ORIGINAL PAPER

Mixed-effects location-scale model based on generalized hyperbolic distribution

Yuki Fujinaga[1] · Hiroki Masuda[2]

Abstract

Motivated by better modeling of *intra-individual* variability in longitudinal data, we propose a class of location-scale mixed-effects models, in which the data of each individual is modeled by a parameter-varying generalized hyperbolic distribution. We first study the local maximum-likelihood asymptotics and reveal the instability in the numerical optimization of the log-likelihood. Then, we construct an asymptotically efficient estimator based on the Newton–Raphson method based on the original log-likelihood function with the initial estimator being naive least-squares-type. Numerical experiments are conducted to show that the proposed one-step estimator is not only theoretically efficient but also numerically much more stable and much less time-consuming compared with the maximum-likelihood estimator.

Keywords Asymptotically efficient estimator · Generalized hyperbolic distribution · Mixed-effects location-scale model

1 Introduction

The key step in the population approach (Lavielle 2015) is modeling dynamics of many individuals to introduce a flexible probabilistic structure for the random vector $Y_i = (Y_i(t_{ij}))_{j=1}^{n_i} \in \mathbb{R}^{n_i}$ representing time series data (supposed to be univariate) from ith individual. Here, $t_{i1} < \cdots < t_{in_i}$ denotes sampling times, which may vary across the individuals with possibly different n_i for $i = 1, \ldots, N$. The model is desired to be tractable from theoretical and computational points of view.

✉ Hiroki Masuda
hmasuda@ms.u-tokyo.ac.jp

1 Graduate School of Mathematics, Kyushu University, 744 Motooka Nishi-ku, Fukuoka 819-0395, Japan

2 Graduate School of Mathematical Sciences, The University of Tokyo, 3-8-1 Komaba Meguro-ku, Tokyo 153-8914, Japan

In the classical linear mixed-effects model (Laird and Ware 1982), the target variable Y_i in $\mathbb{R}^{n_i}$ is described by

$$Y_i = X_i \beta + Z_i b_i + \epsilon_i, \tag{1.1}$$

for $i = 1, \ldots, N$, where the explanatory variables $X_i \in \mathbb{R}^{n_i} \otimes \mathbb{R}^p$ and $Z_i \in \mathbb{R}^{n_i} \otimes \mathbb{R}^q$ are known design matrices, where $\{b_i\}$ and $\{\epsilon_i\}$ are mutually independent centered i.i.d. sequences with covariance matrices $G \in \mathbb{R}^q \otimes \mathbb{R}^q$ and $H_i \in \mathbb{R}^{n_i} \otimes \mathbb{R}^{n_i}$, respectively; typical examples of $H_i = (H_{i,kl})$ include $H_i = \sigma^2 I_{n_i}$ (I_q denotes the q-dimensional identity matrix) and $H_{i,kl} = \sigma^2 \rho^{|k-l|}$ with ρ denoting the correlation coefficient. Although the model (1.1) is quite popular in studying longitudinal data, it is not adequate for modeling *intra-individual* variability. Formally speaking, this means that for each i, conditionally on b_i the objective variable Y_i has the covariance which does not depend on b_i. Therefore, the model is not suitable if one wants to incorporate a random effect across the individuals into the covariance and higher order structures such as skewness and kurtosis.

1.1 Mixed-effects location-scale model

Let us briefly review the previous study which motivated our present study. The paper (Hedeker et al. 2008) introduced a variant of (1.1), called the mixed-effects location-scale (MELS) model, for analyzing ecological momentary assessment (EMA) data; the MELS model was further studied in Hedeker et al. (2009, 2012) and Hedeker and Nordgren (2013) from application and computational points of view. EMA is also known as the experience sampling method, which is not retrospective and the individuals are required to answer immediately after an event occurs. Modern EMA data in mental health research is longitudinal, typically consisting of possibly irregularly spaced sampling times from each patient. To avoid the so-called "recall bias" of retrospective self-reports from patients, the EMA method records many events in daily life at the moment of their occurrence. The primary interest is modeling both between- and within-subjects heterogeneities, hence one is naturally led to incorporate random effects into both trend and scale structures. We refer to Shiffman et al. (2008) for detailed information on EMA data.

In the MELS model, the jth sample Y_{ij} from the ith individual is given by

$$Y_{ij} = x_{ij}^\top \beta + \exp\left(\frac{1}{2} z_{ij}^\top \alpha\right) \epsilon_{1,i} + \exp\left(\frac{1}{2}(w_{ij}^\top \tau + \sigma_w \epsilon_{2,i})\right) \epsilon_{3,ij} \tag{1.2}$$

for $1 \le j \le n_i$ and $1 \le i \le N$. Here, (x_{ij}, z_{ij}, w_{ij}) are non-random explanatory variables, $(\epsilon_{1,i}, \epsilon_{2,i})$ denote the i.i.d. random-effect, and $\epsilon_{3,ij}$ denote the driving noises for each $i \le N$ such that

$$(\epsilon_{1,i}, \epsilon_{2,i}, \epsilon_{3,ij}) \sim N_3\left(0, \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)$$

and that $\epsilon_{3,i1}, \ldots, \epsilon_{3,in_i} \sim$ i.i.d. $N(0, 1)$, with $(\epsilon_{1,i}, \epsilon_{2,i})$ and $(\epsilon_{3,ij})_{j \le n_i}$ being mutually independent. Direct computations give the following expressions: $E[Y_{ij}] = x_{ij}^\top \beta$, $\mathrm{Var}[Y_{ij}] = \exp(w_{ij}^\top \tau + \sigma_w^2/2) + \exp(z_i^\top \alpha)$, and also $\mathrm{Cov}[Y_{ik}, Y_{il}] = \exp(z_i^\top \alpha)$ for $k \ne l$; the covariance structure is to be compared with the one (2.2) of our model. Further, their conditional versions given the random-effect variable $R_i := (\epsilon_{1,i}, \epsilon_{2,i})$ are as follows: $E[Y_{ij}|R_i] = x_{ij}^\top \beta + \exp(z_i^\top \alpha/2)\epsilon_{1,i}$, $\mathrm{Var}[Y_{ij}|R_i] = \exp(w_{ij}^\top \tau + \sigma_w \epsilon_{2,i})$, and $\mathrm{Cov}[Y_{ik}, Y_{il}|R_i] = 0$ for $k \ne l$. We also note that the conditional distribution

$$\mathcal{L}(Y_{i1}, \ldots, Y_{in_i}|R_i) = N_{n_i}\left(X_i\beta + \mathbf{1}_{n_i}e^{z_i^\top \alpha/2}\epsilon_{1,i},\ \mathrm{diag}\left(e^{w_{i1}^\top \tau + \sigma_w \epsilon_{2,i}}, \ldots, e^{w_{in_i}^\top \tau + \sigma_w \epsilon_{2,i}}\right)\right),$$

where $X_i := (x_{i1}, \ldots, x_{in_i})$ and $\mathbf{1}_{n_i} \in \mathbb{R}^{n_i}$ has the entries all being 1. Importantly, the marginal distribution $\mathcal{L}(Y_{i1}, \ldots, Y_{in_i})$ is not Gaussian. See Hedeker et al. (2008) for details about the data-analysis aspects of the MELS model.

The third term on the right-hand side of (1.2) obeys a sort of normal-variance mixture with the variance mixing distribution being log-normal, introducing the so-called leptokurtosis (heavier tail than the normal distribution). Further, the last two terms on the right-hand side enable us to incorporate skewness into the marginal distribution $\mathcal{L}(Y_{ij})$; it is symmetric around $x_{ij}^\top \beta$ if $\rho = 0$.

The optimization of the corresponding likelihood function is quite time-consuming since we need to integrate the latent variables $(\epsilon_{1,ij}, \epsilon_{2,ij})$: the log-likelihood function of $\theta := (\beta, \alpha, \tau, \sigma_w, \rho)$ is given by

$$\theta \mapsto \sum_{i=1}^{N} \log \left\{ \int_{\mathbb{R}^2} \phi_{n_i}\left(Y_i;\ \mu_i(\beta, \alpha, X_i, z_i; x_1),\ \Sigma_i(\tau, \sigma_w, \rho, w_i; x_1, x_2)\right) \right.$$

$$\left. \times \phi_2((x_1, x_2); 0, I_2)dx_1 dx_2 \right\}, \tag{1.3}$$

where $w_i := (w_{ij})_{j \le n_i}$, $z_i := (z_{ij})_{j \le n_i}$, $\phi_m(\cdot; \mu, \Sigma)$ denotes the m-dimensional $N(\mu, \Sigma)$-density, and

$$\mu_i(\beta, \alpha, X_i, z_i; x_1) := X_i\beta + \mathbf{1}_{n_i}e^{z_i^\top \alpha/2}x_1,$$
$$\Sigma_i(\tau, \sigma_w, \rho, w_i; x_1, x_2) :=$$
$$\mathrm{diag}\left(e^{w_{i1}^\top \tau + \sigma_w(\rho x_1 + \sqrt{1-\rho^2}x_2)}, \ldots, e^{w_{in_i}^\top \tau + \sigma_w(\rho x_1 + \sqrt{1-\rho^2}x_2)}\right).$$

Just for reference, we present a numerical experiment by R Software for computing the maximum-likelihood estimator (MLE). We set $N = 1000$ and $n_1 = n_2 = \cdots = n_{1000} = 10$ and generated $x_{ij}, z_{ij}, w_{ij} \sim$ i.i.d. $N_2(0, I_2)$ independently; then, the target parameter is 8-dimensional. The true values were set as follows: $\beta = (0.6, -0.2)$, $\alpha = (-0.3, 0.5)$, $\tau = (-0.5, 0.3)$, $\sigma_w = \sqrt{0.8} \approx 0.894$, and $\rho = -0.3$. The results based on a single set of data are given in Table 1. It took more than 20 h in our R code for obtaining one MLE (Apple M1 Max, memory 64GB; the R function `adaptIntegrate` was used for the numerical integration); we have also run the simulation code for $N = 500$ and $n_1 = n_2 = \cdots = n_{500} = 5$, and then it took

Table 1 MLE results; the computation time for one pair was about 21 hours

	β_0	β_1	α_0	α_1	τ_0	τ_1	σ_w	ρ
True values	0.600	− 0.200	− 0.300	0.500	− 0.500	0.300	0.894	− 0.300
MLE	0.597	− 0.193	− 0.269	0.492	− 0.507	0.285	0.860	− 0.286

about 8 h. The program should run much faster if other software such as Fortran and MATLAB is used instead of R, but we will not deal with that direction here. Though it is cheating, the numerical search started from the true values; it would be much more time-consuming and unstable if the initial values were far from the true ones.

The EM-algorithm type approach for handling latent variables would work at least numerically, while it is also expected to be time-consuming even if a specific numerical recipe is available. Some advanced tools for numerical integration would help to some extent, but we will not pursue it here.

1.2 Our objective

In this paper, we propose an alternative computationally much simpler way of the joint modeling of the mean and within-subject variance structures. Specifically, we construct a class of parameter-varying models based on the univariate generalized hyperbolic (GH) distribution and study its theoretical properties. The model can be seen as a special case of inhomogeneous normal-variance-mean mixtures and may serve as an alternative to the MELS model; see Sect. 1 for a summary of the GH distributions. Recently, the family has received attention for modeling non-Gaussian continuous repeated measurement data (Asar et al. 2020), but ours is constructed based on a different perspective directly by making some parameters of the GH distribution covariate dependent.

This paper is organized as follows. Section 2 introduces the proposed model and presents the local-likelihood analysis, followed by numerical experiments. Section 3 considers the construction of a specific asymptotically optimal estimator and presents its finite-sample performance with comparisons with the MLE. Section 4 gives a summary and potential directions for future issues.

2 Parameter-varying generalized hyperbolic model

2.1 Proposed model

We model the objective variable at jth-sampling time point from the ith-individual by

$$Y_{ij} = x_{ij}^\top \beta + s(z_{ij}, \alpha) v_i + \sqrt{v_i}\, \sigma(w_{ij}, \tau) \epsilon_{ij} \tag{2.1}$$

for $j = 1, \ldots, n_i$ and $i = 1, \ldots, N$, where

- $x_{ij} \in \mathbb{R}^{p_\beta}, z_{ij} \in \mathbb{R}^{p'_\alpha}$, and $w_{ij} \in \mathbb{R}^{p'_\tau}$ are given non-random explanatory variables;

- $\beta \in \Theta_\beta \subset \mathbb{R}^{p_\beta}$, $\alpha \in \Theta_\alpha \subset \mathbb{R}^{p_\alpha}$, and $\tau \in \Theta_\tau \subset \mathbb{R}^{p_\tau}$ are unknown parameters;
- The random-effect variables $v_1, v_2, \ldots \sim$ i.i.d. $GIG(\lambda, \delta, \gamma)$, where GIG refers to the generalized inverse Gaussian distribution (see Sect. 1);
- $\{\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^\top\}_{i \geq 1} \sim$ i.i.d. $N(0, I_{n_i})$, independent of $\{v_i\}_{i \geq 1}$;
- $s : \mathbb{R}^{p'_\alpha} \times \Theta_\alpha \mapsto \mathbb{R}$ and $\sigma : \mathbb{R}^{p'_\tau} \times \Theta_\tau \mapsto (0, \infty)$ are known measurable functions.

As mentioned in the introduction, for (2.1), one may think of the continuous-time model without system noise:

$$Y_i(t_{ij}) = x_i(t_{ij})^\top \beta + s(z_i(t_{ij}), \alpha)v_i + \sqrt{v_i}\, \sigma(w_i(t_{ij}), \tau)\epsilon_i(t_{ij}),$$

where t_{ij} denotes the jth sampling time for the ith individual.

We will write $Y_i = (Y_{i1}, \ldots, Y_{in_i}) \in \mathbb{R}^{n_i}$, $x_i = (x_{i1}, \ldots, x_{in_i}) \in \mathbb{R}^{n_i} \otimes \mathbb{R}^{p_\beta}$, and so on for $i = 1, \ldots, N$, and also

$$\theta := (\beta, \alpha, \tau, \lambda, \delta, \gamma) \in \Theta_\beta \times \Theta_\alpha \times \Theta_\tau \times \Theta_\lambda \times \Theta_\delta \times \Theta_\gamma =: \Theta \subset \mathbb{R}^p,$$

where Θ is supposed to be a convex domain and $p := p_\beta + p_\alpha + p_\tau + 3$. We will use the notation $(P_\theta)_{\theta \in \Theta}$ for the family of distributions of $\{(Y_i, v_i, \epsilon_i)\}_{i \geq 1}$, which is completely characterized by the finite-dimensional parameter θ. The associated expectation and covariance operators will be denoted by E_θ and Cov_θ, respectively.

Let us write $s_{ij}(\alpha) = s(z_{ij}, \alpha)$ and $\sigma_{ij}(\tau) = \sigma(w_{ij}, \tau)$. For each $i \leq N$, the variable $Y_{i1}, \ldots, Y_{in_i}$ are v_i-conditionally independent and normally distributed under P_θ:

$$\mathcal{L}(Y_{ij}|v_i) = N\left(x_{ij}^\top \beta + s_{ij}(\alpha)v_i,\ \sigma_{ij}^2(\tau)v_i\right).$$

For each i, we have the specific covariance structure

$$\mathrm{Cov}_\theta[Y_{ij}, Y_{ik}] = s_{ij}(\alpha)s_{ik}(\alpha)\, \mathrm{Var}_\theta[v_i]. \tag{2.2}$$

The marginal distribution $\mathcal{L}(Y_{i1}, \ldots, Y_{in_i})$ is the multivariate GH distribution; a more flexible dependence structure could be incorporated by introducing the non-diagonal scale matrix (see Sect. 4 for a formal explanation). By the definition of the GH distribution, the variables Y_{ij} and Y_{ik} may be uncorrelated for some (z_{ij}, α) while they cannot be mutually independent.

We can explicitly write down the log-likelihood function of $(Y_1, \ldots, Y_N)$ as follows:

$$\ell_N(\theta) = -\frac{1}{2}\log(2\pi)\sum_{i=1}^N n_i + N\lambda \log\left(\frac{\gamma}{\delta}\right) - N\log K_\lambda(\delta\gamma) - \frac{1}{2}\sum_{i,j}\log\sigma_{ij}^2(\tau)$$

$$+ \sum_{i=1}^N \left(\lambda - \frac{n_i}{2}\right)\log B_i(\beta, \tau, \delta) - \sum_{i=1}^N \left(\lambda - \frac{n_i}{2}\right)\log A_i(\alpha, \tau, \gamma)$$

$$+ \sum_{i,j}\frac{s_{ij}(\alpha)}{\sigma_{ij}^2(\tau)}(Y_{ij} - x_{ij}^\top\beta) + \sum_{i=1}^N \log K_{\lambda-\frac{n_i}{2}}\left(A_i(\alpha, \tau, \gamma)B_i(\beta, \tau, \delta)\right), \tag{2.3}$$

where $\sum_{i,j}$ denotes a shorthand for $\sum_{i=1}^{N} \sum_{j=1}^{n_i}$ and

$$A_i(\alpha, \tau, \gamma) := \sqrt{\gamma^2 + \sum_{j=1}^{n_i} \frac{s_{ij}^2(\alpha)}{\sigma_{ij}^2(\tau)}}, \tag{2.4}$$

$$B_i(\beta, \tau, \delta) := \sqrt{\delta^2 + \sum_{j=1}^{n_i} \frac{1}{\sigma_{ij}^2(\tau)}(Y_{ij} - x_{ij}^\top \beta)^2}. \tag{2.5}$$

The detailed calculation is given in Sect. B.1.

To deduce the asymptotic property of the MLE, there are two typical ways: the global- and the local-consistency arguments. In the present inhomogeneous model where the variables (x_{ij}, z_{ij}, w_{ij}) are non-random, the two asymptotics have different features: on one hand, the global-consistency one generally entails rather messy descriptions of the regularity conditions as was detailed in the previous study (Fujinaga 2021), while entailing theoretically stronger global claims; on the other hand, the local one only guarantees the existence of good local maxima of $\ell_N(\theta)$ while only requiring much weaker local-around-θ_0 regularity conditions.

2.2 Local asymptotics of MLE

In the sequel, we fix a true value $\theta_0 = (\beta_0, \alpha_0, \tau_0, \lambda_0, \delta_0, \gamma_0) \in \Theta$, where $\Theta_\delta \times \Theta_\gamma \subset (0, \infty)^2$; note that we are excluding the boundary (gamma and inverse-gamma) cases for $\mathcal{L}(v_i)$.

For a domain A, let $\mathcal{C}^k(\overline{A})$ denote a set of real-valued $\mathcal{C}^k$-class functions for which the lth-partial derivatives $(0 \le l \le k)$ admit continuous extensions to the boundary of A. The asymptotic symbols will be used for $N \to \infty$ unless otherwise mentioned.

Assumption 2.1 (1) $\sup_{i \ge 1} \left(n_i \vee \max_{1 \le j \le n_i} \max\{|x_{ij}|, |z_{ij}|, |w_{ij}|\} \right) < \infty$.
(2) $\alpha \mapsto s(z, \alpha) \in \mathcal{C}^3(\overline{\Theta_\alpha})$ for each z.
(3) $\tau \mapsto \sigma(w, \tau) \in \mathcal{C}^3(\overline{\Theta_\tau})$ for each w, and $\inf_{(w,\tau) \in \mathbb{R}^{p'_\tau} \times \Theta_\tau} \sigma(w, \tau) > 0$.

We are going to prove the local asymptotics of the MLE by applying the general result (Sweeting 1980, Theorems 1 and 2).

Under Assumption 2.1 and using the basic facts about the Bessel function $K_\cdot(\cdot)$ (see Sect. 1), we can find a compact neighborhood $B_0 \subset \Theta$ of θ_0 such that

$$\forall K > 0, \quad \sup_{i \ge 1} \max_{1 \le j \le n_i} \sup_{\theta \in B_0} E_\theta\big[|Y_{ij}|^K\big] < \infty.$$

Note that $\min\{\delta, \gamma\} > 0$ inside B_0.

Let $M^{\otimes 2} := MM^\top$ for a matrix M, and denote by $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ the largest and smallest eigenvalues of a square matrix M, and by ∂_θ^k the kth-order partial-

differentiation operator with respect to θ. Write

$$\ell_N(\theta) = \sum_{i=1}^{N} \zeta_i(\theta)$$

for the right-hand side of (2.3). Then, by the independence we have

$$E_\theta\left[(\partial_\theta \ell_N(\theta))^{\otimes 2}\right] = \sum_{i=1}^{N} E_\theta\left[(\partial_\theta \zeta_i(\theta))^{\otimes 2}\right];$$

just for reference, the specific forms of $\partial_\theta \ell_N(\theta)$ and $\partial_\theta^2 \ell_N(\theta)$ are given in Sect. B.2. Further by differentiating $\theta \mapsto \partial_\theta^2 \ell_N(\theta)$ with recalling Assumption 2.1, it can be seen that

$$\forall K > 0, \quad \sup_{i \geq 1} \sup_{\theta \in B_0} E_\theta\left[\left|\partial_\theta^m \zeta_i(\theta)\right|^K\right] < \infty \tag{2.6}$$

for $m = 1, 2$, and that

$$\limsup_{N} \sup_{\theta \in B_0} E_\theta\left[\frac{1}{N} \sup_{\theta' \in B_0}\left|\partial_\theta^3 \ell_N(\theta')\right|\right] < \infty. \tag{2.7}$$

These moment estimates will be used later on; unlike the global-asymptotic study (Fujinaga 2021), we do not need the explicit form of $\partial_\theta^2 \ell_N(\theta)$.

We additionally assume the diverging information condition, which is inevitable for consistent estimation:

Assumption 2.2

$$\liminf_{N} \inf_{\theta \in B_0} \lambda_{\min}\left(\frac{1}{N} \sum_{i=1}^{N} E_\theta\left[(\partial_\theta \zeta_i(\theta))^{\otimes 2}\right]\right) > 0.$$

Under Assumption 2.1, we may and do suppose that the matrix

$$A_N(\theta) := \left(E_\theta\left[(\partial_\theta \ell_N(\theta))^{\otimes 2}\right]\right)^{1/2} = \left(\sum_{i=1}^{N} E_\theta\left[(\partial_\theta \zeta_i(\theta))^{\otimes 2}\right]\right)^{1/2}$$

is well-defined, where $M^{1/2}$ denotes the symmetric positive-definite root of a positive definite M. We also have $\sup_{\theta \in B_0} |A_N(\theta)|^{-1} \lesssim N^{-1/2} \to 0$. This $A_N(\theta)$ will serve as the norming matrix of the MLE; see Remark 2.5 below for Studentization. Further, the standard argument through the Lebesgue dominated theorem ensures that $E_\theta[\partial_\theta \ell_N(\theta)] = 0$ and $E_\theta\left[(\partial_\theta \ell_N(\theta))^{\otimes 2}\right] = E_\theta\left[-\partial_\theta^2 \ell_N(\theta)\right]$, followed by $A_N(\theta) = \left(E_\theta\left[-\partial_\theta^2 \ell_N(\theta)\right]\right)^{1/2}$.

For $c > 0$, Assumption 2.2 yields

$$
\sup_{\theta': |\theta' - \theta| \leq c/\sqrt{N}} \left| A_N(\theta)^{-1} A_N(\theta') - I_p \right|
$$

$$
= \sup_{\theta': |\theta' - \theta| \leq c/\sqrt{N}} \left| \left(\frac{1}{\sqrt{N}} A_N(\theta) \right)^{-1} \left(\frac{1}{\sqrt{N}} A_N(\theta') - \frac{1}{\sqrt{N}} A_N(\theta) \right) \right|
$$

$$
\lesssim \sup_{\theta': |\theta' - \theta| \leq c/\sqrt{N}} \left| \frac{1}{\sqrt{N}} A_N(\theta') - \frac{1}{\sqrt{N}} A_N(\theta) \right|
$$

$$
\lesssim \sup_{\theta': |\theta' - \theta| \leq c/\sqrt{N}} \left| \left(\frac{1}{N} \sum_{i=1}^{N} E_{\theta'} \left[(\partial_\theta \zeta_i(\theta'))^{\otimes 2} \right] \right)^{1/2} - \left(\frac{1}{N} \sum_{i=1}^{N} E_\theta \left[(\partial_\theta \zeta_i(\theta))^{\otimes 2} \right] \right)^{1/2} \right|
$$

$$
\to 0. \tag{2.8}
$$

Here, the last convergence holds since the function $\theta \mapsto N^{-1/2} A_N(\theta)$ is uniformly continuous over B_0.

Define the normalized observed information:

$$
\mathcal{I}_N(\theta) := -A_N(\theta)^{-1} \partial_\theta^2 \ell_N(\theta) A_N(\theta)^{-1 \top}.
$$

Then, it follows from Assumption 2.2 that

$$
\left| \mathcal{I}_N(\theta) - I_p \right| = \left| \left(\frac{1}{\sqrt{N}} A_N(\theta) \right)^{-1} \left(\mathcal{I}_N(\theta) - \left(\frac{1}{\sqrt{N}} A_N(\theta) \right)^{\otimes 2} \right) \left(\frac{1}{\sqrt{N}} A_N(\theta) \right)^{-1 \top} \right|
$$

$$
\lesssim \left| \mathcal{I}_N(\theta) - \left(\frac{1}{\sqrt{N}} A_N(\theta) \right)^{\otimes 2} \right|
$$

$$
\lesssim \left| \frac{1}{N} \sum_{i=1}^{N} \left(\partial_\theta^2 \zeta_i(\theta) - E_\theta \left[\partial_\theta^2 \zeta_i(\theta) \right] \right) \right|.
$$

Then, (2.6) ensures that

$$
\sup_{\theta \in B_0} E_\theta \left[|\mathcal{I}_N(\theta) - I_p|^2 \right] \lesssim \frac{1}{N} \left(\frac{1}{N} \sum_{i=1}^{N} \sup_{\theta \in B_0} E_\theta \left[\left| \partial_\theta^2 \zeta_i(\theta) \right|^2 \right] \right) \lesssim \frac{1}{N} \to 0,
$$

followed by the property

$$
\forall \epsilon > 0, \quad \sup_{\theta \in B_0} P_\theta \left[|\mathcal{I}_N(\theta) - I_p| > \epsilon \right] \to 0. \tag{2.9}
$$

Let $\xrightarrow{\mathcal{L}}$ denote the convergence in distribution. Having obtained (2.7), (2.8), and (2.9), we can conclude the following theorem by applying (Sweeting 1980, Theorems 1 and 2).

Theorem 2.3 *Under Assumptions 2.1 and 2.2, we have the following statements under* P_{θ_0}.

(1) *For any bounded sequence $(u_N) \subset \mathbb{R}^p$,*

$$\ell_N \left(\theta_0 + A_N(\theta_0)^{\top -1} u_N \right) - \ell_N(\theta_0) = u_N^\top \Delta_N(\theta_0) - \frac{1}{2} |u_N|^2 + o_p(1),$$

with

$$\Delta_N(\theta_0) := A_N(\theta_0)^{-1} \partial_\theta \ell_N(\theta_0) \xrightarrow{\mathcal{L}} N(0, I_p).$$

(2) *There exists a local maximum point $\hat{\theta}_N$ of $\ell_N(\theta)$ with P_{θ_0}-probability tending to 1, for which*

$$A_N(\theta_0)^\top (\hat{\theta}_N - \theta_0) = \Delta_N(\theta_0) + o_p(1) \xrightarrow{\mathcal{L}} N(0, I_p). \qquad (2.10)$$

Remark 2.4 (Asymptotically efficient estimator) By the standard argument about the local asymptotic normality (LAN) of the family $\{P_\theta\}_{\theta \in \Theta}$, any estimators $\hat{\theta}_N^*$ satisfying that

$$A_N(\theta_0)^\top (\hat{\theta}_N^* - \theta_0) = \Delta_N(\theta_0) + o_p(1) \qquad (2.11)$$

are regular and asymptotically efficient in the sense of Hajék–Le Cam. See Basawa and Scott (1983) and Jeganathan (1982) for details.

Remark 2.5 (Studentization of 2.10) Here is a remark on the construction of approximate confidence sets. Define the statistics

$$\hat{A}_N := \left(\sum_{i=1}^{N} (\partial_\theta \zeta_i(\hat{\theta}_N))^{\otimes 2} \right)^{1/2}. \qquad (2.12)$$

Then, to make inferences for θ_0, we can use the distributional approximations $\hat{A}_N(\hat{\theta}_N - \theta_0) = \Delta_N(\theta_0) + o_p(1) \xrightarrow{\mathcal{L}} N_p(0, I_p)$ and

$$(\hat{\theta}_N - \theta_0)^\top \hat{A}_N^2 (\hat{\theta}_N - \theta_0) \xrightarrow{\mathcal{L}} \chi^2(p). \qquad (2.13)$$

To see this, it is enough to show that under P_{θ_0},

$$A_N(\theta_0)^{-1} \hat{A}_N = I_p + o_p(1). \qquad (2.14)$$

We have $\sqrt{N}(\hat{\theta}_N - \theta_0) = O_p(1)$ by Theorem 2.3 and Assumption 2.2. This together with the Burkholder inequality and (2.6) yield that

$$N^{-1/2} \hat{A}_N = \left(\frac{1}{N} \sum_{i=1}^{N} \left((\partial_\theta \zeta_i(\hat{\theta}_N))^{\otimes 2} - (\partial_\theta \zeta_i(\theta_0))^{\otimes 2} \right) \right.$$

$$+ \frac{1}{N} \sum_{i=1}^{N} \left((\partial_\theta \zeta_i(\theta_0))^{\otimes 2} - E_{\theta_0} \left[(\partial_\theta \zeta_i(\theta_0))^{\otimes 2} \right] \right) + \left(N^{-1/2} A_N(\theta_0) \right)^2 \Big)^{1/2}$$

$$= \left(O_p(N^{-1/2}) + \left(N^{-1/2} A_N(\theta_0) \right)^2 \right)^{1/2}$$

and hence

$$A_N(\theta_0)^{-1} \hat{A}_N = \left(N^{-1/2} A_N(\theta_0) \right)^{-1} \left\{ o_p(1) + \left(N^{-1/2} A_N(\theta_0) \right)^2 \right\}^{1/2} = I_p + o_p(1),$$

concluding (2.14). Note that, instead of (2.12), we may also use the square root of the observed information matrix

$$\widetilde{A}_N := \left(- \sum_{i=1}^{N} \partial_\theta^2 \zeta_i(\hat{\theta}_N) \right)^{1/2} \tag{2.15}$$

for concluding the same weak convergence as in (2.13). In our numerical experiments, we made use of this $\widetilde{A}_N^2$ for computing the confidence interval and the empirical coverage probability. The elements of $\widetilde{A}_N^2$ are explicit while rather lengthy: see Sect. B.2.

Remark 2.6 (Misspecifications) In addition to the linear form $x_{ij}^\top \beta$ in (2.1), misspecification of a parametric form of the function $(s(z_i, \alpha), \sigma(w_{ij}, \tau))$ is always concerned. Using the M-estimation theory (for example, see White (1982) and (Fahrmeir 1990, Section 5)), under appropriate identifiability conditions, it is possible to handle their misspecified parametric forms. In that case, however, the maximum-likelihood-estimation target, say θ_*, is the optimal parameter (to be uniquely determined) in terms of the Kullback–Leibler divergence, and we do not have the LAN property in Theorem 2.3 in the usual sense while an asymptotic normality result of the form $\sqrt{N}(\hat{\theta}_N - \theta_*) \xrightarrow{\mathcal{L}} N(0, \Gamma_0^{-1} \Sigma_0 \Gamma_0^{-1})$ could be given, where (non-random) Σ_0 and Γ_0 are specified by $N^{-1/2} \partial_\theta \ell_N(\theta_*) \xrightarrow{\mathcal{L}} N(0, \Sigma_0)$ and $-N^{-1} \partial_\theta^2 \ell_N(\theta_*) \xrightarrow{P} \Gamma_0$.

Finally, we note that the statistical problem will become non-standard if we allow that the true value of (δ, γ) for the GIG distribution $\mathcal{L}(v_i)$ satisfies that $\delta_0 = 0$ or $\gamma_0 = 0$. We have excluded these boundary cases at the beginning of Sect. 2.2.

2.3 Numerical experiments

For simulation purposes, we consider the following model:

$$Y_{ij} = x_{ij}^\top \beta + \tanh(z_{ij}^\top \alpha) v_i + \sqrt{v_i \exp(w_{ij}^\top \tau)} \, \epsilon_{ij}, \tag{2.16}$$

where the ingredients are specified as follows.

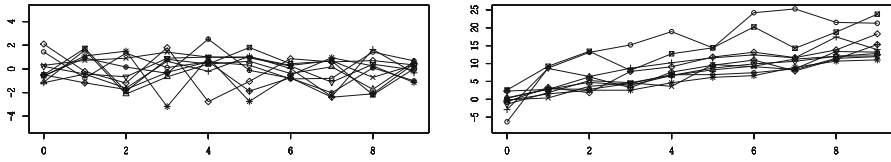- $N = 1000$ and $n_1 = n_2 = \cdots = n_{1000} = 10$.

Fig. 1 Longitudinal-data plots of 10 individuals in case (1) (left) and case (2) (right)

- The two different cases for the covariates $x_{ij}, z_{ij}, w_{ij} \in \mathbb{R}^2$:

(i) $x_{ij}, z_{ij}, w_{ij} \sim$ i.i.d. $N(0, I_2)$;
(ii) The first components of x_{ij}, z_{ij}, w_{ij} are sampled from independent $N(0, 1)$, and all the second ones are set to be $j - 1$.

The setting (ii) incorporates similarities across the individuals; see Fig. 1.

- $v_1, v_2, \cdots \sim$ i.i.d. $GIG(\lambda, \delta, \gamma)$.
- $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i}) \sim N(0, I_{n_i})$, independent of $\{v_i\}$.
- $\theta = (\beta, \alpha, \tau, \lambda, \delta, \gamma) = (\beta_0, \beta_1, \alpha_0, \alpha_1, \tau_0, \tau_1, \lambda, \delta, \gamma) \in \mathbb{R}^9$.
- True values of θ:

(i) $\beta = (0.3, \ 0.5)$, $\alpha = (-0.04, \ 0.05)$, $\tau = (0.05, \ 0.07)$, $\lambda = 1.2$, $\delta = 1.5$, $\gamma = 2$;
(ii) $\beta = (0.3, \ 1.2)$, $\alpha = (-0.4, \ 0.8)$, $\tau = (0.05, \ 0.007)$, $\lambda = 0.9$, $\delta = 1.2$, $\gamma = 0.9$.

We numerically computed the MLE $\hat{\theta}_N$ by optimizing the log-likelihood; the modified Bessel function $K.(\cdot)$ can be efficiently computed by the existing numerical libraries such as `besselK` in R Software. We repeated the Monte Carlo trials 1000 times, computed the Studentized estimates $\widetilde{A}_N(\hat{\theta}_N - \theta_0)$ with (2.15) in each trial, and then drew histograms in Figs. 2 and 3, where the red lines correspond to the standard normal densities. Also given in Figs. 2 and 3 are the histograms of the chi-square approximations based on (2.13).

The computation time for one MLE was about 8 min for case (i) and about 6 min for case (ii). Estimation performance for $(\lambda, \delta, \gamma)$ were less efficient than those for (β, α, τ). It is expected that the unobserved nature of the GIG variables make the standard-normal approximations relatively worse.

It is worth mentioning that case (ii) shows better normal approximations, in particular for $(\lambda, \delta, \gamma)$; case (ii) would be simpler in the sense that the data from each individual have similarities in their trend (mean) structures.

Table 2 shows the empirical 95%-coverage probability for each parameter in both (i) and (ii), based on the confidence intervals $\hat{\theta}_N^{(k)} \pm z_{\alpha/2}[(-\partial_\theta^2 \ell_N(\hat{\theta}_N))^{-1}]_{kk}^{1/2}$ for $k = 1, \ldots, 9$ with $\hat{\theta}_N =: (\hat{\theta}_N^{(k)})_{k \leq 9}$ and $\alpha = 0.05$. We had 365 and 65 numerically unstable cases among 1000 trials, respectively (mostly cased by a degenerate $\det(-\partial_\theta^2 \ell_N(\hat{\theta}_N))$). Therefore, the coverage probabilities were computed based on the remaining cases.

Let us note the crucial problem in the above Monte Carlo trials: the objective log-likelihood is highly non-concave, hence as usual the numerical optimization suffers from the initial-value and local-maxima problems. Here is a numerical example based
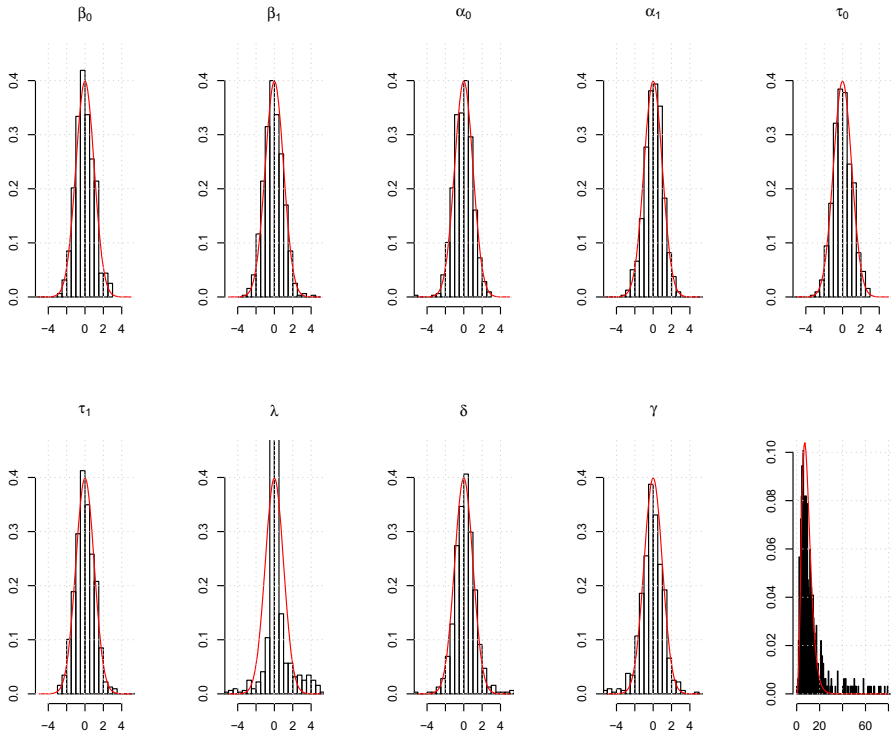
Fig. 2 Standardized distributions of the MLE in case (i). The lower rightmost panel shows the chi-square approximation based on (2.13)

on only a single set of data with $N = 1000$ and $n_1 = n_2 = \cdots = n_{1000} = 10$ as before. The same model as in (2.16) together with the subsequent settings was used, except that we set $\lambda = -1/2$ known from the beginning so that the latent variables $v_1, \ldots, v_N$ have the inverse-Gaussian population $IG(\delta, \gamma) = GIG(-1/2, \delta, \gamma)$. For the true parameter values specified in Table 3, we run the following two cases for the initial values of the numerical optimization:

(i') The true value;
(ii') $(\underbrace{1.0 \times 10^{-8}, \ldots, 1.0 \times 10^{-8}}_{6 \text{ times}}, 1.0 \times 10^{-4}, 1.0 \times 10^{-3})$.

The results in Table 3 clearly show that the inverse-Gaussian parameter (δ, γ) can be quite sensitive to a bad starting point for the numerical search. In the next section, to bypass the numerical instability we will construct easier-to-compute initial estimators and their improved versions asymptotically equivalent to the MLE.

3 Asymptotically efficient estimator

Building on Theorem 2.3, we now turn to global asymptotics through the classical Newton–Raphson type procedure. A systematic account for the theory of the one-step
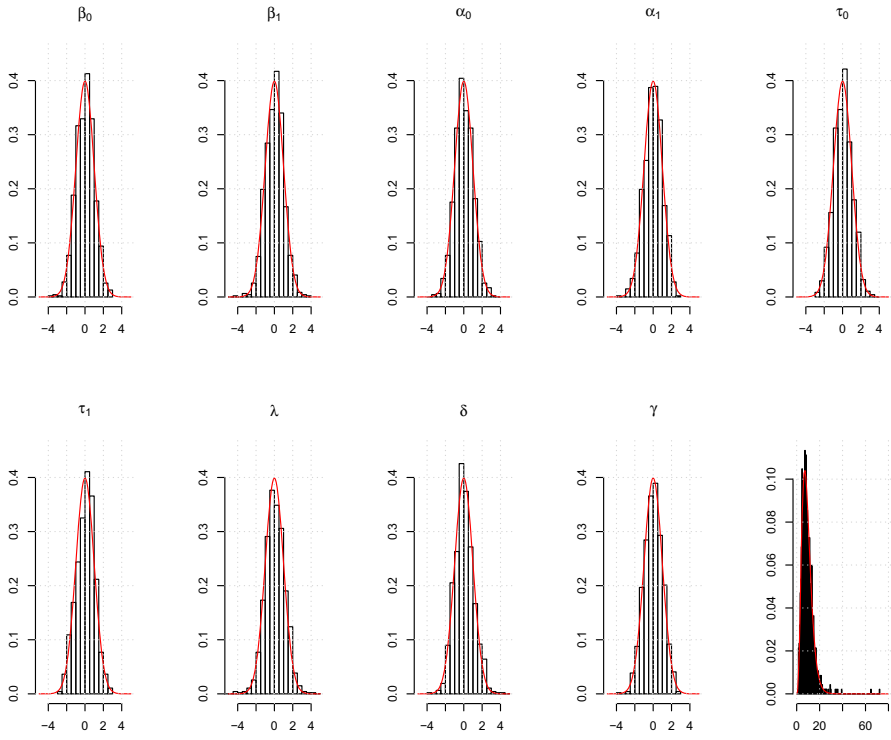
Fig. 3 Standardized distributions of the MLE in case (ii). The lower rightmost panel shows the chi-square approximation based on (2.13)

estimator can be found in many textbooks, such as (van der Vaart 1998, Section 5.7). Let us briefly overview the derivation with the current matrix-norming setting.

Suppose that we are given an initial estimator $\hat{\theta}_N^0 = (\hat{\alpha}_N^0, \hat{\beta}_N^0, \hat{\tau}_N^0, \hat{\lambda}_N^0, \hat{\delta}_N^0, \hat{\gamma}_N^0)$ of θ_0 satisfying that

$$\hat{u}_N^0 := A_N(\theta_0)^\top(\hat{\theta}_N^0 - \theta_0) = O_p(1).$$

By Theorem 2.3 and Assumption 2.2, this amounts to

$$\sqrt{N}(\hat{\theta}_N^0 - \theta_0) = O_p(1). \tag{3.1}$$

We define the one-step estimator $\hat{\theta}_N^1$ by

$$\hat{\theta}_N^1 := \hat{\theta}_N^0 - \left(\partial_\theta^2 \ell_N(\hat{\theta}_N^0)\right)^{-1} \partial_\theta \ell_N(\hat{\theta}_N^0) \tag{3.2}$$

on the event $\{\hat{\theta}_N^1 \in \Theta, \ \det(\partial_\theta^2 \ell_N(\hat{\theta}_N^0)) \neq 0\}$, the P_{θ_0}-probability of which tends to 1. Write $\hat{u}_N^1 = A_N(\theta_0)^\top(\hat{\theta}_N^1 - \theta_0)$ and $\hat{\mathcal{I}}_N^0 = -A_N(\theta_0)^{-1}\partial_\theta^2 \ell_N(\hat{\theta}_N^0)A_N(\theta_0)^{-1\top}$. Using

Table 2 The empirical 95%-coverage probabilities of the MLE in cases (i) and (ii) based on 1000 trials

	β_0	β_1	α_0	α_1	τ_0	τ_1	λ	δ	γ
Case (i)	0.940	0.948	0.953	0.946	0.942	0.953	0.817	0.863	0.839
Case (ii)	0.957	0.952	0.961	0.954	0.952	0.949	0.942	0.945	0.948

Table 3 MLE based on single data set; the running time was about 2 min for case (i') and 8 min for case (ii'); the performance of estimating (δ, γ) in case (ii') shows instability

True value	β_0 -3.000	β_1 5.000	α_0 -3.000	α_1 4.000	τ_0 0.020	τ_1 -0.050	δ 1.600	γ 1.000
(i')	-3.000	4.999	-3.011	4.017	0.023	-0.047	1.603	1.005
(ii')	-3.000	4.999	-2.966	3.947	0.023	-0.052	0.947	0.000

Taylor expansion, we have

$$\hat{\mathcal{I}}_N^0 \hat{u}_N^1 = \hat{\mathcal{I}}_N^0 \hat{u}_N^0 + A_N(\theta_0)^{-1} \partial_\theta \ell_N(\hat{\theta}_N^0). \tag{3.3}$$

By the arguments in Sect. 2.2, it holds that $|\hat{\mathcal{I}}_N^0| \vee |\hat{\mathcal{I}}_N^{0-1}| = O_p(1)$. From (3.1),

$$A_N(\theta_0)^{-1} \partial_\theta \ell_N(\hat{\theta}_N^0) = \Delta_N(\theta_0) - \hat{\mathcal{I}}_N^0 \hat{u}_N^0 + O_p(N^{-1/2}). \tag{3.4}$$

Combining (3.3) and (3.4) and recalling Remarks 2.4 and 2.5, we obtain the asymptotic representation (2.11) for $\hat{\theta}_N^1$, followed by the asymptotic standard normality

$$\hat{u}_N^1 = \Delta_N(\theta_0) + o_p(1) \xrightarrow{\mathcal{L}} N_p(0, I_p)$$

and its asymptotic optimality.

3.1 Construction of initial estimator

This section aims to construct a $\sqrt{N}$-consistent estimator $\hat{\theta}_N^0$ satisfying (3.1) through the stepwise least-squares type estimators for the first three moments of Y_{ij}. We note that the model (2.1) does not have a conventional location-scale structure because of the presence of v_i in the two different terms.

We assume that the parameter space $\Theta_\beta \times \Theta_\alpha \times \Theta_\tau \times \Theta_\lambda \times \Theta_\delta \times \Theta_\gamma$ is a bounded convex domain in $\mathbb{R}^{p_\beta} \times \mathbb{R}^{p_\alpha} \times \mathbb{R}^{p_\tau} \times \mathbb{R} \times (0, \infty)^2$ with the compact closure. Write $\theta' = (\lambda, \delta, \gamma)$ for the parameters contained in $\mathcal{L}(v_1)$, the true value being denoted by $\theta'_0 = (\lambda_0, \delta_0, \gamma_0)$. Let $\mu = \mu(\theta') = E_\theta[v_1], c = c(\theta') := \text{Var}_\theta[v_1]$, and $\rho = \rho(\theta') := E_\theta[(v_i - E_\theta[v_i])^3]$; write $\mu_0 = \mu(\theta'_0), c_0 = c(\theta'_0)$, and $\rho_0 = \rho(\theta'_0)$ correspondingly. Further, we introduce the sequences of the symmetric random matrices:

$$Q_{1,N}(\alpha) := \frac{1}{N} \sum_{i,j} \left(\mu_0 \, \partial_\alpha s_{ij}(\alpha), \, x_{ij}, \, s_{ij}(\alpha_0) \right)^{\otimes 2},$$

$$Q_{2,N}(\tau) := \frac{1}{N} \sum_{i,j} \left(\mu_0 \, \partial_\tau(\sigma_{ij}^2)(\tau), \, s_{ij}^2(\alpha_0) \right)^{\otimes 2}.$$

To state our global consistency result, we need additional assumptions.

Assumption 3.1 In addition to Assumption 2.1, the following conditions hold.

(1) Global identifiability of (α, β, μ):

    (a) $\sup_\alpha |Q_{1,N}(\alpha) - Q_1(\alpha)| \to 0$ for some non-random function $Q_1(\alpha)$;

    (b) $\liminf_N \inf_\alpha \lambda_{\min}(Q_{1,N}(\alpha)) > 0$.

(2) Global identifiability of (τ, c):

    (a) $\sup_\tau |Q_{2,N}(\tau) - Q_2(\tau)| \to 0$ for some non-random function $Q_2(\tau)$;

    (b) $\liminf_N \inf_\tau \lambda_{\min}(Q_{2,N}(\tau)) > 0$.

(3) Global identifiability of ρ: $\liminf_N \frac{1}{N} \sum_{i,j} s_{ij}^6(\alpha_0) > 0$.

(4) There exists a neighborhood of θ_0' on which the mapping $\psi : \Theta_\lambda \times \Theta_\delta \times \Theta_\gamma \to (0,\infty)^2 \times \mathbb{R}$ defined by $\psi(\theta') = (\mu(\theta'), c(\theta'), \rho(\theta'))$ is bijective, and ψ is continuously differentiable at θ_0 with nonsingular derivative.

To construct $\hat{\theta}_N^0$, we will proceed as follows.

Step 1　Noting that $E_\theta[Y_{ij}] = x_{ij}^\top \beta + s_{ij}(\alpha)\mu$, we estimate (β, α, μ) by minimizing

$$M_{1,N}(\alpha, \beta, \mu) := \sum_{i,j} \left(Y_{ij} - x_{ij}^\top \beta - s_{ij}(\alpha)\mu \right)^2. \tag{3.5}$$

Let $(\hat{\alpha}_N^0, \hat{\beta}_N^0, \hat{\mu}_N^0) \in \operatorname{argmin}_{(\alpha,\beta,\mu) \in \overline{\Theta_\beta \times \Theta_\alpha \times \Theta_\mu}} M_{1,N}(\alpha, \beta, \mu)$.
For estimating the remaining parameters, we introduce the (heteroscedastic) residual

$$\hat{e}_{ij} := Y_{ij} - x_{ij}^\top \hat{\beta}_N^0 - s_{ij}(\hat{\alpha}_N^0)\hat{\mu}_N^0, \tag{3.6}$$

which is to be regarded as an estimator of the unobserved quantity $\sqrt{v_i}\, \sigma_{ij}(\tau_0)\epsilon_{ij}$.

Step 2　Noting that $\operatorname{Var}_\theta[Y_{ij}] = \sigma_{ij}^2(\tau)\mu + s_{ij}^2(\alpha)c$, we estimate the variance-component parameter (τ, α) by minimizing

$$M_{2,N}(\tau, c) := \sum_{i,j} \left(\hat{e}_{ij}^2 - \sigma_{ij}^2(\tau)\hat{\mu}_N^0 - s_{ij}^2(\hat{\alpha}_N^0)c \right)^2. \tag{3.7}$$

Let $(\hat{\tau}_N^0, \hat{c}_N^0) \in \operatorname{argmin}_{(\tau,c) \in \overline{\Theta_\tau} \times (0,\infty)} M_{2,N}(\tau, c)$.

Step 3   Noting that $E_\theta[(Y_{ij} - E_\theta[Y_{ij}])^3] = 3s_{ij}(\alpha)\sigma_{ij}^2(\tau)c + s_{ij}^3(\alpha)\rho$, we estimate ρ by the minimizer $\hat{\rho}_N^0$ of

$$M_{3,N}(\rho) := \sum_{i,j} \left(\hat{e}_{ij}^3 - 3s_{ij}(\hat{\alpha}_N^0)\sigma_{ij}^2(\hat{\tau}_N^0)\hat{c}_N^0 - s_{ij}^3(\hat{\alpha}_N^0)\rho \right)^2,$$

that is,

$$\hat{\rho}_N^0 := \left(\sum_{i,j} s_{ij}^6(\hat{\alpha}_N^0) \right)^{-1} \sum_{i,j} \left\{ \hat{e}_{ij}^3 - 3s_{ij}(\hat{\alpha}_N^0)\sigma_{ij}^2(\hat{\tau}_N^0)\hat{c}_N^0 \right\} s_{ij}^3(\hat{\alpha}_N^0). \quad (3.8)$$

Step 4   Finally, under Assumption 3.1(4), we construct $\hat{\theta}_N^{\prime 0} = (\hat{\lambda}_N^0, \hat{\delta}_N^0, \hat{\gamma}_N^0)$ through the delta method by inverting $(\hat{\mu}_N^0, \hat{c}_N^0, \hat{\rho}_N^0)$:

$$\sqrt{N}\left(\hat{\theta}_N^{\prime 0} - \theta_0' \right) = \sqrt{N} \left(\psi^{-1}(\hat{\mu}_N^0, \hat{c}_N^0, \hat{\rho}_N^0) - \psi^{-1}(\mu_0, c_0, \rho_0) \right)$$
$$= \left(\partial_{\theta'} \psi(\theta_0') \right)^{-1} \sqrt{N} \left((\hat{\mu}_N^0, \hat{c}_N^0, \hat{\rho}_N^0) - (\mu_0, c_0, \rho_0) \right) = O_p(1).$$

In the rest of this section, we will go into detail about Steps 1 to 3 mentioned above and show that the estimator $\hat{\theta}_N^0$ thus constructed satisfies (3.1); Step 4 is the standard method of moments (van der Vaart 1998, Chapter 4).

For convenience, let us introduce some notation. The multilinear-form notation

$$M[u] = \sum_{i_1,\dots,i_k} M_{i_1,\dots,i_k} u_{i_1} \dots u_{i_k} \in \mathbb{R}$$

is used for $M = \{M_{i_1,\dots,i_k}\}$ and $u = \{u_{i_1}, \dots u_{i_k}\}$. For any sequence random functions $\{F_N(\theta)\}_N$ and a non-random sequence $(a_N)_N \subset (0, \infty)$, we will write $F_N(\theta) = O_p^*(a_N)$ and $F_N(\theta) = o_p^*(a_N)$ when $\sup_\theta |F_N(\theta)| = O_p(a_N)$ and $\sup_\theta |F_N(\theta)| = o_p(a_N)$ under P_{θ_0}, respectively. Further, we will denote by $m_i = (m_{i1}, \dots, m_{in_i}) \in \mathbb{R}^{n_i}$ any zero-mean (under P_{θ_0}) random variables such that $m_1, \dots, m_N$ are mutually independent and $\sup_{i \geq 1} \max_{1 \leq j \leq n_i} E_{\theta_0}[|m_{ij}|^K] < \infty$ for any $K > 0$; its specific form will be of no importance.

3.1.1 Step 1

Put $a = (\alpha, \beta, \mu)$ and $a_0 = (\alpha_0, \beta_0, \mu_0)$. By (2.1) and (3.5), we have

$$\mathbb{Y}_{1,N}(a) := \frac{1}{N} \left(M_{1,N}(a) - M_{1,N}(a_0) \right)$$
$$= -\frac{2}{N} \sum_{i,j} \left(x_{ij}, s_{ij}(\alpha_0), \mu_0 \right) \cdot \left(\beta - \beta_0, \mu - \mu_0, s_{ij}(\alpha) - s_{ij}(\alpha_0) \right) m_{ij}$$

$$+ \left(\frac{1}{N} \sum_{i,j} \left(\mu_0 \, \partial_\alpha s_{ij}(\tilde{\alpha}), \, x_{ij}, \, s_{ij}(\alpha_0) \right)^{\otimes 2} \right) \left[(a - a_0)^{\otimes 2} \right]$$

$$= -\frac{2}{N} \sum_{i,j} \left(x_{ij}, s_{ij}(\alpha_0), \mu_0 \right) \cdot \left(\beta - \beta_0, \mu - \mu_0, s_{ij}(\alpha) - s_{ij}(\alpha_0) \right) m_{ij}$$

$$+ 2 \left(Q_{1,N}(\tilde{\alpha}) - Q_1(\tilde{\alpha}) \right) \left[(a - a_0)^{\otimes 2} \right] + 2 Q_1(\tilde{\alpha}) \left[(a - a_0)^{\otimes 2} \right],$$

where $\tilde{\alpha} = \tilde{\alpha}(\alpha, \alpha_0)$ is a point lying on the segment joining α and α_0. The first term on the rightmost side equals $O_p^*(N^{-1/2})$. The second term equals $o_p^*(1)$ by Assumption 3.1(1), hence we conclude that $|\mathbb{Y}_{1,N}(a) - \mathbb{Y}_1(a)| = o_p^*(1)$ for $\mathbb{Y}_1(a) := 2 Q_1(\tilde{\alpha}) \left[(a - a_0)^{\otimes 2} \right]$. Moreover, we have $\inf_\alpha \lambda_{\min}(Q_1(\alpha)) > 0$ hence $\operatorname{argmin} \mathbb{Y}_1 = \{a_0\}$, followed by the consistency $\hat{a}_N \xrightarrow{p} a_0$.

To deduce $\sqrt{N}(\hat{a}_N - a_0) = O_p(1)$, we may and do focus on the event $\{\partial_a M_{1,N}(\hat{a}_N) = 0\}$, on which

$$N^{-1} \partial_a^2 M_{1,N}(\tilde{a}_N) \sqrt{N}(\hat{a}_N - a_0) = -N^{-1/2} \partial_a M_{1,N}(a_0), \tag{3.9}$$

where $\tilde{a}_N$ is a random point lying on the segment joining $\hat{a}_N$ and α_0. Observe that

$$-\frac{1}{\sqrt{N}} \partial_a M_{1,N}(a_0) = \frac{2}{\sqrt{N}} \sum_{i,j} \operatorname{diag}\left(\partial_\alpha s_{ij}(\alpha_0), \, I_{p_\beta}, \, 1 \right) \left[(\mu_0, x_{ij}, s_{ij}(\alpha_0)) \right] m_{ij} = O_p(1).$$

Similarly,

$$\frac{1}{N} \partial_a^2 M_{1,N}(\tilde{a}_N) = -\frac{2\mu_0}{N} \sum_{i,j} \Big\{ m_{ij} - \left(x_{ij}, s_{ij}(\alpha_0), \mu_0 \right) \cdot$$

$$\times \left(\tilde{\beta}_N - \beta_0, \tilde{\mu}_N - \mu_0, s_{ij}(\tilde{\alpha}_N) - s_{ij}(\alpha_0) \right) \Big\}$$

$$+ \frac{2}{N} \sum_{i,j} \left(\mu_0 \partial_\alpha s_{ij}(\tilde{\alpha}_N), \, x_{ij}, \, s_{ij}(\alpha_0) \right)^{\otimes 2}.$$

Concerning the right-hand side, the first term equals $o_p(1)$, and the inverse of the second term does $Q_{1,N}(\tilde{\alpha}_N)^{-1} = \{2 Q_{1,N}(\alpha_0) + o_p(1)\}^{-1} = O_p(1)$. The last two displays combined with Assumption 3.1(1) and (3.9) conclude that $\sqrt{N}(\hat{a}_N - a_0) = O_p(1)$; it could be shown under additional conditions that $\sqrt{N}(\hat{a}_N - a_0)$ is asymptotically centered normal, while it is not necessary here.

3.1.2 Step 2

Write $\hat{u}_{\beta,N} = \sqrt{N}(\hat{\beta}_N^0 - \beta_0)$, $\hat{u}_{\mu,N} = \sqrt{N}(\hat{\mu}_N^0 - \mu_0)$, and $\hat{u}_{\alpha,ij}' = \sqrt{N}(s_{ij}(\hat{\alpha}_N^0) - s_{ij}(\alpha_0))$. Let $b := (\tau, c)$ and $b_0 := (\tau_0, c_0)$, and moreover

$$e_{ij} := \sqrt{v_i} \sigma_{ij}(\tau_0) \epsilon_{ij},$$

$$\overline{e}_{ij} := Y_{ij} - E_{\theta_0}[Y_{ij}] = e_{ij} + s_{ij}(\alpha_0)(v_i - \mu_0).$$

We have $\hat{e}_{ij} = e_{ij} - N^{-1/2}\hat{H}_{ij}$ with $\hat{H}_{ij} := x_{ij}^{\top}\hat{u}_{\beta,N} + s_{ij}(\hat{\alpha}_N^0)\hat{u}_{\mu,N} + \mu_0\hat{u}'_{\alpha,ij}$. Introduce the zero-mean random variables $\eta_{ij} := \overline{e}_{ij}^2 - \left(\sigma_{ij}^2(\tau_0) + c_0 s_{ij}^2(\alpha_0)\right)$. Then, we can rewrite $M_{2,N}(b)$ of (3.7) as

$$M_{2,N}(b) = \sum_{i,j}\left(\overline{\eta}_{ij}(b) + \frac{1}{\sqrt{N}}\hat{B}_{ij}\right)^2,$$

where

$$\overline{\eta}_{ij}(b) := \eta_{ij} - \left((\sigma_{ij}^2(\tau) - \sigma_{ij}^2(\tau_0))\hat{\mu}_N^0 + (c - c_0)s_{ij}^2(\hat{\alpha}_N^0)\right),$$

$$\hat{B}_{ij} := -2\hat{H}_{ij} + \frac{1}{\sqrt{N}}\hat{H}_{ij}^2 - \sigma_{ij}^2(\tau_0)\hat{u}_{\mu,N} - c_0\hat{u}'_{\alpha,ij}.$$

As in Sect. 3.1.1, we observe that

$$\begin{aligned}
\mathbb{Y}_{2,N}(b) &:= \frac{1}{N}\left(M_{2,N}(b) - M_{2,N}(b_0)\right) \\
&= O_p^*\left(\frac{1}{\sqrt{N}}\right) + \frac{1}{N}\sum_{i,j}\left(\overline{\eta}_{ij}^2(b) - \eta_{ij}^2\right) \\
&= O_p^*\left(\frac{1}{\sqrt{N}}\right) + \frac{1}{N}\sum_{i,j}\left((\sigma_{ij}^2(\tau) - \sigma_{ij}^2(\tau_0))\hat{\mu}_N^0 + (c - c_0)s_{ij}^2(\hat{\alpha}_N^0)\right)^2 \\
&= o_p^*(1) + \frac{1}{N}\sum_{i,j}\left((\sigma_{ij}^2(\tau) - \sigma_{ij}^2(\tau_0))\mu_0 + (c - c_0)s_{ij}^2(\alpha_0)\right)^2 \\
&= o_p^*(1) + 2Q_{2,N}(\tilde{\tau})\left[(b - b_0)^{\otimes 2}\right]
\end{aligned}$$

for some point $\tilde{\tau} = \tilde{\tau}(\tau, \tau_0)$ lying on the segment joining τ and τ_0. Thus Assumption 3.1(2) concludes the consistency $\hat{b}_N \overset{p}{\to} b_0$: we have $|\mathbb{Y}_{2,N}(b) - \mathbb{Y}_2(b)| = o_p^*(1)$ with $\mathbb{Y}_2(b) := 2Q_2(\tilde{\tau})\left[(b - b_0)^{\otimes 2}\right]$ satisfying that $\inf_\tau \lambda_{\min}(Q_2(\tau)) > 0$, hence $\operatorname{argmin}\mathbb{Y}_2 = \{b_0\}$.

The tightness $\sqrt{N}(\hat{b}_N - b_0) = O_p(1)$ can be also deduced as in Sect. 3.1.1: it suffices to note that

$$\begin{aligned}
\frac{1}{\sqrt{N}}\partial_b M_{2,N}(b_0) &= \frac{2}{\sqrt{N}}\sum_{i,j}\left(\eta_{ij} + \frac{1}{\sqrt{N}}\hat{B}_{ij}\right)\partial_b\overline{\eta}_{ij}^2(b_0) \\
&= -\frac{2}{\sqrt{N}}\sum_{i,j}\left(\mu_0\,\partial_\tau(\sigma_{ij}^2)(\tau),\; s_{ij}(\alpha_0)\right)\eta_{ij} + O_p(1) = O_p(1),
\end{aligned}$$

and that

$$\frac{1}{N}\partial_b^2 M_{2,N}(\tilde{b}_N) = o_p(1) + 2Q_{2,N}(\tau_0)$$

for every random sequence $(\tilde{b}_N)$ such that $\tilde{b}_N \xrightarrow{p} b_0$.

3.1.3 Step 3

By the explicit expression (3.8) and the $\sqrt{N}$-consistency of $(\hat{\alpha}_N^0, \hat{\beta}_N^0, \hat{\mu}_N^0, \hat{c}_N^0)$, we obtain

$$\left(\frac{1}{N}\sum_{i,j} s_{ij}^6(\hat{\alpha}_N^0)\right)\sqrt{N}(\hat{\rho}_N^0 - \rho_0)$$

$$= \frac{1}{\sqrt{N}}\sum_{i,j} s_{ij}^3(\hat{\alpha}_N^0)\left(\hat{e}_{ij}^3 - 3s_{ij}(\hat{\alpha}_N^0)\sigma_{ij}^2(\hat{\tau}_N^0)\hat{c}_N^0 - s_{ij}^3(\hat{\alpha}_N^0)\rho_0\right)$$

$$= O_p(1) + \frac{1}{\sqrt{N}}\sum_{i,j} s_{ij}^3(\hat{\alpha}_N^0)\left\{\left(\overline{e}_{ij} - \frac{1}{\sqrt{N}}\hat{H}_{ij}\right)^3 - 3s_{ij}(\alpha_0)\sigma_{ij}^2(\tau_0)c_0 - s_{ij}^3(\alpha_0)\rho_0\right\}$$

$$= O_p(1) + \frac{1}{\sqrt{N}}\sum_{i,j} s_{ij}^3(\alpha_0)\left(\overline{e}_{ij}^3 - 3s_{ij}(\alpha_0)\sigma_{ij}^2(\tau_0)c_0 - s_{ij}^3(\alpha_0)\rho_0\right) = O_p(1).$$

Hence $\sqrt{N}(\hat{\rho}_N^0 - \rho_0) = O_p(1)$ under Assumption 3.1(3).

We end this section with a few remarks.

Remark 3.2 As an alternative to (3.5), one could also use the profile least-squares estimator (Richards 1961): first, we construct the explicit least-squares estimator of (β, μ) knowing α, and then optimize $\alpha \mapsto M_{1,N}(\alpha, \hat{\beta}_N(\alpha), \hat{\mu}_N(\alpha))$ to get an estimator of α.

Remark 3.3 If one component of $\theta' = (\lambda, \delta, \gamma)$ is known from the very beginning, then it is enough to look at the estimation of (μ, c) and we can remove Assumption 3.1(3) with modifying Assumption 3.1(4).

Remark 3.4 Because of the asymptotic nature, the same flow of estimation procedures (the MLE, the initial estimator, and the one-step estimator) remain valid even if we replace the trend term $x_{ij}^\top \beta$ in (1.2) by some nonlinear one, say $\mu(x_{ij}, \beta)$, with associated identifiability conditions.

Remark 3.5 We can construct a one-step estimator for the MELS model (1.2) in a similar manner to Steps 1 to 3 described in Sect. 3.1. To construct an initial estimator $\hat{\theta}_N^0 = (\hat{\beta}_N^0, \hat{\alpha}_N^0, \hat{\tau}_N^0, \hat{\sigma}_w^{2,0}, \hat{\rho}_N^0)$, we use the identities $E_\theta[Y_{ij}] = x_{ij}^\top \beta$, $\text{Var}_\theta[Y_{ij}] = \exp(w_{ij}^\top \tau + \sigma_w^2/2) + \exp(z_i^\top \alpha)$, and $E_\theta[(Y_{ij} - E_\theta[Y_{ij}])^3] = 3\sigma_w \exp(z_i^\top \alpha/2 + \sigma_w^2/2)\rho$. Then, we can obtain $\hat{\beta}_N^0$ in Step 1, $(\hat{\alpha}_N^0, \hat{\tau}_N^0, \hat{\sigma}_{w,N}^{2,0})$ in Step 2, and then $\hat{\rho}_N^0$ in Step 3 in

this order through the contrast functions to be minimized: denoting $\hat{e}'_{ij} := Y_{ij} - x_{ij}^\top \hat{\beta}_N^0$, we have

$$\beta \mapsto \sum_{i,j} \left(Y_{ij} - x_{ij}^\top \beta \right)^2,$$

$$(\alpha, \tau, \sigma_w^2) \mapsto \sum_{i,j} \left(\hat{e}'^2_{ij} - \exp(w_{ij}^\top \tau + \sigma_w^2/2) - \exp(z_i^\top \alpha) \right)^2,$$

$$\rho \mapsto \sum_{i,j} \left(\hat{e}'^3_{ij} - 3\sqrt{\hat{\sigma}^{2,0}_{w,N}} \exp(z_{ij}^\top \hat{\alpha}_N^0/2 + \hat{\sigma}^{2,0}_{w,N}/2)\rho \right)^2.$$

As in the case of (3.8), $\hat{\rho}_N^0$ is explicitly given while the meaning of the parameter ρ is different in the present context. It is also possible to develop an asymptotic theory for the MLE of the MELS and the related one-step estimator in similar ways to the present study. However, the one-step estimator toward the log-likelihood function (1.3) still necessitates the numerical integration over $\mathbb{R}^2$ with respect to the two-dimensional standard normal random variables; the numerical integration would need to be performed for every $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$, hence the computational load would still be significant.

3.2 Numerical experiments

Let us observe the finite-sample performance of the initial estimator $\hat{\theta}_N^0$, the one-step estimator $\hat{\theta}_N^1$, and the MLE $\hat{\theta}_N$. The setting is as follows:

$$Y_{ij} = x_{ij}^\top \beta + \tanh(z_{ij}^\top \alpha) v_i + \sqrt{v_i \exp(w_{ij}^\top \tau)} \, \epsilon_{ij}, \tag{3.10}$$

where

- $N = 1000, \quad n_1 = n_2 = \cdots = n_N = 10$.
- $x_{ij}, z_{ij}, w_{ij} \in \mathbb{R}^2 \sim$ i.i.d. $N_2(0, I_2)$.
- $v_1, v_2, \ldots \sim$ i.i.d. $IG(\delta, \gamma) = GIG(-1/2, \delta, \gamma)$, the inverse-Gaussian random-effect distribution.
- $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i}) \sim$ i.i.d. $N(0, I_{n_i})$, independent of $\{v_i\}$.
- $\theta = (\beta, \alpha, \tau, \delta, \gamma) = (\beta_0, \beta_1, \alpha_0, \alpha_1, \tau_0, \tau_1, \delta, \gamma) \in \mathbb{R}^8$.
- True values are $\beta = (3, 5), \ \alpha = (-4, 5), \ \tau = (0.05, 0.07), \ \delta = 1.5, \ \gamma = 0.7$.

In this case $\theta' = (\delta, \gamma) \in (0, \infty)^2$ and we need only $(\hat{\mu}_N^0, \hat{c}_N^0)$: we have $\mu = E_{\theta'}[v_i] = \delta/\gamma$ and $c = \text{Var}_{\theta'}[v_i] = \delta/\gamma^3$, namely

$$\gamma = \sqrt{\frac{\mu}{c}}, \qquad \delta = \mu\gamma = \sqrt{\frac{\mu^3}{c}}.$$

As initial values for numerical optimization, we set the following two different cases:

Table 4 The empirical 95%-coverage probabilities of the MLE and the one-step estimators in cases (i') and (ii') based on 1000 trials; MLE of (δ, γ) in case (ii') showed instability in numerical optimizations, while the one-step estimator is stable as in case (i')

		β_0	β_1	α_0	α_1	τ_0	τ_1	δ	γ
Case (i')	$\hat{\theta}_N$	0.960	0.959	0.944	0.942	0.957	0.947	0.948	0.943
	$\hat{\theta}_N^1$	0.960	0.958	0.943	0.943	0.952	0.949	0.943	0.947
Case (ii')	$\hat{\theta}_N$	0.957	0.953	0.917	0.919	0.956	0.944	0	0
	$\hat{\theta}_N^1$	0.960	0.959	0.943	0.943	0.952	0.949	0.943	0.947

(i') The true value;
(ii') $(1.0 \times 10^{-8}, \ldots, 1.0 \times 10^{-8}, 1.0 \times 10^{-4}, 1.0 \times 10^{-3})$.

In each case, we computed $\sqrt{N}(\hat{\xi}_N - \theta_0)$ for $\hat{\xi}_N = \hat{\theta}_N^0, \hat{\theta}_N^1$, and $\hat{\theta}_N$, all being conducted 1000-times Monte Carlo trials. To estimate 95%-coverage probabilities empirically as in Sect. 2.3, we computed the quantities $-\partial_\theta^2 \ell_N(\hat{\theta}_N)$ and $-\partial_\theta^2 \ell_N(\hat{\theta}_N^1)$ through the function $\theta \mapsto -\partial_\theta^2 \ell_N(\theta)$ for the approximately 95%-confidence intervals for each parameter. The results are shown in Table 4; therein, we obtained numerically unstable 4 MLEs and 5 one-step estimators for case (i') and 299 MLEs and 6 one-step estimators for case (ii'), and then computed the coverage probabilities based on the remaining cases. In Figs. 4 and 5 (for cases (i') and (ii'), respectively), we drew histograms of $\hat{\theta}_N^1$ and $\hat{\theta}_N$ together with those of the initial estimator $\hat{\theta}_N^0$ for comparison. In each figure, the histograms in the first and fourth columns are those for $\hat{\theta}_N^0$, those in the second and fifth columns for $\hat{\theta}_N^1$, and those in the third and sixth columns for $\hat{\theta}_N$, respectively; the red solid line shows the zero-mean normal densities with the consistently estimated Fisher information for the variances.

Here is a summary of the important findings.

- Approximate computation times for obtaining one set of estimates are as follows:

(i') 0.2 s for $\hat{\theta}_N^0$; 10 s for $\hat{\theta}_N^1$; 2 min for $\hat{\theta}_N$;
(ii') 0.2 s for $\hat{\theta}_N^0$; 10 s for $\hat{\theta}_N^1$; 9 min for $\hat{\theta}_N$.

A considerable amount of reduction can be seen for $\hat{\theta}_N^1$ compared with $\hat{\theta}_N$.

- About Figs. 4 and 5:

- In both cases (i') and (ii'), the inferior performance of $\hat{\theta}_N^0$ is drastically improved by $\hat{\theta}_N^1$, which in turn shows asymptotically equivalent behaviors to the MLE $\hat{\theta}_N$.
- On one hand, as in Sect. 2.3, the MLE $\hat{\theta}_N$ is much affected by the initial value for the numerical optimization, partly because of the non-convexity of the likelihood function $\ell_N(\theta)$; in Case (ii'), we observed the instability in computing the MLE of (δ, γ) (in the bottom panels in Fig. 5), showing the local maxima problem. On the other hand, we did not observe the local maxima problem in computing $\hat{\theta}_N^0$ and the one-step estimator $\hat{\theta}_N^1$ does not require an initial value for numerical optimization.
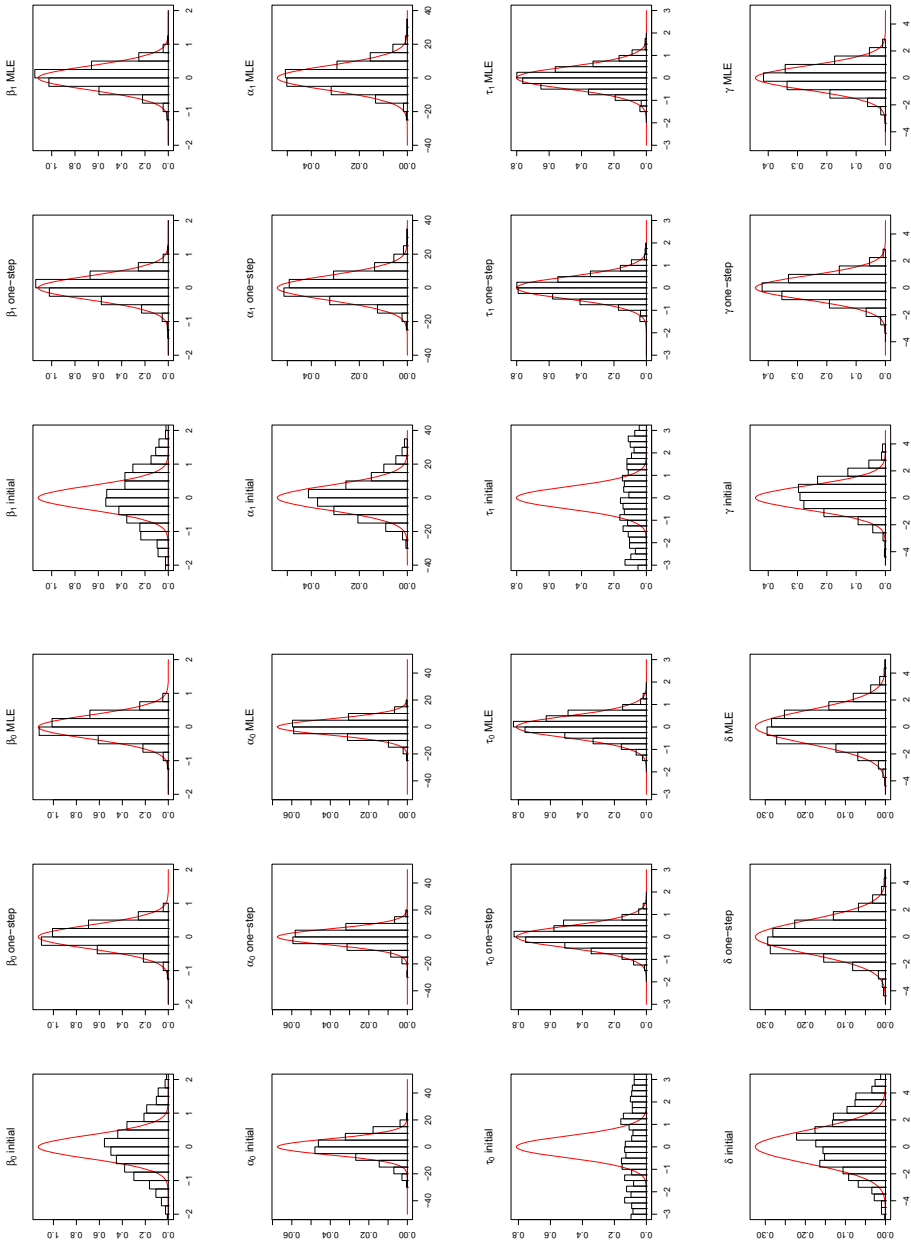
Fig. 4 Case (i'): Histograms of the initial estimator $\hat{\theta}_N^0$ (first and fourth columns), the one-step estimator $\hat{\theta}_N^1$ (second and fifth columns), and the MLE $\hat{\theta}_N$ (third and sixth columns). In each histogram panel, the solid red line shows the estimated asymptotically best possible normal distribution
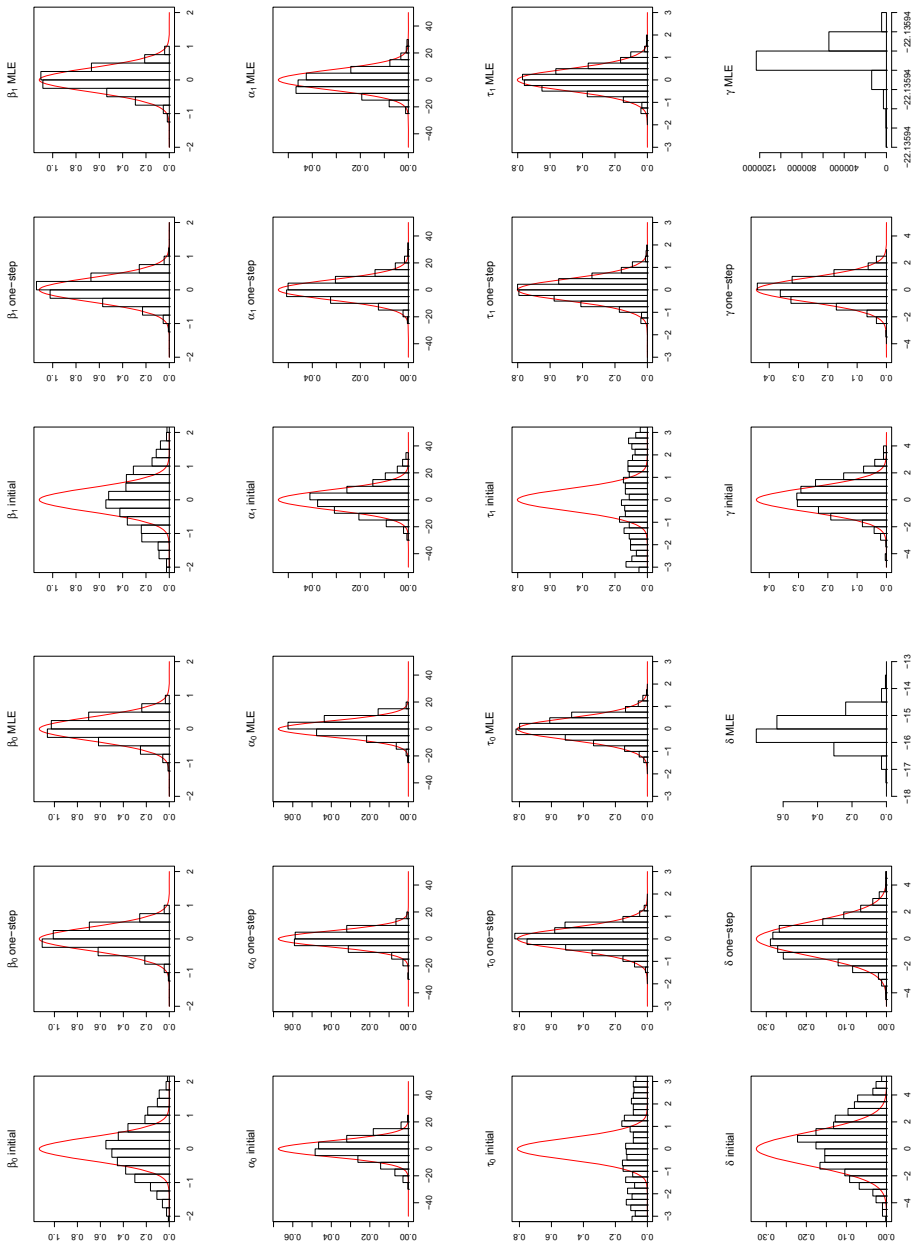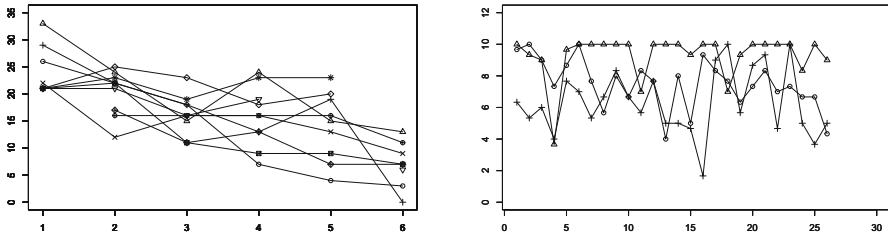
Fig. 5 Case (ii'): Histograms of the initial estimator $\hat{\theta}_N^0$ (first and fourth columns), the one-step estimator $\hat{\theta}_N^1$ (second and fifth columns), and the MLE $\hat{\theta}_N$ (third and sixth columns). In each histogram panel, the solid red line shows the estimated asymptotically best possible normal distribution

Fig. 6 Data plots of `riesby_example.dat` (left) and `posmood_example.dat` (right) borrowed from the supplemental material of Hedeker and Nordgren (2013); the former shows data of 10 patients over 6 time points with a few missing values, and the latter does those of 3 people over 26 time points with no missing value

In sum, $\hat{\theta}_N^1$ is asymptotically equivalent to the efficient MLE and much more robust in numerical optimization than the MLE. It is recommended to use the one-step estimator $\hat{\theta}_N^1$ against the MLE $\hat{\theta}_N$ from both theoretical and computational points of view.

We end this section with applications of the proposed one-step estimator $\hat{\theta}_N^1$ for (3.10) to the two real data sets `riesby_example.dat` and `posmood_example.dat` borrowed from the supplemental material of Hedeker and Nordgren (2013). Here are brief descriptions.

- `riesby_example.dat` contains the Hamiltonian depression rating scale as Y_{ij}. The covariates are given by $x_{ij} = (\text{intercept}, \text{week}, \text{edog}) \in \mathbb{R} \times \{0, 1, 2, \ldots, 5\} \times \{0, 1\}$, $z_{ij} = (\text{intercept}, \text{edog})$, and $w_{ij} = (\text{intercept}, \text{week})$. Here, $N = 66$ and the numbers of sampling times are 6 with a few missing slots, and `edog` denotes the dummy variable for indicating whether the depression of the patient is endogenous ($= 1$) or not ($= 0$).
- `posmood_example.dat` contains the individual mood items as Y_{ij}; the items are pre-processed using factor analysis and take values 1 to 10 with higher ones indicating a higher level of positive mood. The covariates are given by $x_{ij} = (\text{intercept}, \text{alone}, \text{genderf}) \in \mathbb{R} \times \{0, 1\} \times \{0, 1\}$, $z_{ij} = (\text{intercept}, \text{alone})$, and $w_{ij} = (\text{intercept}, \text{alone})$. Here, $N = 515$ with no missing value, with approximately 34 sampling times on average (ranging from 3 to 58). The variable `alone` and `genderf` respectively denote the dummy variables for indicating whether the person is alone ($= 0$) or not ($= 1$), which is time-varying, and whether the person is male ($= 0$) or female ($= 1$).

Figures 6 and 7 show some data plots and histograms, respectively; the former is positively skewed while the latter is negatively skewed. We could apply our one-step estimation methods for these data sets, although they can be seen as categorical data (with a moderately large number of categories). The results are given in Table 5; the parameters β_0, α_0, and τ_0 denote the `intercept`. The skewness mentioned above is reflected in the estimates of α_0 and α_1.
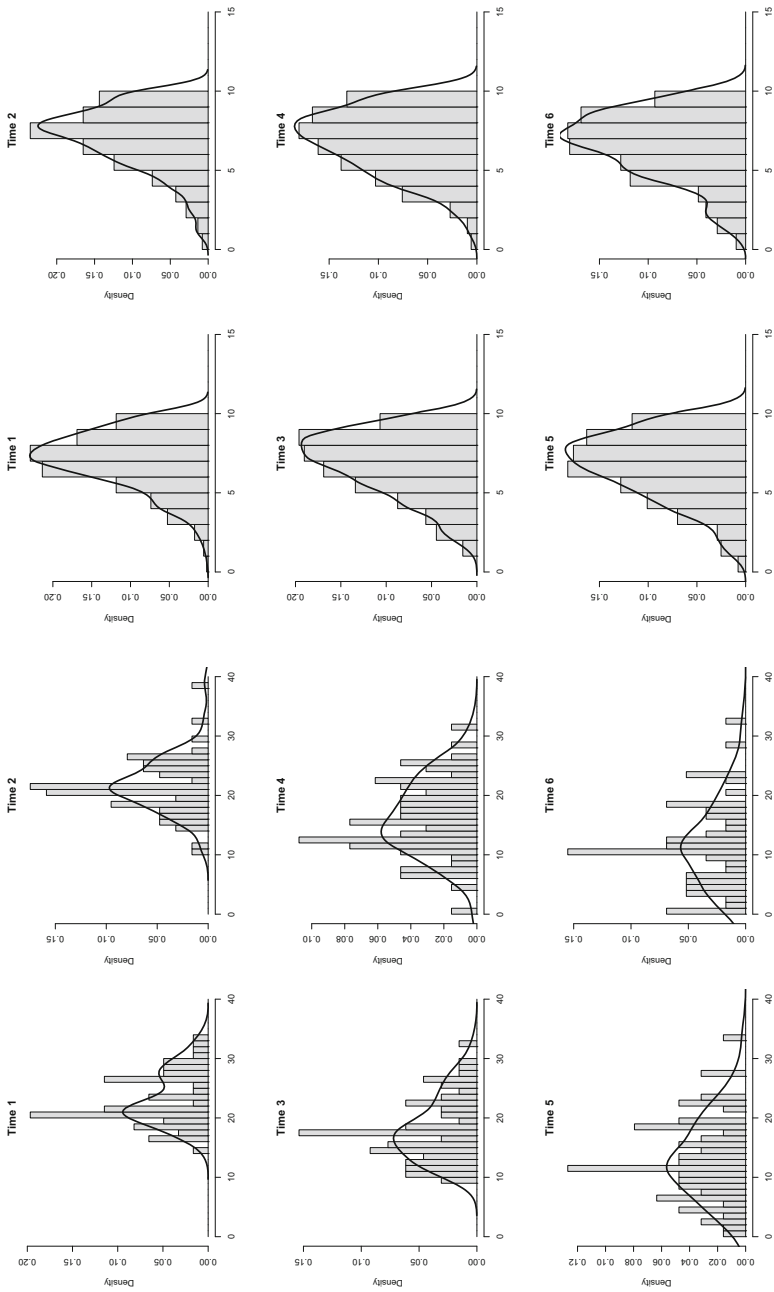
Fig. 7 Histograms of data $(Y_{i1})_{i \leq N}, \ldots, (Y_{i6})_{i \leq N}$ for `riesby_example.dat` (left) and `posmood_example.dat` (right) borrowed from the supplemental material of Hedeker and Nordgren (2013)

Table 5 One-step estimates for the two data sets `riesby_example.dat` and `posmood_example.dat`. It took 0.9 and 6.7 s, respectively

`riesby_example.dat`	β_0	β_1	β_2	α_0	α_1	τ_0	τ_1	δ	γ
	16.930	− 2.443	− 0.833	1.442	5.961	− 0.213	0.396	2.032	0.245
`posmood_example.dat`	β_0	β_1	β_2	α_0	α_1	τ_0	τ_1	δ	γ
	6.608	− 0.229	− 0.295	− 0.125	− 0.016	− 0.332	0.093	4.517	0.606

4 Concluding remarks

We proposed a class of mixed-effects models with non-Gaussian marginal distributions which can incorporate random effects into the skewness and the scale simply and transparently through the normal variance-mean mixture. The associated log-likelihood function is explicit and the MLE is asymptotically efficient (Remark 2.4) while computationally demanding and unstable. To bypass the numerical issue, we proposed the easy-to-use one-step estimator $\hat{\theta}_N^1$, which turned out to not only attain a significant reduction of computation time compared with the MLE but also guarantee the asymptotic efficiency property.

Here are some remarks on important related issues.

(1) *Inter-individual dependence structure.* A drawback of the model (2.1) is that its inter-individual dependence structure is not flexible enough. Specifically, let us again note the following covariance structure for $j, k \leq n_i$:

$$\mathrm{Cov}_\theta[Y_{ij}, Y_{ik}] = s_{ij}(\alpha)s_{ik}(\alpha)\mathrm{Var}_\theta[v_i] = c(\theta')s_{ij}(\alpha)s_{ik}(\alpha).$$

This in particular implies that $Y_{i1}, \ldots, Y_{in_i}$ cannot be correlated as long as $s(z, \alpha) \equiv 0$. Nevertheless, it is formally straightforward to extend the model (2.1) so that the distributional structure of $Y_i \in \mathbb{R}^{n_i}$ obeys the multivariate GH distribution for each $\mathcal{L}(Y_i)$ with a *non-diagonal* scale matrix. To mention it briefly, suppose that the vector of a sample $Y_i = (Y_{i1}, \ldots, Y_{in_i}) \in \mathbb{R}^{n_i}$ from ith individual is given by the form

$$Y_i = x_i\beta + s(z_i, \alpha)v_i + \Lambda(w_i, \tau)^{1/2}\sqrt{v_i}\,\epsilon_i.$$

Here, $v_1, \ldots, v_N \sim$ i.i.d. $GIG(\lambda, \delta, \gamma)$ as before, while we now incorporated the scale matrix $\Lambda(w_i, \tau)$ which should be positive definite and symmetric, but may be *non-diagonal*. Then, the dependence structure of $Y_{i1}, \ldots, Y_{in_i}$ can be much more flexible than (2.1).

(2) *Forecasting random-effect parameters.* In the familiar Gaussian linear mixed-effects model of the form $Y_i = X_i\beta + Z_ib_i + \epsilon_i$, the empirical Bayes predictor of v_i is given by $\hat{b}_i := E_\theta[b_i|Y_i]|_{\theta=\hat{\theta}_N}$. One of the analytical merits of our NVMM framework is that the conditional distribution $\mathcal{L}(v_i|Y_i = y_i)$ of v_i is given by $GIG(v_i, \eta_i, \psi_i)$, where

$$v_i = v_i(\theta) := \lambda - \frac{n_i}{2},$$

$$\eta_i = \eta_i(\theta) := \sqrt{\delta^2 + (y_i - x_i\beta)^\top \Lambda(w, \tau)^{-1}(y_i - x_i\beta)},$$

$$\psi_i = \psi_i(\theta) := \sqrt{\gamma^2 + s_i(\alpha)^\top \Lambda(w_i, \tau)^{-1}s_i(\alpha)}.$$

This is a direct consequence of the general results about the multivariate GH distribution; see Eberlein and Hammerstein (2004) and the references therein for details. As in the Gaussian case mentioned above, we can make use of

$$\hat{v}_i := E_\theta[v_i|Y_i = y_i]|_{\theta=\hat{\theta}_N} = \frac{K_{\hat{v}_i+1}(\hat{\eta}_i\hat{\psi}_i)}{K_{\hat{v}_i}(\hat{\eta}_i\hat{\psi}_i)}\frac{\hat{\eta}_i}{\hat{\psi}_i},$$

where $\hat{v}_i := v_i(\hat{\theta}_N)$, $\hat{\eta}_i := \eta_i(\hat{\theta}_N)$, and $\hat{\psi}_i := \psi_i(\hat{\theta}_N)$; formally $\hat{\theta}_N$ could be replaced by the one-step estimator $\hat{\theta}_N^1$. Then, it would be natural to regard

$$\hat{Y}_{ij} := x'_{ij}\hat{\beta}_N + s(z'_{ij}, \hat{\alpha}_N)\hat{v}_i$$

as a prediction value of Y_{ij} at (x'_{ij}, z'_{ij}). This includes forecasting the value of ith individual at a future time point.

(3) *Lack of fit and model selection.* In relation to Remark 2.6, based on the obtained asymptotic-normality results, we can proceed with lack-of-fit tests, such as the likelihood-ratio test, the score test, and the Wald test; typical forms are $s(z, \alpha) = \sum_{l=1}^{p_\alpha} \alpha_l s_l(z)$ and $\sigma(w, \tau) = \exp\{\sum_{m=1}^{p_\tau} \tau_m \sigma_m(w)\}$, with given basis functions $s_l(z)$ and $\sigma_m(w)$. In that case, we can estimate p-value for each component of θ, say, by $2\Phi(-|\hat{B}_{k,N}\hat{\theta}_{k,N}|)$ for θ_k where $\hat{B}_{k,N} := [(-\partial_\theta^2 \ell_N(\hat{\theta}_N))^{-1}]_{kk}^{-1/2}$. Alternatively, one may consider information criteria such as the conditional AIC (Vaida and Blanchard 2005) and the BIC-type one (Delattre et al. 2014). To develop these devices in rigorous ways, we will need to derive several further analytical results: the uniform integrability of $(\|\sqrt{N}(\hat{\theta}_N - \theta_0)\|^2)_n$ for the AIC, the stochastic expansion for the marginal likelihood function for the BIC, and so on.

Data Availability   We have used the real data of Hedeker and Nordgren (2013) available online.

Declarations

Conflict of interest   The author declares that there is no conflict of interest.

Appendix A: GIG and GH distributions

Let $K_\lambda(t)$ denote the modified Bessel function of the second kind ($\nu \in \mathbb{R}, t > 0$):

$$K_\nu(t) = \frac{1}{2} \int_0^\infty s^{\nu-1} \exp\left\{-\frac{t}{2}\left(s + \frac{1}{s}\right)\right\} ds.$$

We have the following recurrence formulae (Abramowitz and Stegun 1992): $K_{\nu+1}(t) = \frac{2\nu}{t} K_\nu(t) + K_{\nu-1}(t)$ and $K_{\nu-1}(t) + K_{\nu+1}(t) = -2\partial_t K_\nu(t)$. It follows that $K_\nu(t)$ is monotonically decreasing and that $\partial_t K_\nu(t) = -K_{\nu-1}(t) - \frac{\nu}{t} K_\nu(t)$. Further, we have

$$\partial_t \log K_\nu(t) = -\frac{K_{\nu-1}(t)}{K_\nu(t)} - \frac{\nu}{t} =: -R_\nu(t) - \frac{\nu}{t}, \tag{A.1}$$

$$\partial_t^2 \log K_\nu(t) = -\frac{1}{K_\nu^2(t)}\left(K_{\nu-1}^2(t) - K_{\nu-2}(t)K_\nu(t)\right) - \frac{1}{t}\frac{K_{\nu-1}(t)}{K_\nu(t)} + \frac{\nu}{t^2}$$

$$=: -S_\nu(t) - \frac{1}{t}R_\nu(t) + \frac{\nu}{t^2}. \tag{A.2}$$

The following asymptotic behavior holds:

$$K_\nu(t) = \sqrt{\frac{\pi}{2t}} \exp(-t)\{1 + (4\nu^2 - 1)O(t^{-1})\}, \qquad t \to \infty.$$

The generalized inverse Gaussian (GIG) distribution $GIG(\lambda, \delta, \gamma)$ on $\mathbb{R}_+$ is defined by the density:

$$p_{GIG}(z; \lambda, \delta, \gamma) = \frac{(\gamma/\delta)^\lambda}{2K_\lambda(\gamma\delta)} z^{\lambda-1} \exp\left\{-\frac{1}{2}\left(\frac{\delta^2}{z} + \gamma^2 z\right)\right\}, \qquad z > 0.$$

The region of admissible parameters is given by the union of $\{(\lambda, \delta, \gamma) : \lambda > 0, \delta \geq 0, \gamma > 0\}$, $\{(\lambda, \delta, \gamma) : \lambda = 0, \delta > 0, \gamma > 0\}$, and $\{(\lambda, \delta, \gamma) : \lambda < 0, \delta > 0, \gamma > 0\}$, according to the integrability of p_{GIG} at the origin and $+\infty$.

The generalized hyperbolic (GH) distribution denoted by $GH(\lambda, \alpha, \beta, \delta, \mu)$ is defined as the distribution of the normal variance-mean mixture Y with respect to $Z \sim GIG(\lambda, \delta, \gamma)$:

$$Y = \mu + \beta Z + \sqrt{Z}\eta,$$

where $\alpha := \sqrt{\beta^2 + \gamma^2}$ and $\eta \sim N(0, 1)$ independent of Z. By the conditional Gaussianity $\mathcal{L}(Y|Z = z) = N(\mu + \beta z, z)$, the density is calculated as follows:

$$p_{GH}(y; \lambda, \alpha, \beta, \delta, \mu)$$
$$= \int_0^\infty \frac{1}{\sqrt{2\pi z}} \exp\left(-\frac{1}{2z}(y - \mu - \beta z)^2\right) p_{GIG}(z; \lambda, \delta, \gamma)dz$$

$$= \frac{(\alpha^2 - \beta^2)^{\lambda/2} \sqrt{\delta^2 + (y - \mu)^2}^{\lambda - 1/2}}{\sqrt{2\pi} \alpha^{\lambda - 1/2} \delta^\lambda K_\lambda \left(\delta\sqrt{\alpha^2 - \beta^2}\right)} K_{\lambda - \frac{1}{2}} \left(\alpha\sqrt{\delta^2 + (y - \mu)^2}\right) \exp[\beta(y - \mu)].$$

The region of admissible parameters is given by the union of $\{(\lambda, \alpha, \beta, \delta, \mu) : \lambda > 0, \delta \geq 0, \alpha > |\beta|\}$, $\{(\lambda, \alpha, \beta, \delta, \mu) : \lambda = 0, \delta > 0, \alpha > |\beta|\}$, and $\{(\lambda, \alpha, \beta, \delta, \mu) : \lambda < 0, \delta > 0, \alpha \geq |\beta|\}$. The mean and variance of $Y \sim GH(\lambda, \alpha, \beta, \delta, \mu)$ are given by

$$E[Y] = \mu + \frac{\delta\beta K_{\lambda+1}(\delta\gamma)}{\gamma K_\lambda(\delta\gamma)},$$

$$\text{Var}[Y] = \frac{\delta K_{\lambda+1}(\delta\gamma)}{\gamma K_\lambda(\delta\gamma)} + \frac{\beta^2 \delta^2}{\gamma^2} \left[\frac{K_{\lambda+2}(\delta\gamma)}{K_\lambda(\delta\gamma)} - \left(\frac{K_{\lambda+1}(\delta\gamma)}{K_\lambda(\delta\gamma)}\right)^2\right].$$

See Eberlein and Hammerstein (2004) for further details of the GIG and GH distributions.

The normal inverse Gaussian (NIG) distribution is one of the popular subclasses of the GH-distribution family: $NIG(\alpha, \beta, \delta, \mu) := GH(-1/2, \alpha, \beta, \delta, \mu)$, where $GIG(-1/2, \delta, \gamma)$ corresponds to the inverse Gaussian distribution. The $NIG(\alpha, \beta, \delta, \mu)$-density is given by

$$p_{NIG}(x; \alpha, \beta, \delta, \mu) = \frac{\alpha\delta}{\pi} \exp\left(\delta\sqrt{\alpha^2 - \beta^2} + \beta(x - \mu)\right) \frac{K_1\left(\alpha\sqrt{\delta^2 + (x - \mu)^2}\right)}{\sqrt{\delta^2 + (x - \mu)^2}}.$$

All of the mean M, variance V, skewness S, and kurtosis K of $NIG(\alpha, \beta, \delta, \mu)$ are explicitly given:

$$M = \mu + \frac{\beta\delta}{(\alpha^2 - \beta^2)^{\frac{1}{2}}}, \quad V = \frac{\delta\alpha^2}{(\alpha^2 - \beta^2)^{\frac{3}{2}}}, \quad S = \frac{3\beta}{\alpha\sqrt{\delta}(\alpha^2 - \beta^2)^{\frac{1}{4}}},$$

$$K = \frac{3\alpha^2 + 4\beta^2}{\alpha^2 \delta(\alpha^2 - \beta^2)^{\frac{1}{2}}}.$$

Inverting these expressions gives

$$\gamma = \frac{3}{\sqrt{V}\sqrt{3K - 5S^2}}, \quad \beta = \frac{S\sqrt{V}\gamma^2}{3}, \quad \alpha = \sqrt{\gamma^2 + \beta^2}, \quad \delta = \frac{V\gamma^3}{\gamma^2 + \beta^2},$$

$$\mu = M - \frac{\beta\delta}{\gamma},$$

from which one can consider the method-of-moments estimation of $(\alpha, \beta, \delta, \mu)$ based on the empirical counterparts of M, V, S, and K. One should note that the empirical quantity $3\hat{K}_n - 5\hat{S}_n^2$ has to be positive, which may fail in a finite sample and for such a data set the MLE would be also non-computable or unstable. In Yoon et al. (2020), the

estimation problem for the i.i.d. NIG model was studied from the computational point of view; the paper also introduced the change of variables for the parameters to sidestep the positivity restriction, resulting in stabilized results in numerical experiments.

Appendix B: Likelihood function

B.1 Derivation

Writing $\theta_1 = (\beta, \alpha, \tau)$ and $\theta_2 = (\lambda, \delta, \gamma)$, and using the obvious notation, we obtain

$$
\begin{aligned}
\ell_n(\theta) &= \log p_\theta(Y_1, \dots, Y_n) \\
&= \log \int \cdots \int p_{\theta_1}(Y_1, Y_2, \dots, Y_n | v_1, \dots, v_n) \prod_{i=1}^{N} p_{\theta_2}(v_i) dv_i \\
&= \log \left(\int \cdots \int \prod_{i=1}^{N} p_{\theta_1}(Y_i | v_i) \prod_{i=1}^{N} p_{\theta_2}(v_i) dv_i \right) \\
&= \sum_{i=1}^{N} \log \left[\int \left(\prod_{j=1}^{n_i} p_{\theta_1}(Y_{ij} | v_i) \right) p_{\theta_2}(v_i) dv_i \right] \\
&= \sum_{i=1}^{N} \log \left[\int \left(\prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi \sigma_{ij}^2(\tau)}} v_i^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_{ij}^2(\tau)} (Y_{ij} - x_{ij}^\top \beta - s_{ij}(\alpha) v_i)^2 \right] \right) p_{\theta_2}(v_i) dv_i \right] \\
&= \sum_{i=1}^{N} \log \left[(2\pi)^{-\frac{n_i}{2}} \left(\prod_{j=1}^{n_i} \sigma_{ij}^2(\tau) \right)^{-\frac{1}{2}} \int_0^\infty v_i^{-\frac{n_i}{2}} \right. \\
&\quad \times \left. \prod_{j=1}^{n_i} \exp\left[-\frac{1}{2 v_i \sigma_{ij}^2(\tau)} \left\{ (Y_{ij} - x_{ij}^\top \beta)^2 + s_{ij}^2(\alpha) v_i^2 - 2 s_{ij}(\alpha) v_i (Y_{ij} - x_{ij}^\top \beta) \right\} \right] p_{\theta_2}(v_i) dv_i \right] \\
&= \sum_{i=1}^{N} \log \left[(2\pi)^{-\frac{n_i}{2}} \left(\prod_{j=1}^{n_i} \sigma_{ij}^2(\tau) \right)^{-\frac{1}{2}} \int_0^\infty v_i^{-\frac{n_i}{2}} \right. \\
&\quad \times \left. \prod_{j=1}^{n_i} \exp\left[-\frac{1}{2 v_i \sigma_{ij}^2(\tau)} \left\{ (Y_{ij} - x_{ij}^\top \beta)^2 + s_{ij}^2(\alpha) v_i^2 \right\} + \frac{s_{ij}(\alpha)}{\sigma_{ij}^2(\tau)} (Y_{ij} - x_{ij}^\top \beta) \right] p_{\theta_2}(v_i) dv_i \right] \\
&= \sum_{i=1}^{N} \log \left[(2\pi)^{-\frac{n_i}{2}} \left(\prod_{j=1}^{n_i} \sigma_{ij}^2(\tau) \right)^{-\frac{1}{2}} \prod_{j=1}^{n_i} \exp\left(\frac{s_{ij}(\alpha)}{\sigma_{ij}^2(\tau)} (Y_{ij} - x_{ij}^\top \beta) \right) \right. \\
&\quad \times \left. \int_0^\infty v_i^{-\frac{n_i}{2}} \prod_{j=1}^{n_i} \exp\left\{ -\frac{1}{2\sigma_{ij}^2(\tau)} \left(\frac{(Y_{ij} - x_{ij}^\top \beta)^2}{v_i} + s_{ij}^2(\alpha) v_i \right) \right\} p_{\theta_2}(v_i) dv_i \right] \\
&= \sum_{i=1}^{N} \log \left[(2\pi)^{-\frac{n_i}{2}} \left(\prod_{j=1}^{n_i} \sigma_{ij}^2(\tau) \right)^{-\frac{1}{2}} \exp\left(\sum_{j=1}^{n_i} \frac{s_{ij}(\alpha)(Y_{ij} - x_{ij}^\top \beta)}{\sigma_{ij}^2(\tau)} \right) \right. \\
&\quad \times \left. \int_0^\infty v_i^{-\frac{n_i}{2}} \exp\left\{ -\frac{1}{2} \left(\sum_{j=1}^{n_i} \frac{(Y_{ij} - x_{ij}^\top \beta)^2}{v_i \sigma_{ij}^2(\tau)} + \sum_{j=1}^{n_i} \frac{s_{ij}^2(\alpha)}{\sigma_{ij}^2(\tau)} v_i \right) \right\} p_{\theta_2}(v_i) dv_i \right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \log \Bigg[C_i(\alpha, \beta, \tau) \int_0^\infty v_i^{-\frac{n_i}{2}} \exp \Bigg\{ -\frac{1}{2} \Bigg(\sum_{j=1}^{n_i} \frac{(Y_{ij} - x_{ij}^\top \beta)^2}{v_i \sigma_{ij}^2(\tau)} + \sum_{j=1}^{n_i} \frac{s_{ij}^2(\alpha)}{\sigma_{ij}^2(\tau)} v_i \Bigg) \Bigg\} \\
&\qquad \times \frac{(\frac{\gamma}{\delta})^\lambda}{2 K_\lambda(\delta\gamma)} v_i^{\lambda-1} \exp \Bigg\{ -\frac{1}{2} \Bigg(\frac{\delta^2}{v_i} + \gamma^2 v_i \Bigg) \Bigg\} dv_i \Bigg] \\
&= \sum_{i=1}^{N} \log \Bigg[C_i(\alpha, \beta, \tau) \frac{(\gamma/\delta)^\lambda}{2 K_\lambda(\delta\gamma)} \int_0^\infty \exp \Bigg[-\frac{1}{2} \Bigg\{ \frac{1}{v_i} \Bigg(\sum_{j=1}^{n_i} \frac{(Y_{ij} - x_{ij}^\top \beta)^2}{\sigma_{ij}^2(\tau)} + \delta^2 \Bigg) \\
&\qquad + \Bigg(\sum_{j=1}^{n_i} \frac{s_{ij}^2(\alpha)}{\sigma_{ij}^2(\tau)} + \gamma^2 \Bigg) v_i \Bigg\} \Bigg] v_i^{\lambda-1-\frac{n_i}{2}} dv_i \Bigg],
\end{aligned}
$$

where

$$
C_i(\alpha, \beta, \tau) := (2\pi)^{-\frac{n_i}{2}} \Bigg(\prod_{j=1}^{n_i} \sigma_{ij}^2(\tau) \Bigg)^{-\frac{1}{2}} \exp \Bigg(\sum_{j=1}^{n_i} \frac{s_{ij}(\alpha)(Y_{ij} - x_{ij}^\top \beta)}{\sigma_{ij}^2(\tau)} \Bigg).
$$

Making the change of variables $S_i^2 v_i / T_i = u_i$ with

$$
S_i = S_i(\alpha, \tau, \gamma) := \sqrt{\gamma^2 + \sum_{j=1}^{n_i} \frac{s_{ij}^2(\alpha)}{\sigma_{ij}^2(\tau)}},
$$

$$
T_i = T_i(\beta, \alpha, \tau, \delta, \gamma) := S_i \sqrt{\delta^2 + \sum_{j=1}^{n_i} \frac{1}{\sigma_{ij}^2(\tau)} (Y_{ij} - x_{ij}^\top \beta)^2},
$$

we can continue as

$$
\begin{aligned}
\ell_n(\theta) &= \sum_{i=1}^{N} \log \Bigg[C_i(\alpha, \beta, \tau) \frac{(\frac{\gamma}{\delta})^\lambda}{2 K_\lambda(\delta\gamma)} \int_0^\infty \exp \Bigg\{ \\
&\qquad -\frac{1}{2} \Bigg(\frac{S_i^2}{T_i u_i} \frac{T_i^2}{S_i^2} + \frac{T_i}{S_i^2} u_i S_i^2 \Bigg) \Bigg\} \frac{T_i^{\lambda-1-\frac{n_i}{2}}}{2^{\left(\lambda-1-\frac{n_i}{2}\right)}} u_i^{\lambda-1-\frac{n_i}{2}} \frac{T_i}{S_i^2} du_i \Bigg] \\
&= \sum_{i=1}^{N} \log \Bigg[C_i(\alpha, \beta, \tau) \frac{(\frac{\gamma}{\delta})^\lambda}{2 K_\lambda(\delta\gamma)} \int_0^\infty \exp \Bigg\{ -\frac{T_i}{2} \Bigg(\frac{1}{u_i} + u_i \Bigg) \Bigg\} \\
&\qquad \frac{T_i^{\lambda-\frac{n_i}{2}}}{2^{\left(\lambda-\frac{n_i}{2}\right)} S_i} u_i^{\lambda-1-\frac{n_i}{2}} du_i \Bigg] \\
&= \sum_{i=1}^{N} \log \Bigg[(2\pi)^{-\frac{n_i}{2}} \Bigg(\prod_{j=1}^{n_i} \sigma_{ij}^2(\tau) \Bigg)^{-1/2} \frac{(\gamma/\delta)^\lambda}{K_\lambda(\delta\gamma)} \exp \Bigg(\sum_{j=1}^{n_i} \frac{s_{ij}(\alpha)}{\sigma_{ij}^2(\tau)} (Y_{ij} - x_{ij}^\top \beta) \Bigg)
\end{aligned}
$$

$$\times \frac{T_i^{\lambda - \frac{n_i}{2}}}{S_i^{2(\lambda - \frac{n_i}{2})}} \frac{1}{2} \int_0^\infty \exp\left\{-\frac{T_i}{2}\left(\frac{1}{u_i} + u_i\right)\right\} u_i^{\lambda - 1 - \frac{n_i}{2}} du_i\Bigg]$$

$$= \sum_{i=1}^N \log\Bigg[(2\pi)^{-\frac{n_i}{2}}\left(\prod_{j=1}^{n_i} \sigma_{ij}^2(\tau)\right)^{-1/2} \frac{(\gamma/\delta)^\lambda}{K_\lambda(\delta\gamma)} \exp\left(\sum_{j=1}^{n_i} \frac{s_{ij}(\alpha)}{\sigma_{ij}^2(\tau)}(Y_{ij}\right.$$

$$\left. - x_{ij}^\top \beta)\right)\left(\frac{T_i}{S_i^2}\right)^{\lambda - \frac{n_i}{2}} K_{\lambda - \frac{n_i}{2}}(T_i)\Bigg].$$

This leads to the expression (2.3).

B.2 Partial derivatives

Recall the notation: $\ell_N(\theta) = \sum_{i=1}^N \zeta_i(\theta)$, $A_i = A_i(\alpha, \tau, \gamma)$ of (2.4), and $B_i = B_i(\beta, \tau, \delta)$ of (2.5). Let

$$T_i' = T_i'(\theta) := R_{\lambda - \frac{n_i}{2}}(A_i B_i) + \frac{\lambda - \frac{n_i}{2}}{A_i B_i}$$

for $R_\nu(t)$ defined by (A.1). Then, we have the following expressions for the components of $\partial_\theta \ell_N(\theta)$: 10.8

$$\partial_\beta \zeta_i(\theta) = -\sum_{j=1}^{n_i} \frac{s_{ij}(\alpha)}{\sigma_{ij}^2(\tau)} x_{ij} + \left(\lambda - \frac{n_i}{2}\right)\frac{1}{B_i}\partial_\beta B_i - T_i' A_i \partial_\beta B_i$$

$$= -\sum_{j=1}^{n_i} \frac{s_{ij}(\alpha)}{\sigma_{ij}^2(\tau)} x_{ij} - \frac{1}{B_i}\left\{\frac{1}{B_i}\left(\lambda - \frac{n_i}{2}\right) + T_i' A_i\right\}\sum_{j=1}^{n_i} \frac{(Y_{ij} - x_{ij}^\top \beta)}{\sigma_{ij}^2(\tau)} x_{ij},$$

$$\partial_\alpha \zeta_i(\theta) = \sum_{j=1}^{n_i} \frac{\partial_\alpha s_{ij}(\alpha)}{\sigma_{ij}^2(\tau)}(Y_{ij} - x_{ij}^\top \beta) - \left(\lambda - \frac{n_i}{2}\right)\frac{1}{A_i}\partial_\alpha A_i - T_i' B_i \partial_\alpha A_i$$

$$= \sum_{j=1}^{n_i} \frac{\partial_\alpha s_{ij}(\alpha)}{\sigma_{ij}^2(\tau)}(Y_{ij} - x_{ij}^\top \beta) - \frac{1}{A_i}\left\{\frac{1}{A_i}\left(\lambda - \frac{n_i}{2}\right) + T_i' B_i\right\}\sum_{j=1}^{n_i} \frac{s_{ij}(\alpha)}{\sigma_{ij}^2(\tau)}\partial_\alpha s_{ij}(\alpha),$$

$$\partial_\tau \zeta_i(\theta) = -\frac{1}{2}\sum_{j=1}^{n_i} \frac{\partial_\tau(\sigma_{ij}^2(\tau))}{\sigma_{ij}^2(\tau)} - \sum_{j=1}^{n_i} \frac{s_{ij}(\alpha)}{(\sigma_{ij}^2(\tau))^2}(Y_{ij} - x_{ij}^\top \beta)\partial_\tau(\sigma_{ij}^2(\tau))$$

$$+ \left(\lambda - \frac{n_i}{2}\right)\left(\frac{\partial_\tau B_i}{B_i} - \frac{\partial_\tau A_i}{A_i}\right) - T_i' \partial_\tau(A_i B_i)$$

$$= -\frac{1}{2}\sum_{j=1}^{n_i} \frac{\partial_\tau(\sigma_{ij}^2(\tau))}{\sigma_{ij}^2(\tau)} - \sum_{j=1}^{n_i} \frac{s_{ij}(\alpha)}{(\sigma_{ij}^2(\tau))^2}(Y_{ij} - x_{ij}^\top \beta)\partial_\tau(\sigma_{ij}^2(\tau))$$

$$- \frac{1}{2}\left(\lambda - \frac{n_i}{2}\right)\left(\frac{1}{B_i^2}\sum_{j=1}^{n_i} \frac{(Y_{ij} - x_{ij}^\top \beta)^2}{(\sigma_{ij}^2(\tau))^2}\partial_\tau(\sigma_{ij}^2(\tau))\right.$$

$$\left. - \frac{1}{A_i^2}\sum_{j=1}^{n_i} \frac{s_{ij}^2(\alpha)}{(\sigma_{ij}^2(\tau))^2}\partial_\tau(\sigma_{ij}^2(\tau))\right)$$

$$+ \frac{1}{2} T_i' \left(\frac{B_i}{A_i} \sum_{j=1}^{n_i} \frac{s_{ij}^2(\alpha)}{(\sigma_{ij}^2(\tau))^2} \partial_\tau (\sigma_{ij}^2(\tau)) + \frac{A_i}{B_i} \sum_{j=1}^{n_i} \frac{(Y_{ij} - x_{ij}^\top \beta)^2}{(\sigma_{ij}^2(\tau))^2} \partial_\tau (\sigma_{ij}^2(\tau)) \right),$$

$$\partial_\lambda \zeta_i(\theta) = \log \left(\frac{\gamma}{\delta} \right) - \frac{\partial_\lambda K_\lambda(\delta\gamma)}{K_\lambda(\delta\gamma)} + \log B_i - \log A_i + \frac{\partial_\lambda K_{\lambda - \frac{n_i}{2}}(A_i B_i)}{K_{\lambda - \frac{n_i}{2}}(A_i B_i)},$$

$$\partial_\delta \zeta_i(\theta) = \gamma R_\lambda(\delta\gamma) + \left(\lambda - \frac{n_i}{2} \right) \frac{\delta}{B_i^2} - T_i' \frac{A_i}{B_i} \delta,$$

$$\partial_\gamma \zeta_i(\theta) = \frac{2\lambda}{\gamma} + \delta R_\lambda(\delta\gamma) - \left(\lambda - \frac{n_i}{2} \right) \frac{\gamma}{A_i^2} - T_i' \frac{B_i}{A_i} \gamma.$$

As for the second-order derivatives, for brevity, we write

$$U_i = S_{\lambda - \frac{n_i}{2}}(A_i B_i) + \frac{1}{A_i B_i} R_{\lambda - \frac{n_i}{2}}(A_i B_i) - \frac{\lambda - \frac{n_i}{2}}{A_i^2 B_i^2},$$

for $R_\nu(t)$ and $S_\nu(t)$ defined by (A.1) and (A.2). Further, let

$$L_\nu(z) := \frac{1}{K_\nu^2(z)} \left(\partial_\lambda K_{\nu-1}(z) K_\nu(z) - \partial_\lambda K_\nu(z) K_{\nu-1}(z) \right).$$

Below we list the 21 components of $\partial_\theta^2 \zeta_i(\theta)$, which were used to compute the confidence intervals and the one-step estimator; the sizes of the matrices are not confusing, hence we are not taking care of them in notation and use the standard multilinear-form notation such as $(\partial_\beta B_i) \otimes (\partial_\alpha A_i) := \partial_\beta B_i \partial_\alpha^\top A_i \in \mathbb{R}^{p_\beta} \otimes \mathbb{R}^{p_\alpha}$.

$$\partial_\beta^2 \zeta_i(\theta) = \frac{\left(\lambda - \frac{n_i}{2} \right)}{B_i^2} \left(B_i \partial_\beta^2 B_i - (\partial_\beta B_i)^{\otimes 2} \right) - T_i' A_i \partial_\beta^2 B_i - U_i A_i^2 (\partial_\beta B_i)^{\otimes 2},$$

$$\partial_\beta \partial_\alpha \zeta_i(\theta) = -\sum_{j=1}^{n_i} \frac{1}{\sigma_{ij}^2(\tau)} (x_{ij} \otimes \partial_\alpha s_{ij}(\alpha)) - (T_i' + U_i A_i B_i) \{ (\partial_\beta B_i) \otimes (\partial_\alpha A_i) \},$$

$$\partial_\beta \partial_\tau \zeta_i(\theta) = \sum_{j=1}^{n_i} \frac{s_{ij}(\alpha)}{\left(\sigma_{ij}^2(\tau) \right)^2} \left(x_{ij} \otimes \partial_\tau (\sigma_{ij}^2(\tau)) \right) + \frac{\left(\lambda - \frac{n_i}{2} \right)}{B_i^2} \left(B_i (\partial_\tau \partial_\beta B_i) - (\partial_\beta B_i) \otimes (\partial_\tau B_i) \right)$$

$$- U_i A_i (\partial_\beta B_i) \otimes (\partial_\tau (A_i B_i)) - T_i' (\partial_\beta B_i) \otimes (\partial_\tau A_i) - T_i' A_i (\partial_\beta \partial_\tau B_i),$$

$$\partial_\beta \partial_\lambda \zeta_i(\theta) = A_i (\partial_\beta B_i) L_{\lambda - \frac{n_i}{2}}(A_i B_i),$$

$$\partial_\beta \partial_\delta \zeta_i(\theta) = -2 \left(\lambda - \frac{n_i}{2} \right) \frac{\delta}{B_i^3} \partial_\beta B_i - U_i A_i^2 \frac{\delta}{B_i} \partial_\beta B_i + T_i' A_i \frac{\delta}{B_i^2} \partial_\beta B_i,$$

$$\partial_\beta \partial_\gamma \zeta_i(\theta) = -\gamma U_i B_i \partial_\beta B_i - T_i' \frac{\gamma}{A_i} \partial_\beta B_i,$$

$$\partial_\alpha^2 \zeta_i(\theta) = \sum_{j=1}^{n_i} \frac{(y_{ij} - x_{ij}^\top \beta)}{\sigma_{ij}^2(\tau)} \partial_\alpha^2 s_{ij}(\alpha) - \frac{\left(\lambda - \frac{n_i}{2} \right)}{A_i^2} \left\{ A_i \partial_\alpha^2 A_i - (\partial_\alpha A_i)^{\otimes 2} \right\}$$

$$- U_i B_i^2 (\partial_\alpha A_i)^{\otimes 2} - T_i' B_i \partial_\alpha^2 A_i,$$

$$\partial_\alpha \partial_\tau \zeta_i(\theta) = -\sum_{j=1}^{n_i} \frac{(y_{ij} - x_{ij}^\top \beta)}{\sigma_{ij}^4(\tau)} \left\{ (\partial_\alpha s_{ij}(\alpha)) \otimes \partial_\tau (\sigma_{ij}^2(\tau)) \right\}$$

$$- \frac{\left(\lambda - \frac{n_i}{2}\right)}{A_i^2} \left(A_i(\partial_\alpha \partial_\tau A_i) - (\partial_\alpha A_i) \otimes (\partial_\tau A_i)\right)$$

$$- U_i B_i (\partial_\alpha A_i) \otimes (\partial_\tau (A_i B_i)) - T_i' \left((\partial_\alpha A_i) \otimes (\partial_\tau B_i) + B_i \partial_\alpha \partial_\tau A_i)\right),$$

$$\partial_\alpha \partial_\lambda \zeta_i(\theta) = -2 \frac{\partial_\alpha A_i}{A_i} - B_i (\partial_\alpha A_i) L_{\lambda - \frac{n_i}{2}}(A_i B_i),$$

$$\partial_\alpha \partial_\delta \ell_i(\theta) = -\delta U_i A_i \partial_\alpha A_i - T_i' \frac{\delta}{B_i} \partial_\alpha A_i,$$

$$\partial_\alpha \partial_\gamma \zeta_i(\theta) = 2\left(\lambda - \frac{n_i}{2}\right)\frac{\gamma}{A_i^3}\partial_\alpha A_i - U_i B_i^2 \frac{\gamma}{A_i}\partial_\alpha A_i + T_i' B_i \frac{\gamma}{A_i^2}\partial_\alpha A_i,$$

$$\partial_\tau^2 \zeta_i(\theta) = \sum_{j=1}^{n_i} \frac{s_{ij}(\alpha)(y_{ij} - x_{ij}^\top \beta)}{\sigma_{ij}^8(\tau)} \left(2\sigma_{ij(\tau)}^4 \left(\partial_\tau(\sigma_{ij}^2(\tau))\right)^{\otimes 2} - \sigma_{ij}^4(\tau)\partial_\tau^2(\sigma_{ij}^2(\tau))\right)$$

$$- \frac{1}{2}\sum_{j=1}^{n_i}\frac{\partial_\tau^2(\sigma_{ij}^2(\tau))\sigma_{ij}^2(\tau) - (\partial_\tau(\sigma_{ij}^2(\tau)))^{\otimes 2}}{\sigma_{ij}^4(\tau)}$$

$$+ \left(\lambda - \frac{n_i}{2}\right)\left[\frac{1}{B_i^2}\left(B_i(\partial_\tau^2 B_i) - (\partial_\tau B_i)^{\otimes 2}\right) - \frac{1}{A_i^2}\left(A_i(\partial_\tau^2 A_i) - (\partial_\tau A_i)^{\otimes 2}\right)\right]$$

$$- U_i \partial_\tau(A_i B_i) \otimes \partial_\tau(A_i B_i) - T_i' \partial_\tau^2(A_i B_i),$$

$$\partial_\tau \partial_\lambda \zeta_i(\theta) = -\frac{2\partial_\tau A_i}{A_i} - \partial_\tau(A_i B_i) L_{\lambda - \frac{n_i}{2}}(A_i B_i),$$

$$\partial_\tau \partial_\delta \zeta_i(\theta) = -2\left(\lambda - \frac{n_i}{2}\right)\frac{\delta}{B_i^3}\partial_\tau B_i - U_i A_i \frac{\delta}{B_i}\partial_\tau(A_i B_i) + \frac{\delta T_i'}{B_i}\left(\frac{A_i}{B_i}\partial_\tau B_i - \partial_\tau A_i\right),$$

$$\partial_\tau \partial_\gamma \zeta_i(\theta) = 2\left(\lambda - \frac{n_i}{2}\right)\frac{\gamma}{A_i^3}\partial_\tau A_i - U_i B_i \frac{\gamma}{A_i}\partial_\tau(A_i B_i) + \frac{\gamma T_i'}{A_i}\left(\frac{B_i}{A_i}\partial_\tau A_i - \partial_\tau B_i\right),$$

$$\partial_\lambda^2 \zeta_i(\theta) = \frac{\partial_\lambda^2 K_{\lambda - \frac{n_i}{2}}(A_i B_i)}{K_{\lambda - \frac{n_i}{2}}(A_i B_i)} - \left(\frac{\partial_\lambda K_{\lambda - \frac{n_i}{2}}(A_i B_i)}{K_{\lambda - \frac{n_i}{2}}(A_i B_i)}\right)^2 - \frac{\partial_\lambda^2 K_{\lambda - 1}(\delta\gamma)}{K_\lambda(\delta\gamma)} + \left(\frac{\partial_\lambda K_\lambda(\delta\gamma)}{K_\lambda(\delta\gamma)}\right)^2,$$

$$\partial_\lambda \partial_\delta \zeta_i(\theta) = -A_i \frac{\delta}{B_i} L_{\lambda - \frac{n_i}{2}}(A_i B_i) + \gamma L_\lambda(\delta\gamma),$$

$$\partial_\lambda \partial_\gamma \zeta_i(\theta) = \frac{2}{\gamma} - \frac{2\gamma}{A_i^2} + \delta L_\lambda(\delta\gamma) - B_i \frac{\gamma}{A_i} L_{\lambda - \frac{n_i}{2}}(A_i B_i),$$

$$\partial_\delta^2 \zeta_i(\theta) = \left(S_\lambda(\delta\gamma) + \frac{1}{\delta\gamma}R_\lambda(\delta\gamma)\right)\gamma^2 + \left(\lambda - \frac{n_i}{2}\right)\left(\frac{1}{B_i^2} - \frac{2\delta^2}{B_i^4}\right)$$

$$- U_i A_i^2 \frac{\delta^2}{B_i^2} - T_i' A_i \left(\frac{1}{B_i} - \frac{2\delta^2}{B_i^3}\right),$$

$$\partial_\delta \partial_\gamma \zeta_i(\theta) = 2R_\lambda(\delta\gamma) + \delta\gamma S_\lambda(\delta\gamma) - U_i \delta\gamma - T_i' \frac{\delta\gamma}{A_i B_i},$$

$$\partial_\gamma^2 \zeta_i(\theta) = -\frac{2\lambda}{\gamma^2} + \left(S_\lambda(\delta\gamma) + \frac{1}{\delta\gamma}R_\lambda(\delta\gamma)\right)\delta^2 - \left(\lambda - \frac{n_i}{2}\right)\left(\frac{1}{A_i^2} - \frac{2\gamma^2}{A_i^4}\right)$$

$$- U_i B_i^2 \frac{\gamma^2}{A_i^2} - T_i' B_i \left(\frac{1}{A_i} - \frac{\gamma^2}{A_i^3}\right).$$

References

Abramowitz, M., & Stegun, I. A. (Eds.). (1992). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover Publications Inc. Reprint of the 1972 edition.

Asar, O., Bolin, D., Diggle, P. J., & Wallin, J. (2020). Linear mixed effects models for non-Gaussian continuous repeated measurement data. *Journal of the Royal Statistical Society: Series C: Applied Statistics, 69*(5), 1015–1065.

Basawa, I. V., & Scott, D. J. (1983). *Asymptotic optimal inference for nonergodic models* Lecture notes in statistics (Vol. 17). Springer.

Delattre, M., Lavielle, M., & Poursat, M.-A. (2014). A note on BIC in mixed-effects models. *Electronic Journal of Statistics, 8*(1), 456–475.

Eberlein, E., & Hammerstein, E.A.v. (2004). Generalized hyperbolic and inverse Gaussian distributions: limiting cases and approximation of processes. In *Seminar on Stochastic Analysis, Random Fields and Applications IV*, volume 58 of *Progr. Probab.*, pp. 221–264. Birkhäuser.

Fahrmeir, L. (1990). Maximum likelihood estimation in misspecified generalized linear models. *Statistics, 21*(4), 487–502.

Fujinaga, Y. (2021). Asymptotic inference for location-scale mixed-effects model. Master thesis, Kyushu University.

Hedeker, D., Demirtas, H., & Mermelstein, R. J. (2009). A mixed ordinal location scale model for analysis of ecological momentary assessment (EMA) data. *Stat. Interface, 2*(4), 391–401.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics, 64*(2), 627–634, 670.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine, 31*(27), 3328–3336.

Hedeker, D., & Nordgren, R. (2013). Mixregls: A program for mixed-effects location scale analysis. *Journal of Statistical Software, 52*(12), 1–38.

Jeganathan, P. (1982). On the asymptotic theory of estimation when the limit of the log-likelihood ratios is mixed normal. *Sankhyā Series A, 44*(2), 173–212.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*(4), 963–974.

Lavielle, M. (2015). *Mixed effects models for the population approach. Chapman & Hall/CRC biostatistics series. Models, tasks, methods and tools, With contributions by Kevin Bleakley*. CRC Press.

Richards, F. S. G. (1961). A method of maximum-likelihood estimation. *Journal of the Royal Statistical Society Series B, 23*, 469–475.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4*, 1–32.

Sweeting, T. J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *Annals of Statistics, 8*(6), 1375–1381. Corrections: (1982) Annals of Statistics 10, 320.

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika, 92*(2), 351–370.

van der Vaart, A. W. (1998). *Asymptotic statistics, volume 3 of Cambridge series in statistical and probabilistic mathematics*. Cambridge University Press.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*(1), 1–25.

Yoon, J., Kim, J., & Song, S. (2020). Comparison of parameter estimation methods for normal inverse Gaussian distribution. *Communications for Statistical Applications and Methods, 27*(1), 97–108.