Title

Exploratory time varying lagged regression: modeling association of cognitive and functional trajectories with expected clinic visits in older adults.

Permalink

https://escholarship.org/uc/item/2gd3w5pr

Authors

Sentürk, Damla
Ghosh, Samiran
Nguyen, Danh

Publication Date

2014-05-01

DOI

10.1016/j.csda.2013.11.001

Published in final edited form as:
Comput Stat Data Anal. 2014 May 1; 73: 1–15. doi:10.1016/j.csda.2013.11.001.

Exploratory time varying lagged regression: modeling association of cognitive and functional trajectories with expected clinic visits in older adults

Damla Şentürk[a,*], **Samiran Ghosh**[b,c], and **Danh V. Nguyen**[d,e]

[a]Department of Biostatistics, University of California, Los Angeles, CA, USA

[b]Department of Family Medicine and Public Health Sciences, Wayne State University School of Medicine, MI, USA

[c]Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, MI, USA

[d]Department of Medicine, University of California, Irvine, CA, USA

[e]Institute for Clinical and Translation Science, University of California, Irvine, CA, USA

Abstract

Motivated by a longitudinal study on factors affecting the frequency of clinic visits of older adults, an exploratory time varying lagged regression analysis is proposed to relate a longitudinal response to multiple cross-sectional and longitudinal predictors from time varying lags. Regression relations are allowed to vary with time through smooth varying coefficient functions. The main goal of the proposal is to detect deviations from a concurrent varying coefficient model potentially in a subset of the longitudinal predictors with nonzero estimated lags. The proposed methodology is geared towards irregular and infrequent data where different longitudinal variables may be observed at different frequencies, possibly at unsynchronized time points and contaminated with additive measurement error. Furthermore, to cope with the curse of dimensionality which limits related current modeling approaches, a sequential model building procedure is proposed to explore and select the time varying lags of the longitudinal predictors. The estimation procedure is based on estimation of the moments of the predictor and response trajectories by pooling information from all subjects. The finite sample properties of the proposed estimation algorithm are studied under various lag structures and correlation levels among the predictor processes in simulation studies. Application to the clinic visits data show the effect of cognitive and functional impairment scores from varying lags on the frequency of the clinic visits throughout the study.

Keywords

Functional data analysis; Irregular; infrequent and unsynchronized longitudinal design; Lagged effects; Varying coefficient models; Transfer functions

[*]Corresponding address: Department of Biostatistics, UCLA Fielding School of Public Health, Los Angeles, CA 90095-1772, USA. dsenturk@ucla.edu; tel.: +1 310 825 5250; fax: +1 310 267 2113.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Consider the standard varying coefficient model (VCM; Cleveland et al., 1991; Hastie and Tibshirani, 1993),

$$E\{Y(t)-\mu_Y(t)\}=\beta_1(t)\{X(t)-\mu_X(t)\} \quad (1)$$

where the regression function $\beta_1(t)$ is allowed to vary with time. The VCMs have been widely used in longitudinal data analysis in the past decade (e.g., see Fan and Zhang, 2000; 2008; Huang et al., 2002; Hoover et al., 1998; Wu and Chiang, 2000). When the time index is set to the duration of the longitudinal study, the regression function displays the varying relation between the longitudinal response and the predictor throughout the study. Note that in (1) the regression relation is modeled between only the concurrent/ contemporaneous times of the response and the predictor. However, in some applications it is of interest to predict or associate the response at the current time with lagged times of the predictor; e.g., a subset of previous predictor values. For example, Senturk and Mueller (2008) and Koru-Sengul et al. (2007) discovered lagged relations between acute phase protein levels and between a child's growth index and maternal cigarette smoking and alcohol use, respectively.

We propose a time varying lagged regression model to assess the association between predictors, including cognitive and functional impairment scores, with the frequency of the clinic visits of older adults aged 65 to 93. The approach focuses on exploratory modeling of lagged association between previous cognitive and functional impairment statuses with current clinic visits, by sequential conditional modeling where lags are chosen to optimize a normalized covariation between the response and the predictor processes. Such a modeling approach provides important information potentially useful for individual prognosis assessment as well as formulation of managed care strategies. Informative lagged predictors (e.g., based on routine visit assessments of cognitive and functional impairment) can be used as markers to monitor future activities, such as intervention adherence or health care services utilization, for instance. The data which motivates our model development consists of measurements taken annually on multiple health scores (cognitive and functional impairment scores) as well as the total number of clinic visits every three months for four years on 703 older adults. Challenges with the motivating data requires several modeling innovations; although useful in other contexts, existing methods are not directly applicable. The observations are from infrequent time points, where the response and predictors are not necessarily observed at concurrent times. Existing useful models such as the lagged VCM (Koru-Sengul et al., 2007; Senturk and Mueller, 2008) requires equidistant time grid for estimation. Also, although the recent approach of Mueller and Yang (2010) based on transfer functions, can handle irregular and infrequent data, it is not practical as the number of predictors increase. Therefore, several significant modeling challenges are addressed in the current work, including data sparsity, non-synchronicity of measurement times and the curse of dimensionality.

We first briefly review the aforementioned existing models that explore lagged effects. To study lagged predictor effects, Senturk and Mueller (2008) and Koru-Sengul et al. (2007) proposed lagged varying coefficient models

$$E\{Y(t)-\mu_Y(t)\}=\sum_{r=1}^{p}\beta_r(t)\{X(t-r)-\mu_X(t-r)\},$$

where a separate varying coefficient function explains the time dependent effect of the predictor from each lag $t - r$. In this model, an equidistant grid is assumed for the observation times, where r denotes size of the lag on this equidistant grid. Koru-Sengul et al. (2007) also proposed an imputation algorithm to fill occasional missing values in the equidistant grid, although this would be impossible for irregular data where subjects are observed at subject specific observation times. Mueller and Yang (2010) proposed the transfer functions

$$E\{Y(t) - \mu_Y(t)|X(s)\}=\beta(t,s)\{X(s) - \mu_X(s)\}, \quad (2)$$

for jointly Gaussian processes where the transfer function $\beta(t, s)$ reflects the effect of a lag ($s < t$) of the predictor process on the value of the response at the current time t. Unlike the prior proposals of lagged varying coefficient models, this general model can be estimated from irregular and infrequent data which may not be observed concurrently. However the dimension of the transfer function will increase with the number of predictors considered, hence the model is only feasible for a single predictor process in many applications.

In this work, we propose an **e**xploratory time **var**ying **lag**ged (EVarlag) regression model that addresses these challenges to analyze the aforementioned data. The main goal of the proposed model is to embed the classical VCM in a larger class of models to detect deviations from the concurrent nature of the classical VCM via estimated time varying lags. The EVarlag model (for a single predictor) is

$$E\{Y(t) - \mu_Y(t)\}=\beta_1(t)\{X(t-\Delta_t) - \mu_X(t-\Delta_t)\} \quad (3)$$

which relates the response process to time varying lags $t - \Delta_t$, $0 < \Delta_t < t$, of the predictor process. Lagged associations are explored in (3) via estimation of the time dependent lag $t - \Delta_t$ by maximizing the absolute value of a normalized covariance criterion between lags of the predictor process and the response from time t. For homoskedastic predictor processes, the lag search corresponds to finding the lag of the predictor with the highest absolute correlation with the response. This also corresponds to choosing the path in the two-dimensional transfer function $\beta(t, s)$ with the highest absolute value as will be shown in Section 2.1. The classical VCM is a special case of (3) with concurrent relations, i.e. $\Delta_t = 0$. If a nonzero lagged relation is determined from the EVarlag model, this is informative for further investigation of the nature of the lag detected, including whether it is from a specific slice in time or a lagged time interval. Thus, follow-up analysis, such as functional linear models can be used to model the effects of longitudinal predictors from lagged intervals of time on the response (Senturk and Mueller, 2010; Malfait and Ramsay, 2003; Mueller and Zhang, 2005).

The proposed estimation algorithm is designed for irregular and infrequent data and does not require a common grid, similar to estimation procedure for the transfer functions. Also, it can accommodate the response and predictor processes that may not be measured at the same frequency, hence at concurrent times, as encountered in the clinic visits data that will be analyzed in Section 3. Unlike the transfer function approach which is only feasible for a single longitudinal predictor process, or the lagged varying coefficient models which need to select multiple lags for each predictor, the proposed EVarlag model can be feasibly generalized to multiple longitudinal and cross-sectional predictors, as detailed in Section 2.1, where we describe a practical sequential modeling procedure.

For the remainder of the paper, we present the more general EVarlag model with multiple predictors and propose an estimation algorithm in Section 2. Analysis of the clinic visits data and simulations are given in sections 3 and 4, followed by concluding remarks in Section 5.

2. Conditional Model Formulation and Estimation

The model components, namely the time lags and the varying coefficient functions, are estimated based on the moments of the observed predictor and response processes. For ease of exposition, we first consider the EVarlag model with a single longitudinal predictor in detail and then extend it to multiple predictors, including cross-sectional ones.

2.1. Model with a single longitudinal predictor

The EVarlag model introduced earlier with a single longitudinal predictor is

$$E\{Y(t)-\mu_Y(t)\}=\beta_1(t)\{X(t-\Delta_t)-\mu_X(t-\Delta_t)\}. \quad (4)$$

For the proposed lag search algorithm, we assume that var{$X(t)$} > cov{$X(t)$, $X(s)$} for all $t \neq s$, which we will refer to as the decaying covariance assumption. This assumption accommodates typical covariance structures in longitudinal data where measurements further apart in time are less correlated. This includes common longitudinal correlation structures such as autoregressive and exponential-type structures (e.g., see Fitzmaurice et al. (2004), chap. 7). The lag $t - \Delta_t$ is chosen by maximizing the absolute value of the estimator of the two dimensional transfer function $\beta(t, s)$ of Mueller and Yang (2010) which is equal to the covariance between the response $Y(t)$ and the predictor $X(s)$ processes, normalized by the variance of $X(s)$:

$$\frac{\text{cov}\{Y(t),X(s)\}}{\text{var}\{X(s)\}} \quad (5)$$

with respect to $s \leq t$. In applications with homoskedastic predictor processes, this can be interpreted as selecting the lag from the predictor's past trajectory with the highest absolute correlation with the response. Under the EVarlag model (4) with decaying covariance assumption, the quantity in (5) is equal to $\beta_1(t)$cov{$X(s)$, $X(t - \Delta_t)$}/var{$X(s)$}, and the maximizer of its absolute value exists and is unique at lag $s = t - \Delta_t$ with cov{$X(s)$, $X(t - \Delta_t)$}/var{$X(s)$} = 1. Hence the proposed search criterion targets the correct lag under the proposed model.

The goal of the proposed EVarlag model is to embed the classical concurrent VCM in a larger class of models where the proposed exploratory search algorithm is valuable in detecting deviations from a concurrent varying coefficient model, i.e. concurrent relations between the response and the predictors, which is a special case of EVarlag models with $\Delta_t = 0$. Note that independent of the underlying EVarlag model, the lag search algorithm proposed is simply selecting the lag from the predictor's past trajectory with the highest absolute correlation (or conditional correlation in higher dimensions) with the response for homoskedastic predictor processes. We envision the proposed model and estimation algorithms as exploratory tools and acknowledge that more complex models may be sought describing lags from time intervals via functional linear models after detection of potential lagged relations via the use of the EVarlag procedures.

Once the lag Δ_t is estimated, the varying coefficient function $\beta_1(t)$ is estimated based on the following covariance equality

$$\beta_1(t)=\frac{\text{cov}\{Y(t),X(t-\Delta_t)\}}{\text{var}\{X(t-\Delta_t)\}}. \quad (6)$$

Estimation of the moments of the predictor and response processes, leading to an estimator of $\beta_1(t)$ via (6), will be described in Section 2.2.

2.2. Model generalization and estimation

The more general EVarlag model with p longitudinal ($X_1, \ldots, X_p$) and q cross-sectional predictors ($Z_1, \ldots, Z_q$) is given as

$$E\{Y(t)-\mu_Y(t)\}=\sum_{r=1}^{p}\beta_r(t)\{X_r(t-\Delta_{rt})-\mu_{X_r}(t-\Delta_{rt})\}+\sum_{q=1}^{g}\alpha_q(t)(Z_q-\mu_{Z_q}).$$

In order to estimate the time varying lags without increasing the dimension of the estimation procedure, we consider a sequential approach, where the predictors are added into the model one at a time. Appropriate lags are chosen without an increase in the dimension by maximizing a normalized covariance objective criterion that is conditional on the predictors that are already in the model. For models that include cross-sectional predictors, the maximizations for the lags are conditional on the cross-sectional predictors as well. Thus, under this framework, exploration of the lagged association and the (conditional) time varying effects, through the estimated varying coefficient functions, are feasible in higher dimensions. For estimation in higher dimensions, we still assume decaying covariance for the longitudinal predictors, and the existence and uniqueness of the estimated lags in targeting the underlying ones follow similarly to arguments given in Section 2.1 in the univariate case for uncorrelated predictors, where the conditional maximization criteria given in (10) below reduce to their unconditional counterparts. The properties of the proposed algorithm under small to moderate correlations between predictors are studied in the Monte Carlo simulations of Section 4. Note that the algorithm is not guaranteed to target the underlying lags under high multicollinearity between predictors.

The cross-sectional predictors Z_{qi} observed for $i = 1, \ldots, n$ subjects are assumed to have finite second moments. The underlying longitudinal variables X_{ri} and Y_i for $r = 1, \ldots, p$ and $i = 1, \ldots, n$ are square integrable random realizations of smooth random processes X_r and Y, all defined on finite and closed domains. Random processes X_r and Y have smooth mean functions $\mu_{X_r}(t) = EX_r(t)$ and $\mu_Y(t) = EY(t)$ and auto-covariance functions $\mathrm{cov}\{X_r(s), X_r(t)\}$ and $\mathrm{cov}\{Y(s), Y(t)\}$, respectively. The observed longitudinal data are noise contaminated versions of the random processes X_{ri} and Y_i observed at possibly different time points

$$Y_{ij}=Y(t_{ij})+\varepsilon_{ij}, j=1,\ldots,n_i\text{and}X_{rij}=X_r(t_{rij})+\varepsilon_{rij}, j=1,\ldots,m_{ri}, \quad (7)$$

where ε_{ij} and ε_{rij} denote the mean zero finite variance i.i.d additive measurement errors. Note that in formulation (7), longitudinal variables need not be measured at concurrent times and have subject-specific total number of measurements n_i and m_{ri}. Hence, some variables may be measured more frequently than others. For example for the data analyzed in Section 3, total number of clinic visits were measured every three months while the impairment scores were observed yearly. The proposed estimation procedure below utilizes every observation on the longitudinal variables and does not require concurrent measurements. In addition, the proposed estimation procedure, more specifically the estimators of the moments of the random processes proposed in Step 1 below, are consistent even under sparse designs with irregular and infrequent subject specific observation times. Here irregularity means that the longitudinal measurements on each subject do not need to be taken on a common grid and infrequency means that the total repeated measurements on subjects can get low (as low as a single measurement per subject). The moments estimators from Step 1 below, even under these extreme conditions, can be shown to be consistent.

Because the proof follows our previous works (Senturk and Nguyen, 2011) closely, it will not be repeated here.

The estimation procedure involves three main steps, which begins with estimation of the moments of the underlying random processes. In the second step, after estimation of the needed moments, the proposed estimation algorithm chooses the lags of every longitudinal predictor as they enter the model sequentially based on minimization of an expected squared error (ESE) criterion. Once all lags are chosen, the final varying coefficient functions are estimated in the last/third step. We formulate the proposed estimation procedure through the following steps.

Step 1 – Estimation of the moments of the response/predictor processes. We first estimate the mean functions for the longitudinal predictors and response via local polynomial smoothing of the aggregated data $\{(t_{rij}, X_{rij}), i = 1, \ldots, n, j = 1, \ldots, m_{ri}\}$ and $\{(t_{ij}, Y_{ij}), i = 1, \ldots, n, j = 1, \ldots, n_i\}$ yielding $\hat{\mu}_{X_r}(t_{rij})$ and $\hat{\mu}_Y(t_{ij})$, respectively. Raw cross- and auto-covariance estimators obtained from observations on the same subject for $i = 1, \ldots, n$ are given as $G_{Z_{q1}Z_{q2},i} = \{Z_{q1,i} - \overline{Z_{q1}}\}\{Z_{q2,i} - \overline{Z_{q2}}\}$,

$$\begin{aligned} &G_{YZ_q,i}(t_{ij}) = \{Y_{ij} - \widehat{\mu}_Y(t_{ij})\}\{Z_{qi} - \overline{Z}_q\}, j=1,\ldots,n_i, \\ &\quad G_{X_rZ_q,i}(t_{rij}) = \{X_{rij} - \widehat{\mu}_{X_r}(t_{rij})\}\{Z_{qi} - \overline{Z}_q\}, j=1,\ldots,m_{ri}, \\ &\quad G_{YX_r,i}(t_{ij}, t_{rik}) = \{Y_{ij} - \widehat{\mu}_Y(t_{ij})\}\{X_{rik} - \widehat{\mu}_{X_r}(t_{rik})\}, j=1,\ldots,n_i, k=1,\ldots,m_{ri}, \\ &\quad G_{X_{r_1}X_{r_2,i}}(t_{r_1ij}, t_{r_2ik}) = \{X_{r_1ij} - \widehat{\mu}_{X_{r_1}}(t_{r_1ij})\}\{X_{r_2ik} - \widehat{\mu}_{X_{r_2}}(t_{r_2ik})\} \end{aligned}$$

for $j = 1, \ldots, m_{r_1i}$; $k = 1, \ldots, m_{r_2i}$; r, r_1, and $r_2 = 1, \ldots p$; q, q_1, and $q_2 = 1 \ldots, g$ and $\overline{Z_q} = (\sum_{i=1} Z_{qi})/n$. Next, bivariate smoothing of the raw covariances $G_{Y X_{r,i},}(t_{ij}, t_{rik})$ and $G_{X_{r1}X_{r2},i}(t_{r1ij}, t_{r2ik})$ lead to the final smooth covariance surface estimates $\widehat{\text{cov}}\{Y(t), X_r(s)\}$ and $\widehat{\text{cov}}\{X_{r_1}(t), X_{r_2}(s)\}$, respectively. To guarantee positive definiteness of the final auto-covariance estimator, we carry out a functional principle component analysis step on the smooth auto-covariance estimator and exclude the negative estimates of the eigenvalues and the corresponding eigenfunctions. The final auto-covariance estimator is constructed with only a truncated number of positive eigenvalue estimates and their corresponding eigenfunctions. We refer the readers to Senturk and Nguyen (2011) for further details. A one dimensional local polynomial smoothing of the raw estimates $G_{Y Z_q,i}(t_{ij})$ and $G_{X_rZ_q,i}(t_{rij})$ lead to the final smooth estimates $\widehat{\text{cov}}\{Y(t), Z_q\}$ and $\widehat{\text{cov}}\{X_r(t), Z_q\}$, respectively. The variance estimator $\widehat{\text{cov}}(Z_{q_1}, Z_{q_2})$ is equal to $(\sum_{i=1}^{n} G_{Z_{q_1}Z_{q_2},i})/n$. In the above smoothing steps we utilize generalized cross-validation (Liu and Mueller, 2008) in choosing the appropriate bandwidths.

Step 2 – Sequential estimation of lags. Lags are estimated by fitting sequentially a series of EVarlag models where a longitudinal predictor (with the corresponding lag) or a cross-sectional predictor is chosen sequentially based on the smallest estimated expected squared error (ESE); details of moments-based estimates of ESE are provided in the Appendix. We provide in more detail below the procedure for estimating the lags corresponding to each potentially chosen longitudinal predictor in steps 2.1, 2.2. and 2.($p + g$).

Step 2.1 – Selection of the first predictor and its potential lag. Fit $p + g$ univariate models for each cross-sectional and longitudinal predictors with corresponding lags for longitudinal predictors. The first selected predictor is the one with the smallest ESE estimate.

a. *Selecting lags for univariate models with longitudinal predictors.* For selecting the lag for the longitudinal predictor $X_r^{(1)}(\cdot)$, *r* = 1, …, *p*, we maximize the absolute value of

$$\frac{\widehat{\text{cov}}\{Y(t), X_r^{(1)}(s)\}}{\widehat{\text{var}}\{X_r^{(1)}(s)\}} \quad (8)$$

with respect to *s* ≤ *t* as described in Section 2.1. Denote the estimated lag by $\widehat{\Delta}_{rt}^{(1)}$. In (8) and throughout the description of the algorithm, the superscript (*k*), *k* = 1, …, *p* + *g*, will denote quantities from step 2.*k* of the proposed algorithm.

b. *Estimating varying coefficient functions from univariate models.* The estimator of the varying coefficient function for the univariate model with longitudinal predictor $X_r^{(1)}(t-\widehat{\Delta}_{rt}^{(1)})$ is given as $\widehat{\beta}_r^{(1)}(t){=}\widehat{\text{cov}}\{Y(t),X_r^{(1)}(t-\widehat{\Delta}_{rt}^{(1)})\}/\widehat{\text{var}}\{X_r^{(1)}(t-\widehat{\Delta}_{rt}^{(1)})\}$. Varying coefficient function estimator from the univariate model with cross-sectional predictor $Z_q^{(1)}$, *q* = 1, …, *g* is given by $\widehat{\alpha}_q^{(1)}(t){=}\widehat{\text{cov}}\{Y(t),Z_q^{(1)}\}/\widehat{\text{var}}\{Z_q^{(1)}\}$.

c. *Comparing univariate model fits.* To simplify notations, let $W_\ell^{(1)}(t)$, ℓ = 1, …, *p*+*g*, denote a generic predictor that can be longitudinal or cross-sectional, i.e.,

$$W_\ell^{(1)}(t){=}\begin{cases} X_r^{(1)}(t-\widehat{\Delta}_{rt}^{(1)}), & \text{for longitudinal predictor} \\ Z_q^{(1)}, & \text{for cross-sectional predictor} \end{cases}$$

with mean $\mu_{W_\ell}^{(1)}(t){=}\mu_{X_r}^{(1)}(t-\widehat{\Delta}_{rt}^{(1)})$ or $\mu_{W_\ell}^{(1)}(t){=}\mu_{Z_q}^{(1)}$ corresponding to a longitudinal or cross-sectional predictor, respectively. We compare univariate model fits based on the expected squared error criterion

$$\text{ESE}_\ell^{(1)}(t){=}E[\{Y(t)-\mu_Y(t)\}-\gamma_\ell^{(1)}(t)\{W_\ell^{(1)}(t)-\mu_{W_\ell}^{(1)}(t)\}]^2 \quad (9)$$

for each of the *p* + *g* univariate models, where $\gamma_\ell^{(1)}(t)\equiv\beta_r^{(1)}(t)$ for longitudinal predictors and $\gamma_\ell^{(1)}(t)\equiv\alpha_q^{(1)}(t)$ for cross-sectional ones. The predictor which minimizes $\sum_{v=1}^{T}\widehat{\text{ESE}}_\ell^{(1)}(t_v)$ (for a grid of time points t_1, …, t_T), denoted by $W_1^*(t)$ with corresponding mean function $\mu_{W_1}^*(t)$, is the first selected predictor and its respective lag is denoted by $\widehat{\Delta}_{1t}^*$ if it is longitudinal.

Step 2.2 – Selection of the second predictor and its potential lag. Next, a series of bivariate models are fitted and compared, where each fitted model includes the selected first predictor, $W_1^*(t)$, from step 2.1 above; the second predictor in the model is any one of the *p* + *g* − 1 remaining longitudinal and cross-sectional predictors. For comparing model fits, we begin by selecting lags for the longitudinal predictors entering the model as the second covariate.

a. *Selecting lags for bivariate models with longitudinal predictors.* Let $p^{(2)}$ = *p* or *p* − 1 be the number of longitudinal predictors after step 2.1. To select the lag for the longitudinal predictor $X_r^{(2)}(\cdot)$, $r = 1, \ldots, p^{(2)}$, entering the model, we maximize the absolute value of

$$\begin{aligned} &\frac{\widehat{\mathrm{cov}}\{Y(t), X_r^{(2)}(s)|W_1^*(t)\}}{\widehat{\mathrm{var}}\{X_r^{(2)}(s)|W_1^*(t)\}} \\ &=[\chi_2^{-1}(s)\Xi_2(s)]_2 \\ &=\left\{\begin{bmatrix} \widehat{\mathrm{var}}\{W_1^*(t)\} & \widehat{\mathrm{cov}}\{W_1^*(t), X_r^{(2)}(s)\}, \\ \widehat{\mathrm{cov}}\{W_1^*(t), X_r^{(2)}(s)\} & \widehat{\mathrm{var}}\{X_r^{(2)}(t)\} \end{bmatrix}^{-1} \begin{bmatrix} \widehat{\mathrm{cov}}\{Y(t), W_1^*(t)\} \\ \widehat{\mathrm{cov}}\{Y(t), X_r^{(2)}(s)\} \end{bmatrix}\right\}_2, \end{aligned} \tag{10}$$

with respect to $s \leq t$. In (10), we used $[\nu]_k$ to denote the kth element of the vector ν. Denote the estimated lag by $\widehat{\Delta}_{rt}^{(2)}$.

b. *Estimating varying coefficient functions from bivariate models.* Denote a generic covariate by $W_\ell^{(2)}(t)$, $\ell = 1, \ldots, p + q - 1$, which can be equal to one of the remaining longitudinal or cross-sectional predictors, i.e.

$$W_\ell^{(2)}(t)=\begin{cases} X_r^{(2)}(t-\widehat{\Delta}_{rt}^{(2)}), & \mathrm{for\ \ longitudinal\ \ predictor} \\ Z_q^{(2)}, & \mathrm{for\ \ cross{-}sectional\ \ predictor} \end{cases}$$

with mean $\mu_{W_\ell}^{(2)}(t)=\mu_{X_r}^{(2)}(t-\widehat{\Delta}_{rt}^{(2)})$ or $\mu_{W_\ell}^{(2)}(t)=\mu_{Z_q}^{(2)}$, respectively. The estimators of the varying coefficient functions are given by

$$\begin{bmatrix} \widehat{\gamma}_{\ell 1}^{(2)}(t) \\ \widehat{\gamma}_{\ell 2}^{(2)}(t) \end{bmatrix}=\overline{\chi}_2^{-1}(s)\tilde{\Xi}_2(s)=\begin{bmatrix} \widehat{\mathrm{var}}\,\{W_1^*(t)\} & \widehat{\mathrm{cov}}\{W_1^*(t), W_\ell^{(2)}(t)\} \\ \widehat{\mathrm{cov}}\{W_1^*(t), W_\ell^{(2)}(s)\} & \widehat{\mathrm{var}}\{W_\ell^{(2)}(t)\} \end{bmatrix}^{-1} \begin{bmatrix} \widehat{\mathrm{cov}}\{Y(t), W_1^*(t)\} \\ \widehat{\mathrm{cov}}\{Y(t), W_\ell^{(2)}(t)\} \end{bmatrix}.$$

c. *Comparing bivariate model fits.* Similarly, we compare the series of bivariate model fits by estimating the expected squared error

$$\mathrm{ESE}_\ell^{(2)}(t)=E[\{Y(t)-\mu_Y(t)\}-\gamma_{\ell 1}^{(2)}(t)\{W_1^*(t)-\mu_{W_1}^*(t)\}-\gamma_{\ell 2}^{(2)}(t)\{W_\ell^{(2)}(t)-\mu_{W_\ell}^{(2)}(t)\}]^2$$

for $\ell = 1, \ldots, p + g - 1$. The predictor which minimizes $\sum_{v=1}^{T}\widehat{\mathrm{ESE}}_\ell^{(2)}(t_v)$, denoted by $W_2^*(t)$ with mean function $\mu_{W_2}^*(t)$, is the second selected predictor and its respective lag is denoted by $\widehat{\Delta}_{2t}^*$ if it is longitudinal.

Step 2.k – Selecting the kth predictor and its potential lag. The selected predictors after step 2.(k − 1) are $W_1^*(t), \ldots, W_{k-1}^*(t)$. In this step, we fit and compare the series of models with k total predictors; k − 1 predictors selected in the previous steps and the kth predictor is any one of the $p + g - k + 1$ remaining longitudinal and cross-sectional predictors. As in the previous steps, we begin by selecting lags for the longitudinal predictors entering the model as the kth covariate.

a. *Selecting lags for k dimensional models with longitudinal predictors.* To select the lag for a remaining longitudinal predictor $X_r^{(k)}(\cdot)$, we maximize the absolute value of

$$\frac{\widehat{\mathrm{cov}}\{Y(t),X_r^{(k)}(s)|W_1^*(t),\dots,W_{k-1}^*(t)\}}{\widehat{\mathrm{var}}\{X_r^{(k)}(s)|W_1^*(t),\dots W_{k-1}^*(t)\}}=[\chi_k^{-1}(s)\Xi_k(s)]_k$$

with respect to $s \le t$, where

$\Xi_k(s)=[\widehat{\mathrm{cov}}\{Y(t),W_1^*(t)\},\dots,\widehat{\mathrm{cov}}\{Y(t),W_{k-1}^*(t)\},\widehat{\mathrm{cov}}\{Y(t),X_r^{(k)}(s)\}]^{\mathrm{T}}$,

$$\chi_k(s)=\begin{bmatrix}\chi_{k,11}(s) & \chi_{k,12}(s)\\ \chi_{k,12}^{\mathrm{T}}(s) & \widehat{\mathrm{var}}\{X_r^{(k)}(s)\}\end{bmatrix},\ \chi_{k,11}(s)=\begin{bmatrix}\widehat{\mathrm{var}}\{W_1^*(t)\} & \cdots & \widehat{\mathrm{cov}}\{W_1^*(t),W_{k-1}^*(t)\}\\ \vdots & \ddots & \vdots\\ \widehat{\mathrm{cov}}\{W_1^*(t),W_{k-1}^*(t)\} & \cdots & \widehat{\mathrm{var}}\{W_{k-1}^*(t)\}\end{bmatrix},$$

and $\chi_{k,12}(s)=[\widehat{\mathrm{cov}}\{W_1^*(t),X_r^{(k)}(s)\},\dots,\widehat{\mathrm{cov}}\{W_{k-1}^*(t),X_r^{(k)}(s)\}]^{\mathrm{T}}$. Denote this lag by $\widehat{\Delta}_{rt}^{(k)}$.

b. *Estimating varying coefficient functions from models with k predictors.* Let $W_\ell^{(k)}(t)$, $\ell = 1, \dots, p + q - k + 1$, be a generic covariate, which can be equal to one of the remaining longitudinal or cross-sectional predictors, i.e.,

$$W_\ell^{(k)}(t)=\begin{cases}X_r^{(k)}(t-\widehat{\Delta}_{rt}^{(k)}), & \mathrm{for\ \ longitudinal\ \ predictor}\\ Z_q^{(k)}, & \mathrm{for\ \ cross{-}sectional\ \ predictor}\end{cases}$$

with mean $\mu_{W_\ell}^{(k)}(t)=\mu_{X_r}^{(k)}(t-\widehat{\Delta}_{rt}^{(k)})$ or $\mu_{W_\ell}^{(k)}(t)=\mu_{Z_q}^{(k)}$, respectively. The estimators of the varying coefficient functions are given by

$$\begin{bmatrix}\widehat{\gamma}_{\ell 1}^{(k)}(t)\\ \cdots\\ \widehat{\gamma}_{\ell k}^{(k)}(t)\end{bmatrix}=\tilde{\chi}_k^{-1}(s)\tilde{\Xi}_k(s),$$

where $\tilde{\Xi}_k(s)=[\widehat{\mathrm{cov}}\{Y(t),W_1^*(t)\},\dots,\widehat{\mathrm{cov}}\{Y(t),W_{k-1}^*(t)\},\widehat{\mathrm{cov}}\{Y(t),W_\ell^{(k)}(t)\}]^{\mathrm{T}}$ and

$$\tilde{\chi}_k(s)=\begin{bmatrix}\chi_{k,11}(s) & \tilde{\chi}_{k,12}(s)\\ \tilde{\chi}_{k,12}^{\mathrm{T}}(s) & \widehat{\mathrm{var}}\{W_\ell^{(k)}(t)\}\end{bmatrix}$$

with $\tilde{\chi}_{k,12}(s)=[\widehat{\mathrm{cov}}\{W_1^*(t),W_\ell^{(k)}(t)\},\dots,\widehat{\mathrm{cov}}\{W_{k-1}^*(t),W_\ell^{(k)}(t)\}]^{\mathrm{T}}$.

c. *Comparing k dimensional model fits.* We compare the fitted models by

$$\mathrm{ESE}_\ell^{(k)}(t)=E\left[\{Y(t)-\mu_Y(t)\}-\sum_{u=1}^{k-1}\gamma_{\ell u}^{(k)}(t)\{W_u^*(t)-\mu_{W_u}^*(t)\}-\gamma_{\ell k}^{(k)}(t)\{W_\ell^{(k)}(t)-\mu_{W_\ell}^{(k)}(t)\}\right]^2$$

for $\ell = 1, \ldots, p+g-k+1$. The predictor which minimizes $\sum_{v=1}^{T} \widehat{\mathrm{ESE}}_{\ell}^{(k)}(t_v)$, denoted by $W_k^*(t)$, is the selected kth predictor and its respective lag is denoted by $\widehat{\Delta}_{kt}^*$ if it is longitudinal.

Step 3 – Estimation of the varying coefficient functions. Once all the lags of the longitudinal variables are chosen from steps 2.1 to 2.($p + g$), we estimate the final varying coefficient functions for the model with all $p+q$ covariates with the p selected lags. Without loss of generality, let $\{X_1(t - \hat{\Delta}_{1t}), \ldots, X_p(t - \hat{\Delta}_{pt})\}$ and $(Z_1, \ldots, Z_g)$ be the reordered longitudinal and cross-sectional predictors among $\{W_1^*(t), \ldots, W_{p+g}^*(t)\}$ with selected lags $(\widehat{\Delta}_{1t}^*, \ldots, \widehat{\Delta}_{pt}^*)$.

The final varying coefficient function estimators are given as

$$[\widehat{\beta}_1(t), \ldots, \widehat{\beta}_p(t), \widehat{\alpha}_1(t), \ldots, \widehat{\alpha}_g(t)]^{\mathrm{T}} = \chi^{-1}\Xi,$$

where

$$\Xi = [\widehat{\mathrm{cov}}\{Y(t), X_1(t-\widehat{\Delta}_{1t})\}, \ldots, \widehat{\mathrm{cov}}\{Y(t), X_p(t-\widehat{\Delta}_{pt})\}, \widehat{\mathrm{cov}}\{Y(t), Z_1\}, \ldots, \widehat{\mathrm{cov}}\{Y(t), Z_g\}]^{\mathrm{T}}$$

,

$$\chi = \begin{bmatrix} \chi_{11} & \chi_{12} \\ \chi_{12}^{\mathrm{T}} & \chi_{22} \end{bmatrix}, \chi_{11} = \begin{bmatrix} \widehat{\mathrm{var}}_1 & \cdots & \widehat{\mathrm{cov}}_{1,p} \\ \vdots & \ddots & \vdots \\ \widehat{\mathrm{cov}}_{p,1} & \cdots & \widehat{\mathrm{var}}_p \end{bmatrix},$$

$$\chi_{12} = \begin{bmatrix} \widehat{\mathrm{cov}}\{X_1(t-\widehat{\Delta}_{1t}), Z_1\} & \cdots & \widehat{\mathrm{cov}}\{X_1(t-\widehat{\Delta}_{1t}), Z_g\} \\ \vdots & \ddots & \vdots \\ \widehat{\mathrm{cov}}\{X_p(t-\widehat{\Delta}_{pt}), Z_1\} & \cdots & \widehat{\mathrm{cov}}\{X_p(t-\widehat{\Delta}_{pt}), Z_g\} \end{bmatrix},$$

$$\widehat{\mathrm{cov}}_{r_1,r_2} = \widehat{\mathrm{cov}}\{X_{r_1}(t-\widehat{\Delta}_{r_1t}), X_{r_2}(t-\widehat{\Delta}_{r_2t})\}, \widehat{\mathrm{var}}_{r_1} = \widehat{\mathrm{var}}\{X_{r_1}(t-\widehat{\Delta}_{r_1t})\} \mathrm{and} \chi_{22} = \widehat{\mathrm{cov}}(Z).$$

Remarks: *Properties of the Proposed Estimation Algorithm and Implementation.* For the lag choices in step 2, a sequence of regression models is considered that is increasing in the number of predictors. The proposed exploratory conditional regression model starts with one predictor (W_1^*) in step 2.1 and the final model consists of all predictors $(W_1^*, \ldots, W_{p+g}^*)$ in step 2.($p + g$). The sequential approach allows for exploration of complex lags associated with multiple predictors via optimizations using conditional models to avoid the curse of dimensionality.

Note that while the proposed step-wise algorithm may be similar in spirit to step-wise variable selection procedures, it is not proposed for variable selection in this manuscript. It is only used for dimension reduction in the proposed lag exploration process. We illustrate in Section 4 that when the correlation between the predictors is low to moderate, the lags chosen for the longitudinal variables in the conditional models will be close to the lag choices in the underlying full model and the consistency of the varying coefficient function estimators will follow from consistency of the moments estimators proposed in step 1. Consistency of the proposed moments estimators has been established in Senturk and Mueller (2010) and Senturk and Nguyen (2011).

Estimation of the moments of the predictor and response processes in step 1, including the bivariate smoothing and the choice of bandwidths, are carried out with the software package

PACE (http://anson.ucdavis.edu/~ntyang/PACE; Yao et al., 2003; Yao et al., 2005). Computations in steps 2 and 3 involve straight forward matrix evaluations.

3. Data Analysis: Exploration of Lagged Effects Associated with Longitudinal Cognitive and Functional Impairment

The motivating data for the proposed methodology is from an observational cohort study conducted over 4 years on 745 older adults where multiple health scores and the number of clinic visits are recorded. Various subsets of the data have been analyzed to address different clinical hypothesis and more recently Pickett et al. (2011) performed a cross-sectional analysis of only the baseline data to identify factors associated with healthcare utilization. A detailed description of the data collection mechanism and related issues are provided in Seaburn et al. (2005) and Grabovich et al. (2010). In order to model the frequency of the clinic visits as a function of clinical factors observed over time, we analyze here a subset of the data with 703 older adults between ages of 65 and 93 with yearly recorded cognitive (from Mini-Mental State Exam [MMSE]) and physical Karnofsky (KPS) impairment scores > 12 and > 40, respectively. Because our outcome variable of interest is clinic visits, the dataset analyzed excludes individuals who were severely disabled (indicating hospitalization or more severe, e.g., hospitalization required) or with severe cognitive impairment. The response, total number of clinic visits, is originally intended to be recorded every three months but due to frequent missing values even in the baseline measurements, the data is highly irregular and unsynchronized. MMSE and KPS impairment scores and the logarithm of the total clinic visits collected throughout the observation period of four years are displayed in Figure 1. (We add 0.5 to the zero entries of the total number of visits before taking logarithms.) MMSE and KPS scores measure cognitive and physical decline, respectively, as the scores decrease. Cross-sectionally estimated mean trajectories over time are plotted in dark solid, where the estimated mean number of visits has an increasing and the estimated mean physical impairment score has a decreasing trend over the observation period, as expected. On average, cognitive functioning score remains constant over the progression of the study.

Our analysis to explore lagged relations discovers opposing trends in MMSE and KPS scores as they relate to total clinic visits. MMSE scores from baseline and KPS scores from concurrent times have the most pronounced effects on clinic visits throughout the study. Since both predictor processes are close to homoskedasticity in the study period, selected lags can be interpreted as the lags from the predictors' past that has the highest absolute correlation with the response, clinic visits. The results uncover deviations from concurrent modeling of MMSE which would be informative for further investigation of the nature of the lag detected, potentially studying the predictive power of the entire historical trajectory of MMSE in explaining clinic visits. In addition the direction of the regression relations estimated imply that patients with better physical functioning tend to have fewer clinic visits, while patients with better cognitive status are more likely to attend their clinic visits. In the analysis outlined below we include estimated univariate transfer function regressions, univariate EVarlag models for the two predictors separately, and fits from the bivariate EVarlag model with both predictors in the model.

We first explore the time varying lagged regression of the logarithm of the number of clinic visits on the two impairment scores separately. The estimated two dimensional transfer functions $\beta(t, s)$ from separate regressions of the logarithm of the number of visits at time t on MMSE and KPS score from lagged time $s < t$ are given in Figure 2. The plotted transfer functions have a triangular support, where the diagonal $s = t$ represents the regression relation of the response and the predictor from concurrent times. The lag path that maximizes the absolute value of the transfer function is given in thick solid black in Figure

2. Figures 3 (a)–(b) display the estimated varying coefficient functions (solid) from the EVarlag model of the log visits on MMSE and KPS scores, separately. Paths of maximum absolute value of the transfer function highlighted (thick black) in Figure 2 correspond to the chosen time dependent lags plotted in Figure 3 (d) and (e) (thick solid). Also plotted in Figure 3 are the concurrent varying coefficient model fits (dash-dotted) relating the response to the two predictors separately from concurrent times ignoring lag estimation. These fits correspond to the diagonal values of the transfer functions plotted in Figure 2. ±2 bootstrap error bars for both varying coefficient function fits, from the proposed EVarlag model and the concurrent varying coefficient model are plotted in Figure 3 (dotted). Reported bootstrap error bars are based on 200 bootstrap samples drawn via resampling from subjects, where EVarlag model fits are obtained from the bootstrap data at the particular lags chosen for the original data. In order to study the variation in the lag choice, we also plot (thick dashed) in Figures 3 (d) and (e) the median bootstrap lag choice for MMSE and KPS from the EVarlag fits to bootstrap samples where a different time varying lag is chosen each time.

While the lagged and concurrent model fits agree perfectly (Figure 3 (b)) for the regression of log visits on KPS scores with a chosen lag of $t - \Delta_t = t$, i.e. $\Delta_t = 0$, the EVarlag fit shows a more pronounced positive effect (Figure 3 (a)) of the baseline MMSE score throughout the study. More specifically, the estimated transfer function in Figure 2 (b) and the varying coefficient function in Figure 3 (b) suggest that there is a negative regression relation between number of clinic visits and KPS, where the association gets more pronounced with the progression of the study and less pronounced for KPS from further lagged times. The estimated median bootstrap lag plotted in Figure 3 (d) suggests that the concurrent relation is selected consistently for KPS, which corresponds to a lag choice of $t - \Delta_t = t$, i.e. $\Delta_t = 0$. On the other hand, the lagged and concurrent models suggest different fits for MMSE. While the regression fit from concurrent times suggest a decline in the strength of the positive regression relation between MMSE and number of clinic visits with the progression of the study, especially after 1.5 years, the fit from baseline MMSE stays leved without a decline in the strength of the suggested relation (Figure 3 (a)). This difference is reflected in the ±2 bootstrap error bars. The estimated median bootstrap lag also suggests that the selection of the baseline MMSE is consistent over the bootstrap draws (Figure 3 (e)). Hence, the modeling results suggest that while the decline of the (physical) impairment score has a concurrent effect on the frequency of clinic visits, it is cognitive impairment score from lagged times that have more pronounced effects on clinic visits (compared to those from the current time).

Using the sequential estimation algorithm for fitting the EVarlag model with multiple predictors, proposed in Section 2.2, we next regress the logarithm of the total number of visits on both impairment scores, KPS and MMSE. The physical impairment (KPS score) enters the model first, followed by MMSE. The estimated lags (thick solid) and varying coefficient functions (solid) are given in Figure 4. Also presented are ±2 bootstrap error bars (dotted) and median bootstrap lag choices (thick dotted). The results suggest baseline MMSE has a more pronounced effect in the second half of the study in contrast with the declining positive effect of MMSE from concurrent times towards the end of the study. The varying coefficient function for MMSE from the concurrent fit falls outside of the bootstrap error bars towards the end of the study (Figure 4 (b)). We also explored the EVarlag model with the reverse order of predictors and the results are fairly similar and thus omitted here. Also, the chosen lags and overall conclusions from the multiple EVarlag model above are similar to the models with each predictor examined separately. The consistent results across models would be expected for this data, since the correlation between the longitudinal KPS and MMSE measurements range between 0.04 and 0.09 (with a median of 0.05), implying that the impairment scores are close to uncorrelated. This is expected since the two instruments are designed to capture two fairly distinct aspects of overall health, namely cognitive and physical domains.

4. Simulation Studies

We examine the performance of the proposed estimation algorithm in two simulation set-ups with different lag and correlation structures among the predictor processes. The first set-up mimics the clinic visits data using similar predictor and response observation frequency with two longitudinal predictors. The underlying lag structure involves baseline and concurrent lags as observed in the real data analyzed in Section 3. In the second simulation set-up, we study the proposed EVarlag model more generally with four longitudinal and one cross-sectional predictor. We also study the performance of the varying coefficient function and lag estimates under increasing correlation between predictor processes and compare the proposed varying coefficient function estimates with the (a) benchmark (optimal) estimates that uses the true lags and (b) fits from the concurrent VCM that completely ignores lag effects. In estimation of the moments of the predictor and response processes, we utilize the publicly available software package PACE (http://anson.ucdavis.edu/~ntyang/PACE; Yao et al., 2003; Yao et al., 2005) where generalized cross-validation is used for choosing corresponding bandwidths.

4.1 Modeling clinic visits data: Two longitudinal predictor processes

We first consider an EVarlag model with two longitudinal predictors, similar to the clinic visits data analyzed above. The number of measurements for the predictors was randomly chosen with equal probability from {3, 4, 5} and the response from {1, 2, 3, …, 12} for each of the $n = 700$ subjects, similar to the clinic visits data. Locations for the predictors and response measurements for the subject i were generated uniformly from [0, 1] separately. This resulted in concurrent measurements among the predictors which were not concurrent with the response measurements. Since the KPS and MMSE measures are nearly uncorrelated with correlations between their longitudinal measurements ranging between 0.04 and 0.09, we generated the two longitudinal variables independently. The predictor vectors $X_{ri} = [X_{ri1}, \ldots, X_{rim_{ri}}]^{T}$, $r = 1, 2$, $m_{1i} = m_{2i} = m_i$, were generated from multivariate normal distributions with mean vectors $\mu_{X_1}(t_i) = 3 - 2t_i$ and $\mu_{X_2}(t_i) = (1 + t_i)^2$ for $t_i = [t_{i1}, \ldots, t_{im_i}]^{T}$. The m_i by m_i covariance matrices for the longitudinal predictors had (j, j')th elements equal to $\sigma_{i1}(t_{ij}, t_{ij'}) = 6e^{-3|t_{ij}-t_{ij'}|}$ and $\sigma_{i2}(t_{ij}, t_{ij'}) = 4e^{-4|t_{ij}-t_{ij'}|}$ for X_{1i} and X_{2i}, respectively. The response trajectories is

$$Y_i(t_{ij})-\mu_Y(t_{ij})=\beta_1(t_{ij})\{X_1(0)-\mu_{X_1}(0)\}+\beta_2(t_{ij})\{X_2(t_{ij})-\mu_{X_2}(t_{ij})\}+V_i(t_{ij}), \quad (11)$$

where $\beta_1(t) = \cos(\pi t)$, $\beta_2(t) = \sin(2\pi t)$ and $\mu_Y(t) = 2\sin(\pi + \pi t) + \beta_1(t)\mu_{X_1}(0) + \beta_2(t)\mu_{X_2}(t)$.

In model (11), the time varying lags are $t - \Delta_{1t} = 0$, i.e. $\Delta_{1t} = t$ and $t - \Delta_{2t} = t$, i.e. $\Delta_{2t} = 0$, where the baseline measurements of one longitudinal predictor and the measurements from concurrent times with the response of the other longitudinal predictor are included in the regression model; thus, the lag structure is similar to the clinic visits data. The functional error V_i in (11) was equal to $\xi_{V1i}\psi_1(t) + \xi_{V2i}\psi_2(t)$ constructed from the two basis functions $\psi_1(t)=\sqrt{2}\cos(\pi t)$ and $\psi_2(t)=\sqrt{2}\sin(\pi t)$ and subject specific random components ξ_{V1i} and ξ_{V2i} generated independently from zero mean normal distributions with variances 1 and 0.5, respectively. The observed longitudinal predictor and response measurements had time independent mean zero additive measurement errors ε_{rij}, $r = 1, 2$ and ε_{ij} according to (7) where the errors were generated from normal distributions with variances 0.5.

The results from the first simulation study are summarized in Figure 5, which displays the median (dash-dotted) and the 5th and 95th percentiles (dotted) of the varying coefficient function estimates (in Figure 5 (a) and 5 (b)) along with medians of the estimated lags (thick dotted lines in Figures 5 (c) and 5 (d)) for each longitudinal predictor over 200 Monte Carlo runs. The median of the estimated lags and varying coefficient functions trace the true underlying quantities. We evaluate the performance of the estimators via normalized mean integrated squared error for the varying coefficient functions (ME) and for the time varying lags (ML),

$$\mathrm{ME}=\frac{1}{p+g}\left[\sum_{r=1}^{p}\frac{\int\{\beta_r(t)-\widehat{\beta}_r(t)\}^2dt}{\int\beta_r^2(t)dt}+\sum_{q=1}^{g}\frac{\int\{\alpha_q(t)-\widehat{\alpha}_q(t)\}^2dt}{\int\alpha_q^2(t)dt}\right]\quad \mathrm{ML}=\frac{1}{p}\sum_{r=1}^{p}\frac{\int\{\Delta_{rt}-\widehat{\Delta}_{rt}\}^2dt}{1/3},$$

where the maximum squared error for the estimated time varying lags is $\int_0^1 t^2dt=1/3$, $p = 2$ and $g = 0$. The estimated median, 1st quartile and 3rd quartile values from the 200 Monte Carlo runs are (.122, .069, .362) and (.066, .034, .111) for ME and ML, respectively. These values confirm that the proposed estimates are close to the true underlying quantities in this simulation model similar to the clinic visits data, as illustrated in Figure 5.

4.2 Performance of EVarlag under more general simulation settings

In the second simulation set-up, we consider an EVarlag model with four longitudinal and one cross-sectional predictor. The number of repeated measurements for the response and the four longitudinal predictors were chosen with equal probability from 7, 8, 9, 10, for n = 700, 1000 subjects, at concurrent time points ($m_{1i} = m_{2i} = m_{3i} = m_{4i} = n_i$). The locations of the measurements were chosen similarly as in the first simulation study. The combined predictor vector $X_i = [X_{1i1}, \ldots, X_{1in_i}, X_{2i1}, \ldots, X_{2in_i}, X_{3i1}, \ldots, X_{3in_i}, X_{4i1}, \ldots, X_{4in_i}, Z_i]^T$ was generated from a multivariate normal distribution with mean vector $[\mu_{X1}(t_i), \mu_{X2}(t_i), \mu_{X3}(t_i), \mu_{X4}(t_i), E(Z_i)] = [3 - 2t_i, 2 + (t_i - 0.5)^2, t_i, 1 - t_i, 2]$ for $t_i = [t_{i1}, \ldots, t_{in_i}]^T$ and covariance matrix $\Sigma_i=[\Sigma_{i11},\Sigma_{i12};\Sigma_{i12}^T,\mathrm{var}(Z_i)]$, where

$$\Sigma_{i11}=\left[\begin{array}{cccc}\sigma_i & A & A & A\\ A & \sigma_i & A & A\\ A & A & \sigma_i & A\\ A & A & A & \sigma_i\end{array}\right].$$

Here A is a n_i by n_i matrix of contants equal to a, Σ_{i12} is a $4n_i$ by 1 matrix of contants equal to a, σ_i is a n_i by n_i matrix with (j, j')th element equal to $\sigma_i(t_{ij}, t_{ij'}) = 6\exp(-3|t_{ij} - t_{ij'}|)$ and $\mathrm{var}(Z_i) = 6$. We ran simulations with different correlations between the predictor variables by varying the values of a; $a = (0, 0.5, 1, 2)$ corresponding to correlations of (0, 1/12, 1/6, 1/3), respectively. The response were generated from the EVarlag model

$$Y_i(t_{ij}) - \mu_Y(t_{ij}) = \beta_1(t_{ij})\{X_1(t_{ij}) - \mu_{X_1}(t_{ij})\} + \beta_2(t_{ij})\{X_2(0) - \mu_{X_2}(0)\} + \beta_3(t_{ij})\{X_3(t_{ij} - \Delta_{3,t_{ij}}) - \mu_{X_3}(t_{ij} - \Delta_{3,t_{ij}})\} + \beta_4(t_{ij})\{X_4(0) - \mu_{X_4}(0)\} + \alpha(t_{ij})(Z_i - \mu_Z) + V_i(t_{ij}), \qquad (12)$$

where $\beta_1(t) = 3\{\sin(\pi t/2) + \cos(\pi t)\}$, $\beta_2(t) = 2\sin(\pi t)$, $\beta_3(t) = 2\{\sin(\pi t) + \cos(\pi t)\}$, $\beta_4(t) = 3\cos(\pi t)$, $\alpha(t) = 2\sin(0.5\pi t)$ and $\mu_Y(t) = 2\cos(\pi+\pi t)+\beta_1(t)\mu_{X_1}(t)+\beta_2(t)\mu_{X_2}(0)+\beta_3(t)\mu_{X_3}(t_{ij} - \Delta_{3,t_{ij}})+\beta_4(t)\mu_{X_2}(0)+\alpha(t)\mu_Z$. In model (12), the time varying lags are $t - \Delta_{1t} = t$, i.e. $\Delta_{1t} = 0$, $t - \Delta_{2t} = 0$, i.e. $\Delta_{2t} = t$, $t - \Delta_{3t} = (t - 17/29)1_{\{t-17/29\}\geq 0}$ and $t - \Delta_{4t} = 0$, i.e. $\Delta_{4t} = t$. Hence, baseline values of two predictors, concurrent values of one and time varying lagged values of another are included. The functional error V_i in (12) and the additive measurement error on the longitudinal predictor and response trajectories were generated as described above.

We examine the performance of the proposed estimation algorithm under mild to moderate correlation between the predictor variables for $a = (0, 0.5, 1, 2)$. In addition we compare the performance of the proposed varying coefficient function estimators to two alternative estimators: (1) the benchmark estimator using the true lag values and (2) the estimator from the concurrent VCM ignoring lag effects. The benchmark estimator and estimator from the concurrent model are obtained by substituting Δ_{rt} and 0, respectively, in place of $\hat{\Delta}_{rt}$, $r = 1, \ldots, p$, in step 3 of the proposed estimation algorithm. Table 1 summarizes the estimated median, 1st quartile and 3rd quartile values from 200 Monte Carlo runs for ML from the proposed lag estimators and for ME from the varying coefficient function estimators from all three models. Note that although the mean integrated squared errors for the varying coefficient function and lag estimators are getting larger with increasing predictor correlation, the proposed lag and varying coefficient function estimators are still performing well under moderate correlation levels, especially compared to the benchmark and concurrent models. The benchmark estimators are expected to perform the best since they use the true unknown lag values; and the performance of the proposed estimators are fairly close to the performance of the benchmark estimators. In addition, there is substantial gain in estimating the lag relations, as observed from comparisons of the proposed estimators to those from the concurrent model ignoring lag effects. The estimated median of the ME of the proposed varying coefficient function estimators range roughly between 1/5 to 1/2 of the estimated median of the ME from concurrent model estimators.

Figure 6 displays the median (dash-dotted) and the 5th and 95th percentiles (dotted) of the varying coefficient function estimates (Figure 6 (a), (b), (c), (d) and (e)) along with medians of the estimated lags (thick dotted) for each longitudinal predictor (Figure 6 (f), (g), (h) and (i)) over 200 Monte Carlo runs for $n = 700$ and $a = 1$. Even though there are some deviations from the underlying lags in the estimated lags due to correlated predictors, the medians of the estimated varying coefficient functions trace the underlying quantities closely, indicating that the method performs well with cross-sectional as well as longitudinal predictors in the model. Also plotted in thick gray are the medians of the varying coefficient function estimates from the concurrent models; median estimates from the benchmark model are very close to those plotted from the proposed model and hence omitted in the figures. The concurrent fit ignoring lag estimation deviates from the underlying varying coefficient functions considerably especially for those longitudinal predictors that are not related to the response from concurrent times (Figure 6 (b), (c) and (d)). Even for the concurrent longitudinal predictor (Figure 6 (a)) and the cross-sectional covariate (Figure 6 (e)), the estimated varying coefficient functions from the concurrent model visibly deviate from the true values. This is consistent with the larger ME values from the concurrent model estimates reported in Table 1.

5. Discussion

The works described here contain innovations in exploratory modeling of both the varying coefficient functions and the lag effects in the context of challenging longitudinal data conditions (e.g., possibly unsynchronized response-predictor observations, infrequency, and irregularity). The main objective of the proposed exploratory time varying lagged regression model is to detect deviations from the concurrent nature of VCM via the introduction of the time varying lags associated with each longitudinal predictor. In addition to allowing the regression relations to vary over time, the time varying lags enable relating the response to a time varying path from the predictors past. Hence the analysis is not necessarily limited to concurrent relations, instead potentially interesting lagged relationships are explored between the response and predictor processes according to an objective criterion. The flexibility of estimating separate lags for each longitudinal predictor is a strength of the proposal, so that some predictors can be included in the model from concurrent times as the response while others may enter the regression model from lagged times, such as in the analysis of clinic visits data. We note that the proposal is only an exploratory tool and that longitudinal predictors with nonzero estimated lags can be modeled from lagged time intervals or possibly with the entire history of the predictor via functional linear models in follow-up analysis.

A special case of the proposal that may be of interest in particular applications is time invariant lags, i.e. $t - \Delta_{rt} = t - c_r$ for some constant c_r for the rth longitudinal predictor. For selecting a fixed lag c_r for a given longitudinal predictor, the proposed algorithm can be altered to maximize a combined optimization criteria over t such as $\int[\mathrm{cov}\{Y(t), X_r(t-c_r)\}/\mathrm{var}\{X_r(t-c_r)\}]dt$ with respect to the constant c_r. This would be interpreted as maximizing the absolute value of the integrated correlation of the response with the lagged predictor over time for homoskedastic predictor processes.

The proposed estimation procedure addresses new challenges in longitudinal data, specifically data resulting from irregular, infrequent, unsynchronized, and error-prone longitudinal designs. Hence, time varying lags are explored without requiring a common time grid for observations among subjects. The proposed estimation procedure utilizes nonparametric estimation techniques to target the moments of the longitudinal predictor and response processes, which allow for pooling information across subjects. The curse of dimensionality is addressed by taking a conditional modeling approach, which allows for inclusion of multiple longitudinal and cross-sectional predictors. The variance of the proposed estimators are conveyed via bootstrap error bars in Section 3; development of valid inference procedures for the EVarlag model requires further research.

Acknowledgments

We thank Drs. Yolonda Pickett and Jeffrey M. Lyness for providing the data for analysis. The original study generating the data set was supported by the National Institute of Mental Health [R01 MH061429]. This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases [R01 DK092232 to D.S. and D.V.N.], the National Center for Advancing Translational Sciences [UL1 TR000153 to D.V.N.] and the National Institute of Mental Health [P30 MH085943 and R01 MH087562 to S.G.].

Appendix

We provide here the moments-based estimates of ESE used in step 2 of the estimation procedure. We estimate (8) in step 2.1 by

$$\widehat{\text{ESE}}_\ell^{(1)}(t)=\widehat{\text{var}}\{Y(t)\}-2\widehat{\gamma}_\ell^{(1)}(t)\widehat{\text{cov}}\{Y(t),W_\ell^{(1)}(t)\}+\{\widehat{\gamma}_\ell^{(1)}(t)\}^2\widehat{\text{var}}\{W_\ell^{(1)}(t)\}, \quad (13)$$

where $\widehat{\gamma}_\ell^{(1)}(t)=\widehat{\beta}_r^{(1)}(t)$ for longitudinal predictors and $\gamma_\ell^{(1)}(t)=\widehat{\alpha}_q^{(1)}(t)$ for cross-sectional ones.

Similar to (13), $\text{ESE}_\ell^{(2)}(t)$ in step 2.2 is estimated by

$$\begin{aligned}
&\widehat{\text{ESE}}_\ell^{(2)}(t)\\
&\quad=\widehat{\text{var}}\{Y(t)\}\\
&\quad-2\widehat{\gamma}_{\ell1}^{(2)}(t)\widehat{\text{cov}}\{Y(t),W_1^*(t)\}\\
&\quad-2\widehat{\gamma}_{\ell2}^{(2)}(t)\widehat{\text{cov}}\{Y(t),W_\ell^{(2)}(t)\}+2\widehat{\gamma}_{\ell1}^{(2)}(t)\widehat{\gamma}_{\ell2}^{(2)}(t)\widehat{\text{cov}}\{W_1^*(t),W_\ell^{(2)}(t)\}\\
&\quad+\{\widehat{\gamma}_{\ell1}^{(2)}(t)\}^2\widehat{\text{var}}\{W_1^*(t)\}\\
&\quad+\{\widehat{\gamma}_{\ell2}^{(2)}(t)\}^2\widehat{\text{var}}\{W_\ell^{(2)}(t)\}.
\end{aligned}$$

At step 2.*k*, $\text{ESE}_\ell^{(k)}(t)$ is estimated by

$$\begin{aligned}
&\widehat{\text{ESE}}_\ell^{(k)}(t)\\
&\quad=\widehat{\text{var}}\{Y(t)\}\\
&\quad-2\sum_{u=1}^{k-1}\widehat{\gamma}_{\ell u}^{(k)}(t)\widehat{\text{cov}}\{Y(t),W_u^*(t)\}\\
&\quad-2\widehat{\gamma}_{\ell k}^{(k)}(t)\widehat{\text{cov}}\{Y(t),W_\ell^{(k)}(t)\}+2\sum_{u=1}^{k-1}\widehat{\gamma}_{\ell u}^{(k)}(t)\widehat{\gamma}_{\ell k}^{(k)}(t)\widehat{\text{cov}}\{W_u^*(t),W_\ell^{(k)}(t)\}\\
&\quad+\sum_{u=1}^{k-1}\{\widehat{\gamma}_{\ell u}^{(k)}(t)\}^2\widehat{\text{var}}\{W_u^*(t)\}\\
&\quad+\{\widehat{\gamma}_{\ell k}^{(k)}(t)\}^2\widehat{\text{var}}\{W_\ell^{(k)}(t)\}.
\end{aligned}$$

References

Cleveland, WS.; Grosse, E.; Shyu, WM. Local regression models. In Statistical Models in S. Chambers, JM.; Hastie, TJ., editors. Pacific Grove: Wadsworth & Brooks; 1991. p. 309-376.

Fan J, Zhang JT. Two-step estimation of functional linear model with application to longitudinal data. Journal of the Royal Statistical Society Series B. 2000; 62:303–322.

Fan J, Zhang W. Statistical methods with varying coefficient models. Statistics and its Interface. 2008; 1:179–195. [PubMed: 18978950]

Fitzmaurice, GM.; Laird, NM.; Ware, JH. Applied Longitudinal Analysis. Hoboken: New Jersey: John Wiley & Sons, Inc; 2004.

Grabovich A, Lu N, Tang W, Tu X, Lyness JM. Outcomes of subsyndromal depression in older primary care patients. American Journal of Geriatric Psychiatry. 2010; 18:227–235. [PubMed: 20173424]

Hastie T, Tibshirani R. Varying coefficient models. Journal of the Royal Statistical Society Series B. 1993; 55:757–796.

Hoover DR, Rice JA, Wu CO, Yang LP. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika. 1998; 85:809–822.

Huang JZ, Wu CO, Zhou L. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. Biometrika. 2002; 89:111–128.

Koru-Sengul T, Stoffer DS, Day NL. A residuals-based transition model for longitudinal analysis with estimation in the presence of missing data. Statistics in Medicine. 2007; 26:3330–3341. [PubMed: 17124699]

Liu, B.; Mueller, HG. Functional data analysis for sparse auction data. Wiley and Sons Inc; 2008. p. 269-290.

Malfait N, Ramsay JO. The historical functional linear model. Canadian Journal of Statistics. 2003; 31:115–128.

Mueller HG, Yang W. Dynamic relations for sparsely sampled Gaussian processes. Test. 2010; 19:1–29.

Mueller HG, Zhang Y. Time varying functional regression for predicting remaining lifetime distributions from longitudinal trajectories. Biometrics. 2005; 61:1064–1075. [PubMed: 16401280]

Pickett YR, Ghosh S, Anne R, Gary JK, Bruce ML, Lyness JM. Healthcare use in elderly with minor depression. American Journal of Geriatric Psychiatry, submitted. 2011

Seaburn DB, Lyness JM, Eberly S, King DA. Depression, perceived family criticism, and functional status among older, primary-care patients. American Journal of Geriatric Psychiatry. 2005; 13:766–772. [PubMed: 16166405]

Senturk D, Mueller HG. Generalized varying coefficient models for longitudinal data. Biometrika. 2008; 95:653–666.

Senturk D, Mueller HG. Functional varying coefficient models for longitudinal data. Journal of the American Statistical Association. 2010; 105:1256–1264.

Senturk D, Nguyen DV. Varying coefficient models for sparse noise-contaminated longitudinal data. Statistica Sinica. 2011; 21:1831–1856.

Ya F, Mueller HG, Clifford AJ, Dueker SR, Follett J, Lin Y, Buchholz B, Vogel JS. Shrinkage estimation for functional principal component scores, with application to the population kinetics of plasma folate. Biometrics. 2003; 59:676–685. [PubMed: 14601769]

Yao F, Mueller HG, Wang JL. Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association. 2005; 100:577–590.

Wu CO, Chiang CT. Kernel Smoothing on varying coefficient models with longitudinal dependent variable. Statistica Sinica. 2000; 10:433–456.

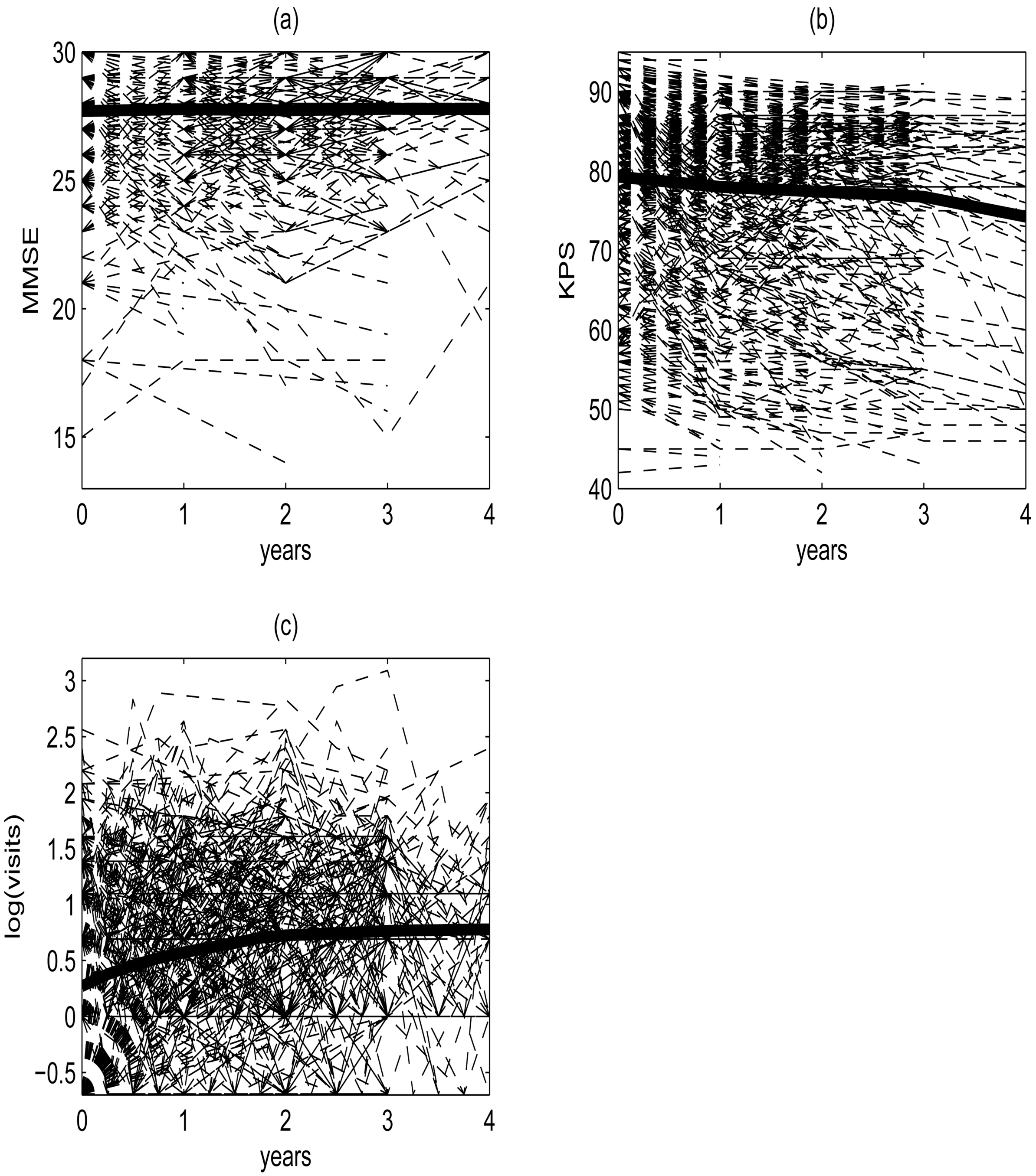

Figure 1.

Observed individual trajectories (dashed) and the smoothed estimate of the cross-sectional mean functions (thick solid) for the (a) cognitive impairment score MMSE, (b) physical impairment score KPS and (c) logarithm of the number of clinic visits.

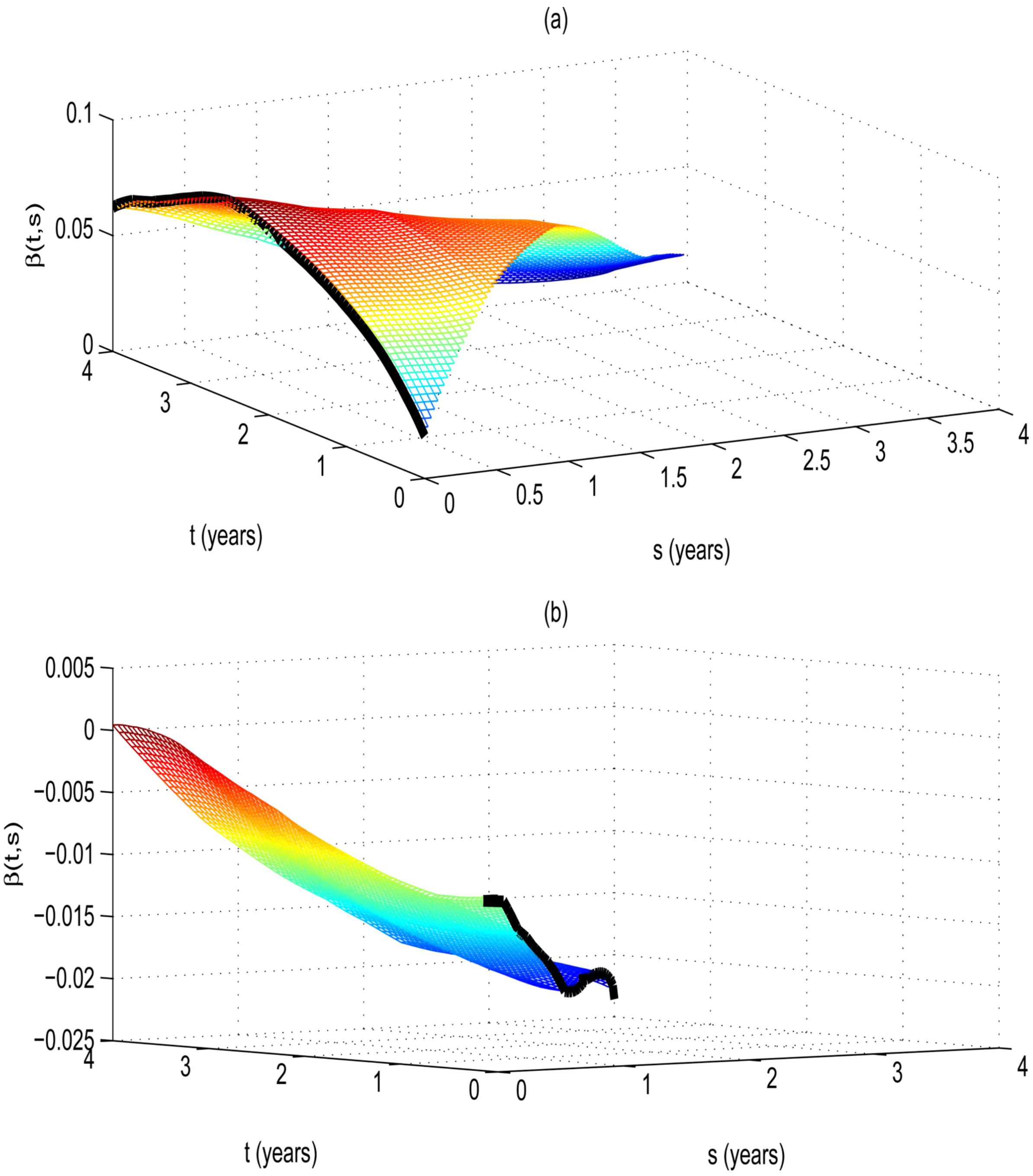

Figure 2.

Estimated two dimensional transfer functions from the separate regressions of the logarithm of the number of clinic visits at time *t* (a) on MMSE and (b) on KPS scores from the lagged times $s \le t$. Plotted in thick solid in black is the path that maximizes the transfer function for each *t*, corresponding to the estimated lags in the proposed exploratory time varying lagged regression.

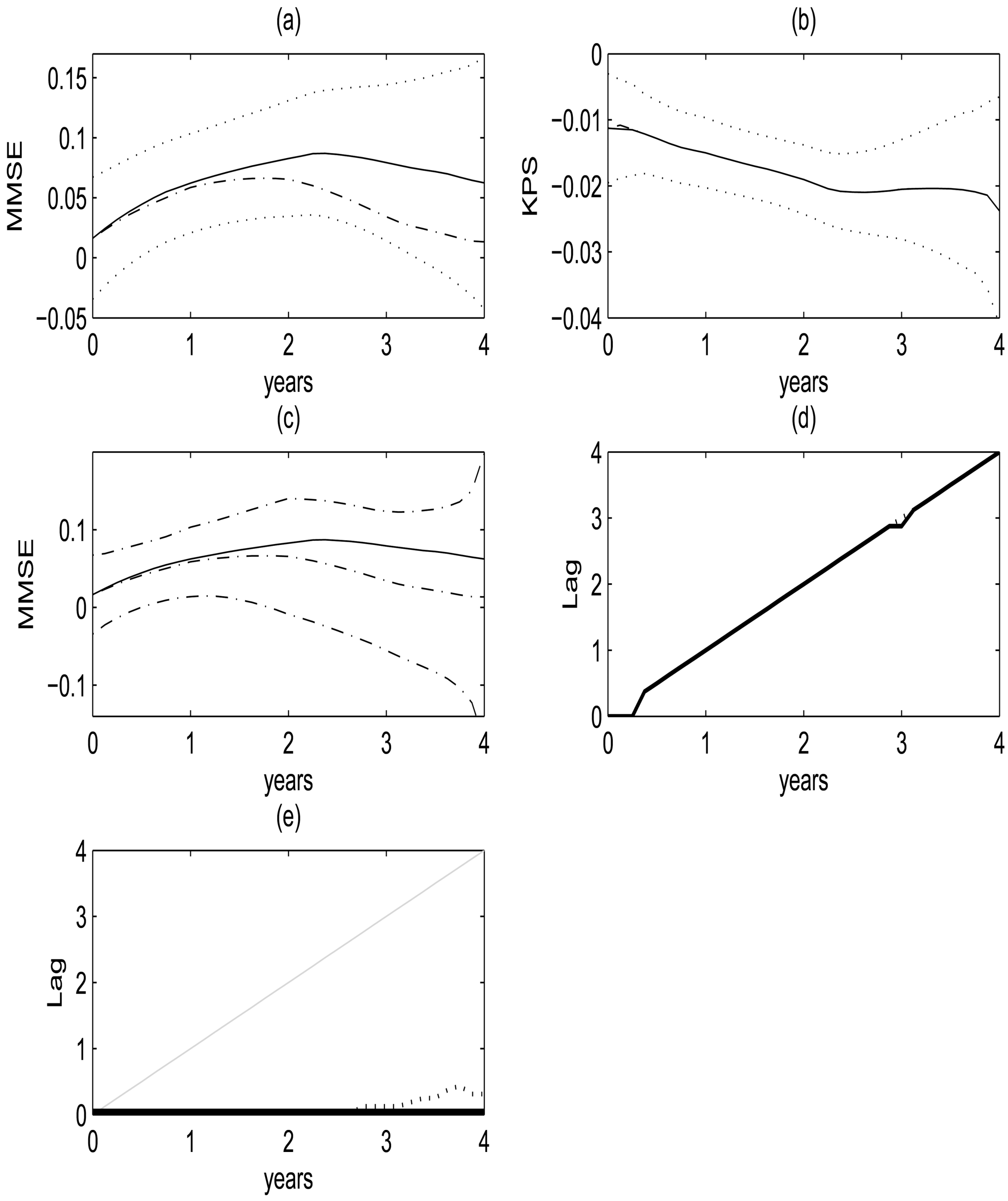

Figure 3.

Two separate (univariate) EVarlag regression models. Given are the estimated varying coefficient functions from the proposed EVarlag regression (solid) and from the varying coefficient regression (dash-dotted) of the logarithm of number of visits (a) on the cognitive impairment score MMSE and (b) on the physical impairment score KPS. ±2 bootstrap error bars from the EVarlag fits (dotted) are given in (a) and (b) and from the varying coefficient model fit (dash-dotted) are given in (c). Also plotted are the estimated lags t vs. $t - \Delta_t$ (thick solid) from the two regressions with MMSE and KPS given in (e) and (d), respectively.

Gray line in (e) denotes the no lag case, i.e. $\Delta_t = 0$ for reference, where thick dotted lines in (d) and (e) are the estimated median bootstrap lag choices.

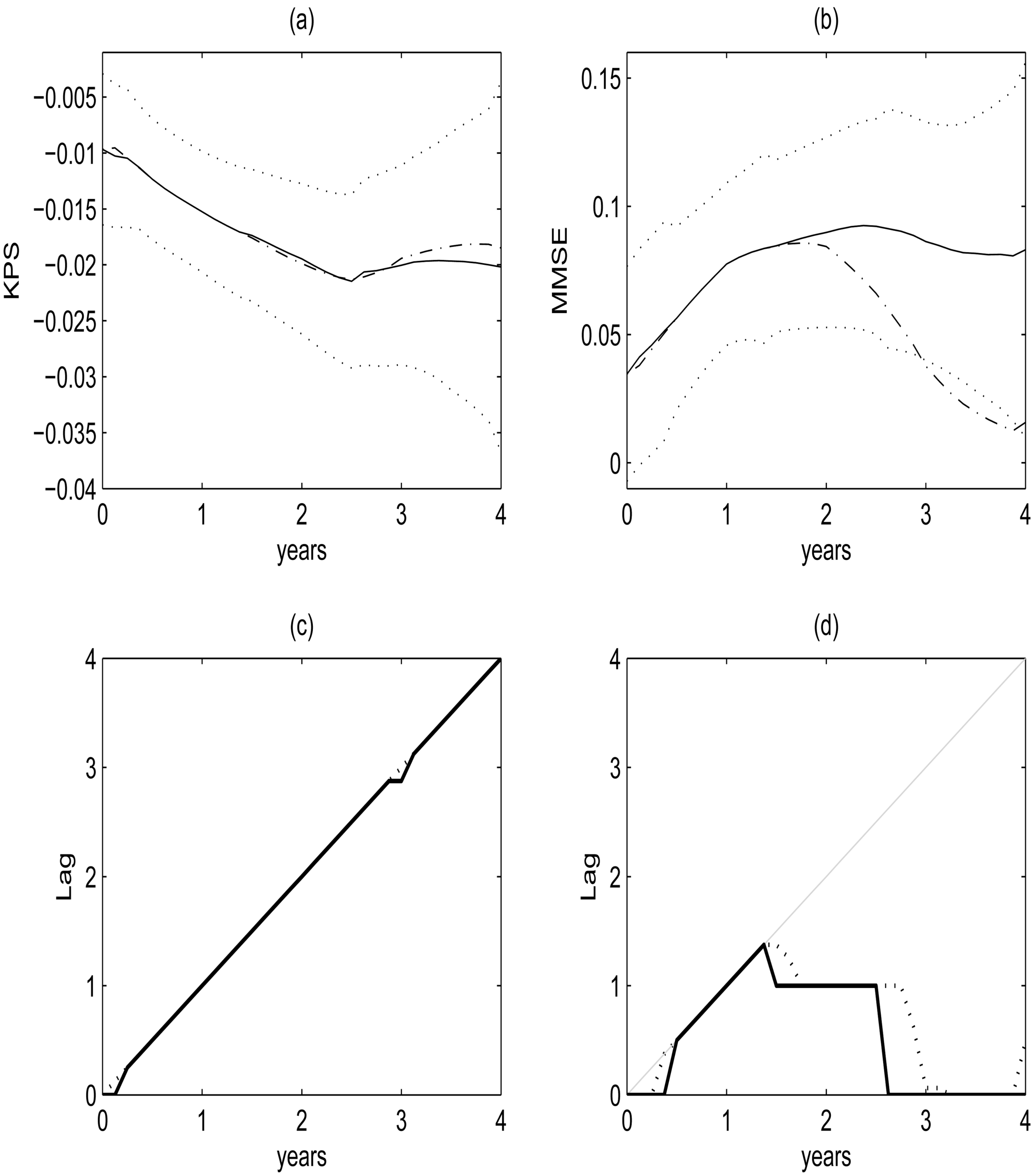

Figure 4.

EVarlag model with multiple predictors (predictor sequence: KPS and MMSE). Plotted are estimated varying coefficient functions from the proposed EVarlag regression (solid) and from the varying coefficient regression (dash-dotted) of the logarithm of number of visits on two predictors, (a) the physical impairment score KPS and (b) the cognitive impairment score MMSE. Also provided are ±2 bootstrap error bars (dotted) for the EVarlag fits. Estimated lags t versus $t - \Delta_t$ (thick solid) for KPS and MMSE are given in (c) and (d), respectively. The gray line in (d) denotes the no lag case, i.e. $\Delta_t = 0$ for reference, where thick dotted lines in (c) and (d) are the estimated median bootstrap lag choices.

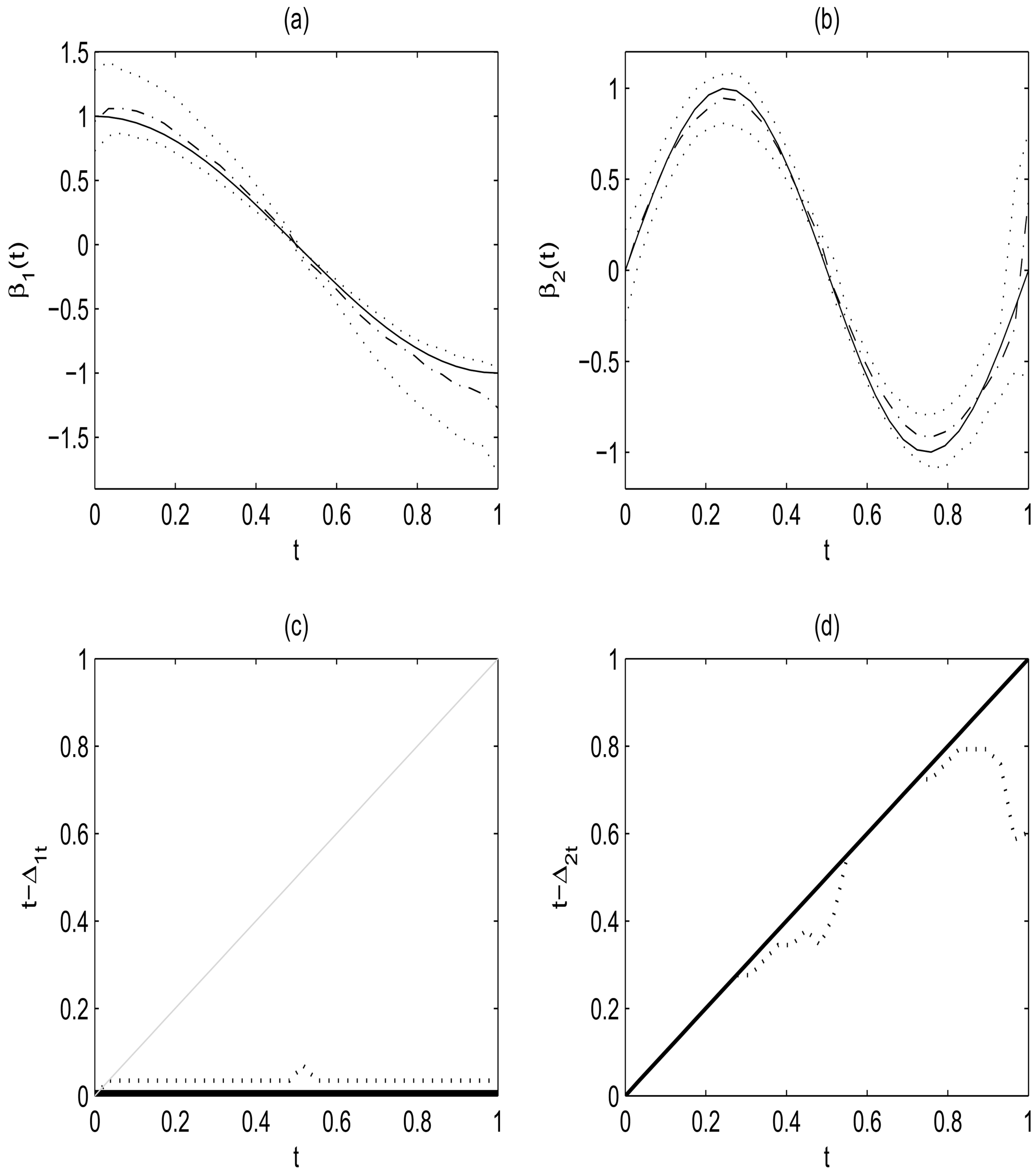

Figure 5.

Results from the first simulation study with data set-up similar to the clinic visits data. The median (dash-dotted) and the 5th and 95th percentiles (dotted) of the varying coefficient function estimates of (a) β$_1$(t) and (b) β$_2$(t) (solid) are plotted based on 200 Monte Carlo runs. Given in (c) and (d) are the medians (thick dotted) of the estimated lags t versus $t - \Delta_t$ (solid) for each longitudinal predictor. The gray line in (c) represents the no lag case, i.e. $\Delta_t = 0$ for reference.

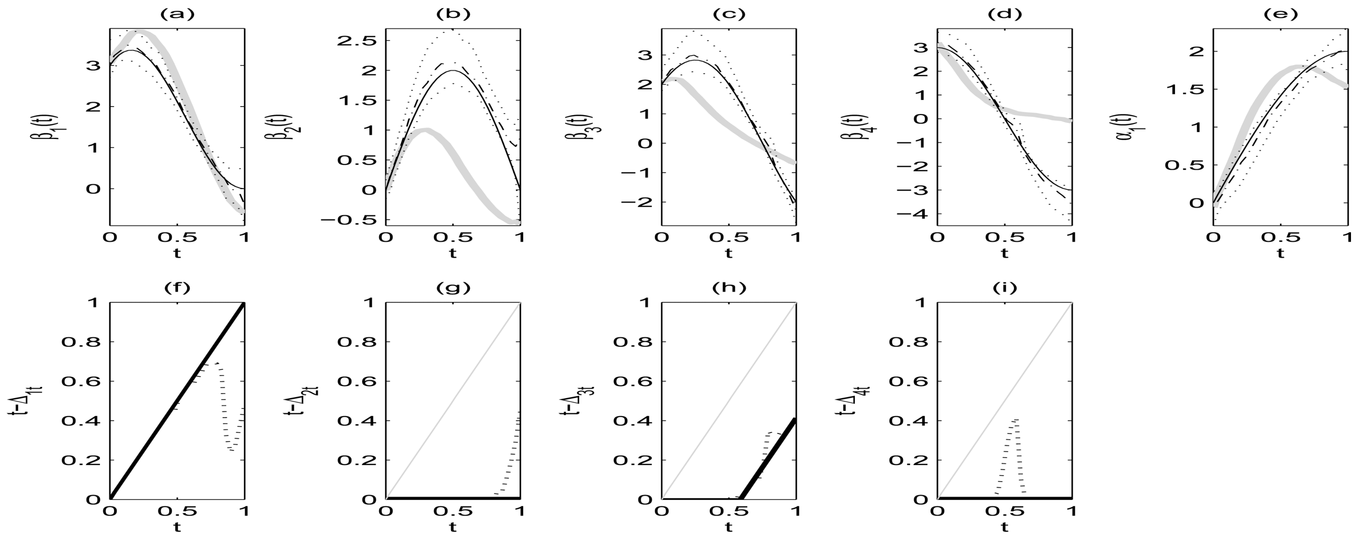

Figure 6.

Results from the second simulation study. The median (dash-dotted) and the 5th and 95th percentiles (dotted) of the varying coefficient function estimates ((a), (b), (c), (d) and (e)) are plotted based on 200 Monte Carlo runs (true values in solid). Plotted in thick gray are the medians of the varying coefficient functions estimates from the concurrent model. Also given ((f), (g), (h) and (i)) are the medians (thick dotted) of the estimated lags t versus $t - \Delta_t$ (solid) for longitudinal predictors. Gray lines represent the no lag case for reference.

Table 1

Normalized mean squared integrated error for the varying coefficient functions (ME) from the proposed estimators, benchmark estimators using true unknown lag values and the concurrent estimators ignoring lag estimation. Also reported are the median normalized mean squared integrated error for lag estimators (ML) from the proposed algorithm. Median and 25% and 75% percentiles of the deviation measures are presented based on 200 Monte Carlo runs/data sets. Varying a values refer to differing degrees of correlation between the predictor variables; a = (0, 0.5, 1, 2) corresponds to correlations of (0, 1/12, 1/6, 1/3), respectively.

n	*a*	ME			ME-benchmark			ME-concurrent			ML
		Median	**25%**	**75%**	**Median**	**25%**	**75%**	**Median**	**25%**	**75%**	**Median**
700	0	.069	.040	.130	.063	.042	.103	.261	.245	.281	.110
700	0.5	.064	.043	.127	.067	.046	.097	.273	.257	.292	.106
700	1	.098	.057	.196	.084	.058	.160	.319	.302	.339	.128
700	2	.276	.113	.809	.156	.083	.385	.533	.510	.567	.165
1000	0	.043	.026	.071	.043	.028	.073	.258	.246	.273	.081
1000	0.5	.055	.033	.086	.054	.035	.084	.265	.251	.279	.105
1000	1	.056	.033	.125	.056	.036	.111	.313	.293	.332	.115
1000	2	.156	.085	.311	.104	.065	.221	.524	.500	.553	.145