Semiparametric Frailty Models for Clustered Failure Time Data

Zhangsheng Yu[1] Xihong Lin[2] and Wanzhu Tu[1,3]

Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, U.S.A.[1,3]

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115 [2]

Regenstrief Institute, Inc., Indianapolis, IN, U.S.A[3]

email: yuz@iupui.edu[1]

SUMMARY: We consider frailty models with additive semiparametric covariate effects for clustered failure time data. We propose a doubly penalized partial likelihood (DPPL) procedure to estimate the nonparametric functions using smoothing splines. We show that the DPPL estimators could be obtained from fitting an augmented working frailty model with parametric covariate effects, whereas the nonparametric functions being estimated as linear combinations of fixed and random effects, and the smoothing parameters being estimated as extra variance components. This approach allows us to conveniently estimate all model components within a unified frailty model framework. We evaluate the finite sample performance of the proposed method via a simulation study, and apply the method to analyze data from a study of sexually transmitted infections (STI).

KEY WORDS: Doubly penalized partial likelihood; smoothing spline; Gaussian frailty; sexually transmitted disease; Smoothing parameter; Variance components.

1. Introduction

Frailty models are commonly used to model correlated failure time data. There is an extensive literature on parametric regression for covariate effects in frailty models (Therneau, Grambsch, and Pankratz 2003). For example, McGilchrist and Aisbett (1991) proposed a penalized partial likelihood approach for parameter estimation in a Gaussian frailty model setting. Following Breslow and Clayton (1993), Ripatti and Palgram (2000) used the Laplace approximation for the integrated likelihood of Gaussian frailty models. Murphy (1995) and Parner (1998) studied the asymptotic properties of Gamma frailty models. To accommodate complicated covariate effects and their unknown functional forms, we propose frailty models with some covariates modeled nonparametrically and some covariates modeled parametrically. To our knowledge, little work has been done on frailty models with semiparametric covariate effects for correlated failure time data.

This research is motivated by a study of sexually transmitted infections (STI). Evidence-based recommendations about the beginning age and frequency of STI screening require an accurate quantification of time from sexual debut to first infection (Meyers et al. 2008). In adolescent women, concurrent infections with multiple organisms are not uncommon, with *Chlamydia trachomatis*, *Neisseria gonorrhoeae* and *Trichomonas vaginalis* being the most frequently detected pathogenes. Once sexually active, young women are at risk of infection via sexual behavior. Evaluation of risk associated with STI acquisition is important as it provides clinically useful markers for more targeted STI screening. Since concurrent and repeated infections with multiple organisms within the same subject are correlated, it is desirable to use frailty models to examine factors related to the timing of STI following the onset of sexual activity.

The number of sex partners is often thought to be indicative of STI risk. However, the assumption of a linear effect for the number of partners is nonetheless questionable.

Intuitively, while a smaller number of partners typically indicates lower risk, a very large number of partners may not present proportionally higher risk because of increased sexual experience and prophylactic activities. Indeed, a closer examination of the fourth order polynomial effects for the lifetime number of sex partners revealed that the coefficients of all polynomial terms were significant at level 0.05. Hence it is desirable to model the effect of the number of sex partners nonparametrically while using the frailty to account for within-subject correlation.

For independent survival data, spline and kernel methods have been used in nonparametric regression. For example, O'Sullivan (1988) estimated the nonparametric covariate functions by smoothing splines. Hastie and Tibshirani (1990) proposed an alternative algorithm for fitting smoothing splines. Gray (1992, 1994) argued for the use of the penalized spline method. Local kernel methods were proposed by Tibshirani and Hastie (1987). Fan et al. (1997) studied the local kernel method and related asymptotic properties. More recently, Duchateau et al. (2004) studied frailty models with a single nonparametric covariate function using a smoothing spline. Cai et al. (2007, 2008) and Yu and Lin (2008) discussed a marginal likelihood estimation associated with local kernel methods for correlated survival data.

Du and Ma (2010) proposed a fully nonparametric hazard model with frailty $\log h(t, u; \boldsymbol{b}) = \eta(t, \boldsymbol{u}) + \boldsymbol{z^T b}$, with a fully nonparametric function of time and covariates $\eta(t, \boldsymbol{u})$, where $\boldsymbol{u}$ is a vector of covaraites and b is a frailty. Our proposed semiparametric model takes the form $\lambda(t; \boldsymbol{x_i}, \boldsymbol{w_i}, \boldsymbol{z_i}, \boldsymbol{b}) = \lambda_0(t) exp\{\sum_{j=1}^{p_1} \theta_j(x_{ij}) + \boldsymbol{w_i^T \gamma} + \boldsymbol{z_i^T b}\}$. It is different from Du and Ma's model in the following two aspects. (i) Modeling structure: our model distangles the baseline hazard and covariates in the same spirit as the Cox model, and models covariate effects using a semiparametric additive function which allows for both parametric and nonparametric covariate effects, with extensions to multiple covariates. Although Du and

Ma's nonparametric model takes a general functional space, models defined in a more restrictive functional space, such as semiparametric additive frailty models, are often much easier to interpret in biomedical applications. Better estimation efficiency associated with the parametric component also increases its appeal compared to the fully nonparametric approach. Further, multi-dimensional nonparametric functions are often difficult to fit. (ii) Parameter estimation: Our approach connects the semiparametric spline Cox model with a frailty model, which is parallel to the connection between mixed models and spline models, as recognized by Green (1984) and Lin and Zhang (1999). As a result, the nonparametric functions can be easily estimated as linear combinations of fixed and random effects. The smoothing parameters can be easily estimated as extra variance components which are naturally imbedded in the frailty model with a parametric covariate function. In contrary to our work, Du and Ma (2010) does not build a connection between the frailty model and spline Cox model and does not take an advantage of the variance components method for estimating the smoothing parameter. Instead, they used a cross-validation method for estimating the smoothing parameter which is computationally more intensive in general.

This article is structured as follows. In Section 2, we present our modeling structure. In Section 3, we discuss the DPPL method for estimation and inference of the semiparametric functions. In Section 4, we describe the estimation of variance components and smoothing parameters. In Section 5, we conduct a simulation study to evaluate the finite sample performance of the method. In Section 6, we illustrate the use of our model by analyzing data from an STI epidemiologic study. We conclude the paper in Section 7 with a few remarks about the proposed method.

2. Semiparametric Frailty Models

To assess the linear and potentially nonlinear effects of covariates in a frailty model setting, we introduce a general frailty model (1) with semiparametric additive covariate functions for clustered, hierarchial, and spatial data. We refer to this model as the semiparametric frailty model for simplicity.

For the *ith* observation, let T_i^* denote the underlying failure time, and C_i denote the censoring time for $i = 1, \cdots, n$, where n is the total number of observations. We write the observed time as $T_i = min(T_i^*, C_i)$. Let δ_i be a censoring indicator. The covariate vector $\boldsymbol{x}_i$ has p_1 components and $\boldsymbol{w}_i$ has p_2 components. The covariates $\boldsymbol{z}_i$ are assumed to be associated with random effects $\boldsymbol{b} = \{b_1, \cdots, b_q\}^T$. Conditioning on the random effects $\boldsymbol{b}$, we assume that survival times T_is are independent, and underlying failure time T_i^* is independent of censoring time C_i. We further assume that the censoring times are independent of the random effects $\boldsymbol{b}$.

We present in (1) a semiparametric frailty model with a general correlation structure. The model can be used to model clustered, hierarchical and spatial survival data,

$$\lambda(t; \boldsymbol{x_i}, \boldsymbol{w_i}, \boldsymbol{z_i}, \boldsymbol{b}) = \lambda_0(t) exp\{\sum_{j=1}^{p_1} \theta_j(x_{ij}) + \boldsymbol{w_i^T \gamma} + \boldsymbol{z_i^T b}\}, \tag{1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard, $\theta_j(x_j)$ is a nonparametric covariate function for the covariate x_j, $\boldsymbol{w}$ is the covariate vector whose effect is modeled linearly, random effects $\boldsymbol{b} \sim \text{MVN}(\boldsymbol{0}, \boldsymbol{D}(\boldsymbol{\nu}))$, with a full rank covariance matrix $\boldsymbol{D}(\boldsymbol{\nu})$, and $\boldsymbol{\nu}$ is a vector of variance components. Examples of the choices of covariance structures for clustered, hierarchical and spatial survival data can be found, e.g., in Breslow and Clayton (1993).

We develop below an estimation and inference procedure for the general semiparametric frailty model (1), of which the model for clustered survival data is a special case. The integrated likelihood of $(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\nu})$ for the observed data under the general additive frailty

model (1) is

$$\begin{aligned} & L\{\lambda_0(t), \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\nu}\} \qquad (2) \\ = & \int \Pi_{i=1}^{n} \lambda(t_i; \boldsymbol{x_i}, \boldsymbol{w_i}, \boldsymbol{z_i}, \boldsymbol{b})^{\delta_i} exp\{-\Lambda(t_i; \boldsymbol{x_i}, \boldsymbol{w_i}, \boldsymbol{z_i}, \boldsymbol{b})\} f\{\boldsymbol{b}, \boldsymbol{D}(\boldsymbol{\nu})\} d\boldsymbol{b} \\ = & \frac{1}{|\mathbf{D}(\boldsymbol{\nu})|^{1/2}} \int \Pi_{i=1}^{n} \{\lambda_0(t_i) e^{\sum_{j=1}^{p_1} \theta_j(x_{ij}) + \boldsymbol{w}_i^T \gamma + \boldsymbol{z}_i^T \boldsymbol{b}}\}^{\delta_i} \\ & \times exp\{-\Lambda_0(t_i) e^{\sum_{j=1}^{p_1} \theta_j(x_{ij}) + \boldsymbol{w}_i^T \gamma + \boldsymbol{z}_i^T \boldsymbol{b}}\} e^{-\frac{1}{2} \boldsymbol{b}^T \boldsymbol{D}(\boldsymbol{\nu})^{-1} \boldsymbol{b}} d\boldsymbol{b}, \end{aligned}$$

where $\boldsymbol{b} \sim MVN(0, \boldsymbol{D}(\boldsymbol{\nu}))$. Since the $\theta_j(\cdot)$ are nonparametric functions, we discuss in the next section their estimation using smoothing splines.

3. Estimation of the Semiparametric Functions Using Smoothing Splines

3.1 *Smoothing Spline Estimation*

In this section, we discuss estimation and inference of model (1) using smoothing splines. We assume the nonparametric covariate functions $\theta_j(\cdot)$ to be smooth and twice-differentiable. Smoothing spline estimation of the $\theta_j(\cdot)$ can proceed by maximizing the penalized integrated loglikelihood

$$\ell\{\lambda_0(t), \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\nu})\} + \sum_{j=1}^{p_1} \frac{1}{2\tau_j} \int \theta_j^{(2)}(x)^2 dx, \qquad (3)$$

where $\ell(\cdot) = logL(\cdot)$ and $L(\cdot)$ is defined in (2), $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_{p_1})^T$ is a vector of smoothing parameters controlling the goodness of fit and the smoothness of the curves.

Let $\boldsymbol{\theta}_j = \{\theta_j(x_{1j}^0), \cdots, \theta_j(x_{r_j j}^0)\}^T$, where x_{kj}^0 $(k = 1, \cdots, r_j)$ are the ordered r_j distinct covariate values for the jth covariate x_j. We assume $\sum_{k=1}^{r_j} \theta_j(x_{kj}^0) = 0$ to ensure identifiability similar to Lin and Zhang (1999). Let $\boldsymbol{N}_j$ be an n by r_j indicator matrix such that $\{\theta_j(x_{1j}), \cdots, \theta_j(x_{nj})\}^T = \boldsymbol{N_j \theta_j}$. The conditional hazard of the ith observation given the random effects $\boldsymbol{b}$ can then be rewritten as

$$\begin{aligned} & \lambda(t; \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{z}_i, \boldsymbol{b}) = \lambda_0(t) \\ & \times exp\{\boldsymbol{N}_{i1}^T \boldsymbol{\theta}_1 + \cdots + \boldsymbol{N}_{ip_1}^T \boldsymbol{\theta}_{p_1} + \boldsymbol{w_i^T \gamma} + \boldsymbol{z}_i^T \boldsymbol{b}\}, \qquad (4) \end{aligned}$$

where $\boldsymbol{N_{ij}}^T$ is the ith row of the indicator matrix $\boldsymbol{N_j}$. Using the results of O'Sullivan (1988) and noting equation (3) is a continuous function of the $\theta_j(\cdot)$, one can show that, for given values of $\lambda_0(t)$, $\boldsymbol{\tau}$ and $\boldsymbol{\nu}$, the maximizer of the integrated penalized loglikelihood (3) is a vector of natural cubic smoothing spline estimators of the $\theta_j(\cdot)$, and (3) can be equivalently written as

$$\ell\{\lambda_0(t), \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\nu})\} + \sum_{j=1}^{p_1} \frac{1}{2\tau_j} \boldsymbol{\theta_j^T K_j \theta_j}, \tag{5}$$

where $\boldsymbol{K_j}$ is the smoothing spline penalty matrix of the covariate x_j. See Section 2.2 of Green and Silverman (1994) for more details of the proof of solution to be smoothing spline.

Since the integrated loglikelihood (2) does not have a closed form expression, following Breslow and Clayton (1993) and Ripatti and Palmgren (2000), we write (2) as $\int exp\{-\boldsymbol{S}(\boldsymbol{b})\}d\boldsymbol{b}$, and apply the Laplace approximation to (2). Using calculations similar to expression (3) and profiling out the baseline hazard in a similar way to Appendix B of Ripatti and Palmgren (2000), one can show that, given $(\boldsymbol{\tau}, \boldsymbol{\nu})$, smoothing spline estimators of the $\theta_j(\cdot)$ and the parametric regression coefficient $\boldsymbol{\gamma}$ can be obtained by jointly maximizing following equation with respect to $\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{b}$.

$$\begin{aligned}
&\ell_D(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{p_1}, \boldsymbol{\gamma}, \boldsymbol{b}; \boldsymbol{\nu}, \boldsymbol{\tau}) \approx \\
&\sum_{i=1}^{n} \delta_i \left[\left\{ \sum_{j=1}^{p_1} \boldsymbol{N_{ij}^T \theta_j} + \boldsymbol{w_i^T \gamma} + \boldsymbol{z_i}^T \boldsymbol{b} \right\} - log \left\{ \sum_{l \in R(t_i)} e^{\sum_{j=1}^{p_1} \boldsymbol{N}_{lj} \boldsymbol{\theta}_j + \boldsymbol{w_i^T \gamma} + \boldsymbol{z_l}^T \boldsymbol{b}} \right\} \right] \\
&-\frac{1}{2} \boldsymbol{b}^T \boldsymbol{D}^{-1} \boldsymbol{b} - \sum_{j=1}^{p_1} \frac{1}{2\tau_j} \boldsymbol{\theta}_j^T \boldsymbol{K}_j \boldsymbol{\theta}_j.
\end{aligned} \tag{6}$$

Given $\boldsymbol{\nu}, \boldsymbol{\tau}$, the DPPL is a partial likelihood with quadratic penalty terms which is concave and has a unique maximizer (up to a constant) as a function of the $\boldsymbol{\theta_j}$ and $\boldsymbol{\gamma}$. Given $\widehat{\boldsymbol{\theta}}_j$ estimated at the design points, the smoothing spline estimators of $\theta_j(x)$ at any point x can be obtained by interpolating the splines using these estimated function values (see Section 2.4.1 of Green and Silverman 1994).

3.2 *The Frailty Model Representation*

In this subsection, we discuss the connection between the DPPL (6) for model (1) and the penalized likelihood of traditional frailty models with parametric covariate functions. Following a similar spirit of Lin and Zhang (1999), one can show that there is a one-to-one transformation between $\boldsymbol{\theta}_j$ and $(\beta_j, \boldsymbol{a}_j^T)^T$, i.e., $\boldsymbol{\theta}_j = \boldsymbol{x}_j^0\beta_j + \boldsymbol{B}_j^T\boldsymbol{a}_j$. Here $\boldsymbol{x}_j^0 = (x_{1j}^0 - \bar{x}_j^0, \cdots, x_{r_jj}^0 - \bar{x}_j^0)^T$, $\bar{x}_j^0 = \sum_{i=1}^{r_j} x_{ij}^0/r_j$, $\boldsymbol{B}_j = \boldsymbol{L}_j(\boldsymbol{L}_j^T\boldsymbol{L}_j)^{-1}$, and $\boldsymbol{L}_j\boldsymbol{L}_j^T = \boldsymbol{K}_j$ (see Green, P. J. 1987 for detail). One therefore can easily show that this new parameterization yields $\boldsymbol{\theta}_j^T\boldsymbol{K}_j\boldsymbol{\theta}_j = \boldsymbol{a}_j^T\boldsymbol{a}_j$.

Applying this transformation to (6), we see that the DPPL is equal to

$$\begin{aligned}
&\ell_D(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\nu}, \boldsymbol{\tau}) = \\
&\sum_{i=1}^{n} \delta_i \Big[\sum_{j=1}^{p_1} \boldsymbol{N}_{ij}^T(\boldsymbol{x}_j^0\beta_j + \boldsymbol{B}_j\boldsymbol{a}_j) + \boldsymbol{w}_i^T\boldsymbol{\gamma} + \boldsymbol{z}_i^T\boldsymbol{b} \\
&-log \left\{ \sum_{l \in R(t_i)} e^{\sum_{j=1}^{p_1} \boldsymbol{N}_{lj}(\mathbf{x}_j^0\beta_j + \boldsymbol{B}_j\boldsymbol{a}_j) + \boldsymbol{w}_l^T\boldsymbol{\gamma} + \boldsymbol{z}_l^T\boldsymbol{b}} \right\} \Big] \\
&\quad -\frac{1}{2}\boldsymbol{b}^T\boldsymbol{D}^{-1}\boldsymbol{b} - \frac{1}{2}\boldsymbol{a}^T\boldsymbol{\Psi}^{-1}\boldsymbol{a}, \qquad (7)
\end{aligned}$$

where $\boldsymbol{a} = (\boldsymbol{a}_1^T, \cdots, \boldsymbol{a}_p^T)^T$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$, and $\boldsymbol{\Psi} = diag(\tau_1\mathbf{I}_{r_1\times r_1}, \cdots, \tau_p I_{r_p\times r_p})$.

A comparison of (7) with equation (4) in Ripatti et al. (2000) shows that the two penalized likelihoods take the same form. Hence the DPPL smoothing spline estimators of $\{\theta_1(\cdot), \cdots, \theta_p(\cdot)\}$ and the regression coefficients $\boldsymbol{\gamma}$ of the semiparametric frailty model (1) can be obtained by fitting the following augmented working frailty model with parametric covariate effects using the penalized likelihood approach of Ripatti, et al. (2000),

$$\lambda_i(t; \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{a}_i, \boldsymbol{z}_i, \boldsymbol{b}) = \lambda_0(t) exp \left\{ \sum_{j=1}^{p_1} \boldsymbol{N}_{ij}\boldsymbol{x}_j^0\beta_j + \boldsymbol{w}_i^T\boldsymbol{\gamma} + \sum_{j=1}^{p_1} \boldsymbol{N}_{ij}\boldsymbol{B}_j\boldsymbol{a}_j + \boldsymbol{z}_i^T\boldsymbol{b} \right\}, \qquad (8)$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of fixed effects, and $\boldsymbol{a}$ and $\mathbf{b}$ are independent random effects distributed as $MVN(0, \boldsymbol{\Psi})$ and $MVN(0, \boldsymbol{D}(\boldsymbol{\nu}))$, respectively.

After obtaining the fixed and random effect estimators $\widehat{\boldsymbol{\gamma}}$, $\widehat{\beta}_j$, and $\widehat{\mathbf{a}}_\mathbf{j}$ by maximizing , the

DPPL smoothing spline estimators of $\boldsymbol{\theta_j}$ can be calculated as $\widehat{\boldsymbol{\theta}}_{\boldsymbol{j}} = \boldsymbol{x}_j^0\widehat{\beta}_j + \boldsymbol{B}_j\widehat{\boldsymbol{a}}_j$, which is a linear combination of fixed and random effect estimators. For inference, we need to estimate the variances of $\widehat{\boldsymbol{\theta}}_{\boldsymbol{j}}$ and $\widehat{\boldsymbol{\gamma}}$. Ripatti et al. (2000) proposed using the inverse of the minus second partial derivative of penalized partial likelihood to estimate covariance matrix of coefficient estimators. Their simulation results showed that the estimators works well. To make inference of $\widehat{\theta}_j$, we take a similar approach. The covariance of $\widehat{\boldsymbol{\theta}}_j$ can then be estimated by the transformation, $\widehat{COV}(\widehat{\boldsymbol{\theta}}_j) = (\boldsymbol{x_j^0}, \boldsymbol{B_j})\widehat{COV}(\widehat{\beta}_j, \widehat{\boldsymbol{a}}_{\boldsymbol{j}}{}^T)(\boldsymbol{x_j^0}, \boldsymbol{B_j})^T$.

3.3 *Frailty Models with Stratified Baseline Hazards*

In the context of the STI data example, infections with different organisms tend to have distinct population prevalence rates and natural histories. We therefore extend model (1) to accommodate stratified baseline hazards as follows

$$\begin{aligned}&\lambda_j(t; \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{z}_i, \boldsymbol{b}) = \lambda_{0j}(t)\\ &\times exp\{\theta_1(x_{i1}) + \cdots + \theta_p(x_{ip_1}) + \boldsymbol{w_i^T\gamma} + \boldsymbol{z}_{ij}^T b\},\end{aligned} \tag{9}$$

where $\lambda_{0j}(t)$ is the baseline hazard for the jth stratum (organism), $j = 1, \cdots, J$. One can write a DPPL similar to (6) with different at risk sets R_{ij}=$\{$All observations with respect to jth organism at risk at time $t_{ij}\}$. Estimation and inference for model (9) then follow a scheme similar to those with a common baseline hazard as presented in Section 3.2.

4. Inference on Smoothing Parameters and Variance Components

We assume in Section 3 that the smoothing parameters $\boldsymbol{\tau}$ and the variance components $\boldsymbol{\nu}$ are known. In practice, they need to be estimated from the data. Motivated by the working parametric frailty model (8), we propose to treat the smoothing parameter τ_j ($j = 1, \cdots, p$) as extra variance components. We estimate $\boldsymbol{\tau}$ and $\boldsymbol{\nu}$ simultaneously as variance components by maximizing the profile likelihood (10) in the working parametric

frailty model (8) as follows:

$$l_D(\beta(\tau_{\mathbf{j}},\nu),\tau_{\mathbf{j}},\nu) = -\frac{1}{2}log|\widetilde{\mathbf{D}}(\theta)| - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{log}|\mathbf{S}''(\boldsymbol{a},\boldsymbol{b})|$$
$$-\frac{1}{2}(\boldsymbol{a},\boldsymbol{b})\widetilde{\mathbf{D}}(\tau_{\mathbf{j}},\nu)^{-\mathbf{1}}(\boldsymbol{a},\boldsymbol{b})^{\mathbf{T}}, \tag{10}$$

where $\widetilde{D} = \begin{pmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}(\nu) \end{pmatrix}$. Following Ripatti and Palgram (2000), The estimating equation of variance components can be derived and simplified from the profile likelihood as follows:

$$-\frac{1}{2}\Big[tr\Big(\widetilde{\mathbf{D}}^{-\mathbf{1}}\frac{\partial\widetilde{\boldsymbol{D}}}{\partial(\tau,\nu)}\Big) + \mathbf{tr}\Big(\mathbf{S}''(\widehat{\boldsymbol{a}},\widehat{\boldsymbol{b}})^{-\mathbf{1}}\frac{\partial\widetilde{\mathbf{D}}^{-\mathbf{1}}}{\partial(\tau,\nu)}\Big)$$
$$-(\widehat{\boldsymbol{a}},\widehat{\boldsymbol{b}})\widetilde{\mathbf{D}}^{-\mathbf{1}}\frac{\partial\widetilde{\mathbf{D}}}{\partial(\tau,\nu)}\widetilde{\mathbf{D}}^{-\mathbf{1}}(\widehat{\boldsymbol{a}},\widehat{\boldsymbol{b}})^{\mathbf{T}}\Big] = \mathbf{0}$$

These equations are in the same spirit as equation (8) of Ripatti and Palmgren (2000) and equation (6.21) of Duchateau et al. (2008) by noting that $\frac{\partial\widetilde{D}^{-1}}{\partial(\tau,\nu)} = \widetilde{\mathbf{D}}^{-\mathbf{1}}\frac{\partial\widetilde{\mathbf{D}}}{\partial(\tau,\nu)}\widetilde{\mathbf{D}}^{-\mathbf{1}}$.

The estimating equations can be simplified as:

$$\widehat{\tau_j} = \frac{\widehat{\boldsymbol{a}}_j'\widehat{\boldsymbol{a}}_j + tr\{[\mathbf{S}''(\widehat{\boldsymbol{a}},\widehat{\boldsymbol{b}})_{\boldsymbol{a}_j\boldsymbol{a}_j}]^{-1}\}}{r_j - 2} \tag{11}$$

and

$$\widehat{\nu} = \frac{\widehat{\boldsymbol{b}}'\widehat{\boldsymbol{b}} + tr\{[\mathbf{S}''(\widehat{\boldsymbol{a}},\widehat{\boldsymbol{b}})_{\boldsymbol{bb}}]^{-1}\}}{n_c}, \tag{12}$$

where $\widehat{\boldsymbol{a}}_j$ is the maximizer of (7); $\boldsymbol{S}''(\widehat{\boldsymbol{a}},\widehat{\boldsymbol{b}})_{\boldsymbol{a}_j\boldsymbol{a}_j}$ is the $(r_j-2)\times(r_j-2)$ diagonal block submatrix of $\mathbf{S}''(\widehat{\boldsymbol{a}},\widehat{\boldsymbol{b}})$ corresponding to $\boldsymbol{a}_j$; and $tr\{\}$ is the trace value of the matrix, and n_c is the number of clusters. Here $\mathbf{S}''(\widehat{\boldsymbol{a}},\widehat{\boldsymbol{b}})$ is calculated by plugging in the corresponding estimators. The estimated value $\widehat{\boldsymbol{a}},\widehat{\boldsymbol{b}},\boldsymbol{D}$ are all functions of τ,ν. The estimates of τ,ν are obtained by iterating equations (11) and (12). The general expression of $\mathbf{S}''(\cdot)$ is given in the Appendix I. The baseline hazard is estimated using Breslow's estimator. An alternative to the $\mathbf{S}''$ is using $\widetilde{\mathbf{S}}''$, the version profiling out the the baseline hazard function.

The variance of $\widehat{\nu}$ for a shared frailty model can be estimated by using

$$var(\widehat{\nu}) = 2\widehat{\nu}^2\Big[n_c + \frac{1}{\widehat{\nu}^2}tr\{[\mathbf{S}''(\widehat{\boldsymbol{a}},\widehat{b})_{bb}]^{-1}[\mathbf{S}''(\widehat{\boldsymbol{a}},\widehat{\boldsymbol{b}})_{\boldsymbol{bb}}]^{-1}\}$$
$$-\frac{2}{\widehat{\nu}}tr\{[\mathbf{S}''(\widehat{\boldsymbol{a}},\widehat{\boldsymbol{b}})_{\boldsymbol{bb}}]^{-1}\}\Big]^{-1}$$

Note that this estimating equation is derived from the Fisher information matrix for τ, ν and has already accounted for the variability due to the estimation of τ. Appendix II summarizes the estimation procedure for all the model components. Sample code for implementing the procedure, along with a sample data set, have been posted on www.biostat.iupui.edu/yuz/Code. The code can be adapted to conveniently fit the semi-parametric additive models for data organized in the manner described in the introduction.txt file and the main function file dpplest.R.

5. Simulation Studies

In this section, we evaluate finite sample performance of the proposed method through a simulation study. We consider the following hazard model

$$\lambda(t_{ij}; x_1, x_2, w_1, b_i) = exp\{\theta_1(x_1) + \theta_2(x_2) + w_1\gamma + b_i\}, \tag{13}$$

where

$$\theta_1(x) = \{2 \times beta(x/10, 8, 8) + beta(x/10, 5, 5)\}/9,$$

$$\theta_2(x) = \{6 \times beta(x/10, 30, 17) + 4 \times beta(x/10, 3, 11)\}/40,$$

and $beta(\cdot)$ is the density function of the BETA distribution, and the b_i are independent random effects following $N(0, \nu)$.

The true value of regression coefficient γ was set as 0.5. Covariate w_1 was generated as a binary random variable with an equal probability to be 0 or 1, x_1 was generated as a cluster-level covariate taking values from 100 equally spaced knots in $[0, 10]$, $x_{ij1} =$

$(i \mod 100)/10$, and x_2 was generated as

$$x_{2ij} = trun\{(i+5)/6\}/10 + 10/5 \times (j-1),$$

where $trun\{\cdot\}$ denotes the truncation operator. The variance component was set as $\nu = 0.25, 0.5$. Censoring time followed an exponential distribution with rate 0.4 and the maximum followup times was set to be 5. The censoring percentage was about 18%. One hundred data sets were generated. In each data set, the number of clusters was 120 with 5 observations per cluster. For each simulated data set, we applied the DPPL method to estimate $\{\theta_1(x_1), \theta_2(x_2), \gamma, \nu\}$.

[Table 1 about here.]

Table 1 gives the averages of the estimates of the regression coefficient γ and the variance component ν. For both settings of $\nu = 0.25$ and $\nu = 0.5$, the estimates of γ are very close to the true values. The estimated standard errors of $\widehat{\gamma}$ are close to their empirical counterparts. The estimates of the variance component ν are also very close to the true value. Their SE estimates are slightly smaller than the empirical ones.

[Figure 1 about here.]

Fig. 1 depicts the performance of the DPPL spline estimates of the two smooth functions $\theta_1(x_1)$ and $\theta_2(x_2)$ when the variance component $\nu = 0.5$. Figures 1 (a) and (d) show that the averages of the estimated DPPL smoothing spline estimates of $\theta_1(x_1)$ and $\theta_2(x_2)$ are close to the true values. There are some biases in the estimated curves when the curvatures of the functions are high, e.g. at the second peak of $\theta_2(x_2)$. Fig. 1(b) and (e) show that the estimated SEs of $\widehat{\theta}_1(x_1)$ and $\widehat{\theta}_2(x_2)$ are close to their empirical counterparts except for the regions where the curvatures of the functions are high. Figure 1(c) and (f) calculate the coverage probabilities of the point-wise estimated 95% confidence intervals of $\widehat{\theta}_1(x_1)$ and $\widehat{\theta}_2(x_2)$. They are close to the nominal value 95% except for the peaks. The results

are similar when $\nu = 0.25$ and are not shown here. Our simulation study shows that the proposed DPPL method works well in finite samples in estimating the nonparametric functions and the variance components.

We also perform simulations with a larger censoring percentage and a smaller number of events. In the simulation with about 66% censoring (200 events, results not shown), the bias of the parametric regression coefficients and the nonparametric function are only slightly larger than when there is 18% censoring. The empirical coverage probability for the parametric regression coefficient is close to 95%. The average coverage probability of the nonparametric function is 86.7% due to the somewhat larger bias at the peak of the curve resulted from a smaller number of events. Performance improves when the sample size increases.

6. Application to the Sexually Transmitted Infection Data

The primary mode of transmission of STI is sexual contact. In the United States, much of the disease burden associated with STI is on women and young people. For example, it is estimated that subjects between 15 and 24 years of age account for about half of new STI cases each year, although they only represent a quarter of the population (Weinstock, Berman, and Cates 2004). STI screening is motivated by the disproportionate morbidity among adolescent women, including pelvic inflammatory disease, ectopic pregnancy, tubal infertility, preterm birth, and increased susceptibility to human immunodeficiency virus infection (CDC 2007; Cates and Wasserheit 1991; Paavonen et al. 2008; and Fleming et al. 1999). Despite the consensus on the need of screening, potentially useful behavioral markers for selected screening have not been adequately assessed because of limited epidemiological data (USPSTF 2007). Among the behavioral markers, the most relevant to the screening practice is the number of sexual partners. An improved understanding

of the effect of the number of partners will help us to identify individuals who are at increased STI risk for selective screening.

A recent study provides an opportunity to examine the effect of sexual partners on the time between sexual debut and the first STI. Briefly, young women between ages 14 and 17 years, attending one of three adolescent medicine clinics were eligible for enrollment regardless of prior sexual activity and STI diagnosis. At enrollment, participants were interviewed for their lifetime and recent (past two months) sexual behaviors, as well as the age of sexual debut (first sex) and the number of sex partners. Cervical and vaginal specimens were collected by a research nurse practitioner for testing of *C trachomatis*, *N gonorrhoeae* and *T vaginalis*. Infected participants were treated at the visit or shortly thereafter. The same procedures were repeated every three months. Information about infections prior to study enrollment or outside of the study venues was extracted from participants' electronic medical records. Detailed study protocol was described elsewhere (Tu et al. 2009). If a participant was sexually active at enrollment, age of first sex was ascertained from enrollment interview; for those who became sexually active during the course of follow-up, age of first sex was determined from quarterly interviews.

There were 387 adolescent women enrolled into the study. The enrolled subjects were followed up to 8.2 years. On average, study participants reported 3 partners in their lifetime at the time of enrollment (median=2,; range 0-28). We focused here on the effects of age of sexual debut, number of unprotected sexual intercourse during the last two months, and the baseline life time number of sexual partners reported at enrollment. The outcome of interest is time from first coitus to the first infection with each of the three organisms *C. trachomatis*, *N. gonorrhoeae* and *T. vaginalis*. Among 387 women, 26% of participants were censored for *C. trachomatis* infection, 51% are censored for *N.*

gonorrhoeae infection, and 47% are censored for *T. vaginalis* infection by the end of follow up.

Times from first coitus to the initial STI with each organism within the same individual are correlated due to the common sexual behavior and physiological environment within the same subject. To model the correlation of the three types of infections of the same subject, we used a stratified frailty model with a common random intercept in (9) by assuming $z_{ij} = 1$ and b_is are identical and independently distributed as a normal distribution with mean zero and an unknown variance ν. Our preliminary study shows that the lifetime number of partners has a polynomial effect up to 4th order. Therefore, it is desirable to fit a model with a nonparametric effect of lifetime number of partners. To accommodate such a nonlinear effect, we introduced a nonparametric component for the lifetime number of sexual partners. Parametric effects were used for age at first intercourse, ethnicity (white), and number of unprotected sex events in past two months.

Specifically, we considered the following model:

$$\lambda_{ij}(t) = \lambda_{0j}(t)exp\{\theta(N\_PARTNERS) + \gamma_1 \times RACE$$
$$+\gamma_2 \times AGE + \gamma_3 \times N\_SEX + b_i\}, \tag{14}$$

where $N\_PARTNERS$ is the lifetime number of partners ascertained at enrollment, AGE is the self-reported age at the first sex, N_SEX is the number of unprotected sex in last two months, $\theta(\cdot)$ is an unknown smooth function and b_i is the subject-specific random effect following $N(0, \nu)$. Preliminary analysis shows a similar effect of race, age, and N_SEX on risk of STI with different organisms. We therefore assume the covariate effects are the same for the three types of infections for model simplicity. As demonstrated in Fig. 2, while there was a generally positive association between the number of partners and STI acquisition, the risk was highest for subjects with four to six partners, which appeared to be the range that the number of partners effect became significantly different from those without sex

partners. This intriguing observation, though previously not reported in literature, is perhaps not entirely surprising. For example, one could speculate that adolescent women with relatively fewer (less than four) partners had lower risk because they are usually younger and are seeing younger male partners, who are unlikely to be sources of infection pathogens. On the other hand, women who had a larger number of sexual partners (more than seven) are likely to be more mature and more cognizant of the STI risk thus using more prophylactics. This may explain the apparent lack of linear increase of STI risk in women with larger number of sexual partners. Additionally, we could not rule out the possibility that these women are more experienced in partner selection. As suggested by the associate editor, we also analyzed the data using an approach similar to that of Du and Ma (2010), using the function *sshzd* in the R package *gss*. Initially, we tried the model $\log \lambda_i(t, N\_PARTNERS, AGE, N\_SEX) = g(t, N\_PARTNERS, AGE, N\_SEX) + b_i$ with a fully nonparametric regression function $g(\cdot)$ and a random effect b_i, as in Du and Ma, but we had numerical difficulties fitting this model. Hence, in an effort to provide a best comparison between the Du-Ma's fully nonparametric approach and our semiparametric additive model approach, we fit a fully nonparametric model of the form $\log \lambda_i(t, N\_PARTNERS, AGE, N\_SEX) = g(t, N\_PARTNERS, AGE, N\_SEX)$ without the random effect b_i. Both analysis point to a generally positive association between number of partners and infection risk. And both models show an attenuation of infection risk when partner number is greater than five (plot not shown). Therefore, the more parsimonious semiparametric model has adequately captured the trend demonstrated by the full nonparametric model.

[Table 2 about here.]

The estimated effects of the covariates are summarized in Table 2. The STI infection risk for subjects with a higher number of unprotected sex in past two months was higher

than those with fewer unprotected sex events although the effect did not reach a level of statistical significance (p-value=0.89). White adolescents tended to have a lower (0.436) STI risk than black teens (p-value<0.001). Teens having later sexual debut at higher age had higher infection rates than others having sexual debut at younger age (p-value<0.001). Again, this seemingly paradoxical observation may reflect the STI risk level presented by the male partners. Although the current study is unable to fully disseminate the reasons behind the observations due to the lack of male partner data, the findings nonetheless highlight the complexity of the STI risk in adolescents and the need for a more careful examination of the behavioral markers for STI screening.

[Figure 2 about here.]

7. Discussion

In this paper, we proposed a semiparametric frailty model for examination of semiparametric covariate effects in correlated survival data. Nonparametric functions are estimated using smoothing splines. Since the observed likelihood involves integration with no closed form, the Laplace method is used to approximate the likelihood. We developed the DPPL method to estimate all of the model components, including the semiparametric covariate effects, the the variance components, and the smoothing parameter within a unified framework using an augmented working frailty model with parametric covariate effects. With the proposed approach, the smoothing spline estimators of the nonparametric functions are obtained as a linear combinations of fixed effects and random effects, and the smoothing parameters are treated as extra variance components.

Alternatively, one may use the Gaussian quadratures or the MCMC-type method to integrate out the cluster-level random effects to reduce bias in some parameters. However, these calculations are computationally intensive. In addition, one still has to estimate the

smoothing parameter. If REML is used to estimate the smoothing parameter, one also needs to deal with high dimensional integration resulted from the random effects associated with the smoothing spline estimator. And these numerical integration methods for the REML likelihood are not feasible. Additionally, inferences for model parameters using numerical integration remain challenging. A key advantage of the DPPL is that estimation and inference of all the model components are easily obtained within a parametric frailty model unified framework. Future research is needed to pursue these numerical integration methods and compare their performance with the DPPL method. For future research, we will consider bias correction similar to what was used by Lin and Breslow (1996) to improve the performance of the DPPL method.

For the STI application, our analysis showed a nonlinear effect of the number of partners on STI acquisition in adolescent women. This finding has painted a more nuanced picture of the partner effect: when the cumulative lifetime number of partners is relatively small, i.e., fewer than five, infection risk clearly increases with the number of partners; but when the number of partners is relatively large, infection risk no longer has proportional increase with the number of partners, possibly due to the increased prophylactic behaviors and immunological maturity in young women with more sexual experience (Ethier and Orr, 2007). As discussed in previous sections, while such a finding is not scientifically unexpected, the complexity of partner effect appears to point to the need of an algorithm to more effectively differentiate the STI risk for screening purposes. In search of meaningful STI screening indicators, the proposed semiparametric frailty model offers an indispensable statistical tool with the necessary modeling flexibility to assess the effects of potential screening variables.

Acknowledgement

This work is supported by National Institutes of Health grants RO1 HD42404 (Tu), RO1 HL095086 (Tu, Yu), R37 CA76404 (Lin, Yu), and P01 CA134294 (Lin, Yu).

References

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Statist. Assoc.*, **88**, 9-25.

Cai, J., Fan, J., Zhou, H. and Zhou, Y. (2007). Marginal hazard models with varying coefficient for multivariate failure time data *Ann. Statist.*, **35**, 324-354.

Cai, J., Fan, J., Jiang, J. and Zhou, H. (2008). Partially linear hazard regression model with varying coefficient for multivariate survival data *J. R. Statist. Soc. B*, **70**, 141-158.

Cates, W. Jr, Wasserheit, J. N. (1991) Genital chlamydial infections: epidemiology and reproductive sequelae. *Am J Obstet Gynecol.* **164**, 1771-1781.

CDC. Trends in reportable sexually transmitted diseases in the United States, 2006. Atlanta, GA: US Dept of Health and Human Services, Centers for Disease Control and Prevention; 12 2007.

Du, P. and Ma, S. (2010). Frailty model with spline estimated nonparametric hazard function. *Statistica Sinica*, **20**, 561-580.

Duchateau, L. and Janssen, P. (2004). Penalized partial likelihood for frailties and smoothing splines in time to first insemination models for dairy cows. *Biometrics*, **60**, 608-614.

Duchateau, L. and Janssen, P. (2008). *The frailty model.* Springer.

Ethier KA, and Orr DP. Behavioral interventions for prevention and control of STDs

among adolescents. In: Aral SO, Douglas JM, Lipshutz JA, eds. Behavioral Interventions for Prevention and Control of Sexually Transmitted Diseases. New York, NY: Springer US; 2007.

Fan, J., Gijbels, I., and King, M. (1997). Local likelihood and local partial likelihood in hazard regression. *Ann. Statist.*, **25**, 1661-1690.

Fleming, D. T., Wasserheit, J. N. (1999) From epidemiological synergy to public health policy and practice: the contribution of other sexually transmitted diseases to sexual transmission of HIV infection. *Sex Transm. Infect.*, **75**, 3-17.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *J. Am. Statist. Assoc.*, **87**, 942-951.

Gray, R. J. (1994). Spline-Based tests in survival analysis. *Biometrics*, **50**, 640-652.

Green, P. J. (1987). Penalized likelihood for General Semi-parametric Regression Models. *Intern. Statist. Rev.*, **55**, 245-260.

Green, P. J. and Silverman, B. W. (1994) Nonparametric regression and generalized linear models. *London: Chapman and hall.*

Hastie, T. and Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, **46**, 1005-1016.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed model by using smoothing splines. *J. R. Statist. Soc. B*, **61**, 381-400.

Lin, X. and Breslow, NE. (1996). Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion. *J. Am. Statist. Assoc.*, **91**, 1007-16.

McGilchrist, C. A. and Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics*, **47**, 461-466.

Meyers, D., Wolff, T., Gregory, K., Marion, L., Moyer, V., Nelson, H., Petitti, D., and Sawaya, G. F. 2008) US Preventive Service Task Force (USPSTF) recommendations for STI screening. *Am Fam Physician.* **77**, 819-824.

Murphy, S. (1995). Asymptotic theory for the frailty model. *Ann. of Statist.*, **23**, 182-198.

O'sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM J. Sci. Stat. Comput.*, **9**, 531-542.

Paavonen, J., Westrom, L., and Eschenbach, D. (2008) Pelvic Inflammatory Disease. In: Holmes KK SP, Stamm WE, Piot P, et al. , ed. *Sexually Transmitted Diseases.* 4th ed. New York, NY: McGraw-Hill; p. 1017-1050.

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Annals of Statistics*, **26**, 183-214.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, **56**, 1016-1022.

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). Penalized survival models and frailty. *J. of Comp. and Graph. Statist.*, **12**, 156-175.

Tibshirani, R. and Hastie, T. (1987). local likelihood estimation. *J. Am. Statist. Assoc.*, **82**, 559-567.

Tu, W., Batteiger, B. E., Wiehe, S., Ofner, S., Van Der Pol, B., Katz, B. P., Orr, D. P., and Fortenberry, J. D. (2009) Time from first intercourse to first sexually transmitted infection diagnoses among adolescent women. *Archives of Pediatrics and Adolescent Medicine.* **163** (12); 1106-1111.

Weinstock, H., Berman, S., and Cates, W., Jr. (2004), Sexually transmitted diseases

among American youth: Incidence and prevalence estimates, 2000, *Perspectives on Sexual and Reproductive Health*, **36**, 610.

US Preventive Services Task Force. (2007) Screening for chlamydial infection: US Preventive Services Task Force recommendation statement. *Ann Intern Med.* **147**(2):128-134.

Yu, Z. and Lin, X. (2008) Nonparametric regression using local kernel estimating equations for correlated failure time data. *Biometrika*, **95**, 123-137.

Appendix I: The Derivatives of the Integrant of the Integrated Likelihood of the General Frailty Model

Consider the general frailty model $\lambda_i(t) = \lambda_0(t)exp\{\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{w_i^T\gamma} + \widetilde{\boldsymbol{z}}_i^T\widetilde{\boldsymbol{b}}\}$, where $\boldsymbol{\beta}$ is a vector of regression coefficients and and $\widetilde{\boldsymbol{z}}_i$ is the covariate vector associated with the random effect $\widetilde{\boldsymbol{b}}$, $\widetilde{\boldsymbol{b}}$ is a vector of random effects following $N(0, \widetilde{\boldsymbol{D}})$ which may include $\boldsymbol{b}, \boldsymbol{a}$ for the additive frailty model (2). Write the integrated loglikelihood as $\int exp\{-S(\widetilde{\boldsymbol{b}})\}d\widetilde{\boldsymbol{b}}$, where

$$\mathbf{S}(\widetilde{\boldsymbol{b}}) = \sum_{\mathbf{i=1}}^{\mathbf{n}} \Big[-\delta_{\mathbf{i}}\{\mathbf{log}(\lambda_{\mathbf{0}}(\mathbf{u_i})) + \sum_{\mathbf{j=1}}^{\mathbf{p}} \boldsymbol{x}_{\mathbf{ij}}^{\mathbf{T}}\boldsymbol{\beta} + \boldsymbol{w_i^T\gamma} + \widetilde{\boldsymbol{z}}_{\mathbf{i}}^{\mathbf{T}}\widetilde{\boldsymbol{b}}\}$$
$$+\Lambda_0(u_i)e^{\sum_{j=1}^p \boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\boldsymbol{w_i^T\gamma}+\widetilde{\boldsymbol{z}}_i^T\widetilde{\boldsymbol{b}}}\Big] + \frac{1}{2}\widetilde{\boldsymbol{b}}^T\widetilde{\boldsymbol{D}}^{-1}\widetilde{\boldsymbol{b}}$$

$$\mathbf{S}'(\widetilde{\boldsymbol{b}}) = \sum_{i=1}^{n} \Big[-\delta_i\widetilde{\boldsymbol{z}}_i + \Lambda_0(u_i)e^{\boldsymbol{x}_i^T\boldsymbol{\beta}+\boldsymbol{w_i^T\gamma}+\widetilde{\boldsymbol{z}}_i^T\widetilde{\boldsymbol{b}}}\widetilde{\boldsymbol{z}}_i\Big] + \widetilde{\boldsymbol{D}}^{-1}\widetilde{\boldsymbol{b}}$$

$$\mathbf{S}''(\widetilde{\boldsymbol{b}}) = \sum_{i=1}^{n} \Lambda_0(u_i)e^{\sum_{j=1}^p \boldsymbol{x}_{ij}\boldsymbol{\beta}+\boldsymbol{w_i^T\gamma}+\widetilde{\boldsymbol{z}}_i^T\widetilde{\boldsymbol{b}}}\widetilde{\boldsymbol{z}}_i\widetilde{\boldsymbol{z}}_i^T + \widetilde{\boldsymbol{D}}^{-1}$$

Appendix II: Algorithm for estimation

- Generate the new covariates $\boldsymbol{x_j^0}$, $\boldsymbol{B_j}$, and indicator matrix $\boldsymbol{N_j}$ for fitting the augment working frailty model (8).

- For fixed variance component $\boldsymbol{\nu}$ and smoothing parameters $\boldsymbol{\tau}$, maximize the (7) with respect to $\beta_j, \boldsymbol{a_j}, \boldsymbol{\gamma}, \boldsymbol{b}$ using a Newton-Raphson algorithm.
- For estimate $\beta_j, \boldsymbol{a_j}, \boldsymbol{\gamma}, \boldsymbol{b}$ obtained from step (2), calculate the smoothing parameters and variance components using equation (11,12).
- Iterate between step (2) and (3) until convergence.

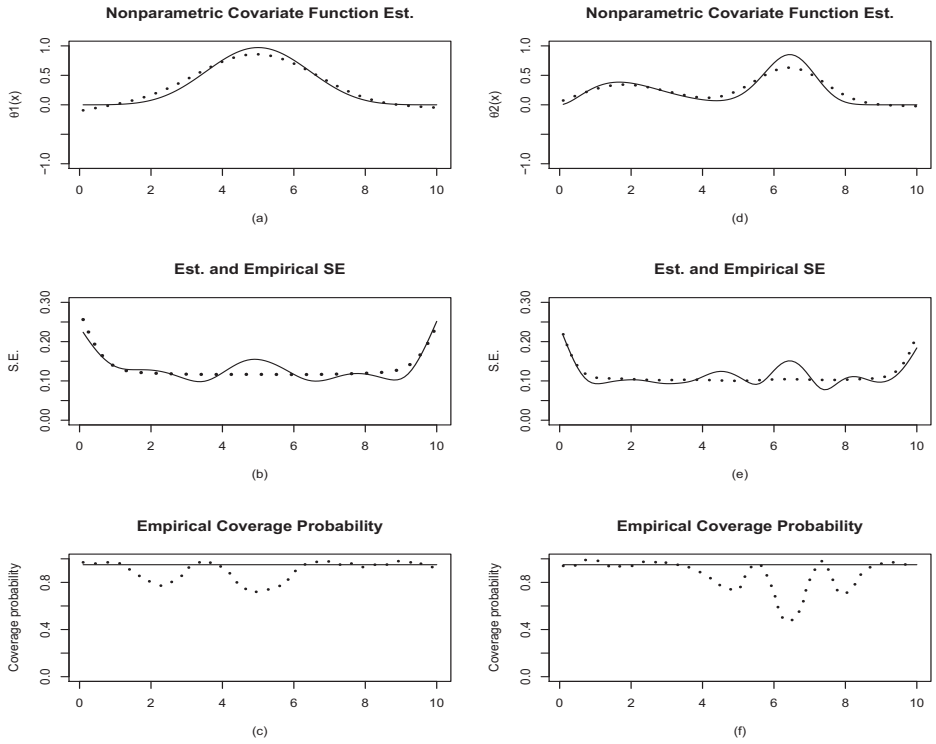

Figure 1. Simulation results for the nonparametric covariate function $\theta_1(x)$ and $\theta_2(x)$: (a) Average of the DPPL spline estimates of $\widehat{\theta}_1(x)$, dotted and true $\theta_1(x)$, solid; (b) Standard errors: estimated, dotted and empirical, solid; (c) Empirical coverage probability: the mean coverage probability is 0.901; (d) Average of the DPPL spline estimates of $\widehat{\theta}_2(x)$, dotted and true $\theta_2(x)$, solid, (e) Standard errors of $\theta_2(x)$: estimated, dotted and empirical, solid; (f) Empirical coverage probability of $\theta_2(x)$: the mean coverage probability is 0.875.

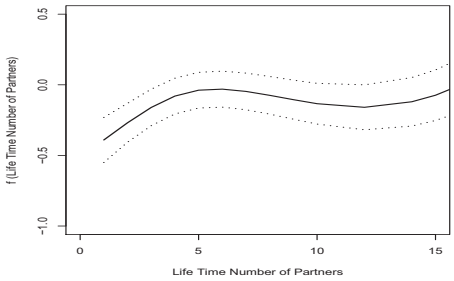

Figure 2. Smoothing spline estimate of the number of lifetime partner effect using model (14) in the sexually transmitted infection study

Table 1
Estimators of regression coefficient coefficient and variance component in the simulation studies based on 100 runs

Parm.	True	Average	Est. SE	Emp. SE	95% CP
			$\nu = 0.25$		
γ	0.5	0.5081	0.1016	0.0923	98%
ν	0.25	0.2445	0.0648	0.0757	90%
			$\nu = 0.5$		
γ	0.5	0.5057	0.1052	0.1003	97%
ν	0.5	0.4884	0.0973	0.1129	93%

Parm: parameter
CP: coverage probability
Average: average of the estimates
Est. SE: Estimated SE
Emp. SE: Monte-Carlo estimated SE

Table 2

Regression Coefficient estimates from fitting the semiparametric frailty model (Time to sexually transmitted infection study

Covariate	Race (white)	Age	N_Sex
Estimate	-0.831	0.326	0.0008
SE	(0.240)	(0.047)	(0.0057)
p-value	<0.001	<0.001	0.89

Age: Age at first sex.