1 Introduction

Data from antenatal clinics (ANCs) in Botswana provide information on age, HIV status, and geographic cluster (i.e. clinic) for women in their child-bearing years. Such data are potentially useful for monitoring the HIV epidemic and for evaluating the impact of HIV prevention strategies in different settings. Using existing methods of surveillance is especially important because randomized studies of HIV prevention strategies are very expensive and can be done only in a very limited number of locations.

The analyses described below rely on publicly available data from ANCs in Botswana (hiv.gov.bw/ uploads). To make use of these data, it is necessary to adjust for correlation that arises due to geographic proximity of clinics and closeness in age among women within clinics. Each antenatal clinic can be viewed as a cluster. Within-cluster correlation may arise from similarity of prevalence rates among women close in age, and between-cluster correlation may arise from similarity in prevalence rates among clinics close in distance to one another. For example, people in districts closer to each other may tend to be more similar in behavior or factors related to susceptibility, or to have contacts within the same sexual networks.

The ANC data from Botswana motivate a generalized regression approach and nonparametric model for describing the relationship between HIV prevalence and covariates of interest. Natural cubic splines are a common choice for nonparametric mean functions in general [2] and GEE approaches in particular [6]. Furthermore, composite likelihood methods can be useful in cases where the sources of correlation are complex [7]. A pairwise composite likelihood approach has previously been applied in a semi-parametric setting for binary data, where it was assumed that the outcome variable was a function of a normally distributed spatial process [5].

2 Nonparametric estimation with an isotropic variance structure

Consider sites $s = 1, ..., S$ and the corresponding $N = \binom{S}{2}$ pairs of sites. We will refer to the pair that includes sites w and t as $pair_{wt}$, and the distance between sites w and t as d_{wt}. Within each site, measurements are available for the binary variable (HIV status $Y \in \{0, 1\}$) and covariate (J age groups $X \in \{1, ..., J\}$) for each of the n_s women in site s, as well as the distance D_s of the site from the approximate center of Gaborone. Our approach groups women within each site by covariate values, and takes the logit of the proportion of HIV positive women as the observed outcome. The model assumes the outcome $logit(p_{wt,j})$ belongs to an exponential family and that the systematic component of the exponential family is nonparametric. This implies that $logit(p_{wt,j}) = \eta_{wt,j} = f(x_{wt,j})$, where η is the canonical link function, and $logit(\hat{p}_{wt,j}) = \hat{\eta}_{wt,j}$ is the observed data. Throughout, we model elements of $\mathbf{f}$ using natural cubic splines, with $f_1, ..., f_J$ corresponding to the J covariate levels. A function is a natural cubic spline on some interval $[a, b]$ if it satisfies three conditions on this interval: i) the function is cubic on all sub intervals, ii) the function has continuous second and third derivatives at each knot of the spline, and iii) the second and third derivatives are 0 at a and b. Furthermore, we will model the relationship between prevalence and distance to Gaborone using a quadratic function, resulting in a parameterization of $\eta(\mu)$ with $\mu = \{f_1, ..., f_J, \beta_1, \beta_2\}$, where β_1 is the coefficient on D_s and β_2 is the coefficient on D_s^2.

For such a model, the pairwise score equations from the penalized quasi-likelihood for $pair_{wt}$ are

$$\Phi_{wt}(\mu) = \mathbf{D}_{wt}^t \mathbf{V}_{wt}^{-1}(\hat{\eta}_{wt} - \eta(\mu)) - \lambda \mathbf{G}\mu$$

where $D_{wt} = \partial\eta_{wt}/\partial\mu$ is a $2J \times (J + 2)$ matrix of partial derivatives. The working covariance matrix for the response from $pair_{wt}$ is denoted $\mathbf{V}_{wt} = Cov(\hat{\eta}_{wt}) = \mathbf{A}_{wt}^{(1/2)} R(\rho)_{wt} \mathbf{A}_{wt}^{(1/2)}$, where $\mathbf{A}_{wt}$ is the diagonal matrix with element $[j, j]$ equal to $1/(n_w \cdot \hat{p}_{w,j} \cdot (1 - \hat{p}_{w,j}))$ and element $[J + j, J + j]$ equal to $1/(n_t \cdot \hat{p}_{t,j} \cdot (1 - \hat{p}_{t,j}))$ for $j \in \{1, ..., J\}$. In our context, the correlation matrix is composed of sub-matrices for covariates X and distances d. Specifically, we apply isotropic models [5], calling $B(\rho_x)$ the $J \times J$ correlation matrix for age with element $B(\rho_x)_{ij} = \rho_x^{||i-j||}$, and $C(\rho_d, \rho_x, wt) = \rho_d^{||d_{wt}||} B(\rho_x)$ the correlation matrix for age and distance. Then $R(\rho)_{wt}$ is the $2J \times 2J$ block matrix

$$R(\rho)_{wt} = \begin{bmatrix} B(\rho_x) & C(\rho_d, \rho_x, wt) \\ C(\rho_d, \rho_x, wt) & B(\rho_x) \end{bmatrix}$$

with $B(\rho_x)$ on the diagonal blocks and $C(\rho_d, \rho_x, wt)$ on the off diagonal blocks. The value of λ adjusts the degree of smoothing, and $\mathbf{G}$ is chosen so that $\mathbf{G_f}$, the upper $J \times J$ portion of $\mathbf{G}$, satisfies $\int [f''(t)]^2 dt = \mathbf{f}^t \mathbf{G_f} \mathbf{f}$ while the remaining elements of $\mathbf{G}$ are zeros [2, 6].

The estimating equations in this context are:

$$\Phi(\mu) = \sum_{w=1}^{S-1} \sum_{t>w}^{S} \Phi_{wt}(\mu) = \mathbf{0}.$$

Applying a modified Fisher scoring algorithm yields the following iterative equation:

$$\hat{\mu}^{p+1} = \hat{\mu}^p + \left[\sum_{w=1}^{S-1}\sum_{t>w}^{S} \mathbf{D}_{wt}^t \tilde{\mathbf{V}}_{wt}^{-1} \mathbf{D}_{wt} + \lambda \mathbf{G}\right]^{-1} \times \left[\sum_{w=1}^{S-1}\sum_{t>w}^{S} \mathbf{D}_{wt}^t \tilde{\mathbf{V}}_{wt}^{-1} \left(\hat{\eta}_{wt} - \eta(\hat{\mu}^p)\right) - \lambda \mathbf{G}\hat{\mu}^p\right]$$

where $\hat{\mu}^p$ is the p^{th} iteration of the estimate for μ. The terms $\tilde{\mathbf{V}}_{wt}$, $\mathbf{D}_{wt}^t$, and $\hat{\eta}_{wt} - \eta(\hat{\mu}^p)$ are evaluated using the p^{th} iteration of μ. In our case, where the outcome variable is given by the logit of the estimated proportion of those with HIV for a specific covariate pattern, the variance and covariance structure can be derived from the delta method and estimated as a function of the estimated p for each unique age within a pairwise cluster. Therefore the only parameters left to be estimated are those given by the within- and between-cluster correlation. Estimation of these parameters can be based on a method-of-moments approach.

3 Large sample properties of the estimator

To demonstrate the consistency and asymptotic normality of an estimator defined as the root of the function $\mathbf{\Phi}(\mathbf{f})$, we make use of regularity conditions proposed by Lindsay [7]. To specify these conditions, we define two matrices. The first is the expected derivative of $\mathbf{\Phi}(\mathbf{f})$,

$$I_0 = \sum_{w=1}^{S-1}\sum_{t=s+1}^{S} E\left(\frac{\partial}{\partial \mathbf{f}} \mathbf{\Phi}_{(w,t)}(\mathbf{f})\right).$$

The second, under regularity conditions, provides the asymptotic variance of $\hat{\mathbf{f}}$:

$$I_N^{-1} = (I_0^{-1}) \times \left[\sum_{w=1}^{S-1}\sum_{t=s+1}^{S} E(\mathbf{\Phi}_{(w,t)}(\mathbf{f})\mathbf{\Phi}_{(w,t)}^T(\mathbf{f}))\right] \times (I_0^{-1})^T.$$

The term I_N is referred to as the Godambe information [7]. We provide the formal proof that $\hat{\mathbf{f}} \longrightarrow \mathbf{f}$ and $I_N^{-1/2}(\hat{\mathbf{f}} - \mathbf{f}) \longrightarrow N(\mathbf{0}, \mathbf{I})$ in the appendix.

Although analytical forms of the desired variances are available, they are not always computationally feasible or convenient. To reduce the complexity of these calculations, several procedures are available. Here we propose an analog to the general empirical variance method [5]. First we make the assumption that $N \times E(\Phi_N(\mathbf{f})\Phi_N(\mathbf{f})^T) \longrightarrow \mathbf{\Sigma}_\infty$, which implies that there is some finite variance covariance matrix for each pairwise cluster. Estimation of this matrix uses composite score evaluations over M subregions where $M < N$. Each subregion is of sample size S_k for $k = 1, ..., M$, and the estimate of $\mathbf{\Sigma}_\infty$ is given by

$$\hat{\mathbf{\Sigma}}_\infty = \frac{1}{M} \sum_{k=1}^{M} S_k \mathbf{\Phi}_{S_k}(\hat{\mathbf{f}}) \mathbf{\Phi}_{S_k}(\hat{\mathbf{f}})^T.$$

The large sample variance of the vector $\hat{\mathbf{f}}$ can then be estimated using

$$\hat{I}_N^{-1} = (\hat{I}_0^{-1}) \times \left[\frac{1}{N} \hat{\mathbf{\Sigma}}_\infty \right] \times (\hat{I}_0^{-1}).$$

4 Simulations

We use simulated data similar to those from ANCs in Botswana to investigate the bias and efficiency of the GEE smoothing spline estimator when ignoring correlation versus when accounting for correlation. We consider $J = 8$ age groups and generate a vector of the log odds of the probability of HIV infection ($log(p/(1-p))$) for each age group as $\mathbf{N}(\mathbf{f}(X_j), \Sigma_j)$ where $f(u) = sin(.5u/\pi)$, and Σ_j is an isotropic covariance matrix with ij^{th} element equal to $\sigma^2 \rho^{||i-j||}$. For simplicity we let $\sigma = 1$ and choose $\rho = .2$. Next, with each value of p serving as a mean parameter, we generated 50 (scenario 1) and 100 (scenario 2) Bernoulli random variables, for each age, to act as binary indicators of HIV prevalence. Finally, we assigned each observation randomly to 1 of 5 clusters, so that all observations exhibit two sources of correlation, between- and within-cluster correlation that are reflected in Σ. From each simulated data set, we used both our nonparametric estimator, and a crude estimator that ignores the correlation and calculates p as a sample proportion for each age. This was process was repeated 500 times.

In order to assess the bias of the GEE smoothing spline that accounts for both sources of correlation and the standard GEE that ignores correlation, we calculated the pointwise sum of absolute deviation (SAD) for each method. We denote the estimate of f, at the j^{th} unique value of X, for the r^{th} repetition as $\hat{f}_j^{(r)}$, and the pointwise average of f for the j^{th} unique value of X from 500 repetitions as $\hat{f}_j^* = \sum_{r=1}^{500} \hat{f}_j^{(r)}/500$. From the true value of f for each value of X (i.e. $f(u) = sin(.5u/\pi)$), we calculated the SAD as $\sum_{j=1}^{J} \mid \hat{f}_j^* - f(x_{(j)}) \mid$ [6]. The values of SAD are 0.0177 and 0.0171 (scenario 1 and scenario 2) for our spline estimator, and 0.0265 and 0.0263 (scenario 1 and scenario 2) for the crude estimator; thus our estimator exhibits less bias than does the estimator that ignores the correlation for both sample size scenarios.

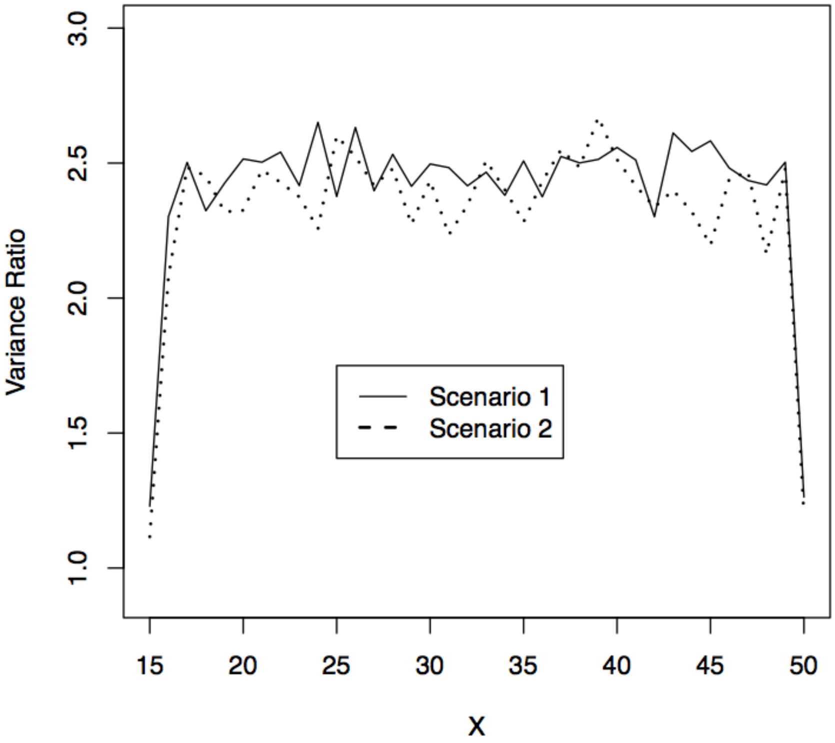

Figure 1: Variance ratio of two smoothing spline estimators, one accounting for correlation and the other assuming independence, from a simulation study.

Figure 1 displays the ratio of the variance the standard GEE estimator that ignores correlation to that of our GEE spline smoothing estimator. For each value of X our spline estimator has smaller variance. Ibrahim and Suliadi [6] show that failing to account for within-cluster correlation also results in a more biased and less efficient estimator.

5 Data analysis

Antenatal clinic data from 2009 are available from for 7331 pregnant women from 24 districts and 264 clinics in Botswana (www.hiv.gov.bw/uploads). We chose a subset of 10 clinics in a region in and around the capital, Gaborone. Figure 2 shows a map of these 10 clinics. Our focus is on the effects of proximity to Gaborone and of age on HIV prevalence.

To estimate these effects we use the semi-parametric GEE spline model from section 2.1. The observed outcome is the logit of the proportion of HIV positive women in age group j and $pair_{wt}$, $logit(\hat{p}_{wt,j}) = \hat{\eta}_{wt,j}$, where $\eta_{ij} = f(age_j) + \beta_1 D_s + \beta_2 D_s^2$, and D_s is distance from the approximate center of Gaborone. This model allows for a baseline age effect [estimate and SE] that varies by distance from Gaborone. We assume an isotropic correlation [estimates and SEs] structure, as used by Heagerty and Lele [5], that takes into account distances among districts, as well as the difference in ages groups. Within a district, the correlation structure reduces to AR(1). Choice of this correlation structure reflects an assumption that women close in age have more similar behaviors regarding both sexual activity and choice of partners than women further apart in age. We also assume that people within same community, or nearby community, have more similar risk than do those from communities further apart. Zeger, Liang and Albert [9] point out that a misspecified working correlation doesn't affect the consistency of the parameter estimates of the mean function; this property also holds for our methods. Also, for the purpose of demonstration, results for several values of the smoothing parameter (1, 25, 100 and 250) are shown. Figure 3 presents the crude estimates of prevalence by age for all clinics combined as well as the smoothed estimates of prevalence by age for Gaborone and districts that are 75 and 150 km from Gaborone.

6 Discussion

This paper combines the methods of Ibrahim and Suliadi [6] regarding spline estimation with those of Heagerty and Lele [5] regarding spatial correlation in order to estimate age-specific prevalence of HIV infection over geographic regions using

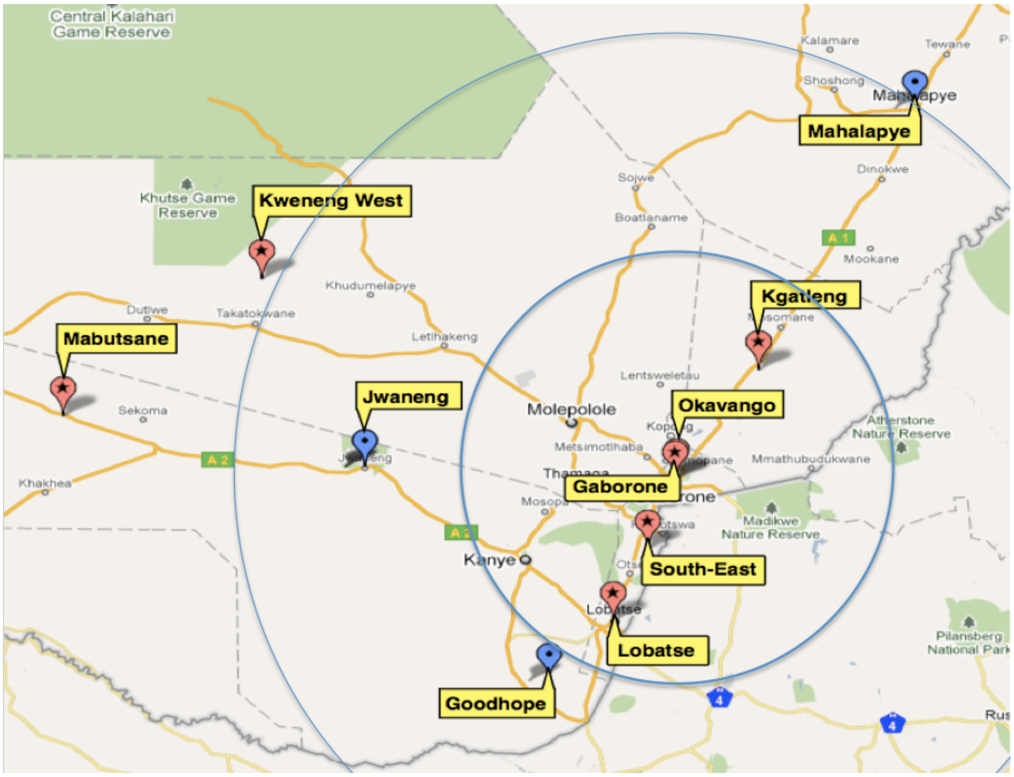

Figure 2: A map of the 10 antenatal clinics in and near Gaborone, Botswana used in this study. Also shown are radii 100km (inner circle) and 200km (outer circle) from Gaborone.

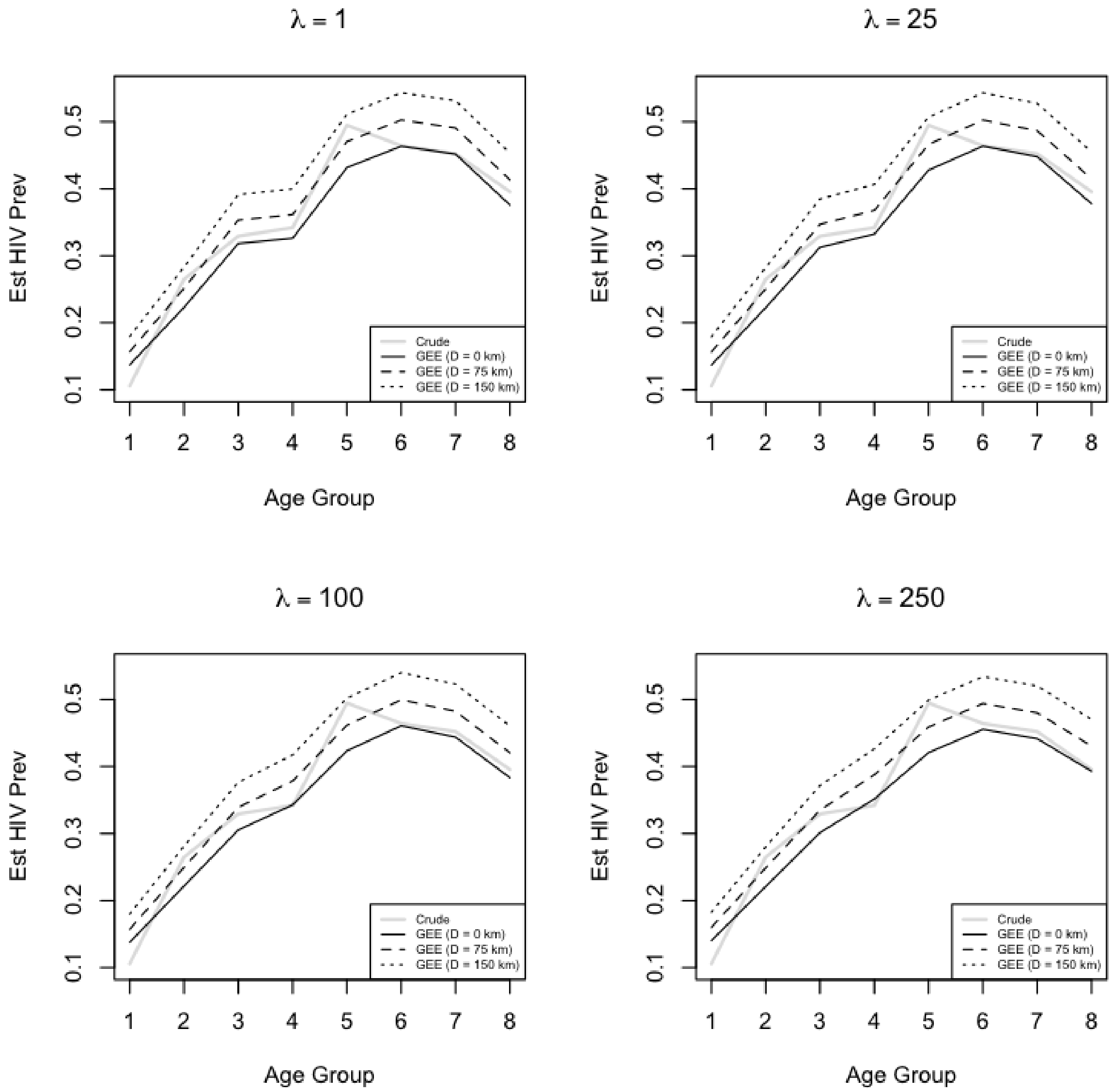

Figure 3: Crude prevalence of HIV by age group across 10 antenatal clinics in Botswana and smoothed estimates for 0, 75, and 150 kilometers from Gaborone. Results are provided for four different values of the smoothing parameter λ.

data from antenatal clinics. Such analyses require models for the age and geographic correlation, and we illustrate our methods assuming isotropic correlation. Further work is required to identify the best correlation structure for such settings. As mentioned above, correlation may arise from the positions in sexual networks occupied by women of different ages and regions as well as age or spatial effects on the nature of sexual behaviors and the prevalence of factors related to susceptibility. Investigation of the properties of such networks and their implications for correlation structure can ultimately provide structures that more fully reflect the process by which the data are generated. Fortunately, as mentioned above, estimates remain consistent even if the correlation structure is not correctly modeled.

Age-specific prevalence is useful for assessing the nature of HIV epidemic across communities, and repeated splines over chronological time in the same community can serve as a basis for estimation of incidence [4]. Future work is needed to develop methods for estimation of incidence from such splines.

In the current application, results were shown for multiple values of the smoothing parameter. Alternatively, the smoothing parameter could be estimated by cross validation. Wu and Zhang [8] suggest leave-one-out cross-validated deviance (SCVD) based on the deviance from a penalized quasi-likelihood (i.e. -2 × penalized quasi-likelihood), and estimated value of the mean for the j^{th} covariate pattern in $pair_{wt}$ using $\mathbf{f}^{(-wt)}$, which is the vector $\mathbf{f}$ obtained without the wt^{th} pairwise cluster. However, this method is computationally intensive, even when using approximations as described by Ibrahim and Suliadi [6].

References

[1] M. Crowder. On consistency and inconsistency of estimating equations. *Econometric Theory*, 2:305–330, 1986.

[2] P. Green and B. Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach.* Chapman and Hall / CRC, 1994.

[3] X. Guyon. *Random fields on a network: modeling, statistics, and applications.* Springer, 1995.

[4] T. Hallett, B. Zaba, J. Todd, B. Lopman, W. Mwita, S. Biraro, S. Gregson, and J. Boerma. Estimating incidence from prevalence in generalised HIV epidemics: methods and validation. *PLoS Medicine*, 5:e80, 2008.

[5] P. Heagerty and S. Lele. A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, 93:1099–1111, 1998.

[6] N. Ibrahim and Suliadi. Nonparametric regression for correlated data. *WSEAS Transactions on Mathematics*, 8:331–340, 2009.

[7] B. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80:221–239, 1988.

[8] H. Wu and J. Zhang. *Nonparametric regression methods for longitudinal data analysis*. Wiley-Blackwell, 2006.

[9] S. Zeger, K. Liang, and P. Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44:1049–1060, 1988.

7 Appendix: large sample properties of $\hat{f}$

7.1 Consistency of $\hat{\mathbf{f}}$

We use composite likelihood theory to establish the consistency of the estimator for $\mathbf{f}$ [1, 3, 5, 7]. In particular, Heagerty and Lele [5] provide the following regularity conditions for establishing consistency. First, define $D_N \subset \mathbb{R}^n$ to be the domain over which $\mathbf{\Phi}_N(\mathbf{f})$ is evaluated and call $\mid D_N \mid$ the cardinality of D_N.
Proposition 1: Assume there exists $\alpha > 0$ and a strictly increasing sequence of intergers, denoted by m_N, such that $\sum_{N\geq 1} N^\alpha \mid D_{m_N} \mid^{-1} < \infty$ and

$$\sum_{N\geq 1} N^\alpha \left(\frac{\mid D_{m_{N+1}} \backslash D_{m_N} \mid}{\mid D_{m_N} \mid} \right)^2 < \infty,$$

and the following conditions hold;

(i) $\mathbf{\Phi}_N(\mathbf{f})$ is continuous

(ii) $supE\left[\mathbf{\Phi}_N(\mathbf{f})\right] < \infty$ and $sup_N E\left[N \times \mathbf{\Phi}^2_N(\mathbf{f})\right] < \infty$

(iii) $\sum_{w=1}^{S-1} \sum_{t=w+1}^{S} \mathbf{D}_{wt} \tilde{\mathbf{V}}^{-1} \longrightarrow \mathbf{B}_\infty$, a finite limit

(iv) $I_N^{(1)} \longrightarrow I_\infty^{(1)}$ where λ_1, the minimum eigenvalue of $I_\infty^{(1)}$, is positive, and where $I_N^{(1)} = \sum_{w=1}^{S-1} \sum_{t=w+1}^{S} E\left[(\partial/\partial \mathbf{f})\mathbf{\Phi}_N(\mathbf{f})\right]$

(v) $(\partial/\partial \mathbf{f})\mathbf{D}_{wt}\tilde{\mathbf{V}}_{wt}$ is bounded uniformly in (w, t)

then $\hat{\mathbf{f}} \longrightarrow \mathbf{f}$.
Proof of Proposition 1: To establish consistency of our estimator, we will show the equivalence between our conditions (i.) - (iv.) and those described by Heagerty and Lele [5]. It is important to note that their asymptotic properties (R1, R2, and R3 of section 4.4), are for a parameter that maximizes a composite likelihood, not a function. But the analogous results will hold.

Condition (ii) is given by Guyon [3], and yields strong convergence of $\mathbf{\Phi}_N(\mathbf{f})$, namely $\mathbf{\Phi}_N(\mathbf{f}) \longrightarrow \mathbf{0}$. Under conditions (i) and (iii), a Taylor series expansion of $E\left[\mathbf{\Phi}_N(\mathbf{f})\right]$ is:

$$(\mathbf{f}_0 - \mathbf{f})^T E\left[\mathbf{\Phi}_N(\mathbf{f})\right] = (\mathbf{f}_0 - \mathbf{f})^T I_N^{(1)} (\mathbf{f}_0 - \mathbf{f}) + o \parallel \mathbf{f}_0 - \mathbf{f} \parallel^2 .$$

This expression and condition (iv) are sufficient for condition (R2). Lemmas 2.2 and 3.2 provided by Crowder [1] show that condition (R3) is equivalent to the same condition on $(\partial/\partial\mathbf{f})\mathbf{\Phi}_N(\mathbf{f})$. Therefore, condition (v) and the strong law of large numbers are jointly sufficient for (R3). This condition can be verified in practice by assuming bounded covariates and a bounded parameter space for $\mathbf{f}$. Therefore all the conditions of Proposition 1 have been met, implying consistency of $\hat{\mathbf{f}}$.

7.2 Asymptotic Normality of $\hat{\mathbf{f}}$

Guyon [3] and Lindsay [7] proposed general conditions for the asymptotic normality of the maximum likelihood estimator from a composite likelihood. Heagerty and Lele [5] have also proposed conditions for a semi-parametric probit model. Here, we restate conditions for our context.
Proposition 2: Consider the following conditions:

(a.) There exists an open neighborhood W of $\mathbf{f} \subset \mathbb{R}^n$ over which $\mathbf{\Phi}_N(\mathbf{f})$ is continuously differentiable, and there exists an integrable random variable q such that for all elements of $(\partial/\partial\mathbf{f})\mathbf{\Phi}_N(\mathbf{f})$ and all $\alpha \in W$, $(\partial/\partial\mathbf{f})\mathbf{\Phi}_N(\alpha, Y) < q(Y)$.

(b.) There exists a limiting covariance matrix $I_\infty^{(2)}$ such that $I_N^{(2)} = N \times E\left[\mathbf{\Phi}_N(\mathbf{f})\mathbf{\Phi}_N(\mathbf{f})^T\right]$ where $I_\infty^{(2)} > 0$ and $I_N^{(2)} > I_\infty^{(2)}$ for $N \geq m$ for some m.

(c.) There exists a sequence of matrices $I_N^{(1)}$ such that $I_N^{(1)} > I_\infty^{(1)}$ for $N \geq m$ for some m and $lim_{N\to\infty}\left[(\partial/\partial\mathbf{f})\mathbf{\Phi}_N(\mathbf{f}) - I_N^{(1)}\right] = \mathbf{0}$ in probability.

Under these conditions

$$\sqrt{N}\left[I_N^{(2)}\right] I_N^{(1)}(\hat{\mathbf{f}} - \mathbf{f}) \longrightarrow N(\mathbf{0}, \mathbf{I}).$$

Proof of Proposition 2: Condition (b.) implies that a central limit theorem can be applied to $\mathbf{\Phi}_N(\mathbf{f})$ and therefore implies $\sqrt{N}\left[I_N^{(2)}\right]\mathbf{\Phi}_N(\mathbf{f}) \longrightarrow N(\mathbf{0}, \mathbf{I})$. By assuming an exponential correlation decay, such as an isotropic correlation used by Heagerty and Lele [5], condition (b.) is met. Conditions (a.) and (c.) are satisfied as a result of (iv.) and (v.) in section 7.1. Under these conditions we can apply (3.4.5) of Guyon [3] to establish the asymptotic distribution of our estimator.