Contents

1

State Space Models

1.1 INTRODUCTION

To a large degree, the origin of statistical state space models can be traced to dynamical systems in engineering branches including automatic control, communications, robotics, and aerospace systems such as spacecraft attitude control. If $\mathbf{U}(t)$, $\mathbf{Y}(t)$, $\mathbf{X}(t)$, represent input, output, and state vectors, respectively, general state-space equations that govern the relationship between these variables are the nonlinear equations [4], [32],

$$\begin{aligned} \mathbf{Y}(t) &= \mathbf{G}(\mathbf{X}(t), \mathbf{U}(t), t) \\ \tfrac{d}{dt}\mathbf{X}(t) &= \mathbf{F}(\mathbf{X}(t), \mathbf{U}(t), t). \end{aligned} \tag{1.1}$$

The corresponding important special discrete-time linear case is

$$\begin{aligned} \mathbf{Y}(t) &= \mathbf{A}(t)\mathbf{X}(t) + \mathbf{B}(t)\mathbf{U}(t) \\ \mathbf{X}(t+1) &= \mathbf{C}(t)(\mathbf{X}(t) + \mathbf{D}(t)\mathbf{U}(t) \end{aligned} \tag{1.2}$$

where the state variables refer to memory variables. A simple example is that of a container into which water flows at a rate $u(t)$ and from which water flows out at rate $y(t)$, and $x(t)$ is the accumulated water. Then for some g, $y = g(x)$ and $dx/dt = u - g(x)$.

The statistical adaptation of equations (1.1) and (1.2) are widely used discrete time regression-like models made of two interconnected equations, the *observation equation* and the *system equation*, which may assume various linear and nonlinear forms and commonly referred to as *state space models*. This chapter discusses linear and nonlinear state space models and their application in prediction, filtering, and smoothing or interpolation.

1.1.1 Historical Note

Early work on linear state space models by R. E. Kalman and others appeared in the late 1950s, however the models owe their widespread use and popularity to NASA's Apollo space program, designed to achieve preeminence in space for the United States including landing humans on the Moon and bringing them safely back to Earth [51], [86]. In March of 1960, Kalman [58] published a seminal paper in which he developed the "Kalman filter" that gives the recursion formulas for filtering and prediction using the linear state space model in discrete time, thus extending the Wiener–Kolmogorov theory of filtering and prediction for stationary time series set forth in the 1940s. As related in [86], in the fall of that year Kalman presented his paper to scientists and engineers at the Ames Research Center (ARC) of NASA. The audience, due to notation and conceptual problems, had great difficulty at first understanding Kalman's work, but pass that stage the value of the state space approach to nonlinear navigation (state estimation) became apparent and a simulation study for validation of the method and in particular the "extended Kalman filter" (linearization about the best state estimate) took place. By early 1961 it was established that on-board optical measurements combined with the equation of motion could yield adequate estimates for navigation and guidance problems, the breakthrough that the NASA scientists were hoping for. Subsequently an early (perhaps the first) Kalman filter application was made circa 1961 during feasibility studies for the Apollo space program at the Instrumentation Laboratory of MIT.

Since then the Kalman filter has been widely used in navigation and guidance systems and in many other control systems. In particular, nowadays Kalman filters are used routinely in inertial navigation systems installed on transoceanic airliners,

submarines, aircraft carriers, ballistic missiles, and certain spacecraft, for the initial alignment and calibration and mid-course update [5].

It seems that the application of a Kalman filter is a simple matter for it appears that once the problem is formulated in terms of equations (1.2), the standard Kalman filter algorithm can be applied in a straightforward manner. This is only ostensibly so on two accounts. First, casting the problem in the right form, especially when the models are nonlinear, is not an easy task, and second, most of the systems are not fully observable ([40], [41]) and thus there are various difficulties in the successful application of the algorithm. Yet, in spite of these difficulties, nowadays the code of the Kalman filter algorithm is a central component in the software of many sophisticated systems[1].

The continuous time analog of the linear state space model and the Kalman filter have been studied in 1961 in another celebrated paper by Kalman and Bucy [59]. In that paper the authors combined and streamlined their ideas developed independently in previous works in the late 1950s. A very similar line of work during roughly the same period was also pursued in the former USSR by the Russian physicist R. L. Stratonovich ([87], [88]) who studied a recursive algorithm for nonlinear least square estimates of the states of nonlinear dynamical systems driven by white noise. A historical account of this and other much related work in the 1940s and 1950s can be found in [57].

State space models started to permeate the statistical literature in the 1960s and 1970s through the work of individuals interested in forecasting and in particular Bayesian forecasting of nonstationary processes–where the assumption of constant coefficients is quite onerous–as is apparent from the accounts in [46] and [90]. Another reason for the interest in state space models by statisticians is the fact that general state space modeling based on recursive relations of probability densities and their integrals are useful for non-Gaussian time series with abrupt discontinuities

[1]The authors are grateful to I. Y. Bar-Itzhack for the last three references and clarifying remarks on the use of Kalman filters in navigation.

and/or outliers [65]. Comprehensive treatments of state space models and their statistical applications can be found in [7], [26], [44], [65], [81], and [90].

1.2 LINEAR GAUSSIAN STATE SPACE MODELS

Let $Y_1, Y_2, ...$ be a sequence of (scalar) observations or responses and $X_1, X_2, ...$ the corresponding covariate sequence, and as before , let $\mathcal{F}_t$ represent the available information to the observer at time t. It is convenient to adopt the convention that

$$\mathcal{F}_0 = \emptyset, \quad \mathcal{F}_t = \{Y_1, \ldots, Y_{t-1}, Y_t\} = \{\mathcal{F}_{t-1}, Y_t\}$$

while the dependence on the covariates $\{X_1, \ldots, X_t\}$ is kept in the background in the sense that the results are interpreted as conditional on the covariates. We have,

$$\text{Observation equation :} \quad \mathrm{Y_t} = \mathbf{z}_\mathrm{t}'\boldsymbol{\beta}_\mathrm{t} + \mathrm{v_t}, \quad \mathrm{v_t} \sim \mathcal{N}(0, \mathrm{V_t}) \tag{1.3}$$

$$\text{System equation :} \quad \boldsymbol{\beta}_\mathrm{t} = \mathbf{F}_\mathrm{t}\boldsymbol{\beta}_{\mathrm{t}-1} + \mathbf{w_t} \quad \mathbf{w_t} \sim \mathcal{N}_\mathrm{p}(\mathbf{0}, \mathbf{W_t}) \tag{1.4}$$

$$\text{Initial information :} \quad \boldsymbol{\beta}_0 \sim \mathcal{N}_\mathrm{p}(\mathbf{b}_0, \mathbf{W}_0) \tag{1.5}$$

where $\mathbf{z}_t$ is a design vector of covariates, such as past observations, supposed known at time t, where $\mathbf{b}_0, \mathbf{W}_0$ are likewise assumed known, and where we first take $\mathbf{F}_t, V_t, \mathbf{W}_t$ as known. In addition, we assume that $\{v_t\}$ and $\{\mathbf{w}_t\}$ each consists of independent random variables and that $\boldsymbol{\beta}_0, \{v_t\}, \{\mathbf{w}_t\}$ are mutually independent. The main departure from the previous GLM models is that the *state* $\boldsymbol{\beta}_t$, a vector of dimension p, is time dependent and random (reminiscent of random effects), and satisfies the autoregression equation (1.4) by means of the time dependent *transition matrix* $\mathbf{F}_t$. Notice that the joint distributions of the observations and the states are determined by the distributions of the initial state $\boldsymbol{\beta}_0$ and of the error sequences $\{v_t\}, \{\mathbf{w}_t\}$. The system of equations (1.3), (1.4), (1.5) is basically a regression model called *linear state space model* or *dynamic linear model*.

Given the observations $Y_1, ..., Y_N$, the linear state space system (1.3)–(1.5) is used in three types of estimation problems at time t referred to as *prediction or forecasting, filtering*, and *smoothing or interpolation* pending on the relationship between t and N: **prediction** for $t > N$, **filtering** for $t = N$, and **smoothing for** $t < N$. Thus,

the estimation of the state $\boldsymbol{\beta}_t$ or its conditional distribution $p(\boldsymbol{\beta}_t \mid \mathcal{F}_N)$ is called prediction if $t > N$, filtering if $t = N$, and smoothing when $t < N$.

1.2.1 Examples of Linear State Space Models

We next illustrate by means of some simple examples the useful fact that, at the price of some redundancy, many linear models and in particular autoregressive moving average models admit state space representations, implying that the estimation theory for linear state space models dubbed filtering, smoothing, and prediction, is quite comprehensive and can cater to stationary as well as to nonstationary data.

A simple example. Consider the linear system,

$$\begin{aligned} Y_t &= X_t + v_t, \quad v_t \sim \mathcal{N}(0, \sigma_v^2) \\ X_t &= \phi_1 X_{t-1} + \phi_2 X_{t-2} + u_t, \quad u_t \sim \mathcal{N}(0, \sigma_u^2) \end{aligned} \tag{1.6}$$

and let,

$$\mathbf{z}_t = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{\beta}_t = \begin{pmatrix} X_t \\ X_{t-1} \end{pmatrix}, \quad \mathbf{F}_t \equiv \mathbf{F} = \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{w}_t = \begin{pmatrix} u_t \\ 0 \end{pmatrix}.$$

Then clearly $Y_t = \mathbf{z}_t' \boldsymbol{\beta}_t + v_t$ and

$$\begin{pmatrix} X_t \\ X_{t-1} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ X_{t-2} \end{pmatrix} + \begin{pmatrix} u_t \\ 0 \end{pmatrix}$$

or, $\boldsymbol{\beta}_t = \mathbf{F}\boldsymbol{\beta}_{t-1} + \mathbf{w}_t$.

Structural time series. A slight extension of (1.6) is the structural time series model

$$\begin{aligned} Y_t &= T_t + S_t + v_t \\ T_t &= \phi T_{t-1} + w_{t1} \\ S_t &= \psi S_{t-1} - S_{t-2} + w_{t2}, \quad \psi = 2\cos(\omega) \end{aligned} \tag{1.7}$$

where T_t and S_t represent trend and seasonal components, respectively. Notice that when $w_{t2} = 0$ then the solution of $S_t = \psi S_{t-1} - S_{t-2}$ is a sinusoid with frequency ω. The structural model (1.7) can be expressed in terms of (1.3) and (1.4) by observing

that,

$$\boldsymbol{\beta}_t = \begin{pmatrix} T_t \\ S_t \\ S_{t-1} \end{pmatrix} = \begin{pmatrix} \phi & 0 & 0 \\ 0 & \psi & -1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} T_{t-1} \\ S_{t-1} \\ S_{t-2} \end{pmatrix} + \begin{pmatrix} w_{t1} \\ w_{t2} \\ 0 \end{pmatrix}.$$

and $Y_t = (1, 1, 0)\boldsymbol{\beta}_t + v_t$.

State space representation for AR(p). Let X_t be an autoregressive process of order p, not necessarily stationary,

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \epsilon_t.$$

Then the same equation can be expressed in matrix form to give the state equation,

$$\boldsymbol{\beta}_t = \begin{pmatrix} X_{t-p+1} \\ \vdots \\ X_{t-1} \\ X_t \end{pmatrix} = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \phi_p & \phi_{p-1} & \cdots & \phi_1 \end{pmatrix} \begin{pmatrix} X_{t-p} \\ \vdots \\ X_{t-2} \\ X_{t-1} \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \epsilon_t \end{pmatrix}$$

or with obvious notation $\boldsymbol{\beta}_t = \mathbf{F}\boldsymbol{\beta}_{t-1} + \mathbf{w}_t$ and $Y_t = (0, 0, \cdots, 0, 1)\boldsymbol{\beta}_t$.

State space representation for ARMA(p, q). There are several state space representations for autoregressive moving averages [7], [65]. To gain insight into a particular representation–for an alternative see Problem 1–consider the polynomials in the backward shift operator B, $BX_t \equiv X_{t-1}$,

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$$

$$\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$$

where $\phi(B)$ has its roots outside the unit circle, the so called stationarity condition. Obtaining a state equation for $\phi(B)X_t = v_t$ using matrices as was done in the previous example and then an observation equation $Y_t = \theta(B)X_t$, then formally

$$Y_t = \theta(B)X_t = \theta(B)\phi^{-1}(B)v_t$$

and $\phi(B)Y_t = \theta(B)v_t$ is ARMA(p, q). Notice that the role of the *state component* X_t is implicit in the ARMA representation but explicit in the state space representation.

This argument gives the necessary clue for going in the reverse direction where we start with a given ARMA(p, q).

Consider the special case where $p = 3, q = 1$,

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + v_t + \theta_1 v_{t-1}.$$

We first obtain a representation for $\phi(B)X_t = v_t$ or $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + v_t$,

$$\begin{pmatrix} X_{t-2} \\ X_{t-1} \\ X_t \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \phi_3 & \phi_2 & \phi_1 \end{pmatrix} \begin{pmatrix} X_{t-3} \\ X_{t-2} \\ X_{t-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ v_t \end{pmatrix}.$$

Next we express $Y_t = \theta(B)X_t$ in matrix form letting $\theta_0 = 1$ and adding $\theta_2 = 0$,

$$Y_t = (\theta_2, \theta_1, \theta_0) \begin{pmatrix} X_{t-2} \\ X_{t-1} \\ X_t \end{pmatrix}$$

It is easy to check that the last two equations represent the above ARMA(3,1) time series.

The same argument holds for the general stationary ARMA(p, q) by augmenting $\phi(B)$ or $\theta(B)$ with higher powers of B with zero coefficients as needed pending on whether $p > q$ or $p \leq q$, respectively. Thus with $r = \max(p, q + 1)$, $\theta_0 = 1$, and adding as needed $\phi_j = 0$, $j > p$, $\theta_j = 0$, $j > q$, define X_t by $Y_t = \theta(B)X_t$ and $\phi(B)X_t = v_t$, and write $\mathbf{X}_t = (X_{t-r+1}, ..., X_t)'$. Then the observation equation is $Y_t = \theta(B)X_t = (\theta_{r-1}, ..., \theta_0)\mathbf{X}_t$, and the state equation is obtained by expressing $\phi(B)X_t = v_t$ as

$$\mathbf{X}_t = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \phi_r & \phi_{r-1} & \phi_{r-2} & \cdots & \phi_1 \end{pmatrix} \mathbf{X}_{t-1} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ v_t \end{pmatrix}.$$

1.2.2 Estimation by Kalman Filtering and Smoothing

Consider the state space system (1.3) - (1.5) for $t = 1, ..., N$, and let

$$\boldsymbol{\beta}_{t|s} = \mathrm{E}[\boldsymbol{\beta}_{\mathrm{t}} \mid \mathcal{F}_{\mathrm{s}}] \tag{1.8}$$

$$\mathbf{P}_{t|s} = \mathrm{E}[(\boldsymbol{\beta}_{\mathrm{t}} - \boldsymbol{\beta}_{\mathrm{t|s}})(\boldsymbol{\beta}_{\mathrm{t}} - \boldsymbol{\beta}_{\mathrm{t|s}})'] \tag{1.9}$$

be the conditional mean and its precision matrix. Observe that the covariance matrix between $\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|s}$ and $Y_1, ..., Y_s$ is zero for all t and s. Therefore, by the normal assumption $\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|s}$ is also independent of $Y_1, ..., Y_s$ for all t and s which implies that $\mathbf{P}_{t|s}$ is also the conditional covariance matrix of $\boldsymbol{\beta}_{t|s}$. Let $\boldsymbol{\beta}_{0|0} = \mathbf{b}_0, \mathbf{P}_{0|0} = \mathbf{W}_0$, and assume the initial condition $\boldsymbol{\beta}_0 \mid \mathcal{F}_0 \sim \mathcal{N}_p(\boldsymbol{\beta}_{0|0}, \mathbf{P}_{0|0})$. Then we have.

Kalman Prediction

$$\begin{aligned} \boldsymbol{\beta}_{t|t-1} &= \mathbf{F}_t \boldsymbol{\beta}_{t-1|t-1} \\ \mathbf{P}_{t|t-1} &= \mathbf{F}_t \mathbf{P}_{t-1|t-1} \mathbf{F}_t' + \mathbf{W}_t \end{aligned} \tag{1.10}$$

Kalman Filtering

$$\begin{aligned} \boldsymbol{\beta}_{t|t} &= \boldsymbol{\beta}_{t|t-1} + \mathbf{K}_t (Y_t - \mathbf{z}_t' \boldsymbol{\beta}_{t|t-1}) \\ \mathbf{P}_{t|t} &= [\mathbf{I} - \mathbf{K}_t \mathbf{z}_t'] \mathbf{P}_{t|t-1} \end{aligned} \tag{1.11}$$

where the so called *Kalman Gain* $\mathbf{K}_t$ is given by

$$\mathbf{K}_t = \mathbf{P}_{t|t-1} \mathbf{z}_t [\mathbf{z}_t' \mathbf{P}_{t|t-1} \mathbf{z}_t + V_t]^{-1}. \tag{1.12}$$

Proof. The prediction equations (1.10) follow from (1.4),

$$\boldsymbol{\beta}_{t|t-1} = \mathrm{E}[\boldsymbol{\beta}_{\mathrm{t}} \mid \mathcal{F}_{\mathrm{t-1}}] = \mathrm{E}[\mathbf{F}_{\mathrm{t}} \boldsymbol{\beta}_{\mathrm{t-1}} + \mathbf{w}_{\mathrm{t}} \mid \mathcal{F}_{\mathrm{t-1}}] = \mathbf{F}_{\mathrm{t}} \boldsymbol{\beta}_{\mathrm{t-1|t-1}}$$

and

$$\begin{aligned} \mathbf{P}_{t|t-1} &= \mathrm{E}[(\boldsymbol{\beta}_{\mathrm{t}} - \boldsymbol{\beta}_{\mathrm{t|t-1}})(\boldsymbol{\beta}_{\mathrm{t}} - \boldsymbol{\beta}_{\mathrm{t|t-1}})'] \\ &= \mathrm{E}\{[\mathbf{F}_{\mathrm{t}}(\boldsymbol{\beta}_{\mathrm{t-1}} - \boldsymbol{\beta}_{\mathrm{t-1|t-1}}) + \mathbf{w}_{\mathrm{t}}][\mathbf{F}_{\mathrm{t}}(\boldsymbol{\beta}_{\mathrm{t-1}} - \boldsymbol{\beta}_{\mathrm{t-1|t-1}}) + \mathbf{w}_{\mathrm{t}}]'\} \\ &= \mathbf{F}_t \mathbf{P}_{t-1|t-1} \mathbf{F}_t' + \mathbf{W}_t \end{aligned}$$

To obtain (1.11), recall the initial condition $\boldsymbol{\beta}_0 \mid \mathcal{F}_0 \sim \mathcal{N}(\boldsymbol{\beta}_{0|0}, \mathbf{P}_{0|0})$, and write

$$\boldsymbol{\beta}_{t-1} \mid \mathcal{F}_{t-1} \sim \mathcal{N}(\boldsymbol{\beta}_{t-1|t-1}, \mathbf{P}_{t-1|t-1}).$$

Then

$$\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1} = \mathbf{F}_t\boldsymbol{\beta}_{t-1} + \mathbf{w}_t \mid \mathcal{F}_{t-1} \sim \mathcal{N}(\boldsymbol{\beta}_{t|t-1}, \mathbf{P}_{t|t-1}),$$

from which

$$Y_t \mid \mathcal{F}_{t-1} \sim \mathcal{N}(\mathbf{z}_t'\boldsymbol{\beta}_{t|t-1}, \mathbf{z}_t'\mathbf{P}_{t|t-1}\mathbf{z}_t + V_t)$$

and

$$\mathrm{Cov}(\boldsymbol{\beta}_\mathrm{t}, \mathrm{Y}_\mathrm{t} \mid \mathcal{F}_{\mathrm{t}-1}) = \mathbf{P}_{\mathrm{t}|\mathrm{t}-1}\mathbf{z}_\mathrm{t}$$

and hence,

$$\left(\begin{array}{c} \boldsymbol{\beta}_t \\ Y_t \end{array} \right) \Bigg| \mathcal{F}_{t-1} \sim \mathcal{N}\left[\left(\begin{array}{c} \boldsymbol{\beta}_{t|t-1} \\ \mathbf{z}_t'\boldsymbol{\beta}_{t|t-1} \end{array} \right), \left(\begin{array}{cc} \mathbf{P}_{t|t-1} & \mathbf{P}_{t|t-1}\mathbf{z}_t \\ \mathbf{z}_t'\mathbf{P}_{t|t-1} & \mathbf{z}_t'\mathbf{P}_{t|t-1}\mathbf{z}_t + V_t \end{array} \right) \right].$$

Therefore, from the conditional multivariate normal distribution (see (1.14) below) and after some algebra,

$$\boldsymbol{\beta}_t \mid Y_t, \mathcal{F}_{t-1} \sim \boldsymbol{\beta}_t \mid \mathcal{F}_t \sim \mathcal{N}[\boldsymbol{\beta}_{t|t-1} + \mathbf{K}_t(Y_t - \mathbf{z}_t'\boldsymbol{\beta}_{t|t-1}), (\mathbf{I} - \mathbf{K}_t\mathbf{z}_t')\mathbf{P}_{t|t-1}]$$

where $\mathbf{K}_t$ is given in (1.12). This completes the proof.

In the above proof we have made use of the important fact that if the p-vector $(\mathbf{x}_1', \mathbf{x}_2')'$ has a multivariate normal distribution with corresponding means $(\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')'$ and covariance matrix partitioned compatibly, $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_{ij})$, $i, j = 1, 2$,

$$\left(\begin{array}{c} \mathbf{x}_1 \\ \mathbf{x}_2 \end{array} \right) \sim \mathcal{N}_p \left[\left(\begin{array}{c} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{array} \right), \left(\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array} \right) \right],$$

then $\mathbf{x}_i$ has a multivariate normal distribution with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_{ii}$, $i = 1, 2$, and the conditional distribution of $\mathbf{x}_2$ given $\mathbf{x}_1$ is again multivariate normal with mean

$$\mathrm{E}[\mathbf{x}_2 \mid \mathbf{x}_1] = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \tag{1.13}$$

and covariance matrix

$$\mathrm{cov}[\mathbf{x}_2 \mid \mathbf{x}_1] = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \tag{1.14}$$

The smoother or interpolator for obtaining $\boldsymbol{\beta}_{t-1|N}$ and its covariance matrix $\mathbf{P}_{t-1|N}$, for $t = N, (N-1), ..., 1$, under normality and the initial filtering conditions

$\boldsymbol{\beta}_{N|N}$, $\mathbf{P}_{N|N}$, is given by the following recursions.

Kalman Smoothing

$$\begin{aligned}
&\boldsymbol{\beta}_{t-1|N} = \boldsymbol{\beta}_{t-1|t-1} + \mathbf{B}_t(\boldsymbol{\beta}_{t|N} - \boldsymbol{\beta}_{t|t-1}) \\
&\mathbf{P}_{t-1|N} = \mathbf{P}_{t-1|t-1} + \mathbf{B}_t(\mathbf{P}_{t|N} - \mathbf{P}_{t|t-1})\mathbf{B}_t' \\
&\mathbf{B}_t \equiv \mathbf{P}_{t-1|t-1}\mathbf{F}_t'\mathbf{P}_{t|t-1}^{-1}
\end{aligned} \tag{1.15}$$

Starting at $t = N$ and going backward in time, the smoothing estimate $\boldsymbol{\beta}_{t-1|N}$ is obtained by adjusting the filtering estimate $\boldsymbol{\beta}_{t-1|t-1}$, adding to it a weighted difference between a smoothing estimate $\boldsymbol{\beta}_{t|N}$ and a prediction estimate $\boldsymbol{\beta}_{t|t-1}$.

Proof. The proof of the smoothing recursions is apparently more complicated than that of the Kalman filtering recursions as was already noted by Kalman (1960). Of the several possible lines of attack, including a proof based on projections, we follow that of maximum likelihood as suggested in [76] and elaborated on in [81]. The idea is to maximize with respect to $\boldsymbol{\beta}_{t-1}, \boldsymbol{\beta}_t$ the conditional Gaussian density

$$p(\boldsymbol{\beta}_{t-1}, \boldsymbol{\beta}_t \mid \mathcal{F}_N), \quad t \leq N, \tag{1.16}$$

upon noting that the values of $\boldsymbol{\beta}_{t-1}, \boldsymbol{\beta}_t$ that maximize (1.16) are the respective conditional means $\boldsymbol{\beta}_{t-1|N}, \boldsymbol{\beta}_{t|N}$. To maximize (1.16), observe that from (1.3) and (1.4)

$$\begin{aligned}
p(\boldsymbol{\beta}_{t-1}, \boldsymbol{\beta}_t \mid \mathcal{F}_N) &\propto p(\boldsymbol{\beta}_{t-1}, \boldsymbol{\beta}_t, \mathcal{F}_{t-1}, Y_t, ...Y_N) \\
&= p(\mathcal{F}_{t-1})p(\boldsymbol{\beta}_{t-1}, \boldsymbol{\beta}_t \mid \mathcal{F}_{t-1})p(Y_t, ...Y_N \mid \boldsymbol{\beta}_{t-1}, \boldsymbol{\beta}_t, \mathcal{F}_{t-1}) \\
&= p(\mathcal{F}_{t-1})p(\boldsymbol{\beta}_{t-1} \mid \mathcal{F}_{t-1})p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1})p(Y_t, ..., Y_N \mid \boldsymbol{\beta}_t)
\end{aligned} \tag{1.17}$$

where $p(\boldsymbol{\beta}_{t-1} \mid \mathcal{F}_{t-1})$ is the density of $\mathcal{N}_p(\boldsymbol{\beta}_{t-1|t-1}, \mathbf{P}_{t-1|t-1})$ and $p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1})$ is the density of $\mathcal{N}_p(\mathbf{F}_t\boldsymbol{\beta}_{t-1}, \mathbf{W}_t)$. Assume now that $\boldsymbol{\beta}_{t|N}$ has already been obtained. Then $\boldsymbol{\beta}_{t-1|N}$ is obtained by *minimizing* $-2\log p(\boldsymbol{\beta}_{t-1}, \boldsymbol{\beta}_{t|N} \mid \mathcal{F}_N)$ with respect to $\boldsymbol{\beta}_{t-1}$. This is equivalent to minimizing

$$\begin{aligned}
&(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{t-1|t-1})'\mathbf{P}_{t-1|t-1}^{-1}(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{t-1|t-1}) \\
&+(\boldsymbol{\beta}_{t|N} - \mathbf{F}_t\boldsymbol{\beta}_{t-1})'\mathbf{W}_t^{-1}(\boldsymbol{\beta}_{t|N} - \mathbf{F}_t\boldsymbol{\beta}_{t-1})
\end{aligned}$$

by differentiating with respect to $\boldsymbol{\beta}_{t-1}$ and equating the derivative to zero. The solution is

$$\boldsymbol{\beta}_{t-1|N} = \left(\mathbf{P}^{-1}_{t-1|t-1} + \mathbf{F}'_t\mathbf{W}^{-1}_t\mathbf{F}_t\right)^{-1}\left(\mathbf{P}^{-1}_{t-1|t-1}\boldsymbol{\beta}_{t-1|t-1} + \mathbf{F}'_t\mathbf{W}^{-1}_t\boldsymbol{\beta}_{t|N}\right) \quad (1.18)$$

This can be simplified using the matrix relations (Problem 5),

$$\left(P^{-1} + F'W^{-1}F\right)^{-1} = P - PF'\left(FPF' + W\right)^{-1}FP$$

$$\left(P^{-1} + F'W^{-1}F\right)^{-1}F'W^{-1} = PF'\left(FPF' + W\right)^{-1}$$

where P, F, W stand for $\mathbf{P}_{t-1|t-1}, \mathbf{F}_t, \mathbf{W}_t$, respectively. This and the prediction expressions (1.10) give the desired smoother/interpolator,

$$\begin{aligned}
&\boldsymbol{\beta}_{t-1|N}\\
&= \boldsymbol{\beta}_{t-1|t-1} + \mathbf{P}_{t-1|t-1}\mathbf{F}'_t\left(\mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}'_t + \mathbf{W}_t\right)^{-1}(\boldsymbol{\beta}_{t|N} - \mathbf{F}_t\boldsymbol{\beta}_{t-1|t-1})\\
&= \boldsymbol{\beta}_{t-1|t-1} + \mathbf{P}_{t-1|t-1}\mathbf{F}'_t\mathbf{P}^{-1}_{t|t-1}(\boldsymbol{\beta}_{t|N} - \boldsymbol{\beta}_{t|t-1})\\
&= \boldsymbol{\beta}_{t-1|t-1} + \mathbf{B}_t(\boldsymbol{\beta}_{t|N} - \boldsymbol{\beta}_{t|t-1}).
\end{aligned}$$

To obtain $\mathbf{P}_{t-1|N} = \mathrm{E}[(\boldsymbol{\beta}_{\mathrm{t-1}} - \boldsymbol{\beta}_{\mathrm{t-1|N}})(\boldsymbol{\beta}_{\mathrm{t-1}} - \boldsymbol{\beta}_{\mathrm{t-1|N}})']$, note that

$$\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{t-1|N} = \boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{t-1|t-1} - \mathbf{B}_t(\boldsymbol{\beta}_{t|N} - \boldsymbol{\beta}_{t|t-1})$$

or by rearranging terms,

$$(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{t-1|N}) + \mathbf{B}_t\boldsymbol{\beta}_{t|N} = (\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{t-1|t-1}) + \mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_{t-1|t-1}, \quad (1.19)$$

and that the following cross terms vanish

$$\mathrm{E}(\boldsymbol{\beta}_{\mathrm{t-1}} - \boldsymbol{\beta}_{\mathrm{t-1|N}})(\mathbf{B}_{\mathrm{t}}\boldsymbol{\beta}_{\mathrm{t|N}})' = \mathrm{E}(\boldsymbol{\beta}_{\mathrm{t-1}} - \boldsymbol{\beta}_{\mathrm{t-1|t-1}})(\mathbf{B}_{\mathrm{t}}\mathbf{F}_{\mathrm{t}}\boldsymbol{\beta}_{\mathrm{t-1|t-1}})' = \mathbf{0},$$

and

$$\mathrm{E}(\boldsymbol{\beta}_{\mathrm{t|s}}\boldsymbol{\beta}'_{\mathrm{t|s}}) = \mathrm{E}(\boldsymbol{\beta}_{\mathrm{t}}\boldsymbol{\beta}'_{\mathrm{t}}) - \mathbf{P}_{\mathrm{t|s}} = \mathbf{F}_{\mathrm{t}}\mathrm{E}(\boldsymbol{\beta}_{\mathrm{t-1}}\boldsymbol{\beta}'_{\mathrm{t-1}})\mathbf{F}'_{\mathrm{t}} + \mathbf{W}_{\mathrm{t}} - \mathbf{P}_{\mathrm{t|s}}.$$

Thus from (1.19),

$$\mathbf{P}_{t-1|N} = \mathbf{P}_{t-1|t-1} + \mathbf{B}_t(\mathbf{P}_{t|N} - \mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}'_t - \mathbf{W}_t)\mathbf{B}'_t$$

which together with (1.10) completes the proof.

1.2.3 Estimation in The Linear Gaussian Model

Estimation of parameters in the linear Gaussian system (1.4)–(1.5) can be carried out by the method of maximum likelihood under a certain parametrization assumption often met in practice. We shall assume that the parameters $\mathbf{b}_0, \mathbf{W}_0, \mathbf{F}_t, V_t, \mathbf{W}_t$ depend completely or in part on a vector $\boldsymbol{\theta}$ of *hyperparameters* which do not depend on t. In this case we write

$$\mathbf{b}_0 = \mathbf{b}_0(\boldsymbol{\theta}),\ \mathbf{W}_0 = \mathbf{W}_0(\boldsymbol{\theta}),\ \mathbf{F}_t = \mathbf{F}_t(\boldsymbol{\theta}),\ V_t = V_t(\boldsymbol{\theta}),\ \mathbf{W}_t = \mathbf{W}_t(\boldsymbol{\theta}),$$

and base the inference on the joint distribution of the observations $Y_1, ..., Y_N$, or equivalently the likelihood of $\boldsymbol{\theta}$. Parametrizations of correlation functions of this type is also used in the next chapter on prediction. Similarly, we may assume that $\mathbf{F}_t, V_t, \mathbf{W}_t$ in the system (1.4)–(1.5) do not depend on t and estimate $\boldsymbol{\theta} = (\mathbf{b}_0, \mathbf{W}_0, \mathbf{F}, V, \mathbf{W})$ by maximum likelihood. In either case, the likelihood is obtained from the joint distribution of the *one-step prediction errors* or *innovations*

$$\epsilon_t = Y_t - \mathrm{E}[Y_t \mid \mathcal{F}_{t-1}] = Y_t - \mathbf{z}_t'\boldsymbol{\beta}_{t|t-1} = Y_t - \mathbf{z}_t'\mathbf{F}_t\boldsymbol{\beta}_{t-1|t-1}, \quad t = 1, ..., N$$

which are independent normal random variables with mean zero and variance

$$\sigma_t^2(\boldsymbol{\theta}) = \mathbf{z}_t'\mathbf{P}_{t|t-1}\mathbf{z}_t + v_t = \mathbf{z}_t'\{\mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t' + \mathbf{W}_t\}\mathbf{z}_t + v_t.$$

The Gaussian assumption implies that $\epsilon_1, ..., \epsilon_N$ is a one-to-one linear transformation of $Y_1, ..., Y_N$ so that up to a constant the log-likelihood of $\boldsymbol{\theta}$ based on $Y_1, ..., Y_N$ is given by

$$\log L_y(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{t=1}^{N} \log \sigma_t^2(\boldsymbol{\theta}) - \frac{1}{2}\sum_{t=1}^{N} \epsilon_t^2/\sigma_t^2(\boldsymbol{\theta}).$$

Maximizing the likelihood with respect to $\boldsymbol{\theta}$ is sometime referred to as the *direct* method. See [44] for more details.

The *indirect* method is based on the joint distribution of both the observed time series and the unobserved states and uses the EM algorithm to maximize the resulting likelihood. From (1.4)–(1.5),

$$p(y_t \mid \boldsymbol{\beta}_t, \mathcal{F}_{t-1}; \boldsymbol{\theta}) = p(y_t \mid \boldsymbol{\beta}_t; \boldsymbol{\theta}) = p_v(y_t - \mathbf{z}_t'\boldsymbol{\beta}_t)$$

and

$$p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}, \boldsymbol{\beta}_{t-2}, ..., \boldsymbol{\beta}_0; \boldsymbol{\theta}) = p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}; \boldsymbol{\theta}) = p_w(\boldsymbol{\beta}_t - \mathbf{F}_t \boldsymbol{\beta}_{t-1})$$

where $v_t \sim p_v$ and $\mathbf{w}_t \sim p_w$. The likelihood is then

$$\begin{aligned} L_{y,\beta}(\boldsymbol{\theta}) &= p(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_N, Y_1, ..., Y_N; \boldsymbol{\theta}) = \\ &= p(\boldsymbol{\beta}_0) \prod_{t=1}^{N} p_w(\boldsymbol{\beta}_t - \mathbf{F}_t \boldsymbol{\beta}_{t-1}) p_v(y_t - \mathbf{z}_t' \boldsymbol{\beta}_t). \end{aligned} \tag{1.20}$$

The application of the EM-algorithm to maximize this likelihood in the presence of unobserved states is described in [26], [81].

A general Bayesian method for the estimation of $\boldsymbol{\theta}$ in dynamic models using a Markov chain Monte Carlo method called permutation sampling is studied in [30]. Other estimation methods and tools in dynamic models are discussed in [26], [44], and [81]. In a related problem, Gerencsér [35] discusses recursive estimation in parameter dependent stochastic systems of the form $Y_n(\theta) = C(\theta)X_n(\theta)$, $X_{n+1}(\theta) = A(\theta)X_n + B(\theta)e_n$.

1.2.3.1 Example of Estimation and Filtering in a Structural Model We apply Kalman filtering and prediction to a monthly time series, recorded in Table **??**, of the number of unemployed women older than 20 years of age from 1997 to 2001. The time series plot in Figure 1.1 points to a slowly decreasing trend and an additional periodic component very characteristic of *structural time series*. It seems therefore that a slight extension of the structural model (1.7) discussed earlier,

$$\begin{aligned} Y_t &= T_t + S_t + v_t \\ T_t &= \phi T_{t-1} + w_{t1} \\ S_t &= \psi_1 S_{t-1} + \psi_2 S_{t-2} + w_{t2}, \end{aligned}$$

where T_t and S_t denote trend and sinusoidal components, respectively, is sensible. Then,

$$\boldsymbol{\beta}_t = \begin{pmatrix} T_t \\ S_t \\ S_{t-1} \end{pmatrix} = \begin{pmatrix} \phi & 0 & 0 \\ 0 & \psi_1 & \psi_2 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} T_{t-1} \\ S_{t-1} \\ S_{t-2} \end{pmatrix} + \begin{pmatrix} w_{t1} \\ w_{t2} \\ 0 \end{pmatrix},$$

and $Y_t = (1, 1, 0)\boldsymbol{\beta}_t + v_t$. Setting $V_t = \sigma_v^2$ for the variance of the independent normal sequence $\{v_t\}$ and

$$\mathbf{W}_t = \mathbf{W} = \begin{pmatrix} \sigma_{w_1}^2 & 0 & 0 \\ 0 & \sigma_{w_2}^2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

for the covariance matrix of $\mathbf{w}_t = (w_{t1}, w_{t2}, 0)'$, the parameters that we need to estimate are $\boldsymbol{\theta} = (\phi, \psi_1, \psi_2, \sigma_v^2, \sigma_{w_1}^2, \sigma_{w_2}^2)'$, $\mathbf{b}_0$ and $\mathbf{W}_0$. Initializing the EM-algorithm (see [81]) at $\mathbf{b}_0 = (3, 25, 20)'$,$\mathbf{W}_0 = \mathbf{I}_3$, the 3×3 identity matrix, and $\boldsymbol{\theta}_0 = (0.3, -0.25, 0.25, 6, 3, 10)'$, a large number of iterations produces the following estimates, $\hat{\mathbf{b}}_0 = (8.198, 18.931, 12.448)'$, and $\hat{\boldsymbol{\theta}} = (0.415, 1.586, -1.026, 1.286, 0.326, 0.572)'$. We note that the estimator of the trend is less than 1, $\hat{\phi} = 0.415 < 1$, confirming that the time series exhibits a slowly decreasing trend, and that the variance of the observed process is larger than the variances of the two unobserved components.

Figure 1.2 shows the filtered estimators of the trend and sinusoidal components $T_{t|t}$ (top) and $S_{t|t}$ (bottom), both evaluated at the maximum likelihood estimates. Prediction of the monthly number of unemployed women during the next 12 months is shown in Figure 1.3. Computations were carried out using the function `kalman` available at `http://lib.stat.cmu.edu/S/`.

1.2.3.2 Software Resources for State Space Models S–Plus functions for fitting state space models include the collection `bts.zip` which is based on algorithms described in [90] and is available in `http://lib.stat.cmu.edu/DOS/S/`, and the dynamic system estimation library whose description is given in

`http://www.bank-banque-canada.ca/pgilbert/dse/dsedesc.htm`.

1.3 NONLINEAR AND NON–GAUSSIAN STATE SPACE MODELS

Prediction, filtering, and smoothing, can be approached more generally by using the laws of conditional probability, including Bayes theorem, and relaxing linearity and the normal assumption. In the general approach the dynamics is captured directly

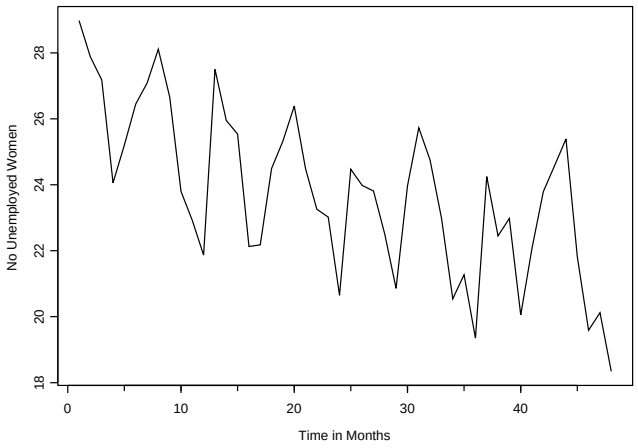

Fig. 1.1 Monthly number of unemployed women between 1997 to 2000. Data in hundreds of thousands, N=48. Source: Bureau of Labor Statistics, Series ID LFU22001702.

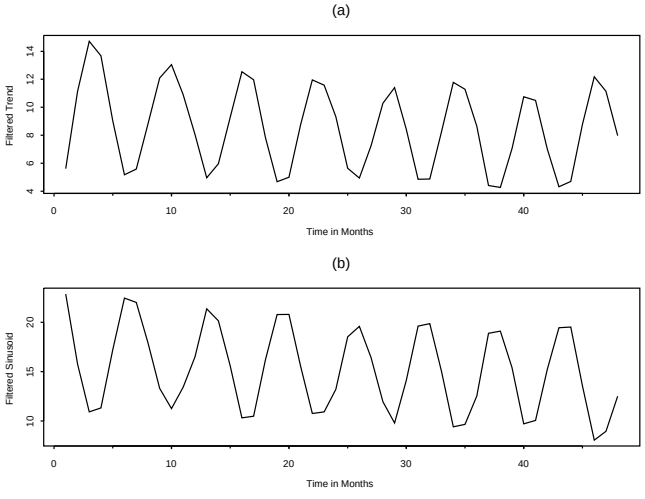

Fig. 1.2 Filtered monthly number of unemployed women between 1997 to 2000. (a) Trend component. (b) Sinusoidal component.

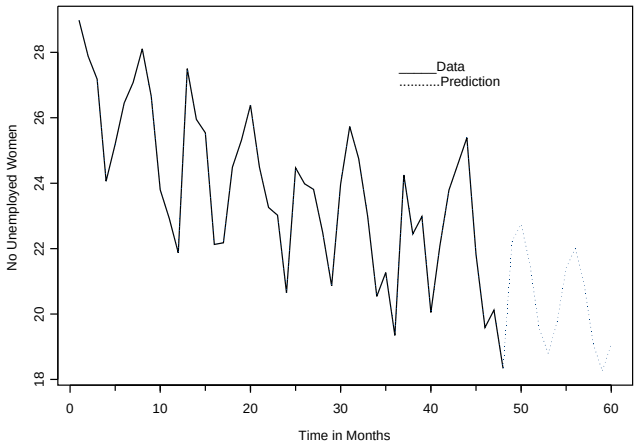

Fig. 1.3 Predicted monthly number of unemployed women for 12 months ahead.

through the conditional densities of the observations and the states, without the formation of any particular system of equations [19], [65], [88].

Analogous to the linear Gaussian system represented by (1.3), (1.4), and (1.5), assume that $\{Y_t\}, t = 1, \ldots, N$ denotes the observed process and let the unobserved state process be $\{\boldsymbol{\beta}_t\},\ t = 0, \ldots, N$. Introducing the generic notation $f()$ for probability densities, the general state space model is formulated as follows:

$$\text{General Observation equation:} \qquad Y_t \mid \boldsymbol{\beta}_t \sim f(y_t \mid \boldsymbol{\beta}_t) \tag{1.21}$$

$$\text{General System equation:} \qquad \boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1} \sim f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}) \tag{1.22}$$

$$\text{Initial information:} \qquad \boldsymbol{\beta}_0 \sim f(\boldsymbol{\beta}_0 \mid \mathcal{F}_0) \equiv f(\boldsymbol{\beta}_0) \tag{1.23}$$

with the understanding that given the states, the responses are independent, and that the sequence of unobserved states forms a Markov process. Thus, equation (1.21) means that given the state sequence $\{\boldsymbol{\beta}_t\}$, the observed process $\{Y_t\}$ forms an independent sequence of random variables, and both (1.22), (1.23) mean that the sequence of unobserved states $\{\boldsymbol{\beta}_t\},\ t = 0, \ldots, N$ is Markov process with initial distribution $f(\boldsymbol{\beta}_0)$.

It is worth pointing out that when the unobserved states assume discrete values then the definition of the general state space model is equivalent to that of a hidden Markov model; see Section **??**. Also, the densities in equations (1.21) and (1.22) may depend on unknown parameters referred to as *hyperparameters*. Estimation methods for hyperparameters are discussed in subsequent sections.

It is easy to verify that the linear normal state space model is a special case of the system represented by (1.21), (1.22) and (1.23) when the corresponding conditional densities are Gaussian. Removing the linearity and normal assumptions, the following representation of a non–linear state space model can be used

$$\begin{aligned} Y_t &= h_t(\boldsymbol{\beta}_t, v_t) \\ \boldsymbol{\beta}_t &= \mathbf{f}_t(\boldsymbol{\beta}_{t-1}, \mathbf{w}_t) \end{aligned} \tag{1.24}$$

where h_t and $\mathbf{f}_t$ are known and suitably defined functions and v_t, $\mathbf{w}_t$ are random sequences, $t = 1, \ldots, N$. A special case of (1.24) is

$$\begin{aligned} Y_t &= h_t(\boldsymbol{\beta}_t) + v_t \\ \boldsymbol{\beta}_t &= \mathbf{f}_t(\boldsymbol{\beta}_{t-1}) + \mathbf{w}_t. \end{aligned} \tag{1.25}$$

Similarly to linear state space models, the problems are estimation of current, future and past states given the data, that is, *filtering*, *prediction* and *smoothing*, respectively. Questions like these pose a great challenge due to non–normality and non–linearity as manifested by (1.21) and (1.22). In what follows we summarize several approaches to these problems before turning to Monte Carlo Markov Chain techniques.

1.3.1 General Filtering and Smoothing

Regardless of any distributional assumptions about the observation and system equations, [60] noted that using the definitions of conditional and marginal distributions and employing Bayes theorem, the following *density* recursions hold in general.

Prediction: General prediction is obtained from the conditional predictive or prediction density $f(\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1})$,

$$f(\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1}) = \int f(\boldsymbol{\beta}_t, \boldsymbol{\beta}_{t-1} \mid \mathcal{F}_{t-1}) d\boldsymbol{\beta}_{t-1}$$

$$\begin{aligned} &= \int f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}, \mathcal{F}_{t-1}) f(\boldsymbol{\beta}_{t-1} \mid \mathcal{F}_{t-1}) d\boldsymbol{\beta}_{t-1} \\ &= \int f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}) f(\boldsymbol{\beta}_{t-1} \mid \mathcal{F}_{t-1}) d\boldsymbol{\beta}_{t-1} \end{aligned} \tag{1.26}$$

for $t = 1, 2, \ldots, N$. In the linear normal case this becomes,

$$f(\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1}) = \int f_{\mathbf{W}}(\boldsymbol{\beta}_t - \mathbf{F}_t \boldsymbol{\beta}_{t-1}) f(\boldsymbol{\beta}_{t-1} \mid \mathcal{F}_{t-1}) d\boldsymbol{\beta}_{t-1} \tag{1.27}$$

where $\mathbf{w}_t \sim f_{\mathbf{W}}$, and as was shown above, $f(\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1})$ reduces to the density of $\mathcal{N}_p(\boldsymbol{\beta}_{t|t-1}, \mathbf{P}_{t|t-1})$.

Filtering: The filtering density $f(\boldsymbol{\beta}_t \mid \mathcal{F}_t)$ is obtained by appealing to Bayes theorem,

$$\begin{aligned} f(\boldsymbol{\beta}_t \mid \mathcal{F}_t) &= f(\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1}, y_t) \\ &= \frac{f(y_t \mid \boldsymbol{\beta}_t, \mathcal{F}_{t-1}) f(\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1})}{f(y_t \mid \mathcal{F}_{t-1})} \\ &= \frac{f(y_t \mid \boldsymbol{\beta}_t) f(\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1})}{f(y_t \mid \mathcal{F}_{t-1})} \end{aligned} \tag{1.28}$$

for $t = 1, 2, \ldots, N$ and $f(y_t \mid \mathcal{F}_{t-1}) = \int f(y_t \mid \boldsymbol{\beta}_t) f(\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1}) d\boldsymbol{\beta}_t$. In the linear normal case this simplifies to

$$f(\boldsymbol{\beta}_t \mid \mathcal{F}_t) = \frac{f_v(y_t - \mathbf{z}_t' \boldsymbol{\beta}_t) f(\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1})}{f(y_t \mid \mathcal{F}_{t-1})} \tag{1.29}$$

where $v_t \sim f_v$, and

$$f(y_t \mid \mathcal{F}_{t-1}) = \int f_v(y_t - \mathbf{z}_t' \boldsymbol{\beta}_t) f(\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1}) d\boldsymbol{\beta}_t,$$

and where $f(\boldsymbol{\beta}_t \mid \mathcal{F}_t)$ is the density of $\mathcal{N}_p(\boldsymbol{\beta}_{t|t}, \mathbf{P}_{t|t})$.

Smoothing: To obtain a general smoothing density $f(\boldsymbol{\beta}_t \mid \mathcal{F}_N)$, $t < N$, consider first the joint density of $\boldsymbol{\beta}_t, \boldsymbol{\beta}_{t+1}$ given the entire history of the process $\mathcal{F}_N$,

$$\begin{aligned} f(\boldsymbol{\beta}_t, \boldsymbol{\beta}_{t+1} \mid \mathcal{F}_N) &= f(\boldsymbol{\beta}_{t+1} \mid \mathcal{F}_N) f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t+1}, \mathcal{F}_N) \\ &= f(\boldsymbol{\beta}_{t+1} \mid \mathcal{F}_N) f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t+1}, \mathcal{F}_t) \\ &= f(\boldsymbol{\beta}_{t+1} \mid \mathcal{F}_N) \frac{f(\boldsymbol{\beta}_t, \boldsymbol{\beta}_{t+1} \mid \mathcal{F}_t)}{f(\boldsymbol{\beta}_{t+1} \mid \mathcal{F}_t)} \\ &= f(\boldsymbol{\beta}_{t+1} \mid \mathcal{F}_N) \frac{f(\boldsymbol{\beta}_{t+1} \mid \boldsymbol{\beta}_t) f(\boldsymbol{\beta}_t \mid \mathcal{F}_t)}{f(\boldsymbol{\beta}_{t+1} \mid \mathcal{F}_t)} \end{aligned} \tag{1.30}$$

The second equality follows from $f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t+1}, \mathcal{F}_N) = f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t+1}, \mathcal{F}_t)$ and the last two equations of (1.30) follow from Bayes theorem. From (1.30) we have,

$$\begin{aligned} f(\boldsymbol{\beta}_t \mid \mathcal{F}_N) &= \int f(\boldsymbol{\beta}_t, \boldsymbol{\beta}_{t+1} \mid \mathcal{F}_N) d\boldsymbol{\beta}_{t+1} \\ &= f(\boldsymbol{\beta}_t \mid \mathcal{F}_t) \int f(\boldsymbol{\beta}_{t+1} \mid \mathcal{F}_N) \frac{f(\boldsymbol{\beta}_{t+1} \mid \boldsymbol{\beta}_t)}{f(\boldsymbol{\beta}_{t+1} \mid \mathcal{F}_t)} d\boldsymbol{\beta}_{t+1}, \quad (1.31) \end{aligned}$$

for $t = N, N-1, \ldots, 1$. In the linear normal case $f(\boldsymbol{\beta}_t \mid \mathcal{F}_N)$ is the density of $\mathcal{N}_p(\boldsymbol{\beta}_{t|N}, \mathbf{P}_{t|N})$.

Implementation of (1.26), (1.28), (1.31) is based on sequential integration which most of the time is complicated. To overcome the computational burden, [60] suggests a numerical method for piecewise linear approximation of the corresponding integrals when the dimension of the state vector is 1. Extension of this method to higher dimensions demands additional computational cost [49]. Some other suggestions for implementation of (1.26), (1.28), (1.31) include the so called Gaussian–sum filter for approximating the integrals with mixtures of Gaussian distributions (Problem 8) and the two–filter formula for smoothing [61], [62] . More recently, a Monte Carlo method for filtering and smoothing was put forth in [63], [65], and is considered in Section 1.4.3.

1.3.2 Dynamic Generalized Linear Models

A special case of non–normal and nonlinear state space models are the so called dynamic generalized linear models (DGLM) for time series data which broaden the class of static generalized linear models–described in Chapter **??**–by allowing time varying random regression coefficients. The definition of DGLM retains the random component– recall (**??**)–while the systematic component (**??**) is augmented by a transition equation for the regression parameters. These models have been applied successfully in several diverse fields such as meteorology, epidemiology and econometrics.

Suppose that $\{Y_t\}$, $t = 1, \ldots, N$ denotes a response time series such that its conditional distribution given the past belongs to the exponential family of distributions

$$f(y_t \mid \theta_t) = \exp\left\{ \frac{y_t\theta_t - b(\theta_t)}{\alpha_t(\phi)} + c(y_t, \phi_t) \right\}, \quad t = 1, \ldots, N \qquad (1.32)$$

where the parametric function $\alpha_t(\phi)$ is assumed known. Unlike the case in (**??**), the random natural parameter θ_t in (1.32) and the conditional expectation of the response $\mu_t = \mathrm{E}[Y_t \mid \theta_t]$ are dynamically linked to a changing sequence of regression or state parameters. That is, with $g(.)$ a monotone link function and a given p-dimensional covariate vector $\{\mathbf{z}_t\}$,

$$g(\mu_t) = \mathbf{z}_t'\boldsymbol{\beta}_t \tag{1.33}$$

where $\{\boldsymbol{\beta}_t\}$, $t = 0, \ldots, N$ is a random sequence of p–dimensional regression parameters obeying the Markov linear transition model

$$\boldsymbol{\beta}_t = \mathbf{F}_t\boldsymbol{\beta}_{t-1} + \mathbf{w}_t, \quad t = 1, 2, \ldots, \tag{1.34}$$

where $\mathbf{F}_t$ is a sequence of $p \times p$ known matrices, and $\{\mathbf{w}_t\}$ is an independent sequence of random variables. It is common in practice to assume that $\mathbf{w}_t$ is normal with zero mean and covariance matrix $\mathbf{W}_t$, and that it is distributed independently of $\mathcal{F}_{t-1}$. In this case we suppose that $\boldsymbol{\beta}_0$ has a normal distribution with mean $\mathbf{b}_0$ and covariance matrix $\mathbf{W}_0$ and that is independent of $\{\mathbf{w}_t\}$ and $\mathcal{F}_{t-1}$. The state may follow other distributions depending on the context of the application.

Equations (1.32) and (1.33) where $f()$ is a member the exponential family of distributions define the *observation equation* of the model corresponding to (1.21). The evolution relation (1.34) corresponds to (1.22) where $f()$ is replaced by the normal density. This coupled with (1.32) and (1.33) conforms to the definition of dynamic generalized linear models for time series data.

Example:A State Space Model for Binary Time Series Suppose that $\{Y_t\}$, $t = 1, \ldots, N$, is a binary time series with π_t denoting the success probability, and consider the following model which includes trend and a lagged value of the response

$$\log\left(\frac{\pi_t}{1-\pi_t}\right) = T_t + \beta_t^1 Y_{t-1}.$$

If

$$T_t = T_{t-1} + w_t^1$$

and

$$\beta_t^1 = \beta_{t-1}^1 + w_t^2$$

with $\{w_t^1\}$ and $\{w_t^2\}$ mutually uncorrelated white noise sequences, then we obtain representation (1.34) by defining $\boldsymbol{\beta}_t = \left(T_t, \beta_t^1\right)'$, $\mathbf{F}_t = \mathbf{I}_2$–the 2×2 identity matrix– and $\mathbf{w}_t = \left(w_t^1, w_t^2\right)'$. In addition set $\mathbf{z}_t = (1, y_{t-1})'$ so that (1.33) is satisfied.

The formulation of DGLM suggests several extensions suitable for non–linear and non–Gaussian state space models as we describe in the following sub–sections regarding conjugate analysis and posterior mode estimation.

1.3.2.1 Conjugate Analysis and Linear Bayes Estimation Binding together conjugate analysis of prior and posterior distributions and linear Bayes estimation (see [43]), we can approach the problems of forecasting, and filtering from a Bayesian perspective [91].

Let the mean and covariance of the state vector $\boldsymbol{\beta}_t$ given $\mathcal{F}_{t-1}$ be $\boldsymbol{\beta}_{t|t-1}$ and $\mathbf{P}_{t|t-1}$, respectively. Dropping any distributional assumptions but specifying the first two moments for the error term in (1.34), and assuming independence assumptions as before, we obtain the following prediction recursions

$$\begin{aligned} \boldsymbol{\beta}_{t|t-1} &= \mathbf{F}_t \boldsymbol{\beta}_{t-1|t-1} \\ \mathbf{P}_{t|t-1} &= \mathbf{F}_t \mathbf{P}_{t-1|t-1} \mathbf{F}_t' + \mathbf{W}_t \\ \boldsymbol{\beta}_{0|0} &= \mathbf{b}_0 \\ \mathbf{P}_{0|0} &= \mathbf{W}_0. \end{aligned} \tag{1.35}$$

An alternative form of the second equation of (1.35) is given by

$$\mathbf{P}_{t|t-1} = \mathbf{B}_t \mathbf{F}_t \mathbf{P}_{t-1|t-1} \mathbf{F}_t' \mathbf{B}_t',$$

where $\mathbf{B}_t$ is a $p \times p$ diagonal matrix of the so called *discount* factors [2], [91].

Recall that for the exponential family (1.32) $b'(\theta_t) = \mu_t$, and suppose we specify a prior for $g^*(\theta_t)$ in terms of the first two moments, where $g^*(.) = (g \circ b')(.)$ such that $g^*(\theta_t) = \mathbf{z}_t' \boldsymbol{\beta}_t$. Then (1.35) implies

$$\begin{aligned} \mathrm{E}\left[g^*(\theta_t) \mid \mathcal{F}_{t-1}\right] &= \mathbf{z}_t' \boldsymbol{\beta}_{t|t-1} \equiv f_t \\ \mathrm{Var}\left[g^*(\theta_t) \mid \mathcal{F}_{t-1}\right] &= \mathbf{z}_t' \mathbf{P}_{t|t-1} \mathbf{z}_t \equiv q_t \end{aligned} \tag{1.36}$$

so that the vector $\left(g^*(\theta_t), \boldsymbol{\beta}_t'\right)'$ given $\mathcal{F}_{t-1}$ has mean vector and covariance matrix

$$\begin{pmatrix} f_t \\ \boldsymbol{\beta}_{t|t-1} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} q_t & \mathbf{z}_t' \mathbf{P}_{t|t-1} \\ \mathbf{P}_{t|t-1} \mathbf{z}_t & \mathbf{P}_{t|t-1} \end{pmatrix}, \tag{1.37}$$

respectively.

Assuming that in (1.32) $\alpha_t(\phi)$ is known for all t, the canonical parameter $\{\theta_t\}$ represents the uncertainty about the distribution of the response given the past of the process. Suppose that $\{\theta_t\}, t = 1, \ldots, N$ follows a conjugate prior of the exponential family type

$$f(\theta_t \mid \mathcal{F}_{t-1}) = \exp\{\gamma_t\theta_t - \delta_t b(\theta_t) + c^\star(\theta_t)\}. \tag{1.38}$$

Then, from (1.32), a straightforward application of Bayes theorem shows that

$$\begin{aligned} f(\theta_t \mid \mathcal{F}_t) &\propto f(y_t \mid \theta_t) f(\theta_t \mid \mathcal{F}_{t-1}) \\ &\propto \exp\left\{(\gamma_t + \frac{y_t}{\alpha_t(\phi)})\theta_t - (\delta_t + \frac{1}{\alpha_t(\phi)})b(\theta_t)\right\}, \end{aligned}$$

which shows that the posterior of θ_t given a new datum at time t belongs to the exponential family. Values for both γ_t and δ_t in (1.38) are chosen based on the fact that the canonical parameter is approximately linked to the covariates and the state parameters by $g^*(\theta_t) \approx \mathbf{z}_t'\boldsymbol{\beta}_t$ [91].

The filtering recursions are obtained by calculating $\boldsymbol{\beta}_{t|t} = \mathrm{E}[\boldsymbol{\beta}_t \mid \mathcal{F}_t]$ and $\mathbf{P}_{t|t} = \mathrm{Var}[\boldsymbol{\beta}_t \mid \mathcal{F}_t]$ using the facts

$$\mathrm{E}[\boldsymbol{\beta}_t \mid \mathcal{F}_t] = \mathrm{E}[\mathrm{E}[\boldsymbol{\beta}_t \mid \theta_t, \mathcal{F}_{t-1}] \mid \mathcal{F}_t], \tag{1.39}$$

$$\mathrm{Var}[\boldsymbol{\beta}_t \mid \mathcal{F}_t] = \mathrm{Var}[\mathrm{E}[\boldsymbol{\beta}_t \mid \theta_t, \mathcal{F}_{t-1}] \mid \mathcal{F}_t] + \mathrm{E}[\mathrm{Var}[\boldsymbol{\beta}_t \mid \theta_t, \mathcal{F}_{t-1}] \mid \mathcal{F}_t]. \tag{1.40}$$

In the absence of a fully specified distribution for $\boldsymbol{\beta}_t$ given $(\theta_t, \mathcal{F}_{t-1})$, (1.37) and an appeal to the linear Bayes methodology (Problem 4) shows that the optimal linear estimator $\mathbf{d}$ of $\mathrm{E}[\boldsymbol{\beta}_t \mid \theta_t, \mathcal{F}_{t-1}]$ in the sense of minimizing the overall quadratic risk

$$r_t(\mathbf{d}) = \mathrm{tr}\{\mathrm{E}[\mathbf{A}_t(\mathbf{d}) \mid \mathcal{F}_{t-1}]\}$$

where tr denotes trace of a matrix, the expectation is taken with respect to $f(\theta_t \mid \mathcal{F}_{t-1})$ and

$$\mathbf{A}_t(\mathbf{d}) = \mathrm{E}[(\boldsymbol{\beta}_t - \mathbf{d})(\boldsymbol{\beta}_t - \mathbf{d})' \mid \theta_t, \mathcal{F}_{t-1}],$$

is given by

$$\hat{\mathrm{E}}[\boldsymbol{\beta}_t \mid \theta_t, \mathcal{F}_{t-1}] = \boldsymbol{\beta}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{z}_t \frac{(g^*(\theta_t) - f_t)}{q_t} \tag{1.41}$$

with an associated minimum value of $r_t(\mathbf{d})$

$$\hat{\text{Var}}\left[\boldsymbol{\beta}_t \mid \theta_t, \mathcal{F}_{t-1}\right] = \mathbf{P}_{t|t-1} - \frac{\mathbf{P}_{t|t-1}\mathbf{z}_t\mathbf{z}_t'\mathbf{P}_{t|t-1}}{q_t}. \tag{1.42}$$

Both of these equations rely on (1.37) and they can be used in both (1.39) and (1.40) in the place of the *true* $\text{E}\left[\boldsymbol{\beta}_t \mid \theta_t, \mathcal{F}_{t-1}\right]$ and $\text{Var}\left[\boldsymbol{\beta}_t \mid \theta_t, \mathcal{F}_{t-1}\right]$. Then, the filtering recursions are as follows:

$$\begin{aligned} \boldsymbol{\beta}_{t|t} &= \boldsymbol{\beta}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{z}_t\tfrac{(f_t^* - f_t)}{q_t} \\ \mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{z}_t\mathbf{z}_t'\mathbf{P}_{t|t-1}\tfrac{(1-q_t^*/q_t)}{q_t} \end{aligned} \tag{1.43}$$

where

$$\begin{aligned} \text{E}\left[g^*(\theta_t) \mid \mathcal{F}_t\right] &= f_t^* \\ \text{Var}\left[g^*(\theta_t) \mid \mathcal{F}_t\right] &= q_t^*. \end{aligned} \tag{1.44}$$

We see that the filtering recursions resemble those of the linear state space model with the exception of the computation of f_t^* and q_t^*.

The smoothing recursions are similar to those derived for the linear state space model and their proof is based on the application of the linear Bayes methodology [90, p. 532].

1.3.2.2 Posterior Mode Estimation Estimation of the state parameters by approximating the mode of their posterior distribution has been suggested in [22] as a generalization of the extended Kalman filter and smoother (see Problem 7). This estimation method is equivalent to the Fisher scoring algorithm.

Consider the vector of the state vector parameters up to time t, that is $\boldsymbol{\beta}_0^t = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{t-1}, \boldsymbol{\beta}_t)'$, $t = 0, \ldots, N$. Then the posterior distribution of $\boldsymbol{\beta}_0^t$ is proportional to

$$f(\boldsymbol{\beta}_0^t \mid \mathcal{F}_t) \propto \left[\prod_{s=1}^{t} f(y_s \mid \boldsymbol{\beta}_s)\right]\left[\prod_{s=1}^{t} f(\boldsymbol{\beta}_s \mid \boldsymbol{\beta}_{s-1})\right] f(\boldsymbol{\beta}_0), \tag{1.45}$$

by Bayes theorem and (1.21), (1.22). Therefore, the posterior log–likelihood is up to a constant equal to

$$\log f(\boldsymbol{\beta}_0^t \mid \mathcal{F}_t) = \sum_{s=1}^{t} \log f(y_s \mid \boldsymbol{\beta}_s) + \sum_{s=1}^{t} \log f(\boldsymbol{\beta}_s \mid \boldsymbol{\beta}_{s-1}) + \log f(\boldsymbol{\beta}_0). \tag{1.46}$$

Formula (1.46) holds quite generally and leads to

$$\begin{aligned} \log f(\boldsymbol{\beta}_0^t \mid \mathcal{F}_t) &= \sum_{s=1}^{t} l_s(\boldsymbol{\beta}_s) - \frac{1}{2}(\boldsymbol{\beta}_0 - \mathbf{b}_0)'\mathbf{W}_0^{-1}(\boldsymbol{\beta}_0 - \mathbf{b}_0) \\ &- \frac{1}{2}\sum_{s=1}^{t}(\boldsymbol{\beta}_s - \mathbf{F}_s\boldsymbol{\beta}_{s-1})'\mathbf{W}_s^{-1}(\boldsymbol{\beta}_s - \mathbf{F}_s\boldsymbol{\beta}_{s-1}) \end{aligned} \quad (1.47)$$

when both $\boldsymbol{\beta}_0$ and $\{\mathbf{w}_t\}$ in (1.34) follow the Gaussian distribution. Maximizing the posterior log likelihood (1.47) yields to the following recursions [22].

Prediction:

$$\begin{aligned} \boldsymbol{\beta}_{t|t-1} &= \mathbf{F}_t\boldsymbol{\beta}_{t-1|t-1} \\ \boldsymbol{\beta}_{0|0} &= \mathbf{b}_0 \\ \mathbf{P}_{t|t-1} &= \mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t' + \mathbf{W}_t \\ \mathbf{P}_{0|0} &= \mathbf{W}_0, \end{aligned} \quad (1.48)$$

for $t = 1, 2, \ldots, N$.

Filtering:

$$\begin{aligned} \boldsymbol{\beta}_{t|t} &= \boldsymbol{\beta}_{t|t-1} + \mathbf{P}_{t|t}\mathbf{S}_t \\ \mathbf{P}_{t|t} &= \left(\mathbf{P}_{t|t-1}^{-1} + \mathbf{G}_t\right)^{-1} \end{aligned} \quad (1.49)$$

for $t = 1, 2, \ldots, N$, where $\mathbf{S}_t = \partial \log f(y_t \mid \boldsymbol{\beta}_t, \mathcal{F}_{t-1})/\partial\boldsymbol{\beta}_t$ and $\mathbf{G}_t = -\mathrm{E}\left[\partial^2 \log f(y_t \mid \boldsymbol{\beta}_t, \mathcal{F}_{t-1})/\partial\boldsymbol{\beta}_t\partial\boldsymbol{\beta}_t'\right]$, that is the score and the expected information matrix of Y_t evaluated at $\boldsymbol{\beta}_{t|t-1}$. When the data follow (1.32) then an application of the matrix inversion lemma in Problem 5 leads to an equivalent representation of (1.49) (Problem 9).

Smoothing:

$$\begin{aligned} \boldsymbol{\beta}_{t-1|N} &= \boldsymbol{\beta}_{t-1|t-1} + \mathbf{B}_t\left(\boldsymbol{\beta}_{t|N} - \boldsymbol{\beta}_{t|t-1}\right) \\ \mathbf{P}_{t-1|N} &= \mathbf{P}_{t-1|t-1} + \mathbf{B}_t\left(\mathbf{P}_{t|N} - \mathbf{P}_{t|t-1}\right)\mathbf{B}_t' \end{aligned} \quad (1.50)$$

for $t = N, N-1, \ldots, 1$ and $\mathbf{B}_t = \mathbf{P}_{t-1|t-1}\mathbf{F}_t'\mathbf{P}_{t|t-1}^{-1}$. Derivations of the recursions (1.49) and (1.50) are worked out in [22], [25].

An application of posterior mode estimation to categorical time series is discussed in [23]. Some additional results regarding hyperparameters can be found in [27] and [82]. Related work in [28] considers the problem of approximating posterior moments by applying a Gauss–Hermite procedure.

1.3.2.3 Summary We have described several recursive estimation methods for non-linear and non-normal state space models some of which are based on approximations. Several additional results are reported in Problems 7 and 8. Over the last decade important advances in computing methods and power have led to a rapid progress in inference for state space models based on simulation methods, a topic that at present is still in a development stage and hence cannot be fully described. The next section discusses some key ideas and results from this unfolding research area.

1.4 SIMULATION BASED METHODS FOR STATE SPACE MODELS

Monte Carlo simulation techniques are frequently employed in the evaluation of integrals of the form

$$\int g(\mathbf{y}, \boldsymbol{\theta}) f(\mathbf{y}, \boldsymbol{\theta}) d\mathbf{y}$$

where g is some integrable function and f denotes a probability density. The basic idea is to approximate the above integral by the following sum

$$\frac{1}{m} \sum_{i=1}^{m} g(\mathbf{y}_i, \boldsymbol{\theta})$$

where $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m$ are independent and identically distributed realizations from the density f. Then the weak law of large numbers states that as m increases the sum converges to the desired integral [77]. Starting with Metropolis and Ulam [72] in 1949, at present there is a vast literature including numerous texts devoted to the study of Monte Carlo simulation techniques.

There are potential difficulties with Monte Carlo based inference especially in the context of Bayesian inference where the posterior distribution might include highly correlated parameters or when the number of parameters is prohibitively large. Recent advances in computing power in connection with the introduction of what is known as Markov Chain Monte Carlo (MCMC) methods bypass these problems satisfactorily.

1.4.1 A Brief MCMC Tutorial

Introduced by Metropolis et al. [71], MCMC is applied whenever we wish to obtain samples from a distribution known up to a constant where the calculation of the constant is formidable. This is done by constructing a Markov chain whose stationary distribution is the desired distribution from which we wish to simulate data. An important application of this is in Bayesian estimation where MCMC methods are used in generating samples from the posterior distribution. In what follows, we review the useful MCMC method referred to as Gibbs sampling.

1.4.1.1 The Gibbs Sampling Algorithm The Gibbs sampling algorithm–introduced by [34]–is perhaps the most extensively used MCMC method. To describe the algorithm, suppose that $f(\boldsymbol{\theta} \mid \mathbf{y})$ denotes a posterior distribution for a p-dimensional parameter vector $\boldsymbol{\theta}$ given data $\mathbf{y}$ and consider the so called full conditionals

$$\begin{array}{ll} f(\theta_1 \mid \theta_2, \theta_3, & \ldots, \quad \theta_p, \mathbf{y}) \\ f(\theta_2 \mid \theta_1, \theta_3, & \ldots, \quad \theta_p, \mathbf{y}) \\ \vdots & \\ f(\theta_p \mid \theta_1, \theta_2, & \ldots, \quad \theta_{p-1}, \mathbf{y}) \end{array} \tag{1.51}$$

Under suitable conditions the joint density $f(\boldsymbol{\theta} \mid \mathbf{y})$ is determined by the set of full conditionals (1.51), and simulation of samples from (1.51) facilitates the approximation of the posterior density [13], [33]. More precisely, let $\boldsymbol{\theta}^0$ be a starting value of the parameter vector $\boldsymbol{\theta}$. Then the value of $\boldsymbol{\theta}^1$ is obtained by successive generation of samples from

$$\begin{array}{ll} f(\theta_1^1 \mid \theta_2^0, \theta_3^0, & \ldots, \quad \theta_p^0, \mathbf{y}) \\ f(\theta_2^1 \mid \theta_1^1, \theta_3^0, & \ldots, \quad \theta_p^0, \mathbf{y}) \\ \vdots & \\ f(\theta_p^1 \mid \theta_1^1, \theta_2^1, & \ldots, \quad \theta_{p-1}^1, \mathbf{y}). \end{array} \tag{1.52}$$

That is, sample θ_1^1 from $f(\theta_1^1 \mid \theta_2^0, \theta_3^0, \ldots, \theta_p^0, \mathbf{y})$, then sample θ_2^1 from $f(\theta_2^1 \mid \theta_1^1, \theta_3^0, \ldots, \theta_p^0, \mathbf{y})$, and so on up to θ_p^1 from $f(\theta_p^1 \mid \theta_1^1, \theta_2^1, \ldots, \theta_{p-1}^1, \mathbf{y})$. This gives the first iteration. As the number of iterations $m \to \infty$, the resulting vector $\boldsymbol{\theta}^m$ is a realization from $f(\boldsymbol{\theta} \mid \mathbf{y})$ [85].

The Gibbs sampler algorithm raises several issues such as the choice of m for the number of iterations, the selection of the *burn–in* period (that is dropping the first few iterations due to dependence), taking into account the Markov dependence of the realizations when forming estimators of the posterior moments, choice of starting values, convergence criteria and diagnostic tools; see [31], [37] among others. Moreover, the algorithm requires knowledge of the full conditional distributions. When the full conditionals do not have a known form then the algorithm may be combined with the *rejection sampling* algorithm [77] or with the *adaptive rejection sampling* [38].

1.4.1.2 The Metropolis–Hastings Algorithm An alternative MCMC method to the Gibbs sampling algorithm when the full conditional distributions are not available in closed from is the Metropolis–Hastings algorithm originally introduced by Metropolis and Ulam [72] and later generalized by Hastings [47]. The algorithm is outlined in [16].

1.4.2 MCMC Inference for State Space Models

Recall that a state space model is defined through the conditional distribution of the response given the unobserved state vector, say $f(y_t \mid \boldsymbol{\beta}_t)$, the transition density for the unobserved states $f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1})$, and the initial distribution for $\boldsymbol{\beta}_0$, $f(\boldsymbol{\beta}_0)$. As before, set $\boldsymbol{\beta}_0^N = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N)'$. The problem of computing the posterior density $f(\boldsymbol{\beta}_0^N \mid \mathcal{F}_N)$ can be approached by appealing to the Gibbs sampling algorithm whose implementation requires the full conditional densities $f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{-t}, \mathcal{F}_N)$ for $t = 1, \ldots, N$, [12]. The notation $\boldsymbol{\beta}_{-t}$ denotes the set of random variables $\{\boldsymbol{\beta}_s, s \neq t\}$. According to (1.45) we obtain that

$$f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{-t}, \mathcal{F}_N) \propto \begin{cases} f(\boldsymbol{\beta}_{t+1} \mid \boldsymbol{\beta}_t) f(\boldsymbol{\beta}_t) & \text{if } t = 0 \\ f(\boldsymbol{\beta}_{t+1} \mid \boldsymbol{\beta}_t) f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}) f(y_t \mid \boldsymbol{\beta}_t) & \text{if } t = 1, \ldots, N-1 \\ f(y_t \mid \boldsymbol{\beta}_t) f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}) & \text{if } t = N \end{cases} \tag{1.53}$$

In some cases it is simple to sample from the above densities. However there are instances when a rejection sampling or the Metropolis–Hastings algorithm is nested

within each Gibbs sampler iteration for obtaining the full conditionals due to the fact that (1.53) is known up to a constant [12].

The following simple but motivating example illustrates the methodology. Consider the so called *random walk* model for $t = 1, \ldots, N$

$$\begin{aligned} Y_t &= \beta_t + u_t \\ \beta_t &= \beta_{t-1} + w_t \end{aligned} \tag{1.54}$$

where u_t are iid $\mathcal{N}(0, \sigma_u^2)$ and w_t are iid $\mathcal{N}(0, \sigma_w^2)$. Here both observed and unobserved states are assumed scalars. Suppose that $\boldsymbol{\beta}_0$ follows a normal distribution with known mean and variance, say b_0 and σ_0^2, respectively. Suppose further that both σ_u^2 and σ_w^2 follow the inverse Gamma distribution with parameters a_u, b_u and a_w, b_w, respectively. That is

$$f(\sigma_u^2) = \frac{1}{b_u^{a_u}\Gamma(a_u)} \left(\frac{1}{\sigma_u^2}\right)^{a_u+1} \exp\left(-\frac{1}{b_u\sigma_u^2}\right),$$

and similarly for σ_w^2. The target posterior distribution is given by $f(\boldsymbol{\beta}_0^N, \sigma^2, \sigma_w^2 \mid \mathcal{F}_N)$. Following the Gibbs sampler, it is sufficient to draw samples from the following densities

- $f(\sigma_u^2 \mid \boldsymbol{\beta}_0^N, \sigma_w^2, \mathcal{F}_N)$
- $f(\sigma_w^2 \mid \boldsymbol{\beta}_0^N, \sigma_v^2, \mathcal{F}_N)$
- $f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{-t}, \sigma_u^2, \sigma_w^2, \mathcal{F}_N)$.

The first two densities are easily sampled since the inverse Gamma distribution is conjugate to the normal with respect to the precision parameter (1/variance). Indeed,

$$\begin{aligned} f(\sigma_u^2 \mid \boldsymbol{\beta}_0^N, \sigma_w^2, \mathcal{F}_N) &\propto f(\sigma_u^2)\left[\prod_{t=1}^N f(y_t \mid \boldsymbol{\beta}_t, \sigma_u^2)\right] \\ &\propto \left(\frac{1}{\sigma_u^2}\right)^{a_u+1} \exp\left(-\frac{1}{b_u\sigma_u^2}\right) \\ &\times \left(\frac{1}{\sigma_u^2}\right)^{N/2} \exp\left(-\frac{1}{2\sigma_u^2}\sum_{t=1}^N (y_t - \beta_t)^2\right). \end{aligned}$$

The first two factors are due to the functional form of the inverse Gamma while the other two appear as a consequence of the observation model from (1.54). This

calculation shows that the full conditional of σ_u^2 is inverse Gamma with parameters $a_u + N/2$ and $\left(1/b_u + \sum_{t=1}^n (y_t - \beta_t)^2/2\right)^{-1}$. Similarly the full conditional of σ_w^2 is inverse Gamma with parameters $a_w + N/2$ and $\left(1/b_w + \sum_{t=1}^n (\beta_t - \beta_{t-1})^2/2\right)^{-1}$. Turning now to the full conditional of the state parameter given the rest, (1.53) shows that

- for $t = 0$

$$\begin{aligned} & f(\beta_0 \mid \boldsymbol{\beta}_{-0}, \sigma_u^2, \sigma_w, \mathcal{F}_N) \propto f(\beta_1 \mid \beta_0) f(\beta_0) \\ & \propto \exp\left[-\frac{1}{2\sigma_w^2}(\beta_1 - \beta_0)^2 - \frac{1}{2\sigma_0^2}(\beta_0 - b_0)^2\right] \\ & \propto \exp\left\{-\frac{1}{2}\left[\left(\frac{1}{\sigma_w^2} + \frac{1}{\sigma_0^2}\right)\beta_0^2 - 2\left(\frac{\beta_1}{\sigma_w^2} + \frac{b_0}{\sigma_0^2}\right)\beta_0\right]\right\}, \end{aligned}$$

  which shows that this is a normal density with mean $(\beta_1/\sigma_w^2 + b_0/\sigma_0^2)(1/\sigma_w^2 + 1/\sigma_0^2)^{-1}$ and variance $(1/\sigma_w^2 + 1/\sigma_0^2)^{-1}$.

- for $t = 1, \ldots, N-1$

$$\begin{aligned} & f(\beta_t \mid \boldsymbol{\beta}_{-t}, \sigma_u^2, \sigma_w, \mathcal{F}_N) \\ & \propto f(\beta_{t+1} \mid \beta_t, \sigma_w^2) f(\beta_t \mid \beta_{t-1}, \sigma_w^2) f(y_t \mid \beta_t, \sigma_u^2) \\ & \propto \exp\left[-\frac{(\beta_{t+1} - \beta_t)}{2\sigma_w^2} - \frac{(\beta_t - \beta_{t-1})}{2\sigma_w^2} - \frac{(y_t - \beta_t)}{2\sigma_u^2}\right] \\ & \propto \exp\left\{-\frac{1}{2}\left[\left(\frac{2}{\sigma_w^2} + \frac{1}{\sigma_u^2}\right)\beta_t^2 - 2\left(\frac{(\beta_{t+1} + \beta_{t-1})}{\sigma_w^2} + \frac{y_t}{\sigma_u^2}\right)\beta_t\right]\right\}, \end{aligned}$$

  that is a normal distribution with mean $((\beta_{t+1} + \beta_{t-1})/\sigma_w^2 + y_t/\sigma_u^2)(2/\sigma_w^2 + 1/\sigma_u^2)^{-1}$ and variance $(2/\sigma_w^2 + 1/\sigma_u^2)^{-1}$.

- For $t = N$, it can be shown in the same way that the conditional distribution of β_N given the rest of the parameters is normal with mean $(\beta_{N-1}/\sigma_w^2 + y_N/\sigma_u^2)(1/\sigma_w^2 + 1/\sigma_u^2)^{-1}$ and variance $(1/\sigma_w^2 + 1/\sigma_u^2)^{-1}$.

It follows that a after a large number of iterations the output of the Gibbs sampler is a random draw from the posterior distribution $f(\boldsymbol{\beta}_0^N, \sigma_u^2, \sigma_w^2 \mid \mathcal{F}_N)$. Running the algorithm only K times leads to $k = 1, \ldots, K$ iid vectors $(\boldsymbol{\beta}_{0,k}^N, \sigma_{u,k}^2, \sigma_{w,k}^2)'$ from the posterior distribution. Hence, an estimate of the smoothing density is given by

$$f(\boldsymbol{\beta}_t \mid \sigma_u^2, \sigma_w^2, \mathcal{F}_N) = \frac{1}{K}\sum_{k=1}^K f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{-t,k}, \sigma_{u,k}^2, \sigma_{w,k}^2, \mathcal{F}_N).$$

For the linear state space model considered above the full conditionals are given in a closed form, a fact which does not hold in general. When closed from conditionals are not available, the Gibbs sampler can be used to approximate posterior means and covariances, provided that these quantities exist.

Although the example is about linear models with random hyperparameters, it also shows that this approach covers a broad class of nonlinear models. However, the above method generates the states in a *single* update, that is the states are generated one at a time which results in slow convergence. Slow convergence of single update methods is a typical problem in the application of this methodology (see [10]). In addition, the method does not allow for sequential updating which means that when a new observation becomes available the algorithm needs to restart from the beginning. Because of these issues alternative approaches have been sought to the problem of simulating samples from the posterior density.

The use of *multiple* update algorithms, where the generation of the states proceeds simultaneously by using the time ordering of the state space model and sampling from $f(\boldsymbol{\beta}_0^N \mid \mathcal{F}_N)$ instead of sampling one at a time from (1.53), has been suggested to address the problem of slow convergence [10], [11], [29] [79]. At least empirically, the convergence of multiple update algorithms is faster when compared to single update algorithms. In particular, efficient MCMC inference for the class of normal dynamic linear models is based on the idea of sampling the entire set of state vectors $\boldsymbol{\beta}_0^N$ which can be accomplished by employing the Markovian structure of the model

$$f(\boldsymbol{\beta}_0^N \mid \mathcal{F}_N) = f(\boldsymbol{\beta}_N \mid \mathcal{F}_N) \prod_{t=1}^{N-1} f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t+1}, \mathcal{F}_{t-1}). \tag{1.55}$$

Equation (1.55) indicates that $\boldsymbol{\beta}_0^N$ can be sampled efficiently by generating $\boldsymbol{\beta}_t$, $t = 0, \ldots, N$ backwards. Furthermore, if the Gaussian linear dynamic model holds, then all the densities in (1.55) are Gaussian so it is sufficient to compute their means and variances. However for $t = N-1, N-2, \ldots, 1, 0$ (see Problem 11)

$$\begin{aligned} \mathrm{E}[\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t+1}, \mathcal{F}_{t-1}] &= \boldsymbol{\beta}_{t|t} + \mathbf{K}_t(\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_{t+1|t}) \\ \mathrm{Var}[\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t+1}, \mathcal{F}_{t-1}] &= \mathbf{P}_{t|t} - \mathbf{K}_t \mathbf{P}_{t+1|t} \mathbf{K}_t' \end{aligned} \tag{1.56}$$

where $\mathbf{K}_t$ is the Kalman gain defined in (1.12). Thus the algorithm for linear dynamic models runs as follows:

1. Run the standard Kalman filter for $t = 0, \ldots, N$ to obtain prediction and filtering estimates.

2. Sample $\boldsymbol{\beta}_N$ form the normal distribution with mean $\boldsymbol{\beta}_{N|N}$ and variance $\mathbf{P}_{N|N}$.

3. Sample $\boldsymbol{\beta}_t$ for $t = N-1, N-2, , \ldots, 1, 0$ from the normal distribution with mean and variance given by (1.56), respectively.

This algorithm is termed forward–filtering (step 1) backward–sampling (steps 2 and 3).

An intermediate approach between single and multiple update algorithms divides the state vector into several blocks and is called *block move* algorithm [80]. Another alternative to the problem of slow convergence reparametrizes the model in terms of independent system disturbances and is applicable to time series that follow generalized linear models [**?**] (see also [**?**]). Both of these methods require Fisher scoring steps which can be avoided according to a block move algorithm proposed in [**?**]. More recently Durbin and Koopman [21] building on an earlier work in [**?**] consider the analysis of non–Gaussian state space models using importance sampling and antithetic variables without resorting to any MCMC methods.

The introduction of MCMC techniques had a profound impact on the analysis of state space model over the last decade or so as we see from the following examples. Thus, Fahrmeir et al. [24] apply the single move Gibbs sampler to non–Gaussian observations, West [**?**] considers importance sampling algorithms and adaptive density estimators for inference in dynamic nonlinear models, Carlin and Polson [**?**] apply MCMC methodology in modeling categorical time series, Albert and Chib [**?**] study autoregressive time series subject to regime switching, and McCulloch and Tsay [**?**] demonstrate the applicability of the Gibbs sampling algorithm to random level–shift models, additive outliers and missing values in autoregression. Furthermore, Jacquier et al. [**?**] employ MCMC inference in the analysis of stochastic volatility models, Chib and Greenberg [**?**] and Chib [15] extend the work of Carter and Kohn [10] on Gibbs sampling, and Muller et al. [**?**] discuss Bayesian mixture models for nonlinear analysis of autoregressive process. More recently Cargnoni et al. [9] apply MCMC inference to multinomial observations, Gerlach et al. [**?**] propose an efficient MCMC

algorithm for the estimation of linear Gaussian state space models generalizing results in [11], and Frühwirth-Schnatter [30] introduces permutation sampling in switching and mixture models.

1.4.3 Sequential Monte Carlo Sampling Methods

The preceding discussion advocates the use of multiple update algorithms instead of single update algorithms as far as convergence is concerned. However multiple update algorithms are not recursive as they do not allow sequential processing of the data. To amend this, Gordon et al. [39] and Kitagawa [63], among others, suggested the so called *particle filter* method whose inception can be traced back to [**?**] and [**?**]. The recent review article by Doucet et al.[19] and the recent collections in [20] and [66] manifest the importance of this methodology and list additional references.

Following [19] and recalling that $\boldsymbol{\beta}_0^t = (\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_t)'$, $\mathcal{F}_t = \{Y_1, \ldots, Y_t\}$ and (1.21) (1.22), the following calculation leads to a recursive formula for $f(\boldsymbol{\beta}_0^{t+1} \mid \mathcal{F}_{t+1})$:

$$\begin{aligned} f(\boldsymbol{\beta}_0^{t+1} \mid \mathcal{F}_{t+1}) &= \frac{f(\mathcal{F}_{t+1} \mid \boldsymbol{\beta}_0^{t+1}) f(\boldsymbol{\beta}_0^{t+1})}{f(\mathcal{F}_{t+1})} \\ &= \frac{f(\mathcal{F}_t \mid \boldsymbol{\beta}_0^t) f(y_{t+1} \mid \boldsymbol{\beta}_{t+1}) f(\boldsymbol{\beta}_0^t) f(\boldsymbol{\beta}_{t+1} \mid \boldsymbol{\beta}_t)}{f(y_{t+1} \mid \mathcal{F}_t) f(\mathcal{F}_t)} \\ &= f(\boldsymbol{\beta}_0^t \mid \mathcal{F}_t) \frac{f(y_{t+1} \mid \boldsymbol{\beta}_{t+1}) f(\boldsymbol{\beta}_{t+1} \mid \boldsymbol{\beta}_t)}{f(y_{t+1} \mid \mathcal{F}_t)}. \end{aligned} \tag{1.57}$$

The first equation is based on Bayes theorem and the second on model assumptions (1.21) and (1.22). The denominator of (1.57) cannot be calculated in closed form but an appeal to Bayesian importance sampling (Problem 12) which consists of obtaining a sample from an importance density $\pi()$ with the property

$$\pi(\boldsymbol{\beta}_0^t \mid \mathcal{F}_t) = \pi(\boldsymbol{\beta}_0) \prod_{s=1}^{t} \pi(\boldsymbol{\beta}_s \mid \boldsymbol{\beta}_0^{s-1}, \mathcal{F}_s) \tag{1.58}$$

yields the following algorithm [19].

Sequential Importance Sampling (SIS):

Let $t = 0, 1, 2, \ldots \ldots, N$.

- For $i = 1, \ldots, n$ draw a sample $\boldsymbol{\beta}_{t,i}$ from $\pi(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{0,i}^{t-1}, \mathcal{F}_t)$ and put $\boldsymbol{\beta}_{0,i}^t = (\boldsymbol{\beta}_{0,i}^{t-1}, \boldsymbol{\beta}_{t,i})$.

- For $i = 1, \ldots, n$ calculate the importance weights up to a normalizing constant

$$\hat{w}_{t,i} = \hat{w}_{t-1,i} \frac{f(y_t \mid \boldsymbol{\beta}_{t,i}) f(\boldsymbol{\beta}_{t,i} \mid \boldsymbol{\beta}_{t-1,i})}{\pi(\boldsymbol{\beta}_{t,i} \mid \boldsymbol{\beta}_{0,i}^{t-1}, \mathcal{F}_t)}. \tag{1.59}$$

- For $i = 1, \ldots, n$, normalize the importance weights by

$$w_{t,i} = \frac{\hat{w}_{t,i}}{\sum_{l=1}^{n} \hat{w}_{t,l}}. \tag{1.60}$$

It is instructive to consider the above recursions in some detail. Suppose that at time $t = 0$, a sample of size n is available from $\pi(\boldsymbol{\beta}_0)$. Then (1.59) implies that the importance weights are given up to a normalizing constant by

$$\hat{w}_{0,i} = \frac{f(\boldsymbol{\beta}_{0,i})}{\pi(\boldsymbol{\beta}_{0,i})}$$

and therefore the normalizing weights are

$$w_{0,i} = \frac{\hat{w}_{0,i}}{\sum_{l=1}^{n} \hat{w}_{0,l}}.$$

That is the standard importance sampling scheme to obtain a random draw from $f(\boldsymbol{\beta}_0)$ since it is the desired target density at $t = 0$. At time $t = 1$, draw a random sample of size n from the importance density $\pi(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_{0,i}, \mathcal{F}_1)$ and set

$$\hat{w}_{1,i} = \hat{w}_{0,i} \frac{f(y_1 \mid \boldsymbol{\beta}_{1,i}) f(\boldsymbol{\beta}_{1,i} \mid \boldsymbol{\beta}_{0,i})}{\pi(\boldsymbol{\beta}_{1,i} \mid \boldsymbol{\beta}_{0,i}, \mathcal{F}_1)}.$$

The recursion is a consequence of (1.57) which states that the joint posterior density of $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$ is proportional to the product of $f(\boldsymbol{\beta}_0)$ and $f(y_1 \mid \boldsymbol{\beta}_1) f(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_0)$. Since the denominator of (1.57) cannot be calculated explicitly, the importance weights needs to be calculated up to a normalizing constant first and then be normalized. Hence (1.59) and (1.60) follow and therefore $(\boldsymbol{\beta}_{0,i}, \boldsymbol{\beta}_{1,i})$ is a random sample from $\pi(\boldsymbol{\beta}_0^1 \mid \mathcal{F}_1)$. The normalized weights at $t = 1$ are given by

$$w_{1,i} = \frac{\hat{w}_{1,i}}{\sum_{l=1}^{n} \hat{w}_{1,l}},$$

Having available the normalized weights $\{w_{1,i}\}$ and the random sample $(\boldsymbol{\beta}_{0,i}, \boldsymbol{\beta}_{1,i})$, $i = 1, \ldots, n$, estimation of the following integral

$$I = \int \mathbf{h}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1) f(\boldsymbol{\beta}_0^1 \mid \mathcal{F}_1) d\boldsymbol{\beta}_0^1$$

is accomplished by the sum

$$\sum_{i=1}^{n} \mathbf{h}(\boldsymbol{\beta}_{0,i}, \boldsymbol{\beta}_{1,i}) w_{1,i},$$

provided that the integral exists.

The iteration process continues until time $t = N$, leading to a sample which is used in the estimation of expectations of $f(\boldsymbol{\beta}_0^N \mid \mathcal{F}_N)$. The above discussion points to the advantage of sequential importance sampling when compared to MCMC methods. Namely, sequential importance sampling is an *on–line* estimation procedure with the advantage that when a new observation becomes available the process need not start from the beginning.

The choice of the importance function is crucial and it is sensible to choose an importance function so that the variance of the importance weights (1.59) conditional on $\boldsymbol{\beta}_{0,i}^{t-1}$ and $\mathcal{F}_t$ is minimized. Otherwise the algorithm is degenerate in the sense that all but one of the normalized weights (1.60) approach zero after a few iterations. It can be proved that $\pi(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{0,i}^{t-1}, \mathcal{F}_t) = f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{0,i}^{t-1}, \mathcal{F}_t)$ is the *optimal* importance function (Problem 13). This choice of importance function has been considered by various authors including [14]. However, there are limitations on the choice of this importance function and several alternative strategies have been considered [6], [19], [63], [69], [89].

To avoid degeneracy of the sequential importance sampling algorithm several authors have considered resampling methods where the key idea is to resample from the generated sample under a predetermined condition [75]. See also [19], [48], [52], [67], [69], and [73] among others.

The simulated sample derived either by sequential importance sampling or by any other resampling method is used for the purpose of prediction, filtering, and smoothing. Consider the problem of filtering, that is estimation of the density $f(\boldsymbol{\beta}_t \mid \mathcal{F}_t)$. Using (1.60), an estimate of $f(\boldsymbol{\beta}_0^t \mid \mathcal{F}_t)$ is given by

$$\hat{f}(\boldsymbol{\beta}_0^t) = \sum_{i=1}^{n} w_{t,i} \delta_{\boldsymbol{\beta}_{0,i}^t}(\boldsymbol{\beta}_0^t),$$

for *any* $t = 0, \ldots, N$, where $\delta_{\boldsymbol{\beta}_{0,i}^t}(\boldsymbol{\beta}_0^t)$ denotes a point mass at the vector point. Consequently an estimator of the filtering density $f(\boldsymbol{\beta}_t \mid \mathcal{F}_t)$ is obtained by keeping

only the corresponding simulated component $\boldsymbol{\beta}_{t,i}$,

$$\hat{f}(\boldsymbol{\beta}_t) = \sum_{i=1}^{n} w_{t,i}\delta_{\boldsymbol{\beta}_{t,i}}(\boldsymbol{\beta}_t).$$

Similarly the prediction density $f(\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1})$ is approximated by

$$\hat{f}(\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1}) = \sum_{i=1}^{n} w_{t-1,i}\delta_{\boldsymbol{\beta}_{t,i}}(\boldsymbol{\beta}_t)$$

where $\boldsymbol{\beta}_{t,i}$ has been drawn from $f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1,i})$. Similar recursions hold for smoothing [19].

1.4.4 Likelihood Inference

In applications both observation and transition densities may depend on unknown parameters as illustrated by the example in Section 1.4.2. Thus, suppose that the observation density $f(y_t; \boldsymbol{\theta}_1 \mid \boldsymbol{\beta}_t, \mathcal{F}_{t-1})$ depends on $\boldsymbol{\theta}_1$, and the transition density $f(\boldsymbol{\beta}_t, \boldsymbol{\theta}_2 \mid \boldsymbol{\beta}_{t-1})$ depends on $\boldsymbol{\theta}_2$. Regarding both $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ as random variables with independent prior distributions $f(\boldsymbol{\theta}_1)$ and $f(\boldsymbol{\theta}_2)$, respectively, then the full conditionals

$$f(\boldsymbol{\theta}_1 \mid \boldsymbol{\beta}_0^N, \mathcal{F}_N, \boldsymbol{\theta}_2) \propto f(\boldsymbol{\theta_1}) \left[\prod_{t=1}^{N} f(y_t; \boldsymbol{\theta}_1 \mid \boldsymbol{\beta}_t)\right]$$

and

$$f(\boldsymbol{\theta}_2 \mid \boldsymbol{\beta}_0^N, \mathcal{F}_N, \boldsymbol{\theta}_2) \propto f(\boldsymbol{\theta_2}) \left[f(\boldsymbol{\beta}_0, \boldsymbol{\theta}_2)\prod_{t=1}^{N} f(\boldsymbol{\beta}_t; \boldsymbol{\theta}_2 \mid \boldsymbol{\beta}_{t-1})\right]$$

can be used in inference about $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. This approach was already discussed in Section 1.4.2 where the MCMC methodology was applied to a linear state space model. When in addition the response belongs to the exponential family then a good choice of the conjugate prior eases the computation of the posterior density.

Regarding the general likelihood

$$\begin{aligned} f(y_1, \ldots, y_N) &= \prod_{t=1}^{N} f(y_t \mid \mathcal{F}_{t-1}) \\ &= \prod_{t=1}^{N} \int f(y_t \mid \boldsymbol{\beta}_t) f(\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1}) d\boldsymbol{\beta}_t, \end{aligned} \tag{1.61}$$

MCMC inference can be avoided as noted by several authors who consider approximating the likelihood by sequential importance sampling as described in Section 1.4.3. See [19], [52], and [64]. The asymptotic behavior of the maximum likelihood estimator for state space models has been discussed in [53].

1.4.5 Longitudinal Data

State space modeling is also useful in the analysis of longitudinal data. Jones [54] shows that the linear mixed model has a state space representation and develops its inference under normality. More recently, multivariate state space models for mixed responses–both continuous and discrete responses–using exponential dispersion models have been studied in [55], and in [56] the authors consider a state space model for *multivariate* count data (Problem 14). The recent volume [17] includes further results on generalized linear models for longitudinal data from a Bayesian perspective.

1.5 KALMAN FILTERING IN SPACE–TIME DATA

Recently there has been a growing interest in the application and extension of dynamic models to space–time data such as environmental data. Non–Bayesian models for space–time data have been considered by Huang and Cressie [50] who analyze snow–water equivalent data using a vector autoregressive process with spatially independent innovations, and by Mardia et al. [70] who combine kriging and state space models. Wikle and Cressie [**?**] (see also [8]) propose a model suitable for a large number of sites whereby the observations are generated as a sum of an unobserved process which incorporates space–time dependence and an unobserved process which is spatially and temporally uncorrelated. Kalman recursions within the Bayesian framework are discussed in [**?**], [78], [84], and [92] among others.

1.6 PROBLEMS AND COMPLEMENTS

1. Verify by substitution that the ARMA$(1,1)$ process

$$Y_t = \phi Y_{t-1} + \theta v_{t-1} + v_t, \quad t = 0, \pm 1, \pm 2, ...,$$

has the state space representation

$$\begin{aligned} Y_t &= \beta_t + v_t \\ \beta_t &= \phi\beta_{t-1} + (\theta + \phi)v_{t-1} \end{aligned}$$

2. Consider a slight reinterpretation of the general state space model (1.22), (1.23) as in

General Observation equation:	$Y_t \mid \boldsymbol{\beta}_t \sim f(y_t \mid \boldsymbol{\beta}_t, \mathcal{F}_{t-1})$
General System equation:	$\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1} \sim f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1})$
Initial information:	$\boldsymbol{\beta}_0 \sim f(\boldsymbol{\beta}_0 \mid \mathcal{F}_0) \equiv f(\boldsymbol{\beta}_0)$.

Prove the following recursions in k:

(a) Prediction densities for the states

$$f(\boldsymbol{\beta}_{t+k} \mid \mathcal{F}_t) = \int f(\boldsymbol{\beta}_{t+k} \mid \boldsymbol{\beta}_{t+k-1}) f(\boldsymbol{\beta}_{t+k-1} \mid \mathcal{F}_t) d\boldsymbol{\beta}_{t+k-1}.$$

(b) Prediction densities for the observations

$$f(y_{t+k} \mid \mathcal{F}_t) = \int f(y_{t+k} \mid \boldsymbol{\beta}_{t+k}) f(\boldsymbol{\beta}_{t+k} \mid \mathcal{F}_t) d\boldsymbol{\beta}_{t+k}.$$

Further results and alternative computational methods for filtering and smoothing densities are given in [68].

3. Suppose that the vector $\mathbf{Y}_t = (Y_{t1}, \ldots, Y_{tk})'$ follows the multinomial distribution

$$f(\mathbf{y}_t \mid \boldsymbol{\pi}_t) = \frac{n_t!}{\prod_{j=1}^k y_{tj}!} \prod_{j=1}^k \pi_{tj}^{y_{tj}},$$

for $t = 1, \ldots, N$, where $n_t = \sum_{j=1}^{k} Y_{tj}$ and $\boldsymbol{\pi}_t = (\pi_{t1}, \ldots, \pi_{tk})'$. Suppose that the distribution of $\boldsymbol{\pi}_{t-1}$ given $\mathcal{F}_{t-1}$ is Dirichlet, that is

$$f(\boldsymbol{\pi}_{t-1} = \boldsymbol{\pi} \mid \mathcal{F}_{t-1}) = C(\mathbf{r}_t) \prod_{j=1}^{k} \pi^{r_{tj}},$$

where $\mathbf{r}_t = (r_{t1}, \ldots, r_{tk})'$ and $\boldsymbol{\pi}$ belongs to k–dimensional simplex.

(a) Calculate the distribution of $\mathbf{Y}_t$ given $\mathcal{F}_{t-1}$.

(b) Calculate the distribution of $\boldsymbol{\pi}_t$ given $\mathcal{F}_t$.

4. *Linear Bayes Methodology* [43], [90, Ch. 4.9.2]. Assume that the joint distribution of data $\mathbf{Y}$ and a parameter $\boldsymbol{\theta}$ is partially specified in terms of the first and second moments such that

$$\mathrm{E}\left\{(\boldsymbol{\theta}, \mathbf{Y})'\right\} = (\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\mu}_{\mathbf{Y}})',$$

and

$$\mathrm{Var}\left\{(\boldsymbol{\theta}, \mathbf{Y})'\right\} = \begin{pmatrix} \boldsymbol{\Sigma}_\theta & \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{Y}} \\ \boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{A}' & \boldsymbol{\Sigma}_{\mathbf{Y}} \end{pmatrix}.$$

Since the posterior risk cannot be calculated, consider the overall risk

$$r(\mathbf{d}) = \mathrm{trace}\mathrm{E}\left[(\boldsymbol{\theta} - \mathbf{d})(\boldsymbol{\theta} - \mathbf{d})'\right],$$

where the expectation is taken *unconditional* on $\mathbf{Y}$. Show that if $\mathbf{d}$ is a linear function of the data such that $\mathbf{d} = \mathbf{h} + \mathbf{H}\mathbf{Y}$, then a linear Bayes estimate of $\boldsymbol{\theta}$ in the sense of minimizing $r(\mathbf{d})$ is given by

$$\mathbf{m} = \boldsymbol{\mu}_{\boldsymbol{\theta}} + \mathbf{A}(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})$$

with associated risk matrix

$$\mathbf{C} = \boldsymbol{\Sigma}_\theta - \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{A}'.$$

Thus the minimum risk is equal to the trace of $\mathbf{C}$.

5. *Matrix inversion lemma.* Let $\mathbf{P}$, $\mathbf{U}$, $\mathbf{R}$ be $p \times p$, $p \times p$ and $k \times k$ symmetric matrices, respectively, and $\mathbf{H}$ an arbitrary $k \times p$ matrix. Assuming that the

indicated inverses exist, prove the following matrix relations [65, p. 85], [74, p. 33].

(a) $\left(\mathbf{P}^{-1}+\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}\right)^{-1}=\mathbf{P}-\mathbf{P}\mathbf{H}'\left(\mathbf{H}\mathbf{P}\mathbf{H}'+\mathbf{R}\right)^{-1}\mathbf{H}\mathbf{P}$

(b) $\left(\mathbf{P}^{-1}+\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{R}^{-1}=\mathbf{P}\mathbf{H}'\left(\mathbf{H}\mathbf{P}\mathbf{H}'+\mathbf{R}\right)^{-1}$

(c) $\left(\mathbf{P}^{-1}-\mathbf{U}^{-1}\right)^{-1}=\mathbf{P}\left(\mathbf{U}-\mathbf{P}\right)^{-1}\mathbf{U}$

(d) $\mathbf{U}+\left(\mathbf{P}^{-1}-\mathbf{U}^{-1}\right)^{-1}=\mathbf{U}\left(\mathbf{U}-\mathbf{P}\right)^{-1}\mathbf{U}$

6. *Lag-One Covariance Smoother* [81, p. 320]. With the Kalman gain $\mathbf{K}_t$ (1.12), $\mathbf{B}_t$ in (1.15), and

$$\mathbf{P}_{t_1,t_2|s}=\mathrm{E}[(\boldsymbol{\beta}_{\mathrm{t_1}}-\boldsymbol{\beta}_{\mathrm{t_1|s}})(\boldsymbol{\beta}_{\mathrm{t_2}}-\boldsymbol{\beta}_{\mathrm{t_2|s}})'],$$

verify that

$$\mathbf{P}_{n,n-1|n}=(\mathbf{I}-\mathbf{K}_n\mathbf{z}_n')\mathbf{F}_t\mathbf{P}_{n-1|n-1}$$

and show that for $t=n,n-1,...,2$,

$$\mathbf{P}_{t-1,t-2|n}=\mathbf{P}_{t-1|t-1}\mathbf{B}_{t-1}'+\mathbf{B}_t[\mathbf{P}_{t,t-1|n}-\mathbf{F}_t\mathbf{P}_{t-1|t-1}]\mathbf{B}_{t-1}'.$$

7. *Extended Kalman Filter* [3, Ch. 8.2], [90, Ch. 13.2]. Consider the following *non-linear* state space model

$$\begin{aligned} Y_t &= \mathbf{h}_t(\boldsymbol{\beta}_t)+v_t \\ \boldsymbol{\beta}_t &= \mathbf{f}_t(\boldsymbol{\beta}_{t-1})+\mathbf{w}_t, \end{aligned} \tag{1.62}$$

for $t=1,\ldots,N$ and known functions $\mathbf{h}_t$, $\mathbf{f}_t$. Assume that $\{v_t\}$ and $\{\mathbf{w}_t\}$ are independent and both follow Gaussian distributions $\mathcal{N}(0,\sigma_t^2)$ and $\mathcal{N}(\mathbf{0},\mathbf{W}_t)$, respectively. Furthermore, suppose that $\boldsymbol{\beta}_0$ follows the normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{W}_0$ and is independent of $\{v_t\}$ and $\{\mathbf{w}_t\}$. By assuming differentiability of $\mathbf{h}_t$ and $\mathbf{f}_t$ and using Taylor expansion show that model (1.62) is equivalent to

$$\begin{aligned} Y_t &= \mathbf{H}_t'\boldsymbol{\beta}_{t-1}+\left(\mathbf{h}_t(\boldsymbol{\beta}_{t|t-1})-\mathbf{H}_t'\boldsymbol{\beta}_{t|t-1}\right)+v_t, \\ \boldsymbol{\beta}_t &= \mathbf{F}_t\boldsymbol{\beta}_{t-1}+\left(\mathbf{f}_t(\boldsymbol{\beta}_{t-1|t-1})-\mathbf{F}_t\boldsymbol{\beta}_{t-1|t-1}\right)+\mathbf{w}_t, \end{aligned} \tag{1.63}$$

where

$$\mathbf{F}_t = \frac{\partial \mathbf{f}_t(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}_{t-1|t-1}} \text{ and } \quad \mathbf{H}_t' = \frac{\partial \mathbf{h}_t(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}_{t|t-1}}.$$

Thus conclude that the so called extended Kalman filter is given by the following recursions.

$$\begin{aligned}
\boldsymbol{\beta}_{t|t-1} &= \mathbf{f}_t(\boldsymbol{\beta}_{t-1|t-1}) \\
\mathbf{P}_{t|t-1} &= \mathbf{F}_t \mathbf{P}_{t-1|t-1} \mathbf{F}_t' + \mathbf{W}_t \\
\boldsymbol{\beta}_{t|t} &= \boldsymbol{\beta}_{t|t-1} + \mathbf{B}_t \left(Y_t - \mathbf{h}_t(\boldsymbol{\beta}_{t|t-1}) \right) \\
\mathbf{P}_{t|t} &= \left(\mathbf{I}_p - \mathbf{B}_t \mathbf{H}_t' \right) \mathbf{P}_{t|t-1},
\end{aligned}$$

where the Kalman gain is given by

$$\mathbf{B}_t = \mathbf{P}_{t|t-1} \mathbf{H}_t \left(\mathbf{H}_t' \mathbf{P}_{t|t-1} \mathbf{H}_t + \sigma_t^2 \right)^{-1}.$$

8. *Gaussian sum approximation* [1], [3, Ch. 8.4], [90, Ch. 12]. The rationale of the Gaussian sum approximation is the fact that any probability density on R^n is approximated by a Gaussian mixture in $L_1(R^n)$ distance. Thus, let

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \boldsymbol{\phi} \left(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \right),$$

where $\boldsymbol{\phi} \left(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \right)$ is the density of an n–dimensional normal random variable with mean $\boldsymbol{\mu}_i$ and positive definite covariance matrix $\boldsymbol{\Sigma}_i$, $\alpha_i > 0$ such that $\sum_i \alpha_i = 1$, and $f(.)$ is a density function. Show that if $\mathbf{X}$ is a random vector with pdf $f(\mathbf{x})$ then

$$\mathrm{E}\left[\mathbf{X}\right] = \sum_{i=1}^{m} \alpha_i \boldsymbol{\mu}_i \tag{1.64}$$

$$\mathrm{Var}\left[\mathbf{X}\right] = \sum_{i=1}^{m} \alpha_i \left(\boldsymbol{\Sigma}_i + \left(\mathrm{E}\left[\mathbf{X}\right] - \boldsymbol{\mu}_i \right)^2 \right). \tag{1.65}$$

The above equations facilitate calculation of posterior means and covariance matrices due to the next two theorems. Indeed, show that:

(a) With state space model (1.62) and with

$$f(\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1}) = \sum_{i=1}^{m} \alpha_{i(t-1)} \boldsymbol{\phi}(\boldsymbol{\beta}_t, \bar{\boldsymbol{\mu}}_{it}, \bar{\boldsymbol{\Sigma}}_{it})$$

the updated density $f(\boldsymbol{\beta}_t \mid \mathcal{F}_t)$ approaches the Gaussian mixture

$$\sum_{i=1}^{m} \alpha_{it} \boldsymbol{\phi}(\boldsymbol{\beta}_t, \boldsymbol{\mu}_{it}, \boldsymbol{\Sigma}_{it})$$

uniformly in $\boldsymbol{\beta}_t$ and Y_t as $\bar{\boldsymbol{\Sigma}}_{it} \to 0$ for $i = 1, 2, \ldots, m$, where

$$\begin{aligned}
\boldsymbol{\mu}_{it} &= \bar{\boldsymbol{\mu}}_{it} + \mathbf{K}_{it}\left(Y_t - \mathbf{h}_t(\bar{\boldsymbol{\mu}}_{it})\right) \\
\boldsymbol{\Sigma}_{it} &= \bar{\boldsymbol{\Sigma}}_{it} - \bar{\boldsymbol{\Sigma}}_{it}\mathbf{H}_{it}\left(\mathbf{H}'_{it}\bar{\boldsymbol{\Sigma}}_{it}\mathbf{H}_{it} + \sigma_t^2\right)^{-1}\mathbf{H}'_{it}\bar{\boldsymbol{\Sigma}}_{it} \\
\mathbf{K}_{it} &= \bar{\boldsymbol{\Sigma}}_{it}\mathbf{H}_{it}\left(\mathbf{H}'_{it}\bar{\boldsymbol{\Sigma}}_{it}\mathbf{H}_{it} + \mathbf{W}_t\right)^{-1} \\
\alpha_{it} &= \frac{\alpha_{i(t-1)}\boldsymbol{\phi}\left(\mathbf{h}_t(\bar{\boldsymbol{\mu}}_{it}), \mathbf{H}'_{it}\bar{\boldsymbol{\Sigma}}_{it}\mathbf{H}_{it} + \mathbf{W}_t\right)}{\sum_{i=1}^{m}\alpha_{i(t-1)}\boldsymbol{\phi}\left(\mathbf{h}_t(\bar{\boldsymbol{\mu}}_{it}), \mathbf{H}'_{it}\bar{\boldsymbol{\Sigma}}_{it}\mathbf{H}_{it} + \mathbf{W}_t\right)} \\
\mathbf{H}'_{it} &= \frac{\partial \mathbf{h}_t(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\bar{\boldsymbol{\mu}}_{ti}}.
\end{aligned} \tag{1.66}$$

(b) With state space model (1.62) and with

$$f(\boldsymbol{\beta}_t \mid \mathcal{F}_t) = \sum_{i=1}^{m} \alpha_{it}\phi(\boldsymbol{\beta}_t, \bar{\boldsymbol{\mu}}_{it}, \bar{\boldsymbol{\Sigma}}_{it})$$

the one–step ahead prediction density $f(\boldsymbol{\beta}_{t+1} \mid \mathcal{F}_t)$ approaches the Gaussian mixture

$$\sum_{i=1}^{m} \alpha_{it}\phi(\boldsymbol{\beta}_t, \boldsymbol{\mu}_{i(t+1)}, \boldsymbol{\Sigma}_{i(t+1)})$$

uniformly in $\boldsymbol{\beta}_t$ as $\bar{\boldsymbol{\Sigma}}_{it} \to 0$ for $i = 1, 2, \ldots, m$, where

$$\begin{aligned}
\boldsymbol{\mu}_{i(t+1)} &= \mathbf{f}_t(\bar{\boldsymbol{\mu}}_{it}) \\
\boldsymbol{\Sigma}_{i(t+1)} &= \mathbf{F}_{it}\bar{\boldsymbol{\Sigma}}_{it}\mathbf{F}'_{it} + \mathbf{W}_t \\
\mathbf{F}_{it} &= \frac{\partial \mathbf{f}_t(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\bar{\boldsymbol{\mu}}_{ti}}.
\end{aligned} \tag{1.67}$$

Some early references are [42] and [83]. Further properties of mixtures in state space modeling are studied in [90, Ch. 12].

9. An alternative form of the smoothing recursions (1.49) can be derived with the aid of the matrix inversion lemma discussed in Problem 5 and the properties of the exponential family. More specifically show that when the data follow the exponential family of distributions then (1.49) are modified as

$$\begin{aligned}
\boldsymbol{\beta}_{t|t} &= \boldsymbol{\beta}_{t|t-1} + \mathbf{B}_t\left(Y_t - \mu_t\right) \\
\mathbf{P}_{t|t} &= \left(\mathbf{I} - \mathbf{B}_t\frac{\partial \mu_t}{\partial \eta_t}\mathbf{z}'_t\right)\mathbf{P}_{t|t-1}
\end{aligned}$$

where μ_t is the conditional expectation of the response and $\eta_t = \mathbf{z}_t'\boldsymbol{\beta}_{t|t-1}$. The matrix $\mathbf{B}_t$ is referred as *Kalman gain* and is equal to

$$\mathbf{P}_{t|t-1}\mathbf{z}_t\frac{\partial\mu_t}{\partial\eta_t}\left(\mathbf{z}_t'\left(\frac{\partial\mu_t}{\partial\eta_t}\right)^2\mathbf{P}_{t|t-1}\mathbf{z}_t+\sigma_t^2\right)^{-1},$$

where σ_t^2 is the conditional variance of Y_t. Notice that a similar recursion holds for multivariate data.

10. Consider a count time series $\{Y_t\}$, $t = 1, \ldots, N$ such that

$$f(y_t \mid \beta_t) = \frac{\beta_t^{y_t} exp(-\beta_t)}{y_t!}$$

where β_t $t = 1, \ldots, N$ is a sequence of unobserved states. Assume further that the distribution of β_{t-1} is Gamma with parameters $a = a_{t-1|t-1}$ and $b = b_{t-1|t-1}$ such that

$$f(\beta_{t-1} \mid \mathcal{F}_{t-1}) = \frac{\exp(-b\beta_{t-1})\beta_{t-1}^{a-1}}{\Gamma(a)b^a}.$$

Suppose that β_t given $\mathcal{F}_{t-1}$ has also a Gamma distribution with parameters $a_{t|t-1} = \omega a_{t-1|t-1}$ and $b_{t|t-1} = \omega b_{t-1|t-1}$ for some $\omega \in (0, 1]$.

(a) Calculate $\mathrm{E}[\beta_t \mid \mathcal{F}_{t-1}]$ and $\mathrm{Var}[\beta_t \mid \mathcal{F}_{t-1}]$ and compare your answer with the standard dynamic linear model recursions.

(b) Compute $f(\beta_t \mid \mathcal{F}_t)$.

(c) Calculate the predictive p.d.f. $f(y_t \mid \mathcal{F}_{t-1})$ and find its mean and variance.

Dynamic models for count time series with conjugate priors have been considered by Harvey and Fernandes [45] who also develop parallel methodology for binomial, multinomial and negative binomial responses.

11. Under the Gaussian linear state space model show that the means and variances of the conditional densities $f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t+1}, \mathcal{F}_{t-1})$ for $t = N-1, \ldots, 0$ are given by (1.56)

12. *Importance Sampling* [36]. Consider the problem of approximating the following expectation

$$I(\theta) = \int g(x;\theta) f(x) dx$$

with respect to f, assuming that it exists. When it is not possible to sample from $f()$ a sensible approach is to sample from another density, say $h()$, which approximates $f()$ and then use the sampled valued $x_1, \ldots, x_n$ to form the estimator

$$\hat{I}(\theta) = \frac{\sum_{i=1}^n w_i g(x_i, \theta)}{\sum_{i=1}^n w_i},$$

where $w_i = f(x_i)/h(x_i)$. Show that if the support of $h(x)$ includes the support of $f(x)$ then the estimator $\hat{I}(\theta)$ is strongly consistent for $I(\theta)$. Obtain an estimator for the Monte Carlo standard error of $\hat{I}(\theta)$.

13. *Optimal Importance Function*. Show that $\pi(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{0,i}^{t-1}, \mathcal{F}_t) = f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{0,i}^{t-1}, \mathcal{F}_t)$ is the importance function which minimizes the variance of the importance weights (1.59) conditional on $\boldsymbol{\beta}_{0,i}^{t-1}$ and $\mathcal{F}_t$. How do the weights (1.59) transform for this choice of importance function?

14. *A state space model for multivariate longitudinal count data* [56]. Consider a multivariate time series of counts $\mathbf{Y}_t = (Y_{t1}, \ldots, Y_{td})'$, $t = 1, \ldots, N$ and suppose that the conditional distribution of Y_{it} given an unobserved process θ_t is Poisson with parameters $a_{it}\theta_t$ with $a_{it} = \exp(\mathbf{x}_t'\boldsymbol{\alpha}_i)$. Here $\mathbf{x}_t$ denotes short–term covariates and $\boldsymbol{\alpha}_i$ are k–dimensional regression parameters, $i = 1, \ldots, d$. Assume that $\theta_0 = 1$ and suppose that the conditional distribution of θ_t given θ_{t-1} is Gamma with mean $b_t\theta_{t-1}$ and squared coefficient of variation equal to σ^2/θ_{t-1}. Here σ^2 denotes a dispersion parameter and b_t depends on the so called long–term covariates $\mathbf{z}_t$ through the model $b_t = \exp(\Delta\mathbf{z}_t'\boldsymbol{\beta})$, where $\Delta\mathbf{z}_t = \mathbf{z}_t - \mathbf{z}_{t-1}$ and $\mathbf{z}_0 = 0$.

    (a) Calculate the conditional expectation and variance of θ_t given $\theta_0, \ldots, \theta_{t-1}$. What do you observe?

    (b) Show that

$$\mathrm{E}[\theta_t] = b_1 \cdots b_t,$$

to conclude

$$\log \mathrm{E}[\theta_t] = \mathbf{z}_t'\boldsymbol{\beta}.$$

In addition show that

$$\mathrm{Var}[\theta_t] = \phi_t \mathrm{E}[\theta_t]\sigma^2$$

and

$$\mathrm{Cov}(\theta_t, \theta_{t+k}) = \phi_t \mathrm{E}[\theta_{t+k}]\sigma^2,$$

where $\phi_t = b_t + b_t b_{t-1} + b_t b_{t-1} \ldots b_1$.

(c) Turning to the moment structure of the observed process, define $\mathbf{a}_t = (a_{1t}, \ldots, a_{kt})'$ and $\boldsymbol{\Lambda}_t = \mathrm{diag}(a_{1t}, \ldots, a_{kt})$ to prove that

$$\mathrm{E}[\mathbf{Y}_t] = \mathbf{a}_t \mathrm{E}[\theta_t]$$

and

$$\mathrm{Var}[\mathbf{Y}_t] = \boldsymbol{\Lambda}_t \mathrm{E}[\theta_t] + \mathbf{a}_t\mathbf{a}_t'\phi_t\sigma^2\mathrm{E}[\theta_t].$$

The last expression shows that the variance of $\mathbf{Y}_t$ consists of two components; the first term is Poisson variance and the second term represents overdispersion. Conclude that $\log \mathrm{E}[Y_{it}] = \mathbf{x}_t'\boldsymbol{\alpha}_i + \mathbf{z}_t'\boldsymbol{\beta}$, a fact that follows from the log–link.

(d) Discuss Kalman prediction and filtering for this model.

References

1. Alspach, D. L. and Sorensen, H. W. (1972). Nonlinear Bayesian estimation using Gaussian sum approximations. *IEEE Transactions on Automatic Control*, 17, 439–448.

2. Ameen, J. R. M. and Harrison, P. J. (1985). Normal discount Bayesian models. In *Bayesian statistics, 2*, 271–298, North-Holland, Amsterdam.

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

4. Aplevich, J. D. (2000). *The Essentials of Linear State-Space Systems*, Wiley, New York.

5. Bar-Itzhack, I. Y. (1990), In-Flight Alignment of Inertial Navigation Systems. In *Control and Dynamic Systems, Vol. 38, Advances in Aeronautical Systems*, C.T. Leondes Ed., Academic Press, San Diego.

6. Berzuini, C. and Best, N. and Gilks, W. and Larizza, C. (1997). Dynamic conditional independence models and Markov chain Monte Carlo methods. *Journal of the American Statistical Association*, 92, 1403–1412.

7. Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Data Analysis and Theory*, 2nd ed., Springer, New York.

8. Brown, P. E. and Diggle, P. J. and Lord, M. E. and Young, P. C. (2001). Space–time calibration of radar rainfall data. *Applied Statistics*, 50, 221–241.

9. Cargnoni, C. and Müller, P. and West, M. (1997). Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models. *Journal of the American Statistical Association*, 92, 640–647.

10. Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81, 541–553.

11. Carter, C. K. and Kohn, R. (1996). Markov chain Monte Carlo in conditionally Gaussian state space models. *Biometrika*, 83, 589–601.

12. Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state space modeling. *Journal of the American Statistical Association*, 75, 493-500.

13. Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *American Statistician*, 3, 167-174.

14. Chen, R. and Liu, J. S. (1996). Predictive updating methods with application to Bayesian classification. *Journal of the Royal Statistical Society*, B, 58, 397–415.

15. Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, 75, 79–97.

16. Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings alsgorithm. *American Statistician*, 49, 327–335.

17. Dey, D. K. and Ghosh, S. K. and Mallick, B. K. (2000). *Generalized Linear Models: A Bayesian Perspective*. Marcell Dekker, New York.

18. Doucet, A. (1998). On sequential simulation-based methods for Bayesian filtering. Technical report CUED/F-INFENG/TR.310, Department of Engineering, University of Cambridge, Cambridge, UK.

19. Doucet, A. and Godsill, S. J. and Andrieu, C. (2000). On sequential Monte Carlo methods for Bayesian Filtering. *Statistics and Computing*, 10, 197–208.

20. *Sequential Monte Carlo methods in practice*, A. Doucet et al., eds., Springer, New York, 2001.

21. Durbin, J. and Koopman, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspective (with discussion). *Journal of the Royal Statistical Society* B, 62, 3–56.

22. Fahrmeir, L. (1992). Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association*, 87, 501-509.

23. Fahrmeir, L. (1992b). State space modeling and conditional mode estimation for categorical time series. In *New Directions in Time Series Analysis, Part I*, 87-109, D. Brillinger et al. eds., Springer, New York.

24. Fahrmeir, L. and Hennevogl, W. and Klemme, K. (1992) Smoothing in dynamic generalized linear models by Gibbs sampling. In *Advances in GLIM and Statistical Modelling*, L. Fahrmeir et al. eds., Springer, New York, 85–90.

25. Fahrmeir, L. and Kaufmann, H. (1991). On Kalman filtering, posterior mode estimation and Fisher scoring in dynamic exponential family regression. *Metrika*, 38, 37-60.

26. Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modeling Based on Generalized Linear Models*, 2nd ed., Springer, New York.

27. Fahrmeir, L. and Wagenpfeil, S. (1997). Penalized likelihood estimation and iterative Kalman smoothing for non–Gaussian dynamic regression models. *Computational Statistics and Data Analysis*, 24, 295-320.

28. Frühwirth-Schnatter, S. (1992). Integration-based Kalman-filtering for a dynamic generalized linear trend model. *Computational Statistics & Data Analysis*, 13, 447-459.

29. Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15, 183-202

30. Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching mixture models. *Journal of the American Statistical Association*, 96, 194–209.

31. Gamerman, D. (1997). *Markov Chain Monte Carlo*. Chapman & Hall, London.

32. Gelb, A. (1974). *Applied Optimal Estimation*, M.I.T. Press, Cambridge, MA.

33. Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85, 398-409.

34. Geman, S. and Geman, D. (1984).Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

35. Gerencsér, L. (2002). Stability of Random Iterative Mappings. In *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, M. Dror et al. eds., 359–371, Kluwer, Boston.

36. Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57, 1317–1339.

[38] Gilks, W. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 4, 337-348.

37. , *Markov chain Monte Carlo in practice*, Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., eds., Chapman & Hall, London.

38. Gilks, W. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 4, 337–348.

39. Gordon, N. J. and Salmond, D. J. and Smith, A. F. M. (1993). Novel approach to nonlinear/non–Gaussian Bayesian state estimation. *IEE Proceedings F*, 1993, 140", 107–113.

40. Goshen-Meskin, D. and Bar-Itzhack, I. Y. (1992a). Observability Analysis of Piece-Wise Constant Systems I: Theory. *IEEE Transactions on Aerospace and Electronic Systems*, AES-28, 1056-1067.

41. Goshen-Meskin, D. and Bar-Itzhack, I. Y. (1992b). Observability Analysis of Piece-Wise Constant Systems II: Application to Inertial Navigation In-Flight Alignment. *IEEE Transactions on Aerospace and Electronic Systems*, AES-28, 1068-1075.

42. Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting (with discussion). *Journal of the Royal Statistical Society*, B, 38, 205–247.

43. Hartigan, J. A. (1969). Linear Bayesian methods. *Journal of the Royal Statistical Society*, B, 31, 446–454.

44. Harvey, A. (1989). *Forecasting, structural time series models and the Kalman Filter*, Cambridge University Press, Cambridge, UK.

45. Harvey, A. C. and Fernandes, C. (1989). Time series models for count or qualitative observations (with discussion). *Journal of Business & Economic Statistics*, 7, 407–422.

46. Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society*, B, 38, 205-247.

47. Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109.

48. Higuchi, T. (1997). Monte Carlo filtering using the genetic algorithm operators. *Journal of Statistical Computation and Simulation*, 59, 1–23.

49. Hodges, P. and Hale, D. (1993). A computational method for estimating densities of non-Gaussian non-stationary univariate time series. *Journal of Time Series Analysis*, 14, 163–178.

50. Huang, H.-C. and Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics & Data Analysis*, 22, 159–175.

51. Hutchinson, C. E. (1984). The Kalman filter applied to aerospace and electronic systems. *IEEE Transactions on Aerospace and Electronic Systems*, AES-20, 500-504.

52. Hürzeler, M. and Künsch, H. R. (1998). Monte Carlo approximations for general state–space models. *Journal of Computational and Graphical Statistics*, 7, 175–193.

53. Jensen, J. L. and Petersen, N. V. (1999). Asymptotic normality of the maximim likelihood estimator in state space models. *Annals of Statistics*, 27, 514–535.

54. R. H. Jones (1993). *Longitudinal Data with Serial Correlation: A State–space aproach*. Chapman & Hall, London.

55. , Jörgensen, B. and Lundbye-Christensen, S. and Song, P. X.-K. and Sun, L. (1996). State-space models for multivariate longitudinal data of mixed types. *The Canadian Journal of Statistics*, 24, 385–402.

56. Jörgensen, B. and Lundbye-Christensen, S. and Song, P. X.-K. and Sun, L. (1999). A state space model for multivariate longitudinal count data. *Biometrika*, 86, 169–181.

57. Kailath, T. (1974). A view of three decades of linear filtering theory. *IEEE Transactions on Information Theory*, IT-20, 146-181

58. Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of ASME, Journal of Basic Engineering*, Series D, 82, 35-45.

59. Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction. *Transactions of the ASME, Journal of Basic Engineering*, Series D, 83, 95-108.

60. Kitagawa, G. (1987). Non-gaussian state space modeling of nonstationary time series (with discussion). *Journal of the American Statistical Association*,82, 1032–1063.

61. Kitagawa, G. (1989). Non-Gaussian seasonal adjustment. *Computer and Mathematics with Applications*, 18, 503–514.

62. Kitagawa, G. (1994). The two-filter formula for smoothing and an implementation ofthe Gaussian-sum smoother. *Annals of the Institute of Statistical Mathematics*, 46, 605-623.

63. Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5, 1-25.

64. Kitagawa, G. (1998). A self–organizing state space model. *Journal of the American Statistical Association*, 93, 1203–1215.

65. Kitagawa, G. and Gersh, W. (1996). *Smoothness Priors Analysis of Time Series*, Springer, New York.

66. Kitagawa, G. and Higuchi, T. (2001). *Nonlinear non-Gaussian models and related filtering methods*. Selected papers from the International Symposium on Frontiers of Time Series Modeling held in Tokyo, February 14–16, 2000, Ann. Inst. Statist. Math. **53** (2001), no. 1. Kluwer Academic Publishers, Boston.

67. Kong, A. and Liu, J. S. and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89, 278–288.

68. Künsch, H. R. (2001). State space and hidden Markov models. In *Complex Stochastic Systems*, E. Barndorff–Nielsen et al. eds., Chapman & Hall, London, 109–117, 2001.

69. Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamix systems. *Journal of the American Statistical Association*, 93, 1032–1044.

70. Mardia, K. V. and Goodall, C. and Redfern, E. J. and Alonso, F. J. (1998). The Kriged Kalman filter (with comments and a rejoinder by the authors). *Test*, 7, 217–285.

71. Metropolis, N., Rosenbluth, A., Rosenbluth, W., Teller, M. and Teller, E. (1953). Equations of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21, 1087-1091.

72. Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical association*, 44, 335-341.

73. Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: auxiliary particle filtering. *Journal of the American Statistical Association*, 94, 590–599.

74. Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed., Wiley, New York.

75. Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In *Bayesian Statistics 3*, J. M. Bernardo, et al. eds. Oxford University Press, Oxford, 1988, 395–402.

76. Rauch, H. E., Tung, F., and Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *American Institute of Aeronautics and Astronomics Journal*, 3, 1445-1450.

77. Ripley, Brian D. (1987). *Stochastic Simulation*, John Wiley, New York.

78. Sansó, B. and Guenni, L. (2000). A nonstationary multisite model for rainfall. *Journal of the American Statistical Association*, 95, 1089–1100.

79. Shephard, N. (1994). Partial non-Gaussian state space. Biometrika, 81, 115–131.

80. Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84, 653–657.

81. Shumway, R. H. and Stoffer, D. S. (2000). *Time Series Analysis and its Applications*, Springer, New York.

82. Singh, A. C. and Roberts, G. R. (1992). State space modeling of cross–classified time series of counts. *International Statistical Review*, 60, 321-336.

83. Smith, A. F. M. and West, M. (1983). Monitoring renal trnsplants: an application of the multi–process Kalman Filter. *Biometrics*, 39, 867-878.

84. Stroud, J. R. and Müller, P. and Sansó, B. (2001). Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society*, 63, 673–689.

85. Tierney, L. (1991). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, 22, 1701-1762.

86. Schmidt, S. F. (1981). The Kalman filter: Its recognition and development for aerospace applications. *Journal of Guidance and Control*, 4, 4-7.

87. Stratonovich, R. L. (1960). Application of the theory of Markov processes for optimum filtration of signals. *Radio Engineering and Electronic Physics* (USSR), 1, 1-19.

88. Stratonovich, R. L. (1970). Detection and estimation of signals in noise when one or both are non-Gaussian. *Proceedings of the IEEE*, 58, 670-679.

89. Tanizaki, H. and Mariano, R. S. (1998). Nonlinear and non–Gaussian state–space modeling with Monte Carlo simulations. *Journal of Econometrics*, 83, 263–290.

90. West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamical Models*, 2nd ed., Springer, New York.

91. West, M, Harrison, P.J. and Migon, H. S. (1985). Dynamic Generalized Linear Models and Bayesian Forecasting (with discussion). *Journal of the American Statistical Association*, 80, 73-97.

92. Wikle, C. K. and Berliner, M. and Cressie, N. (1998). Hierarchical Bayesian space–time models. *Environmental and Ecological Statistics*, 5, 117–124.