Computational Statistics and Data Analysis 189 (2024) 107852

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Standard error estimates in hierarchical generalized linear models

Shaobo Jin [a,*], Youngjo Lee [b]

[a] Department of Mathematics, Department of Statistics, Uppsala University, Ångströmlaboratoriet, Lägerhyddsvägen 1, Uppsala, 75106, Sweden
[b] Department of Statistics, Seoul National University, Gwanak-ro 1, Gwanak-gu, Seoul, 08826, Republic of Korea

ARTICLE INFO

Article history:
Received 29 March 2022
Received in revised form 5 September 2023
Accepted 6 September 2023
Available online 14 September 2023

Keywords:
Laplace approximation
REML
h-likelihood
Sandwich estimator

ABSTRACT

Hierarchical generalized linear models are often used to fit random effects models. However, attention is mostly paid to the estimation of fixed unknown parameters and inference for latent random effects. In contrast, standard error estimators receive less attention than they should be. Currently, the standard error estimators are based on various approximations, even when the mean parameters may be estimated from a higher-order approximation of the likelihood and the dispersion parameters are estimated by restricted maximum likelihood. Existing standard error estimation procedures are reviewed. A numerical illustration shows that the current standard errors are not necessarily accurate. Alternative standard errors are also proposed. In particular, a sandwich estimator that accounts for the dependence between the mean parameters and the dispersion parameters greatly improve the current standard errors.

1. Introduction

A wide range of statistical models consist of the observed data $\boldsymbol{x}$, the latent variables $\boldsymbol{v}$, and the fixed unknown parameters $\boldsymbol{\theta}$. Widely encountered models of this kind include hierarchical generalized linear models (HGLM, Lee and Nelder, 1996), models for missing values (Little and Rubin, 2002), and factor analysis models (Anderson and Rubin, 1956). For models with latent variables, the observed log-likelihood is obtained from the complete likelihood $f(\boldsymbol{x}, \boldsymbol{v}; \boldsymbol{\theta})$ by integrating out the latent variables, i.e.,

$$\ell(\boldsymbol{\theta}) \equiv \log \int \exp\{h(\boldsymbol{v}, \boldsymbol{\theta})\} d\boldsymbol{v}, \tag{1}$$

where $h(\boldsymbol{v}, \boldsymbol{\theta}) = \log f(\boldsymbol{x}, \boldsymbol{v}; \boldsymbol{\theta})$. The joint density $f(\boldsymbol{x}, \boldsymbol{v}; \boldsymbol{\theta})$ is known as the extended likelihood of $\boldsymbol{\theta}$ and $\boldsymbol{v}$. Bjørnstad (1996) showed that, given the statistical model, all information regarding $\boldsymbol{\theta}$ and $\boldsymbol{v}$ is contained in the extended likelihood. However, naively using the extended likelihood can lead to spurious results (Bayarri et al., 1988; Lee and Nelder, 2005). Hence, Lee et al. (2006) proposed the use of the h-likelihood, which is defined as the extended likelihood with $\boldsymbol{v}$ in the (weak) canonical scale. The canonical scale requires that the conditional distribution of $\boldsymbol{v}$ given $\boldsymbol{y}$ carries no information about $\boldsymbol{\theta}$. The weak canonical scale requires that the random effects combine additively with fixed effects in the linear predictor. The reader

* Corresponding author.
E-mail addresses: shaobo.jin@math.uu.se (S. Jin), youngjo@snu.ac.kr (Y. Lee).

https://doi.org/10.1016/j.csda.2023.107852

is directed to Lee et al. (2017) for a thorough discussion of h-likelihood and Jin and Lee (2021) for a recent review. The h-likelihood principle has been applied to generalized linear models (GLMs) with random effects (Lee and Nelder, 1996, 2001a), frailty models (Ha and Lee, 2003; Ha et al., 2001), robust modeling (Lee and Nelder, 2006), multiple testing (Lee and Bjørnstad, 2013), factor analysis (Jin et al., 2018; Wu and Bentler, 2012), just to name a few.

The integral in (1) is often intractable and approximations are necessary. Thus, the use of the Laplace approximation to (1) is proposed for the h-likelihood approaches, which is closely related to the adjusted profile likelihood of Cox and Reid (1987) and remains computationally efficient even when the dimension of $\boldsymbol{v}$ is high. To reduce the bias of the estimators of the dispersion parameters, especially in binary data models, the second-order Laplace approximation is also proposed (Lee and Nelder, 1996, 2001a).

Despite the h-likelihood principle being applied to various models, the issue of standard errors is much less studied. To our knowledge, Lee (2002) is the only study devoted to the standard errors. However, he only considered the mean parameters, not the dispersion parameters. A consequence of using the Laplace approximation is that its Hessian matrix is generally difficult, if not impossible, to compute. Regardless of the order of approximation, the R package `dhglm` (Lee and Noh, 2018) always approximates the Hessian of the first-order Laplace approximation, whereas the package `mdhglm` (Lee et al., 2018) extracts the standard errors of the parameter estimators from the Hessian matrix of $h(\boldsymbol{v}, \boldsymbol{\theta})$. To our knowledge, the standard errors used in the h-likelihood models are not fully justified in the literature, especially for the dispersion parameters. Our main focus is to offer a justification to the current standard error estimators of h-likelihood models and to investigate alternative standard error estimators.

The rest of the paper is organized as follows. First, the inferential procedure of the h-likelihood approach is briefly reviewed. Second, the h-likelihood standard errors are justified. Third, some alternative standard error estimators are studied, followed by numerical illustrations. A conclusion ends the paper.

2. An overview of the inferential procedures

We focus on the HGLM proposed by Lee and Nelder (2001a) because it is general enough to cover various random-effects models (Lee et al., 2017). The HGLM inferential procedure is briefly reviewed here.

Given the random effects, the response variable y_i belongs to the exponential dispersion family, given by

$$f(y_i \mid \boldsymbol{v}) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c_1(y_i, \phi_i)\right\}, \tag{2}$$

where θ_i is the natural parameter, ϕ_i is the dispersion parameter, and $b()$ and $c_1()$ and some functions. Suppose that the linear predictor for the mean model is

$$\boldsymbol{\eta}_\mu = g_\mu(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v},$$

where $\boldsymbol{\mu} = E(\boldsymbol{y} \mid \boldsymbol{v})$, $g_\mu()$ is the link function, $\boldsymbol{X}$ is an $n \times q_\mu$ model matrix for fixed effects $\boldsymbol{\beta}$, and $\boldsymbol{Z}$ is an $n \times q_v$ model matrix for random effects $\boldsymbol{v}$. For the distribution assumption (2), we let $V_\mu(\mu_i)$ be its variance function, and $\boldsymbol{W}_y$ be a diagonal matrix with diagonal entries

$$\frac{1}{V_\mu(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_{\mu,i}}\right)^2,$$

where μ_i is the ith entry of $\boldsymbol{\mu}$ and $\eta_{\mu,i}$ is the ith entry of $\boldsymbol{\eta}_\mu$.

The distribution of the random effects is

$$f(v_k) = \exp\left\{\frac{\psi v_k - k(v_k)}{\sigma_k} + c_2(\sigma_k)\right\}, \tag{3}$$

where v_k is the kth entry in $\boldsymbol{v}$, $c_2()$ is some function, σ_k is the dispersion parameter, and ψ is a known constant that depends on the specific distribution assumption. The reader is directed to Lee and Nelder (2001a) for examples illustrating the correspondence between the distribution assumption and ψ. For the simplicity of notation, we assume that the random effects are mutually independent. In the case of correlated random effects, the method of Lee and Nelder (2001b) can be easily used to transform the correlated random effects into independent ones. For the distribution assumption (3), we will use $V_v()$ and $\boldsymbol{W}_v$ to denote the analogues of $V_y()$ and $\boldsymbol{W}_y$ of distribution (2), respectively.

HGLMs allow regression models for dispersion parameters such as

$$\boldsymbol{\eta}_\sigma = g_\sigma(\boldsymbol{\sigma}) = \boldsymbol{G}_\sigma\boldsymbol{\gamma}_\sigma, \text{ and } \boldsymbol{\eta}_\phi = g_\phi(\boldsymbol{\phi}) = \boldsymbol{G}_\phi\boldsymbol{\gamma}_\phi, \tag{4}$$

where $\boldsymbol{\sigma}$ is the vector that collects all σ_k, $\boldsymbol{\phi}$ is the vector that collects all ϕ_i, $g_\sigma()$ and $g_\phi()$ are the link functions, and $\boldsymbol{G}_\sigma$ and $\boldsymbol{G}_\phi$ are $q_v \times q_\sigma$ and $n \times q_\phi$ model matrices, respectively. Hence, there are two unknowns in the HGLM, namely, random effects $\boldsymbol{v}$ and fixed parameters $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T\right)^T$, where $\boldsymbol{\beta}$ is the mean parameter and $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_\phi^T, \boldsymbol{\gamma}_\sigma^T\right)^T$ is the dispersion parameter. The logarithm of h-likelihood (Lee and Nelder, 1996) is then

$$h(\boldsymbol{v},\boldsymbol{\theta}) = \log f(\boldsymbol{y} \mid \boldsymbol{v}) + \log f(\boldsymbol{v}),$$

since $\boldsymbol{v}$ appears additively in the linear predictor. In the HGLM, we may derive inferential procedures by using $h(\boldsymbol{v},\boldsymbol{\theta})$, which has both unknowns as arguments. Hereafter, we suppress the arguments of h for simplicity, unless confusion may arise.

2.1. Joint maximization

In some special cases, such as the normal linear mixed model and the Poisson regression model with gamma random effects, $\boldsymbol{v}$ and $\boldsymbol{\beta}$ can be jointly estimated using the h-likelihood. In this context, $\hat{\boldsymbol{v}}$ and $\hat{\boldsymbol{\beta}}$ are the zeros of

$$\frac{\partial h}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{\Phi}^{-1} \boldsymbol{W}_y \frac{\partial \boldsymbol{\eta}_\mu}{\partial \boldsymbol{\mu}^T} (\boldsymbol{y} - \boldsymbol{\mu}),$$
$$\frac{\partial h}{\partial \boldsymbol{v}} = \boldsymbol{Z}^T \boldsymbol{\Phi}^{-1} \boldsymbol{W}_y \frac{\partial \boldsymbol{\eta}_\mu}{\partial \boldsymbol{\mu}^T} (\boldsymbol{y} - \boldsymbol{\mu}) + \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\psi} - \frac{\partial k(\boldsymbol{v})}{\partial \boldsymbol{v}} \right),$$

given $\boldsymbol{\gamma}$, where $\boldsymbol{\Phi}$ is a diagonal matrix with diagonal entries ϕ_i, $\partial \boldsymbol{\eta}_\mu / \partial \boldsymbol{\mu}^T$ is a diagonal matrix with diagonal elements $\partial \eta_{\mu,i} / \partial \mu_i$, and $\boldsymbol{\Sigma}$ is a diagonal matrix with diagonal entries σ_k. This is referred to as the joint maximization. Let $\boldsymbol{\delta} = \left(\boldsymbol{v}^T, \boldsymbol{\beta}^T\right)^T$. As shown in Lee and Nelder (2001a), the Hessian of h satisfies

$$E\left(-\frac{\partial^2 h}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} | \boldsymbol{v} \right) = \begin{bmatrix} \boldsymbol{X}^T \boldsymbol{\Phi}^{-1} \boldsymbol{W}_y \boldsymbol{X}^T & \boldsymbol{X}^T \boldsymbol{\Phi}^{-1} \boldsymbol{W}_y \boldsymbol{Z} \\ \boldsymbol{Z}^T \boldsymbol{\Phi}^{-1} \boldsymbol{W}_y \boldsymbol{X}^T & \boldsymbol{Z}^T \boldsymbol{\Phi}^{-1} \boldsymbol{W}_y \boldsymbol{Z} + \boldsymbol{\Sigma}^{-1} \boldsymbol{W}_v \end{bmatrix}. \tag{5}$$

Hence, the estimating equation to update $\boldsymbol{v}$ and $\boldsymbol{\beta}$ is

$$\boldsymbol{T}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{W} \boldsymbol{T} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{v} \end{bmatrix} = \boldsymbol{T}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{W} \begin{bmatrix} \boldsymbol{\eta}_\mu + \frac{\partial \boldsymbol{\eta}_\mu}{\partial \boldsymbol{\mu}^T} (\boldsymbol{y} - \boldsymbol{\mu}) \\ \boldsymbol{z}_1 \end{bmatrix}, \tag{6}$$

for some vector $\boldsymbol{z}_1$, where

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_y & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{W}_v \end{bmatrix}, \quad \boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{\Phi} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix}, \quad \boldsymbol{T} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}.$$

The exact expression of $\boldsymbol{z}_1$ can be found in Lee and Nelder (2001a). In particular, if $\boldsymbol{v}$ is Gaussian, then $\boldsymbol{z}_1 = \boldsymbol{0}$. It can be shown from equation (6) that the equation to update $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} = \left(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \left[\boldsymbol{\eta}_\mu + \frac{\partial \boldsymbol{\eta}_\mu}{\partial \boldsymbol{\mu}^T} (\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{V} \boldsymbol{\Phi}^{-1} \boldsymbol{W}_y \boldsymbol{Z} \boldsymbol{H}^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{W}_v \boldsymbol{z}_1 \right], \tag{7}$$

where $\boldsymbol{V} = \boldsymbol{W}_y^{-1} + \boldsymbol{Z} \boldsymbol{\Sigma} \boldsymbol{Z}^T$ and $\boldsymbol{H} = \boldsymbol{Z}^T \boldsymbol{\Phi}^{-1} \boldsymbol{W}_y \boldsymbol{Z} + \boldsymbol{\Sigma}^{-1} \boldsymbol{W}_v$. Equation (7) works in a similar way as the iterative re-weighted least squares for a regular GLM with $\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}$ being the information matrix, from which the standard errors can be drawn.

Lee and Nelder (2001a) and Lee and Kim (2016) proposed to extract the standard errors of $\hat{\boldsymbol{\beta}}$ from the $(2,2)$th block of the inverse of (5), which is the negative of the inverse of

$$\frac{\partial^2 h}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} - \frac{\partial^2 h}{\partial \boldsymbol{\beta} \partial \boldsymbol{v}^T} \left(\frac{\partial^2 h}{\partial \boldsymbol{v} \partial \boldsymbol{v}^T} \right)^{-1} \frac{\partial^2 h}{\partial \boldsymbol{v} \partial \boldsymbol{\beta}^T}. \tag{8}$$

Equation (8) reduces to $-\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}$ in HGLMs.

2.2. Maximum likelihood

Joint maximization does not always yield valid estimators and Lee et al. (2006) suggested to investigate its adequacy on a model-by-model basis. For example, in the Rasch (1960) model, Haberman (1977) and Haberman (2004) showed that the number of items needs to increase to infinite in order for the joint maximization estimator to be consistent. For the HGLM with binary data, Lee and Nelder (2001a) estimated the mean parameter $\boldsymbol{\beta}$ by joint maximization. Yun and Lee (2004) further proposed to estimate $\boldsymbol{\beta}$ using first-order Laplace approximation to reduce bias in $\boldsymbol{\beta}$. In general, to approximate the integral (1), we may use the first-order Laplace approximation based on the h-likelihood

$$p_v(h) \equiv h(\boldsymbol{v}(\boldsymbol{\beta}), \boldsymbol{\theta}) - \frac{1}{2} \log \left| -\frac{1}{2\pi} \frac{\partial^2 h(\boldsymbol{v},\boldsymbol{\theta})}{\partial \boldsymbol{v} \partial \boldsymbol{v}^T} \right|_{\boldsymbol{v}=\boldsymbol{v}(\boldsymbol{\beta})} \Bigg|.$$

The latent variable $\boldsymbol{v}$ in $p_v(h)$ is evaluated at $\hat{\boldsymbol{v}} = \boldsymbol{v}(\boldsymbol{\beta})$, the solution of $\partial h / \partial \boldsymbol{v} = \boldsymbol{0}$ given $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The notation $\boldsymbol{v}(\boldsymbol{\beta})$ is used to emphasize that $\boldsymbol{v}$ is merely a function of $\boldsymbol{\beta}$ for a given $\boldsymbol{\gamma}$. In contrast, the notation $\boldsymbol{v}$ is used if it is not considered

as a function of $\boldsymbol{\beta}$ when computing the derivatives. Hence, $p_v(h)$ is simply an adjusted profile log-likelihood, profiling out $\boldsymbol{v}$. The subscript v in $p_v(h)$ indicates that the latent variable $\boldsymbol{v}$ is profiled out. The approximated maximum likelihood (ML) estimator of $\boldsymbol{\beta}$ is obtained by maximizing $p_v(h)$ given $\boldsymbol{\gamma}$. Since $\hat{\boldsymbol{v}}$ depends on the value of $\boldsymbol{\beta}$, it is essential to take the dependence between $\hat{\boldsymbol{v}}$ and $\boldsymbol{\beta}$ into consideration in ML. The implicit function theorem implies that

$$\frac{\partial \boldsymbol{v}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} = -\left(\frac{\partial^2 h(\boldsymbol{v},\boldsymbol{\theta})}{\partial \boldsymbol{v}\partial \boldsymbol{v}^T}\right)^{-1} \frac{\partial^2 h(\boldsymbol{v},\boldsymbol{\theta})}{\partial \boldsymbol{v}\partial \boldsymbol{\beta}^T}.$$

Hence, the gradient of the approximation to the observed log-likelihood is

$$\frac{\partial p_v(h)}{\partial \boldsymbol{\theta}} = \frac{\partial h(\boldsymbol{v},\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\beta}} \log\left| -\frac{\partial^2 h(\boldsymbol{v},\boldsymbol{\theta})}{\partial \boldsymbol{v}\partial \boldsymbol{v}^T}\right|_{\boldsymbol{v}=\boldsymbol{v}(\boldsymbol{\beta})}\Bigg|.$$

When $p_v(h)$ is used to estimate $\boldsymbol{\beta}$, the estimating equation is

$$\boldsymbol{T}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{W} \boldsymbol{T} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T \boldsymbol{\Phi}^{-1} \boldsymbol{W}_y \left[\boldsymbol{\eta}_\mu + \frac{\partial \boldsymbol{\eta}_\mu}{\partial \boldsymbol{\mu}^T}(\boldsymbol{y}-\boldsymbol{s}-\boldsymbol{\mu})\right] \\ \boldsymbol{Z}^T \boldsymbol{\Phi}^{-1} \boldsymbol{W}_y \left[\boldsymbol{\eta}_\mu + \frac{\partial \boldsymbol{\eta}_\mu}{\partial \boldsymbol{\mu}^T}(\boldsymbol{y}-\boldsymbol{\mu})\right] + \boldsymbol{\Sigma}^{-1} \boldsymbol{W}_v \boldsymbol{z}_1 \end{bmatrix} \tag{9}$$

for some $\boldsymbol{s}$ (see Noh and Lee, 2007, for its expression). If $\boldsymbol{s}$ is set to zero, then equation (9) reduces to equation (6) of joint maximization. The equation to update $\boldsymbol{\beta}$ is then

$$\begin{aligned} \boldsymbol{\beta} = \left(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \Bigg[& \boldsymbol{\eta}_\mu + \frac{\partial \boldsymbol{\eta}_\mu}{\partial \boldsymbol{\mu}^T}(\boldsymbol{y}-\boldsymbol{\mu}) - \boldsymbol{V}\boldsymbol{\Phi}^{-1}\boldsymbol{W}_y \frac{\partial \boldsymbol{\eta}_\mu}{\partial \boldsymbol{\mu}^T}\boldsymbol{s} \\ & - \boldsymbol{V}\boldsymbol{\Phi}^{-1}\boldsymbol{W}_y \boldsymbol{Z}\boldsymbol{H}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_v \boldsymbol{z}_1 \Bigg], \end{aligned} \tag{10}$$

(Noh and Lee, 2007), which is similar to equation (7) for joint maximization. However, a nonzero $\boldsymbol{s}$ can greatly reduce the bias of joint maximization if the latter is severely biased (Yun and Lee, 2004; Noh and Lee, 2007). Equation (10) is still similar to the iterative re-weighted least squares for a regular GLM. Hence, the R package dhglm (Lee and Noh, 2018) extracts the standard errors of $\hat{\boldsymbol{\beta}}$ from $\left(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1}$, regardless whether h or $p_v(h)$ is used for the estimation of $\boldsymbol{\beta}$.

2.3. Restricted maximum likelihood

In the small sample case, the restricted maximum likelihood (REML) procedure is commonly used to reduce the bias of the dispersion parameter estimators. The REML estimator of $\boldsymbol{\gamma}$ in the h-likelihood approach is often obtained by maximizing the first-order approximation

$$p_{v,\beta}(h) \equiv h - \frac{1}{2}\log\left|-\frac{1}{2\pi}\frac{\partial^2 h}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T}\right|,$$

where $\boldsymbol{\delta} = \left(\boldsymbol{v}^T, \boldsymbol{\beta}^T\right)^T$, $\boldsymbol{v}$ is evaluated at $\hat{\boldsymbol{v}}$, and $\boldsymbol{\beta}$ is evaluated at the approximate ML estimator. It is often the case that the first-order approximation is not accurate enough for the dispersion parameters. Hence, Lee and Nelder (2001a) proposed to use the second-order Laplace approximation

$$s_{v,\beta}(h) \equiv h - \frac{1}{2}\log\left|-\frac{1}{2\pi}\frac{\partial^2 h}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T}\right| + \log\left(1 + r(\boldsymbol{v},\boldsymbol{\theta})\right). \tag{11}$$

We direct the reader to Shun and McCullagh (1995) for the expression of $r(\boldsymbol{v},\boldsymbol{\theta})$. In practice, $\log(1+r)$ is often similar to r. Hence, Lee and Nelder (2001a) among others used $r(\boldsymbol{v},\boldsymbol{\theta})$ directly in $s_{v,\beta}(h)$, instead of $\log(1+r(\boldsymbol{v},\boldsymbol{\theta}))$. To be more specific, $\boldsymbol{v}$ is estimated by maximizing h; $\boldsymbol{\beta}$ is estimated by maximizing h or $p_v(h)$; and $\boldsymbol{\gamma}$ is estimated by maximizing $p_{v,\beta}(h)$ or $s_{v,\beta}(h)$. It is worth mentioning that we are not aware of any higher order approximation (e.g., third-order approximation) in HGLM, partly due to even higher order partial derivatives in higher order approximations.

Lee and Nelder (2001a) also proposed to approximate h in $p_{v,\beta}(h)$ and $s_{v,\beta}(h)$ by the double extended quasi-likelihood given by

$$Q = -\frac{1}{2}\sum_{i=1}^{n}\left\{\frac{d_{0i}}{\phi_i} + \log\left[2\pi\phi_i V_y(y_i)\right]\right\} - \frac{1}{2}\sum_{k=1}^{q}\left\{\frac{d_{1k}}{\sigma_k} + \log\left[2\pi\sigma_k V_v(\psi)\right]\right\},$$

where d_{0i} and d_{1k} are the deviance residuals of $y_i|\boldsymbol{v}$ and v_k, respectively. The reader is directed to Lee and Nelder (2001a) for the corresponding expressions. They argued that using Q is advantageous over h, since it allows overdispersed distribution and a broader class of models can be fitted in the same framework, if no substantial bias is introduced. They showed

Table 1
Objective functions to be optimized for estimation.

Distribution of $y\|v$	Parameters			
	$\boldsymbol{v}$	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}_\sigma$	$\boldsymbol{\gamma}_\phi$
Gaussian	h	h	$p_{v,\beta}(h)$	$p_{v,\beta}(h)$
Gamma	h	$p_v(h)$	$p_{v,\beta}(Q)$	$p_{v,\beta}(h)$
Poisson	h	$p_v(h)$	$p_{v,\beta}(Q)$	-
Bernoulli	h	$p_v(h)$	$s_{v,\beta}(Q)$	-

that, given $\boldsymbol{v}$ and $\boldsymbol{\beta}$, the REML estimation of $\boldsymbol{\gamma}$ reduce to a gamma GLM. For example, the gradients with respect to $\gamma_{\sigma,j}$ and $\gamma_{\phi,j}$ are

$$\sum_{k=1}^{q} \frac{l_{\sigma,k}}{2} \frac{d_{1k}/l_{\sigma,k} - \sigma_k}{\sigma_k^2} \frac{\partial \lambda_k}{\partial \eta_{\sigma,k}} G_{\sigma,kj} \text{ and } \sum_{i=1}^{n} \frac{l_{\phi,i}}{2} \frac{d_{0i}/l_{\phi,i} - \phi_i}{\phi_i^2} \frac{\partial \phi_i}{\partial \eta_{\phi,i}} G_{\phi,ij},$$

respectively, where

$$l_{\sigma,k} = 1 - a_{n+k} + \left(\sum_{j=1}^{n+q} a_j w_j^{-1} \frac{\partial w_j}{\partial \sigma_k} \right) \sigma_k \text{ and } l_{\phi,i} = 1 - a_i + \left(\sum_{j=1}^{n+q} a_j w_j^{-1} \frac{\partial w_j}{\partial \phi_i} \right) \phi_i,$$

with $\boldsymbol{A} = \boldsymbol{T} \left(\boldsymbol{T}^T \boldsymbol{W} \boldsymbol{\Gamma}^{-1} \boldsymbol{T} \right)^{-1} \boldsymbol{T}^T \boldsymbol{W} \boldsymbol{\Gamma}^{-1}$, a_j being the jth diagonal entry of $\boldsymbol{A}$, and w_j being the jth diagonal entry of $\boldsymbol{W}$. The derivative of $s_{v,\beta}(Q)$ has a similar expression as $p_{v,\beta}(Q)$, but another $l_{\sigma,k}$ and $l_{\phi,i}$. Due to its complexity, we skip its exact expression here. The reader is directed to Noh and Lee (2007) for more details.

Because the gradient has a similar form as the gradient of a gamma GLM, the dhglm package estimates the standard errors from the gamma GLM fit. To be more specific, the dhglm package extracts the standard errors of $\boldsymbol{\gamma}_\sigma$ and $\boldsymbol{\gamma}_\phi$ from the inverse of $\boldsymbol{G}_\sigma^T \text{diag}\left(2^{-1} l_{\sigma,k}\right) \boldsymbol{G}_\sigma$ and the inverse of $\boldsymbol{G}_\phi^T \text{diag}\left(2^{-1} l_{\phi,i}\right) \boldsymbol{G}_\phi$, respectively.

2.4. Numerical illustration

To gain insights into the h-likelihood standard errors, four generalized linear mixed models (GLMMs) with random intercepts are considered. All models have the linear predictor

$$\beta_0 + v_j + \beta_1 x_{1jk} + \beta_2 x_{2jk}, \quad j = 1, ..., J, \; k = 1, ..., K. \tag{12}$$

The first two models are the Gaussian GLMM with the identity link and the gamma GLMM with the log link, which have continuous responses. The last two models are the Bernoulli GLMM with the logit link and the Poisson GLMM with the log link, which have discrete responses. In all models, we vary J as 30, 60, 90, and 120, and fix K as 10. The mean parameters are set to $\beta_0 = -1$, $\beta_1 = 1.0$ and $\beta_2 = 1.0$. The random effect v_j follows a normal distribution with mean 0 and variance $\exp(\gamma_0)$ with $\gamma_0 = \log(1.5)$. The covariates are generated from independent standard normal distributions. For the Gaussian model and the gamma model, we let $\phi_i = 0.5$ for all i. The objective functions to be optimized for different parameters are tabulated in Table 1. It is worth mentioning that $h = Q$ in the Gaussian model, and that $p_{v,\beta}(h)$ is used for the gamma model to reduce the bias. If $p_{v,\beta}(Q)$ is used for the Gamma model to estimate $\boldsymbol{\gamma}_\phi$, the resulting estimator is biased.

The standard error methods considered in the illustration are the approach used by the dhglm package ($\left(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1}$ for $\boldsymbol{\beta}$ and the Gamma GLM for $\boldsymbol{\gamma}$) and the approach by equation (8) with $\boldsymbol{\beta}$ replaced by $\boldsymbol{\theta}$. The latter will be referred to as the Hessian of h. The reason of extending equation (8) to $\boldsymbol{\theta}$ is that in the h-likelihood literature the standard errors of the mean parameters are often discussed but not the dispersion parameters. For example, Lee and Kim (2016) studied the joint distribution of the estimated fixed effects and the estimated random effects, but the dispersion parameter is excluded. A justification of such extension will be offered in the later section.

To evaluate the accuracy of the estimated standard errors, we divide the mean of the estimated standard errors by the sample standard deviation for each parameter, where the sample standard deviation is treated as the pseudo true value. In order to reduce the uncertainty in the sample standard deviation, the number of replications is set to 10,000. However, standard errors are computed only for the first 1,000 replications, due to computational complexity. To mitigate the effects of outliers, the point estimates and the standard errors are trimmed before computing the mean and standard deviation. The estimates that are more extreme than 2.5 times the interquartile range are regarded as outliers and discarded from further analysis.

It is seen from Fig. 1 that both methods generally yield accurate standard error estimates for the mean parameters. The method used by the dhglm package is biased for the dispersion parameters. In contrast, using the Hessian of h generally yields accurate standard error estimates for the Gaussian and Gamma models, but is biased for the Poisson and Bernoulli models. Results reported later show that approximately 7% of the estimated covariance matrices using the Hessian of h are not positive definite in the Bernoulli model when $J = 30$. These standard error estimates are removed from the plots.

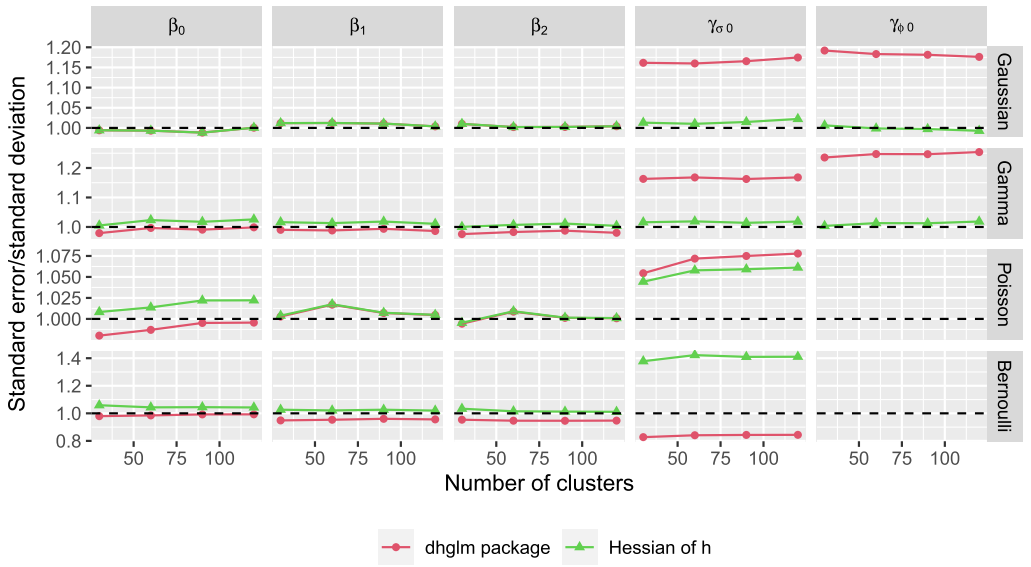


Fig. 1. The mean of the estimated standard errors divided by the sample standard deviation for each parameter.

3. Justification from first-order Laplace approximations

3.1. Standard error for maximum likelihood

When $p_v(h)$ is used to estimate $\boldsymbol{\beta}$ for a fixed $\boldsymbol{\gamma}$, its Hessian may be used to estimate the standard errors. However, the inclusion of the logarithmized determinant makes the exact Hessian difficult to obtain. If all partial derivatives of $\log\left|-\partial^2 h/\partial \boldsymbol{v}\partial \boldsymbol{v}^T\right|$ are ignored and we only focus on the first two terms in $\partial p_v(h)/\partial\boldsymbol{\beta}$, it can be easily shown that

$$\frac{\partial^2 p_v(h)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \approx \frac{\partial^2 h}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} - \frac{\partial^2 h}{\partial\boldsymbol{\beta}\partial\boldsymbol{v}^T}\left(\frac{\partial^2 h}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T}\right)^{-1}\frac{\partial^2 h}{\partial\boldsymbol{v}\partial\boldsymbol{\beta}^T},$$

which is the same as (8). In fact, ignoring $\log\left|-\partial^2 h/\partial\boldsymbol{v}\partial\boldsymbol{v}^T\right|$ can be justified from the first-order Laplace approximation to the observed gradient $\partial\ell(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. For a fixed $\boldsymbol{\gamma}$, the gradient of ℓ with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial\ell}{\partial\boldsymbol{\beta}} = \frac{1}{\int\exp\{h\}\,d\boldsymbol{v}}\cdot\int\frac{\partial h}{\partial\boldsymbol{\beta}}\exp\{h\}\,d\boldsymbol{v}, \tag{13}$$

where both the numerator and denominator involve intractable integrals. Following Evans and Swartz (1995) and Strawderman (2000), the first-order Laplace approximation to the integral of the following form is

$$\int g(\boldsymbol{v})\exp\{h(\boldsymbol{v})\}\,d\boldsymbol{v} \approx g(\boldsymbol{v})\left|-\frac{1}{2\pi}\frac{\partial^2 h(\boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T}\right|^{-1/2}\exp\{h(\boldsymbol{v})\},$$

where $\boldsymbol{v}$ is evaluated at the solution of $\partial h(\boldsymbol{v})/\partial\boldsymbol{v}=\boldsymbol{0}$. If both the numerator and the denominator in (13) are approximated by the first-order Laplace approximations, then

$$\frac{\partial\ell}{\partial\boldsymbol{\beta}} \approx \frac{\frac{\partial h}{\partial\boldsymbol{\beta}}\left|-\frac{1}{2\pi}\frac{\partial^2 h}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T}\right|^{-1/2}\exp\{h\}}{\left|-\frac{1}{2\pi}\frac{\partial^2 h}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T}\right|^{-1/2}\exp\{h\}} = \frac{\partial h}{\partial\boldsymbol{\beta}}, \tag{14}$$

where $\boldsymbol{v}$ is evaluated at $\hat{\boldsymbol{v}}$ after taking all partial derivatives. Hereafter, equation (14) is referred to as the first-order Laplace approximated observed gradient. The approximation (14) is simply the gradient of the logarithm of h-likelihood, used in the joint maximization, which can perform poorly for the dispersion parameter in the models for Bernoulli data as shown in Noh and Lee (2007). Nevertheless, the gradient of equation (14) can be viewed as an approximation to the observed Hessian. In particular,

$$\frac{\partial}{\partial\boldsymbol{\beta}}\left(\frac{\partial h(\boldsymbol{v}(\boldsymbol{\beta}),\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\right) = \frac{\partial^2 h(\boldsymbol{v},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} + \left(\frac{\partial\boldsymbol{v}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\right)^T\frac{\partial^2 h(\boldsymbol{v},\boldsymbol{\beta})}{\partial\boldsymbol{v}\partial\boldsymbol{\beta}^T}, \tag{15}$$

which reduces to equation (8), the approach used by the dhglm package.

Hence, the standard error using equation (8) can also be viewed as the exact Jacobian of the approximated observed gradient when the observed gradient is approximated by the ratio of two first-order Laplace approximations. When estimating the fixed unknown parameters, the h-likelihood approach uses the exact gradient of the approximation to the observed log-likelihood. In contrast, the h-likelihood approach uses the exact derivative of the approximated gradient to compute the standard errors, if the Hessian of $p_v(h)$ is complicated to compute. If equation (14) accurately approximates the gradient of $\ell(\boldsymbol{\beta})$, equation (8) is expected to yield accurate standard error estimates. The pattern in Fig. 1 indicates that the above justification works reasonably well.

3.2. Standard error for restricted maximum likelihood

In the special case where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are asymptotically orthogonal, Cox and Reid (1987) showed that $p_\beta(\ell)$ is an approximation to the conditional log-likelihood $\log L\left(\boldsymbol{\gamma}|\hat{\boldsymbol{\beta}}\right)$. Lee and Nelder (2001a) showed that $p_{v,\beta}(h)$ approximates $p_\beta(\ell)$ to the same order as the first-order Laplace approximation. Consequently, $p_{v,\beta}(h)$ can be viewed as an approximation to $\log L\left(\boldsymbol{\gamma}|\hat{\boldsymbol{\beta}}\right)$ under the orthogonality assumption. In a general case, Meng (2009) suggested that $p_{v,\beta}(h)$ is an approximation to $\int \exp\{h\}\, d\boldsymbol{\delta}$ with $\boldsymbol{\delta} = \left(\boldsymbol{v}^T, \boldsymbol{\beta}^T\right)^T$ under the uniform prior using the first-order Laplace approximation. Hence, the REML estimator of $\boldsymbol{\gamma}$ approximates the solution of

$$\mathbf{0} = \frac{\partial \log \int \exp\{h\}\, d\boldsymbol{\delta}}{\partial \boldsymbol{\gamma}} = \frac{1}{\int \exp\{h\}\, d\boldsymbol{\delta}} \int \frac{\partial h}{\partial \boldsymbol{\gamma}} \exp\{h\}\, d\boldsymbol{\delta}.$$

If the integrals are approximated by the first-order Laplace approximations, then

$$\frac{\partial \log \int \exp\{h\}\, d\boldsymbol{\delta}}{\partial \boldsymbol{\gamma}} \approx \frac{\frac{\partial h}{\partial \boldsymbol{\gamma}} \left| -\frac{1}{2\pi} \frac{\partial^2 h}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} \right|^{-1/2} \exp\{h\}}{\left| -\frac{1}{2\pi} \frac{\partial^2 h}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} \right|^{-1/2} \exp\{h\}} = \frac{\partial h}{\partial \boldsymbol{\gamma}}.$$

Let $\boldsymbol{\delta} = (\boldsymbol{\beta}^T, \boldsymbol{v}^T)^T$. Consequently, the Hessian of h is a first-order Laplace approximation to the Hessian needed for REML. In other words, we may use the curvature of h to estimate the standard errors of $\boldsymbol{\gamma}$. Hence, the negative of inverse of (8) with $\boldsymbol{\beta}$ replaced by $\boldsymbol{\theta}$ can be used for the standard errors of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

4. Alternative approaches

It is seen in Fig. 1 that neither the approach used by the dhglm package nor the Hessian of h is always accurate. Hence, alternative standard error estimators are needed, which will be studied in this section.

4.1. Standard error from Hessian matrix

When the dispersion parameter $\boldsymbol{\gamma}$ is estimated from $p_{v,\beta}(Q)$, Lee and Nelder (2001a) proposed to estimate the standard errors from

$$-\left(\frac{\partial^2 p_{v,\beta}(Q)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T}\right)^{-1},$$

where Q is defined in Section 2.3. They showed how to compute it in their appendix. Since $\boldsymbol{v}$ and $\boldsymbol{\beta}$ are profiled out from Q, estimating $\boldsymbol{\beta}$ should not inflate the standard error of $\hat{\boldsymbol{\gamma}}$, provided that $p_{v,\beta}(Q)$ is sufficient accurate for $\int \exp\{h\}\, d\boldsymbol{\delta}$. However, as suggested by Lee and Nelder (2001a), we still need to account for the dependence between $\hat{\boldsymbol{v}}$ and $\boldsymbol{\gamma}$. The expression of $\partial^2 p_{v,\beta}(Q)/\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T$ for our numerical examples can be found in the appendix. Following the spirit of Lee and Nelder (2001a), if $\boldsymbol{\gamma}$ is estimated from $p_{v,\beta}(h)$, we can estimate the standard errors of $\boldsymbol{\gamma}$ from the negative of the inverse of the Hessian matrix $\partial^2 p_{v,\beta}(h)/\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T$. In the case when $\boldsymbol{\gamma}$ is estimated from a second-order approximation, we still estimate the standard errors using the Hessian of $p_{v,\beta}(\cdot)$, due to the complexity of the higher order derivatives in $r(\boldsymbol{v}, \boldsymbol{\theta})$.

For the ML estimator, a common practice is to use the Fisher information matrix to estimate standard errors. Regarding the standard errors of $\hat{\boldsymbol{\beta}}$, we can consider to get them from

$$-\left(\frac{\partial^2 p_v(h)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)^{-1}.$$

However, such a standard error for $\hat{\boldsymbol{\beta}}$ assumes that $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are orthogonal, since, in $p_v(h)$, only $\boldsymbol{v}$ is profiled out from h. Hence, the uncertainty of estimating $\boldsymbol{\gamma}$ is not accounted for. Consequently, we expect the standard errors for $\hat{\boldsymbol{\beta}}$ to be underestimated if $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are not orthogonal.

Hereafter, the approach described in this subsection will be referred to as Hessian of p, where p stands for various profile log-likelihood.

4.2. Sandwich estimator

The standard error in Section 4.1 essentially assumes that $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are orthogonal, which holds in Gaussian models. However, it is not so in general models. The other alternatives aim to estimate the standard errors without explicitly yielding a covariance matrix of $\left(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T\right)^T$.

In principle, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are the solutions of the system of equations

$$\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \frac{\partial \int \exp\{h\} d\boldsymbol{v}}{\partial \boldsymbol{\beta}} \\ \frac{\partial \int \exp\{h\} d\boldsymbol{\delta}}{\partial \boldsymbol{\gamma}} \end{bmatrix}.$$

By the standard Taylor expansion, the variance of $\hat{\boldsymbol{\theta}}$ can be approximated by

$$\boldsymbol{E}^{-1} \text{var} \begin{bmatrix} \frac{\partial \int \exp\{h\} d\boldsymbol{v}}{\partial \boldsymbol{\beta}} \\ \frac{\partial \int \exp\{h\} d\boldsymbol{\delta}}{\partial \boldsymbol{\gamma}} \end{bmatrix} \left(\boldsymbol{E}^{-1}\right)^T,$$

where

$$\boldsymbol{E} = E \begin{bmatrix} \frac{\partial^2 \int \exp\{h\} d\boldsymbol{v}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 \int \exp\{h\} d\boldsymbol{\delta}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T} \\ \frac{\partial^2 \int \exp\{h\} d\boldsymbol{v}}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 \int \exp\{h\} d\boldsymbol{\delta}}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} \end{bmatrix},$$

and $\partial^2 \int \exp\{h\} d\boldsymbol{\delta} / \partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}^T = \mathbf{0}$ since $\boldsymbol{\beta}$ has been integrated out.

When the approximations $p_v(h)$ and $p_{v,\beta}()$ are used, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are the solutions of the approximated system of equations

$$\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \frac{\partial p_v(h)}{\partial \boldsymbol{\beta}} \\ \frac{\partial p_{v,\beta}()}{\partial \boldsymbol{\gamma}} \end{bmatrix},$$

where $p_{v,\beta}()$ is either $p_{v,\beta}(h)$ or $p_{v,\beta}(Q)$. If $p_{v,\beta}()$ approximates $\int \exp\{h\} d\boldsymbol{\delta}$ well, the variance can be approximated by

$$\begin{bmatrix} \frac{\partial^2 p_v(h)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 p_v(h)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T} \\ \mathbf{0} & \frac{\partial^2 p_{v,\beta}()}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} \end{bmatrix}^{-1} \text{var} \begin{bmatrix} \frac{\partial p_v(h)}{\partial \boldsymbol{\beta}} \\ \frac{\partial p_{v,\beta}()}{\partial \boldsymbol{\gamma}} \end{bmatrix} \left(\begin{bmatrix} \frac{\partial^2 p_v(h)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 p_v(h)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T} \\ \mathbf{0} & \frac{\partial^2 p_{v,\beta}()}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} \end{bmatrix}^T \right)^{-1}. \tag{16}$$

This sandwich estimator will be referred to as Sandwich0 hereafter. On the other hand, we can also compute $\partial^2 p_{v,\beta}() / \partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T$, despite that it is an approximation of the zero term (Hessian of $\int \exp\{h\} d\boldsymbol{\delta}$). In such a case, the variance of $\hat{\boldsymbol{\theta}}$ can then be approximated by

$$\begin{bmatrix} \frac{\partial^2 p_v(h)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 p_v(h)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T} \\ \frac{\partial^2 p_{v,\beta}()}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T} & \frac{\partial^2 p_{v,\beta}()}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} \end{bmatrix}^{-1} \text{var} \begin{bmatrix} \frac{\partial p_v(h)}{\partial \boldsymbol{\beta}} \\ \frac{\partial p_{v,\beta}()}{\partial \boldsymbol{\gamma}} \end{bmatrix} \left(\begin{bmatrix} \frac{\partial^2 p_v(h)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 p_v(h)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T} \\ \frac{\partial^2 p_{v,\beta}()}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T} & \frac{\partial^2 p_{v,\beta}()}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} \end{bmatrix}^T \right)^{-1}. \tag{17}$$

This sandwich estimator will be referred to as Sandwich1 hereafter. Even though $s_{v,\beta}()$ is used for estimation, we still use the Hessian of $p_{v,\beta}()$, due to the complexity of the higher-order derivatives in $r(\boldsymbol{v}, \boldsymbol{\theta})$. At the cost of computational speed, the Hessian of $s_{v,\beta}$ can be computed by numerical differentiation if the analytical Hessian is difficult to obtain.

A common approach to estimate the variance component in (16) or (17) is to use the sample counterpart such as $\sum_j \frac{\partial \ell_j}{\partial \boldsymbol{\theta}} \frac{\partial \ell_j}{\partial \boldsymbol{\theta}^T}$. However, the sample counterpart is not feasible, since $\boldsymbol{\beta}$ appears in every cluster. Instead, we use bootstrap to estimate the variance. Let B be the bootstrap number of replications. For each b in $1, \cdots, B$, we generate $\boldsymbol{v}$ from $f\left(\boldsymbol{v}; \hat{\boldsymbol{\theta}}\right)$ and $\boldsymbol{y}$ from $f\left(\boldsymbol{y}|\boldsymbol{v}; \hat{\boldsymbol{\theta}}\right)$, and compute $\partial p_v(h) / \partial \boldsymbol{\beta}$ and $\partial p_{v,\beta}() / \partial \boldsymbol{\gamma}$ using the bootstrap sample, where $\boldsymbol{\theta}$ is fixed to $\hat{\boldsymbol{\theta}}$. The variance of the bootstrap gradients is then used as an estimate of the variance component in (16) or (17). This parametric bootstrap procedure is similar to that in Flores-Agreda and Cantoni (2019), but in a different context.

Table 2
Percentage of convergence for point estimation.

Dist.	Model	J with $K=5$				J with $K=10$			
		30	60	90	120	30	60	90	120
Poisson	GLMM-I	99.75	99.81	99.87	99.88	99.77	99.86	99.91	99.94
	GLMM-S	97.60	97.08	97.20	97.07	97.09	97.01	96.73	96.68
	HGLM-I	99.52	99.68	99.76	99.76	99.62	99.79	99.82	99.88
Bernoulli	GLMM-I	99.30	99.98	100	100	100	100	100	100
	GLMM-S	92.44	99.13	99.89	99.99	99.78	100	100	100
	HGLM-I	89.55	97.07	98.99	99.63	98.57	99.83	99.99	100

Note: Dist. = Distribution. GLMM-I = GLMM with random intercept (12). GLMM-S = GLMM with random intercept and random slope (18). HGLM-I = HGLM with random intercept (12) and (19).

Table 3
Percentage of valid standard errors from the Hessian of h in different models.

Dist.	Model	J with $K=5$				J with $K=10$			
		30	60	90	120	30	60	90	120
Gamma	GLMM-I	100	100	100	100	100	100	100	100
	GLMM-S	72.19	78.77	81.94	87.05	60.44	59.21	57.27	59.18
	HGLM-I	100	100	100	100	100	100	100	100
Poisson	GLMM-I	100	100	100	100	100	100	100	100
	GLMM-S	97.89	100	100	100	100	100	100	100
	HGLM-I	98.67	100	100	100	100	100	100	100
Bernoulli	GLMM-I	53.82	48.27	47.97	44.88	93.08	98.59	99.60	100
	GLMM-S	16.57	11.81	8.85	6.78	66.29	77.47	85.40	89.77
	HGLM-I	31.94	28.83	27.05	23.93	71.92	84.70	92.72	94.75

Note: Dist. = Distribution. GLMM-I = GLMM with random intercept (12). GLMM-S = GLMM with random intercept and random slope (18). HGLM-I = HGLM with random intercept (12) and (19).

5. Numerical illustration

5.1. Hierarchical generalized linear model

We first revisit the models considered in Section 2.4. We also include the random slope models with the linear predictor

$$\beta_0 + v_j + \left(\beta_1 + u_j\right) x_{1jk} + \beta_2 x_{2jk}, \quad j = 1, ..., J,\ k = 1, ..., K, \tag{18}$$

where u_j is the random effect for the slope. We generate u_j from $N(0, 1.5)$, which is independent of v_j. Besides, we consider an HGLM with the same mean model as in (12), and the dispersion model is given by

$$\log\left(\sigma_j\right) = \gamma_0 + \gamma_1 G_{\sigma,j}, \quad j = 1, ..., J. \tag{19}$$

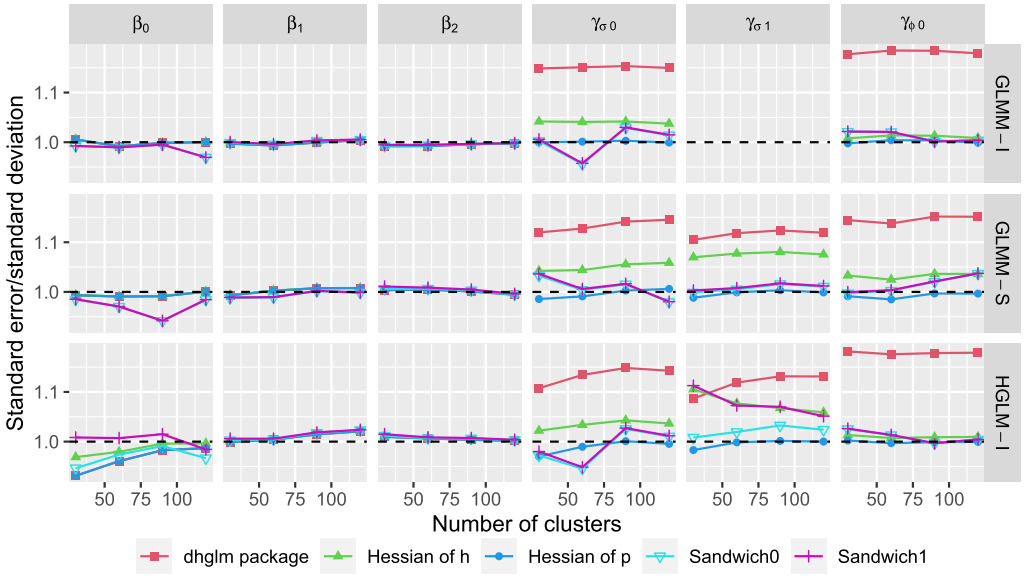
We let $\gamma_0 = \log(1.5)$, $\gamma_1 = 0.5$, and $G_{\sigma,j} \sim N(0, 0.5)$, which is independent of x_{1jk} and x_{2jk}. For the sandwich estimator, the number of bootstrap replication is $1,000$. The objective functions to be optimized for estimation are the same as those in Section 2.4 (tabulated in Table 1). We only considered $K = 10$ in Section 2.4. Here we will also consider $K = 5$.

Nonconvergence is encountered when fitting these models. It is seen from Table 2 that the algorithm converged most of the time in terms of point estimation for the Poisson models and the Bernoulli models. Nonconvergence was not encountered for the Gaussian models and the Gamma models. Hence, they are not tabulated here. Furthermore, it is possible that the covariances ($\mathrm{var}\left(\hat{\boldsymbol{\beta}}\right)$, $\mathrm{var}\left(\hat{\boldsymbol{\gamma}}\right)$, or $\mathrm{var}\left(\hat{\boldsymbol{\theta}}\right)$) estimated by some methods are not positive definite. The corresponding standard errors are considered invalid. Among the methods that we considered, using the Hessian of h can yield a large proportion of invalid standard errors (Table 3). Especially when the Gamma model or the Bernoulli model has a random slope, the percentage can be quite high.

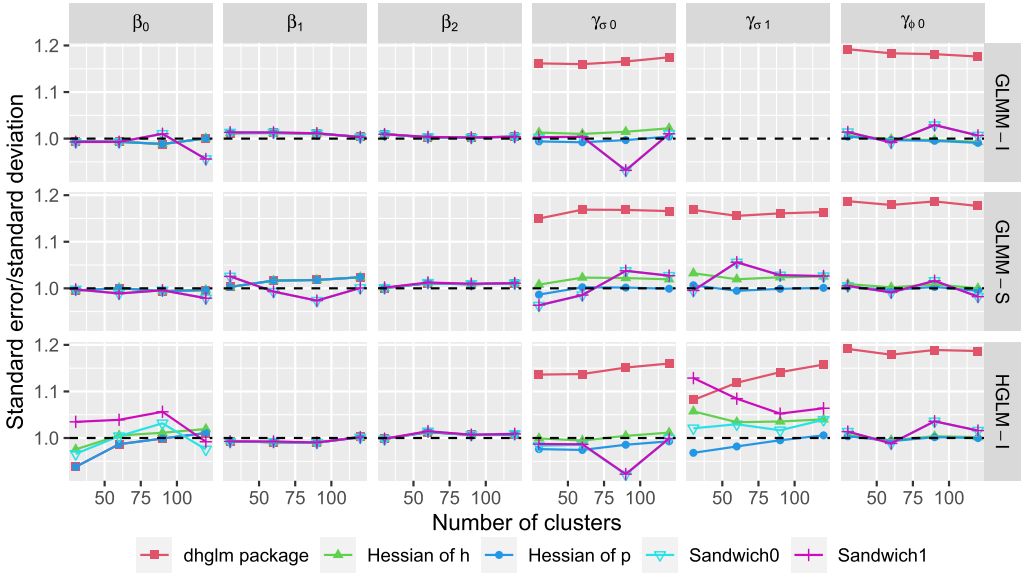
To evaluate the accuracy of the estimated standard errors, we divide the mean of the estimated standard errors by the sample standard deviation, as in Section 2.4. Fig. 2 illustrates the standard error estimates for the Gaussian models. Similar conclusions to those in Fig. 1 can be drawn. While the method used by the dhglm package overestimates the standard errors of the dispersion parameters, the other methods are generally more accurate.

For the gamma models, we exclude the standard errors using Hessian of h from the random slope models, due to their potentially high inadmissible rate, as shown in Table 3. Fig. 3 shows that the dhglm package can still yield largely biased standard errors for the dispersion parameters. Using the Hessian of $p_v(h)$ for the mean parameters and the Hessian of $p_{v,\beta}(h)$ generally yield accurate standard error estimates, so does the sandwich estimators.

For the Poisson models, it is seen from Fig. 4 that the estimated standard errors from the dhglm package and the Hessian of h can be more biased than the other methods for the dispersion parameters. Using the Hessian of $p_v(h)$ generally yields



(a) $K = 5$

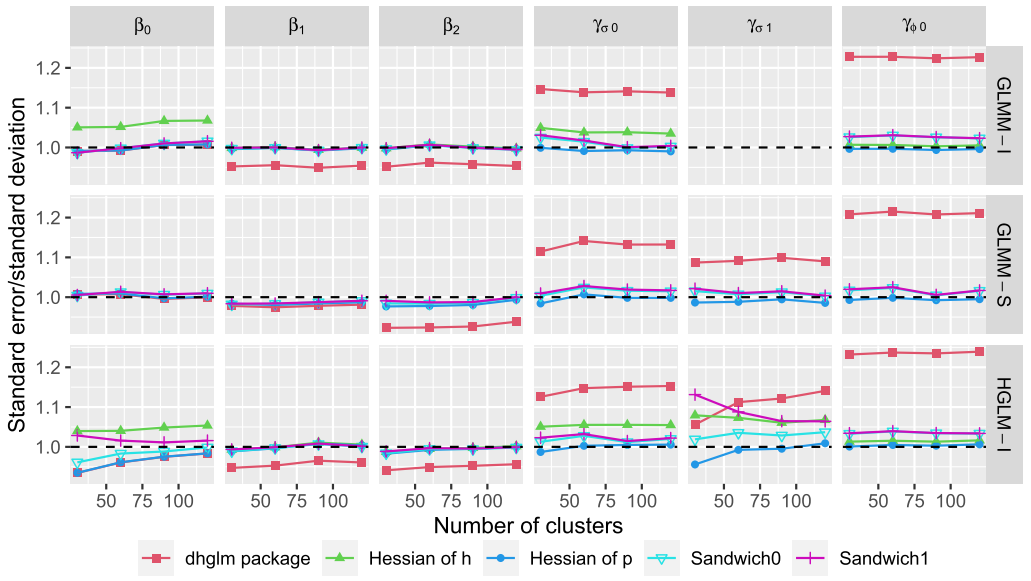


(b) $K = 10$

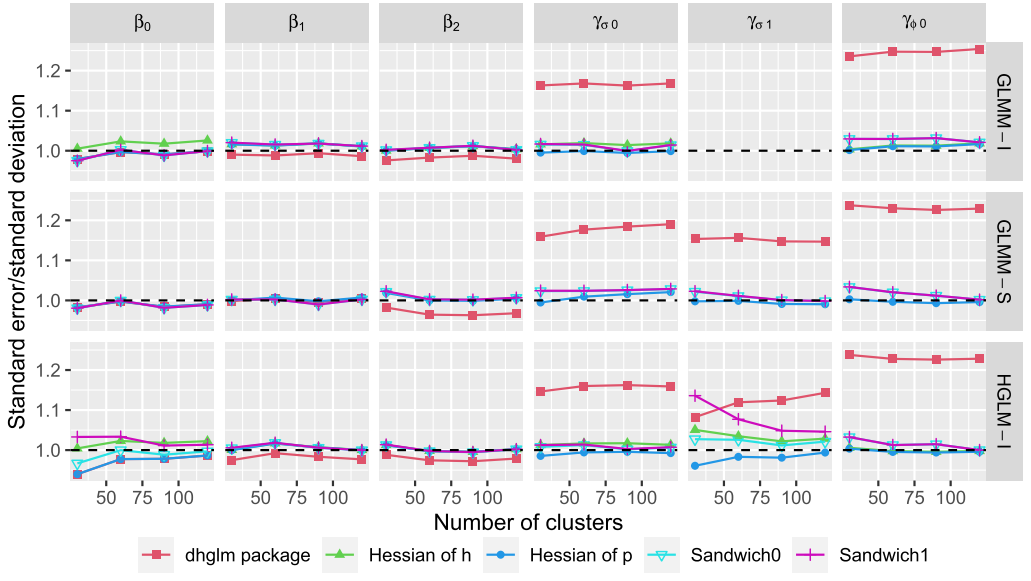
Fig. 2. Mean of the estimated standard errors divided by the sample standard deviation in Gaussian models.

accurate standard errors for the mean parameters. Using the Hessian of $p_{v,\beta}(Q)$ often yields less biased standard errors than using the Hessian of h. When the mean is divided by the standard deviation, the sandwich estimators can yield a large positive ratio for the mean parameters, whereas they tend to be accurate for the dispersion parameters. However, Fig. 5 suggests that the large ratio occurs often when the standard deviation is small, and the sandwich estimators are generally still accurate.

Regarding the Bernoulli models, we exclude the standard errors using the Hessian of h due to their potentially high inadmissible rate and large bias, as seen in Table 3. Results not shown here show that they are even more biased in the random slope model and in the HGLM. It is seen from Fig. 6 that the dhglm package, the Hessian of $p_v(h)$ and $p_{v,\beta}(Q)$, and Sandwich0 often underestimate the standard errors, whereas Sandwich1 tends to be more accurate under most conditions.



(a) $K = 5$

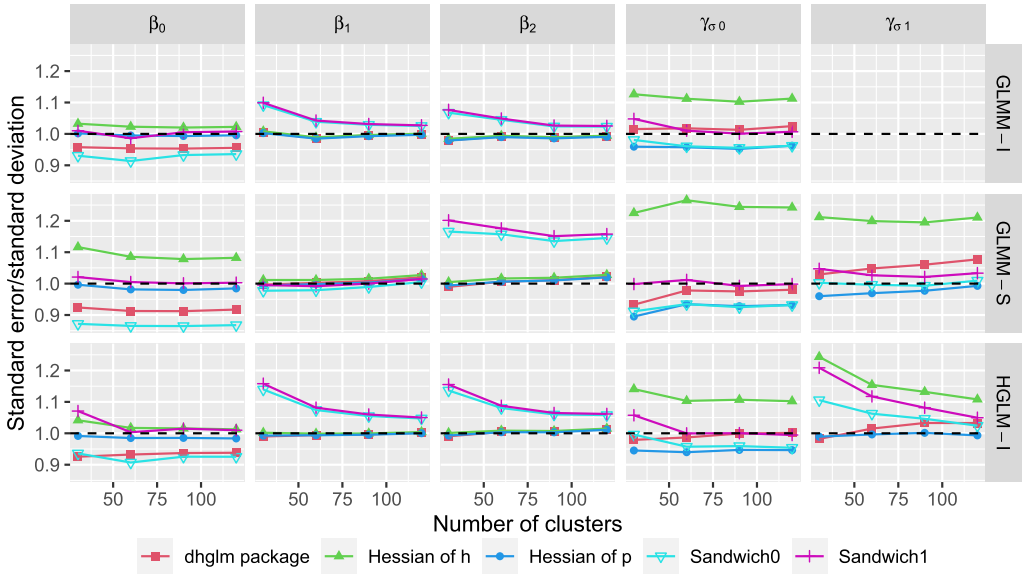


(b) $K = 10$

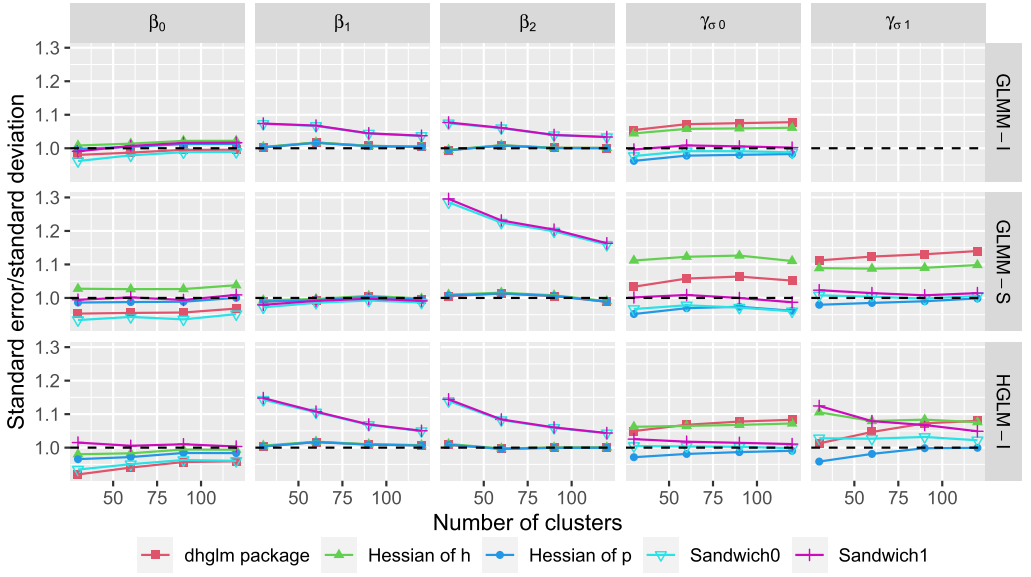
Fig. 3. Mean of the estimated standard errors divided by the sample standard deviation in Gamma models.

5.2. Frailty model

A class of models that are related to the HGLM is the frailty model for survival data. Ha and Lee (2003) and Ha et al. (2001) have developed the h-likelihood approach for the frailty model, which is implemented in the R package frailtyHL (Ha et al., 2018). However, they only investigated the standard errors of the mean parameters, not the frailty parameter. In this part, we conduct a simulation study to investigate the standard errors for the frailty model using the h-likelihood approach. The conditional hazard function of the frailty model is $\lambda_{jk}(t) = \lambda_0(t)\exp\{\eta_{jk}\}$, where the linear predictor η_{jk} remains the same as equation (12) except that β_0 is fixed to zero for identification. In other words, the frailty follows a log-normal distribution. The uncensored failure time t_{jk} is generated from an exponential-parametric frailty model with the baseline hazard set to 1.0. The censoring time c_{jk} is generated from an exponential distribution with an expected value of 4. The observed responses are then $y_{jk} = \min(t_{jk}, c_{jk})$ and $I(t_{jk} \le c_{jk})$, where $I()$ is the indicator function. The number



(a) $K = 5$



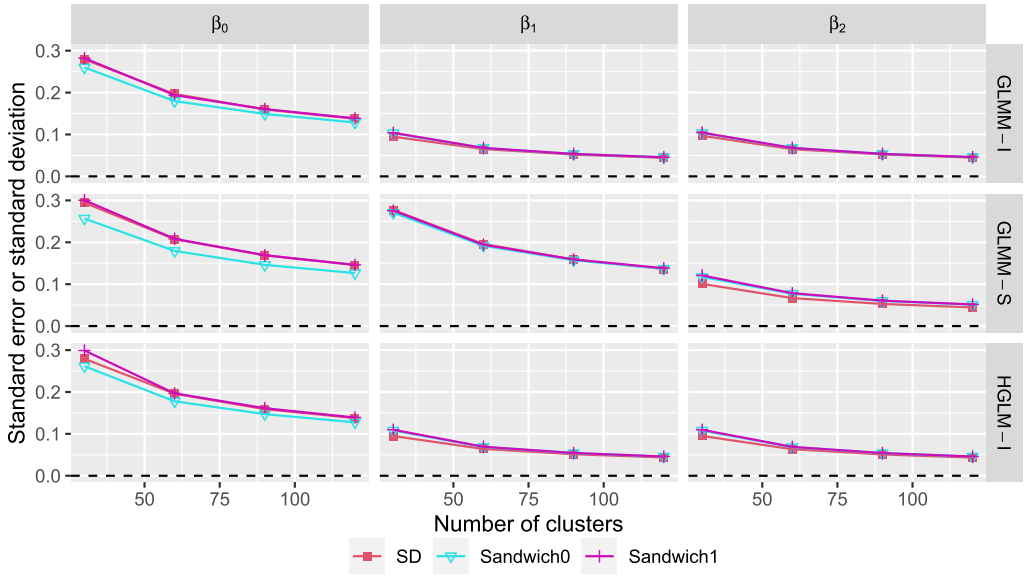
(b) $K = 10$

Fig. 4. Mean of the estimated standard errors divided by the sample standard deviation in Poisson models.

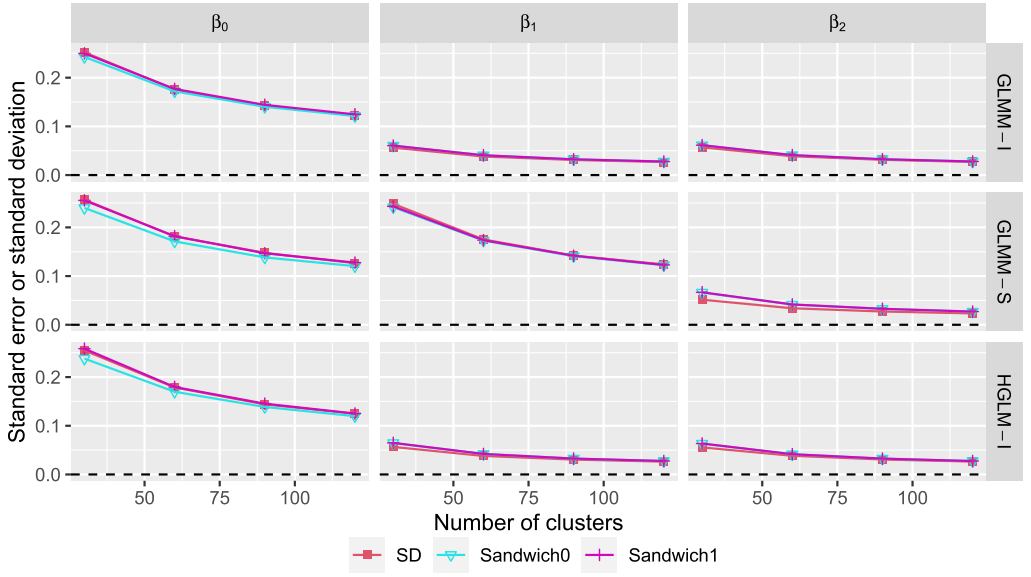
replications remains 1000 for standard errors and 10000 for sample standard deviation. The parametric bootstrap procedure proposed in Massonnet et al. (2006) is used to generate the uncensored failure time and the censoring time.

First, we consider a parametric frailty model. We assume that the uncensored survival time follows an exponential distribution, but the censoring time is left unspecified. As shown in Ha and Lee (2003), the exponential frailty model can be fitted using Poisson HGLM with the response variable $I\left(t_{jk} \leq c_{jk}\right)$ and the offset $\log y_{jk}$. Hence, the standard errors for the HGLMs discussed above can be easily computed. In this simulation, we still let $J = \{30, 60, 90, 120\}$, and keep $K = \{5, 10\}$. It can be observed from Fig. 7 that using the Hessian of $p_v(h)$ and $p_{v,\beta}(Q)$, as well as the sandwich estimators, are generally accurate under most conditions. Whereas the dhglm package and using the Hessian of h are accurate for mean parameters, they can be biased for the dispersion parameter.

Second, we assume that the baseline hazard is unknown and fit a semiparametric frailty model, where the baseline hazard function is profiled out by the nonparametric maximum hierarchical likelihood estimator, as described in Ha et al. (2001), Ha and Lee (2003), and Ha et al. (2017). Semiparametric estimation is performed in the R package frailtyHL.



(a) $K = 5$



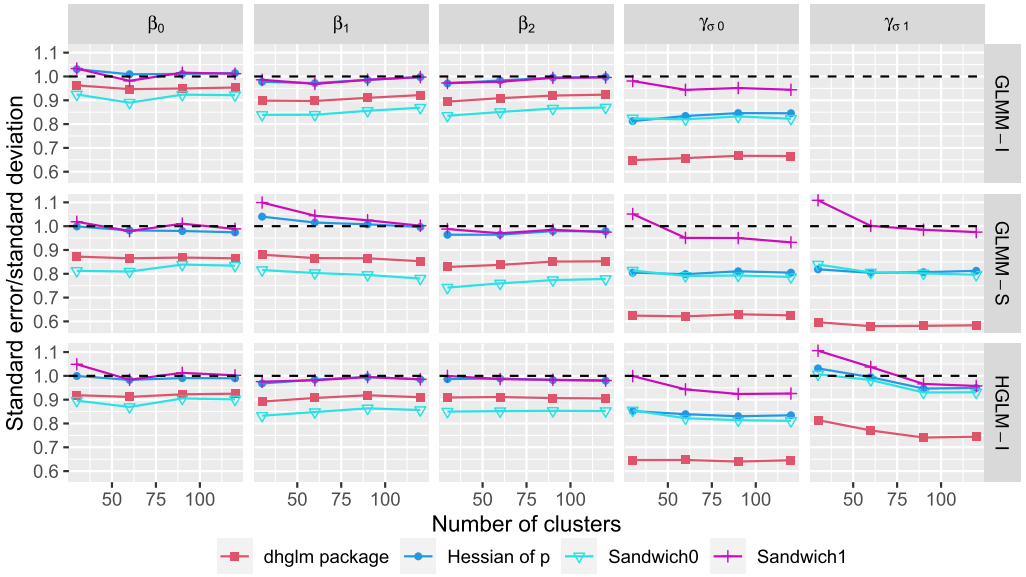
(b) $K = 10$

Fig. 5. Mean of the estimated standard errors and the sample standard deviation in Poisson models.

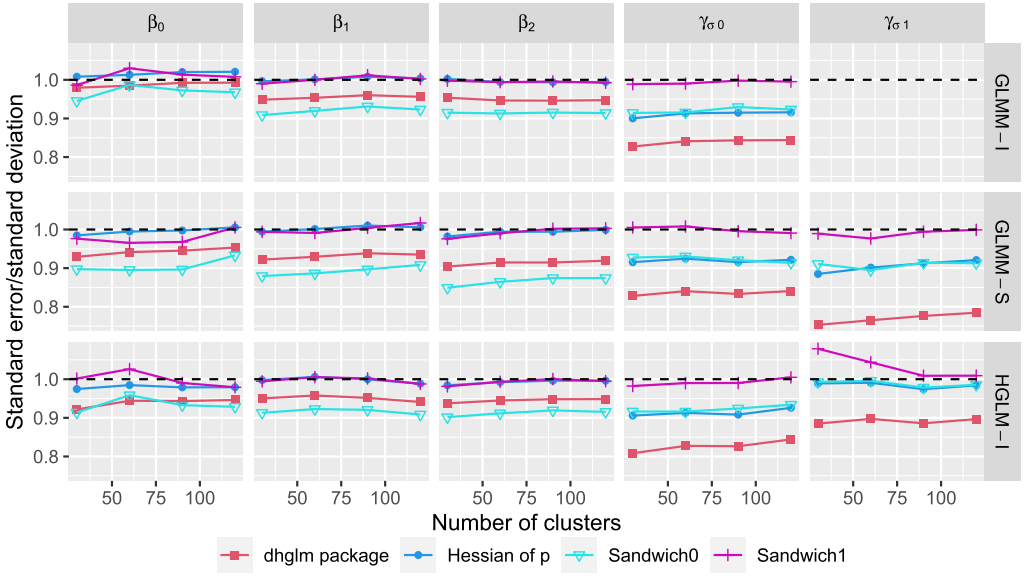
Due to computational cost of the frailty model, we only vary $J = \{30, 40, 50, 60\}$ and the bootstrap replication is set to 200, while K remains $\{5, 10\}$. In frailtyHL, the standard errors are based on Ha and Lee (2003) and Ha et al. (2001). They proposed to estimate the standard errors of $(\boldsymbol{v}, \boldsymbol{\beta})$ and the frailty parameter from the Hessian of h_p and the Hessian of $p_{v,\beta}(h_p)$, respectively, where h_p is the profiled h-likelihood that profiles out the baseline hazard. Hence, we only consider the standard errors used by the frailtyHL packages and the sandwich estimators. It is seen from Fig. 8 that Sandwich1 tends to be more accurate than Sandwich0 and the negative inverse Hessian of the profile h-likelihood used by the package.

6. Conclusion

In this study, we investigate the standard error estimators used by the h-likelihood approach. In HGLM, the standard errors of the mean parameters are often based on the inverse of equation (8). We show that they can be viewed as the exact derivative of the Laplace approximated gradient, taking the dependence between the estimated random effects and



(a) $K = 5$



(b) $K = 10$

Fig. 6. Mean of the estimated standard errors divided by the sample standard deviation in Bernoulli models.

the parameter values into account. Our numerical illustration indicates that they are seemingly accurate, but not in the Bernoulli models.

Among the alternative standard errors, using the inverses of the Hessians of $p_v(h)$ and $p_{v,\beta}(Q)$, as proposed by Lee and Nelder (2001a), generally improve the standard errors of the dhglm package in the investigated models, although they can still yield inaccurate standard errors of the dispersion parameters. Such standard errors assume orthogonality between the mean parameters and the dispersion parameters, which may not hold, except in the normal linear mixed model. The sandwich estimator Sandwich0 often underestimates the standard errors, whereas the sandwich estimator Sandwich1 generally performs well for the discrete response models in our simulation study, showing the importance of accounting for $\partial^2 p_{v,\beta}()/\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^T$, despite the fact that $\int \exp\{h\}\, d\boldsymbol{\delta} = \mathbf{0}$. Our results also show that Sandwich1 based on the Hessian of $p_{v,\beta}(Q)$ works well even though $s_{v,\beta}(Q)$ is used for estimation.

One limitation of the sandwich estimator is that it is computationally more intensive than the other alternatives, due to the bootstrap approximated variance of the gradient. For the sandwich estimator to be accurate, the variance of the

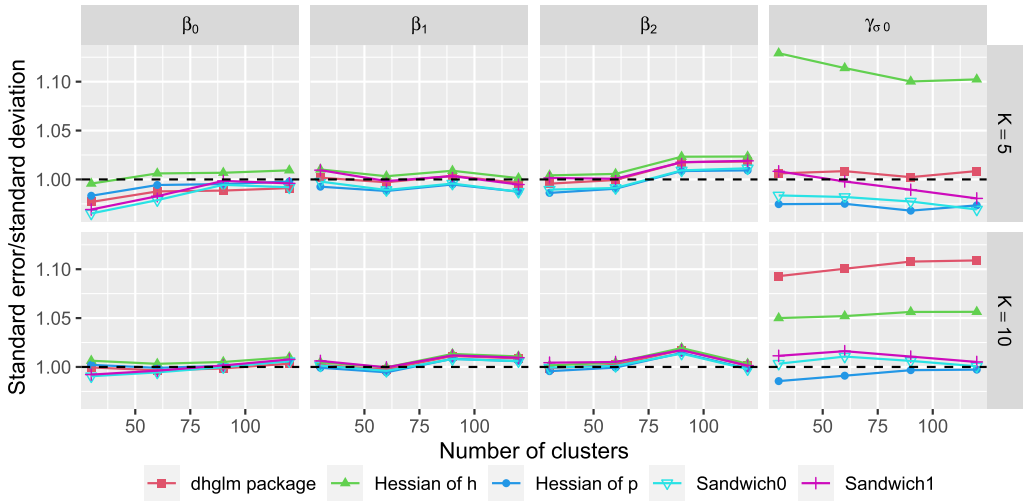


Fig. 7. Mean of the estimated standard errors divided by the sample standard deviation in the exponential frailty model. Note: in this model $\beta_0 = \log \lambda_0$.

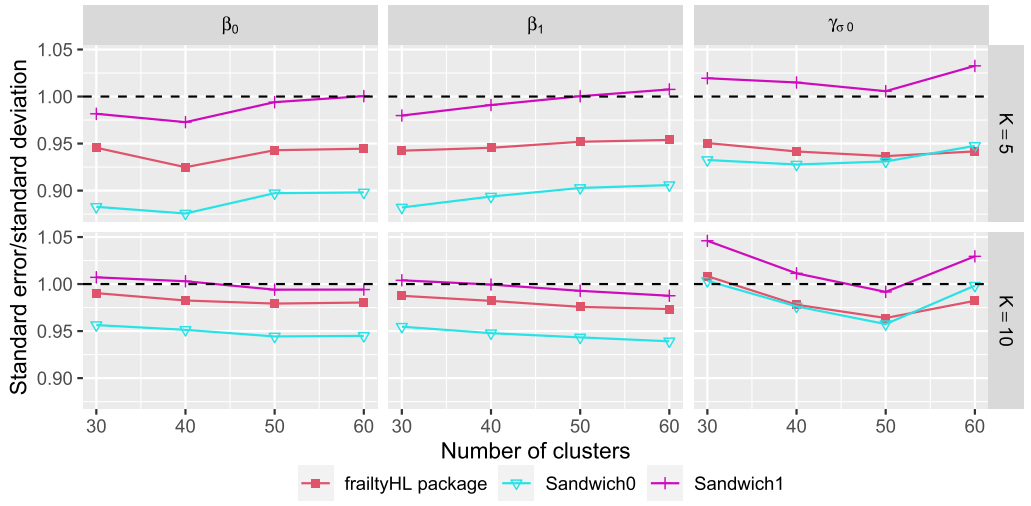


Fig. 8. Mean of the estimated standard errors divided by the sample standard deviation in the semiparametric frailty model.

gradient needs to be well approximated. Another limitation of our study is that we only investigated the performance of standard error estimates in limited models. Lee and Nelder (2006) also introduced random effects in the dispersion models (4), whereas we only considered the fixed effects in them. The random effects considered in the simulation are only normal random effects. In principle, they can belong to a large class of distribution (Lee et al., 2017). We leave these topics as future directions.

Acknowledgements

We are very grateful to the reviewers for their valuable comments and suggestions, which greatly improve the paper.

Funding: Lee was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (grant number 2019R1A2C1002408).

Appendix A. Hessian matrix

In this section, we present the Hessian matrix of $\partial^2 p_{v,\beta}(\mathrm{Q})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$. The notation $\partial \log(\boldsymbol{F})/\partial\theta$ is used to denote a matrix whose (j,k)th entry is $\partial \log\left(F_{jk}\right)/\partial\theta$. Lee and Nelder (2001a) showed how to compute the Hessian matrix $\partial^2 p_{v,\beta}(\mathrm{Q})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T$. The (j,k)th entry of $\partial^2 p_{v,\beta}(\mathrm{Q})/\partial\boldsymbol{\gamma}_\sigma\partial\boldsymbol{\gamma}_\sigma^T$ is

$$\frac{1}{2}\sum_{i=1}^{q}\left(\frac{1}{\sigma_i^2}\frac{\partial d_{1i}}{\partial\gamma_{\sigma,k}}\frac{\partial\sigma_i}{\partial\eta_{\sigma,i}}\right)G_{\sigma,ij}$$
$$+\frac{1}{2}\sum_{i=1}^{q}\left(-\frac{2d_{1i}}{\sigma_i^3}\left(\frac{\partial\sigma_i}{\partial\eta_{\sigma,i}}\right)^2+\frac{1}{\sigma_i^2}\left(\frac{\partial\sigma_i}{\partial\eta_{\sigma,i}}\right)^2+\frac{d_{1i}-\sigma_i}{\sigma_i^2}\frac{\partial^2\sigma_i}{\partial\eta_{\sigma,i}^2}\right)G_{\sigma,ij}G_{\sigma,ik}$$
$$+\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{A}\frac{\partial\log(\boldsymbol{W})-\log(\boldsymbol{\Gamma})}{\partial\gamma_{\sigma,k}}\boldsymbol{A}\frac{\partial\log(\boldsymbol{W})-\log(\boldsymbol{\Gamma})}{\partial\gamma_{\sigma,j}}-\boldsymbol{A}\boldsymbol{W}^{-1}\frac{\partial^2\boldsymbol{W}}{\partial\gamma_{\sigma,j}\partial\gamma_{\sigma,k}}\right\}$$
$$-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{A}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & 2\frac{\partial\log(\boldsymbol{\Sigma})}{\partial\gamma_{\sigma,j}}\frac{\partial\log(\boldsymbol{\Sigma})}{\partial\gamma_{\sigma,k}}-\frac{\partial\log(\boldsymbol{W}_v)}{\partial\gamma_{\sigma,j}}\frac{\partial\log(\boldsymbol{\Sigma})}{\partial\gamma_{\sigma,k}}-\frac{\partial\log(\boldsymbol{W}_v)}{\partial\gamma_{\sigma,k}}\frac{\partial\log(\boldsymbol{\Sigma})}{\partial\gamma_{\sigma,j}}-\boldsymbol{\Sigma}^{-1}\frac{\partial^2\boldsymbol{\Sigma}}{\partial\gamma_{\sigma,j}\partial\gamma_{\sigma,k}}\end{bmatrix}\right\}.$$

The (j,k)th entry of $\partial^2 p_{v,\beta}(Q)/\partial\boldsymbol{\gamma}_\sigma\partial\boldsymbol{\gamma}_\phi^T$ is

$$\frac{1}{2}\sum_{i=1}^{q}\frac{1}{\sigma_i^2}\frac{\partial d_{1i}}{\partial\gamma_{\phi,k}}\frac{\partial\sigma_i}{\partial\eta_{\sigma,j}}G_{\sigma,ij}+\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{A}\begin{bmatrix}\frac{\partial\log(\boldsymbol{W}_v)}{\partial\gamma_{\sigma,j}}\frac{\partial\log(\boldsymbol{\Phi})}{\partial\gamma_{\phi,k}} & \boldsymbol{0}\\ \boldsymbol{0} & \frac{\partial\log(\boldsymbol{W}_v)}{\partial\gamma_{\phi,k}}\frac{\partial\log(\boldsymbol{\Sigma})}{\partial\gamma_{\sigma,j}}\end{bmatrix}\right\}$$
$$-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{A}\frac{\partial\log(\boldsymbol{W})-\log(\boldsymbol{\Gamma})}{\partial\gamma_{\phi,k}}\boldsymbol{A}\frac{\partial\log(\boldsymbol{W})-\log(\boldsymbol{\Gamma})}{\partial\gamma_{\sigma,j}}+\boldsymbol{A}\boldsymbol{W}^{-1}\frac{\partial^2\boldsymbol{W}}{\partial\gamma_{\sigma,j}\partial\gamma_{\phi,k}}\right\}.$$

The (j,k)th entry of $\partial^2 p_{v,\beta}(Q)/\partial\boldsymbol{\gamma}_\phi\partial\boldsymbol{\gamma}_\phi^T$ is

$$\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\phi_i^2}\frac{\partial d_{0i}}{\partial\gamma_{\phi,k}}\frac{\partial\phi_i}{\partial\eta_{\phi,i}}G_{\phi,ij}$$
$$+\frac{1}{2}\sum_{i=1}^{n}\left(-2\frac{d_{0i}}{\phi_i^3}\left(\frac{\partial\phi_i}{\partial\eta_{\phi,i}}\right)^2+\frac{1}{\phi_i^2}\left(\frac{\partial\phi_i}{\partial\eta_{\phi,i}}\right)^2+\frac{d_{0i}-\phi_i}{\phi_i^2}\frac{\partial^2\phi_i}{\partial\eta_{\phi,i}^2}\right)G_{\phi,ij}G_{\phi,ik}$$
$$+\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{A}\frac{\partial\log(\boldsymbol{W})-\log(\boldsymbol{\Gamma})}{\partial\gamma_{\phi,j}}\boldsymbol{A}\frac{\partial\log(\boldsymbol{W})-\log(\boldsymbol{\Gamma})}{\partial\gamma_{\phi,k}}-\boldsymbol{A}\boldsymbol{W}^{-1}\frac{\partial^2\boldsymbol{W}}{\partial\gamma_{\phi,j}\partial\gamma_{\phi,k}}\right\}$$
$$-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{A}\begin{bmatrix}2\frac{\partial\log(\boldsymbol{\Phi})}{\partial\gamma_{\phi,j}}\frac{\partial\log(\boldsymbol{\Phi})}{\partial\gamma_{\phi,k}}-\frac{\partial\log(\boldsymbol{W}_y)}{\partial\gamma_{\phi,j}}\frac{\partial\log(\boldsymbol{\Phi})}{\partial\gamma_{\phi,k}}-\frac{\partial\log(\boldsymbol{W}_y)}{\partial\gamma_{\phi,k}}\frac{\partial\log(\boldsymbol{\Phi})}{\partial\gamma_{\phi,j}}-\boldsymbol{\Phi}^{-1}\frac{\partial^2\boldsymbol{\Phi}}{\partial\gamma_{\phi,j}\partial\gamma_{\phi,k}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\right\}.$$

If $p_\delta(h)$ is used to estimate $\boldsymbol{\gamma}$ in the Gamma models, then

$$\frac{\partial^2 p_\delta(h)}{\partial\gamma_{\phi,j}\partial\gamma_{\phi,k}}=\frac{\partial^2 p_\delta(Q)}{\partial\gamma_{\phi,j}\partial\gamma_{\phi,k}}+\frac{1}{2}\sum_{i=1}^{n}\frac{1+\frac{2}{\phi_i}-\frac{2}{\phi_i^2}\psi_1\left(\frac{1}{\phi_i}\right)}{\phi_i^2}\left(\frac{\partial\phi_i}{\partial\eta_{\phi,i}}\right)^2G_{\phi,ij}G_{\phi,ik}$$
$$+\frac{1}{2}\sum_{i=1}^{n}\frac{\phi_i+2\log(\phi_i)+2\psi\left(\frac{1}{\phi_i}\right)}{\phi_i^2}\left(\frac{\partial^2\phi_i}{\partial\eta_{\phi,i}^2}-\frac{2}{\phi_i}\left(\frac{\partial\phi_i}{\partial\eta_{\phi,i}}\right)^2\right)G_{\phi,ij}G_{\phi,ik},$$

where $\psi()$ is the digamma function and $\psi_1()$ is the trigamma function. The (j,k)th entry of $\partial^2 p_{v,\beta}(Q)/\partial\boldsymbol{\gamma}_\sigma\partial\boldsymbol{\beta}^T$ is

$$\frac{1}{2}\sum_{i=1}^{q}\frac{1}{\sigma_i^2}\frac{\partial d_{1i}}{\partial\beta_k}\frac{\partial\sigma_i}{\partial\eta_{\sigma,i}}G_{\sigma,ij}+\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{A}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \frac{\partial\log(\boldsymbol{W}_v)}{\partial\beta_k}\frac{\partial\log(\boldsymbol{\Sigma})}{\partial\gamma_{\sigma,j}}\end{bmatrix}\right\}$$
$$+\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{A}\frac{\partial\log(\boldsymbol{W})}{\partial\beta_k}\boldsymbol{A}\frac{\partial\log(\boldsymbol{W})-\log(\boldsymbol{\Gamma})}{\partial\gamma_{\sigma,j}}-\boldsymbol{A}\boldsymbol{W}^{-1}\frac{\partial^2\boldsymbol{W}}{\partial\gamma_{\sigma,j}\partial\beta_k}\right\}.$$

The (j,k)th entry of $\partial^2 p_{v,\beta}(Q)/\partial\boldsymbol{\gamma}_\phi\partial\boldsymbol{\beta}^T$ is

$$\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\phi_i^2}\frac{\partial d_{0i}}{\partial\beta_k}\frac{\partial\phi_i}{\partial\eta_{\phi,i}}G_{\phi,ij}+\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{A}\begin{bmatrix}\frac{\partial\log(\boldsymbol{W}_y)}{\partial\beta_k}\frac{\partial\log(\boldsymbol{\Phi})}{\partial\gamma_{\phi,j}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\right\}$$
$$+\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{A}\frac{\partial\log(\boldsymbol{W})}{\partial\beta_k}\boldsymbol{A}\frac{\partial\log(\boldsymbol{W})-\log(\boldsymbol{\Gamma})}{\partial\gamma_{\phi,j}}-\boldsymbol{A}\boldsymbol{W}^{-1}\frac{\partial^2\boldsymbol{W}}{\partial\gamma_{\phi,j}\partial\beta_k}\right\}.$$

Let d_i be either d_{1i} or d_{0i}, and θ be a parameter. Then, to account for the dependence between $\hat{\boldsymbol{v}}$ and θ,

$$\frac{\partial d_i}{\partial \theta} = \frac{\partial d_i}{\partial \boldsymbol{v}^T} \frac{\partial \boldsymbol{v}(\theta)}{\partial \theta}.$$

We also need the Hessian matrix of $p_v(h)$. Recall that $\boldsymbol{H} = \boldsymbol{Z}^T \boldsymbol{W}_y \boldsymbol{\Phi}^{-1} \boldsymbol{Z} + \boldsymbol{\Sigma}^{-1} \boldsymbol{W}_v$. Let $\boldsymbol{K} = \boldsymbol{Z}\boldsymbol{H}^{-1}$ and $\boldsymbol{R} = \boldsymbol{K}\boldsymbol{Z}^T$. For any matrix $\boldsymbol{A}$, we use $\boldsymbol{A}_{i\cdot}$ to denote the ith row of $\boldsymbol{A}$, and $\boldsymbol{A}_{\cdot i}$ the ith column of $\boldsymbol{A}$. Then, for a parameter θ,

$$\begin{aligned}\frac{\partial^2 w_{y,i}}{\partial \beta_j \partial \theta} &= \frac{\partial}{\partial \theta}\left(\frac{\partial w_{y,i}}{\partial \eta_{\mu,i}} \frac{\partial \eta_{\mu,i}}{\partial \beta_j}\right) \\ &= \frac{\partial^2 w_{y,i}}{\partial \eta_{\mu,i}^2} \frac{\partial \eta_{\mu,i}}{\partial \beta_j} \frac{\partial \eta_{\mu,i}}{\partial \theta} + \frac{\partial w_{y,i}}{\partial \eta_{\mu,i}} \boldsymbol{Z}_{i\cdot} \frac{\partial}{\partial \theta}\left(\frac{\partial \boldsymbol{v}(\theta)}{\partial \beta_j}\right),\end{aligned}$$

and

$$\frac{\partial \boldsymbol{\eta}_\mu}{\partial \theta} = \boldsymbol{X}^T \frac{\partial \boldsymbol{\beta}}{\partial \theta} + \boldsymbol{Z} \frac{\partial \boldsymbol{v}(\theta)}{\partial \theta}.$$

The (j,k)th entry of $\partial^2 p_v(h)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T$ is

$$\begin{aligned}&- \boldsymbol{X}_{\cdot j}^T \boldsymbol{W}_y \boldsymbol{\Phi}^{-1} \frac{\partial \boldsymbol{\eta}_\mu}{\partial \beta_k} + \boldsymbol{X}_{\cdot j}^T \boldsymbol{\Phi}^{-1} \mathrm{diag}\left\{\frac{\partial w_{0,i} \frac{\partial \eta_{\mu,i}}{\partial \mu_i}}{\partial \eta_{\mu,i}} \frac{\partial \eta_{\mu,i}}{\partial \beta_k}\right\} (\boldsymbol{y} - \boldsymbol{\mu}) \\ &- \frac{1}{2}\mathrm{tr}\left\{\frac{\partial \boldsymbol{R}}{\partial \beta_k} \frac{\partial \boldsymbol{W}_y}{\partial \beta_j} \boldsymbol{\Phi}^{-1}\right\} - \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{R} \frac{\partial^2 \boldsymbol{W}_y}{\partial \beta_j \partial \beta_k} \boldsymbol{\Phi}^{-1}\right\} - \frac{1}{2} \frac{\partial \mathrm{tr}\left\{\boldsymbol{H}^{-1} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{W}_v}{\partial \beta_j}\right\}}{\partial \beta_k}.\end{aligned}$$

The (j,k)th entry of $\partial^2 p_v(h)/\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}_\sigma^T$ is

$$\begin{aligned}&- \boldsymbol{X}_{\cdot j}^T \boldsymbol{W}_y \boldsymbol{\Phi}^{-1} \frac{\partial \boldsymbol{\eta}_\mu}{\partial \gamma_{\sigma,k}} + \boldsymbol{X}_{\cdot j}^T \boldsymbol{\Phi}^{-1} \mathrm{diag}\left\{\frac{\partial w_{0,i} \frac{\partial \eta_{\mu,i}}{\partial \mu_i}}{\partial \eta_{\mu,i}} \frac{\partial \eta_{\mu,i}}{\partial \gamma_{\sigma,k}}\right\} (\boldsymbol{y} - \boldsymbol{\mu}) \\ &- \frac{1}{2}\mathrm{tr}\left\{\frac{\partial \boldsymbol{R}}{\partial \gamma_{\sigma,k}} \frac{\partial \boldsymbol{W}_y}{\partial \beta_j} \boldsymbol{\Phi}^{-1}\right\} - \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{R} \frac{\partial^2 \boldsymbol{W}_y}{\partial \beta_j \partial \gamma_{\sigma,k}} \boldsymbol{\Phi}^{-1}\right\} - \frac{1}{2} \frac{\partial \mathrm{tr}\left\{\boldsymbol{H}^{-1} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{W}_v}{\partial \beta_j}\right\}}{\partial \gamma_{\sigma,k}}.\end{aligned}$$

The (j,k)th entry of $\partial^2 p_v(h)/\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}_\phi^T$ is

$$\begin{aligned}&- \boldsymbol{X}_{\cdot j}^T \boldsymbol{\Phi}^{-2} \mathrm{diag}\left\{\frac{\partial \phi_i}{\partial \eta_{\phi,i}} G_{\phi,i,k}\right\} \boldsymbol{W}_y \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}^T} (\boldsymbol{y} - \boldsymbol{\mu}) \\ &+ \boldsymbol{X}_{\cdot j}^T \boldsymbol{\Phi}^{-1} \mathrm{diag}\left\{\frac{\partial w_{0,i} \frac{\partial \eta_{\mu,i}}{\partial \mu_i}}{\partial \eta_{\mu,i}} \frac{\partial \eta_{\mu,i}}{\partial \gamma_{\phi,k}}\right\} (\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{X}_{\cdot j}^T \boldsymbol{W}_y \boldsymbol{\Phi}^{-1} \frac{\partial \boldsymbol{\eta}_\mu}{\partial \gamma_{\phi,k}} \\ &- \frac{1}{2}\mathrm{tr}\left\{\frac{\partial \boldsymbol{R}}{\partial \gamma_{\phi,k}} \frac{\partial \boldsymbol{W}_y}{\partial \beta_j} \boldsymbol{\Phi}^{-1}\right\} - \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{R} \frac{\partial^2 \boldsymbol{W}_y}{\partial \beta_j \partial \gamma_{\phi,k}} \boldsymbol{\Phi}^{-1}\right\} - \frac{1}{2} \frac{\partial \mathrm{tr}\left\{\boldsymbol{H}^{-1} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{W}_v}{\partial \beta_j}\right\}}{\partial \gamma_{\phi,k}}.\end{aligned}$$

References

Anderson, T.W., Rubin, H., 1956. Statistical inference in factor analysis. In: Proceedings of Third Berkeley Symposium. University of California Press, Berkeley, pp. 111–150.

Bayarri, M.J., De Groot, M.H., Kanane, J.B., 1988. What is the likelihood function. In: Gupta, S.S., Berger, J.O. (Eds.), Statistical Desision Theory and Related Topics IV. Springer, New York, pp. 3–16.

Bjørnstad, J.F., 1996. On the generalization of the likelihood function and likelihood principle. J. Am. Stat. Assoc. 91, 791–806.

Cox, D.R., Reid, N., 1987. Parameter orthogonality and approximate conditional inference. J. R. Stat. Soc., Ser. B, Methodol. 49, 1–39.

Evans, M., Swartz, T., 1995. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. Stat. Sci. 10, 254–272.

Flores-Agreda, D., Cantoni, E., 2019. Bootstrap estimation of uncertainty in prediction for generalized linear mixed models. Comput. Stat. Data Anal. 130, 1–17.

Ha, I.D., Jeong, J.H., Lee, Y., 2017. Statistical Modelling of Survival Data with Random Effects: H-Likelihood Approach. Springer, Singapore.

Ha, I.D., Lee, Y., 2003. Estimating frailty models via Poisson hierarchical generalized models. J. Comput. Graph. Stat. 12, 663–681.

Ha, I.D., Lee, Y., Song, J.K., 2001. Hierarchical likelihood approach for frailty models. Biometrika 88, 233–243.

Ha, I.D., Noh, M., Kim, J., Lee, Y., 2018. frailtyHL: frailty models via hierarchical likelihood. Version 2.2. https://cran.r-project.org/web/packages/frailtyHL/index.html.

Haberman, S.J., 1977. Maximum likelihood estimates in exponential response models. Ann. Stat. 5, 815–841.

Haberman, S.J., 2004. Joint and conditional maximum likelihood estimation for Rasch model for binary responses. ETS Research report RR-04-20. Princeton, NJ:ETS.

Jin, S., Lee, Y., 2021. Advance online publication in a review of h-likelihood and hierarchical generalized linear model. WIREs: Comput. Stat. 13, e1527.
Jin, S., Noh, M., Lee, Y., 2018. H-likelihood approach to factor analysis for ordinal data. Struct. Equ. Model. 25, 530–540. https://doi.org/10.1080/10705511.2017.1403287.
Lee, Y., 2002. Robust variance estimator for fixed-effect estimates with hierarchical-likelihood. Stat. Comput. 12, 201–207.
Lee, Y., Bjørnstad, J.F., 2013. Extended likelihood approach to large-scale multiple testing. J. Roy. Stat. Soc., Ser. B, Stat. Methodol. 75, 553–575.
Lee, Y., Kim, G., 2016. H-likelihood predictive intervals for unobservables. Int. Stat. Rev. 84, 487–505.
Lee, Y., Molas, M., Noh, M., 2018. mdhglm: multivariate double hierarchical generalized linear models. R package version 1.8. https://CRAN.R-project.org/package=mdhglm.
Lee, Y., Nelder, J.A., 1996. Hierarchical generalized linear models. J. R. Stat. Soc., Ser. B, Methodol. 58, 619–678.
Lee, Y., Nelder, J.A., 2001a. Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. Biometrika 88, 987–1006.
Lee, Y., Nelder, J.A., 2001b. Modelling and anlysing correlated non-normal data. Stat. Model. 1, 3–16.
Lee, Y., Nelder, J.A., 2005. Likelihood for random-effect models. SORT 29, 141–164.
Lee, Y., Nelder, J.A., 2006. Double hierarchical generalized linear models (with discussion). J. R. Stat. Soc., Ser. C, Appl. Stat. 55, 139–185.
Lee, Y., Nelder, J.A., Pawitan, Y., 2006. Generalized Linear Models with Random Effects. Unified Analysis via H-Likelihood. Chapman and Hall, Boca Raton, FL.
Lee, Y., Nelder, J.A., Pawitan, Y., 2017. Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood, second ed. Chapman and Hall, Boca Raton, FL.
Lee, Y., Noh, M., 2018. dhglm: double hierarchical generalized linear models. R package version 2.0. https://CRAN.R-project.org/package=dhglm.
Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data, 2nd ed. Wiley, Hoboken, NJ.
Massonnet, G., Burzykowski, T., Janssen, P., 2006. Resampling plans for frailty models. Commun. Stat., Simul. Comput. 35, 497–514.
Meng, X., 2009. Decoding the h-likelihood. Stat. Sci. 24, 280–293.
Noh, M., Lee, Y., 2007. REML estimation for binary data in GLMMs. J. Multivar. Anal. 98, 896–915. https://doi.org/10.1016/j.jmva.2006.11.009.
Rasch, G., 1960. Probabilistic Models for Some Intelligence and Attainment Tests. Danish Institute for Educational Research, Copenhagen.
Shun, Z., McCullagh, P., 1995. Laplace approximation of high dimensional integrals. J. R. Stat. Soc., Ser. B, Methodol. 57, 749–760.
Strawderman, R.L., 2000. Higher-order asymptotic approximation: Laplace, saddlepoint, and related methods. J. Am. Stat. Assoc. 95, 1358–1364.
Wu, J., Bentler, P.M., 2012. Application of H-likelihood to factor analysis models with binary response data. J. Multivar. Anal. 106, 72–79.
Yun, S., Lee, Y., 2004. Comparison of hierarchical and marginal likelihood estimators for binary outcomes. Comput. Stat. Data Anal. 45, 639–650.