Dynamic Generalized Linear Models and Bayesian Forecasting

MIKE WEST, P. JEFF HARRISON, and HELIO S. MIGON*

Dynamic Bayesian models are developed for application in nonlinear, non-normal time series and regression problems, providing dynamic extensions of standard generalized linear models. A key feature of the analysis is the use of conjugate prior and posterior distributions for the exponential family parameters. This leads to the calculation of closed, standard-form predictive distributions for forecasting and model criticism. The structure of the models depends on the time evolution of underlying state variables, and the feedback of observational information to these variables is achieved using linear Bayesian prediction methods. Data analytic aspects of the models concerning scale parameters and outliers are discussed, and some applications are provided.

KEY WORDS: Nonlinear models; Exponential family; Dynamic regression; Linear Bayes methods.

1. INTRODUCTION

1.1 General

During the 1960s, the introduction of computers enabled static linear models and simple "derivable stationary" state-space models to be used profitably for the analysis and prediction of a wide variety of processes. It was soon evident, however, that the applications of such models often impose unrealistic assumptions. Many inappropriate "static" regressions were, and still are, uncritically applied to production processes or economic series in order to estimate what are really dynamic effects. In forecasting systems, a general assumption, arising from the academic desire for restrictive structures, is that a single simple model adequately represents a particular series at all possible times. Furthermore, the chosen models have often been restricted to the class of stationary or linearly derivable stationary models. There is still much confusion over local and global models in the work of some modelers. In particular, the argument that, locally, the form of a smooth function can be well represented by a low-order Taylor series model has been extended to the totally unjustified and often unconscious assumption that it can always be represented by such a model with constant coefficients. Many of the early applied forecasting approaches recognized the stationarity assumption to be inadequate and embedded the routine statistical model within a complete forecasting system operating according to the principle of *management by exception*. The exceptions are roughly of two kinds: (a) *feedback,* when a model monitoring system indicates a poor performance from the routine model, and (b) *feedforward,* when external subjective information is combined with routine statistical data. However, the methods for incorporating subjective information were ad hoc and rarely reflected the increased uncertainty associated with the occurrence of unusual events.

Towards the end of the decade, in an attempt to overcome some of these disadvantages, Harrison and Stevens (1976a) defined the class of dynamic linear models (DLM's) and developed the Bayesian approach to dynamic modeling and forecasting. The fundamental idea is that at any time t, the process under study is viewed in terms of meaningful parameters θ_t, and that given a history D_t, current knowledge with respect to the future is sufficiently described by a set of probability distributions: for the current state $(\theta_t \mid D_t)$; for future states $(\theta_{t+k+1} \mid \theta_{t+k})$, $k \geq 0$, defining the system model; and for observations $(Y_{t+k} \mid \theta_{t+k})$, $k \geq 0$. This approach is built on existing practice in the sense that many common models can be reformulated as special static and stationary cases. The real power, however, is that the reformulation provides many additional facilities such as are required for operating with little or no data, communicating subjective information, routinely handling "exceptions" using predesigned models, on-line monitoring and model discrimination, on-line estimation of parameters, modeling probabilistic (stochastic) transfer functions, and many more.

Major progress in practical modeling methodology was made in a different direction in the early 1970s based on the theoretical development of the generalized linear model (GLM) (Nelder and Wedderburn 1972). This work provided a unified framework for the study of a wide variety of data structures, and the ensuing development of the interactive GLIM computer program (Baker and Nelder 1978) led to an explosion in the application of formal parametric statistical models. The extension of regression techniques to non-normal sampling models with effects on a nonlinear scale provided what has become in the United Kingdom one of the most important modeling tools of applied statisticians and research workers. More recently, much research effort has been devoted to further theoretical development and refinement of the basic GLM to encompass more complex nonlinear relationships.

As useful as these models are, however, they suffer from some basic and, from a practical viewpoint, highly important drawbacks, some of which are as follows: (a) In common with standard linear regression techniques, the imposition of a static model implies fixed relationships and effects across observations. As noted earlier, this is viewed as an academic and restrictive assumption that leads to inflexibility in modeling and is often contradicted by reality. (b) The complexity of the likelihoods is such that exact inferences about observed rela-

tionships (estimation) and further observations (prediction) are precluded. The GLIM program approximations based on asymptotic theory are generally used, but the extent to which they are adequate in any particular application is usually unclear. (c) Practically important extensions involving sources of stochastic variation other than that in the sampling model (i.e., random effects) increase the complexity of the analysis. These types of models are similar to Bayesian GLM's with structured hierarchical priors that suffer the same problem of intractability (West 1984). (d) The GLIM program does not cater for the sequential processing of observations that is particularly desirable for time series applications. (e) A mainframe computer package, whose use requires relatively advanced mathematical and statistical expertise, is required for application of the GLM technology as it stands. This may have limited the application of these models, particularly in commercial and management areas.

The Bayesian approach to modeling presented here provides an alternative to the standard GLM that does not suffer those drawbacks. Exponential-family sampling models have been used in simple cases by several authors including Smith (1979), Souza (1981), and Azzalini (1983). The general models in Section 3 are related to these works and are, in part, dynamic Bayesian extensions of the GLM, though they go far beyond that particular class to complex nonlinear and even nonexponential family models. An illustration of a nonlinear application was given in Migon and Harrison (1985). Three key features of the models are (a) the sequential analysis and the use of conjugate prior distributions that lead to closed form updating and predictive distributions; (b) the important concept of using a guide relationship (Section 3.1) on which to base the choice of parameters of these prior distributions [this clearly separates the parameters of the sampling model from the underlying mathematical "system model" of the process under study and clarifies the inference problem (Section 3.2) considerably]; and (c) the computations that are no more demanding than those of the standard dynamic linear model. The sequential updating recursions for estimation and prediction are simple to program and may be performed on the smallest microcomputers.

1.2 Generalized Linear Models

A general, non-Bayesian approach to the analysis of static regression problems in the exponential family was given by Nelder and Wedderburn (1972) under the name of generalized linear models, and the application of the related methodology has spread rapidly due to the availability of comprehensive computer packages for such analyses (Baker and Nelder 1978). West (1985) discussed Bayesian inference and data analysis for generalized linear models.

The general exponential family sampling model for the observation Y_t is specified by the density

$$p(Y_t \mid \eta_t, \phi) = \exp[\phi\{Y_t\eta_t - a(\eta_t)\}]b(Y_t, \phi), \quad (1.1)$$

where η_t is the natural parameter of the distribution satisfying

$$E[Y_t \mid \eta_t, \phi] = \mu_t = \dot{a}(\eta_t) \quad (1.2)$$

and ϕ is a scale parameter with

$$V[Y_t \mid \eta_t, \phi] = \ddot{a}(\eta_t)/\phi. \quad (1.3)$$

In Sections 2 and 3 of this article, ϕ is assumed known and the dependence of the sampling density on ϕ is temporarily ignored with (1.1) written as $p(Y_t \mid \eta_t)$. Unknown scale parameters are considered in Section 4, and there the conditioning on ϕ is made explicit.

The static GLM relates η_t to a known $(n \times 1)$ vector of regressors $\mathbf{F}_t$ via

$$g(\eta_t) = \lambda_t = \mathbf{F}_t'\mathbf{\theta}, \quad (1.4)$$

where $\mathbf{\theta}$ is the unknown regression parameter vector and $g(\cdot)$ a specified nonlinear function. In practice, taking $g(\cdot)$ to be the identity function often provides a satisfactory model (e.g., log-linear and logistic-linear in the Poisson and binomial cases, respectively), although there may be several candidate functions to be examined, typically subject to λ_t being real valued. The GLM framework provides a class of models suitable for application in many practical problems, but due to the complexity of the resulting likelihoods, standard analyses provide only approximate inferences based on asymptotic results. In any extension of the GLM framework to encompass dynamic models in which, of course, standard asymptotic results do not apply, it will be necessary to seek some other way of providing precise inferences about regression parameters and future observations. To generalize to dynamic models, it is natural to replace $\mathbf{\theta}$ throughout with $\mathbf{\theta}_t$ as in the DLM; the model discussed in Section 2 has the essential features of such a formulation, yet it avoids the associated problems of intractability.

The model presented takes as a focus the problem of explicit sequential calculation of the forecast distributions for future observations in simple and standard form. By way of notation, define $D_t = \{I_t, D_{t-1}\}$, where for $t \geq 1$, $I_t = \{Y_t, F_t,$ and all other relevant information available at time t but not $t - 1\}$, and D_0 represents all relevant information available at $t = 0$. Then the closed form Bayesian analysis with the forecast distribution for $(Y_t \mid D_{t-1})$ available in standard form is obtained if the prior for the natural parameter has the conjugate form

$$p(\eta_t \mid D_{t-1}) = c(a_t, \beta_t) \cdot \exp[a_t\eta_t - \beta_t a(\eta_t)], \quad (1.5)$$

denoted $(\eta_t \mid D_{t-1}) \sim CP[a_t, \beta_t]$, for some a_t and β_t. For then

$$p(Y_t \mid D_{t-1}) = \frac{c(a_t, \beta_t)}{c(a_t + \phi Y_t, \beta_t + \phi)} b(Y_t, \phi) \quad (1.6)$$

may be obtained directly, and the posterior distribution for $(\eta_t \mid D_t)$ is the updated conjugate form $CP[a_t + \phi Y_t, \beta_t + \phi]$.

Given the restriction of $p(\eta_t \mid D_{t-1})$ to the form specified, the question remaining is that of specifying the structure of the model by relating η_t to the state vector $\mathbf{\theta}_t$. This is considered in Section 3. First, Section 2 provides a reformulation of the standard normal DLM on which the nonlinear models are based.

2. NORMAL DYNAMIC LINEAR MODELS REVISITED

2.1 The DLM and Discounting

In the case of normality, the function $g(\cdot)$ in (1.4) is the identity so that $\eta_t = \lambda_t$ for all t. In this section the sampling variance is assumed known and constant, equal to $1/\phi$. Other models in which ϕ depends on t fall within the general frame-

work but will not be considered here. The normal DLM is specified by setting $\boldsymbol{\theta}$ of (1.4) to $\boldsymbol{\theta}_t$ and defining the state evolution by

$$\boldsymbol{\theta}_t = G_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim N[\mathbf{O}_t, W_t], \qquad (2.1)$$

where G_t is a known $(n \times n)$ state transition matrix, W_t is a known $(n \times n)$ covariance matrix, and $\mathbf{w}_t$ is independent of $(\boldsymbol{\theta}_{t-1} \mid D_{t-1})$. Assuming that $(\boldsymbol{\theta}_0 \mid D_0) \sim N[m_0, C_0]$, and letting $I_t = \{Y_t, \mathbf{F}_t, G_t, W_t\}$, leads to the prior

$$(\boldsymbol{\theta}_t \mid D_{t-1}) \sim N[\mathbf{a}_t, R_t] \qquad (2.2)$$

and the posterior

$$(\boldsymbol{\theta}_t \mid D_t) \sim N[\mathbf{m}_t, C_t], \qquad (2.3)$$

where

$$\mathbf{a}_t = G_t\mathbf{m}_{t-1},$$

$$R_t = G_t C_{t-1} G_t' + W_t,$$

$$\mathbf{m}_t = \mathbf{a}_t + \mathbf{s}_t e_t \phi / (q_t \phi + 1),$$

$$C_t = R_t - \mathbf{s}_t \mathbf{s}_t' \phi / (q_t \phi + 1), \qquad (2.4)$$

with

$$\mathbf{s}_t = R_t \mathbf{F}_t,$$

$$q_t = \mathbf{F}_t' \mathbf{s}_t,$$

$$e_t = Y_t - f_t,$$

$$f_t = \mathbf{F}_t' \mathbf{a}_t. \qquad (2.5)$$

Due to the equivalence of those normal theory results and least squares estimation of an unknown mean when the variance is known, Equations (2.4) define the so-called Kalman-filter recursions (Kalman 1963). These form the basis of the automatic updating of the components of $\mathbf{m}_t$ and C_t, which are used in the calculation of forecast distributions. Concerning the evolution of the state vector in (2.1), the deterministic, linear transformation of $\boldsymbol{\theta}_{t-1}$ to $G_t\boldsymbol{\theta}_{t-1}$ is followed by the addition of the stochastic term $\mathbf{w}_t$, which increases the covariance matrix from $G_t C_{t-1} G_t'$ to R_t to model increasing uncertainty over time. For successful practical implementation the specification of suitable structure and magnitude of W_t is crucial and generally a difficult task. Ameen and Harrison (1985) have suggested a conceptually simple alternative approach using discount parameters in a generalization of simple exponential smoothing techniques. In that paper the matrix R_t was given by

$$R_t = B_t G_t C_{t-1} G_t' B_t, \qquad (2.6)$$

where B_t is an $(n \times n)$ diagonal matrix of positive discount factors $1/\beta_{it}$ $(i = 1, \ldots, n)$ with $\beta_{it} \leq 1$. Generally G_t will be block diagonal with each block representing a particular model component, and then the discount factors corresponding to a given block will usually take the same value. Clearly (2.6) provides for the required increased uncertainty over time using a natural multiplicative discounting of information. Ameen and Harrison also discuss other methods of discounting and the relationship with the original DLM; in this article the form (2.6) is used. Concerning notation, note that the discount factor β_{it} is the square of that used by Ameen and Harrison (1985).

2.2 Reformulation of the DLM

As a basis for the development of non-normal models in Section 3, a reformulation of the standard normal analysis just detailed is discussed here. In this model, $\eta_t = \mathbf{F}_t' \boldsymbol{\theta}_t$ so that, using (2.2) and (2.5),

$$(\eta_t \mid D_{t-1}) \sim N[f_t, q_t]. \qquad (2.7)$$

Now the likelihood for Y_t, given $\boldsymbol{\theta}_t$, depends on $\boldsymbol{\theta}_t$ only through η_t, so by Bayes's Theorem,

$$p(\boldsymbol{\theta}_t, \eta_t \mid D_t) \; \alpha \; p(Y_t \mid \eta_t)p(\boldsymbol{\theta}_t, \eta_t \mid D_{t-1})$$

$$\alpha \; p(Y_t \mid \eta_t)p(\eta_t \mid D_{t-1})p(\boldsymbol{\theta}_t \mid \eta_t, D_{t-1})$$

$$\alpha \; p(\eta_t \mid D_t)p(\boldsymbol{\theta}_t \mid \eta_t, D_{t-1}), \qquad (2.8)$$

where the first term here is the posterior for the scalar quantity η_t, given I_t, based on the prior (2.7); so it has the standard form $(\eta_t \mid D_t) \sim N[g_t, p_t]$, where

$$g_t = f_t + q_t \phi e_t / (q_t \phi + 1),$$

$$p_t = q_t - q_t^2 \phi / (q_t \phi + 1). \qquad (2.9)$$

The second term in (2.8) is calculated from the joint (singular) normal prior distribution of $\boldsymbol{\theta}_t$ and η_t, given by

$$\begin{bmatrix} \eta_t \\ \boldsymbol{\theta}_t \end{bmatrix} \Big| D_{t-1} \sim N\left[\begin{pmatrix} f_t \\ \mathbf{a}_t \end{pmatrix}, \begin{pmatrix} q_t & \mathbf{s}_t' \\ \mathbf{s}_t & R_t \end{pmatrix} \right]. \qquad (2.10)$$

Using standard normal theory, $(\boldsymbol{\theta}_t \mid \eta_t, D_{t-1}) \sim N[\hat{\mathbf{a}}_t, \hat{R}_t]$, where

$$\hat{\mathbf{a}} = \mathbf{a}_t + \mathbf{s}_t(\eta_t - f_t)/q_t,$$

$$\hat{R}_t = R_t - \mathbf{s}_t \mathbf{s}_t' / q_t. \qquad (2.11)$$

The Kalman-filter equations are obtained by noting that from (2.8),

$$E[\boldsymbol{\theta}_t \mid D_t] = E[E[\boldsymbol{\theta}_t \mid \eta_t, D_{t-1}] \mid D_t]$$

$$= E[\hat{\mathbf{a}}_t \mid D_t]$$

$$= \mathbf{a}_t + \mathbf{s}_t \cdot (g_t - f_t)/q_t, \qquad (2.12)$$

$$V[\boldsymbol{\theta}_t \mid D_t] = V[E[\boldsymbol{\theta}_t \mid \eta_t, D_{t-1}] \mid D_t]$$

$$\qquad + E[V[\boldsymbol{\theta}_t \mid \eta_t, D_{t-1}] \mid D_t]$$

$$= V[\hat{\mathbf{a}}_t \mid D_t] + E[\hat{R}_t \mid D_t]$$

$$= R_t - \mathbf{s}_t \mathbf{s}_t'(1 - p_t/q_t)/q_t. \qquad (2.13)$$

Substituting g_t and p_t into these equations gives the required values of $\mathbf{m}_t$ and C_t. So in this case, the updating of the distribution of the state vector is achieved using Bayes's Theorem for the scalar η_t and then filtering back the relevant information via the conditional distribution of $(\boldsymbol{\theta}_t \mid \eta_t, D_{t-1})$. Given the prior for $(\eta_t \mid D_{t-1})$, $\mathbf{m}_t$ depends on I_t only through the change in the mean of η_t; the updated covariance matrix C_t depends on I_t only through the change in the variance of η_t. Similarly, the calculation of the forecast distribution of $(Y_t \mid D_{t-1})$ depends on the prior for $\boldsymbol{\theta}_t$ only through the parameters of $p(\eta_t \mid D_{t-1})$. In this case, since all of the distributions involved are normal, the mean and covariance parameters determine the distribution. In the next section, non-normal analogs of this model are developed in which the prior and posterior distributions for the

state vector are only partially specified in order to obtain a tractable sequential analysis with predictive distributions available in standard form.

3. THE DYNAMIC GENERALIZED LINEAR MODEL

3.1 Basic Structure and Analysis

Suppose that the sampling model at time t has the form (1.1) and that the natural parameter η_t has the conjugate prior $CP[\alpha_t, \beta_t]$ of (1.5) for some α_t and β_t. A direct dynamic extension of the standard GLM would be provided by relating η_t via the nonlinear transformation $g(\eta_t)$ to

$$\lambda_t = \mathbf{F}_t' \boldsymbol{\theta}_t, \qquad (3.1)$$

where $\boldsymbol{\theta}_t$ is an underlying state vector having a time evolution similar to that of the DLM. To do this, however, would impose severe restrictions on the form of the prior for η_t. To avoid this it is recognized that the model involving $\boldsymbol{\theta}_t$ is a fiction, a modeler's "guide" to reality. Accordingly the link between $g(\eta_t)$ and λ_t is used simply as a guide to forming the prior for η_t; the notation $g(\eta_t) \simeq \lambda_t$ will represent this guide relationship.

The prior and posterior distributions of the state vector will not now be normal; the following models assume only that by analogy with the DLM, the first- and second-order moments of the state vector $\boldsymbol{\theta}_t$ are given by

$$(\boldsymbol{\theta}_{t-1} \mid D_{t-1}) \sim [\mathbf{m}_{t-1}, C_{t-1}] \qquad (3.2)$$

and that the time evolution for the model is defined by supposing that

$$(\boldsymbol{\theta}_t \mid D_{t-1}) \sim [\mathbf{a}_t, R_t], \qquad (3.3)$$

where

$$\mathbf{a}_t = G_t \mathbf{m}_{t-1} \quad \text{and} \quad R_t = B_t G_t C_{t-1} G_t' B_t \qquad (3.4)$$

with the transition matrix G_t and the discount matrix B_t known. Notice that the physical representation (2.1) may be used, although of course, $\mathbf{w}_t$ is not necessarily normal. Furthermore, the full distribution of the state vector is unspecified; only the mean and covariance matrix are assumed. Finally, from (3.1),

$$f_t = E[\lambda_t \mid D_{t-1}] = \mathbf{F}_t' \mathbf{a}_t,$$

$$q_t = V[\lambda_t \mid D_{t-1}] = \mathbf{F}_t' R_t \mathbf{F}_t,$$

and

$$\mathbf{s}_t = C[\lambda_t, \boldsymbol{\theta}_t \mid D_{t-1}] = R_t \mathbf{F}_t.$$

At this point the prior for the natural parameter η_t is only partially specified, having the conjugate form

$$(\eta_t \mid D_{t-1}) \sim CP[\alpha_t, \beta_t] \qquad (3.5)$$

with, as yet, no restriction on the values of α_t and β_t. These values are chosen on the basis of the guide relationship $g(\eta_t) \simeq \lambda_t$; fixing the first two moments of $g(\eta_t)$ will determine α_t and β_t. The guide relation suggests the values of f_t and q_t for these moments [and $\mathbf{s}_t$ for the covariance between $g(\eta_t)$ and $\boldsymbol{\theta}_t$], and these are used in this article. It should be stressed that this is not necessary; other values may be chosen to incorporate further subjective information, to allow intervention or to avoid numerical difficulties as noted in the applications in Section 5.

On the basis of this specification, it is immediate that the forecast distribution for $(Y_t \mid D_{t-1})$ has the standard form (1.6) and the posterior for $(\eta_t \mid D_t)$ has the conjugate form $CP[\alpha_t + \phi Y_t, \beta_t + \phi]$. A full Bayesian analysis requires also the posterior for $(\boldsymbol{\theta}_t \mid D_t)$, but this is not available because the prior for $(\boldsymbol{\theta}_t \mid D_{t-1})$ is only partially specified and the model does not provide a likelihood for $\boldsymbol{\theta}_t$. The model as developed so far does not, however, require the complete specification to proceed to time $t + 1$; only the mean and covariance matrix of $(\boldsymbol{\theta}_t \mid D_t)$ are needed and these satisfy the identities

$$\mathbf{m}_t = E[E[\boldsymbol{\theta}_t \mid \eta_t, D_t]] \qquad (3.6)$$

and

$$C_t = V[E[\boldsymbol{\theta}_t \mid \eta_t, D_t]] + E[V[\boldsymbol{\theta}_t \mid \eta_t, D_t]]. \qquad (3.7)$$

Furthermore, as in the normal case, $(\boldsymbol{\theta}_t \mid \eta_t, D_t)$ is conditionally independent of I_t from the factorization in (2.8) so that the conditional moments in the inner expectations in (3.6) and (3.7) are just those of $(\boldsymbol{\theta}_t \mid \eta_t, D_{t-1})$. In general these moments will be unknown, nonlinear functions of η_t; the only information available is that concerning the joint moments of $(g(\eta_t), \boldsymbol{\theta}_t' \mid D_{t-1})$,

$$\begin{bmatrix} g(\eta_t) \\ \boldsymbol{\theta}_t \end{bmatrix} \middle| D_{t-1} \sim \left[\begin{pmatrix} f_t \\ \mathbf{a}_t \end{pmatrix}, \begin{pmatrix} q_t & \mathbf{s}_t' \\ \mathbf{s}_t & R_t \end{pmatrix} \right], \qquad (3.8)$$

where the full covariance matrix here is singular. On the basis of this information alone, an alternative approach is required so that the information in I_t may be filtered back to $\boldsymbol{\theta}_t$. This is discussed in Section 3.2.

3.2 State Vector Update

Hartigan (1969) developed the linear Bayesian approach to problems of inference in essentially linear models when only the first two moments of the prior and likelihood are specified. Goldstein (1976) used a similar approach in partially specified nonlinear regression problems, suggesting a quasi-Bayesian method of providing optimal linear fits to unknown nonlinear functions. The linear Bayesian method may be applied in the preceding model to provide feedback of the information in I_t to $\boldsymbol{\theta}_t$.

The density $p(\boldsymbol{\theta}_t \mid \eta_t, D_{t-1})$ is that of the unknown predictive distribution of $\boldsymbol{\theta}_t$, given η_t; the mean is the optimal predictor in the sense of minimizing the quadratic risk function trace $[A_t(\mathbf{d})]$ with respect to $\mathbf{d}$, where

$$A_t(\mathbf{d}) = E[(\boldsymbol{\theta}_t - \mathbf{d})(\boldsymbol{\theta}_t - \mathbf{d})' \mid \eta_t, D_{t-1}]. \qquad (3.9)$$

The covariance matrix of the distribution is the value of $A_t(\mathbf{d})$ at the mean. Now, since the mean is unknown, an alternative predictor is sought using the linear Bayesian approach; in view of the relation (3.1) and the construction of η_t, it is natural to adopt a linear function of $g(\eta_t)$ as a predictor of $\boldsymbol{\theta}_t$. Specifically, suppose that $\mathbf{d}$ is to be chosen such that $\mathbf{d} = \mathbf{d}_0 + \mathbf{d}_1 g(\eta_t)$ for some $\mathbf{d}_0$ and $\mathbf{d}_1$ and that, in place of (3.9), $\mathbf{d}$ minimizes the overall quadratic risk (or sum of variances) given by

$$r_t(\mathbf{d}) = \text{trace } E[A_t(\mathbf{d}) \mid D_{t-1}],$$

where the expectation is with respect to $p(\eta_t \mid D_{t-1})$.

In this model the joint moments in (3.8) suffice to determine the required predictor. By directly minimizing $r_t(\mathbf{d})$ with re-

spect to $\mathbf{d}_0$ and $\mathbf{d}_1$, it is easily shown that a unique minimum is obtained at $\mathbf{d} = \hat{\mathbf{a}}_t$, where

$$\hat{\mathbf{a}}_t = \mathbf{a}_t + \mathbf{s}_t(g(\eta_t) - f_t)/q_t; \qquad (3.10)$$

the value of $E[A_t(\mathbf{d}) \mid D_{t-1}]$ at the minimum is given by

$$\hat{R}_t = R_t - \mathbf{s}_t \mathbf{s}_t'/q_t. \qquad (3.11)$$

The values $\hat{\mathbf{a}}_t$ and R_t provide an optimal linear predictor of $(\boldsymbol{\theta}_t \mid \eta_t, D_{t-1})$ and the associated measure of risk is an essentially nonlinear problem. The feedback of the information in I_t may now be completed by replacing the conditional mean and covariance matrix in (3.6) and (3.7) by $\hat{\mathbf{a}}_t$ and $\hat{R}_t$ to obtain the expected predictor and related risk given by

$$\mathbf{m}_t = \mathbf{a}_t + \mathbf{s}_t(g_t - f_t)/q_t$$

and

$$C_t = R_t - \mathbf{s}_t \mathbf{s}_t'(1 - p_t/q_t)/q_t,$$

where $g_t = E[g(\eta_t) \mid D_t]$ and $p_t = V[g(\eta_t) \mid D_t]$ may be calculated from the fully conjugate posterior for $(\eta_t \mid D_t)$.

The updating equations for the dynamic generalized linear model (DGLM) are now complete and may be used as the basis for the analysis of further observations; note that they have exactly the same form as the standard Kalman-filter recursions with the relevant values of g_t and p_t substituted.

3.3 Model Summary and Forecasting

For reference purposes the full system of recursions is summarized here.

1. Observation model

$$p(Y_t \mid \eta_t) \alpha \exp[\phi\{Y_t \eta_t - a(\eta_t)\}].$$

2. Prior

$$(\eta_t \mid D_{t-1}) \sim CP[\alpha_t, \beta_t],$$

$$(\boldsymbol{\theta}_t \mid D_{t-1}) \sim [\mathbf{a}_t, R_t].$$

α_t and β_t are chosen in accordance with the guide relationship $g(\eta_t) \simeq \mathbf{F}_t'\boldsymbol{\theta}_t$ so that $E[g(\eta_t) \mid D_{t-1}] \simeq f_t$ and $V[g(\eta_t) \mid D_{t-1}] \simeq q_t$, where $f_t = \mathbf{F}_t'\mathbf{a}_t$, $\mathbf{s}_t = R_t\mathbf{F}_t$, and $q_t = \mathbf{F}_t'\mathbf{s}_t$.

3. Posterior

$$(\eta_t \mid D_t) \sim CP[\alpha_t + \phi Y_t, \beta_t + \phi],$$

with $E[g(\eta_t) \mid D_t] = g_t$ and $V[g(\eta_t) \mid D_t] = p_t$.

4. State update

$$(\boldsymbol{\theta}_t \mid D_t) \sim [\mathbf{m}_t, C_t],$$

with

$$\mathbf{m}_t = \mathbf{a}_t + \mathbf{s}_t(g_t - f_t)/q_t \text{ and } C_t = R_t - \mathbf{s}_t\mathbf{s}_t'(1 - p_t/q_t)/q_t.$$

5. State transition

$$\mathbf{a}_t = G_t\mathbf{m}_{t-1}, \qquad R_t = B_tG_tC_{t-1}G_t'B_t.$$

6. For forecasting purposes, predictive distributions for η_{t+k} and Y_{t+k} ($k > 0$) available in standard forms.

Define

$$D_t(k) = D_{t+k} - \{Y_{t+1}, \ldots, Y_{t+k}\}, \qquad k > 0.$$

Then extending (3.3) and (3.4) to time $t + k$,

$$(\boldsymbol{\theta}_{t+k} \mid D_{t-1}(k)) \sim [\mathbf{a}_t(k), R_t(k)],$$

where for $j \geq 1$,

$$\mathbf{a}_t(j) = G_{t+j}\mathbf{a}_t(j - 1)$$

$$R_t(j) = B_{t,j}G_{t+j}R_t(j - 1)G_{t+j}'B_{t,j},$$

with $\mathbf{a}_t(0) = \mathbf{a}_t$ and $R_t(0) = R_t$. Here $B_{t,j}$ may be taken as B_{t+j} or, to coincide with the representation (2.1), an alternative that ensures that

$$R_t(j) = G_{t+j}R_t(j - 1)G_{t+j}' + W_{t+j}$$

for some W_{t+j}. See Ameen and Harrison (1984, section 3) for further details.

Hence

$$(\eta_{t+k} \mid D_{t-1}(k)) \sim CP[\alpha_t(k), \beta_t(k)],$$

where $\alpha_t(k), \beta_t(k)$ satisfy

$$E[g(\eta_{t+k}) \mid D_{t-1}(k)] = f_{t+k} = \mathbf{F}_{t+k}'\mathbf{a}_t(k)$$

and

$$V[g(\eta_{t+k}) \mid D_{t-1}(k)] = q_t(k) = \mathbf{F}_{t+k}'R_t(k)\mathbf{F}_{t+k}.$$

Finally, $p(Y_{t+k} \mid D_{t-1}(k))$ has the standard form (1.6) with parameters $\alpha_t(k)$ and $\beta_t(k)$.

4. SCALE PARAMETERS AND OUTLIERS

4.1 Normal Models With Unknown Variance

In the normal DLM the observational variance $1/\phi$ is assumed known, and successful implementation of the resulting analysis depends crucially on the choice of ϕ. A simple modification of the model, however, provides a tractable learning procedure for unknown ϕ that is a minor generalization of standard conjugate Bayesian methods (Smith and West 1983).

Again it should be clear that the following analysis extends to more general models with, for example, dynamic variances changing deterministically or stochastically in time.

Suppose that the prior of (2.2) is scaled by ϕ with R_t replaced by R_t/ϕ. This results in both $\mathbf{s}_t$ and q_t being divided by ϕ and hence

$$(\boldsymbol{\theta}_t \mid \phi, D_t) \sim N[\mathbf{m}_t, C_t/\phi],$$

where

$$\mathbf{m}_t = \mathbf{a}_t + \mathbf{s}_t e_t/(q_t + 1) \quad \text{and} \quad C_t = R_t - \mathbf{s}_t\mathbf{s}_t'/(q_t + 1),$$

with $\mathbf{s}_t$, q_t, and e_t as originally defined in (2.5). Thus the standard Kalman-filter recursions apply with the factor ϕ set to unity. The unknown ϕ only appears as a scale parameter multiplying the covariance terms. To learn about ϕ the conjugate prior is a gamma form,

$$(\phi \mid D_{t-1}) \sim G[\nu_{t-1}/2, \delta_{t-1}/2],$$

and standard Bayesian manipulations lead to an updated $G[\nu_t/2, \delta_t/2]$ posterior, where

$$\nu_t = \nu_{t-1} + 1, \qquad \delta_t = \delta_{t-1} + e_t^2/(q_t + 1).$$

Point estimates of ϕ or $1/\phi$ can then be calculated sequentially; for example, $\sigma_t^2 = \delta_t/\nu_t$ is the posterior harmonic mean of the

observational variance. Standard Bayesian results also provide the marginal posterior for $(\theta_t \mid D_t)$ in the form of a multivariate Student distribution. Similarly, the marginal predictive distributions have Student forms; for example,

$$p(Y_{t+1} \mid D_t) \ a \ [\delta_t + e_{t+1}^2/(q_{t+1} + 1)]^{-v_{t+1}/2}.$$

Of course, for v_{t+1} relatively large, these distributions are approximately normal with, for example,

$$(Y_{t+1} \mid D_t) \sim N[f_{t+1}, (q_{t+1} + 1)\sigma_t^2].$$

4.2 Scale Parameters in Non-Normal Models

West (1985) discussed general ideas concerning scale parameters and variance functions in the exponential family of distributions. The continuous sampling models have natural scale parameters, but the discrete forms do not, for then ϕ in (1.1) is fixed usually at unity. Standard inference procedures for GLM, however, provide an estimate of scale that is related to the deviance measure of goodness of fit (Nelder and Wedderburn 1972). A particular exponential family distribution may be embedded in a more general model with an approximate scaled likelihood incorporating a general scale parameter ϕ. This scale parameter can be viewed as a goodness-of-fit measure with $\phi = 1$ corresponding to the original distribution; the general model provides a likelihood that forms the basis for learning about ϕ. Full details were given in West (1985) and will only be summarized here.

Note that the density (1.1) has the form of a power transformation of the standard model for which $\phi = 1$,

$$p(Y_t \mid \eta_t, \phi) \ a \ p(Y_t \mid \eta_t, \phi = 1)^\phi.$$

So the standard models (e.g., binomial, Poisson, and gamma) having variance $\ddot{a}(\eta_t)$ are embedded in the general form with variance $\ddot{a}(\eta_t)/\phi$.

As in the normal case, a conjugate analysis including the unknown scale parameter essentially follows that of Section 3 as far as η_t is concerned.

In the general model defined by (1.1) and (1.5), let X_t be the maximum likelihood value for η_t based on Y_t so that $Y_t = \dot{a}(X_t)$, h_t the prior mode satisfying $a_t = \beta_t \dot{a}(h_t)$, and Z_t the posterior mode satisfying $(a_t + Y_t) = (\beta_t + 1)\dot{a}(Z_t)$.

Then the deviance functions required by the analysis are defined by

1. observation

$$dv(Y_t \mid \eta_t) = -2 \ln[p(Y_t \mid \eta_t, \phi = 1)/p(Y_t \mid X_t, \phi = 1)],$$

2. prior

$$dv(\eta_t \mid D_{t-1}) = -2 \ln[p(\eta_t \mid D_{t-1})/p(h_t \mid D_{t-1})],$$

3. posterior

$$dv(\eta_t \mid D_t) = -2 \ln[p(\eta_t \mid D_t)/p(Z_t \mid D_t)],$$

4. residual

$$dv(Y_t \mid D_{t-1}) = [dv(Y_t \mid \eta_t) + dv(\eta_t \mid D_{t-1})]_{\eta_t=Z_t}.$$

These functions satisfy a generalization of the quadratic form decomposition of normal models that may be deduced directly

from Bayes's Theorem,

$$dv(Y_t \mid \eta_t) + dv(\eta_t \mid D_{t-1}) = dv(Y_t \mid D_{t-1}) + dv(\eta_t \mid D_t).$$

The residual deviance is closely related to the standard deviance measure of residual variation used in the static GLM (Baker and Nelder 1978), and in the normal model it reduces to the squared standardized residual $e_t^2/(q_t + 1)$. The analysis in West (1985) uses a gamma prior for $(\phi \mid D_{t-1})$ in the nonnormal models and leads to an approximate gamma posterior

$$(\phi \mid D_t) \sim G[v_t/2, \delta_t/2],$$

where $v_t = v_{t-1} + 1$ and $\delta_t = \delta_{t-1} + dv(Y_t \mid D_{t-1})$. Then, for example, $\dot{\sigma}_t^2 = \delta_t/v_t$ may be calculated recursively as a point estimate of $1/\phi$. A useful extension to dynamic scale parameters may be achieved simply by discounting v_{t-1} and δ_{t-1} before updating.

In practice the analysis of Section 3 should be performed with ϕ set at unity and the quantities v_t and δ_t calculated after each update. The posterior for ϕ provides a method of monitoring the fit of the standard model; if the model fits the data, then values near $\phi = 1$ will be favored, as simple summaries such as σ_t^2 will indicate. If $\phi < 1$ is indicated, the inference is that the model does not adequately explain observed variation and further elaboration may be necessary. Conversely, if values of $\phi > 1$ are indicated, then there is apparently less variation in the data than is provided for by the standard model and a simpler model may be preferred. Illustration of this data-analytic aid is given in following examples.

4.3 Outlier Modeling and Intervention

The multiprocess extension of the normal DLM (Harrison and Stevens 1976a,b) provides a model capable of automatically accommodating outliers and adapting to structural changes in the observed series. Multiprocess generalizations of the DGLM, and various other monitoring schemes based, for example, on predictive distributions, have been developed and applied, but here a simple intervention approach is mentioned.

West (1981) discussed the modeling of outliers in the DLM using heavy-tailed error distributions as alternatives to normality. A particularly important class of such distributions is that of scale mixtures of normals (West 1984) having densities $p(Y_t \mid \eta_t) = f(Y_t - \eta_t)$ obtained from the conditional normal distribution $(Y_t \mid \eta_t, \phi_t) \sim N[\eta_t, 1/\phi_t]$, with the scale parameter ϕ_t having some prior distribution. The influence function of the density $f(u)$ is given by $-d/du [\ln f(u)] = u\psi(u)$, where $\psi(\cdot)$ is a weight function that acts to limit the influence of outliers. For example, if $f(\cdot)$ is a Student $t - k$ density, $\psi(u) = (k + 1)/(k + u^2)$. An extension of the results in West (1981) can be used to show that the posterior mean of η_t behaves like

$$g_t = f_t + q_t\psi_t e_t/(q_t\psi_t + 1), \qquad (4.1)$$

where $\psi_t = \psi(e_t)$ estimates ϕ_t. Now this is precisely the posterior mean based on a normal observation with variance $1/\psi_t$; the corresponding posterior variance $p_t = q_t/(q_t\psi_t + 1)$ always underestimates the true variance, which often exceeds q_t to reflect the increased uncertainty about η_t due to the large residual. The approximation (4.1) leads to an approximate vari-

ance $p_t^* = p_t(1 + r_t)$, where $r_t = p_t\psi(e_t)e_t$ is always non-negative.

From a practical viewpoint, if the time series exhibits no major structural changes, then simply discounting outliers using (4.1)—that is, treating Y_t as if it was normal with variance $1/\psi_t$—is generally satisfactory. If abrupt changes are to be modeled, however, a large residual $|e_t|$ may represent either a change or an outlier and, based on the data alone, the two events cannot be distinguished at time t. Both the multiprocess approach and intervention ideas of Ameen and Harrison (1985) allow for change by decreasing the discount factors in B_t of (2.6). This increases the uncertainty about the state parameters and leads to greater adaptation to the next observation. A simpler and useful approach is to intervene by considerably increasing p_t on the occurrence of a large residual after processing Y_t as if it was normal with variance $1/\psi_t$. This increased uncertainty is fed back to C_t in (2.13) and hence through to $(\eta_{t+1} \mid D_t)$. This provides a greater degree of adaptation to the next observation (the same procedure being applied at $t + 1$, of course) and automatic discrimination between the outliers and changes.

The generalization of this outlier modeling approach to non-normal models follows West (1985). As with scale mixtures of normal models, the density Y_t is given in (1.1), where ϕ is now replaced by ϕ_t, a scale parameter with a suitable prior distribution. The analysis then proceeds with ϕ_t estimated by the weight function ψ_t, based not on a squared standardized residual, but on the general observational deviance evaluated at the prior mode h_t. The analog of the Student $t - k$ model, for example, has

$$\psi_t = (k + 1)/[k + dv(Y_t \mid h_t)].$$

The resulting posterior for $(\eta_t \mid D_t)$ is then $CP[a_t + \psi_tY_t, \beta_t + \psi_t]$. If Y_t is extreme, the deviance is large, the weight ψ_t is small, and the outlier is discounted.

Finally, incorporating a common scale parameter ϕ as in Section 4.3 simply involves dividing the deviance term by an estimate of $1/\phi$ such as σ_t^2 of Section 4.1, where now

$$dv(Y_t \mid D_{t-1}) = [\psi_t dv(Y_y \mid \eta_t) + dv(\eta_t \mid D_{t-1})]_{\eta_t=z_t}$$

and z_t satisfies $(a_t + \psi_tY_t) = (\beta_t + \psi_t)\dot{a}_t(Z_t)$.

5. SPECIAL CASES AND APPLICATIONS

Two important special cases are the discrete binomial and Poisson models. In application, suitable models are often obtained taking $g(\cdot)$ to be the identity so that the linear evolution is appropriate for the natural parameter η_t. This leads to analogs of logistic-linear and log-linear models in the binomial and Poisson cases, respectively, and these are used in the following examples.

In the Poisson case, the prior $CP[a_t, \beta_t]$ is a log-gamma form for $\eta_t = \ln(\mu_t)$, where μ_t is the Poisson mean or rate. The mean of $(\eta_t \mid D_{t-1})$ is $\gamma(a_t) - \ln(\beta_t)$, the mode is $\ln(a_t/\beta_t)$, and the variance is $\dot{\gamma}(a_t)$, where $\gamma(a) = \dot{\Gamma}(a)/\Gamma(a)$ is the digamma function. At each time t, the values f_t and q_t are used to determine a_t and β_t, as explained in Section 3. Equating f_t to the mean and q_t to the variance of $(\eta_t \mid D_{t-1})$ leads to two equations that may be solved for a_t and β_t by numerically

inverting $\dot{\gamma}$ and γ. As noted in Section 3.1, however, f_t and q_t are just guides to the moments of η_t, and other simpler and more convenient values may be used. In Examples 1–4, f_t is taken as the mode of $(\eta_t \mid D_{t-1})$ and q_t^{-1} as the curvature at the mode, $q_t = a_t^{-1}$. This in effect replaces $\gamma(a_t)$ by $\ln(a_t)$ and leads to $a_t = q_t^{-1}$ and $\beta_t = q_t^{-1}\exp(-f_t)$. Then the posterior moments for $(\eta_t \mid D_t)$ are given by

$$g_t = \gamma(a_t + Y_t) - \ln(\beta_t + 1), \qquad p_t = \dot{\gamma}(a_t + Y_t);$$

these are easily calculated using the simple recursions $\gamma(a) = \gamma(a + 1) - a^{-1}$ and $\dot{\gamma}(a) = \dot{\gamma}(a + 1) + a^{-2}$.

Similar comments apply to the binomial models. With a logistic-linear structure, $\eta_t = \ln[\mu_t/(1 - \mu_t)]$, where μ_t is the binomial probability parameter and μ_t has a beta prior with mean $a_t/(a_t + \beta_t)$. Using the mode and curvature of $(\eta_t \mid D_{t-1})$ for f_t and q_t^{-1} leads to $a_t = q_t^{-1}[1 + \exp(f_t)]$ and $\beta_t = q_t^{-1}[1 + \exp(-f_t)]$. Then the posterior mean and variance are found, using the preceding recursions for γ and $\dot{\gamma}$, from

$$g_t = \gamma(a_t + Y_t) - \gamma(\beta_t + n_t - Y_t),$$

$$q_t = \dot{\gamma}(a_t + Y_t) + \dot{\gamma}(\beta_t + n_t - Y_t),$$

where n_t is the binomial sample size.

In each of the following examples, a relatively uninformative prior is used for θ_0, with C_0 a diagonal matrix having suitably large entries.

Example 1. The 77 observations plotted as asterisks in Figure 1(a) are the monthly sales data of an engineering company discussed by Chatfield and Prothero (1973) with January 1965 as $t = 1$ and May 1971 as $t = 77$. The linear models used for the log observation in that paper perform poorly, considerably overforecasting the next year, $t = 78–89$. Two discussants of the paper considered alternative, and more reasonable, models; one used a less powerful decreasing transformation of the data (G. Tunnicliffe-Wilson 1973), and the other used the original data with a nonconstant variance (P. J. Harrison 1973), although both models still overforecasted the year ahead.

The alternative, and possibly more realistic, approach considered here uses a Poisson log-linear model with a linear growth term and the first five harmonics of the Fourier seasonal description of period 12; thus

$$\mathbf{F}_t = (1, 0; 1, 0, 1, 0, 1, 0, 1, 0, 1, 0),$$

$$G_t = \text{diag}(H_0, H_1, \ldots, H_5),$$

with

$$H_0 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad H_k = \begin{bmatrix} \cos(kp) & \sin(kp) \\ -\sin(kp) & \cos(kp) \end{bmatrix},$$

$$k = 1, \ldots, 5, \qquad p = \pi/6.$$

Thus $E[Y_t \mid \eta_t] = V[Y_t \mid \eta_t] = \exp(\eta_t)$ behaves like $\exp(\lambda_t) = r_ts_t$, where r_t is a log-linear trend term and s_t is a log-linear seasonal term. Finally, a constant rate of discounting is given by taking $B_t = \text{diag}(\beta_0^{-1}I_2, \beta_1^{-1}I_{10})$, where I_k is the $(k \times k)$ identity matrix, $\beta_0^2 = .85$, and $\beta_1^2 = .99$. Here β_0^{-1} is the discount factor corresponding to the trend component and β_1^{-1} to the seasonal component.

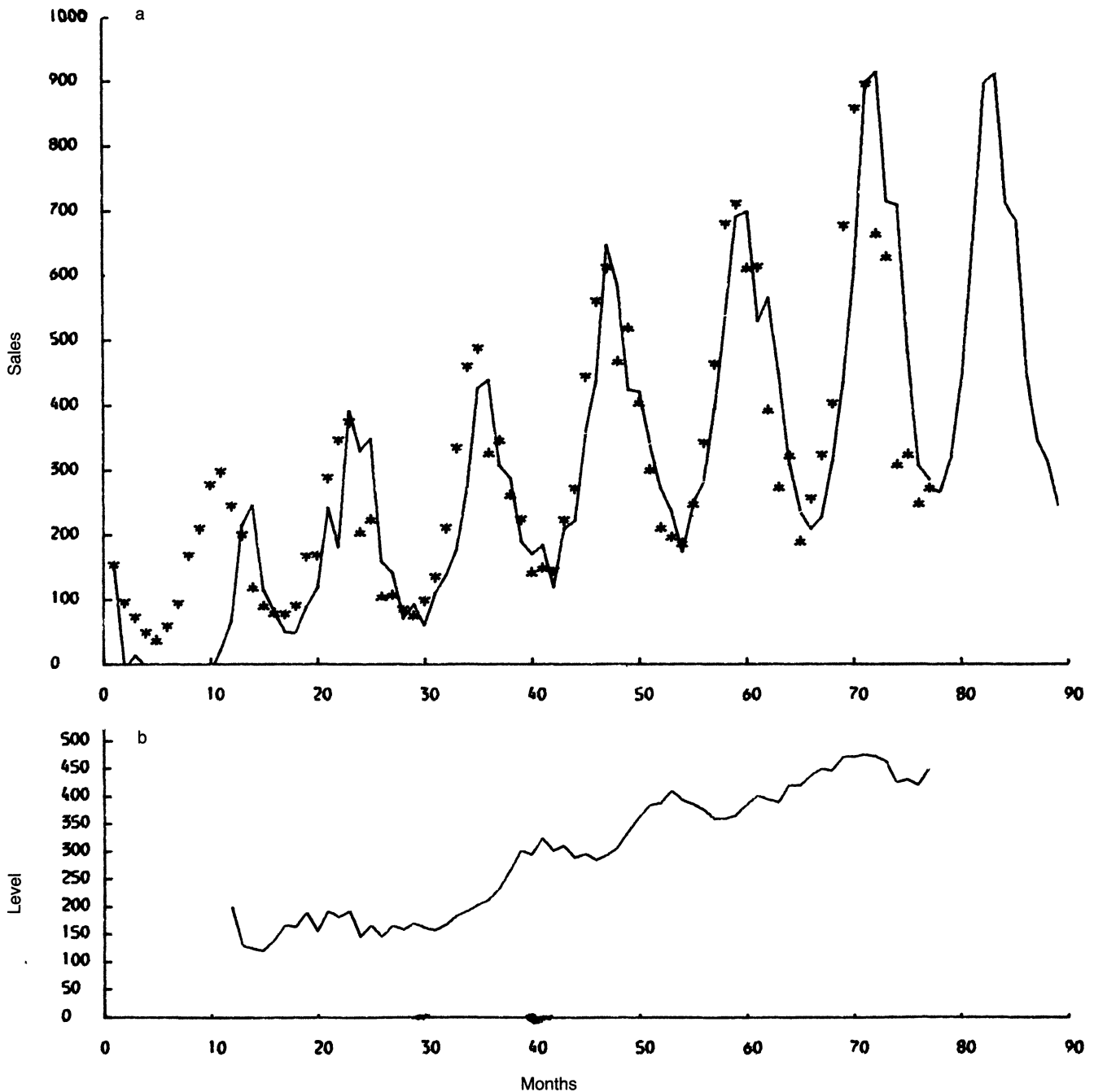The full line in Figure 1(a) provides the one-step forecasts

Figure 1. Example 1: (a) Sales Data and Forecasts; (b) On-Line Estimated Trend.

up to $t = 77$ and then forecasts for the next 12 months from $t = 77$ using this model. The forecast for Y_{t+k}, given data up to time t, is taken to be the mode $\hat{Y}_t(k)$ of the negative binomial predictive distribution $p(Y_{t+k} \mid D_t(k))$. The effect of the weak prior for θ_t is evident in the poor predictions for $t < 14$. The year-ahead forecasts in Figure 1(a) are lower than those of the models mentioned earlier. For $t = 78-84$, the actual observations and forecasts from $t = 77$ are given by

Time t	78	79	80	81	82	83	84
Observation	260	304	390	614	783	872	540
Forecast	266	321	448	655	898	912	712

The forecasts are still overoptimistic. They are, of course,

much improved as further observations are processed and the model adapts further to the change in trend due to the onset of recession in late 1970 and early 1971 ($t > 71$). Figure 1(b) shows the on-line estimated trend. In practice the change in trend signaled around $t = 73, 74$ should lead to immediate intervention to considerably decrease β_0, thus increasing the uncertainty in the model about the trend component and leading to rapid adaptation to the observed change. Finally, note that although r_t is a log-linear trend term, the estimate in Figure 1(b) appears to be piece-wise linear.

This example uses a realistic Poisson sampling model although, due to the range of observations, a normal dynamic model for some transformation of the observations could have been successfully applied. The Poisson model, however, re-
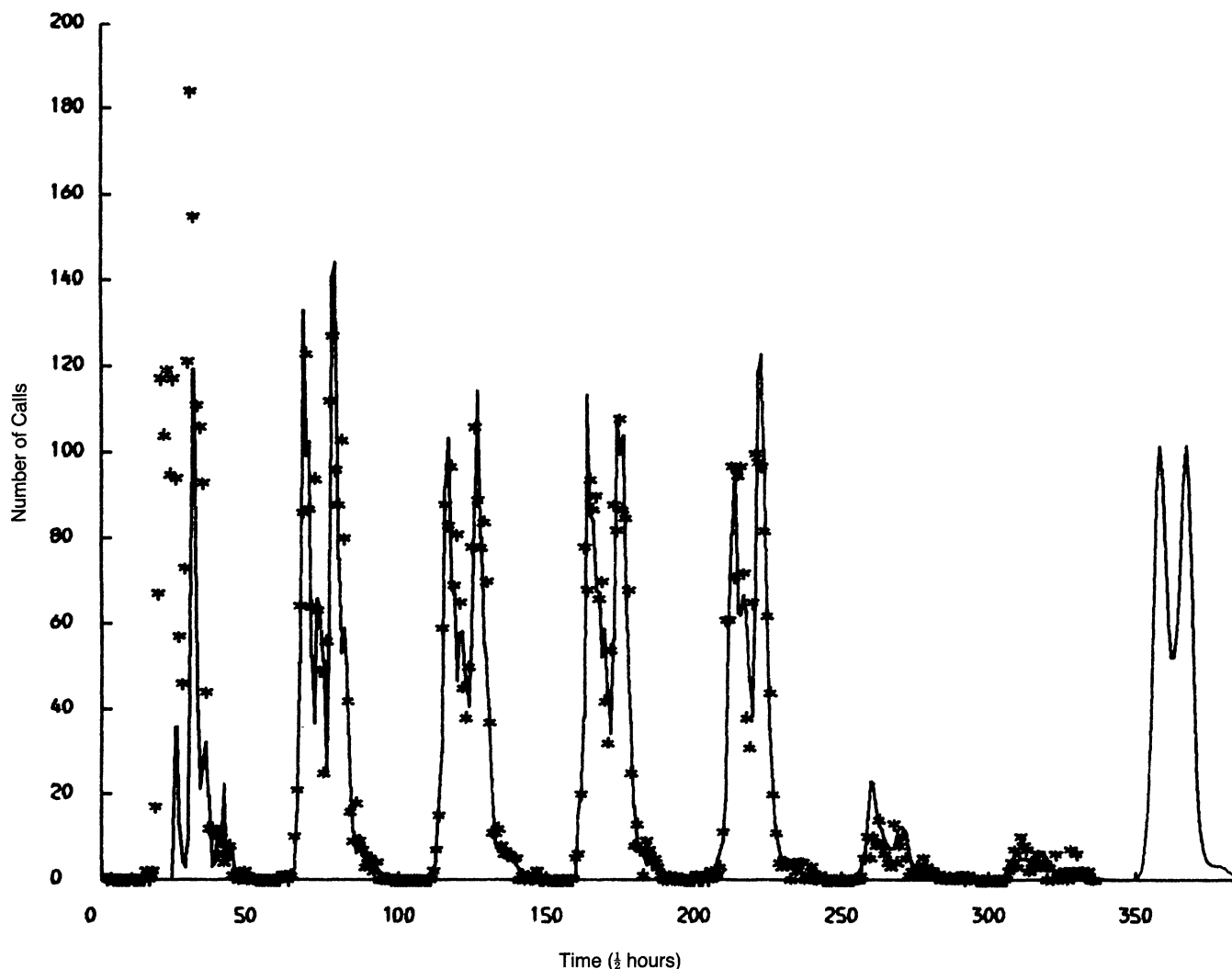
Figure 2. Example 2: Telephone Data and Forecasts.

quires no more computational effort than standard normal linear models, and in some cases, as in the next example, the latter are totally inadequate.

Example 2. The data plotted as asterisks in Figure 2 are the half-hourly counts of incoming telephone calls at the University of Warwick exchange from Monday, September 6, to Sunday, September 12, 1982.

The data and a nonsequential Bayesian analysis appeared in West (1985); note that many of the observations are zero. A Poisson sampling model is clearly appropriate and the model uses a log-linear level plus seasonal term with the first five harmonics of the Fourier description of period 48; thus $F'_t = (1; 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)$ and G_t, as in Example 1, with $H_0 = 1$ and $p = \pi/24$. In this case, $\beta_0^2 = .98$ for the level term and $\beta_1^2 = .99$ for the seasonal component. Further discounting is used between days to model stochastic differences between parameters from day to day as well as to slow changes during the day. This is done using $\beta_0^2 = \beta_1^2 = .7$ at $t = 48$, 96, 144, . . . , 288. The full line in Figure 2 gives the one-step forecasts (again using the modes of the predictive distribution) up to midnight on Sunday ($t = 336$) and then the week-ahead forecast from Friday midnight ($t = 240$) for the following Friday [the latter, longer-term predictions

using the scale-free predictor $E[Y_{t+k} \mid \eta_{t+k} = f_{t+k}] = \exp(f_{t+k})$]. The effect of the weak prior for θ_0 is clear from the forecasts for the first day. The response of the model to the changing form from day to day is also clear, especially at the weekend in which the change is most marked.

Example 3. Migon and Harrison (1985) discussed linear and nonlinear models used to study the relationship between television advertising and consumer awareness and to measure the effectiveness of advertising campaigns. Here a particular DGLM is applied to one of their data sets. The data, marked as asterisks in Figure 3(a), are weekly counts of the number of people, out of a nominal total $n = 66$, who provide a positive response to a standard question concerning the advertising of a popular chocolate bar. A binomial sampling model is clearly appropriate, with $E[Y_t \mid \eta_t] = n\mu_t$, $V[Y_t \mid \eta_t] = n\mu_t(1 - \mu_t)$, and $\eta_t = \ln[\mu_t/(1 - \mu_t)]$, with $n = 66$. There are several weeks when no data were collected, in which case $n = 0$. Weekly advertising is measured on a standard scale, and if x_t denotes the advertising in week t, then a suitable measure of cumulative advertising is the "adstock" $a_t = pa_{t-1} + x_t$, where p is a decay factor, taken as .9 in this example (see Migon and Harrison 1985 for further details). The adstock for this example is plotted as the lower graph in Figure 3(a), and the
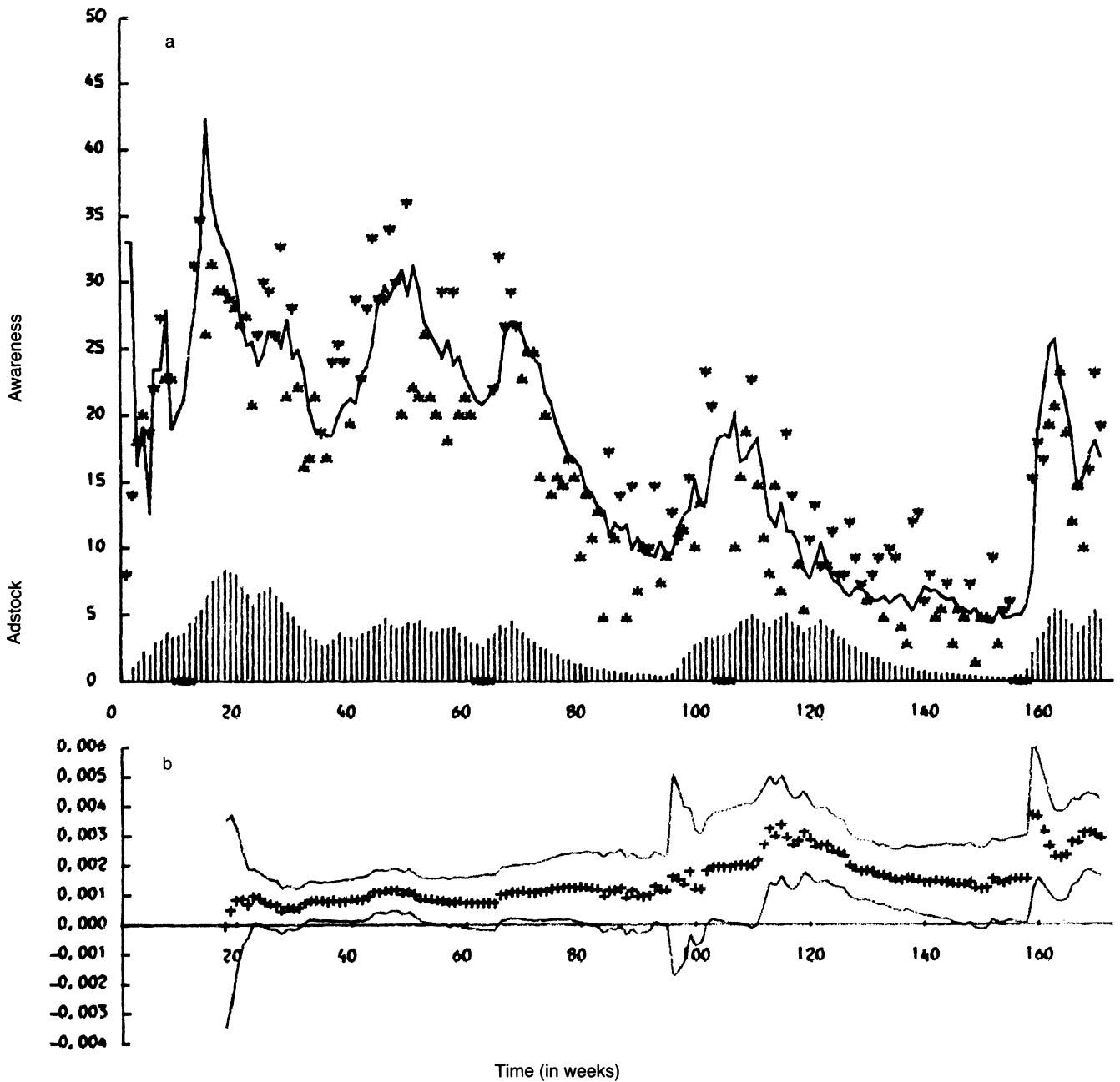
Figure 3. Example 3: (a) Television Data and Forecasts; (b) Adstock Coefficient.

model here is based on logistic-linear regression on a_t with a constant term; thus $F'_t = (1, a_t)$ and $G_t = I_2$, with $B_t = \text{diag}(\beta_0^{-1}, \beta_1^{-1})$, where $\beta_0^2 = .9$ and $\beta_1^2 = .95$. One further feature of the data is that the nature of the advertising campaign was changed at $t = 96$ and $t = 156$. To enable the model to adapt rapidly to any changes in the parameters $\theta'_0 = (\theta_{0t}, \theta_{1t})$ at these times, the discount factors are decreased to .1 at $t = 96$ and $t = 156$, resulting in greater uncertainty [$(.1)^{-.5} 100 \simeq 300\%$ increase in standard deviation] about the state parameters.

The full line in Figure 3(a) is the one-step forecasts using the means of the beta-binomial predictive distributions. Again, the effect of the weak prior for θ_0 is evident in the early stages. Figure 3(b) plots the on-line estimate m_{1t} of the adstock coef-

ficient θ_{1t} and the corresponding two standard-deviation limits. Some change is evident at the onset of the new advertising campaigns at $t = 96$ and $t = 156$. As a final check on the fit of the model, the posterior distribution for the scale parameter ϕ of Section 4.2 on a prior with $\nu_0 = \delta_0 = .01$ is $G[85.5, 100.0]$ with $\sigma^2 = 1.17$. This distribution adequately supports the value $\phi = 1$, indicating that there is no reason to question the model.

Example 4. The final example provides a comparison between the standard GLIM program analysis and a DGLM applied to a supposedly static regression. The data, recorded in Pregibon (1981), consist of 39 binary observations Y_t ($t = 1, \ldots, 39$) and the values of two continuous regressors X_{1t} and

Table 1. Maximum Likelihood Estimates
of the Elements of θ

	GLIM	DGLM
θ_0	−2.85 (1.36)	−2.73 (1.77)
θ_1	5.14 (1.88)	5.26 (1.86)
θ_2	4.53 (1.84)	4.01 (1.72)

NOTE: Approximate standard errors are in parentheses.

X_{2t}. The standard logistic model for this data supposes that $(Y_t \mid \eta_t)$ is Bernoulli with probability $\mu_t = 1/[1 + \exp(-\eta_t)]$, where

$$\eta_t = \mathbf{F}_t'\theta = \theta_0 + \theta_1 X_{1t} + \theta_2 X_{2t}$$

for $t = 1, \ldots, 39$.

The standard maximum likelihood method used in the GLIM program produces an estimate for the constant parameter vector θ and an approximate covariance matrix based on asymptotic normality. In this example the likelihood is unimodal, and the estimate of θ may be viewed as an approximate posterior mode from a Bayesian analysis using an uninformative prior. Of course, in contrast to the DGLM, the intractability of this model precludes the analytic calculation of posterior and predictive distributions for the η_t and Y_t. The maximum likelihood estimates of the elements of θ are given in Table 1 with their approximate standard errors in parentheses. Also given are corresponding "estimates" from a DGLM analysis that, for this static regression, has all discount factors equal to 1 so that $B_t = I_3$ for all t. The prior moments for $(\theta \mid D_0)$ were taken as $\mathbf{m}_0 = \mathbf{0}$ and $C_0 = (100)^2 I_3$ to express relatively vague prior knowledge; the estimates are simply $\mathbf{m}_t$ when $t = 39$, and the corresponding standard errors are the square roots of the diagonal elements of C_t when $t = 39$.

Clearly these values are in close agreement. The "fitted" probabilities in the DGLM are the predictive expectations $1/[1 + \exp(-\mathbf{F}_t'\mathbf{m}_{39})]$ of a further response Y_t at $\mathbf{F}_t$. These also agree closely with those from the GLIM program analysis, although the latter have no such interpretation. A final point concerns the sequential processing, and hence order dependence, in the DGLM of data from a supposedly static regression. We do not see this as a drawback; the full DGLM provides a general and flexible modeling tool, and much of this flexibility stems directly from the sequential approach adopted. To examine order dependence in this example, however, the DGLM analysis was performed again with the data in reverse order; the final results coincided with those of the original analysis with, for example, the elements of $\mathbf{m}_{39}$ and C_{39} agreeing to four decimal places.

This example is provided mainly to demonstrate the com-parability of the standard GLIM program and DGLM analyses using static models. The results are naturally comparable and generally uninfluenced by the order in which the sequential DGLM is performed. The DGLM formulation easily extends from the static to dynamic models to allow for the possibilities of time-varying parameters, outliers, and major changes. It is also, computationally, much simpler and less demanding.

REFERENCES

Ameen, J. R. M., and Harrison, P. J. (1985), "Normal Discount Bayesian Models," in Bayesian Statistics 2, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Amsterdam: North-Holland.

Azzalini, A. (1983), "Approximate Filtering of Parameter Driven Processes," Journal of Time Series Analysis, 3, 219–224.

Baker, R. J., and Nelder, J. A. (1978), GLIM Release 3, Oxford University, Numerical Algorithms Group.

Chatfield, C., and Prothero, D. D. (1973), "Box–Jenkins Seasonal Forecasting: A Case Study" (with discussion), Journal of the Royal Statistical Society, Ser. A, 136, 295–336.

Goldstein, M. (1976), "Bayesian Analysis of Regression Problems," Biometrika, 63, 51–58.

Harrison, P. J. (1973), Discussion of "Box–Jenkins Seasonal Forecasting: A Case Study," by C. Chatfield and D. D. Prothero, Journal of the Royal Statistical Society, Ser. A, 136, 319–324.

Harrison, P. J., and Stevens, C. F. (1976a), "Bayesian Forecasting" (with discussion), Journal of the Royal Statistical Society, Ser. B, 38, 205–247.

——— (1976b), "Case Studies in Bayesian Forecasting," Statistical Research Report 14, University of Warwick.

Hartigan, J. A. (1969), "Linear Bayesian Methods," Journal of the Royal Statistical Society, Ser. B, 31, 446–454.

Kalman, R. E. (1963), "New Methods in Wiener Filtering Theory," in Proceedings of the First Symposium on Engineering Applications of Random Function Theory and Probability, eds. J. L. Bogdanoff and F. Kozin, New York: John Wiley.

Migon, H. S., and Harrison, P. J. (1985), "An Application of Non-Linear Bayesian Forecasting to Television Advertising," in Bayesian Statistics 2, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Amsterdam: North-Holland.

Nelder, J. A., and Wedderburn, R. W. M. (1972), "Generalized Linear Models," Journal of the Royal Statistical Society, Ser. A, 135, 370–384.

Pregibon, D. (1981), "Logistic Regression Diagnostics," Annals of Statistics, 9, 705–724.

Smith, A. F. M., and West, M. (1983), "Monitoring Renal Transplants: An Application of the Multi-process Kalman Filter," Biometrics, 39, 867–878.

Smith, J. Q. (1979), "A Generalization of the Bayesian Steady Forecasting Model," Journal of the Royal Statistical Society, Ser. B, 41, 378–387.

Souza, R. C. (1981), "A Bayesian Entropy Approach to Forecasting; the Multi-state Model," in Time Series Analysis, Houston, TX: North-Holland, 535–542.

Tunnicliffe-Wilson, G. (1973), Discussion of "Box–Jenkins Seasonal Forecasting: A Case Study," by C. Chatfield and D. D. Prothero, Journal of the Royal Statistical Society, Ser. A, 136, 315–319.

West, M. (1981), "Robust Sequential Approximate Bayesian Estimation," Journal of the Royal Statistical Society, Ser. B, 43, 157–166.

——— (1984), "Outlier Models and Prior Distributions in Bayesian Linear Regression," Journal of the Royal Statistical Society.

——— (1985), "Generalized Linear Models; Scale Parameters, Outlier Accommodation and Prior Distributions," in Bayesian Statistics 2, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Amsterdam: North-Holland.