Statistica Sinica **21** (2011), 279-305

COMPOSITE LIKELIHOOD FOR TIME SERIES MODELS WITH A LATENT AUTOREGRESSIVE PROCESS

Chi Tim Ng, Harry Joe, Dimitris Karlis and Juxin Liu

Hong Kong Polytechnic University, University of British Columbia, Athens University of Economics and Business, and University of Saskatchewan

Abstract: Consistency and asymptotic normality properties are proved for various composite likelihood estimators in a time series model with a latent Gaussian autoregressive process. The proofs require different techniques than for clustered data with the number of clusters going to infinity. The composite likelihood estimation method is applied to a count time series consisting of daily car accidents with weather related covariates. A simulation study for the count time series model shows that the performance of composite likelihood estimator is better than Zeger's moment-based estimator, and the relative efficiency is high with respect to approximate maximum likelihood.

Key words and phrases: Asymptotic normality, consistency, count data, Gauss-Hermite quadrature, pairwise likelihood, random effects.

1. Introduction

In recent years, composite likelihood methods, based on sum of log-likelihoods of low-dimensional marginal and conditional densities, have been used for many models for which maximum likelihood estimation is computationally too difficult; see Varin (2008) for an excellent review of the area. In this paper, we study composite likelihood estimation methods for time series models with a latent Gaussian autoregressive process. This is a class of models for which the likelihood consists of a high-dimensional integral.

We consider the data to be of the form $(Y_t, \mathbf{X}_t)$, $t = 1, \ldots, n$, where Y_t is the response variable at time t and $\mathbf{X}_t$ is the $(r+1)$-dimensional vector of covariates (first element is 1 for the intercept) at time t. The Y_t are assumed to be conditionally independent given a latent process $\{\Lambda_t : t = 1, \ldots, n\}$. Using conventional notation for densities with random variables indicated in the subscripts, the joint density of $\{Y_t\}$ is

$$\int \Big\{ \prod_{i=1}^{n} f_{Y_t|\Lambda_t}(y_t|\lambda_t) \Big\} f_{\Lambda_1,\ldots,\Lambda_n}(\lambda_1, \ldots, \lambda_n)\, d\lambda_1 \cdots d\lambda_n.$$

We make further assumptions on $\{Y_t\}$ and $\{\Lambda_t\}$, and consider three cases: Y_t real, Y_t non-negative integer, and Y_t binary 0/1. We assume that the Y_t are exponential family random variables with (conditional) probability density or mass functions:

$$[Y_t|\Lambda_t = \lambda] \ \sim \ \xi(y;\lambda) = \exp\left\{a(\lambda)T(y) + b(\lambda) + S(y)\right\} . \tag{1.1}$$

The parameter λ lies in the set of positive reals or all real numbers depending on the model of interest; see special cases given below. Here, the Λ_t are linked to the covariates via

$$\log\Lambda_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_r X_{rt} + \eta_t = \boldsymbol{\beta}^T \mathbf{X}_t + \eta_t \,. \tag{1.2}$$

Models with many parameters for latent processes (or random effects) become nearly non-identifiable, so we make an assumption that the residuals η_t are modeled by a Gaussian AR(p) process for a small positive integer p:

$$\eta_t = \phi_1\eta_{t-1} + \cdots + \phi_p\eta_{t-p} + V_t \,, \tag{1.3}$$

where $\{V_t\}$ is an independent Gaussian sequence with mean 0 and variance σ_V^2. We are interested in the estimation of the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_V, \phi_1, \ldots, \phi_p)$.

Below are some examples of the models of Y_t.

1. Autogressive stochastic volatility (ARSV) model for financial time series: Y_t normal with mean 0 and random variance/volatility $\sigma_t^2 = \Lambda_t$: $\lambda = \sigma^2$, $a(\lambda) = -\sigma^{-2}/2 = -\lambda^{-1}/2$, $T(y) = y^2$, $b(\lambda) = -\log\sigma = -(\log\lambda)/2$, $S(y) = 0$. Without covariates, different estimation methods for this model has been studied in Harvey, Ruiz and Shephard (1994) and Sandmann and Koopman (1998), among others.
2. Poisson with log link function: Y_t Poisson with random mean Λ_t: $a(\lambda) = \log\lambda$, $T(y) = y$, $b(\lambda) = -\lambda$, $S(y) = -\log\Gamma(y+1)$. This model was used in Zeger (1988) for count time series data.
3. Bernoulli with logit link function: Y_t Bernoulli with random mean $\pi_t = \Lambda_t/(1+\Lambda_t)$: $\lambda = \pi/(1-\pi)$, $a(\lambda) = \log\lambda$, $T(y) = y$, $b(\lambda) = \log(1-\pi) = -\log(1+\lambda)$, $S(y) = 0$.

The likelihood of the models based on (1.1)−(1.3) involve an n-fold integral, so that computation of the maximum likelihood estimator is difficult. However low-dimensional marginal densities such as for (Y_j, Y_{j+m}) or $(Y_j, \ldots, Y_{j+m})$, with $1 \le j \le n-m$ and m a small positive integer can be numerically computed with (adaptive) Gauss-Hermite quadrature or the Laplace approximation (see Pinheiro and Chao (2006); Joe (2008)).

Many applications of composite likelihood methods have been for clustered data, where the proofs of the asymptotics (as number of clusters goes to infinity) use the theory of estimating equations. However for composite likelihood methods applied to a single time series, the proofs of the asymptotics are harder. For a model specified via (1.1)−(1.3), we provide proofs, with some novel techniques, of asymptotic results for composite likelihood estimation. In addition, for the special case where (1.1) is Poisson, we obtain some efficiency results for composite likelihood estimators and the moment-based estimator of Zeger (1988). For the ARSV financial time series model with autoregressive order $p = 1$, the efficiency of composite likelihood methods based on bivariate margins up to lag m decreases as the latent autocorrelation ϕ_1 increases toward 1 (Qu (2008)).

Our main application of (1.1)−(1.3) in Section 5 is for some accident count data time series. With (1.1) being Poisson, the resulting time series model for counts has appeared in Zeger (1988), Chan and Ledolter (1995), Jung, Kukuk and Liesenfeld (2006) with various estimation methods, but not composite likelihood. For count time series, there are other classes of models; see Weiß (2008) for a survey of models such as integer-autoregressive (INAR) models based on thinning operators. As a brief comparison, models based on latent Gaussian processes allow more flexibility in serial dependence patterns including negative dependence, and INAR-type models allow more flexible univariate margins but with restricted types of positive serial dependence. The maximum lag 1 serial correlation depends on the marginal distribution and mean of Y_t, whereas INAR-type models can usually reach a lag 1 serial correlation of 1 in the stationary case.

We outline the remainder of the paper. Section 2 has descriptions of the composite likelihoods that we use. Section 3 has the asymptotic covariance matrices of the composite likelihood estimators and statements of theorems for consistency and asymptotic normality. Appendices A and B contain the proofs. Section 4 summarizes our implementation of Zeger's moment-based estimation method. Section 5 has the example with an accident count data time series. Section 6 summarizes a simulation study to compare composite likelihood estimation with Zeger's method and approximate maximum likelihood via MCMC in WinBUGS. Section 7 concludes with some discussion.

2. Composite Likelihood

A composite likelihood function can be constructed in several ways because there are many choices for the marginal distributions. If all the autocorrelations of $\{\eta_t\}$ up to lag m are involved in the marginal density functions, two ways of constructing the composite likelihood function are given below.

One way is to consider $(m+1)$-variate marginals. We define the $(m+1)$-dimensional multivariate composite log-likelihood (MCL) as

$$Q_n(\boldsymbol{\theta}) = Q_{n:m}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{j=1}^{n-m} q_{j:m}(Y_j, Y_{j+1}, \ldots, Y_{j+m}; \boldsymbol{\theta}), \tag{2.1}$$

where $q_{j:m}(\cdot;\boldsymbol{\theta}) = \log f_{Y_j,Y_{j+1},\ldots,Y_{j+m}}(\cdot;\boldsymbol{\theta})$ and $f_{Y_j,Y_{j+1},\ldots,Y_{j+m}}$ is the unconditional joint density of the $(m+1)$ random variables $Y_j, Y_{j+1}, \ldots, Y_{j+m}$ for $j = 1, 2, \ldots$. The value of $\hat{\boldsymbol{\theta}}$ that maximizes $Q_{n:m}$ is called the MCL or MCL$(m+1)$ estimator.

An alternative approach is to consider the bivariate margins of observations that are adjacent or nearly adjacent. The pairwise log-likelihood or bivariate composite log-likelihood (BCL), up to lag m, is

$$Q_n(\boldsymbol{\theta}) = Q_{nm}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{j=1}^{n-m}\sum_{\ell=1}^{m} q_{j\ell}(Y_j, Y_{j+\ell}; \boldsymbol{\theta}), \tag{2.2}$$

where $q_{j\ell}(y,z;\boldsymbol{\theta}) = \log f_{Y_j,Y_{j+\ell}}(y,z;\boldsymbol{\theta})$ and $f_{Y_j,Y_{j+\ell}}$ is the unconditional joint density function of the random variables Y_j and $Y_{j+\ell}$ for $j = 1, 2, \ldots$ and $\ell = 1, 2, \ldots$. The value of $\hat{\boldsymbol{\theta}}$ that maximizes Q_{nm} is called the BCL or BCL(m) estimator (BCL(1) is the same as MCL(2), and BCL(2) is different from trivariate composite likelihood or MCL(3)). The use of bivariate margins of pairs with small lags for models that are nearly Markovian is studied in Varin and Vidoni (2006) and Joe and Lee (2009). If the dependence is decreasing with lag, then intuitively we can use a subset of pairs with lags $\leq m$ (cardinality $O(n)$) instead of all pairs (cardinality $O(n^2)$) in a composite likelihood.

We use notation $Q_{n:m}, Q_{nm}$ if we have to distinguish (2.1) and (2.2), and Q_n for results that cover both cases.

The above density functions f, and their derivatives with respect to the parameters, are given in the subsequent subsections. Throughout this paper, we assume that the data generating process is obtained from the model with $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. For the proofs of asymptotic results, we let Θ be a compact region containing $\boldsymbol{\theta}^0$.

2.1. Marginals for MCL

Let $\boldsymbol{\alpha} = (\gamma_0, \gamma_1, \ldots, \gamma_m)$ be a given $(m+1)$-dimensional vector for the autocovariances of (1.3), and let

$$\boldsymbol{\Sigma}_{0m} = \boldsymbol{\Sigma}_{0m}(\boldsymbol{\alpha}) = \begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_m \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_m & \gamma_{m-1} & \cdots & \gamma_0 \end{pmatrix} \tag{2.3}$$

be the Toeplitz matrix with these autocovariances. For a function $\psi(\mathbf{y},\mathbf{z})$, where $\mathbf{y}$ and $\mathbf{z}$ are $(m+1)$-dimensional, and with $\boldsymbol{\eta}\sim N(\mathbf{0},\boldsymbol{\Sigma}_{0m}(\boldsymbol{\alpha}))$, write

$$\mathrm{E}\,^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}\psi(\mathbf{y},\boldsymbol{\eta})=\frac{1}{(2\pi)^{(m+1)/2}|\boldsymbol{\Sigma}_{0m}|^{1/2}}\int\psi(\mathbf{y},\mathbf{z})\exp\left\{-\tfrac{1}{2}\mathbf{z}^T\boldsymbol{\Sigma}_{0m}^{-1}\mathbf{z}\right\}d\mathbf{z}.$$

The density function of $(Y_j,Y_{j+1},\dots,Y_{j+m})$ is

$$f_{j:m}(\mathbf{y};\boldsymbol{\theta})=f_{Y_j,Y_{j+1},\dots,Y_{j+m}}(\mathbf{y};\boldsymbol{\theta})=\mathrm{E}\,^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}h_j(\mathbf{y},\boldsymbol{\eta};\boldsymbol{\beta}),\tag{2.4}$$

where

$$h_j(\mathbf{y},\boldsymbol{\eta};\boldsymbol{\beta})=\prod_{i=1}^{m+1}\xi\left(y_i,\exp\{\boldsymbol{\beta}^T\mathbf{X}_{j+i-1}+\eta_i\}\right).\tag{2.5}$$

For a function $\zeta(\mathbf{y})$ and integer $j=1,2,\dots,$ write

$$\mathrm{E}\,^{\mathbf{Y}}\zeta(Y_j,Y_{j+1},\dots,Y_{j+m})=\int\zeta(\mathbf{y})f_{j:m}(\mathbf{y};\boldsymbol{\theta}^0)\,d\mathbf{y}.$$

For any function $\psi(\mathbf{y},\mathbf{z})$, define

$$\begin{aligned}&\mathrm{E}\,^{\mathbf{Y},\boldsymbol{\eta}}\psi(Y_j,Y_{j+1},\dots,Y_{j+m},\eta_1,\eta_2,\dots,\eta_m)\\&=\frac{1}{(2\pi)^{(m+1)/2}|\boldsymbol{\Sigma}_{0m}|^{1/2}}\int\psi(\mathbf{y},\mathbf{z})h_j(\mathbf{y},\mathbf{z};\boldsymbol{\beta})\exp\left\{-\tfrac{1}{2}\mathbf{z}^T\boldsymbol{\Sigma}_{0m}^{-1}\mathbf{z}\right\}d\mathbf{z}d\mathbf{y}.\end{aligned}$$

If Y in (1.1) is discrete, the integration sign for $\mathbf{y}$ should be replaced by the summation sign. For simplicity, only integration signs are used below.

2.2. Marginals for BCL

As in the preceding subsection, let $\boldsymbol{\alpha}=(\gamma_0,\gamma_1,\dots,\gamma_m)$ be a given $(m+1)$-dimensional vector. Let

$$\boldsymbol{\Sigma}_\ell=\begin{pmatrix}\gamma_0&\gamma_\ell\\\gamma_\ell&\gamma_0\end{pmatrix},\tag{2.6}$$

and let $\boldsymbol{\alpha}_\ell=(\gamma_0,\gamma_\ell)$. For any 4-dimensional function $\psi(y,y',z,z')$, and $(\eta,\eta')^T\sim N(\mathbf{0},\boldsymbol{\Sigma}_\ell)$, write

$$\mathrm{E}\,^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}_\ell}\psi(y,y',\eta,\eta')=\frac{1}{2\pi|\boldsymbol{\Sigma}_\ell|^{1/2}}\int\psi(y,y',z,z')\exp\left\{-\tfrac{1}{2}(z,z')\boldsymbol{\Sigma}_\ell^{-1}(z,z')^T\right\}dzdz'.$$

The density function of $(Y_j,Y_{j+\ell})$ is

$$f_{j\ell}(y,y';\boldsymbol{\theta})=f_{Y_j,Y_{j+\ell}}(y,y';\boldsymbol{\theta})=\mathrm{E}\,^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}_\ell}h_{j\ell}(y,y',\eta,\eta';\boldsymbol{\beta}),$$

where

$$h_{j\ell}(y,y',z,z';\boldsymbol{\beta})=\xi\left(y;\exp\{\boldsymbol{\beta}^T\mathbf{X}_j+z\}\right)\xi\left(y';\exp\{\boldsymbol{\beta}^T\mathbf{X}_{j+\ell}+z'\}\right).\tag{2.7}$$

For any 2-dimensional function $\zeta(y,z)$ and integer $j = 1, 2, \ldots,$ define

$$\mathrm{E}^{\mathbf{Y}}\zeta(Y_j, Y_{j+\ell}) = \int \zeta(y, y') f_{j\ell}(y, y'; \boldsymbol{\theta}^0)\, dy\, dy' \,.$$

2.3. Gradient of the marginals

Let $\mathbf{y}$ and $\boldsymbol{\eta}$ be d-dimensional vectors, with $\boldsymbol{\eta} \sim N(0, \boldsymbol{\Sigma})$; $\boldsymbol{\Sigma}$ is one of (2.3), (2.6), and d is $m+1$ or 2, the dimension of $\boldsymbol{\Sigma}$. With h being one of (2.5), (2.7), the marginal density functions for subvectors of $\mathbf{Y}$ in the preceding subsections have the form

$$\mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}} h(\mathbf{y}, \boldsymbol{\eta}; \boldsymbol{\beta}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \int h(\mathbf{y}, \mathbf{z}; \boldsymbol{\beta}) \exp\left\{-\tfrac{1}{2}\mathbf{z}^T \boldsymbol{\Sigma}^{-1}\mathbf{z}\right\} d\mathbf{z} \,.$$

We need the derivatives of $\mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}} h(\mathbf{y}, \boldsymbol{\eta}; \boldsymbol{\beta})$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for analysis of the composite likelihoods.

a. Derivative with respect to $\boldsymbol{\alpha}$: For a square invertible matrix $\boldsymbol{\Omega}$ and a scalar parameter θ, $\partial \log|\boldsymbol{\Omega}|/\partial\theta = \mathrm{tr}(\boldsymbol{\Omega}^{-1}(\partial\boldsymbol{\Omega}/\partial\theta))$ and $(\partial\boldsymbol{\Omega}^{-1}/\partial\theta) = -\boldsymbol{\Omega}^{-1}(\partial\boldsymbol{\Omega}/\partial\theta)\boldsymbol{\Omega}^{-1}$. Then for any $i = 0, \ldots, d-1$,

$$\frac{\partial}{\partial\gamma_i}\mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}} h(\mathbf{y}, \boldsymbol{\eta}; \boldsymbol{\beta}) = \frac{1}{2}\mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}\left\{h(\mathbf{y}, \boldsymbol{\eta}; \boldsymbol{\beta}) \cdot \mathrm{tr}\left[\frac{\partial\boldsymbol{\Sigma}}{\partial\gamma_i}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\right)\right]\right\}.$$

b. Derivative with respect to $\boldsymbol{\beta}$: Let $\mathbf{z} = (z_1, \ldots, z_d)^T$. For $i = 0, 1, \ldots, r$ with $X_{k0} = 1$ for all k,

$$\frac{\partial}{\partial\beta_i} h(\mathbf{y}, \mathbf{z}; \boldsymbol{\beta}) = \sum_{k=1}^{d} X_{ki}\frac{\partial}{\partial z_k} h(\mathbf{y}, \mathbf{z}; \boldsymbol{\beta}) \,.$$

Then, differentiating under the expectation and using integration by parts (and Novikov's theorem as stated in Appendix B),

$$\begin{aligned}
\frac{\partial \mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}} h(\mathbf{y}, \boldsymbol{\eta}; \boldsymbol{\beta})}{\partial\boldsymbol{\beta}_i} &= \frac{-1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\sum_{k=1}^{d} X_{ki}\int h(\mathbf{y}, \mathbf{z}; \boldsymbol{\beta})\frac{\partial}{\partial z_k}\exp\left\{-\tfrac{1}{2}\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\} d\mathbf{z} \\
&= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\sum_{k=1}^{d} X_{ki}\int \left(\mathbf{e}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{z}\right) h(\mathbf{y}, \mathbf{z}; \boldsymbol{\beta})\exp\left\{-\tfrac{1}{2}\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\} d\mathbf{z} \\
&= (X_{1i}, X_{2i}, \ldots, X_{di})\boldsymbol{\Sigma}^{-1}\mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}\left\{\boldsymbol{\eta} \cdot h(\mathbf{y}, \boldsymbol{\eta}; \boldsymbol{\beta})\right\} \,,
\end{aligned}$$

where $\mathbf{e}_k$ is a vector with 1 in the kth position and 0 elsewhere.

c. Derivative with respect to $(\sigma_V^2, \phi_1, \ldots, \phi_{p^*})$: Further suppose that $\boldsymbol{\Sigma}$ is the covariance matrix corresponding to an AR(p^*) process with $p^* \le d$. To find the derivatives with respect to $(\sigma_V^2, \phi_1, \ldots, \phi_{p^*})$, we make use of the Jacobian matrix of the transformation to the autocovariances

$$(\sigma_V^2, \phi_1, \ldots, \phi_{p^*}) \longmapsto (\gamma_0, \gamma_1, \ldots, \gamma_{p^*})\,.$$

The Yule-Walker equations can be written as

$$\begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{p^*} \end{pmatrix} = \mathbf{\Gamma} \begin{pmatrix} \sigma_V^2 \\ \phi_1 \\ \vdots \\ \phi_{p^*} \end{pmatrix}, \qquad \text{where} \quad \mathbf{\Gamma} = \begin{pmatrix} 1 & \gamma_1 & \gamma_2 & \cdots & \gamma_{p^*} \\ 0 & \gamma_0 & \gamma_1 & \cdots & \gamma_{p^*-1} \\ 0 & \gamma_1 & \gamma_0 & \cdots & \gamma_{p^*-2} \\ \vdots & & & & \vdots \\ 0 & \gamma_{p^*-1} & \gamma_{p^*-2} & \cdots & \gamma_0 \end{pmatrix}.$$

Let $\mathbf{J} = \partial(\gamma_0, \ldots, \gamma_{p^*})/\partial(\sigma_V^2, \phi_1, \ldots, \phi_{p^*})$ be the Jacobian matrix. Differentiating the Yule-Walker equations, we have $\mathbf{I} = \Phi^U + \Phi^L + \mathbf{\Gamma}\mathbf{J}^{-1}$, or

$$\frac{\partial(\gamma_0, \ldots, \gamma_{p^*})}{\partial(\sigma_V^2, \phi_1, \ldots, \phi_{p^*})} = (\mathbf{I} - \Phi^U - \Phi^L)^{-1}\mathbf{\Gamma}\,,$$

where

$$\Phi^U = \begin{pmatrix} 0 & \phi_1 & \cdots & \cdots & \phi_{p^*} \\ 0 & \phi_2 & \cdots & \phi_{p^*} & \\ \vdots & \vdots & \nearrow & & \\ 0 & \phi_{p^*} & & & \\ 0 & & & & \end{pmatrix}, \quad \Phi^L = \begin{pmatrix} 0 & & & & \\ \phi_1 & 0 & & & \\ \phi_2 & \phi_1 & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \phi_{p^*} & \phi_{p^*-1} & \cdots & \phi_1 & 0 \end{pmatrix}.$$

For $j = 1, \ldots, d$ and $i > p^*$, we have the recursive relationships:

$$\frac{d\gamma_i}{d\sigma_V^2} = \sum_{k=1}^{p^*} \phi_k \frac{d\gamma_{i-k}}{d\sigma_V^2}, \qquad \frac{d\gamma_i}{d\phi_j} = \gamma_{i-j} + \sum_{k=1}^{p} \phi_k \frac{d\gamma_{i-k}}{d\phi_j}\,.$$

For (2.1), this is applied with $p^* = p$ and $\mathbf{\Sigma}$ the Toeplitz matrix based on $\phi_1, \ldots, \phi_p$ in (1.3); for (2.2), this is applied with $p^* = 1$ and $\mathbf{\Sigma} = \mathbf{\Sigma}_\ell$, where ϕ_1 is the lag ℓ autocorrelation of (1.3).

3. Asymptotic Covariance Matrix of Composite Likelihood Estimators

In this section, the asymptotic covariance matrices of the composite likelihood estimators for (2.1) and (2.2) are expressed in terms of the moments of the derivatives of the log marginals. Formal results of the existence of such moments are provided.

Convention 3.1. *For any s-dimensional real-valued function $g(\theta_1, \ldots, \theta_s)$, let ∇g and $\nabla^2 g$ denote, respectively, the gradient and the Hessian matrix of g,*

$$\nabla g = \left(\frac{\partial g}{\partial \theta_i} \right)_{i=1,\ldots,s} \qquad \textit{and} \quad \nabla^2 g = \left(\frac{\partial^2 g}{\partial \theta_i \partial \theta_j} \right)_{i,j=1,\ldots,s} .$$

3.1. Covariance matrix

m-variate composite likelihood: For (2.1), with $\mathbf{Y}_{j:m} = (Y_j, \ldots, Y_{j+m})$, let

$$\boldsymbol{\Omega}_{1n} = n\mathrm{Var}^{\mathbf{Y}} \nabla Q_n(\boldsymbol{\theta}^0) = n\mathrm{Var}^{\mathbf{Y}} \Big\{ \frac{1}{n} \sum_{j=1}^{n-m} \nabla q_{j:m}(\mathbf{Y}_{j:m}; \boldsymbol{\theta}^0) \Big\} ,$$

$$\boldsymbol{\Omega}_{2n} = -\mathrm{E}^{\mathbf{Y}} \nabla^2 Q_n(\boldsymbol{\theta}^0) = \mathrm{E}^{\mathbf{Y}} \Big\{ \frac{1}{n} \sum_{j=1}^{n-m} \nabla q_{j:m}(\mathbf{Y}_{j:m}; \boldsymbol{\theta}^0) \, \nabla^T q_{j:m}(\mathbf{Y}_{j:m}; \boldsymbol{\theta}^0) \Big\} .$$

Standard arguments yield that the asymptotic covariance matrix of the composite likelihood estimator is

$$n\mathrm{Var}(\hat{\boldsymbol{\theta}}_n) \approx \boldsymbol{\Omega}_{1n}^{-1} \boldsymbol{\Omega}_{2n} \boldsymbol{\Omega}_{1n}^{-1} , \tag{3.1}$$

where $\hat{\boldsymbol{\theta}}_n = \arg\min_\Theta Q_n(\boldsymbol{\theta})$, provided that the expectations in $\boldsymbol{\Omega}_{1n}, \boldsymbol{\Omega}_{2n}$ exist.

BCL(m): For (2.2), $\boldsymbol{\Omega}_{1n}, \boldsymbol{\Omega}_{2n}, \hat{\boldsymbol{\theta}}_n$ are defined differently, but the asymptotic covariance matrix (3.1) has the same form. Let

$$\boldsymbol{\Omega}_{1n} = n\mathrm{Var}^{\mathbf{Y}} \nabla Q_n(\boldsymbol{\theta}^0) = n\mathrm{Var}^{\mathbf{Y}} \Big\{ \frac{1}{n} \sum_{j=1}^{n-m} \sum_{\ell=1}^{m} \nabla q_{j\ell}(Y_j, Y_{j+\ell}; \boldsymbol{\theta}^0) \Big\} ,$$

$$\boldsymbol{\Omega}_{2n} = -\mathrm{E}^{\mathbf{Y}} \nabla^2 Q_n(\boldsymbol{\theta}^0) = \mathrm{E}^{\mathbf{Y}} \Big\{ \frac{1}{n} \sum_{j=1}^{n-m} \sum_{\ell=1}^{m} \nabla q_{j\ell}(Y_j, Y_{j+\ell}; \boldsymbol{\theta}^0) \, \nabla^T q_{j\ell}(Y_j, Y_{j+\ell}; \boldsymbol{\theta}^0) \Big\} .$$

3.2. Existence of the moments

The main results of moment conditions are stated below; the details of the proofs are given in Appendices A and B. The assumptions are listed below.

A1: The expectation

$$\mathrm{E}^{\mathbf{Y}} \mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}} \left\{ \log \xi(Y_j, \exp(\boldsymbol{\beta}^T \mathbf{X}_j + \eta_j)) \right\}$$

exists and is a continuous function of $\boldsymbol{\theta}$, for $j = 1, \ldots, n$.

B1: When $\boldsymbol{\theta} = \boldsymbol{\theta}^0$, we have

$$\operatorname{rank} \begin{pmatrix} \frac{\partial \gamma_1}{\partial \phi_1} & \cdots & \frac{\partial \gamma_1}{\partial \phi_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial \gamma_m}{\partial \phi_1} & \cdots & \frac{\partial \gamma_m}{\partial \phi_p} \end{pmatrix} = p \,.$$

Note that Assumption B1 guarantees that the parameters are identifiable. For BCL(m) and MCL($m+1$), it rules out the cases of $m < p$. It is obvious that the AR parameters are not identifiable when $m < p$.

Let Θ be a compact parameter space satisfying

C1. the true parameter vector $\boldsymbol{\theta}^0$ is an interior point of Θ;

C2. $|\boldsymbol{\Sigma}(\boldsymbol{\theta})|$ is bounded below by a positive constant, where $|\cdot|$ is a matrix norm;

C3. for any $\boldsymbol{\theta} \in \Theta$, both $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}^0) \pm 6[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}^0) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})]$ are positive definite.

Remark 3.1. Assumption A1 is satisfied if $\mathrm{E}^{\mathbf{Y}}T(Y_1)$, $\mathrm{E}^{\mathbf{Y}}S(Y_1)$, $\mathrm{E}^{\boldsymbol{\eta}}a(Ce^{\eta_1})$, and $\mathrm{E}^{\boldsymbol{\eta}}b(Ce^{\eta_1})$ are finite and the last two are continuous functions of $C > 0$ and $\boldsymbol{\theta}$. These conditions can be checked for individual cases of (1.1).

Theorem 3.1. *Suppose that A1 and B1 are satisfied. Then, for MCL*($m+1$) *in* (2.1) *and BCL*(m) *in* (2.2)*, the moment matrices* $\boldsymbol{\Omega}_{1n}$ *and* $\boldsymbol{\Omega}_{2n}$ *exist, and* $\boldsymbol{\Omega}_{2n}$ *is invertible. Furthermore, if the covariates* $\mathbf{X}$ *are stationary, m-dependent, and bounded, then* $\boldsymbol{\Omega}_{1n}$ *and* $\boldsymbol{\Omega}_{2n}$ *converge as* $n \to \infty$.

3.3. The case without covariates

In this subsection, we state the results of consistency and asymptotic normality of the composite likelihood estimator when there are no covariates. In this case, $\mathbf{X}_i$ degenerates to 1.

m-variate composite likelihood: For (2.1), define the limiting matrices $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ as follows,

$$\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_1^{(m)} = \lim_{n\to\infty} n\mathrm{Var}^{\mathbf{Y}} \nabla Q_n(\boldsymbol{\theta}^0) = \lim_{n\to\infty} n\mathrm{Var}^{\mathbf{Y}} \Big\{ \frac{1}{n} \sum_{j=1}^{n-m} \nabla q_{j:m}(\mathbf{Y}_{j:m}; \boldsymbol{\theta}^0) \Big\} \,,$$

$$\boldsymbol{\Omega}_2 = \boldsymbol{\Omega}_2^{(m)} = -\mathrm{E}^{\mathbf{Y}} \nabla^2 q_{j:m}(\mathbf{Y}_{j:m}; \boldsymbol{\theta}^0) \,.$$

BCL(m): For (2.2), define the limiting matrices $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ as follows,

$$\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_1^{(m)} = \lim_{n\to\infty} n\mathrm{Var}^{\mathbf{Y}} \nabla Q_n(\boldsymbol{\theta}^0) = \lim_{n\to\infty} n\mathrm{Var}^{\mathbf{Y}} \Big\{ \frac{1}{n} \sum_{j=1}^{n-m} \sum_{\ell=1}^{m} \nabla q_{j\ell}(Y_j, Y_{j+\ell}; \boldsymbol{\theta}^0) \Big\} \,,$$

$$\boldsymbol{\Omega}_2 = \boldsymbol{\Omega}_2^{(m)} = -\Big\{ \sum_{\ell=1}^{m} \mathrm{E}^{\mathbf{Y}} \nabla^2 q_{j\ell}(Y_j, Y_{j+\ell}; \boldsymbol{\theta}^0) \Big\} \,.$$

Theorem 3.2. *Suppose that A1 and B1 are satisfied. Then, for* (2.1) *and* (2.2) *there exist matrices* $\boldsymbol{\Omega}_1$ *and* $\boldsymbol{\Omega}_2$, *where* $\boldsymbol{\Omega}_2$ *is invertible, such that the Hessian matrix* $-\nabla^2 Q_n(\boldsymbol{\theta}^0) \overset{a.s.}{\to} \boldsymbol{\Omega}_2$ *and* $\sqrt{n}\nabla Q_n(\boldsymbol{\theta}^0) \overset{d}{\to} N(\mathbf{0}, \boldsymbol{\Omega}_1)$. *Let* Θ *be a compact parameter space satisfying* C1−C3. *With* $\hat{\boldsymbol{\theta}}_n = \arg\min_\Theta Q_n(\boldsymbol{\theta})$, *then* $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0 \overset{a.s.}{\to} \mathbf{0}$ *and* $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0) \overset{d}{\to} N(\mathbf{0}, \boldsymbol{\Omega}_2^{-1}\boldsymbol{\Omega}_1\boldsymbol{\Omega}_2^{-1})$.

3.4. The case with covariates

The details are similar with covariates but involve more notation. The assumption of conditionally identically distributed no longer holds, so consistency and asymptotic normality results require that the covariates are well-behaved, such as being stationary and bounded. The conditions CO1 and AN2 in Appendix A must be assumed instead of being proved, because the ergodic theorem does not apply. For the case of covariates $\mathbf{X}_t$ being stationary and m-dependent, Lemma B.1 can be applied to justify the conditions CO1 and AN2.

4. Zeger's Method for Count Time Series

Zeger (1988) assumes (1.1)−(1.3) with Y_t being conditional Poisson. Additional notation is:

$$\sigma_\eta^2 = \mathrm{Var}(\eta_t),$$

$$\rho_{\eta k} = \mathrm{Corr}(\eta_t, \eta_{t+k}),$$

$$\sigma^2 = \exp(\sigma_\eta^2) - 1 = \frac{\mathrm{Var}\,(e^{\eta_t})}{[\mathrm{E}\,(e^{\eta_t})]^2}, \tag{4.1}$$

$$\rho_k = \frac{\exp(\rho_{\eta k}\sigma_\eta^2) - 1}{\exp(\sigma_\eta^2) - 1} = \mathrm{Cov}\,(e^{\eta_t}, e^{\eta_{t-k}}), \tag{4.2}$$

$$\boldsymbol{\beta}^* = (\beta_0 + \tfrac{1}{2}\sigma_\eta^2, \beta_1, \ldots, \beta_r),$$

$$\mu_t = \exp(\mathbf{X}_t^T \boldsymbol{\beta}^*) = \mathrm{E}\,(Y_t), \tag{4.3}$$

$$\mathrm{Var}(Y_t) = \mu_t + \sigma^2 \mu_t^2, \tag{4.4}$$

$$\mathrm{Cov}(Y_t, Y_{t-k}) = \sigma^2 \rho_k \mu_t \mu_{t-k}\,. \tag{4.5}$$

Zeger (1988) suggested a two-step iterative algorithm. The algorithm requires an initial guess for $\boldsymbol{\beta}^*$. In each iteration, the estimation of $\hat{\sigma}_V^2$ and $(\hat{\phi}_1, \ldots, \hat{\phi}_p)$ is updated via a moment matching scheme. Then, $\boldsymbol{\beta}^*$ is updated from a weighted least square equation for which the solution can be computed by a Kalman filter.

For the moment matching for ϕ and σ_V^2, given an initial guess of $\hat{\boldsymbol{\beta}}^*$, then based on (4.3)−(4.5), take

$$\hat{\mu}_t = \exp(\mathbf{X}_t^T \hat{\boldsymbol{\beta}}^*),$$

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{n}\left[(Y_t - \hat{\mu}_t)^2 - \hat{\mu}_t\right]}{\sum_{t=1}^{n}\hat{\mu}_t^2},$$
$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{n}(Y_t - \hat{\mu}_t)(Y_{t-k} - \hat{\mu}_{t-k})}{\hat{\sigma}^2\sum_{t=k+1}^{n}\hat{\mu}_t\hat{\mu}_{t-k}}.$$

Substitute $\hat{\sigma}^2$ and $\hat{\rho}_k$ into equations (4.1) and (4.2) to solve for $\hat{\sigma}_\eta^2$ and $\hat{\rho}_{\eta k}$. Then, $\hat{\sigma}_V^2$ and $\hat{\phi}_1, \ldots, \hat{\phi}_p$ are obtained by the Yule-Walker equations. For $p \geq 2$, it is possible that some estimated $\hat{\rho}_k$'s exceed 1, or $\{\hat{\rho}_k\}$ does not lead to a positive definite Toeplitz matrix.

5. Data Example

In this section, we compare composite likelihood estimation and Zeger's method for some count time series data consisting of daily car accident counts on different major roads in large cities in the Netherlands in 2001; see Brijs, Karlis and Wets (2008) and Sermaidis (2006) for details. One purpose is to study the effects of weather conditions. Initial data analysis shows moderate serial correlations and overdispersion relative to Poisson in some locations. Many covariates were measured each day, but we found only a few of them to be important when fitting regression models that ignore the serial dependence. However to determine the importance of different covariates, the serial dependence should be accounted for.

Full explanations and interpretation of the effects of the weather variables are beyond the scope of the present paper. There is some controversy on how these weather variables affect accident counts, and also there is dependence on the scale of measurement and on local conditions.

To compare estimation methods, we now restrict ourselves to one location that has serial dependence and overdispersion; the location is near Schiphol, the airport in the Netherlands. The covariates that we use for the time series modeling are: (a) WD=cosine of twice the mean wind direction in degrees, (b) RA=mean hourly radiation in Joule/cm^2 as a measurement on the intensity of the sun, (c) PD=mean hourly precipitation duration over units of 0.1 hour, (d) IWD=indicator of weekday (1 for Monday–Friday and 0 for Saturday and Sunday).

For composite likelihood estimation with (2.1) and (2.2), each marginal density was computed with adaptive Gauss-Hermite quadrature, with 3 quadrature points per dimension, and the numerical optimization was done with a quasi-Newton routine (Nash (1990)); see Appendix C for some details. We fitted latent AR(1), AR(2), and AR(3) models for conditional Poisson with the above four covariates and estimation with BCL(3). The estimates of the β's and σ_V

were essentially the same; the estimated AR parameters were 0.56 for AR(1), $(0.36, 0.22)$ for AR(2), and $(0.38, 0.30, -0.13)$ for AR(3). For AR(2), the corresponding $(\rho_{\eta 1}, \rho_{\eta 2})$ estimates were $(0.45, 0.38)$. The β's for the covariates RA and WD seem less important, especially the former, so we also fitted models with three and two covariates. We also fitted latent AR models for conditional negative binomial, but large standard errors for composite likelihood estimates result; this model implies a Poisson mixing distribution which is a combination of gamma and lognormal distributions, with potential near non-identifiability.

Using BCL(3) for the nine fits (three AR orders crossed with three subsets of covariates), we compared the composite likelihood information criterion in Definition 3 of Varin and Vidoni (2005); the penalty term $\mathrm{tr}(JH^{-1})$ is $-\mathrm{tr}(\mathbf{\Omega}_1\mathbf{\Omega}_2^{-1})$ in our notation. Based on this information criterion, the values with three-covariate models with AR(1)$-$AR(3) and the two-covariate model with AR(1) are very close, and the values for the other models are smaller.

We continue with the three-covariate model with AR(2) latent process for further summaries. Table 1 has the estimates based on MCL(3) and BCL(3) and Zeger's method, together with standard errors (SEs) for MCL/BCL. The estimated covariance matrix in (3.1) for composite likelihood was obtained via a parametric bootstrap method. We simulated paths from the parameter $\hat{\theta}_n$, and then the expectation terms in $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ were obtained from Monte-Carlo simulation, with derivatives of $q_{j\ell}$ evaluated using Gauss-Hermite quadrature. Also given are estimates and SEs for approximate maximum likelihood based on Markov chain Monte Carlo (MCMC) using WinBUGS (Lunn et al. (2000)).

For approximate maximum likelihood, the validity of using a Bayesian MCMC method is based on the following. If the prior is flat, then the posterior mode is the same as the maximum likelihood estimate (MLE). If in addition the posterior density is roughly multivariate normal (this holds for large samples via asymptotic theory), then the posterior mode and posterior mean vector are roughly the same, and the posterior covariance matrix matches that inverse Hessian of the negative log-likelihood (or estimated covariance matrix of the MLE). In MCMC, after the Markov chain has reached stationarity (and is thinned if necessary to reduce the serial correlation), the sample distribution of the chain, theoretically, has distribution matching the posterior, and the mean vector and covariance matrix of the chain lead to the MLE and estimated covariance matrix of the MLE.

For MCMC, we modified some WinBUGS code from Meyer and Yu (2000). For a nearly flat prior, we took $\beta_0, \beta_1, \ldots, \beta_r, \sigma_V, (\phi_1, \ldots, \phi_p)$ to be independent and (i) each β parameter with a normal distribution with mean 0 and SD 100; (ii) $1/\sigma_V^2$ having a gamma distribution with mean 1 and SD 100; (iii) the AR

Table 1. Parameter estimates for Poisson regression with latent Gaussian AR(2) model: BCL(3), MCL(3), Zeger's moment method, and maximum likelihood via MCMC in WinBUGS; for the first three methods, SEs are based on parametric bootstrap.

parameter	BCL(3)	SE	MCL(3)	SE	Zeger	SE	MLE	SE
β_0: intercept	1.594	0.061	1.590	0.062	1.790	0.250	1.588	0.065
β_1: WD	-0.056	0.037	-0.058	0.036	-0.040	0.054	-0.053	0.036
β_2: PD	0.175	0.019	0.174	0.018	0.130	0.032	0.174	0.018
β_3: IWD	0.472	0.053	0.478	0.053	0.400	0.130	0.475	0.055
σ_V	0.273	0.016	0.270	0.016	0.231	0.036	0.270	0.031
ϕ_1	0.350	0.130	0.340	0.130	0.360	0.170	0.390	0.150
ϕ_2	0.250	0.140	0.280	0.140	0.320	0.200	0.270	0.140

parameters $(\phi_1, \ldots, \phi_p)$ with a uniform distribution over their parameter space by choosing the partial correlations with appropriate beta distributions (see Jones (1987)). After some checks for insensitivity to parameters in the nearly flat prior and MCMC convergence for the data set, we chose a chain length of 10^5 with a burn-in of 2×10^4 and a thin rate of every 50.

For this and other similar data sets, Brijs, Karlis and Wets (2008) used some models based on binomial thinning, where the innovation term was Poisson with mean depending on the covariates. To get more overdispersion relative to Poisson, other distributions can be used for the innovation, or other thinning operators could be used. For these daily accident data, the latent Gaussian process model is a plausible mechanism for the serial dependence. In general, the latent Gaussian process model can allow for a wider range of autocorrelation structure (relative to lag 1 serial correlation) than models based on thinning operators.

6. Simulation Study

A simulation study was run with the Poisson model for (1.1). We mention the design of the study and then show some representative results to compare estimation via (a) composite likelihood methods such as BCL(2) and BCL(3), (b) Zeger's method, and (c) approximate maximum likelihood via MCMC. For MCMC, we used the control parameters (thin rate, burn-in etc.) mentioned in the preceding section.

Based on experience with other models where composite likelihood estimation has been used, we expect more efficiency loss relative to maximum likelihood when the latent autocorrelation is stronger or when σ_V is smaller. We do not expect the number of covariates or the β parameters to have much effect on relative efficiency. The range of dependence in the observed Y_t, as σ_V^2 changes, depends on (1.1). For a Poisson model, with other parameters held fixed, serial independence is reached in the limit as σ_V or σ_η goes to 0 or ∞; this can be checked based on (4.1)−(4.5).

6.1. Choice of covariates

For the simulation study, we used one continuous covariate and one discrete covariate, with β values near the data example. For the continuous covariate, as a first choice we used the wind direction covariate WD mentioned in Section 5 and as the second choice we used the precipitation duration covariate PD. (WD is in interval -1 to 1, whereas PD is right-skewed). For the discrete covariate, we used the indicator of weekday IWD. We label the covariates as x_1, x_2 with regression parameters β_1, β_2 for the simulation study. We set $(\beta_0, \beta_1, \beta_2) = (1.5, -0.1, 0.4)$ for $x_1 =$WD or $(\beta_0, \beta_1, \beta_2) = (1.5, 0.2, 0.4)$ for $x_1 =$PD; these are values close to those in Table 1.

In order to have arbitrary n, we replicated the WD (or PD) column of the data set for $n > 365$, so that $WD_i = WD_{i-365}$. It was better to increase n in this way because of some serial correlation in the covariate time series; that is, this was a better extension than independent randomly generated covariates. For IWD, the sequence was continued with five 1's and two 0's periodically, for $n > 365$.

6.2. Choice of AR coefficients

We used three sets of (ϕ_1, ϕ_2) for AR(2). The first choice is close to that in Table 1, and the second and third correspond to stronger autocorrelations.

1. $\phi_1 = 0.34$, $\phi_2 = 0.26$, or latent serial correlations $\rho_{\eta 1} = 0.46$, $\rho_{\eta 2} = 0.42$.
2. $\phi_1 = 0.56$, $\phi_2 = 0.06$, or latent serial correlations $\rho_{\eta 1} = 0.60$, $\rho_{\eta 2} = 0.40$.
3. $\phi_1 = 0.55$, $\phi_2 = 0.22$, or latent serial correlations $\rho_{\eta 1} = 0.70$, $\rho_{\eta 2} = 0.60$.

6.3. Choice of σ_V

We chose two levels of $\sigma_V :$ (i) 0.3 near that in Table 1, and (ii) 0.2, a smaller value. A smaller σ_V leads to larger serial correlations for $\exp(\eta_t)$ in (4.2), but smaller correlations for Y_t in (4.5). For σ_V around 0.15 or smaller, the correlations of the Y_t might be small enough that one would not consider a model with time dependence.

6.4. Comparisons

The main design for the simulation study to evaluate composite likelihood estimators is: $3 \times 2 \times 2$: three sets of dependence parameters, two σ_V values and two sample sizes ($n = 365$ and $n = 1{,}095$). This design was used with $x_1 =$WD (or $x_1 =$RD) and $x_2 =$IWD. We could quickly run 500 replications per combination for BCL(2), BCL(3), BCL(4) and MCL(3), and Zeger's method. Because approximate maximum likelihood with MCMC/WinBUGS takes much longer, we ran fewer replications on a subset of the $3 \times 2 \times 2$ design. With an Intel 2.40Ghz processor, a sample size of $n = 1{,}095$ and three covariates, the computing times in a C program for BCL(2), BCL(3), MCL(3), and BCL(4), averaged about 5,5,7, and 10 seconds respectively; the time was much less for Zeger's method and over 70 minutes for approximate maximum likelihood with MCMC/WinBUGS.

The results for MCL(3) are almost the same as BCL(2); conclusions are similar for the two choices of the continuous covariate x_1, so the summary tables include only $x_1 =$WD. BCL(3) is better than BCL(2) in cases of stronger dependence and/or smaller σ_V. BCL(2) leads to efficient estimators of the β parameters but BCL(3) leads to slightly more efficient estimators for σ_V and the AR parameters ϕ_j. The additional improvement from BCL(4) for σ_V and ϕ_j is even smaller. This pattern of needing more lags in BCL with more dependence is similar to what was observed in Qu (2008) for the ARSV model.

In Table 2, root mean squared error (MSE) summaries of estimators for BCL(2), BCL(3), ML/MCMC, and Zeger's method with sample size $n = 1{,}095$ for (a) AR(2): $\phi_1 = 0.34$, $\phi_2 = 0.26$, and $\sigma_V = 0.3$ (close to that in the data set). (b) AR(2): $\phi_1 = 0.55$, $\phi_2 = 0.22$, and $\sigma_V = 0.2$ (stronger latent autocorrelations and smaller σ_V than in the data set). The bias is of the order of 10^{-3} for the β parameters and 10^{-2} (and sometimes 10^{-1}) for σ_V and ϕ_1, ϕ_2, with more bias for the smaller sample size in our design.

Table 2 shows the range of results in the simulation study. For some AR(2) parameter vectors, BCL(2) or MCL(3) are efficient with root MSE very close to ML/MCMC. For other parameter vectors with stronger latent autocorrelations and smaller σ_V, BCL(3) or BCL(4) lead to more efficient estimators than BCL(2). Zeger's moment-based method does not always have a solution; it is worse in efficiency even if we only consider the subset of simulated data sets with estimates. The patterns are confirmed for the AR(1) latent process with parameters close to case (b) above: $\phi_1 = \rho_1 = 0.7$, and $\sigma_V = 0.2$; see Table 3.

7. Summary and Discussion

The simulation study in Section 6 shows the composite likelihood estimation performs very well for (1.1)$-$(1.3) with a conditional Poisson model. BCL with a few lags performed at least as well as trivariate composite likelihood, so we didn't try composite likelihood based on d consecutive observations with $d \geq 4$.

Table 2. Root MSE of parameter estimates for Poisson regression with latent Gaussian AR(2) model, covariates WD, IWD; estimation methods BCL(2), BCL(3), Zeger's moment method, and approximate maximum likelihood via MCMC in WinBUGS; parameters $(\beta_0, \beta_1, \beta_2) = (1.5, -0.1, 0.4)$, and (a) $\phi_1 = 0.34$, $\phi_2 = 0.26$, $\sigma_V = 0.3$; (b) $\phi_1 = 0.55$, $\phi_2 = 0.22$, $\sigma_V = 0.2$. Sample size $n =$1,095; 400 replications. For Zeger's method, 394 out of 400 with solutions in (b). For case (a), estimates of parameters were close to each other for different methods; correlations mostly above 0.9 for BCL(2), BCL(3) and MCMC with each other, and mostly above 0.8 for each with Zeger's method. For case (b), estimates for Zeger's method could be quite different; for other methods, correlations were above 0.9 for β's, above 0.8 for σ_V, and above 0.6 for for ϕ_1, ϕ_2.

	Parameter set (a)				Parameter set (b)			
par.	Zeger	BCL(2)	BCL(3)	ML	Zeger	BCL(2)	BCL(3)	ML
β_0	0.041	0.037	0.037	0.037	0.135	0.036	0.036	0.036
β_1	0.026	0.024	0.024	0.023	0.026	0.022	0.022	0.020
β_2	0.039	0.035	0.035	0.035	0.060	0.031	0.031	0.031
σ_V	0.022	0.022	0.022	0.022	0.044	0.040	0.037	0.032
ϕ_1	0.083	0.082	0.086	0.089	0.260	0.290	0.270	0.210
ϕ_2	0.094	0.094	0.095	0.089	0.260	0.280	0.250	0.180

Table 3. Root MSE of parameter estimates for Poisson regression with latent Gaussian AR(1) model, covariates WD, IWD; estimation methods BCL(2−4), Zeger's moment method, and approximate maximum likelihood via MCMC in WinBUGS; parameters $(\beta_0, \beta_1, \beta_2) = (1.5, -0.1, 0.4)$, $\phi_1 = 0.7$, $\sigma_V = 0.2$. Sample size $n =$1,095; 400 replications. For Zeger's method, 399 out of 400 with solutions. Estimates of parameters were close to each other for MCMC and BCL(2−4), but were more different for Zeger's method. Correlations of parameter estimates with Zeger's method were mostly less than 0.8, but for other methods they were mostly above 0.9.

par.	Zeger	BCL(2)	BCL(3)	BCL(4)	ML/MCMC
β_0	0.135	0.034	0.034	0.034	0.034
β_1	0.026	0.022	0.022	0.022	0.021
β_2	0.061	0.031	0.031	0.031	0.032
σ_V	0.047	0.027	0.024	0.023	0.021
ϕ_1	0.110	0.069	0.059	0.056	0.055

More lags in BCL(m) are needed with stronger latent dependence in order to get comparable efficiency with maximum likelihood. For (1.1)−(1.3) for other conditional distributions, we expect the pattern to be similar, because Joe and Lee (2009) had this pattern for several models where exact efficiency calculations were possible for composite likelihood versus full likelihood. For ARSV models for financial asset return time series, the latent correlation parameter is usually larger than 0.8, and then Qu (2008) found that there was significant efficiency loss even for BCL(m) with m around 4 or 5.

Based on our experience, for good efficiency, a rough rule is to use $m = 2$ or 3 for weak serial dependence; $m = 3$ or 4 for moderate serial dependence, and $m \geq 4$ for stronger dependence. For a particular data set, one could increase m for BCL(m) estimation until the SE estimates have stabilized; further checks can be made for different m with evaluations of asymptotic covariance matrices near the BCL estimate $\hat{\boldsymbol{\theta}}$.

Unless one has reason to believe that the dependence is so strong that composite likelihood methods are inefficient, we recommend composite likelihood methods as they are easier to implement in computer code, and they have faster computational time than other simulation-based methods mentioned below. A fast computational method is useful for deciding on the important covariates and order of the latent autoregressive process. Although Zeger's moment-based estimation method is even computationally faster than composite likelihood methods, we do not recommend it for count data as it can be substantially less efficient, and it can have problems with impossible Toeplitz matrices.

If the efficiency of composite likelihood estimation were worse for approximate maximum likelihood (cf., McCulloch (1997)) with multidimensional integrals, there are variations of simulated likelihood approaches with importance sampling, and these take more effort to implement than composite likelihood. For the ARSV model for financial time series (with no covariates), the Monte Carlo importance sampling method in Sandmann and Koopman (1998) is implemented in Ox `http://www.doornik.com` and has reasonable speed.

For the count time series model that we are using, while approximate ML via MCMC with WinBUGS can be used, it is known that there are large autocorrelations in the Markov chain for models of the form (1.1)−(1.3), and this explains the length of time needed for numerically stable results. Jung, Kukuk and Liesenfeld (2006) have proposed an efficient importance sampling (EIS) method; see also Richard and Zhang (2007). Earlier Chan and Ledolter (1995) proposed a Monte Carlo-EM approach. We did try an implementation of the Monte Carlo-EM and Monte Carlo-EIS approaches, but this was much slower than composite likelihood, and there were more decisions on control parameters affecting the convergence and the number of iterations needed to approximate the likelihood.

There are models for time series based on a latent Gaussian process that do not satisfy (1.1)−(1.3); an example is a binary probit time series model with $Y_t = I(Z_t <= 0)$, where Z_t involves regression on covariates and an error process that is Gaussian. However composite likelihood should be a good estimation method and we expect that some of the techniques of the proofs will apply.

Acknowledgement

This research has been supported by NSERC Canada Discovery Grants, and Start-up grants from Hong Kong Polytechnic University and the University of Saskatchewan. Thanks to Tom Brijs for access to the data set. We are grateful to the referees for detailed comments leading to an improved presentation.

Appendix A. Proofs

Convention A.1. *Let* $\mathbf{i} = \{i_1, i_2, \ldots, i_d\}$ *be an unordered tuples in which the elements are selected from the set* $\{1, \ldots, s\}$ *and* $g(\boldsymbol{\theta})$ *be an s-dimensional real-valued function. We use the notation* $\partial^{\mathbf{i}} g(\boldsymbol{\theta}) = (\partial^d/\partial\theta_{i_1} \cdots \partial\theta_{i_d}) g(\boldsymbol{\theta})$ *.*

Convention A.2. *For any s-dimensional vector* $\mathbf{u}$ *and* $s \times s$ *matrix* $\mathbf{M}$, *we use the notation* u_i *for the ith element of* $\mathbf{u}$ *and* $\mathbf{M}_i$ *for the ith row of* M *.*

Proof of Theorem 3.1. This parallels that of Theorem 3.2 and is omitted.

Proof of Theorem 3.2. The principle to establish consistency and asymptotic normality of quasi-maximum likelihood estimation is standard (see for example, p. 101 of Straumann (2005)). Here, we only give the proof for the m-variate composite likelihood (2.1) case; the bivariate composite likelihood (2.2) can be handled in a similar manner. To show that $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}^0$ almost surely, one way is to establish the following condition.

CO1. With probability 1, the likelihood function $Q_n(\boldsymbol{\theta})$ converges uniformly in Θ to some function $Q(\boldsymbol{\theta})$, i.e., $\sup_\Theta |Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| \overset{a.s.}{\to} 0$.

To establish the asymptotic normality of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0)$, the following conditions are required.

AN1: There exist matrices $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$, where $\boldsymbol{\Omega}_2$ is positive definite, such that $\sqrt{n}\nabla Q_n(\boldsymbol{\theta}^0) \overset{d}{\to} N(\mathbf{0}, \boldsymbol{\Omega}_1)$ and $-\nabla^2 Q_n(\boldsymbol{\theta}^0) \overset{a.s.}{\to} \boldsymbol{\Omega}_2$.

AN2: $\sup_\Theta \sqrt{n}|\nabla Q_n(\boldsymbol{\theta}) - \nabla Q(\boldsymbol{\theta})| \overset{a.s.}{\to} 0$ and $\sup_\Theta |\nabla^2 Q_n(\boldsymbol{\theta}) - \nabla^2 Q(\boldsymbol{\theta})| \overset{a.s.}{\to} 0$.

Lemma A.1 below guarantees that the expectation of $Q(\boldsymbol{\theta})$ exists, and that $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ are well-defined. The convergence of the limit in $\boldsymbol{\Omega}_1$ is established in Lemma A.2. The positive definiteness of $\boldsymbol{\Omega}_2$ is proved in Lemma A.3.

Conditions CO1 and AN2 are established in Lemma A.4 via the Mean Ergodic Theorem. The convergence of $-\nabla^2 Q_n(\boldsymbol{\theta}^0) \overset{a.s.}{\to} \boldsymbol{\Omega}_{2m}$ that appears in AN1 is a consequence of the Mean Ergodic Theorem. Since $\{Y_t, \ldots, Y_{t+m}, \eta_t, \ldots, \eta_{t+m}\}$, $t \geq 1$, is a Markovian process with homogeneous transition probabilities, the Central Limit theorem for a Markov chain (see Theorem 7.5 in Chapter V of Doob (1953)) can be used to establish that $\sqrt{n}\,\nabla Q_n(\boldsymbol{\theta}^0) \overset{d}{\to} N(\mathbf{0}, \boldsymbol{\Omega}_1)$. With conditions

CO1, AN1, and AN2, then $\hat{\boldsymbol{\theta}}_n \overset{a.s.}{\rightarrow} \boldsymbol{\theta}^0$ and $\sqrt{n}\,(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0) \overset{d}{\rightarrow} N(\mathbf{0}, \boldsymbol{\Omega}_2^{-1}\boldsymbol{\Omega}_1\boldsymbol{\Omega}_2^{-1})$ follow from standard arguments. □

Lemmas A.1−A.4 are given below. Some technical lemmas used in the proof of Lemmas A.1−A.4 are given in Appendix B.

Lemma A.1. *Let* Θ *be a compact parameter space satisfying* C1−C3.

(I) *The expectation* $Q(\boldsymbol{\theta}) = \mathrm{E}^{\mathbf{Y}} \log f_{j:m}(Y_j, \ldots, Y_{j+m}; \boldsymbol{\theta})$ *exists for all* $\boldsymbol{\theta} \in \Theta$, *and* $\sup_\Theta |Q(\boldsymbol{\theta})| < K$ *for some* $K > 0$.

(II) $Q(\boldsymbol{\theta})$ *is differentiable with respect to* $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ *up to order 3. For every unordered tuple* $\mathbf{i}$ *with order* ≤ 3 *and the elements selected from the set* $\{\boldsymbol{\beta}, \boldsymbol{\alpha}\}$, *there are constants* $K_{\mathbf{i}} > 0$ *such that* $\sup_\Theta |\partial^{\mathbf{i}} Q(\boldsymbol{\theta})| < K_{\mathbf{i}}$.

Proof. In the following, the notation $\mathbf{Y}_{j:m} = (Y_j, \ldots, Y_{j+m})$ and $\mathbf{y}$ refer to $(m+1)$-dimensional vectors.

(I) To show that $Q(\boldsymbol{\theta})$ exists and is bounded, we only need to establish an upper bound and a lower bound for $f_{j:m}(\mathbf{Y}_{j:m}; \boldsymbol{\theta})$. Jensen's inequality is used. The bounds are as follows:

$$\begin{aligned} Q(\boldsymbol{\theta}) &= \mathrm{E}^{\mathbf{Y}} \log f_{j:m}(\mathbf{Y}_{j:m}; \boldsymbol{\theta}) \leq \log \mathrm{E}^{\mathbf{Y}} f_{j:m}(\mathbf{Y}_{j:m}; \boldsymbol{\theta}) \\ &= \log \int f_{j:m}(\mathbf{y}; \boldsymbol{\theta}) f_{j:m}(\mathbf{y}; \boldsymbol{\theta}^0) d\mathbf{y} \leq \log \max_{\mathbf{y}} f_{j:m}(\mathbf{y}; \boldsymbol{\theta}^0)\,; \\ Q(\boldsymbol{\theta}) &= \mathrm{E}^{\mathbf{Y}} \log \mathrm{E}_{\boldsymbol{\alpha}}^{\boldsymbol{\eta}} h(\mathbf{Y}_{j:m}, \boldsymbol{\eta}; \boldsymbol{\beta}) \geq \mathrm{E}^{\mathbf{Y}} \mathrm{E}_{\boldsymbol{\alpha}}^{\boldsymbol{\eta}} \log h(\mathbf{Y}_{j:m}, \boldsymbol{\eta}; \boldsymbol{\beta}) \\ &= (m+1) \mathrm{E}^{\mathbf{Y}} \mathrm{E}_{\boldsymbol{\alpha}}^{\boldsymbol{\eta}} \log \xi(Y_j, \exp(\beta_0 + \eta_j))\,. \end{aligned}$$

From A1 and the assumption that Θ is compact, the conclusion follows.

(II) We first establish the results for $\partial^{\mathbf{i}} Q(\boldsymbol{\theta})$ in the case that the order of $\mathbf{i}$ is one. The first order derivatives of $Q(\boldsymbol{\theta})$ are

$$\partial^{\mathbf{i}} Q(\boldsymbol{\theta}) = \mathrm{E}^{\mathbf{Y}} \frac{\partial^{\mathbf{i}} f_{j:m}(\mathbf{Y}_{j:m}, \boldsymbol{\theta})}{f_{j:m}(\mathbf{Y}_{j:m}, \boldsymbol{\theta})}\,. \tag{A.1}$$

From the results of Section 2.3, the first order derivatives $\partial^{\mathbf{i}} Q(\boldsymbol{\theta})$ have the form

$$\partial^{\mathbf{i}} Q(\boldsymbol{\theta}) = \mathrm{E}^{\mathbf{Y}} \left\{ \frac{\mathrm{E}_{\boldsymbol{\alpha}}^{\boldsymbol{\eta}}\, g(\boldsymbol{\eta}; \boldsymbol{\theta})\, h(\mathbf{Y}_{j:m}, \boldsymbol{\eta}; \boldsymbol{\beta})}{\mathrm{E}_{\boldsymbol{\alpha}}^{\boldsymbol{\eta}}\, h(\mathbf{Y}_{j:m}, \boldsymbol{\eta}; \boldsymbol{\beta})} \right\}, \tag{A.2}$$

where $g(\boldsymbol{\eta}; \boldsymbol{\theta})$ is a polynomial with order ≤ 2. The required result is a consequence of Lemma B.1. For second and third order derivatives of $Q(\boldsymbol{\theta})$, terms like

$$\mathrm{E}^{\mathbf{Y}} \left\{ \frac{\mathrm{E}_{\boldsymbol{\alpha}}^{\boldsymbol{\eta}}\, g_1(\boldsymbol{\eta}; \boldsymbol{\theta})\, h(\mathbf{Y}_{j:m}, \boldsymbol{\eta}; \boldsymbol{\beta})}{\mathrm{E}_{\boldsymbol{\alpha}}^{\boldsymbol{\eta}}\, h(\mathbf{Y}_{j:m}, \boldsymbol{\eta}; \boldsymbol{\beta})} \right\} \left\{ \frac{\mathrm{E}_{\boldsymbol{\alpha}}^{\boldsymbol{\eta}}\, g_2(\boldsymbol{\eta}; \boldsymbol{\theta})\, h(\mathbf{Y}_{j:m}, \boldsymbol{\eta}; \boldsymbol{\beta})}{\mathrm{E}_{\boldsymbol{\alpha}}^{\boldsymbol{\eta}}\, h(\mathbf{Y}_{j:m}, \boldsymbol{\eta}; \boldsymbol{\beta})} \right\}$$

can be bounded with the Cauchy-Schwarz inequality and, from Section 2.3, differentiating the numerator of (A.2) leads to:

$$\mathrm{E}^{\mathbf{Y}}\left\{\frac{(\partial/\partial\theta_j)\mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}\, g(\boldsymbol{\eta};\boldsymbol{\theta})\, h(\mathbf{Y}_{j:m},\boldsymbol{\eta};\boldsymbol{\beta})}{\mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}\, h(\mathbf{Y}_{j:m},\boldsymbol{\eta};\boldsymbol{\beta})}\right\} = \mathrm{E}^{\mathbf{Y}}\left\{\frac{\mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}\, g^{*}(\boldsymbol{\eta};\boldsymbol{\theta})\, h(\mathbf{Y}_{j:m},\boldsymbol{\eta};\boldsymbol{\beta})}{\mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}\, h(\mathbf{Y}_{j:m},\boldsymbol{\eta};\boldsymbol{\beta})}\right\}$$

for another function g^* which is a polynomial in $\boldsymbol{\eta}$.

Lemma A.2. *The limit*

$$\boldsymbol{\Omega}_1 = \lim_{n\to\infty} n\mathrm{Var}^{\mathbf{Y}}\Big\{\frac{1}{n}\sum_{t=1}^{n-m}\nabla q(Y_t, Y_{t+1},\ldots,Y_{t+m};\boldsymbol{\theta}^0)\Big\}$$

exists and is finite. Here $q = q_{t:m}$ for all t.

Proof. Note that $\mathrm{E}^{\mathbf{Y}}\nabla q(Y_t, Y_{t+1},\ldots,Y_{t+m};\boldsymbol{\theta}^0) = \mathbf{0}$ and $\nabla q(Y_t, Y_{t+1},\ldots,Y_{t+m};\boldsymbol{\theta}^0)$ is stationary. Using the Dominated Convergence Theorem, it can be shown that the limit $\boldsymbol{\Omega}_1$ has the representation $\boldsymbol{\Omega}_1 = v_0 + 2\sum_{i=1}^{\infty} v_i$, where

$$v_i = \mathrm{Cov}^{\mathbf{Y}}\left\{\nabla q(Y_1,\ldots,Y_{m+1};\boldsymbol{\theta}^0), \nabla q(Y_{1+i},\ldots,Y_{m+1+i};\boldsymbol{\theta}^0)\right\}.$$

Using Lemma B.3, the series for $\boldsymbol{\Omega}_1$ converges.

Lemma A.3. *The matrix $\boldsymbol{\Omega}_{2m}$ is positive definite.*

Proof. Let $\boldsymbol{\nu} = (\nu_{\boldsymbol{\beta}}, \nu_{\boldsymbol{\alpha}})$ be a vector satisfying $\boldsymbol{\nu}^T\{\mathrm{E}\,\nabla^2 q(Y_1,\ldots,Y_{m+1};\boldsymbol{\theta}^0)\}\boldsymbol{\nu} = 0$. Here $q = q_{t:m}$ for all t, and the derivatives of q are with respect to $(\boldsymbol{\beta},\boldsymbol{\alpha})$. With Assumption B1, it suffices to show that $\boldsymbol{\nu} = (\nu_{\beta_0}, \boldsymbol{\nu}_{\boldsymbol{\alpha}}) = \mathbf{0}$. By noting that $\mathrm{E}\,\nabla q(Y_1,\ldots,Y_{m+1};\boldsymbol{\theta}^0) = \mathbf{0}$ and

$$-\mathrm{E}\,\nabla^2 q(Y_1,\ldots,Y_{m+1};\boldsymbol{\theta}^0) = \mathrm{E}\,[\nabla q(Y_1,\ldots,Y_{m+1};\boldsymbol{\theta}^0)][\nabla q(Y_1,\ldots,Y_{m+1};\boldsymbol{\theta}^0)]^T,$$

we have $\mathrm{E}\,\boldsymbol{\nu}^T\left\{\nabla q(Y_1,\ldots,Y_{m+1};\boldsymbol{\theta}^0)\right\}\left\{\nabla q(Y_1,\ldots,Y_{m+1};\boldsymbol{\theta}^0)\right\}^T\boldsymbol{\nu} = \mathbf{0}$, which implies

$$\left\{\nabla q(y_1,\ldots,y_{m+1};\boldsymbol{\theta}^0)\right\}^T\boldsymbol{\nu} = \mathbf{0} \quad \forall \mathbf{y}. \tag{A.3}$$

With $\boldsymbol{\Sigma}$ as in (2.3), let

$$\mathbf{V} = \nu_{\gamma_0}\frac{\partial\boldsymbol{\Sigma}}{\partial\gamma_0} + \cdots + \nu_{\gamma_m}\frac{\partial\boldsymbol{\Sigma}}{\partial\gamma_m} = \begin{pmatrix} \nu_{\gamma_0} & \nu_{\gamma_1} & \cdots & \nu_{\gamma_m} \\ \nu_{\gamma_1} & \nu_{\gamma_0} & \cdots & \nu_{\gamma_{m-1}} \\ \vdots & & \ddots & \vdots \\ \nu_{\gamma_m} & \nu_{\gamma_{m-1}} & \cdots & \nu_{\gamma_0} \end{pmatrix}.$$

Let $\omega_1 = -(1/2)\mathrm{tr}\mathbf{V}\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\omega}_2 = \nu_{\beta_0}\boldsymbol{\Sigma}^{-1}\mathbf{1}$, and $\boldsymbol{\omega}_3 = (1/2)\boldsymbol{\Sigma}^{-1}\mathbf{V}\boldsymbol{\Sigma}^{-1}$. From the derivatives in Section 2.3 and (A.3), it can be checked that $\omega_1, \boldsymbol{\omega}_2, \boldsymbol{\omega}_3$ as defined satisfy (B.2) in Appendix B. By Lemma B.4, we have $\omega_1 = 0$, $\boldsymbol{\omega}_2 = \mathbf{0}$, and $\boldsymbol{\omega}_3 = \mathbf{0}$. Therefore $\nu_{\beta_0} = 0$, $\mathbf{V} = \mathbf{0}$, and $\boldsymbol{\nu}_{\boldsymbol{\alpha}} = \mathbf{0}$.

Lemma A.4. *We have* $\sup_{\boldsymbol{\theta}\in\Theta}|Q_n(\boldsymbol{\theta})-Q(\boldsymbol{\theta})|\overset{a.s.}{\to}0$, $\sup_{\boldsymbol{\theta}\in\Theta}|\nabla Q_n(\boldsymbol{\theta})-\nabla Q(\boldsymbol{\theta})|\overset{a.s.}{\to}0$, $\sup_{\boldsymbol{\theta}\in\Theta}|\nabla^2 Q_n(\boldsymbol{\theta})-\nabla^2 Q(\boldsymbol{\theta})|\overset{a.s.}{\to}0$.

Proof. By Lemma A.1, $Q(\boldsymbol{\theta})$ exists. Using the Ergodic Theorem, we have for each $\boldsymbol{\theta}\in\Theta$, $Q_n(\boldsymbol{\theta})\to Q(\boldsymbol{\theta})$. What remains is to show that the convergence is uniform. For $\boldsymbol{\theta}',\boldsymbol{\theta}''\in\Theta$, by the Mean Value Theorem,

$$\frac{|Q_n(\boldsymbol{\theta}')-Q_n(\boldsymbol{\theta}'')|}{|\boldsymbol{\theta}'-\boldsymbol{\theta}''|}\le\frac{1}{n}\sum_{t=1}^{n-m}\left\{\sup_{\boldsymbol{\theta}\in\Theta}\left|\nabla q(Y_t,\ldots,Y_{t+m};\boldsymbol{\theta})\right|\right\},$$

A bound for the right-hand side can be obtained by part (II) of Lemma A.1. Consequently, we have the equicontinuity

$$\sup_{\boldsymbol{\theta}',\boldsymbol{\theta}''\in\Theta}\frac{|Q_n(\boldsymbol{\theta}')-Q_n(\boldsymbol{\theta}'')|}{|\boldsymbol{\theta}'-\boldsymbol{\theta}''|}\le O(1),\ \text{a.s.}.$$

Here, the quantity $O(1)$ does not depend on $\boldsymbol{\theta}',\boldsymbol{\theta}''$. This implies uniform convergence, $\sup_{\boldsymbol{\theta}\in\Theta}|Q_n(\boldsymbol{\theta})-Q(\boldsymbol{\theta})|\overset{a.s.}{\to}0$. Similarly, we obtain the results for the first and second order derivatives of $Q_n(\boldsymbol{\theta})$.

Appendix B. Technical Lemmae

The hard parts of the proof of the asymptotic results are in the lemmas in this appendix. Bounding the covariance of derivatives of the m-dimensional composite likelihood is a key component. Here, for the case of no covariates, we let $q=q_{j:m}$ for all j and, in (2.5), we let $h=h_j$ for all j.

Novikov's theorem. Let $\mathbf{Z}\sim N_d(\mathbf{0},\boldsymbol{\Omega})$ and let ψ be a differentiable function in $\Re^d$. Then

$$\int_{\Re^d}\mathbf{z}\psi(\mathbf{z})e^{-(1/2)\mathbf{z}^T\boldsymbol{\Omega}^{-1}\mathbf{z}}d\mathbf{z}=\int_{\Re^d}\boldsymbol{\Omega}\nabla\psi(\mathbf{z})e^{-(1/2)\mathbf{z}^T\boldsymbol{\Omega}^{-1}\mathbf{z}}d\mathbf{z},$$

assuming the integrals exist, or $\mathrm{E}[\mathbf{Z}\psi(\mathbf{Z})]=\boldsymbol{\Omega}\,\mathrm{E}[\nabla\psi(\mathbf{Z})]$.

Note that this theorem appeared in the Russian physics literature in 1964; a statement is given in Chaturvedi (1983) using different notation. The proof is based on integration by parts. It is also a multivariate version of Stein's identity, for which a general version is given in Arnold, Castillo and Sarabia (2001).

Lemma B.1. *Let* Θ *be a compact space satisfying* C1−C3. *Let* $g_1(\mathbf{z};\boldsymbol{\theta})$ *be a polynomial in the* $(m+1)$*-dimensional vector* $\mathbf{z}$, *and* $g_2(\mathbf{u})$ *be a polynomial in the* $(m+1)$*-dimensional vector* $\mathbf{u}$. *Suppose that* $\mathbf{U}\sim N(0,\boldsymbol{\Sigma}(\boldsymbol{\theta}^0))$, *where* $\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)$ *has form* (2.3). *Then, we have*

$$\sup_{\Theta}\mathrm{E}^{\mathbf{Y},\mathbf{U}}\left\{\frac{|g_2(\mathbf{U})|\mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}\,|g_1(\boldsymbol{\eta};\boldsymbol{\theta})|\,h(\mathbf{Y},\boldsymbol{\eta};\boldsymbol{\beta})}{\mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}\,h(\mathbf{Y},\boldsymbol{\eta};\boldsymbol{\beta})}\right\}^k<\infty$$

for $k = 1, 2, 3$. *In particular, when* $g_2(\mathbf{u}) = 1$, *we have*

$$\sup_{\Theta} \mathrm{E}^{\,\mathbf{Y}} \left\{ \frac{\mathrm{E}_{\boldsymbol{\alpha}}^{\boldsymbol{\eta}}\ |g_1(\boldsymbol{\eta};\boldsymbol{\theta}^0)|\, h(\mathbf{Y},\boldsymbol{\eta};\boldsymbol{\beta})}{\mathrm{E}_{\boldsymbol{\alpha}}^{\boldsymbol{\eta}}\ h(\mathbf{Y},\boldsymbol{\eta};\boldsymbol{\beta})} \right\}^k < \infty.$$

Proof. For non-negative A and real-valued B such that the integrals exist, Hölder's inequality leads to $(\int A|B|)^c \le (\int A)^{c-1}(\int A|B|^c)$ for $c > 1$. This inequality is used twice below with $c = 2k$ and $c = 2k+1$, in a similar manner to its use in Theorem 2.1 of Nie (2006). Also, the inequality $|AB| \le A^2 + B^2$ is used once inside integrals. Let $\boldsymbol{\eta}^* = \boldsymbol{\eta} + \beta_0$, $\mathbf{z}^* = \mathbf{z} + \beta_0$, and $\mathbf{u}^* = \mathbf{u} + \beta_0$. With $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\alpha})$, define

$$\begin{aligned}
g_1^*(\mathbf{z}^*;\boldsymbol{\theta}) &= g_1(\mathbf{z}^* - \beta_0;\boldsymbol{\theta}^0) = g_1(\mathbf{z};\boldsymbol{\theta}^0)\,,\\
g_2^*(\mathbf{u}^*;\boldsymbol{\theta}) &= g_2(\mathbf{u}^* - \beta_0;\boldsymbol{\theta}^0) = g_2(\mathbf{u};\boldsymbol{\theta}^0)\,,\\
g_2^{**}(\mathbf{u}^*) &= \max\{1, |g_2^*(\mathbf{u}^*)|\}\,,\\
\varphi(\mathbf{z}^*;\boldsymbol{\theta}) &= (2\pi)^{-(m+1)/2}|\boldsymbol{\Sigma}|^{-1}\exp\bigl\{-\tfrac{1}{2}(\mathbf{z}^*-\beta_0)^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}^*-\beta_0)\bigr\}\ \text{(density of } \boldsymbol{\eta}^*\text{)}\,,\\
h^*(\mathbf{y},\mathbf{z}^*) &= \prod\nolimits_{i=1}^{m+1}\xi(y_i;\exp\{z_i^*\})\,.
\end{aligned}$$

Then,

$$\begin{aligned}
&\mathrm{E}^{\,\mathbf{Y},\mathbf{U}} \left\{ \frac{|g_2(\mathbf{U})|\mathrm{E}_{\boldsymbol{\alpha}}^{\boldsymbol{\eta}}\ |g_1(\boldsymbol{\eta};\boldsymbol{\theta})|\, h(\mathbf{Y},\boldsymbol{\eta};\boldsymbol{\beta})}{\mathrm{E}_{\boldsymbol{\alpha}}^{\boldsymbol{\eta}}\ h(\mathbf{Y},\boldsymbol{\eta};\boldsymbol{\beta})} \right\}^k\\
&= \int\!\left\{\frac{\int|g_1^*(\mathbf{z}^*;\boldsymbol{\theta})|h^*(\mathbf{y};\mathbf{z}^*)\varphi(\mathbf{z}^*;\boldsymbol{\theta})d\mathbf{z}^*}{\int h^*(\mathbf{y};\mathbf{z}^*)\varphi(\mathbf{z}^*;\boldsymbol{\theta})d\mathbf{z}^*}\right\}^k\cdot\left\{\int |g_2^*(\mathbf{u}^*)|^k h^*(\mathbf{y};\mathbf{u}^*)\varphi(\mathbf{u}^*;\boldsymbol{\theta}^0)d\mathbf{u}^*\right\}d\mathbf{y}\\
&\le \int \frac{\bigl\{\int h^*(\mathbf{y};\mathbf{z}^*)\varphi(\mathbf{z}^*;\boldsymbol{\theta}^0)\bigl(|g_1^*(\mathbf{z}^*)|\varphi(\mathbf{z}^*;\boldsymbol{\theta})/\varphi(\mathbf{z}^*;\boldsymbol{\theta}^0)\bigr)^{2k}d\mathbf{z}^*\bigr\}^{1/2}}{\bigl\{\int h^*(\mathbf{y};\mathbf{z}^*)\varphi(\mathbf{z}^*;\boldsymbol{\theta})d\mathbf{z}^*\bigr\}^k}\\
&\qquad\cdot\left\{\int h^*(\mathbf{y};\mathbf{z}^*)\varphi(\mathbf{z}^*;\boldsymbol{\theta}^0)d\mathbf{z}^*\right\}^{k-1/2}\times\left\{\int |g_2^*(\mathbf{u}^*)|^k h^*(\mathbf{y};\mathbf{u}^*)\varphi(\mathbf{u}^*;\boldsymbol{\theta}^0)d\mathbf{u}^*\right\}d\mathbf{y}\\
&\le \int \frac{\bigl\{\int h^*(\mathbf{y};\mathbf{z}^*)\varphi(\mathbf{z}^*;\boldsymbol{\theta}^0)\bigl(|g_1^*(\mathbf{z}^*)|\varphi(\mathbf{z}^*;\boldsymbol{\theta})/\varphi(\mathbf{z}^*;\boldsymbol{\theta}^0)\bigr)^{2k}d\mathbf{z}^*\bigr\}^{1/2}}{\left\{\int h^*(\mathbf{y};\mathbf{z}^*)\varphi(\mathbf{z}^*;\boldsymbol{\theta})d\mathbf{z}^*\right\}^k}\\
&\qquad\cdot\left\{\int |g_2^{**}(\mathbf{z}^*)|^k h^*(\mathbf{y};\mathbf{z}^*)\varphi(\mathbf{z}^*;\boldsymbol{\theta}^0)d\mathbf{z}^*\right\}^{k+1/2} d\mathbf{y}\\
&\le \int\!\!\int h^*(\mathbf{y};\mathbf{z}^*)\varphi(\mathbf{z}^*;\boldsymbol{\theta}^0)\Bigl(\frac{|g_1^*(\mathbf{z}^*)|\varphi(\mathbf{z}^*;\boldsymbol{\theta})}{\varphi(\mathbf{z}^*;\boldsymbol{\theta}^0)}\Bigr)^{2k}d\mathbf{z}^*d\mathbf{y}\\
&\qquad+\int \frac{\bigl\{\int |g_2^{**}(\mathbf{z}^*)|^k h^*(\mathbf{y};\mathbf{z}^*)\varphi(\mathbf{z}^*;\boldsymbol{\theta}^0)d\mathbf{z}^*\bigr\}^{2k+1}}{\bigl\{\int h^*(\mathbf{y};\mathbf{z}^*)\varphi(\mathbf{z}^*;\boldsymbol{\theta})d\mathbf{z}^*\bigr\}^{2k}}d\mathbf{y}
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}^0}\Big(\frac{|g_1^*(\boldsymbol{\eta}^*)|\varphi(\boldsymbol{\eta}^*;\boldsymbol{\theta})}{\varphi(\boldsymbol{\eta}^*;\boldsymbol{\theta}^0)}\Big)^{2k} \\
&\quad + \int \frac{\big\{\int |g_2^{**}(\boldsymbol{\eta}^*)|^k h^*(\mathbf{y};\mathbf{z}^*)\varphi(\mathbf{z}^*;\boldsymbol{\theta})\big(\varphi(\mathbf{z}^*;\boldsymbol{\theta}^0)/\varphi(\mathbf{z}^*;\boldsymbol{\theta})\big)d\mathbf{z}^*\big\}^{2k+1}}{\big\{\int h^*(\mathbf{y};\mathbf{z}^*)\varphi(\mathbf{z}^*;\boldsymbol{\theta})d\mathbf{z}^*\big\}^{2k}} d\mathbf{y} \\
&\le \mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}^0}\Big(\frac{|g_1^*(\boldsymbol{\eta}^*)|\varphi(\boldsymbol{\eta}^*;\boldsymbol{\theta})}{\varphi(\boldsymbol{\eta}^*;\boldsymbol{\theta}^0)}\Big)^{2k} \\
&\quad + \int\int h^*(\mathbf{y};\mathbf{z}^*)\varphi(\mathbf{z}^*;\boldsymbol{\theta})\Big(\frac{|g_2^{**}(\mathbf{z}^*)|^k\varphi(\mathbf{z}^*;\boldsymbol{\theta}^0)}{\varphi(\mathbf{z}^*;\boldsymbol{\theta})}\Big)^{2k+1} d\mathbf{z}^* d\mathbf{y} \\
&= \mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}^0}\Big(\frac{|g_1^*(\boldsymbol{\eta}^*)|\varphi(\boldsymbol{\eta}^*;\boldsymbol{\theta})}{\varphi(\boldsymbol{\eta}^*;\boldsymbol{\theta}^0)}\Big)^{2k} + \mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}^0}\Big\{|g_2^{**}(\boldsymbol{\eta}^*)|^{k(2k+1)}\Big(\frac{\varphi(\boldsymbol{\eta}^*;\boldsymbol{\theta}^0)}{\varphi(\boldsymbol{\eta}^*;\boldsymbol{\theta})}\Big)^{2k}\Big\}.
\end{aligned}
$$

From conditions C1−C3, the right-hand side is bounded above by some constant.

Lemma B.2. *Suppose that* $\mathbf{U} \sim N(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}^0))$. *Let* $\mathbf{u} = (u_0, u_1, \ldots, u_m)$ *be an* $(m+1)$*-dimensional vector. For each* $\mathbf{i}$, *consider*

$$\varsigma^{\mathbf{i}}(\mathbf{u}) = \int \big\{\partial^{\mathbf{i}} q(\mathbf{y};\boldsymbol{\theta}^0)\big\} h(\mathbf{y}, \mathbf{u}, \boldsymbol{\beta}^0) d\mathbf{y}$$

as a function of $\mathbf{u}$; *this is the expectation of* $\partial^{\mathbf{i}} q(\mathbf{Y};\boldsymbol{\theta}^0)$ *conditional on* $\mathbf{U} = \mathbf{u}$. *Then* $\mathrm{E}\,\big\{\mathbf{U}\,\varsigma^{\mathbf{i}}(\mathbf{U})\big\}$ *exists and is finite.*

Proof. From (2.4), $q(\mathbf{y};\boldsymbol{\theta}) = \log \mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}} h(\mathbf{y}, \boldsymbol{\eta}; \boldsymbol{\beta})$ where, in (2.5), we let $h = h_j$ for all j in the case of no covariates. Then we let $g_1(\boldsymbol{\eta};\boldsymbol{\theta}^0)$ be such that

$$\partial^{\mathbf{i}} q(\mathbf{y};\boldsymbol{\theta}^0) = \frac{\mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}^0}\, g_1(\boldsymbol{\eta};\boldsymbol{\theta}^0)\, h(\mathbf{y}, \boldsymbol{\eta}; \boldsymbol{\beta}^0)}{\mathrm{E}^{\boldsymbol{\eta}}_{\boldsymbol{\alpha}^0}\, h(\mathbf{y}, \boldsymbol{\eta}; \boldsymbol{\beta}^0)},$$

where g_1 is a polynomial (Section 2.3 and proof of Lemma A.1). With $g_2(\boldsymbol{\eta})$ being a component of $\boldsymbol{\eta}$, the conclusion now follows from Lemma B.1.

Lemma B.3. *Suppose that* $\rho_t \approx C\rho^t$ *for a positive constant* C *and* $-1 < \rho < 1$ *when* $t \to \infty$. *Then we have the autocovariance*

$$\mathrm{Cov}^{\mathbf{Y}}(\partial^{i_1} q(Y_1, \ldots, Y_{m+1};\boldsymbol{\theta}^0), \partial^{i_2} q(Y_{t+1}, \ldots, Y_{t+m+1};\boldsymbol{\theta}^0)) = O(\rho^t),$$

where $i_1, i_2 \in \{\boldsymbol{\beta}, \sigma_V^2, \phi_1, \ldots, \phi_p\}$.

Proof. Let $\mathbf{U}_1 \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{11})$ and $\mathbf{U}_2 \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{22})$ be independent $(m+1)$-dimensional Gaussian random vectors. Define $\mathbf{V} = \rho^t \mathbf{A}_t \mathbf{U}_1 + (\mathbf{I} - \rho^{2t}\mathbf{B}_t)^{1/2}\mathbf{U}_2$, where $\mathbf{A}_t = \mathbf{\Sigma}_{21,t}\mathbf{\Sigma}_{11}^{-1}$, $\mathbf{B}_t = \mathbf{\Sigma}_{21,t}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12,t}\mathbf{\Sigma}_{22}^{-1}$. Then $\mathbf{V} \sim N(\mathbf{0}, \mathbf{\Sigma}_{22})$ (same distribution as $\mathbf{U}_2$). Further let $\mathbf{\Sigma}_{11} = \mathbf{\Sigma}_{22} = \mathbf{\Sigma}(\boldsymbol{\alpha})$, as given in (2.3), and let

$$\mathbf{\Sigma}_{12,t} = \mathbf{\Sigma}_{21,t}^T = \rho^{-t} \begin{pmatrix} \gamma_t & \gamma_{t+1} & \cdots & \gamma_{t+m} \\ \gamma_{t-1} & \gamma_t & \cdots & \gamma_{t+m-1} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{t-m} & \gamma_{t-m+1} & \cdots & \gamma_t \end{pmatrix}.$$

Here, the autocovariance vector $\boldsymbol{\gamma}$ is computed based on true distribution with parameter $\boldsymbol{\theta}^0$. Then, $(\mathbf{U}_1, \mathbf{V})$ has the same law as $(\eta_0, \ldots, \eta_m, \eta_t, \ldots, \eta_{t+m})$. For any i and $\mathbf{u}$, define $\varsigma_\ell^i(\mathbf{u})$ as in Lemma B.2. It can be seen that

$$\begin{aligned} &\rho^{-t}\mathrm{Cov}^{\mathbf{Y}}(\partial^{i_1}q(Y_0, \ldots, Y_m; \boldsymbol{\theta}^0), \partial^{i_2}q(Y_t, \ldots, Y_{t+m}; \boldsymbol{\theta}^0)) \\ &= \rho^{-t}\mathrm{E}\left\{\varsigma^{i_1}(\mathbf{U}_1)\varsigma^{i_2}(\mathbf{V}) - \varsigma^{i_1}(\mathbf{U}_1)\varsigma^{i_2}(\mathbf{U}_2)\right\} = [w(\rho^t) - w(0)]/\rho^t, \qquad \text{(B.1)} \end{aligned}$$

where

$$w(\epsilon) = \mathrm{E}\left\{\varsigma^{i_1}(\mathbf{U}_1)\,\varsigma^{i_2}\Big(\epsilon\mathbf{A}_t\mathbf{U}_1 + (\mathbf{I} - \epsilon^2\mathbf{B}_t)^{1/2}\mathbf{U}_2\Big)\right\}, \quad 0 \le \epsilon < 1.$$

We show that the limit of (B.1) can be evaluated under the expectation sign by differentiating w under the expectation sign. Define $\mathbf{A} = \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}$, where $\mathbf{\Sigma}_{12} = \lim_{t\to\infty}\mathbf{\Sigma}_{12,t}$, i.e.,

$$\mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{21}^T = C\gamma_0 \begin{pmatrix} 1 & \rho & \cdots & \rho^m \\ \rho & 1 & \cdots & \rho^{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^m & \rho^{m-1} & \cdots & 1 \end{pmatrix}.$$

Then, the limit of right-hand side of (B.1) becomes

$$\begin{aligned} &\lim_{t\to\infty}\rho^{-t}\mathrm{Cov}^{\mathbf{Y}}(\partial^{i_1}q(Y_0, \ldots, Y_m; \boldsymbol{\theta}^0), \partial^{i_2}q(Y_t, \ldots, Y_{t+m}; \boldsymbol{\theta}^0)) = w'(0) \\ &= \mathrm{E}\left\{\varsigma_\ell^{i_1}(\mathbf{U}_1)\mathbf{U}_1^T \cdot \mathbf{A}^T \cdot \frac{\partial}{\partial\mathbf{u}_2}\varsigma_k^{i_2}(\mathbf{U}_2)\right\} \\ &= \mathrm{E}\left\{\varsigma_\ell^{i_1}(\mathbf{U}_1)\mathbf{U}_1^T\right\} \cdot \mathbf{A}^T \cdot \mathrm{E}\left\{\frac{\partial}{\partial\mathbf{u}_2}\varsigma_k^{i_2}(\mathbf{U}_2)\right\} \\ &= \mathrm{E}\left\{\varsigma_\ell^{i_1}(\mathbf{U}_1)\,\mathbf{U}_1^T\right\} \cdot \mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} \cdot \mathrm{E}\left\{\mathbf{U}_2\,\varsigma_k^{i_2}(\mathbf{U}_2)\right\} = O(1). \end{aligned}$$

In the last line above, we have used Novikov's theorem and Lemma B.2 for the existence of $\mathrm{E}\{\mathbf{U}_1\varsigma_k^{i_1}(\mathbf{U}_1)\}$ and $\mathrm{E}\{\mathbf{U}_2\varsigma_k^{i_2}(\mathbf{U}_2)\}$.

Lemma B.4. *Let* $\boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{\Sigma}(\boldsymbol{\theta}^0))$ *with* $\mathbf{\Sigma}$ *as defined in* (2.3). *Let* $\omega_1, \boldsymbol{\omega}_2, \boldsymbol{\omega}_3$ *be, respectively, a real-valued constant scalar, vector, and matrix satisfying*

$$\mathrm{E}_{\boldsymbol{\theta}}^{\boldsymbol{\eta}}(\omega_1 + \boldsymbol{\omega}_2^T\boldsymbol{\eta} + \boldsymbol{\eta}^T\boldsymbol{\omega}_3\boldsymbol{\eta})h(\mathbf{y}, \boldsymbol{\eta}; \boldsymbol{\beta}^0) = 0 \qquad \text{(B.2)}$$

for any given $(m+1)$*-dimensional vector* $\mathbf{y}$. *Then,* $\omega_1 = 0$, $\boldsymbol{\omega}_2 = \mathbf{0}$ *and* $\boldsymbol{\omega}_3 = \mathbf{0}$.

Proof. For $\mathbf{z} = (z_1, \ldots, z_{m+1})^T$, let $g(\mathbf{z}) = \omega_1 + \boldsymbol{\omega}_2^T\mathbf{z} + \mathbf{z}^T\boldsymbol{\omega}_3\mathbf{z}$. The left-hand side of (B.2) is the integral transform of the function

$$\frac{1}{(2\pi)^{(m+1)/2}|\boldsymbol{\Sigma}|^{1/2}} g(\mathbf{z}) \cdot \exp\Big\{-\tfrac{1}{2}\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z} + b(e^{\beta_0+z_1}) + \cdots + b(e^{\beta_0+z_{m+1}})\Big\} \quad \text{(B.3)}$$

with kernel $\prod_{i=1}^{m+1} \exp\big[a(e^{\beta_0+z_i})T(y_i)\big]$. Since the inverse integral transform of zero must be zero, we have $g \equiv 0$. By noting that g is quadratic, we have $\omega_1 = 0$, $\boldsymbol{\omega}_2 = \mathbf{0}$, $\boldsymbol{\omega}_3 = \mathbf{0}$.

Appendix C. Adaptive Gauss-Hermite Quadrature

The integrals in the composite likelihood have the form $\mathrm{E}\,[g(\mathbf{Z})]$, where $\mathbf{Z} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $d \geq 2$. Let $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$ be the Cholesky decomposition of $\boldsymbol{\Sigma}$ where $\mathbf{A}$ is lower triangular, and let $\mathbf{Z}_0 \sim N_d(\mathbf{0}, \mathbf{I}_d)$. Then $\mathrm{E}\,[g(\mathbf{Z})] = \mathrm{E}\,[g(\boldsymbol{\mu} + \mathbf{A}\mathbf{Z}_0] = \mathrm{E}\,[g_0(\mathbf{Z}_0)]$, where $g_0(\mathbf{z}) = g(\boldsymbol{\mu} + \mathbf{A}\mathbf{z})$. Using d-dimensional Gauss-Hermite quadrature with n_q points per dimension, $\mathrm{E}\,[g_0(\mathbf{Z}_0)]$ is evaluated as

$$\sum_{i_1=1}^{n_q} \cdots \sum_{i_d=1}^{n_q} w^*_{i_1 n_q} \cdots w^*_{i_1 n_q} g_0(x^*_{i_1 n_q}, \ldots, x^*_{i_d n_q}), \quad \text{(C.1)}$$

where $x^*_{in_q} = x_{in_q}\sqrt{2}$, $w^*_{in_q} = \pi^{-1/2} w_{in_q}$, and x_{in_q} are the roots of the Hermite polynomial of order n_q, w_{in_q} are the Gauss-Hermite weights when integrating against e^{-x^2} (see Stroud and Secrest (1966)).

To get around the curse of dimensionality as the dimension d increases, and to reduce n_q, adaptive Gauss-Hermite quadrature can be used when the function g is positive (such as for a term in the composite likelihood). With ϕ_d as the d-variate normal density, write

$$\begin{aligned}\mathrm{E}\,[g(\mathbf{Z})] &= \int g(\mathbf{z})\,\phi_d(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\,d\mathbf{z} = \int g(\mathbf{z})\,\frac{\phi_d(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\phi_d(\mathbf{z}; \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)} \phi_d(\mathbf{z}; \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)\,d\mathbf{z} \\ &= \mathrm{E}\Big[g(\mathbf{Z}^*)\,\frac{\phi_d(\mathbf{Z}^*; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\phi_d(\mathbf{Z}^*; \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)}\Big] = \mathrm{E}\,[g^*(\mathbf{Z}_0)], \qquad \text{(C.2)}\end{aligned}$$

where $\mathbf{Z}^* \sim N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$, $\boldsymbol{\Sigma}_p = \mathbf{A}_p\mathbf{A}_p^T$ and

$$g^*(\mathbf{z}) = g(\boldsymbol{\mu}_p + \mathbf{A}_p\mathbf{z})\,\phi_d(\boldsymbol{\mu}_p + \mathbf{A}\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})/\phi_d(\boldsymbol{\mu}_p + \mathbf{A}\mathbf{z}; \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p).$$

With $g > 0$, $\boldsymbol{\mu}_p$ is chosen as the argmin of $k(\mathbf{z}) = -\log g(\mathbf{z}) - \log \phi_d(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}_p$ is the inverse Hessian of k at $\boldsymbol{\mu}_p$. The parameters $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p^{-1}$ can be obtained via the Newton-Raphson method.

The final expectation in (C.2) can be evaluated like (C.1). For mixed effect models based on multivariate normal random effects, Pinheiro and Chao (2006) and Joe (2008) show that adaptive Gauss-Hermite quadrature often works well with $n_q = 3$ or 1 (latter corresponds to Laplace approximation).

Our implementation is in code in the C programming language, in order to quickly run sets of simulations. We use $n_q = 3$ after comparisons against $n_q = 1$ and $n_q = 5$. Composite likelihood methods for models such as those in this paper can be implemented in any statistical software with a numerical quasi-Newton optimizer. Code for Gauss-Hermite quadrature points and weights are available from several sources including at least one R package (`http://www.r-project.org`).

References

Arnold, B. C., Castillo, E. and Sarabia, J. M. (2001). A multivariate version of Stein's identity with applications to moment calculations and estimation of conditionally specified distributions. *Comm. Statist. Theory Methods* **30**, 2517-2542.

Brijs, T., Karlis, D. and Wets, G. (2008). Studying the effect of weather conditions on daily crash counts using a discrete time-series model. *Accident Analysis and Prevention* **40**, 1180-1190.

Chan, K. S. and Ledolter, J. (1995). Monte Carlo estimation for time series models involving counts. *J. Amer. Statist. Assoc.* **90**, 242-252.

Chaturvedi, S. (1983). Gaussian stochastic processes. In *Stochastic Processes Formalism and Applications* (Edited by G.S. Agarwak and S. Dattagupta), 19-29. Lecture Notes in Physics, Springer, Berlin.

Doob, J. L. (1953). *Stochastic Processes.* Wiley, New York.

Harvey, A. C., Ruiz, E. and Shephard, N. (1994). Multivariate stochastic variance models. *Rev. Econom. Stud.* **61**, 247-264.

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Comput. Statist. Data Anal.* **52**, 5066-5074.

Joe, H. and Lee, Y. (2009). On weighting of bivariate margins in pairwise likelihood. *J. Multivariate Anal.* **100**, 670-685.

Jones, M. C. (1987). Randomly choosing parameters for the stationarity and invertibility region of autoregressive-moving average models. *Appl. Statist.* **36**, 134-148.

Jung, R. C., Kukuk, M. and Liesenfeld, R. (2006). Time series of count data: modeling, estimation and diagnostics. *Comput. Statist. Data Anal.* **51**, 2350-2364.

Lunn, D. J., Thomas, A., Best, N. G. and Spiegelhalter, D. J. (2000). WinBUGS–A Bayesian modelling framework: concepts, structure, and extensibility. *Statist. Comput.* **10**, 325-337.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92**, 162-170.

Meyer, R. and Yu, J. (2000). BUGS for a Bayesian analysis of stochastic volatility models. *Econometrics J.* **3**, 198-215.

Nash, J. C. (1990). *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation.* Second edition. Springer, New York.

Nie, L. (2006). Strong consistency of the maximum likelihood estimator in generalized linear and nonlinear mixed-effects models. *Metrika* **63**, 123-143.

Pinheiro, J. C. and Chao, E. C. (2006). Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *J. Comput. Graph. Statist.* **15**, 58-81.

Qu, J. (2008). Composite Likelihood for a Stochastic Volatility Model for Financial Time Series. *Master's Essay*, Department of Statistics, University of British Columbia.

Richard, J. F. and Zhang, W. (2007). Efficient high-dimensional importance sampling. *J. Econometrics* **141**, 1385-1411.

Sandmann, G. and Koopman, S. J. (1998). Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *J. Econometrics* **87**, 271-301.

Sermaidis, G. I. (2006). Modelling time series of counts with an application on daily car accidents. M.Sc. Thesis, Athens University of Economics and Business.

Straumann, D. (2005). *Estimation in Conditionally Heteroscedastic Time Series Models.* Springer, New York.

Stroud, A. H. and Secrest, D. (1966). *Gaussian Quadrature Formulas.* Prentice-Hall, Englewood Cliffs, NJ.

Varin, C. (2008). On composite marginal likelihoods. *Adv. Statist. Anal.* **92**, 1-28.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* **92**, 519-528.

Varin, C. and Vidoni, P. (2006). Pairwise likelihood inference for ordinal categorical time series. *Comput. Statist. Data Anal.* **51**, 2365-2373.

Weiß, C. H. (2008). Thinning operations for modeling time series counts - a survey. *Adv. Statist. Anal.* **92**, 319-341.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621-629.

Department of Applied Mathematics, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.

E-mail: machitim@inet.polyu.edu.hk

Department of Statistics, University of British Columbia, Vancouver, BC, V6T 1Z2 Canada.

E-mail: harry@stat.ubc.ca

Department of Statistics, Athens University of Economics and Business, Athens, Greece.

E-mail: karlis@aueb.gr

Department of Mathematics and Statistics, University of Saskatchewan, 106 Wiggins Road Saskatoon, SK S7N 5E6 Canada.

E-mail: liu@math.usask.ca

(Received September 2009; accepted August 2010)