Published in final edited form as:
Communications in Statistics - Theory and Methods (2020), **49**(12): 2913–2930
DOI: 10.1080/03610926.2019.1584310

Composite likelihood for aggregate data from clustered multistate processes under intermittent observation

SHU JIANG

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada
E-mail: s64jiang@uwaterloo.ca

RICHARD J. COOK

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada

Summary

Markov processes offer a useful basis for modeling the progression of organisms through successive stages of their life cycle. When organisms are examined intermittently in developmental studies, likelihoods can be constructed based on the resulting panel data in terms of transition probability functions. In some settings however, organisms cannot be tracked individually due to a difficulty in identifying distinct individuals, and in such cases aggregate counts of the number of organisms in different stages of development are recorded at successive time points. We consider the setting in which such aggregate counts are available for each of a number of tanks in a developmental study. We develop methods which accommodate clustering of the transition rates within tanks using a marginal modeling approach followed by robust variance estimation, and through use of a random effects model. Composite likelihood is proposed as a basis of inference in both settings. An extension which incorporates mortality is also discussed. The proposed methods are shown to perform well in empirical studies and are applied in an illustrative example on the growth of the *Arabidopsis thaliana* plant.

Keywords: aggregate data, clustered data, heterogeneity, intermittent observation, Markov process, random effect model

1 INTRODUCTION

1.1 LITERATURE REVIEW

Multistate models are useful when studying chronic disease processes when the disease status can be classified into meaningfully different states. In individuals with hepatitis C infection for example, the extent of liver damage is quantified using a five point scale and we can define state 1 to correspond to no fibrosis, states 2 to 4 to correspond to increasing degrees of fibrosis, with state 5 representing the development of cirrhosis (Sweeting et al., 2006). In arthritis, the extent of joint damage can also be viewed in this way because joints of affected individuals may pass through a sequence of states representing increasing joint damage until total joint destruction occurs (Gladman and Chandran, 2010). In many instances of this sort the precise times that transitions are made between states are unavailable because the state occupied can only be determined upon careful imaging assessments. The resulting data is comprised of the assessment times and the states occupied at these times and is referred to as panel data. Kalbfleisch and Lawless (1985) developed a computationally convenient Fisher-scoring algorithm for obtaining maximum likelihood estimates of transition intensities for Markov processes which is implemented in the *msm* package in R by Jackson (2011).

Multistate models are also useful for modeling the maturation process of organisms since their lifecycle can typically be characterized by transitions through different stages of development (Borror and White, 1970). When organisms are large and can be tagged or otherwise identified, the resulting data has a similar format to panel data encountered in studies of disease processes. In many instances, however, organisms under very study are small and contained in one or more large tanks, and while each organism may be assessed periodically it can be difficult to identify individuals and hence track them over time. In such cases the resulting observations typically take the form of repeated aggregate count data. Such data were available in the study discussed in Munholland and Kalbfleisch (1991) which investigated the lifecycle of a grasshopper *Chorthippus parallelus*. At each assessment time the insects were classified into one of four instars (developmental stages) or as adults, but because they were not labeled the resulting data took the form of aggregate counts of the number of insects in the different maturation stages at each assessment time. In studies of plant growth, the individual organisms (plants) can be tracked of course, but the data may simply be summarized and recorded in aggregate form, resulting in the same type of repeated multinomial count data. This is the case in the motivating growth study of the plant, *Arabidopsis thaliana*, which passes through a series of developmental stages (Gouno et al., 2011).

Organisms within the same tank, or other container, will often exhibit more similar growth patterns than organisms in different tanks due to a shared environmental condition. The resulting within-tank clustering in growth patterns within tanks must be addressed when modeling the multistate processes from several tanks. This can be achieved by the formulation of multivariate models with marginal processes having desired properties, or through use of hierarchical model incorporating random effects. Zeng and Cook (2007) consider a marginal approach to dependence modeling for correlated discrete-time Markov processes using odds ratios as dependence parameters and generalized estimating functions as a basis for inference. Diao and Cook (2014) described an approach for joint analysis of two or more continuous-time Markov processes through use of copula functions which model the dependence between absorption times across correlated progressive processes; likelihood functions can be used in principle for fitting such fully specified models, but composite likelihood (Varin et al., 2011) was proposed for computational convenience. Aalen (1987) discussed the role of random effects for modeling multiple Markov chains with unexplained heterogeneity between processes. Satten (1999) developed a time-homogeneous conditionally Markov model for progressive continuous-time disease processes under panel observation schemes in which random effects accommodated heterogeneity; minor variations of this model were used in Cook et al. (2004), Sutradhar and Cook (2008) and O'Keeffe et al. (2013) to deal with clustered data of this sort. While these and other articles have

dealt with clustered multistate processes, none to our knowledge have dealt with clustered and aggregated multistate data from processes under intermittent observation. We address this here in terms of both marginally Markov models and mixed Markov models using random effects.

The remainder of this paper is organized as follows. In the next subsection, we describe a study of the growth and development of *Arabidopsis thaliana*, a small plant in the mustard family. In Section 2 we define notation and construct the likelihood for clustered aggregate multistate data under a Markov assumption. In Section 3 we review the marginal approach to dependence modeling of multistate processes via copula functions, propose a computationally feasible composite likelihood, describe how to compute robust variance estimates, and evaluate the methods by simulation. An alternative formulation given in Section 4 in which a cluster-level random effect accommodates heterogeneity in the transition intensities between clusters; simulation studies are also reported on there. Both the marginal and the random effect approach are applied to data on the growth and development of *Arabidopsis thaliana* in Section 5, where we present an extension to accommodate mortality. Concluding remarks are given in Section 6.

1.2 Data on the Growth of *Arabidopsis Thaliana*

Arabidopsis thaliana is small flowering plant in the mustard family that is considered a model organism in plant biology because of relatively short life cycle and small genome. The plants pass through seven distinct stages of development characterized by the model depicted in Figure 1. We consider a laboratory cohort study discussed by Gouno et al. (2011) in which a total of 64 tanks, each containing 50 plants, were inspected every 3 days for 3 months. The states occupied at the assessment times are recorded over the 3 month period but since the plants were not monitored continuously the exact transition times between the stages (or states) are unknown. Moreover the data recorded consists only of the total number of plants in each of the development stages at each assessment time for each tank. We use the study reported in Gouno et al. (2011) to illustrate the methods we develop which yield estimates of transition intensities and accommodate a within-tank dependence in growth rates based only on the aggregate counts. Data from a sample tank is given in Table 5 of Gouno et al. (2011).

Figure 1: A 7-state progressive model for the developmental lifecycle of the flowering mustard plant *Arabidopsis thaliana*.

2 Aggregate Data with Independent Units

2.1 Notation and Model Formulation

In order to introduce the notation and computational issues of aggregate panel data, we first consider strictly progressive independent processes with no covariates. The state space is $\{1, \ldots, K\}$ and we assume that only $k \to k+1$ transitions are allowed directly, $k = 1, \ldots, K-1$. We let $Z_j(t)$ denote the state occupied by individual j at time t and $\{Z_j(s), 0 < s\}$ denote their multistate process. If $H_j(t) = \{Z_j(s), 0 \leq s < t\}$ denotes the process history for individual j at time t, the $k \to k+1$

transition intensity is defined as

$$\lambda_k(t|H_j(t)) = \lim_{\triangle t \downarrow 0} \frac{P(Z_j(t+\triangle t^-) = k+1 | Z_j(t^-) = k, H_j(t))}{\triangle t} \ , \tag{1}$$

$k = 1, \ldots, K-1$ (Aalen et al., 2008). Under a Markov assumption the intensity in (1) does not depend on the history $H_j(t)$ other than through the state occupied at t^-, in which case we write it more compactly as $\lambda_k(t)$. Let $\Lambda(t)$ denote a $K \times K$ transition intensity matrix with $(k, k+1)$ entry $\lambda_k(t)$, diagonal entries $-\lambda_k(t)$, $k = 1, \ldots, K-1$, and zeros elsewhere. The $K \times K$ transition probability matrix $P(s,t)$, having (k,l) entry $p_{kl}(s,t) = P(Z_j(t) = l | Z_j(s) = k)$, is obtained by product as integration (Cook and Lawless, 2018) as

$$P(s,t) = \prod_{(s,t]} \{1 + \Lambda(u)du\} \ .$$

Under a time-homogeneous process (i.e. if $\lambda_k(t) = \lambda_k$, $k = 1, \ldots, K-1$) we can write $p_{kl}(s,t) = P(Z_j(t) = l | Z_j(s) = k)$ as

$$p_{kl}(s,t) = \begin{cases} \sum_{h=k}^{l} B(k,h,l) \exp(-\lambda_h(t-s)), & k \leq l \\ 0 & k > l \end{cases} \tag{2}$$

where

$$B(k,h,l) = \prod_{u=k}^{l-1} \lambda_u \Big/ \prod_{\substack{u=k \\ u \neq h}}^{l} (\lambda_u - \lambda_h), \qquad k \leq h \leq l \ , \tag{3}$$

and $B(h,h,h) = 1$ provided $\lambda_k \neq \lambda_l$ for all $k \neq l = 1, \ldots, K-1$ (Satten, 1999). The form of (2) can lead to simplifications of the likelihood which are useful in the setting we discuss in Sections 3 and 4.

We assume here and throughout that the multistate processes begin in state 1 at $t = 0$ with probability one. If assessments on individual j are carried out at times $0 = a_{j0} < a_{j1} < \cdots < a_{jR_j}$ and the state occupied is recorded at these times, we obtain panel data $\{(Z_j(a_{jr}), a_{jr}), r = 0, 1, \ldots, R_j\}$ for individual j. The likelihood for individual j can then be constructed as

$$\prod_{r=1}^{R} \prod_{k=1}^{K-1} \prod_{l \geq k}^{K} [p_{kl}(a_{j,r-1}, a_{jr})]^{N_{jkl}(a_r)}$$

where $N_{jkl}(a_r) = I(Z_j(a_{j,r-1}) = k, Z_j(a_{jr}) = l)$.

2.2 Likelihood with Aggregate Data of Independent Processes

We now consider the setting in which n organisms contained in a tank are examined at a common set of assessment times $0 = a_0 < a_1 < \cdots < a_R$. We let $N_{kl}(a_r) = \sum_{j=1}^{n} I(Z_j(a_{r-1}) = k, Z_j(a_r) = l)$ denote the number of organisms in state k at a_{r-1} and state l at a_r for $k \leq l = 1, \ldots, K$. Table 1 displays the data in matrix form for a progressive K state Markov process for assessment times a_{r-1} and a_r where the entries below the diagonal are zero.

In order to compute $N_{kl}(a_r)$ it is necessary to be able to track the organisms and link their data at a_{r-1} and a_r which requires being able to identify them individually. We consider here the case in which this is not possible because identification of individuals is infeasible, or the data available are summarized in terms of marginal frequencies for other reasons. In this case the entries in the

interior of Table 1 (i.e. $N_{kl}(a_r)$) are missing and we observe only the total number of individuals in each state at each assessment time. That is at a_{r-1} we observe the number of individuals occupying state k denoted by $M_k(a_{r-1}) = \sum_{l=k}^{K} N_{kl}(a_r)$, $k = 1, \ldots, K$, and at a_r we observe the number of individuals occupying state l denoted by $M_l(a_r) = \sum_{k=1}^{l} N_{kl}(a_r)$, $l = 1, \ldots, K$. We let $M(a_r) = (M_1(a_r), \ldots, M_K(a_r))'$ denote the vector of frequencies for the different stages of development (i.e., states) at a_r and $\bar{H}(a_r) = \{M(a_s), s = 1, \ldots, r-1\}$ the history of observed marginal frequencies at a_r, $r = 1, \ldots, R$.

Table 1: Representation of the complete data on transition counts $N_{kl}(a_r)$ over (a_{r-1}, a_r) and the corresponding marginal counts $M_k(a_{r-1})$ and $M_k(a_r)$, $k = 1, \ldots, K$ at a_{r-1} and a_r respectively.

$N_{11}(a_r)$	$N_{12}(a_r)$	$\cdots$	$N_{1K}(a_r)$	$M_1(a_{r-1})$
0	$N_{22}(a_r)$	$\cdots$	$N_{2K}(a_r)$	$M_2(a_{r-1})$
$\cdots$	$\cdots$	$\cdots$		$\cdots$
0	$\cdots$	$\cdots$		$M_{K-1}(a_{r-1})$
0	0	0	$N_{KK}(a_r)$	$M_K(a_{r-1})$
$M_1(a_r)$	$M_2(a_r)$	$\cdots$	$M_K(a_r)$	

We construct the observed data likelihood by first writing the complete data likelihood in the case where tracking is possible and making use of the Markov property that at the individual level the state occupied at a_r depends only on the state occupied at a_{r-1}. We therefore need only consider the data across consecutive assessment times, and build the joint probability of the data over all assessment times as a product of the conditional probabilities of the data at a_r given the data at a_{r-1}, $r = 1, \ldots, R$. If $N_k(a_r) = (N_{kk}(a_r), \ldots, N_{kK}(a_r))'$ denotes the potential non-zero elements in the kth row, these latent transition counts are multinomial with

$$N_k(a_r)|M_k(a_{r-1}) \sim \text{Multinom}(M_k(a_{r-1}); p_{kk}(a_{r-1}, a_r), \ldots, p_{kK}(a_{r-1}, a_r))$$

for $k = 1, \ldots, K-1$ with the probabilities given by (2) and $\sum_{l=k}^{K} p_{kl}(a_{r-1}, a_r) = 1$.

Let $N(a_r) = (N_1'(a_r), \ldots, N_K'(a_r))'$ denote the full vector of latent counts in the interior of Table 1 and $M(a_{r-1}) = (M_1(a_{r-1}), \ldots, M_K(a_{r-1}))'$ denote the vector of marginal counts at a_{r-1}. Then letting $\theta_k = \log \lambda_k$ and $\theta = (\theta_1, \ldots, \theta_{K-1})'$, the observed data likelihood can be constructed as

$$L(\theta) \propto \prod_{r=1}^{R} \sum_{N(a_r) \in \mathcal{N}_r} P(N(a_r)|\bar{H}(a_r)); \theta) = \prod_{r=1}^{R} \prod_{k=1}^{K} P(M_k(a_r)|M(a_{r-1}); \theta) \tag{4}$$

where $\mathcal{N}_r = \{N(a_r) : N_{k\cdot}(a_r) = M_k(a_{r-1}), N_{\cdot l}(a_r) = M_l(a_r), \forall\ (k, l)\}$ is the set of latent transition counts that are compatible with the margins of the table and $N_{k\cdot}(a_r) = \sum_{l=k}^{K} N_{kl}(a_r)$ and $N_{\cdot l}(a_r) = \sum_{k=1}^{l} N_{kl}(a_r)$.

3 Marginal Models for Clustered Aggregate Data

3.1 Composite Likelihood for a Marginal Model

Having discussed likelihood construction in the simple case of the previous section involving independent identical time-homogeneous Markov processes aggregated within a single tank, here we consider the setting with multiple tanks where the developmental patterns of organisms are clustered within tanks and there are tank-level covariates available. We consider the setting with a total of I tanks with

n_i individual organisms in tank i, $i = 1, \ldots, I$. We let X_i denote a $p \times 1$ tank-level covariate vector for tank i, $i = 1, \ldots, I$, and assume there are no covariates at the individual level. If $\{Z_{ij}(s), 0 < s\}$ is the multistate process and $H_{ij}(t) = \{Z_{ij}(s), 0 \leq s < t\}$ is the history for individual j in tank i we adopt a model with

$$\lim_{\Delta t \downarrow 0} \frac{P(Z_{ij}(t + \Delta t^-) = k + 1 | Z_{ij}(t^-) = k, H_{ij}(t))}{\Delta t} = \lambda_k \exp(x_i' \beta) \tag{5}$$

for $k = 1, \ldots, K - 1$. We let $\theta = (\log \lambda_1, \ldots, \log \lambda_{K-1}, \beta)'$ denote the vector of parameters characterizing the dynamics of the marginal processes.

Diao and Cook (2014) develop a model for correlated Markov processes which accommodate a dependence between processes while retaining the marginal Markov property of individual processes. For progressive processes, the dependence is accommodated by selecting a sojourn or entry time of interest and using a copula function to induce a dependence between the corresponding times for different processes within a cluster. We consider here the class of Archimedean copulas of the form

$$C(u_1, u_2, \ldots, u_{n_i}; \eta) = \mathcal{G}^{-1}(\mathcal{G}(u_1; \eta) + \cdots + \mathcal{G}(u_{n_i}; \eta), \eta) \ ,$$

where $\mathcal{G}$:$[0, 1] \to [0, \infty)$ is a continuous, strictly decreasing and convex generator function indexed by a dependence parameter η with $\mathcal{G}(1; \eta) = 0$. For the progressive process of Figure 1 we select the first transition time (i.e. the entry time to state 2) as the time on which to base the dependence modelling. Let T_{ij2} denote the entry time to state 2 for individual j in tank i and $T_{i2} = (T_{1i2}, \ldots, T_{in_i2})'$ denote the vector of all state 2 entry times within tank (cluster) i, $i = 1, \ldots, I$. We use the Clayton copula (Nelsen, 2006) in this setting, with generator $\mathcal{G}(u; \eta) = \eta^{-1}(u^{-\eta} - 1)$ to model the dependence in the state 2 entry times. A common dependence measure for copula models in the Archimedean family is Kendall's τ obtained here by

$$\tau = 1 + 4 \int_0^1 \frac{\mathcal{G}(u; \eta)}{\mathcal{G}'(u; \eta)} du \ .$$

The joint survivor function for $\mathcal{F}(t_{i2}|x_i; \theta, \eta) = P(T_{ij2} \geq t_{ij2}, j = 1, \ldots, n_i | x_i; \theta, \eta)$ is obtained by taking the probability integral transform of T_{ij2} and linking all marginal survivor functions $\mathcal{F}_{ij}(t_{ij2}|x_i) = \exp(-\lambda_1 \exp(x_i' \beta) t_{ij2})$ via the Clayton copula to obtain

$$\mathcal{F}(t_{i2}|x_i; \theta, \eta) = \left(\mathcal{F}(t_{i12}|x_i; \theta)^{-\eta} + \cdots + \mathcal{F}(t_{in_i2}|x_i; \theta)^{-\eta} - (n_i - 1)\right)^{-1/\eta} \ . \tag{6}$$

The key point is that since the individual processes are progressive, the association within tanks in the entry times to state 2 will induce a within-tank association in the entry times to subsequent states. Alternative approaches would be to model the association in the absorption times, as considered in Diao and Cook (2014), or to model dependence in the sojourn time in a particular state.

Generalizing the notation from Section 2 we let $N_{ijkl}(a_{ir}) = I(Z_{ij}(a_{i,r-1} = k, Z_{ij}(a_{ir} = l))$ denote the unobserved indicator of state k occupancy at $a_{i,r-1}$ and l at a_{ir} for individual j in tank i. We also let $M_{il}(a_{ir}) = \sum_{j=1}^{n_i} \sum_{k=1}^{l} N_{ijkl}(a_{ir})$, $l = 1, \ldots, K$ denote the marginal frequencies at a_{ir} and $M_i(a_{ir}) = (M_{i1}(a_{ir}), \ldots, M_{iK}(a_{ir}))'$ denote the vector of marginal counts for tank i, $i = 1, \ldots, I$. We propose inference based on a composite likelihood obtained under two working independence assumptions. The first is a working independence assumption between processes within tanks, which is possible to adopt because of our use of a copula-based dependence model. Under this assumption we have $\{Z_{ij}(s), 0 < s\}$ as independent of $\{Z_{ij'}(s), 0 < s\}$ for $j \neq j' = 1, \ldots, n_i$, $i = 1, \ldots, I$. As a consequence of this working independence assumption no estimate of the dependence parameter is obtained; we propose robust variance estimation to ensure valid inference. A second working independence assumption comes from considering the multinomial frequencies of the states occupied at each assessment time as being independent from the frequencies from the same tank at different

assessment times. That is we use a working independence assumption that states that $M_i(a_{ir}) \perp M_i(a_{is})$ for $s \neq r = 1, \ldots, R_i$. This is clearly violated in reality since there is a dependence induced by the Markov property and the processes are progressive so there are constraints on the possible states occupied at a_{ir} imposed by the states occupied at $a_{i,r-1}$. However this working independence assumption is possible to adopt because the probability contributions for observations at a_{ir} given the data at a_{i0} correspond to legitimate probabilities in scenarios where data from the other time points are not available. We therefore let

$$CL_{ir}(\theta) \propto P(M_i(a_{ir})|M_{i1}(a_{i0}) = n_i, x_i; \theta) \tag{7}$$

be the component likelihood for cluster i containing n_i individuals at each a_{ir}, $r = 1, \ldots, R_i$, $i = 1, \ldots, I$. This is computed in the same way the contributions to (4) are computed. The overall composite likelihood is then given by the product of the component likelihoods,

$$CL(\theta) = \prod_{i=1}^{I} \prod_{r=1}^{R_i} CL_{ir}(\theta) \; . \tag{8}$$

The estimating equations corresponding to the composite likelihood depicted in (8) are

$$S(\theta) = \sum_{i=1}^{I} S_i(\theta) \tag{9}$$

where $S_i(\theta) = \sum_{r=1}^{R_i} S_{ir}(\theta)$ and $S_{ir}(\theta) = \partial \log CL_{ir}(\theta)/\partial\theta$. Since the contributions $CL_{ir}(\theta)$ in (7) are valid likelihood contributions, $E\{S(\theta)\} = 0$ and the estimating function (9) will yield consistent estimates. Under standard regularity conditions (White, 1982), we make use of the result

$$\sqrt{I}(\widehat{\theta} - \theta) \rightarrow N(0, \mathcal{A}^{-1}(\theta)\mathcal{B}(\theta) \left[\mathcal{A}^{-1}(\theta)\right]')$$

where $\mathcal{A}(\theta) = -E\{\partial S_i(\theta)/\partial\theta'\}$ and $\mathcal{B}(\theta) = E\{S_i(\theta)S_i'(\theta)\}$ are $p \times p$ matrices which are elements of the robust sandwich variance. Note that since $S_i(\theta) = \sum_{r=1}^{R_i} S_{ir}(\theta)$ the matrix $\mathcal{B}(\theta)$ contains elements $E\{S_{ir}(\theta)S_{ir}'(\theta)\}$ which deals with the clustering in the transition times within tanks, and $E\{S_{iq}(\theta)S_{ir}'(\theta)\}$ with $q < r$ which deals with dependence between the aggregate counts from the same tank at different assessment times; thus this robust variance addresses both of the dependencies that are neglected in the two working independence assumptions. The matrices $\mathcal{A}(\theta)$ and $\mathcal{B}(\theta)$ can be estimated empirically by

$$\widehat{A}(\theta) = -I^{-1} \sum_{i=1}^{I} \left. \frac{\partial S_i(\theta)}{\partial\theta'} \right|_{\theta=\widehat{\theta}}$$

and

$$\widehat{B}(\theta) = I^{-1} \sum_{i=1}^{I} \left. S_i(\theta)S_i'(\theta) \right|_{\theta=\widehat{\theta}}.$$

and confidence intervals for elements of θ can be obtained using standard errors obtained from the square root of the corresponding diagonal entries of $\widehat{A}(\widehat{\theta})^{-1}\widehat{B}(\widehat{\theta})[\widehat{A}(\widehat{\theta})^{-1}]'$.

3.2 A Simulation Study of the Marginal Dependence Model

Here we consider a strictly progressive 5-state process with all individuals starting at state 1 with probability one. We conceptualize a specified period of interest $[0, 1]$ without loss of generality and set λ_1 such that $P(Z_{ij}(1) = 1|Z_{ij}(0) = 1) = 0.135$ so that we would expect 13.5% of the organisms to still be in state 1 at the end of the period of interest. We set $\lambda_k = 1.1^{k-1}\lambda_1$, $k = 2, 3, 4$, so that

the sojourn times in higher numbered states become shorter until the absorbing state is entered. The data are generated such that the entry times to state 2 are exponential with rate λ_1 but correlated with the other state 2 entry times within the same cluster according to a Clayton copula model given in (6) with the dependence parameter set to give Kendall's $\tau = 0.1$ or 0.2. The entry time to state 3 is the sum of the entry time of state 2 and an independent sojourn time in state 2, and because the former are associated within tanks, a within-tank dependence is induced for the entry times to state 3. In a similar fashion we simulate the sojourn time in state 3 as an independent exponential random variable with hazard λ_3 and add that to the entry time to state 3 to get the entry time to state 4, and proceed similarly for the simulation of the absorption time. The correlations induced in the state entry times for clusters with $x_i = 0$ are 0.091 and 0.177 (state 2), 0.055 and 0.102 (state 3), 0.043 and 0.084 (state 4), 0.031 and 0.065 (state 5) for $\tau = 0.1$ and 0.2 respectively. We consider settings with $n_i = 10$ or 30 individuals per cluster, $R_i = 4$ common assessment times at 0.25, 0.50, 0.75 and 1.0 for each cluster, $i = 1, \ldots, I$ with $I = 25$, 50 or 100 clusters. The design effect $1 + (n_i - 1)\text{corr}(T_{ij2}, T_{ij'2})$ represents the variance inflation due to within-cluster dependence (Donner and Klar, 2000) which are given here for the state 2 entry times by approximately 1.82 and 2.59 for clusters of size 10 and 3.64 and 6.13 for clusters of size 30 for $\tau = 0.1$ and 0.2 respectively. These represent substantial design effects and cover a range of plausible values for many applications.

The simulation results displayed in Table 2 are based on $nsim = 500$ simulated samples. We see that the empirical biases (EBIAS) are all close to zero, the average robust standard errors (ASE) agree well with the empirical standard errors (ESE), and the recent empirical coverage (ECP%) is well within the acceptable range to be compatible with the nominal level. The naive standard error based on reliance of the working independence assumptions (i.e. $\text{cov}(\sqrt{I}(\widehat{\theta} - \theta)) = \widehat{A}(\widehat{\theta})$) is considerably smaller than the robust standard error demonstrating the need for the robust variance estmate. As one would expect that the standard error decreases as the number of organisms per cluster (n_i) increases for a set value of τ, as well as when the number of clusters increases for a given cluster size and within-cluster association. We also observe the anticipated increase in variation in the estimates when the association is increased by specifying a larger value of Kendall's τ.

4 Modeling with-Cluster Dependence via Random Effects

4.1 Dependence Modeling Through Latent Variables

Dependence in growth rates within tanks can also be addressed through modeling the between-tank variation using hierarchical models. We consider the a conditionally time-homogeneous multiplicative intensity Markov model of Satten (1999) and let U_i be a scalar random effect for tank i with $E(U_i) = 1$, $\text{var}(U_i) = \phi$, and distribution function $G(U_i; \phi) = P(U_i < u_i; \phi)$. We further assume $U_i \perp\!\!\!\perp U_{i'}$ for $i \neq i' = 1, \ldots, I$. Given (x_i, u_i) the time-homogeneous transition intensities for individual j in cluster i is defined as

$$\lim_{\Delta t \downarrow 0} \frac{P(Z_{ij}(t + \Delta t^-) = k + 1 | Z_{ij}(t^-) = k, U_i = u_i, X_i = x_i)}{\Delta t} = u_i \lambda_k \exp(x_i' \beta)$$

for $k = 1, \ldots, K - 1$, $j = 1, \ldots, n_i$, and $i = 1, \ldots, I$. We again let $\theta = (\lambda_1, \ldots, \lambda_{K-1}, \beta)'$ but note that in this formulation these parameters have different interpretations than in Section 3 since they have a direct interpretation only given $U_i = u_i$ here. The marginal likelihood is obtained by integrating the joint likelihood for the aggregate data and the random effect with respect to the latent random effect. Maximum likelihood estimates are obtained by maximizing with respect to $\psi = (\theta', \phi)'$. We describe how this can be carried out first for the case where individual-level data are available and then consider the case with aggregation.

Table 2: Empirical performance of maximum composite likelihood estimators from (8) over $nsim = 500$ simulations based on the copula-based dependence model of Section 3.1.

		Kendall's $\tau = 0.1$										Kendall's $\tau = 0.2$									
		$n_i = n = 10$					$n_i = n = 30$					$n_i = n = 10$					$n_i = n = 30$				
I		BIAS	ESE	ASE	ASE[1]	ECP%	BIAS	ESE	ASE	ASE[1]	ECP%	BIAS	ESE	ASE	ASE[1]	ECP%	BIAS	ESE	ASE	ASE[1]	ECP%
25	$\log\lambda_1$	<0.001	0.095	0.090	0.042	93.6	0.002	0.071	0.073	0.024	95.4	<0.001	0.114	0.110	0.042	93.8	-0.006	0.100	0.097	0.024	94.0
	$\log\lambda_2$	-0.005	0.095	0.094	0.063	93.8	0.002	0.057	0.054	0.037	93.4	-0.002	0.099	0.097	0.063	93.6	<0.001	0.065	0.069	0.037	96.6
	$\log\lambda_3$	<0.001	0.095	0.090	0.042	93.6	-0.005	0.069	0.067	0.052	93.0	-0.010	0.121	0.118	0.091	93.0	-0.010	0.072	0.070	0.052	93.4
	$\log\lambda_4$	-0.011	0.116	0.116	0.091	93.9	-0.005	0.069	0.067	0.052	93.0	-0.010	0.121	0.118	0.091	93.0	-0.010	0.072	0.070	0.052	93.4
50	$\log\lambda_1$	-0.005	0.070	0.065	0.030	93.0	0.002	0.051	0.052	0.017	95.8	-0.006	0.079	0.078	0.030	94.8	-0.003	0.075	0.073	0.024	93.8
	$\log\lambda_2$	-0.003	0.064	0.064	0.045	94.8	<0.001	0.041	0.039	0.026	93.6	-0.004	0.065	0.065	0.045	95.2	<0.001	0.044	0.043	0.036	94.2
	$\log\lambda_3$	-0.003	0.088	0.083	0.065	93.0	-0.003	0.048	0.048	0.037	95.4	-0.011	0.084	0.084	0.065	94.6	-0.004	0.051	0.051	0.037	95.6
	$\log\lambda_4$	0.004	0.123	0.118	0.096	94.2	0.003	0.065	0.067	0.055	96.4	-0.001	0.116	0.115	0.096	94.2	-0.002	0.069	0.067	0.055	94.4
100	$\log\lambda_1$	-0.002	0.048	0.046	0.021	93.0	-0.001	0.037	0.037	0.012	94.0	-0.006	0.057	0.055	0.021	93.6	-0.001	0.051	0.049	0.012	94.4
	$\log\lambda_2$	-0.002	0.044	0.045	0.045	95.2	<0.001	0.028	0.028	0.018	94.4	-0.002	0.047	0.047	0.031	95.2	-0.001	0.031	0.030	0.018	93.8
	$\log\lambda_3$	<0.001	0.059	0.059	0.045	95.0	-0.002	0.034	0.035	0.026	96.0	<0.001	0.059	0.059	0.045	94.6	-0.003	0.035	0.036	0.026	95.2
	$\log\lambda_4$	-0.002	0.082	0.081	0.067	94.4	0.001	0.047	0.048	0.039	95.2	-0.005	0.084	0.082	0.067	93.0	-0.001	0.046	0.048	0.039	95.2

[1] Naive ASE based on working independence assumptions

For our strictly progressive process we can make use of (2) to accommodate the cluster-level random effect and covariate and write

$$P(Z_{ij}(a_{ir}) = s_{ijr}|Z_{ij}(a_{i,r-1}) = s_{ij,r-1}; u_i, x_i) = \sum_{h=s_{ij,r-1}}^{s_{ijr}} B(s_{ij,r-1}, h, s_{ijr}) \exp(-u_i \lambda_h e^{x_i'\beta} \triangle a_{ir})$$

where $\triangle a_{ir} = a_{ir} - a_{i,r-1}$ is the lag between the $(r-1)$st and rth assessment times for tank i and s_{ijr} represents the state occupied by individual j in tank i at time a_{ir}; the function $B(\cdot, \cdot, \cdot)$ is defined as in (3).

With individual level panel data, the conditional likelihood contribution for cluster i given (u_i, x_i) can be written as

$$\mathcal{L}_i^c(\theta) \propto \prod_{j=1}^{n_i} \prod_{r=1}^{R_i} P(Z_{ij}(a_{ijr}) = s_{ijr}|Z_{ij}(a_{ij,r-1}) = s_{ij,r-1}; u_i, x_i) \ . \tag{10}$$

Based on the form of (2) we can rewrite (10) as

$$\mathcal{L}_i^c(\theta) \propto \prod_{j=1}^{n_i} \left[\sum_{h_0=s_{ij0}}^{s_{ij1}} \sum_{h_1=s_{ij1}}^{s_{ij2}} \cdots \sum_{h_{R_i-1}=s_{ij,R_i-1}}^{s_{ijR_i}} \left\{ \prod_{r=1}^{R_i} B(s_{ij,r-1}, h_{r-1}, s_{ijr}) \exp(-u_i \lambda_{h_{r-1}} e^{x_i'\beta} \Delta a_{ir}) \right\} \right] \ .$$

We can then get the marginal probabilities for cluster i by averaging over the random effect as

$$\mathcal{L}_i(\psi) \propto \int_0^\infty \mathcal{L}_i^c(\theta) dG(u_i; \phi) \ .$$

A closed-form of the marginal likelihood is obtainable if there exists a Laplace transform $v_\phi(\cdot)$ for the random effect distribution, but not otherwise.

When data are aggregated at the cluster level we must again marginalize over the complete tables as in (7). This summation is infeasible here because even for progressive models the number of possible realizations of individual paths increases at a prohibitive rate with the number of assessment times and the cluster size. We therefore consider an alternative approach, again based on a composite likelihood. Specifically we consider a contribution for the counts of transitions over $(0, a_{ir}]$ given u_i in cluster i, as in Section 3.

4.2 Composite Likelihood Construction

Here we consider composite likelihood contributions from cluster i based on data at times $a_{i0} = 0$ and a_{ir} for $r = 1, \ldots, R_i$. For a particular time a_{ir} in cluster i we write

$$\mathcal{L}_{ir}^c(\theta) \propto \prod_{j=1}^{n_i} P(Z_{ij}(a_{ir}) = s_{ijr}|Z_{ij}(a_{i0}) = s_{ij0}; u_i, x_i) \tag{11}$$

where

$$P(Z_{ij}(a_{ir}) = s_{ijr}|Z_{ij}(a_{i0}) = s_{ij0}; u_i, x_i) = \sum_{h=s_{ij0}}^{s_{ijr}} B(s_{ij0}, h, s_{ir}) \exp(-u_i \lambda_h e^{x_i'\beta} \Delta a_{ir}) \ . \tag{12}$$

Taking the expectation of (11) with respect to the random effect and taking the product of all such terms for $r = 1, \ldots, R_i$ gives a joint probability and composite likelihood for the panel data setting which can be written as

$$\mathcal{L}_i(\theta) \propto \prod_{r=1}^{R_i} \mathcal{L}_{ir}^c(\theta)$$

where

$$\mathcal{L}_{ir}(\psi) = \int_0^\infty \mathcal{L}^c_{ir}(\theta) dG(u_i; \phi) \ . \tag{13}$$

If $v_\phi(\cdot)$ is the Laplace transform of the random effect distribution, a closed-form for the integral is obtained by replacing each exponential factor by the Laplace transform to obtain

$$\mathcal{L}_{ir}(\psi) \propto \sum_{h_1=s_{i10}}^{s_{i1r}} \sum_{h_2=s_{i20}}^{s_{i2r}} \cdots \sum_{h_{n_i}=s_{in_i0}}^{s_{in_ir}} \prod_{j=1}^{n_i} B(s_{ij0}, h_i, s_{ijr}) \cdot v_\phi \left(\textstyle\sum_{j=1}^{n_i} \lambda_{h_i} \exp(x_i'\beta) \Delta a_{ir} \right) \ . \tag{14}$$

When data are aggregated and only marginal totals are available at each assessment time, an observed data composite likelihood is obtained by summing (14) over all possible matrices of transition counts between a_{i0} and a_{ir} to give a contribution

$$L_i(\psi) \propto \prod_{r=1}^{R_i} \sum_{N(a_{ir}) \in \mathcal{N}_{ir}} \mathcal{L}_{ir}(\psi) \tag{15}$$

for each cluster i where $N_{i1l}(a_{ir}) = \sum_{j=1}^{n_i} I(Z_{ij}(a_{ir}) = l | Z_{ij}(a_{i0}) = 1)$ and

$$\mathcal{N}_{ir} = \{ N_i(a_{ir}) : N_{i1\cdot}(a_{ir}) = n_i, N_{i\cdot l}(a_{ir}) = M_{il}(a_{ir}), \forall \ l \} \ .$$

The overall composite likelihood is obtained by multiplying contributions of the form (15) over all I clusters to obtain $L(\psi) = \prod_{i=1}^I L_i(\psi)$ and robust variance estimates can be derived as in Section 3.2 based on the elementary estimating functions given by $S_i(\psi) = \partial \log L_i(\psi) / \partial \psi$.

4.3 A Simulation Study for the Random Effect Model

Here we consider a strictly progressive process with all individuals starting at state 1. We set λ_1 such that $P(Z_{ij}(1) = 1 | Z_{ij}(0) = 1) = 0.135$ and set the other baseline intensities as $\lambda_k = 1.1^{k-1}\lambda_1$, $k = 2, 3, 4$ to reflect more rapid progression through the latter stages as before. The random effect U_i is gamma distributed with $E(U_i) = 1$ and $\text{var}(U_i) = \phi$ and we consider $\phi = 0.4$ and 0.8 to represent modest and larger degrees of between cluster variation. To assess the effect of cluster size, as in Section 3.2 we let $n_j = 10$ or 30 and consider $I = 25$, 50 and 100 clusters. We again consider $R_i = 4$ with common follow-up assessments at $a_{ir} = 0.25r$, $r = 1, \ldots, 4$. The results are displayed in Table 3 for $nsim = 500$ simulations. We see that the bias is generally small for the log intensities but can be more substantial for $\log \phi$ for the setting with $I = 25$; this may be a finite sample bias as it is much smaller when $I = 50$ and 100. The ESE and ASE are in alignment and the empirical coverage probabilities are well within the nominal level. As the number of clusters I increases and the number per cluster $n_i = n$ increases, we see the anticipated decrease in the standard errors.

5 A Developmental Study of *Arabidopsis thaliana*

Gouno et al. (2011) provided data from one tank of the 64 tanks of *Arabidopsis thaliana* in a laboratory based study. The plants are sorted into different clusters according to their origin, with 50 plants in each cluster. The data are recorded every 3 days for 3 months giving a total of 32 assessments per cluster.

So far we have restricted attention to progressive processes depicted in Figure 1 but a key feature of this dataset is that the plants may die at any time from the different stages. To accommodate the fact that individuals may die during the process between any assessment times as depicted in Figure 2 we introduce a transition intensity from state k to the dead state D denoted λ_{kd}; we constrain $\lambda_{kd} = \lambda_d$,

Table 3: Empirical performance of maximum composite likelihood estimators for $nsim = 500$ simulations under a two-way composite likelihood with random effects as described in Section 4.2.

		$\phi = 0.4$								$\phi = 0.8$							
		$n_i = n = 10$				$n_i = n = 30$				$n_i = n = 10$				$n_i = n = 30$			
I		EBIAS	ESE	ASE	ECP%	EBIAS	ESE	ASE	ECP%	EBIAS	ESE	ASE	ECP%	EBIAS	ESE	ASE	ECP%
25	$\log \lambda_1$	0.004	0.167	0.166	93.3	-0.008	0.145	0.141	93.4	-0.014	0.215	0.210	92.7	-0.026	0.216	0.210	93.4
	$\log \lambda_2$	0.019	0.160	0.167	93.5	-0.007	0.149	0.146	93.7	0.012	0.231	0.218	93.5	-0.008	0.207	0.207	93.6
	$\log \lambda_3$	0.021	0.187	0.189	93.9	0.005	0.154	0.157	94.2	0.020	0.245	0.237	93.2	0.002	0.206	0.228	94.0
	$\log \lambda_4$	0.035	0.227	0.226	93.6	0.004	0.163	0.164	94.6	0.027	0.257	0.268	94.1	-0.009	0.227	0.230	95.2
	$\log \phi$	0.225	0.636	0.639	96.3	0.110	0.428	0.420	95.4	0.139	0.483	0.452	95.4	0.130	0.376	0.384	94.4
50	$\log \lambda_1$	-0.009	0.157	0.156	93.6	-0.008	0.145	0.141	93.4	-0.009	0.157	0.156	93.6	-0.013	0.147	0.143	93.4
	$\log \lambda_2$	-0.001	0.112	0.112	95.8	-0.003	0.103	0.101	93.2	<0.001	0.163	0.159	94.6	-0.008	0.153	0.152	93.0
	$\log \lambda_3$	0.009	0.131	0.135	95.0	-0.003	0.115	0.109	93.2	0.005	0.169	0.172	93.6	0.002	0.148	0.145	94.2
	$\log \lambda_4$	0.022	0.152	0.159	94.2	0.007	0.116	0.116	94.4	0.008	0.173	0.189	94.8	0.002	0.154	0.149	93.6
	$\log \phi$	0.095	0.398	0.408	96.6	0.046	0.283	0.275	94.2	0.081	0.289	0.305	96.2	0.059	0.237	0.242	95.0
100	$\log \lambda_1$	<0.001	0.086	0.086	94.2	<0.001	0.708	0.710	94.1	-0.009	0.105	0.127	94.8	-0.004	0.092	0.095	95.6
	$\log \lambda_2$	0.003	0.084	0.084	95.8	0.002	0.073	0.072	93.3	-0.003	0.114	0.116	95.4	0.002	0.095	0.101	96.8
	$\log \lambda_3$	-0.003	0.096	0.096	94.4	0.002	0.074	0.077	94.9	-0.003	0.114	0.128	95.6	0.005	0.095	0.103	95.6
	$\log \lambda_4$	<0.001	0.116	0.117	93.4	<0.001	0.080	0.084	94.5	0.007	0.132	0.140	94.4	0.005	0.097	0.105	96.4
	$\log \phi$	0.042	0.277	0.274	94.6	0.027	0.194	0.191	94.1	0.038	0.196	0.211	96.4	0.021	0.171	0.168	94.0

$k = 1, \ldots, 7$ so that the time of of death is treated as independent of the stage of development. The likelihood for the marginal and random effect approaches can be constructed in a similar way as in Sections 3.1 and 4.2 with death intensities taken into account. Specifically, the composite likelihood for the marginal approach is modified to

$$CL(\theta^\dagger) = \prod_{i=1}^{I} \prod_{r=1}^{R_i} P(M_i(a_{ir}) | M_{i1}(a_{i0}) = n_i, x_i; \theta^\dagger) \tag{16}$$

where $\theta^\dagger = (\theta', \theta_d)'$ with $\theta_d = \log \lambda_d$. For the random effect model we define the likelihood as

$$L(\psi^\dagger) = \prod_{i=1}^{I} \prod_{r=1}^{R_i} \sum_{N(a_{ir}) \in \mathcal{N}_{ir}} \int_0^\infty \mathcal{L}_{ir}^c(\theta^\dagger) dG(u_i; \phi) \tag{17}$$

where $\psi^\dagger = (\theta', \theta_d, \phi)'$ and the probabilities in $\mathcal{L}_{ir}^c(\theta^\dagger)$ are computed based on the 8-state model in Figure 2.

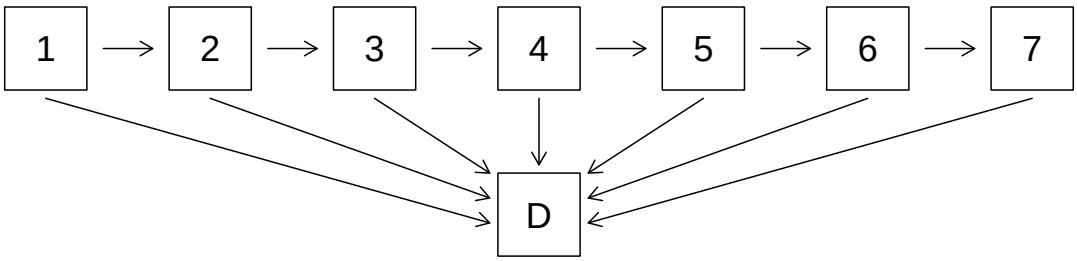

Figure 2: An 8-state progressive model for the developmental lifecycle of the flowering mustard plant *Arabidopsis thaliana* incorporating mortality.

Table 4: Estimates from fitting the marginal model under the composite likelihood for transitions between the developmental stages of *Arabidopsis thaliana* based on the data from a single tank reported in Gouno et al. (2011).

	EST	SE	95% CI
$\log \lambda_1$	-3.53	0.10	(-3.73, -3.34)
$\log \lambda_2$	-1.91	0.20	(-2.30, -1.52)
$\log \lambda_3$	-3.75	0.12	(-3.99, -3.51)
$\log \lambda_4$	-2.08	0.12	(-2.33, -1.84)
$\log \lambda_5$	-1.04	0.22	(-1.47, -0.61)
$\log \lambda_6$	-3.26	0.24	(-3.73, -2.80)
$\log \lambda_d$	-6.38	0.16	(-6.70, -6.07)

Table 4 shows the estimates of the transition intensity of the one tank in Gouno et al. (2011). Time points 12 and 13 in Gouno et al. (2011) are not consistent with the count of the time points

thereafter, hence we drop these points assuming there is a recording error. To assess the goodness of fit of our proposed model in Table 4, we compare our parametric estimates of the state entry time distributions with nonparametric estimates obtained by the pooled-adjacent violators algorithm (Ayer et al., 1955) under a working independence assumption. Let $\mathcal{S}_i = \{a_{ir}, r = 0, 1, \ldots, n_i\}$ denote the set of assessment times from tank i and $\mathcal{S} = \bigcup_{i=1}^{I} \mathcal{S}_i$ denote the set of all assessment times with elements $\mathcal{S} = \{u_r, r = 0, 1, \ldots, R\}$ where $0 = u_0 < u_1 < \cdots < u_R$ are the ordered times. This nonparametric estimate was obtained by constructing a dataset of "pseudo-individuals" where we let m_r denote the total number of individuals who were assessed at time u_r across all tanks, y_r denote the number of those who had experienced the event of interest by u_r, and $m_r - y_r$ as the number who had not. An isotonic regression of $(y_1/m_1, \ldots, y_R/m_R)$ with weights $(m_1, \ldots, m_R)$ gives the nonparametric estimate

$$\widehat{F}(u_r) = \max_{u \le r} \min_{v \ge r} \left(\frac{\sum_{h=u}^{v} y_l}{\sum_{h=u}^{v} m_r} \right) . \tag{18}$$

which is an estimate of the cumulative distribution function for the event of interest. Due to the competing risk of death here we define the events as entry to developmental state k, or death, and let $T_k^{\dagger}$ denote the corresponding time. Figure 3 gives the nonparametric estimates of the corresponding cumulative distribution function for $T_k^{\dagger}$ along with the corresponding estimate from the parametric fit for $k = 6$ and 7. We see good agreement between the fitted values from the proposed model and the nonparametric estimates.

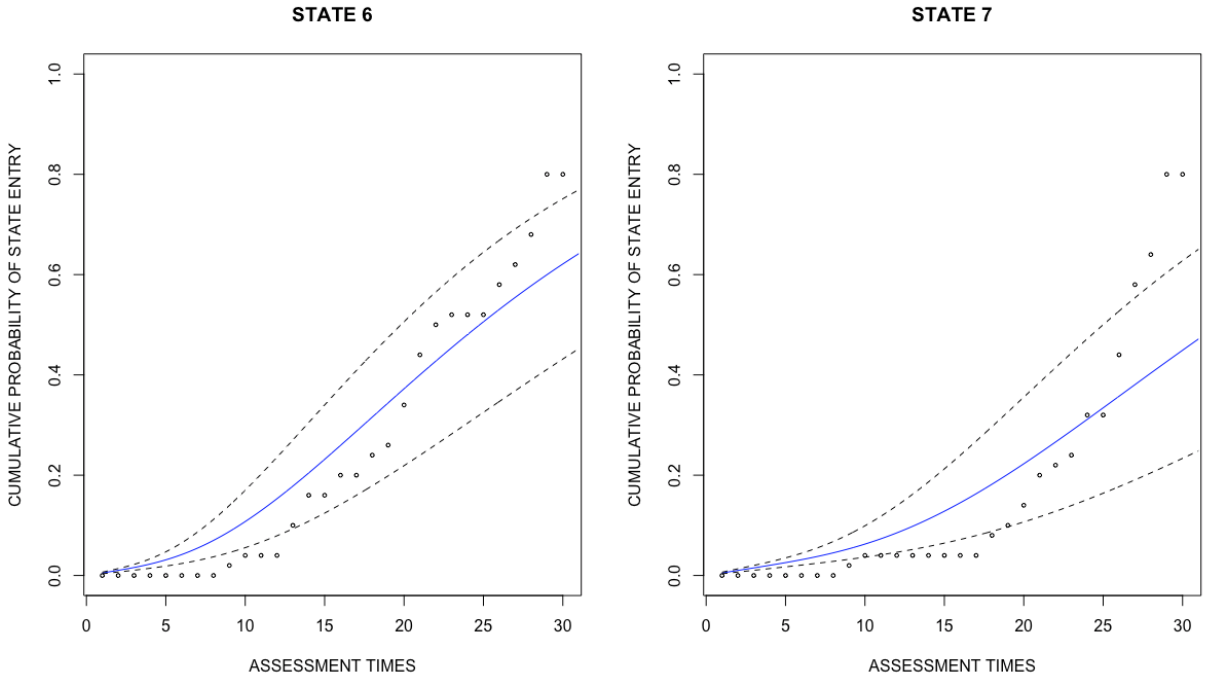

Figure 3: Nonparametric estimates of state entry time distributions obtained by the pooled adjacent violators algorithm plotted with parametric estimates (and 95% confidence limits represented by the dashed lines) obtained from the marginal model via composite likelihood; $F_k(t) = P(T_k^{\dagger} \le t)$, $k = 6, 7$.

We were not able to obtain the data for the 64 clusters of plants mentioned in Gouno et al. (2011) so we used the estimates obtained for the one tank for which the data was provided and specified the inestimable parameters τ for the marginal model and ϕ for the random effect model. We then simulated clustered data based on the estimates in Table 4 with $\tau = 0.2$ for the marginal model, and

$\phi = 0.8$ for the random effect model for the purpose of demonstrating our methods. We simulate data from 100 clusters with 25 organisms per cluster and assumed that the aggregate counts were recorded every 12 days. Estimates from fitting both the robust marginal method and the random effect model by composite likelihood are presented in Table 5.

Table 5: Estimates from fitting the marginal model with robust variance estimation and the random effect model via two-way composite likelihood for the 100 clusters of plants simulated with 25 plants in each cluster.

	Marginal Model			Random Effect Model		
	EST	SE	95% CI	EST	SE	95% CI
$\log \lambda_1$	-3.56	0.05	(-3.65, -3.46)	-3.54	0.12	(-3.78, -3.31)
$\log \lambda_2$	-1.91	0.04	(-1.99, -1.84)	-1.78	0.14	(-2.06, -1.50)
$\log \lambda_3$	-3.75	0.03	(-3.81, -3.69)	-3.73	0.11	(-3.94, -3.51)
$\log \lambda_4$	-2.07	0.04	(-2.15, -1.98)	-1.77	0.13	(-2.02, -1.53)
$\log \lambda_5$	-1.09	0.06	(-1.20, -0.98)	-1.39	0.14	(-1.65, -1.13)
$\log \lambda_6$	-3.27	0.04	(-3.35, -3.19)	-2.94	0.11	(-3.14, -2.73)
$\log \lambda_d$	-6.37	0.06	(-6.50, -6.24)	-5.14	0.16	(-5.45, -4.84)
$\log \phi$	-	-	-	-0.76	0.43	(-1.60, 0.09)

6 DISCUSSION

We have described a composite likelihood-based method for the analysis of clustered aggregate developmental data. The computational feasibility of this approach hinges on the progressive nature of the process which is characteristic of most growth cycles, and the fact that all organisms were observed from the start of the first stage. Use of composite likelihood greatly reduces the size of the sample space that must be marginalized over to compute the probabilities based on the marginal frequencies.

Marginal models and random effect models were used to accommodate clustering of rates within tanks. Estimation of the parameters under the marginal formulation did not involve estimation of the dependence parameter of the copula as this was more of a nuisance parameter in the present setting. In principle, however, one could consider relaxing the working independence assumption within tanks to estimate this parameter as well. We have also restricted attention to time homogeneous transition intensities but this can be relaxed easily to accommodate piecewise constant functions. The plots of the state entry time distributions based on the available data exhibited good agreement with the nonparametric estimates using the pooled adjacent violators algorithm and so models with exponential sojourn times appear reasonable for the data at hand.

In some settings it may be feasible to tag organisms to enable tracking of individuals, but this may incur a cost. If it is possible, it may be of interest to consider the cost-benefit of tracking individual organisms (Jiang and Cook, 2018). In a completely different setting this issue arises in school-based studies of health knowledge, attitudes and behaviour among youth. Here tracking of individuals may require greater effort to get ethics approvals in comparison to repeat cross-sectional studies, which offer data more like the aggregate data in our setting. However school-based studies also feature immigration and emigration which mean any models based on marginal aggregate summaries must accommodate the fact that some new students may have entered the school and some may have left; such data may be available from school administrators.

FUNDING

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN 155849 and RGPIN 04207) and the Canadian Institutes for Health Research (FRN 13887). Richard Cook is a Tier I Canada Research Chair in Statistical Methods for Health Research.

REFERENCES

Aalen, O. O. (1987). Mixing distributions on a Markov chain. *Scandinavian Journal of Statistics*, 14:281–289.

Aalen, O. O., Borgan, O., and Gjessing, H. K. (2008). *Survival and Event History Analysis: A Point Process Point of View*. Springer, New York.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annual of Mathematical and Statistics*, 26:641 – 647.

Borror, D. L. and White, R. E. (1970). *A Field Guide to the Insects of Amnerican North of Mexico*. Boston: Houghton Mifflin.

Cook, R. J. and Lawless, J. F. (2018). *Multistate Models for the Analysis of Life History Data*. Springer: New York.

Cook, R. J., Yi, G. Y., and Lee, K. A. (2004). A conditional Markov model for clustered progressive multistate processes under incomplete observation. *Biometrics*, 60:436–443.

Diao, L. and Cook, R. J. (2014). Composite likelihood for joint analysis of multiple multistate processes via copulas. *Biostatistics*, 15:690 – 705.

Donner, A. and Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. Arnold Publishers, London, U.K.

Gladman, D. D. and Chandran, V. (2010). Observational cohort studies: lessons learnt from the University of Toronto psoriatic arthritis program. *Rheumatology*, 50:25 – 31.

Gouno, E., Courtrai, L., and Fredette, M. (2011). Estimation from aggregate data. *Computational Statistics and Data Analysis*, 55:615 – 626.

Jackson, C. H. (2011). Multi-state models for panel data: The *msm* package for R. *Journal of Statistical Software*, 38:1–29.

Jiang, S. and Cook, R. J. (2018). Cost-effective design of growth studies with aggregation and tracking. *Journal of Biometrics and Biostatistics*, 9:406–411.

Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80:863–871.

Munholland, P. L. and Kalbfleisch, J. D. (1991). A semi-Markov model for insect life history data. *Biometrics*, 47:1117 – 1126.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, New York.

O'Keeffe, A. G., Tom, B. D., and Farewell, V. T. (2013). Mixture distributions in multi-state modelling: some considerations in a study of psoriatic arthritis. *Statistics in Medicine*, 32(4):600–619.

Satten, G. A. (1999). Estimating the extent of tracking in interval-censored chain-of-events data. *Biometrics*, 55:1228 – 1231.

Sutradhar, R. and Cook, R. J. (2008). Analysis of interval-censored data from clustered multistate processes:application to joint damage in psoriatic arthritis. *Journal of Royal Statistical Society*, 57:553 – 566.

Sweeting, M. J., Angelis, D. D., Neal, K. R., Ramsay, M. E., Irving, W. L., Wright, M., Brant, L., and Harris, H. E. (2006). Estimated progression rates in three united kingdom hepatitis c cohorts differed according to method of recruitment. *Journal of Clinical Epidemiology*, 59:144 – 152.

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

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

Zeng, L. and Cook, R. J. (2007). Transition models for multivariate longitudinal binary data. *Journal of the American Statistical Association*, 102(477):211–223.