HIERARCHICAL BAYESIAN TIME SERIES MODELS

L. MARK BERLINER
National Center for Atmospheric Research
& Ohio State University
NCAR, P.O. Box 3000, Boulder, CO 80307-3000, USA. [†]

Abstract. Notions of Bayesian analysis are reviewed, with emphasis on Bayesian modeling and Bayesian calculation. A general hierarchical model for time series analysis is then presented and discussed. Both discrete time and continuous time formulations are discussed. An brief overview of generalizations of the fundamental hierarchical time series model concludes the article.

Key words: Dynamical model, Fokker-Planck equation, Markov process, Prediction, Stochastic differential equation

1. Intorduction

1.1. THE BAYESIAN VIEWPOINT

Much of the Bayesian viewpoint can be argued (as by Jeffreys and Jaynes, for examples) as direct application of the theory of probability. In this article the suggested approach for the construction of Bayesian time series models relies on probability theory to provide decompositions of complex joint probability distributions. Specifically, I refer to the familiar factorization of a joint density into an appropriate product of conditionals.

Let x and y represent two random variables. I will not differentiate between random variables and their realizations. Also, I will use an increasingly popular generic notation for probability densities: $[x]$ represents the density of x, $[x|y]$ is the conditional density of x given y, and $[x, y]$ denotes the joint density of x and y. In this notation we can write "Bayes's Theorem" as

$$[y|x] = [x|y][y]/[x]. \tag{1}$$

Equally important to probability theory and to Bayesian modeling is the relation

$$[x] = \int [x|y][y]dy. \tag{2}$$

[†]Email: berliner@ucar.edu. This work was supported by the NSF under Grant #DMS93-12686.

K. M. Hanson and R. N. Silver (eds.), Maximum Entropy and Bayesian Methods, 15–22.

Of course, relationships such as (1) and (2) hold for conditional densities. For example, if w is a third random variable, then

$$[x|w] = \int [x|y,w][y|w]dy.$$

Furthermore, a variety of nestings, hierarchies, and other relations among various conditional densities are possible. Finally, appropriate factorizations of joint densities are useful. For example,

$$[x,y,w] = [x|y,w][y|w][w] \tag{3}$$

is a familiar result in probability theory.

Equation (3) is particularly relevant to the discussion here. Specifically, it is the basis of *Hierarchical Models.* As modelers, faced with complex structures and a variety of random quantities to be modeled, Bayesians and other "stochastic modelers" break the modeling process of a large collection of variables into the pieces, following (3), and model the required conditional distributions. Hierarchical models have a long history in Bayesian statistics. Some discussion and references may be found in [1] and [2].

Markovian Models form a quintessential example of conditional modeling. In large scale applications of generalizations of (3), various patterns or structures in the conditional distributions are considered. Consider a time series of data, $x_1, \ldots, x_n$. We can write the joint distribution of these values as

$$[x_1, \ldots, x_n] = [x_n|x_{n-1}, \ldots, x_1][x_{n-1}|x_{n-2}, \ldots, x_1] \ldots [x_1]. \tag{4}$$

A common assumption about these conditionals goes something like "the distribution of x_t given 'all' the past x_s's only depends on a restricted subset of the *recent* past." A one-step Markov model is that for each t, $[x_t|\text{the past}] = [x_t|x_{t-1}]$, leading to the joint distribution

$$[x_1, \ldots, x_n] = [x_n|x_{n-1}][x_{n-1}|x_{n-2}] \ldots [x_1].$$

Markovian reasoning has been applied beyond the time series setting. The key observation is that one is free to index a countable collection of random variables in any convenient, meaningful fashion. The modeler may then directly apply (4) and formulate the resulting conditionals. A primary example involves modeling of spatial correlation in spatial statistics and image analysis. See [3] for discussion and references. The Markovian step involves the intuition that the value of a variable at a location, conditional on the values at appropriate collections of other locations, actually depends only on the values in a subset of "nearby" locations. Users of *Markov random field* models use this sort of reasoning, though the usual construction is not a direct use of (4). Hence, some care is taken to insure that the resulting specifications do yield a true joint distribution. *Markov meshes*, special cases of Markov random fields, are constructed by direct use of (4). See [4]. More generally, hierarchical reasoning offers an organized approach to spatio-temporal modeling, but this topic is beyond the scope of this article.

1.2. OUTLINE

Section 2 describes representations of an archetypal hierarchical model for time series. The model is presented in three stages. A casual way to think about these stages is:

Stage 1. [data|process, parameters].
Stage 2. [process|parameters].
Stage 3. [parameters].

In the time series context, the time evolution of the process of interest is primarily modeled in Stage 2. Both discrete and continuous time models are considered. The reader will note relationships between the models described and the so-called Kalman filter formulation; see [5]. In Section 3 a very brief discussion of some natural extensions of the basic models is given.

The purpose of this article is to present notions of modeling strategies useful in time series. For the most part, I will present the models with no concern for computational complexity. The models presented are indeed being developed today. Modern research involving Markov chain Monte Carlo offers a general approach to the approximation of Bayesian results in complex settings. See [6] for some discussion. Also, I will focus on "time domain" modeling; discussion of Bayesian spectral analysis may be found in [7]. Finally, this article is not intended to be a review of Bayesian time series; the reader is referred to [8], [9], and [10] for discussions and further references.

2. Hierarchical Models

2.1. DISCRETE TIME

Assume that a stochastic process, $x_o, x_1, \ldots$, is under study. To allow for measurement error, we allow that the process is observed indirectly as follows:

Stage 1. *Distribution of the Observables.* Assume that a set of n data values, $y_{s_1}, \ldots, y_{s_n}$, are observed. The first stage describes the structure of the conditional distribution of the data, given the underlying x process and any parameters of the model. A common choice is of the form

$$[y_{s_1}, \ldots, y_{s_n} | \{x_t\}_{t \geq 0}, \theta_1] = \prod_{i=1}^{n} [y_{s_i} | x_{s_i} \theta_1]. \tag{5}$$

The subscripts are intended to allow for a wide variety of sampling procedures. For example, in principle Bayesian analysis is unconcerned with issues of "equally spaced" observations. Next, I have assumed that the modeler wishes to allow for model parameters, represented by θ_1. A common example involves a regression formulation in which

$$y_{s_i} = G(x_{s_i}, \eta) + e_{i+1}, i = 1, \ldots, n, \tag{6}$$

where G is a regression function, η represents unknown regression coefficients, and the errors, e_i are uncorrelated, mean zero, random variables with some fixed density. Let the variance of these variables be denoted by v^2. In this context, we have $\theta_1 = (\eta, v^2)$. Another useful error model is a mixture of two distributions, one of which is comparatively longer tailed than the other, thereby presuming to allow for "outliers." The mixing probability can be incorporated into θ_1.

Stage 2. *Structure of the x Process.* A natural model for the evolution of the x process is a dynamic, typically Markovian, model. For example, a one-step Markov model involves the conditioning formula

$$[x_{t+1}|\{x_o, \ldots, x_t\}, \theta_2] = [x_{t+1}|x_t, \theta_2], \tag{7}$$

where θ_2 is a vector of parameters associated with the Stage 2 model. A common structure for this stage involves an autoregressive model,

$$x_{t+1} = F(x_t, \beta) + z_{t+1}, t \geq 0, \tag{8}$$

where F is a dynamical function, β represents unknown regression coefficients, and the z_t are mean zero, random variables. These variables are often suggested to represent unmodeled environmental effects, "noise," and uncertainty concerning the functional form F. Mixture models for the distribution of the effects may also be appropriate. In general, θ_2 is the vector of parameters composed of β and any parameters in the modeler's specification of the distribution of the z_t's. Finally, a prior for the initial condition is proposed: $[x_o]$. (This distribution may depend on the parameters.) Note that the one-step Markov model is merely an example. Higher order time dependencies can of course be modeled. Also, a high order Markovian model can often be written as a one-step, Markov model via *state space representation.*

Stage 3. *Prior on Parameters.* As a final stage for the model, we construct a distribution for the "parameters" introduced above: $[\theta_1, \theta_2]$.

Note that, the presentation of hierarchical models typically involves Markovian like reasoning, but without explicit reference. For example, the Stage 1 distribution described above is actually

$$[y_{s_1}, \ldots, y_{s_n}|\{x_t\}_{t\geq 0}, \theta_1, \theta_2] = \prod_{i=1}^{n}[y_{s_i}|x_{s_i}\theta_1],$$

but the model is that, given the x process and θ_1, the distribution of the data does not depend on θ_2. The specification of the components of these three stages yields a bona fide joint distribution for all the quantities modeled.

Direct computation, that is, probability theory, yield conditional distributions of interesting quantities given the observed data. The main object is

$$[\{x_t\}_{t\geq 0}, \theta_1, \theta_2|y_{s_1}, \ldots, y_{s_n}]. \tag{9}$$

Filtering and interpolation (inference for the x process at times corresponding to observation times and between observation times), backcasting or retrospection (inference for the x process at times before the first observation time), and prediction or forecasting (inference for the x process at times after the last observation time), are based on

$$[\{x_t\}_{t\geq 0}|y_{s_1},\ldots,y_{s_n}] = \int\int[\{x_t\}_{t\geq 0},\theta_1,\theta_2|y_{s_1},\ldots,y_{s_n}]d\theta_1 d\theta_2. \tag{10}$$

(I wrote the above formula as if θ_1 and θ_2 are continuous random variables. The adjustments to the representation in cases involving discrete components are familiar.) I repeatedly used the word "inference" rather than estimation to emphasize the Bayesian view that the conditional distribution of the quantity of interest ought to be the focus. Of course, practical limitations often force summaries of these distributions, though care, including consideration of decision theoretic issues, should be taken.

2.2. CONTINUOUS TIME

A natural starting point for extending the hierarchical model to continuous time is the replacement of (8) with a stochastic differential equation model. (I will not discuss assumptions used to make sense of all the points raised here. See [11].) In particular, consider the model

$$dx = f(x,\beta)dt + \sigma(x,\alpha)dW, \tag{11}$$

where dW represents white noise and f and σ are suitable functions, so that solutions to the equation make (Itô) sense. Define θ_2 to be the collection β and α.

"Kolmogorov Forward" or "Fokker-Planck" analysis based on (11) can be related to Bayesian calculations. The Fokker-Planck analysis solves the following problem: Assume that the initial value of the process described by (11) is a random variable, x_o, with specified density, $[x_o]$. Find the density, $p(x,t|\theta_2)$, of the x process at time t. The result is that $p(x,t|\theta_2)$ is an appropriate solution to the initial value problem

$$\frac{\partial p}{\partial t} = .5\frac{\partial^2}{\partial x^2}(\sigma^2 p) - \frac{\partial}{\partial x}(fp), \tag{12}$$

subject to the initial data $p(x,0|\theta_2) = [x_o]$.

Suppose we can solve the Fokker-Planck equation. Assuming data collection as described in (5), the quantity $p(x,s_1|\theta_2)$ may be viewed as the (conditional on θ_2) prior density for $x(s_1)$. Combining the model, $[y_{s_1}|x(s_1),\theta_1]$, and $p(x,s_1|\theta_2)$ via Bayes's Theorem, we can obtain the conditional posterior density

$$[x(s_1)|y_{s_1},\theta_1,\theta_2] \propto [y_{s_1}|x(s_1),\theta_1]p(x(s_1),s_1|\theta_2). \tag{13}$$

This object then serves as the initial data for the Fokker-Planck equation for the conditional density of $x(t), t > s_1$. We can proceed sequentially as more data is collected. Let

$$D_i = \{y_{s_k} : 1 \leq k \leq i\}.$$

Then, $[x_{s_i}|D_i,\theta_1,\theta_2]$, serves as the initial data for finding the conditional density of $x(t), t > s_i$.

Example: *The Langevin Equation and the Ornstein-Uhlenbeck Process.* The special case of (11),

$$dx = -\beta x dt + \sigma dW, \tag{14}$$

is easily (Itô) integrated. The parameters, $\beta > 0$ and $\sigma > 0$ form θ_2 in this case. Assume that the prior for the initial condition is a normal (Gaussian) distribution with mean $\mu(0)$ and variance $\tau^2(0)$. Under these assumptions this second stage prior leads to a conditionally Gaussian process.

At the first stage, we assume that the n observations, $(y_{s_1}, \ldots, y_{s_n})$, are conditionally independent, normal random variables with means $x(s_i), i = 1, \ldots, n$, and variances, v_i^2.

Next, the Fokker-Planck equation can be solved (or other methods can be called upon) so that the sequential updating can be implemented. Specifically, consider time s_1. Analysis (see [11], pp. 358, 367-68) yields the prior on $x(s_1)$ is Gaussian, with mean

$$\mu(0)\exp(-\beta s_1) \tag{15}$$

and variance

$$\phi^2 = \tau^2(0)\exp(-2\beta s_1) + (\frac{\sigma^2}{2\beta})(1 - \exp(-2\beta s_1)). \tag{16}$$

Next, we combine this with the data point y_{s_1}, as prescribed in (13). The required calculation is familiar in Bayesian analysis ([1], pp. 129-30). The result is that $[x(s_1)|y_{s_1}, \theta_1, \theta_2]$ is a normal density, with mean

$$\mu(s_1) = \{\phi^2/(v_1^2 + \phi^2)\}y_{s_1} + \{v_1^2/(v_1^2 + \phi^2)\}\mu(0)\exp(-\beta s_1), \tag{17}$$

and variance

$$\tau^2(s_1) = \{v_1^2\phi^2\}/\{v_1^2 + \phi^2\}. \tag{18}$$

We can then continue sequentially as more data is collected by recursing $\mu(\cdot)$ and $\tau^2(\cdot)$, being careful to remember to use lengths of time intervals, $s_{k+1} - s_k$, appropriately, including the definition of the ϕ^2 function at each iterate. □

Calculations for dealing with the parameters in the model are direct, in principle. We would sequentially update the distributions, $[\theta_1, \theta_2|D_i]$ of the model parameters via Bayes's Theorem. Based on these distributions, we can compute the quantity,

$$[x(s_i)|D_i] = \int\int [x(s_i)|D_i, \theta_1, \theta_2][\theta_1, \theta_2|D_i]d\theta_1 d\theta_2.$$

The forward analysis described above yield sequential Bayesian predictive and parameter inference analyses. However, for filtering, interpolation, and backcasting based on the full data, one would need the conditional distributions of the x process at all required time points of interest given the full data set.

Implementation of the above analyses is formidible from a computational view. First, the Fokker-Planck equation is seldom tractable enough to be useful in the sense described above. Second, the application of Bayes's Theorem is also typically numerically intensive. Even in the Langevin/Ornstein-Uhlenbeck prior example, updating with priors on β and σ would be difficult. A variety of approximations are available. If the Fokker-Planck analysis appears to be not useful, an interesting possibility is to use a discrete time approximation to the continuous time model. See [12] for an example.

Another issue involves the assertion that an observation is based precisely on the exact value of the underlying process at a specified instant. In many settings uncertainty in the times of observation arise, see [13]. Second, many data collection techniques involve the observation (with error) of weighted time integrals and transforms of the underlying process. In principle, such data can be modeled by appropriate extensions of (6). Finally, in some circumstances it may be appropriate to consider "analog" or continuously sampled data.

3. Extensions

A variety of extensions, including the use of additional stages in hierarchies, are possible. I only allude to a few of these, primarily in the discrete time formulation, but note that this is an active area of research.

An obvious extension to the models described permits time varying parameters, $\theta_1(t)$ and $\theta_2(t)$. Formally, this is no extension at all, since we could append $\theta_1(t)$ and $\theta_2(t)$ to the definition of the variable x modeled in Stage 2. Formalities aside, it may often be sensible from a modeling viewpoint to separate out parameters from process variables. For example, the parameters might be viewed as "slowly varying," compared to the x process.

A more general direction for modeling time varying structure is to suggest that switching from one model paradigm to another occurs. A natural suggestion is to introduce a process variable, say I, where $I = 1, 2, \ldots,$ or K. Stage 2 is then extended as follows:

$$[x_{t+1} | \{x_o, \ldots, x_t\}, \{I_o, \ldots, I_t\}, \theta_2] = [x_{t+1} | x_t, I_t, \theta_2]. \tag{19}$$

The variable I then indicates which of K models,

$$x_{t+1} = F_I(x_t, \eta(I)) + z^I_{t+1}, t \geq 0, \tag{20}$$

and or parameters, $\eta(.)$, is in effect.

The indicator process's evolution itself is then modeled. (The Bayesian use of "mixture models" is of direct use in the sort of models described; see [13].) Two basic approaches are common. First, a (typically, one-step) Markov chain, is used. There is a growing literature on such hierarchical models under the name "Hidden Markov Models." (The "hidden" modifer refers to the fact that the indicators I_t are in fact not typically observed.) A second main class of models involve a variety of "state-dependent" models for the evolution of the indicator process. A key reference in this regard is [14]. In many settings it is natural to believe that the modeling of the dynamics of the x process, say by (8), by defining the autoregressive function relating x_{t+1} to x_t locally depending on the value of x_t. Further, the error process can too be modeled similarly. (H. Tong has been instrumental in making such reasoning, called the "threshold principle," popular in time series analysis; See [15]) Note that we may use the notation of the hidden Markov model, except that the indicator I_t at time t is a function of the state x_t, rather than an automonous process.

The reader may have detected the fact that many of the models described above are not necessarily associated with Bayesian time series. The reader may also have detected that the observation is irrelevant. Good, useful models can be incorporated readily as stages of Bayesian models. Many of us would argue that the hierarchical Bayesian approach therefore subsumes classical modeling in a fashion that both extends the range and enhances the interpretability of time series models.

References

1. J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verla, New York, 1985.
2. J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, John Wiley & Sons, Inc., New York, 1994.
3. Noel Cressie, *Statistics for Spatial Data*, John Wiley & Sons, Inc., New York, 1991.
4. K. Abend and T. J. Harley and L. N. Kanal, "Classification of binary random patterns," *IEEE Trans. Inform. Theory*, **IT-11**, pp. 538-544, 1965.
5. R. J. Meinhold and N. Singpurwalla, "Understanding the Kalman filter," *Amer. Statist.*, **37**, pp. 123-127, 1983.
6. J. Besag and P. Green and D. Higdon and K. Mengersen, "Bayesian computation and stochastic systems," *Statist. Sci.*, **10**, pp. 3-66 (with Discussion), 1995.
7. G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*, Springer-Verlag, New York, 1988.
8. M. West and J. Harrison, *Bayesian Forecasting and Dynamic Models*, Springer-Verlag, New York, 1989.
9. A. Pole and M. West and J. Harrison, *Applied Bayesian Forecasting and Time Series Analysis*, Chapman-Hall, New York, 1994.
10. J. C. Spall(Ed.), *Bayesian Analysis of Time Series and Dynamic Models*, Marcel Dekker, New York, 1988.
11. A. Lasota and M. C. Mackey, *Chaos, Fractal, and Noise*, Springer-Verlag, New York, 1994.
12. C. M. Scipione and L. M. Berliner, "Bayesian inference in nonlinear dynamical systems," 1993 Proc. of the Section on Bayesian Statist. Sci., American Statistical Association, Washington, D.C., 1993.
13. M. West, *Bayesian time series: Models and computations for the analysis of time series in the physical sciences* in *This Volume.*
14. M. Priestley, *Non-Linear and Non-Stationary Time Series Analysis*, Academic Press, New York, 1988.
15. H. Tong, *Non-Linear Time Series*, Oxford University Press, New York, 1990.