Published in final edited form as:
Biostatistics (2016), **17**(3): 437–452
DOI: 10.1093/biostatistics/kxv054

Augmented composite likelihood for copula modeling in family studies under biased sampling

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

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

Summary

The heritability of chronic diseases can be effectively studied by examining the nature and extent of within-family associations in disease onset times. Families are typically accrued through a biased sampling scheme in which affected individuals are identified and sampled along with their relatives who may provide right-censored or current status data on their disease onset times. We develop likelihood and composite likelihood methods for modeling the within-family association in these times through copula models in which dependencies are characterized by Kendall's τ. Auxiliary data from independent individuals are exploited by augmentating composite likelihoods to increase precision of marginal parameter estimates and consequently increase efficiency in dependence parameter estimation. An application to a motivating family study in psoriatic arthritis illustrates the method and provides some evidence of excessive paternal transmission of risk.

Keywords: Auxiliary data; Biased sampling; Composite likelihood; Family study; Gaussian copula.

1 Introduction

The hereditary nature of diseases can be inferred by the structure and extent of within-family dependencies in some feature of the disease process. Family studies employing biased sampling schemes are often advocated as a cost-effective approach to estimate these dependencies and provide a framework for exploring the effects of genetic attributes (Laird and Lange, 2006). In such studies families are typically recruited by selecting an affected individual in a disease registry called the *proband*, and subsequently recruiting their consenting family members for examination (Burton, 2003). The proband often provides more detailed disease history than *non-probands*; it may only be known, for example, whether the non-probands have the condition at their age of assessment.

If there is considerable variation in the age of onset and the age of assessment, analyses based simply on the known disease status of individuals is problematic. Specification of multivariate models for the *time of disease onset* enables one to reflect the time varying nature of the disease status. Mixed-effect models have been studied in this context (Li and Thompson, 1997, Hsu *and others*, 2004) but they do not yield appealing dependence measures in non-linear settings. Copula functions (Nelsen, 2006) yield dependence measures which are functionally independent of the parameters in the marginal onset time distribution and therefore offer a more appealing framework.

We develop marginal models for the disease onset time distribution and use a Gaussian copula to model the role of kinship in the strength of within-family associations (Liang and Beaty, 1991). Covariate effects can be studied in marginal and second-order regression models in the spirit of Prentice and Zhao (1991). Likelihood and composite likelihood (Cox and Reid, 2004) are examined where the latter can offer important simplifications and reduce computational burden when dealing with large families.

The remainder of this paper is organized as follows. In Section 2, we define notation and formulate the joint model for event times of family members. Likelihood and composite likelihood methods for response-biased data are given in Section 3 where asymptotic and empirical studies investigate the relative efficiency of the proposed methods. Extensions are discussed in Section 4 which accommodate a combination of right-censored and current status observation schemes for non-probands. Approaches for making use of auxiliary data on the marginal onset time distribution are also developed and assessed empirically. An application to the motivating family study on the genetic basis for psoriatic arthritis (PsA) is given in Section 5 where important insights are made on excessive paternal transmission of risk. Concluding remarks are given in Section 6.

2 SECOND-ORDER DEPENDENCE MODELS FOR DISEASE ONSET TIMES IN FAMILY STUDIES

Let T_{ij} denote the time of disease onset for individual j in family i comprised of m_i individuals, and Z_{ij} denote the covariate vector, $j = 1, \dots, m_i$; we let $T_i = (T_{i1}, \dots, T_{im_i})'$ and $Z_i = (Z'_{i1}, \dots, Z'_{im_i})'$, $i = 1, \dots, n$. We assume $T_1, \dots, T_n$ are mutually independent given $Z_1, \dots, Z_n$ and $T_i \perp Z^{(-i)}|Z_i$, where $Z^{(-i)} = \{Z_{i^*} : 1 \le i^* \le n, i^* \neq i\}$. The marginal survivor function is $\mathcal{F}(t|Z_{ij};\theta) = P(T_{ij} > t|Z_{ij})$ and we let $F(t|Z_{ij};\theta) = 1 - \mathcal{F}(t|Z_{ij};\theta)$, where θ is a $p \times 1$ parameter vector. A joint model for the event times in family i can be constructed by specifying an m_i dimensional copula function (Joe, 1997), a multivariate cumulative distribution function with uniform $[0, 1]$ margins. If $U_{ij} \sim \text{unif}(0, 1)$ and $U_i = (U_{i1}, \dots, U_{im_i})'$, the joint cumulative distribution function $\mathcal{C}(u_{i1}, \dots, u_{im_i}; \gamma) = P(U_{i1} \le u_{i1}, \dots, U_{im_i} \le u_{im_i}; \gamma)$ defines a copula indexed by a $q \times 1$ parameter vector γ. We construct the survivor function for $T_i|Z_i$ by setting $U_{ij} = \mathcal{F}(T_{ij}|Z_{ij};\theta)$ and defining it through the survival copula specification (Joe, 1997),

$$P(T_{i1} > t_{i1}, \dots, T_{im_i} > t_{im_i}|Z_i;\psi) = \mathcal{C}(\mathcal{F}(t_{i1}|Z_{i1};\theta), \dots, \mathcal{F}(t_{im_i}|Z_{im_i};\theta);\gamma) \;, \tag{2.1}$$

where $\psi = (\theta', \gamma')'$. The Clayton copula, for example, has the form

$$\mathcal{C}(u_{i1}, \dots, u_{im_i};\gamma) = \left(u_{i1}^{-\gamma} + \cdots + u_{im_i}^{-\gamma} - m_i + 1\right)^{-1/\gamma} \;, \quad \gamma \in [-1, \infty) \setminus \{0\} \;, \tag{2.2}$$

where γ is a scalar and Kendall's τ is given by $\tau = \gamma/(\gamma + 2)$ (Nelsen, 2006), having a range over $[-1, 0) \cup (0, 1]$.

The Gaussian copula, which could accommodate different pairwise associations through specification of a general correlation matrix, is given by

$$\mathcal{C}(u_{i1},\ldots,u_{im_i};\gamma)=\Phi_{m_i}(\Phi^{-1}(u_{i1}),\ldots,\Phi^{-1}(u_{im_i});\gamma)\ , \tag{2.3}$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of a standard normal (r.v.) and $\Phi_{m_i}(\cdot;\gamma)$ is the cumulative distribution function of an $m_i \times 1$ multivariate normal r.v. with mean zero and $m_i \times m_i$ covariance matrix $\Sigma_i(\gamma)=\Sigma_i$ with off-diagonal entries σ_{ijk}. This gives

$$P(T_{i1}>t_{i1},\ldots,T_{im_i}>t_{im_i}|Z_i;\psi)=\int_{-\infty}^{r_{i1}}\cdots\int_{-\infty}^{r_{im_i}}\frac{\exp\left(-s_i'\,\Sigma_i^{-1}\,s_i/2\right)}{\sqrt{(2\pi)^{m_i}\,|\Sigma_i|}}\,ds_{i1}\ldots ds_{im_i}\ , \tag{2.4}$$

where $S_i \sim \mathrm{MVN}_{m_i}(0,\Sigma_i)$, s_i is a realization, and $r_{ij}=\Phi^{-1}(\mathcal{F}(t_{ij}|Z_{ij};\theta))$, $j=1,\ldots,m_i$. The association between T_{ij} and T_{ik} conditional on (Z_{ij},Z_{ik}) is measured by Kendall's τ, given here by $\tau_{ijk}=2\arcsin(\sigma_{ijk})/\pi$, $1\le j<k\le m_i$, $i=1,\ldots,n$.

Flexible modeling of the within-cluster association can be achieved by specifying a second-order regression model of the form $g(\tau_{ijk})=v_{ijk}'\gamma$, where $g(\cdot)$ is a 1-1 differentiable link function mapping Kendall's τ onto the real line and v_{ijk} is a $q\times 1$ covariate vector representing family-level or individual-level features, or information on the structural relation between individuals j and k in family i. The Fisher transformation $g(\tau)=\log((1+\tau)/(1-\tau))$ is a natural choice for $g(\cdot)$, giving the second-order model

$$g(\tau_{ijk})=\log((1+\tau_{ijk})/(1-\tau_{ijk}))=v_{ijk}'\gamma\ , \tag{2.5}$$

which determines the structure of the positive definite covariance matrix. For a given v_{ijk} then $\tau_{ijk}=g^{-1}(v_{ijk}'\gamma)=(\exp(v_{ijk}'\gamma)-1)/(\exp(v_{ijk}'\gamma)+1)$ which can be estimated by inserting an estimate of γ on the right-hand side.

3 Likelihood and Composite Likelihood Construction under Biased Sampling

3.1 Maximum Likelihood Estimation and Inference

We consider the setting in which families are sampled through a proband. Without loss of generality, we assign the proband the label 0 and increase the dimension of the response and covariate vectors accordingly to m_i+1, $i=1,\ldots,n$. If T_{i0} denotes the disease onset time for the proband in family i and C_{i0} is the corresponding clinic entry time, the proband enters a registry if $T_{i0}<C_{i0}$. Members of the registry can then be randomly sampled into family study and if the m_i family members of proband i have event times $T_{i1},\ldots,T_{im_i}$ we assume here that they are observed subject to right censoring at their assessment times $C_{i1},\ldots,C_{im_i}$, respectively. We let $X_{ij}=\min(T_{ij},C_{ij})$ and $Y_{ij}=\mathrm{I}(T_{ij}<C_{ij})$, $j=0,\ldots,m_i$. If $Z_i=(Z_{i1}',\ldots,Z_{im_i}')'$, we let $\bar{Z}_i=(Z_{i0}',Z_i')'$ denote the full vector of covariates for family i, and similarly let $\bar{T}_i=(T_{i0},T_i')'$, $\bar{X}_i=(X_{i0},X_i')'$, $\bar{C}_i=(C_{i0},C_i')'$ and $\bar{Y}_i=(Y_{i0},Y_i')'$.

Censoring is assumed to be conditionally independent such that $\bar{T}_i\perp\bar{C}_i|\bar{Z}_i$, and non-informative, so the likelihood contribution from family i is

$$L_i(\psi)\propto P(\bar{X}_i,\bar{Y}_i|\bar{C}_i,\bar{Z}_i,T_{i0}<C_{i0};\psi)=P(\bar{X}_i,\bar{Y}_i|\bar{C}_i,\bar{Z}_i;\psi)/P(T_{i0}<C_{i0}|C_{i0},Z_{i0};\theta)\ , \tag{3.1}$$

which can be expressed in terms of (2.1). An example is given in Appendix A (see supplementary material available at *Biostatistics* online) where we illustrate how to construct the likelihood for response-biased family data based on a copula model for the slightly more general type of data discussed in Section 4.1. From (3.1), the contribution to the score vector and information matrix from family i are

$$S_i(\psi) = \frac{\partial \log L_i(\psi)}{\partial \psi} = \frac{\partial \log P(\bar{X}_i, \bar{Y}_i | \bar{C}_i, \bar{Z}_i; \psi)}{\partial \psi} - \frac{\partial \log F(C_{i0} | C_{i0}, Z_{i0}; \theta)}{\partial \psi} \; , \tag{3.2}$$

and

$$I_i(\psi) = - \, \frac{\partial^2 \log L_i(\psi)}{\partial \psi \partial \psi'} = - \, \left[\frac{\partial^2 \log P(\bar{X}_i, \bar{Y}_i | \bar{C}_i, \bar{Z}_i; \psi)}{\partial \psi \partial \psi'} - \frac{\partial^2 \log F(C_{i0} | C_{i0}, Z_{i0}; \theta)}{\partial \psi \partial \psi'} \right] \; , \tag{3.3}$$

respectively; the age of onset of the proband is contained in the vector $\bar{X}_i$ as $X_{i0} = T_{i0}$ and $Y_{i0} = 1$ by the sampling condition. The maximum likelihood estimator $\widehat{\psi}$ solves $\sum_{i=1}^n S_i(\psi) = 0$ and $\sqrt{n}(\widehat{\psi} - \psi)$ is asymptotically normally distributed with mean zero and variance $\mathcal{I}^{-1}(\psi)$, where $\mathcal{I}(\psi) = E[I_i(\psi)]$. The term $\partial^2 \log F(C_{i0} | C_{i0}, Z_{i0}; \theta) / \partial \psi \partial \psi'$ subtracted in (3.3) represents the loss of "information" about the marginal parameters due to the response-biased sampling.

3.2 Composite Likelihood under Biased Sampling

When family size m_i is large, it can be challenging to compute and maximize the full likelihood; see Appendix A (supplementary material available at *Biostatistics* online). We consider the use of composite likelihood (Lindsay, 1988, Cox and Reid, 2004) comprising contributions based on lower-dimensional subsets of individuals in each family. Working with lower-dimensional distributions leads to considerable simplifications in the analytical expressions and computation. Let $\mathcal{S}_{ir} = \{(0, j_1^{(s)}, \ldots, j_r^{(s)}),\ s = 1, \ldots, m_{ir}\}$ denote the set of $(r+1)-$tuples of individuals in family i containing the proband with size $m_{ir} = m_i!/[r!(m_i - r)!]$, $r = 1, \ldots, m_i$. For example, $\mathcal{S}_{i1} = \{(0, j),\ j = 1, 2, \ldots, m_i\}$, $\mathcal{S}_{i2} = \{(0, j, k),\ 1 \leq j < k \leq m_i\}$ and $\mathcal{S}_{im_i} = \{(0, 1, 2, \ldots, m_i)\}$. We then define a composite likelihood for family i as

$$\mathrm{CL}_{ri}(\psi) \propto \prod_{s=1}^{m_{ir}} P(\bar{W}_{is}^{(r)} | \bar{C}_{is}^{(r)}, \bar{Z}_{is}^{(r)}, T_{i0} < C_{i0}; \psi) \; , \tag{3.4}$$

where $W_{ij} = (X_{ij}, Y_{ij})'$, $\bar{W}_{is}^{(r)} = (W_{i0}', W_{ij_1^{(s)}}', \ldots, W_{ij_r^{(s)}}')'$, and the other vectors are likewise defined. For example, if $r = 1$ a simple "pairwise" conditional composite likelihood

$$\mathrm{CL}_{1i}(\psi) \propto \prod_{j=1}^{m_i} P(\bar{W}_{ij} | \bar{C}_{ij}, \bar{Z}_{ij}, T_{i0} < C_{i0}; \psi) \; , \tag{3.5}$$

is obtained requiring only the use of bivariate distributions, where $\bar{W}_{ij} = (W_{i0}, W_{ij})'$, $\bar{C}_{ij} = (C_{i0}, C_{ij})'$ and $\bar{Z}_{ij} = (Z_{i0}', Z_{ij}')'$. If $r = 2$, a composite likelihood based on all triplets of family members including the proband is obtained:

$$\mathrm{CL}_{2i}(\psi) \propto \prod_{1 \leq j < k \leq m_i} P(\bar{W}_{ijk} | \bar{C}_{ijk}, \bar{Z}_{ijk}, T_{i0} < C_{i0}; \psi) \; , \tag{3.6}$$

where $\bar{W}_{ijk} = (W_{i0}, W_{ij}, W_{ik})'$, $\bar{C}_{ijk} = (C_{i0}, C_{ij}, C_{ik})'$, and $\bar{Z}_{ijk} = (Z_{i0}', Z_{ij}', Z_{ik}')'$. See Appendix A (supplementary material available at *Biostatistics* online) for an illustrative example on composite likelihood construction.

The score functions arising from (3.5) and (3.6) are of the form $U_r(\psi) = \sum_{i=1}^n U_{ri}(\psi)$ where $U_{ri}(\psi) = \partial \log \mathrm{CL}_{ri}(\psi) / \partial \psi$ denotes the corresponding score function contributed from family i. If

$\tilde{\psi}$ denotes the maximum composite likelihood estimator from (3.5) or (3.6), then, under standard regularity conditions, $\sqrt{n}(\tilde{\psi} - \psi)$ converges in distribution to multivariate normal with mean vector zero, and covariance matrix

$$\text{asvar}(\sqrt{n}(\tilde{\psi} - \psi)) = \mathcal{A}^{-1}(\psi)\mathcal{B}(\psi)[\mathcal{A}^{-1}(\psi)]' \,, \tag{3.7}$$

where $\mathcal{A}(\psi) = -E\{\partial^2 \log \text{CL}_{ri}(\psi)/\partial\psi\partial\psi'\}$ and $\mathcal{B}(\psi) = E\{U_{ri}(\psi)U'_{ri}(\psi)\}$. This can be estimated by $\widehat{\text{asvar}}(\sqrt{n}(\tilde{\psi} - \psi)) = A^{-1}(\tilde{\psi})B(\tilde{\psi})[A^{-1}(\tilde{\psi})]'$, where $A(\psi) = -n^{-1}\sum_{i=1}^{n} \partial^2 \log \text{CL}_{ri}(\psi)/\partial\psi\partial\psi'$ and $B(\psi) = n^{-1}\sum_{i=1}^{n} U_{ri}(\psi)U'_{ri}(\psi)$.

3.3 Asymptotic Relative Efficiency of the Composite Likelihoods

Here we examine the asymptotic relative efficiency of the composite likelihood estimators compared to maximum likelihood as a function of the strength of the within-family association. We suppose that ascertained families are comprised of two generations made up of two parents and their children and assume that all family members have a common marginal onset time distribution with $\mathcal{F}(t_{ij};\theta) = \exp(-(\lambda t_{ij})^{\kappa})$, $j = 0, 1, \ldots, m_i$; where $\theta = (\lambda, \kappa)'$. We let $\kappa = 1.2$ and choose λ to give a median of 45 years of age. One of the family members is selected at random as the proband (with equal probability) and assigned the index $j = 0$. Their clinic entry time C_{i0} is normally distributed with mean $\mu = 50$ and variance $\sigma^2 = 20$, and conditional on this right-truncation time we generate $T_{i0}|T_{i0} < C_{i0}$. The latent onset times for the non-probands are then generated as $T_{i1}, \ldots, T_{im_i}|T_{i0}$ and the observed family data are created following the generation of the assessment times. Specifically, for non-probands in the first and second generations of family i, the random age of contact follows $N(\mu = 60, \sigma^2 = 10)$ and $N(\mu = 40, \sigma^2 = 10)$, respectively; the age at contact for all individuals are truncated at 90 years. We consider an exchangeable association structure based on the Clayton copula with Kendall's τ varying from 0.05 to 0.6, reflecting small to strong within-family association; model (2.5) simplifies to $\log\big((1+\tau_{ijk})/(1-\tau_{ijk})\big) = \gamma_0$, $0 \leq j < k \leq m_i$. The expected information matrix $\mathcal{I}(\psi)$ for the likelihood analysis and expected matrices $\mathcal{A}(\psi)$ and $\mathcal{B}(\psi)$ in (3.7) are approximated by Monte Carlo simulation based on 10,000 samples. We define the asymptotic relative efficiency of the composite likelihood approach as the ratio of the asymptotic variance of the maximum likelihood estimator to that of the composite likelihood estimator.

Figure 1 displays the trends in the asymptotic relative efficiencies of the two composite likelihood estimators compared to the maximum likelihood estimator for three settings. We consider the case in which all families are comprised of four individuals ($m_i = 3$, top row), all families are comprised of seven individuals ($m_i = 6$, middle row), and family sizes are random where M_i has a multinomial distribution with $P(M_i = m) = 0.25, 0.25, 0.20, 0.20$ and 0.10, for $m = 2, 3, 4, 5$ and 6, respectively (bottom row). It is apparent that for small families (top row) composite likelihood (3.6) yields quite efficient estimators for all parameters but the estimators based on (3.5) are far less efficient; the greatest relative efficiency arises when the within-family dependence is high. This general trend of greater efficiency under higher within-family association is intuitive because as the association becomes greater, the incremental value of information obtained by using higher dimensional joint models is naturally smaller. With large families (middle row) the relative efficiency of a composite likelihood estimator is lower, which is again intuitively reasonable as much more higher order information is lost when we only consider contributions from bivariate or trivariate models; the second composite likelihood (3.6) retains as much as 70-80% efficiency however. When families are of variable size, the efficiency loss based on (3.6) is greater than in the previous scenario and use of (3.5) again incurs a substantial loss of precision unless the within-family dependence is very strong. We conclude that dependence modeling should be based on (3.6) as it balances computational simplicity with good efficiency.

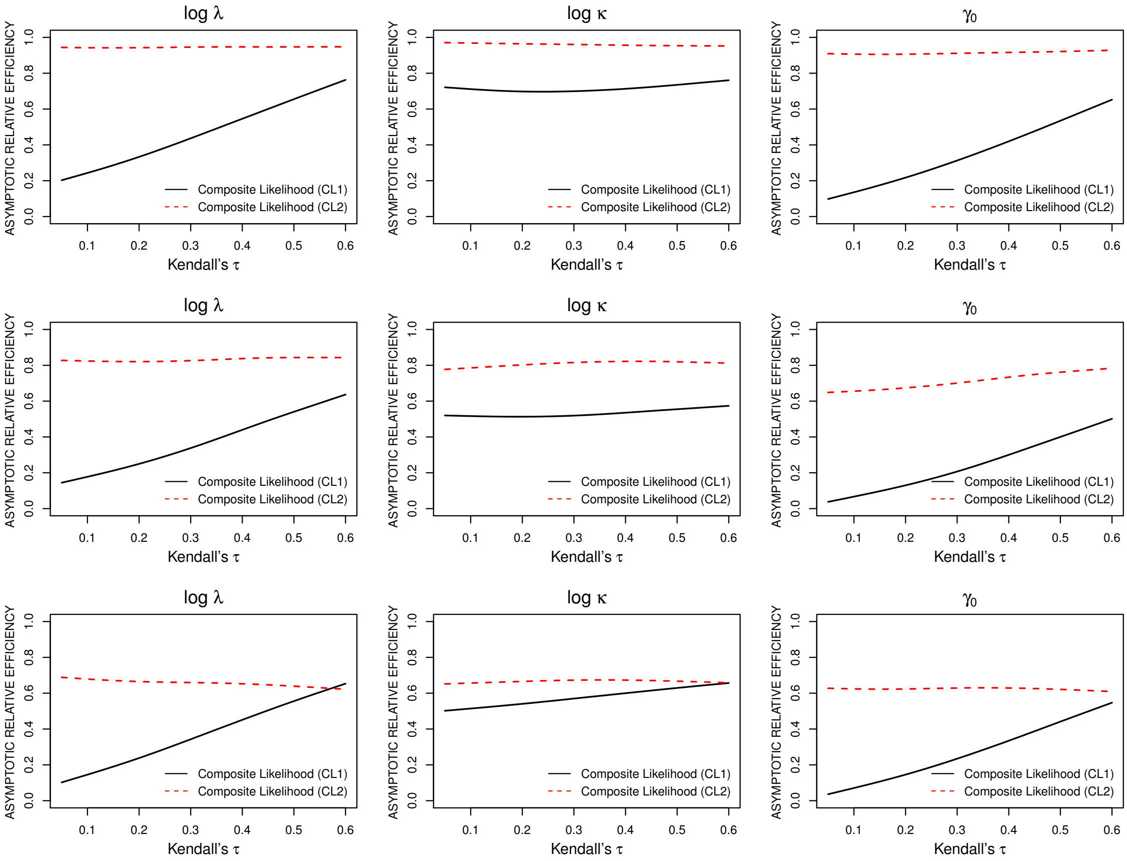

Figure 1: Asymptotic relative efficiency of the first (CL$_1$) and second (CL$_2$) composite likelihood estimators compared to maximum likelihood estimators for all parameters of the Weibull-Clayton copula model as a function of the within-family dependence (Kendall's τ) for family data under response-biased sampling in the presence of random right censoring; Monte Carlo approximations used for the Fisher information and expectations in (3.7) based on 10,000 samples; families have $m_i = 3$ (top row), $m_i = 6$ (middle row) or variable size (bottom row) with $M_i \sim \text{Multinomial}((2, 3, 4, 5, 6), p = (0.25, 0.25, 0.2, 0.2, 0.1))$.

3.4 Finite Sample Study of Composite Likelihood Methods

Here we report on simulation studies designed to assess the validity of the likelihood and two composite likelihoods along with the empirical relative efficiency. The parameter settings are as in Section 3.3 with $m_i = 3$. For the Clayton copula we let Kendall's $\tau = 0.4$, but to accommodate a more general within-family dependence structure, we also consider a Gaussian copula of the form (2.3) involving three types of association: between-parents, between-siblings and parent-child, with Kendall's τ denoted by τ_{pp}, τ_{ss} and τ_{ps}, respectively. We set $\tau_{pp} = 0.1$, $\tau_{ss} = 0.4$ and $\tau_{ps} = 0.2$, with the relative sizes of these measures compatible with the setting where genetic factors may contribute to the aetiology of this disease; the association between parents reflects the possible result of shared enviromental exposures. Therefore, $v_{ijk} = (1, v_{ijk1}, v_{ijk2})'$ in the second-order model (2.5), where $v_{ijk1} = \mathrm{I}((j,k) \text{ pair are siblings})$, $v_{ijk2} = \mathrm{I}((j,k) \text{ pair is parent} - \text{child})$, $0 \leq j < k \leq 3$.

One thousand datasets of $n = 1000$ families were then generated and analysed with likelihood (3.1) and composite likelihoods (3.5) and (3.6). The empirical results are summarized in Table 1 for both dependence structures. For all three methods, the biases are negligible, the empirical standard errors (ESEs) agree with the average standard errors (ASEs), and the empirical coverage probability (ECP) of nominal 95% confidence intervals (the proportion of simulated samples for which the nominal 95% confidence interval contained the true value) are all within an acceptable range. The ASEs are the smallest for all parameters under the likelihood analysis followed by those of the second composite likelihood and then those of the first composite likelihood, in alignment with expectations based on Section 3.3.

In some settings the mechanism for selecting families may be misspecified. We report on further simulation studies in Appendix B (see supplementary material available at *Biostatistics* online) designed to investigate the empirical biases of estimators in three scenarios involving misspecification of the ascertainment.

4 Extensions Dealing with Observation and Sampling Challenges

4.1 Accommodation of Right-censored and Current Status Observation

Information on disease onset time for non-probands is often collected retrospectively by a review of medical records or patient recall. For some non-probands determined to have the disease at the time of recruitment, however, no such information is available; this may arise when they are diagnosed for the first time upon recruitment, or if there are no medical records available. Such individuals furnish current status data with respect to their disease status (Sun, 2006), since all that is known is whether they have the condition at the time of recruitment and clinical examination. We let R_{ij} indicate that individual j in family i is under a right-censored observation scheme (due to the availability of a medical history) where $R_{ij} = 0$ if the individual is under a current status observation scheme; let $R_i = (R_{i1}, \ldots, R_{im_i})'$ and $\bar{R}_i = (R_{i0}, R_i')'$; since the probands are in a clinical registry where detailed information is available; $R_{i0} = 1$, $i = 1, \ldots, n$. For notational convenience we let $X_{ij} = C_{ij}$ if $R_{ij} = 0$, so X_{ij} denotes the time of the assessment for such individuals under a current status observation scheme; as before we let $Y_{ij} = I(T_{ij} < C_{ij})$. We can then write the likelihood as

$$L_i(\psi) \propto P(\bar{X}_i, \bar{Y}_i | \bar{R}_i, \bar{C}_i, \bar{Z}_i, T_{i0} < C_{i0}; \psi) \; , \tag{4.1}$$

and the analogous composite likelihoods as

$$\mathrm{CL}_{ri}(\psi) \propto \prod_{s=1}^{m_{ir}} P(\bar{W}_{is}^{(r)} | \bar{R}_{is}^{(r)}, \bar{C}_{is}^{(r)}, \bar{Z}_{is}^{(r)}, T_{i0} < C_{i0}; \psi) \; , \quad r = 1, 2 \; , \tag{4.2}$$

Table 1: Empirical properties of estimators based on the full likelihood, the first (CL_1) and the second (CL_2) composite likelihoods for family data under response-biased sampling in the context of random right censoring; for the Clayton copula Kendall's $\tau = 0.4$ and for the Gaussian copula $\tau_{pp} = 0.1$, $\tau_{ss} = 0.4$, $\tau_{ps} = 0.2$; $n = 1000$, $nsim = 1000$.

	Composite likelihood											
	CL_1				CL_2				Full likelihood			
	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP
Clayton copula												
$\log\lambda$	-0.004	0.099	0.099	0.936	-0.004	0.075	0.075	0.950	-0.004	0.073	0.074	0.956
$\log\kappa$	0.001	0.022	0.023	0.953	0.001	0.019	0.020	0.952	0.001	0.019	0.019	0.952
γ_0	0.001	0.133	0.133	0.947	0.003	0.089	0.090	0.957	0.003	0.085	0.086	0.965
τ	-0.001	0.055	0.055	0.947	0.000	0.037	0.037	0.957	0.001	0.035	0.036	0.963
Gaussian copula												
$\log\lambda$	-0.001	0.047	0.047	0.942	-0.001	0.041	0.041	0.940	-0.000	0.041	0.041	0.947
$\log\kappa$	0.001	0.020	0.020	0.956	0.001	0.018	0.019	0.956	0.001	0.018	0.019	0.956
γ_0	-0.003	0.075	0.075	0.957	-0.002	0.054	0.054	0.952	-0.001	0.052	0.052	0.951
γ_1	0.007	0.091	0.088	0.944	0.005	0.065	0.063	0.942	0.002	0.061	0.061	0.934
γ_2	0.003	0.064	0.066	0.947	0.002	0.043	0.044	0.951	0.001	0.040	0.042	0.959
τ_{pp}	-0.002	0.037	0.037	0.956	-0.001	0.027	0.027	0.954	-0.001	0.026	0.026	0.949
τ_{ss}	0.001	0.027	0.027	0.948	0.001	0.021	0.021	0.956	0.000	0.020	0.020	0.953
τ_{ps}	-0.000	0.024	0.024	0.957	-0.000	0.019	0.019	0.943	-0.000	0.019	0.019	0.939

ESE is empirical standard error; ASE is the average robust standard error; and ECP is the empirical coverage probability of nominal 95% confidence intervals.

where $\bar{R}_{is}^{(r)} = (R_{i0}, R_{ij_1^{(s)}}, \dots, R_{ij_r^{(s)}})'$. The asymptotic properties of estimators based on the full likelihood and the composite likelihoods are similar to those developed in Section 3. Simulation studies reported in Appendix A (see supplementary material available at *Biostatistics* online) demonstrate good empirical performance of estimators based on (4.1) and (4.2) with a combination of right-censored and current status family data.

4.2 Use of Auxiliary Data on the Marginal Incidence and Two-Stage Estimation

Since the onset times of probands are right-truncated and the prevalence of disease among non-probands is typically low, there is often little information about the marginal onset time distribution in family studies. Auxiliary data are often available, however, which may be exploited to reduce bias and/or improve efficiency (Pitkäniemi *and others*, 2009). Readily available auxiliary data in our setting is the right-truncated disease onset times of individuals not selected from the registry for the family study; we will show how these can be incorporated when participants are randomly selected. We also have current status data on disease onset times from a cross-sectional survey (Gelfand *and others*, 2005) and explore the use of this data under the assumption that the auxiliary processes share parameters with the processes governing the family data.

Let $\mathscr{F}$ denote the set of indices for probands in the family study and $\mathscr{A}$ the set of indices for individuals in an auxiliary sample. The augmented composite likelihood is

$$\mathrm{ACL}_r(\psi) = \prod_{i \in \mathscr{F}} \prod_{s=1}^{m_{ir}} P(\bar{W}_{is}^{(r)} | \bar{R}_{is}^{(r)}, \bar{C}_{is}^{(r)}, \bar{Z}_{is}^{(r)}, T_{i0} < C_{i0}; \psi) \prod_{a \in \mathscr{A}} P(X_a, Y_a | C_a, Z_a, T_a \in B_a; \theta) \ , \tag{4.3}$$

where B_a denotes the selection condition for individual a in the auxiliary sample. If we consider the auxiliary sample as comprised of unselected individuals from the original registry, then individuals in the auxiliary sample have right-truncated onset times like the probands; e.g. $B_a = (0, C_a)$ for $a \in \mathscr{A}$. For current status data from a cross-sectional survey there is no truncation, so $B_a = (0, \infty)$. Note that we can re-express (4.3) as $\mathrm{ACL}_r(\psi) = \mathrm{ACL}_{r1}(\theta) \times \mathrm{ACL}_{r2}(\psi)$, where

$$\mathrm{ACL}_{r1}(\theta) = \prod_{i \in \mathscr{F}} P(T_{i0} | C_{i0}, Z_{i0}, T_{i0} < C_{i0}; \theta) \prod_{a \in \mathscr{A}} P(X_a, Y_a | C_a, Z_a, T_a \in B_a; \theta) \tag{4.4}$$

is comprised of contributions from independent individuals and

$$\mathrm{ACL}_{r2}(\psi) = \prod_{i \in \mathscr{F}} \prod_{s=1}^{m_{ir}} P(W_{is}^{(r)} | \bar{R}_{is}^{(r)}, \bar{C}_{is}^{(r)}, \bar{Z}_{is}^{(r)}, t_{i0}; \psi) \ , \tag{4.5}$$

is based on correlated responses. If θ is high dimensional, we can consider a two-stage estimation (Shih and Louis, 1995) procedure by which (4.4) is maximized to obtain $\breve{\theta}$ (stage 1), $\breve{\theta}$ is plugged into (4.5), which is then maximized with respect to γ to obtain $\breve{\gamma}$ (stage 2). A derivation of the limiting distribution of $\breve{\psi} = (\breve{\theta}, \breve{\gamma})'$ is given in Appendix C (see supplementary material available at *Biostatistics* online).

4.3 Finite Sample Study of Augmented Composite Likelihood Methods

We carry out a simulation study to illustrate the performance of augmented composite likelihood using both simultaneous estimation and two-stage estimation procedures. We consider the same parameter setting of Section 4.1 with two types of auxiliary data: right-truncated individual data (to mimic the PsA registry data) and current status data (to mimic the national PsA survey data). The same marginal disease onset time distribution is assumed for all individuals from the auxiliary samples. The clinic entry times C_r follow the same distribution as that for the proband in the family study for individuals in the right-truncated auxiliary sample and the same distribution is used for the assessment times of the current status auxiliary sample. We generate the right-truncated event time by $T_a \sim T|T < C_a$, and the auxiliary data consist of $\{T_a, C_a, Y_a = 1;\ a = 1, \ldots, n_A\}$, where n_A is size of the auxiliary sample. For the current status sample, the resulting data are $\{C_a, Y_a;\ a = 1, \ldots, n_A\}$. One thousand replicates were generated with the sample size of the family study set to $n_F = 1000$ and the size of the auxiliary sample set to $n_A = 1,000$ or $20,000$. Both simultaneous and two-stage estimation procedures were carried out and the empirical properties of estimators are summarized in Table 2 for the Gaussian copula.

We find that when the size of the auxiliary sample increases, both simultaneous and two-stage estimation can lead to improved precision; simultaneous maximization leads to more efficient estimates than the two-stage procedure in all cases and so is recommended when feasible. When the auxiliary sample is large, the two-stage procedure can yield estimators almost as efficient as those obtained by simultaneous estimation. Current status auxiliary data in our settings lead to more efficient estimators than the right-truncated auxiliary data.

5 Application to the Psoriatic Arthritis Family Study

The incidence of PsA is reported to be between 0.3 and 1.0% (Gladman *and others*, 2005) and some studies have suggested that close blood relatives of individuals affected by PsA have a higher risk of developing this disease compared to the general population. Particular interest lies in assessing whether there is a higher rate of paternal, rather than maternal, transmission of the disease, reflecting the so-called "parent of origin" effect (Burden *and others*, 1998). While no genetic markers for PsA have been linked to the sex chromosomes, it is speculated that there may be sex-linked epigenetic markers which mediate transmission and penetrance (Pollock *and others*, 2015).

Here we consider data from a family study of PsA conducted in the Centre for Prognosis Studies in the Rheumatic Diseases at the University of Toronto. Probands were selected from members of the University of Toronto Psoriatic Arthritis Registry (UTPAR) based on consecutive presentation at the clinic for regularly scheduled appointments as part of an ongoing cohort study. A total of 169 families were recruited which range in size from 2 to 7 individuals; 54 families were comprised of only one non-proband (i.e. $m_i = 1$). There are 369 proband-non-proband pairs that can be constructed in the full dataset and among the 115 families with at least three members, a total of 332 triples can be formed which include the proband. Among the 538 distinct individuals in the dataset only 194 were diagnosed with PsA. The data on the onset time is of a mixed type, since while the event time is available for the proband, for other family members it may only be known whether they are diseased or not at the time of assessment; See Appendix D (supplementary material available at *Biostatistics* online) for more information on the PsA family study.

We begin with a descriptive analysis and plot the estimated cumulative hazard of PsA based on a non-parametric analysis (Sun, 2006) using the current status data from the survey of Gelfand *and others* (2005). A Weibull model is then fitted to the same data as well as the data obtained by pooling the current status survey data with data from the UTPAR; these estimates are plotted in the left panel of Figure 2. Since the onset times of all patients in the UTPAR are right-truncated, there is insufficient information to estimate (λ, κ) based on this data alone; the right panel of Figure 2 illustrates the flatness of the likelihood with respect to λ and hence the critical role of the auxiliary data in this setting. Similarly flat profile composite likelihoods are obtained when incorporating data from the family members of selected probands (not shown).

We maintain the Weibull model for the marginal onset time distribution and use a Gaussian copula

Table 2: Empirical properties of estimators based on augmented composite likelihoods ACL$_1$ and ACL$_2$ for a 50:50 mix of right censored and current status family data under response-biased sampling in the presence of right-truncated or current status auxiliary data; Gaussian copula with Kendall's $\tau_{pp} = 0.1$, $\tau_{ss} = 0.4$, $\tau_{ps} = 0.2$; $n_F = 1000$, $nsim = 1000$.

	Augmented composite likelihood ACL$_1$								Augmented composite likelihood ACL$_2$							
	Simultaneous				Two-stage				Simultaneous				Two-stage			
	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP
Right-truncated auxiliary data; $n_A = 1000$																
$\log \lambda$	-0.001	0.047	0.048	0.956	-0.005	0.157	0.152	0.946	-0.002	0.041	0.042	0.956	-0.005	0.157	0.152	0.946
$\log \kappa$	0.001	0.020	0.020	0.943	0.003	0.035	0.035	0.948	0.001	0.019	0.019	0.943	0.003	0.035	0.035	0.948
γ_0	-0.003	0.083	0.083	0.954	-0.000	0.151	0.151	0.919	-0.001	0.062	0.062	0.952	0.005	0.132	0.131	0.917
γ_1	0.006	0.098	0.096	0.949	0.001	0.105	0.104	0.940	0.004	0.074	0.073	0.953	-0.002	0.082	0.082	0.953
γ_2	0.001	0.072	0.072	0.950	-0.002	0.073	0.074	0.945	-0.000	0.051	0.052	0.952	-0.003	0.052	0.053	0.948
τ_{pp}	-0.002	0.041	0.041	0.953	-0.001	0.074	0.074	0.917	-0.000	0.031	0.031	0.953	0.002	0.064	0.064	0.913
τ_{ss}	0.001	0.030	0.030	0.954	-0.001	0.048	0.049	0.955	0.001	0.024	0.024	0.955	0.001	0.042	0.042	0.948
τ_{ps}	-0.001	0.027	0.026	0.948	-0.002	0.059	0.060	0.932	-0.000	0.021	0.021	0.935	0.000	0.053	0.053	0.925
Right-truncated auxiliary data; $n_A = 20,000$																
$\log \lambda$	-0.001	0.0360	0.035	0.947	0.000	0.046	0.046	0.953	-0.001	0.035	0.033	0.943	0.000	0.046	0.046	0.953
$\log \kappa$	-0.000	0.009	0.009	0.938	0.000	0.011	0.011	0.958	-0.000	0.009	0.009	0.938	0.000	0.011	0.011	0.958
γ_0	-0.000	0.079	0.076	0.944	-0.001	0.082	0.081	0.941	-0.001	0.062	0.059	0.929	-0.000	0.066	0.065	0.944
γ_1	0.001	0.099	0.096	0.941	0.000	0.099	0.096	0.942	-0.000	0.076	0.073	0.940	-0.001	0.077	0.073	0.939
γ_2	0.001	0.074	0.072	0.951	0.001	0.073	0.072	0.952	0.000	0.053	0.052	0.946	0.000	0.053	0.052	0.942
τ_{pp}	-0.000	0.039	0.038	0.944	-0.001	0.040	0.040	0.942	-0.000	0.030	0.029	0.930	-0.000	0.033	0.032	0.943
τ_{ss}	-0.000	0.029	0.029	0.950	-0.001	0.030	0.030	0.953	-0.001	0.024	0.023	0.949	-0.001	0.025	0.025	0.951
τ_{ps}	0.000	0.023	0.030	0.946	-0.000	0.026	0.026	0.954	-0.000	0.020	0.019	0.942	-0.000	0.022	0.022	0.951

(*Continued.*)

Table 2. (*Continued.*)

	Augmented composite likelihood ACL$_1$								Augmented composite likelihood ACL$_2$							
	Simultaneous				Two-stage				Simultaneous				Two-stage			
	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP
Current status auxiliary data; $n_A = 1000$																
$\log\lambda$	-0.000	0.030	0.031	0.953	0.000	0.038	0.038	0.944	-0.001	0.030	0.030	0.949	0.000	0.038	0.038	0.944
$\log\kappa$	0.001	0.024	0.023	0.947	0.001	0.030	0.029	0.942	0.000	0.022	0.022	0.951	0.001	0.030	0.029	0.942
γ_0	-0.003	0.076	0.074	0.948	-0.003	0.078	0.077	0.946	-0.001	0.058	0.058	0.952	-0.001	0.062	0.061	0.947
γ_1	0.006	0.097	0.096	0.947	0.005	0.097	0.096	0.944	0.004	0.074	0.073	0.950	0.004	0.075	0.073	0.949
γ_2	0.001	0.072	0.072	0.946	0.001	0.072	0.072	0.947	0.000	0.051	0.052	0.950	-0.000	0.051	0.052	0.951
τ_{pp}	-0.002	0.037	0.037	0.949	-0.002	0.038	0.038	0.944	-0.001	0.029	0.028	0.951	-0.001	0.031	0.030	0.947
τ_{ss}	0.001	0.028	0.029	0.950	0.001	0.029	0.030	0.949	0.001	0.023	0.023	0.954	0.001	0.024	0.024	0.943
τ_{ps}	-0.001	0.022	0.022	0.949	-0.001	0.024	0.024	0.951	-0.001	0.019	0.019	0.942	-0.001	0.021	0.021	0.947
Current status auxiliary data; $n_A = 20,000$																
$\log\lambda$	0.000	0.009	0.009	0.946	0.001	0.010	0.010	0.951	0.000	0.009	0.009	0.947	0.001	0.010	0.010	0.951
$\log\kappa$	0.001	0.024	0.023	0.938	0.001	0.028	0.027	0.946	0.000	0.022	0.021	0.936	0.001	0.028	0.027	0.946
γ_0	-0.001	0.072	0.069	0.948	-0.001	0.072	0.069	0.948	-0.001	0.056	0.053	0.939	-0.001	0.056	0.053	0.939
γ_1	0.000	0.098	0.095	0.938	0.000	0.098	0.095	0.939	-0.000	0.076	0.072	0.938	-0.001	0.076	0.073	0.938
γ_2	0.001	0.073	0.072	0.952	0.001	0.073	0.072	0.951	0.000	0.053	0.051	0.945	0.000	0.053	0.051	0.944
τ_{pp}	-0.001	0.035	0.034	0.947	-0.001	0.035	0.034	0.949	-0.001	0.028	0.026	0.939	-0.001	0.028	0.026	0.940
τ_{ss}	-0.001	0.028	0.028	0.947	-0.001	0.028	0.028	0.951	-0.001	0.023	0.022	0.952	-0.001	0.023	0.022	0.946
τ_{ps}	0.000	0.018	0.019	0.958	0.000	0.018	0.019	0.953	-0.000	0.016	0.016	0.956	-0.000	0.016	0.016	0.956

True values for parameters: $\log\lambda = -4.112$, $\log\kappa = 0.182$, $\gamma = (0.201, 0.647, 0.205)'$.

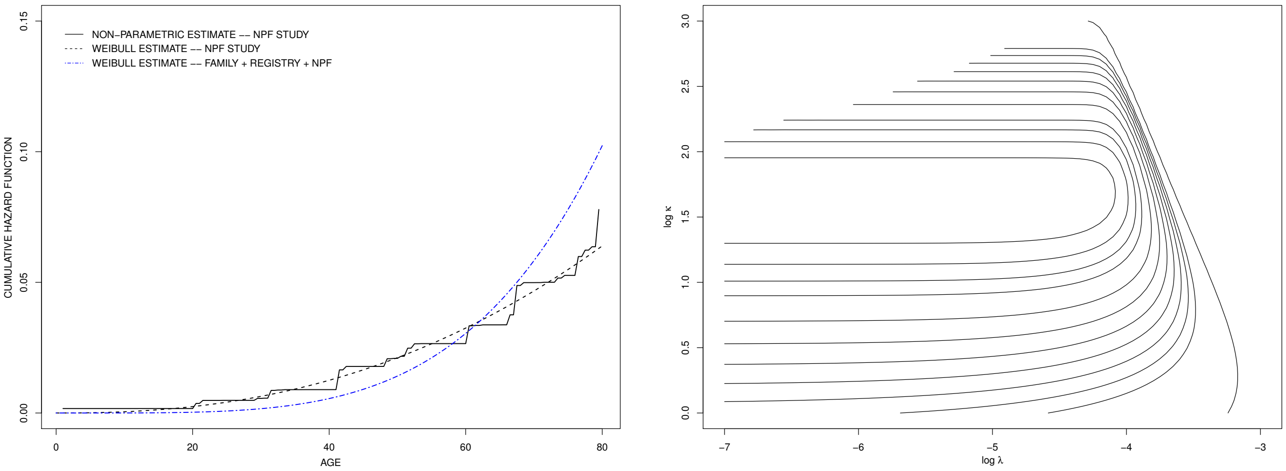

Figure 2: Non-parametric and parametric (Weibull) estimates of the cumulative hazard for the onset of PsA using data from the survey of the National Psoriasis Foundation (Gelfand *and others*, 2005) alone and pooled with data from the UTPAR (left panel); Log-likelihood contours for Weibull parameters (λ, κ) based on right-truncated disease onset times from the UTPAR (right panel).

with a second-order regression model given by

$$\log((1+\tau_{ijk})/(1-\tau_{ijk})) = \gamma_0 + \gamma_1 v_{ijk1} + \gamma_2 v_{ijk2} + \gamma_3 v_{ijk3} \ , \tag{5.1}$$

where $v_{ijk1} = I((j,k) \text{ pair are siblings})$, $v_{ijk2} = I((j,k) \text{ pair is father} - \text{child})$, and $v_{ijk3} = I((j,k) \text{ pair is mother} - \text{child})$. The test of the null hypothesis that the father-child association is the same as the mother-child association can be specified by $H_0 : \gamma_2 = \gamma_3$ vs. $H_A : \gamma_2 \neq \gamma_3$. We consider estimation based on augmented composite likelihoods ACL_1 and ACL_2 making use of auxiliary data on the marginal onset time distribution from $n = 734$ unselected individuals in the UTPAR who provide right-truncated onset times, and the current status data of $n = 15,307$ respondents in the national survey of Gelfand *and others* (2005). The augmentation term in (4.3) based on individuals in the survey has the form

$$P(X_a, Y_a | C_a, Z_a, T_a \in B_a; \theta) = F^{Y_a}(C_a | Z_a, C_a; \theta) \mathcal{F}^{1-Y_a}(C_a | Z_a, C_a; \theta) \ .$$

Even with the augmented data, since only 8 pairs of parents contribute terms to the first composite likelihood it is not possible to estimate the intercept in (5.1), so we fix $\gamma_0 = 0$ (or equivalently, $\tau_{pp} = 0$) to reflect the scenario that there is no environmental familial effect on the occurrence of PsA, and focus on the parent of origin hypothesis.

The top of Table 3 summarizes the estimates for the association parameters based on the augmented composite likelihoods with Weibull margins based on simultaneous and two-stage estimations. The results for the various methods are generally in close agreement. There is moderate association between siblings with Kendall's τ_{ss} around 0.23, suggesting a genetic influence on the PsA onset time. Furthermore, the estimated Kendall's τ for father-child association is quite different from that for mother-child pairs, which suggests that there might be different effect of parents' disease status on children. The Wald statistic for testing a parent of origin effect is given below the estimates along with the associated p-value, and we see that simultaneous or two-stage analyses based on ACL_2 yields borderline significant evidence of a parent-of-origin effect ($p = 0.046$). The bottom of Table 3 summarizes the results of fitting a marginal model with piecewise constant hazards with four cut points chosen to be the $20\%, 40\%, 60\%$ and 80% quantiles of the right-truncated onset time of PsA in the clinical cohort samples giving five pieces (PWC-5); these results are in broad agreement with those based on the model with the Weibull margins. Specifically the p-values for the Wald-based parent-of-origin hypothesis tests are $p = 0.049$ and 0.048 for the simultaneous and two-stage procedures, respectively. Based on these analyses, we reject the null hypothesis and conclude that father-child association in the onset time of PsA is significantly greater than the mother-child association. The corresponding p-values are all larger than 0.05 based on ACL_1 which may be due to the loss of efficiency explored in Section 3.

Table 3: Estimates of association parameters and Wald tests of the parent-of-origin hypothesis based on augmented composite likelihoods ACL_1 and ACL_2, using second-order regression model with $\gamma_0 = 0$; augmentation samples include unselected individuals from the University of Toronto Psoriatic Arthritis Clinic and the data from Gelfand *and others* (2005).

	ACL_1				ACL_2			
	Simultaneous		Two-stage		Simultaneous		Two-stage	
	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E
Weibull model for onset time								
γ_1	0.4387	0.0936	0.4381	0.0935	0.4685	0.1046	0.4673	0.1047
γ_2	0.1764	0.0895	0.1752	0.0892	0.1440	0.0929	0.1441	0.0933
γ_3	-0.0330	0.1081	-0.0340	0.1078	-0.1270	0.0999	-0.1275	0.0998
τ_{ss}	0.2159	0.0446	0.2156	0.0446	0.2301	0.0495	0.2295	0.0496
τ_{fc}	0.0880	0.0444	0.0874	0.0443	0.0719	0.0462	0.0719	0.0464
τ_{mc}	-0.0165	0.0540	-0.0170	0.0538	-0.0634	0.0498	-0.0637	0.0497
Statistic	1.490		1.489		1.995		1.995	
p-value	0.136		0.136		0.046		0.046	
Piecewise constant (PWC-5) model for onset time								
γ_1	0.4137	0.0967	0.4102	0.0965	0.4457	0.1094	0.4406	0.1097
γ_2	0.1891	0.0938	0.1891	0.0933	0.1583	0.0980	0.1622	0.0977
γ_3	-0.0242	0.1092	-0.0242	0.1088	-0.1138	0.0994	-0.1122	0.0997
τ_{ss}	0.2039	0.0464	0.2023	0.0463	0.2192	0.0521	0.2168	0.0523
τ_{fc}	0.0943	0.0465	0.0943	0.0462	0.0790	0.0487	0.0809	0.0485
τ_{mc}	-0.0121	0.0546	-0.0121	0.0544	-0.0568	0.0495	-0.0561	0.0497
Statistic	1.478		1.481		1.967		1.976	
p-value	0.140		0.139		0.049		0.048	

An important issue is whether the population sampled from the survey of Gelfand *and others* (2005) is the same as the population being sampled from for the UTPAR. To investigate the robustness of our findings, we let $\theta_a = (\log \lambda_a, \log \kappa_a)'$ denote the Weibull parameters for the model in the population being sampled from in Gelfand *and others* (2005); we retain $\theta = (\log \lambda, \log \kappa)'$ as the parameters for the population sampled from for the creation of the UTPAR. The contour plot in the right panel of Figure 2 illustrates the paucity of information regarding λ which cannot be estimated based only on the registry data, but we can carry out a 1 d.f. likelihood ratio test of $H_0 : \kappa = \kappa_a$ by fitting a model with separate κ parameters for the registry and survey data. We reject the null hypothesis with $p < 0.05$ and therefore repeat all of the analyses based on the generalized model with common λ but different κ parameters. When testing the parent-of-origin hypothesis based on this more general model, we obtain $p = 0.068$ for both simultaneous and two-stage estimation based on ACL_2. The lack of robustness of the conclusions suggests larger family studies are warranted which will depend less critically on auxiliary data and thereby furnish more robust evidence.

These findings, while mixed, are suggestive of a greater possible father-child association compared to the mother-child association. This is in broad agreement with findings in the current body of literature; see Pollock *and others* (2015) for a recent discussion and causal explanation for their effect.

6 Discussion

One purpose of this paper is to highlight the utility of copula models for obtaining interpretable measures of within-family dependence. Gaussian copula models, in particular, allow one to accommodate elaborate dependence structures that can provide insight into the genetic basis of disease. We feel that dependence modeling is much more natural via copulas than based on models with conditional independence assumptions given shared or correlated frailties because the dependence is functionally independent of the parameters in the marginal model. Moreover, while we have not emphasized this in the simulations or applications, copula models furnish estimates of covariate effects with simple marginal interpretations.

The efficiency loss incurred by use of composite likelihood can be modest when either family sizes are small or the within-family associations are modest. In these, and other settings where the loss can be more appreciable (i.e. when family sizes tend to be larger), this loss can be offset by exploitation of auxiliary data when it is available. In the motivating study this auxiliary data plays a crucial role in that there is limited information about the marginal onset time distribution because the onset times in the registry are all subject to right truncation; this lack of information also means that it is difficult to fully assess the compatability of the onset time distributions in the registry and the survey data. We carried out analyses allowing the trend parameters to differ in the two samples; it was not possible to do this with the rate parameters and we acknowledge this limitation. The lack of information on the onset time distribution is not an issue in analyses based on the binary disease status of participants, and in such settings there is little need for auxiliary data. We feel, however, that such models are typically based on invalid assumptions (i.e. that individuals with identical covariates but very different ages have the same cumulative risk of disease) and therefore yield uninterpretable measures of within-family association when the ages at assessment differ substantially. A generalization of the proposed method accommodating a non-susceptible fraction of individuals offers an alternative representation of this process and would enable one to model the within-family association in both the latent "at risk" indicator along with the time to disease onset among susceptible individuals. The bivariate model of Chatterjee and Shih (2001) could be extended to deal with larger cluster sizes and biased sampling with this in mind.

The construction of the complete data likelihood involving the unknown number of "potential probands" offers an alternative way of conceptualizing the optimization problem through a pseudo-augmentation approach which obviates the need for conditioning (Turnbull, 1976). Actual supplementary data can also be integrated into the complete data likelihood yielding a combination of pseudo and real data augmentation to improve efficiency. This approach can be computationally advantageous as the number of parameters in the marginal disease onset time distributions increases, particularly if software is available for semiparametric maximization of the likelihoods in untruncated samples (Lawless and Yilmaz, 2011). Clayton (2003) proposed an alternative approach to data augmentation for response-biased samples which simply aims to correct naive score functions by estimating their expectation under the selection conditions via simulation. Auxiliary data can also be used for the estimation of sampling weights to facilitate weighted likelihood-based analyses as discussed by Iversen and Chen (2005); we are exploring this in ongoing work but both of the latter approaches are not fully efficient.

Supplementary Material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgements

The authors thank Dr. Michael Siegel and the National Psoriasis Foundation for providing the data from the national survey, as well as Dr. Dafna Gladman and Remy Pollock for stimulating this research and helpful discussions. *Conflict of Interest*: None declared.

Funding

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN 155849) and the Canadian Institutes for Health Research (FRN 13887). R.J.C. is a Canada Research Chair in Statistical Methods for Health Research (950-226626).

References

Burden, A.D., Javed, S., Bailey, M., Hodgins, M., Connor, M. and Tillman, D. (1998). Genetics of psoriasis: paternal inheritance and a locus on chromosome 6p. *Journal of Investigative Dermatology* **110**, 958–960.

Burton, P.R. (2003). Correcting for nonrandom ascertainment in generalized linear mixed models (GLMMs), fitted using Gibbs sampling. *Genetic Epidemiology* **24**, 24–35.

Chatterjee, N. and Shih, J.H. (2001). A bivariate cure-mixture approach for modeling familial association in diseases. *Biometrics* **57**, 779–786.

Clayton, D. (2003). Conditional likelihood inference under complex ascertainment using data augmentation. *Biometrika* **90**, 976–981.

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

Gelfand, J.M., Gladman, D.D., Mease, P.J., Smith, N., Margolis, D.J., Nijsten, T., Stern, R.S., Feldman, S.R. and Rolstad, T. (2005). Epidemiology of psoriatic arthritis in the population of the United States. *Journal of the American Academy of Dermatology* **53**, 573–586.

Gladman, D.D., Antoni, C., Mease, P., Clegg, D.O. and Nash, P. (2005). Psoriatic arthritis: epidemiology, clinical features, course, and outcome. *Annals of the Rheumatic Diseases* **64**, ii14–ii17.

Hsu, L., Chen, L., Gorfine, M. and Malone, K. (2004). Semiparametric estimation of marginal hazard function from case–control family studies. *Biometrics* **60**, 936–944.

IVERSEN, E.S. AND CHEN, S. (2005). Population-calibrated gene characterization. *Journal of the American Statistical Association* **100**, 399–409.

JOE, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall, London.

LAIRD, N.M. AND LANGE, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics* **7**, 385–394.

LAWLESS, J.F. AND YILMAZ, Y.E. (2011). Semiparametric estimation in copula models for bivariate sequential survival times. *Biometrical Journal* **53**, 779–796.

LI, H. AND THOMPSON, E.A. (1997). Semiparametric estimation of major gene and family-specific random effects for age of onset. *Biometrics* **53**, 282–293.

LIANG, K.Y. AND BEATY, T.H. (1991). Measuring familial aggregation by using odds-ratio regression models. *Genetic Epidemiology* **8**, 361–370.

LINDSAY, B.G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 221–239.

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

PITKÄNIEMI, J., VARVIO, S-L. AND CORANDER, J. (2009). Full likelihood analysis of genetic risk with variable age at onset disease - combining population-based registry data and demographic information. *PLoS ONE* **4**, e6836.

POLLOCK, R.A., THAVANESWARAN, A., PELLETT, F., CHANDRAN, V., PETRONIS, A., RAHMAN, P. AND GLADMAN, D.D. (2015). Further evidence supporting a parent-of-origin effect in psoriatic disease. *Arthritis Care & Research*, doi: 10.1002/acr.22625.

PRENTICE, R.L. AND ZHAO, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**, 825–839.

SHIH, J.H. AND LOUIS, T.A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* **51**, 1384–1399.

SUN, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer, New York.

TURNBULL, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)* **38**, 290–295.

Published in final edited form as:
Biostatistics (2016), **17**(3): 437–452
DOI: 10.1093/biostatistics/kxv054

Supplementary Material for Augmented composite likelihood for copula modeling in family studies under biased sampling

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

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

APPENDIX A: LIKELIHOOD AND COMPOSITE LIKELIHOOD CONSTRUCTION

APPENDIX A.1: AN ILLUSTRATION OF LIKELIHOOD AND COMPOSITE LIKELIHOOD CONSTRUCTION

Here we give an illustrative example showing how the likelihood and composite likelihood can be constructed for a particular family under biased sampling. Consider a nuclear family involving two parents and two siblings where

1. the father is the proband with onset time T_0 and clinic entry time C_0.
2. the mother is disease-free at assessment time C_1.
3. the first sibling had a disease onset time prior to his/her age at contact C_2 but the disease onset time is unknown (current status).
4. the second sibling is disease-free at his/her age of contact C_3.

For simplicity, we do no consider covariates, and adopt a Weibull distribution for onset time with survival function $\mathcal{F}(t) = \exp(-(\lambda t)^\kappa)$; the density is $f(t) = \lambda\kappa(\lambda t)^{\kappa-1}\exp(-(\lambda t)^\kappa)$. Using the notation of Section 4, for the proband, $(X_0 = T_0, C_0, Y_0 = \mathrm{I}(T_0 < C_0) = 1)$ and $R_0 = 1$; for the mother, $(X_1 = C_1, Y_1 = 0)$, and $R_1 = 1$; for the first sibling, since he/she is under a current status observation scheme, then $(X_2 = C_2, Y_2 = 1)$ and $R_2 = 0$. The second sibling is also disease-free at their assessment time, so $(X_3 = C_3, Y_3 = 0)$ and $R_3 = 1$. The full likelihood (4.1) for this family can therefore be written as

$$\begin{aligned} L(\psi) &= P(X_0, X_1, X_2, X_3, Y_0 = 1, Y_1 = 0, Y_2 = 1, Y_3 = 0 | T_0 < C_0, \bar{C}, \bar{R}) \\ &= P(T_0, T_1 > C_1, T_2 \le C_2, T_3 > C_3 | T_0 < C_0, \bar{C}) \\ &= \frac{P(T_0, T_1 > C_1, T_2 \le C_2, T_3 > C_3 | \bar{C})}{P(T_0 < C_0 | C_0)} \\ &= \frac{P(T_0, T_1 > C_1, T_3 > C_3 | \bar{C}) - P(T_0, T_1 > C_1, T_2 > C_2, T_3 > C_3 | \bar{C})}{P(T_0 < C_0 | C_0)} . \end{aligned} \tag{A.1}$$

The full likelihood involves a four-dimensional random variable. Under a Gaussian copula (2.3), a joint survival function can be written as

$$P(T_0 > t_0, T_1 > C_1, T_2 > C_2, T_3 > C_3|\bar{C}) = \int_{-\infty}^{r_0} \cdots \int_{-\infty}^{r_3} \phi_4(s_0, s_1, s_2, s_3; \Sigma) ds_0 \dots ds_3 \ , \quad \text{(A.2)}$$

where $r_0 = \Phi^{-1}(\mathcal{F}(t_0))$ and $r_j = \Phi^{-1}(\mathcal{F}(C_j))$, $j = 1, 2, 3$, and $\phi_4(s_0, s_1, s_2, s_3; \Sigma)$ is the density of a r.v. $S = (S_0, S_1, S_2, S_3)' \sim \mathrm{MVN}(0, \Sigma)$. Then by taking the negative derivative of (**??**) with respect to t_0, we obtain

$$\begin{aligned} P(T_0, T_1 > C_1, T_2 > C_2, T_3 > C_3) &= -\frac{d}{dt_0} P(T_0 > t_0, T_1 > C_1, T_2 > C_2, T_3 > C_3) \\ &= \left[\int_{-\infty}^{r_1} \cdots \int_{-\infty}^{r_3} \phi_4(r_0, s_1, s_2, s_3; \Sigma) ds_1 \dots ds_3 \right] \cdot \phi^{-1}(r_0) f(t_0) \\ &= \Phi_3(r_1, r_2, r_3; \mu^*, \Sigma^*) f(t_0) \ , \quad \text{(A.3)} \end{aligned}$$

where $\Phi_3(\cdot\,; \mu^*, \Sigma^*)$ is the CDF of a three-dimensional multivariate normal distribution with mean μ^* and covariance matrix Σ^*. The last equation is because of the fact that the conditional distribution of $(S_1, S_2, S_3)|S_0$ is still multivariate normal if (S_0, S_1, S_2, S_3) is multivariate normal and the explicit expressions of μ^* and Σ^* can be easily obtained by the normal distribution theory. Note that $P(T_0, T_1 > C_1, T_3 > C_3)$ can be derived in a similar way, or one can simply set $r_2 = \infty$ in (**??**). Therefore by plugging (**??**) into (**??**), the full likelihood for this nuclear family can be written as

$$L(\psi) = \frac{[\Phi_3(r_1, \infty, r_3; \mu^*, \Sigma^*) - \Phi_3(r_1, r_2, r_3; \mu^*, \Sigma^*)] \cdot f(t_0)}{F(C_0)} \ . \quad \text{(A.4)}$$

To construct a composite likelihood, we first need to decide on the dimension(s) of the subsets $\mathcal{S}_r$. If $r = 1$, then $\mathcal{S}_1 = \{(0, j), j = 1, 2, 3\}$ and if $r = 2$, then $\mathcal{S}_2 = \{(0, j, k), 1 \le j < k \le 3\}$. Therefore the first composite likelihood (CL$_1$) can be written as

$$\begin{aligned} \mathrm{CL}_1(\psi) &= \prod_{j=1}^{3} P(X_0, Y_0 = 1, X_j, Y_j | \bar{C}_j, \bar{R}_j, T_0 < C_0) \\ &= P(T_0, T_1 > C_1 | \bar{C}_1, T_0 < C_0) P(T_0, T_2 \le C_2 | \bar{C}_2, T_0 < C_0) P(T_0, T_3 > C_3 | \bar{C}_3, T_0 < C_0) \\ &= P(T_0, T_1 > C_1 | \bar{C}_1, T_0 < C_0)(1 - P(T_0, T_2 > C_2 | \bar{C}_2, T_0 < C_0)) P(T_0, T_3 > C_3 | \bar{C}_3, T_0 < C_0) \ , \end{aligned}$$

where

$$P(T_0, T_j > C_j | \bar{C}_j, T_0 < C_0) = \frac{P(T_0, T_j > C_j | C_j)}{P(T_0 < C_0 | C_0)} = \frac{-d\mathcal{F}(t_0, C_j)/dt_0}{F(C_0)} = \Phi\left(\frac{r_j - \sigma_{0j} r_0}{\sqrt{1 - \sigma_{0j}^2}} \right) \cdot \frac{f(t_0)}{F(C_0)} \ ,$$

where σ_{0j} is the corresponding entry of Σ. From the expression of CL$_1$, we know that the first composite likelihood only involve pairs of family members, and it is therefore much easier to determine its closed form. Similarly, the second composite likelihood based on triples (CL$_2$), can be written as

$$\begin{aligned} \mathrm{CL}_2(\psi) &= \prod_{1 \le j < k \le 3} P(X_0, Y_0 = 1, X_j, Y_j, X_k, Y_k | \bar{C}_{jk}, \bar{R}_{jk}, T_0 < C_0) \\ &= P(T_0, T_1 > C_1, T_2 \le C_2 | C_0, C_1, C_2, T_0 < C_0) \times P(T_0, T_1 > C_1, T_3 > C_3 | C_0, C_1, C_3, T_0 < C_0) \\ &\quad \times P(T_0, T_2 \le C_2, T_3 > C_3 | C_0, C_2, C_3, T_0 < C_0) \\ &= \frac{P(T_0, T_1 > C_1 | C_1) - P(T_0, T_1 > C_1, T_2 > C_2 | C_1, C_2)}{F(C_0)} \times \frac{P(T_0, T_1 > C_1, T_3 > C_3 | C_1, C_3)}{F(C_0)} \\ &\quad \times \frac{P(T_0, T_3 > C_3 | C_3) - P(T_0, T_2 > C_2, T_3 > C_3 | C_2, C_3)}{F(C_0)} \ . \end{aligned}$$

We have already shown how to obtain $P(T_0, T_j > C_j|C_j)$ and $P(T_0, T_j > C_j, T_k > C_k|C_j, C_k)$ based on the copula function. The second composite likelihood involves triples of family members and requires to work on three-dimensional distribution. It is more complicated than CL_1 but easier than the full likelihood, in terms of both writing out and maximizing the objective function.

The benefits of using composite likelihood instead of full likelihood is more obvious when the family size is large and variable, and also when the family data are under complex censoring schemes. To illustrate this point, consider a study with families of 7 individuals comprised of two parents and 5 children. Furthermore, we assume the non-probands are all under a current status observation scheme, then the data we have are $\{(T_{i0}, C_{i0}, Y_{i0} = 1, X_{ij}, Y_{ij}); j = 1, \ldots, 6, i = 1, \ldots, n\}$, and the full likelihood is of the form

$$L(\psi) \;=\; \prod_{i=1}^{n} P(T_{i0}, Y_{i1}, \ldots, Y_{i6}|\bar{C}_i, T_{i0} < C_{i0}) = \prod_{i=1}^{n} P(T_{i0}, Y_{i1}, \ldots, Y_{i6}|\bar{C}_i)/F(C_{i0})\ ,$$

where $P(T_{i0}, Y_{i1}, \ldots, Y_{i6}|\bar{C}_i)$ involves $2^6 = 64$ possible combinations. It is tedious and time consuming, although possible, to write out the probability expressions for all these combinations and to maximize the function. Under composite likelihood, however, we require $2^1 = 2$ or $2^2 = 4$ probability expressions corresponding to the first or second composite likelihood respectively. Of course, the ease is at the cost of statistical efficiency loss. We therefore compare the asymptotic relative efficiency of composite likelihoods with full likelihood for different family sizes in the paper to give guidance on the consequences of adopting the composite likelihoods.

Appendix A.2: Composite Likelihood Under Right-Censored and Current Status Observation

Here we conduct a simulation study to assess the performance of the methods with right-censored and current status family data. Again we consider two-generation families comprised of two parents and two children. A Weibull distribution is adopted for the onset times for all family members; $\mathcal{F}(t_{ij}; \theta) = \exp(-(\lambda t_{ij})^{\kappa})$, $j = 0, 1, 2, 3$; $\theta = (\lambda, \kappa)'$. The clinic entry time distribution for the probands and examination time distribution for the non-probands are as in Section 3. We further generate a random binary indicator R_{ij} for non-probands, $j = 1, 2, 3$, which indicate their respective observation scheme with probability $P(R_{ij} = 1) = P(R_{ij} = 0) = 0.5$; if $R_{ij} = 1$, then a medical history is available for this member and we observe $X_{ij} = \min(T_{ij}, C_{ij})$ and $Y_{ij} = I(T_{ij} < C_{ij})$; otherwise, only current status data are available and we observe $Y_{ij} = I(T_{ij} < C_{ij})$ and C_{ij}. For the within-family association structure, a Clayton and a Gaussian copula are used. For the latter, three types of associations (between-parents, between-siblings and parent-child) are specified as they were in Section 3.4. Although the full likelihood is more efficient than the composite likelihood, writing out, computing, and maximizing the full likelihood are burdensome when the family size is large, the within-family association structure is complex or the family data are of a mixed type. Moreover, the second composite likelihood is quite efficient so we only apply the extended composite likelihoods (4.2) with $r = 1$ and 2 to the mixed-type family data with response-biased sampling under the exchangeable and more general within-family structures, respectively; the empirical properties of estimates are summarized in Table **??**. We find that the biases are all negligible, the empirical standard errors (ESE) agree with the average robust standard errors (ASE), and the empirical coverage probabilities (ECP) of nominal 95% confidence intervals are within the acceptable range for all parameters. The ASE under the second composite likelihood are smaller than those under the first composite likelihood. These findings support the validity of the extension of our proposed composite likelihood approaches to the mixed-type family data subject to the response-biased sampling.

Table A.1: Empirical properties of estimators based on composite likelihoods CL_1 and CL_2 for a 50:50 mix of right-censored and current status family data under response-biased sampling; for the Clayton copula Kendall's $\tau = 0.4$ and for the Gaussian copula $\tau_{pp} = 0.1$, $\tau_{ss} = 0.4$, $\tau_{ps} = 0.2$; $n = 1000$, $nsim = 1000$.

		Composite Likelihood CL_1				Composite Likelihood CL_2			
PARAM	TRUE	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP
					Clayton Copula				
$\log\lambda$	-4.112	-0.005	0.112	0.109	0.942	-0.001	0.084	0.081	0.947
$\log\kappa$	0.182	-0.000	0.027	0.027	0.955	-0.000	0.023	0.023	0.953
γ_0	0.847	0.002	0.149	0.148	0.958	-0.001	0.102	0.100	0.944
τ	0.400	-0.001	0.062	0.062	0.948	-0.001	0.042	0.042	0.945
					Gaussian Copula				
$\log\lambda$	-4.112	-0.001	0.049	0.049	0.951	-0.001	0.043	0.043	0.954
$\log\kappa$	0.182	0.000	0.024	0.024	0.952	0.000	0.022	0.022	0.950
γ_0	0.201	-0.001	0.083	0.083	0.950	0.001	0.064	0.063	0.937
γ_1	0.647	0.003	0.096	0.097	0.950	-0.000	0.073	0.073	0.951
γ_2	0.205	0.001	0.071	0.073	0.955	-0.001	0.052	0.052	0.948
τ_{pp}	0.100	-0.000	0.041	0.041	0.950	0.000	0.032	0.031	0.938
τ_{ss}	0.400	0.001	0.029	0.030	0.949	-0.000	0.023	0.024	0.959
τ_{ps}	0.200	0.000	0.028	0.027	0.942	-0.000	0.022	0.021	0.937

APPENDIX B: THE EFFECT OF MISSPECIFYING THE ASCERTAINMENT CONDITION

Here we examine the effect of misspecification of the ascertainment condition on estimation under a composite likelihood. We adopt a Weibull marginal model as before with a Gaussian copula having $\tau_{ss} = 0.4$, $\tau_{ps} = 0.2$ and $\tau_{ss} = 0.1$ and consider studies involving $n = 1000$ families with $m_i = 3$, $i = 1, \dots, n$. The analyses are based on the ascertainment condition $T_{i0} < C_{i0}$, but we generate data under three ascertainment conditions which are different from this.

CASE I: HIGH RISK FAMILIES ARE ASCERTAINED
Here we consider the case in which the proband must satisfy $T_{i0} < C_{i0}$ but in addition at least one non-proband must also be diseased at the time of their assessment. This is somewhat similar to a multiplex sampling scheme but with the proband retaining a special designation. In statistical terms the family ascertainment condition is then $T_{i0} < C_{i0}$ and $\sum_{j=1}^{3} Y_{ij} \geq 1$. The actually ascertained families will suggest a higher risk of developing disease than has been accounted for by the naive model since the condition $T_{i0} < C_{i0}$ is not sufficient. A consequence is that the marginal hazard for disease onset will be over-estimated; this is demonstrated empirically in Figure B.1 where the average estimated hazard is displayed under the misspecified model. The effect of this misspecification on the association parameters is less clear but the empirical biases for this particular model are shown in Table B.1.

CASE II: SAMPLING OF PROBANDS WITH EARLY AGE OF ONSET
Here we suppose that there is an assessment time of the proband generated from normal distribution with mean 40 and variance 50, but sampling of probands requires them to have an early age of onset, so the real ascertainment condition is that $T_{i0} < \min(C_{i0}, E_{i0})$ where we take $E_0 = 40$. Misspecification of the ascertainment condition in this setting should lead to an overestimation of the hazard and hence an underestimation of the survivor function; see Figure B.1. The consequences on the inferences regarding the within-family dependence parameter estimation is less transparent but the empirical results displayed in Table B.1 show that the biases can be substantial. The biases become larger when the early age of onset is less than 40.

CASE III: NO PROBAND IDENTIFIED *a priori*
Here we suppose that while one person is nominally identified a priori as the proband in the analysis, the actual ascertainment condition is simply that one or more individuals must be affected by the condition in the family. i.e. $\sum_{j=0}^{3} Y_{ij} \geq 1$. The requirement that $\sum_{j=0}^{3} Y_{ij} \geq 1$ is less restrictive than the requirement that $T_{i0} < C_{i0}$; note that the index $j = 0$ does not correspond to "*the*" proband in this setting. As a result, families are more likely to be selected under this scheme. The plots of the average cumulative hazards in Figure B.1 show that this misspecification leads to an under-estimation of the cumulative hazard and overestimation of the survivor function. The empirical results on point estimation of the association parameters given in Table B.1 are less intuitive but clearly biases can be appreciable.

In summary there is little robustness that can be claimed for this type of analysis to misspecification of the ascertainment condition, and while the direction of the consequent biases in the marginal parameters can be anticipated in some cases, a general intuition on the nature of the induced bias for association is elusive.

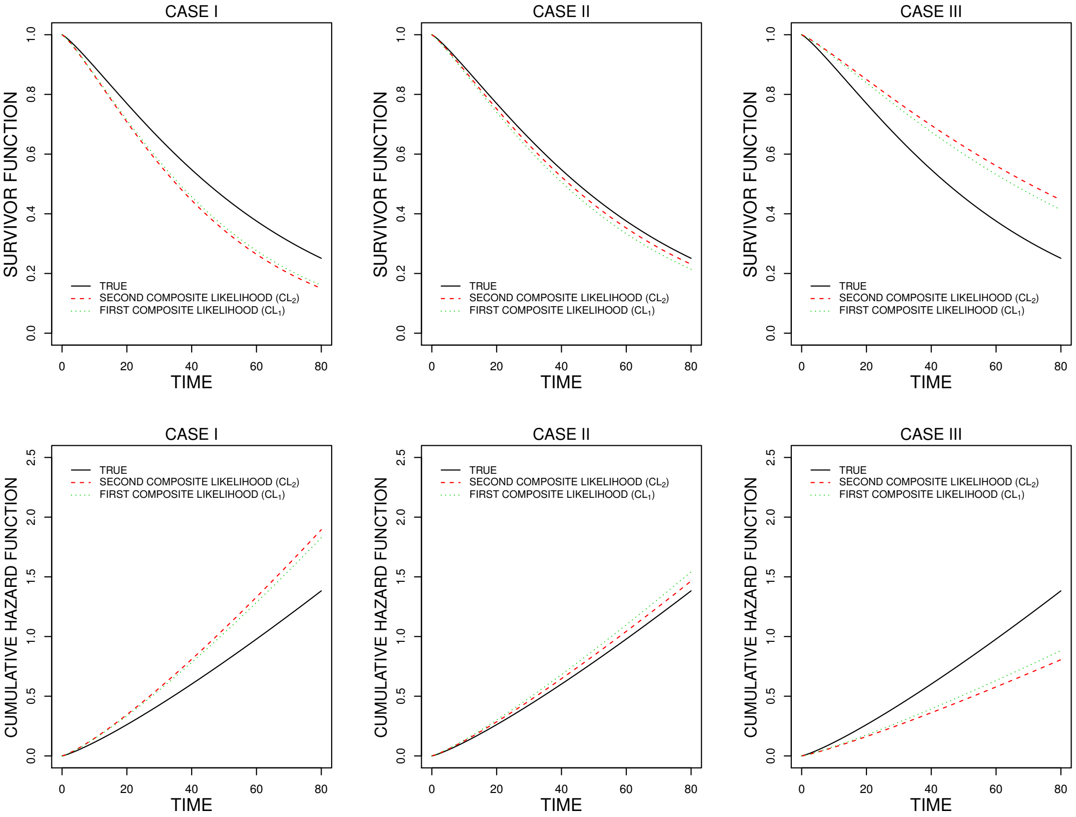

Figure B.1: Estimated survivor and cumulative hazard functions for event time based on composite likelihoods when the ascertainment condition is misspecified; Gaussian copula, $\tau_{pp} = 0.1$, $\tau_{ps} = 0.2$ and $\tau_{ss} = 0.4$; $m_i = 4$, $nsim = 1000$.

Table B.1: Empirical properties of estimates based on composite likelihoods when the ascertainment condition is misspecified; Gaussian copula, $\tau_{pp} = 0.1$, $\tau_{ps} = 0.2$ and $\tau_{ss} = 0.4$; $m_i = 4$, $nsim = 1000$.

		CL$_1$				CL$_2$			
PARAM	TRUE	MEAN	ESE	ASE	ECP	MEAN	ESE	ASE	ECP
CASE I: $T_{i0} < C_{i0}$ and $\sum_{j=1}^{3} Y_{ij} \geq 1$									
$\log \lambda$	-4.1121	-3.8890	0.0319	0.0315	0.000	-3.8613	0.0250	0.0243	0.000
$\log \kappa$	0.1823	0.2017	0.0199	0.0197	0.825	0.2048	0.0183	0.0180	0.749
γ_0	0.2007	0.1609	0.0812	0.0775	0.898	0.1021	0.0524	0.0499	0.500
γ_1	0.6466	0.5912	0.0945	0.0937	0.905	0.6078	0.0648	0.0633	0.908
γ_2	0.2048	0.1604	0.0762	0.0749	0.904	0.1623	0.0469	0.0466	0.854
τ_{pp}	0.1000	0.0802	0.0403	0.0385	0.899	0.0510	0.0261	0.0249	0.504
τ_{ss}	0.4000	0.3590	0.0282	0.0282	0.702	0.3406	0.0216	0.0211	0.197
τ_{ps}	0.2000	0.1592	0.0219	0.0221	0.544	0.1314	0.0171	0.0167	0.026
CASE II: $T_{i0} < C_{i0}$ and $T_{i0} < 40$									
$\log \lambda$	-4.1121	-4.0148	0.0438	0.0459	0.428	-4.0583	0.0404	0.0423	0.730
$\log \kappa$	0.1823	0.1640	0.0192	0.0194	0.859	0.1648	0.0177	0.0180	0.851
γ_0	0.2007	0.1089	0.0693	0.0722	0.771	0.1574	0.0512	0.0526	0.880
γ_1	0.6466	0.6914	0.0845	0.0842	0.909	0.6703	0.0623	0.0615	0.931
γ_2	0.2048	0.2181	0.0624	0.0622	0.947	0.2118	0.0433	0.0430	0.952
τ_{pp}	0.1000	0.0543	0.0345	0.0359	0.775	0.0785	0.0254	0.0261	0.881
τ_{ss}	0.4000	0.3797	0.0270	0.0272	0.905	0.3915	0.0203	0.0206	0.939
τ_{ps}	0.2000	0.1619	0.0240	0.0244	0.646	0.1825	0.0180	0.0187	0.842
CASE III: $\sum_{j=0}^{3} Y_{ij} \geq 1$									
$\log \lambda$	-4.1121	-4.4895	0.0682	0.0645	0.000	-4.5684	0.0695	0.0666	0.000
$\log \kappa$	0.1823	0.1514	0.0209	0.0216	0.704	0.1449	0.0195	0.0202	0.538
γ_0	0.2007	0.1653	0.0675	0.0649	0.897	0.2471	0.0653	0.0630	0.878
γ_1	0.6466	-0.0626	0.2536	0.2490	0.227	0.6557	0.0746	0.0711	0.937
γ_2	0.2048	0.3021	0.0554	0.0539	0.559	0.2361	0.0450	0.0438	0.877
τ_{pp}	0.1000	0.0824	0.0335	0.0322	0.898	0.1228	0.0321	0.0310	0.876
τ_{ss}	0.4000	0.0504	0.1252	0.1224	0.233	0.4228	0.0229	0.0231	0.823
τ_{ps}	0.2000	0.2293	0.0289	0.0276	0.790	0.2369	0.0226	0.0217	0.610

Appendix C: Asymptotic Properties with Augmented Composite Likelihood

Here we prove the asymptotic properties of the two-stage estimator for the augmented composite likelihoods proposed in Section 4. The augmented composite likelihoods can both be expressed as the product of two functions, the first is a function only of the marginal parameter θ as in (4.4), and the second is a function of $\psi = (\theta', \gamma')'$ as in (4.5).

We consider a set of independent individuals $\mathcal{P}$ for whom there is complete data and from which subjects are sampled for inclusion in the family study; let the number of individual in the set $\mathcal{P}$ be n. Let n_R and n_C denote the number of individuals in the registry data and survey data, then $n = n_R + n_C$. Without loss of generality, we assume the first n_R individuals are in the registry and the last $n - n_R$ are in the survey. In Section 4, we assume that individuals are selected by simple random sampling from the registry data and the second part of the augmented composite likelihood is constructed based on the sampled families only of size n_F. We let Δ_i indicate that individual i is sampled for the family study which occurs with probability $\pi = P(\Delta_i = 1)$, where $\pi > 0$, for individuals in the registry; then we let $n_F/n_R \to \pi$, as $n_F \to \infty$ and $n_R \to \infty$. The estimating functions for θ and γ are

$$U_1(\theta) = \sum_{i=1}^{n} U_{i1}(\theta) \; ,$$

and

$$U_2(\psi) = \sum_{i=1}^{n_R} U_{i2}(\psi) = \sum_{i=1}^{n_R} \frac{\Delta_i}{\pi} \cdot U_{i2}^*(\psi) \, ,$$

respectively, with

$$U_{i1}(\theta) = \frac{\partial}{\partial \theta} \log P(X_{i0}, Y_{i0} | C_{i0}, Z_{i0}, T_{i0} \in B_i; \theta) \; ,$$

where $B_i = (0, C_{i0})$ if individual i is in the registry or $B_i = (0, \infty)$ if they belong to the survey and yield current status data. The function $U_{i1}(\theta)$ is just the score function for truncated failure times or current status data, so we have $E[U_{i1}(\theta)] = 0$.

For the first augmented composite likelihood ($r = 1$),

$$U_{i2}^*(\psi) = \sum_{j=1}^{m_i} \frac{\partial}{\partial \gamma} \log P(W_{ij} | \bar{R}_{ij}, \bar{C}_{ij}, \bar{Z}_{ij}, t_{i0}; \psi) \; ,$$

and for the second augmented composite likelihood ($r = 2$),

$$U_{i2}^*(\psi) = \sum_{1 \le j < k \le m_i} \frac{\partial}{\partial \gamma} \log P(W_{ijk} | \bar{R}_{ijk}, \bar{C}_{ijk}, \bar{Z}_{ijk}, t_{i0}; \psi) \; .$$

Under simple random sampling,

$$E[U_{i2}(\psi)] = E_{D_i} \left\{ E_{\Delta_i} \left[\Delta_i U_{i2}^*(\psi)/\pi | D_i \right] \right\} = E_{D_i} \left\{ U_{i2}^*(\psi) \right\} = 0 \; ,$$

where D_i is the data from family i. So both estimating functions are unbiased.

If $\breve{\psi} = (\breve{\theta}', \breve{\gamma}')'$ denotes the solution to $U_1(\theta) = 0$ and $U_2(\gamma\,; \theta) = 0$, then

$$\sqrt{n}(\breve{\theta} - \theta) \approx \left[-\frac{1}{n} \frac{\partial U_1(\theta)}{\partial \theta'} \right]^{-1} \frac{1}{\sqrt{n}} U_1(\theta) \to \mathcal{I}_{11}^{-1} \frac{1}{\sqrt{n}} U_1(\theta) \, , \qquad \text{(C.1)}$$

as $n \to \infty$, where $\mathcal{I}_{11}(\psi) = E[-\partial U_{i1}(\theta)/\partial\theta]$. Also

$$U_2(\,\breve{\gamma};\,\breve{\theta}\,) = U_2(\gamma;\breve{\theta}) + \frac{\partial U_2(\gamma;\,\breve{\theta})}{\partial\gamma'}(\breve{\gamma}-\gamma) + o_p\left(\frac{1}{\sqrt{n_R}}\right) ,$$

and

$$U_2(\gamma;\breve{\theta}) = U_2(\gamma;\theta) + \frac{\partial U_2(\gamma;\theta)}{\partial\theta}(\breve{\theta}-\theta) + o_p\left(\frac{1}{\sqrt{n_R}}\right) .$$

Therefore as $n_F,\ n_R \to \infty$,

$$-\frac{1}{n_R}\cdot\frac{\partial U_2(\gamma;\breve{\theta})}{\partial\gamma} \to E[-\partial U_2(\gamma;\theta)/\partial\gamma] = \mathcal{I}_{22}(\psi) ,$$

and

$$-\frac{1}{n_R}\cdot\frac{\partial U_2(\gamma;\breve{\theta})}{\partial\theta} \to E[-\partial U_2(\gamma;\theta)/\partial\theta] = \mathcal{I}_{21}(\psi) .$$

Moreover as $n_F,\ n_R,\ n \to \infty$, we have

$$\frac{1}{\sqrt{n}}U_1(\theta) \to Z_1 \sim N(0, \mathcal{B}_{11}(\psi)) ,$$

and

$$\frac{1}{\sqrt{n_R}}U_2(\gamma\,;\theta) \to Z_2 \sim N(0, E(\Delta_i^2 U_{2i}^{*} U_{2i}^{*\prime}/\pi^2) = N(0, \mathcal{B}_{22}(\psi)/\pi) ,$$

where $\mathcal{B}_{11}(\psi) = E[U_{i1}(\theta)U_{i1}'(\theta)]$ and $\mathcal{B}_{22}(\psi) = E[U_{i2}^{*}(\psi)U_{i2}^{*\prime}(\psi)]$. Since

$$\begin{aligned}
E\left[\frac{1}{\sqrt{n_R}}U_2(\gamma\,;\theta)\,\frac{1}{\sqrt{n}}U_1'(\theta)\right] &= \frac{1}{\sqrt{n\,n_R}}E\left[\sum_{i=1}^{n_R}\sum_{j=1}^{n}\Delta_i U_{i2}^{*}(\psi)U_{j1}'(\theta)/\pi\right] \\
&= \frac{1}{\sqrt{n\,n_R}}\left\{E\left[\sum_{i=1}^{n_R}\sum_{j=1}^{n_R}\Delta_i U_{i2}^{*}(\psi)U_{j1}'(\theta)/\pi\right] + 0\right\} \\
&= \sqrt{\frac{n_R}{n}}E[U_{i2}^{*}(\psi)U_{i1}'(\theta)],
\end{aligned}$$

then $E[Z_2 Z_1'] = \sqrt{\alpha}\ \mathcal{B}_{21}(\psi)$ where $\mathcal{B}_{21}(\psi) = E[U_{i2}^{*}(\psi)U_{i1}'(\theta)]$, so as $n_F,\ n_R,\ n \to \infty$, and we let $n_R/n \to \alpha$,

$$\begin{aligned}
\sqrt{n_R}\,(\widehat{\gamma}-\gamma) &= \mathcal{I}_{22}^{-1}(\psi)\left\{\frac{1}{\sqrt{n_R}}U_2(\gamma;\theta) - \mathcal{I}_{21}(\psi)\sqrt{n_R}\,(\breve{\theta}-\theta)\right\} \\
&= \mathcal{I}_{22}^{-1}(\psi)\left\{\frac{1}{\sqrt{n_R}}U_2(\gamma;\theta) - \mathcal{I}_{21}(\psi)\sqrt{\frac{n_R}{n}}\,\mathcal{I}_{11}^{-1}(\psi)\left(\frac{1}{\sqrt{n}}U_1(\theta)\right)\right\} \\
&= \mathcal{I}_{22}^{-1}(\psi)\left\{Z_2 - \sqrt{\alpha}\mathcal{I}_{21}(\psi)\mathcal{I}_{11}^{-1}(\psi)Z_1\right\} .
\end{aligned} \tag{C.2}$$

Based on (**??**) and (**??**), we then have

$$\sqrt{n}(\breve{\theta}-\theta) \to N(0,\Sigma) \quad \text{and} \quad \sqrt{n_R}\,(\breve{\gamma}-\gamma) \to N(0,\Gamma) ,$$

where

$$\begin{aligned}
\Sigma &= \mathcal{I}_{11}^{-1}(\psi)\mathcal{B}_{11}(\psi)\left(\mathcal{I}_{11}^{-1}(\psi)\right)' , \\
\Gamma &= \mathcal{I}_{22}^{-1}(\psi)\Big\{\alpha\mathcal{I}_{21}(\psi)\mathcal{I}_{11}^{-1}(\psi)\mathcal{B}_{11}(\psi)\left(\mathcal{I}_{11}^{-1}(\psi)\right)'\mathcal{I}_{21}'(\psi) + \pi^{-1}\mathcal{B}_{22}(\psi) \\
&\qquad - \alpha\mathcal{B}_{21}(\psi)\left(\mathcal{I}_{11}^{-1}(\psi)\right)'\mathcal{I}_{21}'(\psi) - \alpha\mathcal{I}_{21}(\psi)\mathcal{I}_{11}^{-1}(\psi)\mathcal{B}_{12}(\psi)\Big\}\left(\mathcal{I}_{22}^{-1}(\psi)\right)' .
\end{aligned}$$

The asymptotic variance of the two-stage estimator can be consistently estimated by $\hat{\Sigma}$ and $\hat{\Gamma}$, where

$$
\begin{aligned}
\hat{\Sigma} &= \hat{I}_{11}^{-1}(\breve{\psi})\hat{B}_{11}(\breve{\psi})\left(\hat{I}_{11}^{-1}(\breve{\psi})\right)' , \\
\hat{\Gamma} &= \hat{I}_{22}^{-1}(\breve{\psi})\Big\{\alpha\hat{I}_{21}(\breve{\psi})\hat{I}_{11}^{-1}(\breve{\psi})\hat{B}_{11}(\breve{\psi})\left(\hat{I}_{11}^{-1}(\breve{\psi})\right)'\hat{I}'_{21}(\breve{\psi}) + \pi^{-1}\hat{B}_{22}(\breve{\psi}) \\
&\qquad - \alpha\hat{B}_{21}(\breve{\psi})\left(\hat{I}_{11}^{-1}(\breve{\psi})\right)'\hat{I}'_{21}(\breve{\psi}) - \alpha\hat{I}_{21}(\breve{\psi})\hat{I}_{11}^{-1}(\breve{\psi})\hat{B}_{12}(\breve{\psi})\Big\}\left(\hat{I}_{22}^{-1}(\breve{\psi})\right)' .
\end{aligned}
$$

with these expressions easily calculated based on the sample. For example,

$$
\hat{I}_{11}(\breve{\psi}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_{i1}(\theta)}{\partial \theta'}\Big|_{\theta=\breve{\theta}} , \quad \hat{I}_{22}(\breve{\psi}) \;=\; -\frac{1}{n_R}\sum_{i=1}^{n_R}\frac{\Delta_i}{\pi}\cdot\frac{\partial U_{i2}^{*}(\psi)}{\partial \gamma'}\Big|_{\psi=\breve{\psi}} ,
$$

and if π and α are unknown, they can be consistently estimated by $\hat{\pi} = n_F/n_R$ and $\hat{\alpha} = n_R/n$.

APPENDIX D: DESCRIPTION OF PSA FAMILY DATA

There are 169 two-generation families ranging in size from 2 to 7 individuals in the PsA family study and a crude summary of the corresponding data are given in Table **??**. A total of 538 individuals are in the family study and only 194 were diagnosed with PsA. Among the 169 families, 54 have only one non-probands ($m_i = 1$) and 115 have more than one non-probands ($m_i \geq 2$), which lead to 332 proband-involved triples of individuals. Among all of these triples, 86 include both a father and a mother, 156 include a father and a child, 274 include a mother and a child, and 246 include two or more siblings.

There are 369 pairs of family members which can be created including the respective proband and among these 8 include a father and mother, 107 include a father and child, 113 include a mother and child and 141 include two siblings.

Table D.1: Summary of PsA family data

		Number of non-proband siblings							
Mother	Father	0	1	2	3	4	5	6	Frequency
Proband	NA		10	1		1			12
Proband	Non-proband				1				1
NA	Proband		6	5	1				12
Non-proband	Proband		2	3	2				7
NA	NA		27	9		1		1	38
NA	Non-proband	3	2						5
Non-proband	NA	8	7	6	3		1		25
Non-proband	Non-proband	38	17	9	4	1			69
Total									169

NA means information is not available.