University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2003 *Paper* 3

Maximization by Parts in Likelihood Inference

Peter Xuekun Song* Yanqin Fan†

Jack Kalbfleisch‡

*University of Michigan, Ann Arbor, pxsong@umich.edu

†Vanderbilt University, yanqin.fan@vanderbilt.edu

‡University of Michigan, jdkalbfl@umich.edu

Maximization by Parts in Likelihood Inference

Peter Xuekun Song, Yanqin Fan, and Jack Kalbfleisch

Abstract

This paper presents and examines a new algorithm for solving a score equation for the maximum likelyhood estimate in certain problems of practical interest. The method circumvents the need to compute second order derivaties of the full likelihood function. It exploits the structure of certain models that yield a natural decomposition of a very complicated likelihood function. In this decomposition, the first part is a log likelihood from a simply analyzed model and the second part is used to update estimates from the first. Convergence properties of this fixed point algorithm are examined and asymptotics are derived for estimators obtained by using only a finite number of steps. Illustrative examples considered in the paper included bivariate and multivariate Gaussian copula models, nonnormal random effects and state space models. Properties of the algorithm and of estimators are evaluated in simulation studies on a bivariate copula model and a nonnormal linera random effects model.

Maximization by Parts in Likelihood Inference

PETER X.-K. SONG, YANQIN FAN and JOHN D. KALBFLEISCH

ABSTRACT:

This paper presents and examines a new algorithm for solving a score equation for the maximum likelihood estimate in certain problems of practical interest. The method circumvents the need to compute second order derivatives of the full likelihood function. It exploits the structure of certain models that yield a natural decomposition of a very complicated likelihood function. In this decomposition, the first part is a log likelihood from a simply analyzed model and the second part is used to update estimates from the first. Convergence properties of this fixed point algorithm are examined and asymptotics are derived for estimators obtained by using only a finite number of steps. Illustrative examples considered in the paper include bivariate and multivariate Gaussian copula models, nonnormal random effects and state space models. Properties of the algorithm and of estimators are evaluated in simulation studies on a bivariate copula model and a nonnormal linear random effects model.

KEY WORDS: Copula models; Fixed point algorithm; Information dominance; Iterative algorithm; Nonnormal random effects; Score equation; State space models. [1]

[1]P. Song is Associate Professor, Department of Mathematics and Statistics, York University, Toronto, ON M3J 1P3, Canada (Email: *song@mathstat.yorku.ca*). Y. Fan is Professor, Department of Economics, Vanderbilt University, Nashville, TN 37235-1819 (Email: *yanqin.fan@vanderbilt.edu*). J. Kalbfleisch is Professor and Chair, Department of Biostatistics, UM School of Public Health, Ann Arbor, MI 48109-2029 (Email:*jdkalbfl@umich.edu*). The first author's research was supported by the NSERC Operating Grant. The research was done while P. Song was visiting Department of Biostatistics, University of Michigan, and he acknowledges the computing support from the university.

1 INTRODUCTION

Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be independent vectors of random variables, and suppose $\mathbf{y}_i$ has density in the parametric family, $\{p_i(\mathbf{y}|\theta), \theta \in \Theta \subset \mathcal{R}^p\}$. The corresponding log-likelihood function is

$$\ell(\theta) = \sum_{i=1}^{n} \log p_i(\mathbf{y}_i|\theta) = \sum_{i=1}^{n} \log \ell_i(\theta). \tag{1}$$

In the regular case, the maximum likelihood estimate, $\hat{\theta}$, of θ is a solution to the score equation $\dot{\ell}(\theta) = 0$, where $\dot{\ell}(\theta)$ denotes the vector of first order derivatives of $\ell(\theta)$. This solution is unique when the log-likelihood function $\ell(\theta)$ is concave. In some cases, a closed-form solution can be found; more often, a numerical solution is required using iterative methods such as Newton-Raphson, Fisher scoring, the simplex method, quasi-Newton methods, simulated annealing, or the EM algorithm.

The Newton-Raphson algorithm iteratively updates the parameter estimate using,

$$\theta^k = \theta^{k-1} - \left\{\frac{1}{n}\ddot{\ell}\left(\theta^{k-1}\right)\right\}^{-1} \left[\frac{1}{n}\dot{\ell}\left(\theta^{k-1}\right)\right], k = 1, \ldots, \tag{2}$$

where $\ddot{\ell}(\theta)$ is the Hessian matrix of second order derivatives of $\ell(\theta)$ and θ^0 is the initial value. A variation on this is the Fisher scoring algorithm in which $\frac{1}{n}\ddot{\ell}$ is replaced by its expectation or Fisher information. From good starting values, both methods typically converge rapidly to the MLE $\hat{\theta}$, and give rise to estimates of the asymptotic covariance matrix of $\hat{\theta}$.

In many instances, the likelihood function is very complicated and analytic expressions, especially for second order derivatives are not easily obtained or used. One approach to bypass this problem is to replace the Fisher information by an estimate such as $n^{-1}\sum_{i=1}^{n} \dot{\ell}_i(\theta^{k-1})\dot{\ell}_i(\theta^{k-1})^T$. This will work reasonably well if n is relatively large and θ^0 is a consistent estimate. When n is not large, however, this approach can be very unstable due to variation in the estimated information matrix. Another approach is to use the so-called pseudo or empirical derivatives obtained by differencing to approximate first and second derivatives in (2). When the likelihood is changing only slowly, however, the empirical approximation is very sensitive to the choice of grid points for differencing, and an

algorithm built upon this approximation may be very fragile, especially when the dimension of the parameter is high. For a complex likelihood function, algorithms incorporating these approximations often encounter difficulties in invertibility and/or positive definiteness at updated values.

In this paper we propose a new algorithm to solve score equations from some complicated likelihood functions. The proposed algorithm strategically selects a part of the full likelihood function with easily computed second order derivatives. The remaining more difficult part of the likelihood function participates in the algorithm in such a way that its second order derivative is not needed. In this algorithm, as for the quasi-Newton algorithms discussed above, the second order derivatives of the full log-likelihood are not required.

When the full likelihood is so complicated as to be numerically unmanageable, some simplifications may be introduced by using a related estimating equation that is easy to solve. Examples of this strategy are the method of inference functions for margins (IFM) proposed by McLeish and Small (1988), Liang and Zeger's (1986) GEE approach in marginal models, and Breslow and Clayton's (1993) approximate inference in generalized linear mixed models. The major drawback of this strategy is that there is some loss of efficiency in estimation due to the use of the estimating function of an approximate model rather than the exact score. A 'simple analysis' based on the approximate model is useful in some problems, but acquiring efficient estimators is always of often interest. One attractive feature of the algorithm proposed here is that it enables us to link the simple and exact analyses. In doing so, the algorithm uses the residual part of the score equation to correct and improve the efficiency of estimation.

This paper is organized as follows. We present the formulation of the algorithm in Section 2 and briefly discuss examples in Section 3. Asymptotic results are discussed in Section 4 and Section 5 in which a variant of the algorithm is presented. In Section 6, we explore the application of the proposed algorithm in several different problems. Section 7 includes some

discussion and comments and technical details are given in the appendices.

2 FRAMEWORK FOR THE ALGORITHM

Consider a selected additive decomposition of a log-likelihood function

$$\ell(\theta) = \ell_w(\theta) + \ell_e(\theta).$$

The corresponding score function is given by

$$\dot{\ell}(\theta) = \dot{\ell}_w(\theta) + \dot{\ell}_e(\theta).$$

The objective is to find the solution to the score equation $\dot{\ell}(\theta) = 0$, i.e. the maximum likelihood estimate. We assume that the calculation of the $\ddot{\ell}$ is difficult, and aim to avoid it. In contrast, the first part $\ell_w(\theta)$ is selected so that solving the corresponding estimation equation $\dot{\ell}_w(\theta) = 0$ is simple. Throughout this paper we assume $\dot{\ell}_w(\cdot)$ (and hence $\dot{\ell}_e(\cdot)$) is an unbiased inference function. Note that ℓ_w need not be a conditional, marginal or partial likelihood; only the unbiasedness of $\dot{\ell}_w$ is needed.

Now let θ_n^1 denote the solution to $\dot{\ell}_w(\theta) = 0$. Under some mild regularity conditions, the classical theory of estimating functions establishes consistency

$$\theta_n^1 \xrightarrow{p} \theta_0$$

and asymptotic normality,

$$\sqrt{n}(\theta_n^1 - \theta_0) \to N(0, J_1(\theta_0))$$

where $J_1(\theta_0) = \{E(\ddot{\ell}_w^T(\theta_0))\}^{-1} E\{\dot{\ell}_w(\theta_0)\dot{\ell}_w^T(\theta_0)\}\{E(\ddot{\ell}_w(\theta_0))\}^{-1}$.

The estimator θ_n^1 can have low efficiency since only part of the full log-likelihood function is used in estimation. To increase the efficiency, it seems necessary to utilize the information in the second piece $\dot{\ell}_e$. Suppose we are able to evaluate $\dot{\ell}_e$ and consider an iterative algorithm in which the second step is to solve the equation,

$$\dot{\ell}_w(\theta) = -\dot{\ell}_e\left(\theta_n^1\right), \tag{3}$$

for θ_n^2, say. To assure this proposal is worthwhile, we need to answer the following questions:

(a) Is θ_n^2 consistent and asymptotically normal?

(b) Is θ_n^2 more efficient than θ_n^1?

Continuing this approach, consider the following algorithm:

STEP 1 Solve $\dot{\ell}_w(\theta) = 0$ for θ_n^1.

STEP k Solve $\dot{\ell}_w(\theta) = -\dot{\ell}_e(\theta_n^{k-1})$ to produce estimate θ_n^k, $k = 2, 3, \ldots$.

Note that, if the inverse of $\dot{\ell}_w(\cdot)$ exists, the proposed algorithm is a fixed point algorithm since we may write $\theta_n^{k+1} = \dot{\ell}_w^{-1}\left\{-\dot{\ell}_e(\theta_n^k)\right\}$. Classical numerical analysis theory (e.g. Burden and Faires, 1997, Theorems 10.5 and 10.6) gives the condition for the existence and uniqueness of a fixed point. That is, the derivatives of the functional, which in our case is $\dot{\ell}_w^{-1}\left\{-\dot{\ell}_e(\cdot)\right\}$, is bounded by C_0/p where C_0 is a constant less than 1 and p is the dimension of the parameter θ. An equivalent condition, referred to as *information dominance* in Section 4, is required for the convergence of our algorithm. Obviously, if the proposed algorithm converges, it should converge to the MLE. This is because the limiting point θ_n^∞ of θ_n^k as $k \to \infty$ satisfies $\dot{\ell}_w(\theta_n^\infty) + \dot{\ell}_e(\theta_n^\infty) = 0$.

3 EXAMPLES

In this section, we present three examples from very different areas to demonstrate the flexibility of the proposed algorithm and its variant to be introduced in Section 5 in solving complex score equations.

Example 1 (The Gaussian copula) Consider multivariate data $\mathbf{y}_1, \ldots, \mathbf{y}_n$, where the ith observed d-dimensional vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{id})$ follows a d-variate Gaussian copula distribution with distribution function (CDF) $C(F_1(y_1; \alpha_1), \ldots, F_d(y_d; \alpha_d)|\Gamma)$. See, for example,

Joe (1997) and Song (2000). Here, α_j is the (vector) parameter of the marginal distribution of y_{ij}, $j = 1, \ldots, d$ respectively and $C(\cdot)$ is the d-variate Gaussian copula given by

$$C(u_1, \ldots, u_d|\Gamma) = \Phi_d(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)),\ (u_1, \ldots, u_d) \in (0,1)^d,$$

where Φ_d and Φ are respectively the distribution functions of the d-variate normal $N_d(0, \Gamma)$ with a correlation matrix Γ and of the standard normal $N(0,1)$.

Assume the j-th marginal density is $f_j(y_j; \alpha_j), j = 1, \ldots, d$. Let θ denote all the distinct elements in parameters $(\alpha_1, \ldots, \alpha_d, \Gamma)$ in the model. Then the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \left\{ c(F_1(y_{i1}; \alpha_1), \ldots, F_d(y_{id}; \alpha_d)|\Gamma) \prod_{j=1}^{d} f_j(y_{ij}; \alpha_j) \right\}, \tag{4}$$

where $c(\cdot)$ is the density corresponding to $C(\cdot)$,

$$c(u_1, \ldots, u_d|\Gamma) = |\Gamma|^{-1/2} \exp\left\{ \frac{1}{2} \mathbf{z}^T (I_d - \Gamma^{-1}) \mathbf{z} \right\},$$

and $\mathbf{z}^T = (z_1, \ldots, z_d) = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d))$ and I_d denotes the $d \times d$ identity matrix.

The log-likelihood function can be written as

$$\ell(\theta) = \ell_w(\theta) + \ell_e(\theta)$$

where

$$\begin{aligned} \ell_w(\theta) &= \sum_{i=1}^{n} \sum_{j=1}^{d} \ln f_j(y_{ij}; \alpha_j) \\ \ell_e(\theta) &= -\frac{n}{2} \ln |\Gamma| + \frac{1}{2} \sum_{i=1}^{n} \mathbf{z}_i(\theta)^T (I_d - \Gamma^{-1}) \mathbf{z}_i(\theta). \end{aligned}$$

Note that $\ell_w(\theta)$ is the likelihood function under the independence correlation structure ($\Gamma = I_d$) and only involves the marginal parameters α_j, and $\ell_e(\theta)$ contains all parameters. It is often straightforward to handle ℓ_w by computing its first and second order derivatives, but hard to derive analytically the second order derivatives of ℓ_e. Therefore, neither the Newton-Raphson nor the Fisher Scoring algorithm is easily available. Although a quasi-Newton

algorithm may be applied here, it often encounters singularity problem for the matrix of the pseudo-derivatives. Our proposed algorithm provides an alternative approach. As shown in Section 6.1 where a bivariate Gaussian copula model is studied, the proposed algorithm yields closed form expressions for the iteration estimates of both marginal and correlation parameters. This makes the calculation simple and fast.

Example 2 (Non-normal random effects models) The normality assumption in linear random effects models is often made for mathematical convenience and may be violated in practical settings. For example, Zhang and Davidian (2001) report a histogram of subject-specific intercept estimates from individual least squares fits to the Framingham cholesterol data, which clearly indicates that the normality assumption for the random effects is not appropriate. See also Pinheiro et al. (2001) in which they assume a t distribution for random effects to gain robustness in parameter estimation.

Consider a linear random effects model for clustered or longitudinal data,

$$y_{ij} = x_{ij}^T\beta + z_{ij}^T\alpha_i + \varepsilon_{ij},\ j = 1, \ldots, m, \quad i = 1, \ldots, n, \tag{5}$$

where x_{ij} and β are vectors of dimension p, and z_{ij} and α_i are vectors of dimension q. The random effects α_i are *iid* with density $p(\alpha|\eta)$ and the ε_{ij}'s are *iid* $N(0, \sigma)$. The $p(\alpha|\eta)$ may, for example, be a q variate t distribution where each marginal distribution is t on v degrees of freedom (known) and η is the covariance matrix. More generally, the degrees of freedom could also be an unknown parameter included in η.

The model (5) can be rewritten as

$$\mathbf{y}_i = X_i\beta + Z_i\alpha_i + \varepsilon_i, i = 1, \ldots, n, \tag{6}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^T$, $X_i = (x_{i1}, \ldots, x_{im})^T$, and $Z_i = (z_{i1}, \ldots, z_{im})^T$.

Let $\theta = (\beta, \eta, \sigma)$. The likelihood function is

$$L(\theta) = \Pi_{i=1}^n p(\mathbf{y}_i|\theta) = \Pi_{i=1}^n \int p(\mathbf{y}_i, \alpha_i|\theta)d\alpha_i = \Pi_{i=1}^n \int p(\mathbf{y}_i|\alpha_i, \theta)p(\alpha_i|\eta)d\alpha_i, \tag{7}$$

where $p(\mathbf{y}_i|\alpha_i, \theta)$ is the density of the $N(X_i\beta + Z_i\alpha, \sigma I)$ distribution. The maximum likelihood estimates of the parameters are obtained by maximizing L, which is typically difficult due to the presence of the q-dimensional integral. There are several simulation-based methods to find the maximum likelihood estimate of θ_0. For example, the simulated maximum likelihood method (Geyer and Thompson, 1992) approximates the integration in (7) by simulating the random effects according to the 'prior' distribution $p(\alpha_i|\eta)$ either directly or by using importance sampling with a proposal prior distribution such as a multivariate normal distribution. The approximation accuracy, however, is dependent on the choice of the proposal prior (McCulloch and Searle, 2002, p. 180) and can deteriorate for choices far from the truth.

In this paper, we apply the idea of importance sampling in a slightly different way. If $p(\alpha|\eta)$ were normal, the maximization would be simple, because in this case L would be a multivariate normal density. Therefore, consider a working normal distribution, $N_q(0, D)$ with density $\phi(\alpha|\theta)$, for the random effects. In the linear mixed model being considered here, the working normal model leads to consistent estimation of β, D, and σ. This is because the consistency in the linear mixed model requires only the first two moments of the data. Now, under this normal random effects assumption, the working log-likelihood is

$$\ell_w(\theta) = \sum_{i=1}^{n} \ln \phi(\mathbf{y}_i|\theta) = \sum_{i=1}^{n} \ln \int p(\mathbf{y}_i|\alpha_i, \theta)\phi(\alpha_i|\theta)d\alpha_i. \tag{8}$$

We use ϕ to indicate densities under the working model and p to indicate densities under the true model and note that $p(\mathbf{y}_i|\alpha, \theta) = \phi(\mathbf{y}_i|\alpha, \theta)$. Since

$$\phi(\mathbf{y}_i|\theta) = \frac{p(\mathbf{y}_i|\alpha, \theta)\phi(\alpha_i|\theta)}{\phi(\alpha_i|\mathbf{y}_i, \theta)},$$

the ith term in the full likelihood can be written as

$$p(\mathbf{y}_i|\theta) = \phi(\mathbf{y}_i|\theta) \int \frac{p(\alpha_i|\theta)}{\phi(\alpha_i|\theta)} \phi(\mathbf{y}_i|\alpha_i, \theta)d\alpha_i.$$

Substitution into (7) gives the log-likelihood,

$$\ell(\theta) = \ell_w(\theta) + \ell_e(\theta), \tag{9}$$

of the required form where

$$\ell_e(\theta) = \sum_{i=1}^{n} \ln \int \frac{p(\alpha_i|\eta)}{\phi(\alpha_i|\eta)} \phi(\alpha_i|\mathbf{y}_i, \theta) d\alpha_i. \tag{10}$$

Note that ℓ_e may be thought of essentially as a discrepancy measure between the original and working distributions of the random effects, weighted by the working 'posterior' of the random effects.

We make two comments:

(i) ℓ_w is given by equation (8) where a working normal $\phi(\alpha_i|\eta)$ replaces $p(\alpha_i|\eta)$, while ℓ_e represents the error made by this normal working model. Equation (9) provides a natural connection between the log likelihoods of a normal and nonnormal random effects models. With the normal working model, $\phi(\alpha_i|\mathbf{y}_i, \theta)$ is a multivariate normal, and therefore the Monte Carlo evaluation of the integral in (10) is easy to carry out, even when the dimension q of α_i is large. In addition, since ℓ_e is of the form $\int h(\alpha) \exp(-\alpha^T \alpha/2) d\alpha$, Gaussian quadrature can also be used to compute the integral.

(ii) In some settings, the normal random effects model may not be the best working model to choose. Whatever working model is chosen to yield the ℓ_w, this approach allows us to balance conflicting requirements of analytical tractability and the flexibility to model real data.

Example 3 (State space models) High frequency time series of stock transaction records provide valuable information about the stock market. We consider models for the duration process where duration is the time interval between two consecutive trades. It is well-known (e.g. Engle and Russell, 1998 and Bauwens and Veredas, 2003) that the distribution of trading durations is heavy tailed. Let $\{d_i, i = 1, \ldots, n\}$ be a sequence of trading durations. A stochastic conditional duration model proposed by Bauwens and Veredas (2003) takes the form of a state space model

$$\ln(d_i) \quad = \quad \mu + \psi_i + \eta_i$$

$$\psi_i \quad = \quad \beta\psi_{i-1} + \xi_i$$

where η_i and ξ_i are independent errors, η_i follows a heavy tailed distribution $p(\cdot|\alpha)$, such as log-gamma, and ξ_i is Gaussian $\phi(\cdot|\sigma)$. The latent variable ψ_i is of financial interest as it represents the Markovian structure of the log-duration. Note that if η_i were Gaussian, the classical Kalman filter and smoothing technique would be applied to estimate the latent process ψ_i. Let $y_i = \ln(d_i) - \mu$, $\mathbf{y} = (y_1, \ldots, y_n)$, and $\psi = (\psi_1, \ldots, \psi_n)$. Denote all parameters by $\theta = (\beta, \alpha, \sigma)$. Therefore the likelihood function is

$$L(\theta) = \int p(\mathbf{y}|\psi, \theta) p(\psi|\theta) d\psi.$$

Let $L_w(\theta)$ be the corresponding likelihood function under the working assumption that η_i is Gaussian. Following Example 2 above, or Durbin and Koopman (1997), we obtain

$$L(\theta) = L_w(\theta) \int \frac{p(\mathbf{y}|\psi, \theta)}{p_w(\mathbf{y}|\psi, \theta)} p_w(\psi|\mathbf{y}, \theta) d\psi \equiv L_w(\theta) L_e(\theta),$$

where L_e can be thought of as an averaged discrepancy between the approximate and the true distributions of the log-duration over all states. The log-likelihood function is again additive, where ℓ_w is straightforward to analyze and ℓ_e is much more difficult.

4 ASYMPTOTICS

In this section, we study asymptotic properties of the estimators θ_n^k given by the proposed algorithm, including the consistency (Theorem 1) and asymptotic normality (Theorem 3). In fact, Theorem 3 establishes the asymptotic normality for every iteration, which enables us to calculate asymptotic standard errors at any iteration where the algorithm is stopped, such as the case of one-step update. In particular, we give sufficient conditions that assure the convergence and asymptotic normality as the iteration index $k \to \infty$. In order to establish these asymptotic properties, we need the following conditions concerning the log likelihood

function. Let $\Omega_0 = \{\theta : ||\theta - \theta_0|| < \delta\}$ be a neighborhood of the true parameter θ_0, where $||\cdot||$ is the Euclidean norm.

(A) $\ell(\theta)$, $\ell_w(\theta)$ and $\ell_e(\theta)$ are twice continuously differentiable for $\theta \in \Omega_0$;

(B) The matrix $(\mathcal{I}_w^{-1}\mathcal{I}_e)^k \to 0$ as $k \to \infty$, where $\mathcal{I}_w = -n^{-1}E\ddot{\ell}_w(\theta_0)$ and $\mathcal{I}_e = -n^{-1}E\ddot{\ell}_e(\theta_0)$. This condition is referred to as *information dominance.*

We make two comments on condition (B).

(i) That the power series decays to zero implies that $\mathcal{I}_w$ is 'larger' than $\mathcal{I}_e$, meaning that ℓ_w contains more information on θ_0 than ℓ_e. Consequently, the Hessian matrix of ℓ_w directs the movement of the updated values.

(ii) To examine the connection between condition (B) and that required by the fixed point algorithm, let us assume that θ is 1-dimensional. When $\dot{\ell}_w^{-1}$ exists, the information dominance is stochastically equivalent, within a $\sqrt{n}$ neighborbood, to the condition that the derivative of the functional $\dot{\ell}_w^{-1}\{\dot{\ell}_e(\cdot)\}$ is bounded by a constant $c_0 < 1$. This is because given a consistent estimator $\hat{\theta}$ such that $\sqrt{n}(\hat{\theta} - \theta_0) = o_p(1)$, this derivative can be expressed as $\mathcal{I}_w^{-1}\mathcal{I}_e + o_p(1)$. For the case of higher dimensions, a similar argument can be made based on each component of θ.

The proofs of Theorems 1 to 4 are given in Appendix A.

Theorem 1 *Under condition (A), if θ_n^1 is consistent, then θ_n^2 is consistent.*

Thus, if θ_n^1 is consistent, then θ_n^k is consistent for each $k = 2, 3, \ldots$. Let

$$\tau_n(\theta_0) = \left\{-n^{-1}\ddot{\ell}_w(\theta_0)\right\}^{-1}\{n^{-1}\ddot{\ell}_e(\theta_0)\},$$

so that $I_p - \tau_n(\theta_0) = \left\{n^{-1}\ddot{\ell}_w(\theta_0)\right\}^{-1}\{n^{-1}\ddot{\ell}(\theta_0)\}$.

Theorem 2 *Under condition (A), for any integers $k \geq 2$ and $m \geq 1$,*

$$\sqrt{n}\left(\theta_n^{k+m} - \theta_n^k\right) = \{I_p - \tau_n^m(\theta_0)\}\,\tau_n^{k-1}(\theta_0)\left\{-n^{-1}\ddot{\ell}(\theta_0)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\dot{\ell}_1(\theta_0)\tau_n(\theta_0) + \frac{1}{\sqrt{n}}\dot{\ell}_2(\theta_0)\right\} + o_p(1)$$

where I_p is the identity matrix of dimension $p \times p$.

Note that $\text{plim}_{n\to\infty}\tau_n(\theta_0) = -\mathcal{I}_w^{-1}\mathcal{I}_e = \tau$, say. Theorem 2 implies that under condition (B), the difference between two updates θ_n^{k+m} and θ_n^k with m-steps apart will vanish when $k \to \infty$ for large n.

To establish the asymptotic distribution of θ_n^k, we first note that under some mild regularity conditions,

$$n^{-1/2}\begin{bmatrix} \dot{\ell}_w \\ \dot{\ell}_e \end{bmatrix} \to N(0, \Omega),$$

in distribution, where

$$\Omega = \lim_{n\to\infty} n^{-1}\begin{pmatrix} E\dot{\ell}_w\dot{\ell}_w^T & E\dot{\ell}_w\dot{\ell}_e^T \\ E\dot{\ell}_e\dot{\ell}_w^T & E\dot{\ell}_e\dot{\ell}_e^T \end{pmatrix},$$

and $\dot{\ell}_w = \dot{\ell}_w(\theta_0)$ and $\dot{\ell}_e = \dot{\ell}_e(\theta_0)$.

Theorem 3 *Under some regularity conditions as required in the MLE, θ_n^k is asymptotically normally distributed with mean θ_0 and variance $n^{-1}\Sigma_k$ with*

$$\Sigma_k = A_k^T \Omega A_k, \tag{11}$$

where

$$A_k = \begin{pmatrix} \left[I_p - \tau^k\right]\mathcal{I}^{-1} \\ \left[I_p - \tau^{k-1}\right]\mathcal{I}^{-1} \end{pmatrix}.$$

Moreover, under the condition of information dominance as $k \to \infty$, $\Sigma_k \to \mathcal{I}^{-1}$, the inverse of the Fisher information.

It is easy to see that under condition (B), $\tau^k \to 0$ as $k \to \infty$. Hence

$$\lim_{k\to\infty}\Sigma_k = \mathcal{I}^{-1}[\lim n^{-1}E\{\dot{\ell}_w + \dot{\ell}_e\}\{\dot{\ell}_w + \dot{\ell}_e\}^T]\mathcal{I}^{-1} = \mathcal{I}^{-1}.$$

According to Theorem 3, as $k \to \infty$, the asymptotic variance matrix of θ_n^k converges at an exponential rate to the asymptotic variance matrix of the MLE $\hat{\theta}$.

At each iteration, let

$$\widehat{\mathcal{I}}^k = n^{-1} \sum_{i=1}^{n} \dot{\ell}_i(\mathbf{y}_i; \theta_n^k) \dot{\ell}_i(\mathbf{y}_i; \theta_n^k)^T,$$

where $\ell_i(\mathbf{y}_i; \cdot)$ denotes the i-th piece of the likelihood with respect to observation $\mathbf{y}_i$. At convergence, the algorithm yields the MLE $\widehat{\theta}$, and the 'average' Fisher information can be estimated by $\widehat{\mathcal{I}} \equiv \widehat{\mathcal{I}}^\infty$ with the θ_n^k replaced by $\widehat{\theta}$.

An estimate of $\mathcal{I}_w$ is $\widehat{\mathcal{I}}_w = -n^{-1} \sum_{i=1}^{n} \ddot{\ell}_{w,i}(\mathbf{y}_i; \hat{\theta})$. An estimate of $\mathcal{I}_e$ is then given by $\widehat{\mathcal{I}}_e = \widehat{\mathcal{I}} - \widehat{\mathcal{I}}_w$. Similar quantities can be obtained at any iteration, and they allow direct estimation of the asymptotic covariance matrix of θ_n^k.

5 A VARIANT OF THE ALGORITHM

In some cases, such as Example 1 in Section 3, the parameter vector $\theta^T = (\theta_1^T, \theta_2^T)$, where θ_1 and θ_2 are of dimensions p_1 and p_2 with $p_1 + p_2 = p$, and the log-likelihood function can be written as

$$\ell(\theta) = \ell_w(\theta_1) + \ell_e(\theta_1, \theta_2).$$

The resulting score equations are

$$\begin{pmatrix} \dot{\ell}_{w(1)}(\theta_1) + \dot{\ell}_{e(1)}(\theta_1, \theta_2) \\ \dot{\ell}_{e(2)}(\theta_1, \theta_2) \end{pmatrix} = 0,$$

where $\dot{\ell}_{i(j)} = \partial \ell_i(\theta_0)/\partial \theta_j$, $i = w, e;\ j = 1, 2$.

We suppose that $\dot{\ell}_{w(1)}, \dot{\ell}_{e(1)}$ and $\dot{\ell}_{e(2)}$ are all unbiased for parameter θ_1 and θ_2, respectively. A modified version of the algorithm is as follows:

STEP 1 Solve $\dot{\ell}_{w(1)}(\theta_1) = 0$ for $\theta_{1,n}^1$; and

Solve $\dot{\ell}_{e(2)}(\theta_{1,n}^1, \theta_2) = 0$ for $\theta_{2,n}^1$.

STEP k Solve $\dot{\ell}_{w(1)}(\theta_1) = -\dot{\ell}_{e(1)}(\theta_{1,n}^{k-1}, \theta_{2,n}^{k-1})$ for $\theta_{1,n}^k$; and

Solve $\dot{\ell}_{e(2)}(\theta_{1,n}^{k-1}, \theta_2) = 0$ for $\theta_{2,n}^k$, $k = 2, 3, \ldots$.

Let $\theta_n^k = (\theta_{1,n}^k, \theta_{2,n}^k)$. By a similar argument to that in Theorem 1, θ_n^k is consistent for each k. Further, under regularity conditions, θ_n^k will converge to the MLE $\widehat{\theta}$, and hence become fully efficient as $k \to \infty$. In Theorem 4, we establish asymptotic variances of θ_n^k.

Let $\mathcal{I}_{i(jk)} = -n^{-1}E\ddot{\ell}_{i(jk)}(\theta_0), i = w, e;\ j, k = 1, 2,$ where $\ddot{\ell}_{i(jk)}(\theta) = \partial^2 \ell_i(\theta_0)/\partial\theta_j\partial\theta_k^T$. At θ_0, let

$$
\begin{aligned}
D_n &= \begin{bmatrix} -n^{-1}\ddot{\ell}_{w(11)} & 0 \\ 0 & -n^{-1}\ddot{\ell}_{e(22)} \end{bmatrix} = \begin{bmatrix} \mathcal{I}_{w(11)} & 0 \\ 0 & \mathcal{I}_{e(22)} \end{bmatrix} + o_p(1) \equiv D + o_p(1), \\
T_n &= \begin{bmatrix} n^{-1}\ddot{\ell}_{e(11)} & n^{-1}\ddot{\ell}_{e(12)} \\ n^{-1}\ddot{\ell}_{e(21)} & 0 \end{bmatrix} = -\begin{bmatrix} \mathcal{I}_{e(11)} & \mathcal{I}_{e(12)} \\ \mathcal{I}_{e(21)} & 0 \end{bmatrix} + o_p(1) \equiv T + o_p(1), \\
L_n &= \begin{bmatrix} 0 & 0 \\ n[\ddot{\ell}_{e(22)}]^{-1}[\ddot{\ell}_{e(21)}][\ddot{\ell}_{w(11)}]^{-1} & 0 \end{bmatrix} = -\begin{bmatrix} 0 & 0 \\ \mathcal{I}_{e(22)}^{-1}\mathcal{I}_{e(21)}\mathcal{I}_{w(11)}^{-1} & 0 \end{bmatrix} + o_p(1) \equiv L + o_p(1),
\end{aligned}
$$

and $\Gamma = \text{plim}_{n\to\infty} D_n^{-1}T_n = D^{-1}T$.

Let Ω_V be the asymptotic variance matrix of the estimating functions $n^{-1/2}[\dot{\ell}_{w(1)}, \dot{\ell}_{e(2)}, \dot{\ell}_{w(1)}, 0]^T$, where 0 is included only for the sake of dimension.

Theorem 4 *Under the regularity conditions of maximum likelihood, θ_n^k is asymptotically normal with mean θ_0 and variance $n^{-1}\Sigma_k$ where $\Sigma_k = B_k^T\Omega_V B_k$, $B_k = [B_{k1}, B_{k2}]$, and*

$$
\begin{aligned}
B_{k1} &= \left\{I_p - \Gamma^k\right\}\mathcal{I}^{-1} + \Gamma^{k-1}L \\
B_{k2} &= \left\{I_p - \Gamma^{k-1}\right\}\mathcal{I}^{-1}.
\end{aligned}
$$

Moreover, if $\Gamma^k \to 0$ as $k \to \infty$, then $\Sigma_k \to \mathcal{I}^{-1}$.

6 APPLICATIONS

6.1 The Bivariate Gaussian Copula

Consider a bivariate distribution with Gaussian copula, a special case of Example 1 in section 3, with $d = 2$. Thus, $\mathbf{y}_i \equiv (y_{i1}, y_{i2})^T$ has CDF $C(F_1(y_1;\alpha_1), F_2(y_2;\alpha_2);\rho), i = 1, \ldots, n$, where $F_j(\cdot;\alpha_j)$ is the marginal CDF of y_{ij}, $j = 1, 2$ and

$$C(u, v; \rho) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)). \tag{12}$$

In this, $\Phi_\rho(\cdot,\cdot)$ is the CDF of a bivariate normal distribution with zero means, variances 1, and correlation coefficient $|\rho| < 1$.

Let $\theta = (\alpha_1^T, \alpha_2^T, \rho)^T$. The log-likelihood function is

$$\ell(\theta) = \ell_w(\theta_1) + \ell_e(\theta_1, \theta_2), \tag{13}$$

where $\theta_1 = (\alpha_1^T, \alpha_2^T)^T$, $\theta_2 = \rho$,

$$\begin{aligned} \ell_w(\theta_1) &= \sum_{i=1}^n \ln f_1(y_{i1};\alpha_1) + \sum_{i=1}^n \ln f_2(y_{i2};\alpha_2), \\ \ell_e(\theta_1, \theta_2) &= -\frac{n}{2}\ln(1-\rho^2) - \frac{\rho}{2(1-\rho^2)}\{\rho A(\theta_1) - 2B(\theta_1)\}. \end{aligned} \tag{14}$$

In this, $A(\theta_1) = \sum_{i=1}^n [Z_{i1}(\alpha_1)^2 + Z_{i2}(\alpha_2)^2]$, $B(\theta_1) = \sum_{i=1}^n Z_{i1}(\alpha_1)Z_{i2}(\alpha_2)$ and $Z_{ij}(\alpha_j) = \Phi^{-1}(F_j(y_{ij};\alpha_j))$. It follows that

$$\Delta(\theta) \equiv \frac{\partial \ell_e(\theta_1,\theta_2)}{\partial \theta_1} = -\frac{\rho}{1-\rho^2}\left\{\rho\frac{\partial A(\theta_1)}{\partial \theta_1} - 2\frac{\partial B(\theta_1)}{\partial \theta_1}\right\} \tag{15}$$

$$\frac{\partial \ell_e(\theta_1,\theta_2)}{\partial \theta_2} = \frac{n\rho}{1-\rho^2} - \frac{1}{(1-\rho^2)^2}\{\rho A(\theta_1) - (1+\rho^2)B(\theta_1)\}. \tag{16}$$

The algorithm proceeds as follows:

STEP 1 Find $\theta_{1,n}^1$ to maximize $\ell_w(\theta_1)$. Note that $\theta_{1,n}^1$ is the MLE of $\theta_{1,0}$ when $\theta_{2,0} = 0$;

Solve $\partial \ell_e(\theta_{1,n}^1, \theta_2)/\partial\theta_2 = 0$ for $\theta_{2,n}^1$

STEP k Solve $\partial \ell_w(\theta_1)/\partial\theta_1 = -\Delta(\theta_n^{k-1})$ for $\theta_{1,n}^k$, with $\Delta(\theta)$ given by (15); and

Solve $\partial \ell_e(\theta_{1,n}^{k-1}, \theta_2)/\partial\theta_2 = 0$ for $\theta_{2,n}^k$, $k = 2, 3, \ldots$.

This algorithm is applicable to the bivariate Gaussian copula model with any marginal distributions and is easy to implement. For example, computation of $\partial A/\partial\theta_1$ involves calculation of the derivatives of $Z_{ij}(\alpha_j)$ and solving $\partial \ell_e(\theta_1^*, \theta_2)/\partial\theta_2 = 0$ leads to the cubic equation

$$\theta_2^3 - \theta_2^2 B(\theta_1^*) + \theta_2 A(\theta_1^*) - B(\theta_1^*) = 0,$$

which has a unique solution lying between $B(\theta_1^*)$ and $[1 \wedge B(\theta_1^*)/\{A(\theta_1^*) - 1\}]$.

Note that the estimate $\theta_n^1 = (\theta_{1,n}^1, \theta_{2,n}^1)$ from STEP 1 is frequently used in practice due to the complexity of computing the MLE. In general, θ_n^1 may not be asymptotically efficient, since STEP 1 ignores the dependence between y_{i1} and y_{i2} in calculating $\theta_{1,n}^1$. Subsequent steps take account of estimates of the dependence parameter $\theta_{2,0}$ leading to more efficient estimates of $\theta_{1,0}$. In the special case where the marginals are normal distributions, or equivalently where the bivariate distribution is a bivariate Gaussian distribution, STEP 1 generates asymptotically efficient estimates (the MLEs).

Example 6.1: Suppose $y_{ij} \sim N(0, \sigma_j^2)$ for $j = 1, 2$. Then one can easily verify that

$$\begin{aligned} \theta_{1,n}^1 &= (n^{-1}\sum_{i=1}^n y_{i1}^2, n^{-1}\sum_{i=1}^n y_{i2}^2)^T \equiv (\sigma_{1,n}^1, \sigma_{2,n}^1)^T \\ \theta_{2,n}^1 &= n^{-1}\sum_{i=1}^n y_{i1}y_{i2}/(\sigma_{1,n}^1\sigma_{2,n}^1), \end{aligned}$$

the MLEs in the bivariate normal with means zero, variances σ_1^2, σ_2^2 and correlation ρ. The algorithm converges in one step.

Example 6.2: Consider exponential marginal distributions with densities $f_j(y_j; \alpha_j) = \alpha_j \exp(-\alpha_j y_j)$, $\alpha_j > 0$, $j = 1, 2$. The likelihood function for the independence model is

$$\ell_w(\theta_1) = [n \ln \alpha_1 - \alpha_1 \sum_{i=1}^n y_{i1}] + [n \ln \alpha_2 - \alpha_2 \sum_{i=1}^n y_{i2}].$$

Let $\bar{y}_j = n^{-1}\sum_{i=1}^{n} y_{ij}$ for $j = 1, 2$ and $\bar{\Delta}_n^k = (\bar{\Delta}_{1,n}^k, \bar{\Delta}_{2,n}^k)'$, where the bar denotes the sample average of Δ (15) evaluated at the updated values. We find that $\theta_{1,n}^1 = \{\bar{y}_1^{-1}, \bar{y}_2^{-1}\}$ and $\theta_{1,n}^k = \{(\bar{y}_1 + \bar{\Delta}_{1,n}^{k-1})^{-1}, (\bar{y}_2 + \bar{\Delta}_{2,n}^{k-1})^{-1}\}$ for $k \geq 2$. As in general, implementation only requires solving a few third order polynomial equations in θ_2. In contrast, the direct computation of the MLE by a Newton-Raphson algorithm is much more difficult to implement. As shown in Appendix B, the observed and expected Fisher information matrices are very complicated, and there are no closed form expressions for the latter.

To examine speed of convergence, we performed a simulation study to compare $n\Sigma_k^{-1}$, the inverse of the estimated asymptotic variance of θ_n^k, to the observed Fisher information over a number of iterations. The distance between two matrices is defined as the maximum of the absolute entrywise differences.

The simulation study considers 9 parameter combinations with $\rho = 0.3, 0.5, 0.7$ and $\alpha_1/\alpha_2 = 10, 5, 1$. Note that the absolute magnitude of $\alpha = (\alpha_1, \alpha_2)$ does not matter due to scale invariance. Two sample sizes, $n = 20$ and $n = 100$ are considered in Figures 1 and 2 respectively. For each parameter setting, 100 replications were run.

Figure 1 displays the average distances of the updated $n\Sigma_k^{-1}$ matrix relative to the observed information matrix over 100 replications at iteration k. When $\rho = 0.3$, the algorithm takes about five iterations to reach the observed Fisher information; when $\rho = 0.5$, ten iterations is typically enough; when $\rho = 0.7$, it can take as many as 200 iterations. This is a result of ℓ_w being based only on the marginals and not taking correlation into account. When the correlation is high, ℓ_w contributes a smaller portion of the information, and more iterations are needed to recover the full information. If ρ is too large (e.g. $\rho = 0.95$), the marginals-based partition of $\ell(\theta)$ no longer works since the information dominance condition is not satisfied. A new ℓ_w is needed that takes some degree of correlation into account.

Figure 2 is very similar to Figure 1, and the convergence rate of the information does not seem to depend much on sample size. Both figures also suggest that the convergence of the

Figure 1: Distances between the updated information and the observed information matrices with sample size $n = 20$ over 100 replications. Note the iteration numbers shown in the bottom row of panels are the last 50 iterations before convergence.

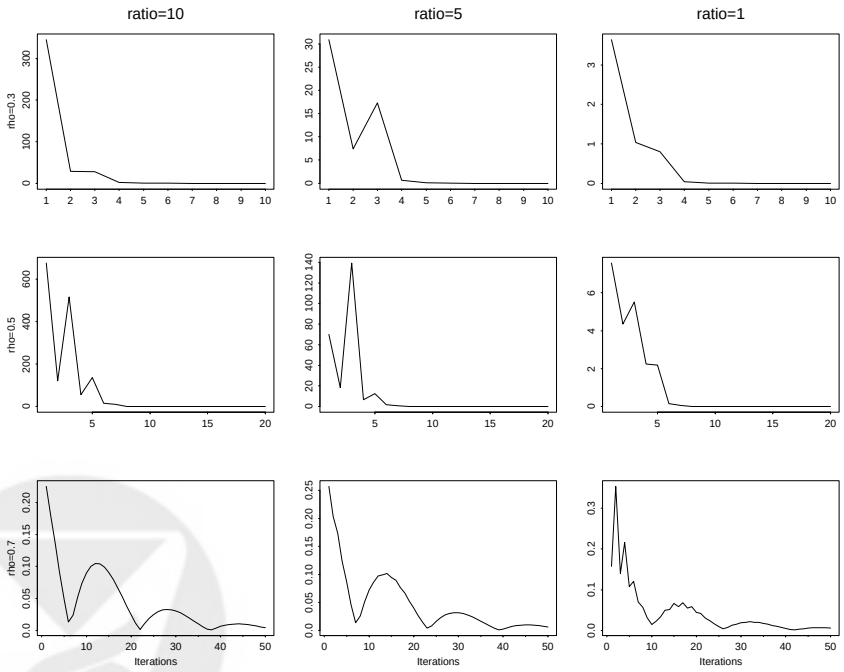
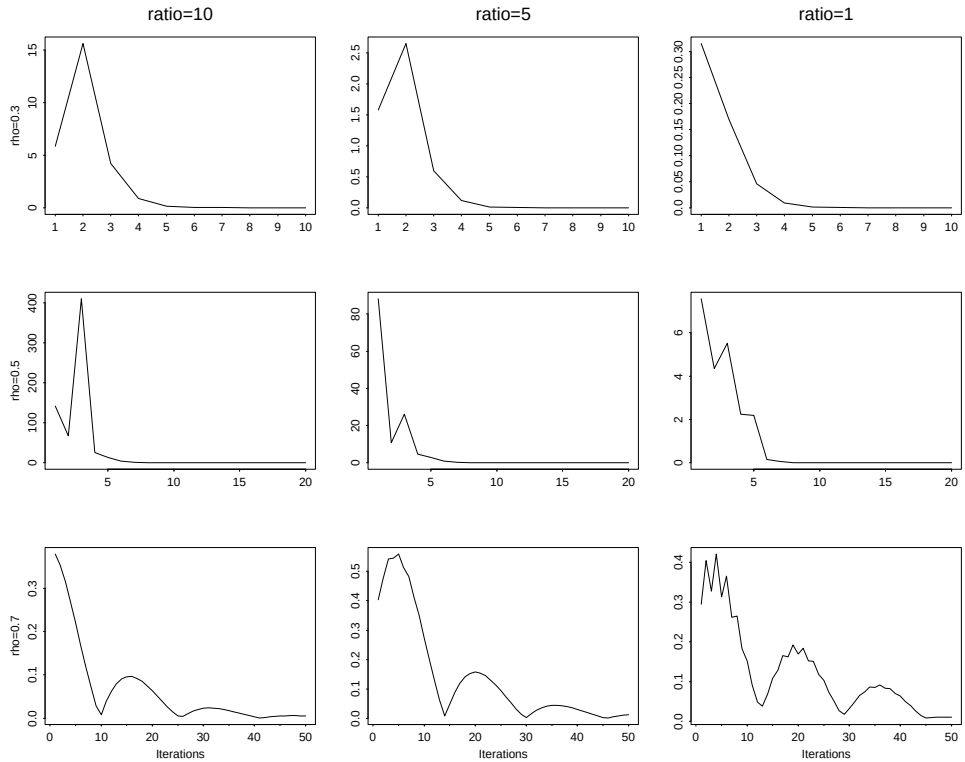


Figure 2: Distances between the updated information and the observed information matrices with sample size $n = 100$ over 100 replications. Note the iteration numbers shown in the bottom row of panels are the last 50 iterations before convergence.



information does not depend much on values of α either.

Table 1 gives the asymptotic relative efficiency (ARE) (first line) and the ratios of sample variances (second line) for the initial $\hat{\theta}_1^0$ and one-step updated estimators $\hat{\theta}_1^1$, and $\hat{\theta}_2^1$ to the sample variance of the MLEs. The ARE is the ratio of the diagonals of $n\Sigma_k^{-1}$ to the diagonals of the inverse of the Fisher information matrix (Appendix B). The Monte Carlo method with sample size $5,000$ was used to evaluate the expectations involved in the information matrix in each replication. We chose $(\alpha_1, \alpha_2) = (1.0, 1.0)$ in this simulation. Other values of the rate parameters with unequal α_1 and α_2 gave similar results.

Table 1: The average ARE's and ratios of the sample variances for the initial and one-step estimators to the sample variance of the MLEs based on 200 simulations.

	$n = 20$					$n = 100$				
	Initial		One-Step			Initial		One-Step		
ρ	α_1^0	α_2^0	α_1^1	α_2^1	ρ^1	α_1^0	α_2^0	α_1^1	α_2^1	ρ^1
0.3	1.093	1.112	1.033	1.029	1.033	1.026	1.012	1.012	1.003	1.033
	1.181	1.045	1.052	1.039	1.183	1.038	1.112	0.998	1.011	1.012
0.6	1.426	1.034	1.001	1.175	1.063	1.517	1.425	1.242	1.062	1.242
	1.156	1.010	1.007	1.076	1.162	1.516	1.505	1.175	1.075	1.298

From this simulation, we find that the MLEs and the one-step estimators for the rate parameters $\alpha_i, i = 1, 2$ are comparable for both small or moderate correlation. On the other hand, the MLE for the correlation parameter ρ is clearly more efficient than the one-step estimation. The current widely used strategy based on the one-step estimation seems reasonable for the estimation of the mean parameters but not for the estimation of the correlation.

6.2 Non-normal random effects models

In this section, we consider a simple non-normal random effects model

$$y_{ij} = x_{ij}^T\beta + \alpha_i + \varepsilon_{ij},\ j = 1, \ldots, m, i = 1, \ldots, n, \tag{17}$$

where the scalar random effects $\alpha_i \sim \sqrt{\omega} t_{(d)}$ has a scaled t distribution with $d > 2$ degrees of freedom and scale parameter $\sqrt{\omega}$. This distribution has mean zero and variance $\eta = \omega d/(d-2)$ with heavier tails than the normal. Parametrization in terms of η gives the density

$$p(\alpha_i|\eta) = \frac{1}{\sqrt{(d-2)\eta} B(\frac{d}{2}, \frac{1}{2})} \left(1 + \frac{\alpha_i^2}{(d-2)\eta}\right)^{-\frac{d+1}{2}}.$$

Linear mixed models with t-distributed random effects have been studied in Bayesian analysis using Markov Chain Monte Carlo methods. See Wakefield et al. (1994) and Wakefield (1995). Pinheiro et al. (2001) developed EM-type algorithms, including ECM, ECME and PX-EM, for maximum likelihood estimation in such a setting. Compared to their methods, the proposed algorithm is much simpler and faster.

As in Example 2, we choose $\alpha_i \sim N(0,\eta)$ and $\varepsilon_{ij} \sim N(0,\sigma), i = 1,\ldots,n, j = 1,\ldots,m$ independent. Let $\theta = (\beta,\eta,\sigma)$ and $\nu = \frac{m\eta}{\sigma+m\eta}$. The second piece $\ell_e(\theta)$ in equation (9) is given by (10), where $\phi(\alpha_i|\mathbf{y}_i,\theta)$ is the density of the univariate normal with mean, $\mu_i^* = \nu\sum_{j=1}^m(y_{ij} - x_{ij}^T\beta)/m$, and variance, $\sigma^* = \sigma\nu/m$.

Let

$$\varphi(\alpha_i|\theta) = \frac{p(\alpha_i|\theta)}{\phi(\alpha_i|\theta)}\phi(\alpha_i|\mathbf{y}_i,\theta).$$

Then, $\ell_e(\theta) = \sum_{i=1}^n \ln\int\varphi(\alpha_i|\theta)d\alpha_i$ and

$$\dot{\ell}_{e(j)}(\theta) = \frac{\partial\ell_e(\theta)}{\partial\theta_j} = \sum_{i=1}^n\left\{\int\varphi(\alpha_i|\theta)d\alpha_i\right\}^{-1}\int\varphi(\alpha_i|\theta)\frac{\partial\ln\varphi(\alpha_i|\theta)}{\partial\theta_j}d\alpha_i, \tag{18}$$

where $\theta_1 = \beta$, $\theta_2 = \eta$, and $\theta_3 = \sigma$. We evaluate the integrals in (18) by the method of Gaussian-Hermite quadrature based on the posterior normal distribution $\phi(\alpha_i|\mathbf{y}_i,\theta)$. Related details are presented in Appendix C.

Let $e_i = \alpha_i - \mu_i^*$ denote the prediction error for the random effect. The derivatives of $\varphi(\alpha_i|\theta)$ are, respectively,

$$\begin{aligned}
\frac{\partial\varphi(\alpha_i|\theta)}{\partial\theta_1} &= -\frac{e_i}{\sigma}\sum_{j=1}^m x_{ij};\\
\frac{\partial\varphi(\alpha_i|\theta)}{\partial\theta_2} &= \frac{1}{2\eta}\left\{\frac{(d+1)\alpha_i^2}{(d-2)\eta+\alpha_i^2} - \frac{\sigma}{\sigma+m\eta} - \frac{\mu_i^{*2}}{\eta}\right\};\\
\frac{\partial\varphi(\alpha_i|\theta)}{\partial\theta_3} &= -\frac{1}{2\sigma}\left\{\nu - \frac{me_i^2}{\sigma} + \frac{2e_i\mu_i^*}{\eta}\right\}.
\end{aligned}$$

For the working normal random effects model, we refer to McCulloch and Searle (Chapter 6, 2001). Some relevant formulas needed in our algorithm are listed as follows. The working

likelihood function in equation (9) is

$$\ell_w(\theta) \quad \propto \quad -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - X_i\beta)^T\Sigma^{-1}(\mathbf{y}_i - X_i\beta) \tag{19}$$

where

$$\Sigma = \sigma I_m + \eta J_m,\ \Sigma^{-1} = \frac{1}{\sigma}\left[I_m - \frac{\nu}{m}J_m\right], |\Sigma| = \sigma^{m-1}(\sigma + m\eta).$$

$J_m = \mathbf{1}\mathbf{1}^T$ is a matrix with all elements equal to 1, and $1-\nu = \frac{\sigma}{\sigma+m\eta}$.

It follows that the working score functions are given by

$$\begin{aligned}
\dot{\ell}_{w(1)} &= \frac{\partial \ell_w(\theta)}{\partial \theta_1} = \sum_{i=1}^{n} X_i^T\Sigma^{-1}(\mathbf{y}_i - X_i\beta), \\
\dot{\ell}_{w(2)} &= \frac{\partial \ell_w(\theta)}{\partial \theta_2} = -\frac{1}{2}\left\{\frac{mn(1-\nu)}{\sigma} - \frac{(1-\nu)^2}{\sigma^2}S_2\right\}, \\
\dot{\ell}_{w(3)} &= \frac{\partial \ell_w(\theta)}{\partial \theta_3} = -\frac{1}{2}\left[\frac{n(m-\nu)}{\sigma} - \frac{1}{\sigma^2}\left\{S_1 - \frac{2\nu}{m}S_2 + \frac{\nu^2}{m}S_2\right\}\right],
\end{aligned}$$

where

$$S_1 = \sum_{i=1}^{n}(\mathbf{y}_i - X_i\beta)^T(\mathbf{y}_i - X_i\beta),\ S_2 = \sum_{i=1}^{n}(\mathbf{y}_i - X_i\beta)^T J_m(\mathbf{y}_i - X_i\beta).$$

Clearly, $\mathrm{E}(S_1) = mn(\sigma+\eta)$ and $\mathrm{E}(S_2) = mn(\sigma + m\eta)$. It is easy to prove that $\mathrm{E}\dot{\ell}_w = 0$, so all working score functions are unbiased. Also, the information matrix $\mathcal{I}_1 = -n^{-1}\mathrm{E}\ddot{\ell}_w$ for the working model is a block diagonal matrix given by

$$\mathcal{I}_1 = \begin{bmatrix} n^{-1}\sum_{i=1}^{n} X_i^T\Sigma^{-1}X_i & 0 & 0 \\ 0 & \frac{m^2(1-\nu)^2}{2\sigma^2} & \frac{m(1-\nu)^2}{2\sigma^2} \\ 0 & \frac{m(1-\nu)^2}{2\sigma^2} & \frac{(m-1)+(1-\nu)^2}{2\sigma^2} \end{bmatrix}.$$

To find the MLE, we proceed as follows:

STEP 1 The consistent initial estimate $\theta_n^1 = (\beta_n^1, \eta_n^1, \sigma_n^1)$ is given by fitting model (17) under a normal distribution $\phi(\cdot|\eta)$, namely by solving $\dot{\ell}_w(\theta) = 0$.

STEP k θ_n^k solves

$$\dot{\ell}_w(\theta) = -\dot{\ell}_e(\theta_n^{k-1}) \tag{20}$$

Note that parameter θ can be entirely updated under ℓ_w, so the asymptotics developed in Section 4 are applicable here.

The solution $\theta_n^k = (\beta_n^k, \eta_n^k, \sigma_n^k)$ to equation (20) can be found using Newton's method or Fisher scoring, or by iteratively solving the following three equations:

$$\beta = \left(\sum_{i=1}^{n} X_i^T \Sigma^{-1} X_i\right)^{-1} \left\{\sum_{i=1}^{n} X_i^T \Sigma^{-1} \mathbf{y}_i + \dot{\ell}_{e(1)}(\theta_n^{k-1})\right\},$$

$$\left\{\frac{2}{mn}\dot{\ell}_{e(2)}(\theta_n^{k-1})\right\}\tau^2 + \tau - \frac{S_2}{mn} = 0,$$

with $\tau = \sigma + m\eta$, and

$$a_0\sigma^2 + a_1\sigma + a_2 = 0,$$

where

$$\begin{aligned} a_0 &= \frac{2}{(m-1)n}\dot{\ell}_{e(3)}(\theta_n^{k-1}) + \frac{2}{m(m-1)n}\dot{\ell}_{e(2)}(\theta_n^{k-1}), \\ a_1 &= -1, \\ a_2 &= \frac{mS_1 - S_2}{m(m-1)n}. \end{aligned}$$

Finally, η is obtained by $\eta = \frac{\tau-\sigma}{m}$.

We conducted a simulation study to investigate how misspecified random effects distributions affect the accuracy and efficiency of parameter estimation. Some studies (e.g. Butler and Louis, 1992; Verbeke and Lesaffre,1997) have revealed that inference on fixed effects is robust against non-normality of random effects. Our simulation results confirm this, but also find that the efficiency of estimates of fixed effects is affected by using a misspecified distribution of random effects.

The simulation is based on the following parameter configurations. First, we chose degrees of freedom $d = 3$ and $d = 20$, where $d = 3$ presents a strong departure from normality whereas $d = 20$ presents little difference from normality. Second, we took $p = 2$ with $\beta_0 = 0.5$ and $\beta_1 = 1.0$. Third, we considered two scenarios for within-cluster correlation: one with $\eta = 1.0, \sigma = 0.25$ (within-cluster correlation of 0.8) and with $\eta = 0.25, \sigma = 1.0$ (within-cluster

Table 2: Simulation results based on $t(3)$ distribution and parameters $\beta_0 = 0.5, \beta_1 = 1.0, \eta = 1.0$, and $\sigma = 0.25$. The within cluster correlation is 0.8. 100 replications are run.

Parameter	Estimate	Iteration	Mean	Empirical std. dev.	Observed std. err.
β_0	naive	0	.5230	.1268	.1554
	MLE	16	.5187	.1101	.1104
β_1	naive	0	.9870	.1732	.2170
	MLE	16	.9714	.1412	.1455
η	naive	0	.8591	.3979	.3845
	MLE	16	.8649	.2113	.2308
σ	naive	0	.2491	.0194	.0182
	MLE	16	.2487	.0191	.0181

correlation of 0.2). Only one covariate was considered with 50 clusters receiving treatment $x = 1$ and the other 50 clusters receiving placebo $x = 0$. We used $n = 100$ clusters, each with $m = 5$ individuals, and 100 replications.

To evaluate the rate of convergence, we monitored the number of iterations required to achieve the MLE, and the averaged number of iterations is reported in Tables 2–5. The algorithm stops when the maximum difference between two consecutive estimates is less than 10^{-4}. In these tables, empirical standard deviation refers to the sample standard deviation and observed standard error refers to the standard error computed from the observed Fisher information given by Theorem 3.

Both naive estimation under the normality assumption for the random effects and maximum likelihood estimation under a t-distribution are reasonable, so far as bias is concerned. However, the standard errors for the fixed effects estimates are affected by the departure from normality. For $d = 3$ and the within cluster correlation of 0.8, based on the observed

Table 3: Simulation results based on $t(20)$ distribution and parameters $\beta_0 = 0.5, \beta_1 = 1.0, \eta = 1.0$, and $\sigma = 0.25$. The within cluster correlation is 0.8. 100 replications are run.

Parameter	Estimate	Iteration	Mean	Empirical std. dev.	Observed std. err.
β_0	naive	0	.4807	.1466	.1474
	MLE	3	.4799	.1445	.1442
β_1	naive	0	.9913	.1847	.2086
	MLE	3	.9938	.1814	.2043
η	naive	0	.9732	.1554	.1598
	MLE	3	.9715	.1514	.1456
σ	naive	0	.2528	.0166	.0185
	MLE	3	.2528	.0166	.0185

Fisher information in Table 2, the MLEs of β_0 and β_1 are 40% to 50% more efficient than the corresponding naive estimates. When the within cluster correlation is 0.2 as in Table 4, the MLEs are 4% to 18% more efficient. From Tables 3 and 5, when $d = 20$, the naive estimation is almost as efficient as the MLE as expected, since $t(20)$ is nearly identical to the normal with the same variance. Similar conclusions hold for the efficiency of the variance components estimates.

7 DISCUSSION

In this paper, we proposed a simple fixed point algorithm that updates an estimator obtained from a simple analysis to obtain the MLE. The choice of the simple analysis is flexible and determined by the specific problem under investigation. We gave three examples where such a likelihood decomposition can arise naturally from the structure of a model itself. We

Table 4: Simulation results based on $t(3)$ distribution and parameters $\beta_0 = 0.5, \beta_1 = 1.0, \eta = 0.25$, and $\sigma = 1.0$. The within cluster correlation is 0.2. 100 replications are run.

Parameter	Estimate	Iteration	Mean	Empirical std. dev.	Observed std. err.
β_0	naive	0	.4991	.0904	.1007
	MLE	7	.5002	.0758	.0965
β_1	naive	0	.9983	.1286	.1449
	MLE	7	1.0022	.1074	.1227
η	naive	0	.2465	.1453	.1521
	MLE	7	.2461	.1231	.1394
σ	naive	0	1.0010	.0791	.0728
	MLE	7	1.0023	.0718	.0724

anticipate there are other settings in which this idea is applicable.

In the study of estimation for bivariate copula distributions, we found that neither the naive estimation based on the marginals nor the one-step updated estimation widely used in the literature, gives satisfactory efficiency for correlation parameters. Our method provides an easy way to obtain the MLE in such a setting. In the study of nonnormal random effects models with t-distributed random effects we found that the misspecified distribution for the random effects can significantly reduce estimation efficiency, especially when the departure from normality or the within class correlation is large.

Finally, as with every numerical algorithm, our algorithm requires a condition for convergence. The key condition, referred to as the information dominance, is conceptually intuitive, and it guides the choice of a model for the simple analysis. Approaches that relax this condition are currently under our investigation.

Table 5: Simulation results based on $t(20)$ distribution and parameters $\beta_0 = 0.5, \beta_1 = 1.0, \eta = 0.25$, and $\sigma = 1.0$. The within cluster correlation is 0.2. 100 replications are run.

Parameter	Estimate	Iteration	Mean	Empirical std. dev.	Observed std. err.
β_0	naive	0	.5075	.0976	.0953
	MLE	2	.5075	.0968	.0949
β_1	naive	0	1.0046	.1575	.1359
	MLE	2	1.0042	.1566	.1352
η	naive	0	.2378	.0765	.0685
	MLE	2	.2378	.0765	.0657
σ	naive	0	1.0091	.0791	.0738
	MLE	2	1.0091	.0791	.0738

REFERENCES

Abramowitz, M. and Stegun, I. (eds.) (1964). *Handbook of Mathematical Functions.* National Bureau of Standards, Washington, D.C.

Burden, R.L. and Faires J. D. (1997). *Numerical Analysis, 6th Ed.* Books/Cole Publishing Company.

Bauwens, L. and Veredas, D. (2003). "The stochastic conditional duration model: a latent factor model for the analysis of financial duration." *Journal of Econometrics* (to appear).

Breslow, N.E. and Clayton, D.G. (1993). "Approximate inference in generalized linear mixed models." *Journal of the American Statistical Association* **88**, 9-25.

Butler, S.M. and Louis, T.A. (1992). "Random effects models with nonparametric priors."

Statistics in Medicine **11**, 1981-2000.

Durbin, J. and Koopman S.J. (1997). "Monte Carlo maximum likelihood estimation for non-Gaussian state space models." *Biometrika* **84**, 669-684

Engle, R.F. and Russell, J. R. (1998). "Autoregressive conditional duration: a new model for irregularly spaced transaction data." *Econometrica* **66**, 1127-1162.

Geyer, C.J. and Thompson, E.A. (1992). "Constrained Monte Carlo maximum likelihood for dependent data." *Journal of the Royal Statistical Society, Series B* **54**, 657-699.

Joe, H. (1997). *Multivariate Models and Dependent Concepts*, London: Chapman and Hall.

Liang, K.-Y. and Zeger, S.L. (1986). "Longitudinal data analysis using generalized linear models." *Biometrika* **73**, 13–22.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models.* New York: Wiley.

McLeish, D.L. and Small, C.G. (1988). *The Theory and Applications of Statistical Inference Functions.* Springer Lecture Notes in Statistics #44, New York: Springer.

Pinheiro, P.C., Liu, C. and Wu, Y. N. (2001). "Efficient algorithm for robust estimation in linear mixed-effects models using the multivariate t distribution." *Journal of Computational and Graphical Statistics* **10**, 249–276.

Song, P.X.-K. (2000). "Multivariate dispersion models generated from Gaussian copula." *Scandinavian Journal of Statistics* **27**, 305–320.

Verbeke, G. and Lesaffre. E. (1997). "The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data." *Computational Statistics and Data Analysis* **23**, 541-556.

Wakefield, J.C., Smith, A.F.M., Racine-Poon, A. and Gelfand, A.E. (1994). "Bayesian analysis of linear and nonlinear population models using the Gibbs sampler." *Applied Statistics* **43**, 201-222.

Wakefield, J.C. (1996). "The Bayesian approach to population pharmacokinetic models." *Journal of the American Statistical Association* **91**, 61-76.

Zhang, D. and Davidian, M. (2001). "Linear mixed models with flexible distributions of random effects for longitudinal data." *Biometrics* **57**, 795-802.

APPENDIX A: PROOFS OF THEOREMS

We use the notation of Sections 4 and 5.

Proof of Theorem 1: Suppose θ_n^1 is consistent, so that $\theta_n^1 = \theta_0 + o_p(1)$. Since θ_n^2 satisfies the equation, $\dot{\ell}_w(\theta_n^2) + \dot{\ell}_e(\theta_n^1) = 0$, the Mean Value Theorem implies that

$$0 = \dot{\ell}_w(\theta_n^2) + \dot{\ell}_e(\theta_n^1) = \dot{\ell}_w(\theta_0) + \ddot{\ell}_w(\theta_n^*)(\theta_n^2 - \theta_0) + \dot{\ell}_e(\theta_n^1), \tag{21}$$

where θ_n^* lies between θ_n^2 and θ_0. It follows that

$$\theta_n^2 - \theta_0 = [-n^{-1}\ddot{\ell}_w(\theta_n^*)]^{-1}[n^{-1}\dot{\ell}_w(\theta_0) + n^{-1}\dot{\ell}_e(\theta_n^1)] \to 0$$

since, under the regularity conditions, $[-n^{-1}\ddot{\ell}_w(\theta_n^*)]$ is bounded and

$$\text{plim}\, n^{-1}\dot{\ell}_w(\theta_0) + n^{-1}\dot{\ell}_e(\theta_n^1) = \lim n^{-1}\mathrm{E}_{\theta_0}\dot{\ell}(\theta_0) = 0.$$

Proof of Theorems 2 and 3: A Taylor expansion of $\dot{\ell}_w$ and $\dot{\ell}_e$ about θ_0 gives

$$\begin{aligned} 0 &= n^{-1/2}\left[\dot{\ell}_w(\theta_n^k) + \dot{\ell}_e(\theta_n^{k-1})\right] \\ &= n^{-1/2}\dot{\ell}_w + n^{-1/2}\dot{\ell}_e + \left(n^{-1}\ddot{\ell}_w\right)\left[\sqrt{n}(\theta_n^k - \theta_0)\right] + \left(n^{-1}\ddot{\ell}_e\right)\left[\sqrt{n}(\theta_n^{k-1} - \theta_0)\right] + o_p(1) \end{aligned}$$

where we have suppressed the dependence on θ_0 in $\dot{\ell}_w = \dot{\ell}_w(\theta_0)$, $\dot{\ell}_e$, $\ddot{\ell}_w$, and $\ddot{\ell}_e$. Hence, for $k = 2, \ldots$, we have

$$\sqrt{n}(\theta_n^k - \theta_0) = \left(-n^{-1}\ddot{\ell}_w\right)^{-1}\left(n^{-1/2}\dot{\ell}_w\right) + \left(-n^{-1}\ddot{\ell}_w\right)^{-1}\left(n^{-1}\ddot{\ell}_e\right)\left[\sqrt{n}(\theta_n^{k-1} - \theta_0)\right] + o_p(1). \quad (22)$$

Iterating (22) yields

$$\begin{aligned}\sqrt{n}(\theta_n^k - \theta_0) &= \sum_{j=0}^{k-1} \tau_n^j \left(-n^{-1}\ddot{\ell}_w\right)^{-1}\left(n^{-1/2}\dot{\ell}_w\right) + \sum_{j=0}^{k-2} \tau_n^j \left(-n^{-1}\ddot{\ell}_w\right)^{-1}\left(n^{-1/2}\dot{\ell}_e\right) + o_p(1) \\ &= \left[A_n^k\right]^T \begin{pmatrix} n^{-1/2}\dot{\ell}_w(\theta_0) \\ n^{-1/2}\dot{\ell}_e(\theta_0) \end{pmatrix} + o_p(1), \end{aligned} \quad (23)$$

where

$$\left[A_n^k\right]^T = \left(\left[-n^{-1}\ddot{\ell}\right]\left[I_p - \tau_n^k\right] \;\middle|\; \left[-n^{-1}\ddot{\ell}\right]\left[I_p - \tau_n^{k-1}\right]\right).$$

Since A_n^k converges in probablitily to A_k and

$$\begin{bmatrix} n^{-1/2}\dot{\ell}_w \\ n^{-1/2}\dot{\ell}_e \end{bmatrix} \to N(0, \Omega)$$

in distribution, it follows that

$$\sqrt{n}(\theta_n^k - \theta_0) \to N(0, \Sigma_k).$$

where $\Sigma_k = A_k^T \Omega A_k$ as stated in Theorem 3.

Proof of Theorem 4: We suppose that consistency of θ_n^k holds and sketch the remainder of the proof. From Step k, we have

$$\dot{\ell}_{w(1)}(\theta_{1,n}^k) + \dot{\ell}_{e(1)}(\theta_{1,n}^{k-1}, \theta_{2,n}^{k-1}) = 0,$$

and

$$\dot{\ell}_{e(2)}(\theta_{1,n}^{k-1}, \theta_{2,n}^k) = 0.$$

A Taylor expansion about $\theta_0 = (\theta_{1,0}, \theta_{2,0})$ gives

$$\dot{\ell}_{w(1)}(\theta_{1,0}) + \ddot{\ell}_{w(11)}(\theta_{1,0})(\theta_{1,n}^k - \theta_{1,0}) + \dot{\ell}_{e(1)}(\theta_0) + \ddot{\ell}_{e(11)}(\theta_0)(\theta_{1,n}^{k-1} - \theta_{1,0}) + \ddot{\ell}_{e(12)}(\theta_0)(\theta_{2,n}^{k-1} - \theta_{2,0}) = 0$$

and

$$\dot{\ell}_{e(2)}(\theta_0) + \ddot{\ell}_{e(21)}(\theta_0)(\theta_{1,n}^{k-1} - \theta_{1,0}) + \ddot{\ell}_{e(22)}(\theta_0)(\theta_{2,n}^{k} - \theta_{2,0}) = 0$$

Rewriting these in a matrix form gives the recurrence relationship

$$\sqrt{n}(\theta_n^k - \theta_0) = D_n^{-1} T_n \sqrt{n}(\theta_n^{k-1} - \theta_0) + D_n^{-1}\left\{n^{-1/2}\dot{\ell}(\theta_0)\right\}. \tag{24}$$

Iterating equation (24) yields

$$\sqrt{n}(\theta_n^k - \theta_0) = \left(D_n^{-1}T_n\right)^{k-1}\sqrt{n}(\theta_n^1 - \theta_0) + \sum_{j=0}^{k-2}\left(D_n^{-1}T_n\right)^j D_n^{-1}\frac{1}{\sqrt{n}}\dot{\ell}(\theta_0).$$

Since $\theta_{1,n}^1$ is used to define $\theta_{2,n}^1$, a Taylor expansion at Step 1 leads to

$$\begin{aligned}
\sqrt{n}(\theta_n^1 - \theta_0) &= \begin{bmatrix} -n^{-1}\ddot{\ell}_{w(11)} & 0 \\ -n^{-1}\ddot{\ell}_{e(21)} & -n^{-1}\ddot{\ell}_{e(22)} \end{bmatrix}^{-1} \begin{bmatrix} n^{-1/2}\dot{\ell}_{w(1)} \\ n^{-1/2}\dot{\ell}_{e(2)} \end{bmatrix} \\
&= D_n^{-1}\begin{bmatrix} n^{-1/2}\dot{\ell}_{w(1)} \\ n^{-1/2}\dot{\ell}_{e(2)} \end{bmatrix} + L_n \begin{bmatrix} n^{-1/2}\dot{\ell}_{w(1)} \\ n^{-1/2}\dot{\ell}_{e(2)} \end{bmatrix}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\sqrt{n}(\theta_n^k - \theta_0) &= \sum_{j=0}^{k-1}\left(D_n^{-1}T_n\right)^j D_n^{-1}\begin{bmatrix} n^{-1/2}\dot{\ell}_{w(1)} \\ n^{-1/2}\dot{\ell}_{e(2)} \end{bmatrix} + \sum_{j=0}^{k-2}\left(D_n^{-1}T_n\right)^j D_n^{-1}\begin{bmatrix} n^{-1/2}\dot{\ell}_{e(1)} \\ 0 \end{bmatrix} \\
&\quad + \left(D_n^{-1}T_n\right)^{k-1} L_n \begin{bmatrix} n^{-1/2}\dot{\ell}_{w(1)} \\ n^{-1/2}\dot{\ell}_{e(2)} \end{bmatrix} \\
&= \left\{I_p - \left(D_n^{-1}T_n\right)^k\right\}\{I_p - D_n^{-1}T_n\}^{-1}D_n^{-1}\begin{bmatrix} n^{-1/2}\dot{\ell}_{w(1)} \\ n^{-1/2}\dot{\ell}_{e(2)} \end{bmatrix} \\
&\quad + \left\{I_p - \left(D_n^{-1}T_n\right)^{k-1}\right\}\{I_p - D_n^{-1}T_n\}^{-1}D_n^{-1}\begin{bmatrix} n^{-1/2}\dot{\ell}_{e(1)} \\ 0 \end{bmatrix} \\
&\quad + \left(D_n^{-1}T_n\right)^{k-1} L_n \begin{bmatrix} n^{-1/2}\dot{\ell}_{w(1)} \\ n^{-1/2}\dot{\ell}_{e(2)} \end{bmatrix}.
\end{aligned}$$

Note that $\{I_p - D_n^{-1}T_n\}^{-1}D_n^{-1} = (D_n - T_n)^{-1} = \left\{-n^{-1}\ddot{\ell}(\theta_0)\right\}^{-1}$.

It follows that, as $n \to \infty$,

$$\begin{aligned}
\sqrt{n}(\theta_n^k - \theta_0) &= \left[\left\{I_p - \left(D_n^{-1}T_n\right)^k\right\}\left\{-n^{-1}\ddot{\ell}(\theta_0)\right\}^{-1} + \left(D_n^{-1}T_n\right)^{k-1} L_n\right]\begin{bmatrix} n^{-1/2}\dot{\ell}_{w(1)} \\ n^{-1/2}\dot{\ell}_{e(2)} \end{bmatrix} \\
&\quad + \left\{I_p - \left(D_n^{-1}T_n\right)^{k-1}\right\}\left\{-n^{-1}\ddot{\ell}(\theta_0)\right\}^{-1}\begin{bmatrix} n^{-1/2}\dot{\ell}_{e(1)} \\ 0 \end{bmatrix} \\
&\to N(0, \Sigma_k),
\end{aligned}$$

where Σ_k is defined in the statement of the theorem.

Under the condition that $\Gamma^k \to 0$, as $k \to \infty$, the algorithm leads to the following asymptotic variance-covariance matrix,

$$\begin{aligned}
\Sigma_\infty &= \mathcal{I}^{-1}\begin{bmatrix} n^{-1}\mathrm{E}(\dot{\ell}_{w(1)} + \dot{\ell}_{e(1)})(\dot{\ell}_{w(1)} + \dot{\ell}_{e(1)})^T & n^{-1}\mathrm{E}(\dot{\ell}_{w(1)} + \dot{\ell}_{e(1)})\dot{\ell}_{e(2)}^T \\ n^{-1}\mathrm{E}\dot{\ell}_{e(2)}(\dot{\ell}_{w(1)} + \dot{\ell}_{e(1)})^T & n^{-1}\mathrm{E}\dot{\ell}_{e(2)}\dot{\ell}_{e(2)}^T \end{bmatrix}\mathcal{I}^{-1} \\
&= \mathcal{I}^{-1}\left\{\lim_n n^{-1}\mathrm{E}\dot{\ell}(\theta_0)\dot{\ell}^T(\theta_0)\right\}\mathcal{I}^{-1} \\
&= \mathcal{I}^{-1},
\end{aligned}$$

where $\mathcal{I}$ is the Fisher information.

Appendix B: Fisher Information for the Bivariate Gaussian Copula

The Fisher information is symmetric with elements $I_{ij} = -\frac{1}{n}\mathrm{E}(\partial S_i/\partial\theta_j)$ where $\theta^T = (\alpha_1, \alpha_2, \rho)$. It is easy to verify that

$$\begin{aligned}
I_{11} &= \alpha_1^{-2} + \rho(1-\rho^2)^{-1}\mathrm{E}\left[\rho\dot{Z}_1^2 + \rho Z_1\ddot{Z}_1 - Z_2\ddot{Z}_1\right] \\
I_{12} &= -\rho(1-\rho^2)^{-1}\mathrm{E}\left[\dot{Z}_1\dot{Z}_2\right]
\end{aligned}$$

$$
\begin{aligned}
I_{13} &= \rho(1-\rho^2)^{-1}\mathrm{E}\left[Z_1\dot{Z}_1\right] \\
I_{22} &= \alpha_2^{-2} + \rho(1-\rho^2)^{-1}\mathrm{E}\left[\rho(\dot{Z}_2^2 + \rho Z_2\ddot{Z}_2 - Z_1\ddot{Z}_2\right] \\
I_{23} &= \rho(1-\rho^2)^{-1}\mathrm{E}\left[Z_2\dot{Z}_2\right] \\
I_{33} &= (1+\rho^2)(1-\rho^2)^{-2}.
\end{aligned}
$$

Also,

$$
\begin{aligned}
Z_j &= \Phi^{-1}(F_j(y_j;\alpha_j)) = \Phi^{-1}\left(1-\exp(-\alpha_j y_j)\right),\ j=1,2 \\
\dot{Z}_j &= y_j\exp(-\alpha_j y_j)[\phi(Z_j(\alpha_j)]^{-1} \\
\ddot{Z}_j &= -y_j^2\exp(-\alpha_j y_j)[\phi(Z_j(\alpha_j)]^{-1} - [\dot{Z}_j]^2 Z_j.
\end{aligned}
$$

Appendix C: Gaussian-Hermite Quadrature

The Gaussian-Hermite quadrature method is used to evaluate the integrals in equation (18). For convenience, we suppress the index i. Let

$$
w(\alpha|\eta) = \frac{p(\alpha|\eta)}{\phi(\alpha|\eta)}, \quad \text{and} \quad \widetilde{w}(\alpha|\theta) = w(\alpha|\eta)\frac{\partial \ln\varphi(\alpha|\theta)}{\partial\theta_j}.
$$

Then, the Gaussian-Hermite quadrature gives the following results.

$$
\int \varphi(\alpha|\theta)d\alpha = \frac{1}{\pi}\sum_j w(\mu^* + \sqrt{2\sigma^*}a_j|\theta)z_j,
$$

and

$$
\int \varphi(\alpha|\theta)\frac{\partial \ln\varphi(\alpha|\theta)}{\partial\theta_j}d\alpha = \frac{1}{\pi}\sum_j \widetilde{w}(\mu^* + \sqrt{2\sigma^*}a_j|\theta)z_j
$$

where a_j and z_j are respectively the abscissae and weight factors given in Abramowitz and Stegun (1965, page 924) and in this application $\mu^* = \frac{\eta}{\sigma+m\eta}\sum(\mathbf{y}_i - X_i\beta)$ and $\sigma^* = \frac{\sigma\eta}{\sigma+m\eta}$.