

Models and inference for on-off data via clipped Ornstein–Uhlenbeck processes

Emil Aas Stoltenberg¹ and Nils Lid Hjort

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF OSLO

AUGUST 24, 2018

ABSTRACT. In this paper we study a model for sequences of nonequidistant binary data. We view each binary (zero-one) sequence as being generated by a latent Ornstein–Uhlenbeck process being above or below a certain threshold. The processes are not monitored continuously in time, but observed at irregular intervals. The states of the observational units between the observation times are unknown and the observational units need not be observed at the same points in time. In our model both features of the latent process and the thresholds are modelled as function of covariates. Inference on the parameters of this model are achieved via what we call a quasi-likelihood approach. We prove consistency and derive the limiting distribution of the quasi-likelihood estimator. An advantage of our method compared to other methods for recurrent event data is that all types of censoring (left, right and intermittent) need not be tended to. The motivation for our model came from a data set pertaining to the incidence of diarrhoea among Brazilian children growing up under rather harsh conditions in the metropolitan area of Salvador. We analyse these data with our model and contrast the results with intensity based counting process analysis of the same data.

KEYWORDS: binary data, zero-one clipped Gaussian process, time series, composite likelihood, quasi-likelihood, pairwise likelihood, Kullback–Leibler, misspecification, model selection.

¹emilas@math.uio.no

1 Introduction

Many phenomena, in the biosciences, epidemiology, engineering and the social sciences, are recorded over time as being in one of two states, ‘on or off’. An epidemiologist has follow-up data on the health history of a number of individuals; an engineer on the stress level of a given component over time; a sociologist on the political preferences of a group of individuals. Such data are often binary (sick/healthy, overheated/temperate, left/right), and the data gathering is not performed continuously in time, but irregularly with differing lengths between the observation times, thus generating sequences of zero and ones where the state of the observational unit between the observation times is unknown. The zero-one data that inspired the present study have previously been analysed in Borgan et al. (2007), and is plotted in Figure 1. The plot shows 925 Brazilian children living in the

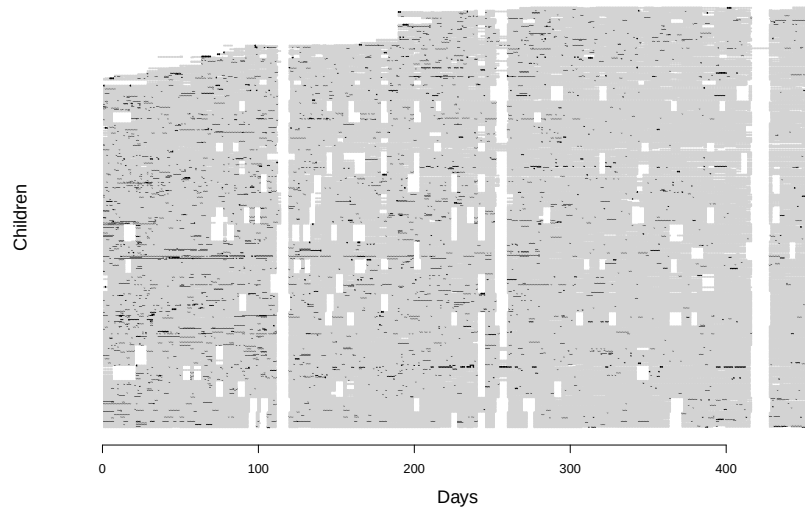


Figure 1: Observation pattern and actual observations for diarrhoea data. Horizontal lines indicate where data for each child are available, while the black dots indicate that an event was observed.

metropolitan area of Salvador, observed over a period of 455 days. In Figure 1 they are plotted according to their identification number, with the horizontal lines indicating whether the child suffered from diarrhoea (black dots) or were in good health (grey dots)

at the time of observation. The blank areas are times at which no observations were made, thus we see that the data are left-, right- and intermittently censored.

Approaches to data such as those in Figure 1 include intensity based counting-process methods, renewal processes, and models for the time between events (Borgan et al., 2007; Andersen et al., 1993). A comprehensive overview of recurrent event methods is given by Cook and Lawless (2007). Borgan et al. (2007) developed a recurrent event version of Aalen’s additive hazard model, and fitted such models to the Brazilian data.

In this paper we study models where the sequences of zeros and ones are generated by latent and independent Ornstein–Uhlenbeck processes being below or above certain thresholds. Differences between the observational units are accounted for by letting features of the Ornstein–Uhlenbeck processes as well as the thresholds be governed by covariates. More precisely, we develop methods for doing inference on the effect of covariates on (i) the covariance structure of the underlying process - allowing us to say something about how different factors affect the tenacity of the child’s health, and on (ii) the effect of these same or other factors on how prone a child is to fall sick. These two features of our a model are rather different, and subject matter knowledge on a case to case basis is required to decide which covariates should enter what regression structure. Given that such subject matter knowledge is available, the fact that our model and methods allow the covariates to affect the data generating process in two different ways is appealing as it potentially permits for a more detailed understanding of the phenomenon under study. Another appealing property of our method, and one which sets it apart from the above mentioned approaches, is that by explicitly modelling the mechanism generating the zeros and ones, all three types of censoring are taken care of without further efforts.

The paper proceeds as follows. Section 2 introduces the latent processes ticking in the background as well as the observational scheme, based on which we derive our model. The complexity of the likelihood of the model presented in Section 2 makes it computationally infeasible to maximise, Section 2.2 therefore presents what we call the quasi-likelihood method of estimation. This and related estimation methods are variably called quasi-, pairwise- and composite-likelihood (Varin and Vidoni, 2005; Hjort and Varin, 2008; Hjort

and Omre, 1994; Nott and Rydén, 1999; Varin, 2008; Varin et al., 2011; Cox and Reid, 2004). An excellent review paper of quasi/composite-likelihoods methods is Varin et al. (2011). The quasi-likelihood theory in this paper is to some extent a special case of the theory presented in that paper. In Section 3.1 consistency of the quasi-maximum likelihood (qml) estimator is proved, and in Section 3.2 we derive its limiting distribution. Section 3.3 contains a study of the qml-estimator under model misspecification, and a model selection criterion is derived along the lines of Varin and Vidoni (2005). Finally, in Section 4 we analyse the Brazilian data using our clipped Ornstein–Uhlenbeck process models, and contrast the results with those obtained by the linear hazard model of Borgan et al.. Most of the proofs are deferred to the appendix.

2 Data and model

In this section we elaborate on the data generating process, and introduce the model derived on the basis of this process as well as the quasi-likelihood method of inference. Since the Brazilian data are analysed in the present study the observational units are often referred to as ‘children’, and the latent processes as ‘health processes’.

2.1 A clipped Ornstein–Uhlenbeck process

The data consist of n children observed at various points in time over a finite interval $[0, T]$. Associated with each of the $i = 1, \dots, n$ children there is a latent stochastic process $\{Z_i(t) : 0 \leq t \leq T\}$ governing the health condition of the child. The child is sick if the process is above a certain threshold, and in good health otherwise, and it is only this zero-one version of the latent process that is actually observed. Moreover, the sample is not continuously monitored, and the health status of a child is only ascertained at certain points in time. These time points might be fixed and different for each of the children, they need not be equidistant, or they might be generated according to a stochastic process independent of the underlying latent processes.

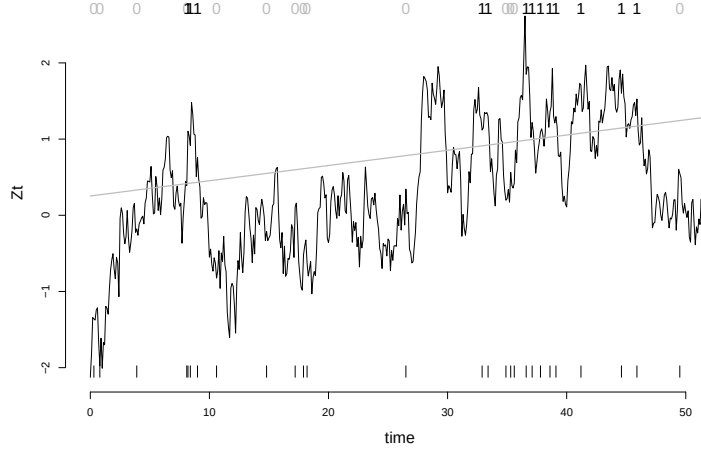


Figure 2: A sample path of a stationary Ornstein–Uhlenbeck process with the values of $Y_{i,0}, \dots, Y_{i,k_i}$ superimposed.

The observation times are denoted by

$$0 \leq t_{i,0} < t_{i,1} < \dots < t_{i,k_i} \leq T, \quad i = 1, \dots, n.$$

Let $\tau_i = \{t_{i,0}, \dots, t_{i,k_i}\}$ be the set of observation times of the i 'th child and define $\Delta_{i,j} = t_{i,j} - t_{i,j-1}$ for $i = 1, \dots, n$ and $j = 1, \dots, k_i$. The zero-one sequences available for analysis are given by

$$Y_{i,j} = I\{Z_i(t_{i,j}) \geq c_i(t_{i,j})\}, \quad j = 0, \dots, k_i,$$

where $c_i(t)$ is the possibly time-varying child specific threshold above which the child is sick. Let $Y_i = (Y_{i,0}, Y_{i,1}, \dots, Y_{i,k_i})$ be the zero-one vector of the i 'th child.

In this paper we take the $Z_1(t), \dots, Z_n(t)$ to be independent and stationary Ornstein–Uhlenbeck processes. More precisely, the health process of the i 'th child is a mean zero Ornstein–Uhlenbeck process with covariance function

$$\text{Cov}(Z_i(t), Z_i(s)) = \exp(-a_i|t - s|),$$

for a non-negative parameter a_i . The children may differ among each other and over time with regard to how prone they are to falling ill (frailty), and they may also differ in how fast their health is changing. Our model captures differences in frailty among the children

and over time through the thresholds $c_1(t), \dots, c_n(t)$. Differences in oscillation of the child specific health processes are accounted for by the parameters $a_1, \dots, a_n$. If covariates are available, say $x_i = (1, x_{i,1}, \dots, x_{i,r_1})^t$ and $z_i(t) = (1, z_{i,1}(t), \dots, z_{i,r_2}(t))^t$, both a_i and $c_i(t)$ can be modelled as functions of these. We propose a model with

$$a_i = \exp(x_i^t \beta), \quad \text{and} \quad c_i(t) = z_i^t(t) \gamma, \quad (2.1)$$

where the vectors x_i and z_i can be identical, overlapping or disjoint, and the vector $z_i(t)$ might contain time varying elements. Notice that there is no loss in generality by letting the marginal variance of the $Z_i(t)$ processes be 1 and by not explicitly introducing a drift parameter. In effect, the introduction of such parameters would lead to an overparametrisation of the model: a drift parameter would be indistinguishable from the time varying thresholds, while a variance parameter would have been absorbed into the γ coefficients. Finally, let

$$D_i = (Y_i, \tau_i, \{z_i(t)\}_{t \in \tau_i}, x_i), \quad i = 1, \dots, n,$$

be the observed data, and $D = (Y, \tau, \{z(t)\}_{t \in \tau}, x)$ denote a generic such observation vector.

Figure 2 displays one sample path of the process defined in this section. The wiggly line is a sample path of a stationary Ornstein–Uhlenbeck process, the sampling times are indicated by the ticks on the time axis, and the values of $Y_{i,j}$ (the zeros and ones) are superimposed on the plot.

The true likelihood contribution of the i 'th child is the multivariate normal probability,

$$L_i(\theta) = \Pr_{\theta}\{Y_{i,0} = y_{i,0}, Y_{i,1} = y_{i,1}, \dots, Y_{i,k_i} = y_{i,k_i}\}, \quad (2.2)$$

which is, for moderate k_i and n , computationally burdensome to compute. Moreover, a clipped Ornstein–Uhlenbeck process is no longer Markov (Slud, 1989), so the likelihood contribution in (2.2) cannot be factorised. In other words, likelihood inference based on $\prod_{i=1}^n L_i(\theta)$ is, in most interesting cases, infeasible. To deal with this issue we propose what we call the quasi-likelihood approach for inference on the parameters governing the underlying Ornstein–Uhlenbeck processes as well as on the parameters determining the varying frailties of the individuals under study. The quasi-likelihood is the topic of the next section.

2.2 The quasi-likelihood approach

Define the $(r_1 + r_2 + 2) \times 1$ vector $\theta = (\beta, \gamma)^t$, and the probabilities

$$p_{uv}(\theta, i, j) = \Pr_{\theta}\{Y_{i,j-1}, Y_{i,j} = (u, v)\},$$

for (u, v) in the set $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$. Throughout the article we write ‘ $\sum_{(u,v)}$ ’ for sums over this set, and follow the same convention for ‘ $\prod_{(u,v)}$ ’, ‘ $\min_{(u,v)}$ ’, and so on. The bivariate normal probabilities p_{uv} can easily be computed with the formulas presented in Section A.1 of the appendix. To not overburden the notation, we may drop some of the arguments from $p_{uv}(\theta, i, j)$ when it is clear from the context what probability we are referring to.

The quasi-likelihood approach consists of approximating the true likelihood in (2.2) with the pairwise construction

$$Q_n(\theta) = \prod_{i=1}^n \prod_{j=1}^{k_i} \prod_{(u,v)} p_{uv}(\theta, i, j)^{I\{(Y_{i,j-1}, Y_{i,j})=(u,v)\}}. \quad (2.3)$$

We define the functions q and q_j by

$$Q_n(\theta) = \prod_{i=1}^n q(\theta, D_i) = \prod_{i=1}^n \prod_{j=1}^{k_i} q_j(\theta, D_i),$$

where the j 'th quasi-likelihood contribution of the i 'th child, that is $q_j(\theta, D_i)$, is given by

$$\begin{aligned} q_j(\theta, D_i) &= p_{00}(\theta, i, j)^{(1-Y_{i,j-1})(1-Y_{i,j})} p_{11}(\theta, i, j)^{Y_{i,j-1}Y_{i,j}} \\ &\quad \times p_{01}(\theta, i, j)^{(1-Y_{i,j-1})Y_{i,j}} p_{10}(\theta, i, j)^{Y_{i,j-1}(1-Y_{i,j})}. \end{aligned}$$

The quasi-maximum likelihood estimator $\hat{\theta}_n$ is the value of θ maximising $Q_n(\theta)$. Notice that $q_j(\theta, Y_i)$ is the probability mass function of a multinomial experiment with four outcomes, hence it is a proper likelihood contribution for such an experiment, and the Bartlett identity

$$\mathbb{E} \frac{\partial}{\partial \theta} \log q_j(\theta_0, D_i) \left(\frac{\partial}{\partial \theta} \log q_j(\theta_0, D_i) \right)^t = -\mathbb{E} \frac{\partial^2}{\partial \theta \partial \theta^t} \log q_j(\theta_0, D_i), \quad (2.4)$$

holds, where θ_0 denotes the true parameter value and the expectation is taken under θ_0 .

In a similar manner the quasi-score function $U_n(\theta) = \partial Q_n(\theta) / \partial \theta$ and its contributions

$u(\theta, D_i)$ and $u_j(\theta, D_i)$ are defined by

$$U_n(\theta) = \sum_{i=1}^n u(\theta, D_i) = \sum_{i=1}^n \sum_{j=1}^{k_i} u_j(\theta, D_i),$$

and the same goes for the second derivative matrices $h(\theta, D_i) = \partial u(\theta, D_i)/\partial \theta$, $h_j(\theta, D_i) = \partial u_j(\theta, D_i)/\partial \theta$ and $H_n(\theta) = n^{-1} \sum_{i=1}^n h(\theta, D_i) = n^{-1} \sum_{i=1}^n \sum_{j=1}^{k_i} h_j(\theta, D_i)$. In the next section we prove consistency of the estimator $\hat{\theta}_n$ and derive its limiting distribution under conditions similar to the regularity conditions in maximum likelihood estimation.

Remark 1. The pairwise quasi-likelihood we consider in this paper can of course be extended to involve triplets of observations, i.e. $(Y_{i,j-1}, Y_{i,j}, Y_{i,j+1})$, quadruples and so on. Such strategies were pursued in Hjort and Varin (2008) for the Markov Chain case. Quasi-likelihood constructions of this type are likely to yield more efficient estimators, at the cost, however, of a larger computational burden. Another pairwise construction not considered in this paper, is to let the quasi-likelihood involve all pairs of child specific observations, that is $(Y_{i,l}, Y_{i,j})$ for all $l \neq j$, and not only the adjacent ones.

3 Large sample properties

For a vector x , $\|x\|$ is the Euclidean norm, while for a function z , $\|z\|_\infty = \sup\{|z(t)| : t \in [0, T]\}$ is the uniform norm. Define the norm

$$\|z_i\|_{\max} = \max\{\|z_{i,1}\|_\infty, \dots, \|z_{i,q}\|_\infty\}.$$

Throughout the paper we assume that the following hold:

- (A1) The parameter space $\Theta \subset \mathbb{R}^{r_1+r_2+2}$ is compact;
- (A2) The matrices $\sum_{i=1}^n x_i x_i^t$ and $\sum_{i=1}^n \sum_{j=0}^{k_i} z_i(t_{i,j-1}) z_i(t_{i,j})^t$ are positive definite and $\|x_i\|$ and $\|z_i\|_{\max}$ are uniformly bounded;
- (A3) None of the processes $Z_1(t), \dots, Z_n(t)$ are degenerate; in particular $\rho(\theta_0, i, j)$ is bounded above by $\rho_{\max} < 1$ for all j and i .

For a given parameter value θ there are $4 \sum_{i=1}^n k_i$ probabilities of the type $p_{uv}(\theta, i, j)$ involved in the full quasi-likelihood. In the following we let $p_{\min}(\theta)$ be the smallest of these probabilities, and define

$$p^*(\theta_0) = \inf_{\theta \in B_\varepsilon(\theta_0)} p_{\min}(\theta),$$

where $B_\varepsilon(\theta_0) = \{\theta: \|\theta - \theta_0\| < \varepsilon\}$ for $\varepsilon > 0$. Assumptions (A1)-(A3) entail that $p^*(\theta_0)$ is bounded below.

3.1 Consistency

Define the Kullback–Leibler divergence of the quasi-likelihood by

$$\text{KL}(\theta, \tau, x, z) = \text{E} \log \frac{q(\theta_0, D)}{q(\theta, D)} = \sum_{j=1}^k \text{E} \log \frac{q_j(\theta_0, D)}{q_j(\theta, D)},$$

and

$$\text{KL}(\theta) = \int \text{KL}(\theta, \tau, x, z) \, d\nu, \quad (3.1)$$

in terms of a hypothesised distribution function ν , seen as the limit of the empirical distribution function (as in Schweder and Hjort (2016, Appendix A.5)). That is, by the law of large numbers

$$\text{KL}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \text{E} \log \frac{q(\theta_0, D_i)}{q(\theta, D_i)} \xrightarrow{p} \text{KL}(\theta),$$

for every θ . Since Q_n consists of proper multinomial likelihood elements, standard techniques (cf. Ferguson (1996, Ch. 17)) can be applied to prove that $\text{KL}(\theta)$ is non-negative and equals zero if and only if $\theta = \theta_0$. This is the content of Lemma A.1 in the appendix.

We have the following result.

Theorem 3.1. *Assume (A1) and (A2). The value $\hat{\theta}_n$ maximising $Q_n(\theta)$ is consistent for θ_0 .*

Proof. The function $q(\theta, D)$ is continuous in θ for all D , and $\log q(\theta, D)/q(\theta_0, D)$ is bounded above by the integrable function $-\sum_{j=1}^k \log q_j(\theta_0, D)$. It follows from Theorem 16(b) in Ferguson (1996, p. 109) that

$$\sup_{\theta \in \Theta} \left| \log \frac{Q_n(\theta)}{Q_n(\theta_0)} + \text{KL}(\theta) \right| \xrightarrow{p} 0.$$

By Lemma A.1 the function $\text{KL}(\theta)$ attains its maximum in the unique point θ_0 . Theorem 5.7 in van der Vaart (1998, p. 45) then gives the result. $\square$

3.2 Asymptotic normality

Assume that the covariates and the sequences of observation times $\tau_i, i = 1, \dots, n$ are fixed. The results are easily extended to the situation where the observation times (or the covariates) are generated according to some stochastic process independent of the underlying Ornstein–Uhlenbeck processes.

Due to the Bartlett identity in (2.4), the variance of the quasi-score function is given by

$$\text{Var}(U_n(\theta_0)) = \sum_{i=1}^n \text{Var}(u(\theta_0, D_i)) = \sum_{i=1}^n \text{E} [h(\theta_0, D_i) + 2c(\theta_0, D_i)],$$

in terms of the matrix

$$c(\theta_0, D_i) = \sum_{1 \leq l < j \leq k_i} u_l(\theta_0, D_i) u_j(\theta_0, D_i)^t.$$

Assumptions (A2) and (A3) entail that both $h(\theta_0, D_i)$ and $c(\theta_0, D_i)$ are integrable for all i . A variant of the multidimensional Lindeberg–Lévy Theorem can now be applied to find the limiting distribution of $n^{-1/2}U_n(\theta_0)$ (see Schweder and Hjort (2016, Appendix A.5)).

Theorem 3.2. *Assume that (A1)–(A3) hold, that $\max_{1 \leq i \leq n} k_i / \sqrt{n} \rightarrow 0$ and that $n^{-1} \sum_{i=1}^n \text{E} (h(\theta_0, D_i) + 2c(\theta_0, D_i)) \rightarrow H + 2C$. Then*

$$n^{-1/2}U_n(\theta_0) \xrightarrow{d} N(0, H + 2C).$$

Proof. From Lemma A.2 we have that $\|u_j(\theta_0, D_i)\| \leq \|x_i\|K_1 + \|z_i\|_{\max}K_2$ for all i and j and finite constants K_1 and K_2 (see Appendix A.3 for details). Hence, for all i

$$\|u(\theta_0, D_i)\| \leq k_i (\|x_i\|K_1 + \|z_i\|_{\max}K_2) \leq \max_{1 \leq i \leq n} k_i (\|x_i\|K_1 + \|z_i\|_{\max}K_2).$$

This gives that

$$\sum_{i=1}^n \text{E} \|u(\theta_0, D_i)\|^2 I\{\|u(\theta_0, D_i)\| \geq \sqrt{n}\varepsilon\} \rightarrow 0,$$

provided $\max_{1 \leq i \leq n} k_i / \sqrt{n} \rightarrow 0$ as n tends to infinity. $\square$

In order to get the limiting distribution of the estimator $\hat{\theta}_n$ we need to prove that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = H^{-1} \frac{1}{\sqrt{n}} U_n(\theta_0) + o_p(1). \quad (3.2)$$

By Theorem 5.23 in van der Vaart (1998, p. 53) the equality in (3.2) holds if the quasi-likelihood functions $\log q(\theta, X)$ are Lipschitz in a neighbourhood of the true θ_0 and H is invertible. That this is indeed the case is proved in Theorem A.4. To summarise, we have that *on the model* the estimator $\hat{\theta}_n$ is consistent for the true value θ_0 , and that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, H^{-1}(H + 2C)H^{-1}).$$

In the next section we consider the state of affairs when the model on which the quasi-likelihood is based is misspecified.

3.3 Outside model conditions and model selection

Suppose that the observed sequences of zeros and ones are generated by Ornstein–Uhlenbeck processes $Z_i(t)$ with covariance functions

$$\text{Cov}(Z_i(t), Z_i(s)) = \exp(-a_0(x_i)|t - s|),$$

for some continuous, non-negative and bounded function $a_0(x)$ (so that (A3) is satisfied), and that the thresholds are $c_0(z_i(t))$, $i = 1, \dots, n$, with $c_0(z(t))$ a bounded function. Otherwise, both functions are unknown. This means that for each t and s , $(Z_i(t), Z_i(s))$ follows a mean zero bivariate normal distribution with covariance matrix

$$\begin{pmatrix} 1 & \exp(-a_0(x_i)|t - s|) \\ \exp(-a_0(x_i)|t - s|) & 1 \end{pmatrix}.$$

In this situation the parametric models of the form given in (2.1) can be viewed as parametric approximations to the true model, perhaps lying out of reach of the parametric models employed for estimation. The quasi-likelihood estimator $\hat{\theta}_n$ then aims at the *least-false* parameter value θ_0 minimising the distance

$$\text{KL}(\theta) = \int \mathbb{E}_0 \log \frac{q_0(\tau, x, z)}{q(\theta, \tau, x, z)} d\nu, \quad (3.3)$$

where $q_0(\Delta, x, z)$ is the hypothetical quasi-likelihood based on the true underlying distribution, and the expectation is taken with respect to the true distribution. Let $\theta_{0,n}$ be the parameter value minimising the hypothetical empirical version of (3.3), namely $n^{-1} \sum_{i=1}^n \mathbb{E}_0 \log q_0(D_i)/q(\theta, D_i)$. We have the following result.

Proposition 3.3. *The estimator $\hat{\theta}_n$ is consistent for the least-false parameter value θ_0 and*

$$\sqrt{n}(\hat{\theta}_n - \theta_{0,n}) \xrightarrow{d} N(0, H^{-1}(K + 2C)H^{-1}).$$

where H and C are the limits of $\mathbb{E}_0 H_n(\theta_{0,n})$ and $n^{-1} \mathbb{E}_0 \sum_{i=1}^n c(\theta_{0,n}, D_i)$, respectively, while K is the limit of $K_n = n^{-1} \sum_{i=1}^n \sum_{j=1}^{k_i} \text{Var}_0(u_j(\theta_{0,n}, D_i))$.

Proof. Follows, with minor modifications, from the results given in Section 3.1 and Section 3.2. $\square$

If the parametric model we are employing happens to be the correct one, then $K = H$, since the Bartlett identity is then back in force.

The model presented in display (2.1) contains two regression models, one on the covariance of the Ornstein–Uhlenbeck processes and one on the thresholds. Since each covariate available for analysis might enter into one, the other or both of these, we are left having to choose between $4^{r_1+r_2}$ different models. Without subject matter knowledge, this quickly becomes an insurmountable task. We therefore introduce the quasi-likelihood information criterion (qlic) which can assist in choosing a good model. The qlic is a special case of the composite likelihood information criterion introduced in Varin and Vidoni (2005), and is given by

$$\text{qlic} = \log Q_n(\hat{\theta}_n) - \text{tr}(\hat{H}_n^{-1}(\hat{K}_n + 2\hat{C}_n)), \quad (3.4)$$

where $\hat{H}_n$, $\hat{K}_n$ and $\hat{C}_n$ are consistent estimators of H , C and K . Among the candidate models, the model with the highest value of the qlic is to be preferred. The derivation of the qlic follows the same principles as the derivation of the AIC in the standard likelihood case (see e.g. Claeskens and Hjort (2008, Ch. 2.3)), and can thus be viewed as an AIC for the quasi-likelihood method of inference. Consistent estimators of the matrices H , C and K are $H_n(\hat{\theta}_n)$, $n^{-1} \sum_{i=1}^n c(\hat{\theta}_n, D_i)$, and $n^{-1} \sum_{i=1}^n \sum_{j=1}^{k_i} u_j(\hat{\theta}_n, D_i)u_j(\hat{\theta}_n, D_i)^t$, respectively.

4 The Brazilian data

The data analysed in this section have previously been studied in Borgan et al. (2007) using a version of Aalen’s linear hazard model. A more elaborate discussion of the data is also found in that paper. For comparative purposes, we briefly present the model of Borgan et al. and fit one such to the data, thereafter, we fit three different latent Ornstein–Uhlenbeck process models.

4.1 The data and previous modelling strategies

As part of a sanitation programme in the metropolitan area of Salvador, Brazil, the Institute of Public Health at the Federal University of Bahia conducted several studies and data gathering efforts. One of these consisted of surveying the extent to which infants in the Salvador area suffered from episodes of diarrhoea. Data collectors were assigned to households and conducted home visits over a period of 455 days from October 2000 to January 2002. One child aged under three years at entry was monitored from each household. A major challenge with these data is the different types of missingness, clearly visible by the white rectangles in Figure 1. Some 16 percent of the children entered late into the study and about 21 percent of the children dropped out of the study before the completion date. In addition, there are the intermittent missingness, whereby observation was interrupted but later resumed. According to Borgan et al., this type of missingness was mainly due to the data collector not being available. Cases where the child was not available for a home visit are more problematic as they can invalidate our assumption of the observation times being independent of the underlying processes. The data collectors were assigned contiguous identification numbers, which explains the white rectangles visible in Figure 1. The periods during which there are no observations are due to periods of vacation and a strike among the data collectors.

Borgan et al. (2007) studied a class of counting process models for recurrent events data in discrete time with linear hazard rates

$$\alpha_i(t) = \beta_0(t) + \beta_1(t)x_{i1}(t) + \cdots + \beta_p(t)x_{ip}(t), \quad i = 1, \dots, n, \quad (4.1)$$

	Estimates	Standard errors
Baseline	3.464	0.283
≥ 3 bedroom	2.560	0.301
≤ 12 months	3.438	0.293
> 24 months	-3.866	0.223
Contaminated water storage	-0.918	0.251
Standing water	0.482	0.299
Contaminated water source	1.799	0.268
Other children ≤ 5 years	0.897	0.215
Male	0.673	0.207
Rain-affected accomodation	2.101	0.251
Mother < 25 years	1.647	0.218
Open sewerage	4.163	0.379
Poor street quality	-0.582	0.241
Low social economic class	0.033	0.233

Table 1: Estimated cumulative regression coefficients, $\sum_{s=1}^t \beta_j(s)$, with $t = 228$. Point-wise 95% confidence intervals in parentheses.

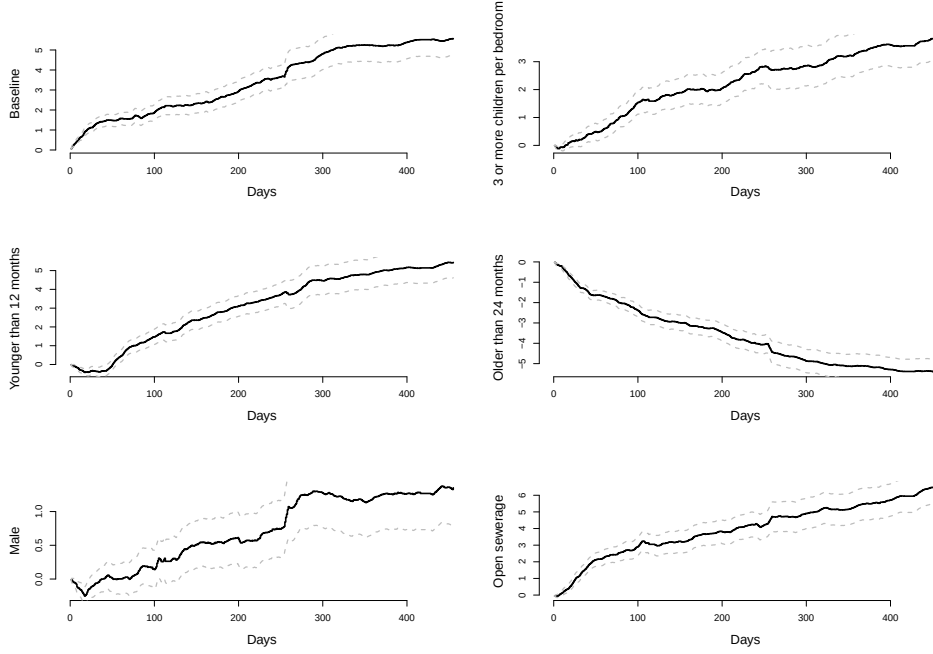


Figure 3: Selected cumulative regression coefficients with pointwise 95% confidence intervals.

and used martingale methods to derive the limiting distribution of the estimators of the cumulative coefficients $B_k(t) = \sum_{s=1}^t \beta_k(s)$, $k = 0, 1, \dots, p$. Due to the three forms of censoring they had to introduce a 'missingness process' indicating whether a child was observed or for what reason the child was not observed (Borgan et al., 2007, p. 57). The value of this process at a given time must be assumed known prior to this time (i.e. it must be predictable), and it is assumed conditionally independent of the counting process, given the past. This latter assumption is similar to our assumption of the observations times being independent of the latent Ornstein–Uhlenbeck processes (given covariates), while the former is immaterial for our model. In our model a missing value is a missing value, the reason for which it is missing being irrelevant (as long as the reason for the observation being missing is independent of the state of $Z_i(t)$).

Table 1 displays estimated cumulative regression coefficients for a model of the form (4.1), and Figure 3 shows the entire sample paths for six of these.

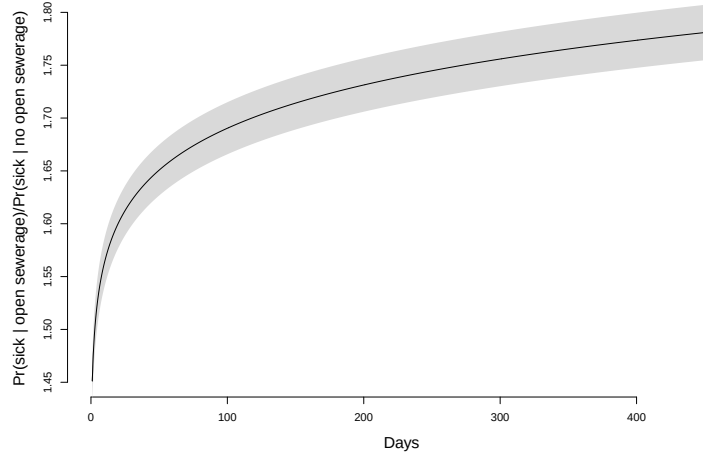


Figure 4: The estimated ratio of the probability of being sick when living and not living in the proximity of open sewerage, for boys whose mothers are below 25 years old living in rain affected accommodation. Estimates are based on the Medium model of Table 2. Pointwise 95% confidence intervals.

4.2 Fitting clipped Ornstein–Uhlenbeck process models

We fitted three different clipped Ornstein–Uhlenbeck process models to the Brazilian data. The results are summarised in Table 2. In the big model we see that six of the estimated coefficients are not significant at the 95 percent level, removing these we obtain the medium model. As one would expect, removing these covariates improves the qlic score somewhat. The small model only includes the two intercepts and the log of time since the start of the study. This model appears to be too parsimonious as its qlic score is inferior to the two bigger models.

The linear hazard model and the clipped Ornstein–Uhlenbeck process models are very different models, so one should be careful in comparing the two. It is, however, interesting to note that some of the covariates seem to have an effect in the linear hazard model of Table 1, but not in the big Ornstein–Uhlenbeck process model of Table 2 (the converse is only true for one of the covariates). Two possible reasons for this could be the efficiency loss due to the quasi-likelihood estimation, or the fact that four of the insignificant coefficients in the big clipped Ornstein–Uhlenbeck model enter the covariance function

of the Ornstein–Uhlenbeck process and therefore play a very different role in this model compared to the linear hazard model, where they work directly on the hazard.

This brief discussion highlights the importance of meticulously thinking through which covariates should enter what regression structure of the clipped Ornstein–Uhlenbeck process model, a task that, admittedly, requires a certain *fingerspitzengefühl* for the phenomenon under study.

According to the qlic the medium model provides the best fit to the data among the three models in Table 2. The negative estimate of the coefficient on the binary age variable (> 24 months), indicates that children older than two years tend to have less oscillating health than those below two years. The effects of consuming water from a contaminated water storage and living in the proximity of standing water seem to work in the same direction, i.e. by attenuating the oscillation of the underlying health processes. These effects likely work in the same direction, but for rather different reasons. Older children are less prone to falling sick and have longer streaks of good health, while children exposed to polluted water are likely to stay ill for longer periods of time when they fall ill.

In Figure 4 we have plotted the estimated ratio

$$\frac{\Pr_{\theta}(\text{ill at } t \mid \text{open sewerage})}{\Pr_{\theta}(\text{ill at } t \mid \text{no open sewerage})},$$

for boys whose mothers are below 25 years old and that are living in rain affected accommodation, along with a pointwise 95 percent confidence band. The upward sloping curve indicates that the sanitation programme had a larger effect in the areas that were not plagued by open sewerage at the start of the programme.

5 Concluding remarks

In this section some possible extensions of the clipped Ornstein–Uhlenbeck process model are briefly introduced, along with some complementing remarks.

Remark 2. Dependency/Contagion. Since diarrhoea is a highly contagious disease, a shortcoming of the model introduced in this paper is that possible dependencies between

	Big			Medium			Small		
	Estimate	St. err.	Wald	Estimate	St. err.	Wald	Estimate	St. err.	Wald
Covariance	β_0	-2.022	0.054				-2.293	0.010	
	≥ 3 bedroom	-0.101	0.065						
	≤ 12 months	-0.048	0.057						
	> 24 months	-0.544	0.088	-6.159	-0.518	0.079	-6.539		
	Contaminated water storage	-0.167	0.068	-2.451	-0.164	0.068	-2.408		
	Standing water	-0.203	0.05	-4.090	-0.216	0.048	-4.473		
	Contaminated water source	-0.092	0.065	-1.426					
	Other children ≤ 5 years	0.084	0.059	1.414					
Threshold	γ_0	1.410	0.034		1.426	0.030		1.244	0.021
	Male	-0.067	0.028	-2.395	-0.065	0.028	-2.320		
	Rain-affected accomodation	-0.103	0.032	-3.208	-0.098	0.032	-3.100		
	Mother < 25 years	-0.149	0.028	-5.254	-0.144	0.028	-5.136		
	Open sewerage	-0.257	0.041	-6.355	-0.245	0.039	-6.225		
	Poor street quality	0.039	0.032	1.209					
	Low social economic class	0.001	0.031	0.029					
	$\log(\text{time})$	0.162	0.003	54.792	0.162	0.003	54.971	0.161	0.003
qlic			-45969.170				-45958.644		
							-46406.492		

Table 2: Parameter estimates and approximate standard errors for three Ornstein–Uhlenbeck process models. The upper panel contains the $\hat{\beta}$ entering the covariance function, while the lower panel shows the $\hat{\gamma}$ of the threshold c .

the children is not taken into account.

An extension of our model that accounts for the dependency between the children is obtained by taking the zero-one processes $Y_i(t) = I\{Z_i(t) \geq c_i(t)\}$ as above, but with $Z_i(t)$ a ‘spatial’ Gaussian process $\{Z_i(t) : i = 1, \dots, n, 0 \leq t \leq T\}$. Such a model does, of course, demand that we have information about the geographical proximity of the children to each other. Recall that for the Brazilian data the data collectors were administered contiguous identification numbers, most likely because children with contiguous identification numbers lived in the same neighbourhood. For the Brazilian data we could for example take

$$\text{Cov}(Z_i(t), Z_l(s)) = \exp(-a_1|t - s| - a_2d(i, j)),$$

for nonnegative constants a_1 and a_2 and some distance d . The child specific covariates would then enter the threshold functions. One extension of the quasi-likelihood appropriate for this model is to consider pairs of observations ‘along the x - and y -axis’, thus letting the extended quasi-likelihood consists of probabilities of the type

$$\Pr_{\theta}\{(Y_{i,j-1}, Y_{i,j}) = (u, v)\}, \quad j = 1, \dots, k_i, \quad i = 1, \dots, n$$

as above, but also

$$\Pr_{\theta}\{(Y_{i,j}, Y_{i+1,j}) = (u, v)\}, \quad i = 1, \dots, n_j - 1, \quad j = 0, 1, \dots, \max_{1 \leq i \leq n} k_i$$

where n_j is the number of children with a j ’th observation. Other constructions along the lines of Hjort and Omre (1994) and Nott and Rydén (1999) could also be worked with.

Remark 3. Focused inference for the quasi-likelihood. The qlic introduced in (3.4) assesses overall goodness of fit issues of the model. In many situation the research question concerns a clearly specified statistical quantity, and the statistical model works as a vehicle in providing inference about this specified quantity. The Focused information criterion (FIC) (Claeskens and Hjort, 2003, 2008; Jullum and Hjort, 2017) takes this into account and aims at selecting the optimal model in terms of mean squared error for a pre-specified statistical quantity. FIC-theory can be developed for the quasi-likelihood worked with in this paper, as well as for the more general composite likelihood case (Varin et al., 2011),

thus yielding the possibility of focused model selection in cases where the full likelihood is computationally infeasible.

Remark 4. Endogenous observation times. The quasi-likelihood theory developed in this paper assumed that the observations times $t_{i,0}, t_{i,1}, \dots, t_{i,k_i}$ were independent of the underlying process $Z_i(t)$. As noted in Section 4.1, this assumption need not always hold. A future research project could be the development of quasi-likelihood theory for the case where the process generating the observation times is dependent on the underlying health process.

Remark 5. Multistate data and several thresholds. Consider multistate data where the states are, at least, on the ordinal level of measurement. Data on the stages of a disease might be of this type. Pursuing the idea of this paper, one could consider models where the indicator function $Y_i(t)$ takes on more than two values, and where the different states corresponds to the level of an underlying continuous process. For example, a three-state model is

$$Y_i(t) = \begin{cases} 0, & Z_i(t) \leq c_{i,1}(t), \\ 1, & c_{i,1}(t) < Z_i(t) \leq c_{i,2}(t), \\ 2, & Z_i(t) > c_{i,2}(t), \end{cases}$$

where $c_{i,1}(t)$ and $c_{i,2}(t)$ are possibly time varying thresholds that do not cross, and $Z_i(t)$ is an Ornstein–Uhlenbeck process of the type introduced in Section 2.1.

Remark 6. Crossings data. Suppose that what we observe are the times at which the underlying process crosses the thresholds, in either direction. Such data would for example arise if each of the children in the Brazil data were continuously monitored. The true likelihood would in this situation consist of probabilities of the type

$$\Pr(\{Z_i(t) \geq c_i(t), t \in [0, t_{i,1})\}, \dots, \{Z_i(t) \leq c_i(t), t \in [t_{i,k_i-1}, t_{i,k_i})\}),$$

with $t_{i,1}, \dots, t_{i,k_i}$ the times of the crossings. This is a rather different object from the likelihood given in (2.2), and a quasi-likelihood of (2.3) would in this situation incur a larger efficiency loss than when the true likelihood is that of (2.2). Since likelihoods

consisting of probabilities such as the one above are computationally burdensome, quasi-likelihood techniques ought to be developed for this kind of data.

Acknowledgements

We thank Robin Henderson for kindly providing us with the Brazil data. The work of E.A. Stoltenberg was supported by the PharmaTox Strategic Research Initiative, Faculty of Mathematics and Natural Sciences, University of Oslo. The work of N.L. Hjort has been supported by the Norwegian Research Foundation, via the five-year project FocuStat (Focus Driven Statistical Inference With Complex Data, led by Hjort).

A Appendix

A.1 Handy formulae

Let

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2) \right\}, \quad (\text{A.1})$$

be the bivariate normal density we are working with. Since $(Z_{i,j}, Z_{i,j-1})$ is a mean zero bivariate normal with covariance matrix

$$\begin{pmatrix} 1 & \rho(i, j) \\ \rho(i, j) & 1 \end{pmatrix},$$

where $\rho(i, j) = \exp\{-a_i \Delta_{i,j}\}$, standard facts about the multivariate normal distribution give that

$$p_{00}(\theta, i, j) = \int_{-\infty}^{c_i(t_{i,j})} \int_{-\infty}^{c_i(t_{i,j-1})} f(x, y) \, dx \, dy = \int_{-\infty}^{c_i(t_{i,j})} \Phi\left(\frac{c_i(t_{i,j-1}) - \rho(i, j)y}{\sqrt{1-\rho(i, j)^2}}\right) \phi(y) \, dy,$$

where $\phi(x)$ and $\Phi(x)$ are the standard normal density and distribution function, respectively. Moreover,

$$\begin{aligned} p_{01}(\theta, i, j) &= \Phi(c(t_{i,j-1})) - p_{00}(\theta, i, j), \\ p_{10}(\theta, i, j) &= \Phi(c(t_{i,j})) - p_{00}(\theta, i, j), \\ p_{11}(\theta, i, j) &= 1 - \Phi(c(t_{i,j-1})) - \Phi(c(t_{i,j})) + p_{00}(\theta, i, j). \end{aligned} \quad (\text{A.2})$$

These identities are used in some of the proofs below.

A.2 Consistency

Lemma A.1. *Assume (A2). The function $\text{KL}(\theta)$ in (3.1) is non-negative, and $\text{KL}(\theta) = 0$ if and only if $\theta = \theta_0$.*

Proof. Since the function $\log x$ is concave,

$$\text{KL}(\theta, \Delta, x, z) = -\mathbb{E} \sum_{j=1}^k \log \frac{q_j(\theta, D)}{q_j(\theta_0, D)} \geq -\sum_{j=1}^k \log \mathbb{E} \frac{q_j(\theta, D)}{q_j(\theta_0, D)} = 0, \quad (\text{A.3})$$

where we have used that

$$\mathbb{E} \frac{q_j(\theta, D)}{q_j(\theta_0, D)} = \sum_{(u,v)} \frac{q_j(\theta, D)}{q_j(\theta_0, D)} q_j(\theta_0, D) = \sum_{(u,v)} p_{uv}(\theta, j) = 1.$$

This shows that $\text{KL}(\theta, \Delta, x, z) \geq 0$, and so is $\text{KL}(\theta)$.

If $\theta = \theta_0$ then $\text{KL}(\theta) = 0$. Assume that $\text{KL}(\theta) = 0$, we must show that this implies that $\theta = \theta_0$. The log-likelihood of an multinomial experiment with four outcomes, say,

$$y_1 \log p_{00} + y_2 \log p_{11} + y_3 \log p_{01} + y_4 \log(1 - p_{00} - p_{11} - p_{01}),$$

is concave, hence it has a unique maximum. In our setting $\text{KL}_n(\theta) = 0$ implies that

$$0 = \text{KL}(\theta, \Delta_i, x_i, z_i) = \mathbb{E} \sum_{j=1}^{k_i} \log \frac{q_j(\theta_0, D_i)}{q_j(\theta, D_i)},$$

for all i . But by (A.3) this may only occur if the ratios $q_j(\theta_0, D)/q_j(\theta, D) = 1$ for each $j = 1, \dots, k_i$. Due to the concavity of the multinomial experiment, this entails that all the individual probabilities must be equal, i.e. $p_{uv}(\theta, i, j) = p_{uv}(\theta_0, i, j)$ for $(u, v) \in \{(0, 0), (1, 0), (0, 1), (1, 1)\}$. But these equalities imply that $c_i(t_{i,j}) = c_{0,i}(t_{i,j})$ and $a_i = a_{0,i}$ for each j and i . To see this use the identity from Section A.1, giving that $p_{01}(\theta, i, j) = p_{01}(\theta_0, i, j)$ is the same as

$$\Phi(c_i(t_{i,j})) - p_{00}(\theta, i, j) = \Phi(c_{0,i}(t_{i,j})) - p_{00}(\theta_0, i, j). \quad (\text{A.4})$$

Since $p_{uv}(\theta, i, j) = p_{uv}(\theta_0, i, j)$ for all (u, v) , the equality in (A.4) gives that $\Phi(c_i(t_{i,j})) = \Phi(c_{0,i}(t_{i,j}))$, which entails $c_i(t_{i,j}) = c_{0,i}(t_{i,j})$ because Φ is an injective function. But then

$a_i = a_{0,i}$ because $p_{00}(\theta, i, j)$, as a function of a_i with $c_{0,i}$ fixed, is also injective. The argument so far gives that

$$c_i(t_{i,j}) = z_i(t_{i,j})^t \gamma = z_i(t_{i,j})^t \gamma_0 = c_{0,i}, \quad \text{and} \quad a_i = e^{x_i^t \beta} = e^{x_i^t \beta_0} = a_{0,i},$$

for all j and $i = 1, \dots, n$. These equalities hold if $\sum_{i=1}^n x_i x_i^t$ and $\sum_{i=1}^n \sum_{j=1}^{k_i} z_i(t_{i,j}) z_i(t_{i,j})^t$ are positive definite, which they are by assumption. Taking limits we obtain the implication $\text{KL}(\theta) = 0 \Rightarrow \theta = \theta_0$. $\square$

A.3 Auxiliary results

In this section we derive some bounds on the quasi score function that are used in the proofs of Lemma A.2 and Theorem 3.2. Due to the identities in (A.2) the j 'th term of the i 'th quasi-score function can be expressed as

$$\begin{aligned} u_j(\theta, D_i) &= \frac{\partial}{\partial \theta} \log q_j(\theta, D_i) \\ &= \left[\begin{aligned} &\frac{\partial}{\partial \beta} p_{00}(\theta, i, j) A_{i,j}(\theta) \\ &\frac{\partial}{\partial \gamma} p_{00}(\theta, i, j) A_{i,j}(\theta) - z_i(t_{i,j-1}) \phi(c_i(t_{i,j-1})) B_{i,j}(\theta) - z_i(t_{i,j}) \phi(c_i(t_{i,j})) C_{i,j}(\theta) \end{aligned} \right] \end{aligned}$$

in terms of the functions

$$\begin{aligned} A_{i,j}(\theta) &= \frac{(1 - Y_{i,j-1})(1 - Y_{i,j})}{p_{00}(\theta, i, j)} + \frac{Y_{i,j-1} Y_{i,j}}{p_{11}(\theta, i, j)} - \frac{(1 - Y_{i,j-1}) Y_{i,j}}{p_{01}(\theta, i, j)} - \frac{Y_{i,j-1} (1 - Y_{i,j})}{p_{10}(\theta, i, j)}, \\ B_{i,j}(\theta) &= \frac{Y_{i,j-1} Y_{i,j}}{p_{11}(\theta, i, j)} - \frac{(1 - Y_{i,j-1}) Y_{i,j}}{p_{01}(\theta, i, j)}, \\ C_{i,j}(\theta) &= \frac{Y_{i,j-1} Y_{i,j}}{p_{11}(\theta, i, j)} - \frac{Y_{i,j-1} (1 - Y_{i,j})}{p_{10}(\theta, i, j)}. \end{aligned}$$

The vector of partial derivatives of p_{00} with respect to β is

$$\frac{\partial}{\partial \beta} p_{00}(\theta, i, j) = -x_i a_i \Delta_{ij} \rho(i, j) \frac{\partial}{\partial \rho} p_{00}(\theta, i, j), \quad (\text{A.5})$$

and using Leibniz integral rule we have that the vector of partial derivatives of p_{00} with respect to γ is

$$\frac{\partial}{\partial \gamma} p_{00}(\theta, i, j) = z_i(t_{i,j}) b_1(i, j) + z_i(t_{i,j-1}) b_2(i, j), \quad (\text{A.6})$$

in terms of the functions

$$b_1(i, j) = \Phi\left(\frac{c_i(t_{i,j-1}) - \rho(i, j)c_i(t_{i,j})}{\sqrt{1 - \rho(i, j)^2}}\right)\phi(c_i(t_{i,j})),$$

$$b_2(i, j) = \frac{1}{\sqrt{1 - \rho(i, j)^2}} \int_{-\infty}^{c_i(t_{i,j})} \phi\left(\frac{c_i(t_{i,j-1}) - \rho(i, j)x}{\sqrt{1 - \rho(i, j)^2}}\right)\phi(x) dx.$$

Now, consider (A.5). It is easy to see that $a_i \Delta_{i,j} \rho(i, j) \leq e^1$. We now derive a bound on $\partial p_{00}(\theta, i, j)/\partial \rho$. Write $g_{i,j}(\rho) = (c_i(t_{i,j-1}) - \rho z)/\sqrt{1 - \rho^2}$. By the Mean Value Theorem there is an ρ_* on the line between $\rho + \varepsilon$ and ρ such that

$$\Phi(h(\rho + \varepsilon)) - \Phi(g_{i,j}(\rho)) = \phi(g_{i,j}(\rho_*))g'_{i,j}(\rho_*)\varepsilon,$$

where $g'_{i,j}(\rho) = (c_i(t_{i,j-1})\rho - z)/(1 - \rho^2)^{3/2}$. Using the results from Section A.1 we can now write

$$\begin{aligned} p_{00}(\rho + \varepsilon, c_i, i, j) - p_{00}(\rho, c_i, i, j) &= \int_{-\infty}^{c_i(t_{i,j})} \{\Phi(g_{i,j}(\rho + \varepsilon)) - \Phi(g_{i,j}(\rho))\} \phi(z) dz \\ &= \int_{-\infty}^{c_i(t_{i,j})} \phi(g_{i,j}(\rho_*))g'_{i,j}(\rho_*)\phi(z) dz \varepsilon. \end{aligned}$$

Recall that we write $f(x, y)$ for the bivariate normal density given in (A.1). A bound on the integral above is

$$\begin{aligned} \left| \int_{-\infty}^{c_i(t_{i,j})} \phi(g_{i,j}(\rho_*))g'_{i,j}(\rho_*)\phi(z) dz \right| &= \left| \int_{-\infty}^{c_i(t_{i,j})} \frac{c_i(t_{i,j-1})\rho - z}{1 - \rho^2} f(c_i(t_{i,j-1}), z) dz \right| \\ &\leq \int_{-\infty}^{c_i(t_{i,j})} \left| \frac{c_i(t_{i,j-1})\rho - z}{1 - \rho^2} \right| f(c_i(t_{i,j-1}), z) dz \\ &\leq \int_{-\infty}^{\infty} \left| \frac{c_i(t_{i,j-1})\rho - z}{1 - \rho^2} \right| f(c_i(t_{i,j-1}), z) dz = \sqrt{\frac{2}{\pi}} \frac{\phi(c_i(t_{i,j-1}))}{\sqrt{1 - \rho^2}} \end{aligned}$$

This means that in a neighbourhood of the true θ_0 we have that for all i and j ,

$$\left| \frac{\partial p_{00}(\theta, i, j)}{\partial \rho} \right| \leq \sqrt{\frac{2}{\pi}} \frac{\phi(c_i(t_{i,j-1}))}{\sqrt{1 - \rho(i, j)^2}} \leq \frac{1}{\pi} \frac{1}{\sqrt{1 - \rho(i, j)^2}}.$$

Next, we derive a bound on $\partial p_{00}(\theta, i, j)/\partial \gamma$ in (A.6),

$$\begin{aligned} \left\| \frac{\partial}{\partial \gamma} p_{00}(\theta, i, j) \right\| &\leq \|z_i(t_{i,j})\| |b_1(i, j)| + \|z_i(t_{i,j-1})\| |b_2(i, j)| \\ &\leq \|z_i\| \phi(0) (1 + 1/\sqrt{1 - \rho(i, j)^2}). \end{aligned}$$

We have the following lemma

Lemma A.2. *For all i and j*

$$\|u_j(\theta, D_i)\| \leq \frac{4}{p_{\min}(\theta)} \left\{ \|x_i\| e^{-1} \frac{1}{\pi \sqrt{1 - \rho(i, j)^2}} + \|z_i\|_{\max} \phi(0) \left(3 + \frac{1}{\sqrt{1 - \rho(i, j)^2}} \right) \right\}.$$

In particular, under assumptions (A1)-(A3), for all $\theta \in B_\varepsilon(\theta_0)$,

$$\|u_j(\theta, D_i)\| \leq \frac{4}{p^*(\theta_0)} \left\{ \|x_i\| e^{-1} \frac{1}{\pi \sqrt{1 - \rho_{\max}^2}} + \|z_i\|_{\max} \phi(0) \left(3 + \frac{1}{\sqrt{1 - \rho_{\max}^2}} \right) \right\} < \infty.$$

Proof. Repeated applications of the triangle inequality and of the bounds derived above, as well as of the fact that $|A_{i,j}(\theta)| \leq 4/p_{\min}(\theta)$, $|B_{i,j}(\theta)| \leq 2/p_{\min}(\theta)$ and $|C_{i,j}(\theta)| \leq 2/p_{\min}(\theta)$. $\square$

Lemma A.3. *The matrix H of Theorem 3.2 is invertible.*

Proof. The Hessian of $\text{KL}(\theta)$ evaluated in θ_0 equals H . Let v be an arbitrary non-zero vector of the same dimension as θ , and α a scalar. Since θ_0 is the unique minimiser of $\text{KL}(\theta)$ (as proven in Lemma A.1), we have that

$$0 = \text{KL}(\theta_0) < \text{KL}(\theta_0 + \alpha v) = \text{KL}(\theta_0) + \frac{1}{2} \alpha^2 v^\top H v + o(\alpha^2) = \frac{1}{2} \alpha^2 v^\top H v + o(\alpha^2),$$

hence $v^\top H v / 2 > o(\alpha^2) / \alpha^2$, which shows that $v^\top H v > 0$ for every non-zero vector v . $\square$

Theorem A.4. *The quasi-likelihood contributions $\log q(\theta, D)$ are Lipschitz in a neighbourhood $B_\varepsilon(\theta_0)$ of the true value. Consequently, (3.2) holds.*

Proof. By the Mean Value Theorem there exists a $\tilde{\theta}$ in $B_\varepsilon(\theta_0)$ such that for $\theta_1, \theta_2 \in B_\varepsilon(\theta_0)$,

$$|\log q(\theta_1, D) - \log q(\theta_2, D)| \leq |u(\tilde{\theta}, D)^\top (\theta_1 - \theta_2)| \leq \|u(\tilde{\theta}, D)\| \|\theta_1 - \theta_2\|,$$

where the second inequality is the Cauchy–Schwarz inequality. Moreover, the quasi-score function $\|u(\tilde{\theta}, D)\| \leq \sup_{\theta \in B_\varepsilon(\theta_0)} \|u(\theta, D)\|$, and by Lemma A.2 the right hand side is bounded by a function that is independent of the data, hence square integrable. Theorem 5.23 in van der Vaart (1998, p. 53) in combination with Lemma A.3 then give the result. $\square$

References

- Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, Berlin.
- Borgan, Ø., Fiaccone, R. L., Henderson, R., and Barreto, M. L. (2007). Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in Brazil. *Scandinavian Journal of Statistics*, 34:53–69.
- Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98:900–916.
- Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.
- Cook, R. J. and Lawless, J. (2007). *The Statistical Analysis of Recurrent Events*. Springer, New York.
- Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91:729–737.
- Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall/CRC, Boca Raton.
- Hjort, N. L. and Omre, H. (1994). Topics in spatial statistics. *Scandinavian Journal of Statistics*, 21:289–357.
- Hjort, N. L. and Varin, C. (2008). ML, PL, QL in Markov chain models. *Scandinavian Journal of Statistics*, 35:64–82.
- Jullum, M. and Hjort, N. L. (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica*, 27:951–981.
- Nott, D. J. and Rydén, T. (1999). Pairwise likelihood methods for inference in image models. *Biometrika*, 86:661–676.

- Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability. Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge.
- Slud, E. (1989). Clipped Gaussian processes are never m -step Markov. *Journal of Multivariate Analysis*, 29:1–14.
- van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.
- Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, 92:1–28.
- Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42.
- Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92:519–528.