Statistica Neerlandica (2008) Vol. 62, nr. 1, pp. 131–154
doi:10.1111/j.1467-9574.2007.00379.x

Continuous-time modelling of irregularly spaced panel data using a cubic spline model

Sy-Miin Chow*

University of North Carolina, CB#3270 Davie Hall, Chapel Hill, NC 27599-3270, USA

Guangjian Zhang†

University of Notre Dame, 108 Haggar Hall, Notre Dame, IN 46556-5611, USA

Continuous-time modelling remains a somewhat 'idealized' representation tool. Even though conceptualizing a dynamic process as a continuous process has clear appeal from a theoretical standpoint, practical tools that allow researchers to effectively map an idealized continuous model onto a set of discrete-time observed data are still lacking observed data. Irregularly spaced longitudinal data frequently arise in empirical settings because of the prevalence of longitudinal studies with partially randomized measurement intervals and other related designs. We present a practical approach that capitalizes on a nonparametric spline interpolation approach to impute the gaps in irregularly spaced panel data. Simulated and empirical examples are provided to demonstrate the applicability of the proposed approach to studies of group-based dynamics using panel data.

Keywords and Phrases: stochastic differential equation, state-space modelling, Kalman filter, smoothing, exact discrete time, panel data.

1 Introduction

Consider the case of a swinging pendulum: a researcher who is attempting to measure the behaviour of the pendulum can only take snapshots of the pendulum from time to time, thus getting experimental data that are discrete in nature. Even though identifying the points at which the pendulum exhibits certain critical changes can help identify the regularities in the pendulum's behaviour, it is apparent that the behaviour of the pendulum itself is by no means discrete. In the physical sciences and engineering, dynamic processes are often construed and represented as continuous-time processes. Discrete-time models, in contrast, are still the prominent modelling tools in statistics and social sciences because of the kinds of data that are

*symiin@email.unc.edu
†gzhang3@nd.edu

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

typically available to researchers in such disciplines. Thus, discrete-time models are adopted with the understanding that the time evolution of human behaviours is, in most instances, continuous in nature.

In the present article, we echo the hunt for continuity put forth by other researchers in this special issue for one practical reason: most longitudinal data in social sciences, despite our best efforts, are often sampled at irregularly spaced intervals. For instance, the last decade has seen a surge of daily diary designs used, e.g. in the studies of affect (Larsen and Ketelaar, 1991) and interpersonal processes (Laurenceau, Barrett and Pietromonaco, 1998; Thompson and Bolger, 1999). In more recent adaptations of such designs, participants are further prompted to provide momentary reports at random times within several predesignated time blocks throughout the day (Hawkley *et al.*, 2003; Sbarra, 2006; Ong, Horn and Walsh, 2007), with the aim of obtaining more accurate assessments of the participants' underlying status or feelings *at the moment*.

We provide a practical approach to representing continuous processes using *SsfPack*, a suite of C routines for implementing state–space modelling techniques, including numerical routines for fitting models in state–space form (Koopman, Shephard and Doornik, 1999). *SsfPack* is one of the many statistical packages implemented using Ox, an object-oriented matrix programming language with a comprehensive mathematical and statistical function library (Doornik, 1998). Specifically, we present a panel-data extension of a nonparametric interpolation technique suited for interpolating irregularly spaced missing data (Wecker and Ansley, 1983; Kohn and Ansley, 1987; Durbin and Koopman, 2001). The nonparametric nature of the approach makes it especially suited for representing very diverse patterns of intra-individual changes.

The article is organized as follows. We begin by presenting a set of stochastic differential equations associated with the cubic spline model (Wahba, 1978; Durbin and Koopman, 2001), which serves as the basis of our modelling examples. We show how this model can be fitted to panel data with irregularly spaced intervals as an exact discrete-time model through simulation examples. An empirical example is then provided to illustrate one possible way in which the cubic spline model can be combined with other models aimed at capturing momentary deviations from a group-based, global trend. To conclude, we discuss how such techniques can be used to represent group-based changes in panel data, including changes that unfold at different rates.

2 Continuous-time spline-smoothing model

In this paper, we utilize a function in *SsfPack* termed *GetSsfSpline* to interpolate the 'gaps' in irregularly spaced panel data by means of a cubic spline function. Specifically, we define panel data as data that have a much larger number of participants (e.g. $N = 100$, 200) than number of measurement occasions (denoted as T; e.g.

$T \leq 10$). The continuous-time model on which the discretized interpolation model is based is constructed from a continuous-time spline-smoothing model (Wahba, 1978; Wecker and Ansley, 1983; Kohn and Ansley, 1987). We summarize below the key features of these two models and where appropriate, refer interested readers to elsewhere for further derivation details (Koopman *et al.*, 1999; Durbin and Koopman, 2001).

Consider a scenario where person i's continuous state, $\eta_i(t)$, is observed at discrete time points $t_{1i}, t_{2i}, \ldots, t_{Ti}$, with $\delta_{t_{ki}}$ representing the time elapsed between t_{ki} and $t_{k+1,i}$ for person i. In discrete-time dynamic models (e.g. difference equation models), $\delta_{t_{ki}}$ is assumed to be constant for all time points (i.e. $\delta_{t_{ki}} = \delta$ for $t_{1i}, \ldots, t_{Ti}$). Their continuous-time analogues have the added flexibility of incorporating $\delta_{t_{ki}}$ as a time-varying entity to accommodate irregular sampling intervals. A particularly flexible way of representing $\eta_i(t)$ and its continuous changes without necessarily imposing a theoretically driven model is to model $\eta_i(t)$ as composed of a local intercept, $\mu_i(t)$, and a local slope, $\beta_i(t)$. Letting $\eta_i(t) = [\mu_i(t)\ \beta_i(t)]'$, a set of linear stochastic differential equations in Itô form is used to represent person i's trajectory on a construct of interest as

$$\mathrm{d}\begin{bmatrix}\mu_i(t)\\ \beta_i(t)\end{bmatrix} = \begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}\begin{bmatrix}\mu_i(t)\\ \beta_i(t)\end{bmatrix}\mathrm{d}t + \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}0\\ \mathrm{d}w_i(t)\end{bmatrix}, \tag{1}$$

written more compactly as

$$\mathrm{d}\eta_i(t) = A\eta_i(t)\mathrm{d}t + G\mathrm{d}W_i(t), \tag{2}$$

where the component $w_i(t)$ is a Wiener process[1] (also referred to as Brownian motion) whose increments $\mathrm{d}w_i(t)$ follow a normal distribution with a mean 0 and a variance that increases linearly with the length of time interval, namely, $\sigma_w^2 \mathrm{d}t$. Specifically, the increment in $W_i(t)$ between two person-specific time points, denoted below as $W_i(s_i) - W_i(r_i)$, is assumed to be normally distributed with

$$W_i(s_i) - W_i(r_i) \sim N\left([0 \quad 0]', (s_i - r_i)\Psi\right), \quad \text{where } s_i > r_i, \tag{3}$$

and

$$\Psi = \begin{bmatrix}0 & 0\\ 0 & \sigma_w^2\end{bmatrix}. \tag{4}$$

The matrix A is known as the drift matrix and it governs the drift rate or local change in $\eta_i(t)$ over the interval $\mathrm{d}t$. σ_w^2 governs the rate of variance change, often referred to as the *volatility factor* in economics (Tsay, 2005). Note that no process noise is added directly to $\mu_i(t)$, but rather, the process noise associated with the local slope is manifested indirectly on $\mu_i(t)$ through $\beta_i(t)$. Constraining the model this way forces $\mu_i(t)$ to form a smooth trajectory, which is one plausible representation of trends or more systematic growth patterns. As we explicate at the end of this section, the model is nonparametric because fitting the model to a set of data essentially involves performing cubic spline smoothing on the data.

Equation (1) yields the solution (see also DURBIN and KOOPMAN, 2001, p. 59)

$$\beta_i(t) = \beta_i(0) + \sigma_w w_i(t), \tag{5}$$

$$\mu_i(t) = \mu_i(0) + \beta_i(0)t + \int_0^t w_i(s)\mathrm{d}s. \tag{6}$$

In fact, any process in the form of Equation (2), when discretized over (possibly) irregular time intervals, has a general solution that can be expressed in an alternative form, namely, as a 'one-step-ahead' or first-order autoregressive process. In subsequent sections, we denote η_{it_k} as the discrete-time counterpart of $\eta_i(t_{ki})$, where the person index associated with η is retained but that associated with t_{ki} is omitted to ease presentation. The dynamics of η_{it_k} can then be written as a one-step-ahead model in which the discrete measurement times (e.g. t_{ki} and $t_{k-1,i}$) and the corresponding time intervals (i.e. $\delta_{t_{ki}}$ for $t_{1i},\ldots,t_{Ti}$), are individual-specific. The relationship between the state vector at time t_{ki} and $t_{k+1,i}$ can be expressed as (see for instance, HARVEY, 2001 p. 484, Equation 9.1.18)

$$\eta_{it_{k+1}} = \mathrm{e}^{A\delta_{t_{ki}}}\eta_{it_k} + \int_0^{\delta_{t_{ki}}} \mathrm{e}^{A(\delta_{t_{ki}}-s)} G\mathrm{d}W_i(t_{ki}+s), \tag{7}$$

where $\delta_{t_{ki}} = t_{k+1,i} - t_{ki}$. Equation (7) can be re-expressed as

$$\eta_{it_{k+1}} = B_{it_k}\eta_{it_k} + \zeta_{it_k}, \tag{8}$$

where

$$B_{it_k} = \mathrm{e}^{A\delta_{t_{ki}}}, \tag{9}$$

and ζ_{it_k} is a multivariate white-noise disturbance term with zero mean and covariance matrix

$$\Psi_{it_k} = \int_0^{\delta_{t_{ki}}} \mathrm{e}^{A(\delta_{t_{ki}}-s)} G\Psi G' \mathrm{e}^{A'(\delta_{t_{ki}}-s)}\mathrm{d}s. \tag{10}$$

In general, any matrix exponential function can be expressed as a power series as

$$\mathrm{e}^A = I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \cdots, \tag{11}$$

which is equivalent to performing a Taylor series expansion on the function e^A around 0. Keeping only the first-order term from the Taylor series expansion, B_{it_k} in Equation (9) can now be expressed as (see HARVEY, 2001, Equation 9.2.3)

$$B_{it_k} = \exp\left(\delta_{t_{ki}}\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}\right) = I + \begin{bmatrix}0 & \delta_{t_{ki}}\\ 0 & 0\end{bmatrix} = \begin{bmatrix}1 & \delta_{t_{ki}}\\ 0 & 1\end{bmatrix}, \tag{12}$$

and Equation (8) now appears as a discrete-time linear dynamic model,[2] written as

$$\begin{bmatrix}\mu_{it_{k+1}}\\ \beta_{it_{k+1}}\end{bmatrix} = \begin{bmatrix}1 & \delta_{t_{ki}}\\ 0 & 1\end{bmatrix}\begin{bmatrix}\mu_{it_k}\\ \beta_{it_k}\end{bmatrix} + \begin{bmatrix}\zeta_{\mu,it_k}\\ \zeta_{\beta,it_k}\end{bmatrix}. \tag{13}$$

As mentioned earlier, the discrete-time transition matrix, B_{it_k}, only varies across individuals because of the possibly person-varying measurement intervals, $\delta_{t_{ki}}$. In

addition, even though the original continuous-time stochastic differential equation model shown in Equation (1) (see also the expression for Ψ in Equation 4) does not have a distinct source of process noise for the change in local intercept, $\mathrm{d}\mu_i(t)$, the process noise associated with the change in local slope, $\mathrm{d}\beta_i(t)$, is manifested indirectly on $\mu_i(t)$ (see Equation 6). As a result, in the discrete-time version of the model, the variance of $\zeta_{\mu,\, it_k}$ is not zero but rather, a constrained function of σ_w^2, the volatility factor of the Wiener process governing how much the local slope changes over a specific time interval.

Here, we derive the constraints that have to be imposed on the covariance structure of ζ_{it_k} to establish the equivalence between the linear state–space model and the original continuous-time model. Substituting the matrix exponential function defined in Equation (12) into Equation (10) and setting $G = I_2$ as defined in Equation (1) yields

$$\Psi_{it_k} = \int_0^{\delta_{t_{ki}}} \begin{bmatrix} 1 & \delta_{t_{ki}} - s \\ 0 & 1 \end{bmatrix} I_2 \begin{bmatrix} 0 & 0 \\ 0 & \sigma_w^2 \end{bmatrix} I_2 \begin{bmatrix} 1 & 0 \\ \delta_{it_{ki}} - s & 1 \end{bmatrix} \mathrm{d}s \tag{14}$$

$$= \sigma_w^2 \int_0^{\delta_{t_{ki}}} \begin{bmatrix} \delta_{t_{ki}}^2 - 2\delta_{t_{ki}} s + s^2 & \delta_{t_{ki}} - s \\ \delta_{t_{ki}} - s & 1 \end{bmatrix} \mathrm{d}s \tag{15}$$

$$= \sigma_w^2 \delta_{t_{ki}} \begin{bmatrix} \frac{\delta_{t_{ki}}^2}{3} & \frac{\delta_{t_{ki}}}{2} \\ \frac{\delta_{t_{ki}}}{2} & 1 \end{bmatrix}. \tag{16}$$

In sum, $\zeta_{\mu,\, it_k}$ now has occasion-specific variance that is a constrained function of the process noise variance for the local slope, scaled by the time interval between two successive measurement occasions. The process noise variance associated with the local slope is proportional to σ_w^2, scaled by the time interval between two successive measurements. The local intercept, μ_{it_k} – now defined over possibly irregularly spaced, discrete-time intervals – is typically apprehended through a set of discrete-time manifest observations. In particular, the measurement model is expressed as

$$y_{it_k} = \mu_{it_k} + \epsilon_{it_k}, \tag{17}$$

where y_{it_k} represents person i's manifest observation on measurement occasion t_{ki} and ϵ_{it_k} the corresponding measurement error on that occasion. In this way, μ_{it_k} will take on the form of a smooth trajectory whose variability stems from changes in the local slope only. The local slope, β_{it_k}, is assumed to give rise to local changes in the manifest observations *indirectly* through μ_{it_k}.

To facilitate model estimation, the measurement error variance, σ_ϵ^2, is usually fixed at a prespecified value (e.g. at 1.0) in practice and the ratio between σ_w^2 and σ_ϵ^2 – namely, the signal-to-noise ratio, denoted as $q = \sigma_w^2/\sigma_\epsilon^2$ – is estimated instead. This yields

$$\Psi_{it_k} = \begin{bmatrix} \frac{q\delta_{t_{ki}}^3}{3} & \frac{q\delta_{t_{ki}}^2}{2} \\ \frac{q\delta_{t_{ki}}^2}{2} & q\delta_{t_{ki}} \end{bmatrix}. \tag{18}$$

If no constraints are imposed on the covariance structure of ζ_{it_k}, the discrete–time model shown in Equation (13) is generally referred to as the local linear trend model (e.g. Harvey, 1989; Durbin and Koopman, 2001). Its constrained, continuous-time analogue as detailed above is commonly referred to as the cubic spline model. This is because fitting the constrained version of the model in Equations (13–18) is equivalent to smoothing the data by means of a cubic spline (see also Kohn and Ansley, 1987). Specifically, fitting the constrained model has been shown to be equivalent to minimizing the penalty function

$$\sum_{k=1}^{T_i}[y_{it_k}-\mu_{it_k}]^2+q^{-1}\int_{t=a_i}^{b_i}\left(\frac{\partial^2\mu_i(t)}{\partial t^2}\right)^2 \mathrm{d}t, \tag{19}$$

where $k=1,\ldots,T_i$, $a\leq t_{1i}$ and $b\geq t_{T_i}$. The parameter q^{-1} in this case serves as a smoothing parameter and it is the only parameter that has to be estimated. If the signal-to-noise ratio, q, is small and thus q^{-1} is large, the estimated trajectory of $\hat{\mu}_{it_k}$ will be smooth but the estimates may not be close enough to the actual observations y_{it_k}. In contrast, if q is large and q^{-1} is small, the estimated $\hat{\mu}_{it_k}$ will be close to y_{it} but $\hat{\mu}_{it_k}$ may not be smooth enough (Wahba, 1978; Wecker and Ansley, 1983; Durbin and Koopman, 2001). The order of the polynomial spline is determined by the order m of the partial derivative term in Equation (19), which in this case, $m=2$. Fitting the state–space model in Equation (13) is then equivalent to fitting a polynomial spline of degree $2m-1$, namely, a cubic spline through the data by minimizing the penalty function in Equation (19). Therefore, this continuous-time version of the constrained local linear trend model is often referred to as the cubic spline model (Koopman *et al.*, 1999).

In sum, the exact discrete-time (EDM) approach provides a general framework for fitting stochastic differential equation models as constrained discrete-time models to (possibly) irregularly spaced data. This approach was explicated in detail by Bergstrom (1988) and it has been used to fit other stochastic differential equations in structural equation modelling and state–space modelling frameworks (e.g. Oud and Jansen, 2000; Hamerle, Nagl and Singer, 1990). The cubic spline model discussed in the present article is in particular a very flexible data-tracking and missing data interpolation device. The approach used to construct the cubic spline model in the present article and elsewhere (e.g. Durbin and Koopman, 2001) is similar to the conventional approach of using person-specific occasions as variables to accommodate irregular time intervals within and across individuals in standard mixed effects and hierarchical linear models (e.g. Bryk and Raudenbush, 1987; Verbeke and Molenberghs, 2000). Here, we fit the cubic spline model within a state–space framework, the details of which are summarized in the next section.

3 State–space representation in *SsfPack*

The state–space modelling framework provides a general way of representing the relationships among a set of dynamic processes and their associated observed data.

A common linear representational form is to express a process of interest in terms of a dynamic model written as

$$\eta_{it_{k+1}} = \alpha_{it_k} + B_{it_k}\eta_{it_k} + \zeta_{it_k}, \tag{20}$$

and a measurement model expressed as

$$y_{it_k} = \tau_{it_k} + \Lambda_{it_k}\eta_{it_k} + \epsilon_{it_k}, \tag{21}$$

where η_{it_k} is person i's $w \times 1$ vector of latent variables apprehended through a $p \times 1$ vector of manifest variables y_{it_k}; α_{it_k} a $w \times 1$ vector of constants on measurement occasion t_k for person i, B_{it_k} a $w \times w$ transition matrix depicting the transition of the system from time t_{ki} to $t_{k+1,i}$, ζ_{it_k} a $w \times 1$ vector of residuals or process noise for person i representing uncertainties not accounted for by the assumed model. Λ_{it_k} is a $p \times w$ matrix of factor loadings, τ_{it_k} a $p \times 1$ vector of intercepts and ϵ_{it_k} person i's p-variate vector of measurement errors. The process and measurement noises (ζ_{it_k} and ϵ_{it_k}) are assumed to be normally distributed with zero means and covariance matrices Ψ_{it_k} and Θ_{it_k} respectively. In the present context, the only parameter to be estimated is the person- and time-invariant signal-to-noise ratio, q.

Ψ_{it_k} and B_{it_k} are assumed to be occasion- and person-specific in the cubic spline model because of the unique time interval, $\delta_{t_{ki}}$, associated with each person and measurement occasion. The components Λ_{it_k}, α_{it_k}, τ_{it_k} and Θ_{it_k} are assumed in the present context to be person- and time-invariant (i.e. $\Lambda_{it_k} = \Lambda$, $\alpha_{it_k} = \alpha$, $\tau_{it_k} = \tau$, and $\Theta_{it_k} = \Theta$) but we include the person and time indices in presenting the associated state–space algorithms to highlight the generality of the state–space modelling framework.

Details pertaining to state–space modelling techniques are well documented (see for instance, Harvey, 1989; Durbin and Koopman, 2001). In *SsfPack*, the program used to fit the models considered herein, an alternative formulation is used. We focus here on elucidating the key differences between the modelling framework implemented in *SsfPack* and other, arguably more familiar state–space modelling frameworks, including the one postulated in Equations (20–21). Specifically, *SsfPack* concatenates the latent variable vector (e.g. $\eta_{it_{k+1}}$ in Equation 20) and the observed measurement vector (e.g. y_{it_k} in Equation 21) into one augmented vector. The key modelling equation is now written as

$$\begin{bmatrix} \eta_{it_{k+1}} \\ y_{it_k} \end{bmatrix} = d_{it_k} + \Phi_{i_t}\eta_{it_k} + u_{it_k}, u_{it_k} \sim N(0, \Omega_{it_k}) \tag{22}$$

where

$$d_{it_k} = \begin{bmatrix} \alpha_{it_k} \\ \tau_{it_k} \end{bmatrix}, \tag{23}$$

$$\Phi_{it_k} = \begin{bmatrix} B_{it_k} \\ \Lambda_{it_k} \end{bmatrix}, \tag{24}$$

$$u_{it_k} = \begin{bmatrix} \zeta_{it_k} \\ \epsilon_{it_k} \end{bmatrix} \tag{25}$$

and

$$\Omega_{it_k} = \begin{bmatrix} \Psi_{it_k} & \text{cov}(\zeta_{it_k}, \epsilon_{it_k}) \\ \text{cov}(\epsilon_{it_k}, \zeta_{it_k}) & \Theta_{it_k} \end{bmatrix}. \tag{26}$$

The latent state vector, $\eta_{it_{k+1}}$, and the observed measurement, y_{it_k} is offset by one lag caused by the nature of the one-step-ahead prediction postulated in Equation (20). Furthermore, the matrices B_{it_k} and Λ_{it_k} are vertically concatenated to yield a $(p+w)\times w$ matrix of transition and factor loadings (Equation 24). We adopt this general presentation in our illustrative examples because it gives a concise way to show the dynamic and measurement models in one equation.

This alternative representation form serves as the basis for constructing state–space models in *SsfPack*. In the *SsfPack* library, functions ranging from some of the more standard state–space modelling options to more advanced techniques that are currently unavailable in standard statistical programs (e.g. Markov chain Monte Carlo and importance sampling techniques; Koopman *et al.*, 1999) are included. (A selected summary of estimation procedures most pertinent to the present article is outlined here.)

4 The Kalman filter, Kalman smoothers and prediction error decomposition

The Kalman filter (KF) and the related Kalman smoothers are common approaches used to estimate the state (namely, the 'true score' or factor score) of a system at time t. Suppose the data $Y_{it_k} = [y_{1,it_k}, y_{2,it_k}, \ldots, y_{p,it_k}]'$ are available for estimation purposes, where y_{j,it_k} represents person i's time series on the jth manifest variable from $t_1, \ldots t_{k_i}$ and p the number of manifest variables available from person i on occasion t_{ki}. The KF can be used to derive state estimates based on manifest observations up to time t_{ki} [i.e. yielding $E(\eta_{it_k} | Y_{it_k})$]. In one alternative Kalman smoother, the fixed interval smoother (FIS; Anderson and Moore, 1979), the aim is to derive $E(\eta_{it_k} | y_{it_T})$ using all available data (Anderson and Moore, 1979). Applications of the FIS in psychology can be found in Oud (2002) and the reader is referred to the extant literature for the procedures involved in implementing this smoother. *SsfPack* implementation of the KF and the FIS (or the moment smoother) is detailed in Koopman *et al.* (1999; pp. 128–132).

For parameter estimation purposes, a log-likelihood function can be constructed using by-products from the KF via the prediction error decomposition (Schweppe, 1965; Caines and Rissanen, 1974; Harvey, 1989). The log-likelihood function can then be optimized with respect to model parameters to yield maximum likelihood estimates of the parameters. This log-likelihood function and a concentrated version of the log-likelihood function have been implemented in *SsfPack* together with several 'canned' numerical optimization routines from GAUSS for model-fitting purposes. Procedures associated with the KF, the FIS and prediction error decomposition are outlined in Appendix A. A collection of several key commands used in

Table 1. Selected SsfPack and Ox commands used for model fitting in Examples 1 and 2.

Ox commands	Functions
GetSsfSpline	Formulate the exact discrete time version of the nonparametric cubic spline model in state–space form
SsfLik	Return the log-likelihood function for a given state–space model
AddSsfReg	Add regressors to an existing state–space model with time-invariant parameters
MaxBGFS	Canned routine from GAUSS for performing numerical optimization
SsfMomentEst	Return output from prediction, forecasting or smoothing

the two examples considered in the present article is summarized in Table 1. Selected Ox scripts used in the present article are available from the first author's website at http://www.nd.edu/∼schow.

5 Illustrative examples

5.1 Example 1: Cubic spline model as a missing data interpolation technique

In the present example, we used simulated panel data to demonstrate the utility of the cubic spline model. We generated data from $N = 100$ individuals with unequally spaced data, formulated such that a range of one to four measurement occasions could be missing between two observed time points. The missingness was generated completely at random (MCAR; Little and Rubin, 1987). Using a signal-to-noise ratio of $q = 0.3$ for each individual, we generated each person's time series sequentially until 10 non-missing observations were obtained. This resulted in time series of different lengths (ending T including missing and non-missing occasions ranged from 20 to 42) and with different percentages of missingness (percentage of missingness within individuals ranged from 50% to 76%). As a whole, the ratio between missing observations and total (including missing and non-missing) observations collapsed across all individuals was 0.6. Plots of the complete true scores with irregularly spaced missingness are shown in Figure 1.

The FIS was used with the cubic spline model to yield true score estimates of the time series, including the portions for which actual observations were unavailable. Prediction error decomposition was used to estimate q, the signal-to-noise ratio appearing explicitly in the cubic spline model. The true score estimates obtained using the FIS are shown in Figure 2. The cubic spline model and its ability to capture local changes make it particularly suited as a tool for representing systematic or non-systematic trends associated with group-based changes. When used in conjunction with the FIS, the cubic spline model served as a flexible and effective 'stepping stone' for interpolating the gaps in unequally spaced data even with 60% of total missingness in the data. In our next example, we show that the cubic spline model can be used as an effective data interpolation tool in empirical data whose parametric form of change is unknown.

Properties of the prediction error decomposition when used in conjunction with panel data were examined using Monte Carlo simulations. We simulated data with

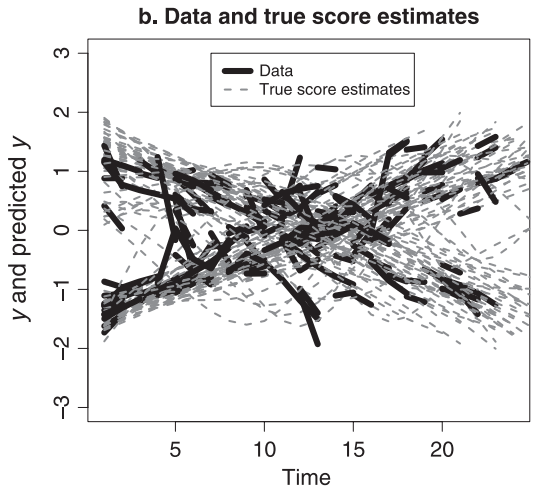

Fig. 1. A plot of the true scores of 100 simulated trajectories and the associated irregularly spaced observations.

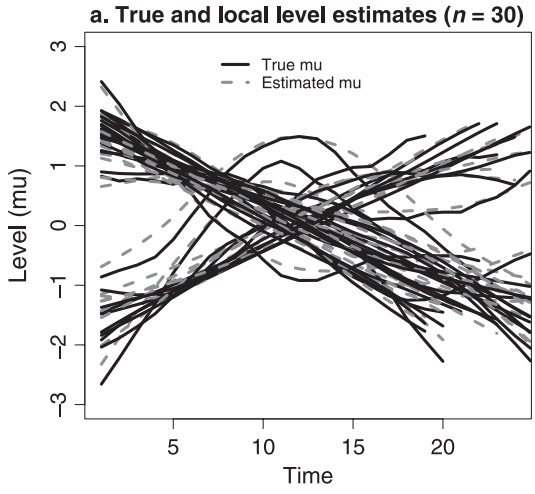

Fig. 2. A plot of the estimated true scores trajectories and the corresponding observed data from 30 randomly selected subjects.

different numbers of non-missing measurement occasions and subjects, including $T = 8$ and 15 non-missing occasions for each subject, and with $N = 50$, 100 and 200 participants. The results for each condition were based on estimates obtained from 10,000 Monte Carlo replications so as to minimize the possibility of random sampling error. Three observations can be made regarding the mean and the variability of the signal-to-noise estimates across the Monte Carlo runs. First, the mean values of the maximum likelihood estimates were close to the true value of $q = 0.3$ under all

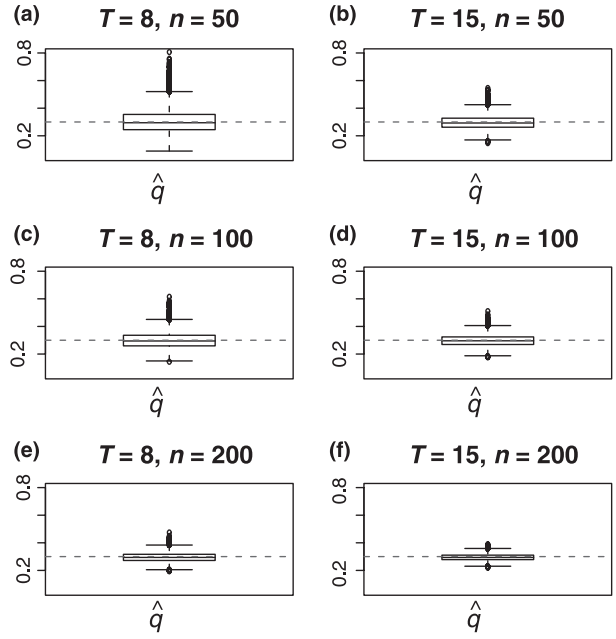

Fig. 3. Boxplots of the signal-to-noise ratios estimated by means of the prediction error decomposition with (a) $T=8$, $N=50$, (b) $T=15$, $N=50$, (c) $T=8$, $N=100$, (d) $T=15$, $N=100$, (e) $T=8$, $N=200$ and (f) $T=15$, $N=200$. The true value of q was set to 0.3 in all simulations and this value is marked with a dashed line in each plot.

six conditions, suggesting that the maximum likelihood (ML) estimator is essentially unbiased (see Figure 3).

Second, when $T=8$, the ratio between the true standard errors (namely, the standard deviation in $\hat{q}$ across Monte Carlo runs within each condition) corresponding to $N=50$ and $N=100$ was $0.086/0.059=1.458$. This value was close to, but slightly higher than the value $\sqrt{100/50}=1.414$, namely, the amount of reduction in variability to be expected based on the amount of increase in sample size (Casella and Berger, 2001). This indicated that approximately the same amount of precision was gained than was expected by increasing N from 50 to 100. The ratio between the true standard deviations for $N=100$ and $N=200$ was 1.439, which was even closer to the theoretical value of $\sqrt{200/100}=1.414$). When $T=15$, the ratio between the standard errors with $N=50$ and $N=100$ was 1.441, and the ratio between the standard errors with $N=100$ and $N=200$ was 1.417. In short, when $T=8$ or 15, the gain in precision by increasing N from 100 to 200 was in line with expected gain. When $N=50$, unbiased estimates were obtained on average, but the estimates showed a slight trend towards disproportionately lower precision.

Third, when $N=50$, the ratio between the standard errors at $T=8$ and $T=15$ was 1.755; when $N=100$, the ratio between the standard errors at $T=8$ and $T=15$ was 1.735; when $N=200$, the corresponding ratio was 1.708. The ratios in all three cases were larger than $\sqrt{15/8}=1.369$. In other words, increasing N from 50 to 200

Table 2. Summary statistics of the signal-to-noise ratio estimates obtained across 10,000 Monte Carlo runs.

T	N	Mean MC $\hat{q}$	SD MC $\hat{q}$	Range $\widehat{SE}$
8	50	0.306	0.086	0.028–0.250
8	100	0.300	0.059	0.060–0.081
8	200	0.298	0.041	0.045–0.055
15	50	0.297	0.049	0.024–0.091
15	100	0.296	0.034	0.028–0.034
15	200	0.295	0.024	0.018–0.022

Notes: The true value of q was set to 0.30 in all simulations.
T, number of non-missing occasions per participant; N, total number of participants; mean MC $\hat{q}$, average $\hat{q}$ across Monte Carlo runs; SD MC $\hat{q}$, empirical standard deviation of $\hat{q}$ across Monte Carlo runs; range $\widehat{SE}$, range of asymptotic standard errors given by the observed Fisher information matrix.

resulted in reductions of estimation standard error by about half in all instances as expected, but increasing T from 8 to 15 reduced the estimation standard error by approximately 1.7 times. This is not surprising as the specification of diffuse prior information for the state can give rise to large innovation values for the first few measurement occasions. Including additional measurement occasions can thus be helpful in down-weighting the inflation in variability in the first few measurement occasions, especially when T is small in the first place. Thus, increasing the number of time points within the constraints of typical panel designs can often help improve estimation precision, at least in the context of the present simulations.

We also compared the asymptotic standard error estimates based on the observed Fisher information matrix with the 'true' variability in $\hat{q}$ as indicated by the empirical standard deviations across Monte Carlo runs. Specifically, we computed standard errors associated with q by taking the square-root of the inverse of the numerical negative second derivative evaluated at $\hat{q}$. All pertinent information is summarized in Table 2. When $N = 50$, the asymptotic standard errors clearly overestimated the true standard errors for both $T = 8$ and $T = 15$, particularly when $T = 8$. When N was as small as 50, the parameter estimates for $\hat{q}$ were unbiased but the variability around the estimates was high, as indicated by the large values of the true standard errors. In this case, increasing T from 8 to 15 led to a substantial increase in precision. The asymptotic standard errors became closer to the true standard errors as N increased. Currently, we are exploring the feasibility of using a modified bootstrap approach based on that proposed by Stoffer and Wall (1991) to obtain more accurate standard error estimates in the context of panel data with small to moderate sample sizes and time lengths.

5.2 Example 2: Application of the cubic spline model to epidemiological data

As illustrated in the previous example, the nonparametric nature of the cubic spline model affords researchers a way to represent trends or patterns of intra-individual

change that do not necessarily conform to a prespecified form (linear, exponential, etc.). Either the interpolated time series can be used as input data for fitting other dynamic models, or the cubic spline model can be combined with other time-invariant or time-varying models to capture the fluctuations around a relatively smooth trend. In the present example, we used an epidemiological data set from a follow-up study to the original Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. The cubic spline model is particularly suited for use with this data set because the intervals for successive assessments were random both within and across individuals. A brief theoretical introduction to this data set is first provided before we proceed to presenting other modelling details.

Primary biliary cirrhosis is a liver disease that slowly causes liver failure. Albumin is a protein produced by the liver and is a good marker of liver functioning. Decreased serum albumin concentrations are often associated with serious damage to the liver. From 1974 to 1984, 424 PBC patients participated in a longitudinal study in the Mayo Clinic to determine the effectiveness of the drug D-penicillamine (for details, see Appendix D of Fleming and Harrington, 1991; Murtaugh *et al.*, 1994). This example was aimed at representing changes in serum albumin concentrations as the disease progressed.

Of the total 424 PBC patients, the first 312 cases in the data set participated in the randomized trial and contained largely complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded. We used repeated assessment data of patients in the randomized trial for illustrative purposes. Retaining only participants who contributed at least three measurement occasions, the total number of measurement occasions available from each individual ranged from three to 16, with a total of 259 participants. In addition, two participants showed an unusual elevation in albumin level on one measurement occasion (i.e. 6.82 g/100 ml and 8.01 g/100 ml, compared with the normal range of 3.4–5.4 g/100 ml, and the decreased albumin range of 1.17–4.64 g/100 ml within this sample). Elevated albumin levels are typically associated with dehydration and congestive heart failure (see for instance, Corti *et al.*, 1994). Data from these two participants on these occasions were therefore discarded because of the possible complications by other confounds. The remaining participants' standardized (i.e. within individuals over time) albumin levels were used for all subsequent model fitting. Plots of the participants' standardized albumin concentrations as grouped by drug condition are shown in Figure 4.

An autoregressive component of order 1 [i.e., an AR(1) model] was added to the stochastic differential equation model in Equation (1). The associated continuous-time dynamic model is expressed as

$$\mathrm{d}\begin{bmatrix}\mu_i(t)\\ \beta_i(t)\\ \alpha_i(t)\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 0\\ 0 & 0 & \phi_1\end{bmatrix}\begin{bmatrix}\mu_i(t)\\ \beta_i(t)\\ \alpha_i(t)\end{bmatrix}\mathrm{d}t + \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}0\\ \mathrm{d}w_{1i}(t)\\ \mathrm{d}w_{2i}(t)\end{bmatrix}, \tag{27}$$

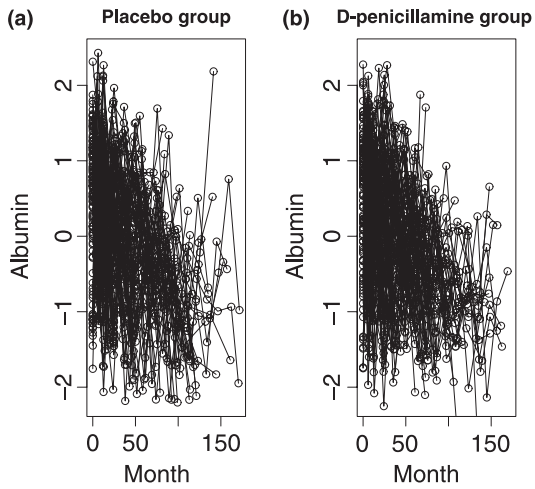

Fig. 4. Within-person standardized changes in albumin concentration level over time in (a) the placebo group and (b) the treatment group.

where $\alpha_i(t)$ is a continuous-time autoregressive process of order 1 [i.e. AR(1) process] used to accommodate additional variability in the participants' albumin levels, and $w_{2i}(t)$ is a Wiener process associated with the AR(1) process. The parameter ϕ_1 is the associated AR coefficient whose magnitude, if negative, determines how promptly previous deviations from a baseline dissipate over time. In the present context, the 'baseline' was essentially each person's systematic trend as constituted by the cubic spline model. The time scale in this case was constructed in months such that $\delta_{tk_i} = 1$ represents a 1-month interval between measurement occasions t_{ki} and $t_{k+1,i}$. The measurement intervals in the study ranged from 1.6 to 70.2 months.

Discrete-time representation of Equation (27) can be derived based on Equation (7) by using the matrix exponential function. Note that the transition matrix in this specific model is not nilpotent. Instead of using the power series to approximate the matrix exponential function, one common approach, if the transition matrix is diagonalizable, is to compute the matrix exponential function as $\exp(A) = v \exp[\mathrm{diag}(D)]v^{-1}$, where $\mathrm{diag}(D)$ is a diagonal matrix containing the eigenvalues of A, v is a matrix whose columns are the right eigenvectors of A, and the exponential function on the right-hand-side of the equation involves computing element-by-element exponents of the diagonal entries of D. A general approach constructed along this line is detailed by Harvey (2001, pp. 503–505). This was the approach adopted here. The resultant matrix exponential of the transition matrix is defined as

$$\exp\left(\delta_{t_{ki}}\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 0\\ 0 & 0 & \phi_1\end{bmatrix}\right) = \begin{bmatrix}1 & \delta_{t_{ki}} & 0\\ 0 & 1 & 0\\ 0 & 0 & e^{\phi_1\delta_{t_{ki}}}\end{bmatrix}. \tag{28}$$

The discrete-time analogue of the dynamic model is thus expressed as

$$\begin{bmatrix} \mu_{i,\,t_{k+1}} \\ \beta_{i,\,t_{k+1}} \\ \alpha_{i,\,t_{k+1}} \end{bmatrix} = \begin{bmatrix} 1 & \delta_{t_{ki}} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \mathrm{e}^{\phi_1 \delta_{t_{ki}}} \end{bmatrix} \begin{bmatrix} \mu_{i,\,t_k} \\ \beta_{i,\,t_k} \\ \alpha_{i,\,t_k} \end{bmatrix} + \begin{bmatrix} \zeta_{\mu,\,it_k} \\ \zeta_{\beta,\,it_k} \\ \zeta_{\alpha,\,it_k} \end{bmatrix}. \tag{29}$$

The process noise vector $[\zeta_{\mu,\,it_k}\ \zeta_{\beta,\,it_k}\ \zeta_{\alpha,\,it_k}]'$ is multivariate normally distributed with a mean vector of zeros and covariance matrix given by

$$\begin{aligned} \Psi_{it_k} &= \int_0^{\delta_{t_{ki}}} \begin{bmatrix} 1 & \delta_{t_{ki}} - s & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \mathrm{e}^{\phi_1(\delta_{t_{ki}} - s)} \end{bmatrix} I_3 \begin{bmatrix} 0 & 0 & 0 \\ 0 & \sigma^2_{w_1} & 0 \\ 0 & 0 & \sigma^2_{w_2} \end{bmatrix} \\ &\quad \times I_3 \begin{bmatrix} 1 & 0 & 0 \\ \delta_{t_{ki}} - s & 1 & 0 \\ 0 & 0 & \mathrm{e}^{\phi_1(\delta_{t_{ki}} - s)} \end{bmatrix} \mathrm{d}s \\ &= \int_0^{\delta_{t_{ki}}} \begin{bmatrix} (\delta_{t_{ki}} - s)^2 \sigma^2_{w_1} & (\delta_{t_{ki}} - s)\sigma^2_{w_1} & 0 \\ (\delta_{t_{ki}} - s)\sigma^2_{w_1} & \sigma^2_{w_1} & 0 \\ 0 & 0 & \sigma^2_{w_2} e^{2\phi_1(\delta_{t_{ki}} - s)} \end{bmatrix} \mathrm{d}s \\ &= \begin{bmatrix} \frac{\sigma^2_{w_1}\delta^3_{t_{ki}}}{3} & \frac{\sigma^2_{w_1}\delta^2_{t_{ki}}}{2} & 0 \\ \frac{\sigma^2_{w_1}\delta^2_{t_{ki}}}{2} & \sigma^2_{w_1} & 0 \\ 0 & 0 & \frac{\sigma^2_{w_2}(\mathrm{e}^{2\phi_1\delta_{t_{ki}}} - 1)}{2\phi_1} \end{bmatrix}. \end{aligned} \tag{30}$$

where $\sigma^2_{w_1}$ and $\sigma^2_{w_2}$ denote the rates of variance change associated with $\mathrm{d}w_{1i}(t)$ and $\mathrm{d}w_{2i}(t)$ respectively. Notice from Equation (30) that the variance for $\zeta_{\alpha,\,it_k}$ is a constrained function of the AR(1) coefficient, ϕ_1, in addition to $\sigma^2_{w_2}$ and $\delta_{t_{ki}}$.

The measurement model used in the present example is expressed as

$$y_{it_k} = b * isStage34_{it_k} + [1 \quad 0 \quad 1] \begin{bmatrix} \mu_{it_k} \\ \beta_{it_k} \\ \alpha_{i,\,t_k} \end{bmatrix} + \epsilon_{it_k}. \tag{31}$$

On each measurement occasion, each participant was coded as being in one of four distinct stages of the disease, ranging from 1 = lowest severity to 4 = highest severity. The covariate $isStage34_{it_k}$ is a time-varying dummy code which, if equal to 1, indicates that person i was in one of the more severe stages (stage 3 or 4) of the disease at time t. The term ϵ_{it_k} represents measurement error that only influences a specific measurement occasion, t_{k_i}. In short, the manifest albumin level for person i at time t_k, y_{it_k}, is now represented as a linear combination of a local level component (through which the local slope exerts an indirect influence), dynamic variability around the local level which fluctuates over time according to a continuous-time AR(1) process, and measurement error which shows no continuity across measurement occasions. To aid estimation, the measurement error variance was fixed at 1.0 and the parameter vector $\theta = [\phi_1, \sigma^2_{w_1}, \sigma^2_{w_2}, b]$ estimated by means of the prediction error decomposition procedure. With the addition of the AR(1) component, the ratio $\sigma^2_{w_1}/\sigma^2_e$ is no longer a pure signal-to-noise ratio. We therefore denote this

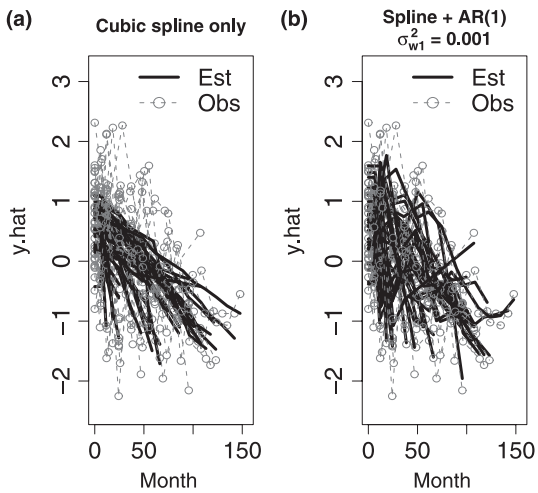

Fig. 5. Plots of 35 randomly selected participants' estimated measurements, $\hat{y}_{it_k}$, based on (a) the cubic spline model, (b) the cubic spline model with the addition of an AR(1) component and the time-varying covariate, $isStage34_{it_k}$, which served as a marker of whether these participants were in the more severe stage of the disease. Est, estimated trajectories; Obs, observed trajectories.

ratio below using the more general notation of σ^2_{w1}, instead of q as in our earlier simulation.

The difference in albumin level between the placebo and treatment groups over time was minimal ($M_{\text{placebo}} = 3.39$, $\text{SD}_{\text{placebo}} = 0.48$; $M_{\text{drug}} = 3.40$, $\text{SD}_{\text{placebo}} = 0.48$), suggesting the lack of effect of the drug D-penicillamine. We thus performed all subsequent model fitting on the pooled data of both groups. We began by fitting a model with only the cubic spline to the pooled data. This yielded a process noise variance estimate that was approximately zero, $\hat{\sigma}^2_{w1} \approx 0.000$ (SE = 0.000). The estimated observations, $\hat{y}_{it_k}$, of 35 randomly selected participants are plotted in Figure 5a. It can be seen from the plot that the near-zero smoothing parameter gave rise to deterministic downward declines in the participants' albumin levels.

We then proceeded to fitting the full model with the AR(1) component and the time-varying covariate, $isStage34_{it_k}$. Instead of estimating $\hat{\sigma}^2_{w1}$, however, we fixed it to a small value, 0.001, which was close to the estimated $\hat{\sigma}^2_{w1}$ obtained earlier from fitting the cubic spline-only model. Based on this model, we obtained an AR(1) estimate of $\hat{\phi}_1 = -2.53$ (SE = 2.18), a process noise variance for the AR(1) process of $\hat{\sigma}^2_{w2} = 0.98$ SE = 1.48 and a regression coefficient of $\hat{b} = -0.04$ (SE = 0.01). The resultant asymptotic standard errors indicated that all the parameter estimates, except for $\hat{b}$, were not reliably different from zero. The value of $\hat{b}$ indicated that on average, being in stage 3 or 4 of the disease was associated with a significantly lower albumin level over and above an individual's general decline in albumin level as captured by the cubic spline. The estimated observations of the same 35 randomly selected participants, now based on the full model, are plotted in Figure 5b.

Results from model fitting suggested that the AR(1) parameter, ϕ_1, was not significantly different from zero. Setting this parameter to zero yields an interesting picture on the participants' variability in albumin level, however. Specifically, setting ϕ_1 to zero in Equation (29) would turn the equation for α_{it_k} into a random walk model. In other words, when an individual's albumin level deviated from its overall decline trajectory, his/her albumin level might continue to deviate more from this trajectory rather than returning to it. This was the case for some, but not all of the participants. The relatively large intervals between successive measurement occasions (ranging from 1.6 months to 6 years) might have contributed to this deviation pattern as well. With more individualized information, other more appropriate models can be used to represent how and why some individuals' albumin levels deviated from their overall decline trajectories but others' did not.

In sum, the present example serves to illustrate the strengths and flexibility of the cubic spline model in summarizing group-based dynamics as well as interindividual differences in trends. We showed that on the one hand, inferences regarding group-based dynamics can be made on the basis of the person-invariant parameters, including $\hat{\phi}_1$ and $\hat{b}$. On the other hand, sufficient flexibility for representing interindividual differences in change is accommodated through the use of the nonparametric cubic spline model. When appropriate, other dynamic models can be constructed and combined with the cubic spline model to better suit the theoretical framework of interest to a researcher. Alternatively, the cubic spline model and its variations can be used as a flexible missing data interpolation tool.

6 Discussion

In the last few decades, a growing number of social scientists, particularly sociologists, have turned to differential equations for exploring new ways of capturing the dynamics of social processes. Although the use of differential equations is by no means novel in the realm of physical sciences, and has long been advocated by some social scientists (e.g. Coleman, 1968), the increased popularity of this approach marks an important milestone for other researchers, as they are beginning to re-conceptualize the unfolding of a social phenomenon as continuous in nature.

For model-fitting purposes, we capitalized on functions implemented in the *SSfPack* package to perform continuous-time modelling of unequally spaced epidemiological data. (Ox codes for all the model-fitting examples considered in the present paper, including codes documenting how linear and nonlinear constraints are imposed to implement different variations of the cubic spline model are available from the first author's website.) Because Ox basically consists of different libraries of C routines and its particular syntax structure deviates slightly but is still close to C, it provides a very flexible programming language for implementing many statistical and mathematical algorithms. We showed in the present context that the language structure of Ox is flexible enough to accommodate all the nonlinear and time-varying constraints in the different cubic spline models.

Results from our Monte Carlo simulations indicated that standard error estimates based on asymptotic results overestimated the true variability of the model parameters when sample sizes are small, more notably when $N = 50$. However, the parameter estimates were still unbiased. As N increased, the asymptotic standard errors converged to the true empirical standard errors as expected. Furthermore, in the designs that we considered, increasing the number of measurement occasions from eight to 15 when N is small to moderate helped improve the precision of the estimates more so than increasing N. It is thus beneficial for researchers to consider increasing T within the constraints of their panel designs. To this end, bootstrap approaches (e.g. Stoffer and Wall, 1991; Efron and Tibshirani, 1993) may also provide a practical alternative for examining the finite-sample properties of model parameters of interest. Efforts to adapt some of the existing bootstrap approaches for use with panel data that are characterized by irregularly spaced intervals are currently underway by the authors of this article.

In our empirical example, the cubic spline model was used to represent the changes in albumin levels among PBC patients. Here, the cubic spline model was used to extract the overall declines in albumin level whereas a continuous-time AR(1) model was used to capture the variability around the trends. Results from model fitting suggested that on average, when the participants' albumin levels did show deviations from their overall decline trends, the deviations tended to persist, rather than diminish exponentially over successive measurement occasions. Of course, our interpretation of 'trend' in this case was somewhat subjective because the cubic spline model can also be used to accommodate non-monotonic changes and consequently, the nonstationary deviation pattern noted here. However, we limited ourselves to using a relatively small value of σ^2_{w1} in conjunction with the time-varying covariate, $isStage34_{it_k}$, so as to extract relatively smooth downward decline trends. Our results can serve as a stepping stone to derive other more theoretically meaningful representations of trends. More individual-based information is certainly helpful as well in explaining the nonstationary deviation pattern seen in some patients but not others.

In sum, we showed that the cubic spline model in constrained discrete form provides a flexible, data-driven model for interpolating missing observations even (and especially) when the choice of a theoretical model is not immediately clear. This approach is practical and is well suited as a tool for representing changes that are less structured. This model is also useful as a way to combine gradual patterns of intra-individual change with relatively short-lived fluctuations in the form of *state fluctuations*. Such models provide a convenient way to depict processes that span different time scales and they have direct applications, e.g. in the study of ageing, wherein researchers have long sought to isolate irreversible developmental changes from short-term (e.g. day-to-day) performance variability (e.g. Lindenberger and Oertzen, 2006). This is particularly relevant for the study of human development, as the gains and losses that occur across different time scales can serve very different roles (Newell, Liu and Meyer-Kress, 2001).

As a whole, models that can readily incorporate time-varying intervals (i.e. $\delta_{t_{ki}}$) provide a direct way to handle irregularly spaced data, and more importantly, to evaluate the timing of change in data that are either equally or unequally spaced. Often, the theoretical importance of the timing of change in a process can be as critical as the occurrence of change itself. In fact, some researchers have attempted to formulate longitudinal models that emphasize not only the magnitude, but also the timing of change. For instance, the proposition of Davidson (1998) regarding the 'rise time' of emotion places great emphasis on how the rise time of an individual's emotion can reveal signs of emotional dysfunctions. The timing of change is thus critical in this case, and perhaps, in any study of intra-individual change. Continuous-time modelling tools help free researchers from the burden of collecting equally spaced longitudinal data.

Finally, we emphasize that discrete-time models undoubtedly have their own strengths in modelling events in which the timing is ignored or predetermined. Unfortunately, they have often been misused in cases where continuous-time models are indeed the more appropriate modelling tool (c.f. Sorenson, 1978) and are sometimes chosen for no clear theoretical reason. In this case, the underlying dynamics of a process can be heavily distorted or even masked by the artifacts of a discrete sampling process, or an inappropriately spaced discrete-time model. To this end, we see continuous-time models as a way of supplementing other existing data analytic tools, rather than as a marker for distinguishing continuous processes from discrete ones. As asserted by Arminger (1986), continuous-time models have their own strengths in capturing the interdependence among a system of variables, and unlike difference equations, are not limited in applications to processes of equally spaced time intervals. We hope to have presented the reader with a practical approach for fitting stochastic differential equations in exact discrete form to irregularly spaced data. Ultimately, the question of choice between continuous-and discrete-time models should be a question of one's theory and the nature of the data at hand.

Acknowledgements

We would like to thank two anonymous reviewers and the two guest editors for their thoughtful comments on earlier versions of this paper. We have also benefited from discussions with our colleagues Ke-Hai Yuan and Ji-Yun Zu. Selected OxMetrics codes used for model fitting in this article can be downloaded from the first author's website at http://www.unc.edu/~symiin.

Appendix A The Kalman filter, the fixed interval smoother and prediction error decomposition

A.1 The Kalman filter

We outline here the basic KF algorithm implemented within the *SSfpack* framework with a diffuse initial condition. For generality, all parameter vectors or matrices are

marked with a person index but they can be person-invariant depending on one's model. The steps involved are as follows:

step 1. Set i to 1. At $t=0$, set $\eta_{i,0|0}$ to a $w\times 1$ vector of zeros and $P_{0|0}=\kappa I_w$, where κ is a large positive scalar and I_w is a $w\times w$ identity matrix.

step 2. Person i's state estimates (e.g. factor scores) at time t_{k+1}, the associated covariance matrix and several by-products are computed as

$$\begin{bmatrix} \bar{\eta}_{i,t_{k+1}} \\ \hat{y}_{it_k} \end{bmatrix} = d_{it_k} + \Phi_{it_k}\eta_{i,t_k|t_{k-1}}, \tag{32}$$

$$\begin{bmatrix} \bar{P}_{i,t_{k+1}} & M_{it_k} \\ M'_{it_k} & F_{it_k} \end{bmatrix} = \Phi_{it_k}P_{i,t_k|t_{k-1}}\Phi'_{it_k} + \Omega_{it_k}, \tag{33}$$

$$K_{it_k} = M_{it_k}F^{-1}_{it_k}, \tag{34}$$

and

$$M_{it_k} = B_{t_k}P_{t_k|t_{k-1}}\Lambda'_{it_k} + \mathrm{cov}(\zeta_{it_k}, \epsilon_{it_k}). \tag{35}$$

The one-step-ahead estimates $\eta_{i,t_{k+1}|t_k}$ and $P_{i,t_{k+1}|t_k}$ for time t_{k+1} are then derived based on manifest information available at time t_{k+1} as

$$v_{it_k} = y_{it_k} - \hat{y}_{it_k} = y_{it_k} - \tau_{it_k} - \Lambda_{it_k}\eta_{t|t-1}, \tag{36}$$

$$\eta_{i,t_{k+1}|t_k} = \bar{\eta}_{i,t_{k+1}} + K_{it_k}v_{it_k}, \tag{37}$$

$$p_{i,t_{k+1}|t_k} = \bar{P}_{t_{k+1}} - K_{it_k}M_{it_k}. \tag{38}$$

The element v_{it_k} is often referred to as the innovation (vector) for person i at time t_k and F_{it_k} is the associated innovation or 'prediction error' covariance matrix. Note, however, that K_{it_k} in Equations (34) is somewhat different from what is typically referred to as the gain matrix in the KF literature, denoted herein as $Gain_{it_k}$. $\bar{\eta}_{t_{k+1}}$ and $\bar{P}_{t_{k+1}}$ can be viewed as intermediate state estimates and their associated covariance matrix because they are composed of estimates for time t_{k+1} based on observation from up to time t_{k-1}. This specific formulation combines a few computational steps such that $\eta_{t_k|t_k}$, $P_{t_k|t_k}$ are calculated as part of Equation (37) and Equation (38), respectively, and once completed, the one-step-ahead estimates $\eta_{t_{k+1}|t_k}$ and $P_{t_{k+1}|t_k}$ are now available. In addition, combining the dynamic and the measurement equations into one single model has the advantage of allowing $\mathrm{cov}(\zeta_{it_k}, \epsilon_{it_k})$ to be estimated, when deemed necessary.

step 3. Set t_k to t_{k+1} and repeat step (2) until $t_k = T$. Then repeat steps 1—3 for $i = i+1, \ldots N$.

A.2 The fixed interval smoother

Once the KF is completed, an additional backward smoothing can be used to yield further refine the state estimates using, e.g. the FIS. Many subsequent researchers

considered reformulated versions of the classical FIS algorithm DURBIN and KOOPMAN (2001), which primarily involve reorganizations of the terms to minimize the number of matrix inversions that have to be performed to reduce computational costs (e.g. BRYSON and HO, 1969; KOHN and ANSLEY, 1989). One such version, considered in detail by KOOPMAN (1993), is readily available in *SSfpack* through the function *KalmanSmo*. This particular smoother is structured by defining additional components, denoted below as L_{it_k}, r_{it_k} and N_{it_k}, in a backward recursion for $t_k = t_{T-1}, \ldots, t_1$ with

$$L_{it_k} = B_{it_k} - K_{it_k}\Lambda_{it_k}, \tag{39}$$

$$r_{i,\,t_{k-1}} = P^{-1}_{i,\,t_k|t_{k-1}}(\eta_{i,\,t_k|t_T} - \eta_{i,\,t_k|t_{k-1}}), \tag{40}$$

$$= \Lambda'_{it_k} F^{-1}_{it_k} v_{it_k} + L'_{it_k} r_{i,\,t_k}, \tag{41}$$

$$N_{i,\,t_{k-1}} = \Lambda'_{it_k} F^{-1}_{it_k} \Lambda_{it_k} + L'_{it_k} N_{it_k} L_{it_k}. \tag{42}$$

The recursion is initiated by first setting $r_{it_T} = 0$ and $N_{it_T} = 0$. The smoothed state vector and its associated smoothed covariance matrix can then be derived using r_{it_k} and N_{it_k} as

$$\eta_{i,\,t_k|t_T} = \eta_{it_k|t_{k-1}} + P_{i,\,t_k|t_{k-1}} r_{i,\,t_{k-1}} \tag{43}$$

and

$$P_{i,\,t_k|t_T} = P_{i,\,t_k|t_{k-1}} - P_{i,\,t_k|t_{k-1}} N_{i,\,t_{k-1}} P_{i,\,t_k|t_{k-1}}. \tag{44}$$

The primary appeal of formulating the FIS this way is to avoid the need to invert the covariance matrix, $P_{i,\,t_k|t_{k-1}}$, $T-1$ times, which is a necessary step in the classical version of the smoother. Computing $F^{-1}_{it_k}$ is the only matrix inversion operation involved, but $F^{-1}_{it_k}$ is already computed as part of the KF implementation (see Equation 34).

A.3 Parameter estimation in state–space models via maximum likelihood

As the KF is cycling through the estimation across and T time points and N persons, several by-products from the KF can be substituted into a prediction error decomposition function (SCHWEPPE, 1965; CAINES and RISSANEN, 1974; HARVEY, 1989) to yield maximum-likelihood (ML) estimates of all time–invariant parameters. Theoretically speaking, time-varying parameters can also be estimated using the prediction error decomposition approach if data from multiple measurement occasions are incorporated into the measurement vector of one particular time point and the state-space model is reformulated accordingly.

The log-likelihood function can be written as a function of the individual innovation vector, v_{it_k}, and its associated covariance matrix, F_{it_k}, defined earlier, yielding

$$\log f(\theta) = \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} -p_{it_k} \log(2\pi) - \log |F_{it_k}| - v'_{it_k} F_{it_k}^{-1} v_{it_k}, \tag{45}$$

where p_{it_k} is the number of complete manifest variables at time t_k for person i. The person index in p_{it_k} indicates that a person i may have missing responses on certain variables at a particular time t_k. Maximizing Equation (45) with respect to all time- and person-invariant parameters in Λ_{it_k}, τ_{it_k}, B_{it_k}, α_{it_k}, Ψ_{it_k} and Θ_{it_k}, if any, results in ML estimates of these parameters.

Notes

1. One important property of a Wiener process is that its paths are not differentiable with respect to time. Thus, integration of such stochastic equations constitutes a topic of extensive research in itself (Arnold, 1974; Higham, 2001.)
2. In the present context, given $A = \begin{bmatrix} 0 & \delta_{t_{ki}} \\ 0 & 0 \end{bmatrix}$ is nilpotent with $A^2 = \mathbf{0}$, the power series terminates after the first two terms. $A^* = \begin{bmatrix} 1 & \delta_{t_{ki}} \\ 0 & 1 \end{bmatrix}$ thus gives an exact expression of the matrix exponential function in Equation (12), namely, $\exp\left(\delta_{t_{ki}} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\right)$.

References

Anderson, B. D. O. and J. B. Moore (1979), *Optimal filtering*, Prentice Hall, Englewood Cliffs, NJ.

Arminger, G. (1986), Linear stochastic differential equation models for panel data with unobserved variables, in: N. Tuma (ed.), *Sociological methodology 1986*, Jossey-Bass, San Francisco, CA, 187–212.

Arnold, L. (1974), *Stochastic differential equations*, Wiley, New York.

Bergstrom, A. R. (1988), The history of continuous-time econometric models, *Econometric Theory* **4**, 365–383.

Bryk, A. S. and S. W. Raudenbush (1987), Application of hierarchical linear models to assessing change, *Psychological Bulletin* **101**, 147–158.

Bryson, A. E. and Y. C. Ho (1969), *Applied optimal control, optimization, estimation and control*, Blaisdell Publishing Company, Waltham.

Caines, P. E. and J. Rissanen (1974), Maximum likelihood estimation of parameters, *IEEE Transactions on Information Theory* **IT-20**, 102–104.

Casella, G. and R. L. Berger (2001), *Statistical inference*, 2nd edn, Duxbury Press, Pacific Grove, CA.

Coleman, J. S. (1968), The mathematical study of change, in: H. M. Blalock Jr and A. Blalock (eds.), *Methodology in social research*, McGraw-Hill, New York, 428–478.

Corti, M. C., J. M. Guralnik, M. E. Salive and J. D. Sorkin (1994), Serum albumin level and physical disability as predictors of mortality in older persons, *Journal of the American Medical Association* **272**, 1036–1042.

Davidson, R. J. (1998), Affective style and affective disorders: perspectives from affective neuroscience, *Cognition and Emotion* **12**, 307–330.
Doornik, J. A. (1998), *Object-oriented matrix programming using Ox 2.0*, Timberlake Consultants Press, London.
Durbin, J. and S. J. Koopman (2001), *Time series analysis by state–space methods*, Oxford University Press, New York.
Efron, B. and R. Tibshirani (1993), *An introduction to the bootstrap*, Chapman & Hall, New York.
Fleming, T. R. and D. P. Harrington (1991), *Counting processes and survival analysis*, Wiley, New York.
Hamerle, A., W. Nagl and H. Singer (1990), Problems with the estimation of stochastic differential equation using structural equations models, *Journal of Mathematical Sociology* **16**, 201–220.
Harvey, A. C. (1989), *Forecasting, structural time series models and the Kalman filter*, Princeton University Press, Princeton, NJ.
Harvey, A. C. (2001), *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press, Cambridge.
Hawkley, L. C., M. H. Burleson, G. G. Berntson and J. T. Cacioppo (2003), Stress in everyday life: cardiovascular activity, psychosocial context and health behaviors, *Journal of Personality and Social Psychology: Personality and Individual Differences* **85**, 105–120.
Higham, D. J. (2001), An algorithmic introduction to numerical simulation of stochastic differential equations, *SIAM Review* **43**, 525–546.
Kohn, R. and C. F. Ansley (1987), A new algorithm for spline smoothing based on smoothing a stochastic process, *SIAM Journal of Scientific and Statistical Computing* **8**, 33–48.
Kohn, R. and C. F. Ansley (1989), A fast algorithm for signal extraction, influence and cross-validation, *Biometrika* **76**, 65–79.
Koopman, S. J. (1993), Disturbance smoother for state–space models, *Biometrika* **80**, 117–126.
Koopman, S. J., N. Shephard and J. A. Doornik (1999), Statistical algorithms for models in state space using ssfpack 2.2, *Econometrics Journal* **2**, 113–166.
Larsen, R. J. and T. Ketelaar (1991), Personality and susceptibility to positive and negative emotional states, *Journal of Personality ans Social Psychology* **61**, 132–140.
Laurenceau, J. P., L. F. Barrett and P. R. Pietromonaco (1998), The importance of self–disclosure partner discplore, and perceived partner responsiveness in interpersonal exchanges, *Journal of Personality and Social Psychology* **74**, 1238–1251.
Lindenberger, U. and T. V. Oertzen (2006), Variability in cognitive aging: from taxonomy to theory, in: F. I. M. Craik and E. Bialystok (eds.), *Lifespan cognition: mechanisms of change*, Oxford University Press, Oxford, 297–314.
Little, R. J. A. and D. B. Rubin (1987), *Statistical analysis with missing data*, Wiley, New York.
Murtaugh, P. A., E. R. Dickson, G. M. Van dam, M. Malinchoc, P. M. Grambsch, A. L. Langwirthy and C. H. Gips (1994), Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits, *Hepatology* **20**, 126–134.
Newell, K. M., Y. T. Liu and G. Meyer-Kress (2001), Time scales in motor learning and development, *Psychological Review* **108**, 57–82.
Ong, A. D., J. L. Horn and D. A. Walsh (2007), Stepping into the light: modeling the intra-individual dimensions of hedonic and eudaemonic well–being, in: A. D. Ong and M. H. M. Van Dulmen (eds.), *Oxford Handbook of Methods in Positive Psychology*, Oxford University Press, New York, 12–25.
Oud, J. H. L. and R. A. R. G. Jansen (2000), Continuous time state space modeling of panel data by means of SEM, *Psychometrika* **65**, 199–215.
Oud, J. H. L. (2002), Continuous time modeling of the cross-lagged panel design, *Kwantitatieve Methoden* **69**, 1–26.
Sbarra, D. A. (2006), Predicting the onset of emotional recovery following nonmarital relationship dissolution: survival analyses of sadness and anger, *Personality and Social Psychology Bulletin* **32**, 298–312.

Schweppe, F. (1965), Evaluation of likelihood functions for gaussian signals, *IEEE Transactions on Information Theory* **11**, 61–70.
Sorenson, A. B. (1978), Mathematical models in sociology, *American Review of Sociology* **4**, 345–371.
Stoffer, D. S. and K. D. Wall (1991), Bootstraping state-space models: Gaussian maximum likelihood estimation and the Kalman filter, *Journal of the American Statistical Association* **86**, 1024–1033.
Thompson, A. and N. Bolger (1999), Emotional transmission in couples under stress, *Journal of Marriage & the Family* **61**, 38–48.
Tsay, R. S. (2005), *Analysis of financial time series*, 2nd edn, Wiley Interscience, New Jersey.
Verbeke, G. and G. Molenberghs (2000), *Linear mixed models for longitudinal data*, Springer-Verlag, New York.
Wahba, G. (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regression, *Journal of the Royal Statistical Society, Series B (Methodological)* **40**, 364–372.
Wecker, W. E. and C. F. Ansley (1983), The signal extraction approach to nonlinear regression and spline smoothing, *Journal of the American Statistical Association* **78**, 81–89.

Received: December 2006. Revised: June 2007.