# On Measurement Error Problems with Predictors Derived from Stationary Stochastic Processes and Application to Cocaine Dependence Treatment Data

Yongtao Guan, Yehua Li and Rajita Sinha

**Abstract**

In a cocaine dependence treatment study, we use linear and nonlinear regression models to model post-treatment cocaine craving scores and first cocaine relapse time. A subset of the covariates are summary statistics derived from baseline daily cocaine use trajectories, such as baseline cocaine use frequency and average daily use amount. These summary statistics are subject to estimation error and can therefore cause biased estimators for the regression coefficients. Unlike classical measurement error problems, the error we encounter here is heteroscedastic with an unknown distribution, and there are no replicates for the error-prone variables or instrumental variables. We propose

Yongtao Guan and Yehua Li are joint first authors of this paper. Yongtao Guan is Associate Professor, Division of Biostatistics, Yale School of Public Health, Yale University, New Haven, CT 06520, e-mail yongtao.guan@yale.edu. Yehua Li is Assistant Professor, Department of Statistics, University of Georgia, Athens, GA 30602, e-mail yehuali@uga.edu. Rajita Sinha is Professor of Psychiatry and Child Study Center, Yale University, New Haven, CT 06511, e-mail rajita.sinha@yale.edu. Yongtao Guan's research was supported by NSF grant DMS-0845368 and by NIH grant 1R01DA029081-01A1. Yehua Li's research was supported by NSF grant DMS-0806131. Rajita Sinha's research was supported by NIH grants P50DA016556 and UL1DE019586.

two robust methods to correct for the bias: a computationally efficient method-of-moment-based method for linear regression models and a subsampling extrapolation method that is generally applicable to both linear and nonlinear regression models. Simulations and an application to the cocaine dependence treatment data are used to illustrate the efficacy of the proposed methods. Asymptotic theory and variance estimation for the proposed subsampling extrapolation method and some additional simulation results are described in the online web supplementary material.
KEY WORDS: Bias Correction, Measurement Error, Method-of-Moment Correction, Subsampling Extrapolation.

# 1 Introduction

Cocaine dependence remains a serious public health problem in the US with more than one million individuals meeting criteria for current dependence (SAMSHA, 2004). Cocaine abuse causes serious concerns to the society because of its association with criminal activities as well as the high cost to treat health problems related to it. In the battle against cocaine addiction, cocaine craving and relapse pose the greatest challenges. Despite the availability of efficacious behavioral treatments, relapse rates remain high (Sinha, 2001, 2007).

Previous studies have revealed that one's baseline cocaine use behavior is predictive of cocaine relapse (Fox et al., 2006), along with many other risk factors such as age and gender, cocaine withdrawal severity (Kampman et al., 2001), and stress and negative mood (Sinha, 2001, 2007). To describe the baseline cocaine use behavior, daily cocaine use trajectory data are often collected in a short period prior to treatment. Summary statistics derived from these trajectories are then used as predictors in a subsequent analysis to explain cocaine craving and cocaine relapse. Among these, the most popular choices are the baseline cocaine use frequency and average daily use amount (Carroll et al., 1993; Sinha et al., 2006).

Because the baseline trajectories are random, we view them each as a realization from a stochastic process. The parameters governing the distribution of the process then character-

ize one's true baseline cocaine use behavior. However, the summary statistics derived from an observed realization are only estimates of these parameters and may often contain large error, leading to a measurement error problem. In a regression setting, the use of error-prone variables as predictors may cause severe bias to regression coefficients that are associated with these variables and those that are correlated with them (Carroll et al., 2006). Thus, the effect of measurement error due to the summary statistics must be accounted for.

We stress that the kind of measurement error problem faced here is different from the classical measurement error problems. For the latter, the measurement error is often assumed to have a known and constant variance and also follow certain distribution, e.g., a normal distribution. When the variance is unknown, replicates of the error-prone variable or some instrumental variables are needed in order to estimate the variance (Carroll et al., 2006). In the current setting, the summary statistics are derived based on only one single realization of the underlying process for each subject, i.e., there are no replicates, and the variances of the summary statistics are typically unknown and unequal across subjects. In addition, the errors in the summary statistics are usually non-Gaussian.

In the joint modeling literature, many authors have considered the so-called second level regression models, where the covariates are subject-specific regression parameters of some longitudinal measurements. A partial list of such work includes Wang et al. (2000), Song et al. (2002) and Li et al. (2004, 2007). In all these papers, strong parametric assumptions are made for the measurement error, such as that they are normally distributed. In a recent application in the Sleep Heart Health Study, Crainiceanu et al. (2009) proposed building a regression model to predict coronary heart disease status using summary statistics derived from the electroencephalogram (EEG) $\delta$-power spectra. They adopted a functional data analysis approach and proposed a simulation extrapolation (SIMEX) method to account for the measurement error due to smoothing the EEG curves, where the simulation step was done by bootstrapping the (independent) residuals. To account for the randomness in

each experiment, they proposed a Bayesian hierarchical model that jointly models the EEG data from multiple visits. Similar to the authors mentioned above, we are also interested in building a second level model where some covariates are summary statistics derived from the baseline cocaine use trajectories. It is becoming an increasingly more important problem in statistical practice to account for the estimation error in such summary statistics.

In this paper, we develop two robust bias-correction methods to account for the effect of measurement error when a subset of the predictors are summary statistics derived from some stationary stochastic processes. We consider both linear and nonlinear regression models. For the former, we propose a computationally simple method-of-moment bias-correction method, which is closely related to its counterpart in classical measurement problems. For the latter, we propose a novel subsampling extrapolation (SUBEX) method, which can yield approximately unbiased estimator in the presence of measurement error for a wide range of problems. The SUBEX is motivated by the widely used SIMEX method proposed by Cook and Stefanski (1994), which involves a simulation step to increase the variance in the measurement error and an extrapolation step to correct for the bias. The simulation step in SIMEX is difficult to apply here because of the challenges highlighted earlier. We instead develop a novel subsampling algorithm to achieve the variance inflation. The proposed methods do not assume any distribution for the measurement error, nor require knowledge on its variance or the availability of replicates.

We organize the remainder of the paper as follows. In Section 2 we describe a real dataset that has motivated our research. We discuss the proposed bias-correction methods in Section 3, and assess their numerical properties through simulation in Section 4. We then apply them to analyze the motivating data example in Sections 5. We conclude this article with a discussion in Section 6. Additional theoretical and simulation results are given in the online Supplementary Materials.

# 2 The Motivating Dataset

## 2.1 Description of data

One hundred and forty two cocaine dependent individuals between the age of 19 and 51 were admitted to the Clinical Neuroscience Research Unit (CNRU) of the Connecticut Mental Health Center, for two to four weeks of inpatient treatment for cocaine dependence. The CNRU is a locked inpatient treatment and research facility with no access to alcohol or drugs and limited access to visitors. The study was conducted in two separate periods with 59 enrolling in the first period and 83 enrolling in the second.

Upon treatment entry, all subjects were interviewed by means of the Structured Clinical Interview for DSM-IV (First et al., 1995) to collect baseline demographic variables and daily cocaine use history in the 90 days prior to admission. The latter was documented using a 90-day time-line follow-back (TLFB; Sobell and Sobell, 1993) Substance Use Calendar, which is a reliable instrument for assessing self-report drug use in alcoholic and drug abusing populations (Fals-Stewart et al., 2000; see Section 2.3 for more discussion on the TLFB). In addition, the Tiffany Cocaine Craving Questionnaire-Brief (CCQ-Brief) (Tiffany et al., 1993), which measures "desire for using cocaine at this moment", was also collected.

After completion of the inpatient treatment, all participants were invited back for follow-up interview(s) to assess cocaine use outcomes. For the 59 subjects in the first study, only one interview was administrated at day 90 after treatment. For the 83 subjects in the second study, four interviews were given at days 14, 30, 90 and 180 after treatment. Daily cocaine record was collected based on the TLFB procedure during each interview for the period before. Urine toxicology screen was conducted to check the accuracy of a reported relapse or abstinence in the same period. Some subjects reported abstinence but their urine samples tested positive. Hence, these subjects had relapsed in the period proceeding the first positive urine sample test date. However, the exact relapse date is unknown. Thus, our relapse data

are interval censored. In addition, the CCQ-Brief scores were also measured, typically only at day 90 for the first 59 subjects but at days 14, 90 and 180 for second 83 subjects.

## 2.2 Exploratory analysis of baseline trajectories

The 90-day baseline daily cocaine use trajectory data were given in the form of actual daily use amount (in gram equivalents) that was estimated by the study participants. However, the data may contain large errors due to two reasons. First, it is generally difficult to recall the exact use amount over a long period of time, at least more difficult than to recall use/no use, say. Second, it is challenging to develop a common scale to assess the amount used because of different methods of consumption. For example, some subjects may smoke cocaine in the form of crack, others may use it in powder form intranasally, and others may inject it; requiring all subjects to convert the amounts into the exact gram equivalents would simply be unrealistic. Given these reasons, we also consider dichotomized trajectories, i.e., trajectories comprised of no use (=0) and any use (=1). We illustrate the main ideas in the remainder of Section 2 based on such trajectories. However, we will also consider the trajectories of actual use amounts in Section 5.

Our primary goal here is to assess the effect of one's long-term cocaine use behavior on posttreatment cocaine craving and cocaine relapse. The vast majority of the study participants had had multiple years of cocaine use history before enrolling in the study. Because the baseline period is relatively short compared to the cocaine use history, it is sensible to assume that the observed baseline cocaine use pattern was stable and also reflective of one's long-term cocaine use behavior. However, the cocaine use behavior could have been altered temporarily because of the impending admission to treatment. To be sure, we have examined the plot of daily percentages of individuals reporting cocaine use (Figure 1, (a)). The plot indicates that there were more subjects using cocaine on the second last day before treatment admission but much fewer on the last day. It also reveals that the daily percent-

ages in the first six days appeared to be lower than average, which could be due to, e.g., recalling error. Since these data may not be most representative for one's long-term cocaine use behavior, we have removed them from our analysis. This leaves us with an 80-day long baseline period. Figure 1 (b) plots the baseline trajectories for four representative subjects. The trajectories suggest that the daily cocaine use patterns are reasonably consistent over the baseline period for these subjects.

We next study the dependence within each trajectory. Let $B_{ij}$ be a binary variable to denote a cocaine use (=1) or no use (=0) for the $i$th subject on the $j$th day, $i = 1, \cdots, 142$ and $j = 1, \cdots, 80$. Let $r \leq 40$ be a positive integer. Define

$$g_i(r) = \frac{1}{80} \sum_{j=1}^{80} B_{ij} B_{i(j-r)}.$$

If $j < r$, we set $B_{i(j-r)} = B_{i(80+j-r)}$. This is equivalent to assuming that the first and last days are next to each other. Note that $g_i(r)$ is a second-order statistic and thus may contain useful information about the dependence in a trajectory.

Figure 2 (a) shows values of $g_i(r)$ for the four selected trajectories shown in Figure 1 (b). It reveals a large variability among the individual $g_i(r)$, which range from consistently near-zero values for Trajectory 1 and a constant value of one for Trajectory 4. The former suggests a very low baseline frequency whereas the latter implies cocaine use everyday. The plot of $g_i(r)$ for Trajectory 2 reaches a peak every seven days, suggesting a weekly cocaine use pattern; however, the same plot for Trajectory 3 shows a slightly decreasing trend, implying that the dependence becomes weaker as the time lag $r$ increases. Combined together, the plot illustrates the highly complex nature of the dependence in these trajectories. Consequently, the distribution of the summary statistics and hence that of the resulting measurement error are also expected to be highly complex. A strong parametric assumption for the measurement error may then be unrealistic and potentially restrictive.

## 2.3 Construction of baseline summary statistics

We illustrate the construction of baseline summary statistics based on the four trajectories given in Figure 1 (b). A most obvious choice is the baseline frequency, expressed as

$$W_i = \frac{1}{80}\sum_{j=1}^{80} B_{ij}, \tag{1}$$

where $B_{ij}$ is as defined in Section 2.2. For the four given trajectories, the resulting baseline frequencies are 0.125, 0.725, 0.763 and 1, respectively. Note that Trajectories 2 and 3 yield similar baseline frequencies, although they look different both visually and as indicated by the plots of $g_i(r)$.

To better distinguish the patterns between Trajectories 2 and 3, we may consider

$$W_i^* = \sum_{r=1}^{r^*} \psi(r) g_i(r) = \frac{1}{80}\sum_{j=1}^{80}\left[\sum_{r=1}^{r^*} \psi(r) B_{ij} B_{i(j-r)}\right] \equiv \frac{1}{80}\sum_{j=1}^{80} B_{ij}^*, \tag{2}$$

where $r^*$ is a positive integer and $\psi(r)$ is a properly defined function. For example, if $r^* = 7$ and $\psi(r) = 0$ except when $r = 7$, then $W_i^*$ simply quantifies the strength of a weekly pattern. We may also set $\psi(r)$ as an (estimated) principal component of $g_i(r)$. For our data, the second and third principal components (Figure 3) describe variabilities due to weekly patterns and short-range dependence, respectively. The obtained summary statistics are -0.1945 and 0.0037 for Trajectory 2, and are 0.0688 and -0.0474 for Trajectory 3. Hence, Trajectory 2 exhibits more evidence for a weekly pattern while Trajectory 3 appears to have a stronger short-range dependence. The first principal component essentially yields a summary statistic that is comparable with the baseline frequency. Note that the summary statistics defined in (1) and (2) can both be viewed as the sample mean of a stochastic process associated with the $i$th subject. We will focus on this case in our paper.

The obtained summary statistics are random mainly due to two reasons. First, the baseline cocaine use trajectories, even if they can be accurately observed, are only realizations of some stochastic processes. Suppose that the baseline period could be repeated, then the

resulting new trajectories, and hence the summary statistics based on them, would most likely be different. This will be especially true for Trajectories 1 and 3. Such kind of randomness cannot be reduced, because it is an inherent part of the stochastic nature of one's cocaine use behavior. Second, the accuracy of a reconstructed baseline trajectory will depend on a number of factors, such as patient characteristics and their patterns of cocaine use as well as methods and settings in which data were collected. Some of these factors are random and therefore may introduce additional variability in the obtained baseline trajectories. The TLFB is designed to minimize such variabilities. It uses a calendar prompt and a number of other memory aids (e.g., the use of key dates such as holidays, birthdays, newsworthy events and other personal events as anchor points) to render more accurate recalls of daily cocaine uses. The subjects are also asked to recall lengthy periods of time when they completely abstained or used in a very regular pattern (e.g., used consecutively or weekly); patterns in such periods can often be identified more accurately. It has been shown that the TLFB can provide reliable daily cocaine use data that have high 1) retest reliability, 2) correlation with other cocaine use measures and 3) agreement with collateral informants' reports of patients' cocaine use as well as results obtained from urine assays (Fals-Stewart et al., 2000). Hence, the variability due to self-report should be small, especially for those with marked regular cocaine use patterns such as Trajectories 2 and 4.

Despite the well recognized effectiveness of TLFB, underreporting could still be a concern as with most other self-report data on use of illicit drugs. However, there are considerably less incentives for one to deny a use that occurred in a baseline period compared to during treatment and followup. Several studies have in fact reported a high degree of agreement between patients' self-report recent drug use and urine testing results obtained at the time of admission to treatment (Brown et al., 1992; Sherman and Bigelow, 1992). In lights of these facts, we do not consider the potential underreporting in the baseline data but will do so when analyzing the relapse data.

Another potential source of variability in the summary statistics is due to the use of an estimated function $\psi(r)$. However, this is often small when the sample size is large and can be eliminated when $\psi(r)$ is decided in advance (e.g., as is the case with baseline frequency). Our main focus of the paper is to account for the variability caused by the stochastic nature in one's cocaine use behavior, although our approach can also be applied when the baseline trajectories each are contaminated with a stationary error process of mean zero, i.e., when there is no systematic underreporting. The major challenge is that the distribution of the summary statistics may be very difficult to model, as a result of the highly complex within-trajectory dependence shown in Section 2.2. We will develop methods without having to specify the within-trajectory dependence.

# 3 Statistical Methods

## 3.1 Notation and setup

Let $N_i$ be a stationary stochastic process that has generated the $i$th cocaine use trajectory over the baseline period $(0, \tau]$, where $\tau > 0$. Let $W_i$ be a summary measure that is calculated from a realization of $N_i$. We assume that $W_i$ takes the following general form:

$$W_i = \frac{1}{\tau} \int_0^\tau N_i^*(dt), \tag{3}$$

where $N_i^*$ is either $N_i$ or a different stationary process that is derived from $N_i$. In other words, $W_i$ is the sample mean of $N_i^*$. We present $W_i$ in a more general integral form as given in (3), but its discrete counterpart can be easily derived; for example, see the baseline frequency and the second-order summary statistics given in (1) and (2). For ease of presentation but without losing generality, we simply write $N_i^*$ as $N_i$.

Suppose that the distribution of $N_i$ depends on a set of parameters $\Lambda_i$. Let $X_i = \mathrm{E}(W_i|\Lambda_i)$ be a desired predictor that describes some aspect of the long-term cocaine use behavior for

the $i$th subject, e.g., the long-term cocaine use frequency. Let $U_i$ be a measurement error such that $W_i = X_i + U_i$; this specifies the classical measurement error model (Carroll et al., 2006). The magnitude of $U_i$ can vary greatly across subjects. For example, for the four baseline trajectories that are given in Figure 1 (b), $U_i$ is expected to be small for Trajectories 2 and 4, but may be much larger for Trajectories 1 and 3. In regression analyses, the use of $W_i$ as a substitute for $X_i$ may lead to biased estimators for the regression coefficients.

To correct for the bias, it is often necessary to estimate the variability in $W_i$. When independent replicates of $W_i$ are available, this is rather straightforward (Carroll et al., 2006). However, independent replicates are not available for our data, because only one baseline trajectory was observed for each subject. However, for any given baseline period, we may obtain multiple equal-length subsets. Let $S(t,p) = (t, t + p\tau]$ be one such subset, where $0 < p \leq 1$ and $0 \leq t \leq (1-p)\tau$. The summary statistics defined on $S(t,p)$, denoted by $W_i(t,p)$, can then be treated as replicates of each other. We will derive our proposed bias-correction methods based on these replicates. However, $W_i(t,p)$ are not independent because of the within-trajectory dependence and also possible overlapping in the subsets.

Let $U_i(t,p)$ denote the measurement error that is associated with $W_i(t,p)$. By stationarity and condition (3), $W_i(t,p)$ is an unbiased estimator for $X_i$ and hence $U_i(t,p)$ has an expectation equal to zero when conditional on $\Lambda_i$. We next consider the variance of $U_i(t,p)$. Define $\tilde{\sigma}_i^2 = \lim_{\tau\to\infty}(\tau\sigma_{u_i}^2)$, where $\sigma_{u_i}^2$ is the variance of $U_i$ conditional on $\Lambda_i$. Let $\alpha_i$ be a constant that is typically nonnegative. We assume that

$$(p\tau)Var[U_i(t,p)|\Lambda_i] = \tilde{\sigma}_i^2 - \frac{1}{p\tau}\alpha_i + o\left(\frac{1}{p\tau}\right). \tag{4}$$

Condition (4) is a standard assumption for sample-mean-like statistics (Politis et al., 1999). For the motivating data given in Section 2, it essentially requires that the baseline trajectories are weakly correlated, and that the rate of decay in the correlation is dominated by the inverse of the length of the time interval. This condition is used mainly to ensure that our proposed methods are asymptotically valid as both $n$ and $\tau$ increase to infinity, where $n$ is

the total number of subjects. However, in a practical setting, $\tau$ is often small relative to the dependence range and hence the last term of (4) can be quite large compared to the two leading terms. We investigate the performance of our proposed methods under different levels of dependence through simulation (see Section 4).

Based on condition (4), we can link the variance of the measurement error in $W_i(t,p)$ to that in $W_i$. This link will be critical for the development of our proposed bias-correction methods based on $W_i(t,p)$. Specifically, it follows from straightforward algebra that

$$Var[U_i(t,p)|X_i] \approx \frac{\sigma_{u_i}^2}{p} - \frac{1-p}{p^2\tau^2}\alpha_i, \tag{5}$$

if $p > p_0$ for some $p_0 > 0$ such that $p_0\tau$ is sufficiently large. In the special case that $N_i$ is an independent process, then (5) holds exactly with $\alpha_i = 0$ and consequently $\sigma_{u_i}^2 = pVar[U_i(t,p)]$. However, this is unrealistic for our data given the evidence of within-trajectory dependence shown in Figure 2. In a more general setting, condition (5) implies that $pVar[U_i(t,p)]$ is biased for $\sigma_{u_i}^2$, but the bias can be reduced by adding back $(1-p)\alpha_i/(p\tau^2)$.

We introduce the following notation. Let $Y_i$ be the response and let $\mathbf{Z}_i$ be some covariates observed without error, both for the $i$th subject. Let $\mathbf{Y} = (Y_1, \cdots, Y_n)$ be an $n \times 1$ column vector. Similarly define $\mathbf{W}$, $\mathbf{X}$ and $\mathbf{U}$ as the $n \times 1$ vectors based on $W_i$, $X_i$ and $U_i$, respectively, and define $\mathbf{Z}$ as the matrix obtained by stacking together the row vectors $\mathbf{Z}_i^T$, where $\mathbf{Z}_i^T$ denotes the transpose of $\mathbf{Z}_i$.

## 3.2 Method-of-moment bias correction for linear models

We illustrate our proposed bias-correction method through the following simple model:

$$Y_i = X_i\beta + \mathbf{Z}_i^T\boldsymbol{\eta} + \epsilon_i,$$

where $\boldsymbol{\theta} = (\beta, \boldsymbol{\eta})$ is a vector of regression coefficients and $\epsilon_i : i = 1, \cdots, n$ are independent random variables with mean zero. Based on the observed data $(\mathbf{Y}, \mathbf{W}, \mathbf{Z})$, the least-squared

estimator for $\boldsymbol{\theta}$ can be written as

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{W}^T\mathbf{W} & \mathbf{W}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{W} & \mathbf{Z}^T\mathbf{Z} \end{bmatrix}^{-1} \left[ \begin{pmatrix} \mathbf{W}^T \\ \mathbf{Z}^T \end{pmatrix} \mathbf{Y} \right] \equiv \mathbf{A}^{-1}\mathbf{B}.$$

Note that $\mathbf{W} = \mathbf{X} + \mathbf{U}$. It follows from some simple algebra that

$$\mathrm{E}(\mathbf{A}|\mathbf{X}, \mathbf{Z}) = \begin{pmatrix} \mathbf{X}^T\mathbf{X} + \sigma^2 & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} \end{pmatrix},$$

$$\mathrm{E}(\mathbf{B}|\mathbf{X}, \mathbf{Z}) = \begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} \end{pmatrix} \boldsymbol{\theta},$$

where $\sigma^2 = \mathrm{E}(\mathbf{U}^T\mathbf{U}) = \sum_{i=1}^n \sigma_{u_i}^2$. Clearly $\hat{\boldsymbol{\theta}}$ is biased for $\boldsymbol{\theta}$ because $\sigma^2 \neq 0$. If an estimator for $\sigma^2$ is available, we can then obtain a method-of-moment bias-corrected estimator, say $\hat{\boldsymbol{\theta}}_{mom}$, by subtracting the variance estimator from the first element of $\mathbf{A}$. Thus, the problem has become a variance estimation problem.

We develop a nonparametric variance estimator for $\sigma^2$. A nonparametric estimator is desirable because of the complex within-trajectory dependence. To develop our method, let $t$ be an arbitrary time point satisfying $0 \leq t < \tau - kp\tau$, where $k = [1/p]$ and $[x]$ denotes the integer part of a positive constant $x$. We require $p \leq 1/2$ so that $k \geq 2$. We then partition the interval $(t, t+kp\tau]$ into $k$ nonoverlapping subintervals each of length $p\tau$. Let $W_i(t+jp\tau, p)$ be the summary statistic defined on the $(j+1)$th subinterval for the $i$th trajectory, where $j = 0, 1, \cdots, k-1$. Define

$$\tilde{\sigma}_{u_i}^2(p) = \frac{1}{\tau - kp\tau} \int_0^{\tau - kp\tau} \tilde{\sigma}_{u_i}^2(t, p)dt,$$

where

$$\tilde{\sigma}_{u_i}^2(t, p) = \frac{1}{k} \sum_{j=0}^{k-1} \left[ W_i(t + jp\tau, p) - W_i(t, kp) \right]^2.$$

In the special case that $\tau = kp\tau$, we simply set $\tilde{\sigma}_{u_i}^2(p) = \tilde{\sigma}_{u_i}^2(0, p)$.

By condition (3), the average of the summary statistics $W_i(t + jp\tau, p)$ is $W_i(t, kp)$. Thus, $\tilde{\sigma}^2_{u_i}(t, p)$ intends to estimate the variance of $U_i(t, p)$. Define $\alpha = \sum_{i=1}^n \alpha_i$. Recall that $W_i(t + jp\tau, p)$ is an unbiased estimator for $X_i$. It then follows from condition (5) that

$$\mathrm{E}\left[\sum_{i=1}^{n} \tilde{\sigma}^2_{u_i}(p)\right] \approx \frac{1}{p}\left(1 - \frac{1}{k}\right)\sigma^2 - \frac{1}{p^2\tau^2}\left(1 - p - \frac{1 - kp}{k^2}\right)\alpha. \tag{6}$$

In light of (6), define

$$\tilde{Y}(p) = \frac{p\sum_{i=1}^n \tilde{\sigma}^2_{u_i}(p)}{1 - 1/k} \equiv \sum_{i=1}^{n} \tilde{Y}_i(p) \text{ and } \tilde{X}(p) = \frac{1 - p - (1 - kp)/k^2}{p\tau^2(1 - 1/k)}. \tag{7}$$

We propose to regress $\tilde{Y}(p)$ on $\tilde{X}(p)$ at some preselected values of $p$, say $(p_1, \cdots, p_J)$ where $J \geq 2$. The resulting intercept and (minus) slope estimators, denoted by $\hat{\sigma}^2$ and $\hat{\alpha}$, can then be treated as estimators for $\sigma^2$ and $\alpha$, respectively. Moreover, a scatter plot of $\tilde{X}(p)$ and $\tilde{Y}(p)$ can be used as a practical tool to assess whether condition (4) is reasonable.

To understand the contributions of individual trajectories to the calculation of $\hat{\sigma}^2$, we plot $\tilde{Y}_i(p)$ against $\tilde{X}(p)$ in Figure 2 (b) for the four trajectories shown in Figure 1 (b); here $W_i$ is the baseline frequency. Note that Trajectory 3 has significantly larger $\tilde{Y}_i(p)$ values than the other three trajectories. This reflects the greatly increased uncertainty in the empirical baseline frequency given the strong within-trajectory dependence for Trajectory 3. It is also interesting to see that Trajectories 2 and 4 have the smallest $\tilde{Y}_i(p)$ values, as a result of their highly regular patterns. Combined together, these observations suggest that $\hat{\sigma}^2$ is mostly affected by trajectories that exhibit strong within-trajectory dependence, but much less so by those that possess highly regular patterns.

### 3.3 A subsampling extrapolation approach

The method-of-moment bias-correction method discussed in the last section is limited to linear models. Cook and Stefanski (1994) proposed a SIMEX method which can yield approximately unbiased estimators in the presence of measurement error for both linear and

nonlinear models. The key idea is to generate some pseudo "remeasurements", $W_i(\zeta)$, where $\text{Var}[W_i(\zeta)|X_i] = (1+\zeta)\sigma^2_{u_i}$, $i = 1, \cdots, n$. This is typically achieved through simulation by adding to each error-prone variable $W_i$ an independent error with variance equal to $\zeta\sigma^2_{u_i}$ when $\sigma^2_{u_i}$ is known, or by taking suboptimal linear combinations of the replicate measurements when $\sigma^2_{u_i}$ is unknown but independent replicates of $W_i$ are available (Devanarayan and Stefanski, 2002). Regression analysis without any bias correction is then conducted in terms of $W_i(\zeta)$ to obtain $\hat{\boldsymbol{\theta}}(\zeta)$. Because $\zeta = -1$ corresponds to the case of no measurement error, a bias-corrected estimator can then be obtained by extrapolating a fitted model to $\hat{\boldsymbol{\theta}}(\zeta)$ at $\zeta = -1$.

It is difficult to apply the simulation step of SIMEX here, because there is only one baseline trajectory observed per subject and the variance of the summary statistic defined on it is also unknown. The lack of independent replicates is common for data arising from stochastic processes such as time series and spatial data. If the within-process dependence is reasonably weak, one can create approximately independent subreplicates using subsampling techniques. The obtained subreplicates can then be used in applications such as estimating the variance and distribution of a general statistic calculated from the data (Politis et al., 1999). In the current setting, the use of subsampling entails considering the summary statistics $W_i(t,p)$ that are defined over the subintervals $S(t,p) = (t, t+p\tau]$, where $0 \le t \le (1-p)\tau$ and $p_0 < p < 1$ for some $p_0 > 0$. By condition (5), we can immediately conclude

$$\text{Var}[W_i(t,p)|\Lambda_i] \approx \frac{\sigma^2_{u_i}}{p} - \frac{1-p}{p^2\tau^2}\alpha_i = \left(\frac{1}{p} - \frac{1-p}{p^2\tau^2}\frac{\alpha_i}{\sigma^2_{u_i}}\right)\sigma^2_{u_i} \equiv x_i(p)\sigma^2_{u_i}. \tag{8}$$

Because $x_i(p)$ is typically larger than one, $W_i(t,p)$ has a variance that is larger than $\sigma^2_{u_i}$.

The above interesting observation, together with the fact that $W_i(t,p)$ is an unbiased estimator for $X_i$, has inspired us to treat $W_i(t,p)$ as a pseudo remeasurement for $W_i$. Regression analysis without accounting for measurement error can be done based on data $\{Y_i, W_i(t_i,p), \mathbf{Z}_i\}$, where $t_i \in [0,(1-p)\tau]$, $i = 1, \cdots, n$. Let $\hat{\boldsymbol{\theta}}(p; t_1, \cdots, t_n)$ denote the

resulting estimator. We then integrate out the $t_i$'s and obtain

$$\hat{\boldsymbol{\theta}}(p) = \frac{1}{[(1-p)\tau]^n} \int_0^{(1-p)\tau} \cdots \int_0^{(1-p)\tau} \hat{\boldsymbol{\theta}}(p; t_1, \cdots, t_n) dt_1 \cdots dt_n. \tag{9}$$

It is generally difficult to calculate (9) exactly. We use Monte-Carlo average of $\hat{\boldsymbol{\theta}}(p; t_1, \cdots, t_n)$ based on multiple randomly selected $t_i$'s as an estimate for (9), where $0 \leq t_i \leq (1-p)\tau$.

To correct for the bias, we also need to obtain $x_i(p)$, where

$$x_i(p) = \frac{1}{p}\left(1 - \frac{1-p}{p\tau^2}\frac{\alpha_i}{\sigma_{u_i}^2}\right).$$

Because $\alpha_i$ and $\sigma_{u_i}^2$ are both unknown, $x_i(p)$ must be estimated. One quick approximation is to set $x_i(p) \approx 1/p$. This is valid if $\alpha_i/(\tau^2\sigma_{u_i}^2)$ is small and/or $p$ is large so that $(1-p)\alpha_i/(p\tau^2\sigma_{u_i}^2)$ is much smaller than one. Alternatively we may estimate $\alpha_i/\sigma_{u_i}^2$ in $x_i(p)$ by $\hat{\alpha}/\hat{\sigma}^2$. This can lead to less biased estimators if $\alpha_i/\sigma_{u_i}^2$ is approximately constant. In either case, the estimate is free of $i$ so we rewrite it as $\hat{x}(p)$.

Following Cook and Stefanski (1994), we propose to fit a parametric model to $(\hat{x}(p), \hat{\boldsymbol{\theta}}(p))$ based on some preselected values of $p$, where $\hat{\boldsymbol{\theta}}(p)$ is as defined in (9). We then use the extrapolated value at $p = \infty$ from the fitted model as our bias-corrected estimate for $\boldsymbol{\theta}$. The validity of extrapolating at $p = \infty$ can be justified in two ways. First, note that $p = \infty$ means that each $N_i$ is observed over an infinitely long time interval. This is equivalent to that $X_i$'s are observed without error, since $W_i$ converges to $X_i$ almost surely assuming ergodicity (Daley and Vere-Jones, 2008). Second, note that the extra variance added to $W_i$ is approximately $(1/p-1)\sigma_{u_i}^2$ under our approach, but would be $\zeta\sigma_{u_i}^2$ if SIMEX could be used. Solving $1/p - 1 = \zeta$ for $\zeta = -1$ leads to $p = \infty$. From this perspective, our extrapolation criterion coincides with that used for SIMEX.

To conduct the extrapolation, we also need a parametric model for $(\hat{x}(p), \hat{\boldsymbol{\theta}}(p))$. We consider a quadratic model defined in terms of $\hat{x}(p)$,

$$\mathcal{G}_Q(p, \boldsymbol{\Gamma}) = \gamma_1 + \hat{x}(p)\gamma_2 + \hat{x}(p)^2\gamma_3,$$

where $\mathbf{\Gamma} = (\gamma_1, \gamma_2, \gamma_3)$. Carroll et al. (2006) suggested that the SIMEX version of the above model often works well in practice. In our simulation, we also consider a cubic model,

$$\mathcal{G}_C(p, \mathbf{\Gamma}) = \gamma_1 + \hat{x}(p)\gamma_2 + \hat{x}(p)^2\gamma_3 + \hat{x}(p)^3\gamma_4,$$

with the understanding that now $\mathbf{\Gamma} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)$.

The asymptotic theory for the proposed SUBEX estimator is derived in our online appendix. Following similar lines as those in Carroll et al. (1996), we show that the SUBEX estimator has an asymptotically normal distribution. Inspired by the theory, we also propose a sandwich formula in the appendix to estimate the variance of the estimator. Our simulation results confirm that the sandwich formula works well.

## 3.4 Comparison to two naive alternatives

It is intuitively appealing to assume independence between subsets of a baseline cocaine use trajectory observed over the intervals $(0, \tau/2]$ and $(\tau/2, \tau]$. Recall that $W_i(0, 1/2)$ and $W_i(\tau/2, 1/2)$ are the counterparts of $W_i$ based on the data in $(0, \tau/2]$ and $(\tau/2, \tau]$, respectively. In the linear model case, this assumption leads to the naive variance estimator,

$$\hat{\sigma}^2_{naive} = \sum_{i=1}^{n} \tilde{\sigma}^2_{u_i}(1/2) = \frac{1}{4}\sum_{i=1}^{n} \left[W_i(0, 1/2) - W_i(\tau/2, 1/2)\right]^2.$$

In the nonlinear model case, we may instead use the pseudo "remeasurements":

$$W_i(\zeta) = W_i + \frac{\sqrt{\zeta}}{2}\left[W_i(0, 1/2) - W_i(\tau/2, 1/2)\right], \text{ for each } \zeta > 0.$$

A bias-corrected estimator can then be obtained based on $\hat{\sigma}^2_{naive}$ instead of $\hat{\sigma}^2_{mom}$ in the linear case, and SIMEX can be applied based on the pseudo "remeasurements" $W_i(\zeta)$ in the nonlinear case. The latter is simply a special case of the empirical SIMEX when two independent replicates are available for the error-prone variable (Devanarayan and Stefanski, 2002).

For the naive variance estimator, it follows from (6) that

$$E\left(\hat{\sigma}^2_{naive}\right) \approx \sigma^2 - \frac{2\alpha}{\tau^2}.$$

Thus, $\hat{\sigma}^2_{naive}$ underestimates $\sigma^2$ if $\alpha > 0$, where the bias is of order $\tau^{-2}$ and can be approximated by $-2\hat{\alpha}/(\tau^2\hat{\sigma}^2)$. This immediately implies that $\hat{\sigma}^2_{naive}$ is not consistent for $\sigma^2$ unless $\tau \to \infty$. However, for our motivating data $\tau$ cannot be too large, because of the difficulty to reconstruct a reliable baseline trajectory over an extensively long period of time.

In contrast, our proposed estimator $\hat{\sigma}^2$ has a much smaller bias, which in turn will yield less biased regression coefficient estimates. To see this point, suppose that $N_i$ is $m$-dependent, i.e., the dependence in $N_i$ completely vanishes beyond a lag distance $m$. Then it is easy to show that (4) becomes

$$(p\tau)Var[U_i(t,p)|\Lambda_i] = \tilde{\sigma}_i^2 - \frac{1}{p\tau}\alpha_i, \text{ if } p\tau \geq m.$$

Suppose that $m/\tau \leq p \leq 1/2$. Then (6) becomes

$$E\left[\sum_{i=1}^{n} \tilde{\sigma}^2_{u_i}(p)\right] = \frac{1}{p}\left(1 - \frac{1}{k}\right)\sigma^2 - \frac{1}{p^2\tau^2}\left(1 - p - \frac{1-kp}{k^2}\right)\alpha.$$

The above relationship implies that $\hat{\sigma}^2$ is an unbiased estimator for $\sigma^2$. Write $\hat{\sigma}^2_n$ for $\hat{\sigma}^2$ to make explicit its dependence on $n$. By the strong law of large numbers, $\frac{1}{n}\sum_{i=1}^{n} \tilde{\sigma}^2_{u_i}(p)$ converges to its expected value almost surely. It then follows that $\hat{\sigma}^2_n$ is consistent for $\sigma^2$, in the sense that $\hat{\sigma}^2_n/\sigma^2 \to 1$ almost surely as $n \to \infty$. This in turn implies that the bias-corrected estimator $\hat{\boldsymbol{\theta}}_{mom}$ is consistent for $\boldsymbol{\theta}$. Note that the consistency result requires only $\tau > 2m$ and $n \to \infty$ but not $\tau \to \infty$.

To further illustrate their differences, we estimate $\sigma^2_{u_i}$ by our proposed variance estimator and the naive estimator for the four baseline trajectories given in Figure 1 (b), where $W_i$ is the baseline frequency. Specifically, the former estimator is obtained by regressing $\tilde{Y}_i(p)$ against $\tilde{X}(p)$, where $\tilde{Y}_i(p)$ and $\tilde{X}(p)$ are as defined in (7), and the latter estimator is simply $\tilde{\sigma}^2_{u_i}(1/2)$. Our proposed variance estimator yields a much larger estimate (=0.0308) than

the naive one (=0.0264) for Trajectory 2; this is expected given its relatively strong within-trajectory dependence. The estimates for the other three trajectories are all similar. These results demonstrate the difference by accounting for the negative bias in the naive estimator when there is a strong within-trajectory dependence.

For the new pseudo "remeasurements" $W_i(\zeta)$, it follows from condition (5), the definition of $W_i$ and stationarity of $N_i$ that

$$\mathrm{Var}\left[W_i(\zeta)|\Lambda_i\right] = \mathrm{Var}(W_i|\Lambda_i) + \frac{\zeta}{4}\mathrm{Var}\left[W_i(0,1/2) - W_i(\tau/2,1/2)|\Lambda_i\right] \approx \left(1 + \zeta - \frac{2\zeta\alpha_i}{\tau^2\sigma_{u_i}^2}\right)\sigma_{u_i}^2.$$

Clearly, the amount of variance inflation is smaller than the desired amount, i.e., $\zeta\sigma_{u_i}^2$. Let $\zeta = 1/p - 1$. Then, we can rewrite the above as

$$\mathrm{Var}\left[W_i(\zeta)|\Lambda_i\right] \approx \left[\frac{1}{p} - \frac{2(1-p)\alpha_i}{p\tau^2\sigma_{u_i}^2}\right]\sigma_{u_i}^2.$$

For our proposed SUBEX, we use $W_i(t,p)$ as the pseudo "remeasurements", where $W_i(t,p)$ is the counterpart of $W_i$ defined on $(t, t+p\tau]$. Recall from (8) that

$$\mathrm{Var}\left[W_i(t,p)|\Lambda_i\right] \approx \left[\frac{1}{p} - \frac{(1-p)\alpha_i}{p^2\tau^2\sigma_{u_i}^2}\right]\sigma_{u_i}^2.$$

If $\alpha_i \geq 0$ and $p \geq 1/2$, then $\mathrm{Var}\left[W_i(t,p)|\Lambda_i\right] \geq \mathrm{Var}\left[W_i(\zeta)|\Lambda_i\right]$, where $\zeta = 1/p - 1$. Thus, the variance of $W_i(t,p)$ is closer to the target (inflated) variance $(1+\zeta)\sigma_{u_i}^2$ than that of $W_i(\zeta)$. This implies that the proposed SUBEX with $\hat{x}(p) = 1/p$ can be more effective in bias correction than the empirical SIMEX. Moreover, it suggests that we should set $p \geq 1/2$. The proposed SUBEX with $\hat{x}(p) = 1/p - (1-p)\hat{\alpha}/(p^2\tau^2\hat{\sigma}^2)$ is generally even more effective in bias reduction. This can be seen from our simulation results in Section 4.

### 3.5 Extension to nonstationary processes

We argued in Section 2.2 that it is reasonable to assume stationarity for the baseline trajectories in our motivating data. However, there may be situations where such an assumption is questionable. To generalize our methods, let $N_i$ be a nonstationary stochastic process with

mean given by $X_i + \lambda(t)$, where $\lambda(t)$ changes over time but is fixed across processes at any given $t$. For identifiability let $\int_0^\tau \lambda(t)dt = 0$. Define $W_i$, $U_i$ and $W_i(t,p)$ as in Section 3.1. It is easy to see that $\mathrm{E}(U_i|\Lambda_i) = 0$. Also define

$$W_i^*(t,p) = W_i(t,p) - \frac{1}{p\tau}\int_t^{t+p\tau} \lambda(t)dt,$$

and

$$U_i(t,p) = W_i^*(t,p) - X_i.$$

Clearly, $\mathrm{E}[W_i^*(t,p)|\Lambda_i] = X_i$ and hence $\mathrm{E}[U_i(t,p)|\Lambda_i] = 0$.

Suppose that the variance of $U_i(t,p)$ satisfies condition (4). In the linear model case, our proposed method-of-moment bias-correction method can be extended to the current setting by replacing $W_i(t,p)$ therein with $W_i^*(t,p)$. The validity of this approach follows analogously as in the stationary case. In the nonlinear model case, observe the facts that $\mathrm{E}[W_i^*(t,p)|\Lambda_i] = \mathrm{E}(W_i|\Lambda_i)$ and

$$\mathrm{Var}\,[W_i^*(t,p)|\Lambda_i] = \mathrm{Var}\,[U_i(t,p)|\Lambda_i] \approx \left[\frac{1}{p} - \frac{(1-p)\alpha_i}{p^2\tau^2\sigma_{u_i}^2}\right]\sigma_{u_i}^2.$$

We therefore use $W_i^*(t,p)$ as the pseudo "remeasurements". In practice we replace $\lambda(t)$ by its consistent estimator $\hat{\lambda}(t)$, given that such an estimator exists.

# 4 Simulation Study

## 4.1 Simulation 1: linear model

We simulate independent tuples $\{Y_i, N_i, Z_i\}$, for $i = 1, \cdots, n$. Here $N_i$ is a stationary point process with intensity $X_i$, where $X_i \sim 1+$ Log Normal $(0, 0.5)$; $Z_i$ is an error-free Bernoulli random variable independent of $X_i$ and with mean equal to 0.5; and $Y_i$ is given by

$$Y_i = \theta_0 + \theta_1 X_i + \theta_2 Z_i + \epsilon_i,$$

where $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)^T = (1, 1, 1)^T$ and $\epsilon_i \sim \text{Normal}(0, 0.5^2)$. We assume that $X_i$ is unknown but can be estimated by $W_i = N_i((0, \tau])/\tau$, where $N_i((0, \tau])$ denotes the number of events of $N_i$ contained in the time interval $(0, \tau]$. Note that $W_i$ takes the general form given in (3).

We consider two scenarios for $N_i$. Under the first scenario, $N_i$ is a homogeneous Poisson process. To simulate $N_i$, we first generate a Poisson random variable $M_i$ with mean equal to $\tau X_i$, and then randomly simulate $M_i$ events in $(0, \tau]$ according to a uniform distribution. Under the second scenario, $N_i$ is a homogeneous Poisson cluster process. To simulate $N_i$, we first generate a homogeneous Poisson process with intensity $\rho_i = X_i/3$ as the parent process. For each parent, we then generate $m = 3$ children on average according to a Poisson distribution, and disperse the children locations independently following a normal distribution centered at the parent location and with standard deviation $\omega = 2$. The final process contains only the children event times. For Poisson processes, all events in disjoint time intervals are independent. However, this is not true for Poisson cluster processes.

We repeat the simulation 200 times in each case. For each simulated data set, we apply our proposed MOM and SUBEX estimators to estimate $\boldsymbol{\theta}$. For the former, we obtain $\hat{\alpha}$ and $\hat{\sigma}^2$ based on 20 equally spaced $p$ values in $[1/3, 1/2]$ when regressing $\tilde{Y}(p)$ on $\tilde{X}(p)$. For the latter, we assign $\hat{x}(p) = 1/p$ and $\hat{x}(p) = 1/p - (1-p)\widehat{\alpha}/(p^2\tau^2\widehat{\sigma}^2)$ and refer to the resulting estimators as $\text{SUBEX}_1$ and $\text{SUBEX}_2$, respectively. We set the $p$ values used for both estimators to be from 0.6 to 0.95 with an increment of 0.05, and use a quadratic function for the extrapolation. In the online appendix, we show additional results for SUBEX with a finer choice of $p$ values and also with a cubic extrapolating function. These results suggest that the SUBEX method is not sensitive to the choice of $p$ values, since increasing the number of $p$ does not change the results much. Using the cubic function can reduce the bias but also increases the variance of the SUBEX estimator, and leads to a less efficient estimator in terms of mean squared error.

For comparison, we also apply the naive estimator by treating $W_i$ as if it was $X_i$, and the

naive MOM (referred to as $\text{MOM}_1$) and the empirical SIMEX estimators described in Section 3.4. For the last two estimators, we assume that $N_i$ in $(0, \tau/2]$ and $(\tau/2, \tau]$ are independent. This is true for the Poisson processes, but not for the Poisson cluster processes. In Table 1, we report the bias, standard error, mean of the estimated standard error and relative efficiency of the considered estimators. The relative efficiency of an estimator is defined as the mean squared error of the naive estimator divided by that of the estimator under consideration.

In the Poisson process case, we set $\tau = 10$ and $n = 200$. From Table 1, we see that the naive estimator for $\theta_1$ has a large bias due to the estimation error in $W_i$. However, the biases in the other estimators are much smaller. Since $N_i$ is Poisson, there is no need to consider within-trajectory dependence. As a result, the naive $\text{MOM}_1$ estimator, the empirical SIMEX, and $\text{SUBEX}_1$ have slightly higher relative efficiencies than the proposed $\text{MOM}_2$ and $\text{SUBEX}_2$ estimators, both of which are designed to account for within-trajectory dependence. However, none of the estimators for $\theta_2$ suffers from attenuation, and the naive estimator has the smallest standard error. This is because $Z_i$ is measured without error and is also independent with $X_i$. This phenomena has been repeatedly observed throughout our simulation. We will therefore discuss only the results for $\theta_1$ from now on.

In the Poisson cluster process case, we set $\tau = 10, 20$ and $n = 200$. From Table 1, we see that the naive estimator is even more biased than in the Poisson process case. This is because of the increased estimation error in $W_i$ due to correlation. Our proposed $\text{MOM}_2$ estimator has the smallest bias among all estimators: the bias when $\tau = 10$ is $-0.1869$ whereas the next smallest bias in magnitude is $-0.4009$; the bias when $\tau = 20$ further reduces to $-0.1014$, which is again the smallest among all estimators. The relative efficiency for the proposed $\text{MOM}_2$ estimator is also the highest. In particular, it can be eight times as efficient as the naive estimator when $\tau = 20$. The two SUBEX estimators perform significantly better than the empirical SIMEX. This is expected following our discussion in Section 3.4. Between the

two, $\text{SUBEX}_2$ works slightly better than $\text{SUBEX}_1$, because it better accounts for within-trajectory dependence. In general, the amount of information used to estimate $X_i$ increases as $\tau$ increases. This fact explains the better results for all estimators when $\tau = 20$.

## 4.2 Simulation 2: survival model

An important part of the project is to model the first relapse time after treatment. As explained in Section 2, the data are partially interval censored. We use the second simulation study to mimic the real data and to check the performance of the proposed estimators.

We simulate $X_i$, $N_i$ and $Z_i$ as in Section 4.1, for $i = 1, \cdots, n$, and simulate the failure time $T_i$ through a Cox proportional hazard model with the hazard function

$$\lambda(t|X_i, Z_i) = \lambda_0(t) \exp(X_i\beta + Z_i\eta), \tag{10}$$

where $\lambda_0(t) = t$ is the (unobserved) baseline hazard function. The regression coefficients are $\boldsymbol{\theta} = (\beta, \eta)^T = (1, 1)^T$. We assume censoring at random and set the censoring times to be $(0.2, 0.5, 1)$. Let the censoring indicator $\delta_i$ be a binary variable independent of $X_i$ and $Z_i$, with $P(\delta_i = 1) = 0.5$. When $\delta_i = 1$, the event time $T_i$ is censored in the interval between the two censoring times closest to $T_i$; if $T_i$ is less than 0.2, it is censored in $[T_i^l = 0, T_i^r = 0.2]$; if an event time is over 1, it is automatically right censored at 1. Overall, about 12% of the observations are right censored, 43% are interval censored, and the rest 45% are observed.

For interval censored data, it is difficult to separate estimating the baseline hazard function from estimating $\boldsymbol{\theta}$ by using approaches such as the partial likelihood (Cai and Betensky, 2003; Sun 2006). Many authors have considered modeling the baseline hazard as spline functions. We follow Cai and Betensky (2003) and Ruppert et al. (2003) to model the log baseline hazard as a linear spline function:

$$\phi(t) = \log \lambda_0(t) = a_0 + a_1 t + \sum_{k=1}^{K} b_k (t - \kappa_k)_+, \tag{11}$$

where $x_+ \equiv \max(x, 0)$ and $\kappa_k$'s are the knots. There are two immediate benefits for using this model. First, $\lambda_0(\cdot)$ is guaranteed to be nonnegative, so that we do not need any constraint on the parameters when maximizing the likelihood. Second, since $\phi(\cdot)$ is modeled as piecewise linear polynomial, the cumulative hazard function can be written out in an explicit form. For higher-order spline functions, such explicit expressions are not available.

Let $\Theta$ be the collection of all parameters, including both $\boldsymbol{\theta}$ and the spline coefficients in (11). Then, $\Theta$ can be estimated by maximizing the penalized likelihood

$$\ell_{\text{pen}}(\Theta) = \sum_{i=1}^{n} \ell_i(\Theta) - \frac{1}{2\sigma_b^2}\mathbf{b}^T\mathbf{b}, \tag{12}$$

where $\ell_i(\Theta)$ is the log likelihood function for the $i$-th subject, $\sigma_b^2$ is a tuning parameter that controls the smoothness of $\phi(t)$, and $\mathbf{b} = (b_1, b_2, \cdots, b_K)^T$. Cai and Betensky (2003) gave the detailed expression for $\ell_i(\Theta)$ and also suggested a data-driven procedure to estimate $\sigma_b^2$. As pointed out by Apanasovich et al. (2009), estimation of the parametric component in a semiparametric setting is not sensitive to the choice of the smoothing parameter. We therefore use a fixed yet reasonable $\sigma_b^2$ in our simulation so as to reduce the computational burden.

As in the linear model case, we use both homogeneous Poisson processes and Poisson cluster processes for $N_i$. We consider five estimators: the estimator using the true $X_i$, the naive estimator that replaces $X_i$ by $W_i$, the two SUBEX estimators, i.e., $\text{SUBEX}_1$ and $\text{SUBEX}_2$, and the empirical SIMEX. The results in Table 2 are obtained by modeling $\phi(t)$ with a 10-knot spline function. The spline function provides a sufficient approximation, as can be seen from the fact that the estimator based on the true $X_i$ is almost unbiased. Thus, all biases in the other four estimators are mainly caused by the estimation error in $W_i$. We have also tried $K = 25$ knots and obtained very similar results.

In the Poisson process case, we set $\tau = 10$ and $n = 200$. Similar to the linear model case, the naive estimator is severely biased, but the SUBEX and the empirical SIMEX estimators can significantly reduce the bias. Since there is no within-trajectory dependence in $N_i$, the

empirical SIMEX perform better than the two SUBEX estimators, and there is no benefit for using $\text{SUBEX}_2$ over $\text{SUBEX}_1$.

In the Poisson cluster process case, we set $\tau = 10, 20$ and $n = 200$. We also consider $n = 400$ when $\tau = 20$. From Table 2, we see that both $\text{SUBEX}_1$ and $\text{SUBEX}_2$ have much smaller bias in magnitude than the empirical SIMEX but are also more variable. Thus, there is a trade-off between reducing the bias and increasing the variance. Between the two SUBEX estimators, $\text{SUBEX}_2$ is slightly more effective in reducing the bias than $\text{SUBEX}_1$. In terms of relative efficiency, the two SUBEX estimators perform significantly better than the empirical SIMEX except when $\tau = 20$ and $n = 200$. In this case, the relative efficiency of all three estimators are similar.

When $\tau = 10$, the bias is quite large after the bias correction. This is because the within-trajectory dependence is strong in this case. When $\tau$ increases to 20, the biases decrease significantly in all cases. However, when $n$ increases from 200 to 400, the biases are nearly unchanged for all estimators, although the standard errors decrease. In general, the bias of the SUBEX estimators depends on the strength of the within-trajectory dependence relative to the length of the observation interval, i.e., $\tau$, but not on the sample size $n$. The statistic $2\hat{\alpha}/(\tau^2\hat{\sigma}^2)$ given in Section 3.4 can be used to gauge the level of dependence. Specifically, it is approximately equal to 0.2519 when $\tau = 10$ and equal to 0.1228 when $\tau = 20$.

In the online appendix, we show additional simulation results with a finer choice of $p$ values and also with a cubic extrapolating function. Similar to the linear model case, we find that using the cubic function can reduce the bias but inflate the variance of the final estimator. Increasing the number of $p$ values does not change the results much, although variability of the new estimator does become smaller.

# 5 Application to Cocaine Dependence Treatment Data

## 5.1 Definition of variables

We apply our proposed bias-correction methods to analyze the motivating dataset given in Section 2. The main outcomes of interest are the (log) CCQ-Brief scores collected at days 90 and 180 (when available) after treatment and the time to first relapse. For the latter, we focus only on data in the first 90 days from participants in the second study, because the self-report cocaine use data obtained in this period are believed to be more reliable and are also supplemented with more frequent urine samples for these subjects.

The predictor variables used in both analysis include age, gender (=1 for female and 0 for male), race (=1 for African American and 0 for the rest), number of cocaine use years and number of anxiety disorders present at the baseline interview. For the first analysis involving the CCQ-Brief scores, we also include an indicator variable which is equal to one if the measurement was take at day 90 and equal to zero otherwise. For a given baseline cocaine use trajectory, we calculate the baseline frequency and baseline average daily use amount. We include one of them each time as an error-prone predictor variable in the analyses. We have also conducted a separate analysis to include the second-order statistics defined in (2) with $\psi(r)$ therein being the second and third principal components of $g_i(r), r = 1, \cdots, 7$; these summary statistics characterize one's weekly use pattern and short-range dependence, respectively. Figure 4 shows histograms for these four summary statistics. However, neither the weekly pattern nor the short-range dependence is significant, likely because the associated second-order summary statistics are tightly distributed around their means (see Figure 4 (c) and (d)) and also the sample size is relatively small. A significant relationship could be obtained if more subjects are included. In light of this result, we focus on the baseline frequency and baseline average daily use amount in our discussion below.

## 5.2 Analysis of CCQ-Brief scores

Our analysis is based on 126 participants whose CCQ-Brief scores were available at days 90 and/or 180. Eight of the remaining sixteen participants never returned for any interview nor provided any cocaine use record. It is difficult to assess the causes of dropout for these individuals. The other eight participants dropped out before day 90 with partial cocaine use records reported. Among these, six dropped out on or before day 15, but only two of them relapsed before dropout. Thus, there is no strong evidence suggesting a connection between the missing mechanism and the success of treatment. Given the relatively small number of missing observations, we conduct our analysis based on the complete data.

Figure 5 shows a scatter plot of $\tilde{X}(p)$ and $\tilde{Y}(p)$ for $p = k/\tau$, where $k = 30, 31, \cdots, 40$ and $\tau = 80$. There is clearly a linear trend between $\tilde{X}(p)$ and $\tilde{Y}(p)$, which suggests that condition (4) is reasonable. By regressing $\tilde{Y}(p)$ on $\tilde{X}(p)$, we obtain $\hat{\sigma}^2 = 2.269$ and $\hat{\alpha}/\tau^2 = 0.4261$. Recall that $2\hat{\alpha}/(\tau^2\hat{\sigma}^2)$ can be regarded as the amount of bias of the naive variance estimator $\hat{\sigma}^2_{naive}$ relative to $\sigma^2$. Based on $\hat{\alpha}$ and $\hat{\sigma}^2$, we conclude that $\hat{\sigma}^2_{naive}$ approximately underestimates $\sigma^2$ by 37.56%. This bias is quite large and in turn will yield biased regression coefficient estimates. As we discussed in Section 3.3, the bias is mainly caused by underestimation of $\sigma^2_{u_i}$ for trajectories that are similar to the third trajectory given in Figure 1 (b).

Table 3 lists the estimated regression coefficients and their standard errors for our proposed estimators based on the proposed MOM and the two SUBEX estimators used in the simulation. The standard errors are obtained by bootstrap. The estimated coefficients from $\text{SUBEX}_1$ are very close to those from the MOM estimator, whereas those from $\text{SUBEX}_2$ are slightly more different. For baseline frequency, the resulting estimates are equal to 0.565, 0.577 and 0.662 for these three estimators, respectively. For comparison, we also apply the naive estimator, the naive MOM estimator based on $\hat{\sigma}^2_{naive}$, and the empirical SIMEX estimator. For baseline frequency, the estimates are equal to 0.476, 0.529 and 0.537, which are all smaller than those given by our proposed methods.

We interpret our findings based on the proposed MOM estimator, in light of its better performance over the SUBEX as suggested by our simulation. The results suggest that both race and baseline frequency are significant at the 0.01 level. Because the signs for these estimated coefficients are positive, we conclude that African Americans and heavy baseline cocaine user tend to have high (log) CCQ-Brief scores three/six months after treatment (95% confidence intervals 0.1116 to 0.5027 for race and 0.1523 to 0.9782 for baseline frequency). In addition, the number of current anxiety disorders is significant at the 0.10 level but not at the 0.05 level. Thus, the (log) CCQ-Brief scores appear to increase with the number of current anxiety disorders. The remaining variables are all insignificant. In particular, the (log) CCQ-Brief scores are not significantly different at days 90 and 180 after treatment.

We have conducted additional analysis by defining the baseline average daily use amount as the error-prone variable but have found it insignificant at the 0.10 level; this is likely due to the significantly larger measurement error associated with this particular summary statistic. Among the remaining variables, Race still remains significant at the 0.01 level but all the others are not significant at the 0.10 level. We have also repeated our analysis by using the (log) CCQ-Brief score collected at day 14 as the response variable. In this case, no variable is significant at the 0.10 level.

## 5.3 Analysis of first relapse time

We model the time to first relapse data by the proportional hazard model given in (10). We use data from 79 subjects participating in the second study; data for the remaining 4 subjects in the same study were completely missing. The first relapse time was determined from the self-report posttreatment cocaine use trajectories. To ensure the quality of the self-report data, urine samples were tested on days 14, 30, 90 and 180 to check whether a patient had lied. As a result, the relapse time data are partially interval-censored. If a positive urine test was obtained before the self-report relapse time, then the true relapse time was interval

censored between the first positive urine sample test date and the previous negative test date (or the discharge date if no previous test was available). Since data are less reliable after the first 90 days, any relapse time greater than 90 is considered as right censored. This data structure is similar to that considered in the simulation study in Section 4.2: about 50.6% of the subjects have observed relapse time, 31.6% are interval censored and 17.8% are right censored.

We use the baseline average daily use amount as the error-prone variable. Table 4 lists the estimated regression coefficients and their standard errors from the proposed SUBEX and the empirical SIMEX estimators. As we can see from the table, both age and cocaine use years have a significant effect on the time to first relapse. Specifically, cocaine use years has a positive effect on the hazard function, meaning that a longer cocaine use history results in a quicker relapse. The variable, age, on the other hand, has a negative effect. This means that among users with the same length of cocaine use history, those who are older tend to remain abstinent for a longer period of time.

For the baseline average daily use amount, the naive estimator surprisingly shows a significant negative effect on the hazard function. A similar result was also reported in Sinha et al. (2006). However, this is very counterintuitive, since it implies a longer time to relapse for those who used more during the baseline period. The empirical SIMEX estimator yields results very similar to the naive estimator and hence fails to correct for the bias in this case. In contrast, our proposed SUBEX estimators yield estimates that are much closer to zero than the naive and the empirical SIMEX estimators; $\text{SUBEX}_2$, which accounts for within-trajectory dependence, even gives an insignificant positive estimate. Both estimators suggest that the effect due to baseline average daily use amount is insignificant after accounting for effects due to the other variables. This is intuitively more reasonable.

# 6 Discussion

In many scientific problems in particular substance use research, it is common to have summary statistics derived from some stochastic processes as covariates. The cocaine dependence treatment dataset considered in this paper is one example of such problems. The estimation error in these summary statistics causes estimation bias in the regression coefficients like in classical measurement error problems. As we have illustrated through simulation and data analysis, the estimated coefficients can be severely attenuated and can even result in wrong inference, if the measurement error is not properly accounted for.

The fact that the summary statistics are derived from stochastic processes presents new challenges. Unlike in classical measurement error problems, the error in the summary statistics is heteroscedastic and depends on individual stochastic processes. To correct for the bias in the estimators, we propose a new method-of-moment approach for linear models and a subsampling extrapolation method that is generally applicable to both linear and nonlinear models. The methods we have proposed are based on novel subsampling techniques that take into account of the correlation within individual processes, and have shown good performance in both simulation and real data analysis.

It is important to note that there could be significant bias (typically underreporting) associated with self-report cocaine use data. The TLFB is no exception to that limitation. Other than the TLFB, cocaine use in addicted individuals has been measured in other ways such as using toxicology analysis of cocaine metabolites in urine, blood or hair samples. Indeed all patients included in this study were tested with urine samples upon admission to the inpatient treatment unit. However, detection of cocaine in blood or urine samples is limited to 2-3 days since use and hence requires multiple sampling in an ongoing basis to evaluate pattern of drug use over a defined period. As such, it cannot be used to assess pattern of use in the recent past such as would be needed as patients enter a treatment facility. While hair samples can provide presence or absence of drug over a 90 day period,

they do not provide any information on when a subject used drugs and at what frequency. Furthermore, hair samples are not reliable for all types of hair and often it is problematic to detect drug in high curly or wavy hair. Cost of hair sample testing is also high and hence it presents a feasibility challenge. Finally, new daily self-report methods such as ecological momentary assessment, which involves moment-to-moment monitoring of daily drug use with electronic diaries while patients are in the real world, has also been used recently (Epstein and Preston, 2010; Preston et al., 2009). However, this method has been used in community and outpatient treatment samples of drug abusers and it would be difficult to implement to assess recent drug use pattern when subjects are entering a treatment facility.

Given the limitations with other data collection methods as being explained above, we relied on structured interview techniques (i.e., the TLFB) to obtain self-report baseline cocaine uses in our study. To ensure data quality, we followed strictly the well-developed TLFB interview procedures and also took additional measures. For example, all research assistants who collected the data had been trained by PhD level psychologists and each one of them had already had over three-year experience in administration of similar assessments; they were also closely supervised when conducting these interviews. Moreover, all subjects had been informed upfront that all data would be only coded by a number and be kept strictly confidential, and that the federal certificate of confidentiality granted to this study would protect the research information from being legally summoned by the courts. Hence there would be no legal consequences to them by providing the most honest responses. They were also informed that they would be removed from the study if they were found out not being truthful about their drug use. All interviews were conducted in a quiet and comfortable testing room, and excellent rapport was established with subjects.

With the aforementioned additional measures we believe that we have minimized the chance of underreporting, however this problem cannot be completely eliminated. The effect of measurement error due to self report has also been studied in other settings, most notably

in nutritional epidemiology. For example, Kipnis et al. (1999) modeled the self-report dietary intake by $Q_{ij} = \alpha_0 + \alpha_1 T_i + r_i + \epsilon_{ij}$, where $Q_{ij}$ is the $j$th replicate of the self-report dietary intake from the $i$th subject, $T_i$ is the true intake, $r_i$ is a subject-specific bias and $\epsilon_{ij}$ is a random error. Different from the classical measurement error models, the model proposed by Kipnis et al. allows a subject-specific bias, $r_i$, which is modeled as a random effect with distribution $N(0, \sigma_r^2)$ for some $\sigma_r^2 > 0$. They argued that this subject-specific bias if ignored would cause further attenuation to the regression coefficients and make them even less significant. We expect a similar phenomenon to happen in our setting as well if there is serious underreporting. To account for the self-report error, Kipnis et al. assumed that a reference instrument was available and could be modeled as $F_{ij} = \mu_0 + T_i + s_i + u_{ij}$, where $s_i$ is another subject-specific random effect possibly correlated with $r_i$ and $u_{ij}$ is a random error in the instrument variable. As pointed out by the authors, even with the additional information in $F_{ij}$, this model is complicated and some of the parameters may have identifiability issues. Neither such a reference instrument nor replicates of the baseline trajectories are available in our data. Pushing for additional baseline measurements is key in order to better account for the potential bias due to underreporting for substance use research. However, this is beyond the scope of this paper.

# 7 Supplementary Materials

**Theory for SUBEX and simulation:** Asymptotic theory and variance estimation for the proposed SUBEX estimator and some additional simulation results. (pdf file)

# References

Apanasovich, T. V., Carroll, R. J. and Maity, A. (2009). SIMEX and standard error estimation in semiparametric measurement error model, *Electronic Journal of Statistics*, 3,

318-348.

Brown, J., Kranzler, H. R. and Del Boca, F. K. (1992). Self-reports by alcohol and drug abuse inpatients: Factors affecting reliability and validity, *British Journal of Addictions*, 87, 1013-1024.

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline, *Biometrics*, 59, 570-579.

Carroll, K. C., Power, M., Bryant, K. and Rounsaville, B. J. (1993). One year follow-up status of treatment-seeking cocaine abusers: Psychopathology and dependence severity as predictors of outcome, *Journal of Nervous and Mental Disease*, 181(2), 71-79.

Carroll, R. J., Küchenhoff, H., Lombard, F. and Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models, *Journal of the American Statistical Association*, 91, 433, 242-250.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, Chapman and Hall, New York, NY.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models, *Journal of the American Statistical Association*, 89, 1314-1328.

Crainiceanu, C. M., Caffo, B. S., Di, C.-Z. and Punjabi, N. M. (2009). Nonparametric signal extraction and measurement error in the analysis of electroencephalographic activity during sleep, *Journal of the American Statistical Association*, 104, 541-555.

Daley, D. J. and Vere-Jones, D. (1988). *An introduction to the theory of point processes*, Springer-Verlag Inc, New York.

Devanarayan, V. and Stefanski, L. A. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements, *Statistics & Probability Letters*, 59,

219-225.

Epstein D. H. and Preston K. L. (2010). Daily life hour by hour, with and without cocaine: an ecological momentary assessment study. *Psychopharmacology* (Berl), 211(2):223-232.

Fals-Stewart, W., O'Farrell, T. J., Freitas, T. T., McFarlin, S. K. and Rutigliano, P. (2000). The timeline follow-back reports of psychoactive substance use by drug-abusing patients: Psychometric properties, *Journal of Consulting and Clinical Psychology*, 68,134144.

First, M., Spitzer, R., Gibbon, M., and Williams, J. (1995). *Structured Clinical Interview for DSMIV: Patient Edition*, American Psychiatric Press Inc, Washington, DC.

Fox, H. C., Garcia, M., Milivojevic, V., Kreek, M. J. and Sinha, R. (2006). Gender differences in cardiovascular and corticoadrenal response to stress and drug cues in cocaine dependent individuals, *Psychopharmacology*, 185(3), 34857.

Kampman, K. M., Volpicelli, J. R., Mulvaney, F., Alterman, A. I., Cornish, J., Gariti, P., Cnaan, A., Poole, S., Muller, E., Acosta, T., Luce, D. and O'Brien, C. (2001). Effectiveness of propranolol for cocaine dependence treatment may depend on cocaine withdrawal symptom severity, *Drug and Alcohol Dependence*, 63, 6978.

Kipnis, V., Carroll, R. J., Freedman, L. S. and Li, L. (1999). Implication of a new dietary measurement error model for estimation of relative risk: application to four calibration studies, *American Journal of Epidemiology*, 150, 642-651.

Li, E., Zhang, D. and Davidian, M. (2004). Conditonal estimation for generalized linear models when covariates are subject-specific parameters in a mixed model for longitudinal measurements, *Biometrics*, 60, 1-7.

Li, E., Wang, N. and Wang, N.-Y. (2007). Joint models for a primary endpoint and multiple longitudinal covariate processes, *Biometrics*, 63, 1068-1078.

Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling*, Springer-Verlag Inc, New

York.

Preston, K. L., Vahabzadeh, M., Schmittner, J., Lin, J. L., Gorelick, D. A. and Epstein, D. H. (2009). Cocaine craving and use during daily life, *Psychopharmacology* (Berl), 207(2):291-301.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, New York, NY.

SAMHSA (2004). National household survey on drug abuse, http://www.oas.samhsa.gov/nhsda.htm.

Sherman, M. F. and Bigelow, G. E. (1992). Validity of patients self-reported drug use as a function of treatment status, *Drug and Alcohol Development*, 30, 1-11.

Sinha, R. (2001). How does stress increase risk of drug abuse and relapse? *Psychopharmacology*, 158, 343-359.

Sinha, R. (2007). The role of stress in addiction relapse, *Current Psychiatry Reports*, 9, 388-395.

Sinha, R., Garcia, M., Paliwal, P., Kreek, M. J. and Rounsaville, B. J. (2006). Stress-induced cocaine craving and hypothalamic-pituitary-adrenal responses are predictive of cocaine relapse outcomes, *Archives of General Psychiatry*, 63(3), 324-31.

Sobell, L. and Sobell, M. (1993). Timeline follow back: A technique for assessing self-reported ethanol consumption. In J. Allen & R. Litten (Eds.), *Techniques to Assess Alcohol Consumption*. New Jersey: Humana Press, Inc.

Song, X., Davidian, M. and Tsiatis, A. A. (2002). An estimator for the proportional hazard model with multiple longitudinal covariates measured with error, *Biostatistics*, 3, 4, 511-528.

Tiffany, S. T., Singleton, E., Haertzen, C. A. and Henningfield, J. E. (1993). The develop-

ment of a cocaine craving questionnaire, *Drug and Alcohol Dependence*, 34, 19-28.

Wang, C. Y., Wang, N. and Wang, S. (2000). regression analysis when covariates are regression parameters of a random effect model for observed longitudinal measurements, *Biometrics*, 56, 487-495.

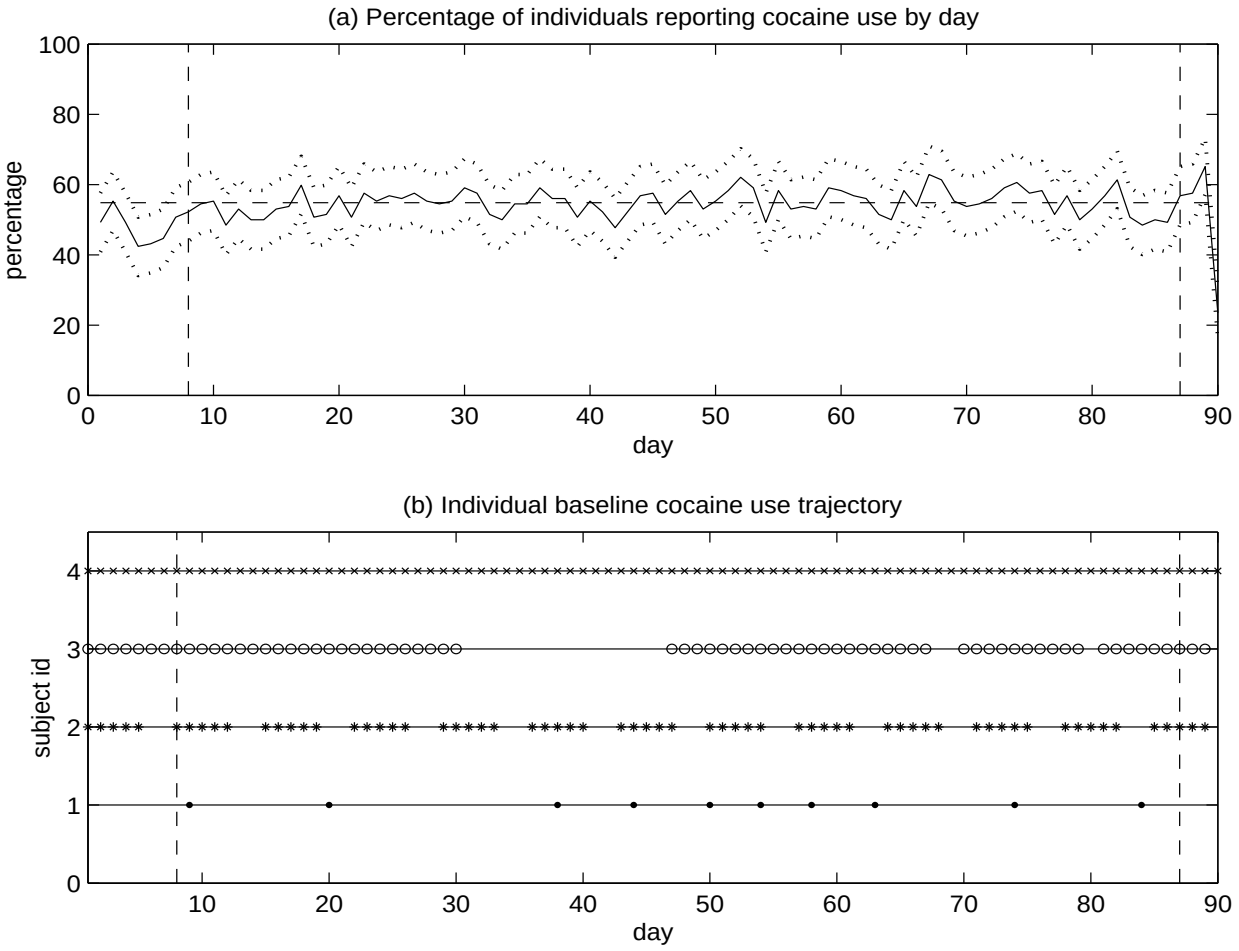


Figure 1: Plot (a): Percentage of individuals reporting cocaine use by day; the solid line shows the daily percentage, the dotted lines are the pointwise 95% confidence envelope, the horizontal dashed line is the average daily percentage. Plot (b): Individual baseline cocaine use trajectories for four selected subjects, where each symbol on a given day signifies a cocaine use that day. The vertical dashed lines in both plots show the selected baseline period.

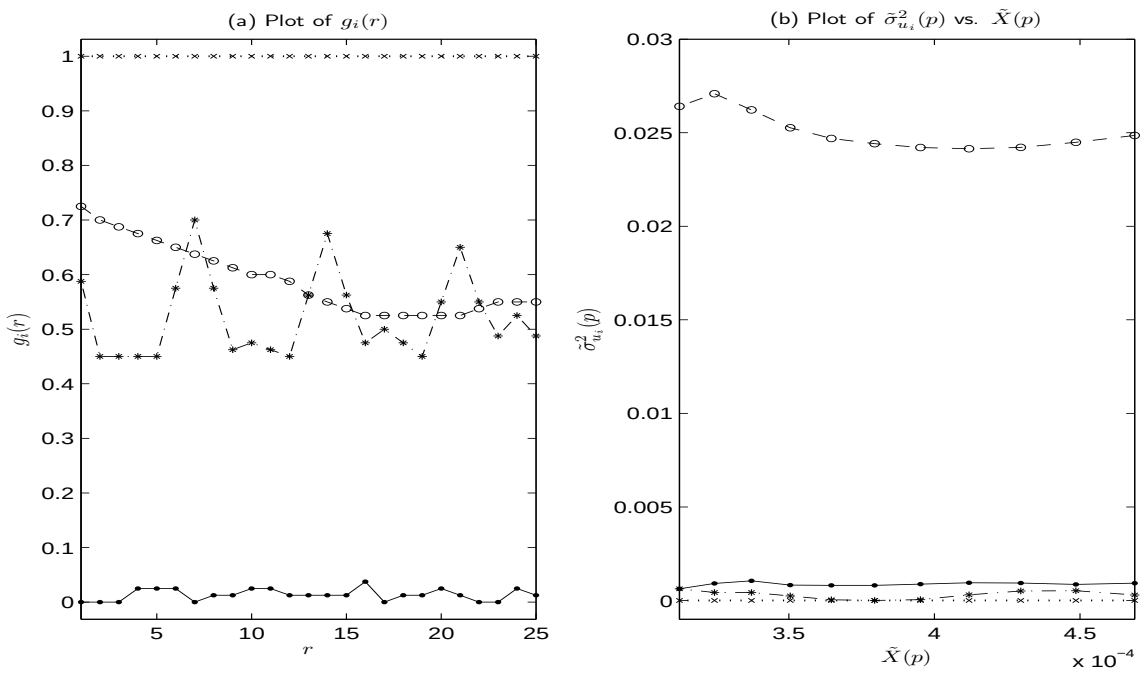


Figure 2: Plots (a) and (b) show plots of $g_i(r)$ and of $\tilde{\sigma}^2_{u_i}(p)$ vs $\tilde{X}(p)$ for the four individual trajectories given in Figure 1 (b). The symbols in each plot match those in Figure 1 (b).

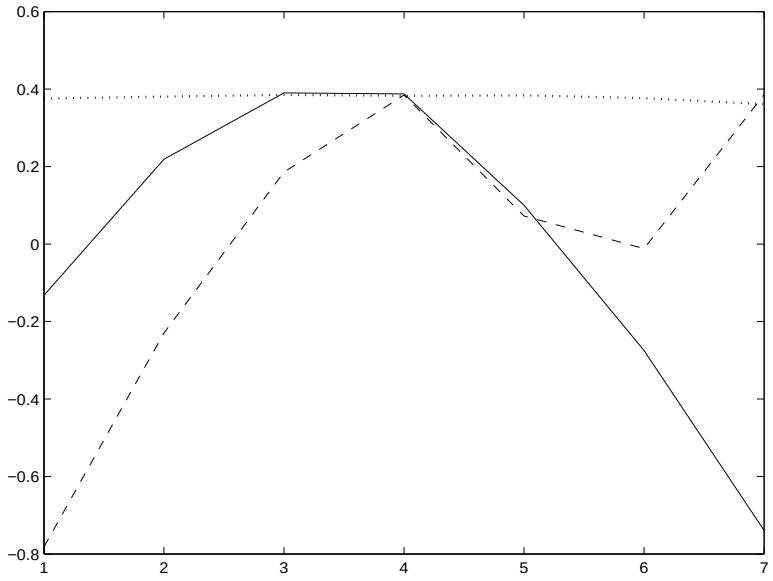

Figure 3: Plot of the (estimated) first, second and third principal components of $g_i(r)$, denoted by the dotted, solid and dashed lines, respectively.

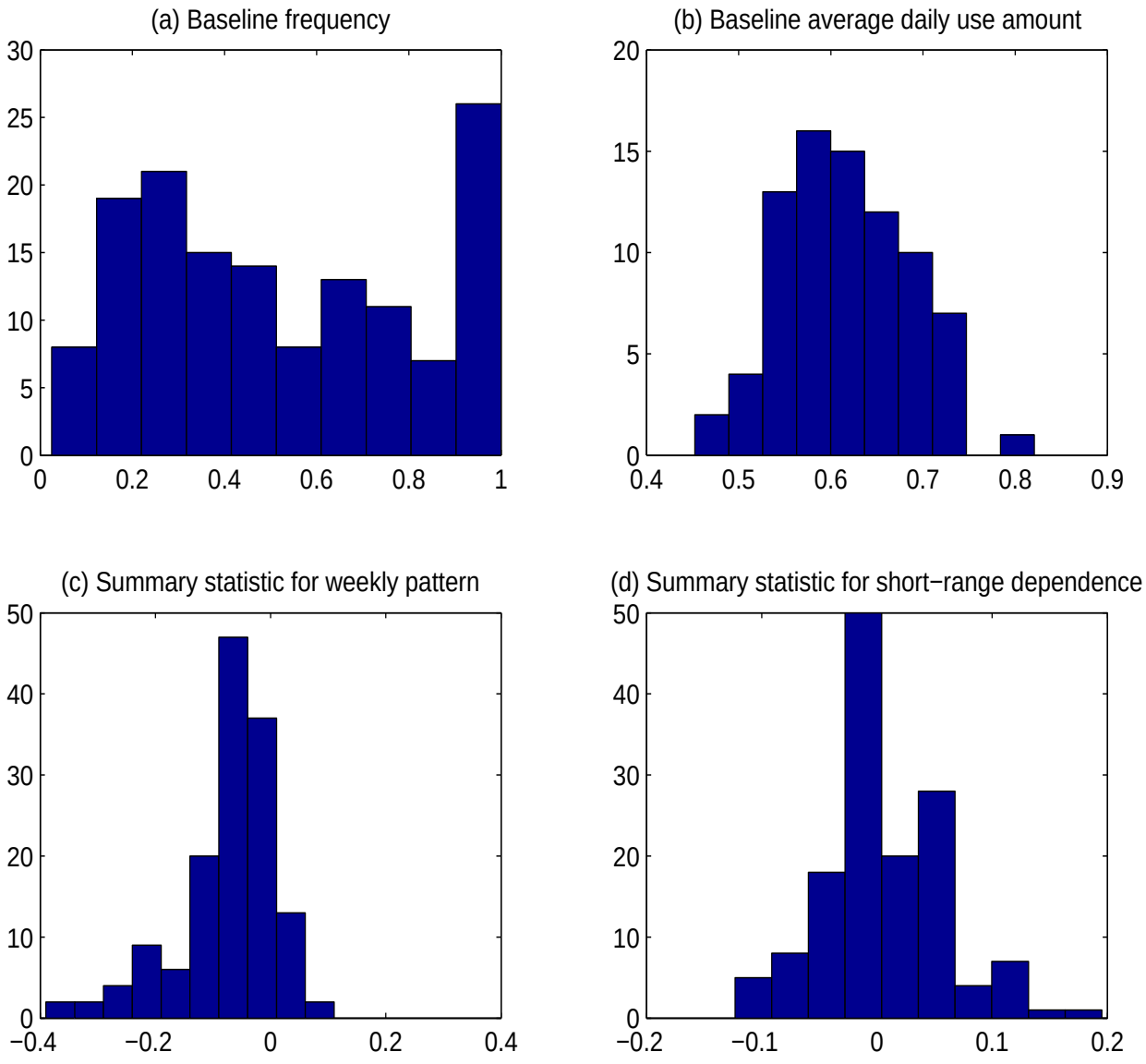


Figure 4: Histograms of the four summary statistics derived from the baseline cocaine use trajectories.

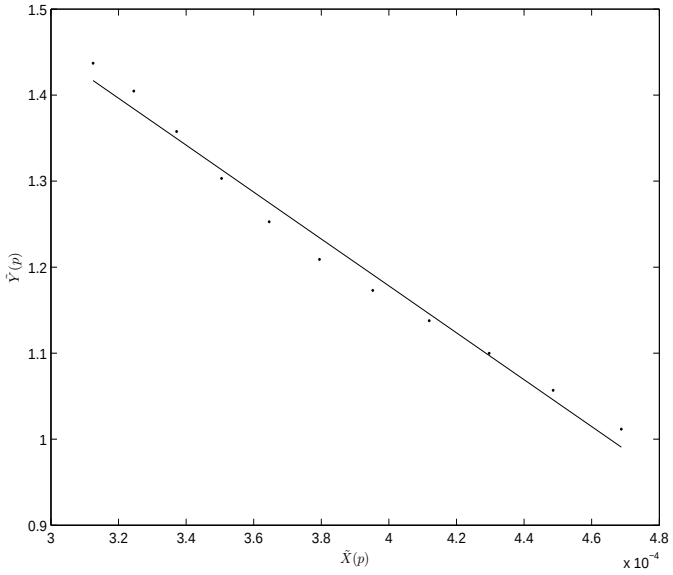


Figure 5: Scatter plot of $\tilde{X}(p)$ and $\tilde{Y}(p)$ for $p = k/\tau$ where $k = 30, 31, \cdots, 40$ and $\tau = 80$.

Table 1: Simulation results for linear model. The values in the parenthesis are the mean of the estimated standard errors (SEs). Relative efficiency (Rel. Eff.) is defined as the ratio between the MSE of the naive estimator and that of the estimator under consideration. $MOM_1$ and $MOM_2$ are the naive and proposed method-of-moment methods, respectively.

| Scenario 1: Poisson process, $\tau = 10$ | | | | | | |
|---|---|---|---|---|---|---|
| | $\theta_1$ | | | $\theta_2$ | | |
| | Bias | SE | Rel. Eff. | Bias | SE | Rel. Eff. |
| Naive | -.3710 | .0686 | | .0075 | .0893 | |
| $MOM_1$ | .0134 | .1590 | 5.6190 | .0086 | .0973 | .8406 |
| $MOM_2$ | .0163 | .1765 | 4.5517 | .0086 | .0975 | .8376 |
| $SUBEX_1$ | -.0999 | .1317 (.1334) | 5.2228 | .0052 | .1255 (.1308) | .5084 |
| $SUBEX_2$ | -.1073 | .1346 (.1317) | 4.8206 | .0063 | .1246 (.1297) | .5158 |
| SIMEX | -.1348 | .0989 (.0896) | 5.0983 | .0075 | .0929 (.0913) | .9237 |
| Scenario 2: Poisson cluster process | | | | | | |
| | $\theta_1$, $\tau = 10$ | | | $\theta_1$, $\tau = 20$ | | |
| | Bias | SE | Rel. Eff. | Bias | SE | Rel. Eff. |
| Naive | -.6678 | .0616 | | -.5163 | .0670 | |
| $MOM_1$ | -.4205 | .1106 | 2.3798 | -.1388 | .1519 | 6.4177 |
| $MOM_2$ | -.1869 | .2974 | 3.6586 | -.1014 | .1551 | 7.9226 |
| $SUBEX_1$ | -.4417 | .1245 (.1227) | 2.1361 | -.2263 | .1508 (.1463) | 3.6700 |
| $SUBEX_2$ | -.4009 | .1544 (.1556) | 2.4387 | -.1997 | .1750 (.1692) | 3.8524 |
| SIMEX | -.5189 | .0831 (.0802) | 1.6292 | -.2959 | .0958 (.0913) | 2.8029 |

Table 2: Simulation results in the interval-censored failure time data case. Only the estimation results on $\theta_1$ are presented. The values in the parenthesis are the mean estimated SE. Relative efficiency (Rel. Eff.) is defined as the ratio between the MSE of the naive estimator and that of the estimator under consideration.

| | Poisson, $\tau = 10$, $n = 200$ | | | Poisson cluster, $\tau = 10$, $n = 200$ | | |
|---|---|---|---|---|---|---|
| | Bias | SE | Rel. Eff. | Bias | SE | Rel. Eff. |
| No Error | .0103 | .1471 (.1469) | | .0053 | .1480 (.1462) | |
| Naive | -.4273 | .1168 (.1129) | | -.7225 | .0890 (.0800) | |
| $SUBEX_1$ | -.1424 | .2787 (.2587) | 2.0113 | -.5073 | .2176 (.2016) | 1.7402 |
| $SUBEX_2$ | -.1486 | .2789 (.2570) | 1.9709 | -.4687 | .2768 (.2590) | 1.7908 |
| SIMEX | -.1674 | .1830 (.1732) | 3.1990 | -.5931 | .1360 (.1237) | 1.4318 |
| | Poisson cluster, $\tau = 20$, $n = 200$ | | | Poisson cluster, $\tau = 20$, $n = 400$ | | |
| | Bias | SE | Rel. Eff. | Bias | SE | Rel. Eff. |
| No Error | .0186 | .1545 (.1465) | | .0081 | .1041 (.1015) | |
| Naive | -.5825 | .1067 (.0971) | | -.5923 | .0730 (.0695) | |
| $SUBEX_1$ | -.2967 | .2717 (.2602) | 2.1715 | -.3007 | .2001 (.1843) | 2.7329 |
| $SUBEX_2$ | -.2698 | .3223 (.3052) | 1.9915 | -.2713 | .2351 (.2178) | 2.7691 |
| SIMEX | -.3750 | .1712 (.1558) | 2.0654 | -.3871 | .1193 (.1107) | 2.1713 |

Table 3: Results for the analysis of the CCQ-Brief score data. The table shows the estimated regression coefficients (standard errors) of the naive estimator (NAIVE) and the bias-corrected estimators from the naive method-of-moment method ($MOM_1$), the empirical SIMEX, the proposed method-of-moment method ($MOM_2$), the two proposed estimators based on subsampling extrapolation ($SUBEX_1$ $SUBEX_2$). cocyrs and curanxs below denote the numbers of cocaine use years and of current anxiety symptoms at baseline interview, respectively.

| | NAIVE | $MOM_1$ | SIMEX | $MOM_2$ | $SUBEX_1$ | $SUBEX_2$ |
|---|---|---|---|---|---|---|
| frequency | .476 (.176) | .529 (.195) | .537 (.223) | .565 (.211) | .577 (.204) | .662 (.257) |
| indicator | -.046 (.062) | -.044 (.062) | -.049 (.064) | -.043 (.062) | -.040 (.064) | -.033 (.065) |
| gender | -.108 (.093) | -.107 (.093) | -.112 (.091) | -.106 (.093) | -.096 (.092) | -.083 (.095) |
| race | .305 (.099) | .306 (.100) | .311 (.101) | .307 (.100) | .289 (.100) | .271 (.103) |
| age | -.088 (.081) | -.096 (.082) | -.087 (.081) | -.101 (.083) | -.095 (.081) | -.101 (.083) |
| cocyrs | -.013 (.087) | -.011 (.087) | -.019 (.085) | -.010 (.088) | -.015 (.083) | -.013 (.086) |
| curanxs | .139 (.085) | .142 (.085) | .134 (.079) | .145 (.086) | .163 (.084) | .178 (.086) |

Table 4: Results for the analysis of the time to first relapse data. The table shows the estimated regression coefficients (standard errors) of the naive estimator (NAIVE), the empirical SIMEX, and the proposed two estimators based on subsampling extrapolation ($SUBEX_1$ $SUBEX_2$). cocyrs, curanxs and cocuse below denote the number of cocaine use years, the number of current anxiety symptoms at baseline interview and the baseline average daily use amount, respectively.

| | NAIVE | SIMEX | $SUBEX_1$ | $SUBEX_2$ |
|---|---|---|---|---|
| cocuse | -.203 (.081) | -.210 (.096) | -.097 (.095) | .017 (.206) |
| gender | -.377 (.299) | -.387 (.301) | -.351 (.319) | -.301 (.347) |
| race | -.196 (.274) | -.182 (.277) | -.187 (.272) | -.162 (.276) |
| age | -.053 (.024) | -.053 (.024) | -.054 (.024) | -.056 (.024) |
| cocyrs | .110 (.028) | .111 (.028) | .111 (.028) | .109 (.029) |
| curanxs | .279 (.221) | .275 (.224) | .308 (.218) | .352 (.232) |