Received: 10 March 2019 | Revised: 9 May 2020 | Accepted: 16 May 2020

DOI: 10.1111/sjos.12472

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

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

Emil Aas Stoltenberg | **Nils Lid Hjort**

Department of Mathematics, University of Oslo, Norway

Corresponding author:
Emil Aas Stoltenberg PB 1053, Blindern, 0316 Oslo, Norway.
Email: emilas@math.uio.no

Abstract

We introduce a model for recurrent event data subject to left-, right-, and intermittent-censoring. The observations consist of binary sequences (along with covariates) for each individual under study. These sequences are modeled as generated by latent Ornstein–Uhlenbeck processes being above or below certain thresholds. Features of the latent process and the thresholds are taken as functions of covariates, allowing the researcher to distinguish factors that have an effect on the frailty, from those that have an effect on the variability, of the observational unit. Inference is achieved by a quasi-likelihood approach, for which consistency and asymptotic normality is established. An advantage of our model is that particularities regarding the censoring need not be taken actively into account, and that it is well suited for situations where the individuals under study are irregularly and asynchronously observed. The motivation for our model came from a dataset pertaining to the incidence of diarrhoea among Brazilian children growing up under rather harsh conditions. We analyze these data with our model and contrast the results with an intensity-based counting process analysis of the same data.

KEYWORDS

binary time series, censoring, clipped Gaussian process, composite-likelihood, misspecification, model selection, recurrent events

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; an economist has data on the employment status of group of individuals surveyed at various points in time. Such data are often binary (sick/healthy, overheated/temperate, employed/unemployed), and the data gathering is not performed continuously in time, but irregularly with differing lengths between the observation times. Observational schemes of this kind abound in many fields of science, and have in common that they generate sequences of zeros and ones where the state of the observational unit between the observation times is unknown (i.e., censored). The statistical modeling and analysis of this kind of data can be quite challenging because one needs to tend to the dependence in time as well as account for rather complicated patterns of censoring. An appealing feature of the model we introduce in this paper is that both the temporal dependence, as well as all types of censoring, are accommodated in a straightforward and neat manner requiring minimal statistical ingenuity on the part of the applied researcher doing the analysis.

The data that inspired the present study have previously been analyzed in Borgan, Fiaccone, Henderson, and Barreto (2007), and is shown in Figure 1. The plot shows the health status of a sample of Brazilian children living in the metropolitan area of Salvador, observed at discrete and irregular times over a period of 455 days. In the figure the children are plotted according to their identification number (the y-axis), with the color of the dots indicating whether the child suffered from diarrhoea (black dots) or were in good health (grey dots) at the time of observation. The blank dots are times at which no observation was made, thus we see that the data are left-, right-, and intermittently censored. That the observations took place at discrete and irregular times, means that the data consist of a time stamp and an indicator of the state of the child (sick/healthy) at that time, along with covariates. (In Figure 1 we have only included every 10th child in the sample of 925, because the resolution becomes problematic when the entire dataset is included.) The pattern depicted among the 93 children in the plot is characteristic for the entire sample.

Approaches to data such as those in Figure 1 include intensity-based counting-process methods, renewal processes, and models for the time between events (Andersen, Borgan, Gill, & Keiding, 1993; Borgan et al., 2007). 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 introduce a class of 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. Each covariate might enter into one or both of the regression structures of the model, thus allowing for inference on the effect of each covariate on the temporal correlation of the underlying process, its effect on the level of the process, or its effect on both these features. In the setting of the Brazilian data, this means that our model enables us to say something about how different factors affect the fluctuation of a child's health, and on how these same or other factors affect the frailty of a child.

These two features of the model are rather different, and subject matter knowledge on a case to case basis is required to decide which covariates should enter where. Given that such knowledge is available, the fact that our model allows the covariates to affect the data generating process in

FIGURE 1 Observation pattern and actual observations for diarrhoea data (for every 10th child in the sample of 925). The grey dots indicate that the child was healthy at the observation time, and the black dots indicate that the child was sick. The white areas are time points at which no observations were made

two different ways is appealing as it potentially permits for a more detailed understanding of the phenomenon under study. As already mentioned, another appealing property of our method, and one which sets it apart from the above-mentioned approaches, is that by explicitly modeling the mechanism generating the zeros and ones, all three types of censoring are taken care of without further efforts.

The paper proceeds as follows. In Section 2 we introduce 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 maximize; in Section 2.2 we therefore present what we call the quasi-likelihood method of estimation. This and related estimation methods are variably called quasi-, pairwise- and composite-likelihood (Cox & Reid, 2004; Hjort & Omre, 1994; Hjort & Varin, 2008; Nott & Rydén, 1999; Varin, 2008; Varin, Reid, & Firth, 2011; Varin & Vidoni, 2005). A lucid review paper of quasi/composite-likelihoods methods is Varin et al. (2011). In Sections 3.2–3.3 we prove consistency of the maximum quasi-likelihood estimator, and derive its limiting distribution. Section 3.4 contains a study of the maximum quasi-likelihood estimator when the chosen parametric model is misspecified, a model selection criterion is derived, and we introduce a goodness of fit measure based on the ratio of two nested quasi-likelihoods. In Section 3.5 we conduct a small simulation study to assess the asymptotic results in a finite-sample setting. Finally, in Section 4 we analyze the Brazilian data using clipped Ornstein–Uhlenbeck process models, and contrast the results with those obtained by the linear hazard model of Borgan et al. (2007). Most of the proofs are deferred to Appendix.

2 | DATA AND MODEL

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 $\{\xi_i(t) : 0 \le t \le T\}$ 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.

The observation times are denoted by

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

Let $\tau_i = \{t_{i,0}, \dots, t_{i,k_i}\}$ be the set of observation times of the ith child. The zero-one sequences available for analysis are given by

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

for $i = 1, \dots, n$, 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 ith child.

In this paper we take the $\xi_1(t), \dots, \xi_n(t)$ to be independent Ornstein–Uhlenbeck processes. More precisely, the health process of the ith child is a mean zero Gaussian process with covariance function $\mathrm{Cov}(\xi_i(t), \xi_i(s)) = \exp(-a_i|t-s|)$, for a nonnegative 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 $x_i = (1, x_{i,1}, \dots, x_{i,r_1})^{\mathrm{t}}$ and $z_i(t) = (1, z_{i,1}(t), \dots, z_{i,r_2}(t))^{\mathrm{t}}$, both a_i and $c_i(t)$ can be modeled as functions of these. We propose a model with

$$a_i = \exp(x_i^{\mathrm{t}}\beta), \quad c_i(t) = z_i^{\mathrm{t}}(t)\gamma, \tag{1}$$

where the vectors x_i and $z_i(t)$ can be identical, overlapping or disjoint, and the vector $z_i(t)$ might contain time-varying elements. In particular, in this model the probability of the ith child being ill at time t is $1 - \Phi(c_i(t))$, with $\Phi(x)$ the standard normal distribution function. Notice that there is no loss in generality by letting the marginal variance of the $\xi_i(t)$ processes be 1 and by not explicitly introducing a drift parameter. In effect, the introduction of such parameters would lead to an overparametrization of the model: A drift parameter would be indistinguishable from the time-varying thresholds, while the introduction of a variance parameter would just result in a scaling of 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.

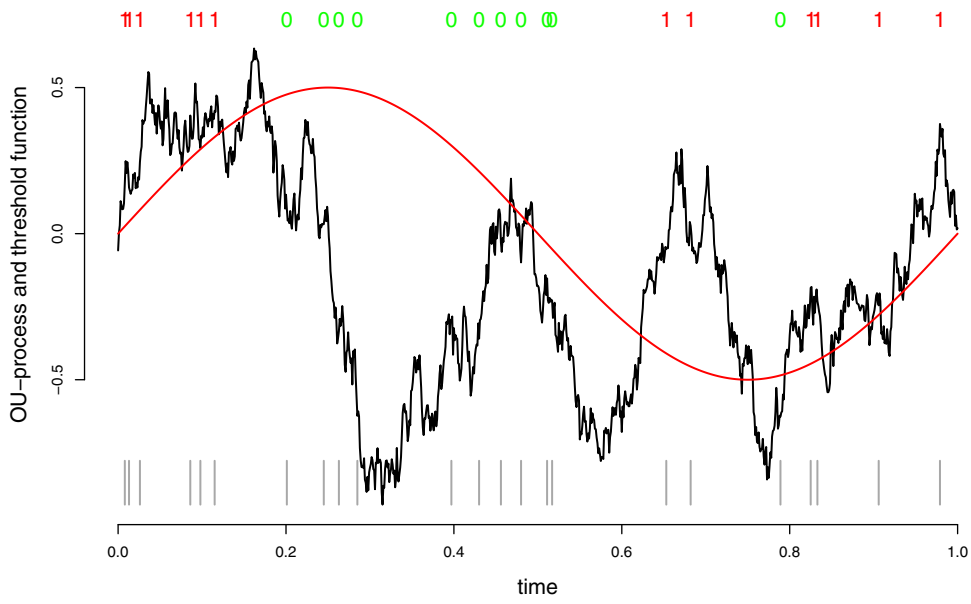

FIGURE 2 A sample path of a stationary Ornstein–Uhlenbeck process with the values of $Y_{i,0}, \ldots, Y_{i,k_i}$ superimposed. The red sine-curve is one realization of the time-varying threshold used in the simulations of Section 3.5. The grey ticks on the x-axis are the times at which observations were made [Color figure can be viewed at wileyonlinelibrary.com]

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 ith child is the multivariate normal probability,

$$L_i(\beta, \gamma) = \Pr_{\beta,\gamma}(Y_{i,0} = y_{i,0}, Y_{i,1} = y_{i,1}, \ldots, Y_{i,k_i} = y_{i,k_i}), \tag{2}$$

which is, for moderate k_i and n, computationally burdensome to compute. Moreover, contrary to an Ornstein–Uhlenbeck process itself, a *clipped* Ornstein–Uhlenbeck process is no longer Markov (Slud, 1989), so the likelihood contribution in (2) cannot be factorized. In other words, likelihood inference based on $\prod_{i=1}^{n} L_i(\beta, \gamma)$ is, excluding the trivial cases, for example, $k_i \leq 4$, 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 $\theta = (\beta^{\mathrm{t}}, \gamma^{\mathrm{t}})^{\mathrm{t}}$, 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 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) with the pairwise construction

$$Q_n(\theta)=\prod_{i=1}^{n}\prod_{j=1}^{k_i}\prod_{(u,v)}p_{uv}(\theta,i,j)^{I\{(Y_{i,j-1},Y_{i,j})=(u,v)\}}. \tag{3}$$

We define the functions q and q_j via $Q_n(\theta)=\prod_{i=1}^{n}q(\theta,D_i)=\prod_{i=1}^{n}\prod_{j=1}^{k_i}q_j(\theta,D_i)$, where the jth quasi-likelihood contribution of the ith child, that is $q_j(\theta,D_i)$, is given by

$$q_j(\theta,D_i)=p_{00}(\theta,i,j)^{(1-Y_{i,j-1})(1-Y_{i,j})}p_{11}(\theta,i,j)^{Y_{i,j-1}Y_{i,j}}p_{01}(\theta,i,j)^{(1-Y_{i,j-1})Y_{i,j}}p_{10}(\theta,i,j)^{Y_{i,j-1}(1-Y_{i,j})}.$$

The quasi-maximum likelihood estimator $\hat{\theta}_n$ is the value of θ maximizing $Q_n(\theta)$. Notice that $q_j(\theta,D_i)$ 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

$$E_{\theta_0}\frac{\partial}{\partial\theta}\log q_j(\theta_0,D_i)\Big(\frac{\partial}{\partial\theta}\log q_j(\theta_0,D_i)\Big)^{\mathrm{t}}=-E_{\theta_0}\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{t}}}\log q_j(\theta_0,D_i), \tag{4}$$

holds, where θ_0 denotes the true parameter value. 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 $U_n(\theta)=\sum_{i=1}^{n}u(\theta,D_i)=\sum_{i=1}^{n}\sum_{j=1}^{k_i}u_j(\theta,D_i)$. The jth quasi-likelihood score contribution of the ith child is

$$u_j(\theta,D_i)=\begin{pmatrix}-x_ia_i(t_{i,j}-t_{i,j-1})e^{-a_i(t_{i,j}-t_{i,j-1})}A_{i,j}(\theta)\\ z_i(t_{i,j})B_{i,j}(\theta)+z_i(t_{i,j-1})C_{i,j}(\theta)\end{pmatrix}, \tag{5}$$

in terms of the random variables

$$\begin{aligned}A_{i,j}(\theta)&=\sum_{(u,v)}I\{(Y_{i,j-1},Y_{i,j})=(u,v)\}\frac{\partial p_{uv}(\theta,i,j)/\partial\rho}{p_{uv}(\theta,i,j)},\\ B_{i,j}(\theta)&=\sum_{(u,v)}I\{(Y_{i,j-1},Y_{i,j})=(u,v)\}\frac{\partial p_{uv}(\theta,i,j)/\partial c_i(t_{i,j})}{p_{uv}(\theta,i,j)},\\ C_{i,j}(\theta)&=\sum_{(u,v)}I\{(Y_{i,j-1},Y_{i,j})=(u,v)\}\frac{\partial p_{uv}(\theta,i,j)/\partial c_i(t_{i,j-1})}{p_{uv}(\theta,i,j)},\end{aligned} \tag{6}$$

where all three derivatives are bounded as long as the correlations $\rho(\theta,i,j)$ are bounded away from 1. Note also that the derivatives appearing in $B_{i,j}$ and $C_{i,j}$ are not the same function evaluated in different time points, but derivatives with respect to differentarguments of p_{uv}.

Remark 1. The pairwise quasi-likelihood we consider in this paper can of course be extended to involve triplets of observations, that is, $(Y_{i,j-1},Y_{i,j},Y_{i,j+1})$, 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

3.1 | Assumptions and notation

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. For the covariance function on the observation grids we write $\rho(\theta, i, j)$ for $j = 1, \ldots, k_i$, $i = 1, \ldots, n$, that is $\rho(\theta, i, j) = \mathrm{Cov}_\theta(\xi_i(t_{i,j}), \xi_i(t_{i,j-1})) = \exp(-a_i|t_{i,j} - t_{i,j-1}|)$. Throughout the paper we assume that the following hold: (i) the vectors (x_i, z_i, k_i) for $i = 1, \ldots, n$ are i.i.d. from a distribution ν independent of the latent Ornstein–Uhlenbeck processes; (ii) ν is such that with probability one $(x_{i,1}, \ldots, x_{i,r_1}) \in [-K, K]^{r_1}$, z_i is continuous on $[0,T]$, the number of observations is $2 \leq k_i \leq k_{\max} < \infty$ for all i, and that given k_i, the observation times $t_{i,0} < t_{i,1} < \ldots t_{i,k_i}$ are generated from a distribution that is continuous on $[0,T]$; (iii) the matrices $n^{-1}\sum_{i=1}^n (x_i, z_i(t_{i,j-1}) + z_i(t_{i,j}))^{\mathrm{t}}(x_i, z_i(t_{i,j-1}) + z_i(t_{i,j}))$ become positive definite with probability one under ν for all j; and (iv) the parameter space $\Theta \subset \mathbb{R}^{r_1+r_2+2}$ is compact.

We note that assumption (ii) entails that none of the processes $\xi_1(t), \ldots, \xi_n(t)$ are degenerate, in particular, the correlation function $\rho(\theta_0, i, j)$, evaluated in the true parameter value, is strictly smaller than one with ν-probability one for all i and j. This implies that none of the probabilities $p_{00}, p_{01}, p_{10}, p_{11}$ approaches zero, ensuring that the random variables $A_{i,j}(\theta_0), B_{i,j}(\theta_0), C_{i,j}(\theta_0)$ are bounded when evaluated in θ_0. In particular, for any ν-integrable real valued function g we have $n^{-1}\sum_{i=1}^n g(x_i, \{z_i(t)\}_{t\in\tau_i}, \tau_i) \to \int g(x, \{z(t)\}_{t\in\tau}, \tau)\, \mathrm{d}\nu$ as $n \to \infty$, by the law of large numbers. The compactness assumption on the parameter space is used in the consistency proof below.

3.2 | Consistency

Define the Kullback–Leibler divergence $\mathrm{KL}(\theta)$ of the quasi-likelihood by

$$\mathrm{KL}(\theta) = \int \mathrm{KL}(\theta, \tau, x, z)\, \mathrm{d}\nu, \tag{7}$$

where $\mathrm{KL}(\theta, \tau, x, z) = E_{\theta_0}\ \log\{q(\theta_0, D)/q(\theta, D)\} = \sum_{j=1}^k E_{\theta_0}\ \log\{q_j(\theta_0, D)/q_j(\theta, D)\}$. Since Q_n consists of proper multinomial likelihood elements, standard techniques (e.g., Ferguson, 1996, chapter 17) can be applied to prove that $\mathrm{KL}(\theta)$ is nonnegative and equals zero if and only if $\theta = \theta_0$. This is the content of Lemma 1 in Appendix. We have the following result.

Theorem 1. *The value $\hat\theta_n$ maximizing $Q_n(\theta)$ 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 $-\sum_{j=1}^k \log q_j(\theta_0, D)$. By compactness of Θ, it follows from Theorem 16(b) in Ferguson (1996, p. 109) that $\sup_{\theta\in\Theta} |\log(Q_n(\theta)/Q_n(\theta_0)) + \mathrm{KL}(\theta)| = o_p(1)$. By

Lemma 1, the function $\mathrm{KL}(\theta)$ attains its maximum in the unique point θ_0. Theorem 5.7 in van der Vaart (1998) then gives the result. ∎

3.3 | Asymptotic normality

Due to the Bartlett identity in (4), the variance of the quasi-score function given the covariates $(x_i, z_i, k_i)_{i\leq n}$ is given by

$$\mathrm{Var}_\theta\ U_n(\theta) = \sum_{i=1}^{n}\left\{\sum_{j=1}^{k_i} E_\theta\ u_j(\theta, D_i)u_j(\theta, D_i)^{\mathrm{t}} + 2\sum_{j<l} E_\theta\ u_j(\theta, D_i)u_l(\theta, D_i)^{\mathrm{t}}\right\}.$$

Under the assumptions imposed above, this variance is finite, and assumption (i) ensures the existence of matrices

$$\begin{aligned} H(\theta) &= \int H(\theta, x, z, \tau)\ \mathrm{d}\nu = \int \sum_{j=1}^{k} E_\theta\ u_j(\theta, D)u_j(\theta, D)^{\mathrm{t}}\ \mathrm{d}\nu; \\ C(\theta) &= \int C(\theta, x, z, \tau)\ \mathrm{d}\nu = \int \sum_{j<l} E_\theta\ u_j(\theta, D)u_l(\theta, D)^{\mathrm{t}}\ \mathrm{d}\nu, \end{aligned} \tag{8}$$

such that

$$\frac{1}{n}\mathrm{Var}_\theta\ U_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\theta_0}\ u(\theta_0, D_i) \xrightarrow{\nu} H(\theta_0) + 2C(\theta_0),$$

in probability under ν. When evaluated in the true parameter values, we write $H = H(\theta_0)$ and $C = C(\theta_0)$.

Theorem 2. *The sequence $n^{-1/2}U_n(\theta_0)$ converges to a mean zero normal distribution with covariance matrix $H + 2C$.*

Proof. By assumption (i) and (ii), $n^{-1}\mathrm{Var}_{\theta_0}U_n(\theta_0)$ converges to $H + 2C$. From Lemma 2 we have that $\|u_j(\theta_0, D_i)\|$ is bounded by something that is proportional to $\|x_i\| + 2\|z_i\|_\infty$ (see Lemma 2 for details). Hence, for all i

$$\|u(\theta_0, D_i)\| \lesssim k_i(\|x_i\| + 2\|z_i\|_\infty) \leq k_{\max}(\|x_i\| + 2\|z_i\|_\infty),$$

where the right-hand side is finite with ν-probability one. The $u(\theta_0, D_1), \ldots, u(\theta_0, D_n)$ are then i.i.d. and bounded random variables, and the central limit theorem yields the result. ∎

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

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = H^{-1}n^{-1/2}U_n(\theta_0) + o_p(1). \tag{9}$$

By theorem 5.23 in van der Vaart (1998, p. 53) the equality in (9) 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 4. To summarize, we have that *at the model*, the estimator $\hat{\theta}_n$ is consistent for the true value θ_0, and that $n^{1/2}(\hat{\theta}_n - \theta_0)$ converges to a mean zero normal distribution with covariance matrix $H^{-1}(H + 2C)H^{-1}$. In the next section we consider the state of affairs when the model on which the quasi-likelihood is based is misspecified, and derive a model selection criterion and a goodness-of-fit test.

3.4 | Model selection

Suppose that the observed sequences of zeros and ones are generated by Ornstein–Uhlenbeck processes $\xi_i(t)$ with covariance functions $\mathrm{Cov}(\xi_i(t), \xi_i(s)) = \exp(-a_0(x_i)|t - s|)$ for some continuous, non-negative and bounded function $a_0(x)$; and that the thresholds are $c_0(z_i(t))$, $i = 1, \ldots, n$, with $c_0(z(t))$ a bounded function. Otherwise, both functions are unknown. In this situation the parametric models of the form given in (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_{lf} minimizing the distance

$$\mathrm{KL}(\theta) = \int E_0 \ \log \frac{q_0(\tau, x, z)}{q(\theta, \tau, x, z)} \ \mathrm{d}\nu, \tag{10}$$

where $q_0(\tau,\ x,\ z)$ is the hypothetical quasi-likelihood based on the true underlying distribution, and the expectation is taken with respect to the true distribution.

Theorem 3. *The estimator $\hat{\theta}_n$ is consistent for the least-false parameter value θ_{lf}; and*

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

where $H = \int \sum_{j=1}^{k} E_0 \ \partial u_j(\theta_{\mathrm{lf}}, D)/\partial\theta \ \mathrm{d}\nu$ *and* $C = \int \sum_{j<l}^{k} E_0 \ u_j(\theta_{\mathrm{lf}}, D)u_l(\theta_{\mathrm{lf}}, D)^{\mathrm{t}} \ \mathrm{d}\nu$, *and* $K = \int \mathrm{Var}_0(\sum_{j=1}^{k} \ u_j(\theta_{\mathrm{lf}}, D)) \ \mathrm{d}\nu$.

Proof. Follows, with minor modifications, from the results given in Sections 3.2 and 3.3. ∎

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 (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 a insurmountable task, as exploring all models would quickly exhaust the computing power at one's disposal. Given that there are only a handful of models that are deemed plausible, however, the model selection criterion and the test statistic that we now introduce can assist in choosing the best among these.

(1) The quasi-likelihood information criterion (QLIC) is a special case of the composite likelihood information criterion introduced in Varin and Vidoni (2005). It is given by

$$QLIC = \log Q_n(\hat{\theta}_n) - \mathrm{tr}\ \{\hat{H}_n^{-1}(\hat{K}_n + 2\hat{C}_n)\}, \tag{11}$$

where $\hat{H}_n$, $\hat{K}_n$, $\hat{C}_n$ are consistent estimators of H, C, 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 Akaike information criterion in the standard likelihood case (e.g., Claeskens & Hjort, 2008, chapter 2.3), and can thus be viewed as an AIC-type criterion for the quasi-likelihood method of inference. Consistent estimators of the matrices H, C and K are $\hat{H}_n = n^{-1}\sum_{i=1}^n \sum_{j=1}^{k_i} \partial u_j(\hat{\theta}_n, D_i)/\partial\theta$, $\hat{C}_n = n^{-1}\sum_{i=1}^n \sum_{j<l} u_j(\hat{\theta}_n, D_i)u_l(\hat{\theta}_n, D_i)^{\mathrm{t}}$, and $\hat{K}_n = n^{-1}\sum_{i=1}^n \sum_{j=1}^{k_i} u_j(\hat{\theta}_n, D_i)u_j(\hat{\theta}_n, D_i)^{\mathrm{t}}$, respectively.

(2) The quasi-likelihood ratio test. The adequacy of different models whose parameters have been estimated by the quasi-likelihood methods can also be assessed by the likelihood ratio inspired statistic

$$M_n = 2\{\log Q_{n,\mathrm{wide}}(\hat{\eta}_1, \hat{\eta}_2) - \log Q_{n,\mathrm{narrow}}(\tilde{\eta}_1, 0)\}. \tag{12}$$

Here we have partitioned the $(r_1 + r_2 + 2)$ parameter vector $\theta = (\beta, \gamma)$ as (η_1, η_2), with η_1 and η_2 of dimension p and q, respectively. We are interested in testing whether the subset η_2 of parameters is equal to zero. Here, $Q_{n,\mathrm{wide}}(\hat{\eta}_1, \hat{\eta}_2)$ and $Q_{n,\mathrm{narrow}}(\tilde{\eta}_1, 0)$ are the quasi-likelihoods of the model including the full parameter vector (η_1, η_2), and a narrow model where $\eta_2 = 0$; both evaluated in their respective maximizers $(\hat{\eta}_1, \hat{\eta}_2)$ and $(\tilde{\eta}_1, 0)$. For a given choice of (η_1, η_2), corresponding to a wide and a narrow model, write

$$H = \begin{pmatrix} H_{00} & H_{01} \\ H_{10} & H_{11} \end{pmatrix},$$

for the $(p+q)\times(p+q)$ matrix where H_{00}, $H_{01} = H_{10}^{\mathrm{t}}$, and H_{11} are the probability limits of $n^{-1}\sum_{i=1}^n \sum_{j=1}^{k_i} \partial^2 q_j(\eta_1, \eta_2, D_i)/(\partial\eta_1\partial\eta_1^{\mathrm{t}})$, $n^{-1}\sum_{i=1}^n \sum_{j=1}^{k_i} \partial^2 q_j(\eta_1, \eta_2, D_i)/(\partial\eta_1\partial\eta_2^{\mathrm{t}})$, and $n^{-1}\sum_{i=1}^n \sum_{j=1}^{k_i} \partial^2 q_j(\eta_1, \eta_2, D_i)/(\partial\eta_2\partial\eta_2^{\mathrm{t}})$, respectively, all evaluated in $(\eta_1, \eta_2) = (\eta_1, 0)$. Provided the conditions of Theorem 2 are satisfied, we have

$$\log Q_n(\eta_1, 0) = \log Q_n(\hat{\eta}_1, \hat{\eta}_2) - \frac{1}{2} n \begin{pmatrix} \hat{\eta}_1 - \eta_1 \\ \hat{\eta}_2 - 0 \end{pmatrix}^{\mathrm{t}} H \begin{pmatrix} \hat{\eta}_1 - \eta_1 \\ \hat{\eta}_2 - 0 \end{pmatrix} + o_p(1),$$

for estimation in the wide model, as well as

$$\log Q_n(\eta_1, 0) = \log Q_n(\tilde{\eta}_1, 0) - \frac{1}{2} n(\tilde{\eta}_1 - \eta_1)^{\mathrm{t}} H_{00}(\tilde{\eta}_1 - \eta_1) + o_p(1),$$

for estimation in the narrow model. Coupling this with

$$n^{1/2}\begin{pmatrix} \hat{\eta}_1 - \eta_1 \\ \hat{\eta}_2 - \eta_2 \end{pmatrix} = H^{-1} n^{1/2} \begin{pmatrix} U_n \\ V_n \end{pmatrix} + o_p(1), \quad \text{and} \quad n^{1/2}(\tilde{\eta}_1 - \eta_1) = H_{00}^{-1} n^{1/2} U_n + o_p(1),$$

we find that under the null hypothesis $\eta_2 = 0$,

$$M_n \xrightarrow{d} M = \begin{pmatrix} U \\ V \end{pmatrix}^{\mathrm{t}} H^{-1} \begin{pmatrix} U \\ V \end{pmatrix} - U^{\mathrm{t}} H_{00}^{-1} U, \tag{13}$$

where (U, V) is mean zero multinormal with covariance matrix $H + 2C$, with C the probability limit of $n^{-1}\sum_{i=1}^n \sum_{j<l} u_j(\eta_1, 0, D_i)u_l(\eta_1, 0, D_i)^{\mathrm{t}}$. Contrary to the standard likelihood case, the limiting

random variable M will not be a chi-square (e.g., chapter 22 of Ferguson, 1996 for Wilks theorem), but can relatively easily be simulated, via multinormal realizations of (U, V) from the $N_{p+q}(0, H + 2C)$ distribution, inserting the consistent estimators for H and C, similar to those introduced above.

3.5 | A simulation study

In this section we assess the asymptotic results in a finite-sample setting. To do so, we simulated data of the form presented in Section 2 for $n = 1,000$ individuals. Specifically, exploiting the Markovian property of the Ornstein–Uhlenbeck processes, we simulated these over a fine partition of the unit interval according to

$$\xi_i(j\Delta) = \rho_i(\Delta)\xi_i((j-1)\Delta) + (1-\rho_i(2\Delta))^{1/2}\varepsilon_{i,j}, \quad \text{for} \quad j = 1, \ldots, 1/\Delta,$$

with $\Delta = 1/10^3$; the $\varepsilon_{i,j}, j = 0, \ldots, 1/\Delta$, $i = 1, \ldots, n$ being independent standard normals; and $\xi_i(0) = \varepsilon_{i,0}$. The correlation and threshold functions were set to

$$\rho_i(\Delta) = \exp(-\exp(-0.55 + 1.23\, x_i)\Delta), \quad \text{and} \quad c_i(j\Delta) = 1.04 - 0.70\, z_{i,1} + 1.55\, z_{i,2}(j\Delta),$$

with $z_{i,2}(j\Delta) = z_{i,2}\sin(2\pi j\Delta)$, $j = 0, 1, \ldots, 1/\Delta$, and $x_i, z_{i,1}, z_{i,2}$, $i = 1, \ldots, n$ taken as independent standard normals. The time points, say $J_i \subset \{t_j : j = 0, 1, \ldots, 1/\Delta\}$, at which observations were made, were determined by independent Bernoulli sampling with success probability $1/60$, which means that the number of observations per individual is about 17. Finally, the observed zero-one processes were generated by taking $Y_{i,j_i} = I\{\xi_i(j_i\Delta) \geq c_i(j_i\Delta)\}$, $j_i \in J_i$, for $i = 1, \ldots, n$.

To these data, we fit three models; one small, one correctly specified, and one big; with the misspecification in both the small and the big model taking place in the threshold function. In the small model the time varying covariate is excluded, that is, $c_{\text{small},i} = \gamma_0 + \gamma_1 z_{i,1}$; while the big model contains an extra covariate $z_{i,3}$, also taken as independent standard normal, independent from the three other covariates; thus $c_{\text{big},i}(t) = \gamma_0 + \gamma_1 z_{i,1} + \gamma_2 z_{i,2}(t) + \gamma_3 z_{i,3}$.

In Figure 3 we display quantile–quantile plots of the estimates from the correctly specified model, that is, for $n^{1/2}(\hat\beta_j - \beta_j)/\hat\sigma_{j,j}$, $j = 1, 2$ and $n^{1/2}(\hat\gamma_j - \gamma_j)/\hat\sigma_{j+2,j+2}$, $j = 1, 2, 3$, with $\hat\sigma^2_{j,j}$ being the jth diagonal element of $\hat H_n^{-1}(\hat H_n + 2\hat C_n)\hat H_n^{-1}$. The estimators $\hat H_n$ and $\hat C_n$ are the consistent plug-in estimators introduced at the end of Section 3.4.

The simulations and the estimation were conducted in the R programming language (R Core Team, 2013). Implementing the quasi log-likelihood functions only requires computing bivariate normal probabilities of the type $\Pr_\rho\{\xi_i(t_{j-1}) \leq c_i(t_{j-1}), \xi_i(t_j) \leq c_i(t_j)\}$, which can be done by numerical integration using the formula displayed in (A2). The remaining three probabilities are then taken care of by the identities in (A3). Note, however, that these probabilities may differ for each individual and for each time point. The quasi log-likelihoods were then maximized using the `nlm()`-function in R. With the amount of data used for the simulations, that is about $n \times (n^{-1}\sum_{i=1}^n k_i) \approx 1{,}000 \times 17$ individual data points, the optimization took about 10 min on a standard portable computer.

For each simulation we computed the QLIC score for each of the three models. In 86 of 100 simulations, the QLIC selected the true model above the big model, and it always chose the true model over the smaller model.

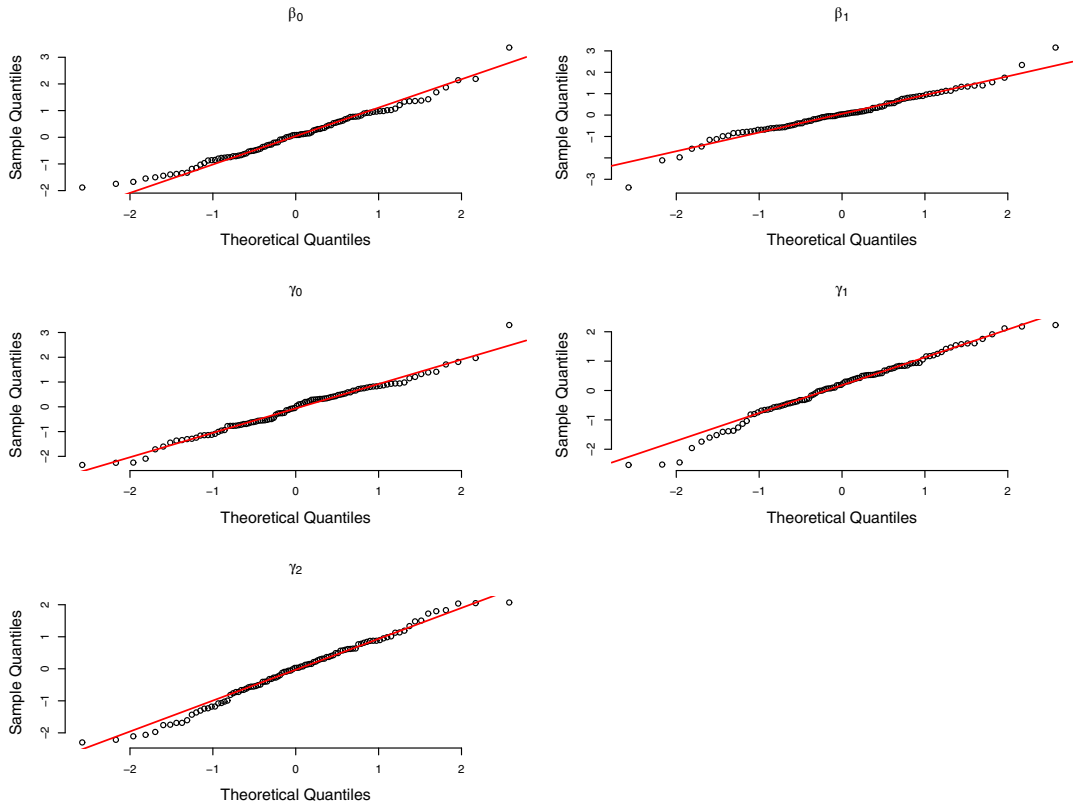

FIGURE 3 Quantile–quantile plots of $\sqrt{n}(\hat{\theta}_j - \theta_j)/\hat{\sigma}_{j,j}$, $j = 1, \ldots, 5$, for the estimated parameters of the model presented in Section 3.5 [Color figure can be viewed at wileyonlinelibrary.com]

When studying the quantile–quantile plots in Figure 3, it should be kept in mind that the number of observations per individual is small (i.e., low sampling frequency), and that the underlying truth, involving the time varying covariate $z_{i,2}(t) = z_{i,2}\sin(2\pi t)$, is quite complicated. The estimates and the QLIC do, however, appear to be behaving as expected.

Remark 2. *Sampling frequency.* In Section 3.3, consistency and limiting normality were derived under the $n \to \infty$ regime. One could, however, consider other types of asymptotic regimes, either of the infill-variety where n is held constant and the k_is tend to infinity; or where a ratio of $\max_{i \le n} k_i$ and n tends to some constant. Even in the $n \to \infty$ setup of this paper, a pertinent question is whether the sampling frequency (the size of the k_is) affects the precision of the quasi-likelihood estimates. As a first stab at this question, we estimated the asymptotic relative efficiency of the estimates of the true model estimated above, compared to the same model but with higher sampling frequency. Recall that in the former, we sampled the 1001 equidistant time points partitioning the unit interval, with probability 1/60. In a new set of simulations, we sampled them with probability 1/40. This means that the expected number of observations per child increases from about 17 to about 25. Looking at the average over 100 simulations of estimated ratios of the type Var(low freq. estimate)/Var(high freq. estimate), for the estimates of $\beta_0, \beta_1, \gamma_0, \gamma_1$, and γ_2 in the correctly specified model, we got 1.164, 1.340, 1.047, 1.007, and 1.057, respectively. These numbers indicate that the variance decreases as the sampling frequency increases, notably, the rather modest increase in sampling frequency appear to have a large effect on the precision with which

we are able to estimate the parameters entering the covariance function, that is, β_0 and β_1. These results merit further investigation.

4 | THE BRAZILIAN DATA

4.1 | The data and previous modeling strategies

The data analyzed in this section have previously been studied in Borgan et al. (2007) using a version of Aalen's linear hazard regression model. A more elaborate discussion of the data is found in that paper. For comparative purposes, we briefly present the model of Borgan et al. (2007) and fit one such to the data, thereafter, we fit three different latent Ornstein–Uhlenbeck process models. The adequacy of the Ornstein–Uhlenbeck process models compared to the linear hazard models may be evaluated using the focused information criterion introduced in Jullum and Hjort (2017), and extended to regression models in Cunen, Walløe, and Hjort (2020) and Claeskens, Cunen, and Hjort (2019). Since such comparisons must be rather elaborate and would lead us too far afield, we do not pursue such a study of different model classes here.

As part of a sanitation program 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 3 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% of the children entered late into the study and about 21% 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. (2007), this type of missingness was mainly due to data collectors 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. Such breaches of the independence assumption occur if the children could not be visited, or were not at home, because of their health condition. The data do not contain information about the reasons for which censoring occurs, and in our analysis we have assumed that censoring is independent of the underlying health processes. See Remark 5 for further discussion of this point. 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

$$\alpha_i(t) = \beta_0(t) + \beta_1(t)x_{i,1}(t) + \ldots + \beta_p(t)x_{i,p}(t), \quad i = 1, \ldots, n, \tag{14}$$

and used martingale methods to derive the limiting distribution of the estimators of the cumulative coefficients $B_k(t) = \sum_{s=1}^{t} \beta_k(s),\ k = 0, \ldots, p$. Due to the three forms of censoring they had to introduce a 'missingness process' indicating whether a child was observed, not observed, or had dropped out of the study (Borgan et al., 2007). The value of this process at a given time must

TABLE 1 Estimated cumulative regression coefficients, $\sum_{s=1}^{228} \beta_j(s)$ for the model in (14), along with standard errors (*SE*s)

	Estimates	*SE*s
Baseline	3.46	0.28
$\geq$ 3 bedroom	2.56	0.30
$\leq$ 12 months	3.49	0.29
$>$ 24 months	−3.86	0.22
Contaminated water storage	−0.99	0.25
Standing water	0.48	0.30
Contaminated water source	1.80	0.27
Other children $\leq$ 5 years	0.90	0.22
Male	0.67	0.21
Rain-affected accommodation	2.10	0.25
Mother $<$ 25 years	1.65	0.22
Open sewerage	4.16	0.38
Poor street quality	−0.58	0.24
Low social economic class	0.03	0.23

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, while the former is immaterial for our model. In our model, the reason for which a value is missing is irrelevant, as long as it is independent of the state of $\xi_i(t)$.

Table 1 displays estimated cumulative regression coefficients for a model of the form (14), along with standard errors (*SE*s). Estimators have approximately normal distributions, so Wald ratio tests may be read off from the table, pointing to those cumulative coefficients which are significantly present.

4.2 | Fitting clipped Ornstein–Uhlenbeck process models

We fitted three clipped Ornstein–Uhlenbeck process models to the Brazilian data. The results are shown in Table 2. This table contains the parameter estimates and the estimated standard errors, along with the Wald statistics; the latter are the parameter estimates divided by their approximate standard errors, that is, the statistic testing the null-hypothesis of no effect against its two-sided alternative. For illustrative purposes we also include the QLIC score of the three models, along with the M_n statistics of (12), comparing a big model to a medium model, and a medium model to a small model. As the three models in Table 2 are arrived at in a rather ad hoc fashion, the two model selection criteria should not be taken too seriously. In the big model we see that six of the estimated coefficients are not significant at the 0.05 percent level, removing these we obtain the medium model, with a somewhat superior QLIC score. The small model, that only includes the two intercepts and the log of time since the start of the study, appears to be too parsimonious as its QLIC score is inferior to the two bigger models. The M_n-statistic of (12) comparing the big and

TABLE 2 Parameter estimates and approximate *SE*s for three Ornstein–Uhlenbeck process models, along with the ratio of the two (i.e., testing the null-hypothesis of no effect against its two-sided alternative). The upper panel contains estimates of the β entering the covariance function, while the lower panel contains estimates of the γ of the threshold $c(t)$. The last line is the M_n of (12), of Big versus Medium and of Medium versus Small

		Big			Medium			Small		
		Estimate	***SE***	**Wald**	**Estimate**	***SE***	**Wald**	**Estimate**	***SE***	**Wald**
Covariance	β_0	−2.022	0.054		−2.053	0.030		−2.293	0.065	
	≥ 3 bedroom	−0.101	0.065	−1.561						
	≤ 12 months	−0.048	0.057	−0.830						
	> 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		1.245	0.060	
	Male	−0.067	0.028	−2.395	−0.065	0.028	−2.320			
	Rain-affected accommodation	−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(time)	0.162	0.003	54.792	0.162	0.003	54.971	0.161	0.011	14.957
	QLIC			−45, 969.2			−45, 958.6			−46, 434.1
	M_n (approximately 95% quantile)			20.74 (118.99)			980.58 (274.14)			

the medium model does not lead to rejection of the null-hypothesis of the medium model being true; the M_n-statistic pitting the medium against the small model does reject the hypothesis of the small model being true.

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 model of Table 2, while 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 thereby play a 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 part of the clipped Ornstein–Uhlenbeck process model, a task that, admittedly, requires a certain intuition for the phenomenon under study.

According to the QLIC, the medium model provides a better fit to the data than the big and the small models. The negative estimate of the coefficient on the binary age variable (> 24 months) indicates that children older than 2 years tend to have less oscillating health than those below 2 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, that is, by attenuating the oscillation of the underlying health processes, but it is likely that they do so 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 medium model quasi-likelihood estimate of the ratio

$$\frac{\Pr_\theta(\text{ill at } t|\text{open sewerage})}{\Pr_\theta(\text{ill at } t|\text{no open sewerage})} = \frac{1-\Phi(\gamma_0+\gamma_1+\gamma_2+\gamma_3+\gamma_4+\gamma_5\log t)}{1-\Phi(\gamma_0+\gamma_1+\gamma_2+\gamma_3+\gamma_5\log t)}.$$

This is the ratio for boys whose mothers are below 25 years old and that are living in rain-affected accommodation, in the proximity or not to open sewerage. The upward sloping curve indicates that the sanitation program had a larger effect in the areas that were not plagued by open sewerage at the start of the program. The estimate displayed in the plot was obtained by plugging in the quasi-likelihood parameter estimates, and then using the delta method to obtain a pointwise confidence band.

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 3. *Dependency/Contagion*. Since diarrhoea is a highly contagious disease, a shortcoming of the model introduced in this paper is that possible dependencies between the children is not taken into account. An extension of our model that accounts for dependency between the children is obtained by taking the zero-one processes $Y_i(t) = I\{\xi_i(t) \geq c_i(t)\}$ as above, but with $\xi_i(t)$ a spatial Gaussian process $\{\xi_i(t) : i = 1, \ldots, n, 0 \leq t \leq T\}$. In the case of the Brazilian data, such a model

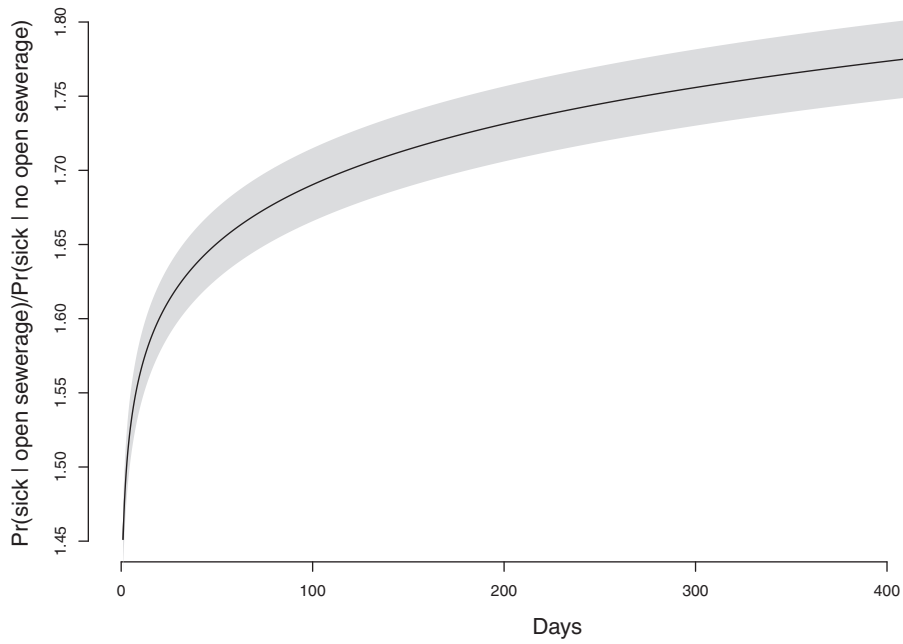

FIGURE 4 The estimated ratio of the probability of being sick when living and not living in the proximity of open sewerage, for boys whose mothers are below 25 years old living in rain affected accommodation. Estimates are based on the Medium model of Table 2. Pointwise 95% confidence intervals

demands information about the geographical proximity of the children to each other. Given that such information is available, one could take

$$\mathrm{Cov}(\xi_i(t), \xi_j(s)) = \exp(-a_1|t-s| - a_2 d(i,j)),$$

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 horizontally and vertically, so to speak, thus letting the extended quasi-likelihood consist of probabilities of the type $\Pr_\theta\{(Y_{i,j-1}, Y_{i,j}) = (u,v)\}$, for $j = 1, \ldots, k_i, i = 1, \ldots, n$, as above, but also

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

where n_j is the number of children with a jth observation. Other constructions along the lines of Hjort and Omre (1994) and Nott and Rydén (1999) could also be worked with. Other applications where this type of extension of our model is of interest include random effects type models, where the zero-one sequences are clustered in subgroups of various sizes, and associated with each individual in a group there is an unobservable threshold, $c_{i,k}$ say, centered around a group specific threshold c_k; or in genomics, where the assessment of genomic co-occurrence—which comes down to assessing the similarity of genome-wide binary vectors—is an active field of research. See the recent PhD thesis of Rand (2019), and in particular Salvatore et al. (2019), where it is argued that genome-wide binary vectors ought to be regarded as generated by correlated Gaussian processes, clipped at various thresholds.

Remark 4. *Focused inference for the quasi-likelihood*. The QLIC introduced in (11) assesses overall goodness-of-fit issues of the model. In many situations 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 et al., 2019; Claeskens & Hjort, 2003, 2008; Cunen et al., 2020; Jullum & Hjort, 2017) takes this into account and aims at selecting the optimal model in terms of mean squared error for a prespecified statistical quantity. FIC-theory can be developed for the quasi-likelihood worked with in this paper, as well as for the 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 5. *Endogenous observation times*. As noted in Section 4.1 the assumption of the observations times being independent of the underlying process $\xi_i(t)$ is in many application untenable. An interesting future research project is the development of methods of inference, for example, the quasi-likelihood approach, for clipped Ornstein–Uhlenbeck processes when the observation times are endogenous. As an example, consider the following modification of the model studied in this paper: For the ith child, the process $\xi_i(t)$ defined in Section 2.1 can be characterized as the solution to the stochastic differential equation $\mathrm{d}\xi_i(t) = -a_i\xi_i(t)\,\mathrm{d}t + (2a_i)^{1/2}\,\mathrm{d}B_i(t)$, where B_t is a standard Brownian motion. Let $B_i'(t)$ be a Brownian motion correlated with $B_i(t)$ and suppose that the observation times are generated by a point process whose intensity $\lambda_i(t)$ is the solution to $\mathrm{d}\lambda_i(t) = \mu_i(\overline{\lambda}_i - \lambda_i(t))\,\mathrm{d}t + \nu_i\lambda_i(t)^{1/2}\,\mathrm{d}B_i'(t)$, for positive parameters $\overline{\lambda}_i$, μ_i and ν_i satisfying the Feller condition.

Remark 6. *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 correspond to the level of an underlying continuous process. For example, a three-state model takes $Y_i(t) = I\{c_{i,1}(t) < \xi_i(t) \le c_{i,2}(t)\} + 2I\{c_{i,2}(t) < \xi_i(t)\}$, where $c_{i,1}(t) < c_{i,2}(t)$, and these are possibly time-varying thresholds, and $\xi_i(t)$ is an Ornstein–Uhlenbeck process of the type introduced in Section 2.1.

Remark 7. *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

$$\Pr[\{\xi_i(t) \ge c_i(t), t \in [0, t_{i,1})\}, \ldots, \{\xi_i(t) \le c_i(t), t \in [t_{i,k_i-1}, t_{i,k_i})\}],$$

with $t_{i,1}, \ldots, t_{i,k_i}$ the times of the crossings. This is a rather different object from the likelihood given in (2), and the quasi-likelihood of (3) would in this situation incur a larger efficiency loss than when the true likelihood is that of (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. Note that the model we sketch here is closely related to first hitting time models in survival analysis, see Caroni (2017) for a booklength treatise such models.

ACKNOWLEDGMENTS

We thank Robin Henderson for kindly providing us with the Brazil data. We are also grateful to the editor and two anonymous referees for constructive comments which have contributed to a stronger paper. The work of E.A.S toltenberg 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 project FocuStat (Focus Driven Statistical Inference With Complex Data, led by Hjort).

ORCID

Emil Aas Stoltenberg https://orcid.org/0000-0001-6825-4670

REFERENCES

Andersen, P. K., Borgan, Ø., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. Berlin, Germany: Springer.

Borgan, Ø., Fiaccone, R. L., Henderson, R., & 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.

Caroni, C. (2017). *First hitting time regression models: Lifetime data analysis based on underlying stochastic processes*. Hoboken, NJ: John Wiley & Sons.

Claeskens, G., Cunen, C., & Hjort, N. L. (2019). Model selection via focused information criteria for complex data in ecology and evolution. *Frontiers in Ecology and Evolution*, *7*, 415.

Claeskens, G., & Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, *98*, 900–916.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press.

Cook, R. J., & Lawless, J. (2007). *The statistical analysis of recurrent events*. New York, NY: Springer.

Cox, D. R., & Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, *91*, 729–737.

Cunen, C., Walløe, L., & Hjort, N. L. (2020). Focused model selection for linear mixed models, with an application to whale ecology. *The Annals of Applied Statistics*, *14*, 872–904.

Ferguson, T. S. (1996). *A course in large sample theory*. Boca Raton, FL: Chapman & Hall.

Hjort, N. L., & Omre, H. (1994). Topics in spatial statistics [with discussion, comments and rejoinder]. *Scandinavian Journal of Statistics*, *21*, 289–357.

Hjort, N. L., & Varin, C. (2008). ML, PL, QL in Markov chain models. *Scandinavian Journal of Statistics*, *35*, 64–82.

Jullum, M., & Hjort, N. L. (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica*, *27*, 951–981.

Nott, D. J., & Rydén, T. (1999). Pairwise likelihood methods for inference in image models. *Biometrika*, *86*, 661–676.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rand, K. D. (2019). *Towards inference with graph based genome representations* (PhD thesis). University of Oslo.

Salvatore, S., Rand, K. D., Grytten, I., Ferkingstad, E., Domanska, D., Holden, L., … Sandve, G. K. (2019). Beware the Jaccard: The choice of similarity measure is important and non-trivial in genomic colocalisation analysis. *Briefings in Bioinformatics*.

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: Cambridge University Press.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, *92*, 1–28.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*, 5–42.

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

How to cite this article: Stoltenberg EA, Hjort NL. Models and inference for on–off data via clipped Ornstein–Uhlenbeck processes. *Scand J Statist*. 2020;1–22. https://doi.org/10.1111/sjos.12472

APPENDIX A

Let

$$f_\rho(x,y) = \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left\{-\frac{(x^2-2\rho xy+y^2)}{2(1-\rho^2)}\right\}, \tag{A1}$$

be the bivariate normal density we are working with. Since $\{\xi_i(t_{i,j}), \xi_i(t_{i,j-1})\}$, as defined in Section 2.1, has the density $f(x,y)$ with $\rho = \rho(i,j) = \exp(-a_i|t_{i,j} - t_{i,j-1}|)$, standard facts about the multivariate normal distribution give that

$$p_{00}(\theta,i,j) = \int_{-\infty}^{c_i(t_{i,j})} \int_{-\infty}^{c_i(t_{i,j-1})} f_\rho(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{-\infty}^{c_i(t_{i,j})} \Phi\left(\frac{c_i(t_{i,j-1}) - \rho(i,j)y}{\{1-\rho(i,j)^2\}^{1/2}}\right)\phi(y)\,\mathrm{d}y, \tag{A2}$$

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

$$\begin{aligned} p_{01}(\theta,i,j) &= \Phi(c(t_{i,j-1})) - p_{00}(\theta,i,j),\\ p_{10}(\theta,i,j) &= \Phi(c(t_{i,j})) - p_{00}(\theta,i,j),\\ p_{11}(\theta,i,j) &= 1 - \Phi(c(t_{i,j-1})) - \Phi(c(t_{i,j})) + p_{00}(\theta,i,j). \end{aligned} \tag{A3}$$

These identities are used in some of the proofs below.

Lemma 1. *The function* $\mathrm{KL}(\theta)$ *in (7) is nonnegative, and* $\mathrm{KL}(\theta) = 0$ *if and only if* $\theta = \theta_0$.

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

$$\mathrm{KL}(\theta,\tau,x,z) = -E_{\theta_0} \sum_{j=1}^{k} \log \frac{q_j(\theta,D)}{q_j(\theta_0,D)} \geq -\sum_{j=1}^{k} \log E_{\theta_0} \frac{q_j(\theta,D)}{q_j(\theta_0,D)} = 0, \tag{A4}$$

where we have used that $E_{\theta_0}\, q_j(\theta,D)/q_j(\theta_0,D) = \sum_{(u,v)} p_{uv}(\theta,j) = 1$. This establishes that $\mathrm{KL}(\theta,\tau,x,z)$ is nonnegative, and so is $\mathrm{KL}(\theta)$. Clearly, if $\theta = \theta_0$ then $\mathrm{KL}(\theta) = 0$. Assume that $\mathrm{KL}(\theta) = 0$. The inequality in (A4) entails that $\mathrm{KL}(\theta) = 0$ implies $\int E_{\theta_0} \log\{q_j(\theta_0,D)/q_j(\theta,D)\} = 0$ for each j. That θ_0 is the unique maximizer of the limiting multinomial likelihood function $\int E_{\theta_0} \log q_j(\theta,D)\,\mathrm{d}\nu$ follows from assumption (iii), which ensures the concavity of this function. ∎

Lemma 2. *Let* $p_{\min}(\theta)$ *be the smallest of all the probabilities* $p_{uv}(\theta,i,j)$ *that enter the quasi log-likelihood function. Then, for all* i *and* j

$$\|u_j(\theta,D_i)\| \leq K \frac{\|x_i\| + 2\|z_i(t)\|_\infty}{p_{\min}(\theta)\{1-\rho(i,j)^2\}^{1/2}},$$

where the constant K *does not depend on* i,j, *or* n. *The right-hand side of this equality is finite with* ν*-probability one when evaluated in the true value.*

Proof. An expression for $u_j(\theta,D_i)$, the jth quasi-score contribution of the ith individual, is given in (5). We have that

$$\|u_j(\theta,D_i)\| \leq \|x_i\|\,|a_i\Delta_{i,j}e^{-a_i\Delta_{i,j}}A_{i,j}| + \|z_i(t_{i,j})\|\,|B_{i,j}| + \|z_i(t_{i,j-1})\|\,|C_{i,j}|.$$

Since $a_i\Delta_{i,j}e^{-a_i\Delta_{i,j}} \le e^{-1}$ and the covariates are bounded by assumption, we need bounds on $A_{i,j}, B_{i,j}$, and $C_{i,j}$ as defined in (6). From (A3) we see that the absolute values of the derivates of p_{00}, p_{01}, p_{10}, and p_{11} with respect to ρ, are all equal to $|\partial p_{00}/\partial\rho|$. We now derive a bound on $\partial p_{00}(\theta, i, j)/\partial\rho$.

$$\begin{aligned}
|\partial p_{00}(\theta,i,j)/\partial\rho| &= \left| \int_{-\infty}^{c_i(t_{i,j})} \frac{c_i(t_{i,j-1})\rho - z}{1-\rho^2} f_{\rho(i,j)}(c_i(t_{i,j-1}), z)\, \mathrm{d}z \right| \\
&\le \int_{-\infty}^{c_i(t_{i,j})} \left| \frac{c_i(t_{i,j-1})\rho - z}{1-\rho^2} \right| f_{\rho(i,j)}(c_i(t_{i,j-1}), z)\, \mathrm{d}z \\
&\le \int_{-\infty}^{\infty} \left| \frac{c_i(t_{i,j-1})\rho - z}{1-\rho^2} \right| f_{\rho(i,j)}(c_i(t_{i,j-1}), z)\, \mathrm{d}z = (2/\pi)^{1/2} \frac{\phi(c_i(t_{i,j-1}))}{(1-\rho^2)^{1/2}}. \qquad \text{(A5)}
\end{aligned}$$

To justify the interchange of differentiation and integration taking place here: Fix ρ and let $h_1, h_2, \ldots$ be a sequence decreasing to 0, with $h_1 < 1-\rho$. With ζ_{h_j} a value in $(\rho, \rho + h_j)$,

$$\Phi\left(\frac{c-(\rho+h_j)z}{(1-(\rho+h_j)^2)^{1/2}}\right) - \Phi\left(\frac{c-\rho z}{(1-\rho^2)^{1/2}}\right) = \frac{c\,\zeta_{h_j} - z}{(1-\zeta_{h_j}^2)^{3/2}} \phi\left(\frac{c-\zeta_{h_j} z}{(1-\zeta_{h_j}^2)^{1/2}}\right) h_j.$$

The right-hand side divided by h_j converges pointwise to $(c\rho - z)/(1-\rho^2) f_\rho(c,z)$ as $\zeta_{h_j} \to \rho$, and

$$\left| \frac{c\,\zeta_{h_j} - z}{(1-\zeta_{h_j}^2)^{3/2}} \phi\left(\frac{c-\zeta_{h_j} z}{(1-\zeta_{h_j}^2)^{1/2}}\right) \right| \le \frac{|c|+|z|}{(1-(\rho+h_1)^2)^{3/2}} \phi(0),$$

where the right-hand side is integrable w.r.t. $\phi(z)\,\mathrm{d}z$. The first equality in (A5) then follows from dominated convergence. This means that in a neighborhood of the true θ_0 we have that for all i and j,

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

The derivatives of $\partial p_{uv}/\partial c_i(t_{i,j})$ and $\partial p_{uv}/\partial c_i(t_{i,j-1})$ appearing in $B_{i,j}$ and $C_{i,j}$ can all be expressed in terms of the standard normal density $\phi(\cdot)$ and the functions $\psi_1(i,j)$ and $\psi_2(i,j)$ given by

$$\begin{aligned}
\psi_1(i,j) &= \Phi\left\{ \frac{c_i(t_{i,j-1}) - \rho(i,j)c_i(t_{i,j})}{(1-\rho(i,j)^2)^{1/2}} \right\} \phi\{c_i(t_{i,j})\}, \\
\psi_2(i,j) &= \frac{1}{\{1-\rho(i,j)^2\}^{1/2}} \int_{-\infty}^{c_i(t_{i,j})} \phi\left\{ \frac{c_i(t_{i,j-1}) - \rho(i,j)x}{(1-\rho(i,j)^2)^{1/2}} \right\} \phi(x)\, \mathrm{d}x, \qquad \text{(A6)}
\end{aligned}$$

where for ψ_2 the interchange of differentiation and integration can be justified by an argument similar to that above. Note that $|\psi_1(i,j)| \le \phi(0)$, and that

$$|\psi_2(i,j)| \le \frac{\Phi(c_i(t_{i,j}))\phi(0)}{(1-\rho(i,j)^2)^{1/2}} \le \frac{\phi(0)}{(1-\rho(i,j)^2)^{1/2}}. \qquad \text{(A7)}$$

By several applications of the triangle inequality, this gives

$$\|u_j(\theta, D_i)\| \le K \frac{\|x_i\| + \|z_i(t_{i,j})\| + \|z_i(t_{i,j-1})\|}{p_{\min}(\theta)\{1-\rho(i,j)^2\}^{1/2}}.$$

Since $\|z_i(t)\| \le \|z(t)\|_\infty$ the first claim follows. Moreover, by assumption (ii), $p_{\min}(\theta_0)$ is bounded away from zero, and the $\rho(i,j,\theta_0)$ are bounded away from one with probability one under ν. ∎

Lemma 3. *The matrix H of Theorem 2 is invertible.*

Proof. The Hessian of $\mathrm{KL}(\theta)$ 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 minimizer of $\mathrm{KL}(\theta)$ (as proven in Lemma 1), we have that

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

hence $v^{\mathrm{t}} H v/2 > o(\alpha^2)/\alpha^2$, which shows that $v^{\mathrm{t}} H v > 0$ for every nonzero vector v. ∎

Theorem 4. *The quasi-likelihood contributions* $\log q(\theta, D)$ *are Lipschitz in a neighborhood $B_\varepsilon(\theta_0)$ of the true value. Consequently, (9) holds.*

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

$$|\log q(\theta_1, D) - \log q(\theta_2, D)| \le |u(\tilde{\theta}, D)^{\mathrm{t}}(\theta_1 - \theta_2)| \le \|u(\tilde{\theta}, D)\| \|\theta_1 - \theta_2\|,$$

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