Original Article

Journal of Time Series Analysis

First Version received September 2011

Published online in Wiley Online Library: 14 March 2012

(wileyonlinelibrary.com) DOI: 10.1111/j.1467-9892.2012.00784.x

A state space model approach for HIV infection dynamics

Jiabin Wang[a], Hua Liang[b], and Rong Chen[a,*,†]

Mathematical models have been proposed and developed to model HIV dynamics over the past few decades, achieving remarkable gains. However, it remains a challenging problem to accurately and efficiently estimate all key parameters in these models from clinical data. In this article, we propose a state space approach to link these models with clinical data and estimate the corresponding parameters. Specifically, we approximate the target nonlinear differential equations by difference equations and add a stochastic perturbation in the dynamic of all cells and virions. The proposed model and methodology provide an alternative tool in HIV dynamic modeling and can be easily applied in other biomedical system characterized by dynamic systems. Simulation studies are conducted to compare the performance of the proposed model with some previous approaches and show superior performance. With moderate sample size, the approach is successful in estimating all HIV viral dynamic parameters without many constraints on the parameters. To illustrate the proposed method, we apply it on the clinical data of two individual HIV infected patients treated with antiretroviral therapies.

Keywords: Difference equation; Kalman filter; missing data; model selection; Runge–Kutta; varying-coefficient models.

1. INTRODUCTION

HIV infection dynamics have been developed and investigated by biomathematicians and biologists since the end of the 1980s (Merrill, 1987; Anderson and May, 1992; Perelson, Kirschner and Boer, 1993; Perelson and Nelson, 1999). One major breakthrough in the study of viral dynamics was to use simplified differential equation models to fit actual clinical data (Ho *et al.*, 1995; Wei *et al.*, 1995; Perelson *et al.*, 1997), which resulted in a revolution in our understanding of HIV pathogenesis. Such approaches made it possible to determine many quantitative features of the interaction between HIV, the virus that causes AIDS and the immune cells that are infected by the virus. Some other important findings on the behavior of HIV and its host cells were also obtained from recent viral dynamic studies (Mittler, 1997; Perelson *et al.*, 1997; Wu and Ding, 1999). Such approaches make it possible to evaluate the design of a trial, to identify any flaws in the design and to optimize the design for future studies. The results can also be used to conduct sensitivity analysis to ascertain the key factors that contribute most to the uncertainty in the results. The design may then be modified, or pilot studies conducted to narrow the range of plausible inputs to the model.

The main target cell of HIV virus is CD4+T cells. To study the effect of a certain HIV drug, a sequence of measurements are usually taken on viral loads and CD4+T cell counts after initialization of the therapy. Many different mathematical models have been proposed to explore HIV dynamics with drug effect. In this article we consider the following dynamic model Perelson and Nelson (1999) for patients under long term treatments:

$$\begin{aligned} \frac{dT_U}{dt} &= \lambda - \rho T_U - \gamma(t) T_U V, \\ \frac{dT_I}{dt} &= \gamma(t) T_U V - \delta T_I, \\ \frac{dV}{dt} &= N\delta T_I - cV, \end{aligned} \tag{1}$$

where T_U, T_I and V denote the concentration of uninfected target CD4+T cells, the concentration of infected cells and the virus load respectively; λ represents the rate at which new T-cells are created from sources within the body, such as the thymus; N is the number of new virions produced from each of the infected cells during their life-time; c is the death (clearance) rate of free virions; ρ is the death rate of uninfected T-cell; δ is the death rate of infected cells. The antiviral drug efficacy was characterized by the time-varying infection rate $\gamma(t)$. As argued by former researchers (Liang and Wu, 2008; Liang *et al.*, 2010), model provides a flexible yet

[a]Rutgers University
[b]University of Rochester Medical Center
*Correspondence to: Rong Chen, Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA.
†E-mail: rongchen@stat.rutgers.edu

simple approach for studying the long-term viral dynamics. From this basic model, several extensions have been proposed (Perelson and Nelsen, 1999; Nowak and May, 2000; Callaway and Perelson, 2002).

It is a non-trivial task to simultaneously estimate all parameters in this model, including the time-varying HIV viral dynamic parameters for individual patients. Liang and Wu (2008) and Liang et al. (2010) approximated $\gamma(t)$ with a spline function and estimate the model parameters by minimizing mean square error between the numerical solution of the system and actual clinical observations. In this article, we use the same spline approximation to $\gamma(t)$ and convert the system into a state-space model. This approach has the advantage of more model flexibility, noise incorporation and accurate statistical inferences.

A state-space model (SSM) describes the underlying process of the observed time-series data as being driven by some unobservable latent state variables. It consists of an observation equation and a state transitional equation. A standard linear state-space model with Gaussian noise is generally written as:

$$\begin{aligned} \mathbf{X}_t &= \mathbf{\Lambda}_t + \mathbf{F}_t\mathbf{X}_{t-1} + \mathbf{U}_t \\ \mathbf{Y}_t &= \mathbf{G}_t\mathbf{X}_t + \mathbf{W}_t \qquad \begin{pmatrix} \mathbf{U}_t \\ \mathbf{W}_t \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} Q & 0 \\ 0 & H \end{pmatrix} \right\}, \end{aligned} \tag{2}$$

where $\mathbf{X}_t$ is the unobserved state vector, $\mathbf{Y}_t$ is the observation vector, $\mathbf{\Lambda}_t$, $\mathbf{F}_t$ and $\mathbf{G}_t$ are matrices known at time t, except a set of parameters. Here the observation noise $\mathbf{W}_t$ and the state noise $\mathbf{U}_t$ are assumed to be normally distributed and independent. Nonlinear and non-Gaussian state space models have been extensively studied (Harvey, 1989; West and Harrison, 1997; Durbin and Koopman, 2001; Tsay, 2005), though they pose computational challenges.

In this article, we propose to convert model (1) into a corresponding varying-coefficient state-space model to simultaneously estimate all the parameters and the time-varying drug effect. This conversion step is very important in our study, especially when time-intervals are unequal as in most of clinical trials.

The advantages of our approach can be summarized in three major points: (i) our approach is computationally efficient since we effectively adopt Kalman filter in the evaluation of loglikelihood and optimization. Sensible starting values are generally available from previous study and biological knowledge, (ii) our model is very feasible and incorporates observational noises and random dynamics fluctuations in the model: It allows proper statistical inference and model estimation and validation. In addition, missing data can be easily handled and the model can be extended to more complex situations such as nonlinear non-gaussian SSM and (iii) our approach provides more accurate estimation than existing methods, as shown in the simulation study.

The rest of the paper is organized as follows. In Section 2, we present the SSM formulation for HIV dynamic model (1), with emphasis on the conversion from ordinary differential equation (ODE) to SSM and spline approximation to time-varying coefficients. In Section 3, we introduce estimation procedure using Kalman filer and discuss some computation aspects. Simulation studies are presented in Section 4 to illustrate the performance of the proposed approach. In Section 5, we apply the proposed method to estimate the parameters and the time-varying drug effect for two clinical data sets from two AIDS patients. We conclude the paper with discussions in Section 6.

2. MODEL REPRESENTATION

2.1. ODE to SSM

Model (1) can be represented by a system of first order ODEs of the form:

$$\dot{\mathbf{X}} + \mathbf{P}[X,t]\mathbf{X} = \mathbf{Q}[X,t], \tag{3}$$

where

$$\mathbf{X} = \begin{pmatrix} T_U \\ T_I \\ V \end{pmatrix}, \mathbf{Q}[X,t] = \mathbf{Q} = \begin{pmatrix} \lambda \\ 0 \\ 0 \end{pmatrix}, \mathbf{P}[X,t] = \begin{pmatrix} \rho + \gamma(t)V & 0 & 0 \\ -\gamma(t)V & \delta & 0 \\ 0 & -N\delta & c \end{pmatrix}. \tag{4}$$

Note that although $\mathbf{Q}[X,t]$ is independent of $\mathbf{X}$, $\mathbf{P}[X, t]$ is function of $\mathbf{X}$ and t, hence the system is nonlinear. To obtain a corresponding SSM for this ODE system, we first convert the ODE into an ordinary difference equation. Following Freed and Walker (1991), a linear approximation yields:

$$\mathbf{X}[t+\Delta_t] = \exp(-\mathbf{P}[X_{t^*},t]\Delta_t)\mathbf{X}[t] + \{\mathbf{I} - \exp(-\mathbf{P}[X_{t^*},t]\Delta_t)\}\mathbf{P}^{-1}[X_{t^*},t]\mathbf{Q} + O\left[\frac{\partial}{\partial t}\frac{\mathbf{Q}}{\mathbf{P}[X_{t^*},t]}\Delta_t^2\right]. \tag{5}$$

where t^* can either be $t + \Delta_t$ or t, which corresponds to implicit and explicit approximation respectively. In the following, we use explicit approximation. This approximation is exact when the coefficient matrices $\mathbf{P}$ and $\mathbf{Q}$ are constant, i.e. the ODE is linear. Approximation error in our case will be discussed in section 2.3.

Random perturbations in the state process lead to adding a noise term in the difference equation:

$$\mathbf{X}[t+\Delta_t] = \exp(-\mathbf{P}[X_t,t]\Delta_t)\mathbf{X}[t] + \{\mathbf{I} - \exp(-\mathbf{P}[X_t,t]\Delta_t)\}\mathbf{P}^{-1}[X_t,t]\mathbf{Q} + \mathbf{\Sigma}(\mathbf{X},t)v_t. \tag{6}$$

where v_t is a Gaussian noise with variance Δ_t and the whole random perturbation term has a variance of $\Delta_t \boldsymbol{\Sigma}(\mathbf{X}, t)\boldsymbol{\Sigma}(\mathbf{X}, t)'$ that depends on current state, time and the length of time-interval. Here we assume $\boldsymbol{\Sigma}(\mathbf{X},t) = \boldsymbol{\Sigma}$. The above representation corresponds to a stochastic differential equation (SDE) of the state process, i.e.,

$$\dot{\mathbf{X}}_t = -\mathbf{P}[X, t]\mathbf{X}_t + \mathbf{Q}[X, t] + \boldsymbol{\Sigma}\, \mathrm{d}B_t, \tag{7}$$

where B_t is a standard Brownian process.

Denote $\mathbf{Y}_t = (Y_{1,t}, Y_{2,t})$ the observed count data for total CD4+T and viral load of the patient at time t, then we have

$$\begin{aligned} Y_{1t} &= T_{u,t} + T_{i,t} + w_{1,t}, \\ Y_{2t} &= V_t + w_{2,t} \end{aligned} \tag{8}$$

where $w_{1,t}$ and $w_{2,t}$ are measurement errors.

Exponential decrease in viral load invalidates the possible assumption of constant variance in viral measurement error. In Liang et al. (2010) and Liu *et al.* (2011), log-transformation was used to stabilize the variance. In our case, one possible way is to assume the nonlinear model $\log(Y_{2,t}) = \log(V_t) + w_{2,t}, w_{2,t} \sim N(0, \sigma^2)$ for the measurement equation of V_t. However, we found that when the variance is relatively small compared to the mean in a log-normal distribution, as in this case, a normal distribution with the variance proportional to the square of the mean is a reasonable choice. Hence we assume $w_{1,t} \sim N(0, \sigma_1^2)$ and $w_{2,t} \sim N(0, \sigma_2^2 V_t^2)$. It is equivalent to assuming that the signal noise ratio of V_t is constant over the measurements.

Putting all the terms in the form of model (2), we have an SSM for the HIV dynamic after drug initialization as:

$$\begin{aligned} &\mathbf{X}_t = \boldsymbol{\Lambda} + \mathbf{F}_t\mathbf{X}_{t-1} + \mathbf{U}_t, \\ &\mathbf{Y}_t = \mathbf{G}\mathbf{X}_t + \mathbf{W}_t, \qquad \begin{pmatrix} \mathbf{U}_t \\ \mathbf{W}_t \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} H_1 & 0 \\ 0 & H_2 \end{pmatrix} \right\}, \end{aligned}$$

where

$$\begin{aligned} &\mathbf{X}_t = \begin{pmatrix} T_{U,t} \\ T_{I,t} \\ V_t \end{pmatrix}, \mathbf{G} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{F}_t = \exp(-\mathbf{P}[X_t, t]\Delta_t) \\ &\boldsymbol{\Lambda} = \{\mathbf{I} - \exp(\mathbf{P}[X_t, t]\Delta_t)\}\mathbf{P}^{-1}[X_t, t]\mathbf{Q}, \\ &H_1 = \Delta_t \begin{pmatrix} \eta_1^2 & 0 & 0 \\ 0 & \eta_2^2 & 0 \\ 0 & 0 & \eta_3^2 \end{pmatrix}, H_2 = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 V_t^2 \end{pmatrix}, \end{aligned} \tag{9}$$

and Δ_t is the time-interval between observed time t and $t-1$, H_1 and H_2 are assumed to be diagonal, i.e., all random perturbations are uncorrelated and $\mathbf{P}[X_t, t]$ and $\mathbf{Q}$ are defined in (4).

Parlett (1976) demonstrated that the exponential of a triangular matrix can be explicitly written out with a recurrence formula on the matrix elements. Hence $\mathbf{F}_t$ has an explicit lower-triangle form. Specifically, its elements are

$$\begin{aligned} &\mathbf{F}_{t,(1,1)} = \exp(-\rho\Delta_t - \gamma(t)V_{t-1}\Delta_t), \mathbf{F}_{t,(2,2)} = \exp(-\delta\Delta_t), \mathbf{F}_{t,(3,3)} = \exp(-c\Delta_t), \\ &\mathbf{F}_{t,(2,1)} = \gamma(t)V_{t-1}\frac{\mathbf{F}_{t,(1,1)} - \mathbf{F}_{t,(2,2)}}{-\rho - \gamma(t)V_{t-1} + \delta}, \mathbf{F}_{t,(3,1)} = N\delta\gamma(t)V_{t-1}\Delta_t\frac{\mathbf{F}_{t,(2,1)} - \mathbf{F}_{t,(3,2)}}{-\rho - \gamma(t)V_{t-1} + c}, \\ &\mathbf{F}_{t,(3,2)} = N\delta\frac{\mathbf{F}_{t,(2,2)} - \mathbf{F}_{t,(3,3)}}{-\delta + c}. \end{aligned}$$

Notice that when Δ_t is very small, $\mathbf{F}(t)$ can be further approximated using Taylor expansion as:

$$\boldsymbol{\Lambda}_t = (\lambda\Delta_t, 0, 0)^T, \mathbf{F}_t = \begin{pmatrix} 1 - \rho\Delta_t - \gamma(t)V_{t-1}\Delta_t & 0 & 0 \\ \gamma(t)V_{t-1}\Delta_t & 1 - \delta\Delta_t & 0 \\ 0 & N\delta\Delta_t & 1 - c\Delta_t \end{pmatrix}. \tag{10}$$

Such a representation corresponds to Euler method of solving an ODE.

There are several complexities. One is the nonlinearity of the ODE, since $\mathbf{P}[X, t]$ depends on $\mathbf{X}$. Because of this, $\mathbf{F}_t$ in (9) and (10) depends on the unobserved state V_{t-1}, making the SSM nonlinear. Instead of using more complicated methods to deal with such a nonlinear SSM, we choose to use another layer of approximation. Specifically, we can replace V_{t-1} in $\mathbf{F}_t$ with a non-parametric function estimate based on entire observed sequence of $Y_{2,t}$. A second choice is to use the conditional mean of V_{t-1} given all the information up to time $t-1$. More detail of this implementation when Δ_t is large is given in real data analysis section. Another difficulty is that $\mathbf{P}[X,t]$ also depends on the drug effect function $\gamma(t)$. Following Liang et al. (2010), we model $\gamma(t)$ with a B-spline approximation (de Boor, 2001). Specifically we assume $\gamma(t) \approx \sum_{j=1}^{s} a_j b_{j,k}(t)$, where $\{b_{j,k}\}_{j=1}^{s}$ are B-spline basis function of kth order piecewise polynomial with $s-k$ interior knots and $\mathbf{a} = (a_1, \ldots, a_s)$ are constant B-spline coefficients.

Note that we did not start directly from a state space model. Our approach preserves the link between the mathematical model and the SSM, hence the interpretations of original parameters are preserved and all the methods developed based on the ODE could be employed to provide insights about the system in the new modeling framework. Also, when the observation times are not equally spaced, some prespecified form of state space model is no longer applicable, but this case can be easily dealt by our method.

2.2. SSM vs. Runge–Kutta

Runge–Kutta methods are a family of explicit and implicit methods for solving ODE in numerical analysis. Essentially, most of them can be seen as special cases of using equation (5). Liang et al. (2010) studied HIV dynamics using Runge–Kutta to solve an ordinary differential system, with parameters optimized by minimizing the squared error between observed data and the numerical solutions. A simple experiment in the simulation part shows that when the time-interval is reasonably small, our solution using model (9) and that by Runge–Kutta are close to each other. This validates our state space model representation of the original ODE system. Based on this fact, we can use simpler but crude methods for solving the ODE to obtain the starting values for optimization in our approach.

2.3. Approximation error

Equation (5) gives a rough magnitude of the approximation error. Specifically, if we denote $f(t) = \gamma(t)V(t)$ (the only time-dependent part in the transition matrix), the magnitude of the error term could be further expanded as follows:

$$\begin{aligned}\frac{\partial}{\partial t}\frac{\mathbf{Q}}{\mathbf{P}[X_t,t]}\Delta_t^2 &= \frac{\partial}{\partial t}\left\{\begin{pmatrix}\rho+f(t) & 0 & 0\\ -f(t) & \delta & 0\\ 0 & -N\delta & c\end{pmatrix}^{-1}\begin{pmatrix}\lambda\\0\\0\end{pmatrix}\right\}\Delta_t^2\\ &= \frac{\partial}{\partial t}\left\{\lambda\begin{pmatrix}\frac{1}{\rho+f(t)}\\ \frac{f(t)}{\rho+f(t)}\frac{1}{\delta}\\ \frac{f(t)}{\rho+f(t)}\frac{N}{c}\end{pmatrix}\right\}\Delta_t^2 = \Delta_t^2\lambda f'(t)\begin{pmatrix}-\frac{1}{(\rho+f(t))^2}\\ \frac{\rho}{(\rho+f(t))^2}\frac{1}{\delta}\\ \frac{\rho}{(\rho+f(t))^2}\frac{N}{c}\end{pmatrix}.\end{aligned} \tag{11}$$

V_t is observed in real data to have large but quickly decreasing values in the early phase and then stay around some stable small value in the latter phase. As $\gamma(t)$ has roughly the same magnitude in the process, it yields that when $f'(t)$ is large, $f(t)$ also has large values (early phase) and that when $f(t)$ is small, $f'(t)$ is also small (latter phase). Therefore for the whole process, the error term (11) is expected to have a small magnitude.

3. MODEL ESTIMATION

Model (9) is a time-varying coefficients state-space model, with the structure of $\mathbf{F}_t$ known at each observed time-point. Let $\mathbf{\Theta} = (\lambda, \rho, N, \delta, c, \{\eta_i\}_{i=1}^3, \{\sigma_i\}_{i=1}^2, \mathbf{a})$ be the parameters we need to estimate, where $\mathbf{a}$ are the B-Spline coefficients. For a given parameter configuration, we use Kalman filter to compute the optimal observation predictions and the corresponding prediction errors, then the likelihood function of the model is calculated by the prediction-error decomposition. All parameters are estimated by maximizing the likelihood function.

Kalman filter (Kalman, 1960) is a recursive algorithm for filtering and estimation of linear and Gaussian state space models. Here we give a brief review of this algorithm. Details can be found in Harvey (1989) and Durbin and Koopman (2000). Since the distribution of $\mathbf{X}_t \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_t$ is Gaussian, one only needs to obtain its mean and covariance matrix. Denote $\boldsymbol{\mu}_{t|t} = E(X_t \mid Y_1, \ldots, Y_t)$ and $\mathbf{\Sigma}_{t|t} = Cov(X_t \mid Y_1, \ldots, Y_t)$, the Kalman recursion for model (2) proceeds by updating $\boldsymbol{\mu}_{t|t}$ and $\mathbf{\Sigma}_{t|t}$ as follows:

$$\begin{aligned}\mathbf{P}_t &= \mathbf{F}_t\mathbf{\Sigma}_{t|t}\mathbf{F}_t^{\mathrm{T}} + \mathbf{W}_t, \quad \mathbf{S}_t = \mathbf{G}\mathbf{P}_t\mathbf{G}^T + \mathbf{V}_t,\\ \boldsymbol{\mu}_{t+1|t} &= \mathbf{\Lambda} + \mathbf{F}_t\boldsymbol{\mu}_{t|t}, \quad \mathbf{e}_{t+1} = \mathbf{Y}_{t+1} - \boldsymbol{\mu}_{t+1|t},\\ \boldsymbol{\mu}_{t+1|t+1} &= \boldsymbol{\mu}_{t+1|t} + \mathbf{P}_t\mathbf{G}^{\mathrm{T}}\mathbf{S}_t^{-1}e_{t+1}, \quad \mathbf{\Sigma}_{t+1|t+1} = \mathbf{P}_t - \mathbf{P}_t\mathbf{G}^T\mathbf{S}_t^{-1}\mathbf{G}\mathbf{P}_t,\end{aligned} \tag{12}$$

where $\mathbf{e}_t$ and $\mathbf{S}_t$ are the prediction error and prediction error variance respectively. When the initial state vector has density $N(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$, the likelihood by prediction error decomposition is written as

$$L(\mathbf{Y} \mid \mathbf{\Theta}) = \sum_{t=1}^{n} \log p(\mathbf{Y}_t \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_{t-1}) = -\frac{np}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}(\log \mid \mathbf{S}_t \mid + \mathbf{e}_t^T\mathbf{S}_t^{-1}\mathbf{e}_t). \tag{13}$$

When little information is known about the initial state vector, diffuse initialization ($\mathbf{\Sigma}_1 \to \infty$) could be used and a modified version of the above likelihood could be applied (Durbin and Koopman, 2001).

We maximize the likelihood by the Broyden–Fletcher–Goldfarb–Shannon method (Broyden, 1970) which is a generalized Newton's method with the Hessian matrix approximated by the function value. Several computational aspects need to be addressed here. First, model (9) involves N, δ and $N\delta$. It is often more computationally stable to use $N\delta$ as a parameter, which has the biological interpretation as the average rate of viral production. Second, sensible starting values are important for locating the true MLE. Some parameters have biological meanings and can often be set to certain reasonable starting values using biological knowledge. There is less knowledge about the spline coefficients. We use a two stage procedure. First, we estimate model (9) assuming constant $\gamma(t)$. The results are used as the starting values for a more complex structure of $\gamma(t)$ in a refined estimation step. This produces more stable and accurate results in our simulation. Third, when the magnitude of a parameter is expected to be smaller than that of other parameters,

it should be scaled in the optimization function for finding the MLE and estimating the hessian matrix around MLE, as is suggested by Nash (2010).

4. SIMULATION STUDIES

To evaluate the performance of the proposed methods, we carried out a simulation study with multiple sets of signal to noise ratio. Parameter set-up in model (9) is as follows: the initial values $(T_{U0}, T_{I0}, V_0) = (600, 30, 10^5)$, the true parameters $(\lambda_0, \rho_0, N_0, \delta_0, c_0) = (36, 0.108, 10^3, 0.5, 3)$ and $\gamma(t) = 9{\cdot}10^{-5}\{1 - 0.9\cos(\pi t/1000)\}$. First we study the difference between the solution of the difference equation (5) and that of the ODE (3). On the time-domain [0,20], four time-intervals settings: $h = 0.1, 0.2, 0.4, 0.667$ are used to obtain the solution paths. They are further compared to the solution given by Runge–Kutta Methods. The difference is measured by absolute relative difference (ARD), defined as $1/T\sum_{t=1}^{T}|Y_{t,ssm} - Y_{t,rk}|/Y_{t,rk}$, where Y_t could be $T_{U,t} + T_{I,t}$ or V_t, the count number path generated by the corresponding method. The comparison is given in Table 1. It is seen that the two methods give similar solutions to the same ODE when the time-interval is reasonably small, hence the approximation, as well as the conversion from ODE to SSM, is reasonably accurate.

The performance of the proposed method is compared to SNLS method in Liang et al. (2010). The data are simulated with very small state noise $(\eta_1 = \eta_2 = \eta_3 = 10^{-2})$. We vary the level of (constant) measurement error variances (σ_1^2 and σ_2^2). The sequences of the underlying states $(T_{U,t}, T_{I,t}, V_t)$ are obtained and the observed data are simulated by further adding the measurement white noises. This set-up is the same as the deterministic simulation setting in Liang et al. (2010) for the purpose of a numerical comparison. Two sampling schedules were used: (i) at every 0.1 time-units on the interval [0,20] and (ii) at every 0.2 time-units on the interval [0,20] which correspond to sample size of 200 and 100 respectively. To obtain the parameters estimates, Liang et al. (2010) proposed a multi-stage smoothing-based (MSSB) approach to search starting values. Based on initial values obtained from MSSB, our final parameters are estimated by optimizing on the loglikelihood function (13). Such procedures are repeated on 200 simulated data sets. The drug effect function $\gamma(t)$ is approximated by a spline of order two with three knots (a straight line, i.e. $\gamma(t) = \sum_{j=1}^{2} a_j b_{j,2}$). The estimation performance is evaluated by the average relative estimation error (ARE), defined as:

$$\text{ARE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\theta - \hat{\theta}_i|}{\theta} \times 100\%,$$

where when θ stands for one of λ, ρ, N, δ, c, it is the true value and $\hat{\theta}$ is the corresponding estimate; when θ stands for spline coefficient $\{a_j j = 1,2\}$, θ corresponds to the optimal spline coefficient with respect to the true function $\gamma(t)$.

Table 2 reports the simulation results, from which we observe that (i) ARE of the parameter estimates by our approach are smaller than the simulation results of the SNLS method, given in Table 1 of Liang et al. (2010) and greatly improve the rough estimates obtained from MSSB methods, (ii) ARE of the spline coefficient estimates relative to the optimal spline coefficients are very small, showing relatively accurate estimation of the function $\gamma(t)$ and (iii) as expected, higher signal noise ratio and larger sample size yield more accurate estimates.

5. REAL DATA ANALYSIS

In this section, the proposed SSM model (9) is applied on clinical HIV data of two patients. Each data set comprises the virus concentration measurements, CD4+T cell measurements and the time-points at which the measurements are collected. As is known,

Table 1. ARD for $T_U + T_I$ and V approximated by model (9) and Runge–Kutta

h	0.1	0.2	0.4	0.667
$T_U + T_I$	0.005	0.005	0.009	0.012
V	0.002	0.004	0.008	0.013

Table 2. ARE of parameters estimated by state space model approach under different settings

			ARE(%):MSSB					ARE(%):SSM						
N	σ_1^2	σ_2^2	λ	ρ	N	δ	c	λ	ρ	N	δ	c	$\mathbf{a}_1$	$\mathbf{a}_2$
200	40	200	90.09	19.13	46.89	84.88	13.10	1.04	3.36	1.29	0.55	0.93	0.77	0.52
	30	150	89.96	15.23	51.87	85.01	14.98	0.94	3.19	0.97	0.46	0.72	0.76	0.52
	20	100	89.96	13.40	47.98	84.45	13.10	0.88	2.93	0.85	0.42	0.64	0.76	0.50
100	40	200	80.80	29.94	41.42	67.38	17.74	1.12	3.77	1.21	0.56	0.85	0.74	0.95
	30	150	80.55	25.27	42.30	68.80	16.92	1.01	3.31	1.19	0.50	0.83	0.74	0.83
	20	100	80.20	21.78	44.53	69.92	17.01	0.94	3.13	0.97	0.49	0.69	0.71	0.69

CD4 cells protect the body from infection and are one of the most important defence mechanisms of the body. HIV virus attacks CD4 cells to disable them, and uses CD4 cells to make more copies of HIV virus. Hence CD4 cell count is one of most important biomarkers in assessment of antiretroviral. If a HIV patient's CD4 cell count drops below 200, the patient is classified as having AIDS. In our data, virus load is scheduled to have 13 measurements and 14 measurements for the first 2 weeks and then one measurement at weeks 4,8, 12, 14, 20, 24, 28, 32, 36, 40, 44, 48, 52, 56, 64, 74 and 76. Total CD4+T cell count, including uninfected target cells T_U and productively infected cell T_I, are measured at weeks 2, 4 and monthly thereafter (weeks 8,12, . . . ,76). Note that not all of the measurements of virus load and CD4 cell counts were taken at the same time. Figure 1 shows the data. The data has been analyzed by Liang and Wu (2008) and Liang et al. (2010).

5.1. Model specification and estimation

Due to the nonlinear nature of the state space model under consideration, we insert intermediate time-steps between the time of observations. At these intermediate time-points, we process (filter) the states, but the observations are treated as missing. Because of the asynchronous nature of our observed data, at some points, one of the V (HIV virus) or $T_U + T_I$ (CD4+T) is observed but the other is missing. The intermediate time-steps are used to minimize the discretizing approximation error of the differential equations, a method commonly used in numerical methods as well. Filtering through these intermediate time-steps updates the state mean and adjust the covariance matrix using the state equation only. Such a procedure reduces the approximation error, and does not artificially reduce the variability in the estimate. However, it does increase the computational costs.

On the other hand, the transition matrix $\mathbf{F}_t$ depends on unknown state V_{t-1}. For simplification and ease of computation, we replace it with an estimate. As mentioned earlier, there are two possible choices. One is to use a smoothing estimate of V_t as a function of t, based on the observed $Y_{2,t}$. The second is to use the filtered mean estimate of $E[V_{t-1} \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_{t-1}]$, obtained through Kalman filter recursion. Since the observed viral load sequence $Y_{2,t}$ is a direct observation of V_t with measurement noise, our experience shows that the smoothing estimator is better, especially when Δ_t is large. Of course, if the prediction is the objective, one has to use the predicted mean. In this study, since the main objective is estimating all parameters instead of prediction, lowess curve is used as to estimate the path of viral load so that $\mathbf{F}_t$ depends only on Θ. The smoothing parameter in lowess curve fit is chosen such that the

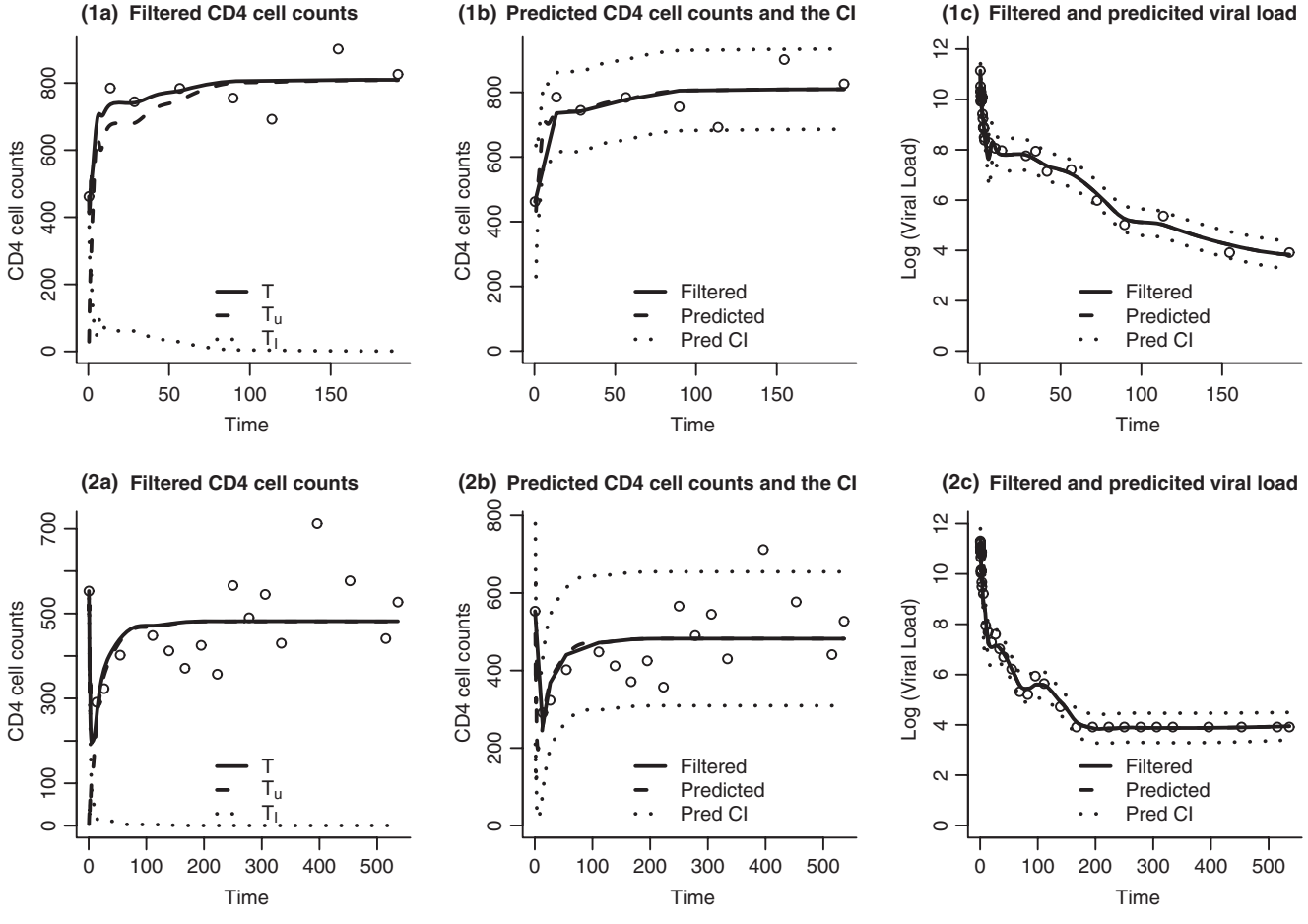

Figure 1. The observed values (circle), fitted states (solid line) and the associated pointwise confidence intervals for one-step predictions (dotted line) of the CD4 cells (left two columns) and viral load (the 3rd column) for patients 1 (upper panel) and 2 (lower panel)

fitted values at observed times are very close to the observed values. Such a substitution will certainly create approximation bias and also reduces the overall variability in the estimates and caution should be exercised. A nonlinear filtering algorithm such as the Sequential Monte Carlo methods (Liu and Chen, 1998; Doucet *et al.*, 2001) may be used to overcome such difficulties.

We initialize the system as follows. Let $(T_{U,0}, T_{I,0}, V_0) \sim N(\mu_0, \Sigma_0)$, where $\mu_0 = (\alpha Y_{1,1}, (1-\alpha)Y_{1,1}, Y_{1,2})$, $(Y_{1,1}, Y_{1,2})$ is the first observed CD4+T cell counts and viral load. Σ_0 is set to be $(10^4, 10^8)$. The initial distribution introduced α, the initial ratio between $T_{U,1}$ and $T_{U,1} + T_{I,1}$. We treat it as a parameter to be estimated. Our experience shows that the likelihood function can be quite flat at places and the MLE may fit the data well but lacks of the biological meanings. In this case, we fix δ and c at the estimated value from Perelson's model (Perelson *et al.*, 1996) based on the viral load data within the 1 week and estimate the rest of the parameters. The resulting likelihood is very close to the one without the constraint. However, the estimated parameters with the constraint make more biological sense.

The dynamic of the drug effect $\gamma(t)$ is expected to have more dynamic in the beginning, hence the interior knots are increased to be equally-spaced on the scale of logarithm time-scale. To avoid local oscillations, only splines with orders three and four, with interior knots number up to four, are considered. Here the sample size is small relative to the maximum order of the models in the candidate class, BIC (Schwarz, 1978) and AICc (Hurvich and Tsai, 1989) are used to choose the best model.

5.2. Results and discussions

The BIC and AIC_c values for two patients with $\gamma(t)$ approximated by different orders and knot numbers are shown in Table 3. Among all the models considered, $\gamma(t)$ approximated by a spline of order three and knots in the boundary (no interior knots) is shown to be the best one, in terms of both BIC and AIC_c. Hence $\gamma(t)$ is estimated with a cubic polynomial.

Under this optimal model, the estimated values and the associated 95% confidence intervals of the parameters in model (9) are shown in Table 4. Apart from N, all other parameters have similar estimated values to those obtained by using method SNLS (Liang et al., 2010). The estimated variances of the noise in the state equation (9) are very small, which essentially make our state equation in the model a deterministic one. The proliferation rates of uninfected CD4+T cells are estimated as 415 and 40 cells per day for patient 1 and 2 respectively; the death rates are .51 and .08 per day, which means the half-life of 1.35 and 8.66 days for these 2 patients; the numbers of virions produced by each of the infected cells are 90 and 770 per cell for patient 1 and 2 respectively. Such a rapid proliferation rate for patient 1 may partially explain why the CD4 curve of this patient exceeds that of patient 2. This finding also suggests that the antiretroviral activity in patient 1 is superior to that in patient 2 in terms of building patients' immune system.

Model fitting on the observed data is shown in Figure 1. Fitting for the viral load is shown in log-scale and the standard error used in constructing the CI is estimated by the delta method. The filtered states give a good fit to the data and one-step prediction CI covers almost all the next observed data. It can be seen that all three filtered states have big changes in the beginning and stay in a steady state after about 3 months. The curve of uninfected CD4 cell counts for patient 1 takes a rapid upward direction initially and maintains a steady increase around 800 copies per cubic millimeter (mm^3) of blood. In contrast, the curve of uninfected CD4 cell counts for patient 2 takes a dramatically down-up pattern and then maintains a flat trend around 500 copies. Nevertheless the curves of T_I for both patients decrease dramatically in the early phase of treatment and becomes close to 0 in the steady state.

Compared to the drug effect directly after drug initialization, long term drug effect in Figure 2 shows a decreased drug effect in the latter phase of treatment. The trend of drug effect agrees with former study. It should be noted that since we essentially model the dynamic process after the first measurement is taken, drug effect before that is not shown and starts from a non-zero value.

Table 3. The BIC and AIC_c values with $\gamma(t)$ approximated by B-spline with different orders and knot numbers

	Patient 1				Patient 2			
	Order 3		Order 4		Order 3		Order 4	
Interior knots	BIC	AICc	BIC	AICc	BIC	AICc	BIC	AICc
0	615.6	610.9	619.0	616.3	829.1	810.0	832.9	813.4
1	617.1	614.3	620.4	620.1	831.9	812.4	835.3	815.6
2	619.9	619.6	624.1	626.8	836.2	816.5	840.7	820.9
3	624.1	626.9	628.0	634.5	838.7	819.0	842.2	822.7
4	627.2	633.7	630.1	641.2	843.1	823.6	847.7	828.6

Table 4. Estimated values and 95% confidence interval (smaller font) of the parameters λ,ρ,N using the proposed method SSM and SNLS for 2 patients. The estimated values of δ and c were obtained from Perelson et al.(1996) and fixed at (1.09, 2.46) and (0.43, 3.78) for patients 1 and 2 respectively

	Patient 1		Patient 2	
Parameter	SNLS	SSM	SNLS	SSM
λ	$397.09_{(216.43,594.30)}$	$414.57_{(219.87,609.26)}$	$45.45_{(29.78,81.48)}$	$40.37_{(17.40,63.33)}$
ρ	$0.49_{(0.26,0.75)}$	$0.51_{(0.26,0.76)}$	$0.10_{(0.06,0.18)}$	$0.08_{(0.03,0.13)}$
N	$264.74_{(203.40,350.00)}$	$90.31_{(75.43,105.19)}$	$1114.37_{(856.62,1428.93)}$	$770.84_{(602.03,939.65)}$

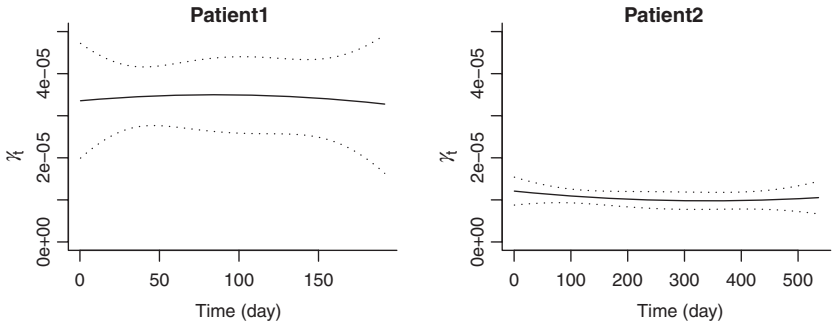

Figure 2. The estimated drug effect γ_t for 2 patients and its 95% pointwise CIs

6. CONCLUSION AND DISCUSSION

In this article, we have proposed a state space model approach for modeling the dynamics of HIV after initialization of drug therapy. This approach involves a transformation from some acknowledged HIV dynamic models ODE and a specification of the noise term in the stochastic process of the states to form a state space model, which is further estimated utilizing available algorithm and methods. Such approach generalizes the deterministic ODE to a stochastic representation so that estimation and prediction could be done under the framework of the state space model. Instead of starting from a prespecified state space model, the proposed SSM is derived from the ODE and all biologically meaningful parameters are reserved and estimated.

Besides the direct link with original ODE model, our approach has several other advantages. As the dynamic is modeled using SSM, many benefits of SSM modeling are attained. For example, Kalman filter recursion leads to easy estimation as well as prediction. Missing values can also be easily dealt with. In both simulation and real data application, only routine optimization scheme are needed to obtain accurate results. Forming the dynamic process in a SDE form also provides more flexibility in developing more sophisticated models. One potential extension might be letting the noise variance $\mathbf{\Sigma}$ in the SDE of state process take some explicit form or depend on the state, which could help model the usually noisy data of CD4+T cell counts, as believed by AIDS investigators. Another interesting extension might be a mixed effect model, which was once explored in Liu et al. (2011), but how the state space model could be set up from the ODE was not discussed.

The proposed method involves various approximations to form and calibrate the state space model from the ODE system. There are essentially two types of approximation error involved: discretization error when changing the ODE to the state space form and calibration error when forming the transition matrix $\mathbf{F}(X_t, t)$ with some "plug-in" estimates of V_t. Discretization error is controlled by the step size, typically in the order of the square root of the step size for Euler–Maruyama type of approximations. This is the reason that we inserted intermediate time-steps between observation time-points to reduce the discretization error, at the expense of heavier computation. Table 1 in the simulation study shows small discretization error in the transformation from ODE to SSM. Table 2 further validates that parameters estimation error induced by such discretization error can be effectively controlled even when $F(X_t, t) = F(V_{t-1}, t)$ is approximated by $F(Y_{2,t-1}, t)$ in the filtering, though in the simulation the noise variance in $Y_{2,t}$ is relatively small comparing to the scale of V_t. Therefore, the main estimate error would come from plugging in some estimates of V_{t-1} to form $F(X_t,t)$ and treating it as true. When $E(V_{t-1})$ is used, the procedure is essentially a variant of the Extended Kalman Filter (Harvey, 1989). The error accumulation problem is usually not severe as the system is constantly corrected by the observations. The Kalman gain tends to reduce approximation biases, though the variability of the estimate can be artificially reduced. The problem can be a severe problem if the nonlinearity is severe and the state variance is large. Fortunately in our problem we do observe a noisy version of the V_t (though not at all points). Empirically we found that replacing V_t with a smoothing estimator of Y_{2t} is satisfactory.

By applying the presented approach on both viral load and CD4+T cell data of 2 patients, we are able to estimate all HIV viral dynamic parameters simultaneously, as well as the long-term drug effect. The estimated set of parameters, depending on the individual patient, is able to provide very good fit to the observed data. Due to the small sample size and the high dimensional parameter space, there may exist over-fitting problem in the fitted model. Fixing δ and c at estimations from simpler model helps alleviate it, but more data points are desired to avoid this problem as well as to estimate δ and c at the same time.

Acknowledgements

Chen's research was partially supported by NSF grants DMS-0800183, DMS-0905763, DMS-0915139, Liang's research was partially supported by NSF grant DMS 1007167. The authors thank an associate editor and a referee for valuable comments and suggestions which resulted in significant improvement of the paper.

REFERENCES

Anderson, R. M. and May, R. M. (1992) *Infectious Diseases of Humans: Dynamics and Control*. new edn. New York: Oxford University Press.
de Boor, C. (2001) *A Practical Guide to Splines*. Vol. **27**, revised edn. New York: Springer-Verlag.
Broyden, C. G. (1970) The convergence of a class of double-rank minimization algorithms. *IMA Journal of Applied Mathematics* **6**, 222–31.

Callaway, D. S. and Perelson, A. S. (2002) HIV-1 infection and low steady state viral loads. *Bulletin of Mathematical Biology* **64**, 29–64.
Doucet, A., de Freitas, N. and Gordon, N., et al. (2001) *Sequential Monte Carlo Methods in Practice, Statistics for Engineering and Information Science*. New York: Springer-Verlag.
Durbin, J. and Koopman, S. J. (2000) Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society, Series B* **62**, 3–56.
Durbin, J. and Koopman, S. J. (2001) *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.
Freed, A. D. and Walker, K. P. (1991) From differential to difference equations for first order ODEs. Technical Memorandum 104530, NASA.
Harvey, A. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.
Ho, D., Neumann, A., Perelson, A., Chen, W., Leonard, J. and Markowitz, M. (1995) Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**, 123–6.
Hurvich, C. and Tsai, C. (1989) Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.
Kalman, R. E. (1960) A new approach to linear filtering and prediction problems. *Trans ASME, Journal of Basic Engineering* **82**, 35–45.
Liang, H., Miao, H. and Wu, H. (2010) Estimation of constant and time varying dynamic parameters of HIV infection in a nonlinear differential equation model. *The Annals of Applied Statistics* **4**, 460–83.
Liang, H. and Wu, H. (2008) Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association* **103**, 1570–83.
Liu, J. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association* **93**, 1032–44.
Liu, D., Lu, T., Niu, X. F. and Wu, H. (2011) Mixed-effects state-space models for analysis of longitudinal dynamic systems. *Biometrics* **67**, 476–85.
Merrill, S. (1987) *AIDS: Background and the Dynamics of the Decline of Immunocompetence*. Reading, MA: Addison-Wesley, chapter Theoretical Immunology, part II, pp. 59–75.
Mittler, J. (1997) Dynamics of HIV-1-infected cell turnover evaluated using mathematical models, international workshop on HIV drug resistance, treatment strategies and eradication. Abstract 101. Antiviral Therapy.
Nash, J. C. (2010) Tutorial: optimization and related nonlinear modelling computations in R, *http://user2010.org/tutorials/Nash.pdf*. Accessed on 10 November 2010.
Nowak, M. A. and May, R. M. (2000) *Virus Dynamics: Mathematical Principles of Immunology and Virology*. Oxford: Oxford University Press.
Parlett, B. N. (1976) A recurrence among the elements of functions of triangular matrices. *Linear Algebra and its Applications* **14**, 117–21.
Perelson, A., Essunger, P., Cao, Y., Vesanen, M.m, Hurley, A., Saksela, K., Markowitz, M. and Ho, D. (1997) Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature* **387**, 188–91.
Perelson, A. S., Kirschner, D. E. and Boer, R. D. (1993) Dynamics of HIV infection of CD4+T cells. *Mathematical Biosciences* **114**, 81–125.
Perelson, A. S. and Nelson, P. W. (1999) Mathematical analysis of HIV-1 dynamics in vivo. *SIAM Review* **41**, 3–44.
Perelson, A., Neumann, A., Markowitz, M., Leonard, J. and Ho, D. (1996) HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span and viral generation time. *Science* **271**, 1582–6.
Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–4.
Tsay, R. S. (2005) *Analysis of Financial Time Series, Series in Probability and Statistics*. 2nd edn. Hoboken, NJ: John Wiley & Sons.
Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., Lifsonparallel, J. D., Bonhoeffer, S., Nowak, M. A., Hahn, B. H., Saag, M. and Shaw, G. M. (1995) Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* **373**, 117–22.
West, M. and Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models, Springer Series in Statistics*. 2nd edn. New York: Springer-Verlag.
Wu, H. and Ding, A. (1999) Population HIV-1 dynamics in vivo: Applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics* **55**, 410–8.