Published in final edited form as:

J Am Stat Assoc. 2009 September 1; 104(487): 1144–1154. doi:10.1198/jasa.2009.tm08160.

Cox Models With Smooth Functional Effect of Covariates Measured With Error

Yu-Jen Cheng, Ph.D. and

Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205
(ycheng3@jhsph.edu).

Ciprian M. Crainiceanu[Associate Professor]

Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205
(ccrainic@jhsph.edu).

Abstract

We propose, develop, and implement a fully Bayesian inferential approach for the Cox model when the log hazard function contains unknown smooth functions of the variables measured with error. Our approach is to model nonparametrically both the log-baseline hazard and the smooth components of the log-hazard functions using low-rank penalized splines. Careful implementation of the Bayesian inferential machinery is shown to produce remarkably better results than the naive approach. Our methodology was motivated by and applied to the study of progression time to chronic kidney disease as a function of baseline kidney function and applied to the Atherosclerosis Risk in Communities study, a large epidemiological cohort study. This article has supplementary material online.

Keywords

Measurement error; Smoothing; Survival analysis

1. INTRODUCTION

Survival analysis has developed from the analysis of life tables in actuarial sciences and has enjoyed remarkable success with modern applications in medicine, epidemiology, and the social sciences. The popularity of survival analysis models is probably surpassed only by the popularity of standard linear regression models.

Survival data are the product of a continuous death process coupled with a censoring mechanism. Typically, the death rate depends on a number of factors and time to death is

© 2009 American Statistical Association

SUPPLEMENTAL MATERIALS

Supplemental Materials.pdf: Supplemental materials pdf file for this article.

Supplemental Materials.tex: Supplemental materials latex file for this article.

MCMC.pdf: Figure 1 in the supplemental materials.

multibeta1_3.pdf: Figure 2 in the supplemental materials.

multibeta4_6.pdf: Figure 3 in the supplemental materials.

SIMEX_4.pdf: Figure 4 in the supplemental materials.

boxplotSIMEX.pdf: Figure 5 in the supplemental materials.

SIMEX50.pdf: Figure 6 in the supplemental materials.

SIMEXZ50.pdf: Figure 7 in the supplemental materials.

ECA_jasa.bst, jasa_harvard.sty, and JASA_manu.sty are files for helping to run Supplemental Materials.tex.

only partially observed for those subjects with censored observations. Standard analyses of survival data assume that all covariates affecting survival rates are observed without error. However, in many applications some of the covariates are subject to measurement error or are available without error only for a subsample of the population.

Substantial methodological and applied research has been dedicated in recent years to survival analysis with covariates subject to measurement error, starting with the seminal paper by Prentice (1982). The regression calibration approach was expanded and refined by Pepe, Self, and Prentice (1989) and Wang et al. (1997). Clayton (1991) used regression calibration within risk sets thus avoiding the rare disease assumption. For data containing a validation sample, Zhou and Pepe (1995) and Zhou and Wang (2000) proposed nonparametric estimators of the induced hazard function. For data with at least two replicates Huang and Wang (2000) proposed a consistent nonparametric estimator based on a modification of the partial likelihood score equation. Augustin (2004) showed that the likelihood adjustment methods developed by Nakamura (1992) can be applied to the Breslow likelihood to provide a corrected likelihood. This result circumvented the impossibility result derived by Stefanski (1989) for the partial likelihood. Hu, Tsiatis, and Davidian (1998) proposed likelihood maximization algorithms for parametric and nonparametric specifications of the distribution of the unobserved variables. Greene and Cai (2004) established the asymptotic properties of the SIMEX estimators for models with measurement error and multivariate failure time data. Hu and Lin (2004) introduced a modified score equation and established the asymptotic properties of the estimators for multivariate failure time data. Li and Lin (2003) used the a combination of the EM and SIMEX algorithms to provide maximum likelihood estimators for frailty models with variables observed with error. Song and Huang (2005) compare the conditional score estimation of Tsiatis and Davidian (2001) with Nakamura's (1992) parametric adjustment.

Current statistical approaches for Cox regression with covariates measured with error are not designed for unknown exposure/outcome functions. Addressing these limitations is highly relevant both from a methodological and a practical perspective, as we illustrate with our application to the Atherosclerosis Risk in Communities (ARIC) study in Section 5. To address these problems we propose a new inferential method that models nonparametrically the baseline hazard and the exposure/outcome functions using log penalized splines. We assume that the exposure variable is observed with error with a known measurement error distribution and use a fully Bayesian approach to inference. We show that the resulting likelihood is log-concave which leads to important computational advantages that make the implementation tractable for large data sets. To ensure reproducibility of our results we posted our Bayesian software developed in R and additional simulation results at www.biostat.jhsph.edu/~ccrainic/webpage/software/Surv_ME.zip.

Our proposed methodology was motivated by the analysis of time to event data from the ARIC study. ARIC is a large multi-purpose epidemiological study conducted in four U.S. communities (Forsyth County, NC; suburban Minneapolis, MN; Washington County, MD; and Jackson, MS). A detailed description of the ARIC study is provided by Investigators (1989). In short, from 1987 through 1989, 15,792 male and female volunteers aged 45 through 64 were recruited from these communities for a baseline and three subsequent visits. Time to event data is observed continuously for multiple end points, but we focus here on the event *detection of primary chronic kidney disease* (CKD), the least severe phase of CKD. For the purpose of this study all primary CKD events up to January 1, 2003 were included and the time-to-event data was obtained from annual participant interviews and review of local hospital discharge lists and county death certificates.

The relationship between various risk factors, such as race, age, or sex, and progression time to primary CKD may be confounded by a series of baseline confounders. An important confounder is the baseline kidney function as measured by the glomerular filtration rate (GFR). Because GFR can only be obtained through a long and awkward procedure, the estimated GFR (eGFR) is used in practice. eGFR is obtained from a prediction equation based on creatinine, sex, gender, and age and is subject to regression and biological measurement error. Primary CKD is defined as either achievement of follow-up eGFR < 60 ml/min/1.73 m² or a post-baseline CKD hospitalization or death (Hsu et al. 2005; Kottgen et al. 2007). As is the case in many biological applications, subjects with lower-baseline GFR (higher risk) are expected to progress faster towards primary CKD (clinical endpoint), but the effect of GFR on the risk of CKD is likely to be nonlinear. Given the measurement error in eGFR, estimating the exposure/outcome curve is a difficult inferential problem, while its result is of interest to nephrologists, internists, and researchers (Stevens et al. 2006).

The paper is organized as follows. Section 2 describes our proposed modeling framework. Section 3 presents the calculation of full conditional distributions and introduces a refined computational approach that ultimately leads to successful inference. Section 4 provides simulation results indicating the large statistical advantages of our method over a naive approach. Section 5 provides our results and their interpretation for the ARIC study. Section 6 provides the summary and discussions.

2. MODELING FRAMEWORK

In this paper we follow the notations in Miller (1998). Let T_i and C_i be the failure and censoring times for subject i , $i = 1, \dots, n$, respectively, $\delta_i = I(T_i < C_i)$ be the noncensoring indicator, and $Y_i = \min(T_i, C_i)$ be the time of failure or censoring for subject i . We assume that the survival time, T , depends on the covariates X and $\mathbf{Z}$, where X is a scalar and $\mathbf{Z}$ is a vector. We also assume that we only measure a proxy, W , of X , that W is contaminated by classical measurement error (Carroll et al. 2006), and that $\mathbf{Z}$ is measured without error. Our aim is to estimate the exposure/outcome function adjusted for covariates, $\mathbf{Z}$, where the exposure is the unobserved covariate X and the outcome is the partially observed survival time, T .

Such data structures are common in epidemiological studies. For example, in the ARIC study we are interested in estimating the adjusted exposure/outcome function where the outcome, Y , is time to primary CKD and the exposure, X , is the true GFR of the kidney. The GFR is not directly measured and the eGFR, W , is used in practice. The eGFR is predicted from the following equation (Hsu et al. 2005):

$$\text{eGFR} = 186.3 * (\text{Serum Creatinine})^{-1.154} * (\text{Age})^{-0.203} \\ * (0.742)^{I(\text{Female})} * (1.21)^{I(\text{African American})}.$$

Thus, the eGFR measurement contains at least two nonignorable sources of error: (1) the biological variability unaccounted for by the prediction equation; and (2) the laboratory variability associated with urine serum creatinine. To assess the variability of eGFR a replication study was conducted in the Third National Health and Nutrition Examination Survey (NHANES III). Duplicate eGFR measurements were obtained for each of 513 participants aged 45 to 64 with eGFR ≥ 60 from two visits at a median of 17 days apart (Coresh et al. 2002). We assumed a classical measurement error model for log eGFR and calculated the measurement error variance as

$$\widehat{\sigma}_u^2 = \frac{1}{2} \sum_{i=1}^{513} (W_{i1} - W_{i2})^2,$$

where W_{ij} is the observed log eGFR for subject i at visit j . The estimated measurement error variance was $\widehat{\sigma}_u^2 = 0.00688$ corresponding to a reliability of 0.85 in the ARIC data set and will be treated as a constant in our subsequent analyses. In the remainder of the paper we use the log transformed eGFR instead of eGFR.

Our proposed model will have to be complex enough to account for the complexity of the data structure while remaining computationally tractable. To achieve this we design an overall model with four simple components: the outcome model, the measurement error model, the smoothing assumptions and the hyperpriors. We present each of these components in one subsection as follows.

2.1 The Outcome Model

The Cox proportional hazard model (Cox 1972) assumes that the hazard function of the failure time, T , given covariates X and Z is

$$\lambda(t|x, z) = \lambda_0(t) \exp \{f(x; \beta_x) + \mathbf{z}^T \beta_z\}, \quad (1)$$

where $\lambda_0(t)$ is an unspecified function and $f(x; \beta_x) = x\beta_{x1}$, where β_{x1} is a scalar component of β_x . Traditionally, inference for Cox models when x is observed without error has been based on the partial likelihood (Cox 1975). Moreover, spline-based estimation of the function $f(x; \beta_x)$ has been discussed by several authors. For example, Sleeper and Harrington (1990) proposed a regression spline approach. Hastie and Tibshirani (1990) and Gray (1992) considered a penalized partial likelihood approach with a spline with a small number of knots but did not model the baseline hazard function. However, none of these approaches considered covariates measured with error. Recently, some extensions based on the full likelihood that are relevant to our problem have been proposed. Using a frequentist perspective, Li and Ryan (2006) proposed an imputation-based approach for fitting proportional hazard models with covariates measured with error for the case of a linear function $f(x; \beta_x) = x\beta_x$. For the same model Tadesse et al. (2005) proposed a Bayesian inferential approach. Hennerfeind, Brezger, and Fahrmeir (2006) proposed a geoadditive survival model that treats $f(x; \beta_x)$ as a smooth unknown function of x , which is observed without measurement error. Kauermann, Xu, and Vaida (2008) proposed a stacked Laplace-EM algorithm to estimate the baseline hazard and the smooth effects in the case when frailties are also included. In this article, we model both $\lambda_0(t)$ and $f(x; \beta_x)$ as smooth functions and assume that the covariate X is subject to classical measurement error. To simplify the presentation we reparameterize the baseline hazard, $\lambda_0(t)$ as $\lambda_0(t) = \exp\{g(t; \gamma)\}$, where γ is a vector of parameters associated with the baseline hazard function. Then Equation (1) can be rewritten as

$$\lambda(t|x, z) = \exp \{g(t; \gamma) + f(x; \beta_x) + \mathbf{z}^T \beta_z\}. \quad (2)$$

We make the following assumptions throughout the remainder of the paper:

(A1) *Noninformative censoring*: the failure time, T , and the censoring time, C , are independent given the covariates X and Z , that is, $P(T, C|X, Z) = P(T|X, Z)P(C|X, Z)$.

(A2) *Nondifferential error*: the distribution of the outcome, (Y, δ) , depends only on the true covariates X and $\mathbf{Z}$, that is, $P(Y, \delta | X, W, \mathbf{Z}) = P(Y, \delta | X, \mathbf{Z})$.

If assumptions (A1) and (A2) hold then the likelihood function of model (2) is

$$L_1(\gamma, \beta_x, \beta_z) \propto \exp \left[\sum_{i=1}^n \delta_i \{g(y_i; \gamma) + f(x_i; \beta_x) + \mathbf{z}_i^T \beta_z\} - \exp \{f(x_i; \beta_x) + \mathbf{z}_i^T \beta_z\} \int_0^{y_i} \exp \{g(t; \gamma)\} dt \right]. \quad (3)$$

2.2 The Measurement Error Model

We assume the following classical measurement error model (Carroll et al. 2006) for the proxy measurements, W_i ,

$$W_i = X_i + U_i, \quad (4)$$

where $U_i | \mathbf{N}(0, \sigma_u^2)$, $X_i | \mathbf{N}(\mu_{x|z}, \sigma_x^2)$, $\mu_{x|z} = \alpha_0 + \mathbf{z}^T \alpha_z$, $\alpha = (\alpha_0, \alpha_z)^T | \mathbf{N}(0, \sigma_\alpha^2 \mathbf{I}_2)$, and U_i are assumed to be independent of X_i and $\mathbf{Z}_i$. In this paper we assume that the variance of the measurement error, σ_u^2 , is known. However, our methodology is not limited by this assumption and we could easily incorporate a model for a replication study on the entire dataset, a subsample of the data or a sample from a different dataset. This might be necessary in some applications where the estimation variability of σ_u^2 is an issue. However, in the ARIC application this is not a problem because σ_u^2 is estimated using data from a rigorous replication study on more than 500 subjects.

The likelihood function, L_2 , of the measurement model, (4), is

$$L_2(\mathbf{X}, \alpha, \sigma_x, \sigma_\alpha) \propto \frac{1}{\sigma_u^n \sigma_x^n \sigma_\alpha^n} \exp \left[-\frac{1}{2} \sum_{i=1}^n \left\{ \frac{(w_i - x_i)^2}{\sigma_u^2} + \frac{(x_i - \mu_{x|z})^2}{\sigma_x^2} + \frac{\alpha^T \alpha}{\sigma_\alpha^2} \right\} \right], \quad (5)$$

where $\mathbf{X} = (x_1, \dots, x_n)$, $\alpha = (\alpha_0, \alpha_z)$, and σ_x^2 is the variance of X conditional on Z , and σ_α^2 is the prior variance of the components of α , respectively.

2.3 Smoothing Assumptions

Our approach is to model flexibly both the baseline hazard and the exposure/outcome functions using log-penalized splines. While many alternatives were considered both in terms of spline basis and smoothness penalties, we ended up making a few choices that proved crucial to ensure convergence and good mixing properties of our MCMC simulations. We will point out these choices while keeping our presentation general.

We started with the following models:

$$g(t; \gamma) = \mathbf{B}_g^T(t) \gamma, \quad f(x; \beta_x) = \mathbf{B}_f^T(x) \beta_x, \quad (6)$$

where $\mathbf{B}_g^T(t) = \{B_1^g(t), \dots, B_{K_1}^g(t)\}$ and $\mathbf{B}_f^T(x) = \{B_1^f(x), \dots, B_{K_2}^f(x)\}$ are vectors of known basis functions evaluated at t and x , respectively, while $\gamma = (\gamma_1, \dots, \gamma_{K_1})$ and $\beta_x = (\beta_{x1}, \dots, \beta_{xK_2})$ are the corresponding vectors of regression parameters. The number of basis functions, K_1 and K_2 , control the maximum complexity of the functions.

Here we focus on B-spline basis functions because of their good numerical properties and assume that a number of knots were fixed. The simplest choice of B-spline basis is a piecewise constant function, that is, $B_k(x) = 1$ over the k th between-knot interval and $B_k(x) = 0$ elsewhere, as shown in the top panels of Figure 1.

We propose to use such a piecewise constant function to fit the log baseline hazard, $g(t; \gamma)$. This choice has a long and successful history in the statistical literature (e.g., Leonard 1978; Gamerman 1991; and Gray 1994). One advantage is that it avoids numerical integration and the cumulative baseline hazard has a closed form. More precisely, if $g(t; \gamma) = \gamma_k$ for

$t \in [\kappa_{k-1}^g, \kappa_k^g)$, where $\kappa_k^g, k=1, \dots, K_1, k=1, \dots, K_1$, are a set of knots, then $\int_0^y \exp \{g(t; \gamma)\} dt$ is equal to

$$\sum_{k: \kappa_k^g \leq y} e^{\gamma_k} (\kappa_k^g - \kappa_{k-1}^g) + e^{\gamma_{l+1}} (y - \kappa_l^g)$$

for $y \in [\kappa_l^g, \kappa_{l+1}^g]$. In contrast, we found that the particular choice of basis function for $f(x; \beta_x)$, $\mathbf{B}_f^T(x)$, is not as important from a computational perspective and we used a B-spline basis of degree 1. An example of B-spline basis functions of degree 1 is shown in the bottom panels of Figure 1.

To avoid overfitting the unknown functions $g(t; \gamma)$ and $f(x; \beta_x)$, we use a penalized approach. In a Bayesian framework the difference penalty introduced by Eilers and Marx (1996) for B-splines is equivalent to a random walk prior (Lang and Brezger 2004). In our setting, the first and second-order random walk priors are used to induce smoothing of $g(t; \gamma)$ and $f(x; \beta_x)$, respectively. More precisely, we use

$$\begin{aligned} [\gamma_1 | \sigma_{0\gamma}^2] &= N(0, \sigma_{0\gamma}^2), \\ [\gamma_k | \gamma_{k-1}, \sigma_{\gamma}^2] &= N(\gamma_{k-1}, \sigma_{\gamma}^2), k=2, \dots, K_1, \\ [\beta_{x1} | \sigma_{0\beta}^2] &= N(0, \sigma_{0\beta}^2), \\ [\beta_{x2} | \sigma_{0\beta}^2] &= N(0, \sigma_{0\beta}^2), \\ [\beta_{xk} | \beta_{x(k-1)}, \beta_{x(k-2)}, \sigma_{\beta_x}^2] &= N(2\beta_{x(k-1)} - \beta_{x(k-2)}, \sigma_{\beta_x}^2), k=3, \dots, K_2, \end{aligned} \quad (7)$$

where $[\cdot | \cdot]$ denotes the conditional distribution. The prior variances $\sigma_{0\gamma}^2, \sigma_{0\beta}^2$ are assumed to be very large to ensure that the prior distributions of γ_1, β_{x1} , and β_{x2} are noninformative. The variance components, σ_{γ}^2 and $\sigma_{\beta_x}^2$, are the smoothing parameters for the functions $g(t; \gamma)$ and $f(x; \beta_x)$, respectively, and are estimated from the data together with the other unknowns in the model. To better understand the effect of the smoothing parameters on functional estimation, note that $\sigma_{\gamma}^2=0$ corresponds to a constant γ_k , or a constant baseline hazard function. In contrast, $\sigma_{\gamma}^2=\infty$ corresponds to a piecewise constant function without any information borrowing across temporal segments. When the number of knots used is moderate or large this leads to a rough fit. By estimating σ_{γ}^2 our model will provide a compromise between these two extremes based on the laws of probability and data.

A good reference for the choice of prior distribution for the baseline hazard function can be found in Ibrahim, Chen, and Sinha (2001). From our review of the literature it seems that the choice we made is also the most popular choice; see Leonard (1978), Gamerman (1991), and Gray (1994). An alternative that seems equally promising, but not tried in this paper, was

proposed by Arjas and Gasbarra (1994) who specified a gamma prior distribution on the baseline hazard function and not on the log-baseline hazard.

The second-order random walk prior for degree 1 B-spline basis function is equivalent to penalize deviations from a linear trend. It also corresponds to an independent normal prior on the coefficients of the linear truncated power basis. The relationship between B-spline and truncated power basis function was described, for example, by Ruppert, Wand, and Carroll (2003).

2.4 Hyperprior Distributions

Our choices of hyperprior distributions are in line with standard choices. For example, we assumed that $[\alpha|\sigma_\alpha^2] = N(0, \sigma_\alpha^2 \mathbf{I}_2)$, where σ_α^2 is very large. For the variance components we assumed independent inverse gamma priors as follows:

$[\sigma_x^2|\rho_{1\mu}, \rho_{2\mu}] = \text{IG}(\rho_{1\mu}, \rho_{2\mu})$, $[\sigma_\gamma^2|\rho_{1\gamma}, \rho_{2\gamma}] = \text{IG}(\rho_{1\gamma}, \rho_{2\gamma})$, and $[\sigma_{\beta_x}^2|\rho_{1\beta}, \rho_{2\beta}] = \text{IG}(\rho_{1\beta}, \rho_{2\beta})$, where $\text{IG}(a, b)$ denotes the inverse of a Gamma distribution with mean a/b and variance a/b^2 . We chose $\rho_{1\mu}, \rho_{2\mu}, \rho_{1\gamma}, \rho_{2\gamma}, \rho_{1\beta}, \rho_{2\beta}$ to be very small to ensure that the priors are noninformative. For a detailed discussion of noninformative priors of smoothing parameters for penalized splines see Crainiceanu et al. (2007).

3. STATISTICAL INFERENCE

Let $\mathbf{D} = (\mathbf{Y}, \boldsymbol{\delta}, \mathbf{W}, \mathbf{Z})$ be the observed data, where $\mathbf{Y}$, $\boldsymbol{\delta}$, $\mathbf{W}$, and $\mathbf{Z}$ are vectors of minimum of failure and censoring time, noncensoring indicator, covariates measured with error, and covariates measured without error, respectively. Because the vector of true covariates, $\mathbf{X}$, is not directly observed it will be treated as a vector of unknown parameters.

If noninformative censoring (A1) and nondifferential error (A2) assumptions hold then the joint posterior distribution of the parameters given the data is

$$[\boldsymbol{\Omega}|\mathbf{D}] \propto [\mathbf{D}|\boldsymbol{\Omega}] [\mathbf{X}|\alpha, \sigma_x^2] [\gamma + \sigma_\gamma^2] [\beta_x|\sigma_{\beta_x}^2] [\beta_z] [\alpha] [\sigma_x^2] [\sigma_\gamma^2] [\sigma_{\beta_x}^2],$$

where $\boldsymbol{\Omega} = (\mathbf{X}, \gamma, \beta_x, \beta_z, \alpha, \sigma_x^2, \sigma_\gamma^2, \sigma_{\beta_x}^2)$ is the collection of all parameters, and the distribution $[\mathbf{D}|\boldsymbol{\Omega}]$ is the product of likelihood functions L_1 and L_2 .

We use Gibbs sampling (Geman and Geman 1984) to simulate $[\boldsymbol{\Omega}|\mathbf{D}]$. A particularly convenient partition of the parameter space and the associated full conditional distributions are described below.

$$[X_i|\mathbf{D}, \boldsymbol{\Omega}_{-X_i}] \propto \exp \left[\delta_i f(X_i; \beta_x) - \int_0^{\gamma_i} \exp \{g(t; \gamma)\} dt \times \exp \{f(x_i; \beta_x) + \mathbf{z}_i \beta_z\} \right] [X_i|\alpha, \sigma_x^2];$$

$$[\gamma_k|\mathbf{D}, \boldsymbol{\Omega}_{-\gamma_k}] \propto \exp \left[\gamma_k \sum_{i=1}^n \delta_i B_k^g(y_i) - e^{\gamma_k} \sum_{i=1}^n R(i, k) \times \exp \{f(x_i; \beta_x) + \mathbf{z}_i \beta_z\} \right] [\gamma_k|\gamma_{-k}, \sigma_\gamma^2];$$

$$[\beta_{xk}|\mathbf{D}, \boldsymbol{\Omega}_{-\beta_{xk}}] \propto \exp \left[\beta_{xk} \sum_{i=1}^n \delta_i B_k^f(x_i) - \sum_{i=1}^n \int_0^{\gamma_i} \exp \{g(t; \gamma)\} dt \times \exp \{f(x_i; \beta_x) + \mathbf{z}_i \beta_z\} \right] [\beta_{xk}|\beta_{-xk}, \sigma_{\beta_x}^2];$$

$$[\beta_{zk}|\mathbf{D}, \mathbf{\Omega}_{-\beta_{zk}}] \propto \exp \left[\beta_{zk} \sum_{i=1}^n \delta_i z_i - \sum_{i=1}^n \int_0^{\gamma_i} \exp \{g(t; \gamma)\} dt \times \exp \{f(x_i; \beta_x) + \mathbf{z}_i \beta_z\} \right] [\beta_z];$$

$$[\alpha|\mathbf{D}, \mathbf{\Omega}_{-\alpha}] \propto N \left\{ \left(H^T H + \sigma_x^2 / \sigma_\alpha^2 \mathbf{I} \right)^{-1} H^T \mathbf{X}, \left(H^T H + \sigma_x^2 / \sigma_\alpha^2 \mathbf{I} \right)^{-1} \sigma_x^2 \right\};$$

$$[\sigma_x^2|\mathbf{D}, \mathbf{\Omega}_{-\sigma_x^2}] \propto \text{IG} \left\{ \frac{n}{2} + \rho_{1\mu}, \sum_{i=1}^n \frac{x_i^2}{2} + \rho_{2\mu} \right\};$$

$$[\sigma_\gamma^2|\mathbf{D}, \mathbf{\Omega}_{-\sigma_\gamma^2}] \propto \text{IG} \left\{ \frac{K_1 - 2}{2} + \rho_{1\gamma}, \sum_{k=2}^{K_1} \frac{(\gamma_k - \gamma_{k-1})^2}{2} + \rho_{2\gamma} \right\};$$

$$[\sigma_{\beta_x}^2|\mathbf{D}, \mathbf{\Omega}_{-\sigma_{\beta_x}^2}] \propto \text{IG} \left\{ \frac{K_2 - 1}{2} + \rho_{1\beta}, \sum_{k=3}^{K_2} \frac{\{\beta_{xk} - 2\beta_{x(k-1)} + \beta_{x(k-2)}\}^2}{2} + \rho_{2\beta} \right\}.$$

We used $\mathbf{\Omega}_{-\theta}$ to denote the set of all parameters with the exception of θ , γ_{-k} to denote the vector of γ parameters with the exception of γ_k , β_{-xk} to denote the vector of β_x parameters with the exception of β_{xk} , β_{-zk} to denote the vector of β_z parameters with the exception of

β_{zk} and $H_i = (1Z_i^T)^T$ where H is the matrix with i th row equal to H_i^T . We also denoted by $R(i, k) = \min(y_i, \kappa_k^g) - \kappa_{k-1}^g$. The distributions $[\gamma_k|\gamma_{-k}, \sigma_\gamma^2]$ and $[\beta_{xk}|\beta_{-xk}, \sigma_{\beta_x}^2]$ are the products of the likelihood terms that are depend only on γ_k and β_{xk} , respectively. For example,

$[\gamma_k|\gamma_{-k}, \sigma_\gamma^2]$ varies with k as

$$\begin{aligned} [\gamma_1|\gamma_{-1}, \sigma_\gamma^2] &\propto \exp \left\{ \frac{-(\gamma_1 - \gamma_2)^2}{2\sigma_\gamma^2} \right\}, \\ [\gamma_k|\gamma_{-k}, \sigma_\gamma^2] &\propto \exp \left\{ \frac{-\{(\gamma_{k-1} - \gamma_k)^2 - (\gamma_k - \gamma_{k+1})^2\}}{2\sigma_\gamma^2} \right\}, k=2, \dots, K_1 - 1, \\ [\gamma_{K_1}|\gamma_{-K_1}, \sigma_\gamma^2] &\propto \exp \left\{ \frac{-(\gamma_{K_1-1} - \gamma_{K_1})^2}{2\sigma_\gamma^2} \right\}. \end{aligned} \quad (8)$$

In large datasets, such as the one in the ARIC study, updating the missing observations, X_i , $i = 1, \dots, n$, is the computational bottleneck. However, $[X_i|\mathbf{D}, \mathbf{\Omega}_{-X_i}]$ depends only on subject i data which leads important software simplifications.

The full conditional distribution of $\mathbf{X}$, γ , β_x , and β_z do not have a closed form and direct simulation is not possible. A standard approach to solving this problem is the Metropolis-Hastings (MH) algorithm. However, our implementation of the MH algorithm led to very poor convergence and mixing of the simulated chains. We believe that the main reason for this behavior is the difficulty of tuning the proposal distributions for all missing observations, X_i . To avoid this problem, we employed a different approach. Because the full conditional distributions are log-concave (see Theorem), the adaptive rejection sampling (ARS) algorithm (Gilks and Wild 1992) can be applied. This algorithm is an accept-reject

algorithm that ensures high acceptance rates by adaptively constructing an envelop of the target distribution. ARS is known to perform well for measurement error problems with a linear exposure/outcome function in the Cox model (Tadesse et al. 2005). We report a similar good behavior when the exposure/outcome function in the variable measured with error is nonlinear.

Theorem. $[X_i|\mathbf{D}, \mathbf{\Omega}_{-X_i}]$, $\gamma_k|\mathbf{D}, \mathbf{\Omega}_{-\gamma_k}$, $[\gamma_{xk}|\mathbf{D}, \mathbf{\Omega}_{-\beta_{xk}}]$, and $[\beta_{zk}|\mathbf{D}, \mathbf{\Omega}_{-\beta_{zk}}]$ are log-concave functions for all i and k .

Given the definition in Equation (6), it follows that $g(t; \gamma)$ and $f(x; \beta_x)$ are linear functions in γ and β_x , respectively. In addition, $f(x; \beta_x)$ is everywhere differentiable in x with the exception of a finite number of knots because $f(x; \beta_x)$ is a continuous piecewise linear function in x . Thus, it is enough to show that the second derivative of the log likelihood functions are negative. This is true since, with a slight abuse of notation,

$$\begin{aligned}\frac{\partial^2 \log [X_i|\mathbf{D}, \mathbf{\Omega}_{-X_i}]}{\partial^2 X_i} &= -\left\{f'(x_i; \beta_x)\right\}^2 \int_0^{\gamma_i} \exp\{g(t; \gamma)\} dt \times \exp\{f(x_i; \beta_x) + z_i \beta_z\} - \frac{1}{\sigma_x^2}, \\ \frac{\partial^2 \log [\gamma_k|\mathbf{D}, \mathbf{\Omega}_{-\gamma_k}]}{\partial^2 \gamma_k} &= -e^{\gamma_k} \sum_{i=1}^n R(i, k) \exp\{f(x_i; \beta_x) + z_i \beta_z\} - \frac{\partial^2 \log [\gamma_k|\gamma_{-k}, \sigma_\gamma^2]}{\partial^2 \gamma_k}, \\ \frac{\partial^2 \log [\beta_{xk}|\mathbf{D}, \mathbf{\Omega}_{-\beta_{xk}}]}{\partial^2 \beta_{xk}} &= -\sum_{i=1}^n \left\{B_k^f(x_i)\right\}^2 \int_0^{\gamma_i} \exp\{g(t; \gamma)\} dt \times \exp\{f(x_i; \beta_x) + z_i \beta_z\} - \frac{\partial^2 \log [\beta_{xk}|\beta_{-xk}, \sigma_{\beta_x}^2]}{\partial^2 \beta_{xk}}, \\ \frac{\partial^2 \log [\beta_{zk}|\mathbf{D}, \mathbf{\Omega}_{-\beta_{zk}}]}{\partial^2 \beta_{zk}} &= -\sum_{i=1}^n z_{ik}^2 \int_0^{\gamma_i} \exp\{g(t; \gamma)\} dt \times \exp\{f(x_i; \beta_x) + z_i^T \beta_z\} - \frac{1}{\sigma_z^2}.\end{aligned}$$

4. SIMULATIONS

In this section, we present two simulation studies to compare the performance of our fully Bayesian measurement error methodology (denoted BME in the remainder of the article) and a naïve Bayesian method (denoted by BN in the remainder of the article). As their names suggest, both methods are Bayesian. Moreover, they both use the full likelihood instead of the partial likelihood and model the log-baseline hazard and exposure/outcome functions nonparametrically using penalized splines. However, the main difference is that the naïve method ignores the measurement error mechanism and uses the proxy measurements, W , instead of the true measurements, X . We have also conducted simulations based on SIMEX using regression splines (Carroll et al. 2006). As BME and BN provided better results in cases when the number and placement of knots is unknown, we only report these results in the web supplement.

For each subject i , we simulated mutual independent observations X_i and Z_i from a standard normal distribution. We treated X_i as the covariates observed with error and Z_i as the covariates observed without error. The observed data vectors were simulated according to the following algorithm:

- Use a constant baseline hazard function $\lambda_0(t) = \zeta$, where $\zeta \sim 0.5$.

- Conditional = on X_i and Z_i , simulated independent failure times, T_i , with a exponential distribution with failure rate $\zeta \exp\{\sin(x_i) + z_i\beta_z\}$, where $\beta_z = 1$.
- Simulate W_i by adding normal independent noise to X_i , that is,

$$W_i \sim N(X_i, \sigma_u^2).$$

Two simulation scenarios are conducted in this study: σ_u^2 is 0.3 and 1 corresponding to reliability 0.77 and 0.5, respectively.

- Simulate independent censoring time, C_i , from an uniform distribution on the interval (0, 50), which corresponds to roughly 10% censoring rate.
- The observed data vector for subject i is $(Y_i, \delta_i, W_i, Z_i)_{i=1}^n$, where $Y_i = \min(T_i, C_i)$, $\delta_i = I(T_i < C_i)$.

Using this algorithm, we simulated $M = 50$ data sets, each of sample size $n = 500$. For each data set, we ran 10,000 simulations using the Gibbs sampling methodology described in Section 3. The first 5,000 simulations were discarded as burn-in and inference was based on the remaining 5,000 simulations.

Figure 2 displays results for the two simulation scenarios corresponding to moderate and large measurement error. The top panel corresponds to moderate measurement error (reliability = 0.77) and the bottom panel corresponds to large measurement error (reliability = 0.5). In each panel, we plotted the true exposure/outcome function, $f(x) = \sin(x)$ (solid line), and the averages over the 50 simulation data sets of BME_{10} , BME_{20} , and BN_{20} functional estimators, where the subscripts of BME and BN denote the number of knots. More precisely, the dashed line corresponds to the function $x \rightarrow \bar{f}^{BN_{20}}(x)$, where

$$\bar{f}^{BN_{20}}(x) = \frac{1}{M} \sum_{m=1}^M \hat{f}_m^{BN_{20}}(x),$$

and $\hat{f}_m^{BN_{20}}(x)$ is the BN_{20} estimator of $f(x)$ using dataset m , where $m = 1, \dots, M$. Similarly, the dotted and dot-dash lines correspond to the function $x \rightarrow \bar{f}^{BME_{20}}(x)$ and $x \rightarrow \bar{f}^{BME_{10}}(x)$,

where $\bar{f}^{BME_{20}}(x) = \frac{1}{M} \sum_{m=1}^M \hat{f}_m^{BME_{20}}(x)$ and $\bar{f}^{BME_{10}}(x) = \frac{1}{M} \times \sum_{m=1}^M \hat{f}_m^{BME_{10}}(x)$. Figure 2 indicates that there is a tendency of the BN_{20} estimator to be biased towards no functional effect. This situation has been reported in other linear and nonlinear regression with measurement error (e.g., Berry, Carroll, and Ruppert 2002 and Carroll et al. 2006). In contrast, both BME_{10} and BME_{20} estimators tend, at least on average, to perform better than the Bayesian naive method especially when measurement error is large.

To provide a more detailed comparison we computed the average mean square error (AMSE) over a grid of points. More precisely, for a dataset m we define

$$AMSE(m, \hat{f}^{BN_{20}}) = \frac{1}{G} \sum_{g=1}^G \left\{ \hat{f}_m^{BN_{20}}(x_g) - \sin(x_g) \right\}^2,$$

$$AMSE(m, \hat{f}^{BME_{20}}) = \frac{1}{G} \sum_{g=1}^G \left\{ \hat{f}_m^{BME_{20}}(x_g) - \sin(x_g) \right\}^2,$$

$$\text{AMSE}(m, \widehat{f}^{\text{BME}_{10}}) = \frac{1}{G} \sum_{g=1}^G \left\{ \widehat{f}_m^{\text{BME}_{10}}(x_g) - \sin(x_g) \right\}^2,$$

where $G = 100$ and $x_g, g = 1, \dots, G$, is an equally spaced grid of points in the interval $-2, 2$ that covers the area of high probability for X . Similarly, we computed

$\text{AMSE}(m, \widehat{\lambda}_0^{\text{BN}_{20}})$, $\text{AMSE}(m, \widehat{\lambda}_0^{\text{BME}_{20}})$, and $\text{AMSE}(m, \widehat{\lambda}_0^{\text{BME}_{10}})$ for the baseline hazard $\lambda_0(t)$, where notations should be self explanatory. In this case, we used an equally spaced grid of $G = 100$ points in the interval $[0, y_{\max}]$, where $y_{\max}$ is the largest observed failure time.

The top two panels of Figure 3 display the boxplots of $\text{AMSE}(m, \widehat{f}^{\text{BN}_{20}})$, $\text{AMSE}(m, \widehat{f}^{\text{BME}_{20}})$, and $\text{AMSE}(m, \widehat{f}^{\text{BME}_{10}})$ over all simulated datasets. The left panels corresponds to moderate measurement error (reliability = 0.77) while the right panels correspond to large measurement error (reliability = 0.5). Both panels indicate that the BME method outperforms the BN method, no matter how many knots, especially when the measurement error variability is large. Note, for example, that when $\sigma_u^2=1$ the worst fit using both BME_{10} and BME_{20} are better than the best fit using BN_{20} .

The middle and bottom plots tell a similar story for the estimation of the baseline hazard function and β_z , respectively. The middle plots display the AMSE while the lower plots display the MSE because β_z is a parameter, not a function.

These simulations were possible due the computational efficiency of our algorithms. For example, it takes about 0.74 seconds to complete one sampling cycle for one of the datasets (10 knots, $\sigma_u^2=0.3$) on a PC (dual core processor 3 GHz CPU, 8 GB RAM), thus resulting in roughly 2.05 hours for 10,000 iterations for one MCMC chain.

5. TIME TO CKD IN THE ARIC STUDY

We applied the Bayesian measurement error (BME) and a Bayesian naïve methods (BN) to the ARIC study. We have already provided a brief description of this study, but it is worth emphasizing one more time the sample size. Recall that 15,792 men and women were recruited with a median follow-up time of roughly 14 years and a total of 1,602 incidence CKD cases were recorded. As we will illustrate in the following, our methods handle data of this size and complexity with ease.

Considering the following Cox model for the hazard function of time to CKD:

$$\lambda_i(t) = \exp \left[g(t; \gamma) + f \{ \log(\text{GFR}_i), \beta_x \} + \beta_{z1} \text{AA}_i + \beta_{z2} \text{Age}_i + \beta_{z3} \text{Sex}_i \right],$$

where $g(t, \gamma) = \log \lambda_0(t, \gamma)$, $f \{ \log(\text{GFR}_i), \beta_x \}$ is an unspecified function of log GFR, AA is an indicator for African Americans, and Age and Sex are self explanatory. The model is completed by adding the classical measurement error model

$$\log(\text{eGFR})_i = \log(\text{GFR})_i + U_i,$$

where eGFR is the estimated eGFR, GFR is the true GFR, $E\{\log(\text{GFR})|\text{AA}, \text{Age}, \text{Sex}\} = \alpha_0 + \alpha_1 \text{AA} + \alpha_3 \text{Sex}$ and $U_i \sim N(0, \sigma_u^2)$, where $\sigma_u^2 = (0.083)^2$.

The function $g(t, \gamma)$ is modeled as a piecewise constant B-spline with 20 knots placed at equally spaced quantiles of failure times. The function $f\{\log(\text{GFR}_i), \beta_x\}$ is modeled as a linear B-spline. To gain insight into the effect of the number of knots we used both 30 and 50 knots placed at equally spaced quantiles of the observed log eGFR. We centered age and log eGFR around their means, 54 and 4.53 ml/min/1.73 m², respectively. Thus, the reference group in our analysis is white females of age 54 and with a GFR equal to $91.8 = \exp(4.52)$ ml/min/1.73 m².

The initial values for the BN method were obtained by fitting a Cox model without penalization. The initial values for the BME method were obtained by using the BN method. We used 10,000 iterations to simulate the joint posterior distributions of the model parameters given the data using the methodology described in Section 3. We discarded the first 5,000 iterations as burn-in and used the remaining 5,000 iterations for inference. Mixing of the chains was good as illustrated in the attached supplementary materials (www.amstat.org/publications/jasa/supplemental_materials). The entire program was implemented in R. It took 21.3 seconds to complete one sampling cycle and 59.1 hours to obtain 10,000 samples for one MCMC chain on a PC (dual core processor 3 GHz CPU, 8 GB RAM).

Figure 4 displays the posterior means of $f\{\log(\text{GFR}_i), \beta_x\}$ using the BN and BME methods, respectively. We relabeled the x axis in terms of true GFR instead of log GFR for reasons of interpretability. We used notations similar to those in Section 4. For example, BN₅₀ indicates the Bayesian naïve estimator using 50 knots for log GFR. The dotted line indicates the posterior mean of the BN₅₀ estimator. The solid and dash lines indicate the posterior mean of our BME₅₀ and BME₃₀ estimators, while the dot-dash lines indicate the pointwise 95% credible intervals for the BME₅₀ functional estimator. Both approaches (BN and BME) show that risk of progression to CKD is a decreasing function of GFR when GFR is roughly less than 90 but when $\text{GFR} > 90$ there is no statistical evidence of increased risk of progression to CKD. An important difference is that the BME methodology, both for 30 and 50 knots, estimates a much higher relative risk of progression to CKD when $\text{GFR} < 90$ than the BN approach. Moreover, there is no practical difference between using 30 and 50 knots to model the effect of log GFR.

Figure 5 displays the the posterior means of $g(t, \gamma)$ using the BN and BME methods, respectively. The lines are coded as in Figure 4 indicating that both methods estimate a much higher baseline hazard around year 9 and slightly higher around year 3. The higher baseline hazard around year 9, as well as year 3, is a consequence of the sampling mechanism. More precisely, these are the times corresponding to the visits where serum creatinine was collected. Serum creatinine is used as laboratory evidence of CKD. This explains the additional hazard on top of the risk detected by hospitalization and death surveillance which was present at all years. Another interesting finding is that the BN₅₀ estimator of the baseline hazard function tends to be larger than the BME₃₀ and BME₅₀ estimators. In contrast, the BN₅₀ estimator of $f(\text{GFR}, \beta_x)$ tends to be smaller than the BME₃₀ and BME₅₀ estimators.

Table 1 provides a comparison between the BN and BME methods for the parameters of the covariates measurement without error. Both methods estimate a statistically significant African American race and age effects and insignificant sex effect. As a final point, note that the BN approach produces narrower confidence intervals for all covariates.

Carroll et al. (2006) also analyzed the same dataset in their chapter 14 using the SIMEX approach and a regression spline with 4 knots. This procedure works well when the number of knots is small and their placement is informed by science. However, our proposed

approach can avoid the discussion of how many knots should be used and where they should be placed. This has important implications for our scientific collaborators by confirming that the levelling of the risk around $\text{GFR} = 90$ does not depend on the knot placement. On the other hand, the SIMEX approach provides less satisfactory results when using regression splines with a large number of knots, say 50. The reason is that the naïve as well as the noise-inflated estimators are seriously undersmoothed. Extrapolating these undersmoothed estimators results in even more undersmoothed SIMEX estimators. To show these effects, the SIMEX functional estimator using regression splines with 50 knots is compared with our proposed BME_{50} method as well as with the naïve Bayesian smooth method BN_{50} in the attached supplementary materials (www.amstat.org/publications/jasa/supplemental_materials). Some potential solutions could be to use SIMEX for our BN_{50} method or using smoothing methods specifically designed for SIMEX curve estimation (Staudenmayer and Ruppert 2004). Both these methods exceed the purpose of the current paper.

Despite of these differences, it is instructive to compare our results with the SIMEX results in table 14.2 in Carroll et al. (2006). We estimate the effect of African American race to be 0.60 with a standard error of 0.064 compared to 0.63 with a standard error of 0.062 in Carroll et al. (2006). We estimated the effect of age to be 0.069 with a standard error of 0.0051 compared to 0.054 with a standard error of 0.0049 in Carroll et al. (2006). As Carroll et al. (2006), we report an statistically nonsignificant effect of sex on time to progression to CKD.

6. DISCUSSION

This methodological research was motivated by a complex scientific application to estimate the hazard function of time to CKD in a large cohort study. We developed a computationally tractable method that accounts for the obvious and hidden complexities of the data set. In particular, we provided a coherent penalized approach to model flexibly both the baseline hazard and the exposure/outcome function. This model was coupled with a measurement error approach to account for the effects of imperfect observations on inference.

Our simulation studies and data analysis show that the naïve approach tends to alter both the functional shape of functions and the properties of the point estimators for the linear component parameters. In this paper, we only observe attenuation of effects, but this need not be true in general (see Carroll et al. 2006 for more details).

From a scientific perspective we have shown that there is no statistical evidence of a decreased risk of progression to CKD when the true GFR is greater than 90. We have also provided bias corrected risk of progression estimates for the case when GFR is less than 90. Last, but not least, because we allowed for a large maximum complexity we provided evidence that there are no hidden shape complexities of the exposure/outcome and baseline hazard functions. These results complement the classification of kidney function as normal if eGFR is greater than 90 and suggest that despite limitations in GFR estimation there is a risk gradient for onset of CKD with greater risk at lower estimated GFR between 60 and 89 among middle-aged adults (National Kidney Foundation 2002; Coresh et al. 2003).

There are many potentially important directions for future methodological development. For example, the measurement error models could be refined to incorporate non-Gaussian distribution of the measurement errors or unobserved true GFR measurements. Also, incorporating a mixture of Berkson and classical measurement error as described in Mallick, Hoffman, and Carroll (2002).

Acknowledgments

Crainiceanu's research was supported by the grant R01NS060910 from the National Institute of Neurological Disorders and Stroke. The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts N01-HC-55015, N01-HC-55016, N01-HC-55018, N01-HC-55019, N01-HC-55020, N01-HC-55021, and N01-HC-55022. The authors thank the staff and participants of the ARIC study for their important contributions.

REFERENCES

- Arjas E, Gasbarra D. Nonparametric Bayesian Inference From Right Censored Survival Data, Using Gibbs Sampler. *Statistica Sinica*. 1994; 4:505–524.
- Augustin T. An Exact Corrected Log-Likelihood Function for Cox's Proportional Hazards Model Under Measurement Error and Some Extensions. *Scandinavian Journal of Statistics*. 2004; 31:43–50.
- Berry S, Carroll R, Ruppert D. Bayesian Smoothing and Regression Splines for Measurement Error Problems. *Journal of the American Statistical Association*. 2002; 97:160–169.
- Carroll, R.; Ruppert, D.; Stefanski, L.; Crainiceanu, C. *Measurement Error in Nonlinear Models, A Modern Perspective*. 2nd ed.. Chapman & Hall; New York: 2006.
- Clayton, D. *Statistical Models for Longitudinal Studies of Health*. Oxford University Press; New York: 1991. *Models for the Analysis of Cohort and Case-Control Studies With Inaccurately Measured Exposures*.
- Coresh J, Astor B, Greene T, Eknoyan G, Levey A. Prevalence of Chronic Kidney Disease and Decreased Kidney Function in the Adult US Population: Third National Health and Nutrition Examination Survey. *American Journal of Kidney Diseases*. 2003; 41:1–12. [PubMed: 12500213]
- Coresh J, Astor B, McQuillan G, Kusek J, Greene T, Van Lente F, Levey A. Calibration and Random Variation of the Serum Creatinine Assay as Critical Elements of Using Equations to Estimate Glomerular Filtration Rate. *American Journal of Kidney Diseases*. 2002; 39:920–929. [PubMed: 11979335]
- Cox D. Regression Models and Life Tables. *Journal of the Royal Statistical Society, Ser. B*. 1972; 34:187–220. (with discussion).
- Cox D. Partial Likelihood. *Biometrika*. 1975; 62:269–276.
- Crainiceanu C, Ruppert D, Carroll R, Joshi A, Goodner B. Spatially Adaptive Bayesian Penalized Splines With Heteroscedastic Errors. *Journal of Computational and Graphical Statistics*. 2007; 16:265–288.
- Eilers P, Marx B. Flexible Smoothing Using B-Splines and Penalized Likelihood. *Statistical Science*. 1996; 11:89–121.
- Gamerman D. Dynamic Bayesian Models for Survival Data. *Applied Statistics*. 1991; 40:63–79.
- Geman S, Geman D. Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1984; 6:721–741.
- Gilks W, Wild P. Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics*. 1992; 41:337–348.
- Gray RJ. Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis. *Journal of the American Statistical Association*. 1992; 87:942–951.
- Gray RJ. A Bayesian Analysis of Institutional Effects in a Multicenter Cancer Clinical Trial. *Biometrics*. 1994; 50:244–253. [PubMed: 8086608]
- Greene W, Cai J. Measurement Error in Covariates in the Marginal Hazards Model for Multivariate Failure Time Data. *Biometrics*. 2004; 60:987–996. [PubMed: 15606419]
- Hastie, T.; Tibshirani, R. *Generalized Additive Models*. Chapman & Hall; London: 1990.
- Hennerfeind A, Brezger A, Fahrmeir L. Geoadditive Survival Models. *Journal of the American Statistical Association*. 2006; 101:1065–1075.
- Hsu C, Kao W, Coresh J, Pankow J, Marsh-Manzi J, Boerwinkle E, Bray M. Apolipoprotein E and Progression of Chronic Kidney Disease. *Journal of the American Medical Association*. 2005; 293(23):2892–2899. [PubMed: 15956634]

- Hu C, Lin D. Semiparametric Failure Time Regression With Replicates of Mismeasured Covariates. *Journal of the American Statistical Association*. 2004; 99:105–118.
- Hu P, Tsiatis A, Davidian M. Estimating the Parameters of Cox Model When Covariate Variables Are Measured With Errors. *Biometrics*. 1998; 54:1407–1419. [PubMed: 9883541]
- Huang Y, Wang C. Cox Regression With Accurate Covariates Unascertainable: A Nonparametric Correction Approach. *Journal of the American Statistical Association*. 2000; 96:1469–1482.
- Ibrahim, J.; Chen, M-H.; Sinha, D. *Bayesian Survival Analysis*. Springer-Verlag; New York: 2001.
- Investigators A. The Atherosclerosis Risk in Communities (ARIC) Study: Design and Objectives. *American Journal of Epidemiology*. 1989; 129:689–702.
- Kauermann G, Xu R, Vaida F. Stacked Laplace-EM Algorithm for Duration Models With Time-Varying and Random Effects. *Computational Statistics and Data Analysis*. 2008; 52(5):2514–2528.
- Kottgen A, Russell S, Loehr L, Crainiceanu C, Rosamond W, Chang PP, Chambless L, Coresh J. Reduced Kidney Function as a Risk Factor for Incident Heart Failure: The Atherosclerosis Risk in Communities (ARIC) Study. *Journal of the American Society of Nephrology*. 2007; 18:1307–1315. [PubMed: 17344421]
- Lang S, Brezger A. Bayesian P-Splines. *Journal of Computational and Graphical Statistics*. 2004; 13:181–212.
- Leonard T. Density Estimation, Stochastic Processes and Prior Information. *Journal of the Royal Statistical Society, Ser. B*. 1978; 40:113–146.
- Li Y, Lin X. Functional Inference in Frailty Models for Clustered Survival Data Using the Simex Approach. *Journal of the American Statistical Association*. 2003; 98:191–203.
- Li Y, Ryan L. Inference on Survival Data With Covariate Measurement Error an Imputation-Based Approach. *Scandinavian Journal of Statistics*. 2006; 33:169–190.
- Mallick B, Hoffman F, Carroll R. Semiparametric Regression Modelling With Mixtures of Berkson and Classical Error, With Application to Fallout From the Nevada Test Site. *Biometrics*. 2002; 58:13–20. [PubMed: 11890308]
- Miller, R. *Survival Analysis*. Wiley; New York: 1998.
- Nakamura T. Proportional Hazards Models With Covariates Subject to Measurement Error. *Biometrics*. 1992; 48:829–838. [PubMed: 1420844]
- National Kidney Foundation. K/DOQI Clinical Practice Guidelines for Chronic Kidney Disease: Evaluation, Classification and Stratification. *American Journal of Kidney Disease*. 2002; 39(1):S1–S266.
- Pepe M, Self S, Prentice R. Further Results in Covariate Measurement Errors in Cohort Studies With Time to Response Data. *Statistics in Medicine*. 1989; 8:1167–1178. [PubMed: 2799136]
- Prentice R. Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model. *Biometrika*. 1982; 69:331–342.
- Ruppert, D.; Wand, M.; Carroll, R. *Semiparametric Regression*. Cambridge University Press; Cambridge: 2003.
- Sleeper LA, Harrington DP. Regression Splines in the Cox Model With Application to Covariate Effects in Liver Disease. *Journal of the American Statistical Association*. 1990; 85:941–948.
- Song X, Huang Y. On Corrected Score Approach for Proportional Hazards Model With Covariate Measurement Error. *Biometrics*. 2005; 61:702–714. [PubMed: 16135021]
- Staudenmayer J, Ruppert D. Local Polynomial Regression and Simulation-Extrapolation. *Journal of the Royal Statistical Society, Ser. B*. 2004; 66:17–30.
- Stefanski L. Unbiased Estimation of a Nonlinear Function of a Normal Mean With Application to Measurement Error Models. *Communications in Statistics, Ser. A*. 1989; 18:4335–4358.
- Stevens L, Coresh J, Greene T, Levey A. Assessing Kidney Function—Measured and Estimated Glomerular Filtration Rate. *New England Journal of Medicine*. 2006; 354(23):2473–2483. [PubMed: 16760447]
- Tadesse M, Ibrahim J, Gentleman R, Chiaretti S, Ritz J, Foa R. Bayesian Error-in Variable Survival Model for the Analysis of GeneChip Arrays. *Biometrics*. 2005; 61:488–497. [PubMed: 16011696]

- Tsiatis A, Davidian M. A Semiparametric Estimator for the Proportional Hazards Model With Longitudinal Covariates Measured With Error. *Biometrika*. 2001; 88:447–458.
- Wang C, Hsu L, Feng Z, Prentice R. Regression Calibration in Failure Time Regression. *Biometrics*. 1997; 53:131–145. [PubMed: 9147589]
- Zhou H, Pepe M. Auxiliary Covariate Data in Failure Time Regression. *Biometrika*. 1995; 82:139–149.
- Zhou H, Wang C. Failure Time Regression With Continuous Covariates Measured With Error. *Journal of the Royal Statistical Society, Ser. B*. 2000; 62:657–665.

Figure 1. The top panel displays the degree 0 B-spline basis function for a particular choice of knots. We shifted up by a small amount every other basis function to ensure a better visualization. The bottom panel displays the degree 1 B-spline basis function with the same knots.

Figure 2.

The solid line denotes true function, $\sin(x)$. The dashed, dotted, and dot-dash lines denote the average of BN_{20} , BME_{20} , and BME_{10} functional estimators over 50 simulation datasets when $\sigma_u^2=0.3$ (top panel) and $\sigma_u^2=1$ (bottom panel).

Figure 3.

The boxplots of Average Mean Square Error (AMSE) of BN_{20} , BME_{20} , and BME_{10} functional estimator for $\sin(x)$ (top two panels) and $\lambda_0(t)$ (middle two panels). The bottom two panels display the MSE for the BN_{20} , BME_{20} , and BME_{10} estimators of β_z . The left panels correspond to moderate measurement error, $\sigma_u^2=0.3$, while the right panels correspond to large measurement error, $\sigma_u^2=1$. Results are based on inference for 50 simulated datasets.

Figure 4.

BME₅₀, BME₃₀, and BN₅₀ estimators of the exposure/outcome function, $f\{\log(\text{GFR}), \beta_x\}$ using the model described in Section 5 for the ARIC data set. The x axis was relabeled in terms of true GFR instead of log GFR for reasons of interpretability. The solid and dot-dash lines indicate the posterior mean and 95% pointwise credible intervals based on the BME₅₀ estimator. The dashed line indicates the posterior mean based on the BME₃₀ estimator. The dotted line indicates the posterior mean based on the BN₅₀ estimator.

Figure 5.

BME_{50} , BME_{30} , and BN_{50} of log-baseline hazard function, $g(t, \gamma)$, using the model described in Section 5 for the ARIC dataset. The solid and dot-dash lines indicate the posterior mean and 95% pointwise credible intervals based on the BME_{50} estimator. The dashed line indicates the posterior mean based on the BME_{30} estimator. The dotted line indicates the posterior mean based on the BN_{50} estimator.

Table 1

Posterior mean and 95% credible interval (CI) for the parameters of AA, Age, and Sex using the BN and BME methods

	AA	Age	Sex
BN ₅₀	0.496	0.075	0.015
95% CI	(0.375, 0.611)	(0.065, 0.083)	(−0.084, 0.111)
BME ₃₀	0.583	0.071	0.004
95% CI	(0.437, 0.721)	(0.059, 0.083)	(−0.140, 0.139)
BME ₅₀	0.596	0.069	0.003
95% CI	(0.466, 0.721)	(0.059, 0.079)	(−0.104, 0.111)