Journal of Statistical Planning and Inference 238 (2025) 106271

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

On cross-validated estimation of skew normal model

Jian Zhang [a],[*], Tong Wang [b]

[a] School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury CT2 7NF, UK
[b] Novatis Pharmaceuticals UK, The WestWorks, London W12 7FQ, UK

ARTICLE INFO

Keywords:
Skew normal
Maximum penalised likelihood estimator
Multifold cross-validation
Asymptotics

ABSTRACT

Skew normal model suffers from inferential drawbacks, namely singular Fisher information when it is close to symmetry and diverging of maximum likelihood estimation. This causes a large variation of the conventional maximum likelihood estimate. To address the above drawbacks, Azzalini and Arellano-Valle (2013) introduced maximum penalised likelihood estimation (MPLE) by subtracting a penalty function from the log-likelihood function with a pre-specified penalty coefficient. Here, we propose a cross-validated MPLE to improve its performance when the underlying model is close to symmetry. We develop a theory for MPLE, where an asymptotic rate for the cross-validated penalty coefficient is derived. We further show that the proposed cross-validated MPLE is asymptotically efficient under certain conditions. In simulation studies and a real data application, we demonstrate that the proposed estimator can outperform the conventional MPLE when the model is close to symmetry.

1. Introduction

Skewness, which measures the asymmetry of a distribution, is an important data feature to characterise. Change of data skewness can serve as a basis for detecting an attack upon a sensor network, for providing an early warning for abrupt climate changes, for estimating aggregates of small domain business, for modelling equity excess returns, for characterising sensitivity of anti-cancer drugs, among others (Buttyan et al., 2006; He et al., 2013; Colacito et al., 2016; Ferrante and Pacei, 2017; Dhar et al., 1996). For example, in cancer research, people are interested in characterising drug sensitivity and development of novel therapeutics. The data considered in this study consist of the measurements of median inhibition concentrations, IC50s, of 227 drugs in 111 cancer cell lines (Iorio et al., 2016). IC50 is a measure of how much drug is needed to inhibit the multiplication of that cell line by 50%. The log-IC50 informs the drug sensitivity against cancer cells. The location, dispersion and skewness parameters of the log-IC50 can be used to search for a combined drug therapy. For example, histogram plots for log-IC50s of drugs Erlotinib and Paclitaxel in Fig. 1.1 demonstrate that these two drugs have contrasting data features, one has positive drug response and the other has drug resistance. Erlotinib is an inhibitor of the epidermal growth factor receptor (EGFR) tyrosine kinase pathway while Paclitaxel is a chemotherapy drug. Cancer stem cells are often enriched after chemotherapy and induce tumor recurrence, which poses a significant clinical challenge. Combining Erlotinib with Paclitaxel can overcome paclitaxel-resistant cervical cancer (Lv et al., 2019).

By introducing a shape parameter α in a normal distribution, the skew-normal and more generally, skew symmetry distributions, can provide a better fit for asymmetric data than does the normal (Azzalini, 1985). Because of their appealing mathematical properties and usefulness in practice, skew-normal distributions have received considerable attention in the past two decades. Extensions have been made in various directions, including multivariate skew-normal, skew t- and skew elliptical distributions and finite mixtures of skew normals (Azzalini and Capitanio, 1999, 2003; Lin, 2009; Azzalini and Capitanio, 2014; Wang et al.,

[*] Correspondence to: School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7NF, United Kingdom.
E-mail address: J.Zhang-79@kent.ac.uk (J. Zhang).

https://doi.org/10.1016/j.jspi.2025.106271
Received 19 April 2024; Received in revised form 21 October 2024; Accepted 15 January 2025

Available online 25 January 2025

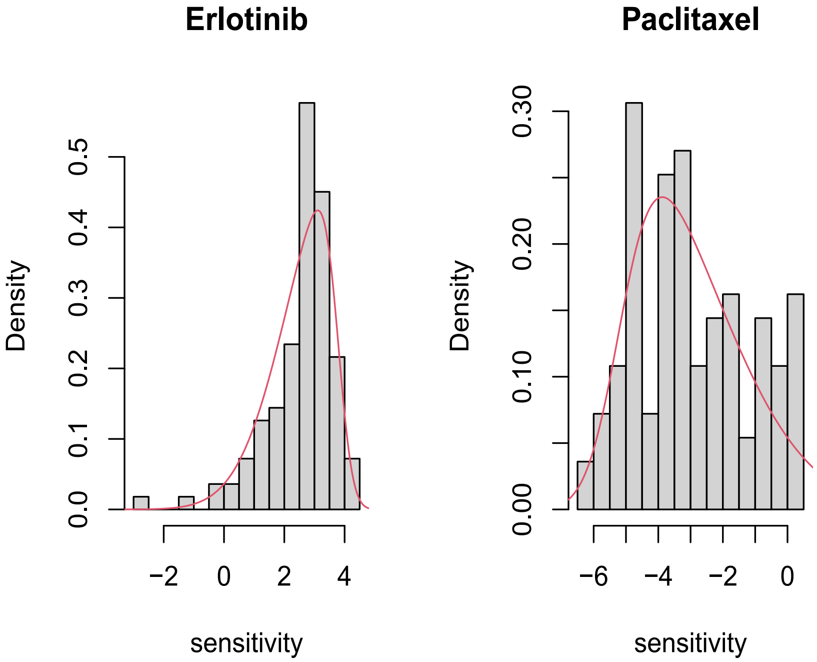


Fig. 1.1. Histogram plots and skew normal fits of log-IC50 data for drugs Erlotinib and Paclitaxel, where the solid curves are the fits produced by the proposed MPLE.

2020 and references therein). Some of these developments, however, suffer from inferential drawbacks, namely singular Fisher information in the vicinity of symmetry and diverging of maximum likelihood estimator of α (Azzalini, 1985; Hallin and Ley, 2012). To eliminate the singularity, a tentative remedy called centered parametrisation (CP) was put forward by Azzalini (1985), where α is reparametrised to Pearson's skewness index. Unfortunately, the likelihood function under the CP is not explicitly available as the Jacobian factor of the transformation is unbounded when the underlying model is symmetric. Chiogna (2005) showed that even using reparameterisation, the resulting maximum likelihood estimator (MLE) of α has a rate of $n^{-1/6}$ lower than the usual rate of root-n at $\alpha = 0$. The above remedy never really caught upon, partly because the mechanism of skewness is unknown in practice and the resulting skew-normal family, under the new parametrisation, loses much of its simplicity (Azzalini and Capitanio, 2014). To develop an alternative remedy to handle the above inferential drawbacks, Azzalini and Arellano-Valle (2013) consider a penalised log-likelihood by subtracting $\lambda Q(\alpha)$ from the log-likelihood, where penalty $Q(\alpha)$ is used to control the magnitude of α. By using Firth's bias-correction technique (Firth, 1993), Azzalini and Capitanio (2014) chose a fixed penalty coefficient $\lambda \approx 0.87591$ for $Q(\alpha) = \log(1 + 0.85625\alpha^2)$. However, this choice is against our intuition that when the underlying value of α is close to zero, the penalty coefficient should increase to infinity. In this paper, we aim to develop a data-driven procedure for improving the choice of the penalty coefficient with a theoretical guarantee.

In literature, there are two approaches for determining the penalty coefficient, one is information criterion and the other is cross-validation. However, the former, usually working for non-degenerate models, may be invalid for the singular models in which the penalised likelihood (or posterior) cannot be approximated by any normal distribution. Although in a singular model, the generalisation error of an inference procedure may not be estimated well by information criteria, it can be estimated by the cross-validation (Watanabe, 2021). This motivates us to investigate a multifold cross-validation procedure for skew normal estimation. Our contributions to the research field are two-fold. Firstly, we develop a hyperbolic parametrisation to understand the nature of α from a point view of active function. Under the new parametrisation, we show that the cross-validated estimator asymptotically attains the Cramer–Rao lower bound to estimating error. By simulation studies and a real data application, we demonstrate that the proposed estimation can outperform the bias-correction approach used in the software **SN** (https://CRAN.R-project.org/package=sn) in terms of bias and standard error. Secondly, we unveil a super efficiency for the cross-validated estimation at $\alpha = 0$, namely after an appropriate tuning, the cross-validated estimator can recover the true $\alpha = 0$ exactly with a probability tending to one. This is in striking contrast to the MLE which has a very slow convergence rate at $\alpha = 0$. By theoretical analysis and simulations, we show that Firth's bias-correction technique may under-regularise the estimation at $\alpha = 0$ in the sense that the penalty coefficient $\lambda = 0.8759$ is too small to reduce the variability of estimation. Furthermore, we demonstrate a filtering effect of penalisation by simulations that any small skewness will be filtered out by the cross-validated MPLE.

The rest of the article is organised as follows. In Section 2 we develop the cross-validation procedure for skew normal models. We establish theoretical properties of the proposed procedure in Section 3. We develop a penalised Expectation-Maximisation (EM) algorithm in Section 4. We conduct simulation studies and a real data analysis in Section 5. We conclude with a discussion in Section 6. The proofs are relegated to the Appendix.

2. Methodology

In this section, we first review the location-scale model for skew normals and reparameterisation. Then after a short discussion of its inherent inferential issues we develop a multifold cross-validation procedure for the penalised likelihood inference.

2.1. Location-scale model

Let $\mathrm{SN}(0,1,\alpha)$ denote the standard skew normal distribution with density $f(x;\alpha)=2\phi(x)\Phi(\alpha x)$, $x\in\mathbb{R}$, where $\phi(x)$ and $\Phi(x)$ are the standard normal density and cumulative distribution function respectively, and α is the shape parameter to regulate the skewness. The skew normal distribution family includes the standard normal as a special when $\alpha=0$. The skew normal random variable X can be expressed as a linear combination of two independent variables, a half standard normal X_+ with density $2\phi(x), x\geq 0$ and the standard normal X_0, in the form $X=\delta X_+ + \sqrt{1-\delta^2}X_0$, where $\delta=\alpha/\sqrt{1+\alpha^2}$. Consider the location-scale model $Y=\mu+\sigma(X-\delta\sqrt{2/\pi})$ with location parameter $\mu\in\mathbb{R}$, scale parameter $\sigma\geq 0$ and density $f(y;\mu,\sigma,\alpha)=\sigma^{-1}f((y-\mu)/\sigma+\delta\sqrt{2/\pi};\alpha)$. We have $E[Y]=\mu$, $\mathrm{var}(Y)=\sigma^2\mathrm{var}(X)=\sigma^2(1-2\delta^2/\pi)$, $E[X]=\delta\sqrt{2/\pi}$, and $\mathrm{var}(X)=1-2\delta^2/\pi$. Note that $Y=\mu-\sigma\delta\sqrt{2/\pi}+\sigma X$ reduces to the standard skew location-scale normal model. The term $\delta\sqrt{2/\pi}$ shows the effect of skewness on the location parameter μ. Under the above location-scale model, Pearson's skewness index γ_1 is related to parameters α and δ via

$$\gamma_1 = \frac{E\left[(Y-\mu)^3\right]}{\mathrm{var}(Y)^{3/2}} = \frac{4-\pi}{2}\frac{\delta^3(2/\pi)^{3/2}}{(1-2\delta^2/\pi)^{3/2}}, \quad |\gamma_1|\leq 0.9952.$$
$$\delta = \sqrt{\frac{\pi}{2}}\frac{(2\gamma_1/(4-\pi))^{1/3}}{\sqrt{1+(2\gamma_1/(4-\pi))^{2/3}}}, \quad \alpha=\delta/\sqrt{1-\delta^2}. \tag{2.1}$$

Here, α and δ can be viewed as activation functions of γ_1 with derivatives

$$\frac{d\delta}{d\gamma_1} = \sqrt{\frac{\pi}{2}}\frac{2}{3(4-\pi)}\left(\frac{2\gamma_1}{4-\pi}\right)^{-2/3}\left(1+\left(\frac{2\gamma_1}{4-\pi}\right)^{2/3}\right)^{-3/2},$$
$$\frac{d\alpha}{d\gamma_1} = \sqrt{\frac{\pi}{2}}\frac{2}{3(4-\pi)}\left(\frac{2\gamma_1}{4-\pi}\right)^{-2/3}\left(1+(1-\pi/2)\left(\frac{2\gamma_1}{4-\pi}\right)^{2/3}\right)^{-3/2}$$

which are unbounded at $\gamma_1=0$. This explains why the Fisher information matrix is singular at $\alpha=0$. Fig. 2.1 demonstrates that the skewness is introduced into the model via the activation function δ of the input γ_1, where γ_1 is restricted to the interval $(-0.9952, 0.9952)$. By the activation function, the negative (positive) γ_1 will be mapped onto strongly negative (positive) δ while zero γ_1 will be mapped onto zero δ. In the CP, Azzalini (1985) used γ_1 to reparametrise α and δ through Eq. (2.1).

2.2. Hyperbolic reparameterisation

To tackle the above issue of unboundedness, we reparametrise α by the scaled inverse hyperbolic transformation $\theta=\mathrm{arcsinh}(\alpha)/a$, where $a>0$ is a pre-selected constant. This gives rise to a family of activation functions

$$\alpha=\sinh(a\theta)=\frac{1}{2}(e^{a\theta}-e^{-a\theta}), \qquad \delta=\tanh(a\theta)=\frac{e^{2a\theta}-1}{e^{2a\theta}+1}, \qquad \alpha=\frac{\delta}{\sqrt{1-\delta^2}} \tag{2.2}$$

with derivatives

$$\frac{d\alpha}{d\theta}=a\cosh(a\theta)=\frac{a}{2}(e^{a\theta}+e^{-a\theta}), \qquad \frac{d\delta}{d\theta}=a(1-\delta^2).$$

When $a=1$, θ reduces to the Fisher transformation of δ. Fig. 2.1 shows that as a tends to infinity $\sinh(a\theta)$ is close to the CP. However, for finite fixed a's, these activation functions give bounded derivatives. In the following, we focus on the simple case $a=1$. To remove the constraint $\sigma\geq 0$, we reparametrise σ by $\eta=\log(\sigma)$, which has a range of $(-\infty,\infty)$ and $\frac{d\sigma}{d\eta}=\sigma$.

2.3. Initial estimation

Given an i.i.d. sample $\boldsymbol{y}=(y_i)_{1\leq i\leq n}$ drawn from the above location-scale model, we use the method of moments to construct initial estimators as follows. Setting the first three moments of Y equal to their corresponding sample moments and using the relationships between γ_1, δ, α and θ, we have the initial estimates $\mu^{(0)},\sigma^{(0)},\eta^{(0)}$, $\theta^{(0)},\alpha^{(0)}$ and $\delta^{(0)}$. Let $\mu_0,\sigma_0,\eta_0,\theta_0,\alpha_0$ and δ_0 be the ground-truth of the parameters in the model. It follows from the central limit theorem that $\mu^{(0)}=\mu_0+O_p(1/\sqrt{n})$ and $\gamma_1^{(0)}=\gamma_0+O_p(1/\sqrt{n})$. Using Eq. (2.1), for $\theta_0=0$, we have $\delta^{(0)}=O_p(n^{-1/6})$, $\theta^{(0)}=O_p(n^{-1/6})$, $\alpha^{(0)}=O_p(n^{-1/6})$, and $\eta^{(0)}=\eta_0+O_p(n^{-1/3})$. In contrast, for $\theta_0\neq 0$, we have the following standard root-$\sqrt{n}$ convergence rates, $\delta^{(0)}=\delta_0+O_p(1/\sqrt{n})$, $\theta^{(0)}=\theta_0+O_p(1/\sqrt{n})$, $\alpha^{(0)}=\alpha_0+O_p(1/\sqrt{n})$, and $\eta^{(0)}=\eta_0+O_p(1/\sqrt{n})$. The above estimation will be used to develop consistent maximum penalised likelihood estimators in Section 2.3 and as the starting point for the penalised Expectation-Maximisation in Section 2.4 below.

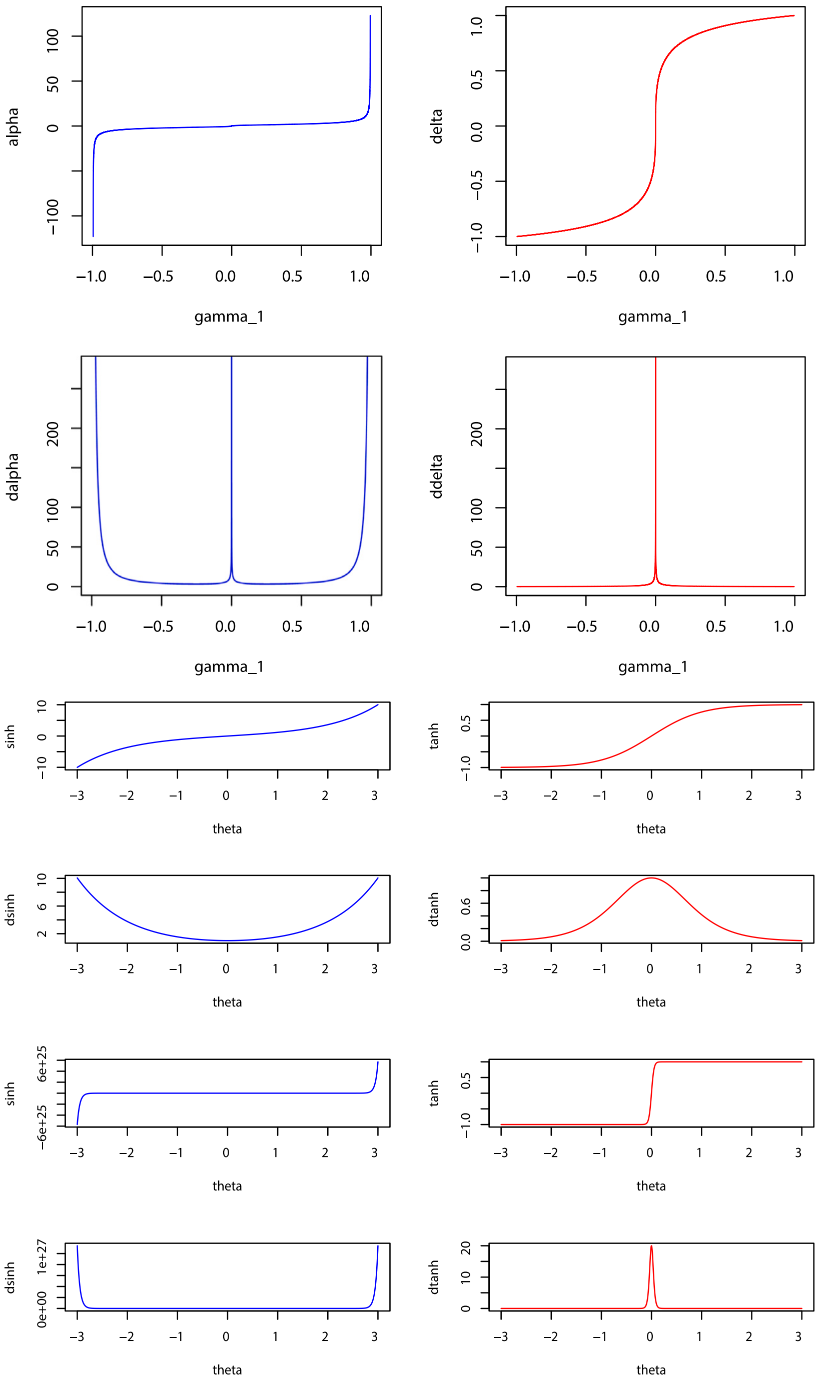


Fig. 2.1. Activation functions for α and δ. The plots in the top row are for $\alpha(\gamma_1)$ and $\delta(\gamma_1)$ respectively. The plots in the 2nd row are for the derivatives of α and δ (namely, $dalpha$ and $ddelta$) with respect to γ_1 respectively. The plots in the 3rd row are for $\alpha(\theta)$ and $\delta(\theta)$ respectively. The plots in the 4th row are for the derivatives of $\alpha(\theta)$ and $\delta(\theta)$ with respect to θ respectively. The plots in the 5th row are for $\alpha(20\theta)$ and $\delta(20\theta)$ respectively. The plots in the bottom row are for the derivatives of $\alpha(20\theta)$ and $\delta(20\theta)$ with respect to θ respectively.

2.4. Maximum penalised likelihood estimation

Given an i.i.d. sample $\boldsymbol{y}$ of size n drawn from a skew normal distribution and letting $z_i = (y_i - \mu)/\sigma$, we have the following log-likelihood

$$l_{inc} = l_{inc}\left(\mu, \eta, \theta \mid \boldsymbol{y}\right) = \frac{n}{2}\log\left(\frac{2}{\pi\sigma^2}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(z_i + \delta\sqrt{2/\pi}\right)^2 + \sum_{i=1}^{n}\log\Phi\left(\alpha\cdot\left(z_i + \delta\sqrt{2/\pi}\right)\right),$$

where the dependence of δ and α on θ and σ on η are suppressed. When $|\theta|$ tends to infinity, the skew normal distribution reduces to a half-normal distribution whose support may depend on the parameters. To prevent this, we consider maximum likelihood estimate of (μ, η, θ) defined on a bounded open subset of $\mathbb{R}^3$. For $(\mu_0, \eta_0, \theta_0)$ in the above bounded subset, it follows from Chiogna (2005) that the maximum likelihood estimates of μ, η and θ are of convergence rates $O_p(n^{-1/2})$, $O_p(n^{-1/3})$ and $O_p(n^{-1/6})$ respectively, despite that the Fisher information matrix is degenerate at $\theta = 0$. This implies these maximum likelihood estimates are asymptotically in the restricted parametric space

$$\Omega_n = \{(\mu, \eta, \theta) : \mid \mu - \mu^{(0)} \mid\leq c_{\mu_0} n^{-1/2}, \mid \eta - \eta^{(0)} \mid\leq c_{\eta_0} n^{-1/3}, \mid \theta - \theta^{(0)} \mid\leq c_{\theta_0} n^{-1/6}\}$$

for some arbitrary large positive constants c_{μ_0}, c_{η_0} and c_{θ_0}. We asymptotically have the maximum likelihood estimate (MLE)

$$(\hat{\mu}, \hat{\eta}, \hat{\theta}) = \underset{(\mu,\eta,\theta)\in\Omega_n}{\arg\max}\ l_{inc}\left(\mu, \eta, \theta \mid \boldsymbol{y}\right).$$

The above likelihood is singular at $\theta = 0$ with a stationary point at $\theta = 0$ regardless values of the other parameters as the Fisher information matrix at $(\mu, \eta, 0)$,

$$-E\left(\frac{\partial^2 \log l_{inc}}{\partial(\mu, \eta, \theta)\partial(\mu, \eta, \theta)^T}\right)|_{(\mu,\eta,0)} = \mathrm{diag}\left(\frac{n}{\sigma^2}, 2n, 0\right)$$

is degenerate. This results in a slow convergence rate of $\hat{\theta}$ and non-standard asymptotic behavior of the MLE when the underlying value of θ is zero or near zero. As noted previously, the MLE of the shape parameter of the skew normal diverges with a probability that is non-negligible for small and moderate sample sizes. To address these issues, following Azzalini and Arellano-Valle (2013), we maximise the penalised log-likelihood

$$l_{incp}(\mu, \eta, \theta \mid \boldsymbol{y}) = l_{inc}(\mu, \eta, \theta \mid \boldsymbol{y}) - \lambda\ \mathrm{pen}(\theta),$$

where a penalty is used to control the size of θ, satisfying the condition

$$C1:\ \ \mathrm{pen}(\theta) \geq 0, \quad \mathrm{pen}(0) = \mathrm{pen}'(0) = 0, \quad \mathrm{pen}''(0) = 2, \quad \lim_{|\theta|\to\infty} \mathrm{pen}(\theta) \to \infty.$$

For example, hyperbolic penalty $\mathrm{pen}_1(\theta) = \alpha^2 = (e^\theta - e^{-\theta})^2/4$, ridge penalty $\mathrm{pen}_2(\theta) = \theta^2$ and log-Cauchy $\mathrm{pen}_3(\theta) = \log(1 + c_2\alpha^2) = \log(1 + c_2(e^\theta - e^{-\theta})^2/4)$ meet these conditions, where the log-Cauchy was proposed by Azzalini and Capitanio (2014). All these penalties encourage shrinkage of the skewness parameter towards zero while preventing it from diverging to infinity. Under the penalisation, the Fisher information matrix of the penalised likelihood is not singular. For each $0 \leq \lambda/n \leq \omega_0$, define maximum penalised likelihood estimate (MPLE) of (μ, η, θ) on any bounded subset of $\mathbb{R}^3$. Similar to Chiogna (2005), we can show that the MPLE is asymptotically equal to

$$(\hat{\mu}_\lambda, \hat{\eta}_\lambda, \hat{\theta}_\lambda) = \arg \max_{(\mu,\eta,\theta)\in\Omega_n} l_{incp}(\mu, \eta, \theta \mid \boldsymbol{y}).$$

The larger the penalty coefficient, the greater the accuracy of estimating θ when the true value of θ is zero while the larger estimating bias when the true value of θ is not zero. Multifold cross-validation below strikes a balance between the accuracy and the bias by tuning the penalty coefficient and therefore achieves a better prediction for new samples.

2.5. Multifold cross validation

Given the sample $\boldsymbol{y}$, for a pre-specified positive constant ω_0, define the expected out-of-sample generalisation error,

$$\mathrm{CV}(\lambda) = -E[l_{inc}(\hat{\mu}_\lambda, \hat{\eta}_\lambda, \hat{\theta}_\lambda|\boldsymbol{y}^*)],$$

if we were to apply the model based on estimator $(\hat{\mu}_\lambda, \hat{\eta}_\lambda, \hat{\theta}_\lambda)$ to predict a new set of observations $\boldsymbol{y}^*$ drawn independently from the same distribution as that of $\mathbf{y}$. The above expectation is taken with respect to both $\boldsymbol{y}$ and $\boldsymbol{y}^*$. The generalisation error $\mathrm{CV}(\lambda)$ can be used as a criterion to compare candidate estimators $(\hat{\mu}_\lambda, \hat{\eta}_\lambda, \hat{\theta}_\lambda)$, $0 \leq \lambda/n \leq \omega_0$. We estimate the expected out-of-sample generalisation error $\mathrm{CV}(\lambda)$ by multifold cross-validation as follows.

For a pre-specified integer $K > 0$, divide the data $\boldsymbol{y}$ into K groups $\boldsymbol{y}_j, 1 \leq j \leq K$ with corresponding index groups $[j], 1 \leq j \leq K$. For each λ and $1 \leq j \leq K$, we calculate estimates $\left(\hat{\mu}_{[-j]\lambda}, \hat{\eta}_{[-j]\lambda}, \hat{\theta}_{[-j]\lambda}\right)$, based on the training set $\boldsymbol{y}_{[-j]}$, by maximising log-likelihood $l_{incp}\left(\mu, \eta, \theta \mid \boldsymbol{y}_{[-j]}\right)$. For each j, taking $\boldsymbol{y}_j$ as the validation sample, we can estimate the out-of-sample generalisation error by $-l_{inc}\left(\hat{\mu}_{[-j]\lambda}, \hat{\eta}_{[-j]\lambda}, \hat{\theta}_{[-j]\lambda} \mid \boldsymbol{y}_j\right)$. Averaging these estimated errors, we have the following average generalisation error

$$\mathrm{CV}_a(\lambda) = -\frac{1}{K}\sum_{j=1}^{K} l_{inc}\left(\hat{\mu}_{[-j]\lambda}, \hat{\eta}_{[-j]\lambda}, \hat{\theta}_{[-j]\lambda} \mid \boldsymbol{y}_j\right),$$

where K is a positive integer. The optimal tuning parameter $\lambda_{op} = \arg\min_{0\le\lambda/n\le\omega_0} E[\mathrm{CV}_a(\lambda)]$ is estimated by

$$\lambda_{cv} = \arg\min_{0\le\lambda/n\le\omega_0} \mathrm{CV}_a(\lambda).$$

3. Asymptotic theory

Similar to Chiogna (2005), we can show the consistency of the penalised MLE defined on any bounded open ball containing the true values of (μ,η,θ). As pointed out before, for simplicity, we focus on the penalised MLE defined on Ω_n. However, the following asymptotic theory holds for the above general case. Note that for each λ, the penalised MLE is obtained by solving simultaneous equations

$$\begin{aligned}&\frac{1}{n}\frac{\partial l_{inc}((\hat{\mu},\hat{\eta},\hat{\theta}) \mid \mathbf{y})}{\partial\mu} = 0, \quad \frac{1}{n}\frac{\partial l_{inc}((\hat{\mu},\hat{\eta},\hat{\theta}) \mid \mathbf{y})}{\partial\eta} = 0,\\ &\frac{1}{n}\frac{\partial l_{inc}((\hat{\mu},\hat{\eta},\hat{\theta}) \mid \mathbf{y})}{\partial\theta} - \frac{\lambda}{2n}(e^{2\hat{\theta}} - e^{-2\hat{\theta}}) = 0,\end{aligned} \tag{3.1}$$

where

$$\begin{aligned}\frac{1}{n}\frac{\partial l_{inc}((\mu,\eta,\theta) \mid \mathbf{y})}{\partial\mu} &= \frac{1}{n}\sum_{i=1}\left(z_i + \delta\sqrt{2/\pi}\right)\frac{1}{\sigma} - \frac{1}{n}\sum_{i=1}^{n}\frac{\phi(A_i)}{\Phi(A_i)}\frac{\alpha}{\sigma},\\ \frac{1}{n}\frac{\partial l_{inc}((\mu,\eta,\theta) \mid \mathbf{y})}{\partial\eta} &= -1 + \frac{1}{n}\sum_{i=1}\left(z_i + \delta\sqrt{2/\pi}\right)z_i - \frac{\alpha}{n}\sum_{i=1}^{n}\frac{\phi(A_i)}{\Phi(A_i)}z_i,\\ \frac{1}{n}\frac{\partial l_{inc}((\mu,\eta,\theta) \mid \mathbf{y})}{\partial\theta} &= -\frac{1}{n}\sum_{i=1}\left(z_i + \delta\sqrt{2/\pi}\right)(1-\delta^2)\sqrt{2/\pi}\\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\frac{\phi(A_i)}{\Phi(A_i)}\left(\frac{(e^{\theta}+e^{-\theta})}{2}\left(z_i + \delta\sqrt{2/\pi}\right) + \alpha(1-\delta^2)\sqrt{2/\pi}\right)\end{aligned}$$

with $A_i = \alpha(z_i + \delta\sqrt{2/\pi})$. To develop the theory, denote by $\mathbf{C} = (c_{ij})_{3\times3} = \mathbf{C}_{(\mu,\eta,\theta)} = (c_{ij}(\mu,\eta,\theta))_{3\times3}$ the second derivative matrix of the log-likelihood with respect to (μ,η,θ). Applying the Taylor expansion to the functions in (3.1) at the ground-truth (μ_0,η_0,θ_0), we have

$$\begin{aligned}0 = &\frac{1}{\sqrt{n}}\frac{\partial l_{inc}(y_i)}{\partial(\mu_0,\eta_0,\theta_0)^T} - \frac{\lambda}{2\sqrt{n}}(e^{2\theta_0} - e^{-2\theta_0})\boldsymbol{e}_3\\ &+ (\mathbf{C}-\mathbf{D})_{(\mu^*,\eta^*,\theta^*)}\sqrt{n}(\hat{\mu}-\mu_0, \hat{\eta}-\eta_0, \hat{\theta}-\theta_0)^T,\end{aligned} \tag{3.2}$$

where $\boldsymbol{e}_3 = (0,0,1)^T$ and $(\mu^*,\eta^*,\theta^*) = (\mu_0,\eta_0,\theta_0) + t(\hat{\mu}-\mu_0, \hat{\eta}-\eta_0, \hat{\theta}-\theta_0)$, $0 \le t \le 1$. Denote $\mathbf{D}_{\theta\lambda/n} = \mathrm{diag}\left(0, 0, \frac{\lambda}{n}(e^{2\theta}+e^{-2\theta})\right)$ and $(\mathbf{C}-\mathbf{D})_{(\mu,\eta,\theta)} = \mathbf{C}_{(\mu,\eta,\theta)} - \mathbf{D}_{\theta\lambda/n}$. For the simplicity of notation, let $\mathbf{D}_{0\lambda/n}$ denote $\mathbf{D}_{\theta_0\lambda/n}$ and $\mathbf{C}_0$ denote $\lim_{n\to\infty}\mathbf{C}_{(\mu_0,\eta_0,\theta_0)}$. Then $\mathbf{C}_{(\mu_0,\eta_0,\theta_0)} = \mathbf{C}_0 + O_p(1/\sqrt{n})$, where $-\mathbf{C}_0$ is the Fisher information matrix at (μ_0,η_0,θ_0). For a pre-specified positive constant ω_0, consider $0 \le \lambda/n \le \omega_0$. Taking $\{c_{ij} = c_{ij}(\mu,\eta,\theta) : (\mu,\eta,\theta) \in \Omega_1\}$ as empirical processes to which we apply the weak large law, we have, uniformly for $0 \le \lambda/n \le \omega_0$ and bounded (μ_0,η_0,θ_0),

$$|| (\mathbf{C}-\mathbf{D})_{(\mu^*,\eta^*,\theta^*)} - (\mathbf{C}_0 - \mathbf{D}_{0\lambda/n}) ||= o_p(1)$$

in terms of the Frobenius norm. We have

Proposition 3.1. *Assume that the MPLE of θ is in Ω_1 and that the penalty $pen(\theta)$ satisfies the condition $(C1)$. Then, when the true value $\theta_0 \neq 0$, as $\lambda/\sqrt{n} \to 0$, the MPLE $(\hat{\theta}_\lambda, \hat{\sigma}_\lambda, \hat{\theta}_\lambda)$ is asymptotically optimal in the sense that it is asymptotically unbiased and attains the Cramer–Rao low bound to estimation error.*

As in practice, the true value θ_0 is unknown, we have to use a data-driven cross-validation to tune the penalty. In the following theorem, we show that $\lambda_{cv}/\sqrt{n} \to 0$ and the cross-validated MPLE is asymptotically optimal in terms of mean square error when the underlying $\theta_0 \neq 0$.

Theorem 1. *Assume that the MPLE of θ is in Ω_1 and that the penalty $pen(\theta)$ satisfies the condition $(C1)$. Then, when the true value $\theta_0 \neq 0$, we have $\lambda_{cv}/\sqrt{n} \to 0$ in probability and the MPLE $(\hat{\theta}_{\lambda_{cv}}, \hat{\sigma}_{\lambda_{cv}}, \hat{\theta}_{\lambda_{cv}})$ is asymptotically unbiased and attains the Cramer–Rao low bound to estimation error.*

Let $z_{i0} = (y_i - \mu_0)/\sigma_0$. In the next proposition, we show that a fixed λ in the Q penalty may give rise to a biased estimate of α.

Proposition 3.2. *Assume that the MPLE of θ is in Ω_1 and that the penalty $pen(\theta)$ satisfies the condition $(C1)$. Then, when the true value $\theta_0 = 0$, we have $\hat{\theta}_\lambda = 0$ for $\lambda \ge \max\left\{\sum_{i=1}^{n}\left(1 - z_{i0}^2\right)/\pi, 0\right\}$. And on $\sum_{i=1}^{n}\left(1 - z_{i0}^2\right) > 0$, $\hat{\theta}_{\lambda_{cv}}$ is non-zero for $0 \le \lambda < \sum_{i=1}^{n}\left(1 - z_{i0}^2\right)/\pi$.*

The above proposition implies that when the true value $\theta_0 = 0$, we have

(i) As $\lambda/\sqrt{n} \to \infty$, we have $\hat{\theta}_\lambda = 0$.
(ii) On $\frac{1}{\sqrt{n}} \sum_{i=1}^n \left(1 - z_{i0}^2\right) \leq 0$, for any $\lambda \geq 0$, we have $\hat{\theta}_\lambda = 0$.
(iii) On $\frac{1}{\sqrt{n}} \sum_{i=1}^n \left(1 - z_{i0}^2\right) > 0$, there is λ such that $\frac{1}{n} l_{incp}(\hat{\mu}_\lambda, \hat{\eta}_\lambda, \hat{\theta}_\lambda | \mathbf{y})$ attains the maximum at non-zero $\hat{\theta}_\lambda$.

Let $\hat{\mu}_j, \hat{\eta}_j$ be the MLEs of μ and η based on the subsample $\mathbf{y}_j$ when θ is known to be zero. Let $\hat{\mu}_{[-j]\lambda}$, $\hat{\eta}_{[-j]\lambda}$ and $\hat{\theta}_{[-j]\lambda}$ be the penalised MLEs based on the remaining observations $(\boldsymbol{y}_i)_{i\neq j}$ after removing $\boldsymbol{y}_j$ from $\boldsymbol{y}$. For the simplicity of notation, let $\mathbf{I}_{11\lambda}^{*[-j]}$, $\mathbf{I}_{12\lambda}^{*[-j]}$, $\mathbf{I}_{21\lambda}^{*[-j]}$ and $\mathbf{I}_{22\lambda}^{*[-j]}$ respectively denote $\mathbf{I}_{11} \mid_{(\mu^*_{[-j]\lambda}, \eta^*_{[-j]\lambda}, \theta^*_{[-j]\lambda})}$, $\mathbf{I}_{12} \mid_{(\mu^*_{[-j]\lambda}, \eta^*_{[-j]\lambda}, \theta^*_{[-j]\lambda})}$, $\mathbf{I}_{21} \mid_{(\mu^*_{[-j]\lambda}, \eta^*_{[-j]\lambda}, \theta^*_{[-j]\lambda})}$ and $\mathbf{I}_{22} \mid_{(\mu^*_{[-j]\lambda}, \eta^*_{[-j]\lambda}, \theta^*_{[-j]\lambda})}$. Then $\mathbf{I}_{11} \mid_{(\mu^*_{[-j]\lambda}, \eta^*_{[-j]\lambda}, \theta^*_{[-j]\lambda})} = \mathbf{I}_{110}(1 + o_p(1))$. Let $\lambda_r = \max_{1\leq j\leq K} \max\{\sum_{i\in[-j]}(1 - z_{i0}^2)/\pi, 0\}$. The next theorem shows that if the underlying value of θ is zero, then the $\mathrm{CV}_a(\lambda)$ attains a local minimum when $\lambda \geq \lambda_r$.

Theorem 2. *Assume that the true value $\theta_0 = 0$, the MPLE of θ is in Ω_1 and that the penalty $pen(\theta)$ satisfies the condition $(C1)$. Then, for $\lambda \geq \lambda_r$, the function $CV_a(\lambda)$ is asymptotically flat, that is*

$$
\begin{aligned}
CV_a(\lambda) \;=\; & -\frac{1}{K}\sum_{j=1}^K l_{inc}(\hat{\mu}_j, \hat{\eta}_j, 0 \mid \mathbf{y}_j) + \frac{1}{2K}\sum_{j=1}^K n_j \left(\frac{1}{n_{[-j]}} \sum_{i\in[-j]} \boldsymbol{u}_i - \frac{1}{n_j}\sum_{i\in[j]} \boldsymbol{u}_i \right)^T \\
& \times (-\mathbf{I}_{110})^{-1} \left(\frac{1}{n_{[-j]}} \sum_{i\in[-j]} \boldsymbol{u}_i - \frac{1}{n_j}\sum_{i\in[j]} \boldsymbol{u}_i \right)(1 + o_p(1)).
\end{aligned}
$$

4. Penalised EM algorithm

We first introduce a positive latent random variable W such that (Y, W) has the following easily calculated **joint** density

$$
g(y, w) = \frac{2}{\sigma}\phi\left(z + \delta\sqrt{2/\pi}\right)\phi\left(w - \alpha\left(z + \delta\sqrt{2/\pi}\right)\right), \qquad y \in \mathbb{R}, w \in (0, \infty)
$$

with marginal density of Y,

$$
g(y) = \frac{2}{\sigma}\phi\left(z + \delta\sqrt{2/\pi}\right)\int_0^\infty \phi\left(w - \alpha\left(z + \delta\sqrt{2/\pi}\right)\right)dw = f(y; \mu, \sigma, \alpha)
$$

and conditional density of W given Y,

$$
g(w|y) = \phi\left(w - \alpha\left(z + \delta\sqrt{2/\pi}\right)\right)\left(\Phi\left(\alpha\left(z + \delta\sqrt{2/\pi}\right)\right)\right)^{-1}, \tag{4.1}
$$

where z denotes $(y - \mu)/\sigma$. Letting z_i denote $(y_i - \mu)/\sigma$ as before and augmenting $\mathbf{y}$ by $\boldsymbol{w} = (w_i)_{1\leq i\leq n}$, we form the complete data $(\boldsymbol{y}, \boldsymbol{w}) = (y_i, w_i)_{1\leq i\leq n}$ with the penalised complete-data log-likelihood

$$
\begin{aligned}
l_{comp}(\mu, \eta, \theta \mid \boldsymbol{y}, \boldsymbol{w}) \;=\; & -\frac{n}{2}\log\left(\sigma(\eta)^2\pi^2\right) - \frac{1}{2}\sum_{i=1}^n \left(z_i + \delta(\theta)\cdot\sqrt{2/\pi}\right)^2 \\
& -\frac{1}{2}\sum_{i=1}^n \left(w_i - \alpha(\theta)\cdot\left(z_i + \delta(\theta)\cdot\sqrt{2/\pi}\right)\right)^2 - \lambda(e^\theta - e^{-\theta})^2/4,
\end{aligned}
$$

where $\sigma(\eta) = \exp(\eta)$ and $\alpha(\theta)$ and $\delta(\theta)$ are defined by Eq. (2.2). Let $z_i^{(v)} = (y_i - \hat{\mu}^{(v)})/\sigma(\hat{\eta}^{(v)})$ and $b_i = z_i + \delta(\theta)\sqrt{2/\pi}$ and $b_i^{(v)} = z_i^{(v)} + \delta(\hat{\theta}^{(v)})\sqrt{2/\pi}$. We define **E-step** and **M-step** for $(v+1)$-iteration as follows.

E-Step: Given the estimates $\hat{\mu}^{(v)}$, $\hat{\eta}^{(v)}$ and $\hat{\theta}^{(v)}$ obtained in the vth iteration, use the conditional density in Eq. (4.1) to compute the conditional expectation of the complete log-likelihood:

$$
\begin{aligned}
\Psi(\mu, \eta, \theta \mid \hat{\mu}^{(v)}, \hat{\eta}^{(v)}, \hat{\theta}^{(v)}) &= E_{\mathbf{w}|\mathbf{y}, \hat{\mu}^{(v)}, \hat{\eta}^{(v)}, \hat{\theta}^{(v)}}\left[l_{com}(\mu, \eta, \theta \mid \mathbf{y}, \boldsymbol{w})\right] - \lambda(e^\theta - e^{-\theta})^2/4 \\
&= -n\log(\pi) - n\eta - \frac{1}{2}\sum_{i=1}^n b_i^2 \\
&\quad -\frac{1}{2}\sum_{i=1}^n \left\{ 1 + \left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)} - 2\alpha(\theta)\cdot b_i\right)\frac{\phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)} \right. \\
&\quad \left. + \left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)} - \alpha(\theta)\cdot b_i\right)^2 \right\} - \lambda(e^\theta - e^{-\theta})^2/4.
\end{aligned}
$$

M-Step: For the simplicity of notation, we denote $\Psi(\mu,\eta,\theta \mid \hat{\mu}^{(v)},\hat{\eta}^{(v)},\hat{\theta}^{(v)})$ by Ψ. To maximise the conditional expectation Ψ, compute partial derivatives of Ψ w.r.t. (μ,η,θ) and set these derivatives equal to 0. We solve the above partial derivative equations by alternative iterations as follows.

Firstly, fixing $(\eta,\theta)=(\hat{\eta}^{(v)},\hat{\theta}^{(v)})$, we update μ. It follows from $\frac{\partial\Psi}{\partial\mu}=0$ that given $\theta=\hat{\theta}^{(v)}$ and $\eta=\hat{\eta}^{(v)}$, the $(v+1)$-th update of μ,

$$\begin{aligned}\hat{\mu}^{(v+1)} &= \bar{y}+\sigma(\hat{\eta}^{(v)})\cdot\delta(\hat{\theta}^{(v)})\cdot\sqrt{\frac{2}{\pi}}-\frac{\sigma(\hat{\eta}^{(v)})}{n}\cdot\frac{\alpha(\hat{\theta}^{(v)})}{1+\alpha(\hat{\theta}^{(v)})^2}\\ &\quad\times\sum_{i=1}^{n}\left\{\frac{\phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)}+\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right\}.\end{aligned}$$

Secondly, fixing $(\mu,\theta)=(\hat{\mu}^{(v+1)},\hat{\theta}^{(v)})$, we update η. Let $\boldsymbol{b}^{(v)}=(b_i^{(v)})_{1\le i\le n}$. It follows from $\frac{\partial\Psi}{\partial\eta}=0$ that

$$e^{2\eta}-T(\mu,\theta,\boldsymbol{b}^{(v)})\cdot e^{\eta}-\frac{(1+\alpha(\theta)^2)e^{2\eta}}{n}\sum_{i=1}^{n}z_i^2=0$$

with

$$\begin{aligned}T(\mu,\theta,\boldsymbol{b}^{(v)}) &= \left(1+\alpha(\theta)^2\right)\cdot\delta(\theta)\cdot\sqrt{\frac{2}{\pi}}e^{\eta}\bar{z}-\frac{\alpha(\theta)}{n}\sum_{i=1}^{n}e^{\eta}z_i\\ &\quad\times\left[\frac{\phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)}+\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right].\end{aligned}$$

Solving the above quadratic equation, we update η (and $\sigma=e^{\eta}$) via

$$\begin{aligned}\hat{\sigma}^{(v+1)} &= \sigma(\hat{\eta}^{(v+1)})=e^{\hat{\eta}^{(v+1)}}=\frac{1}{2}T(\hat{\mu}^{(v+1)},\hat{\theta}^{(v)},\boldsymbol{b}^{(v)})\\ &\quad+\sqrt{\frac{1}{4}T(\hat{\mu}^{(v+1)},\hat{\theta}^{(v)},\boldsymbol{b}^{(v)})^2+\frac{1+\alpha(\hat{\theta}^{(v)})^2}{n}\sum_{i=1}^{n}(y_i-\hat{\mu}^{(v+1)})^2}.\end{aligned}$$

Finally, fixing $(\mu,\eta)=(\hat{\mu}^{(v+1)},\hat{\eta}^{(v+1)})$ and letting $f(\theta)=\frac{\partial\Psi}{\partial\theta}$, $f'(\theta)=\frac{\partial^2\Psi}{\partial\theta^2}$, we obtain $\theta^{(v+1)}$ by using the Newton–Raphson iteration to solve the equation $f(\theta)=0$.

In each update, we need to verify whether the incomplete data likelihood is increasing. The PEM algorithm iteration alternates between **E-step** and **M-step** until

$$\left|\frac{l_{incp}(\hat{\mu}^{(v+1)},\hat{\eta}^{(v+1)},\hat{\theta}^{(v+1)}\mid\boldsymbol{y})-l_{incp}(\hat{\mu}^{(v)},\hat{\eta}^{(v)},\hat{\theta}^{(v)}\mid\boldsymbol{y})}{l_{incp}(\hat{\mu}^{(v)},\hat{\eta}^{(v)},\hat{\theta}^{(v)}\mid\boldsymbol{y})}\right|<\varepsilon$$

where ε is the tolerance with default value of 10^{-8}. We take the moment estimates of $(\mu^{(0)},\eta^{(0)},\theta^{(0)})$ as the initial values in the PEM.

It is easy to prove that the PEM has a non-decreasing property similar to that of the standard EM.

5. Numerical results

In this section, we report the results of simulation studies designed to assess the performance of our cross-validated MPLE and to compare it to some existing methods (MLE and Q-based MPLE in the R-package SN, https://CRAN.R-project.org/package=sn) in terms of median bias and standard error of differences between $(\hat{\mu},\hat{\sigma},\hat{\alpha})$ and the ground truth $(\mu_0,\sigma_0,\alpha_0)$.

5.1. Behavior of λ_{cv}

We first examine the asymptotic behavior of λ_{cv} by conducting the following simulation study.

Setting 1: Assume that Y follows a skew normal with unknown parameters (μ,σ,α), where the underlying values $(\mu_0,\sigma_0)=(0,1)$ and $\alpha_0\in\{0,2,3,4\}$. We draw a sample of size n for Y for each combination of (α_0,n), $\alpha_0\in\{0,2,3,4\}$ and $n\in\{50,100,200,300,400,500,600,1000\}$. We repeat this sampling process 20 times, obtaining 20 replicates.

We applied the proposed cross-validation procedure to each sample, obtaining the value of $(\hat{\mu},\hat{\sigma},\hat{\alpha})$ and the value of λ_{cv}. The results are displayed in Figs. 5.1 and 5.2. The results show that when $\alpha_0=0$ (that is, the underlying model is a normal), the sample means and variances of these 20 simulated λ_{cv}/n tend to a constant (≈ 0.0035) and zero respectively; when $\alpha_0\neq 0$ (that is, the underlying model is a skew-normal), both the sample means and variances of these 20 simulated $\lambda_{cv}/\sqrt{n}$ tend to zero. It follows from Chebyshev's inequality that λ_{cv}/n tends to a positive constant in probability when the underlying $\alpha_0=0$, while $\lambda_{cv}/\sqrt{n}$ tends to zero in probability when $\alpha_0\neq 0$. Therefore, the numerical results support the theory we develop in the previous section.

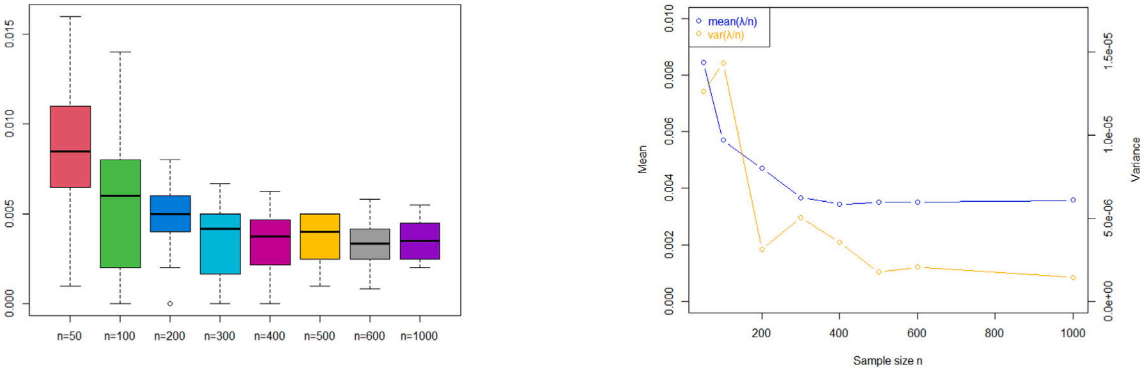


Fig. 5.1. Trend plots of λ_{cv} when the underlying $(\mu_0, \sigma_0^2, \alpha_0) = (0, 1, 0)$, $n = 50, 100, 200{,}300, 400, 500, 600$ and 1000. The box plots on the left, mean–variance chart on the right for 20 simulated λ_{cv}/n.

5.2. Estimation error

The log-likelihood function is non-quadratic at the stationary point $\alpha = 0$, which makes it non-trivial to estimate. Moreover, the MLE of α is diverging with a positive probability. Azzalini and Arellano-Valle (2012) proposed Q-based MPLE by maximising the penalised likelihood $l_p(\mu, \sigma, \alpha) = l(\mu, \sigma, \alpha) - \lambda \log\left(1 + c_2\alpha^2\right)$ in order to tackle the divergent behavior of estimate $\hat{\alpha}$. Using Firth's bias correction technique, they fixed λ and c_2 as constants with $\lambda \approx 0.875913$ and $c_2 \approx 0.856250$. In contrast, we determine the penalty coefficient λ by 10−fold cross-validation. Note that when $\alpha \neq 0$, $l_p(\mu, \sigma, \alpha)$ tends to $-\infty$ as the penalty coefficient λ tends to $+\infty$. So, intuitively, when the true value α_0 is approaching 0, the penalty coefficient λ should be relatively larger compared to the case where α_0 is away from 0. Cross-validation chooses the penalty coefficient by letting dataset speak for itself. In the next simulation study, we demonstrate that our cross-validated MPLE can outperform the MLE and Q-based MPLE procedures when the underlying value $\alpha_0 = 0$. The result confirms the theory developed in the previous section.

Setting 2: We first generate $\mu_0 \sim U(-2, 2)$, $\sigma_0 \sim U(0.5, 1.5)$ and choose $\alpha_0 \in \{0, 1, 2, 3, 5\}$. Then for each combination of (α_0, n), $\alpha_0 \in \{0, 1, 2, 3, 5\}$ and $n \in \{50, 100,\ 200, 400\}$ and given the value of $(\mu_0, \sigma_0, \alpha_0)$, we draw samples of size n from a skew normal with parameters $(\mu_0, \sigma_0, \alpha_0)$. We repeat this sampling process $m = 20$ times, obtaining m replicates.

In Fig. 5.3, the boxplots of estimates of (μ, σ, α) suggest that both the proposed cross-validated MPLE and the Q-based MPLE performed substantially better than the MLE in all cases in terms of mean square error. For the sample size ≥ 500, the cross-validated MPLE does have a strong superiority over the Q-based MPLE, demonstrating that fixing the penalty coefficient to a constant in the Q-based MPLE can compromise the performance of the MPLE. Figures 5.4 and 5.5 demonstrate that the cross-validated MPLE and MPLE virtually coincide when the underlying value α_0 is not zero. When $\alpha_0 = 1$, both the cross-validated MPLE and the Q-based MPLE tend to shrink to zero, suggesting that weak skewness will be filtered out after the penalisation (see Figs. 5.4 and 5.5).

5.3. IC50 data

In this subsection, we considered an IC50 dataset derived from the experiment in Iorio et al. (2016). We first fit the proposed skew-normal model to log-IC50 measurements of each anti-cancer drugs over 111 cancer cell lines and then characterise these drugs by their estimated location, scale and skewness parameters $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$. The results are displayed in Fig. 5.6. There are 170 out of 227 anti-cancer drugs with evident skewness $|\hat{\alpha}| > 1$. We applied K-means to these estimates, obtaining four clusters with centers $(-1.47, 2.50, 3.95)$, $(2.56, 1.50, -0.43)$, $(2.88, 1.91, -5.03)$, $(2.07, 2.54, -27.8)$ and of sizes $45, 107, 65, 10$ respectively. The distributions in Cluster 4 are very negatively skewed, consisting of cancer growth blockers including inhibitors of LCK, BRAF, C-RAF-1, receptor tyrosine kinase and SRC kinase. Cluster 3 is positive log-response group, where the distribution of $\hat{\alpha}$ is negatively skewed with $\hat{\mu}$ mainly taking positive values. The distributions in Cluster 2 are close to normal while the distributions in Cluster 1 (resistance group) are positively skewed with log-response mainly taking negative values. Therefore, classification of drugs to four clusters indicate different patterns of drug response while drugs in the same cluster show a similar mechanism of action.

Here, the numerical result demonstrated the role of the skewness in discriminating functions of drugs. Given empirical distributions of the skewness of the log-IC50 in each drug group described as in Fig. 5.6. the function role of a new drug can be tested again the existing group by comparing its estimated skewness parameter to the expected value of the skewness of the existing group.

6. Discussion and conclusion

We have proposed a novel approach for determining penalty coefficients in the maximum penalised likelihood estimation for skew normal distribution families by using the multifold cross-validation. The proposed procedure has addressed the problem of under-regularisation in the Q-based MPLE caused by fixing the penalty coefficient to a constant. We have conducted an asymptotic

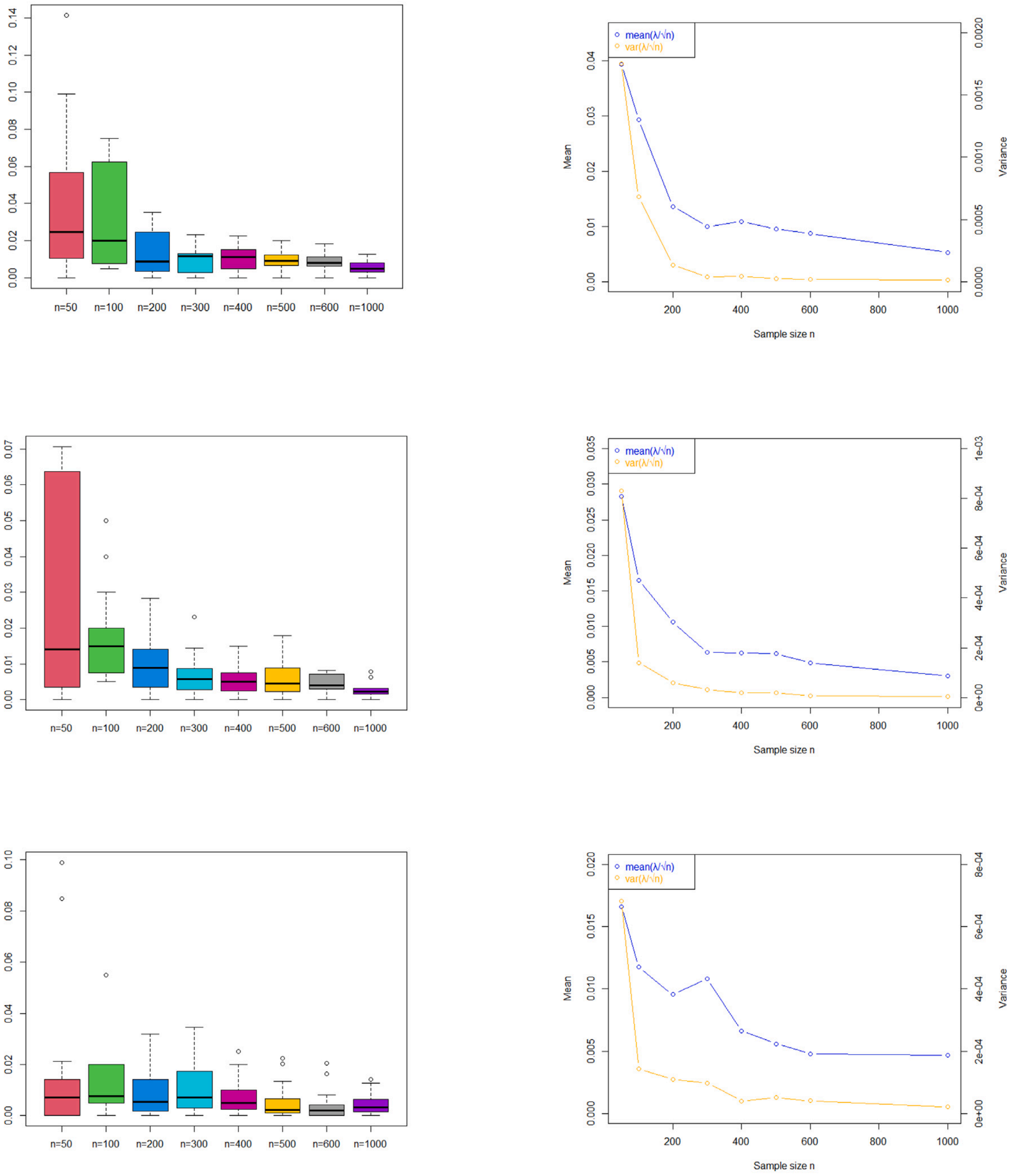


Fig. 5.2. Trend plots of λ_{cv} when the underlying $(\mu_0, \sigma_0^2) = (0, 1)$, $n = 50, 100, 200,300, 400, 500, 600$ and 1000. In each row, the box plots on the left, mean–variance chart on the right for 20 simulated $\lambda_{cv}/\sqrt{n}$. Row 1 for $\alpha_0 = 2$. Row 2 for $\alpha_0 = 3$. Row 3 for $\alpha_0 = 4$.

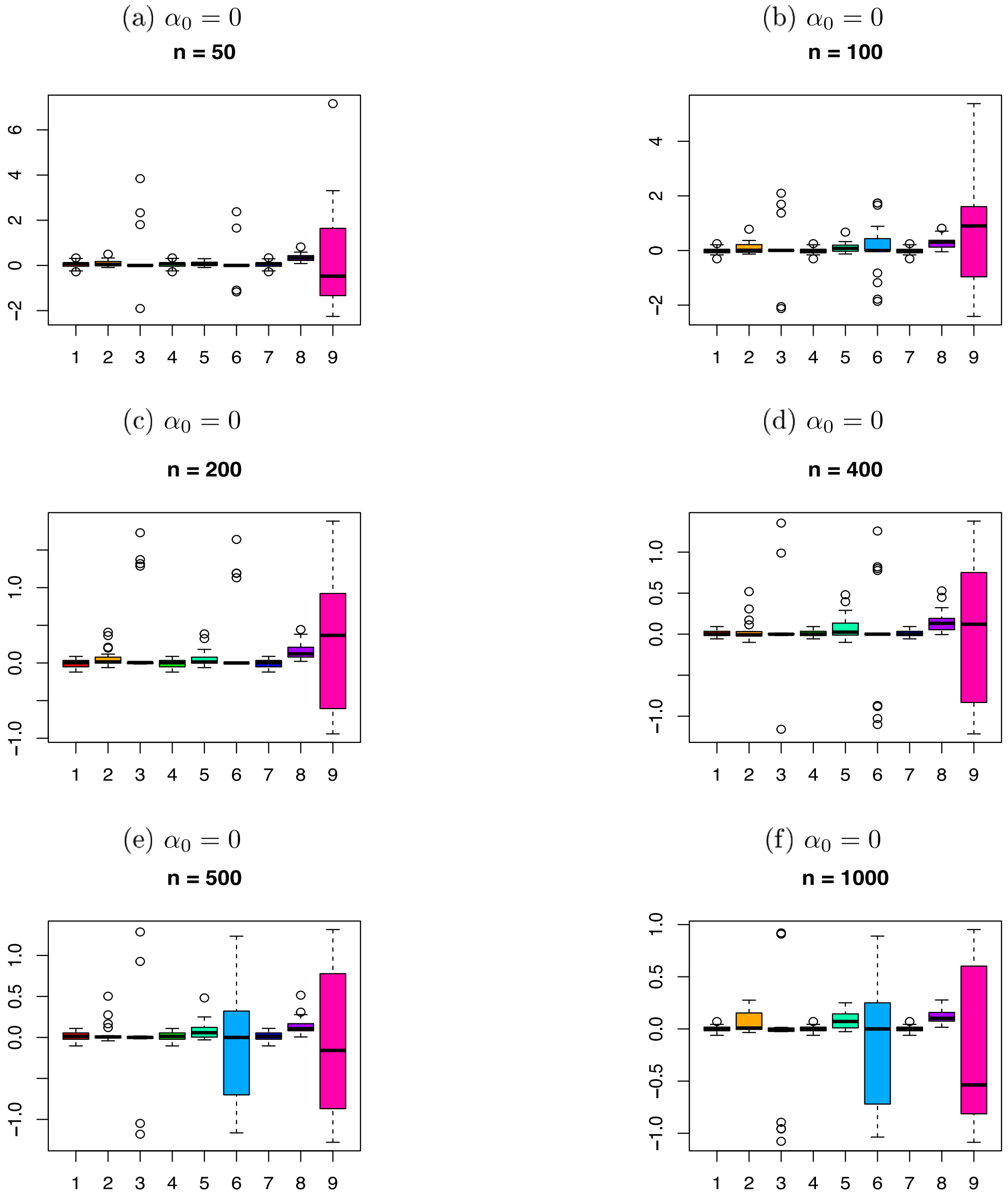


Fig. 5.3. Box plots of $\hat{\mu} - \mu_0$, $\hat{\sigma} - \sigma_0$ and $\hat{\alpha} - \alpha_0$ for 20 replicates when $\alpha_0 = 0$, $\mu_0 \sim U(-2, 2)$, $\sigma_0 \sim U(0.5, 1.5)$ and $n = 50$, 100, 200, 400, 500 and 1000 respectively. In each plot, the first three box-plots, the second three box-plots and the last three box-plots are for the cross-validated MPLE, the MPLE and the MLE, respectively. The plots demonstrate that the cross-validated MPLE outperforms the other two methods in terms of bias and variability, in particular, when the sample size is increasing. Note that the scale of vertical axis in these plots is decreasing when the sample size is increasing.

analysis on the behavior of the proposed procedure. In particular, under some regularity condition, we have shown that the cross-validated MPLE can make a sharp improvement over the Q-based approach. This has resulted in the asymptotic efficiency of the proposed estimators.

We have assessed the performance of the proposed procedure by use of simulated and real data. The simulations have demonstrated that our new procedure can substantially outperform the Q-based MPLE in terms of bias and standard error in a range of scenarios. We have applied the proposed procedure to the analysis of an anti-cancer drug sensitivity dataset, identifying two clusters that have contrasting behavior of drug resistance, one with strong negatively skew drug sensitivity and the other with strong positively skew drug sensitivity. The result is consistent with the existing finding about the role of drug Erlotinib in reducing cancer cell lines resistance to drug Paclitaxel.

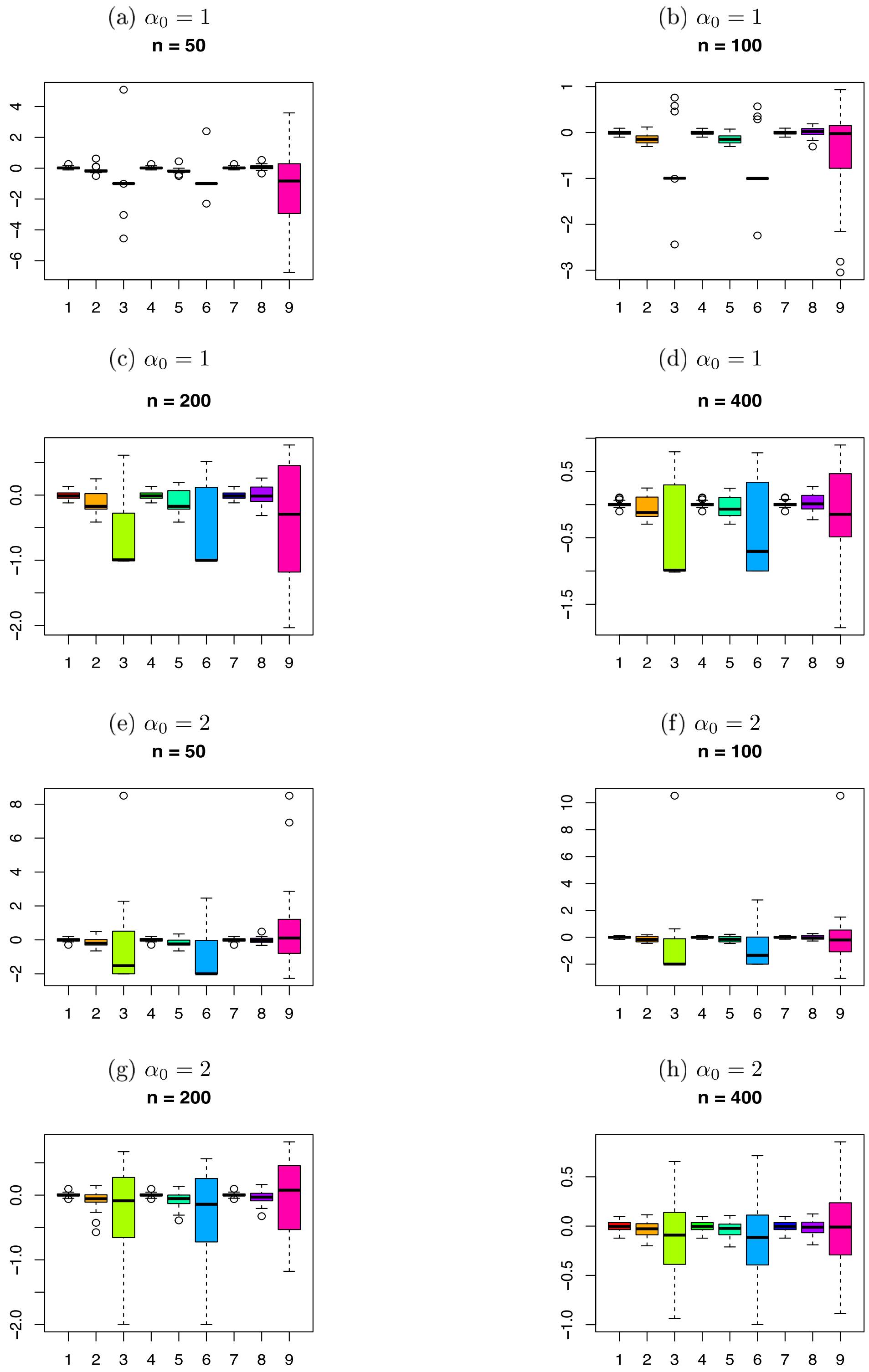
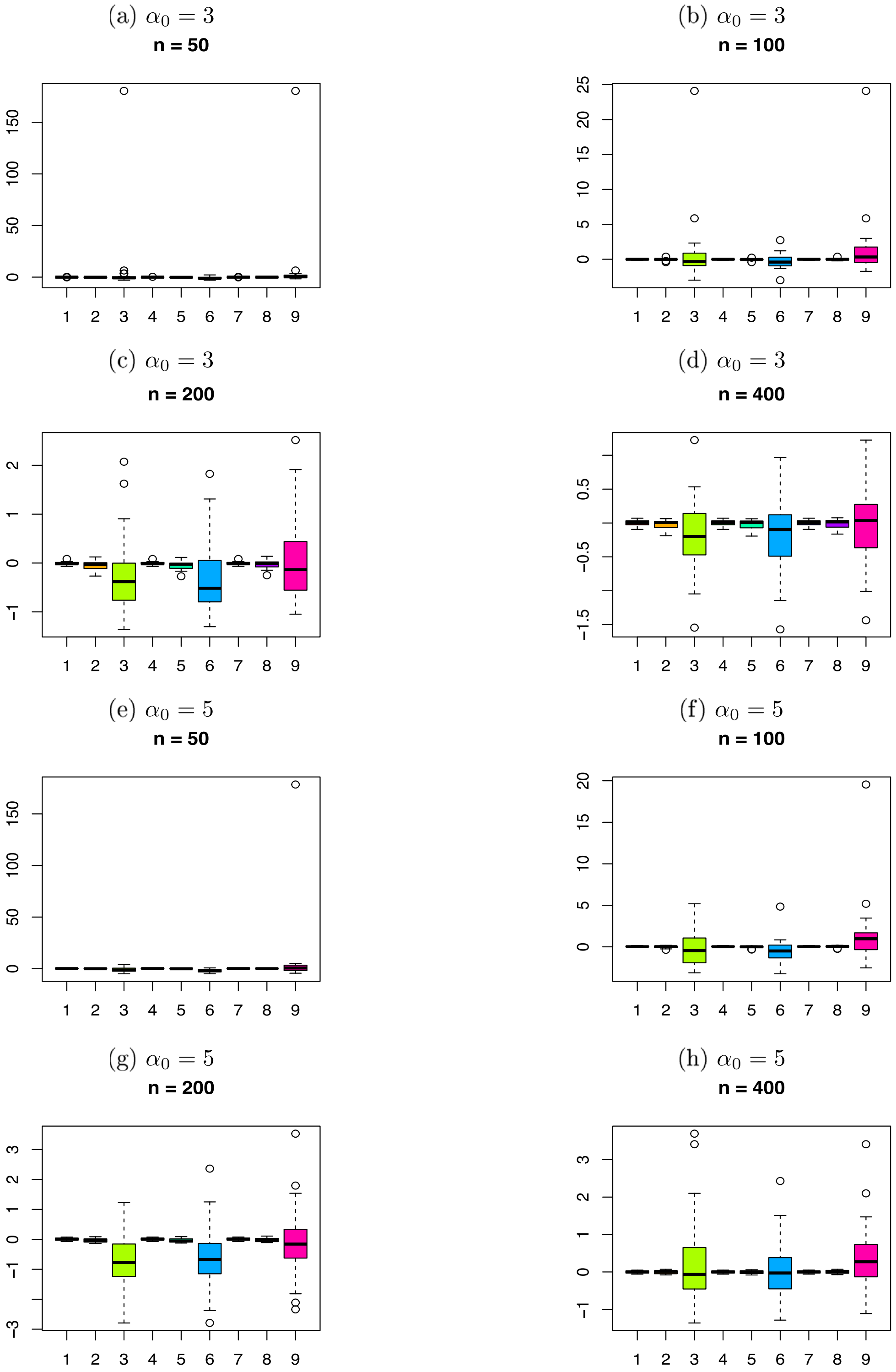


Fig. 5.4. Box plots of $\hat{\mu} - \mu_0$, $\hat{\sigma} - \sigma_0$ and $\hat{\alpha} - \alpha_0$ for 20 replicates when $\alpha_0 = 1$ and 2, $\mu_0 \sim U(-2, 2)$, $\sigma_0 \sim U(0.5, 1.5)$ and $n = 50, 100, 200, 400$ respectively. In each plot, the first three box-plots, the second three box-plots and the last three box-plots are for the cross-validated MPLE, the Q-based MPLE and the MLE, respectively. The plots demonstrate that the cross-validated MPLE performs similarly to the other two methods in terms of bias and variability. Note that the scale of vertical axis in these plots is decreasing when the sample size is increasing.



(a) $\alpha_0 = 3$

(b) $\alpha_0 = 3$

(c) $\alpha_0 = 3$

(d) $\alpha_0 = 3$

(e) $\alpha_0 = 5$

(f) $\alpha_0 = 5$

(g) $\alpha_0 = 5$

(h) $\alpha_0 = 5$

Fig. 5.5. Box plots of $\hat{\mu}-\mu_0$, $\hat{\sigma}-\sigma_0$ and $\hat{\alpha}-\alpha_0$ for 20 replicates when $\alpha_0 = 3$ and 5, $\mu_0 \sim U(-2,2)$, $\sigma_0 \sim U(0.5,1.5)$ and $n = 50, 100, 200, 400$ respectively. In each plot, the first three box-plots, the second three box-plots and the last three box-plots are for the cross-validated MPLE, the Q-based MPLE and the MLE, respectively. The plots demonstrate that the cross-validated MPLE performs similarly to the other two methods in terms of bias and variability. Note that the scale of vertical axis in these plots is decreasing when the sample size is increasing.

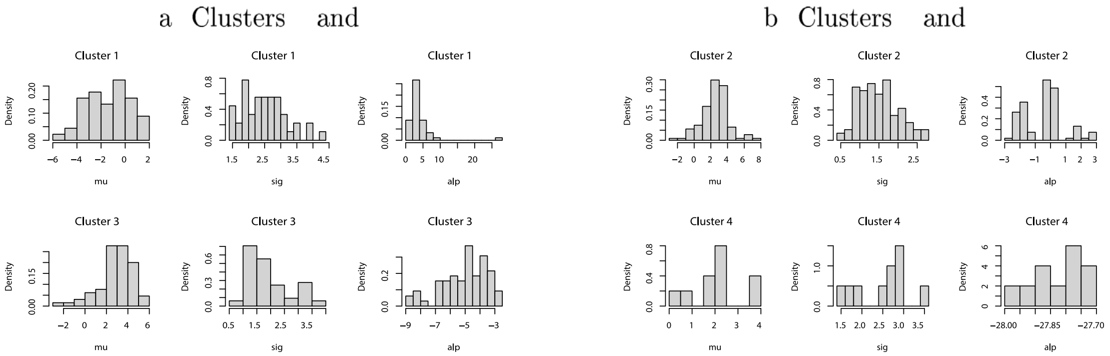


Fig. 5.6. (a) Distribution patterns of estimated $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\alpha}$ in the clusters 1 and 3. (b) Distribution patterns of estimated $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\alpha}$ in the clusters 2 and 4.

Acknowledgements

The research of the first author was supported by the Engineering and Physical Sciences Research Council (EPSRC), United Kingdom under grant (EP/X038297/1) while the research of the second author was supported by the Graduate Teaching Assistant (GTA) scholarship, University of Kent, United Kingdom.

Appendix. Proofs

For the simplicity of notation, in the following let $\boldsymbol{\psi}$ denote (μ, η, θ).

Proof of Proposition 3.1. Without loss of generality, assume that $\mathrm{pen}(\theta) = \mathrm{pen}_1(\theta)$. Note that when $\theta_0 \neq 0$, $\mathbf{C}_0$ is invertible which implies invertibility of $(\mathbf{C}_0 - \mathbf{D}_{0\lambda/n})^{-1}$. We have

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^T &= \frac{1}{\sqrt{n}}(-(\mathbf{C}_0 - \mathbf{D}_{0\lambda/n})^{-1})\frac{\partial l_{inc}(y_i)}{\partial \boldsymbol{\psi}_0^T}(1 + o_p(1)) \\ &+ \frac{\lambda}{2\sqrt{n}}(e^{2\theta_0} - e^{-2\theta_0})(\mathbf{C}_0 - \mathbf{D}_{0\lambda/n})^{-1}\boldsymbol{e}_3(1 + o_p(1))\end{aligned}$$

which is asymptotically normal with asymptotic mean

$$\mathbf{M}_{0\lambda/n} = (\mathbf{C}_0 - \mathbf{D}_{0\lambda/n})^{-1}\frac{\lambda}{2\sqrt{n}}(e^{2\theta_0} - e^{-2\theta_0})\boldsymbol{e}_3$$

and asymptotic covariance matrix

$$\mathbf{V}_{0\lambda/n} = \left(\mathbf{C}_0 - \mathbf{D}_{0\lambda/n}\right)^{-1}(-\mathbf{C}_0)\left(\mathbf{C}_0 - \mathbf{D}_{0\lambda/n}\right)^{-1}.$$

Let $\boldsymbol{d}_{0\lambda/n} = \sqrt{\lambda/n}\sqrt{e^{2\theta_0} + e^{-2\theta_0}}\boldsymbol{e}_3$ and

$$\mathbf{C}_0^{-1} = \begin{pmatrix} c_0^{11} & c_0^{12} & c_0^{13} \\ c_0^{21} & c_0^{22} & c_0^{23} \\ c_0^{31} & c_0^{32} & c_0^{33} \end{pmatrix}, \quad c_0^{\cdot 3} = \begin{pmatrix} c_0^{13} \\ c_0^{23} \\ c_0^{33} \end{pmatrix}, \quad c_0^{3\cdot} = (c_0^{31} \quad c_0^{32} \quad c_0^{33}).$$

Then $\mathbf{D}_{0\lambda/n} = \boldsymbol{d}_{0\lambda/n}\boldsymbol{d}_{0\lambda/n}^T = (\lambda/n)(e^{2\theta_0} + e^{-2\theta_0})\boldsymbol{e}_3\boldsymbol{e}_3^T$ and

$$\begin{aligned}\left(\mathbf{C}_0 - \mathbf{D}_{0\lambda/n}\right)^{-1} &= \left(\mathbf{C}_0 - \boldsymbol{d}_{0\lambda/n}d_{0\lambda/n}^T\right)^{-1} \\ &= \mathbf{C}_0^{-1} + \frac{\mathbf{C}_0^{-1}\boldsymbol{d}_{0\lambda/n}\boldsymbol{d}_{0\lambda/n}^T\mathbf{C}_0^{-1}}{1 + \boldsymbol{d}_{0\lambda/n}^T\mathbf{C}_0^{-1}\boldsymbol{d}_{0\lambda/n}}.\end{aligned}$$

This together with the definitions of $\mathbf{M}_{0\lambda/n}$ and $\mathbf{V}_{0\lambda/n}$ yields

$$\begin{aligned}\mathbf{V}_{0\lambda/n} &= -\left(\mathbf{C}_0^{-1} + \frac{\frac{\lambda}{n}(e^{2\theta_0} + e^{-2\theta_0})}{1 + \frac{\lambda}{n}(e^{2\theta_0} + e^{-2\theta_0})c_0^{33}}\boldsymbol{c}_0^{\cdot 3}\boldsymbol{c}_0^{3\cdot}\right)\mathbf{C}_0 \\ &\times \left(\mathbf{C}_0^{-1} + \frac{\frac{\lambda}{n}(e^{2\theta_0} + e^{-2\theta_0})}{1 + \frac{\lambda}{n}(e^{2\theta_0} + e^{-2\theta_0})c_0^{33}}\boldsymbol{c}_0^{\cdot 3}\boldsymbol{c}_0^{3\cdot}\right) \\ \mathbf{M}_{0\lambda/n} &= \left(\mathbf{C}_0^{-1} + \frac{\frac{\lambda}{n}(e^{2\theta_0} + e^{-2\theta_0})}{1 + \frac{\lambda}{n}(e^{2\theta_0} + e^{-2\theta_0})c_0^{33}}c_0^{\cdot 3}c_0^{3\cdot}\right)\frac{\lambda}{2\sqrt{n}}(e^{2\theta_0} - e^{-2\theta_0})\boldsymbol{e}_3.\end{aligned}$$

So, when $\lambda/\sqrt{n} \to 0$, estimate $(\hat{\mu}, \hat{\eta}, \hat{\theta})$ is asymptotically unbiased and efficient in the sense that its variance asymptotically achieves the Cramer–Rao lower bound. The proof is completed.

Proof of Theorem 1. Without loss of generality, assume that n is a multiple of K and that $n_1 = \cdots = n_K = n/K$. Let $n_{-j} = n - n_j = (K-1)n/K$. Let $\mathbf{D}_{0\lambda/n_{-j}}$ denote $(\lambda/n_{-j})(e^{2\theta_0} + e^{-2\theta_0})\boldsymbol{e}_3\boldsymbol{e}_3^T$. It follows from Eq. (3.2) that

$$\begin{aligned}(\hat{\boldsymbol{\psi}}_{[-j]\lambda} - \boldsymbol{\psi}_0)^T &= \frac{1}{n_{-j}}\left(\sum_{i\in[n]} - \sum_{i\in[j]}\right)(-(\mathbf{C}_0 - \mathbf{D}_{0\lambda/n_{-j}})^{-1})\frac{\partial l_{inc}(y_i)}{\partial \boldsymbol{\psi}_0^T}(1+o_p(1)) \\ &\quad + (\mathbf{C}_0 - \mathbf{D}_{0\lambda/n_{-j}})^{-1}\frac{\lambda}{2n_{-j}}(e^{2\theta_0} - e^{-2\theta_0})\boldsymbol{e}_3(1+o_p(1))\end{aligned}$$

Similarly,

$$(\hat{\boldsymbol{\psi}}_{[j]} - \boldsymbol{\psi}_0)^T = \frac{1}{n_j}\sum_{i\in[j]}(-\mathbf{C}_0^{-1})\frac{\partial l_{inc}(y_i)}{\partial \boldsymbol{\psi}_0^T}(1+o_p(1)).$$

Consequently,

$$\begin{aligned}(\hat{\boldsymbol{\psi}}_{[-j]\lambda} - \hat{\boldsymbol{\psi}}_{[j]})^T &= \frac{K}{K-1}(-(\mathbf{C}_0 - \mathbf{D}_{0\lambda/n_{-j}})^{-1})\frac{1}{n}\sum_{i\in[n]}\frac{\partial l_{inc}(y_i)}{\partial \boldsymbol{\psi}_0^T}(1+o_p(1)) \\ &\quad + (\mathbf{C}_0 - \mathbf{D}_{0\lambda/n_{-j}})^{-1}\frac{\lambda}{2n_{-j}}(e^{2\theta_0} - e^{-2\theta_0})\boldsymbol{e}_3(1+o_p(1)) \\ &\quad - \left(-(\mathbf{C}_0 - \mathbf{D}_{0\lambda/n_{-j}})^{-1}\frac{1}{n_{-j}} - \mathbf{C}_0^{-1}\frac{1}{n_j}\right)\sum_{i\in[j]}\frac{\partial l_{inc}(y_i)}{\partial \boldsymbol{\psi}_0^T}(1+o_p(1))\end{aligned}$$

$$\begin{aligned}&= \frac{K}{K-1}(-(\mathbf{C}_0 - \mathbf{D}_{0\lambda/n_{-j}})^{-1})\frac{1}{n}\sum_{i\in[n]}\frac{\partial l_{inc}(y_i)}{\partial \boldsymbol{\psi}_0^T}(1+o_p(1)) \\ &\quad + \frac{\lambda}{n}\frac{K}{K-1}(\mathbf{C}_0 - \mathbf{D}_{0\lambda/n_{-j}})^{-1}\frac{1}{2}(e^{2\theta_0} - e^{-2\theta_0})(1+o_p(1)) \\ &\quad - \left(-(\mathbf{C}_0 - \mathbf{D}_{0\lambda/n_{-j}})^{-1}\frac{1}{K-1} - \mathbf{C}_0^{-1}\right)\frac{1}{n_j}\sum_{i\in[j]}\frac{\partial l_{inc}(y_i)}{\partial \boldsymbol{\psi}_0^T}(1+o_p(1))\end{aligned}$$

Let $\boldsymbol{v}_i^T = \frac{\partial l_{inc}(y_i)}{\partial \boldsymbol{\psi}_0}(-\mathbf{C}_0)^{-1/2}$ and $\mathbf{W}_{0\lambda/n_{-j}} = (-\mathbf{C}_0)^{1/2}(-(\mathbf{C}_0 - \mathbf{D}_{0\lambda/n_{-j}})^{-1})(-\mathbf{C}_0)^{1/2}$. Then

$$\begin{aligned}&n_j(\hat{\boldsymbol{\psi}}_{[-j]\lambda} - \hat{\boldsymbol{\psi}}_{[j]})\mathbf{C}_0(\hat{\boldsymbol{\psi}}_{[-j]\lambda} - \hat{\boldsymbol{\psi}}_{[j]})^T \\ &= -\frac{K}{(K-1)^2}\left(\frac{1}{\sqrt{n}}\sum_{i\in[n]}\boldsymbol{v}_i^T\right)\mathbf{W}^2_{0\lambda/n_{-j}}\left(\frac{1}{\sqrt{n}}\sum_{i\in[n]}\boldsymbol{v}_i\right) \\ &\quad - \frac{2K}{(K-1)^2}\frac{\lambda}{\sqrt{n}}\left(\frac{1}{\sqrt{n}}\sum_{i\in[n]}\boldsymbol{v}_i^T\right)\mathbf{W}^2_{0\lambda/n_{-j}}(-\mathbf{C}_0)^{-1/2}\frac{1}{2}(e^{2\theta_0} - e^{-2\theta_0})\boldsymbol{e}_3 \\ &\quad + \frac{2\sqrt{K}}{K-1}\left(\frac{1}{\sqrt{n}}\sum_{i\in[n]}\boldsymbol{v}_i^T\right)\left(\mathbf{W}^2_{0\lambda/n_{-j}}\frac{1}{K-1} + \mathbf{W}_{0\lambda/n_{-j}}\right)\left(\frac{1}{\sqrt{n_j}}\sum_{i\in[j]}\boldsymbol{v}_i\right) \\ &\quad - \frac{\lambda^2}{n}\frac{K}{(K-1)^2}\frac{1}{2}(e^{2\theta_0} - e^{-2\theta_0})\boldsymbol{e}_3^T(-\mathbf{C}_0)^{-1/2}\mathbf{W}^2_{0\lambda/n_{-j}}(-\mathbf{C}_0)^{-1/2}\frac{1}{2}(e^{2\theta_0} - e^{-2\theta_0})\boldsymbol{e}_3 \\ &\quad + \frac{2\sqrt{K}}{K-1}\frac{\lambda}{\sqrt{n}}\frac{1}{2}(e^{2\theta_0} - e^{-2\theta_0})\boldsymbol{e}_3^T(-\mathbf{C}_0)^{-1/2}\left(\mathbf{W}^2_{0\lambda/n}\frac{1}{K-1} + \mathbf{W}_{0\lambda/n_{-j}}\right)\frac{1}{\sqrt{n_j}}\sum_{i\in[j]}\boldsymbol{v}_i \\ &\quad - \left(\frac{1}{\sqrt{n_j}}\sum_{i\in[j]}\boldsymbol{v}_i^T\right)\left(\mathbf{W}_{0\lambda/n_{-j}}\frac{1}{K-1} + \mathbf{I}\right)^2\left(\frac{1}{\sqrt{n_j}}\sum_{i\in[j]}\boldsymbol{v}_i\right).\end{aligned}$$

Expanding the jth validated log-likelihood function $l_{inc}(\hat{\boldsymbol{\psi}}_{[-j]\lambda} \mid \boldsymbol{y}_j)$ at the MLE $\hat{\boldsymbol{\psi}}_j$ of $l_{inc}(\boldsymbol{\psi} \mid \boldsymbol{y}_j)$, we have

$$\begin{aligned}l_{inc}(\hat{\boldsymbol{\psi}}_{[-j]\lambda} \mid \boldsymbol{y}_j) &= l_{inc}(\hat{\boldsymbol{\psi}}_j \mid \boldsymbol{y}_j) \\ &\quad + 0.5\sqrt{n_j}(\hat{\boldsymbol{\psi}}_{[-j]\lambda} - \hat{\boldsymbol{\psi}}_j)\frac{1}{n_j}\frac{\partial^2 l_{inc}(\boldsymbol{\psi} \mid \boldsymbol{y}_j)}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^T}\mid_{\hat{\boldsymbol{\psi}}^*_{[-j]\lambda}}\sqrt{n_j}(\hat{\boldsymbol{\psi}}_{[-j]\lambda} - \hat{\boldsymbol{\psi}}_j)^T \\ &= l_{inc}(\hat{\boldsymbol{\psi}}_j \mid \boldsymbol{y}_j) \\ &\quad + 0.5\sqrt{n_j}(\hat{\boldsymbol{\psi}}_{[-j]\lambda} - \hat{\boldsymbol{\psi}}_j)(1+o_p(1))\mathbf{C}_0\sqrt{n_j}(\hat{\boldsymbol{\psi}}_{[-j]\lambda} - \hat{\boldsymbol{\psi}}_j)^T,\end{aligned}$$

where

$$\hat{\boldsymbol{\psi}}^*_{[-j]\lambda} = \hat{\boldsymbol{\psi}}_j + t(\hat{\boldsymbol{\psi}}_{[-j]\lambda} - \hat{\boldsymbol{\psi}}_j), 0 \le t \le 1.$$

Consequently,

$$\begin{aligned}
&\mathrm{CV}_a(\lambda) + \frac{1}{K}\sum_{j=1}^{K} l_{inc}(\hat{\boldsymbol{\psi}}_j \mid \boldsymbol{y}_j) \\
&= -\frac{1}{K}\sum_{j=1}^{K} 0.5\sqrt{n_j}(\hat{\boldsymbol{\psi}}_{[-j]\lambda} - \hat{\boldsymbol{\psi}}_j)\mathbf{C}_0(1+o_p(1))\sqrt{n_j}(\hat{\boldsymbol{\psi}}_{[-j]\lambda} - \hat{\boldsymbol{\psi}}_j)^T \\
&= -\frac{1}{2K}\sum_{j=1}^{K} \sqrt{n_j}(\hat{\boldsymbol{\psi}}_{[-j]\lambda} - \hat{\boldsymbol{\psi}}_j)\mathbf{C}_0\sqrt{n_j}(\hat{\boldsymbol{\psi}}_{[-j]\lambda} - \hat{\boldsymbol{\psi}}_j)^T(1+o_p(1)) \\
&= \left(\frac{\lambda^2}{n}a_{\lambda/n} + 2\frac{\lambda}{\sqrt{n}}b_{\lambda/n} + c_{\lambda/n}\right)(1+o_p(1)) \\
&= \left[a_{\lambda/n}\left(\frac{\lambda}{\sqrt{n}} + b_{\lambda/n}/a_{\lambda/n}\right)^2 - b^2_{\lambda/n}/a_{\lambda/n} + c_{\lambda/n}\right](1+o_p(1)) \\
&\geq \left(-b^2_{\lambda/n}/a_{\lambda/n} + c_{\lambda/n}\right)(1+o_p(1)),
\end{aligned}$$

where

$$\begin{aligned}
a_{\lambda/n} &= \frac{K}{2(K-1)^2}\frac{1}{4}(e^{2\theta_0} - e^{-2\theta_0})^2\boldsymbol{e}_3^T(-\mathbf{C}_0)^{-1/2}\mathbf{W}^2_{0\lambda/n_{-j}}(-\mathbf{C}_0)^{-1/2}\boldsymbol{e}_3, \\
b_{\lambda/n} &= \frac{1}{2(K-1)}\frac{1}{2}(e^{2\theta_0} - e^{-2\theta_0})\left(\frac{1}{\sqrt{n}}\sum_{i\in[n]}\boldsymbol{v}_i^T\right)W_{0\lambda/n_{-j}}(\mathbf{I} - \mathbf{W}_{0\lambda/n_{-j}})(-\mathbf{C}_0)^{-1/2}\boldsymbol{e}_3 \\
c_{\lambda/n} &= \left(\frac{1}{\sqrt{n}}\sum_{i\in[n]}\boldsymbol{v}_i^T\right)\left(\frac{K-2}{2(K-1)^2}\mathbf{W}^2_{0\lambda/n_{-j}} - \frac{1}{K-1}\mathbf{W}_{0\lambda/n_{-j}}\right)\left(\frac{1}{\sqrt{n}}\sum_{i\in[n]}\boldsymbol{v}_i\right) \\
&\quad + \frac{1}{2K}\sum_{j=1}^{K}\left(\frac{1}{\sqrt{n_j}}\sum_{i\in[j]}\boldsymbol{v}_i^T\right)\left(W_{0\lambda/n_{-j}}\frac{1}{K-1} + \mathbf{I}\right)^2\left(\frac{1}{\sqrt{n_j}}\sum_{i\in[j]}\boldsymbol{v}_i\right).
\end{aligned}$$

Over $\lambda/\sqrt{n} \in [0,\infty)$, when $\lambda/\sqrt{n} = \max\{0, -b_{\lambda/n}/a_{\lambda/n}\}$, $\mathrm{CV}_a(\lambda)$ asymptotically attains the minimum

$$-b^2_{\lambda/n}/a_{\lambda/n} + c_{\lambda/n} = -\frac{1}{2K}\frac{\left(\left(\frac{1}{\sqrt{n}}\sum_{i\in[n]}\boldsymbol{v}_i^T\right)\mathbf{W}_{0\lambda/n_{-j}}(\mathbf{I} - \mathbf{W}_{0\lambda/n_{-j}})(-\mathbf{C}_0)^{-1/2}\boldsymbol{e}_3\right)^2}{\boldsymbol{e}_3^T(-\mathbf{C}_0)^{-1/2}W^2_{0\lambda/n_{-j}}(-\mathbf{C}_0)^{-1/2}\boldsymbol{e}_3} + c_{\lambda/n}$$

which is independent of θ_0.

Note that

$$\begin{aligned}
\mathbf{W}_{0\lambda/n_{-j}} &= \mathbf{I} + \frac{\lambda}{n}\frac{K}{K-1}\frac{(e^{2\theta_0} + e^{-2\theta_0})(-\mathbf{C}_0)^{-1/2}\boldsymbol{e}_3\boldsymbol{e}_3^T(-\mathbf{C}_0)^{-1/2}}{1 - \frac{\lambda}{n}\frac{K}{K-1}(e^{2\theta_0} + e^{-2\theta_0})\boldsymbol{e}_3^T(-\mathbf{C}_0)^{-1}\boldsymbol{e}_3}. \\
\mathbf{W}_{0\lambda/n_{-j}}(\mathbf{I} - \mathbf{W}_{0\lambda/n_{-j}}) &= \frac{\lambda}{n}O(1).
\end{aligned}$$

For fixed $\theta_0 \neq 0$, λ satisfying $\lambda/\sqrt{n} \to \infty$ and $0 \leq \lambda/n \leq \omega_0$, it follows from the above equations that $\mathrm{CV}_a(\lambda)$ tends to infinity. While for bounded $\lambda/\sqrt{n}$, λ/n tends to zero, $\mathbf{W}_{0\lambda/n_{-j}} \to I$ and

$$\begin{aligned}
a_{\lambda/n} &\to \frac{K}{2(K-1)^2}\frac{1}{4}(e^{2\theta_0} - e^{-2\theta_0})^2\boldsymbol{e}_3^T(-\mathbf{C}_0)^{-1}\boldsymbol{e}_3, \\
b_{\lambda/n_{-j}}/a_{\lambda/n_{-j}} &= -\frac{\lambda}{\sqrt{n}}O(1)\frac{2(e^{2\theta_0} + e^{-2\theta_0})}{\sqrt{n}(e^{2\theta_0} - e^{-2\theta_0})} \\
&\quad\times\frac{\frac{1}{\sqrt{n}}\sum_{i\in[n]}\boldsymbol{v}_i^T(-\mathbf{C}_0)^{-1/2}\boldsymbol{e}_3\boldsymbol{e}_3^T(-\mathbf{C}_0)^{-1/2}}{1 - (\lambda K/(n(K-1)))(e^{2\theta_0} + e^{-2\theta_0})\boldsymbol{e}_3^T(-\mathbf{C}_0)^{-1}\boldsymbol{e}_3}. \\
\left(\frac{\lambda}{\sqrt{n}} + b_{\lambda/n_{-j}}/a_{\lambda/n_{-j}}\right)^2 &= \frac{\lambda^2}{n}(1 - o_p(1))^2. \\
c_{\lambda/n} &\to \frac{K}{2(K-1)^2}(-\chi_3^2 + \chi_{3K}^2),
\end{aligned}$$

where χ_3^2 and χ_{3K}^2 are two dependented chi-squared random variables. Therefore, for $\lambda/n \in [0,\omega_0]$ $\mathrm{CV}_a(\lambda)$ asymptotically attains the minimum $-\frac{K}{2(K-1)^2}\chi_3^2 + \frac{K}{2(K-1)^2}\chi_{3K}^2$ when $\lambda/\sqrt{n} = 0$. This implies that $\lambda_{cv}/\sqrt{n}$ tends to zero in probability when the true value of $\theta_0 \neq 0$. The proof is completed.

Proof of Proposition 3.2. Note that when $\theta_0 = 0$, $\mathbf{C}_0 = \mathrm{diag}(\sigma_0^{-2}, -2, 0)$ is degenerate. Let

$$\boldsymbol{u}_i = \left(z_{i0}, z_{i0}^2 - 1\right)^T .$$

Then

$$\frac{1}{n}\frac{\partial l_{inc}}{\partial \boldsymbol{\psi}^T} \mid_{(\mu_0,\eta_0,0)} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{u}_i^T, 0)^T$$

Letting $z_i^* = (y_i - \mu_\lambda^*)/\sigma_\lambda^*$, we have

$$\begin{aligned}
c_{11}(\boldsymbol{\psi}_\lambda^*) &= -\frac{1}{\sigma_\lambda^{*2}} - \frac{\alpha_\lambda^{*2}}{\sigma_\lambda^{*2}}\frac{2}{\pi}(1 + o_p(1)) \\
c_{12}(\boldsymbol{\psi}_\lambda^*) &= -\frac{2}{\sigma_\lambda^*}\frac{1}{n}\sum_{i=1}^{n} z_i^* + \frac{\alpha_\lambda^*}{\sigma_\lambda^*}\sqrt{2/\pi}O_p(\theta_\lambda^* + 1/\sqrt{n}).
\end{aligned}$$

$$\begin{aligned}
c_{13}(\boldsymbol{\psi}_\lambda^*) &= -\frac{\delta_\lambda^{*2}}{\sigma_\lambda^*}\sqrt{2/\pi} - \frac{\theta_\lambda^{*2}}{2\sigma_\lambda^*}\sqrt{2/\pi} - \frac{\alpha_\lambda^*}{\sigma_\lambda^*}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi(A_i^*)}{\Phi(A_i^*)} \\
&\quad \times\left(\alpha_\lambda^*\left(z_i^* + \delta_\lambda^*\sqrt{2/\pi}\right) + \frac{\phi(A_i^*)}{\Phi(A_i^*)}\right) \\
&\quad \times\left(\alpha_\lambda^*(1 - \delta_\lambda^{*2})\sqrt{2/\pi}\right) \\
&= O_p(\theta_\lambda^{*2}).
\end{aligned}$$

$$\begin{aligned}
\frac{c_{13}(\boldsymbol{\psi}_\lambda^*)}{\theta_\lambda^{*2}} &\to -\frac{1}{\sigma_\lambda^*}\sqrt{2/\pi}\left(\frac{3}{2} + 2/\pi\right) \text{ as } \theta_\lambda^* \to 0. \\
c_{21}(\boldsymbol{\psi}_\lambda^*) &= c_{12} \mid_{\boldsymbol{\psi}_\lambda^*} \\
c_{22}(\boldsymbol{\psi}_\lambda^*) &= \theta_\lambda^*\frac{4}{\pi}\sum_{i=1}^{n} z_i^{*2} + \theta_\lambda^* O_p(\theta_\lambda^* + 1/\sqrt{n}). \\
\frac{c_{23}(\boldsymbol{\psi}_\lambda^*)}{\theta_\lambda^*} &\to \frac{4}{\pi} \text{ as } n \to \infty. \\
c_{31}(\boldsymbol{\psi}_\lambda^*) &= c_{13}(\boldsymbol{\psi}_\lambda^*), \quad c_{32}(\boldsymbol{\psi}_\lambda^*) = c_{23}(\boldsymbol{\psi}_\lambda^*) \\
c_{33}(\boldsymbol{\psi}_\lambda^*) &= \frac{2}{\pi}\frac{1}{n}\sum_{i=1}^{n}\left(1 - z_i^{*2}\right) + \theta_\lambda^* O_p(\theta_\lambda^* + 1/\sqrt{n}).
\end{aligned}$$

Let

$$\mathbf{I}_{11} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}, \mathbf{I}_{12} = \begin{pmatrix} c_{13} \\ c_{23} \end{pmatrix}, \mathbf{I}_{21} = \mathbf{I}_{12}^T, \mathbf{I}_{22} = c_{33} - \frac{\lambda}{n}(e^{2\theta} + e^{-2\theta}).$$

Let $\mathbf{I}_{110} = \mathrm{diag}(-1/\sigma_0^2, -2)$, $\mathbf{I}_{220} = \frac{2}{\pi}\frac{1}{n}\sum_{i=1}^{n}(1 - z_{i0}^2) - 2\lambda/n$, $\mathbf{I}_{11\lambda}^* = \mathbf{I}_{11} \mid_{(\mu_\lambda^*,\eta_\lambda^*,\theta_\lambda^*)}$, $\mathbf{I}_{12\lambda}^* = \mathbf{I}_{12} \mid_{(\mu_\lambda^*,\eta_\lambda^*,\theta_\lambda^*)}$, $\mathbf{I}_{21\lambda}^* = \mathbf{I}_{21} \mid_{(\mu_\lambda^*,\eta_\lambda^*,\theta_\lambda^*)}$ and $\mathbf{I}_{22\lambda}^* = \mathbf{I}_{22} \mid_{(\mu_\lambda^*,\eta_\lambda^*,\theta_\lambda^*)}$. Then

$$\begin{aligned}
\mathbf{I}_{11}^{-1} &= \frac{1}{c_{11}c_{22} - c_{21}c_{12}}\begin{pmatrix} c_{22} & -c_{12} \\ -c_{21} & c_{11} \end{pmatrix}, \\
\mathbf{I}_{21}\mathbf{I}_{11}^{-1}\mathbf{I}_{12} &= \frac{c_{31}^2 c_{22} - 2c_{32}c_{21}c_{13} + c_{32}^2 c_{11}}{c_{11}c_{22} - c_{21}c_{12}}. \\
\mathbf{I}_{21\lambda}^*\mathbf{I}_{110}^{-1}\mathbf{I}_{12\lambda}^* &= \frac{\theta_\lambda^{*2}O_p(1)}{\frac{1}{\sigma_\lambda^{*2}}\frac{2}{n}\sum_{i=1}^{n} z_i^{*2} + O_p(1/n) + \theta_\lambda^* O_p(\theta^* + 1/\sqrt{n})}. \\
\mathbf{I}_{22\lambda}^* - \mathbf{I}_{21\lambda}^*\mathbf{I}_{110}^{-1}\mathbf{I}_{12\lambda}^* &= \frac{2}{\pi}\frac{1}{n}\sum_{i=1}^{n}\left(1 - z_i^{*2}\right) - \frac{2\lambda}{n} + \theta^* O_p(\theta_\lambda^* + 1/\sqrt{n}). \\
\mathbf{I}_{12\lambda}^* &= \frac{4}{\pi}\theta_\lambda^*\left(O_p(\theta_\lambda^*), \frac{1}{n}\sum_{i=1}^{n} z_i^{*2} + O_p(\theta_\lambda^* + 1/\sqrt{n})\right)^T.
\end{aligned}$$

We have

$$\begin{aligned}
-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{u}_i &= \mathbf{I}_{11\lambda}^*\begin{pmatrix} \sqrt{n}(\hat{\mu}_\lambda - \mu_0) \\ \sqrt{n}(\hat{\eta}_\lambda - \eta_0) \end{pmatrix} + \mathbf{I}_{12\lambda}^*\sqrt{n}\hat{\theta}_\lambda \\
0 &= \mathbf{I}_{21\lambda}^*\begin{pmatrix} \sqrt{n}(\hat{\mu}_\lambda - \mu_0) \\ \sqrt{n}(\hat{\eta}_\lambda - \eta_0) \end{pmatrix} + \mathbf{I}_{22\lambda}^*\hat{\theta}_\lambda
\end{aligned}$$

which implies

$$\begin{aligned}
-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{u}_i - \sqrt{n}\hat{\theta}_\lambda \mathbf{I}^*_{12\lambda} \;&=\; \mathbf{I}_{110}\begin{pmatrix}\sqrt{n}(\hat{\mu}_\lambda-\mu_0)\\ \sqrt{n}(\hat{\eta}_\lambda-\eta_0)\end{pmatrix}(1+o_p(1))\\
(-\mathbf{I}_{110})^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{u}_i + \sqrt{n}\hat{\theta}_\lambda(-\mathbf{I}_{110})^{-1}\mathbf{I}^*_{12\lambda} \;&=\; \begin{pmatrix}\sqrt{n}(\hat{\mu}-\mu_0)\\ \sqrt{n}(\hat{\eta}-\eta_0)\end{pmatrix}(1+o_p(1)),
\end{aligned}$$

$$0=\mathbf{I}^*_{21\lambda}(-\mathbf{I}_{110})^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i + \left(\mathbf{I}^*_{21\lambda}(-\mathbf{I}_{110})^{-1}\mathbf{I}^*_{12\lambda}+\mathbf{I}^*_{22\lambda}\right)\hat{\theta}_\lambda$$

to which $\hat{\theta}_\lambda = 0$ is a solution, since $\mathbf{I}_{21}\mid_{(\mu^*_\lambda,\eta^*_\lambda,0)}=0$ and $\mathbf{I}_{12}\mid_{(\mu^*_\lambda,\eta^*_\lambda,0)}=0$. Note that

$$\begin{aligned}
\frac{1}{n}l_{incp}(\hat{\boldsymbol{\psi}}_\lambda|\boldsymbol{y}) \;=\;& \frac{1}{n}l_{incp}(\mu_0,\eta_0,0|\boldsymbol{y}) + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i^\tau(\hat{\mu}_\lambda-\mu_0,\hat{\eta}_\lambda-\eta_0)^\tau\\
&+\frac{1}{2}(\hat{\mu}_\lambda-\mu_0,\hat{\eta}_\lambda-\eta_0,\hat{\theta}_\lambda)\mathrm{diag}(\mathbf{I}_{110},\mathbf{I}_{220})(\hat{\mu}_\lambda-\mu_0,\hat{\eta}_\lambda-\eta_0,\hat{\theta}_\lambda)^\tau\\
&\times(1+o_p(1))\\
=\;& \frac{1}{n}l_{incp}(\mu_0,\eta_0,0|\boldsymbol{y}) + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i^\tau(\hat{\mu}_\lambda-\mu_0,\hat{\eta}_\lambda-\eta_0)^\tau\\
&+\frac{1}{2}(\hat{\mu}_\lambda-\mu_0,\hat{\eta}_\lambda-\eta_0)\mathbf{I}_{110}(\hat{\mu}_\lambda-\mu_0,\hat{\eta}_\lambda-\eta_0)^\tau(1+o_p(1))\\
&+\frac{1}{2}\mathbf{I}_{220}\hat{\theta}^2_\lambda(1+o_p(1))
\end{aligned}$$

$$\begin{aligned}
=\;& \frac{1}{n}l_{incp}(\mu_0,\eta_0,0|\boldsymbol{y}) + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i^\tau(-\mathbf{I}_{110})^\tau\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i(1+o_p(1))\\
&+\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i^\tau(-\mathbf{I}_{110})^{-1}\mathbf{I}^*_{12\lambda}\hat{\theta}_\lambda(1+o_p(1))\\
&-\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i^\tau(-\mathbf{I}_{110})^{-1}+\mathbf{I}^*_{21\lambda}(-\mathbf{I}_{110})^{-1}\hat{\theta}_\lambda\right)(-\mathbf{I}_{110})\\
&\quad\times\left((-\mathbf{I}_{110})^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i+(-\mathbf{I}_{110})^{-1}\mathbf{I}^*_{12\lambda}\hat{\theta}_\lambda\right)\\
&+\frac{1}{2}\mathbf{I}_{220}\hat{\theta}^2_\lambda(1+o_p(1))\\
=\;& \frac{1}{n}l_{incp}(\mu_0,\eta_0,0|\boldsymbol{y}) + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i^\tau(-\mathbf{I}_{110})^\tau\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i(1+o_p(1))\\
&-\frac{1}{2}\left(-\mathbf{I}_{220}+\mathbf{I}_{21*}(-\mathbf{I}_{110})^{-1}\mathbf{I}^*_{12\lambda}\right)\hat{\theta}^2_\lambda(1+o_p(1))
\end{aligned}$$

which attains the maximum at $\hat{\theta}_\lambda = 0$ when $\mathbf{I}_{220}\le 0$, that is, when

$$\frac{\lambda}{\sqrt{n}} \ge \max\left\{\frac{1}{\pi}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-z^2_{i0}\right),0\right\}.$$

This implies:

- When $\lambda/\sqrt{n}\to\infty$, we have $\hat{\theta}_\lambda = 0$.
- When $\frac{1}{\pi}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-z^2_{i0}\right)\le 0$, for any $\lambda \ge 0$, we have $\hat{\theta}_\lambda = 0$.

When $\hat{\theta}_\lambda = 0$, we have

$$\begin{pmatrix}\sqrt{n}(\hat{\mu}-\mu_0)/\sigma_0\\ \sqrt{n}(\hat{\eta}-\eta_0)\end{pmatrix}. = \begin{pmatrix}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}z_{i0}\\ \frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\left(z^2_{i0}-1\right)\end{pmatrix}(1+o_p(1))$$

which is asymptotically normal with mean zero and covariance matrix diag$(1,1/2)$. When

$$0\le\frac{\lambda}{\sqrt{n}}<\frac{1}{\pi}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-z^2_{i0}\right)-\frac{\sqrt{n}}{2}\mathbf{I}^*_{21\lambda}(-\mathbf{I}_{110})^{-1}\mathbf{I}^*_{12\lambda},$$

we have

$$\frac{1}{n}l_{incp}(\hat{\boldsymbol{\psi}}_\lambda|\boldsymbol{y}) > \frac{1}{n}l_{incp}(\mu_0,\eta_0,0|\boldsymbol{y}) + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i^\tau(-\mathbf{I}_{110})^\tau\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i(1+o_p(1)).$$

Therefore, when $\frac{1}{\pi}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-z_{i0}^2\right)>0$, there is λ such that $\frac{1}{n}l_{incp}(\hat{\mu}_\lambda,\hat{\eta}_\lambda,\hat{\theta}_\lambda|\boldsymbol{y})$ attains the maximum at non-zero $\hat{\theta}_\lambda$ satisfying

$$\frac{\sqrt{n}}{2}\mathbf{I}_{21\lambda}^{*}(-\mathbf{I}_{110})^{-1}\mathbf{I}_{12\lambda}^{*}<\frac{1}{\pi}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-z_{i0}^2\right).$$

The proof is completed.

Proof of Theorem 2. It follows from

$$\frac{\partial l_{inc}(\hat{\boldsymbol{\psi}}_{[-j]\lambda}\mid \boldsymbol{y}_{[-j]})}{\partial \boldsymbol{\psi}^T}=0$$

and Taylor's theorem that

$$-\frac{1}{n_{[-j]}}\sum_{i\in[-j]}(\boldsymbol{u}_i^T,0)^T=\begin{pmatrix}\mathbf{I}_{11} & \mathbf{I}_{12}\\ \mathbf{I}_{21} & \mathbf{I}_{22}\end{pmatrix}_{\boldsymbol{\psi}^*_{[-j]\lambda}}(\hat{\boldsymbol{\psi}}_{[-j]\lambda}-(\mu_0,\eta_0,0)),$$

where $\boldsymbol{\psi}^*_{[-j]\lambda}=t\times(\hat{\boldsymbol{\psi}}_{[-j]\lambda})$ for some $0\le t\le 1$. Furthermore, if $\hat{\theta}_{[-j]\lambda}\neq 0$, then

$$\frac{\mathbf{I}_{21\lambda}^{*[-j]}}{\hat{\theta}_{[-j]\lambda}}(-\mathbf{I}_{110})^{-1}\frac{1}{n_{[-j]}}\sum_{i\in[-j]}\boldsymbol{u}_i+\mathbf{I}_{21\lambda}^{*[-j]}(-\mathbf{I}_{110})^{-1}\mathbf{I}_{12\lambda}^{*[-j]}+\mathbf{I}_{22\lambda}^{*[-j]}=0.$$

Consider λ at which $\mathbf{I}_{22\lambda}^{*[-j]}-\mathbf{I}_{21\lambda}^{*[-j]}I_{110}^{-1}\mathbf{I}_{12\lambda}\neq 0$. We have

$$\begin{aligned}(\hat{\boldsymbol{\psi}}_{[-j]\lambda}-(\mu_0,\eta_0,0))^T&=(-\mathbf{I}_{110})^{-1}\left(\frac{1}{n_{[-j]}}\sum_{i\in[-j]}\boldsymbol{u}_i+\mathbf{I}_{12\lambda}^{*[-j]}\hat{\theta}_\lambda\right)(1+o_p(1))\\&=\Big((-\mathbf{I}_{110})^{-1}-(-\mathbf{I}_{110})^{-1}\mathbf{I}_{12\lambda}^{*[-j]}\\&\quad\times\left(\mathbf{I}_{22\lambda}^{*[-j]}+\mathbf{I}_{21\lambda}^{*[-j]}(-\mathbf{I}_{110})^{-1}\mathbf{I}_{12*}\right)^{-1}\mathbf{I}_{21*}(-I_{110})^{-1}\Big)\\&\quad\times\frac{1}{n_{[-j]}}\sum_{i\in[-j]}\boldsymbol{u}_i(1+o_p(1)).\end{aligned}\tag{A.1}$$

Similarly, it follows from

$$\frac{\partial l_{inc}(\hat{\mu}_j,\hat{\eta}_j,0\mid \boldsymbol{y}_j)}{\partial \boldsymbol{\psi}^T}=0$$

and Taylor's theorem that

$$(\hat{\mu}_j-\mu_0,\hat{\eta}_j-\eta_0)^T=(-\mathbf{I}_{110})^{-1}\frac{1}{n_j}\sum_{i\in[j]}\boldsymbol{u}_i(1+o_p(1)).$$

Consequently, we have

$$\begin{aligned}&(\hat{\mu}_{[-j]\lambda}-\hat{\mu}_j,\hat{\eta}_{[-j]\lambda}-\hat{\eta}_j)^T\\&=(-\mathbf{I}_{110})^{-1}\left(\frac{1}{n_{[-j]}}\sum_{i\in[-j]}\boldsymbol{u}_i+\mathbf{I}_{12\lambda}^{*[-j]}\hat{\theta}_\lambda-\frac{1}{n_j}\sum_{i\in[j]}\boldsymbol{u}_i\right)(1+o_p(1))\\&=\frac{1}{n_{[-j]}}\sum_{i\in[-j]}\Big\{(-\mathbf{I}_{110})^{-1}-(-\mathbf{I}_{110})^{-1}\mathbf{I}_{12\lambda}^{*[-j]}\\&\quad\times\left(\mathbf{I}_{22\lambda}^{*[-j]}+\mathbf{I}_{21\lambda}^{*[-j]}(-\mathbf{I}_{110})^{-1}\mathbf{I}_{12\lambda}^{*[-j]}\right)^{-1}\mathbf{I}_{21\lambda}^{*[-j]}(-\mathbf{I}_{110})^{-1}\Big\}\boldsymbol{u}_i(1+o_p(1))\\&\quad-\frac{1}{n_j}\sum_{i\in[j]}(-\mathbf{I}_{110})^{-1}\boldsymbol{u}_i(1+o_p(1)).\end{aligned}$$

It follows from Proposition 3.2 that for $\lambda\ge\max\left\{\frac{1}{\pi}\sum_{i\in[-j]}\left(1-z_{i0}^2\right),0\right\}$, we have $\hat{\theta}_{[-j]\lambda}=0$, $1\le j\le K$.

On other hand, using the Taylor expansion, we have

$$\begin{aligned}&l_{inc}(\hat{\mu}_{[-j]\lambda},\hat{\eta}_{[-j]\lambda},0\mid \boldsymbol{y}_j)-l_{inc}(\hat{\mu}_j,\hat{\eta}_j,0\mid \boldsymbol{y}_j)\\&=\frac{n_j}{2}(\hat{\boldsymbol{\psi}}_{[-j]\lambda}-\hat{\boldsymbol{\psi}}_j)\frac{1}{n_j}\frac{\partial^2 l_{inc}(\hat{\boldsymbol{\psi}}^*_{[-j]\lambda})}{\partial\boldsymbol{\psi}^T\partial\boldsymbol{\psi}}(\hat{\boldsymbol{\psi}}_{[-j]\lambda}-\hat{\boldsymbol{\psi}}_j)^T\\&=\frac{n_j}{2}(\hat{\boldsymbol{\psi}}_{[-j]\lambda}-\hat{\boldsymbol{\psi}}_j)\mathrm{diag}(-\sigma_0^{-2},-2,0)(\hat{\boldsymbol{\psi}}_{[-j]\lambda}-(\hat{\mu}_j,\hat{\eta}_j,0))(1+o_p(1))\\&=-\frac{n_j}{2}(\hat{\boldsymbol{\psi}}_{[-j]\lambda}-\hat{\boldsymbol{\psi}}_j)\mathbf{I}_{110}(\hat{\boldsymbol{\psi}}_{[-j]\lambda}-\hat{\boldsymbol{\psi}}_j)^T(1+o_p(1))\\&=-\frac{n_j}{2}\left(\frac{1}{n_{[-j]}}\sum_{i\in[-j]}\boldsymbol{u}_i-\frac{1}{n_j}\sum_{i\in[j]}\boldsymbol{u}_i\right)^T(-\mathbf{I}_{110})^{-1}\end{aligned}$$

$$\times\left(\frac{1}{n_{[-j]}}\sum_{i\in[-j]}\boldsymbol{u}_i-\frac{1}{n_j}\sum_{i\in[j]}\boldsymbol{u}_i\right)(1+o_p(1)). \tag{A.2}$$

Therefore,

$$\begin{aligned}
&\mathrm{CV}(\lambda)+\frac{1}{K}\sum_{j=1}^{K}l_{inc}(\hat{\mu}_j,\hat{\eta}_j,0\mid \boldsymbol{y}_j)\\
&=\frac{1}{2K}\sum_{j=1}^{K}n_j\left(\frac{1}{n_{[-j]}}\sum_{i\in[-j]}\boldsymbol{u}_i-\frac{1}{n_j}\sum_{i\in[j]}\boldsymbol{u}_i\right)^T(-I_{110})^{-1}\\
&\quad\times\left(\frac{1}{n_{[-j]}}\sum_{i\in[-j]}\boldsymbol{u}_i-\frac{1}{n_j}\sum_{i\in[j]}\boldsymbol{u}_i\right)(1+o_p(1)).
\end{aligned}$$

The proof is completed.

References

Azzalini, A., 1985. A class of distributions which includes the normal ones. Scand. J. Stat. 12, 171–178.

Azzalini, A., Arellano-Valle, R., 2013. Maximum penalized likelihood estimation for skew-normal and skew-t distributions. J. Stat. Plan. & Inference 143, 419–433.

Azzalini, A., Capitanio, A., 1999. Statistical applications of the multivariate skew normal distribution. J. R. Stat. Soc. Ser. B 61, 579–602.

Azzalini, A., Capitanio, A., 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. J. R. Stat. Soc. Ser. B 65, 367–389.

Azzalini, A., Capitanio, A., 2014. The Skew-Normal and Related Families. IMS Monographs. Cambridge University Press.

Buttyan, L., Schaffer, P., Vajda, I., 2006. Resilient aggregation with attack detection in sensor networks. In: The 4th IEEE Conference on Pervasive Computing and Communications Workshops (PerCom 2006 Workshops), Pisa, Italy. pp. 332–336.

Chiogna, M., 2005. A note on the asymptotic distribution of the maximum likelihood estimator for the scalar skew-normal distribution. Stat. Methods Appl. 14, 331–342.

Colacito, R., et al., 2016. Skewness in expected macro fundamentals and the predictability of equity returns: evidence and theory. Rev. Financ. Stud. 29, 2069–2109.

Dhar, S., et al., 1996. Anti-cancer drug characterisation using a human cell line panel representing defined types of drug resistance. Br. J. Cancer 74, 888–896.

Ferrante, M., Pacei, S., 2017. Small domain estimation of business statistics by using multivariate skew normal models. J. Roy. Statist. Soc. Ser. A 180, 1057–1088.

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

Hallin, M., Ley, C., 2012. Skew-symmetric distributions and Fisher information-a tale of two densities. Bernoulli 18, 747–763.

He, W., et al., 2013. Detecting abrupt change on the basis of skewness: numerical tests and applications. Int. J. Climatol. 33, 2713–2727.

Iorio, et al., 2016. A landscape of pharmacogenomic interactions in cancer. Cell 166, 740–754.

Lin, T., 2009. Maximum likelihood estimation for multivariate skew normal mixture models. J. Multivariate Anal. 100, 257–265.

Lv, Y., et al., 2019. Erlotinib overcomes paclitaxel-resistant cancer stem cells by blocking the EGFR-CREB/GRβ-IL-6 axis in MUC1-positive cervical cancer. Oncogenesis 8 (70).

Wang, S., Zimmerman, D., Breheny, P., 2020. Sparsity-regularized skewness estimation for the multivariate skew normal and multivariate skew t distributions. J. Multi. Anal. 179, 1046.39.

Watanabe, S., 2021. Information criteria and cross validation for Bayesianinference in regular and singular cases. Jpn. J. Stat. Data Sci. 4, 1–19.