Statistica Sinica **29** (2019), 47-66
doi:https://doi.org/10.5705/ss.202015.0369

ADDITIVE MEAN RESIDUAL LIFE MODEL WITH LATENT VARIABLES UNDER RIGHT CENSORING

Haijin He, Deng Pan, Xinyuan Song and Liuquan Sun

Shenzhen University, Huazhong University of Science and Technology, The Chinese University of Hong Kong and Chinese Academy of Sciences

Abstract: We propose a novel additive mean residual life model to examine the effects of observable and latent risk factors on the mean residual life function of interest in the presence of right censoring. We use factor analysis to characterize the latent risk factors on the basis of multiple observed variables. We develop a borrow-strength estimation procedure that incorporates an asymptotically distribution-free generalized least square method and a corrected estimating equation approach. We establish the asymptotic properties of the proposed estimators. We develop a goodness-of-fit test for model checking. We report on simulations to evaluate the finite sample performance of the method. The application to a study on chronic kidney disease for type 2 diabetic patients reveals insights into the prevention of such common diabetic complications.

Key words and phrases: Borrow-strength estimation, corrected estimating equations, distribution-free factor analysis, latent variables, mean residual life function, model checking.

1. Introduction

In medical studies, patients and physicians are often interested in how much a new treatment potentially affects the mean residual life (MRL) rather than the hazard. It is thus appealing to directly investigate a surviving patient's remaining life. A useful alternative to the hazard-based approach is the MRL model (Oakes and Dasu (1990); Maguluri and Zhang (1994); Chen et al. (2005); Chen (2007)). For a nonnegative survival time T with finite expectation, the MRL function at time $t \geq 0$ is $m(t) = E(T - t|T > t)$, which measures the remaining life expectancy of a subject who has survived until time t. The MRL function has a one-to-one correspondence with the survival function of T; thus, in theory, it also characterizes the stochastic behavior of T. The MRL function is widely applied in many substantive fields. For example, under the term 'life expectancy', demographers studied the MRL function in human population research. In industrial reliability studies, the MRL function is highly effective for

enhancing system reliability and developing maintenance policies. In biomedical studies involving survivorship, a so-far survived patient may wish to know how much longer he/she can expect to live. The MRL function is also useful in actuarial studies relating to life insurance, health science, and so on. We refer interested readers to Guess and Proschan (1988) for a detailed discussion on the applications of the MRL function.

Various models have been proposed for regression analysis that assesses the effects of covariates $\mathbf{Z}$ on the MRL function $m(t|\mathbf{Z})$. For instance, Oakes and Dasu (1990) and Maguluri and Zhang (1994) studied the proportional MRL model without censoring. Chen and Cheng (2005) and Chen et al. (2005) developed semiparametric estimation procedures for the proportional MRL model with censored data. Chen and Cheng (2006) and Chen (2007) proposed the additive MRL model and discussed various estimation procedures with or without right censoring. Sun and Zhang (2009) and Sun, Song and Zhang (2012) proposed a class of transformed MRL models with time-independent and time-dependent coefficients, and developed inference procedures for estimating model parameters under right censoring.

These regression models assume that risk factors are observable and directly assessable. However, unobservable traits, 'latent variables', are common in many applications. Such latent variables cannot be fully measured by a single observed variable, but are characterized by several highly correlated observed indicators from different perspectives. Typical examples include 'blood pressure', summarized by systolic blood pressure (SBP) and diastolic blood pressure (DBP) and 'lipid' is measured by total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), and/or triglycerides (TG). A number of limitations are evident if the correlated indicators of a latent trait are simply regarded as independent variables in a regression model (Roy and Lin (2000, 2002)). Major limitations include incomplete measurement of latent traits, multicollinearity incurred by high correlations among multiple indicators, and incapability of providing insights into the overall effects of latent risk factors or the determinants of latent responses. For instance, the application reported in Section 7 shows that simultaneously incorporating the two indicators of blood pressure, such as SBP and DBP, or the three indicators of lipid, such as TC, HDL-C, and TG, in regression can cause multicollinearity and produce misleading results.

Factor analysis is a widely used statistical tool that measures latent variables on the basis of multiple observed indicators (Lawley and Maxwell (1971)).

Various latent variable models have been developed based on the factor analysis model (Jöreskog and Sörbom (1996); Lee (2007); Song and Lee (2012)). Sammel and Ryan (1996) analyzed the effects of anticonvulsant medication during pregnancy with the use of a joint modeling approach, wherein the response (the overall severity of birth defects) in their regression model was characterized by various adverse effects through a confirmatory factor analysis model (Bollen (1989)). Roy and Lin (2000) extended the model of Sammel and Ryan (1996) to accommodate longitudinal latent variables in a linear mixed model, where the latent response was summarized by three longitudinal outcomes relating to the effectiveness of the treatment practice. Roy and Lin (2002) considered similar models that further accounted for non-ignorable dropouts and missing covariates.

The development of latent variable modeling in survival analysis is rather limited. Pan et al. (2015) proposed an additive hazards model with latent variables to investigate the observed and latent risk factors of failure time and demonstrated the advantages of their model over the conventional additive hazards model. However, the construction of latent risk factors has never been introduced into the MRL regression framework. The MRL function is an attractive alternative to hazard when the ordering assumption on a hazard function is violated (Sun and Zhang (2009)), and it can provide straightforward knowledge about the remaining life expectancy for events of interest. Despite the developments of Pan et al. (2015), considering the construction of the MRL model with latent variables remains important because the MRL model represents another type of model that has been commonly used in survival and reliability analyses. Moreover, Pan et al. (2015) assumed that latent risk factors and residual errors were normally distributed, and this may not hold in a substantive study, misspecification leading to erroneous inference. To advance matters, we propose a joint model that comprises a distribution-free factor analysis model for measuring latent risk factors via multiple observed indicators and a MRL model for examining the effects of latent and observed risk factors on the MRL function of interest. Two types of MRL models, additive and proportional, can be considered. We focus on the former because its additive structure complies with the intrinsic constraint that $m(t|\mathbf{Z}) + t$ is monotonically nondecreasing for all values of covariates $\mathbf{Z}$, and its regression parameters allow interpretation in mean differences. The proportional MRL model can violate the intrinsic constraint without a monotonically nondecreasing baseline MRL function that is not always compatible with the underlying process (Oakes and Dasu (1990); Chen and Cheng (2006)).

In conducting inference in the proposed model, the EM-type algorithms that are commonly used in the latent variable modeling literature (Sammel and Ryan (1996); Roy and Lin (2000, 2002)) are not directly applicable. The difficulty lies in the fact that the likelihood function is unavailable because the distributions of the latent variables and residual errors in the factor analysis model are unspecified. To circumvent this, we propose the use of an asymptotically distribution-free generalized least square (ADF-GLS, Browne (1984)) approach to estimate the parameters and latent variables in our factor analysis model. We develop a borrow-strength estimation procedure that copes with the ideas of the corrected score method (Carroll, Ruppert and Stefanski (1995)) and of the estimating equation methods (Chen and Cheng (2005, 2006); Sun and Zhang (2009)) to estimate the parameters in the MRL model.

Although the basic ideas of constructing a joint model with latent risk factors are similar, ours differs from the study of Pan et al. (2015) in several aspects. In particular, the model frameworks are different.

The rest of the paper is organized as follows. Section 2 describes the proposed joint model. Section 3 presents the borrow-strength estimation procedure for regression parameters of interest. The asymptotic properties of the proposed estimators are established in Section 4. A goodness-of-fit test is developed for checking the adequacy of the proposed model, in Section 5. Section 6 reports on simulation studies to evaluate the empirical performance of the proposed methods. Section 7 reports on an application to a study on chronic kidney disease (CKD) for type 2 diabetic patients. Section 8 presents the conclusions drawn by the paper. Technical details are in Supplementary Materials.

2. Model

Let $\mathbf{V}_i$ $(i = 1, 2, \ldots, n)$ be a $p \times 1$ vector of observed variables, and $\boldsymbol{\xi}_i$ be a $q \times 1$ vector of latent variables. The latent variables in $\boldsymbol{\xi}_i$ are measured by the observed variables in $\mathbf{V}_i$ via a distribution-free factor analysis model

$$\mathbf{V}_i = \mathbf{B}\boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i, \tag{2.1}$$

where $\mathbf{B}$ is $p \times q$ factor loading matrix, $\boldsymbol{\xi}_i$ has mean zero and covariance matrix $\boldsymbol{\Phi}$, $\boldsymbol{\epsilon}_i$ is a $p \times 1$ vector of random errors independent of $\boldsymbol{\xi}_i$, and $\boldsymbol{\epsilon}_i$ is assumed to have mean zero and diagonal covariance matrix $\boldsymbol{\Psi}_\epsilon$. In this study, we consider model (2.1) as a confirmatory factor analysis model, where the numbers of observed variables and latent factors, p and q, as well as the structure of the factor loading matrix, $\mathbf{B}$, are pre-determined based on substantive theory, expert knowledge,

and/or existing literature (Bollen (1989); Lee (2007)). If such information is unavailable, one can conduct an exploratory factor analysis to determine p, q, and the structure of $\mathbf{B}$ based on the data (Jöreskog and Sörbom (1996)).

Let $\mathbf{Z}_i$ be an $s \times 1$ vector of observed covariates. To investigate the effects of $\mathbf{Z}_i$ and $\boldsymbol{\xi}_i$ on the failure time T_i, we propose the additive MRL model

$$m(t|\mathbf{Z}_i, \boldsymbol{\xi}_i) = m_0(t) + \boldsymbol{\beta}^T \mathbf{Z}_i + \boldsymbol{\gamma}^T \boldsymbol{\xi}_i, \tag{2.2}$$

where $m_0(t)$ is the unspecified baseline MRL function, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are $s \times 1$ and $q \times 1$ vectors of unknown regression parameters. Assume that $m_0(t) + \boldsymbol{\beta}^T \mathbf{Z}_i + \boldsymbol{\gamma}^T \boldsymbol{\xi}_i$ is nonnegative. The joint model defined by (2.1) and (2.2) preserves the embedding constraint of the MRL function that $m(t|\mathbf{Z}_i, \boldsymbol{\xi}_i) + t = E(T_i|T_i > t)$ is nondecreasing (Chen and Cheng (2006)) and the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ directly explain the effects of $\mathbf{Z}_i$ and $\boldsymbol{\xi}_i$ on the MRL function. Here the $\boldsymbol{\xi}_i$ in (2.2) include latent traits (e.g., lipid) that truly exist but cannot be characterized by a single observed variable.

It is possible to use other techniques such as principle component analysis (PCA) to project groups of correlated variables into independent basis functions. If, however, all the principal components are included in a regression, the resulting model is equivalent to the one obtained using least squares, and the large variances caused by multicollinearity do not diminish (Jolliffe (2002)). Selection of the number of principal components is an important issue (Jolliffe (2002); Bair et al. (2006)). But latent variables in a factor analysis represent traits that have intuitive meanings not available through PCA.

To identify the proposed model and obtain a clear interpretation of latent factors, we follow the common practice (Bollen (1989); Song and Lee (2012)) of imposing the identifiability constraints that the diagonal elements of $\boldsymbol{\Phi}$ are fixed to 1 to unify the scales of latent factors and a non-overlapping structure of $\mathbf{B}$ is assigned to ensure that each observed variable does not measure two or more latent factors.

3. Inference Procedure

In this section, we describe the statistical inference of the proposed model. Section 3.1 proposes a corrected estimating equation approach under both covariate-independent censoring and covariate-dependent censoring (CEE1). We also develop a corrected estimating equation method in the presence of covariate-independent censoring only (CEE2). We present CEE1 in Section 3.1 and defer CEE2 to Supplementary Material S3. Section 3.2 proposes the use of an ADF-

GLS approach for the inference of the factor analysis model.

3.1. Corrected estimating equation procedure (CEE1)

Denote the unknown parameters in $(\mathbf{B}, \mathbf{\Phi}, \mathbf{\Psi}_\epsilon)$ by a $p^* \times 1$ vector $\boldsymbol{\theta}$. Based on $\{\mathbf{V}_i, i = 1, \ldots, n\}$, by adopting an ADF-GLS approach (Browne (1984)), we obtain the estimator of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$. The implementation of the ADF-GLS approach in (2.1) and the asymptotic properties of $\hat{\boldsymbol{\theta}}$ are deferred to Section 3.2 and Supplementary Material S1, respectively.

Let C_i be the censoring time. The support of C_i is assumed to be longer than that of the survival time T_i to ensure that the MRL function is estimable, and T_i and C_i are assumed to be independent given $\mathbf{Z}_i$ and $\boldsymbol{\xi}_i$. Let $X_i = \min\{T_i, C_i\}$ be the observed time, and $\Delta_i = I(T_i \leq C_i)$ be the censoring indicator, $I(\cdot)$ the indicator function. $\{(\mathbf{V}_i, \mathbf{Z}_i, \boldsymbol{\xi}_i, X_i, \Delta_i), i = 1, \ldots, n\}$ are independent and identically distributed copies. The observed data consist of $\{(\mathbf{V}_i, \mathbf{Z}_i, X_i, \Delta_i), i = 1, \ldots, n\}$.

Let $N_i(t) = I(X_i \leq t)\Delta_i$ be the observed failure counting process, and $Y_i(t) = I(X_i \geq t)$ be the at-risk process. Let $\lambda(t|\mathbf{Z}_i, \boldsymbol{\xi}_i)$ and $\Lambda(t|\mathbf{Z}_i, \boldsymbol{\xi}_i)$ be the hazard function and cumulative hazard function of T_i given $\mathbf{Z}_i$ and $\boldsymbol{\xi}_i$, respectively. Then (Sun and Zhang (2009))

$$\lambda(t|\mathbf{Z}_i, \boldsymbol{\xi}_i) = \frac{m'(t|\mathbf{Z}_i, \boldsymbol{\xi}_i) + 1}{m(t|\mathbf{Z}_i, \boldsymbol{\xi}_i)},$$

which is equivalent to

$$\{m_0(t) + \boldsymbol{\beta}^T\mathbf{Z}_i + \boldsymbol{\gamma}^T\boldsymbol{\xi}_i\} d\Lambda(t|\mathbf{Z}_i, \boldsymbol{\xi}_i) = d\{m_0(t) + t\}.$$

Take $\boldsymbol{\alpha} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$. If $\boldsymbol{\xi}_i$ are observable, for given $\boldsymbol{\alpha}$, $m_0(t)$ can be estimated by $\hat{m}_{a0}(t; \boldsymbol{\alpha})$ which satisfies the estimating equation (Chen and Cheng (2006))

$$\sum_{i=1}^{n} \{\hat{m}_{a0}(t; \boldsymbol{\alpha}) + \boldsymbol{\beta}^T\mathbf{Z}_i + \boldsymbol{\gamma}^T\boldsymbol{\xi}_i\} dN_i(t) - \sum_{i=1}^{n} Y_i(t) d\{\hat{m}_{a0}(t; \boldsymbol{\alpha}) + t\} = 0. \quad (3.1)$$

As the $\boldsymbol{\xi}_i$ are unobservable, the estimating equation (3.1) is intractable. To address this, we consider the estimator of $\boldsymbol{\xi}_i$ based on (2.1) with known $\boldsymbol{\theta}$, which takes the form (Lee (2007))

$$\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta}) = \mathbf{\Gamma}(\boldsymbol{\theta})\mathbf{V}_i, \quad (3.2)$$

where $\mathbf{\Gamma}(\boldsymbol{\theta}) = (\mathbf{B}^T\mathbf{\Psi}_\epsilon^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{\Psi}_\epsilon^{-1}$ is a $q \times p$ matrix function of $\boldsymbol{\theta}$. For a given $\boldsymbol{\theta}$, we can estimate $m_0(t)$ using $\hat{m}_{a0}(t; \boldsymbol{\alpha}, \boldsymbol{\theta})$ which satisfies

$$\sum_{i=1}^{n}\{\hat{m}_{a0}(t;\boldsymbol{\alpha},\boldsymbol{\theta})+\boldsymbol{\beta}^T\mathbf{Z}_i+\boldsymbol{\gamma}^T\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta})\}dN_i(t)-\sum_{i=1}^{n}Y_i(t)d\{\hat{m}_{a0}(t;\boldsymbol{\alpha},\boldsymbol{\theta})+t\}=0. \tag{3.3}$$

Solving the first order linear differential equation (3.3), we have

$$\hat{m}_{a0}(t;\boldsymbol{\alpha},\boldsymbol{\theta})=\hat{S}_{NA}(t)^{-1}\int_t^{\tau}\hat{S}_{NA}(u)\frac{\sum_{i=1}^{n}\left[Y_i(u)du-\{\boldsymbol{\beta}^T\mathbf{Z}_i+\boldsymbol{\gamma}^T\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta})\}dN_i(u)\right]}{\sum_{i=1}^{n}Y_i(u)}, \tag{3.4}$$

where

$$\hat{S}_{NA}(t)=\exp\left(-\int_0^t\frac{\sum_{i=1}^{n}dN_i(u)}{\sum_{i=1}^{n}Y_i(u)}\right),$$

and $0<\tau=\inf\{t:P(X_i\geq t)=0\}<\infty$. Define

$$d\Omega_i(t;\boldsymbol{\alpha},\boldsymbol{\theta})=\{m_0(t)+\boldsymbol{\beta}^T\mathbf{Z}_i+\boldsymbol{\gamma}^T\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta})\}dN_i(t)-Y_i(t)d\{m_0(t)+t\},$$

or equivalently

$$\Omega_i(t;\boldsymbol{\alpha},\boldsymbol{\theta})=\int_0^t\left[\{m_0(s)+\boldsymbol{\beta}^T\mathbf{Z}_i+\boldsymbol{\gamma}^T\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta})\}dN_i(s)-Y_i(s)d\{m_0(s)+s\}\right],$$

then the $\Omega_i(t;\boldsymbol{\alpha},\boldsymbol{\theta})$ are zero-mean processes under the true model. Subtracting $\sum_{i=1}^{n}d\Omega_i(t;\boldsymbol{\alpha},\boldsymbol{\theta})$ from both sides of (3.3) results in

$$\begin{aligned}&\sum_{i=1}^{n}d\Omega_i(t;\boldsymbol{\alpha},\boldsymbol{\theta})+\sum_{i=1}^{n}\{\hat{m}_{a0}(t;\boldsymbol{\alpha},\boldsymbol{\theta})-m_0(t)\}dN_i(t)\\&\quad-\sum_{i=1}^{n}Y_i(t)d\{\hat{m}_{a0}(t;\boldsymbol{\alpha},\boldsymbol{\theta})-m_0(t)\}=0,\end{aligned}$$

which gives the solution

$$\hat{m}_{a0}(t;\boldsymbol{\alpha},\boldsymbol{\theta})-m_0(t)=-\hat{S}_{NA}(t)^{-1}\int_t^{\tau}\hat{S}_{NA}(u)\frac{\sum_{i=1}^{n}d\Omega_i(u;\boldsymbol{\alpha},\boldsymbol{\theta})}{\sum_{i=1}^{n}Y_i(u)}. \tag{3.5}$$

Consistency and asymptotic normality of $\hat{m}_{a0}(t;\boldsymbol{\alpha},\boldsymbol{\theta})$ can be derived based on (3.5). We can check that

$$\frac{\partial\hat{m}_{a0}(t;\boldsymbol{\alpha},\boldsymbol{\theta})}{\partial\boldsymbol{\alpha}^T}=-\hat{S}_{NA}(t)^{-1}\int_t^{\tau}\hat{S}_{NA}(u)\frac{\sum_{i=1}^{n}\{\mathbf{Z}_i^T,\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta})^T\}dN_i(u)}{\sum_{i=1}^{n}Y_i(u)}. \tag{3.6}$$

If the $\boldsymbol{\xi}_i$ are observable, given $m_0(t)$ estimated by $\hat{m}_{a0}(t;\boldsymbol{\alpha})$, the estimating equation for $\boldsymbol{\alpha}$ (Chen and Cheng (2006)) is

$$\mathbf{U}_a^*(\boldsymbol{\alpha})=\sum_{i=1}^{n}\int_0^{\tau}\begin{pmatrix}\mathbf{Z}_i-\bar{\mathbf{Z}}_a(t)\\ \boldsymbol{\xi}_i-\bar{\boldsymbol{\xi}}_a(t)\end{pmatrix}\{\hat{m}_{a0}(t;\boldsymbol{\alpha})+\boldsymbol{\beta}^T\mathbf{Z}_i+\boldsymbol{\gamma}^T\boldsymbol{\xi}_i\}dN_i(t)=0,$$

where $\bar{\mathbf{Z}}_a(t)=\sum_{i=1}^{n}Y_i(t)\mathbf{Z}_i/\sum_{i=1}^{n}Y_i(t)$, $\bar{\boldsymbol{\xi}}_a(t)=\sum_{i=1}^{n}Y_i(t)\boldsymbol{\xi}_i/\sum_{i=1}^{n}Y_i(t)$.

Notably, $\mathbf{U}_a^*(\boldsymbol{\alpha})$ is not directly tractable because the $\boldsymbol{\xi}_i$ are unobservable in the proposed model. Although $E\{\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta})|\boldsymbol{\xi}_i\} = \boldsymbol{\xi}_i$, a simple replacement of $\boldsymbol{\xi}_i$ by $\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta})$ in $\mathbf{U}_a^*(\boldsymbol{\alpha})$ results in a biased estimator because

$$E\{\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta})^{\otimes 2}|\boldsymbol{\xi}_i\} = \mathbf{D}(\boldsymbol{\theta}) + \boldsymbol{\xi}_i^{\otimes 2},$$

where $\mathbf{D}(\boldsymbol{\theta}) = (\mathbf{B}^T\boldsymbol{\Psi}_\epsilon^{-1}\mathbf{B})^{-1}$ is a $q \times q$ matrix function of $\boldsymbol{\theta}$ and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ for a vector $\mathbf{a}$. Plugging in $\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta})$ as an estimate of $\boldsymbol{\xi}_i$ introduces error in the covariates, requiring correction as in the measurement error literature. A common technique to manage covariate measurement error is the corrected score method, of which the key idea is to remove the incurred bias through correction of the score function. The applications of such kinds of methods can be found in Carroll, Ruppert and Stefanski (1995) for non-censored data, in Nakumura (1992), Kong and Gu (1999), and Huang and Wang (2000) for Cox models with censored data, and in Kulich and Lin (2000) and Song and Huang (2006) for additive hazards models with censored data. This motivates us to develop the corrected estimating equation approach for the MRL model with latent variables. To deduct the bias, for given $\boldsymbol{\theta}$, we propose the corrected estimating function $\mathbf{U}_a(\boldsymbol{\alpha};\boldsymbol{\theta}) = (\mathbf{U}_{a1}(\boldsymbol{\alpha};\boldsymbol{\theta})^T, \mathbf{U}_{a2}(\boldsymbol{\alpha};\boldsymbol{\theta})^T)^T$, with

$$\mathbf{U}_{a1}(\boldsymbol{\alpha};\boldsymbol{\theta}) = \sum_{i=1}^{n}\int_0^\tau \{\mathbf{Z}_i - \bar{\mathbf{Z}}_a(t)\}\big\{\hat{m}_{a0}(t;\boldsymbol{\alpha},\boldsymbol{\theta}) + \boldsymbol{\beta}^T\mathbf{Z}_i + \boldsymbol{\gamma}^T\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta})\big\}dN_i(t),$$

$$\begin{aligned}\mathbf{U}_{a2}(\boldsymbol{\alpha};\boldsymbol{\theta}) = &\sum_{i=1}^{n}\int_0^\tau \{\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta}) - \bar{\boldsymbol{\xi}}_a^*(t;\boldsymbol{\theta})\}\big\{\hat{m}_{a0}(t;\boldsymbol{\alpha},\boldsymbol{\theta}) + \boldsymbol{\beta}^T\mathbf{Z}_i + \boldsymbol{\gamma}^T\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta})\big\}dN_i(t)\\ &- \mathbf{D}(\boldsymbol{\theta})\boldsymbol{\gamma}\sum_{i=1}^{n}\int_0^\tau dN_i(t),\end{aligned}$$

where $\bar{\boldsymbol{\xi}}_a^*(t;\boldsymbol{\theta}) = \sum_{i=1}^n Y_i(t)\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta})/\sum_{i=1}^n Y_i(t)$. Now $\mathbf{U}_a(\boldsymbol{\alpha};\boldsymbol{\theta})$ meets the need for correcting the bias because it can be shown that (Kulich and Lin (2000)) $E\{\mathbf{U}_a(\boldsymbol{\alpha};\boldsymbol{\theta})|(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\ldots,\boldsymbol{\xi}_n)\} = \mathbf{U}_a^*(\boldsymbol{\alpha}) + o_p(n^{1/2})$.

If $\boldsymbol{\theta}$ is estimated by $\hat{\boldsymbol{\theta}}$, we can solve $\mathbf{U}_a(\boldsymbol{\alpha};\hat{\boldsymbol{\theta}}) = 0$ to obtain the estimate of $\boldsymbol{\alpha}$. The explicit form of the estimator is

$$\hat{\boldsymbol{\alpha}}_a = -\left\{\frac{\partial \mathbf{U}_a(\boldsymbol{\alpha};\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\alpha}^T}\right\}^{-1}\mathbf{U}_a(0;\hat{\boldsymbol{\theta}}),$$

with detailed expression provided in Supplementary Material S2. The estimator of the baseline MRL function $m_0(t)$ on $t \in [0,\tau]$ is then given by $\hat{m}_{a0}(t) = \hat{m}_{a0}(t;\hat{\boldsymbol{\alpha}}_a,\hat{\boldsymbol{\theta}})$, where $\hat{m}_{a0}(t;\boldsymbol{\alpha},\boldsymbol{\theta})$ is defined in (3.4).

3.2. Inference of distribution-free factor analysis model

The true value of $\boldsymbol{\theta}$ (the vector of unknowns of $(\mathbf{B}, \boldsymbol{\Phi}, \boldsymbol{\Psi}_\epsilon)$ in model (2.1)) is denoted by $\boldsymbol{\theta}_0$. To estimate $\boldsymbol{\theta}$, under the normality assumption of $\boldsymbol{\xi}_i$ and $\boldsymbol{\epsilon}_i$, Pan et al. (2015) employed a commonly used EM algorithm to obtain the maximum likelihood estimator, it is not directly applicable here because the distributions of $\boldsymbol{\xi}_i$ and $\boldsymbol{\epsilon}_i$ are unspecified. We propose the use of an ADF-GLS approach based on the discrepancy function (Browne (1974, 1984); Lee (2007))

$$F(\boldsymbol{\theta}) = \frac{1}{2}\{\text{vec}(\mathbf{S} - \boldsymbol{\Pi}(\boldsymbol{\theta}))\}^T \mathbf{W}^{-1} \{\text{vec}(\mathbf{S} - \boldsymbol{\Pi}(\boldsymbol{\theta}))\},$$

where $vec(\cdot)$ denotes the operation that converts a matrix into a column vector by stacking the rows sequentially; $\mathbf{S} = (n-1)^{-1}\sum_{i=1}^n (\mathbf{V}_i - \bar{\mathbf{V}})^{\otimes 2}$ and $\boldsymbol{\Pi}(\boldsymbol{\theta}) = \mathbf{B}\boldsymbol{\Phi}\mathbf{B}^T + \boldsymbol{\Psi}_\epsilon$ are the sample and the theoretical covariance matrices of $\mathbf{V}_i$, respectively; $\bar{\mathbf{V}} = n^{-1}\sum_{i=1}^n \mathbf{V}_i$, $\mathbf{W}$ is a $p^2 \times p^2$ random weight matrix; $n^{1/2}\text{vec}(\mathbf{S} - \boldsymbol{\Pi}(\boldsymbol{\theta}_0))$ has an asymptotic covariance matrix, denoted by $\boldsymbol{\Sigma}^*(\boldsymbol{\theta}_0)$, and $\mathbf{W}$ is chosen to converge in probability to $\boldsymbol{\Sigma}^*(\boldsymbol{\theta}_0)$.

For $a, b, c, d = 1, \ldots, p$, let $\boldsymbol{\Sigma}^*(\boldsymbol{\theta}_0)[ab, cd]$ denote the $\{(a-1)p + b, (c-1)p + d\}$ element of $\boldsymbol{\Sigma}^*(\boldsymbol{\theta}_0)$, $\boldsymbol{\Pi}_{ab}(\boldsymbol{\theta}_0)$ denote the (a, b) element of $\boldsymbol{\Pi}(\boldsymbol{\theta}_0)$, and $\mathbf{V}_i[a]$ denote the ath element of $\mathbf{V}_i$. Then (Lee (2007)), $\boldsymbol{\Sigma}^*(\boldsymbol{\theta}_0)[ab, cd] = \sigma_{abcd} - \boldsymbol{\Pi}_{ab}(\boldsymbol{\theta}_0)\boldsymbol{\Pi}_{cd}(\boldsymbol{\theta}_0)$, where

$$\sigma_{abcd} = E\{\mathbf{V}_i[a] - E\mathbf{V}_i[a]\}\{\mathbf{V}_i[b] - E\mathbf{V}_i[b]\}\{\mathbf{V}_i[c] - E\mathbf{V}_i[c]\}\{\mathbf{V}_i[d] - E\mathbf{V}_i[d]\},$$

in which the expectation is taken with respect to $\mathbf{V}_i$.

A natural choice of $\mathbf{W}$ is $\mathbf{W}[ab, cd] = \mathbf{S}_{abcd} - \mathbf{S}_{ab}\mathbf{S}_{cd}$, where

$$\mathbf{S}_{abcd} = n^{-1}\sum_{i=1}^n \{\mathbf{V}_i[a] - \bar{\mathbf{V}}[a]\}\{\mathbf{V}_i[b] - \bar{\mathbf{V}}[b]\}\{\mathbf{V}_i[c] - \bar{\mathbf{V}}[c]\}\{\mathbf{V}_i[d] - \bar{\mathbf{V}}[d]\},$$

$\mathbf{S}_{ab}$ is the (a, b) element of $\mathbf{S}$, and $\bar{\mathbf{V}}[a] = n^{-1}\sum_{i=1}^n \mathbf{V}_i[a]$. Notably, $\mathbf{S}_{abcd}$ and $\mathbf{S}_{ab}$ are the empirical counterparts of σ_{abcd} and $\boldsymbol{\Pi}_{ab}(\boldsymbol{\theta}_0)$, respectively. Thus, $\mathbf{W}$ is basically the empirical estimate of $\boldsymbol{\Sigma}^*(\boldsymbol{\theta}_0)$.

The ADF-GLS estimator of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, is defined as the minimizer of $F(\boldsymbol{\theta})$. Asymptotic properties of $\hat{\boldsymbol{\theta}}$ are stated in Lemma 1 of Supplementary Material S1. Its proof in Browne (1984) is an extension to the proof of the asymptotic results of the maximum likelihood estimator provided in Browne (1974). The Newton-Raphson, Gauss-Newton, or Fletcher-Powell algorithm (Lee (2007)) can be employed to carry out the estimation. These algorithms involve the gradient and the Hessian matrix of $F(\boldsymbol{\theta})$ that have closed forms $\dot{F}(\boldsymbol{\theta}) = -\dot{\boldsymbol{\Pi}}(\boldsymbol{\theta})\mathbf{W}^{-1}\text{vec}(\mathbf{S} - \boldsymbol{\Pi}(\boldsymbol{\theta}))$ and $\ddot{F}(\boldsymbol{\theta}) = \dot{\boldsymbol{\Pi}}(\boldsymbol{\theta})\mathbf{W}^{-1}\dot{\boldsymbol{\Pi}}(\boldsymbol{\theta})^T - \ddot{\boldsymbol{\Pi}}(\boldsymbol{\theta})[\mathbf{I}_{p^*} \otimes \{\mathbf{W}^{-1}\text{vec}(\mathbf{S} - \boldsymbol{\Pi}(\boldsymbol{\theta}))\}]$, where $\dot{\boldsymbol{\Pi}}(\boldsymbol{\theta}) =$

$\partial(\text{vec}\boldsymbol{\Pi}(\boldsymbol{\theta}))^T/\partial\boldsymbol{\theta}$, $\ddot{\boldsymbol{\Pi}}(\boldsymbol{\theta}) = \partial(\text{vec}\dot{\boldsymbol{\Pi}}(\boldsymbol{\theta}))^T/\partial\boldsymbol{\theta}$. We use the Newton-Raphson algorithm to compute $\hat{\boldsymbol{\theta}}$ as follows. Set an initial value $\boldsymbol{\theta}^{(0)}$, for the current value $\boldsymbol{\theta}^{(r)}$, the estimate is updated by $\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} - \{\ddot{F}(\boldsymbol{\theta}^{(r)})\}^{-1}\dot{F}(\boldsymbol{\theta}^{(r)})$. The iteration stops, for example, when the difference of the two successive estimates is less than 0.001 for each element of $\boldsymbol{\theta}$.

4. Asymptotic Properties

The asymptotic properties of the ADF-GLS estimator $\hat{\boldsymbol{\theta}}$ was studied by Browne (1984). We adjust its statement and make use of the delta method, see Supplementary Material S1. In conjunction with the asymptotic theory for the MRL model (Chen and Cheng (2006)), we are able to obtain the asymptotics of CEE1 estimators. Let $\boldsymbol{\alpha}_0$ be the true value of $\boldsymbol{\alpha}$.

Theorem 1. *Under the regularity conditions* (C1)–(C4) *in Supplementary Materials* S1 *and* S2, $\hat{\boldsymbol{\alpha}}_a$ *is consistent for* $\boldsymbol{\alpha}_0$, *and* $n^{1/2}(\hat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_0)$ *is asymptotically normal with mean zero and covariance matrix consistently estimated by* $\hat{\mathbf{A}}_a^{-1}\hat{\boldsymbol{\Sigma}}_a\hat{\mathbf{A}}_a^{-1T}$, *with explicit expressions given in Supplementary Material* S2.

Theorem 2. *Under the regularity conditions* (C1)–(C4) *in Supplementary Materials* S1 *and* S2, $\hat{m}_{a0}(t)$ *converges in probability to* $m_0(t)$ *uniformly in* $t \in [0, \tau]$, *and* $n^{1/2}\{\hat{m}_{a0}(t) - m_0(t)\}$ *converges weakly on* $[0, \tau]$ *to a zero-mean Gaussian process whose covariance function at* (t, s) *can be consistently estimated by* $\hat{\Upsilon}_a(t, s) = n^{-1}\sum_{i=1}^n \hat{O}_{ai}(t)\hat{O}_{ai}(s)$, *with explicit expressions given in Supplementary Material* S2.

5. Goodness-of-fit Test

We propose a test procedure for assessing the goodness-of-fit of the additive MRL model (2.2). Let $\hat{\mathbf{Z}}_i^* = \left(\mathbf{Z}_i^T, (\boldsymbol{\Gamma}(\hat{\boldsymbol{\theta}})\mathbf{V}_i)^T\right)^T$, and

$$\hat{\Omega}_i(t) = \int_0^t \left[\{\hat{m}_{a0}(s) + \hat{\boldsymbol{\beta}}_a^T\mathbf{Z}_i + \hat{\boldsymbol{\gamma}}_a^T\boldsymbol{\Gamma}(\hat{\boldsymbol{\theta}})\mathbf{V}_i\}dN_i(s) - Y_i(s)d\{\hat{m}_{a0}(s) + s\}\right].$$

Lin, Wei and Ying (1993) used martingale-based residuals to check the adequacy of the Cox model and applied a similar idea to the model checking for the mean and rate models of recurrent events (Lin et al. (2000)). Here $\hat{\Omega}_i(t)$ is analogous to the martingale residual. We consider the process

$$\varphi(t; \mathbf{z}) = n^{-1/2}\sum_{i=1}^n I(\hat{\mathbf{Z}}_i^* \le \mathbf{z})\hat{\Omega}_i(t),$$

where $I(\hat{\mathbf{Z}}_i^* \le \mathbf{z})$ is 1 when each component of $\hat{\mathbf{Z}}_i^*$ is less than or equal to the

corresponding component of $\mathbf{z}$, and 0 otherwise. In Supplementary Material S4, we define $\tilde{\varphi}_i(t;\mathbf{z})$ and show that $\varphi(t;\mathbf{z})$ is asymptotically equivalent to the zero-mean Gaussian process $\tilde{\varphi}(t;\mathbf{z}) = n^{-1/2}\sum_{i=1}^{n}\tilde{\varphi}_i(t;\mathbf{z})$.

As estimating the asymptotic covariance function of $\varphi(t;\mathbf{z})$ analytically is difficult, we propose resampling (Lin, Wei and Ying (1993); Lin et al. (2000); Sun, Song and Zhang (2012)) and approximate $\varphi(t;\mathbf{z})$ with $\hat{\varphi}(t;\mathbf{z}) = n^{-1/2}\sum_{i=1}^{n}\tilde{\varphi}_i(t;\mathbf{z})\eta_i$, where $(\eta_1,\ldots,\eta_n)$ are independent standard normal variables independent of $\{(\mathbf{V}_i,\mathbf{Z}_i,\, X_i,\Delta_i)\}$ $(i=1,\ldots,n)$. Thus, we can obtain a large number of realizations from $\hat{\varphi}(t;\mathbf{z})$ by repeatedly generating standard normal random samples $(\eta_1,\ldots,\eta_n)$ while holding the observed data $\{(\mathbf{V}_i,\mathbf{Z}_i,X_i,\Delta_i)\}(i=1,\ldots,n)$ fixed to approximate the null distribution of $\varphi(t;\mathbf{z})$.

To assess the overall fit of the additive MRL model (2.2), we can plot the observed $\varphi(t;\mathbf{z})$ along with a few realizations from $\hat{\varphi}(t;\mathbf{z})$, and see how unusual the observed $\varphi(t;\mathbf{z})$ is under the posited model. Quantitatively, we can apply the supremum test statistic $\sup_{t;\mathbf{Z}}|\varphi(t;\mathbf{z})|$. The p-value of the test can be estimated by drawing a large number of realizations from $\sup_{t;\mathbf{Z}}|\hat{\varphi}(t;\mathbf{z})|$.

6. Simulation Study

In this section we report on simulations conducted to assess the finite sample performance of CEE1 in the presence of covariate-independent censoring in Section 6.1 and covariate-dependent censoring in Section 6.2.

6.1. Simulation 1

We considered a model defined by (2.1)) and (2.2)) with $p=6$, $q=2$, and $s=1$. The true values of parameters in model (2.1)) were set as

$$\mathbf{B}^T = \begin{bmatrix} 0.8 & 0.8 & 0.8 & 0^* & 0^* & 0^* \\ 0^* & 0^* & 0^* & 0.8 & 0.8 & 0.8 \end{bmatrix}, \quad \mathbf{\Phi} = \begin{bmatrix} 1^* & \phi_{12} \\ \phi_{21} & 1^* \end{bmatrix},$$

where the elements with asterisk are fixed. The $\boldsymbol{\epsilon}_i$ were $N(\mathbf{0},\mathbf{\Psi}_\epsilon)$ with $\mathbf{\Psi}_\epsilon = \mathrm{diag}(0.3,0.3,0.3,0.3,0.3,0.3)$. The diagonal elements of $\mathbf{\Phi}$ were fixed at 1.0 to identify the factor analysis model.

In the MRL model (2.2)), we generated Z_i independently from the Bernoulli distribution with a success probability of 0.5. To evaluate the empirical performance of the proposed methods under different distributions of $\boldsymbol{\xi}_i$, we considered the three cases: (I) $\boldsymbol{\xi}_i \sim N(\mathbf{0},\mathbf{\Phi})$; (II) $\boldsymbol{\xi}_i \sim \{\Gamma(4,2)-2\}\mathbf{I}$, where $\mathbf{I}$ is a two-dimensional identity matrix; and (III) $\boldsymbol{\xi}_i \sim 2/3N(\boldsymbol{\mu}_1,\mathbf{\Sigma}) + 1/3N(\boldsymbol{\mu}_2,\mathbf{\Sigma})$, where $\boldsymbol{\mu}_1 = (-0.5,-0.5)^T$, $\boldsymbol{\mu}_2 = (1.0,1.0)^T$, and $\{\sigma_{11},\sigma_{12},\sigma_{22}\}$ in $\mathbf{\Sigma}$ are $\{0.5,0.3,0.5\}$.

Table 1. Results of the MRL model in Simulation 1.

			$n = 500$				$n = 1{,}500$			
Case	CR	Par	Bias	SE	SEE	CP	Bias	SE	SEE	CP
(I)	10%	β	−0.021	0.081	0.082	0.957	−0.007	0.048	0.048	0.947
		γ_1	−0.001	0.048	0.046	0.943	−0.003	0.026	0.027	0.953
		γ_2	−0.013	0.048	0.046	0.936	−0.007	0.028	0.027	0.933
	30%	β	−0.023	0.097	0.098	0.952	−0.008	0.058	0.056	0.939
		γ_1	−0.001	0.058	0.055	0.941	−0.003	0.032	0.032	0.950
		γ_2	−0.012	0.056	0.055	0.943	−0.007	0.034	0.032	0.935
(II)	10%	β	−0.020	0.086	0.082	0.935	−0.006	0.046	0.048	0.953
		γ_1	−0.006	0.049	0.047	0.937	−0.003	0.028	0.027	0.947
		γ_2	−0.021	0.053	0.050	0.938	−0.012	0.030	0.030	0.933
	30%	β	−0.021	0.100	0.098	0.943	−0.007	0.055	0.056	0.947
		γ_1	−0.006	0.060	0.057	0.928	−0.003	0.034	0.033	0.940
		γ_2	−0.020	0.064	0.061	0.930	−0.012	0.037	0.036	0.938
(III)	10%	β	−0.021	0.083	0.082	0.933	−0.008	0.048	0.048	0.940
		γ_1	0.000	0.046	0.047	0.939	0.000	0.028	0.027	0.939
		γ_2	−0.011	0.048	0.047	0.948	−0.006	0.028	0.027	0.934
	30%	β	−0.026	0.100	0.098	0.935	−0.010	0.057	0.057	0.953
		γ_1	0.000	0.057	0.057	0.945	0.001	0.035	0.033	0.930
		γ_2	−0.011	0.058	0.056	0.944	−0.007	0.034	0.033	0.937

The elements ϕ_{12} and ϕ_{21} were 0.2 in (I) and (III), and 0 in (II). Notably, the distribution of $\mathbf{V}_i$ is determined by the distributions of $\boldsymbol{\xi}_i$ and $\boldsymbol{\epsilon}_i$ through the factor analysis model, and is thus normal in Case (I) but non-normal in Cases (II) and (III). To generate failure time T, the true value of $\boldsymbol{\alpha} = (\beta, \gamma_1, \gamma_2)$ was set as $(1, 0.2, 0.5)$ and $m_0(t)$ was taken from the Hall-Wellner family such that $m(t|Z_i, \boldsymbol{\xi}_i) = D_1 t + D_2 + \beta Z_i + \boldsymbol{\gamma}^T \boldsymbol{\xi}_i$ for $0 \le t \le -(D_2 + \beta Z_i + \boldsymbol{\gamma}^T \boldsymbol{\xi}_i)/D_1$, where $D_1 > -1$ and $D_2 > 0$. We set $D_1 = -0.9$ and $D_2 = 3$ in this simulation. The independent censoring time C was generated as Uniform$(0, c)$, where c was selected to yield censoring rates of 10% and 30%, respectively, and to ensure that the support of C be longer than the support of T for all Z_i and $\boldsymbol{\xi}_i$.

The simulation results were based on 1,000 replications with $n = 500$ and $n = 1{,}500$. The parameter estimates in the MRL model are summarized in Table 1, whereas those in the factor analysis model are reported in Supplementary Material S3. In these tables, Bias is the sampling mean of the estimate minus the true value, SE is the sampling standard error of the estimate, SEE is the sampling mean of standard error estimate, and CP is the 95% empirical coverage probability based on the normal approximation. For the methods and sample sizes considered, the biases of the estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are small, their estimated

Table 2. Comparison between the GLS and the EM-type methods.

	$\boldsymbol{\xi}_i \sim N(\mathbf{0}, \boldsymbol{\Phi})$, $n = 500$, CR = 10%							
	GLS				EM Algorithm			
Par	Bias	SE	SEE	CP	Bias	SE	SEE	CP
b_{11}	−0.004	0.039	0.038	0.934	0.001	0.039	0.038	0.945
b_{21}	−0.004	0.040	0.037	0.940	0.001	0.038	0.038	0.946
b_{31}	−0.005	0.039	0.037	0.940	0.001	0.038	0.038	0.950
b_{42}	−0.004	0.039	0.038	0.934	−0.000	0.040	0.038	0.939
b_{52}	−0.004	0.039	0.037	0.942	−0.001	0.039	0.038	0.949
b_{62}	0.003	0.037	0.037	0.950	0.000	0.038	0.038	0.949
	$\boldsymbol{\xi}_i \sim \{\Gamma(4,2) - 2\}\mathbf{I}$, $n = 1{,}500$, CR = 10%							
	GLS				EM Algorithm			
Par	Bias	SE	SEE	CP	Bias	SE	SEE	CP
b_{11}	−0.001	0.026	0.025	0.941	0.001	0.025	0.022	0.928
b_{21}	−0.002	0.026	0.025	0.940	0.001	0.025	0.022	0.909
b_{31}	−0.002	0.026	0.025	0.940	0.001	0.025	0.022	0.915
b_{42}	−0.001	0.025	0.025	0.960	0.001	0.025	0.022	0.918
b_{52}	−0.001	0.026	0.025	0.939	0.001	0.025	0.022	0.911
b_{62}	−0.001	0.026	0.025	0.950	0.001	0.025	0.022	0.924

standard errors are close to the sampling standard errors, and the CP's are close to the nominal level.

In addition, we examined how much efficiency the GLS would lose comparing to the EM-type method when the normal assumption of the factor analysis model was satisfied and how biased the EM-type method would be when the normality assumption was violated. We re-analyzed the simulated datasets using the EM-type method for the factor analysis in the two settings of Simulation 1: (1) $\boldsymbol{\xi}_i \sim N(\mathbf{0}, \boldsymbol{\Phi})$ with $n = 500$ and censoring rate 10%, and (2) $\boldsymbol{\xi}_i \sim \{\Gamma(4,2) - 2\}\mathbf{I}$ with $n = 1{,}500$ and censoring rate 10%. In the first setting, the GLS and EM-type methods perform almost the same except that the biases of the factor loadings (see the upper panel of Table 2) are slightly larger for the GLS approach than for the EM-type method. Thus, the efficiency loss of the GLS is not substantial. In the second setting, although both methods perform similarly for the MRL model, the EM-type method cannot provide correct 95% empirical coverage probabilities for the factor loadings (see the lower panel of Table 2). Thus, routinely using the EM-type method for a factor analysis with non-normal latent variables is problematic.

6.2. Simulation 2

To assess the performances of CEE1 in the presence of covariate-dependent

Table 3. Results of the MRL model in Simulation 2.

			$\boldsymbol{\xi}_i \sim N(\mathbf{0}, \boldsymbol{\Phi})$				$\boldsymbol{\xi}_i \sim \{\Gamma(4,2)-2\}\mathbf{I}$			
n	CR	Par	Bias	SE	SEE	CP	Bias	SE	SEE	CP
500	10%	β	−0.017	0.083	0.081	0.941	−0.015	0.085	0.080	0.940
		γ_1	−0.003	0.047	0.045	0.941	0.001	0.049	0.047	0.943
		γ_2	−0.008	0.046	0.047	0.944	−0.006	0.063	0.055	0.929
	50%	β	−0.013	0.147	0.143	0.943	−0.006	0.148	0.139	0.930
		γ_1	0.005	0.073	0.070	0.941	0.009	0.089	0.083	0.924
		γ_2	0.005	0.079	0.078	0.943	0.022	0.122	0.118	0.920
	73%	β	0.013	0.264	0.257	0.932	0.005	0.307	0.285	0.924
		γ_1	0.012	0.104	0.104	0.940	0.022	0.152	0.141	0.925
		γ_2	0.024	0.125	0.119	0.926	0.054	0.220	0.205	0.923
1,500	10%	β	−0.007	0.047	0.047	0.940	−0.006	0.047	0.047	0.950
		γ_1	0.002	0.027	0.026	0.947	0.001	0.028	0.028	0.951
		γ_2	−0.001	0.027	0.027	0.951	−0.001	0.034	0.033	0.941
	50%	β	−0.003	0.085	0.082	0.930	−0.002	0.085	0.080	0.944
		γ_1	0.008	0.041	0.041	0.959	0.009	0.050	0.049	0.937
		γ_2	0.007	0.049	0.045	0.939	0.016	0.068	0.066	0.938
	73%	β	0.015	0.150	0.145	0.944	0.008	0.164	0.160	0.943
		γ_1	0.016	0.059	0.060	0.932	0.014	0.080	0.080	0.942
		γ_2	0.016	0.074	0.068	0.937	0.025	0.126	0.118	0.926
3,000	10%	β	−0.003	0.032	0.033	0.962	−0.003	0.031	0.033	0.962
		γ_1	0.001	0.019	0.019	0.954	0.003	0.020	0.020	0.951
		γ_2	−0.000	0.020	0.019	0.942	0.002	0.024	0.023	0.943
	50%	β	−0.000	0.058	0.058	0.952	−0.003	0.060	0.057	0.939
		γ_1	0.007	0.028	0.029	0.955	0.010	0.035	0.035	0.931
		γ_2	0.006	0.033	0.032	0.927	0.008	0.044	0.046	0.952
	73%	β	−0.005	0.103	0.102	0.951	−0.007	0.119	0.113	0.938
		γ_1	0.009	0.041	0.042	0.950	0.009	0.058	0.057	0.950
		γ_2	0.010	0.051	0.048	0.934	0.015	0.084	0.083	0.937

censoring, we considered the same joint model under cases (I): $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2})^T \sim N(\mathbf{0}, \boldsymbol{\Phi})$ and case (II): $\boldsymbol{\xi}_i \sim \{\Gamma(4,2)-2\}\mathbf{I}$, as in Section 6.1. The parameter setup was the same, except for the covariate-dependent censoring

$$\lambda_C(t|Z_i, \boldsymbol{\xi}_i) = \lambda_{C0}(t) \exp(\kappa Z_i + \eta_1 \xi_{i1} + \eta_2 \xi_{i2}),$$

where $\lambda_C(t|Z_i, \boldsymbol{\xi}_i)$ is the hazard function of censoring time given Z_i and $\boldsymbol{\xi}_i$. We set $\kappa = 1.5$, $\eta_1 = 0.5$, and $\eta_2 = 1$. $\lambda_{C0}(t)$ was chosen to yield censoring rates of 10%, 50%, and 73%, where the censoring rate of 73% mimics that of the CKD study presented in Section 7.

We considered $n = 500, 1{,}500$, and 3,000, where the size 3,000 mimics that of the CKD study. The results of parameter estimates in the MRL model are

summarized in Table 3. The parameter estimates associated with the factor analysis model are similar to those presented in Section 6.1 and are not reported. We observe from Table 3 that the sample size, censoring rate, and the distribution of $\boldsymbol{\xi}_i$ all have impact on the estimation results. For instance, all the CP's are slightly less than the nominal level when $n = 500$. This is improved when $n =$ 1,500 and further improved when $n = 3{,}000$. Likewise, the performance of CP's is enhanced with a decrease of the censoring rate and/or a normally distributed $\boldsymbol{\xi}_i$.

To investigate the impact of the misspecification of $\mathbf{B}$ on the inference results, we conducted an additional simulation based on this model setup except that the factor loading matrix $\mathbf{B}^*$ was overlapping with $b^*_{12} = 0.3$. Based on the proposed model with $\mathbf{B}^*$, we generated 1,000 datasets under the setting of $n = 500$, CR $= 73\%$, and $\boldsymbol{\xi}_i \sim \{\Gamma(4, 2) - 2\}\mathbf{I}$. We analyzed the datasets by assuming a non-overlapping $\mathbf{B}$ ($b_{12} = 0$). The results (not reported) show that the parameter estimates in the CFA model are sensitive to the misspecification of $\mathbf{B}$, but those in the MRL model are relatively less sensitive.

7. Application

We applied the proposed method to the CKD study described in the Introduction. A main goal was to identify the potential risk factors that might influence the MRL function of CKD for diabetic patients. A total of 3,586 Chinese type 2 diabetic patients entered a 10-year prospective cohort study conducted by Hong Kong Diabetes Registry. The failure (clinical endpoint) of CKD was defined by DNP plus follow up estimate glomerular filtration rate (eGFR) < 60 (Song et al. (2009)). The failure time of CKD was calculated as the period from enrollment to the date of the first clinical endpoint or 31 January 2009, whichever came first. Thus, the data were censored at 31 January 2009 with 73% censoring rate. The information of patients was collected as follows: age at enrollment (Age), duration of diabetes (Duration), Sex (1 = female, 0 = male), WAIST, BMI, SBP, DBP, HbAlc, FPG, TC, HDL-C, and TG, where WAIST and BMI characterize the latent factor 'Obesity, ξ_1'; SBP and DBP summarize the latent factor 'Blood pressure, ξ_2'; HbAlc and FPG measure the latent factor 'Glycemia, ξ_3'; and TC, HDL-C, and TG group the latent factor 'Lipid, ξ_4'.

To analyze the data, we let $\mathbf{V} = (V_1, \ldots, V_9)^T =$ (WAIST, BMI, SBP, DBP, HbAlc, FPG, TC, HDL-C, TG)^T, $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3, \xi_4)^T =$ (Obesity, Blood pressure, Glycemia, Lipid)^T, and $\mathbf{Z} = (Z_1, Z_2, Z_3)^T =$ (Age, Duration, Sex)^T.

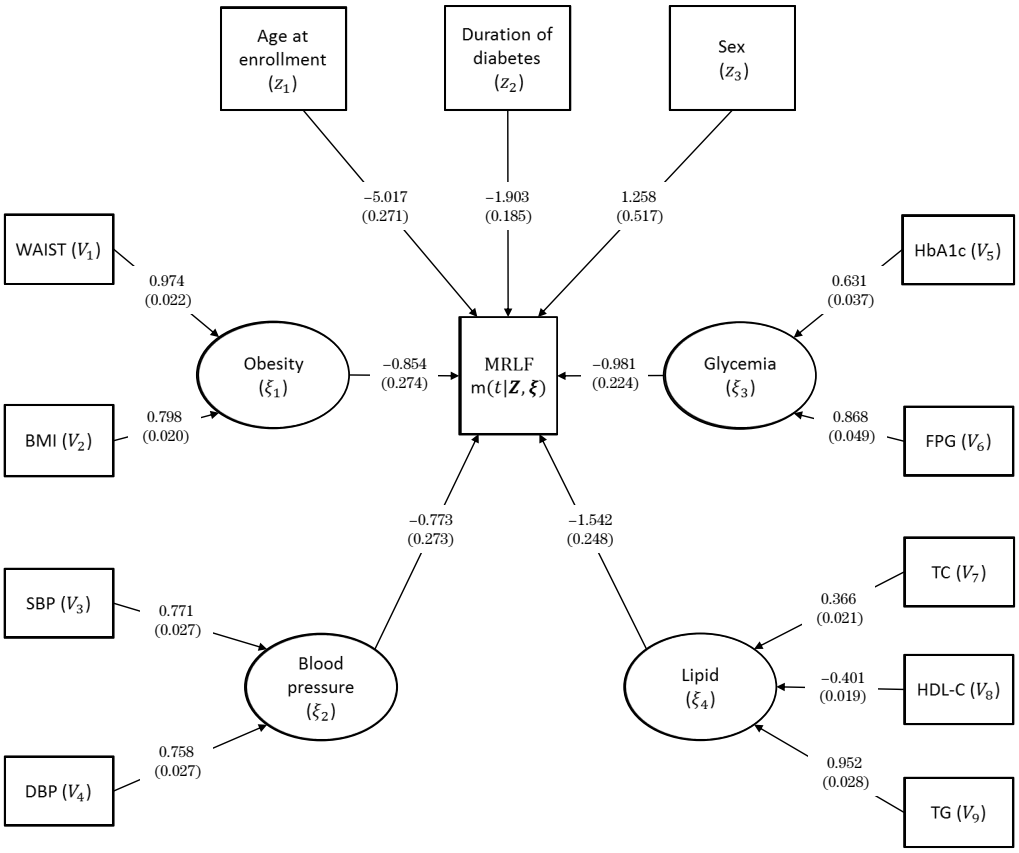

Figure 1. Path diagram of the proposed joint model, along with parameter estimates and their standard error estimates (in parentheses), in the analysis of the CKD data. In the path diagram, the latent variables are enclosed by ellipses, whereas the observed variables are enclosed by squares.

The observed variables in $\mathbf{V}$, as well as covariates Z_1 and Z_2, were standardized prior to the analysis. The factor loading matrix $\mathbf{B}$ was set as

$$\mathbf{B}^T = \begin{bmatrix} b_{11} & b_{21} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & b_{32} & b_{42} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & b_{53} & b_{63} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & b_{74} & b_{84} & b_{94} \end{bmatrix}. \tag{7.1}$$

This setting of the factor analysis model was well cross-validated by the result of an exploratory factor analysis (see Supplementary Material S5). Parameter estimates of primary interest, along with their standard error estimates obtained using CEE1, are reported in Figure 1. The results are interpreted as follows. Age at enrollment and duration of diabetes all have a significantly negative effect on

the MRL function of CKD, implying that elder patients and those suffering from diabetes in a longer duration are likely to develop CKD in a shorter time span. Sex has a significantly positive effect on the MRL function of CKD. The time span of evolving CKD is generally shorter for males than for females. As for latent risk factors, obesity, blood pressure, glycemia, and lipid all have a significantly negative effect on the MRL function. Thus, overweight, hypertension, bad glycemia control, and worse lipid profile would shorten the time span of developing CKD for diabetic patients. The estimated factor loadings are all highly significant, suggesting strong associations among latent factors and their corresponding observed indicators.

We re-conducted the analysis using a conventional additive MRL model by regarding the multiple observed indicators of the latent risk factors as independent covariates and directly incorporating them into the regression. Confounding results were obtained. The effect on the MRL function was significant for WAIST [−1.232 (0.486)] but nonsignificant for BMI [0.303 (0.507)], negative for SBP [−1.482 (0.310)] but positive (and nonsignificant) for DBP [0.480 (0.330)], significant for HbA1c [−1.438 (0.288)] but nonsignificant for FPG [−0.150 (0.295)], negative for TG [−1.994 (0.287)] and HDL-C [−0.405 (0.262)] but positive for TC [0.676 (0.311)]. As shown in the factor analysis (see Figure 1), all the factor loadings are substantially different from zero, implying that each observed indicator significantly contributes to characterization of the associated latent variable. Further, the factor loadings, except that corresponding to HDL-C, are all positive, implying that the observed indicators except HDL-C measure the associated latent variables in the same direction, and should likewise influence the MRL function of CKD in the same direction. These diverse effects are misleading. When checking the data, we found that the sample correlations between WAIST and BMI, SBP and DBP, HbA1c and FPG, TC and TG, as well as TC and HDL-C were 0.827, 0.605, 0.682, 0.364, and −0.437, respectively. Simultaneously incorporating these highly correlated variables into the regression elicits multicollinearity, thereby leading to the confounding results.

We further used the test procedure in Section 5 to examine the goodness-of-fit of the proposed model. We obtained $\sup_{t;\mathbf{Z}} |\varphi(t;\mathbf{z})| \approx 1.212$ and a p-value of 0.822 based on 1,000 realizations of the statistic $\sup_{t;\mathbf{Z}} |\hat{\varphi}(t;\mathbf{z})|$. This indicates that the proposed model fits the observed data well.

8. Discussion

Our joint modeling approach incorporates a distribution-free factor analysis

model and an additive MRL model. This joint modeling includes the capability to reveal both observed and latent risk factors of the MRL function of interest, to avoid multicollinearity caused by highly correlated covariates, and to provide comprehensible interpretation for the effects of those that really exist but cannot be measured by a single observed variable. We develop a borrow-strength estimation procedure by combining an ADF-GLS approach for the factor analysis model and a corrected estimating equation approach for the MRL model. Asymptotic properties of the proposed estimators are established. Model checking techniques are developed. The simulation results indicate that the method works well. The joint model is general and robust because the distributions of the latent risk factors and random errors are left unspecified. The utility of the methodology is demonstrated by an application to the CKD dataset.

The present study has several extensions. We can consider a joint model that consists of a factor analysis model and a proportional MRL model,

$$m(t|\mathbf{Z}_i, \boldsymbol{\xi}_i) = m_0(t) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i + \boldsymbol{\gamma}^T \boldsymbol{\xi}_i),$$

but this is beyond the scope of this paper. We have not thoughtfully discussed the efficiency of the method in the present study. An augmented IPCW estimating equation approach (Sun and Zhang (2009)), or a weighted estimating equation approach (Lin and Ying (1994); Chen and Cheng (2005, 2006)), could be used to obtain more efficient estimators. Sun and Zhang (2009) and Sun, Song and Zhang (2012) have studied a general class of transformed MRL models. The extension of incorporating latent variables into these modeling frameworks is of interest, but the feasibility of developing the corrected estimating equation approach requires further investigation.

Supplementary Materials

The supplementary materials contain five sections. Section S1 presents asymptotic results of the ADF-GLS estimator. Section S2 provides proofs of asymptotic results of CEE1 estimator. Section S3 presents CEE2 and additional simulation results. Section S4 shows weak convergence of $\varphi(t, \mathbf{z})$. Section S5 reports the exploratory factor analysis in the CKD study.

Acknowledgment

This research was supported by NSFC 11471277, 11231010, 11690015, 11601171, 11701387, and 11601343 from the National Natural Science Foundation of China, GRF grants 14305014 and 14601115 from the Research Grant Council

of the HKSAR, Foundation of Shenzhen University (2016011), Fundamental Research Funds for the Central Universities (2016YXMS005), and the direct grants of Chinese University of Hong Kong. The authors thank the Editor, an associate editor, and the referee for their valuable comments.

References

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association* **101**, 119–137.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, New York.

Browne, M. W. (1974). Generalized least square estimators in the analysis of covariance structures. *South African Statistical Journal* **8**, 1–24.

Browne, M. W. (1984). Asymptotic distribution free methods in analysis of covariance structure. *British Journal of Mathematical and Statistical Psychology* **37**, 62–83.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Chen, Y. Q. (2007). Additive regression of expectancy. *Journal of the American Statistical Association* **102**, 153–166.

Chen, Y. Q. and Cheng, S. C. (2005). Semiparametric regression analysis of mean residual life with censored survival data. *Biometrika* **92**, 19–29.

Chen, Y. Q. and Cheng, S. C. (2006). Linear life expectancy regression with censored data. *Biometrika* **93**, 303–313.

Chen, Y. Q., Jewell, N. P., Lei, X. and Cheng, S. C. (2005). Semiparametric estimation of proportional mean residual life model in presence of censoring. *Biometrics* **61**, 170–178.

Guess, F. and Proschan, F. (1988). Mean residual life: theory and application. in *Handbook of Statistics*, **7** (Edited by P. R. Krishnaiah and C. R. Rao) 215–224. North-Holland, New York.

Huang, Y. and Wang, C. Y. (2000). Cox regression with accurate covariates unascertainable: a nonparametri-correction approach. *Journal of the American Statistical Association* **95**, 1209–1219.

Jolliffe, I. T. (2002). *Principal component analysis*. 2nd Edition. Springer-Verlag, New York.

Jöreskog, K. G. and Sörbom, D. (1996). *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Scientific Software International.

Kong, F. H. and Gu, M. (1999). Consistent estimation in Cox proportional hazards model with covariate measurement errors. *Statistica Sinica* **9**, 953–969.

Kulich, M. and Lin, D. Y. (2000). Additive hazards regression with covariate measurement error. *Journal of the American Statistical Association* **95**, 238–248.

Lawley, D. N. and Maxwell, A. E. (1971). *Factor Analysis as A Statistical Method*. Butterworths, London.

Lee, S. Y. (2007). *Structural Equation Modeling: A Bayesian Approach*. John Wiley & Sons, New York.

Lin, D. Y., Wei, L. J. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–72.

Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Ser. B* (Statistical Methodology) **62**, 711–730.

Lin, D. Y. and Ying, Z. L. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61–71.

Maguluri, G. and Zhang, C.-H. (1994). Estimation in the mean residual life regression model. *Journal of the Royal Statistical Society, Ser. B* (Statistical Methodology) **56**, 477–489.

Nakumura, T. (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics* **48**, 829–838.

Oakes, D. and Dasu, T. (1990). A note on residual life. *Biometrika* **77**, 409–410.

Pan, D., He, H. J., Song, X. Y. and Sun, L. Q. (2015). Regression analysis of additive hazards model with latent variables. *Journal of the American Statistical Association* **110**, 1148–1159.

Roy, J. and Lin, X. (2000). Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics* **56**, 1047–1054.

Roy, J. and Lin, X. (2002). Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: changes in methadone treatment practices. *Journal of the American Statistical Association* **97**, 40–52.

Sammel, M. D. and Ryan, L. M. (1996). Latent variable models with fixed effects. *Biometrics* **52**, 650–663.

Song, X. and Huang, Y. (2006). A corrected pseudo-score approach for additive hazards model with longitudinal covariates measured with error. *Lifetime Data Analysis* **12**, 97–110.

Song, X. Y. and Lee, S. Y. (2012). *Basic and Advanced Bayesian Structural Equation Modeling: With Applications in the Medical and Behavioral Sciences*. John Wiley & Sons, London.

Song, X. Y., Lee, S. Y., Ma, R. C. W, So, W. Y., Cai, J. H., Tam, C., Lam, V., Ying, W., Ng, M. C. Y. and Chan, J. C. N. (2009). Phenotype-genotype interactions on renal function in type 2 diabetes: an analysis using structural equation modelling. *Diabetologia* **52**, 1543–1553.

Sun, L., Song, X. Y. and Zhang, Z. (2012). Mean residual life models with time-dependent coefficients under right censoring. *Biometrika* **99**, 185–197.

Sun, L. and Zhang, Z. (2009). A class of transformed mean residual life models with censored survival data. *Journal of the American Statistical Association* **104**, 803–815.

College of Mathematics and Statistics, Shenzhen University, Shenzhen, China.

E-mail: hehj@szu.edu.cn

School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan, China.

E-mail: pand@hust.edu.cn

Department of Statistics, The Chinese University of Hong Kong, Hong Kong.

E-mail: xysong@sta.cuhk.edu.hk

Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China.

E-mail: slq@amt.ac.cn

(Received October 2015; accepted February 2017)