Handbook of Statistics, Vol. 25
ISSN: 0169-7161

DOI 10.1016/S0169-7161(05)25019-8

Dynamic Models

Helio S. Migon, Dani Gamerman, Hedibert F. Lopes and Marco A.R. Ferreira

1. Model structure, inference and practical aspects

Dynamic Bayesian modelling and forecasting of time series is one of the most important areas emerged in Statistics at the end of the last century. Taking as starting point the regression model, an extension is provided by the introduction of an equation governing the regression coefficients evolution through time. Many of the most important problems in Statistics will be encapsulated in this structure. Some special characteristics of Bayesian Forecasting include: (i) all relevant information sources are used, including history, factual or subjective experiences, and knowledge of forthcoming events. (ii) routine forecasting is produced by a statistical model and exceptions can be considered as an anticipation or in a retrospective base, (iii) prospective (what happened) and retrospective (what if) analysis are easily accommodate, (iv) model decomposition: a full Bayesian forecasting model may be decomposed into independent dynamic linear models (DLM), each one describing particular features of the process under analysis. We present in this chapter an overview of dynamic Bayesian models. A more comprehensive treatment of the subject can be found in books such as West and Harrison (1997), Pole et al. (1994), Durbin and Koopman (2002) and Harvey (1989).

In Section 1.1 we describe the class of DLM both to set the notation and to introduce important arguments of Bayesian dynamic models such as model superposition. Inference and practical aspects of Bayesian forecasting are discussed in Sections 1.2 and 1.3, respectively. The nonlinear and nonnormal models are discussed in Section 1.4, giving particular attention to the class of dynamic generalized linear models, an extension of dynamic normal linear models to the exponential family. Section 1.5 briefly summarizes the class of dynamic hierarchical models.

1.1. Dynamic linear models: General notation

Dynamic linear models are a broad class of models with time varying parameters, useful to modeling time series data and regression. It was introduced by Harrison and Stevens (1976) and is very well documented in the book by West and Harrison (1997). In this section some fundamental aspects of dynamic models will be introduced and some examples in time series as well as in regression will be addressed.

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 2
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

Dynamic linear models are parametric models where the parameter variation and the available data information are described probabilistically. They are characterized by a pair of equations, named observational equation and parameter evolution or system equation. The DLM can be seen as a generalization of the regression models allowing changes in parameters values throughout time. The observational and system equations are respectively given by

$$y_t = \mathrm{F}_t'\theta_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, V_t), \tag{1.1}$$

$$\theta_t = G_t\theta_{t-1} + \omega_t, \quad \omega_t \sim N(0, W_t), \tag{1.2}$$

where y_t is a time sequence of scalar observations, conditionally independent given the sequence of parameters θ_t, F_t is a $p \times 1$ vector of explanatory variables, θ_t is a $p \times 1$ vector of parameters, G_t is a $p \times p$ matrix describing the parameter evolution and, finally, V_t and W_t are the variances of the errors associated with the unidimensional observation and with the p-dimensional vector of parameters, respectively. This class includes many of the models found in the statistical literature. For example, if $G = I_p$, the identity matrix of order p and $\omega_t = 0$, $\forall t$, all linear regression models can be represented. On the other hand if F_t, V_t and W_t are constant $\forall t$, then the model covers the linear time series models such as ARIMA processes of Box and Jenkins (1976).

Summarizing, a dynamic linear model is completely specified by the quadruple $\{F_t, G_t, V_t, W_t\}$. Two special cases are, respectively, time series models characterized by $F_t = F$ and $G_t = G$, $\forall t$, and dynamic regression models, described by $G_t = I_p$.

EXAMPLE 1.1 (*1st-order polynomial model*). The simplest model in time series is the 1st-order polynomial model, which corresponds to a 1st-order Taylor series approximation of a smooth time function, named the time series trend. This model is completely defined by the quadruple $\{1, 1, V_t, W_t\}$. The above equations specialize to $y_t = \theta_t + \varepsilon_t$, $\varepsilon_t \sim N(0, V_t)$, $\theta_t = \theta_{t-1} + \omega_t$, $\omega_t \sim N(0, W_t)$, where θ_t is unidimensional and describes the underlying trend of the process. Although this model is very simple, it can be applied in many short-term forecasting systems involving a large number of time series such as in stock control or production planning. The observational and parameters evolution variance can also evolve in time, offering a broad scope for modeling.

A slightly more elaborated model, named linear growth model (LGM, in short), is derived after including an extra parameter $\theta_{2,t}$ to describe the underlying growth of the process. Then, after some minor modifications in state space equations, it follows that $\theta_{1,t} = \theta_{1,t-1} + \theta_{2,t-1} + \omega_{1,t}$, $\theta_{2,t} = \theta_{2,t-1} + \omega_{2,t}$ $\omega_t \sim N(0, W_t)$. The parameter $\theta_{1,t}$ is interpreted as the current level of the process and it is easy to verify that $F_t = (1, 0)$ and $G_t = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$, $\forall t$, characterizing a time series model.

EXAMPLE 1.2 (*Simple dynamic linear regression*). Suppose, in this example, that pairs of values (x_t, y_t) are observed through time and that it is wished to model the existing relationship between x_t and y_t. Assuming that the linear model is a good approximation for the relationship between these values, a simple linear regression model can be set. Since the linear relationship is only a local approximation for the true functional dependence involving x and y, a model with varying parameters is appropriate. For example, the omission of some variables, the nonlinearity of the functional relationship

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 3
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

connecting x and y or some structural changes occurring in the process under investigation, can be responsible for the parameter instability. These situations can be modeled as $y_t = \mathrm{F}_t'\theta_t + \varepsilon_t$, $\theta_t = \theta_{t-1} + \omega_t$ where $\mathrm{F}_t = (1, x_t)'$ and $\omega_t \sim N(0, W_t)$. Note that, in this case, $G_t = I_2$.

As we can observe, the choice of F_t and G_t depends on the model and the nature of the data that is being analyzed. To complete the model specification the variances V_t and W_t must be set. The latter describes the speed of the parameters evolution. In applications V_t is often larger than the elements of W_t. To make the parameter estimation method easier for the conjugate analysis, W_t is scaled by V_t and the conditional variances of the ω_t becomes $V_t W_t$. Therefore, the matrix W_t can be interpreted as a matrix of relative weights with respect to the observational variance. The parameter evolution variance matrix must be assessed subjectively by the user of the method and, in order to do that, the notion of discount factor will be useful (see Ameen and Harisson, 1985). Alternatively, it can be estimated by one of the methods described in the next sections. Therefore the equations presented before can be rewritten as

$$y_t|\theta_t \sim N\big(\mathrm{F}_t'\theta_t, V_t\big), \tag{1.3}$$

$$\theta_t|\theta_{t-1} \sim N(G_t\theta_{t-1}, V_t W_t). \tag{1.4}$$

EXAMPLE 1.3 (*Component models*). It is a good practice to build a model step by step. A component model is structured as a composition of say r DLM's $\{F_i, G_i, 0, W_i\}_t$, $i = 1, \ldots, r$, and a noise model $\{0, 0, V, 0\}_t$. Then the model elements are expressed as $\theta' = (\theta_1', \ldots, \theta_r')$, $\mathrm{F}_t' = (F_1', \ldots, F_r')$, $G = \mathrm{diag}(G_1, \ldots, G_r)$ and $W = \mathrm{diag}(W_1, \ldots, W_r)$. An example, assuming that the data is observed quarterly, with $r = 3$ components includes the *linear trend* and the *regression component* (the 2 examples above) and the quarterly *seasonal component* $\{(1, 0, 0, 0), G_{3,t}, 0, W_{t,3}\}$, where $G_{3,t} = \left(\begin{smallmatrix} 0 & I_3 \\ 1 & 0 \end{smallmatrix}\right)$, named also form free seasonal component model in opposition to the Fourier form representation of seasonality (West and Harrison, 1997).

It is worth noting that it is assumed that for any time t, the current observation y_t is independent of the past observations given the knowledge of θ_t. This means that the temporal dynamics are summarized in the state parameters evolution. This linear structure for modeling data observed through time combines very well with the principles of Bayesian inference because it is possible to describe subjectively the involved probabilities and because of its sequential nature. Therefore, subjective information is coherently combined with past information to produce convenient inference.

1.2. Inference in DLM

The inference in DLM follows the usual steps in Bayesian inference. It explores the sequential aspects of Bayesian inference combining two main operations: *evolution* to build up the prior and *updating* to incorporate the new observation arrived at time t. Let $D_t = D_{t-1} \cap \{y_t\}$ denote the information until time t, including the values of x_t and G_t, $\forall t$, which are supposed to be known, with D_0 representing the prior information. Then

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 4
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

for each time t the prior, predictive and posterior distribution are respectively given by

$$p(\theta_t|D_{t-1}) = \int p(\theta_t|\theta_{t-1})\,p(\theta_{t-1}|D_{t-1})\,\mathrm{d}\theta_{t-1}, \tag{1.5}$$

$$p(y_t|D_{t-1}) = \int p(y_t|\theta_t)\,p(\theta_t|D_{t-1})\,\mathrm{d}\theta_t, \tag{1.6}$$

$$p(\theta_t|D_t) \propto p(\theta_t|D_{t-1})\,p(y_t|D_{t-1}), \tag{1.7}$$

where the last one is obtained via Bayes theorem. The constant of integration in the above specification is sometimes easily obtained. This is just the case when $(F, G, V, W)_t$ are all known and normality is assumed. The resulting algorithm in this very narrow case is known as Kalman filter (Anderson and Moore, **?**). Usually the above matrices depend on some unknown parameters denoted generically by ψ.

The forecasting function and DLM design

In time series model it is very useful to obtain the mean response forecasting function $E\{E[y_{t+h}|\theta_{t+h}]|D_t\} = F'E[\theta_{t+h}|D_t] = F'G^h m_t$ where h is the forecasting horizon, $m_t = E[\theta_t|D_t]$. The structure of this function depends mainly on the eigenvalues of the design matrix G. The interested reader must consult West and Harrison (1997) to learn how to apply these theoretical results in model design.

Evolution and updating equations

The equations described before enable a joint description of (y_t, θ_t) given the past observed data D_{t-1} via $p(y_t, \theta_t|D_{t-1}) = p(y_t|\theta_t)\,p(\theta_t|D_{t-1})$. This leads to the predictive distribution after integrating out θ_t.

One of the main characteristics of the dynamic linear model is that, at each instant of time, all the information available is used to describe the posterior distribution of the state vector. The theorem that follows shows how to evolve from the posterior distribution at time $t-1$ to the posterior at t.

THEOREM 1.1. *Consider a normal dynamic linear model with $V_t = V, \forall t$. Denote the posterior distribution at $t-1$ by $(\theta_{t-1}|D_{t-1}, V) \sim N(m_{t-1}, VC_{t-1})$ and the marginal posterior distribution of $\phi = V^{-1}$ as $\phi|D_{t-1} \sim G(n_{t-1}/2, n_{t-1}s_{t-1}/2)$. Then,*

(1) *Conditionally on V, it follows*
   (a) *Evolution – the prior distribution at t will be $\theta_t|V, D_{t-1} \sim N(a_t, VR_t)$, with $a_t = G_t m_{t-1}$ and $R_t = G_t C_{t-1} G_t' + W_t$.*
   (b) *the one step ahead predictive distribution will be $y_t|V, D_{t-1} \sim N(f_t, VQ_t)$, with $f_t = \mathrm{F}_t' a_t$ and $Q_t = \mathrm{F}_t' R_t \mathrm{F}_t + 1$.*
   (c) *Updating – the posterior distribution at t will be $\theta_t|V, D_t \sim N(m_t, VC_t)$, with $m_t = a_t + A_t e_t$ and $C_t = R_t - A_t A_t' Q_t$, where $A_t = R_t \mathrm{F}_t'/Q_t$ and $e_t = y_t - f_t$.*
(2) *The precision ϕ is updated by the relation $\phi|D_t \sim G(n_t/2, n_t s_t/2)$, with $n_t = n_{t-1} + 1$ and $n_t s_t = n_{t-1}s_{t-1} + e_t^2/Q_t$.*
(3) *Unconditionally on V, we will have*: (a) $\theta_t|D_{t-1} \sim t_{n_{t-1}}(a_t, s_{t-1}R_t)$, (b) $y_t|D_{t-1} \sim t_{n_{t-1}}(f_t, Q_t^*)$, *with $Q_t^* = s_{t-1}Q_t$ and* (c) $\theta_t|D_t \sim t_{n_{t-1}}(m_t, s_t C_t)$.

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 5
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

PROOF. Item (1)(a) follows immediately using the parameter evolution equation and standard facts from the normal theory. With respect to (1)(b), using the prior distribution in (a), it follows that $f_t = E[E(y_t|\theta_t)|V, D_{t-1}] = \mathrm{F}_t' a_t$, $Q_t = V[E(y_t|\theta_t)|V, D_{t-1}] + E[V(y_t|\theta_t)|V, D_{t-1}] = V(\mathrm{F}_t' R_t \mathrm{F}_t + 1)$ and the normality is a consequence of the fact that all the distributions involved are normal. To prove part (1)(c), suppose that the posterior distribution at $t-1$ is as given in the theorem. We wish to show that (c) follows from the application of Bayes theorem, that is, $p(\theta_t|V, D_t) \propto p(\theta_t|V, D_{t-1})p(y_t|\theta_t, V)$. To show that it is sufficient to use Theorem 2.1 in Migon and Gamerman (**?**) and the identity $C_t^{-1} = R_t^{-1} + \phi \mathrm{F}_t \mathrm{F}_t'$.

If V is unknown, it will follow

- By hypothesis, $\phi|D_{t-1} \sim G(n_{t-1}/2, n_{t-1}s_{t-1}/2)$, and $y_t|\phi, D_{t-1} \sim N(f_t, Q_t/\phi)$. Then, by Bayes theorem, $p(\phi|D_t) \propto \phi^{(n_{t-1}+1)/2-1} \exp\{-\frac{\phi}{2}(n_{t-1}s_{t-1} + \frac{e_t^2}{Q_t})\}$ and therefore, $\phi|D_t \sim G(n_t/2, n_t s_t/2)$.
- Finally, for part (3) of the theorem, the proofs of items (a)–(c) follow from the results about conjugacy of the normal–gamma to the normal model and from the marginal distributions obtained after integrating out V.

□

1.3. Practical aspects of Bayesian forecasting

In this subsection, some special aspects involved in dynamic Bayesian modeling will be briefly discussed, including variance law, discount factor, smoothing, intervention and monitoring.

Variance law

The possibility of modeling the observational variance deserves special attention among the special aspects involved in the dynamic Bayesian modeling. For example, the observational variance can be modeled as a power law, $V_t = V v(\mu_t)$, where $v(\mu_t)$ is predictable, as for example: $v(\mu_t) = \mu_t^{b_1}(1-\mu_t)^{b_2}$ where $\mu_t = \mathrm{F}_t'\theta_t$ is the process mean level. With $b_1 = b_2 = 1$, this mimics the binomial law and fixing $b_1 = 1, b_2 = 0$ it follows a Poisson law. An explanation of this law in a commercial environment is that the number of orders is Poisson ($b_1 = 1, b_2 = 0$) but the amounts per order vary so that demand follows a compound Poisson distribution. For economic indices, the log transformation is often applied, which is equivalent to using $b_1 \simeq 2, b_2 = 0$. The constant b_1 can be chosen in parallel to the well-known Box–Cox family of transformation. The scale factor V can be sequentially estimated as stated in the theorem or more generally assuming a dynamic evolution governed by the transformation $\phi_t = \gamma_t \phi_{t-1}$, where $\gamma_{t-1} \sim Ga(n_{t-1}/2, n_{t-1}s_{t-1}/2)$, $\gamma_t \sim Be(\delta_v n_{t-1}/2, (1-\delta_v)n_{t-1}/2)$, $\delta_v \in (0,1)$ is a (variance) discount factor and s_t is an approximate point estimate of the variance. The main advantage in this form of modeling is to avoid the transformation of the original data, keeping in this way, the interpretation of the parameters, which is very useful, for example, when one wishes to perform some subjective intervention.

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 6
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

Discount factor

The use of discount factor is recommended to avoid the difficult task of directly setting the state parameters evolution matrix. These are fixed numbers between zero and one describing subjectively the loss of information through time. Remember that the prior variance of the state vector is obtained as $R_t = P_t + W_t$ where $P_t = G_t C_{t-1} G_t'$. Denoting the discount factor by δ, we can rewrite $R_t = P_t/\delta$, showing clearly that there is a relationship between W_t and δ. This is given by $W_t = (\delta^{-1} - 1)P_{t-1}$, showing that the loss of information is proportional to the posterior variance of the state parameters. For example, if $\delta = 0.9$, only about 90% of the information passes through time. The case of multiple discount factors is easily incorporated in the DLM and is very useful in practice since DLM components typically lose information at different rates.

Missing observation

Other relevant aspects of dynamic linear models are to easily take care of missing observations and to automatically implement subjective interventions. In the first case, it suffices not to use the updating equations at the time the observations are missing. In this way, the uncertainties increase with the evaluation of the new prior distribution and the recurrence equation continues to be valid without any additional problem. From the intervention point of view, the simplest proposal is to use a small discount factor, close to zero, at the time of announced structural changes in the data generation process. In this way the more recent observations will be strongly considered in the updating of the prior distribution and the system can be more adaptive to possible changes. An application of this idea in financial data analysis can be found in Reyna et al. (1999).

Retrospective analysis

It is worth mentioning that parameters distribution at any time t can be revised with the arrival of new observations. We can generically obtain the parameter distributions $p(\theta_t|D_{t+k})$, $\forall k$ integer. If $k > 0$, this is named the smoothed distribution, if $k = 0$, it is just the posterior and if $k < 0$ it is the prior distribution. In dynamic modeling it is common to use the distributions $p(\theta_t|D_T)$, $\forall t = 1, \ldots, T$, where T is the size of the series, to retrospectively analyze the parameter behavior. For example, one may want to quantify the change in behavior induced by some measure of economic policy. The future data would inform about the change occurred in any particular parameter of the model describing the behavior of the involved economic agents.

Monitoring and interventions

A continual assessment of the performance of any forecast system is vital for its effective use. Model monitoring is concerned with detecting inadequacy in the current model due to any major unanticipated event. One of the most important features of the Bayesian forecasting methodology is the sequential monitoring (West and Harrison, 1986) and intervention process described by West and Harrison (1997), which is based on the analysis of the standardized one step ahead forecasting error, $e_t = (y_t - f_t)/Q_t^{1/2}$, and allows them to identify the existence of level and variance changes. Suppose that at time t a suitable alternative model exists and is denoted by A. Based on the cumulative Bayes factor a sequential monitoring algorithm, like a Cusum test, follows as: (i) the

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 7
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

model A is accepted if $H_t^1(r) \geqslant \tau^{-1}$, the current model is accepted if $H_t^1(r) \leqslant 1$ and otherwise go on cumulating some more evidence, where $H_t(r) = \frac{p(e_t,\ldots,e_{t-r+1}|D_{t-r})}{p_A(e_t,\ldots,e_{t-r+1}|D_{t-r})}$ and τ and r are constants chosen in advance.

An implicit decision problem is associated with the above algorithm, where the state of the nature is $M = \{A_0, A\}$, with A_0 representing the current model and A an alternative one. The decision space is given by $D = \{d_0, d_1, d_2\}$, with d_0 indicating that the standard model is accepted, d_1, the alternative model is accepted and d_2 standing for the fact that the available information is not sufficient to decide between A_0 and A. Finally, assume that the loss function is proportional to the following payoffs: $l(d_0, A_0) = l(d_1, A) = 1 - l(d_0, A) = 0$, and $l(d_1, A_0) = l_1$ and $l(d_2, A_0) = l(d_2, A) = l_2$, with $l_2 < 1 < l_1$. Letting π_0 be the prior probability of the current model, it follows that the West and Harrison sequential monitoring algorithm corresponds to the solution of this decision problem with $l_1 = \pi_1(1 + \pi_1/(\tau\pi_0))$, $l_2 = \pi_1$ and $\pi_1 > \tau^{1/2}/(1 + \tau^{1/2})$, where $\pi_1 = 1 - \pi_0$.

In a recent paper, Harrison and Lai (1999) discuss a sequential Bayesian decision monitor for application in quality control, allowing to monitor the natural parameters of many exponential family models with conjugate prior. This proposal is able to discern whether the deterioration is due to change in level, slope, regression coefficient or seasonal pattern, individually. An extended version of the West and Harrison monitoring algorithm, allowing for, jointly, monitoring of several types of shocks, is proposed in Gargallo and Salvador (2002), where a frequentist evaluation of the proposed procedure is investigated and some guidelines as to how to choose its parameters are given.

Multiprocess models

The multi-process was introduced by Harrison and Stevens (1976) and involves a set of sub-models which together provide individual component models for outlying observations and changes in any or all of the components of the parameter vector. This is an alternative approach for model monitoring and intervention, where the alternative sub-models are fully stated. The multi-process approach is quite demanding computationally but proves extremely effective in identifying and adapting to changes with minimum loss of predictive performance. The model structure involves k states, $S_t(i), i = 1, \ldots, k$. The first state $S_t(1)$ refers to the no change, nonoutlier model, the other one $S_t(2)$ allows for an outlying observation and the remaining ones, $S_t(i), i = 3, \ldots, k-2$, to changes in particular components of the parameter vector, θ_t. At each time the probability of any particular state is assumed to be independent of what has occurred up to time $t - 1$ and are denoted by $\pi_i = P[S_t(i)]$. A very useful example, in short term forecasting, includes a four states multi-process model. The first state is a linear growth model, the others represent the outlier model and changes in level and growth respectively.

The posterior distribution of the parameters at time $t - 1$ is defined conditionally on the states at that time, that is: $p(\theta_{t-1}|S_{t-1}(i), D_{t-1})$. The unconditional posterior of the parameters at time $t - 1$ can be easily obtained as a mixture since the posterior probability over the states is known.

The prior distribution for the parameters is evaluated conditional on the states $S_t(j)$ and $S_{t-1}(i)$, that is: $p(\theta_t|S_{t-1}(i), S_t(j), D_{t-1})$. As soon as a new observa-

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 8
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

tion is obtained, k^2 posterior distributions are obtained. The full conditional posterior distribution of θ_t is given by the k^2 component mixture. The transitions probabilities involved in the mixture are obtained from Bayes theorem as: $p_t(i, j) \propto \pi_j p_{t-1}(i) p(y_t|S_t(j), S_{t-1}(i), D_{t-1})$, where $p_t(i, j) = P[S_t(j), S_{t-1}(i)|D_t]$. From this, $p_t(j) = \sum_{i=1}^{k} p_t(i, j)$ and $p_{t-1}(i) = \sum_{j=1}^{k} p_t(i, j)$. These probabilities prove useful in diagnosing the particular type of change or outlier after the event. Then, to avoid getting involved with an explosive number of alternative distributions, a collapsing procedure is in order.

The specification of the alternative models and many practical aspects involved in the multi-process implementation can be found in West and Harrison (1997), Chapter 12. The original multi-process model of Harrison and Stevens was developed for normal linear processes. A version based on the ideas of discounting was successfully implemented in the eighties. The extension to nonnormal and nonlinear models was developed in West (1986). An alternative method, computationally more efficient, is to use a simple monitoring scheme coupled with a multi-process approach, that is applied only when the monitor detects model breakdown (Ameen and Harrison, 1983).

1.4. Dynamic nonlinear/nonnormal models

There are many practical situations where nonlinearity is implicit in the modeling process. An initial very simple example in time series is the seasonal multiplicative model. The mean effect is defined as $\mu_t = \alpha_t(1 + \phi_t)$ where α_t is the trend component and ϕ_t the multiplicative seasonal effect. As a second example let us consider the first-order transfer response. This is useful to describe the dynamic relationship involved in a pulse response random variable x_t. This response function is given by $E_t = \phi E_{t-1} + \gamma x_t$, where $\gamma > 0$ represents the instantaneous "gain" and $|\phi| < 1$ models the decay rate. The mean response can be phrased as $\mu_t = \alpha_t + E_t$. Extensive applications of nonlinear and nonnormal models is reported on Migon and Harrison (1985), where the impact of consumer awareness of the advertising of various products is assessed.

The case of nonnormal models is often met in the practice of dynamic modeling. Some examples include count data and nonnegative observations and also the desire to use heavy tails density to protect oneself against spurious observations.

There are many different ways to approach both the above questions. The simplest one is to use some sort of transformation, with the undesirable cost of losing interpretability. Moreover the nonlinearity can be handled through linearization, which consists in the use of a low-order Taylor approximation.

Let the model be $y_t = F_t(\theta_t) + \varepsilon_t$, with $\theta_t = g(\theta_{t-1}) + \omega_t$. Assuming that the terms other than the linear one are negligible, the evolution becomes $g_t(\theta_{t-1}) \simeq g_t(m_{t-1}) + G_t(\theta_{t-1} - m_{t-1}) + \omega_t = h_t + G_t\theta_{t-1} + \omega_t$, with $G_t = \frac{\partial}{\partial\theta_{t-1}} g_t(\theta_{t-1})|_{m_{t-1}}$, where $m_{t-1} = E[\theta_{t-1}|D_{t-1}]$ and $h_t = g_t(m_{t-1}) - G_t(m_{t-1})$. The same sort of first-order approximation can also be applied to $F(\theta_t) \simeq F_t(a_t) + F_t'(\theta_t - a_t)$, where $F_t' = \frac{\partial}{\partial\theta_t} F_t(\theta_t)|_{a_t}$. Then the observation equation simplifies to $y_t = (F_t(a_t) - F_t'a_t) + F_t'\theta_t + \varepsilon_t$, where $a_t = E[\theta_t|D_{t-1}]$. The inference follows using the theorem stated for the linear case.

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 9
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

The interested reader is referred to West and Harrison (1997), Chapter 13, and Durbin and Koopman (2001), Chapter 10.

EXAMPLE 1.4. A broad class of nonlinear growth curve models, including the modified exponential ($\lambda = 1$), the logistic ($\lambda = -1$) and the Gompertz ($\lambda \to 0$), was introduced in Migon and Gamerman (1993). The global mean function of the growth process is defined by $\mu_t = [a + b\exp(\gamma t)]^{1/\lambda}$, with parameterization (a, b, γ, λ). The major advantage of this approach is to keep the measurements in the original scale, making the interpretation easier. A similar local model is easily obtained and can be represented as a dynamic nonlinear model with $F_t(\theta_t) = \theta_{1,t}$ and $G(\theta_t) = (\theta_1 + \theta_2, \theta_2\theta_3, \theta_3)'$, the link function is $g(\mu_t) = F_t(\theta_t)$, where g is an invertible function and a variance law $v(\mu_t)$.

The non-Gaussian state space model is characterized by the same structure as in the linear Gaussian case, with $p(y_t|\mu_t)$, $\mu_t = F_t'\theta_t$, and the states evolving as $\theta_t = G_t\theta_t + \omega_t, \omega_t \sim p(\omega_t)$, where $\omega_t, t = 1, \ldots, n$, are serially independent and either $p(y_t|\mu_t)$ or $p(\omega_t)$ or both can be non-Gaussian. Two special cases are worth considering: (a) observation from the exponential family, and (b) observation generated by the relation $y_t = \mu + \varepsilon_t$, with $\varepsilon_t \sim p(\varepsilon_t)$ non-Gaussian. The first case will be considered in the next subsection and follows West et al. (1985), while the case (b) will be considered in the next sections of this chapter. The case where only $p(\varepsilon_t)$ is nonnormal will be considered in the following example.

EXAMPLE 1.5 (*A non-Gaussian model*). In finance V_t may depend upon the time interval between trading days, the time to a bond's maturity, trading volume and so forth. A broad class of models was presented in the literature of the end of the last century. An excellent example is the stochastic volatility model. Let us denote the first differences of a particular series of asset log prices by y_t. A basic stochastic volatility model is given by: $y_t = V_t^{1/2}\varepsilon_t$, where $\varepsilon_t \sim N[0, 1]$. Taking the square and the logarithm we obtain: $\log(y_t^2) = h_t + \log(\varepsilon_t^2)$, where $h_t = \log(V_t)$. Assuming that $h_t = \alpha + \beta h_{t-1} + \omega_t$ we obtain a non-Gaussian first-order dynamic model named, in the recent literature of finance, as the log-stochastic volatility model. The estimation can be done using Bayes' linear method, to be described in the next subsection, since $\log(\varepsilon_t^2)$ is log-χ^2 distribution. This approach is similar to the quasi-likelihood method proposed in Harvey et al. (1994). Various extensions of the SV models can be considered. For example the normality assumption can be replaced by some heavy tail distribution such as the t-distribution, as supported by many empirical studies.

Triantafyllopoulos and Harrison (2002) discuss stochastic volatility forecasting in the context of dynamic Bayesian models. They criticize the GARCH models showing some of their limitations and presenting dynamic models for forecasting the unknown stochastic volatility. The dynamic variance law produces a volatility estimate in the form of a discount weight moving average. A zero drift model ($F_t = 1, G_t = 0$ and $W_t = 0$) produces the volatility estimate $n_t\widehat{V}_t = \beta n_{t-1}\widehat{V}_{t-1} + y_t^2$, which looks like a GARCH recursion. Previous application of Bayesian forecasting in financial time series is presented in Migon and Mazuchelli (1999). A dynamic Bayesian GARCH model was introduced and the volatility of four different kinds of assets were predicted.

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 10
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

Dynamic generalized linear models

The extension of DLMs to allow observations in the exponential family was introduced by West et al. (1985) based on the generalized linear models of Nelder and Wedderburn (1972). The observation Eq. (1.1) is replaced by

$$p(y_t|\eta_t) \propto \exp\big[\big(y_t\eta_t - b(\eta_t)\big)/\phi_t\big] \tag{1.8}$$

and, in addition, a suitable link function is introduced, relating the mean $\mu_t = E[y_t|\eta_t] = b'(\eta_t)$ to the regressors F_t through $g(\mu_t) = \lambda_t = F_t'\theta_t$. A conjugate prior for the (1.8) is given as $p(\eta_t|D_{t-1}) \propto \exp[(r_t\eta_t - b(\eta_t))/s_t]$. The integral in Eqs. (1.5)–(1.7) cannot be obtained in closed form, and so the inference must be done in an approximate way. A procedure allowing the sequential analysis of DGLM was implement in West et al. (1985) using linear Bayes estimation.

The evolution equation (1.2) is only partially specified. This means that the distributions of $\theta_{t-1}|D_{t-1}$ and ω_t are only specified by the first- and second-order moments, that is: $\theta_{t-1}|D_{t-1} \sim [m_{t-1}, C_{t-1}]$ and $\omega_t \sim [0, W_t]$. Then the prior distribution for the state parameters is also partially specified as $\theta_t|D_t \sim [a_t, R_t]$ as given before. Then the prior distribution for $\lambda_t = g(\mu_t)$ is $\lambda_t|D_{t-1} \sim [f_t, q_t]$ where $f_t = F_t'a_t$ and $q_t = F_t'R_tF_t$. The parameters (r_t, s_t) in the prior distribution for η_t must be related to f_t and q_t through the $E[g(b'(\eta_t))|D_{t-1}] = f_t$ and $\text{var}[g(b'(\eta_t))|D_{t-1}] = q_t$. Then the posterior for η_t is in the same form of its (conjugate) prior distribution with parameters $(r_t/s_t + y_t/\phi_t, 1/s_t + 1/\phi_t)$. The posterior distribution of the linear predictor is $\lambda_t|D_t \sim [f_t^*, q_t^*]$, where, again, $f_t^* = E[g(b'(\eta_t))|D_t]$ and $q_t^* = \text{var}[g(b'(\eta_t))|D_t]$. Moreover, to complete the analysis, the posterior distribution of the state parameters must be obtained. The linear Bayes estimation is used to approximate the first- and second-order moments of this distribution, leading to: $\widehat{E}[\theta_t|\eta_t, D_{t-1}] = a_t + R_tF_t[\eta_t - f_t]/q_t$ and $\widehat{\text{var}}[\theta_t|\eta_t, D_{t-1}] = R_t - R_tF_tF_t'R_t/q_t$. The moments of $\theta_t|D_t$ are calculated using the iterated expectation law given $\theta_t|D_t \sim [m_t, C_t]$, where $m_t = a_t + R_tF_t[f_t^* - f_t]/q_t$ and $C_t = R_t - R_tF_tF_t'R_t(1 - q_t^*/q_t)/q_t$.

A practical example

This example describes a model to predict the value of the Brazilian industrialized exports; for more details the interested reader should see Migon (2000). The model is built up step by step starting with a supply and demand equilibrium model. Dynamics are introduced via the general adaptive expectations hypothesis. After some simplifications we end up with a dynamic nonlinear model: $\log y_t = \eta_t + \varepsilon_t$, where, $\varepsilon_t \sim N[0, \sigma_t^2]$ and $\eta_t = \mu_t + E_t$. The parameters evolution is given by: $\theta_t = \text{diag}(G_1, G_2, G_3)\theta_{t-1} + \omega_t$, where $\theta_t = (\mu, \beta, E, \gamma, \lambda)_t$, with $G_1 = \left(\begin{smallmatrix}1 & 1\\ 0 & 1\end{smallmatrix}\right)$, $G_2 = \left(\begin{smallmatrix}\lambda_{t-1} & \log r_t\\ 0 & 1\end{smallmatrix}\right)$ and $G_3 = 1$. The scale factor $\phi_t = \sigma_t^{-2}$ evolves through time keeping its mean and increasing the uncertainty via a discount factor. Next we will present a summary description of the data used and the main findings. The period of the data used in this application, Jun./79 up to Dec./84, is characterized by a deep recession in the world trade and many devaluations of the Brazilian exchange rate, which make it attractive to illustrate the capabilities of the models developed. In Figure 1(a), the one step ahead point forecasting, posterior mode, obtained using a simple linear growth model *with and without* subjective intervention

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 11
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

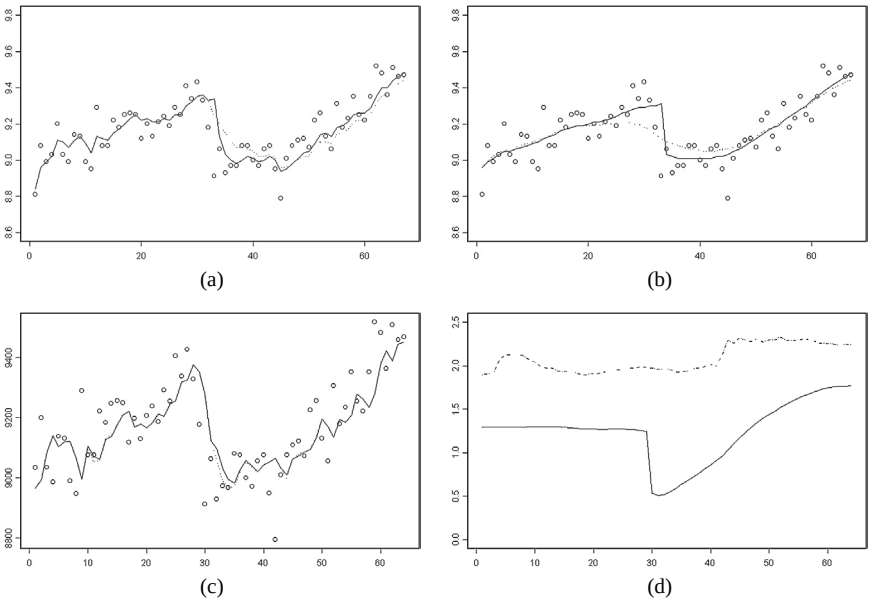

Fig. 1. (a) On-line fitting of exports linear growth model without and with intervention; (b) smoothed fitting of exports linear growth model without and with intervention; (c) on line fitting of exports transfer function model without and with intervention; (d) exchange rate effect.

can be observed and also the smoothed mean of the level is plotted. The interventions improve considerably the predictive performance. The mean absolute error (MAD) decreases from .12 to .09 and the logarithm of the predictive likelihood (LPL) raises from 35 to 47, supporting the need of intervention. It is worth pointing out that the interventions were necessary to cope with change in the growth at Feb./82 and Jun./82 as is evident from Figure 1(b), where the on line fitting ($E[\mu_t|D_t]$) is shown. Examining the residuals of those fittings two alternatives can take place: to include a seasonal component or some of the omitted variables. In this study the exchange rate r_t is the candidate suggested by the theoretical model developed before. In fact seasonality is not supported by the data (MAD $=$ .12 and LPL $=$ 37). On the other hand, inclusion of the control variable, exchange rate, improves considerably the results. The best prediction results were obtained with the first-order transfer response model, which is simple and sophisticated. In the graph below, Figure 1(c), one can appreciate the predictive capability of the model and in Figure 1(d), the smoothed effect of the exchange rate is plotted.

An application of transfer response model in hydrology can be found in Monteiro and Migon (1997) where the dynamic relationship between rainfall and runoff is modeled.

1.5. Dynamic hierarchical models

This is a very important class of models useful for modeling several concomitant time series of observations. The methodology was developed in Gamerman and Migon (1993), combining the dynamic models previously presented and the hierarchical mod-

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 12
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

els of Lindley and Smith (1972). These models appear in the Statistical literature under the name of panel data or longitudinal studies (Jones, 1993). The dynamic evolution provides a general framework for analysis of multivariate time series. The model specification includes a set of structural equations besides the usual observation and evolution equations. The structural equations progressively reduce the dimension of the parameter space as the level becomes higher.

The dynamic hierarchical model, with three levels, can be written as: $y_t = F_{1,t}\theta_{1,t} + \nu_{1,t}$ where $\nu_{i,t} \sim N[0, V_{1,t}]$, $\theta_{i,t} = F_{i+1,t}\theta_{i,t-1} + \nu_{i,t}$, $i = 1, 2$, and, finally, $\theta_{3,t} = G_t\theta_{3,t-1} + \omega_t$ where $\nu_{i,t} \sim N[0, V_{1,t}]$, $\omega_t \sim N[0, W_t]$, all the disturbance terms are independent with known variance and $F_{i,t}$, $i = 1, 2, 3$, known full rank matrices. It is mandatory that the evolution equation is applied to the higher level of the hierarchy.

The example includes cross-section of random samples with linear growing exchangeable means and of regression models with steady parameters, where all the observations at time t are explained by the same regressors with known values for each observation i. Inference for dynamic hierarchical models can be found in Gamerman and Migon (1993). Extension to the case of multivariate time series in presented in Landim and Gamerman (2000). A very interesting application for modeling animal growth curves is presented in Barbosa and Migon (1996).

2. Markov Chain Monte Carlo

Dynamic models introduced in the previous section allow for full inference only when the F_t's, G_t's and W_t's are entirely known and a conjugate form is imposed on the $V = V_t$, $\forall t$. In general, these quantities or other quantities used in their definition are unknown and inference about them must be based on their posterior distribution. This distribution is generally not analytically tractable.

In this section, the problem of making full (about all model parameters) inference about DLM's is considered. We start by considering normal DLM's and then generalize ideas to nonnormal DLM's. Ideally analytically unsolved problems in Bayesian inference can be approximated solved by sampling from the relevant posterior distribution. The distributions involved here are too complicated for directly drawing samples from. The tool presented in this section to solve this problem is MCMC, a powerful collection of sampling techniques that has revolutionized Bayesian inference for complex models in the last decades. Space constraints allow only a few descriptive lines here about MCMC. Full consideration of MCMC at an expository level can be seen in Gamerman (1997).

MCMC stands for Markov chain Monte Carlo and deals with sampling from a complicated distribution (in our case, the joint posterior of all model parameters) when direct sampling is not available. It is a two step procedure. In the first step, a Markov chain with a transition kernel such that the limiting distribution is given by the joint posterior distribution is constructed. In the second step, a trajectory is sampled from this chain. For a suitably large iteration, values are virtually sampled from the posterior distribution.

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 13
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

The most used transition kernels are derived from the Gibbs sampling (Geman and Geman, 1984) and componentwise Metropolis–Hastings steps (Metropolis et al., 1953; Hastings, 1970). In the first group, the kernel is composed of product of the posterior full conditionals of parameters divided in blocks. In the second group, the kernel is still composed of products of proposal full conditionals, not necessarily derived from the posterior. A correction term must then be imposed to blocks not drawn from the posterior to ensure convergence of the chain to the posterior. This scheme can also be applied if all parameters are gathered into a single block. In this case, the proposal replaces the joint posterior. In theory, any partition of the parameters into blocks ensures convergence. In practice, however, this partition choice plays an important role in providing a computationally feasible solution.

2.1. Normal DLM

This subsection considers the solution to full inference in normal DLM by sampling. Alternative sampling schemes are presented and compared in theoretical and empirical terms. The basic difference between them lies in the specification of blocks of parameters. Schemes are presented in the order of increasing dimension of the blocks, or in other words, decreasing number of blocks.

The normal dynamic models considered here are given by (1.1)–(1.2) with known F_t and G_t and constant variances of the observation and system disturbances, i.e., $V_t = V$ and $W_t = W$, for all t. This restriction is aimed mainly at presentation clarity. It also provides for a more parsimonious model. The extension to the general case of unequal variances is not difficult but can hinder meaningful inference unless substantial prior information is available for the different variance parameters. Extensions to cases where unknown quantities are present in F_t and G_t are not difficult to implement with the methodology described below.

The previous section showed how to perform exact inference in normal DLM when the evolution variance matrix W was known. It also showed a few simple alternatives to inference with respect to an unknown W, namely comparison of a few values for W and its specification through the use of discount factors, based on scaling on V.

Componentwise sampling schemes

Now, a sample from the posterior of $(\theta_1, \ldots, \theta_T, V, W)$ after observing the series up to time n is drawn. The joint posterior is

$$\pi(\theta, V, W|D_T) \propto \prod_{i=1}^{T} p(y_t|\theta_t, V) \prod_{i=2}^{T} p(\theta_t|\theta_{t-1}, W)\, p(\theta_1)\, p(V, W),$$

where $\theta = (\theta_1, \ldots, \theta_T)$. The specification of the model can be taken as a basis for blocking parameters. So, an initial blocking choice is $\theta_1, \ldots, \theta_T$, V and W. The full conditional of the θ_t's, denoted by $\pi(\theta_t|\theta_{-t}, V, W, D_T)$, are normal distributions that are easily obtained. If the joint prior for (V, W) is a product of an inverse Gamma for V and inverse Wishart for W, then their full conditional will also be so. So, sampling from their full conditionals is easily available.

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 14
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

These full conditional distributions complete a cycle of the Gibbs sampler. They are all easy to sample from and a MCMC-based Bayesian inference can be performed. This approach was introduced by Carlin et al. (1992) in the context on nonlinear and scale mixture of normal models and is summarized in Algorithm 1.

The prior for θ is highly dependent on the value of W. For small values of W, large correlations are induced and preserved through to the posterior. Unfortunately, this is usually the case observed in practice with system parameters typically experiencing small disturbances through time. This means that sampling the θ components separately may imply a kernel that will be very slow and may only achieve equilibrium after an unnecessarily large number of iterations was drawn. Figure 1(b) illustrates the convergence of state parameters with this sampling scheme.

ALGORITHM 1 (*Initial MCMC for DLM*).

a. Initialization: *set initial values* $(\theta^{(0)}, V^{(0)}, W^{(0)})$ *and iteration counter* $j = 1$;
b. Sampling θ**:** $\theta^{(j)}$ *is sampled componentwise as follows*

1. *set* $t = 1$;
2. *sample* $\theta_t^{(j)}$ *from* $\pi(\theta_t|\theta_1^{(j)}, \ldots, \theta_{t-1}^{(j)}, \theta_{t+1}^{(j-1)}, \ldots, \theta_T^{(j-1)}, V^{(j-1)}, W^{(j-1)}, D_T)$;
3. *set* $t \to t+1$ *and return to* ***b2****, if* $t < T$;

c. Sampling V **and** W**:** $V^{(j)}$ *and* $W^{(j)}$ *are sampled successively from the respective full conditionals* $\pi(V|\theta^{(j)}, W^{(j-1)}, D_T)$ *and* $\pi(W|\theta^{(j)}, V^{(j)}, D_T)$;
d. Updating: *set* $j \to j+1$ *and return to* ***b*** *until convergence.*

Block sampling schemes

One alternative to avoid this computational problem is to sample θ is a single block. There are a number of ways to obtain a sample for the posterior full conditional of $(\theta|V, W, D_T)$. A statistical way explores the properties of the model by noting that

$$\pi(\theta|V, W, D_T) = p(\theta_T|V, W, D_n)\prod_{t=1}^{T} p(\theta_t|\theta_{t+1}, V, W, D_t).$$

The distributions of $(\theta_t|\theta_{t+1}, V, W, D_t)$ are obtained from $p(\theta_{t+1}|\theta_t, V, W, D_t)$ and $p(\theta_t|V, W, D_t)$ and are given by

$$(\theta_t|\theta_{t+1}, V, W, D_t) \sim N\big[\big(G_t'W^{-1}G_t + C_t^{-1}\big)^{-1}\big(G_t'W^{-1}\theta_{t+1} + C_t^{-1}m_t\big), \big(G_t'W^{-1}G_t + C_t^{-1}\big)^{-1}\big] \tag{2.1}$$

for $t = 1, \ldots, n-1$. So, a scheme for sampling from the full conditional of θ is given by incorporation of this backward (in time) sampling step in the forward (in time) Kalman filter presented in the previous section. The above sampling scheme was independently proposed by Carter and Kohn (1994) and Frühwirth-Schnatter (1994), that appropriately named it forward filtering backward smoothing (FFBS). The sampling algorithm is summarized as:

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 15
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

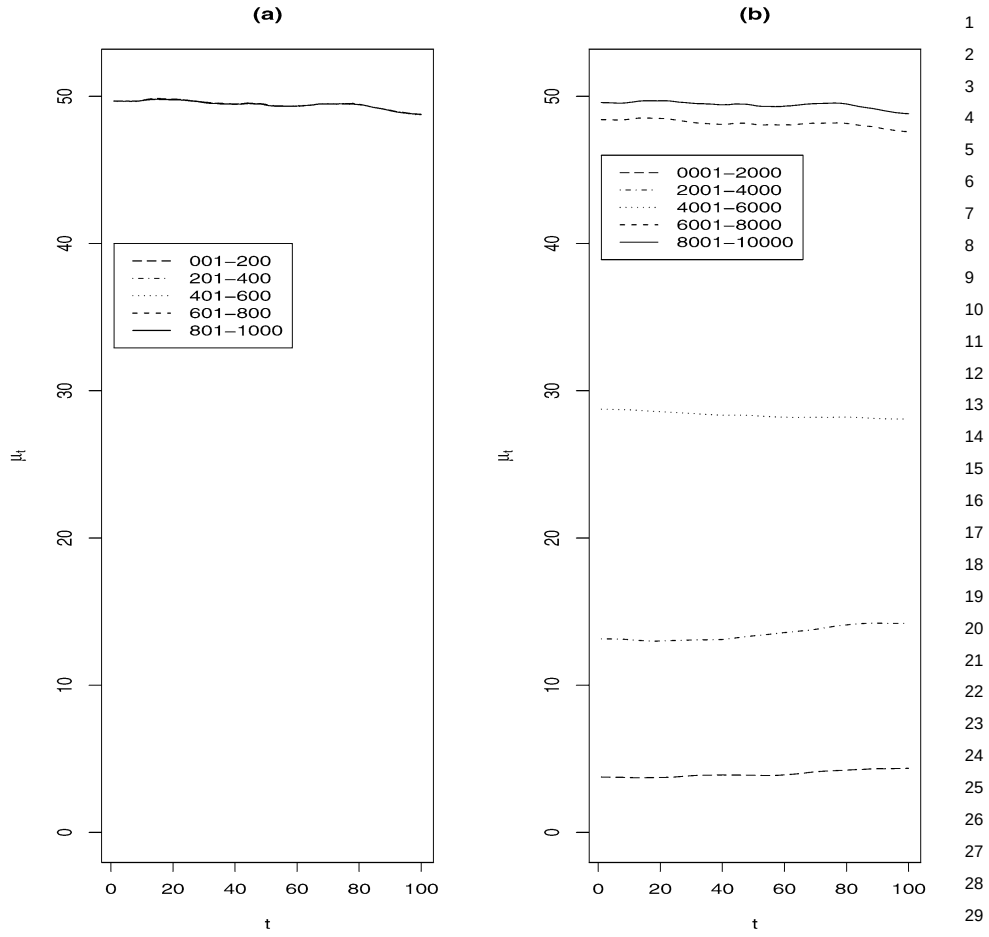

Fig. 2. Average trajectory of μ_t, $t = 1, \ldots, 100$, over batches of M successive iterations under different sampling schemes: (a) – sampling $(\mu_1, \ldots, \mu_{100})$ in a block ($M = 200$); (b) – sampling each μ_t separately ($M = 2000$). Data was generated from a first-order model with $V = 1$ and $W = 0.01$. Chains were started from 0.

ALGORITHM 1* (*FFBS*).

Exactly as Algorithm 1 *but for replacement of* θ *sampling* (*step* **b**) *by:*

b1*. *sample* θ_T *from its updated distribution (given in previous section) and set* $t = T - 1$.

b2*. *Sample* θ_t *from the distribution* (2.1).

b3*. *Decrease* t *to* $t - 1$ *and return to step* **b2*** *until* $t = 1$.

Step **b1*** is obtained by running the Kalman filter from $t = 1$ to $t = T$ with given values of V and W. When running the filter, the updated means m_t and variances C_t, $t = 1, \ldots, T$, are stored for use in step **b2***.

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 16
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

A more numerical way to obtain a sample from θ is given by direct evaluation of the prior full conditional of $(\theta|V, W) \sim N(A, P^{-1})$ with

$$A = \begin{pmatrix} I \\ G_2 \\ G_3G_2 \\ \vdots \\ \prod_{t=2}^{T} G_t \end{pmatrix} a \quad \text{and}$$

$$P = \begin{pmatrix} P_{11} & P_{12} & 0 & \cdots & \cdots & \cdots & 0 \\ P_{21} & P_{22} & P_{23} & \ddots & & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & \ddots & \ddots & P_{T-1,T} \\ 0 & \cdots & \cdots & \cdots & 0 & P_{T,T-1} & P_{TT} \end{pmatrix}.$$

The precision matrix P is a symmetric, block tridiagonal matrix with main diagonal elements $P_{11} = R^{-1} + G_2' W^{-1} G_2$, $P_{TT} = W^{-1}$ and $P_{tt} = W^{-1} + G_{t+1}' W^{-1} G_{t+1}$, for $t = 2, \ldots, n-1$ and secondary diagonal elements given by $P_{t,t+1}' = P_{t+1,t} = W^{-1} G_t$, for $t = 1, \ldots, T-1$.

The likelihood is now given in matrix form by $y|\theta, V \sim N(F\theta, VI_T)$, where $y = (y_1, \ldots, y_T)$ and $F = \operatorname{diag}(F_1', \ldots, F_T')$. Combining prior with likelihood leads to the posterior $\theta|V, W, D_T \sim N(M, Q^{-1})$ where $M = Q^{-1}(V^{-1}Fy + PA)$ and $Q = P + V^{-1}F'F$. Note that, like P, Q is also block tridiagonal since $F'F$ is block diagonal.

Great computational advantages can be obtained from the sparseness of Q. In particular, fast inversion algorithms can be used ensuring that samples from θ are quickly drawn. Thus, an alternative algorithm is given by

ALGORITHM $1^{\dagger}$ (*Block sampling*).

Exactly as Algorithm 1^{*} *but for replacement of* θ *sampling* (*step* $\mathbf{b}^{*}$) *by:*

$\mathbf{b}^{\dagger}$. *sample* $\theta^{(j)}$ *from its joint full conditional* $N(M, Q^{-1})$, *using a fast inversion algorithm.*

Similar comments carry over to models with dependence of k lags in the system equation (1.2). Instead of bandwidth of 3 $(= 1 + 2.1)$ blocks in Q, one would have bandwidth of $(1 + 2.k)$ blocks but algorithms are still fast, provided k is orders of magnitude smaller than T. Results in this direction are more common in the context of Markov random fields (MRF) used in spatial models (Gamerman et al., 2003) but can be equally applied in dynamic modeling settings. Connections between MRF and DLM are described in Section 4.3.

A comparison between the sampling schemes is shown in Figure 1. The computational advantages are clearly seen from this figure. Blocking θ turns the convergence

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 17
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

issue into a much faster task. The computing time required for each iteration is higher but it is compensated by the smaller number of iterations required.

So far, we concentrated on θ sampling, with the hyperparameters V and W still sampled separately (from their full conditionals). It is possible however to sample all parameters in a single block. This task is achieved by use of

$$\pi(\theta, V, W|D_T) = \pi(\theta|V, W, D_T)\pi(V, W|D_T).$$

The joint density of all model parameters is known up to a proportionality constant and the distribution of $(\theta|V, W, D_T)$ was derived above. Therefore, the marginal density $\pi(V, W|D_T)$ is also known up to a proportionality constant. It does not have a known form and direct sampling becomes difficult. MCMC sampling can be used instead with the Metropolis–Hastings algorithm. Subsequent sampling of θ is done through its full conditional. This sampling strategy was discussed in Gamerman and Moreira (2002) and is summarized in Algorithm 2.

Algorithm 2 avoids MCMC convergence problems associated with the posterior correlation between θ and (V, W). Note that the algorithm uses $\pi(V, W|D_T)$ only in ratio form. Therefore, the unknown proportionality constant of this density is not required. The algorithm requires suitable choices of the proposal density q in **c1** to yield large enough values of α, thus ensuring computational efficiency. This may become a potentially important drawback, specially if the dimensionality of W becomes large. Usual choices of proposals are provided by a product of log random walk forms where $q(V, W|\theta^{(j-1)}, V^{(j-1)}, W^{(j-1)}) = q_1(V|\theta^{(j-1)}, V^{(j-1)})q_2(W|\theta^{(j-1)}, W^{(j-1)})$, with q_1 given by an inverse Gamma density centered around $V^{(j-1)}$ and q_2 given by an inverse Wishart density centered around $W^{(j-1)}$.

ALGORITHM 2 (*Joint sampling of all parameters*).

Exactly as Algorithm $1^\dagger$ *but for replacement of* (V, W) *sampling* (*step* **c**) *by:*

c1. *Sample* (V^*, W^*) *from* $q(V, W|\theta^{(j-1)}, V^{(j-1)}, W^{(j-1)})$;
c2. *set* $(V^{(j)}, W^{(j)}) = (V^*, W^*)$*, with probability* α *and* $(V^{(j-1)}, W^{(j-1)})$ *with probability* $1 - \alpha$*, where*

$$\alpha = \min\left\{1, \frac{\pi(V^*, W^*|D_T)}{\pi(V^{(j-1)}, W^{(j-1)}|D_T)} \frac{q(V^{(j-1)}, W^{(j-1)}|\theta^{(j-1)}, V^*, W^*)}{q(V^*, W^*|\theta^{(j-1)}, V^{(j-1)}, W^{(j-1)})}\right\}.$$

The case where the evolution disturbance matrix W is scaled by V can be easily accommodated within the framework here. Algorithms 1^*, $1^\dagger$ and 2 would sample (θ, V) jointly as now their full conditional is known to be in Normal-inverse Gamma form. Algorithm 2 would still consider sampling W via a proposal while Algorithms 1^* and $1^\dagger$ would sample W from its inverse Wishart full conditional.

The case where hyperparameters include also unknowns in the expression of F_t and G_t (examples of such models are given in Section 2) would require an additional step in the algorithms to handle sampling from these unknowns. Their full conditional will typically not be tractable for direct sampling and Metropolis proposals should be used. Unfortunately, the diversity of options available make it very hard to design multi-purpose proposals that would fit any model. Again, usual choices for proposals are

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 18
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

random walk forms centered around previous chain values on some suitable transformation leading to real-valued parameters and rescaled priors. Either way, some tuning of the precision of this proposal is needed to allow appropriate chain movement. This in turn, will hopefully lead to fast convergence.

2.2. Nonnormal models

This subsection considers the solution to full inference in nonnormal DLM by sampling. Presentation moves progressively away from normality, starting with (scale) mixture of normals, then moving to close approximation to scale mixture of normals and finally to exponential-family models.

Mixture of normals

The first case of nonnormality of the time series is associated with deviations from Gaussianity due to mixtures in the scale. In these cases, auxiliary variables ψ_t can be introduced such that $y_t|\psi_t, \theta, V \sim N(\mu_t, \psi_t V)$. The observation model is completed with specification of the mixing distribution for the scales ψ_t. These classes of models were studied in the context of DLM by Carlin et al. (1992) and include the important family of t-Student distribution obtained when ψ_t has inverse Gamma prior. This family is possibly the most common choice made by practitioners wanting to guard against marked observational departures from the standard or outliers.

An even simpler form of mixture of normals is provided by discrete mixture of normals where the observational density is given by $\sum_{i=1}^{k} p_i f_i(y|\mu_i, V_i)$ with the nonnegative weights p_i summing to 1 and $f_i(\cdot|\mu, V)$ denoting the $N(\mu, V)$ density. Defining mixture indicators z_t having multinomial distributions leads to normal models for $y_t|z_t$. An application of this setting can be presented in the context of Example 6. Kim et al. (1998) suggested that the $\log -\chi^2$ for the squared residual observation can be well approximated by a finite mixture of normals.

Another source of data nonnormality is provided by time series that can become normal after suitable transformations are performed to observations. Among normalizing transformations, the most common ones are provided by the Box–Cox family $g_\lambda(y) = (y^\lambda - 1)/\lambda$, for $\lambda \neq 0$ and $g_\lambda(y) = \log y$, for $\lambda = 0$. The model is completed with specification of a prior distribution for the λ_t's.

These three classes are qualitatively similar since they are examples of distributions that can are conditionally normalized. Conditional of the additional parameters Ψ respectively given by the ψ_t's, z_t's or λ_t's, the sampling schemes for all the other model parameters proceed as in the previous subsection. Additional sampling steps are required for the extra parameters.

Exponential-family models

The extensions above imply a MCMC solution that is based on extra steps in the sampling algorithms for normal DLMs. The observation models considered here in more detail are provided by the class of DGLM and do not fall into the categories above. The main problem of the previous section was the computational task of sampling θ efficiently but its full conditional was always directly available.

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 19
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

In the case of DGLM, the full conditional of θ does not fall into any recognizable form and can not be directly sampled from or integrated out. Not even the full conditional of θ_t is recognizable. Therefore, sampling from them is ruled out and proposals must be employed. Suitable proposals for θ with efficient correction terms are extremely hard to find due to the dimension of θ. Suitable proposals for θ_t are easier to find and can be constructed. However, they lead to slow convergence of the chain for the same reasons presented for normal models.

Shephard and Pitt (1997) and Knorr-Held (1999) suggested the use of random blocks containing some of the components of θ. Their proposals were based on normal approximations to the likelihood and random walk moves, respectively. Gamerman (1998) suggested the use of time-specific blocks but used a reparametrization in terms of the disturbances w_t's to avoid the strong correlation between θ_t's. The former can be rewritten as $w_t = \theta_t - G_t\theta_{t-1}$, $t = 2, \ldots, T$, in terms of the system parameters θ_t, with $w_1 = \theta_1$. The system parameters θ_t are easily recovered by the inverse relation

$$\theta_t = \sum_{l=1}^{t} \left(\prod_{k=1}^{t-l} G_{t-k+1} \right) w_l$$

for $t = 2, \ldots, n$ and $\theta_1 = w_1$. Note that $\pi(w|W, D_T) = \pi(\theta(w)|W, D_T)$ for $w = (w_1, \ldots, w_T)$, since the Jacobian of the transformation is 1. The sampling algorithm is given by

ALGORITHM 3 (*MCMC for DGLM based on state reparametrization*).

a. Initialization: *set initial values* $(\theta^{(0)}, W^{(0)})$ *and iteration counter* $j = 1$;
b. Sampling θ**:** $\theta^{(j)}$ *is sampled componentwise as follows*

1. *set* $t = 1$;
2. *Sample* w_t^* *from* $q_t(w_t|w_t^{(j-1)}, w_{-t}^{(j-1)}, W^{(j-1)}, D_T)$, *with* $w_{-t}^{(j-1)} = (w_1^{(j)}, \ldots, w_{t-1}^{(j)}, w_{t+1}^{(j-1)}, \ldots, w_T^{(j-1)})$;
3. *set* $w_t^{(j)} = w_t^*$ *with probability* α_t *and* $w_t^{(j)} = w_t^{(j-1)}$ *with probability* $1 - \alpha_t$, *where* $\alpha_t = min(1, A)$ *where* A *is*

$$\frac{\pi(w_t^*|w_{-t}^{(j)}, W^{(j-1)}, D_T)}{\pi(w_t^{(j-1)}|w_{-t}^{(j)}, W^{(j-1)}, D_T)} \quad \frac{q_t(w_t^{(j-1)}|w_t^*, w_{-t}^{(j-1)}, W^{(j-1)}, D_T)}{q_t(w_t^*|w_t^{(j-1)}, w_{-t}^{(j-1)}, W^{(j-1)}, D_T)}$$

4. *reconstruct* $\theta_t^{(j)} = G_t\theta_{t-1}^{(j)} + w_t^{(j)}$;
5. *set* $t \to t + 1$ *and return to* ***b2****, if* $t < T$;

c. Sampling W**:** $W^{(j)}$ *is sampled from its full conditional* $\pi(W|\theta^{(j)}, D_T)$;
d. Updating: *set* $j \to j + 1$ *and return to* ***b*** *until convergence.*

The full conditional densities of w_t required in the expression of α_t are simply obtained from the joint density $\pi(w|W, D_T)$ after eliminating terms not involving w_t and are known up to a proportionality constant. As it appears in ratio form, knowledge of the constant is unnecessary. The proposal densities q_t used by Gamerman (1998)

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 20
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

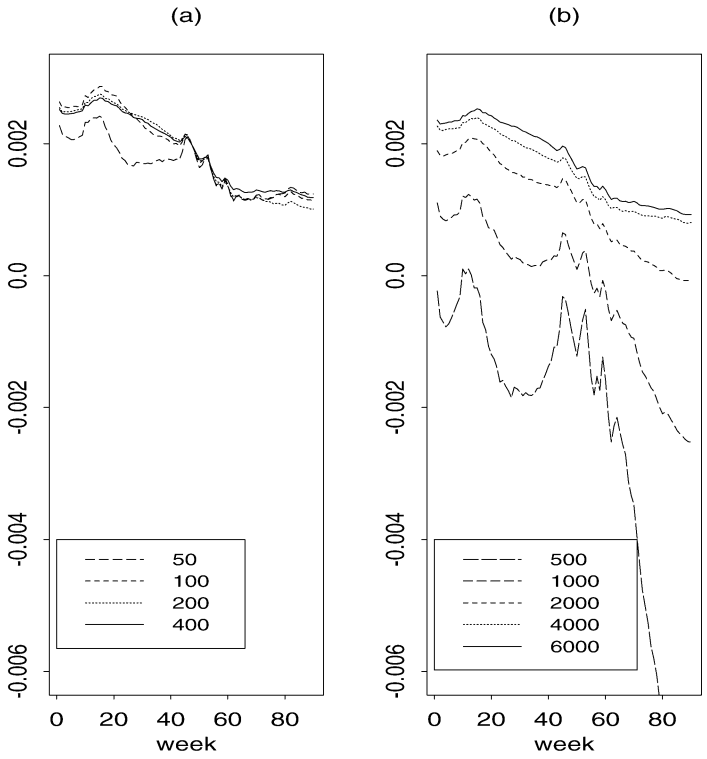

Fig. 3. Gamerman (1998) Average trajectory of a regression coefficient in 500 parallel chains with number of iterations for sampling from: (a) – w_t's; (b) – θ_t's. The data application is a dynamic logistic regression of expenditure on counts of advertising awareness.

are based on the working variables used by Singh and Roberts (1992) for mode evaluation is DGLM. They are given by the (normal) full conditional of w_t in a DLM with same system equation and modified observation equation $\tilde{y}_t \sim N(F_t'\theta_t, \widetilde{V}_t)$, with $\tilde{y}_t = g(\mu_t) + g'(\mu_t)(y_t - \mu_t)$ and $\widetilde{V}_t = [g'(\mu_t)]^2 \operatorname{Var}(y_t|\theta_t)$. These variables are an extension of the working variables used in maximum likelihood evaluation of generalized linear models. Ferreira and Gamerman (2000) detail and illustrate the use of this sampling scheme.

The iterations are more costly here because of the reconstruction equations above but provide savings in computing time. Figure 2 shows a comparison between convergence with and without this reparametrization. It shows a significant improvement of the reparametrization in terms of number of iterations to convergence. The savings observed are comparable to the ones observed in Section 2.1 in the comparison between sampling θ_t's separately or in a single block.

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 21
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

3. Sequential Monte Carlo

Let us start this section with a fairly general dynamic model:

$$\text{Observational Equation: } p(y_t|\theta_t, \psi), \tag{3.1}$$

$$\text{System Equation: } p(\theta_t|\theta_{t-1}, \psi), \tag{3.2}$$

where y_t is the vector of observable time series, θ_t is the vector of unobservable state parameters and ψ is the vector of static parameters, sometimes referred to as the *hyper-parameters*. For the moment suppose that ψ is known and omitted from the notation. Later on we will show how to include ψ in the analysis. Therefore, the *evolution equation* at each time t is given by (1.5) while the *updating equation* at each time t is given by (1.7). Apart from the general normal DLM case (see Section 1), the filtering densities cannot be obtained analytically and numerical methods must be used.

The advent of MCMC methods generated a number of algorithms to sample from $p(\theta_1, \ldots, \theta_T|D_T)$, described in the previous section. Therefore, in principle, one could simply rerun the MCMC algorithm as data become available. However, in many practical situations, the cost of rerunning MCMC routines are prohibitive and real-time sequential algorithms play a crucial role in assessing $p(\theta_t|D_t)$. Doucet (2003) illustrates this issue pointing to a situation in wireless communication where you do not want to wait for a MCMC chain to converge in order to listen to your correspondent's next bit of conversation. Another common situation is when a financial analyst needs to forecast a few steps ahead, say for the next five hours, the volatilities for several assets that compound her portfolio.

Analytical approximations have been proposed in the literature. Kitagawa (1987), for instance, has introduced an algorithm where piecewise linear functions approximates the densities in both Eqs. (1.5) and (1.7). Similarly, Pole and West (1990) apply quadrature techniques to the class of conditionally conjugate models where one, or more, of the four elements that defines the general normal linear dynamic model is function of unknowns, say ψ. Quadrature techniques has recently reappeared in Bolviken and Storvik (2001). Both algorithms, based on linear splines and Gaussian quadratures, become computationally prohibitive as either the sample size or the dimension of the state vector increases.

The remainder of this section is dedicated to introducing several *particle filters*, as most sequential Monte Carlo algorithms are commonly referred to (Kitagawa, 1996). We start with the *bootstrap filter* or simply the *SIR filter*, in Section 3.1. Sequential importance sampling is also introduced, but immediately discarded because of an inherent degeneracy problem as t increases. In Section 3.2, Pitt and Shephard's (1999) extensively used *auxiliary particle filter* is introduced. Section 3.3 presents an extension of Pitt and Shephard's filter, proposed by Liu and West (2001), to deal with the sequential estimation of fixed parameters, ψ in Eqs. (3.1) and (3.2). Alternative schemes for parameter estimation are briefly discussed.

Before we proceed, let us introduce some notation that will facilitate algorithmic expositions. $\{(\theta_t^{(1)}, w_t^{(1)}), \ldots, (\theta_t^{(N)}, w_t^{(N)})\} \overset{a}{\sim} p(\theta_t|D_t)$ is used to denote that the probability density function, $p(\theta_t|D_t)$, of the continuous random variable, θ_t, is approximated by a discrete variable with random support. Therefore, if one is interested,

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 22
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

for instance, in computing $E(g(\theta_t)|D_t)$, an approximation based on the set of points $\theta_t^{(1)}, \ldots, \theta_t^{(N)}$ is $\sum_{i=1}^{N} w_t^{(i)} g(\theta_t^{(i)})$.

3.1. SIR- and SIS-based filters

Simultaneous rediscovery of related approaches can be tracked back to Gordon et al. (1993), West (1993a, 1993b), and Kitagawa (1993, 1996) with the first becoming a seminal paper in the field of sequential Monte Carlo methods, at least for (Bayesian) statisticians. Doucet (2003) argues that Hetherington (1984) was the first to introduce, in physics, a multinomial resampling step. Gordon et al. (1993) use iterative sampling importance resampling (SIR, in short) (Rubin, 1988; Smith and Gelfand, 1993) in what turned out to be worldly known as the *bootstrap filter*. For consistency with the other filters, we will simply call it the *SIR filter*.

More specifically, if $\theta_{t-1}^{(1)}, \ldots, \theta_{t-1}^{(N)}$ represents a sample, commonly called *particles*, from $p(\theta_{t-1}|D_{t-1})$ a natural step is to use the evolution equation (3.2) to sample a new set of particles, $\tilde{\theta}_t^{(1)}, \ldots, \tilde{\theta}_t^{(N)}$. The new set of particles represent a sample from $p(\theta_t|D_{t-1})$, which is the state prior distribution at time t. Finally, the particles $\tilde{\theta}_t^{(i)}$ are resampled, for simplicity N times, with weights proportional to $p(y_t|\tilde{\theta}_t^{(i)})$ to produce a new set of particles $\theta_t^{(1)}, \ldots, \theta_t^{(M)}$ that are approximately distributed as $p(\theta_t|D_t)$. This is schematically represented in Algorithm 4.

The evolution and updating steps are also known as importance sampling and selection steps, respectively. It can be easily seen that $I_1 = \sum_{i=1}^{N} \omega_t^{(i)} g(\tilde{\theta}_t^{(i)})$ converges to $I = \int g(\theta_t) p(\theta_t|D_t)\, \mathrm{d}\theta_t$ as $N \to \infty$. Theoretically, the updating step in the previous algorithm is unnecessary and, in fact, $\mathrm{Var}(I_1) \leqslant \mathrm{Var}(I_2)$, where $I_2 = \sum_{i=1}^{N} g(\theta_t^{(i)})/N$ is also a consistent estimator of I (Geweke, 1989). Lopes et al. (1999) and Schmidt et al. (1999) investigate empirically the performance of both procedures in estimating fixed parameters in a class of nonlinear/nonnormal dynamic model.

ALGORITHM 4 (*SIR filter*).

Posterior at $t-1$**:** $\{(\theta_{t-1}^{(1)}, \frac{1}{N}), \ldots, (\theta_{t-1}^{(N)}, \frac{1}{N})\} \overset{a}{\sim} p(\theta_{t-1}|D_{t-1})$
Evolution: *For* $i = 1, \ldots, n$*, sample* $\tilde{\theta}_t^{(i)}$ *from* $p(\theta_t|\theta_{t-1}^{(i)})$
Weights: *For* $i = 1, \ldots, n$*, compute* $\omega_t^{(i)} \propto p(y_t|\tilde{\theta}_t^{(i)})$
Updating: *For* $i = 1, \ldots, n$*, sample* $\theta_t^{(i)}$ *from* $\{(\tilde{\theta}_t^{(1)}, \omega_t^{(i)}), \ldots, (\theta_t^{(N)}, \omega_t^{(i)})\}$
Posterior at t**:** $\{(\theta_t^{(1)}, \frac{1}{N}), \ldots, (\theta_t^{(N)}, \frac{1}{N})\} \overset{a}{\sim} p(\theta_t|D_t)$

If the updating step is deleted from the SIR filter, then sample $\theta_{t-1}^{(1)}, \ldots, \theta_{t-1}^{(N)}$ must be accompanied by the weights $\omega_{t-1}^{(1)}, \ldots, \omega_{t-1}^{(N)}$, and the previous algorithm becomes Algorithm 5.

ALGORITHM 5 (*SIS filter*).

Posterior at $t-1$**:** $\{(\theta_{t-1}^{(1)}, \omega_{t-1}^{(1)}), \ldots, (\theta_{t-1}^{(N)}, \omega_{t-1}^{(N)})\} \overset{a}{\sim} p(\theta_{t-1}|D_{t-1})$

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 23
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

Evolution: *For* $i = 1, \ldots, n$, *sample* $\theta_t^{(i)}$ *from* $p(\theta_t|\theta_{t-1}^{(i)})$
Weights: *For* $i = 1, \ldots, n$, *compute* $\omega_t^{(i)} \propto \omega_{t-1}^{(i)} p(y_t|\theta_t^{(i)})$
Posterior at t**:** $\{(\theta_t^{(1)}, \omega_t^{(1)}), \ldots, (\theta_t^{(N)}, \omega_t^{(N)})\} \overset{a}{\sim} p(\theta_t|D_t)$

At any given time t, the importance function in the SIS filter is the prior distribution for the whole state vector up to time t, i.e., $p(\theta_0)\prod_{j=1}^{t} p(\theta_j|\theta_{j-1})$. The dimension of $(\theta_1, \ldots, \theta_t)$ increases with t and, as it is well known, importance sampling techniques become practically infeasible even for moderately large t (Geweke, 1989). It is not unusual, after a few step in the SIS algorithm, to wind up with only a handful of particles with nonnegligible weights. One of the neatest aspects of the SIR filter is that at each time point only particles with high weights are kept, or using *genetic algorithm* terminology, at each generation only strong strings are selected. See Higuchi (1997, 2001) for further comparisons between *Monte Carlo filters*, as he calls sequential Monte Carlo methods, and genetic algorithms.

Finally, since at any given time t what we are effectively trying to do is to sample from the posterior distribution, $p(\theta_t|D_t)$, rejection sampling and MCMC methods could be alternatively used. Their main drawback is the mostly invariable need of evaluation the prior distribution, $p(\theta_t|D_{t-1})$, which can be troublesome. Müller (1991, 1992) proposes a Metropolis algorithm where $p(\theta_t|D_{t-1})$ is reconstructed by a mixture of Dirichlet process model based on the prior sample, $(\tilde{\theta}_t^{(1)}, \ldots, \tilde{\theta}_t^{(N)})$. West (1993a, 1993b) uses mixture of multivariate normals to reconstructed the prior and posterior distributions of the state vector at each time.

Below we introduce yet another filter, the *auxiliary particle filter*, introduced by Pitt and Shephard (1999) to tackle the main weakness of the SIR filter, which is sampling from the prior, $p(\theta_t|D_t)$ and perform poorly when the next observation, y_t, is on the tail of $p(y_t|\theta_t)$.

3.2. Auxiliary particle filter

Assuming that $\{(\theta_{t-1}^{(1)}, \omega_{t-1}^{(1)}), \ldots, (\theta_{t-1}^{(N)}, \omega_{t-1}^{(N)})\} \overset{a}{\sim} p(\theta_{t-1}|D_{t-1})$, a natural Monte Carlo approximation (as $M \to \infty$) for the prior (1.5) is

$$\hat{p}(\theta_t|D_{t-1}) = \sum_{j=1}^{N} p\big(\theta_t|\theta_{t-1}^{(j)}\big) w_{t-1}^{(j)} \tag{3.3}$$

which, following Pitt and Shephard's (1999) terminology, is called the *empirical prediction density*. Combining this approximate prior with the observation equation produces, by Bayes' theorem, the following approximation for the state space vector posterior distribution at time t is

$$\begin{aligned}\hat{p}(\theta_t|D_t) &\propto p(y_t|\theta_t) \sum_{j=1}^{N} p\big(\theta_t|\theta_{t-1}^{(j)}\big) w_{t-1}^{(j)} \\ &\propto \sum_{j=1}^{N} p(y_t|\theta_t)\, p\big(\theta_t|\theta_{t-1}^{(j)}\big) w_{t-1}^{(j)}\end{aligned} \tag{3.4}$$

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 24
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

the *empirical filtering density* according to Pitt and Shephard (1999). A sampling scheme from the approximate posterior (3.4) is needed in order to complete the evolution/update cycle, very much like the SIR and the SIS filters. However, as it was discussed previously, the SIR method becomes ineffective either when the prior is relatively diffuse or the likelihood is highly informative.

ALGORITHM 6 (*Auxiliary particle filter*).

Posterior at $t-1$**:** $\{(\theta_{t-1}^{(1)}, \omega_{t-1}^{(1)}), \ldots, (\theta_{t-1}^{(N)}, \omega_{t-1}^{(N)})\} \overset{a}{\sim} p(\theta_{t-1}|D_{t-1})$
Sampling (k, θ_t)**:** *For* $i = 1, \ldots, N$

Indicator: *sample* k^i *such that* $Pr(k^i = k) \propto p(y_t|\boldsymbol{\mu}_t^{(k)}) w_{t-1}^{(k)}$

Evolution: *sample* $\theta_t^{(i)}$ *from* $p(\theta_t|\theta_{t-1}^{(k^i)})$

Weights: *compute* $w_t^{(i)} \propto p(y_t|\theta_t^{(i)}) / p(y_t|\boldsymbol{\mu}_t^{(k^j)})$

Posterior at t**:** $\{(\theta_t^{(1)}, \omega_t^{(1)}), \ldots, (\theta_t^{(N)}, \omega_t^{(N)})\} \overset{a}{\sim} p(\theta_t|D_t)$

Pitt and Shephard improve on particle filter methods by addressing to practical and important issues: (i) efficiently sampling from the approximate posterior distribution (Eq. (3.4)) and (ii) efficiently approximating tails' behavior of the approximate prior (Eq. (3.3)). They developed a generic filtering algorithm that is currently well known as *auxiliary particle filtering*. The basic feature of a auxiliary particle filter is to take advantage of the mixture of densities (3.4) to obtain draws from $p(\theta_t|D_t)$ by introducing latent indicator variables to identify the terms in the mixture (an idea commonly used in mixture modeling, Diebolt and Robert, 1994). In other words, if (θ_t, k) is sampled from $p(\theta_t, k) \propto p(y_t|\theta_t) p(\theta_t|\theta_{t-1}^{(k)}) w_{t-1}^{(k)}$, the resulting θ_t is a sample from (3.4). If $\{(\theta_{t-1}^{(1)}, \omega_{t-1}^{(1)}), \ldots, (\theta_{t-1}^{(N)}, \omega_{t-1}^{(N)})\} \overset{a}{\sim} p(\theta_{t-1}|D_{t-1})$, then for $i = 1, \ldots, N$, sample $(\theta_t^{(i)}, k^i)$ from $g(\theta_t, k|D_t)$ and compute weights $w_t^{(i)} \propto p(y_t|\theta_t^{(i)}) p(\theta_t^{(i)}|\theta_{t-1}^{(k^i)}) g(\theta_t^{(i)}, k^i|D_t)$. By following these steps, which are essentially SIR steps, $\{(\theta_t^{(1)}, \omega_t^{(1)}), \ldots, (\theta_t^{(N)}, \omega_t^{(N)})\} \overset{a}{\sim} p(\theta_t|D_t)$.

Pitt and Shephard use $g(\theta_t, k|D_t) \propto p(y_t|\boldsymbol{\mu}_t^{(k)}) p(\theta_t|\theta_{t-1}^{(k)}) w_{t-1}^{(k)}$, as a generic importance function, where $\boldsymbol{\mu}_t^{(k)}$ is an estimate of θ_t given $\theta_{t-1}^{(k)}$, for instance the mean, the mode or any other likely value from $p(\theta_t|\theta_{t-1})$), such that $g(k|D_t) \propto p(y_t|\boldsymbol{\mu}_t^{(k)}) w_{t-1}^{(k)}$.

Choosing $g(\cdot)$ is a nontrivial task and this is inherent to all Monte Carlo methods (SIR, adaptive, MCMC, etc.). Pitt and Shephard argue that the simulation algorithm will favor particles with larger predictive likelihoods. By doing so, the resampling step will have lower computational cost and will improve on statistical efficiency of the procedure.

3.3. Parameter estimation and sequential Monte Carlo

In this section the uncertainty about the hyperparameter ψ is incorporated in the filtering analysis. The main algorithm we present here is due to Liu and West (2001), who

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 25
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

extend Pitt and Shephard's auxiliary particle filter. They combine kernel density estimation techniques with artificial parameter evolution and propose a novel algorithm to sequentially treat fixed parameters in general dynamic model settings.

Initially, let (1.7) be rewritten as

$$p(\theta_t, \psi|D_t) \propto p(y_t|\theta_t, \psi)p(\theta_t|\psi, D_{t-1})p(\psi|D_{t-1}), \tag{3.5}$$

where the uncertainty about ψ is assessed by adding the term $p(\psi|D_{t-1})$. As before, and conditional on ψ, the evolution density $p(\theta_t|\psi, D_{t-1})$ can be approximated by

$$\hat{p}(\theta_t|\psi, D_{t-1}) = \sum_{j=1}^{N} p\big(\theta_t|\psi, \theta_{t-1}^{(j)}\big)w_{t-1}^{(j)}, \tag{3.6}$$

where $\{(\theta_{t-1}^{(1)}, \psi_{t-1}^{(1)}, w_{t-1}^{(1)}), \ldots, (\theta_{t-1}^{(N)}, \psi_{t-1}^{(N)}, w_{t-1}^{(N)})\} \overset{a}{\sim} p(\theta_{t-1}, \psi|D_{t-1})$.

A natural solution, firstly explored by Gordon et al. (1993), is to pretend that the fixed parameters are states in the dynamic modeling, for instance, by adding small random disturbances to artificial evolutions, and proceed the analysis with the auxiliary particle filters presented in the previous section. Such artificial evolution reduces the sample degeneracy problems. However, it imputes unnecessary uncertainty into the model and also creates artificial loss of information resulting on overdispersion of the posterior distributions.

Liu and West (2001) reinterpret Gordon, Salmond, and Smith's artificial parameter evolution idea and combine it with West's kernel smoothing techniques. Approximations for $p(\psi|D_t)$ based on mixtures of multivariate normals were suggested by West (1993b),

$$\hat{p}(\psi|D_{t-1}) = \sum_{j=1}^{M} N\big(\psi|\boldsymbol{m}_{t-1}^{(j)}, h^2\boldsymbol{V}_{t-1}\big)w_{t-1}^{(j)}, \tag{3.7}$$

where h is a smoothing parameter, $\boldsymbol{V}_{t-1} = \mathrm{Var}(\psi|D_{t-1})$, and $\boldsymbol{\mu}_{t-1}^{(j)}$ are the locations of the components of the mixture. In standard kernel methods, $\boldsymbol{\mu}_{t-1}^{(j)} = \psi_{t-1}^{(j)}$. Also, for large M, it is also common practice to have h as a decreasing function of M. West (1993b) introduces a shrinkage rule for the locations,

$$\boldsymbol{m}_{t-1}^{(j)} = a\psi_{t-1}^{(j)} + (1-a)\bar{\boldsymbol{\psi}}_{t-1}, \tag{3.8}$$

where $\bar{\boldsymbol{\psi}}_{t-1} = E(\psi_{t-1}|D_{t-1})$. The variance of the resulting mixture of normals is $(a^2 + h^2)\boldsymbol{V}_{t-1}$, which is always larger then $\boldsymbol{V}_{t-1}$ for $a^2 + h^2 > 1$. West (1993b) suggests using $a^2 = 1 - h^2$ to guarantee that the correct variance is used in the approximation, crucial in sequential schemes. Liu and West (2001) show that if δ is the discount factor used in Gordon, Salmond, and Smith's artificial evolution method, then defining $h^2 = 1 - [(3\delta - 1)/2\delta]^2$ produces an algorithm that links the kernel estimation (3.6) with the shrinkage idea (3.8).

Liu and West (2001) apply their novel strategy in a couple of situations, including a simpler version of Aguilar and West's (2000) dynamic factor model for financial time series, which are similar to models previously described in this chapter. Among other

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 26
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

empirical findings, they argue that MCMC methods should be combined with sequential algorithms in real applications. They show that, when performed for longer periods of time the filtering algorithm starts to deteriorate and diverges from the "gold standard" MCMC results.

ALGORITHM 7 (*Liu–West filter*).

At $t-1$**:** $\{(\theta_{t-1}^{(1)}, \psi_{t-1}^{(1)}, w_{t-1}^{(1)}), \ldots, (\theta_{t-1}^{(N)}, \psi_{t-1}^{(N)}, w_{t-1}^{(N)})\} \overset{a}{\sim} p(\theta_{t-1}, \psi | D_{t-1})$.

Kernel: $\boldsymbol{V}_{t-1} = \sum_{j=1}^{M} (\psi_{t-1}^{(j)} - \bar{\boldsymbol{\psi}}_{t-1})(\psi_{t-1}^{(j)} - \bar{\boldsymbol{\psi}}_{t-1})' w_{t-1}^{(j)}$ *and* $\bar{\boldsymbol{\psi}}_{t-1} = \sum_{j=1}^{M} \psi_{t-1}^{(j)} w_{t-1}^{(j)}$

Shrinkage: *For* $j = 1, \ldots, N$, $\boldsymbol{\mu}_t^{(j)} = E(\theta_t | \theta_{t-1}^{(j)}, \psi^{(j)})$ *and* $\boldsymbol{m}_{t-1}^{(j)} = a\psi_{t-1}^{(j)} + (1-a)\bar{\boldsymbol{\psi}}_{t-1}$

Sampling (k, θ_t, ψ)**:** *For* $i = 1, \ldots, N$

- **Indicator:** *Sample* k^i *such that* $Pr(k^i = k) \propto w_{t-1}^{(k)} p(y_t | \boldsymbol{\mu}_t^{(k)}, \boldsymbol{m}_{t-1}^{(k)})$
- **Parameter:** *Sample* $\psi_t^{(i)}$ *from* $N(\boldsymbol{m}_{t-1}^{(k^i)}; h^2 \boldsymbol{V}_{t-1})$
- **State:** *Sample* $\theta_t^{(i)}$ *from the* $p(\theta_t | \theta_{t-1}^{(k^i)}, \psi_t^{(i)})$
- **Weights:** $w_t^{(i)} \propto p(y_t | \theta_t^{(i)}, \psi_t^{(i)}) / p(y_t | \boldsymbol{\mu}_t^{(k^i)}, \boldsymbol{m}_{t-1}^{(k^i)})$

At t**:** $\{(\theta_t^{(1)}, \psi_t^{(1)}, w_t^{(1)}), \ldots, (\theta_t^{(N)}, \psi_t^{(N)}, w_t^{(N)})\} \overset{a}{\sim} p(\theta_t, \psi | D_t)$.

Computing predictive densities

Regardless of which filter is used, the predictive density

$$p(y_t | D_{t-1}) = \int p(y_t | \theta_t, \psi) p(\theta_t, \psi | D_{t-1}) \, \mathrm{d}\theta_t \, \mathrm{d}\psi$$

can be approximated by

$$\hat{p}(y_t | D_{t-1}) = \frac{1}{N} \sum_{j=1}^{N} p\big(y_t | \tilde{\theta}_t^{(l)}, \tilde{\psi}_t^{(l)}\big),$$

where $\{(\tilde{\theta}_t^{(1)}, \tilde{\psi}_t^{(1)}, \frac{1}{N}), \ldots, (\tilde{\theta}_t^{(N)}, \tilde{\psi}_t^{(N)}, \frac{1}{N})\} \overset{a}{\sim} p(\theta_t, \psi | D_{t-1})$, the joint prior distribution at time t of the states and hyperparameters. The pair $(\tilde{\theta}_t^{(i)}, \tilde{\psi}_t^{(i)})$ can be sampled from $p(\theta_t, \psi | D_{t-1})$ by following three simple steps: (i) sample k^i such that $\Pr(k^i = j) \propto w_t^{(j)}$, (ii) make $\tilde{\psi}_t^{(i)} = \psi_{t-1}^{(k^i)}$, and (iii) sample $\tilde{\theta}_t^{(i)}$ from the evolution equation $p(\theta_t | \theta_{t-1}^{(k^i)}, \tilde{\psi}_t^{(i)})$. If one is interested, for instance, in computing $E(g(y_t) | D_{t-1})$, an approximation based on $\{(\tilde{\theta}_t^{(1)}, \tilde{\psi}_t^{(1)}, \frac{1}{N}), \ldots, (\tilde{\theta}_t^{(N)}, \tilde{\psi}_t^{(N)}, \frac{1}{N})\}$ is $\widehat{E}(g(y_t) | D_{t-1}) = \frac{1}{N} \sum_{i=1}^{N} E(g(y_t) | \tilde{\theta}_t^{(i)}, \tilde{\psi}_t^{(i)})$. These approximations are extremely useful for sequential model comparison.

EXAMPLE 3.1. Lopes and Marinho (2002) studied the Markov switching stochastic volatility (MSSV) model, where $y_t = \mathrm{e}^{\lambda_t/2} u_t$ and $\lambda_t = \alpha_{S_t} + \phi\lambda_{t-1} + \varpi_t$, for $u_t \sim N(0,1)$ and $\varpi_t \sim N(0, \sigma^2)$ serially and temporally uncorrelated. S_t followed a homogeneous k-state first-order Markov process with transition matrix P where

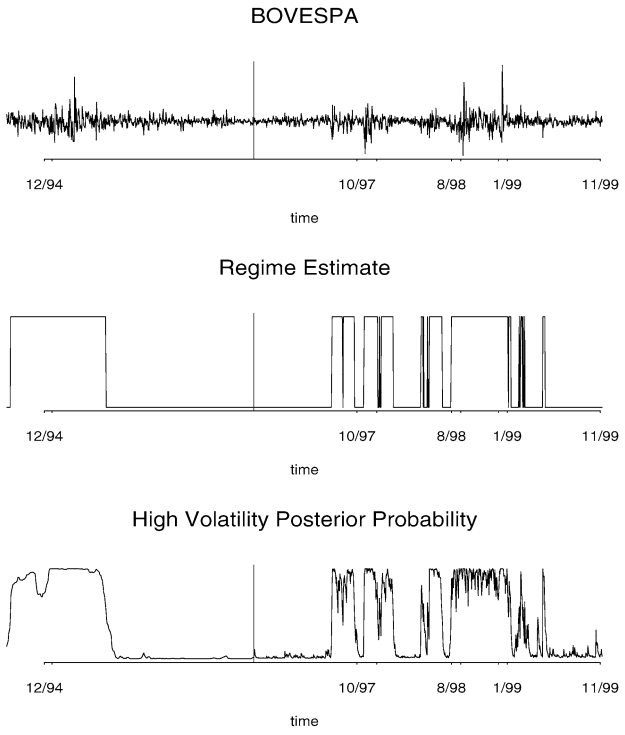

Fig. 4. Brazilian Bovespa: time series (top), regimes posterior mean (middle) and posterior mean of probability that the series in the "high volatility" regime at any given time t.

$p_{ij} = P[S_t = j|S_{t-1} = i]$, for $i, j = 1, \ldots, k$. The $k^2 + 2$-dimensional parameter vector is $\psi = (\alpha_1, \ldots, \alpha_k, \phi, \sigma^2, p_{11}, \ldots, p_{1,k-1}, \ldots, p_{k1}, \ldots, p_{k,k-1})$, while the state vector is $\theta_t = (S_t, \lambda_t)$. Figure 4 exhibits the sequential Monte Carlo estimates of $p(y_t|D_{t-1})$ for all t when y_t is the Bovespa stock index (São Paulo Stock Exchange). An extension of this model appears in Lopes (2002) who uses the MSSV structure to model common factor log-volatilities in a factor stochastic volatility model (Lopes, 2000).

3.4. Recent developments

There have been several theoretical developments that try to explain the behavior of sequential Monte Carlo filters as the number of particles, M in our notation, tends to infinity (see, for instance, Berzuini et al., 1997; Crisan and Doucet, 2002; Künsch, 2001). However, little is formally known about particle filter behavior when M is kept fixed and the sample size increases.

Alternative approaches to deal with parameter estimation in sequential particle filtering is already flourishing. Storvik (2002) proposed histogram-like filters that take advantage of sufficient statistics present in the state equations of several existing dynamic models. Polson et al. (2002) generalizes Storvik (2002) by using a fixed-lag

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 28
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

approximation that simplifies the computation and improves the performance of the filter both in terms of estimation and computational speed.

Doucet and Tadić (2003) combines gradient algorithms with particle filters to estimate, by recursive maximum likelihood, the fixed parameters, ψ in our notation, in a fairly broad class of dynamic models.

4. Extensions

In the last decade, there have been new areas of statistical research derived from the extension of dynamic linear models to topologies other than time as, for example, space-time, and dyadic and quad trees. Moreover, there has been increasing understanding of the connections between DLMs and Gaussian Markov random fields. These developments have pointed out to exciting new applications to areas such as finance, epidemiology, climatology and image analysis.

In this final section of the chapter, the development of dynamic models for spatio-temporal processes is addressed in Section 4.1, the development of multi-scale models for processes living at different levels of resolution is addressed in Section 4.2, and some connections between Markov random fields and DLMs is addressed in Section 4.3.

4.1. Dynamic spatio-temporal models

There has been increasing interest in the statistics community on the development of models for spatio-temporal processes. Within the statistical community, initial efforts to analyze spatio-temporal data include the work of Haslett and Raftery (1989) on the assessment of wind power resource in Ireland and Handcock and Wallis (1994) on the analysis of meteorological fields. Cressie and Huang (1999) have worked on the development of nonseparable spatio-temporal variograms.

Ghil et al. (1981) were the first to consider the use of state space models to spatio-temporal data. Let $y(s_1, t_1), \ldots, y(s_N, t_T)$ be data obtained from a spatially continuous process $y(s, t)$ where s belongs to some spatial domain R^2 or R^3 and t indexes discrete times. Let $y_t = (y(s_1, t), \ldots, y(s_N, t))$ be the vectorized observed spatial data at time t. A dynamic linear spatio-temporal model can be written as (1.1)–(1.2), where the interpretation of the different elements of the model is analogous to the interpretation in traditional multivariate dynamic linear models. θ_t is the latent vectorized spatio-temporal process. F_t is the matrix that connects the latent process with the observations. ε_t is an idiosyncratic random error at time t. G_t describes the evolution of the latent spatio-temporal process through time. ω_t is a random error with effect not only at time t but also at subsequent times. The matrices V_t and W_t play the very important role of describing the spatial dependence of the process at the observation and state space levels. Eqs. (1.1) and (1.2) define a very flexible class of models for spatio-temporal processes, the key feature for the successful application of these models being the specification of the latent process θ_t and of the matrices F_t, G_t, V_t and W_t.

In the simplest dynamic linear spatio-temporal model, F_t and G_t are identity matrices, that is, θ_t is the level of the observed process and follows a random walk process

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 29
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

through time. More interesting cases include some type of dimension reduction from the observed field y_t to the latent process θ_t. When there is dimension reduction, the latent process has much smaller dimension than the observations, and the matrix F_t makes the connection between them. This dimension reduction can be achieved through some type of factor modeling or through some type of spatial aggregation. Other interesting spatio-temporal dynamic models use the matrix G_t to incorporate expert knowledge of the physical characteristics of the process under study (e.g., Wikle et al., 2001).

The specification of V_t and W_t as covariance matrices of spatial processes is quite a challenge because spatio-temporal datasets are usually large and the computations necessary for the estimation of the parameters of the model become prohibitive. Thus, specifying V_t and W_t using geostatistical models such as the Matérn class is unfeasible because of the computational burden. An alternative approach that is being investigated by one of the authors is the use of Markov random fields in the specification of V_t and W_t (Vivar-Rojas and Ferreira, 2003).

Other alternative approaches that reduce the computational burden include some type of dimension reduction from the observations to the latent process through the matrix F_t. In this direction, the literature includes approaches based on principal components and factor analysis (Cressie, 1994; Goodall and Mardia, 1994; Mardia et al., 1998; Wikle and Cressie, 1999). In the other hand, dimension reduction based on multiscale ideas is being developed by Johannesson et al. (2003). As the approaches based on principal components and factor analysis are at this time better established, let us now describe them in more detail. In particular, let us review the quite general approach proposed by Wikle and Cressie (1999). Their model is the following:

$$y(s,t) = z(s,t) + \varepsilon(s,t), \tag{4.1}$$

$$z(s,t) = \theta(s,t) + \gamma(s,t), \tag{4.2}$$

$$\theta(s,t) = \int w_s(u)\theta(u,t-1)\,\mathrm{d}u + \eta(s,t), \tag{4.3}$$

where $y(s,t)$ are data from a continuous spatial process observed at discrete times, $z(s,t)$ is a nonobservable smooth process, $\varepsilon(s,t)$ is the measurement error, $\gamma(s,t)$ is a variance component, $\theta(s,t)$ is the latent spatio-temporal process that evolves through time, $w_s(u)$ is an interaction function between site s and its *neighbors* through time, and $\eta(s,t)$ are independent Gaussian errors. They assume that $z(s,t)$ can be decomposed into K dominant components $z(s,t) = \sum_{k=1}^{K} \phi_k(s)a_k(t)$ where $a_k(\cdot), k = 1,\ldots,K$, are zero mean time series and $\phi_1(\cdot), \phi_2(\cdot), \ldots$ are complete and orthonormal basis functions. Moreover, taking advantage of the completeness of the $\phi's$, they expand the interaction function as $w_s(u) = \sum_{l=1}^{\infty} b_l(s)\phi_l(u)$. In addition, using the orthonormality property and after some algebraic manipulation, they arrive at a DLM-like formulation:

$$y(s,t) = \phi(s)'a(t) + \gamma(s,t) + \varepsilon(s,t), \tag{4.4}$$

$$a(t) = Ha(t-1) + J\eta(t), \tag{4.5}$$

where H and J are matrices that depend on the ϕ's and the b's. The identifiability issue on the roles of $\varepsilon(s,t)$ and $\gamma(s,t)$ is resolved by assuming that $\varepsilon(s,t)$ represents a nugget effect and $\gamma(\cdot,t)$ follows a L_2-continuous random field process. In order to

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 30
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

save computational time, they propose an empirical Bayes procedure to estimate the unknown matrices and then use the Kalman filter to estimate the state space parameters. See Wikle and Cressie (1999) for more details.

Wikle et al. (2001) propose a dynamic spatio-temporal model to tropical ocean surface winds. They incorporate scientific knowledge about tropical winds into the model through the evolution matrix G_t. Moreover, they take a hierarchical approach to incorporate two different types of data: (1) satellite data obtained with a scatterometer; (2) wind fields generated by global scale numerical weather prediction models from sparse in situ observations. Motivated also by the combination of data from different sources, Brown et al. (2001) develop a space-time model for the calibration of radar rainfall data. Another example of incorporation of scientific knowledge in a dynamic spatio-temporal model is given by Wikle (2003), where a diffusion process is included in the evolution equation to model the spread of a species population.

Wikle (2003) proposes a class of dynamic spatio-temporal models with evolution equation based on kernel convolutions.

Stroud et al. (2001) propose a locally weighted mixture of linear regressions, allowing the regression surfaces to change over time. Using a similar idea, Huerta et al. (2004) present a nice application of spatio-temporal models to the analysis of ozone levels, with F_t being regressors and θ_t temporally varying regressor coefficients.

These ideas involve either spatially or time varying regression coefficients. Banerjee et al. (2003) propose the use of spatially varying coefficients evolving in time according to a multivariate random walk process while making use of general forms for the evolution disturbance processes through coregionalization ideas. Related to that work, Gamerman (2002) considers spatio-temporally varying coefficients and also nonnormality, irregularities in the process of data collection and structural changes in the spatio-temporal process.

Dynamic spatio-temporal modeling is an area of active research, and we expect several important new developments on the subject in the following years.

4.2. Multi-scale modeling

Multi-scale modeling has mainly appeared in the engineering literature. Basseville et al. (1992) introduced isotropic multi-scale models defined on dyadic trees using a definition analogous to the one of autoregressive time series models. Chou et al. (1994) introduced the state-space representation of multi-scale models as an extension of the dynamic linear models, inference being efficiently carried out with a variant of the Kalman filter. Basseville et al. (1992) and Chou et al. (1994) define the model from coarser to finer levels and assume that the nodes of a given level are conditionally independent given the immediate coarser level. Consider a dyadic or quad tree with a corresponding latent process denoted by θ and observations y. Denote by $t = (m, n)$ the index of the nth node of the mth scale and by $\Delta(t-1)$ its parent node. The model proposed by Chou et al. (1994) can be written in a DLM formulation as follows:

$$y_t = F_t\theta_t + v_t, \tag{4.6}$$

$$\theta_t = G_t\theta_{\Delta(t-1)} + w_t, \tag{4.7}$$

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 31
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

where as usual w_t and v_t are independent zero mean processes with covariance matrices W_t and V_t respectively. The matrix F_t connects the state process θ_t to the observation y_t. The matrix G_t relates the tth node of the tree with its parent.

If the matrices F_t, G_t, W_t and V_t are known, then the estimation of the state parameters can be performed with a Kalman filter-like algorithm with some important modifications. First, while for processes on time the algorithm is composed by a filtering step forward in time and a smoothing step backward in time, the algorithm for processes on trees is composed by a filtering sweep from finer to coarser levels followed by a coarser to finer levels smoothing sweep. Second, the tree structure of the model and thus the algorithm have a pyramidal structure that lends itself to efficient parallel implementation.

While the assumption of conditionally independence of the nodes of a given level given the immediate coarser level leads to an efficient estimation algorithm, it also leads to blocky behavior, as pointed out by Irving et al. (1997). Recently, Ferreira (2002) has proposed a new class of multi-scale random fields also defined from coarser to finer levels, but assuming that conditional on its neighbors at the same resolution level and on its parent a node is independent of the remaining nodes at the same level. Full Bayesian analysis was developed with the help of Markov chain Monte Carlo techniques and the novel multi-scale framework was applied to the estimation of multi-scale permeability fields.

Other developments related to multi-scale models include: extension of the multi-scale models of Willsky by allowing arbitrary numbers of children and parents for each node (Huang and Cressie, 1997); multi-scale models for discrete valued latent processes considered on quadtrees (Laferté et al., 2000) and on pyramidal structures (Bouman and Shapiro, 1994; Kato et al., 1996a, 1996b), in the context of image segmentation. Johannesson and Cressie (2003) have considered the modeling and estimation of variances and covariances for multi-scale spatial models.

4.3. Connections between Gaussian Markov random fields and DLMs

Gaussian Markov random fields (GMRF) can be seen as generalizations of dynamic linear models to two or more dimensions. Even more interesting, Lavine (1999) pointed out that a Gaussian Markov random field prior on a lattice is the same as the posterior distribution of the state parameters of a particular dynamic linear model updated with special observations. Besides allowing fast computation of likelihoods, this idea leads to a fast exact simulation algorithm for Gaussian Markov random fields. For first-order neighborhood structures, this simulation algorithm is as efficient as the one proposed by Rue (2001) and is much more efficient than the Gibbs sampler simulation of GMRFs. For general discussion of Markov Chain Monte Carlo techniques and the Gibbs sampler in particular, see **?** (**?**). This chapter is concluded with presentation of Lavine's idea.

Let us consider a first-order isotropic Gaussian Markov random field on a finite two-dimensional lattice (Besag, 1974). Let θ_{ij} be the latent process at site (i, j), $i = 1, \ldots, I$, $j = 1, \ldots, J$. Let $\theta_i = (\theta_{i1}, \ldots, \theta_{iJ})'$ be the latent process at row i. Note that because of the Markovian property, θ_{i-1} and θ_{i+1} are conditionally independent given θ_i. This statement can be made more explicit by considering the system

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 32
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

equation

$$\theta_i = \theta_{i-1} + w_i, \quad w_i \sim N\big(0, \psi^{-1} I\big), \tag{4.8}$$

where ψ is the partial autocovariance of neighbor sites. Together with a flat prior for θ_1, this system equation takes care of the interactions between neighbors in the same column.

In order to take care of the interactions between neighbors in the same row, Lavine (1999) considers pseudo-observations $Y = \{Y_{i,j}\colon\ i = 1, \ldots, I; j = 1, \ldots, J-1\}$, $Y_i = (Y_{i1}, \ldots, Y_{iJ})'$ and the observation equation:

$$Y_i = F\theta_i + v_i, \quad v_i \sim N\big(0, \psi^{-1} I\big), \tag{4.9}$$

where

$$F = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix},$$

that is, $E(Y_{ij}|\theta) = \theta_{ij} - \theta_{i,j+1}$.

Taking the pseudo-observations to be equal to zero, then the posterior distribution of θ in the DLM defined by Eq. (4.8) and (4.9) is the same as the prior distribution of the considered GMRF.

Acknowledgements

The first, second and fourth authors wish to thank the financial support of research grants from CNPq. The authors thank Romy Rodriguez and Leonardo Bastos for the computations leading to Figure 1 and 2, respectively.

References

Aguilar, O., West, M. (2000). Bayesian dynamic factor models and variance matrix discounting for portfolio allocation. *J. Business Econom. Statist.* **18**, 338–357.

Ameen, J.R.M., Harisson, P. (1985). Normal discount Bayesian models. In: Bernardo, D.M.H.L.D.V., Smith, A.F.M. (Eds.), *Bayesian Statistics, vol. 2*. North-Holland and Valencia University Press.

Ameen, J.R.M., Harrison, P.J. (1983). Discount Bayesian models. In: Anderson, O. (Ed.), *Time Series Analysis: Theory and Practice, vol. 5*. North-Holland.

Banerjee, S., Gamerman, D., Gelfand, A.E. (2003). Spatial process modelling for univariate and multivariate dynamic spatial data. Technical Report, Department of Statistical Methods, Federal University of Rio de Janeiro.

Barbosa, E.P., Migon, H.S. (1996). Longitudinal data analysis of animal growth via multivariate dynamic models. *Comm. Statist. Simulation* **25**, 369–380.

Basseville, M., Benveniste, A., Willsky, A.S. (1992). Multiscale autoregressive processes, part i: Schur–Levinson parameterizations. *IEEE Trans. Signal Process.* **40**, 1915–1934.

Berzuini, C., Best, N., Gilks, W., Larizza, C. (1997). Dynamic conditional independence models and Markov chain Monte Carlo methods. *J. Amer. Statist. Assoc.* **92**, 1403–1412.

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 33
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc., Ser. B* **36**, 192–236.

Bolviken, E., Storvik, G. (2001). Deterministic and stochastic particle filters in state-space models. In: Doucet, A., de Freitas, N., Gordon, N. (Eds.), *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.

Bouman, C.A., Shapiro, M. (1994). A multiscale random field model for Bayesian image segmentation. *IEEE Trans. Image Process.* **3** (2), 162–177.

Box, G., Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*, second ed. Holden-Day, San Francisco.

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

Carlin, B.P., Polson, N.G., Stoffer, D.S. (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *J. Amer. Statist. Assoc.* **87**, 493–500.

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

Chou, K.C., Willsky, A.S., Benveniste, A. (1994). Multiscale recursive estimation, data fusion, and regularization. *IEEE Trans. Automat. Control* **39**, 464–478.

Cressie, N. (1994). Comment on "An approach to statistical spatial-temporal modeling of meteorological fields" by M.S. Handcock and J.R. Wallis. *J. Amer. Statist. Assoc.* **89** (426), 379–382.

Cressie, N., Huang, H.C. (1999). Classes of nonseparable spatio-temporal stationary covariance functions. *J. Amer. Statist. Assoc.* **94** (448), 1330–1340.

Crisan, D., Doucet, A. (2002). A survey of convergence results on particle filtering for practitioners. *IEEE Trans. Signal Process.* **50** (3), 736–746.

Diebolt, J., Robert, C. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc., Ser. B* **56**, 163–175.

Doucet, A. (2003). Sequential Monte Carlo methods for Bayesian analysis. *ISBA Bull.* **10** (1), 2–4.

Doucet, A., Tadić, V. (2003). Particle filters in state space models with the presence of unknown static parameters. *Ann. Inst. Statist. Math.* Submitted for publication.

Durbin, J., Koopman, S.J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press, London.

Durbin, J., Koopman, S. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika* **89**, 603–615.

Ferreira, M.A.R. (2002). Bayesian multi-scale modelling. Ph.D. thesis, Institute of Statistics and Decision Sciences, Duke University.

Ferreira, M.A.R., Gamerman, D. (2000). Dynamic generalized linear models. In: Ghosh, J.K., Dey, D., Mallick, B. (Eds.), *Generalized Linear Models: A Bayesian Perspective*. Marcel Dekker.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *J. Time Series Anal.* **15**, 183–202.

Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall, London.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalized linear models. *Biometrika* **85**, 215–227.

Gamerman, D. (2002). A latent approach to the statistical analysis of space-time data. In: *Proceedings of the XV International Workshop on Statistical Modeling*, pp. 1–15.

Gamerman, D., Migon, H.S. (1993). Dynamic hierarchical models. *J. Roy. Statist. Soc., Ser. B* **55**, 629–642.

Gamerman, D., Moreira, A.R.B. (2002). Bayesian analysis of econometric time series models using hybrid integration rules. *Comm. Statist.* **31**, 49–72.

Gamerman, D., Moreira, A.R.B., Rue, H. (2003). Space-varying regression models: Specifications and simulation. *Comput. Statist. Data Anal.* **42**, 513–533.

Gargallo, P., Salvador, M. (2002). Joint monitoring of several types of shocks in dynamic linear models: A Bayesian decision approach. Technical Report, Faculdad de Ciencias Económicas y Empresariales, Universidad de Zaragoza.

Geman, S., Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intelligence* **6**, 721–741.

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 34
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

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

Ghil, M., Cohn, S., Tavantzis, J., Bube, K., Isaacson, E. (1981). Applications of estimation theory to numerical weather prediction. In: Bengtsson, L., Ghil, M., Källén, E. (Eds.), *Dynamic Meteorology: Data Assimilation Methods*. Springer-Verlag, New York, pp. 139–224.

Goodall, C., Mardia, K. (1994). Challenges in multivariate spatio-temporal modeling. In: *Proceedings of the XVIIth International Biometric Conference*. Hamilton, Ontario, Canada, p. 17.

Gordon, N., Salmond, D., Smith, A. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F* **140**, 107–113.

Handcock, M., Wallis, J. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *J. Amer. Statist. Assoc.* **89** (426), 368–390.

Harrison, P.J., Lai, I.C. (1999). Statistical process control and model monitoring. *J. Appl. Statist.* **26**, 273–292.

Harrison, P.J., Stevens, C. (1976). Bayesian forecasting (with discussion). *J. Roy. Statist. Soc., Ser. B* **38**, 205–247.

Harvey, A. (1989). *Forecasting Structural Time Series Models and the Kalman Filter*. Cambridge University Press, London.

Harvey, A., Ruiz, E., Shephard, N. (1994). Multivariate stochastic variance models. *Rev. Econom. Studies* **61**, 247–264.

Haslett, J., Raftery, A. (1989). Space-time modeling with long-memory dependence: Assessing ireland's wind power resource (with discussion). *J. Roy. Statist. Soc., Ser. C* **38**, 1–50.

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

Hetherington, J. (1984). Observations on the statistical iteration of matrices. *Phys. Rev. A* **30**, 2713–2719.

Higuchi, T. (1997). Monte Carlo filter using the genetic algorithm operator. *J. Statist. Comput. Simulation* **59**, 1–23.

Higuchi, T. (2001). Self-organizing time series model. In: Doucet, A., de Freitas, N., Gordon, N. (Eds.), *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.

Huang, H.-C., Cressie, N. (1997). Multiscale spatial modeling. In: *ASA Proceedings of the Section on Statistics and the Environment*. Amer. Statist. Assoc., Alexandria, VA, pp. 49–54.

Huerta, G., Sansó, B., Stroud, J.R. (2004). A spatio-temporal model for Mexico city ozone levels. *Appl. Statist.* Submitted for publication.

Irving, W.W., Fieguth, P.W., Willsky, A.S. (1997). An overlapping tree approach to multiscale stochastic modeling and estimation. *IEEE Trans. Image Process.* **6**, 1517–1529.

Johannesson, G., Cressie, N. (2003). Variance–covariance modeling and estimation for multi-resolution spatial models. Technical Report, Department of Statistics, Ohio State University.

Johannesson, G., Cressie, N., Huang, H.-C. (2003) Dynamic multi-resolution spatial models. In: Higuchi, T., Iba, Y., Ishiguro, M. (Eds.), *Science of Modeling – The 30th Anniversary of the AIC*.

Jones, R. (1993). *Longitudinal Data with Serial Correlation – A State Space Approach*. Chapman-Hall, London.

Kato, Z., Berthod, M., Zerubia, J. (1996a). A hierarchical Markov random field model and multi-temperature annealing for parallel image classification. *Graph. Models Image Process.* **58** (1), 18–37.

Kato, Z., Zerubia, J., Berthod, M. (1996b). Unsupervised parallel image classification using Markovian models. *Graph. Models Image Process.* **58** (1), 18–37.

Kim, S., Shephard, N., Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Rev. Econom. Studies* **65**, 361–393.

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series (with discussion). *J. Amer. Statist. Assoc.* **82**, 1032–1063.

Kitagawa, G. (1993). A Monte Carlo filtering and smoothing method for non-Gaussian nonlinear state space models. In: *Proceedings of the 2nd US–Japan Joint Seminar on Statistical Time Series Analysis*. Honolulu, Hawaii, pp. 110–131.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graph. Statist.* **5**, 1–25.

Knorr-Held, L. (1999). Conditional prior proposal in dynamic models. *Scand. J. Statist.* **26**, 129–144.

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 35
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

Künsch, H. (2001). State space and hidden Markov models. In: Barndorff-Nielsen, D., Klüppelberg, C. (Eds.), *Complex Stochastic Systems*. Chapman and Hall, Boca Raton.

Laferté, J.-M., Pérez, P., Heitz, F. (2000). Discrete Markov image modeling and inference on the quadtree. *IEEE Trans. Image Process.* **9** (3), 390–404.

Landim, F., Gamerman, D. (2000). Dynamic hierarchical models – an extension to matricvariate observations. *Comput. Statist. Data Anal.* **35**, 11–42.

Lavine, M. (1999). Another look at conditionally Gaussian Markov random fields. In: Bernardo, J., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), *Bayesian Statistics, vol. 6*. Oxford Universtity Press, pp. 577–585.

Lindley, D.V., Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc., Ser. B*, 1–43.

Liu, J., West, M. (2001). Combined parameter and state estimation in simulation-based filtering. In: Doucet, A., de Freitas, N., Gordon, N. (Eds.), *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.

Lopes, H. (2000). Bayesian analysis in latent factor and longitudinal models. Ph.D. thesis, Institute of Statistics and Decision Sciences, Duke University.

Lopes, H. (2002). Sequential analysis of stochastic volatility models: Some econometric applications. Technical Report, Department of Statistical Methods, Federal University of Rio de Janeiro.

Lopes, H., Marinho, C. (2002). A particle filter algorithm for the Markov switching stochastic volatility model. Technical Report, Department of Statistical Methods, Federal University of Rio de Janeiro.

Lopes, H., Moreira, A., Schmidt, A. (1999). Hyperparameter estimation in forecast models. *Comput. Statist. Data Anal.* **29**, 387–410.

Mardia, K., Goodall, C., Redfern, E., Alonso, F. (1998). The Kriged Kalman filter (with discussion). *Test* **7**, 217–285.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953). Equation of state calculations by fast computing machine. *J. Chem. Phys.* **21**, 1087–1091.

Migon, H.S. (2000). The prediction of Brazilian exports using Bayesian forecasting. *Investigation Operativa* **9**, 95–106.

Migon, H.S., Gamerman, D. (1993). An application of nonlinear Bayesian forecasting to television advertising. *J. Forecasting* **12**, 573–584.

Migon, H.S., Harrison, P.J. (1985). An application of nonlinear Bayesian forecasting to TV advertising. In: Bernardo, J.M., DeGroot, M.H., Lindley, D., Smith, A.F.M. (Eds.), *Bayesian Statistics, vol. 2*. North-Holland and Valencia University Press.

Migon, H.S., Mazuchelli, J. (1999). Bayesian garch models: Approximated methods and applications. *Rev. Brasil. Econom.* **1999**, 111–138.

Monteiro, A., Migon, H.S. (1997). Dynamic models: An application to rainfall-runoff modelling. *Stochastic Hydrology and Hydraulics* **11**, 115–127.

Müller, P. (1991). Monte Carlo integration in general dynamic models. *Contemp. Math.* **115**, 145–163.

Müller, P. (1992). Posterior integration in dynamic models. *Comput. Sci. Statist.* **24**, 318–324.

Nelder, J.A., Wedderburn, R.W.M. (1972). Generalized linear models. *J. Roy. Statist. Soc., Ser. A* **135**, 370–384.

Pitt, M., Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *J. Amer. Statist. Assoc.* **94**, 590–599.

Pole, A., West, M. (1990). Efficient Bayesian learning in nonlinear dynamic models. *J. Forecasting* **9**, 119–136.

Pole, A., West, M., Harrison, P.J. (1994). *Applied Bayesian Forecasting and Time Series Analysis*. Chapman-Hall, London.

Polson, N., Stroud, J. Müller, P. (2002). Practical filtering with sequential parameter learning. Technical Report, Graduate School of Business, University of Chicago.

Reyna, F.R.Q., Migon, H.S., Duarte, A. (1999). Missing data in optimization models for financial planning. *Stochastic Hydrology and Hydraulics* **8**, 9–30.

Rubin, D. (1988). Using the SIR algorithm to simulate posterior distributions. In: Bernardo, J., DeGroot, M., Lindley, D., Smith, A. (Eds.), *Bayesian Statistics, vol. 3*.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *J. Roy. Statist. Soc., Ser. B* **65**, 325–338.

hs25 v.2005/04/21 Prn:20/05/2005; 12:07 F:hs25019.tex; VTEX/GIT p. 36
aid: 25019 pii: S0169-7161(05)25019-8 docsubty: REV

Schmidt, A.M., Gamerman, D., Moreira, A.R.B. (1999). An adaptive resampling scheme for cycle estimation. *J. Appl. Statist.* **26**, 619–641.

Shephard, N., Pitt, M.K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84**, 653–667.

Singh, A.C., Roberts, G.R. (1992). State space modelling of cross-classified time series of counts. *Internat. Statist. Rev.* **60**, 321–336.

Smith, A., Gelfand, A. (1993). Bayesian statistics without tears: A sampling–resampling perspective. *Amer. Statist.* **46**, 84–88.

Storvik, G. (2002). Particle filters in state space models with the presence of unknown static parameters. *IEEE Trans. Signal Process.* **50** (2), 281–289.

Stroud, J.R., Müller, P., Sansó, B. (2001). Dynamic models for spatio-temporal data. *J. Roy. Statist. Soc., Ser. B* **63**, 673–689.

Triantafyllopoulos, K., Harrison, P. (2002). Stochastic volatility forecasting with dynamic models. Research Report sta02-06, School of Mathematics and Statistics, University of Newcastle.

Vivar-Rojas, J.C., Ferreira, M.A.R. (2003). A new class of spatio-temporal models. Technical Report, Universidade Federal do Rio de Janeiro.

West, M. (1986). Bayesian model monitoring. *J. Roy. Statist. Soc., Ser. B* **48**, 70–78.

West, M. (1993a). Approximating posterior distributions by mixtures. *J. Roy. Statist. Soc., Ser. B* **55**, 409–422.

West, M. (1993b). Mixture models, Monte Carlo, Bayesian updating and dynamic models. In: Newton, J. (Ed.), *Computing Science and Statistics: Proceedings of the 24th Symposium of the Interface*. Interface Foundation of North America, Fairfax Station, VA, pp. 325–333.

West, M., Harrison, P.J. (1986). Monitoring and adaptation in Bayesian forecasting models. *J. Amer. Statist. Assoc.* **81**, 741–750.

West, M., Harrison, P.J. (1997). *Bayesian Forecasting and Dynamic Models*, second ed. Springer-Verlag, London.

West, M., Harrison, P.J., Migon, H. (1985). Dynamic generalized linear model and Bayesian forecasting. *J. Amer. Statist. Assoc.* **80**, 73–97.

Wikle, C., Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika* **86**, 815–829.

Wikle, C.K. (2003). Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology* **84**, 1382–1394.

Wikle, C.K. (2003). A kernel-based spectral approach for spatiotemporal dynamic models. Technical Report, Department of Statistics, University of Missouri.

Wikle, C.K., Milliff, R.F., Nychka, D., Berliner, L.M. (2001). Spatio-temporal hierarchical Bayesian modeling: Tropical ocean surface winds. *J. Amer. Statist. Assoc.* **96**, 382–397.