Generalized varying coefficient models for longitudinal data

By DAMLA ŞENTÜRK

Department of Statistics, Penn State University, University Park, Pennsylvania 16802, U.S.A

dsenturk@stat.psu.edu

And HANS-GEORG MÜLLER

Department of Statistics, University of California, Davis, California 95616, U.S.A

mueller@wald.ucdavis.edu

Summary

We propose a generalization of the varying coefficient model for longitudinal data to cases where not only current but also recent past values of the predictor process affect current response. More precisely, the targeted regression coefficient functions of the proposed model have sliding window supports around current time t. A variant of a recently proposed two-step estimation method for varying coefficient models is proposed for estimation in the context of these generalized varying coefficient models, and is found to lead to improvements, especially for the case of additive measurement errors in both response and predictors. The proposed methodology for estimation and inference is also applicable for the case of additive measurement error in the common versions of varying coefficient models that relate only current observations of predictor and response processes to each other. Asymptotic distributions of the proposed estimators are derived, and the model is applied to the problem of predicting protein concentrations in a longitudinal study. Simulation studies demonstrate the efficacy of the proposed estimation procedure.

Some key words: Linear regression; Measurement error model; Prediction; Smoothing; Two-step procedure.

1. Introduction

Longitudinal data are encountered frequently in medical studies, one example being the study of Kaysen et al. (2001) on 64 haemodialysis patients. Repeated measurements were taken on each subject to investigate the relationship between the levels of long-lived acute phase proteins such as serum albumin concentration, C-reactive protein, ceruloplasmin, $\alpha 1$ acid glycoprotein and transferrin. One aim of the study is to predict future concentrations for one of the proteins from present or past levels of another.

Let $\{t_{ij}, j = 1, \ldots, T_i\}$ denote the time-points at which the measurements for the ith of n subjects were taken. Also, let y_{ij} and x_{ij} denote the response and predictor values observed for the ith subject at time t_{ij}. A useful way of modelling longitudinal data is provided by varying coefficient regression models (Hastie & Tibshirani, 1993),

$$y_i(t_{ij}) = \beta_0(t_{ij}) + \beta_1(t_{ij})x_i(t_{ij}) + \epsilon_i(t_{ij}), \tag{1}$$

where $y_i(t_{ij}) = y_{ij}$, $x_i(t_{ij}) = x_{ij}$ and $\epsilon_i(t_{ij})$ is a zero-mean stochastic process with covariance function $\delta(t, t') = \text{cov}\{\epsilon_i(t), \epsilon_i(t')\}$. The time-dependent relationship between the response and predictor, both of which are repeatedly measured, is modelled through the coefficient functions $\beta_0(\cdot)$ and $\beta_1(\cdot)$. For a fixed time t_{ij}, (1) reduces to a simple linear model. Varying coefficient models are appealing as they present a parsimonious and easily interpreted approach for the modelling of the functional relationship between predictor and response trajectories. Estimation of the time-varying coefficient functions involves not more than one-dimensional smoothing (Hoover et al., 1998; Fan & Zhang, 2000; Wu et al., 2000; Wu & Chiang, 2000; Chiang et al., 2001; and Huang et al., 2004). A thorough literature review of applications to longitudinal data can be found in Wu & Yu (2002). Note that in general longitudinal data can be viewed as observed at a common set of time-points, where missing values, which are missing completely at random, might be present. Let $\{t_j, j = 1, \ldots, T\}$ be the distinct time-points among $\{t_{ij}, j = 1, \ldots, T_i, i = 1, \ldots, n\}$. The varying coefficient model in (1) can then be rewritten as

$$y_i(t_j) = \beta_0(t_j) + \beta_1(t_j)x_i(t_j) + \epsilon_i(t_j), \tag{2}$$

where not all n subjects might be observed at every t_j.

Model (2) assumes that responses $y_i(t_j)$ at current time t_j are only influenced by current predictor values $x_i(t_j)$. This might not be a fully adequate way to model the dynamics of many systems, in biology for example. A range of past predictor values, in addition to current values, might play a role in predicting a response in these cases. For example, proteins considered as predictors may have long half-lives. In this paper we therefore propose a generalization of the varying coefficient model for longitudinal data to cases where not only current but also recent past levels of the predictor process affect the current response:

$$y_i(t_j) \;=\; \beta_0(t_j) + \sum_{r=1}^{p} \beta_r(t_j) x_i(t_{j-q-(r-1)}) + \epsilon_i(t_j). \tag{3}$$

Here, p denotes the number of time-points, i.e. the window width into the past, of the predictor process that is considered to affect the response at the current time. The influence of past predictor values is modelled through p separate varying coefficient functions, $\beta_1(\cdot), \ldots, \beta_p(\cdot)$. In order to include prediction of future values, a time lag of size $q > 0$ is included in (3). The varying coefficient model is a special case of (3), in which $q = 0$ and $p = 1$.

The formulation in (3) applies to longitudinal designs with equidistant time-points. Nevertheless, the proposed estimation method will be shown in simulations to be easily adapted to missing values. This property, coupled with a pre-binning step used to synchronize the measurements across subjects, makes the proposed methodology applicable to a broader class of longitudinal designs. In addition, to avoid singularities in (3), we assume that predictor trajectories are not constant on any interval, as this would lead to nonidentifiability of local regressions.

Model (3) is appealing, as the linear regression coefficient function extends beyond the point-wise relationship that characterizes the usual varying coefficient model, to include also data in a window prior to and up to time t. This characteristic is suitable for scenarios in which the response at a fixed time t is likely to depend on the behaviour of the predictor process not only at t but also at times before t.

Model (3) also has features reminiscent of time series models. In fact, related varying coefficient models for time series have been developed and referred to as 'functional coefficient autoregressive models' (Chen & Tsay, 1993) and 'functional coefficient regression models' (Cai et al., 2000), and these models also incorporate smooth coefficient functions and regression modelling. While there are similarities, the data structures to which these models pertain, and consequently their implementation and analysis, are quite different. While the time series models deal with only one time series, models for longitudinal data, considered here, focus on situations where one has repeated measurements for each of a sample of subjects.

2. Estimation in the generalized varying coefficient model under measurement errors

2.1 *Model specifications*

In many longitudinal studies, not only the response but also the predictor variables are contaminated by measurement errors. Let y_{ij}, x_{ij} denote the underlying response and predictor, and y'_{ij}, x'_{ij} denote the observed response and predictor, so that

$$\begin{aligned} y'_{ij} &= y_{ij} + e_{yij}, \\ x'_{ij} &= x_{ij} + e_{xij}, \end{aligned}$$

where e_{yij} and e_{xij} are independently and identically distributed, over i and j, zero mean additive measurement errors with variances σ_x^2 and σ_y^2, respectively. This results in observed longitudinal data of the form

$$(t_j, x'_{ij}, y'_{ij}), \qquad j = 1, \ldots, T_i, \qquad i = 1, \ldots, n.$$

The varying coefficient functions β_j, $j = 0, \ldots, p$ are defined in the error-free model

$$y_i(t_j) = \beta_0(t_j) + \sum_{r=1}^{p} \beta_r(t_j) x_i(t_{j-q-(r-1)}) + \epsilon_i(t_j), \tag{4}$$

for $j = q + p, \ldots, T$, but in the contaminated situation must be targeted based on the observed contaminated response and predictor. The error $\epsilon_i(t_j)$ is the realization of a zero-mean stochastic process with covariance function $\delta(t', t) = \mathrm{cov}\{\epsilon_i(t'), \epsilon_i(t)\}$, which will be denoted by $\delta_j = \delta(t_j, t_j)$ when evaluated at the same time-points.

The proposed estimation algorithm is an extension of the two-step estimation procedure for longitudinal data that was developed by Fan & Zhang (2000). Noting that a different linear regression between the observed response and the predictors applies for each time point in a varying coefficient model, as given by (2), Fan & Zhang (2000) regress the observed response on the observed predictor at a fixed time point t_j to obtain the raw estimates for the smooth coefficient functions $\beta_0(t_j)$ and $\beta_1(t_j)$ in a first step. In a second step, the scatter-plots of the raw estimates for the coefficient functions are smoothed against the time-points, for each component separately, to obtain the final smooth estimates for the coefficient functions. This two-step estimation procedure is intuitively appealing and easy to implement, involving only linear regression fits and one-dimensional smoothing procedures.

2.2 *Proposed estimates*

The observed predictors and their error-free unobserved counterparts considered for the response at a fixed time t_j are $x'_i(t_{j-q}), \ldots, x'_i(t_{j-q-p+1})$ and $x_i(t_{j-q}), \ldots, x_i(t_{j-q-p+1})$, respectively. We collect the observed predictors and response into the matrix $X'_{qpj} = (X'_{1,q,p,j}, \ldots, X'_{n_j,q,p,j})^{\mathrm{T}}$ and the vector $Y'_j = (y'_{1j}, \ldots, y'_{n_j j})^{\mathrm{T}}$, where $X'_{i,q,p,j} = \{1, x'_i(t_{j-q}), \ldots, x'_i(t_{j-q-p+1})\}^{\mathrm{T}}$. Here, n_j denotes the number of subjects observed at time t_j and $(t_{j-q}, \ldots, t_{j-q-p+1})$. Let C_j denote the set of corresponding subject indices. Auxiliary parameters for the method are the lag value q and the window width p. Analogously let X_{qpj} and Y_j denote the unobserved error free data at time t_j.

It follows from (4) that the response at time t_j is modelled through the linear form

$$Y_j = X_{qpj}\beta(t_j) + \epsilon(t_j), \qquad (5)$$

where $\beta(t_j) = \{\beta_0(t_j), \ldots, \beta_p(t_j)\}^{\mathrm{T}}$. The error process observed at time t_j is denoted by $\epsilon(t_j)$. Since the responses observed at time t_j come from different subjects, $E\{\epsilon(t_j)\} = 0_{n_j}$ and $\mathrm{cov}\{\epsilon(t_j)\} = \delta_j I_{n_j}$, where 0_{n_j} denotes a vector of n_j zeros and I_{n_j} the identity matrix of dimension $n_j \times n_j$. Fan & Zhang (2000) obtain their raw estimates at the first step by fitting the linear model in (5) at each time point. However, we do not observe Y_j and X_{qpj}, but only observe their noisy counterparts Y'_j and X'_{qpj}. The Fan & Zhang (2000) raw estimates $(X'^{\mathrm{T}}_{qpj}X'_{qpj})^{-1}X'^{\mathrm{T}}_{qpj}Y'_j$ which correspond to a linear regression fit can easily handle the additive measurement error in the response. However, because of the measurement error in the predictors, these raw estimates in general will not target the $\beta(t_j)$ in (5). More explicitly, consider the target of Fan & Zhang (2000) raw estimates for the simple case of $p = 1$, which is

$$\frac{\mathrm{cov}\{y'_j, x'(t_{j-q})\}}{\mathrm{var}\{x'(t_{j-q})\}} = \frac{\beta_{1j}\mathrm{var}\{x(t_{j-q})\}}{\mathrm{var}\{x'(t_{j-q})\}} = \beta_{1j}\left(\frac{\mathrm{var}\{x(t_{j-q})\}}{\mathrm{var}\{x(t_{j-q})\} + \mathrm{var}(e_{xj-q})}\right) = \beta_{1j}\zeta_j,$$

where $e_{xj-q} = x'(t_{j-q}) - x(t_{j-q})$. As the values of ζ_j range between 0 and 1, Fan & Zhang (2000) raw estimates potentially underestimate the target function $\beta_1(t_j)$. The resulting bias can become arbitrarily large as the error variance increases and ζ_j moves close to zero.

An alternative is therefore needed for the case of contaminated predictors. We note that this problem can be equivalently viewed as finding an instrumental variable for the problem at hand. We demonstrate that the following estimator (6) indeed provides a construction of such an instrumental variable; compare with Carroll et al., (2004). Our proposed estimator $\beta(t_j)$ is

$$b_{qp}(t_j) = (b_{0j}, b_{1j}, \ldots, b_{pj})^{\mathrm{T}} = (X'^{\mathrm{T}}_{qpj-p}M_{j-p,j}X'_{qpj})^{-1}X'^{\mathrm{T}}_{qpj-p}M_{j-p,j}Y'_j, \tag{6}$$

for $j = q + 2p, \ldots, T$. Here $M_{j-p,j}$ denotes a $n_{j-p} \times n_j$ matrix for which the (a, b)th entry equals 1 if the ath entry of Y'_{j-p} and the bth entry of Y'_j come from the same subject, and equals 0 otherwise.

The estimator in (6) targets the correct value $\beta(t_j)$ since

$$\begin{aligned} E(X'^{\mathrm{T}}_{qpj-p}M_{j-p,j}Y'_j) &= E(X^{\mathrm{T}}_{qpj-p}M_{j-p,j}Y_j), \\ E(X'^{\mathrm{T}}_{qpj-p}M_{j-p,j}X'_{qpj}) &= E(X^{\mathrm{T}}_{qpj-p}M_{j-p,j}X_{qpj}) \end{aligned}$$

are not affected by the measurement error in the predictors. This is because X'^{T}_{qpj-k} and X'_{qpj} do not contain any predictors evaluated at the same time-points if $k \geq p$. In most situations it is plausible to assume that $\mathrm{cov}\{x(t_{j-q-(r-1)}), x(t_{j-q-k-(r-1)})\}$, $r = 1, \ldots, p$, which is known to be inversely proportional to the variance of $b_{qp}(t_j)$ by standard least-squares theory, becomes smaller as the jump k between the time-points increases. Therefore, the size of the jump is chosen as $k = p$ in (6), the smallest acceptable value to deal with measurement error, with the aim of keeping the variance of $b_{qp}(t_j)$ as small as possible. With the same argument, it is clear that the choice of no jump, for which the proposed raw estimator reduces to the raw estimator of Fan & Zhang (2000), entails an estimator that has smallest variance compared to other choices of k. However, as pointed out earlier, the estimator with no jump is extremely vulnerable to measurement error in the predictors. Thus, there is a trade-off between variance and robustness to measurement error in the predictors. A recommended strategy is to compute both estimators in applications and to choose the robust estimator with jumps if the estimator with no jump yields smaller estimates in absolute value consistently for all the time-points, indicating presence of measurement error. Otherwise, the estimator with no jump should be preferred since it has smaller variance.

2.3 *Asymptotic properties and finite sample inference*

The following results are obtained assuming that the window width p and the lag parameter q are known. Recall that C_j contains the subject indices of those subjects observed at time t_j and $(t_{j-q}, \ldots, t_{j-q-p+1})$. Let $n_{j-p,j}$ denote the number of subjects in $C_{j-p} \cap C_j$. Further define

$$
\begin{aligned}
\mathcal{X}_j &= E(n_{j-p,j}^{-1} X_{qpj-p}^{\prime\mathrm{T}} M_{j-p,j} X'_{qpj}) \\
&= \begin{bmatrix}
1 & E\{x'(t_{j-q})\} & \cdots & E\{x'(t_{j-q-p+1})\} \\
E\{x'(t_{j-q-p})\} & E\{x'(t_{j-q-p})x'(t_{j-q})\} & \cdots & E\{x'(t_{j-q-p})x'(t_{j-q-p+1})\} \\
\vdots & & \ddots & \vdots \\
E\{x'(t_{j-q-2p+1})\} & E\{x'(t_{j-q})x'(t_{j-q-2p+1})\} & \cdots & E\{x'(t_{j-q-p+1})x'(t_{j-q-2p+1})\}
\end{bmatrix},
\end{aligned}
$$

and $(\Sigma_j)_{s,s'}$ to be equal to

$$
\left\{
\begin{array}{cc}
E\{x'(t_{j-q-p-s+2})x'(t_{j-q-p-s'+2})\}\eta_{qpj} & \text{for} \quad 2 \le s, s' \le p+1 \\
E\{x'(t_{j-q-p-s'+2})\}\eta_{qpj} & \text{for} \quad s = 1,\ 2 \le s' \le p+1 \\
E\{x'(t_{j-q-p-s+2})\}\eta_{qpj} & \text{for} \quad s' = 1,\ 2 \le s \le p+1 \\
\eta_{qpj} & \text{for} \quad s = s' = 1
\end{array}
\right\},
$$

for all time-points t_j such that $j = q + 2p, \ldots, T$, where σ_y^2, σ_x^2 and δ_j are defined in §2.1, and $\eta_{qpj} = \delta_j + \sigma_y^2 + \sum_{r=1}^p \beta_r^2(t_j)\sigma_x^2$. Here, $(\Sigma_j)_{s,s'}$ denotes the (s,s')th element of Σ_j. The following result gives the asymptotic distribution of the proposed estimates assuming the case of missing completely at random.

Theorem 1. *Under the technical conditions* A1 − A3 *given in the Appendix, it holds that*

$$\sqrt{n_{j-p,j}}\{b_{qp}(t_j) - \beta(t_j)\} \to N(0_{p+1}, \mathcal{X}_j^{-1}\Sigma_j\mathcal{X}_j^{-1})$$

in distribution as $n_{j-p,j} \to \infty$ *for all time-points* t_j *such that* $j = q+2p, \ldots, T$.

The estimates given in (6) are not necessarily smooth, and a second smoothing step in the estimation procedure may be beneficial in improving the efficiency of the estimates, as well as imputing occasional missing values. The smoothing step would be carried out for each of the r coefficients separately for $r = 0, \ldots, p$,

$$\hat{\beta}_{rqp}(t) = \sum_{j=1}^{T} w(t_j, t) b_{rj}, \tag{7}$$

where the b_{rj} are as in (6) and depend on q and p. The smoothing weights $w(t_j, t)$ can be obtained from any linear smoothing technique, such as local polynomial smoothing, as used by Fan & Zhang (2000), or spline smoothing, as used by Wu et al. (2000). In the implementation of the smoothing step (7) by local linear smoothing, let h denote the bandwidth and $K(\cdot)$ the weight function or equivalent kernel of the local polynomial fit (Fan & Gijbels, 1996). For $n_0 = \inf_j n_{j-p,j}$ assume that $n_0 \to \infty$. The following result establishes the asymptotic bias behaviour of the smoothed estimates $\hat{\beta}_{rqp}$, and in particular asymptotic unbiasedness.

Theorem 2. *Under the technical conditions* A1 − A5 *given in the Appendix, when* $h \to 0$, $Th \to \infty$, *and* $n_0 h^4 \to \infty$ *as* $T \to \infty$ *and* $n_0 \to \infty$, *it holds that*

$$\hat{\beta}_{rqp}(t) = \beta_r(t) + \frac{h^2 \beta^{(2)}(t) \int K(x) x^2 dx}{2} + o_p(h^2).$$

Similarly to Fan & Zhang's proposed bands for their smooth estimates, approximate error bands indicating the size of standard errors can be constructed around the estimators in (7), based on standard error estimates of the $b_{qp}(t_j)$. It follows from standard least squares theory that

$$\mathrm{cov}(b_{rj}, b_{rj'}|\mathcal{D}) = \begin{cases} \delta(t_j, t_{j'}) c_{r,p}^{\mathrm{T}} (X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} X'_{qpj})^{-1} X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} & \\ \times M_{j,j'} M_{j',j'-p} X'_{qpj'-p} (X'^{\mathrm{T}}_{qpj'-p} M_{j'-p,j'} X'_{qpj'})^{-1} c_{r,p} & \text{for} \quad j \neq j', \\ (\delta_j + \sigma_y^2) c_{r,p}^{\mathrm{T}} (X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} X'_{qpj})^{-1} X'^{\mathrm{T}}_{qpj-p} & \\ \times M_{j-p,j} M_{j,j-p} X'_{qpj-p} (X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} X'_{qpj})^{-1} c_{r,p} & \text{for} \quad j = j', \end{cases} \tag{8}$$

where $e_{xj} = X'_{qpj} - X_{qpj}$, $\mathcal{D} = \{(X'_{qpj}, X_{qpj}, t_j), j = 1, \ldots, T\}$ and $c_{r,p}$ denotes a p-dimensional unit vector with 1 at its rth entry.

An estimator for $\mathrm{cov}(b_{rqpj}, b_{rqpj'}|\mathcal{D})$ can be constructed based on (8) once estimators for $\delta(t_j, t_{j'})$ and $\delta_j + \sigma_y^2$ are available. To obtain estimators for $\delta(t_j, t_{j'})$ and $\delta_j + \sigma_y^2$, define $\hat{e}_{qpj} = (I_{n_j} - P_{qpj}) Y'_j$ to be the residuals from the proposed regression at time t_j, where $P_{qpj} = X'_{qpj} (X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} X'_{qpj})^{-1} X'^{\mathrm{T}}_{qpj-p} M_{j-p,j}$. If we assume that $\mathrm{tr}\{(I_j - P_{qpj}) M_{j,j'} (I_{j'} - P_{qpj'})\} \neq 0$ and $n_j > p$, a set of estimators for $\delta(t_j, t_{j'})$ and $\Delta_j := \delta_j + \sigma_y^2$ is

$$\hat{\delta}(t_j, t_{j'}) = \mathrm{tr}(\hat{e}_{qpj} \hat{e}^{\mathrm{T}}_{qpj'}) / \mathrm{tr}\{(I_j - P_{qpj}) M_{j,j'} (I_{j'} - P_{qpj'})^{\mathrm{T}}\} \tag{9}$$

$$\hat{\Delta}_j = \hat{e}^{\mathrm{T}}_{qpj} \hat{e}_{qpj} / (n_j - p), \tag{10}$$

which use the fact that

$$\mathrm{tr}\{\mathrm{cov}(\hat{e}_{qpj}, \hat{e}_{qpj'}|\mathcal{D})\} = \begin{cases} \delta(t_j, t_{j'}) \mathrm{tr}\{(I_j - P_{qpj}) M_{j,j'} (I_{j'} - P_{qpj'})\} & \text{for} \quad j \neq j' \\ (\delta_j + \sigma_y^2)(n_j - p) & \text{for} \quad j = j'. \end{cases}$$

Plugging (9) and (10) into (8) yields an estimator for $\mathrm{cov}(b_{rj}, b_{rj'}|\mathcal{D})$. Finally

$$\mathrm{var}\{\hat{\beta}_{rqp}(t)|D\} = \sum_{j=1}^{T} \sum_{j'=1}^{\mathrm{T}} w_{rqp}(t_j, t) w_{rqp}(t_{j'}, t) \mathrm{cov}(b_{rqpj}, b_{rqpj'}|\mathcal{D})$$

can be estimated by plugging in respective estimators of $\mathrm{cov}(b_{rqpj}, b_{rqpj'}|\mathcal{D})$. If we are willing to assume that the smoothers we employ use fixed smoothing windows and ignore bias, then we find from the above that ± 2 error bars can be constructed for the final smooth estimators as

$$\hat{\beta}_{rqp}(t) \pm 2\widehat{\mathrm{var}}\{\hat{\beta}_{rqp}(t)|D\}^{1/2}. \tag{11}$$

2.4 *Choosing window widths and lags*

We view the choice of the window width p and lag q as a variable selection problem as these quantities determine which predictors will be included in the proposed model. Nevertheless, one difference from a standard variable selection situation is the restriction that the final sequence of predictor times chosen as predictors has to be consecutive. For example, reasonable choices of predictors for the response at time t_j would not include $x(t_{j-1})$ and $x(t_{j-3})$ as predictors, and not $x(t_{j-2})$.

Therefore, we use a variation of the backward stepwise deletion technique of Fan et al. (2003), for our implementation. We make use of a modified AIC and partial F-statistics. We start by identifying an initial group of predictor times for modelling the response at time t_j, say $\{x(t_{j-q}), \ldots, x(t_{j-q-p+1})\}$. We then identify the least significant predictor among the two candidates which are the smallest and largest time lags, namely $x(t_{j-q})$ and $x(t_{j-q-p+1})$, according to their partial F-statistic values. This yields a reduced and a full model, where the best model would be chosen by $\text{AIC} = \log\{\text{RSS}/(n_{j,j-p} - p)\} + 2p/n_{j,j-p}$. Here RSS stands for the residual sum of squares of the fitted model at time t_j,

$$\text{RSS}_{qpj} = \sum_{i=1}^{n_{j,j-p}} \left\{ y'(t_{ij}) - b_{0j} - \sum_{r=1}^{p} b_{rj} x_i'(t_{j-q-(r-1)}) \right\}^2$$

and p, the number of predictors considered, would be equal to p in the full model of the above example.

The F-statistics considered for the coefficient estimates in the linear model at times t_j are

$$F_{rqp} = \frac{\{\text{RSS}_{qpj}(R) - \text{RSS}_{qpj}(F)\}/1}{\text{RSS}_{qpj}(F)/(n_{j,j-p} - p)},$$

for $r = 0, \ldots, p-1$, where $\text{RSS}_{qpj}(F)$ and $\text{RSS}_{qpj}(R)$ denote the residual sum of squares of the full and reduced models, with or without $x(t_{j-q-r})$, respectively. Assume for example that the F-value of $x(t_{j-q})$ is smaller than that of $x(t_{j-q-p+1})$ in the above example. In that case $x(t_{j-q})$ is deleted from the full model containing all p predictors to form the reduced model, and it is finally deleted from our set of considered predictors if the AIC of the reduced model is smaller than that of the full model. If $x(t_{j-q})$ is deleted from the original set, we restart the deletion process, this time having $\{x(t_{j-q+1}), \ldots, x(t_{j-q-p+1})\}$

as our initial set of predictors. We continue comparing the partial F-statistic values of the coefficient estimators corresponding to $x(t_{j-q+1})$ and $x(t_{j-q-p+1})$. This backward stepwise deletion is repeated until we cannot delete any further predictors. The lag, q^*, and the window width, p^*, of the final model are the chosen values for these parameters.

3. Simulation study

The goal of this simulation study is to assess the effectiveness of the proposed procedure for dealing with measurement error through the jump in the time-points. Hence, we compare the proposed estimator to an alternative estimator which can be derived from the Fan & Zhang (2000) raw estimates, with the measurement error ignored. We also explore the performance of the backward stepwise deletion technique proposed in §2.4 for the choice of window width and lag.

The data are generated from the model

$$y_i(t_j) = \beta_0(t_j) + \beta_1(t_j)x_i(t_{j-1}) + \beta_2(t_j)x_i(t_{j-2}) + \epsilon_i(t_j),$$

for $j = 1, \ldots, 20$ and $i = 1, \ldots, 64$, with lag $q = 1$ and window width $p = 2$. The time-points $t_1, \ldots, t_{20}$ are are chosen to be equidistant between 0.01 and 1, and the coefficient functions are $\beta_0(t) = 250 + 200\sin(3\pi t)$, $\beta_1(t) = -200 - 180t$ and $\beta_2(t) = 50 + 150t^2$. Predictor and error processes are both generated from multivariate normal distributions with decaying covariance structures,

$$\text{cov}\{x_i(t_j), x_i(t_{j'})\} = 6e^{-8|t_j - t_{j'}|^2}, \qquad \text{cov}\{\epsilon_i(t_j), \epsilon_i(t_{j'})\} = 0.15e^{-0.3|t_j - t_{j'}|},$$

and means $20 + 180t_j^2$ and 0, respectively. The predictor and response are observed with additive measurement error, and are denoted by

$$x_i'(t_j) = x_i(t_j) + e_{xij}, \qquad y_i'(t_j) = y_i(t_j) + e_{yij}.$$

The measurement errors e_{yij} and e_{xij} are simulated to be independently and identically distributed, both over i and j, normal random variables with means 0 and standard deviations 0.15 and 0.1, respectively.

The number of repeated measurements for each subject is generated randomly between 1 and 20. Thus, potentially unequal numbers of observations are taken on each subject, and on average 30% of the data are missing. This yields 14 repetitions per subject on average, and roughly 45 data-points observed at a given time t_j.

To explore the performance of the backward stepwise deletion technique in §2.4 for the choice of window widths and lags, we apply the method to variable selection from the initial set of predictors $\{x(t_{j-1}), x(t_{j-2}), x(t_{j-3})\}$ at each time point. We ran 1000 simulations to estimate the deletion frequencies of these three predictors and these are given in Fig. 1 (c). The simulations indicate that we should keep two time-points in the predictor model, but not three, which is in line with the simulation model. The downward trend in the deletion frequency of the second predictor is also as expected, since $\beta_2(t)$ increases substantially as t moves from 0 to 1.

To assess the effectiveness of the estimation strategy as implemented in estimator (6), we compare the two algorithms, one with the jump in the time-points as in (6), and the other directly derived from the Fan & Zhang (2000) raw estimates, with the additive measurement error in the predictor ignored. The means of resulting estimates and their ± 2 error bars over 1000 Monte Carlo runs are shown in Fig. 2. The estimates that ignore the measurement error, dash-dotted, clearly deviate considerably further from the target function: the ± 2 error bars for the two slope estimates do not contain the true coefficient functions, solid curve. The proposed estimators (6), dotted, which are reasonably close to the target functions have wider error bars, as expected, since their variance is larger than that of the unadjusted estimates.

Other measures of the performance of the fits obtained by the two estimates are the mean absolute deviation error and the weighted average squared error, defined as

$$\text{MADE} = \left(3\sum_j 1\right)^{-1} \sum_{r=0}^{2}\sum_j \frac{|\beta_r(t_j) - \hat{\beta}_r(t_j)|}{\text{range}(\beta_r)}, \; \text{WASE} = \left(3\sum_j 1\right)^{-1} \sum_{r=0}^{2}\sum_j \frac{\{\beta_r(t_j) - \hat{\beta}_r(t_j)\}^2}{\text{range}^2(\beta_r)},$$

where range(β_r) is the range of the function $\beta_r(t)$, and the sums over j are taken over $j = q + 2p, \ldots, T$. We also consider unweighted average squared error, UASE which is defined in the same way as WASE, but without any weights in the denominator. Box-plots of the ratios of the values of MADE, WASE and UASE of the proposed method over the unadjusted estimator from 1000 Monte Carlo runs are given in Fig. 1 (a). The plots indicate that the proposed estimators indeed handle measurement error in the predictors much better than do the unadjusted estimators. We have also compared the two estimators under no measurement error and the respective box plots of MADE, WASE and UASE ratios are given in Fig. 1 (b). In this case the estimates ignoring the measurement error perform better in this case than the proposed robust estimates, as expected, because of their smaller variance.

4. Application to protein data

A motivation for this study was the investigation of longitudinal relationships between the levels of positive acute phase proteins such as C-reactive protein, CRP, and negative acute phase proteins such as transferrin, TRF; see Kaysen et al. (2001) for background in the context of haemodialysis. In the Kaysen et al. study, the levels of acute phase proteins were recorded for 64 hemodialysis patients. The number of repeated measurements for the 64 patients ranged from 9 to 39 per patient, and the visits were on average a month apart. Of particular interest are relationships between negative and positive acute phase proteins. We aim at predicting transferrin from previous C-reactive protein levels and consider models that regress transferrin levels at time j on past values of C-reactive protein levels, recorded possibly at times t_{j-1}, t_{j-2} and t_{j-3}. Here the unit of time is one month.

Accordingly, we start with the initial model

$$\mathrm{TRF}_i(t_j) = \beta_0(t_j) + \beta_1(t_j)\mathrm{CRP}_i(t_{j-1}) + \beta_2(t_j)\mathrm{CRP}_i(t_{j-2}) + \beta_3(t_j)\mathrm{CRP}_i(t_{j-3}) + \epsilon_i(t_j),$$

for which $q = 1$ and $p = 3$, and then apply the proposed backward variable selection technique to choose the final predictors, which corresponds to choosing q and p. The predictor $\mathrm{CRP}(t_{j-2})$ turns out to be the only one that is significant for more than half of the time-points considered. Thus we eliminate $\mathrm{CRP}(t_{j-1})$ and $\mathrm{CRP}(t_{j-3})$ from the initial model, and choose the one with $q = 2$ and $p = 1$, leading to the model

$$\mathrm{TRF}_i(t_j) = \beta_0(t_j) + \beta_1(t_j)\mathrm{CRP}_i(t_{j-2}) + \epsilon_i(t_j).$$

The proposed estimates for coefficient functions $\beta_0(t)$ and $\beta_1(t)$ obtained for this model are illustrated in Fig. 3, including the corresponding unadjusted estimates that ignore measurement error. The ± 2 error bars (11) for the coefficient functions are also shown.

The coefficient functions obtained from the proposed and the unadjusted estimates clearly differ. The proposed method leads to more pronounced coefficient functions for C-reactive protein, consistently for all time-points. This indicates that measurement error is indeed present in the predictors, which is also consistent with the underlying biology, and that it is masking the true predictive effects of C-reactive protein. The error bands indicate a degree of significance of the coefficient function β_1 at $t \simeq 600$ days, if we ignore the fact that the bars are pointwise and approximate. In contrast to the previous analysis of Kaysen et al. (2001), in which only lags of one month were considered, the proposed model indicates that lags of two months are particularly relevant. This points to lingering effects of C-reactive protein levels on negative acute phase proteins such as transferrin that extend well beyond one month.

Acknowledgement

We are extremely grateful to anonymous referees and the editor for many helpful remarks that improved the exposition of the paper. This research was supported in part by grants from the U.S. National Science Foundation.

Appendix

Technical details

We introduce some technical conditions. For a fixed time-point t_j such that $q + 2p \leq j \leq T$, we impose the following conditions.

Condition A1. The matrices $\tilde{\mathcal{X}}_j = (n_{j-p,j}^{-1} X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} X'_{qpj})$ and $\mathcal{X}_j = E(n_{j-p,j}^{-1} X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} X'_{qpj})$ are nonsingular.

Condition A2. The variances of $x'(t_{j-q-s})$, $x'(t_{j-q-p-s'})$, $\{x'(t_{j-q-s})x'(t_{j-q-p-s'})\}$ and the expected values of $x'(t_{j-q-p-s'})$, $\{x'(t_{j-q-p-s'})x'(t_{j-q-p-s})\}$ are finite for all $s, s' = 0, \ldots, p-1$.

Condition A3. It holds that $E\{\epsilon_i^2(t_j)\}$, σ_y^2 and σ_x^2 are finite.

Condition A4. Conditions and bounds in A1, A2 and A3 hold uniformly in j. For condition A1, this implies that $\inf_j \det|\tilde{\mathcal{X}}_j| > 0$ and $\inf_j \det|\mathcal{X}_j| > 0$.

Condition A5. The functions β_r are twice continuously differentiable, and the kernel K is a continuous density function with finite second moment.

Proof of Theorem 1. Define the vectors $\epsilon_j = (\epsilon_1(t_j), \ldots, \epsilon_{n_j}(t_j))^{\mathrm{T}}$, $e_{yj} = (e_{y1j}, \ldots, e_{yn_jj})^{\mathrm{T}}$ and the matrices

$$e_{xj} = X'_{qpj} - X_{qpj} = \begin{bmatrix} 0 & e_{x,1,j-q} & \cdots & e_{x,1,j-q-p+1} \\ \vdots & \vdots & & \vdots \\ 0 & e_{x,n_j,j-q} & \cdots & e_{x,n_j,j-q-p+1} \end{bmatrix}, \quad Z_j = (n_{j-p,j}^{-1} X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} X'_{qpj}).$$

Then

$$\begin{aligned} b_{qp}(t_j) &= (X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} X'_{qpj})^{-1} X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} Y'_j \\ &= (X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} X'_{qpj})^{-1} X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} X'_{qpj} \beta(t_j) \\ &\quad + (X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} X'_{qpj})^{-1} X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} \{Y'_j - X'_{qpj}\beta(t_j)\} \\ &= \beta(t_j) + (X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} X'_{qpj})^{-1} X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} \{\epsilon_j + e_{yj} - e_{xj}\beta(t_j)\} \\ &= \beta(t_j) + (n_{j-p,j}^{-1} X'^{\mathrm{T}}_{qpj-p} M_{j-p,j} X'_{qpj})^{-1} \end{aligned}$$

$$\times \begin{bmatrix} n_{j-p,j}^{-1} \sum_{i \in C_{j-p,j}} \{\epsilon_i(t_j) + e_{yij} - \sum_{r=1}^{p} e_{x,i,j-q-(r-1)}\beta_r(t_j)\} \\ n_{j-p,j}^{-1} \sum_{i \in C_{j-p,j}} x'_i(t_{j-q-p}) \{\epsilon_i(t_j) + e_{yij} - \sum_{r=1}^{p} e_{x,i,j-q-(r-1)}\beta_r(t_j)\} \\ \vdots \\ n_{j-p,j}^{-1} \sum_{i \in C_{j-p,j}} x'_i(t_{j-q-2p+1}) \{\epsilon_i(t_j) + e_{yij} - \sum_{r=1}^{p} e_{x,i,j-q-(r-1)}\beta_r(t_j)\} \end{bmatrix}$$

$$\equiv \beta(t_j) + (Z_j)^{-1}_{(p+1)\times(p+1)} (R_j)_{(p+1)\times 1},$$

where $C_{j-p,j}$ is the set of subject indices such that $C_{j-p,j} = C_{j-p} \cap C_j$, and $E(R_j) = 0_{p+1} \equiv R$. It holds that

$$\begin{aligned} Z_j^{-1} R_j &= \mathcal{X}_j^{-1}(R_j - R) - \mathcal{X}_j^{-1}(Z_j - \mathcal{X}_j) Z_j^{-1} R_j + \mathcal{X}_j^{-1} R \\ &= \mathcal{X}_j^{-1}(R_j - 0_{p+1}) - \mathcal{X}_j^{-1}(Z_j - \mathcal{X}_j) Z_j^{-1} R_j, \end{aligned}$$

where $\mathcal{X}_j$ is as defined before Theorem 1. Therefore,

$$\sqrt{n_{j-p,j}}(Z_j^{-1} R_j) = \mathcal{X}_j^{-1} \sqrt{n_{j-p,j}}(R_j - 0_{p+1}) - \mathcal{X}_j^{-1} \sqrt{n_{j-p,j}}(Z_j - \mathcal{X}_j) Z_j^{-1} R_j. \qquad \text{(A1)}$$

If we use (A1) and if we can show that

$$\sqrt{n_{j-p,j}}(R_j - 0_{p+1}) \to N(0_{p+1}, \Sigma_j), \qquad \text{(A2)}$$

in distribution,

$$\sqrt{n_{j-p,j}}(Z_j - \mathcal{X}_j) = O_p(1)(1_{p+1}1_{p+1}^{\mathrm{T}}), \tag{A3}$$

$$Z_j^{-1} \to \mathcal{X}_j^{-1}, \tag{A4}$$

in probability where 1_{p+1} denotes a vector of $p+1$ ones, then Theorem 1 follows. Since

$$Z_j = n_{j-p,j}^{-1} \sum_{i \in C_{j-p,j}} \begin{bmatrix} 1 & x_i'(t_{j-q}) & \dots & x_i'(t_{j-q-p+1}) \\ x_i'(t_{j-q-p}) & x_i'(t_{j-q-p})x_i'(t_{j-q}) & \dots & x_i'(t_{j-q-p})x_i'(t_{j-q-p+1}) \\ \vdots & & \ddots & \vdots \\ x_i'(t_{j-q-2p+1}) & x_i'(t_{j-q})x_i'(t_{j-q-2p+1}) & \dots & x_i'(t_{j-q-p+1})x_i'(t_{j-q-2p+1}) \end{bmatrix},$$

by condition A2, $E(Z_j - \mathcal{X}_j)_{s,s'} = 0$ and $\mathrm{var}(Z_j - \mathcal{X}_j)_{s,s'} = O_p(n_{j-p,j}^{-1})$, for $s, s' = 1, \dots, p+1$, (A3) follows. It follows from the Law of Large Numbers that $Z_j \to \mathcal{X}_j$, in probability. Now consider

$$\det(Z_j) = \sum_{\ell=1}^{(p+1)!} (-1)^{\mathrm{sign}(\tau)} (Z_j)_{1\tau_\ell(1)} \dots (Z_j)_{p+1,\tau_\ell(p+1)},$$

where the sum is taken over all permutations τ_ℓ of $(1, \dots, p+1)$, and $\mathrm{sign}(\tau)$ equals $+1$ or -1, depending on whether τ can be written as the product of an even or odd number of transpositions. Let $Z_j^{-ss'}$ denote the matrix obtained after deleting the sth row and s'th column of Z_j. Then the cofactor of $(Z_j)_{ss'}$ is defined by $(-1)^{s+s'}$ times the determinant of $Z_j^{-ss'}$ and thus equals $(-1)^{s+s'} \sum_{\ell=1}^{p!} (-1)^{\mathrm{sign}(\tau)} (Z_j^{-ss'})_{1\tau_\ell(1)} \dots (Z_j^{-ss'})_{p,\tau_\ell(p)}$. Therefore, the (s, s')th element of Z_j^{-1} is equal to

$$(Z_j^{-1})_{ss'} = \frac{(-1)^{s+s'} \sum_{\ell=1}^{p!} (-1)^{\mathrm{sign}(\tau)} (Z_j^{-ss'})_{1\tau_\ell(1)} \dots (Z_j^{-ss'})_{p,\tau_\ell(p)}}{\sum_{\ell=1}^{(p+1)!} (-1)^{\mathrm{sign}(\tau)} (Z_j)_{1\tau_\ell(1)} \dots (Z_j)_{p+1,\tau_\ell(p+1)}}$$

for $s, s' = 1, \dots, p+1$. Since $Z_j \to \mathcal{X}_j$, in probability,

$$(Z_j^{-1})_{ss'} \to \frac{(-1)^{s+s'} \sum_{\ell=1}^{p!} (-1)^{\mathrm{sign}(\tau)} (\mathcal{X}_j^{-ss'})_{1\tau_\ell(1)} \dots (\mathcal{X}_j^{-ss'})_{p,\tau_\ell(p)}}{\sum_{\ell=1}^{(p+1)!} (-1)^{\mathrm{sign}(\tau)} (\mathcal{X}_j)_{1\tau_\ell(1)} \dots (\mathcal{X}_j)_{p+1,\tau_\ell(p+1)}} = (\mathcal{X}_j^{-1})_{ss'},$$

in probability and (A4) follows.

Result (A2) follows by the Central Limit Theorem, given conditions A2 and A3, where Σ_j is as defined before Theorem 1. Theorem 1 then follows from (A1).

Proof of Theorem 2. We fix the index r and suppress it in the following. Consider

$$\begin{aligned} |\hat{\beta}(t) - \beta(t)| &= \left| \sum_{j=1}^{T} w(t_j, t) b_{qp}(t_j) - \beta(t) \right| \\ &\le \left| \sum_{j=1}^{T} w(t_j, t) b_{qp}(t_j) - \sum_{j=1}^{T} w(t_j, t)\beta(t_j) \right| + \left| \sum_{j=1}^{T} w(t_j, t)\beta(t_j) - \beta(t) \right| = A + B, \end{aligned}$$

say. From the Cauchy-Schwarz inequality, A can be bounded as follows:

$$\begin{aligned} A &= \left|\sum_{j=1}^{T} w(t_j,t)\{b_{qp}(t_j)-\beta(t_j)\}\right| \le \sum_{j=1}^{T} |w(t_j,t)|\mathcal{I}_{\{w(t_j,t)\neq 0\}} \sup_k |b_{qp}(t_k)-\beta(t_k)| \\ &= O_p(n_0^{-1/2}) \sum_{j=1}^{T} |w(t_j,t)|\mathcal{I}_{\{w(t_j,t)\neq 0\}} \le O_p(n_0^{-1/2}) \left\{\sum_{j=1}^{T} w^2(t_j,t)\right\}^{1/2} (Th)^{1/2} = O_p(n_0^{-1/2}). \end{aligned}$$

Here, $\mathcal{I}_{\{w(t_j,t)\neq 0\}}$ denotes the indicator function that $w(t_j,t)$ is not zero and we have used $\sum_j w^2(t_j,t) = O\{(Th)^{-1}\}$ and $\sup_k |b_{qp}(t_k)-\beta(t_k)| = O_p(n_0^{-1/2})$, which follows from condition A4 and arguments in the proof of Theorem 1. Since $n_0 h^4 \to \infty$, we conclude that $A= o_p(h^2)$. From well-known facts about local linear fits for equidistant designs, $B= h^2\beta^{(2)}(t)\int K(x)x^2dx/2 + o(h^2)$ and Theorem 2 follows.

References

Cai, Z., Fan, J. & Yao, Q. (2000). Functional coefficient regression models for nonlinear time series. *J. Am. Statist. Assoc.* **95,** 941-56.

Carroll, R., Ruppert, D., Tosteson, T., Crainiceanu, C. & Karagas, M. (2004). Instrumental variables and nonparametric regression. *J. Am. Statist. Assoc.* **99,** 736-50.

Chen, R. & Tsay, R. S. (1993). Functional coefficient autoregressive models. *J. Am. Statist. Assoc.* **88,** 298-308.

Chiang, C., Rice, J. A. & Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *J. Am. Statist. Assoc.* **96,** 605-17.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* London: Chapman and Hall.

Fan, J. & Zhang, J. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J.R.S.S.* B **62,** 303-22.

Fan, J., Yao, Q. & Cai, Z. (2003). Adaptive varying coefficient linear models. *J.R.S.S.* B **65,** 57-80.

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient model (with Discussion). *J.R.S.S.* B **55,** 757-96.

Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data.

Biometrika **85,** 809-22.

Huang, J. Z., Wu, C. O. & Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14,** 763-88.

Kaysen, G. A., Dubin, J. A., Müller, H. G., Mitch, W. E., Levin, N. W. & HEMO Study Group (2001). Levels of $\alpha 1$ acid glycoprotein and ceruloplasmin predict future albumin levels in hemodialysis patients. *Kidney Int.* **60,** 2360-6.

Wu, C. O. & Chiang, C. T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statist. Sinica* **10,** 433-56.

Wu, C. O. & Yu, K. F. (2002). Nonparametric varying coefficient models for the analysis of longitudinal data. *Int. Statist. Rev.* **70,** 373-93.

Wu, C. O., Yu, K. F. & Chiang, C. T. (2000). A two-step smoothing method for varying coefficient models with repeated measurements. *Ann. Inst. Statist. Math.* **52,** 519-43.

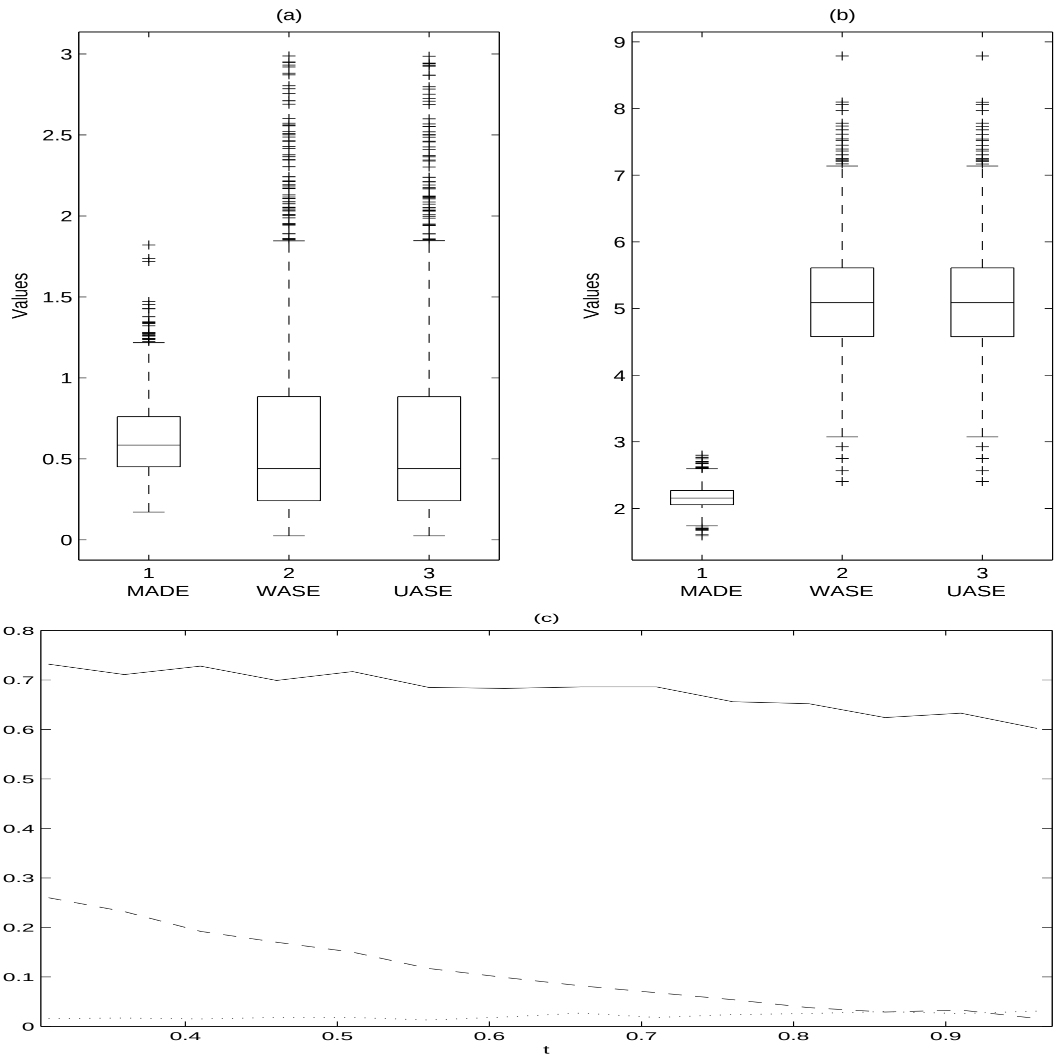

Figure 1: (a) Box-plots for the ratios of error measures for proposed estimates over unadjusted estimates, for MADE, WASE and UASE. Quotients smaller than one show that the proposed method is superior in the presence of measurement error. The box plots are based on ratios obtained from 1000 Monte Carlo runs. (b) Box-plots for the ratios of MADE, WASE and UASE for the case of no measurement error. (c) Deletion frequencies of the predictors $x(t_{j-1})$, dotted, $x(t_{j-2})$, dashed, and $x(t_{j-3})$, solid, from §3, based on 1000 simulation runs.

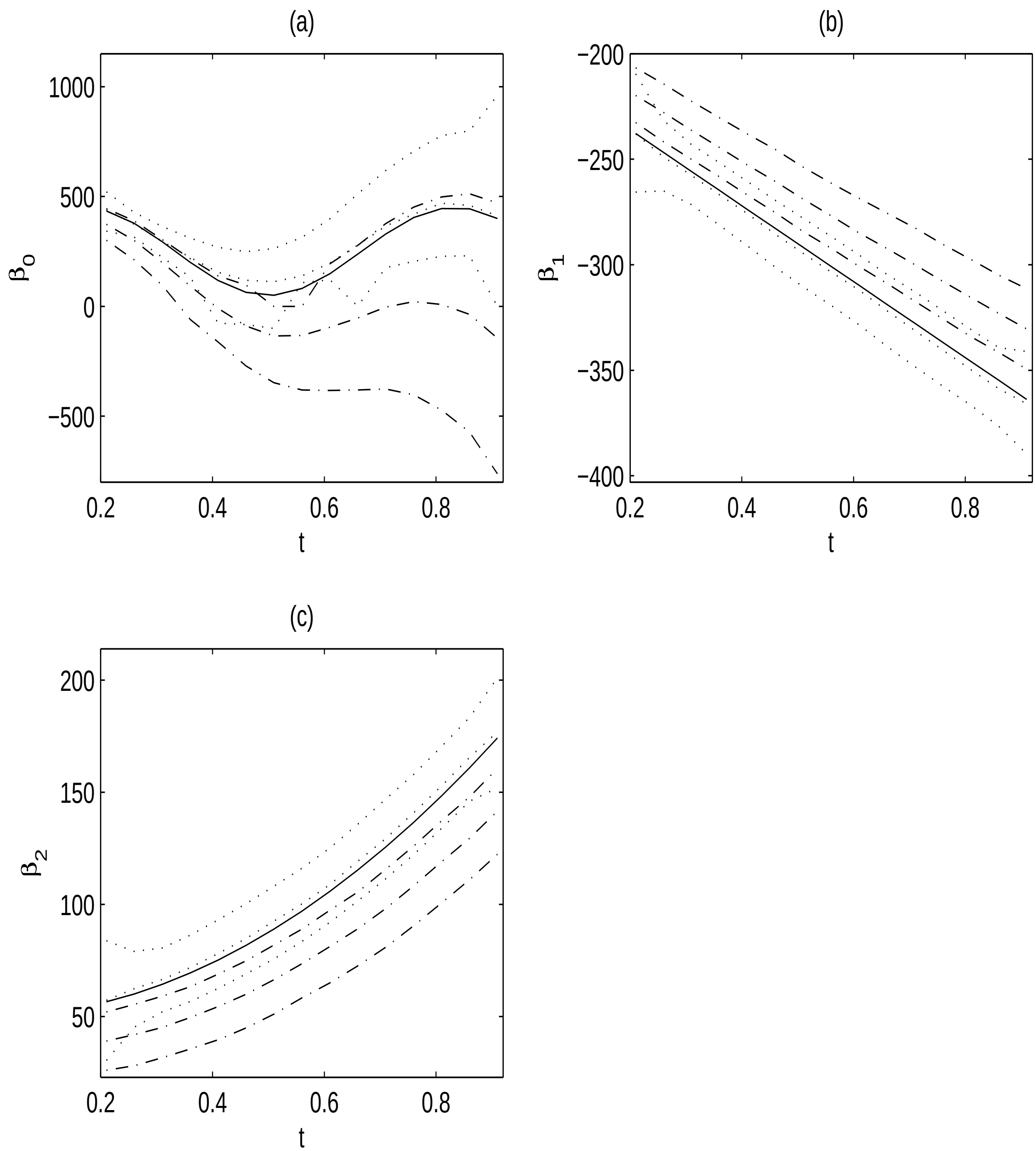

Figure 2: The cross-sectional mean curves of the proposed estimates with jump in the time-points, dotted, and the corresponding unadjusted estimates, dash-dotted, along with their ± 2 error bars for the true coefficient functions, for (a) $\beta_0(t)$, (b) $\beta_1(t)$, (c) $\beta_2(t)$.

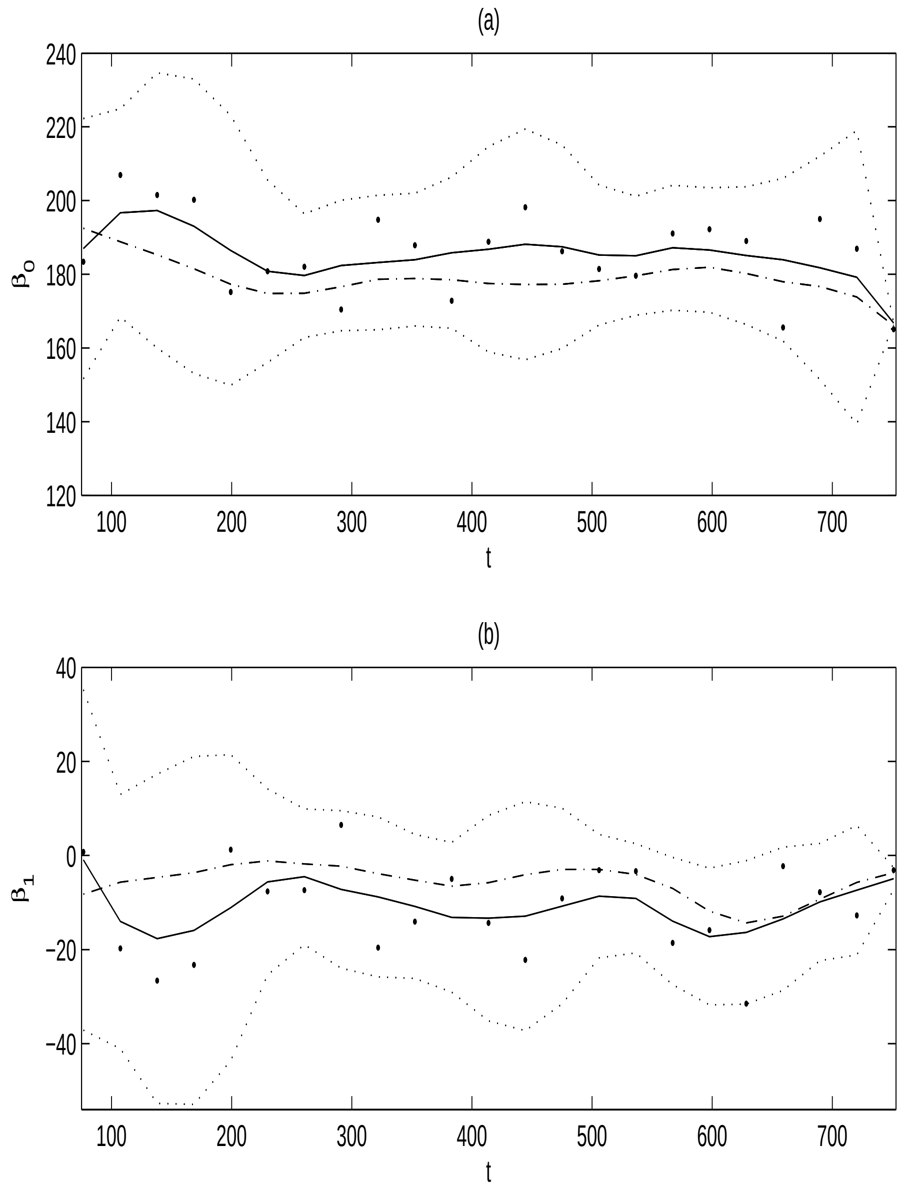

Figure 3: Smooth fits, solid, fitted to the proposed estimates, dots, and to the estimates that do not adjust for measurement error, dash-dotted, along with proposed ± 2 error bars, dotted, for the true coefficient functions (a) $\beta_0(t)$ and (b) $\beta_1(t)$ in the model $\mathrm{TRF}_i(t_j) = \beta_0(t_j) + \beta_1(t_j)\mathrm{CRP}_i(t_{j-2}) + \epsilon_i(t_j)$. The crossvalidation bandwidth choices for local polynomial fits are 80 and 70 for β_0 and β_1, respectively.