1)})_b(\mathbf{l}_0^{(1)})_c(\mathbf{l}_0^{(1)})_d(\mathbf{l}_0^{(1)})_e(\mathbf{l}_0^{(1)})_f\}\\
&\quad=\sum_{a,b,c,d,e,f=1}^{q}(\boldsymbol{\lambda}^{(1)})_a(\boldsymbol{\lambda}^{(1)})_b(\boldsymbol{\lambda}^{(2)})_{cd}(\boldsymbol{\lambda}^{(2)})_{ef}\\
&\qquad\times\sum_{(a,b,c,d,e,f)}^{15}n\kappa_2\{(\bar{\mathbf{s}})_a,(\bar{\mathbf{s}})_b\}n\kappa_2\{(\bar{\mathbf{s}})_c,(\bar{\mathbf{s}})_d\}n\kappa_2\{(\bar{\mathbf{s}})_e,(\bar{\mathbf{s}})_f\}+O(n^{-1})\\
&\quad=\alpha_{\mathrm{ML2}}\left\{\alpha_{\mathrm{ML1}}^2+\frac{1}{2}(\mathbf{I}_0^{-1})_{\theta\cdot}\mathbf{I}_0^{(\mathrm{D1})}\mathbf{A}_{\mathrm{ML2}}^{<2>}\mathbf{I}_0^{(\mathrm{D1})'}(\mathbf{I}_0^{-1})_{\cdot\theta}\right\}\ (\#3)\\
&\qquad-2\alpha_{\mathrm{ML1}}(\mathbf{A}_{\mathrm{ML2}})_{\theta\cdot}^{<2>}\mathbf{I}_0^{(\mathrm{D1})'}(\mathbf{I}_0^{-1})_{\cdot\theta}\ (\#4)\\
&\qquad+2(\mathbf{I}_0^{-1})_{\theta\cdot}\mathbf{I}_0^{(\mathrm{D1})}[\{(\mathbf{A}_{\mathrm{ML2}})_{\cdot\theta}(\mathbf{A}_{\mathrm{ML2}})_{\theta\cdot}\}\otimes\mathbf{A}_{\mathrm{ML2}}]\mathbf{I}_0^{(\mathrm{D1})'}(\mathbf{I}_0^{-1})_{\cdot\theta}\ (\#8)+O(n^{-1}),
\end{aligned}
$$

$$
\begin{aligned}
&n^3\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(1)'}\mathbf{l}_0^{(1)})^3\boldsymbol{\lambda}^{(3)'}\mathbf{l}_0^{(3)}\}\\
&\quad=\sum_{a,b,c,d,e,f=1}^{q}(\boldsymbol{\lambda}^{(1)})_a(\boldsymbol{\lambda}^{(1)})_b(\boldsymbol{\lambda}^{(1)})_c(\boldsymbol{\lambda}^{(3)})_{def}\\
&\qquad\times n^3\mathrm{E}_{\mathrm{T}}\{(\mathbf{l}_0^{(1)})_a(\mathbf{l}_0^{(1)})_b(\mathbf{l}_0^{(1)})_c(\mathbf{l}_0^{(1)})_d(\mathbf{l}_0^{(1)})_e(\mathbf{l}_0^{(1)})_f\}\\
&\quad=\sum_{a,b,c,d,e,f=1}^{q}(\boldsymbol{\lambda}^{(1)})_a(\boldsymbol{\lambda}^{(1)})_b(\boldsymbol{\lambda}^{(1)})_c(\boldsymbol{\lambda}^{(3)})_{def}\\
&\qquad\times\sum_{(a,b,c,d,e,f)}^{15}n\kappa_2\{(\bar{\mathbf{s}})_a,(\bar{\mathbf{s}})_b\}n\kappa_2\{(\bar{\mathbf{s}})_c,(\bar{\mathbf{s}})_d\}n\kappa_2\{(\bar{\mathbf{s}})_e,(\bar{\mathbf{s}})_f\}+O(n^{-1})\\
&\quad=-3\alpha_{\mathrm{ML2}}\{(\mathbf{A}_{\mathrm{ML2}})_{\theta\cdot}\otimes\boldsymbol{\alpha}'_{\mathrm{ML1}}\}\mathbf{I}_0^{(\mathrm{D1})'}(\mathbf{I}_0^{-1})_{\cdot\theta}\ (\#3)\\
&\qquad+3\alpha_{\mathrm{ML2}}\{\mathrm{vec}'(\mathbf{A}_{\mathrm{ML2}})\otimes(\mathbf{A}_{\mathrm{ML2}})_{\theta\cdot}\}\{\mathbf{I}_{(q)}^*\otimes(\mathbf{I}_0^{(\mathrm{D1})'}\mathbf{I}_0^{-1})\}\mathbf{I}_0^{(\mathrm{D1})'}(\mathbf{I}_0^{-1})_{\cdot\theta}\ (\#6)\\
&\qquad+3[(\mathbf{A}_{\mathrm{ML2}})_{\theta\cdot}\otimes\{(\mathbf{A}_{\mathrm{ML2}})_{\theta\cdot}^{<2>}\mathbf{I}_0^{(\mathrm{D1})'}\mathbf{I}_0^{-1}\}]\mathbf{I}_0^{(\mathrm{D1})'}(\mathbf{I}_0^{-1})_{\cdot\theta}\ (\#6)\\
&\qquad-\frac{3}{2}\alpha_{\mathrm{ML2}}\{(\mathbf{A}_{\mathrm{ML2}})_{\theta\cdot}\otimes\mathrm{vec}'(\mathbf{A}_{\mathrm{ML2}})\}\mathbf{I}_0^{(\mathrm{D2})'}(\mathbf{I}_0^{-1})_{\cdot\theta}\ (\#9)\\
&\qquad-(\mathbf{A}_{\mathrm{ML2}})_{\theta\cdot}^{<3>}\mathbf{I}_0^{(\mathrm{D2})'}(\mathbf{I}_0^{-1})_{\cdot\theta}\ (\#6)+O(n^{-1}).
\end{aligned}
$$

A.2 Asymptotic cumulants of the studentized estimators for the canonical parameters in the exponential family under possible model misspecification

A.2.1 MLE

Define $t_{\mathrm{ML}} \equiv n^{1/2}\{(\hat{\mathbf{I}}_{\mathrm{ML}}^{-1})_{\theta\theta}\}^{-1/2}(\hat{\theta}_{\mathrm{ML}} - \theta_0) \equiv n^{1/2}(\hat{\bar{i}}_{\mathrm{ML}}^{\theta\theta})^{-1/2}(\hat{\theta}_{\mathrm{ML}} - \theta_0)$, where $(\cdot)_{\theta\theta}$ is the diagonal element corresponding to θ in a matrix. Expand

$$
\begin{aligned}
(\hat{\bar{i}}_{\mathrm{ML}}^{\theta\theta})^{-1/2} &= (\bar{i}_0^{\theta\theta})^{-1/2} + \frac{1}{2}\sum_{a=1}^{q}\left(\mathbf{I}_0^{-1}\frac{\partial \mathbf{I}_0}{\partial(\boldsymbol{\theta}_0)_a}\mathbf{I}_0^{-1}\right)_{\theta\theta}(\bar{i}_0^{\theta\theta})^{-3/2}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_0)_a \\
&\quad + \sum_{a,b=1}^{q}\left\{\left(\frac{1}{4}\mathbf{I}_0^{-1}\frac{\partial^2 \mathbf{I}_0}{\partial(\boldsymbol{\theta}_0)_a\partial(\boldsymbol{\theta}_0)_b}\mathbf{I}_0^{-1} - \frac{1}{2}\mathbf{I}_0^{-1}\frac{\partial \mathbf{I}_0}{\partial(\boldsymbol{\theta}_0)_a}\mathbf{I}_0^{-1}\frac{\partial \mathbf{I}_0}{\partial(\boldsymbol{\theta}_0)_b}\mathbf{I}_0^{-1}\right)_{\theta\theta}(\bar{i}_0^{\theta\theta})^{-3/2}\right. \\
&\quad \left. + \frac{3}{8}\left(\mathbf{I}_0^{-1}\frac{\partial \mathbf{I}_0}{\partial(\boldsymbol{\theta}_0)_a}\mathbf{I}_0^{-1}\right)_{\theta\theta}\left(\mathbf{I}_0^{-1}\frac{\partial \mathbf{I}_0}{\partial(\boldsymbol{\theta}_0)_b}\mathbf{I}_0^{-1}\right)_{\theta\theta}(\bar{i}_0^{\theta\theta})^{-5/2}\right\}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_0)_a(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_0)_b \\
&\quad + O_p(n^{-3/2}) \\
&\equiv (\bar{i}_0^{\theta\theta})^{-1/2} + \mathbf{i}_0^{(1)'}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_0) + \mathbf{i}_0^{(2)'}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_0)^{<2>} + O_p(n^{-3/2}) \\
&= (\bar{i}_0^{\theta\theta})^{-1/2} + \mathbf{i}_0^{(1)'}\boldsymbol{\Lambda}^{(1)}\mathbf{l}_0^{(1)} + \{\mathbf{i}_0^{(1)'}\boldsymbol{\Lambda}^{(2)}\mathbf{l}_0^{(2)} + \mathbf{i}_0^{(2)'}(\boldsymbol{\Lambda}^{(1)}\mathbf{l}_0^{(1)})^{<2>}\} + O_p(n^{-3/2}) \\
&= (\bar{i}_0^{\theta\theta})^{-1/2} + \mathbf{i}_0^{(1)'}\mathbf{I}_0^{-1}\mathbf{l}_0^{(1)} + (\mathbf{i}_0^{(1)'}\boldsymbol{\Lambda}^{(2)} + \mathbf{i}_0^{(2)'}\mathbf{I}_0^{-1<2>})\mathbf{l}_0^{(1)<2>} + O_p(n^{-3/2}).
\end{aligned}
$$

Then,

$$
\begin{aligned}
t_{\mathrm{ML}} &= n^{1/2}\sum_{j=1}^{3}\boldsymbol{\lambda}^{(j)'}\mathbf{l}_0^{(j)}\{(\bar{i}_0^{\theta\theta})^{-1/2} + \mathbf{i}_0^{(1)'}\mathbf{I}_0^{-1}\mathbf{l}_0^{(1)} + (\mathbf{i}_0^{(1)'}\boldsymbol{\Lambda}^{(2)} + \mathbf{i}_0^{(2)'}\mathbf{I}_0^{-1<2>})\mathbf{l}_0^{(1)<2>}\} \\
&\quad + O_p(n^{-1}) \\
&\equiv n^{1/2}\sum_{j=1}^{3}\boldsymbol{\lambda}^{(tj)'}\mathbf{l}_0^{(j)} + O_p(n^{-1}),
\end{aligned}
$$

where

$$
\begin{aligned}
&\boldsymbol{\lambda}^{(t1)} = \boldsymbol{\lambda}^{(1)}(\bar{i}_0^{\theta\theta})^{-1/2}, \boldsymbol{\lambda}^{(t2)} = \{\boldsymbol{\lambda}^{(2)'}(\bar{i}_0^{\theta\theta})^{-1/2} + \boldsymbol{\lambda}^{(1)'} \otimes (\mathbf{i}_0^{(1)'}\mathbf{I}_0^{-1})\}', \\
&\boldsymbol{\lambda}^{(t3)} = \{\boldsymbol{\lambda}^{(3)'}(\bar{i}_0^{\theta\theta})^{-1/2} + \boldsymbol{\lambda}^{(2)'} \otimes (\mathbf{i}_0^{(1)'}\mathbf{I}_0^{-1}) + \boldsymbol{\lambda}^{(1)'} \otimes (\mathbf{i}_0^{(1)'}\boldsymbol{\Lambda}^{(2)} + \mathbf{i}_0^{(2)'}\mathbf{I}_0^{-1<2>})\}'.
\end{aligned}
$$

From the above and previous results,

$$
\begin{aligned}
\kappa_1(t_{\mathrm{ML}}) &= n^{-1/2}\boldsymbol{\lambda}^{(t2)'}n\mathrm{E}_{\mathrm{T}}(\mathbf{l}_0^{(2)}) + O(n^{-3/2}) \\
&= n^{-1/2}\{(\bar{i}_0^{\theta\theta})^{-1/2}\alpha_{\mathrm{ML1}} + \boldsymbol{\lambda}^{(1)'}\mathbf{A}_{\mathrm{ML2}}\mathbf{I}_0^{-1}\mathbf{i}_0^{(1)}\} + O(n^{-3/2}) \\
&\equiv n^{-1/2}\alpha_{\mathrm{ML1}}^{(t)} + O(n^{-3/2}), \\
\kappa_2(t_{\mathrm{ML}}) &= \boldsymbol{\lambda}^{(t1)'}n\mathrm{E}_{\mathrm{T}}(\mathbf{l}_0^{(1)}\mathbf{l}_0^{(1)'})\boldsymbol{\lambda}^{(t1)} + n^{-1}[\boldsymbol{\lambda}^{(t2)'}n^2\mathrm{E}_{\mathrm{T}}(\mathbf{l}_0^{(2)}\mathbf{l}_0^{(2)'})\boldsymbol{\lambda}^{(t2)} \\
&\quad + 2\boldsymbol{\lambda}^{(t1)'}n^2\mathrm{E}_{\mathrm{T}}(\mathbf{l}_0^{(1)}\mathbf{l}_0^{(2)'})\boldsymbol{\lambda}^{(t2)} + 2\boldsymbol{\lambda}^{(t1)'}n^2\mathrm{E}_{\mathrm{T}}(\mathbf{l}_0^{(1)}\mathbf{l}_0^{(3)'})\boldsymbol{\lambda}^{(t3)}] + O(n^{-2})
\end{aligned}
$$

$$
\begin{aligned}
&\equiv \alpha_{\mathrm{ML}2}^{(t)} + n^{-1}\alpha_{\mathrm{ML}\Delta 2}^{(t)} + O(n^{-1/2}),\\
\kappa_3(t_{\mathrm{ML}}) &= n^{-1/2}[n^2\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(t1)'}\mathbf{l}_0^{(1)})^3\} + 3n^2\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(t1)'}\mathbf{l}_0^{(1)})^2\boldsymbol{\lambda}^{(t2)'}\mathbf{l}_0^{(2)}\} - 3\alpha_{\mathrm{ML}1}^{(t)}\alpha_{\mathrm{ML}2}^{(t)}]\\
&\quad + O(n^{-3/2})\\
&\equiv n^{-1/2}\alpha_{\mathrm{ML}1}^{(t)} + O(n^{-3/2}),\\
\kappa_4(t_{\mathrm{ML}}) &= \mathrm{E}_{\mathrm{T}}(t_{\mathrm{ML}1}^4) - 3\{\alpha_{\mathrm{ML}1}^{(t)}\}^2 - n^{-1}\{4\alpha_{\mathrm{ML}1}^{(t)}\alpha_{\mathrm{ML}3}^{(t)} + 6\alpha_{\mathrm{ML}2}^{(t)}\alpha_{\mathrm{ML}\Delta 2}^{(t)} + 6\{\alpha_{\mathrm{ML}1}^{(t)}\}^2\alpha_{\mathrm{ML}2}^{(t)}\}\\
&\quad + O(n^{-2})\\
&= n^{-1}[n^3\boldsymbol{\lambda}^{(t1)'<4>}\kappa_4(\mathbf{l}_0^{(1)}) + 4n^3\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(t1)'}\mathbf{l}_0^{(1)})^3\boldsymbol{\lambda}^{(t2)'}\mathbf{l}_0^{(2)}\}\\
&\quad + 6n^3\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(t1)'}\mathbf{l}_0^{(1)})^2(\boldsymbol{\lambda}^{(t2)'}\mathbf{l}_0^{(2)})\} + 4n^3\mathrm{E}_{\mathrm{T}}\{(\boldsymbol{\lambda}^{(t1)'}\mathbf{l}_0^{(1)})^3\boldsymbol{\lambda}^{(t3)'}\mathbf{l}_0^{(3)}\}\\
&\quad - 4\alpha_{\mathrm{ML}1}^{(t)}\alpha_{\mathrm{ML}3}^{(t)} - 6\alpha_{\mathrm{ML}2}^{(t)}\alpha_{\mathrm{ML}\Delta 2}^{(t)} - 6\{\alpha_{\mathrm{ML}1}^{(t)}\}^2\alpha_{\mathrm{ML}2}^{(t)}] + O(n^{-2})\\
&\equiv n^{-1}\alpha_{\mathrm{ML}4}^{(t)} + O(n^{-2}).
\end{aligned}
$$

A.2.2 Estimators by the weighted score method

Define $t_{\mathrm{W}} \equiv n^{1/2}(\hat{\bar{i}}_{\mathrm{W}}^{\theta\theta})^{-1/2}(\hat{\theta}_{\mathrm{W}} - \theta_0)$, where

$$
\hat{\theta}_{\mathrm{W}} - \theta_0 = n^{-1}(\mathbf{I}_0^{-1}\mathbf{q}_0^*)_\theta + \sum_{j=1}^{3}\boldsymbol{\lambda}^{(j)'}\mathbf{l}_0^{(j)} + n^{-1}\{(\hat{\mathbf{I}}_{\mathrm{W}}^{-1}\hat{\mathbf{q}}_{\mathrm{W}}^* - \mathbf{I}_0^{-1}\mathbf{q}_0^*)_\theta\}_{O_p(n^{-1/2})} + O_p(n^{-2}).
$$

Expand

$$
\begin{aligned}
(\hat{\bar{i}}_{\mathrm{W}}^{\theta\theta})^{-1/2} &= (\bar{i}_0^{\theta\theta})^{-1/2} + \mathbf{i}_0^{(1)'}(\hat{\boldsymbol{\theta}}_{\mathrm{W}} - \boldsymbol{\theta}_0) + \mathbf{i}_0^{(2)'}(\hat{\boldsymbol{\theta}}_{\mathrm{W}} - \boldsymbol{\theta}_0)^{<2>} + O_p(n^{-3/2})\\
&= (\bar{i}_0^{\theta\theta})^{-1/2} + \mathbf{i}_0^{(1)'}\mathbf{I}_0^{-1}\mathbf{l}_0^{(1)} + n^{-1}\mathbf{i}_0^{(1)'}\mathbf{I}_0^{-1}\mathbf{q}_0^* + (\mathbf{i}_0^{(1)'}\boldsymbol{\Lambda}^{(2)} + \mathbf{i}_0^{(2)'}\mathbf{I}_0^{-1<2>})\mathbf{l}_0^{(1)<2>}\\
&\quad + O_p(n^{-3/2}).
\end{aligned}
$$

Recalling $\hat{\theta}_{\mathrm{ML}} - \theta_0 = \sum_{j=1}^{3}\boldsymbol{\lambda}^{(j)'}\mathbf{l}_0^{(j)} + O_p(n^{-2})$ and $t_{\mathrm{ML}} = n^{1/2}\sum_{j=1}^{3}\boldsymbol{\lambda}^{(tj)'}\mathbf{l}_0^{(j)} + O_p(n^{-3/2})$,

$$
\begin{aligned}
t_{\mathrm{W}} &= n^{1/2}\sum_{j=1}^{3}\boldsymbol{\lambda}^{(tj)'}\mathbf{l}_0^{(j)} + n^{-1/2}(\bar{i}_0^{\theta\theta})^{-1/2}(\mathbf{I}_0^{-1}\mathbf{q}_0^*)_\theta + n^{-1/2}(\bar{i}_0^{\theta\theta})^{-1/2}(\hat{\mathbf{I}}_{\mathrm{W}}^{-1}\hat{\mathbf{q}}_{\mathrm{W}}^* - \mathbf{I}_0^{-1}\mathbf{q}_0^*)_\theta\\
&\quad + n^{-1/2}(\mathbf{I}_0^{-1}\mathbf{q}_0^*)_\theta\mathbf{i}_0^{(1)'}\mathbf{I}_0^{-1}\mathbf{l}_0^{(1)} + n^{-1/2}\mathbf{i}_0^{(1)'}\mathbf{I}_0^{-1}\mathbf{q}_0^*(\mathbf{I}_0^{-1})_\theta.\mathbf{l}_0^{(1)} + O_p(n^{-3/2}),\\
\kappa_1(t_{\mathrm{W}}) &= n^{-1/2}\{\alpha_{\mathrm{ML}1}^{(t)} + (\bar{i}_0^{\theta\theta})^{-1/2}(\mathbf{I}_0^{-1}\mathbf{q}_0^*)_\theta\} + O(n^{-3/2}) \equiv n^{-1/2}\alpha_{\mathrm{W}1}^{(t)} + O(n^{-3/2}),\\
\kappa_2(t_{\mathrm{W}}) &= \alpha_{\mathrm{ML}2}^{(t)} + n^{-1}[\alpha_{\mathrm{ML}\Delta 2}^{(t)} + 2(\bar{i}_0^{\theta\theta})^{-1}n\,\mathrm{acov}_{\mathrm{T}}\{(\hat{\mathbf{I}}_{\mathrm{ML}}^{-1}\hat{\boldsymbol{q}}_{\mathrm{ML}}^*)_\theta,\, \hat{\theta}_{\mathrm{ML}}\}\\
&\qquad + 2(\bar{i}_0^{\theta\theta})^{-1/2}\{(\mathbf{I}_0^{-1}\mathbf{q}_0^*)_\theta\mathbf{i}_0^{(1)'}(\mathbf{A}_{\mathrm{ML}2})_{\cdot\theta} + \mathbf{i}_0^{(1)'}\mathbf{I}_0^{-1}\mathbf{q}_0^*\alpha_{\mathrm{ML}2}\}] + O(n^{-2})\\
&\equiv \alpha_{\mathrm{ML}2}^{(t)} + n^{-1}\alpha_{\mathrm{W}\Delta 2}^{(t)} + O(n^{-2})(\alpha_{\mathrm{W}2}^{(t)} = \alpha_{\mathrm{ML}2}^{(t)}),\\
\kappa_3(t_{\mathrm{W}}) &= n^{-1/2}\alpha_{\mathrm{ML}3}^{(t)} + O(n^{-3/2})(\alpha_{\mathrm{W}3}^{(t)} = \alpha_{\mathrm{ML}2}^{(t)}),\\
\kappa_4(t_{\mathrm{W}}) &= n^{-1}\alpha_{\mathrm{ML}4}^{(t)} + O(n^{-2})(\alpha_{\mathrm{W}4}^{(t)} = \alpha_{\mathrm{ML}4}^{(t)}).
\end{aligned}
$$

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Proceedings of the 2nd international symposium on information theory* (pp.267–281). Budapest: Académiai Kiado.

Amari, S., & Nagaoka, H. (2000). *Methods of information geometry* (D. Harada, Trans.). Providence, RI: American Mathematical Society and Oxford University Press.

Chen, M.-H., Ibrahim, J. G., & Kim, S. (2008). Properties and implementation of Jeffreys's prior in binomial regression models. *Journal of the American Statistical Association, 103*, 1659–1664.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80*, 27–38.

Giles, D. E. A., & Rayner, A. C. (1979). The mean squared errors of the maximum likelihood and natural-conjugate Bayes regression estimators. *Journal of Econometrics, 11*, 319–334.

Gruber, M. H. J. (1998). *Improving efficiency by shrinkage: The James-Stein and ridge regression estimators.* New York: Marcel Dekker.

Hall, P. (1992). *The bootstrap and Edgeworth expansion.* New York: Springer. Corrected printing, 1997.

Ibrahim, J. G., & Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science, 15*, 46–60.

Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2003). On optimality properties of the power prior. *Journal of the American Statistical Association, 98*, 204–213.

Ibrahim, J. G., & Laud, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association, 86*, 981–986.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A, 186*, 453–461.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Clarendon.

Kass, R. E. (1989). The geometry of asymptotic inference. *Statistical Science, 4*, 188–234.

Kass, R. E., & Vos, P. W. (1997). *Geometrical foundations of asymptotic inference.* New York: Wiley.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association, 91*, 1343–1370.

le Cessie, S., & van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics, 41*, 191–201.

Ogasawara, H. (2010). Asymptotic expansions for the pivots using log-likelihood derivatives with an application in item response theory. *Journal of Multivariate Analysis, 101*, 2149–2167.

Ogasawara, H. (2012). Cornish-Fisher expansions using sample cumulants and monotonic transformations. *Journal of Multivariate Analysis, 103*, 1–18.

Ogasawara, H. (2013). Asymptotic cumulants of the estimator of the canonical parameter in the exponential family. *Journal of Statistical Planning and Inference, 143*, 2142–2150.

Ogasawara, H. (2014). Bias adjustment minimizing the asymptotic mean square error. *Communications in Statistics – Theory and Methods.* DOI:10.1080/03610926.2013.786788.

Ogasawara, H. (2015). An expository supplement to the paper "Optimization of the Gaussian and Jeffreys power priors with emphasis on the canonical parameters in the exponential family". To appear in *Economic Review (Otaru University of Commerce).* http://www.res.otaru-uc.ac.jp/~hogasa/, http://barrel.ih.otaru-uc.ac.jp/

Poirier, D. (1994). Jeffreys' prior for logit models. *Journal of Econometrics, 63*, 327–339.

Rubin, D. B., & Schenker, N. (1987). Logit-based interval estimation for binomial data using the Jeffreys prior. In C. C. Clogg (Ed.), *Sociological Methodology Vol.17 (1987)* (pp.131–144). Washington D. C.: American Sociological Association.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior

distributions. In P. Goel and A. Zellner (Eds.), *Bayesian inference and decision techniques* (pp.233–243). Amsterdam: Elsevier.

(Received April 14 2014, Revised July 14 2014)