Prepared for **The Handbook of Economic Forecasting**, edited by Graham Elliott, Clive W. J. Granger, and Allan Timmerman. Amsterdam: North-Holland. Forthcoming, 2006.

Bayesian Forecasting

John Geweke and Charles Whiteman
Department of Economics
University of Iowa
Iowa City, IA 52242-1000

September 17, 2004

Abstract

Bayesian forecasting is a natural product of a Bayesian approach to inference. The Bayesian approach in general requires explicit formulation of a model, and conditioning on known quantities, in order to draw inferences about unknown ones. In Bayesian forecasting, one simply takes a subset of the unknown quantities to be future values of some variables of interest. This paper presents the principles of Bayesian forecasting, and describes recent advances in compuational capabilities for applying them that have dramatically expanded the scope of applicability of the Bayesian approach. It describes historical developments and the analytic compromises that were necessary prior to recent developments, the application of the new procedures in a variety of examples, and reports on two long-term Bayesian forecasting exercises.

...in terms of forecasting ability, ... a good Bayesian will beat a non-Bayesian, who will do better than a bad Bayesian.

(C.W.J. Granger, 1986, p. 16)

1 Introduction

Forecasting involves the use of information at hand—hunches, formal models, data, etc.—to make statements about the likely course of future events. In technical terms, conditional on what one knows, what can one say about the future? The Bayesian approach to inference, as well as decision-making and forecasting, involves conditioning on what is known to make statements about what is not known. Thus "Bayesian Forecasting" is a mild redundancy, because forecasting is at the core of the Bayesian approach to just about anything. The parameters of a model, for example, are no more known than future values of the data thought to be generated by that model, and indeed the Bayesian approach treats the two types of unknowns in symmetric fashion. The future values of

an economic time series simply constitute another function of interest for the Bayesian analysis.

Conditioning on what is known, of course, means using prior knowledge of structures, reasonable parameterizations, etc., and it is often thought that it is the use of prior information that is the salient feature of a Bayesian analysis. While the use of such information is certainly a distinguishing feature of a Bayesian approach, it is merely an implication of the principles that one should fully specify what is known and what is unknown, and then condition on what is known in making probabilistic statements about what is unknown.

Until recently, each of these two principles posed substantial technical obstacles for Bayesian analyses. Conditioning on known data and structures generally leads to integration problems whose intractability grows with the realism and complexity of the problem's formulation. Fortunately, advances in numerical integration that have occurred during the past fifteen years have steadily broadened the class of forecasting problems that can be addressed routinely in a careful yet practical fashion. This development has simultaneously enlarged the scope of models that can be brought to bear on forecasting problems using either Bayesian or non-Bayesian methods, and significantly increased the quality of economic forecasting. This chapter provides both the technical foundation for these advances, and the history of how they came about and improved economic decision-making.

2 Bayesian inference and forecasting: a primer

Bayesian methods of inference and forecasting all derive from two simple principles.

1. *Principle of explicit formulation.* Express all assumptions using formal probability statements about the joint distribution of future events of interest and relevant events observed at the time decisions, including forecasts, must be made.

2. *Principle of relevant conditioning.* In forecasting, use the distribution of future events conditional on observed relevant events and an explicit loss function.

The fun (if not the devil) is in the details. Technical obstacles can limit the expression of assumptions and loss functions or impose compromises and approximations. These obstacles have largely fallen with the advent of posterior simulation methods described in Section 3, methods that have themselves motivated entirely new forecasting models. In practice those doing the technical work with distributions (investigators, in the dichotomy drawn by Hildreth (1963)) and those whose decision-making drives the list of future events and the choice of loss function (Hildreth's clients) may not be the same. This poses the question of what investigators should report, especially if their clients are anonymous, an issue to which we return in Section 3.3. In these and a host of other tactics, the two principles provide the strategy.

2.1 Models for observables

Bayesian inference takes place in the context of one or more models that describe the behavior of a $p \times 1$ vector of observable random variables $\mathbf{y}_t$ over a sequence of discrete time units $t = 1, 2, \ldots$. The history of the sequence at time t is given by $\mathbf{Y}_t = \{\mathbf{y}_s\}_{s=1}^{t}$. The sample space for $\mathbf{y}_t$ is ψ_t, that for $\mathbf{Y}_t$ is Ψ_t, and $\psi_0 = \Psi_0 = \{\emptyset\}$. A model, A, specifies a corresponding sequence of probability density functions

$$p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}_A, A\right) \tag{1}$$

in which $\boldsymbol{\theta}_A$ is a $k_A \times 1$ vector of unobservables, and $\boldsymbol{\theta}_A \in \Theta_A \subseteq \mathbb{R}^k$. The vector $\boldsymbol{\theta}_A$ includes not only parameters as usually conceived, but also latent variables convenient in model formulation. This extension immediately accommodates non-standard distributions, time varying parameters, and heterogeneity across observations; Carter and Kohn (1994) provide examples of this flexibility in the context of Bayesian time series modeling.

The notation $p(\cdot)$ indicates a generic probability density function (p.d.f.) with respect to a generic measure $\nu(\cdot)$, and $P(\cdot)$ the corresponding cumulative distribution function (c.d.f.). For simplicity the measure notation can nearly always be suppressed and $\mathbf{y}_t$ treated as a continuously distributed random vector, but the results apply to discrete and mixed continuous-discrete random vectors as well. Then the probability density function (p.d.f.) for $\mathbf{Y}_T$, conditional on the model and unobservables vector $\boldsymbol{\theta}_A$, is

$$p\left(\mathbf{Y}_T \mid \boldsymbol{\theta}_A, A\right) = \prod_{t=1}^{T} p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}_A, A\right). \tag{2}$$

When used alone, expressions like $\mathbf{y}_t$ and $\mathbf{Y}_T$ denote random vectors. In equations (1) and (2) $\mathbf{y}_t$ and $\mathbf{Y}_T$ are arguments of functions. These uses are distinct from the observed values themselves. To preserve this distinction explicitly, denote observed $\mathbf{y}_t$ by $\mathbf{y}_t^o$ and observed $\mathbf{Y}_T$ by $\mathbf{Y}_T^o$. In general, the superscript o will denote the observed value of a random vector. For example, the *likelihood function* is $L\left(\boldsymbol{\theta}_A; \mathbf{Y}_T^o, A\right) \propto p\left(\mathbf{Y}_T^o \mid \boldsymbol{\theta}_A, A\right)$.

2.1.1 An example: vector autoregressions

Following Sims (1980) and Litterman (1979) (which are discussed below), vector autoregressive models have been utilized extensively in forecasting macroeconomic and other time series owing to the ease with which they can be used for this purpose and their apparent great success in implementation. Adapting the notation of Litterman (1979), the VAR specification for

$$p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}_A, A\right)$$

is given by

$$\mathbf{y}_t = \mathbf{B}_D D_t + \mathbf{B}_1 \mathbf{y}_{t-1} + \mathbf{B}_2 \mathbf{y}_{t-2} + \ldots + \mathbf{B}_m \mathbf{y}_{t-m} + \varepsilon_t \tag{3}$$

where A now signifies the autoregressive structure, D_t is a deterministic component of dimension d, and $\varepsilon_t \overset{iid}{\sim} N(0, \mathbf{\Psi})$. In this case,

$$\boldsymbol{\theta}_A = (\mathbf{B}_D, \mathbf{B}_1, ..., \mathbf{B}_m, \mathbf{\Psi}) .$$

2.1.2 An example: stochastic volatility

Models with time-varying volatility are standard tools in portfolio allocation problems. Jacquier, Polson and Rossi (1994) developed such a model utilizing a time series of latent volatilities $\mathbf{h} = (h_1, \ldots, h_T)'$:

$$h_1 \sim N\left[0, \sigma_\eta^2 / \left(1 - \phi^2\right)\right], \tag{4}$$

$$h_t = \phi h_{t-1} + \sigma_\eta \eta_t \ \ (t = 2, \ldots, T) . \tag{5}$$

An observable sequence of asset returns $\mathbf{y} = (y_1, \ldots, y_T)'$ is then conditionally independent,

$$y_t = \beta \exp\left(h_t/2\right) \varepsilon_t; \tag{6}$$

$(\varepsilon_t, \eta_t)' \overset{iid}{\sim} N\left(\mathbf{0}, \mathbf{I}_2\right)$. The $(T+3) \times 1$ vector of unobservables is

$$\boldsymbol{\theta}_A = \left(\beta, \sigma_\eta^2, \phi, h_1, \ldots, h_T\right)' . \tag{7}$$

It is conventional to speak of $\left(\beta, \sigma_\eta^2, \phi\right)$ as a parameter vector and $\mathbf{h}$ as a vector of latent variables, but in Bayesian inference this distinction is a matter only of language, not substance. The unobservables $\mathbf{h}$ can be any real numbers, whereas $\beta > 0$, $\sigma_\eta > 0$, and $\phi \in (-1, 1)$. If $\phi > 0$ then the observable sequence $\left\{y_t^2\right\}$ exhibits the positive serial correlation characteristic of many sequences of asset returns.

2.1.3 The forecasting vector of interest

Models are means, not ends. A useful link between models and the purposes for which they are formulated is a vector of interest, which we denote $\boldsymbol{\omega} \in \Omega \subseteq \mathbb{R}^q$. The vector of interest may be unobservable, for example the monetary equivalent of a change in welfare, or the change in an equilibrium price vector, following a hypothetical policy change. In order to be relevant, the model must not only specify (1), but also

$$p\left(\boldsymbol{\omega} \mid \mathbf{Y}_T, \boldsymbol{\theta}_A\right) . \tag{8}$$

In a forecasting problem, by definition, $\{y_{T+1}, \ldots, y_{T+F}\} \in \boldsymbol{\omega}$ for some $F > 0$. In some cases $\boldsymbol{\omega}' = (y_{T+1}, \ldots, y_{T+F})'$ and it is possible to express $p\left(\boldsymbol{\omega} \mid \mathbf{Y}_T, \boldsymbol{\theta}_A\right) \propto p\left(\mathbf{Y}_{T+F} \mid \boldsymbol{\theta}_A, A\right)$ in closed form, but in general this is not so. Suppose, for example, that a stochastic volatility model is a means to the solution of a portfolio allocation problem with a 20-day horizon so that $\boldsymbol{\omega} = (y_{T+1}, \ldots, y_{T+20})'$. Then there is no analytical expression for $p\left(\boldsymbol{\omega} \mid \mathbf{Y}_T, \boldsymbol{\theta}_A\right)$ with $\boldsymbol{\theta}_A$ defined as it is in (7). If $\boldsymbol{\omega}$ is extended to include $(h_{T+1}, \ldots, h_{T+20})'$ as well as $(y_{T+1}, \ldots, y_{T+20})'$, then the expression is simple. Continuing with an analytical approach then

confronts the original problem of integrating over $(h_{T+1}, \ldots, h_{T+20})'$ to obtain $p(\boldsymbol{\omega} \mid \mathbf{Y}_T, \boldsymbol{\theta}_A)$. But it also highlights the fact that it is easy to simulate from this extended definition of $\boldsymbol{\omega}$ in a way that is, today, obvious:

$$h_t \sim N\left(\phi h_{t-1},\ \sigma_\eta^2\right),\ \ y_t \sim N\left[0, \beta^2 \exp\left(h_t\right)\right] \quad (t = T+1, \ldots, T+20).$$

Since this produces a simulation from the joint distribution of $(h_{T+1}, \ldots, h_{T+20})'$ and $(y_{T+1}, \ldots, y_{T+20})'$, the "marginalization" problem simply amounts to discarding the simulated $(h_{T+1}, \ldots, h_{T+20})'$.

A quarter-century ago, this idea was far from obvious. Wecker (1979), in a paper on predicting turning points in macroeconomic time series, appears to have been the first to have used simulation to access the distribution of a problematic vector of interest $\boldsymbol{\omega}$ or functions of $\boldsymbol{\omega}$. His contribution was the first illustration of several principles that have emerged since and will appear repeatedly in this survey. One is that while producing marginal from joint distributions analytically is demanding and often impossible, in simulation it simply amounts to discarding what is irrelevant. (In Wecker's case the future y_{T+s} were irrelevant in the vector that also included indicator variables for turning points.) A second is that formal decision problems of many kinds, from point forecasts to portfolio allocations to the assessment of event probabilities can be solved using simulations of $\boldsymbol{\omega}$. Yet another insight is that it may be much simpler to introduce intermediate conditional distributions, thereby enlarging $\boldsymbol{\theta}_A$, $\boldsymbol{\omega}$, or both, retaining from the simulation only that which is relevant to the problem at hand. The latter idea was fully developed in the contribution of Tanner and Wong (1987).

2.2 Model completion with prior distributions

The generic model for observables (2) is expressed conditional on a vector of unobservables, $\boldsymbol{\theta}_A$, that includes unknown parameters. The same is true of the model for the vector of interest $\boldsymbol{\omega}$ in (8), and this remains true whether one simulates from this distribution or provides a full analytical treatment. Any workable solution of a forecasting problem must, in one way or another, address the fact that $\boldsymbol{\theta}_A$ is unobserved. A similar issue arises if there are alternative models A—different functional forms in (2) and (8)—and we return to this matter in Section 2.3.

2.2.1 The role of the prior

The Bayesian strategy is dictated by the first principle, which demands that we work with $p(\boldsymbol{\omega} \mid \mathbf{Y}_T, A)$. Given that $p(\mathbf{Y}_T \mid \boldsymbol{\theta}_A, A)$ has been specified in (2) and $p(\boldsymbol{\omega} \mid \mathbf{Y}_T, \boldsymbol{\theta}_A)$ in (8), we meet the requirements of the first principle by specifying

$$p(\boldsymbol{\theta}_A \mid A), \tag{9}$$

because then

$$p(\boldsymbol{\omega} \mid \mathbf{Y}_T, A) \propto \int_{\Theta_A} p(\boldsymbol{\theta}_A \mid A)\, p(\mathbf{Y}_T \mid \boldsymbol{\theta}_A, A)\, p(\boldsymbol{\omega} \mid \mathbf{Y}_T, \boldsymbol{\theta}_A)\, d\boldsymbol{\theta}_A.$$

The density $p(\boldsymbol{\theta}_A \mid A)$ defines the *prior distribution* of the unobservables. For many practical purposes its proves useful to work with an intermediate distribution, the *posterior distribution* of the unobservables whose density is

$$p(\boldsymbol{\theta}_A \mid \mathbf{Y}_T^o, A) \propto p(\boldsymbol{\theta}_A \mid A)\, p(\mathbf{Y}_T^o \mid \boldsymbol{\theta}_A, A)$$

and then $p(\boldsymbol{\omega} \mid \mathbf{Y}_T^o, A) = \int_{\Theta_A} p(\boldsymbol{\theta}_A \mid \mathbf{Y}_T^o, A)\, p(\boldsymbol{\omega} \mid \mathbf{Y}_T^o, \boldsymbol{\theta}_A)\, d\boldsymbol{\theta}_A$.

Much of the prior information in a complete model comes from the specification of (1): for example, Gaussian disturbances limit the scope for outliers regardless of the prior distribution of the unobservables; similarly in the stochastic volatility model outlined in Section 2.1.2 there can be no "leverage effects" in which outliers in period $T+1$ are more likely following a negative return in period T than following a positive return of the same magnitude. The prior distribution further refines what is reasonable in the model.

There are a number of ways that the prior distribution can be articulated. The most important, in Bayesian economic forecasting, have been the closely related principles of shrinkage and hierarchical prior distributions, which we take up shortly. Substantive expert information can be incorporated, and can improve forecasts. For example DeJong, Ingram and Whiteman (2000) and Ingram and Whiteman (1994) utilize dynamic stochastic general equilibrium models to provide prior distributions in vector autoregressions to the same good effect that Litterman (1979) did with shrinkage priors (see Section 4.3 below). Chulani et al. (1999) construct a prior distribution, in part, from expert information and use it to improve forecasts of the cost, schedule and quality of software under development. Heckerman (1997) provides a closely related approach to expressing prior distributions using Bayesian belief networks.

2.2.2 Prior predictive distributions

Regardless of how the conditional distribution of observables and the prior distribution of unobservables are formulated, together they provide a distribution of observables with density

$$p(\mathbf{Y}_T \mid A) = \int_{\Theta_A} p(\boldsymbol{\theta}_A \mid A)\, p(\mathbf{Y}_T \mid \boldsymbol{\theta}_A)\, d\boldsymbol{\theta}_A, \tag{10}$$

known as the *prior predictive density*. It summarizes the whole range of phenomena consistent with the complete model and it is generally very easy to access by means of simulation. Suppose that the values $\boldsymbol{\theta}_A^{(m)}$ are drawn i.i.d. from the prior distribution, an assumption that we denote $\boldsymbol{\theta}_A^{(m)} \overset{iid}{\sim} p(\boldsymbol{\theta}_A \mid A)$, and then successive values of $\mathbf{y}_t^{(m)}$ are drawn independently from the distributions whose densities are given in (1),

$$\mathbf{y}_t^{(m)} \overset{id}{\sim} p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}^{(m)}, \boldsymbol{\theta}_A^{(m)}, A\right) \quad (t = 1, \ldots, T;\ m = 1, \ldots, M). \tag{11}$$

Then the simulated samples $\mathbf{Y}_T^{(m)} \overset{iid}{\sim} p(\mathbf{Y}_T \mid A)$. Notice that so long as prior distributions of the parameters are tractable, this exercise is entirely straight-

forward. The vector autoregression and stochastic volatility models introduced above are both easy cases.

The prior predictive distribution summarizes the substance of the model and emphasizes the fact that the prior distribution and the conditional distribution of observables are inseparable components, a point forcefully argued a quarter-century ago in a seminal paper by George Box (1980). It can also be a very useful tool in understanding a model – one that can greatly enhance research productivity, as emphasized in recent papers by Geweke (1998), Geweke and McCausland (2001) and Gelman (2003) as well as in recent Bayesian econometrics texts by Lancaster (2004, Section 2.4) and Geweke (2005, Section 5.3.1). This is because simulation from the prior predictive distribution is generally much simpler than formal inference (Bayesian or otherwise) and can be carried out relatively quickly when a model is first formulated. One can readily address the question of whether an observed function of the data $g\left(\mathbf{Y}_T^o\right)$ is consistent with the model by checking to see whether it is within the support of $p\left[g\left(\mathbf{Y}_T\right) \mid A\right]$ which in turn is represented by $g\left(\mathbf{Y}_T^{(m)}\right)$ $(m = 1, \ldots M)$. The function g could, for example, be a unit root test statistic, a measure of leverage, or the point estimate of a long-memory parameter.

2.2.3 Hierarchical priors and shrinkage

A common technique in constructing a prior distribution is the use of intermediate parameters to facilitate expressing the distribution. For example suppose that the prior distribution of a parameter μ is Student-t with location parameter $\underline{\mu}$, scale parameter $\underline{h}^{-1}$ and ν degrees of freedom. The underscores, here, denote parameters of the prior distribution, constants that are part of the model definition and are assigned numerical values. Drawing on the familiar genesis of the t-distribution, the same prior distribution could be expressed $\left(\underline{\nu}/\underline{h}\right) h \sim \chi^2\left(\underline{\nu}\right)$, the first step in the hierarchical prior, and then $\mu \mid h \sim N\left(\underline{\mu}, h^{-1}\right)$, the second step. The unobservable h is an intermediate device useful in expressing the prior distribution; such unobservables are sometimes termed *hyperparameters* in the literature. A prior distribution with such intermediate parameters is a *hierarchical prior*, a concept introduced by Lindley and Smith (1972) and Smith (1973). In the case of the Student-t distribution this is obviously unnecessary, but it still proves quite convenient in conjunction with the posterior simulators discussed in Section 3.

In the formal generalization of this idea the complete model provides the prior distribution by first specifying the distribution of a vector of hyperparameters $\boldsymbol{\theta}_A^*$, $p\left(\boldsymbol{\theta}_A^* \mid A\right)$, and then the prior distribution of a parameter vector $\boldsymbol{\theta}_A$ conditional on $\boldsymbol{\theta}_A^*$, $p\left(\boldsymbol{\theta}_A \mid \boldsymbol{\theta}_A^*, A\right)$. The distinction between a hyperparameter and a parameter is that the distribution of the observable is expressed, directly, conditional on the latter: $p\left(\mathbf{Y}_T \mid \boldsymbol{\theta}_A, A\right)$. Clearly one could have more than one layer of hyperparameters and there is no reason why $\boldsymbol{\theta}_A^*$ could not also appear in the observables distribution.

In other settings hierarchical prior distributions are not only convenient, but

essential. In economic forecasting important instances of hierarchical prior arise when there are many parameters, say $\theta_1, \ldots, \theta_r$ that are thought to be similar but about whose common central tendency there is less information. To take the simplest case, that of a multivariate normal prior distribution, this idea could be expressed by means of a variance matrix with large on-diagonal elements $\underline{h}^{-1}$, and off-diagonal elements $\underline{\rho}$, with $\underline{\rho}$ close to 1. Equivalently, this idea could be expressed by introducing the hyperparameter θ^*, then taking

$$\theta^* \sim N\left(0, \underline{\rho}\underline{h}^{-1}\right) \tag{12}$$

followed by

$$\theta_i \mid \theta^* \sim N\left[\theta^*, \left(1-\underline{\rho}\right)\underline{h}^{-1}\right], \tag{13}$$

$$\mathbf{y}_t \sim p\left(\mathbf{y}_t \mid \theta_1, \ldots, \theta_r\right) \quad (t = 1, \ldots, T). \tag{14}$$

This idea could then easily be merged with the strategy for handling the Student-t distribution, allowing some outliers among θ_i (a Student-t distribution conditional on θ^*), thicker tails in the distribution of θ^*, or both.

The application of hierarchical priors in (12)-(13) is an example of shrinkage. The concept is familiar in non-Bayesian treatments as well (for example, ridge regression) where its formal motivation originated with James and Stein (1961). In the Bayesian setting shrinkage is toward a common unknown mean θ^*, for which a posterior distribution will be determined by the data, given the prior.

This idea has proven to be vital in forecasting problems in which there are many parameters. Section 4 reviews its application in vector autoregressions and its critical role in turning mediocre into superior forecasts in that model. Zellner and Hong (1989) used this strategy in forecasting growth rates of output for 18 different countries, and it proved to minimize mean square forecast error among eight competing treatments of the same model. More recently Tobias (2001) applied the same strategy in developing predictive intervals in the same model. Zellner and Chen (2001) approached the problem of forecasting US real GDP growth by disaggregating across sectors and employing a prior that shrinks sector parameters toward a common but unknown mean, with a payoff similar to that in Zellner and Hong (1989). In forecasting long-run returns to over 1,000 initial public offerings Brav (2000) found a prior with shrinkage toward an unknown mean essential in producing superior results.

2.2.4 Latent variables

Latent variables, like the volatilities h_t in the stochastic volatility model of Section 2.1.2, are common in econometric modelling. Their treatment in Bayesian inference is no different from the treatment of other unobservables, like parameters. In fact latent variables are, formally, no different from hyperparameters. For the stochastic volatility model equations (4)-(5) provides the distribution of the latent variables (hyperparameters) conditional on the parameters, just as (12) provides the hyperparameter distribution in the illustration of shrinkage.

Conditional on the latent variables $\{h_t\}$, (6) indicates the observables distribution, just as (14) indicates the distribution of observables conditional on the parameters.

In the formal generalization of this idea the complete model provides a conventional prior distribution $p(\boldsymbol{\theta}_A \mid A)$, and then the distribution of a vector of latent variables $\mathbf{z}$ conditional on $\boldsymbol{\theta}_A$, $p(\mathbf{z} \mid \boldsymbol{\theta}_A, A)$. The observables distribution typically involves both $\mathbf{z}$ and $\boldsymbol{\theta}_A$: $p(\mathbf{Y}_T \mid \mathbf{z}, \boldsymbol{\theta}_A, A)$. Clearly one could also have a hierarchical prior distribution for $\boldsymbol{\theta}_A$ in this context as well.

Latent variables are convenient, but not essential, devices for describing the distribution of observables, just as hyperparameters are convenient but not essential in constructing prior distributions. The convenience stems from the fact that the likelihood function is otherwise awkward to express, as the reader can readily verify for the stochastic volatility model. In these situations Bayesian inference then has to confront the problem that it is impractical, if not impossible, to evaluate the likelihood function or even to provide an adequate numerical approximation. Tanner and Wong (1987) provided a systematic method for avoiding analytical integration in evaluating the likelihood function, through a simulation method they described as data augmentation. Section 5.2.2 provides an example.

This ability to use latent variables in a routine and practical way in conjunction with Bayesian inference has spawned a generation of Bayesian time series models useful in prediction. These include state space mixture models (see Carter and Kohn (1994, 1996) and Gerlach et al. (2000)), component models (see West (1995) and Huerta and West (1999)) and factor models (see Geweke and Zhou (1996) and Aguilar and West (2000)). The last paper provides a full application to the applied forecasting problem of foreign exchange portfolio allocation.

2.3 Model combination and evaluation

In applied forecasting and decision problems one typically has under consideration not a single model A, but several alternative models $A_1, \ldots, A_J$. Each model is comprised of a conditional observables density (1), a conditional density of a vector of interest $\boldsymbol{\omega}$ (8) and a prior density (9). For a finite number of models, each fully articulated in this way, treatment is dictated by the principle of explicit formulation: extend the formal probability treatment to include all J models. This extension requires only attaching prior probabilities $p(A_j)$ to the models, and then conducting inference and addressing decision problems conditional on the universal model specification

$$\left\{p(A_j), p\left(\boldsymbol{\theta}_{A_j} \mid A_j\right), p\left(\mathbf{Y}_T \mid \boldsymbol{\theta}_{A_j}, A_j\right), p\left(\boldsymbol{\omega} \mid \mathbf{Y}_T, \boldsymbol{\theta}_{A_j}, A_j\right)\right\} \quad (j = 1, \ldots, J). \tag{15}$$

The J models are related by their prior predictions for a common set of observables $\mathbf{Y}_T$ and a common vector of interest $\boldsymbol{\omega}$. The models may be quite similar: some, or all, of them might have the same vector of unobservables $\boldsymbol{\theta}_A$ and the same functional form for $p(\mathbf{Y}_T \mid \boldsymbol{\theta}_A, A)$, and differ only in their

specification of the prior density $p(\boldsymbol{\theta}_A \mid A_j)$. At the other extreme some of the models in the universe might be simple or have a few unobservables, while others could be very complex with the number of unobservables, which include any latent variables, substantially exceeding the number of observables. There is no nesting requirement.

2.3.1 Models and probability

The penultimate objective in Bayesian forecasting is the distribution of the vector of interest $\boldsymbol{\omega}$, conditional on the data $\mathbf{Y}_T^o$ and the universal model specification $A = \{A_1, \ldots, A_J\}$. Given (15) the formal solution is

$$p(\boldsymbol{\omega} \mid \mathbf{Y}_T^o) = \sum_{j=1}^{J} p(\boldsymbol{\omega} \mid \mathbf{Y}_T^o, A_j)\, p(A_j \mid \mathbf{Y}_T^o), \tag{16}$$

known as *model averaging*. In expression (16),

$$\begin{aligned} p(A_j \mid \mathbf{Y}_T^o) &= p(\mathbf{Y}_T^o \mid A_j)\, p(A_j) / p(\mathbf{Y}_T^o) && (17) \\ &\propto p(\mathbf{Y}_T^o \mid A_j)\, p(A_j). && (18) \end{aligned}$$

Expression (17) is the posterior probability of model A_j. Since these probabilities sum to 1, the values in (18) are sufficient. Of the two components in (18) the second is the prior probability of model A_j. The first is the *marginal likelihood*

$$p(\mathbf{Y}_T^o \mid A_j) = \int_{\Theta_{A_j}} p\left(\mathbf{Y}_T^o \mid \boldsymbol{\theta}_{A_j}, A_j\right) p\left(\boldsymbol{\theta}_{A_j} \mid A_j\right) d\boldsymbol{\theta}_{A_j}. \tag{19}$$

Comparing (19) with (10), note that (19) is simply the prior predictive density, evaluated at the realized outcome $\mathbf{Y}_T^o$ – the data.

The ratio of posterior probabilities of the models A_j and A_k is

$$\frac{P(A_j \mid \mathbf{Y}_T^o)}{P(A_k \mid \mathbf{Y}_T^o)} = \frac{P(A_j)}{P(A_k)} \cdot \frac{p(\mathbf{Y}_T^o \mid A_j)}{p(\mathbf{Y}_T^o \mid A_k)}, \tag{20}$$

known as the *posterior odds ratio* in favor of model A_j versus model A_k. It is the product of the *prior odds ratio* $P(A_j)/P(A_k)$, and the ratio of marginal likelihoods $p(\mathbf{Y}_T^o \mid A_j)/p(\mathbf{Y}_T^o \mid A_k)$, known as the *Bayes factor*. The Bayes factor, which may be interpreted as updating the prior odds ratio to the posterior odds ratio, is independent of the other models in the universe $\{A_1, \ldots, A_J\}$. This quantity is central in summarizing the evidence in favor of one model, or theory, as opposed to another one, an idea due to Jeffreys (1939). The significance of this fact in the statistics literature was recognized by Roberts (1965), and in econometrics by Leamer (1978). The Bayes factor is now a practical tool in applied statistics; see the reviews of Draper (1995), Chatfield (1995), Kass and Raftery (1995) and Hoeting et al. (1999).

2.3.2 A model is as good as its predictions

It is through the marginal likelihoods $p\left(\mathbf{Y}_T^o \mid A_j\right)$ $(j = 1, \ldots, J)$ that the observed outcome (data) determines the relative contribution of competing models to the posterior distribution of the vector of interest $\boldsymbol{\omega}$. There is a close and formal link between a model's marginal likelihood and the adequacy of its out-of-sample predictions. To establish this link consider the specific case of a forecasting horizon of F periods, with $\boldsymbol{\omega}' = \left(\mathbf{y}'_{T+1}, \ldots, \mathbf{y}'_{T+F}\right)$. The *predictive density* of $\mathbf{y}_{T+1}, \ldots, \mathbf{y}_{T+F}$, conditional on the data $\mathbf{Y}_T^o$ and a particular model A is

$$p\left(\mathbf{y}_{T+1}, \ldots, \mathbf{y}_{T+F} \mid \mathbf{Y}_T^o, A\right). \tag{21}$$

The predictive density is relevant after formulation of the model A and observing $\mathbf{Y}_T = \mathbf{Y}_T^o$, but before observing $\mathbf{y}_{T+1}, \ldots, \mathbf{y}_{T+F}$. Once $\mathbf{y}_{T+1}, \ldots, \mathbf{y}_{T+F}$ are known, we can evaluate (21) at the observed values. This yields the *predictive likelihood* of $\mathbf{y}_{T+1}^o, \ldots, \mathbf{y}_{T+F}^o$ conditional on $\mathbf{Y}_T^o$ and the model A, the real number $p\left(\mathbf{y}_{T+1}^o, \ldots, \mathbf{y}_{T+F}^o \mid \mathbf{Y}_T^o, A\right)$. Correspondingly, the *predictive Bayes factor* in favor of model A_j, versus the model A_k, is

$$p\left(\mathbf{y}_{T+1}^o, \ldots, \mathbf{y}_{T+F}^o \mid \mathbf{Y}_T^o, A_j\right) / p\left(\mathbf{y}_{T+1}^o, \ldots, \mathbf{y}_{T+F}^o \mid \mathbf{Y}_T^o, A_k\right).$$

There is an illuminating link between predictive likelihood and marginal likelihood that dates at least to Geisel (1975). Since

$$\begin{aligned} p\left(\mathbf{Y}_{T+F} \mid A\right) &= p\left(\mathbf{Y}_{T+F} \mid \mathbf{Y}_T, A\right) p\left(\mathbf{Y}_T \mid A\right) \\ &= p\left(\mathbf{y}_{T+1}, \ldots, \mathbf{y}_{T+F} \mid \mathbf{Y}_T, A\right) p\left(\mathbf{Y}_T \mid A\right), \end{aligned}$$

the predictive likelihood is the ratio of marginal likelihoods

$$p\left(\mathbf{y}_{T+1}^o, \ldots, \mathbf{y}_{T+F}^o \mid \mathbf{Y}_T^o, A\right) = p\left(\mathbf{Y}_{T+F}^o \mid A\right) / p\left(\mathbf{Y}_T^o \mid A\right).$$

Thus the predictive likelihood is the factor that updates the marginal likelihood, as more data become available.

This updating relationship is quite general. Let the strictly increasing sequence of integers $\{s_j,\ (j = 0, \ldots, q)\}$ with $s_0 = 1$ and $s_q = T$ partition T periods of observations $\mathbf{Y}_T^o$. Then

$$p\left(\mathbf{Y}_T^o \mid A\right) = \prod_{\tau=1}^{q} p\left(\mathbf{y}_{s_{\tau-1}+1}^o, \ldots, \mathbf{y}_{s_\tau}^o \mid \mathbf{Y}_{s_{\tau-1}}^o, A\right). \tag{22}$$

This decomposition is central in the updating and prediction cycle that

1. Provides a probability density for the next $s_\tau - s_{\tau-1}$ periods

$$p\left(\mathbf{y}_{s_{\tau-1}+1}, \ldots, \mathbf{y}_{s_\tau} \mid \mathbf{Y}_{s_{\tau-1}}^o, A\right),$$

2. After these events are realized evaluates the fit of this probability density by means of the predictive likelihood

$$p\left(\mathbf{y}^o_{s_{\tau-1}+1}, \ldots, \mathbf{y}^o_{s_\tau} \mid \mathbf{Y}^o_{s_{\tau-1}}, A\right),$$

3. Updates the posterior density

$$p\left(\boldsymbol{\theta}_A \mid \mathbf{Y}^o_{s_\tau}\right) \propto p\left(\boldsymbol{\theta}_A \mid \mathbf{Y}^o_{s_{\tau-1}}\right) p\left(\mathbf{y}^o_{s_{\tau-1}+1}, \ldots, \mathbf{y}^o_{s_\tau} \mid \mathbf{Y}^o_{s_{\tau-1}}, \boldsymbol{\theta}_A, A\right),$$

4. Provides a probability density for the next $s_{\tau+1} - s_\tau$ periods

$$\begin{aligned} & p\left(\mathbf{y}_{s_\tau+1}, \ldots, \mathbf{y}_{s_{\tau+1}} \mid \mathbf{Y}^o_{s_\tau}, A\right) \\ = & \int_{\Theta_A} p\left(\boldsymbol{\theta}_A \mid \mathbf{Y}^o_{s_\tau}\right) p\left(\mathbf{y}_{s_\tau+1}, \ldots, \mathbf{y}_{s_{\tau+1}} \mid \mathbf{Y}^o_{s_\tau}, \boldsymbol{\theta}_A, A\right) d\boldsymbol{\theta}_A. \end{aligned}$$

This system of updating and probability forecasting in real time was termed *prequential* (a combination of probability forecasting and sequential prediction) by Dawid (1984). Dawid carefully distinguished this process from statistical forecasting systems that do not fully update: for example, using a "plug-in" estimate of $\boldsymbol{\theta}_A$, or using a posterior distribution for $\boldsymbol{\theta}_A$ that does not reflect all of the information available at the time the probability distribution over future events is formed.

Each component of the multiplicative decomposition in (22) is the realized value of the predictive density for the following $s_\tau - s_{\tau-1}$ observations, formed after $s_{\tau-1}$ observations are in hand. In this, well-defined, sense the marginal likelihood incorporates the out-of-sample prediction record of the model A. Equations (16), (18) and (22) make precise the idea that in model averaging, the weight assigned to a model is proportional to the product of its out-of-sample predictive likelihoods.

2.3.3 Posterior predictive distributions

Model combination completes the Bayesian structure of analysis, following the principles of explicit formulation and relevant conditioning set out at the start of this section (p. 2). There are many details in this structure important for forecasting, yet to be described. A principal attraction of the Bayesian structure is its internal logical consistency, a useful and sometimes distinguishing property in applied economic forecasting. But the external consistency of the structure is also critical to successful forecasting: a set of bad models, no matter how consistently applied, will produce bad forecasts. Evaluating external consistency requires that we compare the set of models with unarticulated alternative models. In so doing we step outside the logical structure of Bayesian analysis. This opens up an array of possible procedures, which cannot all be described here. One of the earliest, and still one of the most complete descriptions of these possible procedures is the seminal 1980 paper by Box (1980) that appears with

comments by a score of discussants. For a similar more recent symposium, see Bayarri and Berger (1998) and their discussants.

One of the most useful tools in the evaluation of external consistency is the *posterior predictive distribution.* Its density is similar to the prior predictive density, except that the prior is replaced by the posterior:

$$p\left(\widetilde{\mathbf{Y}}_T \mid \mathbf{Y}_T^o, A\right) = \int_{\Theta_A} p\left(\boldsymbol{\theta}_A \mid \mathbf{Y}_T^o, A\right) p\left(\widetilde{\mathbf{Y}}_T \mid \mathbf{Y}_T^o, \boldsymbol{\theta}_A, A\right) d\boldsymbol{\theta}_A. \tag{23}$$

In this expression $\widetilde{\mathbf{Y}}_T$ is a random vector: the outcomes, given model A and the data $\mathbf{Y}_T^o$, that might have occurred but did not. Somewhat more precisely, if the time series "experiment" could be repeated, (23) would be the predictive density for the outcome of the repeated experiment. Contrasts between $\widetilde{\mathbf{Y}}_T$ and $\mathbf{Y}_T^o$ are the basis of assessing the external validity of the model, or set of models, upon which inference has been conditioned. If one is able to simulate unobservables $\boldsymbol{\theta}_A^{(m)}$ from the posterior distribution (more on this in Section 3) then the simulation $\widetilde{\mathbf{Y}}_T^{(m)}$ follows just as the simulation of $\mathbf{Y}_T^{(m)}$ in (11).

The process can be made formal by identifying one or more subsets S of the range Ψ_T of Y_T. For any such subset $P\left(\widetilde{\mathbf{Y}}_T \in S \mid \mathbf{Y}_T^o, A\right)$ can be evaluated using the simulation approximation $M^{-1}\sum_{m=1}^{M} I_S\left(\widetilde{\mathbf{Y}}_T^{(m)}\right)$. If $P\left(\widetilde{\mathbf{Y}}_T \in S \mid \mathbf{Y}_T^o, A\right) = 1-\alpha$, α being a small positive number, and $\mathbf{Y}_T^o \notin S$, there is evidence of external inconsistency of the model with the data. This idea goes back to the notion of "surprise" discussed by Good (1956): we have observed an event that is very unlikely to occur again, were the time series "experiment" to be repeated, independently, many times. The essentials of this idea were set out by Rubin (1984) in what he termed "model monitoring by posterior predictive checks." As Rubin emphasized, there is no formal method for choosing the set S (see, however, Section 2.4.1 below). If S is defined with reference to a scalar function g as $\left\{\widetilde{\mathbf{Y}}_T : g_1 \le g\left(\widetilde{\mathbf{Y}}_T\right) \le g_2\right\}$ then it is a short step to reporting a "p-value" for $g\left(\mathbf{Y}_T^o\right)$. This idea builds on that of the probability integral transform introduced by Rosenblatt (1952), stressed by Dawid (1984) in prequential forecasting, and formalized by Meng (1994); see also the comprehensive survey of Gelman et al. (1995).

The purpose of posterior predictive exercises of this kind is not to conduct hypothesis tests that lead to rejection or non-rejection of models; rather, it is to provide a diagnostic that may spur creative thinking about new models that might be created and brought into the universe of models $A = \{A_1, \ldots, A_J\}$. This is the idea originally set forth by Box (1980). Not all practitioners agree: see the discussants in the symposia in Box (1980) and Bayarri and Berger (1998), as well as the articles by Edwards et al. (1963) and Berger and Delampady (1987). The creative process dictates the choice of S, or of $g\left(\widetilde{\mathbf{Y}}_T\right)$, which can be quite flexible, and can be selected with an eye to the ultimate application of the model, a subject to which we return in the next section. In general the function $g\left(\widetilde{\mathbf{Y}}_T\right)$ could be a pivotal test statistic (e.g., the difference between the first

order statistic and the sample mean, divided by the sample standard deviation, in an i.i.d. Gaussian model) but in the most interesting and general cases it will not (e.g., the point estimate of a long-memory coefficient). In checking external validity, the method has proven useful and flexible; for example see the recent work by Koop (2001) and Geweke and McCausland (2001) and the texts by Lancaster (2004, Section 2.5) and Geweke (2005, Section 5.3.2). Brav (2000) utilizes posterior predictive analysis in examining alternative forecasting models for long-run returns on financial assets.

2.4 Forecasting

To this point we have considered the generic situation of J competing models relating a common vector of interest $\boldsymbol{\omega}$ to a set of observables $\mathbf{Y}_T$. In forecasting problems $(Y_{T+1}, \ldots, Y_{T+F})' \in \boldsymbol{\omega}$. Sections 2.1 and 2.2 showed how the principle of explicit formulation leads to a recursive representation of the complete probability structure, which we collect here for ease of reference. For each model A_j, a prior model probability $p(A_j)$, a prior density $\mathrm{p}\left(\boldsymbol{\theta}_{A_j} \mid A_j\right)$ for the unobservables $\boldsymbol{\theta}_{A_j}$ in that model, a conditional observables density $p\left(\mathbf{Y}_T \mid \boldsymbol{\theta}_{A_j}, A_j\right)$, and a vector of interest density $p\left(\boldsymbol{\omega} \mid \mathbf{Y}_T, \boldsymbol{\theta}_{A_j}, A_j\right)$ imply

$$\begin{aligned} & p\left\{\left[A_j, \boldsymbol{\theta}_{A_j} \;\; (j = 1, \ldots, J)\right], \mathbf{Y}_T, \boldsymbol{\omega}\right\} \\ = & \sum_{j=1}^{J} p(A_j) \cdot p\left(\boldsymbol{\theta}_{A_j} \mid A_j\right) \cdot p\left(\mathbf{Y}_T \mid \boldsymbol{\theta}_{A_j}, A_j\right) \cdot p\left(\boldsymbol{\omega} \mid \mathbf{Y}_T, \boldsymbol{\theta}_{A_j}, A_j\right). \end{aligned}$$

The entire theory of Bayesian forecasting derives from the application of the principle of relevant conditioning to this probability structure. This leads, in order, to the posterior distribution of the unobservables in each model

$$p\left(\boldsymbol{\theta}_{A_j} \mid \mathbf{Y}_T^o, A_j\right) \propto p\left(\boldsymbol{\theta}_{A_j} \mid A\right) p\left(\mathbf{Y}_T^o \mid \boldsymbol{\theta}_{Aj}, A_j\right) \;\; (j = 1 \ldots, J), \tag{24}$$

the predictive density for the vector of interest in each model

$$p\left(\boldsymbol{\omega} \mid \mathbf{Y}_T^o, A_j\right) = \int_{\Theta_{A_j}} p\left(\boldsymbol{\theta}_{A_j} \mid \mathbf{Y}_T^o, A_j\right) p\left(\boldsymbol{\omega} \mid \mathbf{Y}_T^o, \boldsymbol{\theta}_{A_j}\right) d\boldsymbol{\theta}_{A_j}, \tag{25}$$

posterior model probabilities

$$p\left(A_j \mid \mathbf{Y}_T^o\right) \propto p(A_j) \cdot \int_{\Theta_{A_j}} p\left(\mathbf{Y}_T^o \mid \boldsymbol{\theta}_{A_j}, A_j\right) p\left(\boldsymbol{\theta}_{A_j} \mid A_j\right) d\boldsymbol{\theta}_{A_j} \;\; (j = 1 \ldots, J), \tag{26}$$

and, finally, the predictive density for the vector of interest,

$$p\left(\boldsymbol{\omega} \mid \mathbf{Y}_T^o\right) = \sum_{j=1}^{J} p\left(\boldsymbol{\omega} \mid \mathbf{Y}_T^o, A_j\right) p\left(A_j \mid \mathbf{Y}_T^o\right). \tag{27}$$

The density (25) involves one of the elements of the recursive formulation of the model and consequently, as observed in Section 2.2.2, simulation from

the corresponding distribution is generally straightforward. Expression (27) involves not much more than simple addition. Technical hurdles arise in (24) and (26), and we shall return to a general treatment of these problems using posterior simulators in Section 3. Here we emphasize the incorporation of the final product (27) in forecasting – the decision of what to report about the future. In Sections 2.4.1 and 2.4.2 we focus on (24) and (25), suppressing the model subscripting notation. Section 2.4.3 returns to issues associated with forecasting using combinations of models.

2.4.1 Loss functions and the subjective decision maker

The elements of Bayesian decision theory are isomorphic to those of the classical theory of expected utility in economics. Both Bayesian decision makers and economic agents associate a cardinal measure with all possible combinations of relevant random elements in their environment – both those that they cannot control, and those that they do. The latter are called *actions* in Bayesian decision theory and choices in economics. The mapping to a cardinal measure is a *loss function* in the Bayesian decision theory and a utility function in economics, but except for a change in sign they serve the same purpose. The decision maker takes the *Bayes action* that minimizes the expected value of his loss function; the economic agent makes the choice that maximizes the expected value of her utility function.

In the context of forecasting the relevant elements are those collected in the vector of interest $\boldsymbol{\omega}$, and for a single model the relevant density is (25). The Bayesian formulation is to find an action $\mathbf{a}$ (a vector of real numbers) that minimizes

$$E\left[L\left(\mathbf{a},\boldsymbol{\omega}\right)\mid\mathbf{Y}_{T}^{o},A\right]=\int_{\Omega}\int_{\Theta_{A}}L\left(\mathbf{a},\boldsymbol{\omega}\right)p\left(\boldsymbol{\omega}\mid\mathbf{Y}_{T}^{o},A\right)d\boldsymbol{\omega}. \tag{28}$$

The solution of this problem may be denoted $\mathbf{a}\left(\mathbf{Y}_{T}^{o},A\right)$. For some well-known special cases these solutions take simple forms; see Bernardo and Smith (1994, Section 5.1.5) or Geweke (2005, Section 2.5). If the loss function is quadratic, $L\left(\mathbf{a},\boldsymbol{\omega}\right)=\left(\mathbf{a}-\boldsymbol{\omega}\right)'\mathbf{Q}\left(\mathbf{a}-\boldsymbol{\omega}\right)$, where $\mathbf{Q}$ is a positive definite matrix, then $\mathbf{a}\left(\mathbf{Y}_{T}^{o},A\right)=E\left(\mathbf{a}\mid\mathbf{Y}_{T}^{o},A\right)$; point forecasts that are expected values assume a quadratic loss function. A zero-one loss function takes the form $L\left(\mathbf{a},\boldsymbol{\omega};\varepsilon\right)=1-\int_{N_{\varepsilon}\left(\mathbf{a}\right)}\left(\boldsymbol{\omega}\right)$, where $N_{\varepsilon}\left(\mathbf{a}\right)$ is an open ε-neighborhood of $\mathbf{a}$. Under weak regularity conditions, as $\varepsilon\rightarrow0$, $\mathbf{a}\rightarrow\arg\max_{\boldsymbol{\omega}}p\left(\boldsymbol{\omega}\mid\mathbf{Y}_{T}^{o},A\right)$.

In practical applications asymmetric loss functions can be critical to effective forecasting; for one such application see Section 6.2 below. One example is the linear-linear loss function, defined for scalar ω as

$$L\left(a,\omega\right)=\left(1-q\right)\cdot\left(a-\omega\right)I_{\left(-\infty,a\right)}\left(\omega\right)+q\cdot\left(\omega-a\right)I_{\left(a,\infty\right)}\left(\omega\right), \tag{29}$$

where $q\in\left(0,1\right)$; the solution in this case is $a=P^{-1}\left(q\mid\mathbf{Y}_{T}^{o},A\right)$, the q'th quantile of the predictive distribution of ω. Another is the linear-exponential loss function studied by Zellner (1986):

$$L\left(a,\omega\right)=\exp\left[r\left(a-\omega\right)\right]-r\left(a-\omega\right)-1,$$

where $r \neq 0$; then (28) is minimized by

$$a = -r^{-1} \log \left\{ E \left[\exp \left(-r\omega \right) \right] \mid \mathbf{Y}_T^o, A \right\};$$

if the density (25) is Gaussian, this becomes

$$a = E \left(\omega \mid \mathbf{Y}_T^o, A \right) - (r/2) \, var \left(\omega \mid \mathbf{Y}_T^o, A \right).$$

The extension of both the quantile and linear-exponential loss functions to the case of a vector function of interest $\boldsymbol{\omega}$ is straightforward.

Forecasts of discrete future events also emerge from this paradigm. For example, a business cycle downturn might be defined as $\omega = y_{T+1} < y_T^o > y_{T-1}^o$ for some measure of real economic activity y_t. More generally, any future event may be denoted $\Omega_0 \subseteq \Omega$. Suppose there is no loss given a correct forecast, but loss L_1 in forecasting $\omega \in \Omega_0$ when in fact $\omega \notin \Omega_0$, and loss L_2 in forecasting $\omega \notin \Omega_0$ when in fact $\omega \in \Omega_0$. Then the forecast is $\omega \in \Omega_0$ if

$$\frac{L_1}{L_2} < \frac{P \left(\omega \in \Omega_0 \mid \mathbf{Y}_T^o, A \right)}{P \left(\omega \notin \Omega_0 \mid \mathbf{Y}_T^o, A \right)}$$

and $\omega \notin \Omega_0$ otherwise. For further details on event forecasts and combinations of event forecasts with point forecasts see Zellner et al. (1990).

In simulation-based approaches to Bayesian inference a random sample $\boldsymbol{\omega}^{(m)}$ $(m = 1, \ldots, M)$ represents the density $p \left(\boldsymbol{\omega} \mid \mathbf{Y}_T^o, A \right)$. Shao (1989) showed that

$$\arg\max_{\mathbf{a}} M^{-1} \sum_{m=1}^{M} L \left(\mathbf{a}, \boldsymbol{\omega}^{(m)} \right) \overset{a.s.}{\rightarrow} \arg\max_{\mathbf{a}} E \left[L \left(\mathbf{a}, \boldsymbol{\omega} \right) \mid \mathbf{Y}_T^o, A \right]$$

under weak regularity conditions that serve mainly to assure the existence and uniqueness of $\arg\max_{\mathbf{a}} E \left[L \left(\mathbf{a}, \boldsymbol{\omega} \right) \mid \mathbf{Y}_T^o, A \right]$. This result opens up the scope of tractable loss functions to those that can be optimized for fixed $\boldsymbol{\omega}$.

Once in place, loss functions often suggest candidates for the sets S or functions $g \left(\widetilde{\mathbf{Y}}_T \right)$ used in posterior predictive distributions as described in Section 2.3.3. A generic set of such candidates stems from the observation that a model provides not only the optimal action $\mathbf{a}$, but also the predictive density of $L \left(\mathbf{a}, \boldsymbol{\omega} \right) \mid \left(\mathbf{Y}_T^o, A \right)$ associated with that choice. This density may be compared with the realized outcomes $L \left(\mathbf{a}, \boldsymbol{\omega}^o \right) \mid \left(\mathbf{Y}_T^o, A \right)$. This can be done for one forecast, or for a whole series of forecasts. For example, $\mathbf{a}$ might be the realization of a trading rule designed to minimize expected financial loss, and L the financial loss from the application of the trading rule; see Geweke (1989b) for an early application of this idea to multiple models.

Non-Bayesian formulations of the forecasting decision problem are superficially similar but fundamentally different. In non-Bayesian approaches it is necessary to introduce the assumption that there is a data generating process $f \left(\mathbf{Y}_T \mid \boldsymbol{\theta} \right)$ with a fixed but unknown vector of parameters $\boldsymbol{\theta}$, and a corresponding generating process for the vector of interest $\boldsymbol{\omega}$, $f \left(\boldsymbol{\omega} \mid \mathbf{Y}_T, \boldsymbol{\theta} \right)$. In so doing these approaches condition on unknown quantities, sewing the seeds of internal

logical contradiction that subsequently re-emerge, often in the guise of interesting and challenging problems. The formulation of the forecasting problem, or any other decision-making problem, is then to find a mapping from all possible outcomes $\mathbf{Y}_T$, to actions $\mathbf{a}$, that minimizes

$$E\left\{L\left[\mathbf{a}\left(\mathbf{Y}_T\right),\boldsymbol{\omega}\right]\right\}=\int_{\boldsymbol{\Omega}}\int_{\Psi_T}L\left[\mathbf{a}\left(\mathbf{Y}_T\right),\boldsymbol{\omega}\right]f\left(\mathbf{Y}_T\mid\boldsymbol{\theta}\right)f\left(\boldsymbol{\omega}\mid\mathbf{Y}_T,\boldsymbol{\theta}\right)d\mathbf{Y}_T d\boldsymbol{\omega}. \tag{30}$$

Isolated pedantic examples aside, the solution of this problem invariably involves the unknown $\boldsymbol{\theta}$. The solution of the problem is infeasible because it is ill-posed, assuming that which is unobservable to be known and thereby violating the principle of relevant conditioning. One can replace $\boldsymbol{\theta}$ with an estimator $\widehat{\boldsymbol{\theta}}\left(\mathbf{Y}_T\right)$ in different ways and this, in turn, has led to a substantial literature on an array of procedures. The methods all build upon, rather than address, the logical contradictions inherent in this approach. Geisser (1993) provides an extensive discussion; see especially Section 2.2.2.

2.4.2 Probability forecasting and remote clients

The formulation (24)-(25) is a synopsis of the prequential approach articulated by Dawid (1984). It summarizes all of the uncertainty in the model (or collection of models, if extended to (27)) relevant for forecasting. From these densities remote clients with different loss functions can produce forecasts $\mathbf{a}$. These clients must, of course, share the same collection of (1) prior model probabilities, (2) prior distributions of unobservables, and (3) conditional observables distributions, which is asking quite a lot. However, we shall see in Section 3.3.2 that modern simulation methods allow remote clients some scope in adjusting prior probabilities and distributions without repeating all the work that goes into posterior simulation. That leaves the collection of observables distributions $p\left(\mathbf{Y}_T\mid\boldsymbol{\theta}_{A_j},A_j\right)$ as the important fixed element with which the remote client must work, a constraint common to all approaches to forecasting.

Because of the substantial non-Bayesian literature on probability forecasting and the expression of uncertainty about probability forecasts, it is necessary to emphasize the point that there is no uncertainty about the predictive density $p\left(\boldsymbol{\omega}\mid\mathbf{Y}_T^o\right)$ given the specified collection of models; this is a consequence of consistency with the principle of relevant conditioning. The probability integral transform of the predictive distribution $P\left(\boldsymbol{\omega}\mid\mathbf{Y}_T^o\right)$ provides candidates for posterior predictive analysis. Dawid (1984, Section 5.3) pointed out that not only is the marginal distribution of $P^{-1}\left(\boldsymbol{\omega}\mid\mathbf{Y}_T\right)$ uniform on $(0,1)$, but in a prequential updating setting of the kind described in Section 2.3.2 these outcomes are also i.i.d. This leads to a wide variety of functions $g\left(\widetilde{\mathbf{Y}}_T\right)$ that might be used in posterior predictive analysis. (Kling (1987) and Kling and Bessler (1989) applied this idea in their assessment of vector autoregression models.) Some further possibilities were discussed in recent work by Christoffersen (1998) that addressed interval forecasts; see also Chatfield (1993).

Non-Bayesian probability forecasting addresses a superficially similar but

fundamentally different problem, that of estimating the predictive density inherent in the data generating process, $f(\boldsymbol{\omega} \mid \mathbf{Y}_T^o, \boldsymbol{\theta})$. The formulation of the problem in this approach is to find a mapping from all possible outcomes $\mathbf{Y}_T$ into functions $p(\boldsymbol{\omega} \mid \mathbf{Y}_T)$ that minimizes

$$\begin{aligned} & E\left\{L\left[p\left(\boldsymbol{\omega} \mid \mathbf{Y}_T\right), f\left(\boldsymbol{\omega} \mid \mathbf{Y}_T, \boldsymbol{\theta}\right)\right]\right\} \\ = & \int_{\Omega} \int_{\Psi_T} L\left[p\left(\boldsymbol{\omega} \mid \mathbf{Y}_T\right), f\left(\boldsymbol{\omega} \mid \mathbf{Y}_T, \boldsymbol{\theta}\right)\right] \\ & \cdot f\left(\mathbf{Y}_T \mid \boldsymbol{\theta}\right) f\left(\boldsymbol{\omega} \mid \mathbf{Y}_T, \boldsymbol{\theta}\right) d\mathbf{Y}_T d\boldsymbol{\omega}. \end{aligned} \tag{31}$$

In contrast with the predictive density, the minimization problem (31) requires a loss function, and different loss functions will lead to different solutions, other things the same, as emphasized by Weiss (1996).

The problem (31) is a special case of the frequentist formulation of the forecasting problem described at the end of Section 2.4.1. As such, it inherits the internal inconsistencies of this approach, often appearing as challenging problems. In their recent survey of density forecasting using this approach Tay and Wallis (2000, p. 248) pinpointed the challenge, if not its source: "While a density forecast can be seen as an acknowledgement of the uncertainty in a point forecast, it is itself uncertain, and this second level of uncertainty is of more than casual interest if the density forecast is the direct object of attention ... How this might be described and reported is beginning to receive attention."

2.4.3 Forecasts from a combination of models

The question of how to forecast given alternative models available for the purpose is a long and well-established one. It dates at least to the 1963 work of Barnard (1963) in a paper that studied airline data. This was followed by a series of influential papers by Granger and coauthors (Bates and Granger (1969), Granger and Ramanathan (1984), Granger (1989)); Clemen (1989) provides a review of work before 1990. The papers in this and the subsequent forecast combination literature all addressed the question of how to produce a superior forecast given competing alternatives. The answer turns in large part on what is available. Producing a superior forecast, given only competing point forecasts, is distinct from the problem of aggregating the information that produced the competing alternatives (see Granger and Ramanathan (1984, p. 198)) and Granger (1989, pp. 168-169)). A related, but distinct, problem is that of combining probability distributions from different and possibly dependent sources, taken up in a seminal paper by Winkler (1981).

In the context of Section 2.3, forecasting from a combination of models is straightforward. The vector of interest $\boldsymbol{\omega}$ includes the relevant future observables $(\mathbf{Y}_{T+1}, \ldots, \mathbf{Y}_{T+F})$, and the relevant forecasting density is (16). Since the minimand $E\left[L\left(\mathbf{a}, \boldsymbol{\omega}\right) \mid \mathbf{Y}_T^o, A\right]$ in (28) is defined with respect to this distribution, there is no substantive change. Thus the combination of models leads to a single predictive density, which is a weighted average of the predictive densities of the individual models, the weights being proportional to the posterior probabilities of those models. This predictive density conveys all uncertainty about

$\boldsymbol{\omega}$, conditional on the collection of models and the data, and point forecasts and other actions derive from the use of a loss function in conjunction with it.

The literature acting on this paradigm has emerged rather slowly, for two reasons. One has to do with computational demands, now largely resolved and discussed in the next section; Draper (1995) provides an interesting summary and perspective on this aspect of prediction using combinations of models, along with some applications. The other is that the principle of explicit formulation demands not just point forecasts of competing models, but rather (1) their entire predictive densities $p(\boldsymbol{\omega} \mid \mathbf{Y}_T^o, A_j)$ and (2) their marginal likelihoods. Interestingly, given the results in Section 2.3.2, the latter requirement is equivalent to a record of the one-step-ahead predictive likelihoods $p\left(\mathbf{y}_t^o \mid \mathbf{Y}_{t-1}^o, A_j\right)$ $(t = 1, \ldots, T)$ for each model. It is therefore not surprising that most of the prediction work based on model combination has been undertaken using models also designed by the combiners. The feasibility of this approach was demonstrated by Zellner and coauthors (Zellner and Palm (1992), Min and Zellner (1993)) using purely analytical methods. Petridis et al. (2001) provide a successful forecasting application utilizing a combination of heterogeneous data and Bayesian model averaging.

2.4.4 Conditional forecasting

In some circumstances, selected elements of the vector of future values of $\mathbf{y}$ may be known, making the problem one of conditional forecasting. That is, restricting attention to the vector of interest $\boldsymbol{\omega} = (\mathbf{y}_{T+1}, \ldots, \mathbf{y}_{T+F})'$, one may wish to draw inferences regarding $\boldsymbol{\omega}$ treating $\left(S_1\mathbf{y}'_{T+1}, \ldots, S_F\mathbf{y}'_{T+F}\right) \equiv \mathbf{S}\boldsymbol{\omega}$ as known for $q \times p$ "selection" matrices $(S_1, \ldots, S_F)$, which could select elements or linear combinations of elements of future values. The simplest such situation arises when one or more of the elements of $\mathbf{y}$ become known before the others, perhaps because of staggered data releases. More generally, it may be desirable to make forecasts of some elements of $\mathbf{y}$ given views that others follow particular time paths as a way of summarizing features of the joint predictive distribution for $(\mathbf{y}_{T+1}, \ldots, \mathbf{y}_{T+F})$.

In this case, focusing on a single model, A, (25) becomes

$$p(\boldsymbol{\omega} \mid \mathbf{S}\boldsymbol{\omega}, \mathbf{Y}_T^o, A) = \int_{\Theta_A} p(\boldsymbol{\theta}_A \mid \mathbf{S}\boldsymbol{\omega}, \mathbf{Y}_T^o, A)\, p(\boldsymbol{\omega} \mid \mathbf{S}\boldsymbol{\omega}, \mathbf{Y}_T^o, \boldsymbol{\theta}_A)\, d\boldsymbol{\theta}_A \qquad (32)$$

As noted by Waggoner and Zha (1999), this expression makes clear that the conditional predictive density derives from the *joint* density of $\boldsymbol{\theta}_A$ and $\boldsymbol{\omega}$. Thus it is not sufficient, for example, merely to know the conditional predictive density $p(\boldsymbol{\omega} \mid \mathbf{Y}_T^o, \boldsymbol{\theta}_A)$, because the pattern of evolution of $(\mathbf{y}_{T+1}, \ldots, \mathbf{y}_{T+F})$ carries information about which $\boldsymbol{\theta}_A$ are likely, and vice versa.

Prior to the advent of fast posterior simulators, Doan, Litterman, Sims (1984) produced a type of conditional forecast from a Gaussian vector autoregression (see (3)) by working directly with the mean of $p\left(\boldsymbol{\omega} \mid \mathbf{S}\boldsymbol{\omega}, \mathbf{Y}_T^o, \bar{\boldsymbol{\theta}}_A\right)$, where $\bar{\boldsymbol{\theta}}_A$ is the posterior mean of $p(\boldsymbol{\theta}_A \mid \mathbf{Y}_T^o, A)$. The former can be obtained

as the solution of a simple least squares problem. This procedure of course ignores the uncertainty in $\boldsymbol{\theta}_A$.

More recently, Waggoner and Zha (1999) developed two procedures for calculating conditional forecasts from VARs according to whether the conditions are regarded as "hard" or "soft". Under "hard" conditioning, $\mathbf{S}\boldsymbol{\omega}$ is treated as known, and (32) must be evaluated. Waggoner and Zha (1999) develop a Gibbs sampling procedure to do so. Under "soft" conditioning, $\mathbf{S}\boldsymbol{\omega}$ is regarded as lying in a pre-specified interval, which makes it possible to work directly with the unconditional predictive density (25), obtaining a sample of $\mathbf{S}\boldsymbol{\omega}$ in the appropriate interval by simply discarding those samples $\mathbf{S}\boldsymbol{\omega}$ which do not. The advantage to this procedure is that (25) is generally straightforward to obtain, whereas $p\left(\boldsymbol{\omega} \mid \mathbf{S}\boldsymbol{\omega}, \mathbf{Y}_T^o, \boldsymbol{\theta}_A\right)$ may not be.

Robertson, Tallman, and Whiteman (2004) provide an alternative to these conditioning procedures by approximating the relevant conditional densities. They specify the conditioning information as a set of moment conditions (e.g., $E\mathbf{S}\boldsymbol{\omega} = \hat{\boldsymbol{\omega}}_{\mathbf{S}}$; $E(\mathbf{S}\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}_{\mathbf{S}})(\mathbf{S}\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}_{\mathbf{S}})' = \mathbf{V}_\omega$), and work with the density (i) that is closest to the unconditional in the information-theoretic sense and that also (ii) satisfies the specified moment conditions. Given a sample $\{\boldsymbol{\omega}^{(m)}\}$ from the unconditional predictive, the new, minimum-relative-entropy density is straightforward to calculate; the original density serves as an importance sampler for the conditional. Cogley, Morozov, and Sargent (2004) have utilized this procedure in producing inflation forecast fan charts from a time-varying parameter VAR.

3 Posterior simulation methods

The principle of relevant conditioning in Bayesian inference requires that one be able to access the posterior distribution of the vector of interest $\boldsymbol{\omega}$ in one or more models. In all but simple illustrative cases this cannot be done analytically. A posterior simulator yields a pseudo-random sequence $\left\{\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(M)}\right\}$ that can be used to approximate posterior moments of the form $E\left[h\left(\boldsymbol{\omega}\right) \mid \mathbf{Y}_T^o, A\right]$ arbitrarily well: the larger is M, the better is the approximation. Taken together, these algorithms are known generically as posterior simulation methods. While the motivating task, here, is to provide a simulation representative of $p\left(\boldsymbol{\omega} \mid \mathbf{Y}_T^o, A\right)$, this section will suppress the conditioning, in most cases, and work with the density $p\left(\boldsymbol{\theta}\right)$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$, and $p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}\right)$, $\boldsymbol{\omega} \in \boldsymbol{\Omega} \subseteq \mathbb{R}^q$, in order to simplify notation. Consistent with the motivating problem, we shall assume that there is no difficulty in drawing $\boldsymbol{\omega}^{(m)} \overset{iid}{\sim} p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}\right)$.

The methods described in this section all utilize as building blocks the set of distributions from which it is possible to produce pseudo-i.i.d. sequences of random variables or vectors. We shall refer to such distributions as conventional distributions. This set includes, of course, all of those found in standard mathematical applications software. There is a grey area beyond these distributions; examples include the Dirichlet (or multivariate beta) and Wishart distributions. What is most important, in this context, is that posterior distributions in all

but the simplest models lead almost immediately to distributions from which it is effectively impossible to produce pseudo-i.i.d. sequences of random vectors. It is to these distributions that the methods discussed in this section are addressed.

3.1 Simulation methods before 1990

The applications of simulation methods in statistics and econometrics before 1990, including Bayesian inference, were limited to sequences of independent and identically distributed random vectors. The state of the art by the mid-1960s is well summarized in Hammsersly and Handscomb (1964) and the early impact of these methods in Bayesian econometrics is evident in Zellner (1971). A survey of progress as of the end of this period is Geweke (1991) written at the dawn of the application of Markov chain Monte Carlo (MCMC) methods in Bayesian statistics.[1] Since 1990 MCMC methods have largely supplanted i.i.d. simulation methods. MCMC methods, in turn, typically combine several simulation methods, and those developed before 1990 are important constituents in MCMC.

3.1.1 Direct sampling

In direct sampling $\boldsymbol{\theta}^{(m)} \overset{iid}{\sim} p(\boldsymbol{\theta})$. If $\boldsymbol{\omega}^{(m)} \sim p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}^{(m)}\right)$ is a conditionally independent sequence, then $\left\{\boldsymbol{\theta}^{(m)}, \boldsymbol{\omega}^{(m)}\right\} \overset{i.i.d.}{\sim} p(\boldsymbol{\theta})\, p(\boldsymbol{\omega} \mid \boldsymbol{\theta})$. Then for any existing moment $E[h(\boldsymbol{\theta}, \boldsymbol{\omega})]$, $M^{-1} \sum_{m=1}^{M} h\left(\boldsymbol{\theta}^{(m)}, \boldsymbol{\omega}^{(m)}\right) \overset{a.s.}{\to} E[h(\boldsymbol{\theta}, \boldsymbol{\omega})]$; this property, for any simulator, is widely termed *simulation-consistency*. An entirely conventional application of the Lindberg-Levy central limit theorem provides a basis of assessing the accuracy of the approximation. The conventional distributions p from which direct sampling is possible coincide, more or less, with those for which a fully analytical treatment of Bayesian inference and forecasting is possible. An excellent example is the fully Bayesian and entirely analytical solution of the problem of forecasting turning points by Min and Zellner (1993).

The Min-Zellner treatment addresses only one-step-ahead forecasting. Forecasting successive steps ahead entails increasingly nonlinear functions that rapidly become intractable in a purely analytical approach. This problem was taken up in Geweke (1988) for multiple-step-ahead forecasts in a bivariate Gaussian autoregression with a conjugate prior distribution. The posterior distribution, like the prior, is normal-gamma. Forecasts F steps ahead based on a quadratic loss function entail linear combinations of posterior moments of order F from a multivariate Student-t distribution. This problem plays to the comparative advantage of direct sampling in the determination of posterior expectations

[1] Ironically, MCMC methods were initially developed in the late 1940's in one of the first applications of simulation methods using electronic computers, to the design of thermonuclear weapons (see Metropolis et al. (1953)). Perhaps not surprisingly, they spread first to disciplines with the greatest access to computing power: see the application to image restoratoin by Geman and Geman (1984).

of nonlinear functions of random variables with conventional distributions. It nicely illustrates two variants on direct sampling that can dramatically increase the speed and accuracy of posterior simulation approximations.

1. The first variant is motivated by the fact that the conditional mean of the F-step ahead realization of $\mathbf{y}_t$ is a deterministic function of the parameters. Thus, the function of interest $\boldsymbol{\omega}$ is taken to be this mean, rather than a simulated realization of $\mathbf{y}_t$.

2. The second variant exploits the fact that the posterior distribution of the variance matrix of the disturbances (denoted $\boldsymbol{\theta}_2$, say) in this model is inverted Wishart, and the conditional distribution of the coefficients ($\boldsymbol{\theta}_1$, say) is Gaussian. Corresponding to the generated sequence $\boldsymbol{\theta}_1^{(m)}$, consider also $\widetilde{\boldsymbol{\theta}}_1^{(m)} = 2E\left(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^{(m)}\right) - \boldsymbol{\theta}_1^{(m)}$. Both $\boldsymbol{\theta}^{(m)\prime} = \left(\boldsymbol{\theta}_1^{(m)\prime}, \boldsymbol{\theta}_2^{(m)\prime}\right)$ and $\widetilde{\boldsymbol{\theta}}^{(m)\prime} = \left(\widetilde{\boldsymbol{\theta}}_1^{(m)\prime}, \boldsymbol{\theta}_2^{(m)\prime}\right)$ are i.i.d. sequences drawn from $p(\boldsymbol{\theta})$. Take $\boldsymbol{\omega}^{(m)} \sim p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}^{(m)}\right)$ and $\widetilde{\boldsymbol{\omega}}^{(m)} \sim p\left(\boldsymbol{\omega} \mid \widetilde{\boldsymbol{\theta}}^{(m)}\right)$. (In the forecasting application of Geweke (1988) these latter distributions are deterministic functions of $\boldsymbol{\theta}^{(m)}$ and $\widetilde{\boldsymbol{\theta}}^{(m)}$.) The sequences $h\left(\boldsymbol{\omega}^{(m)}\right)$ and $h\left(\widetilde{\boldsymbol{\omega}}^{(m)}\right)$ will also be i.i.d. and, depending on the nature of the function h, may be negatively correlated because $cov\left(\boldsymbol{\theta}_1^{(m)}, \widetilde{\boldsymbol{\theta}}_1^{(m)}\right) = -var\left(\boldsymbol{\theta}_1^{(m)}\right) = -var\left(\widetilde{\boldsymbol{\theta}}_1^{(m)}\right)$. In many cases the approximation error using $(2M)^{-1}\sum_{m=1}^{M}\left[h\left(\boldsymbol{\omega}^{(m)}\right) + h\left(\widetilde{\boldsymbol{\omega}}^{(m)}\right)\right]$ may be much smaller than that using $M^{-1}\sum_{m=1}^{M} h\left(\boldsymbol{\omega}^{(m)}\right)$.

The second variant is an application of antithetic sampling, an idea well established in the simulation literature (see Hamersly and Morton (1956) and Geweke (1996, Section 5.1)). In the posterior simulator application just described, given weak regularity conditions and for a given function h, the sequences $h\left(\boldsymbol{\omega}^{(m)}\right)$ and $h\left(\widetilde{\boldsymbol{\omega}}^{(m)}\right)$ become more negatively correlated as sample size increases (see Geweke (1988, Theorem 1)); hence the term *antithetic acceleration*. The first variant has acquired the monicker *Rao-Blackwellization* in the posterior simulation literature, from the Rao-Blackwell Theorem, which establishes $var\left[E\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}\right)\right] \leq var\left(\boldsymbol{\omega}\right)$. Of course the two methods can be used separately. For one-step ahead forecasts, the combination of the two methods drives the variance of the simulation approximation to zero; this is a close reflection of the symmetry and analytical tractability exploited in Min and Zellner (1993). For near-term forecasts the methods reduce variance by more than 99% in the illustration taken up in Geweke (1988); as the forecasting horizon increases the reduction dissipates, due to the increasing nonlinearity of h.

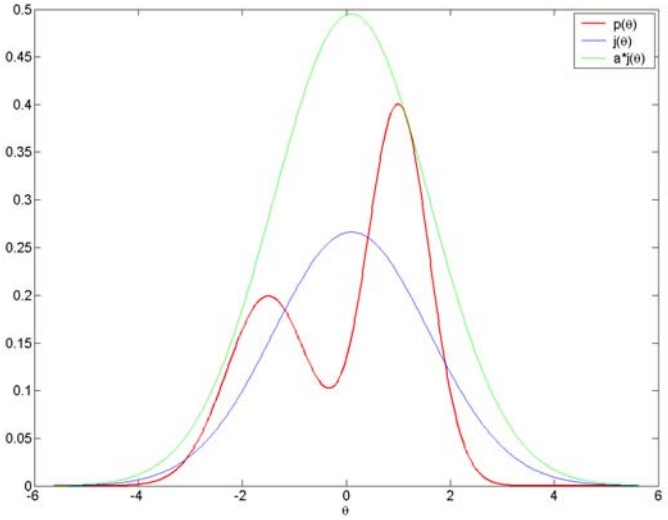

Figure 1: Acceptance sampling

3.1.2 Acceptance sampling

Acceptance sampling relies on a conventional source density $q(\boldsymbol{\theta})$ that approximates $p(\boldsymbol{\theta})$, and then exploits an acceptance-rejection procedure to reconcile the approximation. The method yields a sequence $\boldsymbol{\theta}^{(m)} \overset{iid}{\sim} p(\boldsymbol{\theta})$; as such, it renders the density p conventional, and in fact acceptance sampling is the "black box" that produces pseudo-random variables in most mathematical applications software; for a review see Geweke (1996).

Figure 1 provides the intuition of acceptance sampling. The heavy curve is the target density $p(\theta)$, and the lower bell-shaped curve is the source density $q(\theta)$. The ratio $p(\theta)/q(\theta)$ is bounded above by a constant a. In Figure 1, $p(1.16)/q(1.16) = a = 1.86$, and the lightest curve is $a \cdot q(\theta)$. The idea is to draw θ^* from the source density, which has kernel $a \cdot q(\theta^*)$, but to accept the draw with probability $p(\theta^*)/a \cdot q(\theta^*)$. For example if $\theta^* = 0$, then the draw is accepted with probability 0.269, whereas if $\theta^* = 1.16$ then the draw is accepted with probability 1. The accepted values in fact simulate i.i.d. drawings from the target density $p(\theta)$.

While Figure 1 is necessarily drawn for scalar θ it should be clear that the principle applies for vector $\boldsymbol{\theta}$ of any finite order. In fact this algorithm can be implemented using a kernel $k_p(\boldsymbol{\theta})$ of the density $p(\boldsymbol{\theta})$ i.e., $k_p(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta})$, and this can be important in applications where the constant of integration is not known. Similarly we require only a kernel $k_q(\boldsymbol{\theta})$ of $q(\boldsymbol{\theta})$, and let $a_k = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} k_p(\boldsymbol{\theta})/k_q(\boldsymbol{\theta})$. Then for each draw m the algorithm works as follows.

1. Draw u uniform on $[0,1]$.
2. Draw $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta})$.
3. If $u > k_p(\boldsymbol{\theta}^*)/a_k k_q(\boldsymbol{\theta}^*)$ return to step 1.
4. Set $\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^*$.

To see why the algorithm works, let Θ^* denote the support of q; $a < \infty$ implies $\Theta \subseteq \Theta^*$. Let $c_p = k_p/p$ and $c_q = k_q/q$. The unconditional probability of proceeding from step 3 to step 4 is

$$\int_{\boldsymbol{\Theta}^*} \{k_p(\boldsymbol{\theta}) / [a_k k_q(\boldsymbol{\theta})]\} q(\boldsymbol{\theta}) \, d\boldsymbol{\theta} = c_p/a_k c_q. \tag{33}$$

Let A be any subset of Θ. The unconditional probability of proceeding from step 3 to step 4 with $\boldsymbol{\theta} \in A$ is

$$\int_A \{k_p(\boldsymbol{\theta}) / [a_k k_q(\boldsymbol{\theta})]\} q(\boldsymbol{\theta}) \, d\boldsymbol{\theta} = \int_A k_p(\boldsymbol{\theta}) \, d\boldsymbol{\theta}/a_k c_q. \tag{34}$$

The probability that $\boldsymbol{\theta} \in A$, conditional on proceeding from step 3 to step 4, is the ratio of (34) to (33), which is $\int_A k_p(\boldsymbol{\theta}) \, d\boldsymbol{\theta}/c_p = \int_A p(\boldsymbol{\theta}) \, d\boldsymbol{\theta}$.

Regardless of the choices of kernels the unconditional probability in (33) is $c_p/a_k c_q = \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} q(\boldsymbol{\theta})/p(\boldsymbol{\theta})$. If one wishes to generate M draws of $\boldsymbol{\theta}$ using acceptance sampling, the expected number of times one will have to draw u, draw $\boldsymbol{\theta}^*$, and compute $k_p(\boldsymbol{\theta}^*) / [a_k k_q(\boldsymbol{\theta}^*)]$ is $M \cdot \sup_{\boldsymbol{\theta} \in \Theta} p(\boldsymbol{\theta})/q(\boldsymbol{\theta})$. The computational efficiency of the algorithm is driven by those $\boldsymbol{\theta}$ for which $q(\boldsymbol{\theta})$ has the greatest relative undersampling. In most applications the time consuming part of the algorithm is the evaluation of the kernels $k_q(\boldsymbol{\theta})$ and $k_p(\boldsymbol{\theta})$, especially the latter. (If $p(\boldsymbol{\theta})$ is a posterior density, then evaluation of $k_p(\boldsymbol{\theta})$ entails computing the likelihood function.) In such cases this is indeed the relevant measure of efficiency.

Since $\boldsymbol{\theta}^{(m)} \overset{iid}{\sim} p(\boldsymbol{\theta})$, $\boldsymbol{\omega}^{(m)} \overset{iid}{\sim} p(\boldsymbol{\omega}) = \int_\Theta p(\boldsymbol{\theta}) p(\boldsymbol{\omega} \mid \boldsymbol{\theta}) \, d\boldsymbol{\theta}$. Acceptance sampling is limited by the difficulty in finding an approximation q that is efficient, in the sense just described, and by the need to find $a_k = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} k_p(\boldsymbol{\theta}) / k_q(\boldsymbol{\theta})$. While it is difficult to generalize, these tasks are typically more difficult the greater the number of elements of $\boldsymbol{\theta}$.

3.1.3 Importance sampling

Rather than accept only a fraction of the draws from the source density, it is possible to retain all of them, and consistently approximate the posterior moment by appropriately weighting the draws. The probability density function of the source distribution is then called the *importance sampling density*, a term due to Hammersly and Handscomb (1964), who were among the first to propose the method. It appears to have been introduced to the econometrics literature by Kloek and van Dijk (1978).

To describe the method, denote the source density by $q(\boldsymbol{\theta})$ with support Θ^*, and an arbitrary kernel of the source density by $k_q(\boldsymbol{\theta}) = c_q \cdot q(\boldsymbol{\theta})$ for any $c_q \neq 0$. Denote an arbitrary kernel of the target density by $k_p(\boldsymbol{\theta}) = c_p \cdot p(\boldsymbol{\theta})$ for any $c_p \neq 0$, the i.i.d. sequence $\boldsymbol{\theta}^{(m)} \sim q(\boldsymbol{\theta})$, and the i.i.d. sequence $\boldsymbol{\omega}^{(m)} \mid \boldsymbol{\theta}^{(m)} \sim p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}^{(m)}\right)$. Define the weighting function $w(\boldsymbol{\theta}) = k_p(\boldsymbol{\theta}) / k_q(\boldsymbol{\theta})$. Then the approximation of $\overline{h} = E[h(\boldsymbol{\omega})]$ is

$$\overline{h}^{(M)} = \frac{\sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right) h\left(\boldsymbol{\omega}^{(m)}\right)}{\sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right)}. \tag{35}$$

Geweke (1989a) showed that if $E[h(\boldsymbol{\omega})]$ exists and is finite, and $\Theta^* \supseteq \Theta$, then $\overline{h}^{(M)} \overset{a.s.}{\to} \overline{h}$. Moreover if $var[h(\boldsymbol{\omega})]$ exists and is finite, and if $w(\boldsymbol{\theta})$ is bounded above on Θ, then the accuracy of the approximation can be assessed using the Lindberg-Levy central limit theorem with an appropriately approximated variance (see Geweke (1989a, Theorem 2)). In applications of importance sampling, this accuracy can be summarized in terms of the *numerical standard error* of $\overline{h}^{(M)}$, its sampling standard deviation in independent runs of length M of the importance sampling simulation, and in terms of the *relative numerical efficiency* of $\overline{h}^{(M)}$, the ratio of simulation size in a hypothetical direct simulator to that required using importance sampling to achieve the same numerical standard error. These summaries of accuracy can be used with other simulation methods as well, including the Markov chain Monte Carlo algorithms described in Section 3.2.

To see why importance sampling produces a simulation-consistent approximation of $E[h(\boldsymbol{\omega})]$, let $E_q(\cdot)$ and $var_q(\cdot)$ denote mean and variance, respectively, with respect to the density $q(\boldsymbol{\theta})\, p(\boldsymbol{\omega} \mid \boldsymbol{\theta})$. Then

$$E_q[w(\boldsymbol{\theta})] = \int_{\Theta} \frac{k_p(\boldsymbol{\theta})}{k_q(\boldsymbol{\theta})} q(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \frac{c_p}{c_q} \equiv \overline{w}.$$

Since $\left\{\boldsymbol{\omega}^{(m)}\right\}$ is i.i.d. the strong law of large numbers implies

$$M^{-1} \sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right) \overset{a.s.}{\to} \overline{w}. \tag{36}$$

The sequence $\left\{w\left(\boldsymbol{\theta}^{(m)}\right),\ h\left(\boldsymbol{\omega}^{(m)}\right)\right\}$ is also i.i.d., and

$$\begin{aligned} E_q[w(\boldsymbol{\theta})\, h(\boldsymbol{\omega})] &= \int_{\Theta} w(\boldsymbol{\theta}) \left[\int_{\Omega} h(\boldsymbol{\omega})\, p(\boldsymbol{\omega} \mid \boldsymbol{\theta})\, d\boldsymbol{\omega}\right] q(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= (c_p/c_q) \int_{\Theta} \int_{\Omega} h(\boldsymbol{\omega})\, p(\boldsymbol{\omega} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\omega} d\boldsymbol{\theta} \\ &= (c_p/c_q)\, E[h(\boldsymbol{\omega})] = \overline{w} \cdot \overline{h}. \end{aligned}$$

By the strong law of large numbers,

$$M^{-1}\sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right) h\left(\boldsymbol{\omega}^{(m)}\right) \overset{a.s.}{\to} \overline{w}\cdot\overline{h}. \tag{37}$$

The fraction in (35) is the ratio of the left side of (37) to the left side of (36).

An attraction of importance sampling is that it requires only that p/q be bounded, whereas acceptance sampling requires that the supremum of this ratio (or that for kernels of the densities) be known. Moreover the known supremum is required in order to implement acceptance sampling, whereas the boundedness of p/q is utilized in importance sampling only to exploit a central limit theorem to assess numerical accuracy. An important application of importance sampling is in providing remote clients with a simple way to revise prior distributions, as discussed below in Section 3.3.2.

3.2 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are generalizations of direct sampling. The idea is to construct a Markov chain $\left\{\boldsymbol{\theta}^{(m)}\right\}$ with continuous state space Θ and unique invariant probability density $p\left(\boldsymbol{\theta}\right)$. Following an initial transient or *burn-in* phase, the distribution of $\boldsymbol{\theta}^{(m)}$ is approximately that of the density $p\left(\boldsymbol{\theta}\right)$. The exact sense in which this approximation holds is important. We shall touch on this only briefly; for full detail and references see Geweke (2005, Section 3.5). We continue to assume that $\boldsymbol{\omega}$ can be simulated directly from $p\left(\boldsymbol{\omega}\mid\boldsymbol{\theta}\right)$, so that given $\left\{\boldsymbol{\theta}^{(m)}\right\}$ the corresponding $\boldsymbol{\omega}^{(m)} \sim p\left(\boldsymbol{\omega}\mid\boldsymbol{\theta}^{(m)}\right)$ can be drawn.

Markov chain methods have a history in mathematical physics dating back to the algorithm of Metropolis et al. (1953). This method, which was described subsequently in Hammersly and Handscomb (1964, Section 9.3) and Ripley (1987, Section 4.7), was generalized by Hastings (1970), who focused on statistical problems, and was further explored by Peskun (1973). A version particularly suited to image reconstruction and problems in spatial statistics was introduced by Geman and Geman (1984). This was subsequently shown to have great potential for Bayesian computation by Gelfand and Smith (1990). Their work, combined with data augmentation methods (see Tanner and Wong (1987)) has proven very successful in the treatment of latent variables in econometrics. Since 1990 application of MCMC methods has grown rapidly: new refinements, extensions, and applications appear constantly. Accessible introductions are Gelman et al. (1995) and Geweke (2005); a good collection of applications is Gilks et al. (1996). Section 5 provides several applications of MCMC methods in Bayesian forecasting models.

3.2.1 The Gibbs sampler

Most posterior densities $p\left(\boldsymbol{\theta}_A\mid\mathbf{Y}_T^o, A\right)$ do not correspond to any conventional family of distributions. On the other hand, the conditional distributions of

subvectors of $\boldsymbol{\theta}_A$ often do, which is to say that the conditional posterior distributions of these subvectors are conventional. This is partially the case in the stochastic volatility model described in Section 2.1.2. If, for example, the prior distribution of ϕ is truncated Gaussian and those of β^2 and σ_η^2 are inverted gamma, then the conditional posterior distribution of ϕ is truncated normal and those of β^2 and σ_η^2 are inverted gamma. (The conditional posterior distributions of the latent volatilities h_t are unconventional, and we return to this matter in Section 5.5.)

This motivates the simplest setting for the Gibbs sampler. Suppose $\boldsymbol{\theta}' = \left(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'\right)$ has density $p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right)$ of unconventional form, but that the conditional densities $p\left(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2\right)$ and $p\left(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1\right)$ are conventional. Suppose (hypothetically) that one had access to an initial drawing $\boldsymbol{\theta}_2^{(0)}$ taken from $p\left(\boldsymbol{\theta}_2\right)$, the marginal density of $\boldsymbol{\theta}_2$. Then after iterations $\boldsymbol{\theta}_1^{(m)} \sim p\left(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^{(m-1)}\right)$, $\boldsymbol{\theta}_2^{(m)} \sim p\left(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{(m)}\right)$ $(m = 1, \ldots, M)$ it would be the case that $\boldsymbol{\theta}^{(m)} = \left(\boldsymbol{\theta}_1'^{(m)}, \boldsymbol{\theta}_2'^{(m)}\right)' \sim p\left(\boldsymbol{\theta}\right)$. The extension of this idea to more than two components of $\boldsymbol{\theta}$, given a *blocking* $\boldsymbol{\theta}' = \left(\boldsymbol{\theta}_1', \ldots \boldsymbol{\theta}_B'\right)$ and an initial $\boldsymbol{\theta}^{(0)} \sim p\left(\boldsymbol{\theta}\right)$, is immediate, cycling through

$$\boldsymbol{\theta}_b^{(m)} \sim p\left[\boldsymbol{\theta}^{(b)} \mid \boldsymbol{\theta}_a^{(m)}\left(a < b\right), \boldsymbol{\theta}_a^{(m-1)}\left(a > b\right)\right]\left(b = 1, \ldots, B;\, m = 1, 2, \ldots\right). \tag{38}$$

Of course, if it were possible to make an initial draw from the density p, then independent draws directly from p would also be possible. The purpose of that assumption here is to marshal an informal argument that the density $p\left(\boldsymbol{\theta}\right)$ is an invariant density of this Markov chain: that is, if $\boldsymbol{\theta}^{(m)} \sim p\left(\boldsymbol{\theta}\right)$, then $\boldsymbol{\theta}^{(m+s)} \sim p\left(\boldsymbol{\theta}\right)$ for all $s > 0$.

It is important to elucidate conditions for $\boldsymbol{\theta}^{(m)}$ to converge in distribution to $p\left(\boldsymbol{\theta}\right)$ given any $\boldsymbol{\theta}^{(0)} \in \Theta$. Note that even if $\boldsymbol{\theta}^{(0)}$ were drawn from p, the argument just given demonstrates only that any single $\boldsymbol{\theta}^{(m)}$ is also drawn from p. It does not establish that a single sequence $\left\{\boldsymbol{\theta}^{(m)}\right\}$ is representative of p. Consider the example shown in Figure 2(a), in which $\Theta = \Theta_1 \bigcup \Theta_2$, and the Gibbs sampling algorithm has blocks $\boldsymbol{\theta}_{(1)} = \theta_1$ and $\boldsymbol{\theta}_{(2)} = \theta_2$. If $\boldsymbol{\theta}^{(0)} \in \Theta_1$, then $\boldsymbol{\theta}^{(m)} \in \Theta_1$ for $m = 1, 2, \ldots$. Any single $\boldsymbol{\theta}^{(m)}$ is just as representative of p as is the single drawing $\boldsymbol{\theta}^{(0)}$, but the same cannot be said of the collection $\left\{\boldsymbol{\theta}^{(m)}\right\}$. Indeed, $\left\{\boldsymbol{\theta}^{(m)}\right\}$ could be highly misleading. In the example shown in Figure 2(b), if $\boldsymbol{\theta}^{(0)}$ is the indicated point at the lower left vertex of the triangular closed support of $p\left(\boldsymbol{\theta}\right)$, then $\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(0)}\ \forall\ m$. What is required is that the Markov chain $\left\{\boldsymbol{\theta}^{(m)}\right\}$ be ergodic. That is, if $\boldsymbol{\omega}^{(m)} \sim p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}\right)$ and $E\left[h\left(\boldsymbol{\theta}, \boldsymbol{\omega}\right)\right]$ exists, then we require $M^{-1}\sum_{m=1}^{M} \overset{a.s.}{\to} E\left[h\left(\boldsymbol{\theta}, \boldsymbol{\omega}\right)\right]$. Careful statement of the weakest sufficient conditions demands considerably more theoretical apparatus than can be developed here; for this, see Tierney (1994). Somewhat stronger, but still widely applicable, conditions are easier to state. For example, if for any p-measurable A with $\int_A p\left(\boldsymbol{\theta}\right) d\theta > 0$ it is the case that in the Markov chain

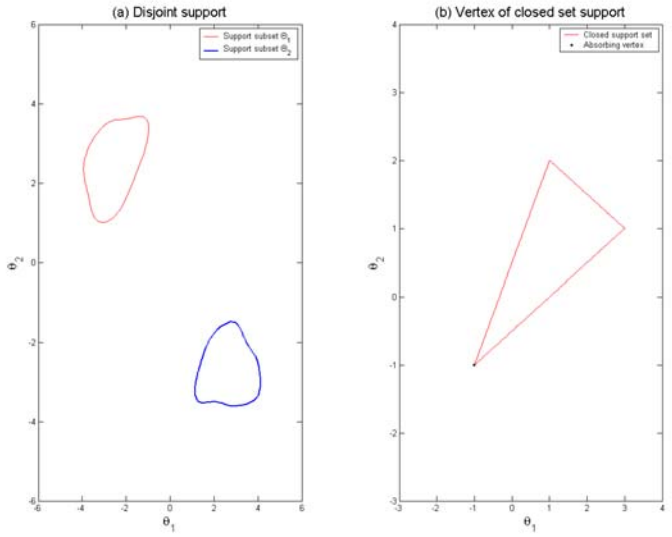

Figure 2: Two examples in which a Gibbs sampling Markov chain will be reducible

(38) $P\left(\boldsymbol{\theta}^{(m+1)} \in A \mid \boldsymbol{\theta}^{(m)}\right) > 0$ for any $\boldsymbol{\theta}^{(m)} \in \Theta$, then the Markov chain is ergodic. (Clearly neither example in Figure 2 satisfies this condition.) For this and other simple conditions see Geweke (2005, Section 3.5).

3.2.2 The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is defined by a probability density function $q\left(\boldsymbol{\theta};\boldsymbol{\theta}^*\right)$ indexed by $\boldsymbol{\theta} \in \Theta$ and with density argument $\boldsymbol{\theta}^*$. The random vector $\boldsymbol{\theta}^*$ generated from $q\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^*\right)$ is a candidate value for $\boldsymbol{\theta}^{(m)}$. The algorithm sets $\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^*$ with probability

$$\alpha\left(\boldsymbol{\theta}^{(m)};\boldsymbol{\theta}^*\right) = \min\left\{\frac{p\left(\boldsymbol{\theta}^*\right)/q\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^*\right)}{p\left(\boldsymbol{\theta}^{(m-1)}\right)/q\left(\boldsymbol{\theta}^*;\boldsymbol{\theta}^{(m-1)}\right)}, 1\right\}; \tag{39}$$

otherwise, $\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)}$. Conditional on $\boldsymbol{\theta} = \boldsymbol{\theta}^{(m-1)}$ the distribution of $\boldsymbol{\theta}^*$ is a mixture of a continuous distribution with density $u\left(\boldsymbol{\theta};\boldsymbol{\theta}^*\right) = q\left(\boldsymbol{\theta};\boldsymbol{\theta}^*\right)\alpha\left(\boldsymbol{\theta};\boldsymbol{\theta}^*\right)$, corresponding to the accepted candidates, and a discrete distribution with probability mass $r\left(\boldsymbol{\theta}\right) = 1 - \int_\Theta u\left(\boldsymbol{\theta};\boldsymbol{\theta}^*\right) d\nu\left(\boldsymbol{\theta}^*\right)$ at the point $\boldsymbol{\theta}$, which is the probability of drawing a $\boldsymbol{\theta}^*$ that will be rejected. The entire transition density can

be expressed using the Dirac delta function as

$$p_H\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^{(m)}\right) = u\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^{(m)}\right) + r\left(\boldsymbol{\theta}^{(m-1)}\right)\delta_{\boldsymbol{\theta}^{(m-1)}}\left(\boldsymbol{\theta}^{(m)}\right). \quad (40)$$

The intuition behind this procedure is evident on the right side of (39), and is in many respects similar to that in acceptance and importance sampling. If the transition density q makes a move from $\boldsymbol{\theta}^{(m-1)}$ to $\boldsymbol{\theta}^*$ quite likely, relative to the target density p at $\boldsymbol{\theta}^*$, and a move back from $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}^{(m-1)}$ quite unlikely, relative to the target density at $\boldsymbol{\theta}^{(m-1)}$, then the algorithm will place a low probability on actually making the transition and a high probability on staying at $\boldsymbol{\theta}^{(m-1)}$. In the same situation, a prospective move from $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}^{(m-1)}$ will always be made because draws of $\boldsymbol{\theta}^{(m-1)}$ are made infrequently relative to the target density p.

This is the most general form of the Metropolis-Hastings algorithm, due to Hastings (1970). The Metropolis et al. (1953) form takes $q\left(\boldsymbol{\theta};\boldsymbol{\theta}^*\right) = q\left(\boldsymbol{\theta}^*;\boldsymbol{\theta}\right)$, which leads to the simplification $\alpha\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^*\right) = \min\left[p\left(\boldsymbol{\theta}^*\right)/p\left(\boldsymbol{\theta}^{(m-1)}\right), 1\right]$. A leading example of this form is the *Metropolis random walk*, in which $q\left(\boldsymbol{\theta};\boldsymbol{\theta}^*\right) = q\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}\right)$ and the latter density is symmetric about $\mathbf{0}$, for example that of the multivariate normal distribution with mean $\mathbf{0}$. Another special case is the *Metropolis independence chain* (see Tierney (1994)) in which $q\left(\boldsymbol{\theta};\boldsymbol{\theta}^*\right) = q\left(\boldsymbol{\theta}^*\right)$. This leads to $\alpha\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^*\right) = \min\left[w\left(\boldsymbol{\theta}^*\right)/w\left(\boldsymbol{\theta}^{(m-1)}\right), 1\right]$, where $w\left(\boldsymbol{\theta}\right) = p\left(\boldsymbol{\theta}\right)/q\left(\boldsymbol{\theta}\right)$. The independence chain is closely related to acceptance sampling and importance sampling. But rather than place a low probability of acceptance or a low weight on a draw that is too likely relative to the target distribution, the independence chain assigns a low probability of transition to that candidate.

There is a simple two-step argument that motivates the convergence of the sequence $\left\{\boldsymbol{\theta}^{(m)}\right\}$, generated by the Metropolis-Hastings algorithm, to the target distribution. (This approach is due to Chib and Greenberg (1994).) First, note that if a transition probability density function $p\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^{(m)}\right)$ satisfies the *reversibility condition*

$$p\left(\boldsymbol{\theta}^{(m-1)}\right)p\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^{(m)}\right) = p\left(\boldsymbol{\theta}^{(m)}\right)p\left(\boldsymbol{\theta}^{(m)},\boldsymbol{\theta}^{(m-1)}\right)$$

for stated $p\left(\cdot\right)$ with support Θ, then

$$\begin{aligned}
&\int_\Theta p\left(\boldsymbol{\theta}^{(m-1)}\right)p\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^{(m)}\right)d\nu\left(\boldsymbol{\theta}^{(m-1)}\right) \\
&= \int_\Theta p\left(\boldsymbol{\theta}^{(m)}\right)p\left(\boldsymbol{\theta}^{(m)};\boldsymbol{\theta}^{(m-1)}\right)d\nu\left(\boldsymbol{\theta}^{(m-1)}\right) \quad (41) \\
&= p\left(\boldsymbol{\theta}^{(m)}\right)\int_\Theta p\left(\boldsymbol{\theta}^{(m)};\boldsymbol{\theta}^{(m-1)}\right)d\nu\left(\boldsymbol{\theta}^{(m-1)}\right) = p\left(\boldsymbol{\theta}^{(m)}\right).
\end{aligned}$$

Expression (41) indicates that if the unconditional distribution of $\boldsymbol{\theta}^{(m-1)}$ corresponds to the density $p\left(\cdot\right)$, then the same is true of the unconditional distribu-

tion of $\boldsymbol{\theta}^{(m)}$. The density $p(\cdot)$ is an invariant density of the Markov chain with transition density $p\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^{(m)}\right)$.

The second step in this argument is to consider the implications of the requirement that $p_H\left(\boldsymbol{\theta}^{(m-1)},\boldsymbol{\theta}^{(m)}\right)$ be reversible for $p(\boldsymbol{\theta})$:

$$p\left(\boldsymbol{\theta}^{(m-1)}\right)p_H\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^{(m)}\right)=p\left(\boldsymbol{\theta}^{(m)}\right)p_H\left(\boldsymbol{\theta}^{(m)};\boldsymbol{\theta}^{(m-1)}\right).$$

For $\boldsymbol{\theta}^{(m-1)}=\boldsymbol{\theta}^{(m)}$ the requirement holds trivially. For $\boldsymbol{\theta}^{(m-1)}\neq\boldsymbol{\theta}^{(m)}$ it implies that

$$p\left(\boldsymbol{\theta}^{(m-1)}\right)q\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^{*}\right)\alpha\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^{*}\right)=p\left(\boldsymbol{\theta}^{*}\right)q\left(\boldsymbol{\theta}^{*};\boldsymbol{\theta}^{(m-1)}\right)\alpha\left(\boldsymbol{\theta}^{*};\boldsymbol{\theta}^{(m-1)}\right). \tag{42}$$

Suppose without loss of generality that

$$p\left(\boldsymbol{\theta}^{(m-1)}\right)q\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^{*}\right)>p\left(\boldsymbol{\theta}^{*}\right)q\left(\boldsymbol{\theta}^{*};\boldsymbol{\theta}^{(m-1)}\right).$$

If $\alpha\left(\boldsymbol{\theta}^{*};\boldsymbol{\theta}^{(m-1)}\right)=1$ and

$$\alpha\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^{*}\right)=p\left(\boldsymbol{\theta}^{*}\right)q\left(\boldsymbol{\theta}^{*};\boldsymbol{\theta}^{(m-1)}\right)/p\left(\boldsymbol{\theta}^{(m-1)}\right)q\left(\boldsymbol{\theta}^{(m-1)};\boldsymbol{\theta}^{*}\right),$$

then (42) is satisfied.

3.2.3 Metropolis within Gibbs

Different MCMC methods can be combined in a variety of rich and interesting ways that have been important in solving many practical problems in Bayesian inference. One of the most important in econometric modelling has been the Metropolis within Gibbs algorithm. Suppose that in attempting to implement a Gibbs sampling algorithm, a conditional density $p\left[\boldsymbol{\theta}_{(b)}\mid\boldsymbol{\theta}_{(a)}\ \ (a\neq b)\right]$ is intractable. The density is not of any known form, and efficient acceptance sampling algorithms are not at hand. This occurs in the stochastic volatility example, for the volatilities $h_1,\ldots,h_T$.

This problem can be addressed by applying the Metropolis-Hastings algorithm in block b of the Gibbs sampler while treating the other blocks in the usual way. Specifically, let $q(\boldsymbol{\theta};\boldsymbol{\theta}_b^*)$ be the density (indexed by $\boldsymbol{\theta}$) from which candidate $\boldsymbol{\theta}_b^*$ is drawn. At iteration m, block b, of the Gibbs sampler draw $\boldsymbol{\theta}_b^*\sim q\left[\boldsymbol{\theta}_a^{(m)}(a<b),\boldsymbol{\theta}_a^{(m-1)}(a\geq b);\boldsymbol{\theta}_b^*\right]$, and set $\boldsymbol{\theta}_b^{(m)}=\boldsymbol{\theta}_b^*$ with probability

$$\begin{aligned}&\alpha\left[\boldsymbol{\theta}_a^{(m)}(a<b),\boldsymbol{\theta}_a^{(m-1)}(a\geq b);\boldsymbol{\theta}_b^*\right]\\ =\ &\min\left\{\frac{p\left[\boldsymbol{\theta}_a^{(m)}(a<b),\boldsymbol{\theta}_b^*,\boldsymbol{\theta}_a^{(m-1)}(a>b)\right]}{q\left[\boldsymbol{\theta}_a^{(m)}(a<b),\boldsymbol{\theta}_a^{(m-1)}(a\geq b);\boldsymbol{\theta}_b^*\right]}\Big/\right.\\ &\left.\frac{p\left[\boldsymbol{\theta}_a^{(m)}(a<b),\boldsymbol{\theta}_a^{(m-1)}(a\geq b)\right]}{q\left[\boldsymbol{\theta}_a^{(m)}(a<b),\boldsymbol{\theta}_b^*,\boldsymbol{\theta}_a^{(m-1)}(a>b);\boldsymbol{\theta}_b^{(m-1)}\right]},\ \ 1\right\}.\end{aligned}$$

If $\boldsymbol{\theta}_b^{(m)}$ is not set to $\boldsymbol{\theta}_b^*$, then $\boldsymbol{\theta}_b^{(m)} = \boldsymbol{\theta}_b^{(m-1)}$. The procedure for $\boldsymbol{\theta}_b$ is exactly the same as for a standard Metropolis step, except that $\boldsymbol{\theta}_a\,(a \neq b)$ also enters the density p and partially indexes the candidate density q. It is usually called a *Metropolis within Gibbs step*.

To see that $p(\boldsymbol{\theta})$ is an invariant density of this Markov chain, consider the simple case of two blocks with a Metropolis within Gibbs step in the second block. Adapting the notation of (40) denote the Metropolis step for the second block

$$p_H(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{\theta}_2^*) = u(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{\theta}_2^*) + r(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2)\, \delta_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_2^*),$$

where

$$u(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{\theta}_2^*) = \alpha(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{\theta}_2^*)\, q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{\theta}_2^*)$$

and

$$r(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2) = 1 - \int_{\Theta_2} u(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{\theta}_2^*)\, d\nu(\boldsymbol{\theta}_2^*). \tag{43}$$

The one-step transition density for the entire chain is

$$p_M(\boldsymbol{\theta}; \boldsymbol{\theta}^*) = p(\boldsymbol{\theta}_1^* \mid \boldsymbol{\theta}_2)\, p_H(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2; \boldsymbol{\theta}_2^*).$$

Then p is an invariant density of p_M if

$$\int_{\Theta} p(\boldsymbol{\theta})\, p_M(\boldsymbol{\theta}; \boldsymbol{\theta}^*)\, d\nu(\boldsymbol{\theta}) = p(\boldsymbol{\theta}^*). \tag{44}$$

To establish (44), begin by expanding the left side,

$$\begin{aligned}
\int_{\Theta} p(\boldsymbol{\theta})\, p_M(\boldsymbol{\theta}; \boldsymbol{\theta}^*)\, d\nu(\boldsymbol{\theta}) &= \int_{\Theta_2} \int_{\Theta_1} p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\, d\nu(\boldsymbol{\theta}_1)\, p(\boldsymbol{\theta}_1^* \mid \boldsymbol{\theta}_2) \\
&\quad \cdot \left[u(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2; \boldsymbol{\theta}_2^*) + r(\boldsymbol{\theta}_1^*; \boldsymbol{\theta}_2)\, \delta_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_2^*)\right] d\nu(\boldsymbol{\theta}_2) \\
&= \int_{\Theta_2} p(\boldsymbol{\theta}_2)\, p(\boldsymbol{\theta}_1^* \mid \boldsymbol{\theta}_2)\, u(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2; \boldsymbol{\theta}_2^*)\, d\nu(\boldsymbol{\theta}_2) \qquad (45) \\
&\quad + \int_{\Theta_2} p(\boldsymbol{\theta}_2)\, p(\boldsymbol{\theta}_1^* \mid \boldsymbol{\theta}_2)\, r(\boldsymbol{\theta}_1^*; \boldsymbol{\theta}_2)\, \delta_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_2^*)\, d\nu(\boldsymbol{\theta}_2). \qquad (46)
\end{aligned}$$

In (45) and (46) we have used the fact that $p(\boldsymbol{\theta}_2) = \int_{\Theta_1} p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\, d\nu(\boldsymbol{\theta}_1)$. Using Bayes rule (45) is the same as

$$p(\boldsymbol{\theta}_1^*) \int_{\Theta_2} p(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^*)\, u(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2; \boldsymbol{\theta}_2^*)\, d\nu(\boldsymbol{\theta}_2). \tag{47}$$

Carrying out the integration in (46) yields

$$p(\boldsymbol{\theta}_2^*)\, p(\boldsymbol{\theta}_1^* \mid \boldsymbol{\theta}_2^*)\, r(\boldsymbol{\theta}_1^*; \boldsymbol{\theta}_2^*). \tag{48}$$

Recalling the reversibility of the Metropolis step,

$$p(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^*)\, u(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2; \boldsymbol{\theta}_2^*) = p(\boldsymbol{\theta}_2^* \mid \boldsymbol{\theta}_1^*)\, u(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*; \boldsymbol{\theta}_2),$$

and so (47) becomes

$$p\left(\boldsymbol{\theta}_1^*\right) p\left(\boldsymbol{\theta}_2^* \mid \boldsymbol{\theta}_1^*\right) \int_{\Theta_2} u\left(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*; \boldsymbol{\theta}_2\right) d\nu\left(\boldsymbol{\theta}_2\right). \tag{49}$$

We can express (48) as

$$p\left(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*\right) r\left(\boldsymbol{\theta}_1^*; \boldsymbol{\theta}_2^*\right). \tag{50}$$

Finally, recalling (43), the sum of (49) and (50) is $p\left(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*\right)$, thus establishing (44).

This demonstration of invariance applies to the Gibbs sampler with b blocks, with a Metropolis within Gibbs step for one block, simply through the convention that Metropolis within Gibbs is used in the last block of each iteration. Metropolis within Gibbs steps can be used for several blocks, as well. The argument for invariance proceeds by mathematical induction, and the details are the same.

Sections 5.2.1 and 5.5 provide applications of Metropolis within Gibbs in Bayesian forecasting models.

3.3 The full Monte

We are now in a position to complete the practical Bayesian agenda for forecasting by means of simulation. This process integrates several sources of uncertainty about the future. These are summarized in the most widely used graduate econometrics textbook (Greene (2003, p 576)) as

1. Uncertainty about parameters ("which will have been estimated");
2. Uncertainty about forecasts of exogenous variables; and
3. Uncertainty about unobservables realized in the future;

   To these we must surely add, along with Diebold (1998, pp 291-292 who includes (1) and (3) but not (2) in his list),

4. Uncertainty about the model itself.

In commenting on this list Greene (2003) observed, "In practice handling the second of these errors is largely intractable while the first is merely extremely difficult." We take strong exception to this assessment, whose dour tone can be traced directly to violations of the two principles set forth at the outset of this chapter. The problem with parameters originates in the violation of the principle of relevant conditioning, as discussed in the conclusions of Sections 2.4.2 and 2.4.3. The difficulty with exogenous variables is grounded in violation of the principle of explicit formulation: a so-called exogenous variable in this situation is one whose joint distribution with the forecasting vector of interest $\boldsymbol{\omega}$ should have been expressed explicitly, but was not.[2] This problem is resolved

[2] The formal problem is that "exogenous variables" are not ancillary statistics when the vector of interest includes future outcomes. In other applications of the same model, they may be. This distinction is clear in the Bayesian statistics literature; see, e.g. Bernardo and Smith (1994, Section 5.1.4) or Geweke (2005, Section 2.2.2).

every day in decision-making, either formally or informally, in any event. If there is great uncertainty about the joint distribution of some relevant variables and the forecasting vector of interest, that uncertainty should be incorporated in the prior distribution, or in uncertainty about the appropriate model.

We turn first to the full integration of the first three sources of uncertainty using posterior simulators (Section 3.3.1) and then to the last source (Section 3.3.2).

3.3.1 Predicitve distributions and point forecasts

Section 2.4 summarized the probability structure of the recursive formulation of a single model A: the prior density $p(\boldsymbol{\theta}_A \mid A)$, the observables density $p(\mathbf{Y}_T \mid \boldsymbol{\theta}_A, A)$, and the probability density of future observables $\boldsymbol{\omega}$, $p(\boldsymbol{\omega} \mid \mathbf{Y}_T, \boldsymbol{\theta}_A, A)$. It is straightforward to simulate from the corresponding distributions, and this is useful in the process of model formulation as discussed in Section 2.2. The principle of relevant conditioning, however, demands that we work instead with the distribution of the unobservables ($\boldsymbol{\theta}_A$ and $\boldsymbol{\omega}$) conditional on the observables, $\mathbf{Y}_T$, and the assumptions of the model, A:

$$p(\boldsymbol{\theta}_A, \boldsymbol{\omega} \mid \mathbf{Y}_T, A) = p(\boldsymbol{\theta}_A \mid \mathbf{Y}_T, A)\, p(\boldsymbol{\omega} \mid \boldsymbol{\theta}_A, \mathbf{Y}_T, A).$$

Substituting the observed values (data) $\mathbf{Y}_T^o$ for $\mathbf{Y}_T$, we can access this distribution by means of a posterior simulator for the first component on the right, followed by simulation from the predictive density for the second component:

$$\boldsymbol{\theta}_A^{(m)} \sim p(\boldsymbol{\theta}_A \mid \mathbf{Y}_T^o, A), \;\; \boldsymbol{\omega}^{(m)} \sim p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}_A^{(m)}, \mathbf{Y}_T^o, A\right). \tag{51}$$

The first step, posterior simulation, has become practicable for most models by virtue of the innovations in MCMC methods summarized in Section 3.2. The second simulation is relatively simple, because it is part of the recursive formulation. The simulations $\boldsymbol{\theta}_A^{(m)}$ from the posterior simulator will not necessarily be i.i.d. (in the case of MCMC) and they may require weighting (in the case of importance sampling) but the simulations are *ergodic*: i.e., so long as $E[h(\boldsymbol{\theta}_A, \boldsymbol{\omega}) \mid \mathbf{Y}_T^o, A]$ exists and is finite,

$$\frac{\sum_{m=1}^M w^{(m)} h\left(\boldsymbol{\theta}_A^{(m)}, \boldsymbol{\omega}^{(m)}\right)}{\sum_{m=1}^M w^{(m)}} \overset{a.s.}{\to} E[h(\boldsymbol{\theta}_A, \boldsymbol{\omega}) \mid \mathbf{Y}_T^o, A]. \tag{52}$$

The weights $w^{(m)}$ in (52) come into play for importance sampling. There is another important use for weighted posterior simulation, to which we return in Section 3.3.2.

This full integration of sources of uncertainty by means of simulation appears to have been applied for the first time in the unpublished thesis of Litterman (1979) as discussed in Section 4. The first published full application of simulation methods in this way in a published paper appears to have been Thompson and Miller (1986), which built on Thompson (1984). This study

applied an autoregressive model of order 2 with a conventional improper diffuse prior (see Zellner (1971, p 195)) to quarterly US unemployment rate data from 1968 through 1979, forecasting for the period 1980 through 1982. Section 4 of their paper outlines the specifics of (51) in this case. They computed posterior means of each of the 12 predictive densities, corresponding to a joint quadratic loss function; predictive variances; and centered 90% predictive intervals. They compared these results with conventional non-Bayesian procedures (see Box and Jenkins (1976)) that equate unknown parameters with their estimates, thus ignoring uncertainty about these parameters. There were several interesting findings and comparisons.

1. The posterior means of the parameters and the non-Bayesian point estimates are similar: $y_t = .441 + 1.596y_{t-1} - 0.669y_{t-2}$ for the former and $y_t = 0.342 + 1.658y_{t-1} - 0.719y_{t-2}$ for the latter.

2. The point forecasts from the predictive density and the conventional non-Bayesian procedure depart substantially over the 12 periods, from unemployment rates of 5.925% and 5.904%, respectively, one-step-ahead, to 6.143% and 5.693%, respectively, 12 steps ahead. This is due to the fact that an F-step-ahead mean, conditional on parameter values, is a polynomial of order F in the parameter values: predicting farther into the future involves an increasingly non-linear function of parameters, and so the discrepancy between the mean of the nonlinear function and the non-linear function of the mean also increases.

3. The Bayesian 90% predictive intervals are generally wider than the corresponding non-Bayesian intervals; the difference is greatest 12 steps ahead, where the width is 5.53% in the former and 5.09% in the latter. At 12 steps ahead the 90% intervals are (3.40%, 8.93%) and (3.15%, 8.24%)

4. The predictive density is platykurtic; thus a normal approximation of the predictive density (today a curiosity, in view of the accessible representation (51)) produces a 90% predictive density that is too wide, and the discrepancy increases for predictive densities farther into the future: 5.82% rather than 5.53%, 12 steps ahead.

Thompson and Miller did not repeat their exercise for other forecasting periods, and therefore had no evidence on forecasting reliability. Nor did they employ the shrinkage priors that were, contemporaneously, proving so important in the successful application of Bayesian vector autoregressions at the Federal Reserve Bank of Minneapolis. We return to that project in Section 6.1.

3.3.2 Model combination and the revision of assumptions

Incorporation of uncertainty about the model itself is rarely discussed, and less frequently acted upon; Greene (2003) does not even mention it. This lacuna is rational in non-Bayesian approaches: since uncertainty cannot be integrated in

the context of one model, it is premature, from this perspective, even to contemplate this task. Since model-specific uncertainty has been resolved, both as a theoretical and as a practical matter, in Bayesian forecasting, the problem of model uncertainty is front and center. Two variants on this problem are integrating uncertainty over a well-defined set of models, and bringing additional, but similar, models into such a group in an efficient manner.

Extending the expression of uncertainty to a set of J specified models is straightforward in principle, as detailed in Section 2.3. From (24)-(27) it is clear that the additional technical task is the evaluation of the marginal likelihoods

$$p\left(\mathbf{Y}_T^o \mid A_j\right) = \int_{\Theta_{A_j}} p\left(\mathbf{Y}_T^o \mid \boldsymbol{\theta}_{A_j}, A_j\right) p\left(\boldsymbol{\theta}_{A_j} \mid A_j\right) d\boldsymbol{\theta}_{A_j} \quad (j = 1, \ldots . J) .$$

Specific cases aside, simulation approximation of the marginal likelihood turns out not to be a special case of approximating a posterior moment in the model A_j. One such specific case of practical importance involves models A_j and A_k with a common vector of unobservables $\boldsymbol{\theta}_A$ and likelihood $p\left(\mathbf{Y}_T^o \mid \boldsymbol{\theta}_A, A_j\right) = p\left(\mathbf{Y}_T^o \mid \boldsymbol{\theta}_A, A_k\right)$ but different prior densities $p\left(\boldsymbol{\theta}_A \mid A_j\right)$ and $p\left(\boldsymbol{\theta}_A \mid A_k\right)$. (For example, one model might incorporate a set of inequality restrictions while the other does not.) If $p\left(\boldsymbol{\theta}_A \mid A_k\right) / p\left(\boldsymbol{\theta}_A \mid A_j\right)$ is bounded above on the support of $p\left(\boldsymbol{\theta}_A \mid A_j\right)$, and if $\boldsymbol{\theta}_A^{(m)} \sim p\left(\boldsymbol{\theta}_A \mid \mathbf{Y}_T^o, A_j\right)$ is ergodic then

$$M^{-1} \sum_{m=1}^{M} p\left(\boldsymbol{\theta}_A \mid A_k\right) / p\left(\boldsymbol{\theta}_A \mid A_j\right) \stackrel{a.s.}{\rightarrow} p\left(\mathbf{Y}_T^o \mid A_k\right) / p\left(\mathbf{Y}_T^o \mid A_j\right) ; \qquad (53)$$

see Geweke (2005, Section 5.2.1).

For certain types of posterior simulators, simulation-consistent approximation of the marginal likelihood is also straightforward: see Geweke (1989b, Section 5 or Geweke (2005, Section 5.2.2) for importance sampling, Chib (1995) for Gibbs sampling, Chib and Jeliazkov (2001) for the Metropolis-Hastings algorithm, and Meng and Wong (1996) for a general theoretical perspective. An approach that is more general, but often computationally less efficient in these specific cases, is the density ratio method of Gelfand and Dey (1994), also described in Geweke (2005, Section 5.2.4). These approaches, and virtually any conceivable approach, require that it be possible to evaluate or approximate with substantial accuracy the likelihood function. This condition is not necessary in MCMC posterior simulators, and this fact has been central to the success of these simulations in many applications, especially those with latent variables. This, more or less, defines the rapidly advancing front of attack on this important technical issue at the time of this writing.

Some important and practical modifications can be made to the set of models over which uncertainty is integrated, without repeating the exercise of posterior simulation. These modifications all exploit reweighting of the posterior simulator output. One important application is updating posterior distributions with new data. In a real-time forecasting situation, for example, one might wish to update predictive distributions minute-by-minute, whereas as a full posterior

simulation adequate for the purposes at hand might take more than a minute (but less than a night). Suppose the posterior simulation utilizes data through time T, but the predictive distribution is being formed at time $T^* > T$. Then

$$\begin{aligned} p\left(\boldsymbol{\omega} \mid \mathbf{Y}_{T^*}^o, A\right) &= \int_{\Theta_A} p\left(\boldsymbol{\theta}_A \mid \mathbf{Y}_{T^*}^o, A\right) p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}_A, \mathbf{Y}_{T^*}^o, A\right) d\boldsymbol{\theta}_A \\ &= \int_{\Theta_A} p\left(\boldsymbol{\theta}_A \mid \mathbf{Y}_T^o, A\right) \frac{p\left(\boldsymbol{\theta}_A \mid \mathbf{Y}_{T^*}^o, A\right)}{p\left(\boldsymbol{\theta}_A \mid \mathbf{Y}_T^o, A\right)} p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}_A, \mathbf{Y}_{T^*}^o, A\right) d\boldsymbol{\theta}_A \\ &\propto \int_{\Theta_A} p\left(\boldsymbol{\theta}_A \mid \mathbf{Y}_T^o, A\right) p\left(\mathbf{y}_{T+1}^o, \ldots, \mathbf{y}_{T^*}^o \mid \boldsymbol{\theta}_A, A\right) \\ &\quad \cdot p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}_A, \mathbf{Y}_{T^*}^o, A\right) d\boldsymbol{\theta}_A. \end{aligned}$$

This suggests that one might use the simulator output $\boldsymbol{\theta}^{(m)} \sim p\left(\boldsymbol{\theta}_A \mid \mathbf{Y}_T^o, A\right)$, taking $\boldsymbol{\omega}^{(m)} \sim p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}_A^{(m)}, \mathbf{Y}_{T^*}^o, A\right)$ but reweighting the simulator output to approximate $E\left[h\left(\boldsymbol{\omega}\right) \mid \mathbf{Y}_{T^*}^o, A\right]$ by

$$\sum_{m=1}^{M} p\left(\mathbf{y}_{T+1}^o, \ldots, \mathbf{y}_{T^*}^o \mid \boldsymbol{\theta}_A^{(m)}, A\right) h\left(\boldsymbol{\omega}^{(m)}\right) / \sum_{m=1}^{M} p\left(\mathbf{y}_{T+1}^o, \ldots, \mathbf{y}_{T^*}^o \mid \boldsymbol{\theta}_A^{(m)}, A\right). \tag{54}$$

This turns out to be correct; for details see Geweke (2000). One can show that (54) is a simulation-consistent approximation of $E\left[h\left(\boldsymbol{\omega}\right) \mid \mathbf{Y}_{T^*}^o, A\right]$ and in many cases the updating requires only spreadsheet arithmetic, as illustrated in Foster and Whiteman (2004). There are central limit theorems on which to base assessments of the accuracy of the approximations; these require more advanced, but publicly available, software; see Geweke (1999) and Geweke (2005, Sections 4.1 and 5.4).

The method of reweighting can also be used to bring into the fold models with the same likelihood function but different priors, or to explore the effect of modifying the prior, as (53) suggests. In that context A_k denotes the new model, with a prior distribution that is more informative in the sense that $p\left(\boldsymbol{\theta}_A \mid A_k\right) / p\left(\boldsymbol{\theta}_A \mid A_j\right)$ is bounded above on the support of Θ_{A_j}. Reweighting the posterior simulator output $\boldsymbol{\theta}_{A_j}^{(m)} \sim p\left(\boldsymbol{\theta}_{A_j} \mid \mathbf{Y}_T^o, A_j\right)$ by $p\left(\boldsymbol{\theta}_{A_j}^{(m)} \mid A_k\right) / p\left(\boldsymbol{\theta}_{A_j}^{(m)} \mid A_j\right)$ provides the new simulation-consistent set of approximations. Moreover, the exercise yields the marginal likelihood of the new model almost as a by-product, because

$$M^{-1} \sum_{m=1}^{M} p\left(\boldsymbol{\theta}_{A_j}^{(m)} \mid A_k\right) / p\left(\boldsymbol{\theta}_{A_j}^{(m)} \mid A_j\right) \xrightarrow{a.s.} p\left(\mathbf{Y}_T^o \mid A_k\right) / p\left(\mathbf{Y}_T^o \mid A_j\right) \tag{55}$$

This suggests a pragmatic reason for investigators to use prior distributions $p\left(\boldsymbol{\theta}_A \mid A_j\right)$ that are uninformative, in this sense: clients can tailor the simulator output to their more informative priors $p\left(\boldsymbol{\theta}_A \mid A_k\right)$ by reweighting.

4 'Twas not always so easy: a historical perspective

The procedures outlined in the previous section accommodate, at least in principle (and much practice), very general likelihood functions and prior distributions, primarily because numerical substitutes are available for analytic evaluation of expectations of functions of interest. But prior to the advent of inexpensive desktop computing in the mid-1980's, Bayesian prediction was an analytic art. The standard econometric reference for Bayesian work of any such kind was Zellner (1971), which treats predictive densities at a level of generality similar to that in Section 1.2 above, and in detail for Gaussian location, regression, and multiple regression problems.

4.1 In the beginning, there was diffuseness, conjugacy, and analytic work

In these specific examples, Zellner's focus was on the diffuse prior case, which leads to the usual normal-gamma posterior. To illustrate his approach to prediction in the normal regression model, let $p = 1$ and write the model (a version of equation (1)) as

$$\mathbf{Y}_T = \mathbf{X}_T\boldsymbol{\beta} + \mathbf{u}_T \tag{56}$$

where

$\mathbf{X}_T$ = a $T \times k$ matrix, with rank k, of observations on the independent variables,

$\boldsymbol{\beta}$ = a $k \times 1$ vector of regression coefficients,

$\mathbf{u}_T$ = a $T \times 1$ vector of error terms, assumed Gaussian with mean zero and variance matrix $\sigma^2\mathbf{I}_T$.

Zellner (1971, Section 3.2) employs the "diffuse" prior specification $p(\boldsymbol{\beta},\sigma) \propto \frac{1}{\sigma}$. With this prior, the joint density for the parameters and the q-step prediction vector $\tilde{\mathbf{Y}} = \{y_s\}_{s=T+1}^{T+q}$, assumed to be generated by

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{u}},$$

(a version of (8)) is given by

$$p(\tilde{\mathbf{Y}}, \boldsymbol{\beta},\sigma|\mathbf{Y}_T, \mathbf{X}_T, \tilde{\mathbf{X}}) = \; p(\tilde{\mathbf{Y}}|\boldsymbol{\beta},\sigma, \tilde{\mathbf{X}})p(\boldsymbol{\beta},\sigma|\mathbf{Y}_T, \mathbf{X}_T)$$

which is the product of the conditional Gaussian predictive for $\tilde{\mathbf{Y}}$ given the parameters, and independent variables and the posterior density for $\boldsymbol{\beta}$ and σ, which is given by

$$p(\boldsymbol{\beta},\sigma|\mathbf{Y}_T, \mathbf{X}_T) \propto \sigma^{-(T+1)} \exp\{-(\mathbf{Y}_T - \mathbf{X}_T\beta)'(\mathbf{Y}_T - \mathbf{X}_T\beta)/2\sigma^2\} \tag{57}$$

and which in turn can be seen to be the product of a conditional Gaussian density for $\boldsymbol{\beta}$ given σ and the data and an inverted gamma density for σ given the data. In fact, the joint density is

$$\begin{aligned} p(\tilde{\mathbf{Y}}, \boldsymbol{\beta}, \sigma | \mathbf{Y}_T, \mathbf{X}_T, \tilde{\mathbf{X}}) \quad \propto \quad & \sigma^{-(T+q+1)} \exp\left\{ \left[(\mathbf{Y}_T - \mathbf{X}_T\beta)'(\mathbf{Y}_T - \mathbf{X}_T\beta) \right.\right. \\ & \left.\left. + (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\beta)'(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\beta) \right] / 2\sigma^2 \right\} \end{aligned}$$

To obtain the predictive density (21), $p(\tilde{\mathbf{Y}},|\mathbf{Y}_T, \mathbf{X}_T, \tilde{\mathbf{X}})$, Zellner marginalizes analytically rather than numerically. He does so in two steps: first, he integrates with respect to σ to obtain

$$p(\tilde{\mathbf{Y}}, \boldsymbol{\beta} | \mathbf{Y}_T, \mathbf{X}_T, \tilde{\mathbf{X}}) \propto \left[(\mathbf{Y}_T - \mathbf{X}_T\boldsymbol{\beta})'(\mathbf{Y}_T - \mathbf{X}_T\boldsymbol{\beta}) + (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})'(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}) \right]^{-(T+q)/2}$$

and then completes the square in $\boldsymbol{\beta}$, rearranges, integrates and obtains

$$\begin{aligned} p(\tilde{\mathbf{Y}},|\mathbf{Y}_T, \mathbf{X}_T, \tilde{\mathbf{X}}) \quad \propto \quad & \left[\mathbf{Y}_T'\mathbf{Y}_T + \tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} \right. \\ & \left. - (\mathbf{X}_\mathbf{T}'\mathbf{Y}_\mathbf{T} + \tilde{\mathbf{X}}'\tilde{\mathbf{Y}})' M^{-1} (\mathbf{X}_T'\mathbf{Y}_T + \tilde{\mathbf{X}}'\tilde{\mathbf{Y}}) \right]^{-(T-k+q)/2} \end{aligned}$$

where $\mathbf{M} = \mathbf{X}_T'\mathbf{X}_T + \tilde{\mathbf{X}}'\tilde{\mathbf{X}}$. After considerable additional algebra to put this into "a more intelligible form", Zellner obtains

$$p(\tilde{\mathbf{Y}},|\mathbf{Y}_T, \mathbf{X}_T, \tilde{\mathbf{X}}) \propto \left[T - k + (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}})' \mathbf{H} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}) \right]^{-(T-k+q)}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}_T'\mathbf{X}_T)^{-1}\mathbf{X}_T'\mathbf{Y}_T$ is the in-sample ordinary least squares estimator, $\mathbf{H} = (1/s^2)(I - \tilde{\mathbf{X}}\mathbf{M}^{-1}\tilde{\mathbf{X}}')$, and $s^2 = \frac{1}{T-k}(\mathbf{Y}_T - \mathbf{X}_T\hat{\boldsymbol{\beta}})'\mathbf{H}(\mathbf{Y}_T - \mathbf{X}_T\hat{\boldsymbol{\beta}})$. This formula is then recognized as the multivariate Student-t density, meaning that $\tilde{\mathbf{Y}}$ is distributed as such with mean $\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}$ (provided $T - k > 1$) and covariance matrix $\frac{T-k}{T-k-2}\mathbf{H}^{-1}$ (provided $T-k > 2$). Zellner notes that a linear combination of the elements of $\tilde{\mathbf{Y}}$ (his example of such a function of interest is a discounted sum) will be distributed as univariate Student-t, so that expectations of such linear combinations can be calculated as a matter of routine, but he does not elaborate further. In the multivariate regression model (Zellner, 1971, Section 8.2), similar calculations to those above lead to a generalized multivariate-t predictive distribution.

Zellner's treatment of the Bayesian prediction problem constituted the state of the art at the beginning of the 1970's. In essence, linear models with Gaussian errors and flat priors could be utilized, but not much more generality than this was possible. Slightly greater generality was available if the priors were *conjugate.* Such priors leave the posterior in the same form as the likelihood. In the Gaussian regression case, this means a normal-gamma prior (normal for the regression coefficients, inverted gamma for the residual standard deviation) and a normal likelihood. As Section 2 makes clear, there is no longer need for conjugacy and simple likelihoods, as developments of the past 15 years have made it possible to replace "integration by Arnold Zellner" with "integration by Monte Carlo."

4.2 The dynamic linear model

In 1976, P. J. Harrison and C. F. Stevens (Harrison and Stevens, 1976) read a paper with a title that anticipates ours before the Royal Statistical Society in which they remarked that "[c]ompared with current forecasting fashions our views may well appear radical". Their approach involved the dynamic linear model (see also **Harvey Chapter in this Volume**), which is a version of a state-space observer system:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{x}_t'\boldsymbol{\beta}_t + \mathbf{u}_t; \\ \boldsymbol{\beta}_t &= G\boldsymbol{\beta}_{t-1} + \mathbf{w}_t \end{aligned}$$

with $\mathbf{u}_t \overset{iid}{\sim} N(0, \mathbf{U}_t)$ and $\mathbf{w}_t \overset{iid}{\sim} N(0, \mathbf{W}_t)$. Thus the slope parameters are treated as latent variables, as in Section 2.2.4. As Harrison and Stevens note, this generalizes the standard linear Gaussian model (one of Zellner's examples) by permitting time variation in $\boldsymbol{\beta}$ and the residual covariance matrix. Starting from a prior distribution for $\boldsymbol{\beta}_0$ Harrison and Stevens calculate posterior distributions for $\boldsymbol{\beta}_t$ for $t = 1, 2, ...$ via the (now) well-known Kalman filter recursions. They also discuss prediction formulae for $\mathbf{y}_{T+k}$ at time T under the assumption (i) that $\mathbf{x}_{T+k}$ is known at T, and (ii) $\mathbf{x}_{T+k}$ is unknown at T. They note that their predictions are "distributional in nature, and derived from the current parameter uncertainty" and that "[w]hile it is natural to think of the expectations of the future variate values as "forecasts" there is no need to single out the expectation for this purpose ... if the consequences of an error in one direction are more serious that an error of the same magnitude in the opposite direction, then the forecast can be biased to take this into account" (cf Section 2.4.1).

Harrison and Stevens take up several examples, beginning with the standard regression model, the "static case". They note that in this context, their Bayesian–Kalman filter approach amounts to a

> computationally neat and economical method of revising regression coefficient estimates as fresh data become available, without effectively re-doing the whole calculation all over again and without any matrix inversion. This has been previously pointed out by Plackett (1950) and others but its practical importance seems to have been almost completely missed. (p. 215)

Other examples they treat include the linear growth model, additive seasonal model, periodic function model, autoregressive models, and moving average models. They also consider treatment of multiple possible models, and integrating across them to obtain predictions, as in Section 2.3.

Note that the Harrison-Stevens approach generalized what was possible using Zellner's 1971 book, but priors were still conjugate, and the underlying structure was still Gaussian. The structures that could be handled were more general, but the statistical assumptions and nature of prior beliefs accommodated were

quite conventional. Indeed, in his discussion of Harrison-Stevens, Chatfield (1976) remarks that

> ... you do not need to be Bayesian to adopt the method. If, as the authors suggest, the general purpose default priors work pretty well for most time series, then one does not need to supply prior information. So, despite the use of Bayes' theorem inherent in Kalman filtering, I wonder if *Adaptive Forecasting* would be a better description of the method. (p.231)

The fact remains, though, that latent-variable structure of the forecasting model does put uncertainty about the parameterization on a par with the uncertainty associated with the stochastic structure of the observables themselves.

4.3 The Minnesota revolution

During the mid- to late-1970's, Christopher Sims was writing what would become "Macroeconomics and Reality", the lead article in the January 1980 issue of *Econometrica*. In that paper, Sims argued that identification conditions in conventional large-scale econometric models that were routinely used in (non Bayesian) forecasting and policy exercises, were "incredible" – either they were normalizations with no basis in theory, or "based" in theory that was empirically falsified or internally inconsistent. He proposed, as an alternative, an approach to macroeconomic time series analysis with little theoretical foundation other than statistical stationarity. Building on the Wold decomposition theorem, Sims argued that, exceptional circumstances aside, vectors of time series could be represented by an autoregression, and further, that such representations could be useful for assessing features of the data even though they reproduce only the first and second moments of the time series and not the entire probabilistic structure or "data generation process."

With this as motivation, Robert Litterman (1979) took up the challenge of devising procedures for forecasting with such models that were intended to compete directly with large-scale macroeconomic models then in use in forecasting. Betraying a frequentist background, much of Litterman's effort was devoted to dealing with "multicollinearity problems and large sampling errors in estimation". These "problems" arise because in (3), each of the equations for the p variables involves m lags of each of p variables, resulting in mp^2 coefficients in $\mathbf{B}_1, ..., \mathbf{B}_m$. To these are added the parameters $\mathbf{B}_0$ associated with the deterministic components, as well as the $p(p+1)$ distinct parameters in $\mathbf{\Psi}$.

Litterman (1979) treats these problems in a distinctly classical way, introducing "restrictions in the form of priors" in a subsection on "Biased Estimation". While he notes that "each of these methods may be given a Bayesian interpretation," he discusses reduction of sampling error in classical estimation of the parameters of the normal linear model (56) via the standard ridge regression estimator (Hoerl and Kennard, 1970)

$$\boldsymbol{\beta}_R^k = (\mathbf{X}_T'\mathbf{X}_T + \varrho\mathbf{I}_k)^{-1}\mathbf{X}_T'\mathbf{Y}_T,$$

the Stein (1974) class

$$\boldsymbol{\beta}_S^k = (\mathbf{X}_T'\mathbf{X}_T + \varrho\mathbf{X}_T'\mathbf{X}_T)^{-1}\mathbf{X}_T'\mathbf{Y}_T,$$

and, following Maddala (1977), the "generalized ridge"

$$\boldsymbol{\beta}_S^k = (\mathbf{X}_T'\mathbf{X}_T + \varrho\boldsymbol{\Delta}^{-1})^{-1}(\mathbf{X}_T'\mathbf{Y}_T + \varrho\boldsymbol{\Delta}^{-1}\theta). \tag{58}$$

Litterman notes that the latter "corresponds to a prior distribution on β of $N(\theta, \lambda^2\Delta)$ with $\varrho = \sigma^2/\lambda^2$." (Both parameters σ^2 and λ^2 are treated as known.) Yet Litterman's next statement is frequentist: "The variance of this estimator is given by $\sigma^2(\mathbf{X}_T'\mathbf{X}_T + \varrho\boldsymbol{\Delta}^{-1})^{-1}$". It is clear from his development that he has the "Bayesian" shrinkage in mind as a way of reducing the sampling variability of otherwise frequentist estimators.

Anticipating a formulation to come, Litterman considers two shrinkage priors (which he refers to as "generalized ridge estimators") designed specifically with lag distributions in mind. The canonical distributed lag model for scalar y and x is given by

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1+\ldots+}\beta_l x_{t-m} + u_t. \tag{59}$$

The first prior, due to Leamer (1972), shrinks the mean and variance of the lag coefficients at the same geometric rate with the lag, and covariances between the lag coefficients at a different geometric rate according to the distance between them:

$$\begin{aligned} E\beta_i &= \upsilon\rho^i \\ cov(\beta_i, \beta_j) &= \lambda^2\omega^{|i-j|}\rho^{i+j-2} \end{aligned}$$

with $0 < \rho, \omega < 1$. The hyperparameters ρ, and ω control the decay rates, while υ and λ control the scale of the mean and variance. The spirit of this prior lives on in the "Minnesota" prior to be discussed presently.

The second prior is Shiller's (1973) "smoothness" prior, embodied by

$$\mathbf{R}[\beta_1 \ldots \beta_m]' = \mathbf{w}; \qquad \mathbf{w} \sim N(0, \sigma_w^2 I_{m-2}) \tag{60}$$

where the matrix R incorporates smoothness restrictions by "differencing" adjacent lag coefficients; for example, to embody the notion that second differences between lag coefficients are small (that the lag distribution is quadratic), R is given by

$$R = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & & \ldots & & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & \ldots & & 0 \\ & & & & & \ddots & & & \\ 0 & & \ldots & & & & 1 & -2 & 1 \end{bmatrix}$$

Having introduced these priors, Litterman dismisses the latter, quoting Sims: "... the whole notion that lag distributions in econometrics ought to be smooth is ... at best weakly supported by theory or evidence" (Sims, 1974, p. 317). In place of a smooth lag distribution, Litterman (1979, p. 20) assumed that

"a reasonable approximation of the behavior of an economic variable is a random walk around an unknown, deterministic component." Further, Litterman operated equation by equation, and therefore assumed that the parameters for equation i of the autoregression (3) were centered around

$$y_{it} = y_{i,t-1} + d_{it} + \varepsilon_{it}.$$

Litterman goes on to describe the prior:

> The parameters are all assumed to have means of zero except the coefficient on the first lag of the dependent variable, which is given a prior mean of one. The parameters are assumed to be uncorrelated with each other and to have standard deviations which decrease the further back they are in the lag distributions. In general, the prior distribution on lag coefficients of the dependent variable is much looser, that is, has larger standard deviations, than it is on other variables in the system. (p. 20)

A footnote explains that while the prior represents Litterman's opinion, "it was developed with the aid of many helpful suggestions from Christopher Sims." (Litterman, 1979, p. 96) Inasmuch as these discussions and the prior development took place during the course of Litterman's dissertation work at the University of Minnesota under Sims's direction, the prior has come to be known as the "Minnesota" or "Litterman" prior. Prior information on deterministic components is taken to be diffuse, though he does use the simple first order stationary model

$$y_{1t} = \alpha + \beta y_{1,t-1} + \varepsilon_{1t}$$

to illustrate the point that the mean $M_1 = E(y_{1t})$ and persistence (β) are related by $M_1 = \alpha/(1-\beta)$, indicating that priors on the deterministic components independent of the lag coefficients are problematic. This notion was taken up by Schotman and Van Dijk (1991) in the unit root literature.

The remainder of the prior involves the specification of the standard deviation of the coefficient on lag l of variable j in equation i: δ_{ij}^l. This is specified by

$$\delta_{ij}^l = \begin{cases} \frac{\lambda}{l^{\gamma_1}} & \text{if } i = j \\ \frac{\lambda \gamma_2 \hat{\sigma}_i}{l^{\gamma_1} \hat{\sigma}_j} & \text{if } i \neq j \end{cases} \tag{61}$$

where γ_1 is a hyperparameter greater than 1.0, γ_2 and λ are scale factors, and $\hat{\sigma}_i$ and $\hat{\sigma}_j$ are the estimated residual standard deviations in unrestricted ordinary least squares estimates of equations i and j of the system. (In subsequent work, e.g., Litterman, 1986, the residual standard deviation estimates were from univariate autoregressions.) Alternatively, the prior can be expressed as

$$\mathbf{R}_i \boldsymbol{\beta}_i = \mathbf{r}_i + \mathbf{v}_i; \qquad \mathbf{v}_i \sim N(0, \lambda^2 \mathbf{I}_{mp}) \tag{62}$$

where $\boldsymbol{\beta}_i$ represents the lag coefficients in equation i (the i^{th} row of $B_1, B_2, ..., B_l$ in equation (3)), R_i is a diagonal matrix with zeros corresponding to deterministic components and elements λ/δ_{ij}^l corresponding to the l^{th} lag of variable j, and r_i is a vector of zeros except for a one corresponding to the first lag of variable i. Note that specification of the prior involves choosing the prior hyperparameters for "overall tightness" λ, the "decay" γ_1, and the "other's weight" γ_2. Subsequent modifications and embellishments (encoded in the principal software developed for this purpose, RATS) involved alternative specifications for the decay rate (harmonic in place of geometric), and generalizations of the meaning of "other" (some "others" are more equal than others).

Litterman is careful to note that the prior is being applied equation by equation, and that he will "indeed estimate each equation separately." Thus the prior was to be implemented one equation at a time, with known parameter values in the mean and variance; this meant that the "estimator" corresponded to Theil's (1963) mixed estimator, which could be implemented using the generalized ridge formula (58). With such an estimator, $\tilde{\mathbf{B}} = (\tilde{B}_0, \tilde{B}_1, ..., \tilde{B}_m)$, forecasts were produced recursively via (3). thus the one-step-ahead forecast so produced will correspond to the mean of the predictive density, but ensuing steps will not owing to the nonlinear interactions between forecasts and the $B_j s$. (For an example of the practical effect of this phenomenon, see Section 3.3.1.)

Litterman noted a possible loss of "efficiency" associated with his equation-by-equation treatment, but argued that the loss was justified because of the "computational burden" of a full system treatment, due to the necessity of inverting the large cross-product matrix of right-hand-side variables. This refers to the well-known result that equation-by-equation ordinary least squares estimation is sampling-theoretic efficient in the multiple linear regression model when the right-hand-side variables are the same in all equations. Unless $\boldsymbol{\Psi}$ is diagonal, this does not hold when the right-hand-side variables differ across equations. This, coupled with the way the prior was implemented led Litterman to reason that a system method would be more "efficient". To see this, suppose that $p > 1$ in (3), stack observations on variable i in the $T \times 1$ vector $\mathbf{Y}_{iT}$, the $T \times pm + d$ matrix with row t equal to $(D_t', y_{t-1}', ..., y_{t-m}')$ as $\mathbf{X}_T$ and write the equation i analogue of (56) as

$$\mathbf{Y}_{iT} = \mathbf{X}_T \boldsymbol{\beta}_i + \mathbf{u}_{iT}. \tag{63}$$

Obtaining the posterior mean associated with the prior (62) is straightforward using a "trick" of mixed estimation: simply append "dummy variables" r_i to the bottom of $\mathbf{Y}_{iT}$ and R_i to the bottom of $\mathbf{X}_T$, and apply OLS to the resulting system. This produces the appropriate analogue of (58). But now the right-hand-side variables for equation i are of the form

$$\begin{bmatrix} \mathbf{X}_T \\ R_i \end{bmatrix}$$

which are of course not the same across equations. In a sampling-theory context with multiple equations with explanatory variables of this form, the "efficient"

estimator is the seemingly-unrelated-regression (see Zellner, 1971) estimator, which is not the same as OLS applied equation-by-equation. In the special case of diagonal $\mathbf{\Psi}$, however, equation-by-equation calculations are sufficient to compute the posterior mean of the VAR parameters. Thus Litterman's (1979) "loss of efficiency" argument suggests that a perceived computational burden in effect forced him to make unpalatable assumptions regarding the off-diagonal elements of $\mathbf{\Psi}$

Litterman also sidestepped another computational burden (at the time) of treating the elements of the prior as unknown. Indeed, the use of estimated residual standard deviations in the specification of the prior is an example of the "empirical" Bayesian approach. He briefly discussed the difficulties associated with treating the parameters of the prior as unknown, but argued that the required numerical integration of the resulting distribution (the diffuse prior version of which is Zellner's (57) above) was "not feasible." As is clear from Section 2 above (and 5 below), ten years later, feasibility was not a problem.

Litterman implemented his scheme on a three-variable VAR involving real GNP, M1, and the GNP price deflator using a quarterly sample from 1954:1 to 1969:4, and a forecast period 1970:1 to 1978:1. In undertaking this effort, he introduced a recursive evaluation procedure. First, he estimated the model (obtained $\tilde{\mathbf{B}}$) using data through 1969:4 and made predictions for 1 through K steps ahead. These were recorded, the sample updated to 1970:1, the model re-estimated, and the process was repeated for each quarter through 1977:4. Various measures of forecast accuracy (mean absolute error, root mean squared error, and Theil's U–the ratio of the root mean squared error to that of a no-change forecast) were then calculated for each of the forecast horizons 1 through K. Estimation was accomplished by the Kalman filter, though it was used only as a computational device, and none of its inherent Bayesian features were utilized. Litterman's comparison to McNees's (1975) forecast performance statistics for several large-scale macroeconometric models suggested that the forecasting method worked well, particularly at horizons of about two to four quarters.

In addition to traditional measures of forecast accuracy, Litterman also devoted substantial effort to producing Fair's (1980) "estimates of uncertainty". These are measures of forecast accuracy that embody adjustments for changes in the variances of the forecasts over time. In producing these measures for his Bayesian VARs, Litterman anticipated much of the essence of posterior simulation that would be developed over the next fifteen years. The reason is that Fair's method decomposes forecast uncertainty into several sources, of which one is the uncertainty due to the need to estimate the coefficients of the model. Fair's version of the procedure involved simulation from the frequentist *sampling* distribution of the coefficient estimates, but Litterman explicitly indicated the need to stochastically simulate from the posterior distribution of the VAR parameters as well as the distribution of the error terms. Indeed, he generated 50 (!) random samples from the (equation-by-equation, empirical Bayes' counterpart to the) predictive density for a six variable, four-lag VAR. Computations required 1024 seconds on the CDC Cyber 172 computer at the University of

Minnesota, a computer that was fast by the standards of the time.

Doan, Litterman, Sims (DLS, 1984) built on Litterman, though they retained the equation-by-equation mode of analysis he had adopted. Key innovations included accommodation of time variation via a Kalman filter procedure like that used by West and Harrison (1976) for the dynamic linear model discussed above, and the introduction of new features of the prior to reflect views that sums of own lag coefficients in each equation equal unity, further reflecting the random walk prior. (Sims, 1992, subsequently introduced a related additional feature of the prior reflecting the view that variables in the VAR may be cointegrated.)

After searching over prior hyperparameters (overall tightness, degree of time variation, etc.) DLS produced a "prior" involving small time variation and some "bite" from the sum-of-lag coefficients restriction that improved psuedo-real time forecast accuracy modestly over univariate predictions for a large (10 variable) model of macroeconomic time series. They conclude the improvement is "... substantial relative to differences in forecast accuracy ordinarily turned up in comparisons across methods, even though it is not large relative to total forecast error." (pp. 26-27)

4.4 After Minnesota: subsequent developments

Like DLS, Kadiyala and Karlsson (1993) studied a variety of prior distributions for macroeconomic forecasting, and extended the treatment to full system-wide analysis. They began by noting that Litterman's (1979) equation-by-equation formulation has an interpretation as a multivariate analysis, albeit with a Gaussian prior distribution for the VAR coefficients characterized by a diagonal, known, variance-covariance matrix. (In fact, this "known" covariance matrix is data determined owing to the presence of estimated residual standard deviations in equation (61).) They argue that diagonality is a more troublesome assumption (being "rarely supported by data") than the one that the covariance matrix is known, and in any case introduce four alternatives that relax them both.

Horizontal concatenation of equations of the form (63) and then vertically stacking (vectorizing) yields the Kadiyala-Karlsson (1993) formulation

$$\mathbf{y}_T = (\mathbf{I}_p \otimes \mathbf{X}_T)\mathbf{b} + \mathbf{U}_T \tag{64}$$

where now $\mathbf{y}_T = vec(\mathbf{Y}_{1T}, \mathbf{Y}_{2T}, ..., \mathbf{Y}_{pT})$, $\mathbf{b} = vec(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_p)$, and $\mathbf{U}_T = vec(\mathbf{u}_{1T}, \mathbf{u}_{2T}, ..., \mathbf{u}_{pT})$. Here $\mathbf{U}_T \sim N(0, \boldsymbol{\Psi} \otimes \mathbf{I}_T)$. The Minnesota prior treats $var(\mathbf{u}_{iT})$ as fixed (at the unrestricted OLS estimate $\hat{\sigma}_i$) and $\boldsymbol{\Psi}$ as diagonal, and takes, for autoregression model A,

$$\boldsymbol{\beta}_i | A \sim N(\underline{\boldsymbol{\beta}}_i, \underline{\boldsymbol{\Sigma}}_i)$$

where $\underline{\boldsymbol{\beta}}_i$ and $\underline{\boldsymbol{\Sigma}}_i$ are the prior mean and covariance hyperparameters. This formulation results in the Gaussian posteriors

$$\boldsymbol{\beta}_i | \mathbf{y}_T, A \sim N(\bar{\boldsymbol{\beta}}_i, \bar{\boldsymbol{\Sigma}}_i)$$

where (recall (58))

$$\bar{\boldsymbol{\beta}}_i = \bar{\boldsymbol{\Sigma}}_i(\underline{\boldsymbol{\Sigma}}_i^{-1}\underline{\boldsymbol{\beta}}_i + \hat{\sigma}_i^{-1}\mathbf{X}_T'\mathbf{Y}_{iT})$$

$$\bar{\boldsymbol{\Sigma}}_i = (\underline{\boldsymbol{\Sigma}}_i^{-1} + \hat{\sigma}_i^{-1}\mathbf{X}_T'\mathbf{X}_T)^{-1}.$$

Kadiyala and Karlsson's first alternative is the "normal-Wishart" prior, which takes the VAR parameters to be Gaussian conditional on the innovation covariance matrix, and the covariance matrix not to be known but rather given by an inverted Wishart random matrix:

$$\mathbf{b}|\boldsymbol{\Psi} \sim N(\underline{\mathbf{b}}, \boldsymbol{\Psi} \otimes \underline{\boldsymbol{\Omega}}) \tag{65}$$

$$\boldsymbol{\Psi} \sim IW(\underline{\boldsymbol{\Psi}}, \alpha)$$

where the inverse Wishart density for $\boldsymbol{\Psi}$ given degrees of freedom parameter α and "shape" $\underline{\boldsymbol{\Psi}}$ is proportional to $|\boldsymbol{\Psi}|^{-(\alpha+p+1)/2}\exp\{-0.5tr\boldsymbol{\Psi}^{-1}\underline{\boldsymbol{\Psi}}\}$ (see, e.g., Zellner, 1971, p. 395.) This prior is the natural conjugate prior for $\mathbf{b}$, $\boldsymbol{\Psi}$. The posterior is given by

$$\mathbf{b}|\boldsymbol{\Psi}, \mathbf{y_T}, A \sim N(\bar{\mathbf{b}}, \boldsymbol{\Psi} \otimes \bar{\boldsymbol{\Omega}})$$

$$\boldsymbol{\Psi}|\mathbf{y_T}, A \sim IW(\bar{\boldsymbol{\Psi}}, T + \alpha)$$

where the posterior parameters $\bar{\mathbf{b}}$, $\bar{\boldsymbol{\Omega}}$, and $\bar{\boldsymbol{\Psi}}$ are simple (though notationally cumbersome) functions of the data and the prior parameters $\underline{\mathbf{b}}$, $\underline{\boldsymbol{\Omega}}$, and $\underline{\boldsymbol{\Psi}}$. Simple functions of interest can be evaluated analytically under this posterior, and for more complicated functions, evaluation by posterior simulation is trivial given the ease of sampling from the inverted Wishart (see, e.g., Geweke, 1988).

But this formulation has a drawback, noted long ago by Rothenberg (1963), that the Kroneker structure of the prior covariance matrix enforces an unfortunate symmetry on ratios of posterior variances of parameters. To take an example, suppress deterministic components ($d = 0$) and consider a 2-variable, 1-lag system ($p = 2$, $m = 1$) :

$$\begin{aligned} y_{1t} &= B_{1,11}y_{1t-1} + B_{1,12}y_{2t-1} + \varepsilon_{1t} \\ y_{2t} &= B_{1,21}y_{1t-1} + B_{1,22}y_{2t-1} + \varepsilon_{2t} \end{aligned}$$

Let $\boldsymbol{\Psi} = [\psi_{ij}]$ and $\bar{\boldsymbol{\Omega}} = [\bar{\sigma}_{ij}]$Then the posterior covariance matrix for $\mathbf{b} = (B_{1,11}\ B_{1,12}\ B_{1,21}\ B_{1,22})'$ is given by

$$\boldsymbol{\Psi} \otimes \bar{\boldsymbol{\Omega}} = \begin{bmatrix} \psi_{11}\bar{\sigma}_{11} & \psi_{11}\bar{\sigma}_{12} & \psi_{12}\bar{\sigma}_{11} & \psi_{12}\bar{\sigma}_{12} \\ \psi_{11}\bar{\sigma}_{21} & \psi_{11}\bar{\sigma}_{22} & \psi_{12}\bar{\sigma}_{21} & \psi_{12}\bar{\sigma}_{22} \\ \psi_{21}\bar{\sigma}_{11} & \psi_{21}\bar{\sigma}_{12} & \psi_{22}\bar{\sigma}_{11} & \psi_{22}\bar{\sigma}_{12} \\ \psi_{21}\bar{\sigma}_{21} & \psi_{21}\bar{\sigma}_{22} & \psi_{22}\bar{\sigma}_{21} & \psi_{22}\bar{\sigma}_{22} \end{bmatrix},$$

so that

$$\begin{aligned} var(B_{1,11})/var(B_{1,21}) &= \psi_{11}\bar{\sigma}_{11}/\psi_{22}\bar{\sigma}_{11} \\ &= var(B_{1,12})/var(B_{1,22}) = \psi_{11}\bar{\sigma}_{22}/\psi_{22}\bar{\sigma}_{22}. \end{aligned}$$

That is, under the normal-Wishart prior, the ratio of the posterior variance of the "own" lag coefficient in equation 1 to that of the "other" lag coefficient in equation 2 is identical to the ratio of the posterior variance of the "other" lag coefficient in equation 1 to the "own" lag coefficient in equation 2: ψ_{11}/ψ_{22} This is a very unattractive feature in general, and runs counter to the spirit of the Minnesota prior view that there is greater certainty about each equation's "own" lag coefficients than the "others". As Kadiyala and Karlsson (1993) put it, this "force(s) us to treat all equations symmetrically."

Like the normal-Wishart prior, the "diffuse" prior

$$p(\mathbf{b}, \mathbf{\Psi}) \propto |\mathbf{\Psi}|^{-(p+1)/2} \tag{66}$$

results in a posterior with the same form as the likelihood, with

$$\mathbf{b}|\mathbf{\Psi} \sim N(\hat{\mathbf{b}}, \mathbf{\Psi} \otimes (\mathbf{X}_{\mathbf{T}}^{'}\mathbf{X}_{\mathbf{T}})^{-1})$$

where now $\hat{\mathbf{b}}$ is the ordinary least squares (equation-by-equation, of course) estimator of $\mathbf{b}$, and the marginal density for $\mathbf{\Psi}$ is again of the inverted Wishart form. Symmetric treatment of all equations is also feature of this formulation owing to the product form of the covariance matrix. Yet this formulation has found application (see, e.g., section 5.2) because its use is very straightforward.

With the "Normal-diffuse" prior

$$\mathbf{b} \sim N(\underline{\mathbf{b}}, \underline{\mathbf{\Sigma}})$$
$$p(\mathbf{\Psi}) \propto |\mathbf{\Psi}|^{-(p+1)/2}$$

of Zellner (1971, p. 239), Kadiyala and Karlsson (1993) relaxed the implicit symmetry assumption at the cost of an analytically intractable posterior. Indeed, Zellner had advocated the prior two decades earlier, arguing that "the price is well worth paying". Zellner's approach to the analytic problem was to integrate $\mathbf{\Psi}$ out of the joint posterior for $\mathbf{b}$, $\mathbf{\Psi}$ and to approximate the result (a product of generalized multivariate Student t and multivariate Gaussian densities) using the leading (Gaussian) term in a Taylor series expansion. This approximation has a form not unlike (65), with mean given by a matrix-weighted average of the OLS estimator and the prior mean. Indeed, the similarity of Litterman's initial attempts to treat residual variances in his prior as unknown, which he regarded as computationally expensive at the time, to Zellner's straightforward approximation apparently led Litterman to abandon pursuit of a fully Bayesian analysis in favor of the mixed estimation strategy. But by the time Kadiyala and Karlsson (1993) appeared, initial development of fast posterior simulators (e.g., Drèze, 1977; Kloek and van Dijk, 1978; Drèze and Richard, 1983; and Geweke, 1989) had occurred, and they proceeded to utilize importance-sampling-based Monte Carlo methods for this normal-diffuse prior and a fourth, extended natural conjugate prior (Drèze and Morales, 1976), with only a small apology: "Following Kloek and van Dijk (1978), we have chosen to evaluate equation (5) using Monte Carlo integration instead of standard numerical integration techniques. Standard numerical integration is relatively inefficient when the integral has a high dimensionality ...."

A natural byproduct of the adoption of posterior simulation is the ability to work with the correct predictive density without resort to the approximations used by Litterman (1979), Doan, Litterman, and Sims (1984), and other successors. Indeed, Kadiyala and Karlsson's (1993) equation "(5)" is precisely the posterior mean of the predictive density (our (23)) with which they were working. (This is not the *first* such treatment, as production forecasts from full predictive densities have been issued for Iowa tax revenues (see Section 6.2) since 1990, and the shell code for carrying out such calculations in the diffuse prior case appeared in the RATS manual in the late 1980's.)

Kadiyala and Karlsson (1993) conducted three small forecasting "horse race" competitions amongst the four priors, using hyperparameters similar to those recommended by Doan, Litterman, and Sims (1984). Two experiments involved quarterly Canadian M2 and real GNP from 1955 to 1977; the other involved monthly data on the U.S. price of wheat, along with wheat export shipments and sales, and an exchange rate index for the U.S. dollar. In a small sample of the Canadian data, the normal-diffuse prior won, followed closely by the extended-natural-conjugate and Minnesota priors; in a larger data set, the normal-diffuse prior was the clear winner. For the monthly wheat data, no one procedure dominated, though priors that allowed for dependencies across equation parameters were generally superior.

Four years later, Kadiyala and Karlsson (1997) analyzed the same four priors, but by then the focus had shifted from the pure forecasting performance of the various priors to the numerical performance of posterior samplers and associated predictives. Indeed, Kadiyala and Karlsson (1997) provide both importance sampling and Gibbs sampling schemes for simulating from each of the posteriors they considered, and provide information regarding numerical efficiencies of the simulation procedures.

Sims and Zha (1999), which was submitted for publication in 1994, and Sims and Zha (1998), completed the Bayesian treatment of the VAR by generalizing procedures for implementing prior views regarding the structure of cross-equation errors. In particular, they wrote (3) in the form

$$\mathbf{C}_0\mathbf{y}_t = \mathbf{C}_D D_t + \mathbf{C}_1\mathbf{y}_{t-1} + \mathbf{C}_2\mathbf{y}_{t-2} + ... + \mathbf{C}_m\mathbf{y}_{t-m} + \mathbf{u}_t \tag{67}$$

with

$$E\mathbf{u}_t\mathbf{u}_t' = \mathbf{I}$$

which accommodates various identification schemes for $\mathbf{C}_0$. For example, one route for passing from (3) to (67) is via "Choleski factorization" of $\mathbf{\Sigma}$ as $\mathbf{\Sigma} = \mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2'}$ so that $\mathbf{C}_0 = \mathbf{\Sigma}^{-1/2}$ and $u_t = \Sigma^{-1/2}\varepsilon_t$. This results in exact identification of parameters in $\mathbf{C}_0$, but other "overidentification" schemes are possible as well. Sims and Zha (1999) worked directly with the likelihood, thus implicitly adopting a diffuse prior for $\mathbf{C}_0, \mathbf{C}_D, \mathbf{C}_1, ..., \mathbf{C}_m$. They showed that conditional on $\mathbf{C}_0$, the posterior ("likelihood") for the other parameters is Gaussian, but the marginal for $\mathbf{C}_0$ is not of any standard form. They indicated how to sample from it using importance sampling, but in application used a random walk

Metropolis-chain procedure utilizing a multivariate-t candidate generator. Subsequently, Sims and Zha (1998) showed how to adopt an informative Gaussian prior for $\mathbf{C}_D, \mathbf{C}_1, ..., \mathbf{C}_m|\mathbf{C}_0$ together with a general (diffuse or informative) prior for $\mathbf{C}_0$ and concluded with the "hope that this will allow the transparency and reproducibility of Bayesian methods to be more widely available for tasks of forecasting and policy analysis." (p. 967)

5 Some Bayesian forecasting models

The vector autoregression (VAR) is the best known and most widely applied Bayesian economic forecasting model. It has been used in many contexts, and its ability to improve forecasts and provide a vehicle for communicating uncertainty is by now well established. We return to a specific application of the VAR illustrating these qualities in Section 6. In fact Bayesian inference is now widely undertaken with many models, for a variety of applications including economic forecasting. This section surveys a few of the models most commonly used in economics. Some of these, for example ARMA and fractionally integrated models, have been used in conjunction with methods that are not only non-Bayesian but are also not likelihood-based because of the intractability of the likelihood function. The technical issues that arise in numerical maximization of the likelihood function, on the one hand, and the use of simulation methods in computing posterior moments, on the other, are distinct. It turns out, in these cases as well as in many other econometric models, that the Bayesian integration problem is easier to solve than is the non-Bayesian optimization problem. We provide some of the details in Sections 5.2 and 5.3 below.

The state of the art in inference and computation is an important determinant of which models have practical application and which do not. The rapid progress in posterior simulators since 1990 is an increasingly important influence in the conception and creation of new models. Some of these models would most likely never have been substantially developed, or even emerged, without these computational tools, reviewed in Section 3. An example is the stochastic volatility model, introduced in Section 2.1.2 and discussed in greater detail in Section 5.5 below. Another example is the state space model, often called the dynamic linear model in the statistics literature, which is described briefly in Section 4.2 and in more detail in Chapter **Harvey chapter** of this volume. The monograph by West and Harrison (1997) provides detailed development of the Bayesian formulation of this model, and that by Pole, West and Harrison (1994) is devoted to the practical aspects of Bayesian forecasting.

These models all carry forward the theme so important in vector autoregressions: priors matter, and in particular priors that cope sensibly with an otherwise profligate parameterization are demonstrably effective in improving forecasts. That was true in the earliest applications when computational tools were very limited, as illustrated in Section 4 for VARs, and here for autoregressive leading indicator models (Section 5.1). This fact has become even more striking as computational tools have become more sophisticated. The review

of cointegration and error correction models (Section 5.4) constitutes a case study in point. More generally models that are preferred, as indicated by Bayes factors, should lead to better decisions, as measured by ex post loss, for the reasons developed in Sections 2.3.2 and 2.4.1. This section closes with such a comparison for time-varying volatility models.

5.1 Autoregressive leading indicator models

In a series of papers (Garcia-Ferrer et al. (1987), Zellner and Hong (1989), Zellner et al. (1990), Zellner et al. (1991), Min and Zellner (1993)) Zellner and coauthors investigated the use of leading indicators, pooling, shrinkage, and time-varying parameters in forecasting real output for the major industrialized countries. In every case the variable modeled was the growth rate of real output; there was no presumption that real output is cointegrated across countries. The work was carried out entirely analytically, using little beyond what was available in conventional software at the time, which limited attention almost exclusively to one-step-ahead forecasts. A principal goal of these investigations was to improve forecasts significantly using relatively simple models and pooling techniques.

The observables model in all of these studies is of the form

$$y_{it} = \alpha_0 + \sum_{s=1}^{3} \alpha_s y_{i,t-s} + \boldsymbol{\beta}' \mathbf{z}_{i,t-1} + \varepsilon_{it}, \;\; \varepsilon_{it} \overset{iid}{\sim} N\left(0, \sigma^2\right), \tag{68}$$

with y_{it} denoting the growth rate in real GNP or real GDP between year $t-1$ and year t in country i. The vector $\mathbf{z}_{i,t-1}$ comprises the leading indicators. In Garcia-Ferer et al. (1987) and Zellner and Hong (1989) $\mathbf{z}_{it}$ consisted of real stock returns in country i in years $t-1$ and t, the growth rate in the real money supply between years $t-1$ and t, and world stock return defined as the median real stock return in year t over all countries in the sample. Attention was confined to nine OECD countries in Garcia-Ferer et al. (1987). In Zellner and Hong (1989) the list expanded to 18 countries but the original group was reported separately, as well, for purposes of comparison.

The earliest study, Garcia-Ferer et al. (1987), considered five different forecasting procedures and several variants on the right-hand-side variables in (68). The period 1954-1973 was used exclusively for estimation, and one-step-ahead forecast errors were recorded for each of the years 1974 through 1981, with estimates being updated before each forecast was made. Results for root mean square forecast error, expressed in units of growth rate percentage, are as fol-

lows.

Summary of forecast RMSE for 9 countries in Garcia-Ferer et al. (1987)					
Estimation method:	(None)	OLS	TVP	Pool	Shrink 1
Growth rate = 0	3.09				
Random walk growth rate	3.73				
AR(3)		3.46			
AR(3)-LI1		2.70	2.52	3.08	
AR(3)-LI2		2.39		2.62	
AR(3)-LI3		2.23	1.82	2.22	1.78

The model LI1 includes only the two stock returns in $\mathbf{z}_{it}$; LI2 adds the world stock return and LI3 adds also the growth rate in the real money supply. The time varying parameter (TVP) model utilizes a conventional state-space representation in which the variance in the coefficient drift is $\sigma^2/2$. The pooled models constrain the coefficients in (68) to be the same for all countries. In the variant "Shrink 1" each country forecast is an equally- weighted average of the own country forecast and the average forecast for all nine countries; unequally-weighted averages (unreported here) produce somewhat higher root mean square error of forecast.

The subsequent study by Zellner and Hong (1989) extended this work by adding nine countries, extending the forecasting exercise by three years, and considering an alternative shrinkage procedure. In the alternative, the coefficients estimates are taken to be a weighted average of the least squares estimates for the country under consideration, and the pooled estimates using all the data. The study compared several weighting schemes, and found that a weight of one-sixth on the country estimates and five-sixths on the pooled estimates minimized the out-of-sample forecast root mean square error. These results are reported in the column "Shrink 2" in the following table.

Summary of forecast RMSE for 18 countries in Zellner and Hong (1989)					
Estimation method:	(None)	OLS	Pool	Shrink 1	Shrink 2
Growth rate = 0	3.07				
Random walk growth rate	3.02				
Growth rate = Past average	3.09				
AR(3)		3.00			
AR(3)-LI3		2.62	2.14	2.32	2.13

Garcia-Ferer et al. (1987) and Zellner and Hong (1989) demonstrated the returns both to the incorporation of leading indicators and to various forms of pooling and shrinkage. Combined, these two methods produce root mean square errors of forecast somewhat smaller than those of considerably more complicated OECD official forecasts (see Smyth (1983)), as described in Garcia-Ferer et al. (1987) and Zellner and Hong (1989). A subsequent investigation by Min and Zellner (1993) computed formal posterior odds ratios between the most competitive models. Consistent with the results described here, they found that odds rarely exceeded 2:1 and that there was no systematic gain from combining forecasts.

5.2 Stationary linear models

Many routine forecasting situations involve linear models of the form $y_t = \boldsymbol{\beta}'\mathbf{x}_t + \boldsymbol{\varepsilon}_t$, in which $\boldsymbol{\varepsilon}_t$ is a stationary process, and the covariates $\mathbf{x}_t$ are ancillary – for example they may be deterministic (e.g., calendar effects in asset return models), they may be controlled (e.g. traditional reduced form policy models), or they may be exogenous and modelled separately from the relationship between $\mathbf{x}_t$ and y_t.

5.2.1 The stationary AR(p) model

One of the simplest models of serial correlation in $\boldsymbol{\varepsilon}_t$ is an autoregression of order p. The contemporary Bayesian treatment of this problem (see Chib and Greenberg (1994b) or Geweke (2005, Section 4.8)) exploits the structure of MCMC posterior simulation algorithms, and the Gibbs sampler in particular, by decomposing the posterior distribution into manageable conditional distributions for each of several groups of parameters.

Suppose

$$\varepsilon_t = \sum_{s=1}^{p} \phi_s \varepsilon_{t-s} + u_t,\ u_t \mid (\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots) \overset{iid}{\sim} N\left(0, h^{-1}\right),$$

and $\boldsymbol{\phi} = \left(\phi_1, \ldots, \phi_p\right)' \in S_p = \{\boldsymbol{\phi} : |1 - \sum_{s=1}^{p} \phi_s z^s| \neq 0 \ \forall\ z : |z| \leq 1\} \subseteq \mathbb{R}^p$. There are three groups of parameters: $\boldsymbol{\beta}, \boldsymbol{\phi}$, and h. Conditional on $\boldsymbol{\phi}$, the likelihood function is of the classical generalized least squares form and reduces to that of ordinary least squares by means of appropriate linear transformations. For $t = p+1, \ldots, T$ these transformations amount to $y_t^* = y_t - \sum_{s=1}^{p} \phi_s y_{t-s}$ and $\mathbf{x}_t^* = \mathbf{x}_t - \sum_{s=1}^{p} \mathbf{x}_{t-s}\phi_s$. For $t = 1, \ldots, p$ the p Yule-Walker equations

$$\begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \cdots & \rho_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \cdots & 1 \end{bmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{pmatrix}$$

can be inverted to solve for the autocorrelation coefficients $\boldsymbol{\rho} = \left(\rho_1, \ldots, \rho_p\right)'$ as a linear function of $\boldsymbol{\phi}$. Then construct the $p \times p$ matrix $\mathbf{R}_p(\boldsymbol{\phi}) = \left[\rho_{|i-j|}\right]$, let $\mathbf{A}_p(\boldsymbol{\rho})$ be a Choleski factor of $[\mathbf{R}_p(\boldsymbol{\phi})]^{-1}$, and then take $\left(y_1^*, \ldots, y_p^*\right)' = \mathbf{A}_p(\boldsymbol{\rho})\left(y_1, \ldots, y_p\right)'$. Creating $\mathbf{x}_1^*, \ldots, \mathbf{x}_p^*$ by means of the same transformation, the linear model $y_t^* = \boldsymbol{\beta}'\mathbf{x}_t^* + \varepsilon_t^*$ satisfies the assumptions of the textbook normal linear model. Given a normal prior for $\boldsymbol{\beta}$ and a gamma prior for h, the conditional posterior distributions come from these same families; variants on these prior distributions are straightforward; see Geweke (2005, Sections 2.1 and 5.3).

On the other hand, conditional on $\boldsymbol{\beta}$, h, $\mathbf{X}$ and $\mathbf{y}^o$,

$$\mathbf{e} = \begin{pmatrix} \varepsilon_{p+1} \\ \varepsilon_{p+2} \\ \vdots \\ \varepsilon_T \end{pmatrix} \quad \text{and} \quad \mathbf{E} = \begin{bmatrix} \varepsilon_p & \cdots & \varepsilon_1 \\ \varepsilon_{p+1} & \cdots & \varepsilon_2 \\ \vdots & & \vdots \\ \varepsilon_{T-1} & \cdots & \varepsilon_{T-p} \end{bmatrix}$$

are known. Further denoting $\mathbf{X}_p = [\mathbf{x}_1, \ldots, \mathbf{x}_p]'$ and $\mathbf{y}_p = (y_1, \ldots, y_p)'$, the likelihood function is

$$p\left(\mathbf{y}^o \middle| \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\phi}, h\right) = (2\pi)^{-T/2} h^{T/2} \exp\left[-h\left(\mathbf{e} - \mathbf{E}\boldsymbol{\phi}\right)'\left(\mathbf{e} - \mathbf{E}\boldsymbol{\phi}\right)/2\right] \tag{69}$$

$$\cdot \left|\mathbf{R}_p\left(\boldsymbol{\phi}\right)\right|^{-1/2} \exp\left[-h\left(\mathbf{y}_p^o - \mathbf{X}_p\boldsymbol{\beta}\right)'\mathbf{R}_p\left(\boldsymbol{\phi}\right)^{-1}\left(\mathbf{y}_p^o - \mathbf{X}_p\boldsymbol{\beta}\right)/2\right]. \tag{70}$$

The expression (69), treated as a function of $\boldsymbol{\phi}$, is the kernel of a p-variate normal distribution. If the prior distribution of $\boldsymbol{\phi}$ is Gaussian, truncated to S_p, then the same is true of the product of this prior and (69). (Variants on this prior can be accommodated through reweighting as discussed in Section 3.3.2.) Denote expression (70) as $r\left(\boldsymbol{\beta}, h, \boldsymbol{\phi}\right)$, and note that, interpreted as a function of $\boldsymbol{\phi}$, $r\left(\boldsymbol{\beta}, h, \boldsymbol{\phi}\right)$ does not correspond to the kernel of any tractable multivariate distribution. This apparent impediment to an MCMC algorithm can be addressed by means of a Metropolis within Gibbs step, as discussed in Section 3.2.3. At iteration m a Metropolis within Gibbs step for $\boldsymbol{\phi}$ draws a candidate $\boldsymbol{\phi}^*$ from the Gaussian distribution whose kernel is the product of the untruncated Gaussian prior distribution of $\boldsymbol{\phi}$ and (69), using the current values $\boldsymbol{\beta}^{(m)}$ of $\boldsymbol{\beta}$ and $h^{(m)}$ of h. From (70) the acceptance probability for the candidate is

$$\min\left[\frac{r\left(\boldsymbol{\beta}^{(m)}, h^{(m)}, \boldsymbol{\phi}^*\right) I_{S_p}\left(\boldsymbol{\phi}^*\right)}{r\left(\boldsymbol{\beta}^{(m)}, h^{(m)}, \boldsymbol{\phi}^{(m-1)}\right)}, \ 1\right].$$

5.2.2 The stationary ARMA(p,q) model

The incorporation of a moving average component

$$\varepsilon_t = \sum_{s=1}^{p} \phi_s \varepsilon_{t-s} + \sum_{s=1}^{q} \theta_s u_{t-s} + u_t$$

adds the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ and complicates the recursive structure. The first broad-scale attack on the problem was Monahan (1983) who worked without the benefit of modern posterior simulation methods and was able to treat only $p + q \leq 2$. Nevertheless he produced exact Bayes factors for five alternative models, and obtained up to four-step ahead predictive means and standard deviations for each model. He applied his methods in several examples developed originally in Box and Jenkins (1976). Chib and Greenberg (1994b) and Marriott et al. (1996) approached the problem by means of data

augmentation, adding unobserved pre-sample values to the vector of unobservables. In Marriott et al. (1996) the augmented data are $\boldsymbol{\varepsilon}_0 = (\varepsilon_0, \ldots, \varepsilon_{1-p})'$ and $\mathbf{u}_0 = (u_0, \ldots, u_{1-q})'$. Then (see Marriott et al. (1996, pp. 245-246))

$$p(\varepsilon_1, \ldots, \varepsilon_T \mid \boldsymbol{\phi}, \boldsymbol{\theta}, h, \boldsymbol{\varepsilon}_0, \mathbf{u}_0) = (2\pi)^{-T/2} h^{T/2} \exp\left[-h \sum_{t=1}^{T} (\varepsilon_t - \mu_t)^2 / 2\right] \quad (71)$$

with

$$\mu_t = \sum_{s=1}^{p} \phi_s \varepsilon_{t-s} - \sum_{s=1}^{t-1} \theta_s (\varepsilon_{t-s} - \mu_{t-s}) - \sum_{s=t}^{q} \theta_s \varepsilon_{t-s}. \quad (72)$$

The data augmentation scheme is feasible because the conditional posterior density of $\mathbf{u}_0$ and $\boldsymbol{\varepsilon}_0$,

$$p(\boldsymbol{\varepsilon}_0, \mathbf{u}_0 \mid \boldsymbol{\phi}, \boldsymbol{\theta}, h, \mathbf{X}_T, \mathbf{y}_T) \quad (73)$$

is that of a Gaussian distribution and is easily computed (see Newbold (1974)). The product of (73) with the density corresponding to (71)-(72) yields a Gaussian kernel for the presample $\boldsymbol{\varepsilon}_0$ and $\mathbf{u}_0$. A draw from this distribution becomes one step in a Gibbs sampling posterior simulation algorithm. The presence of (73) prevents the posterior conditional distribution of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ from being Gaussian. This complication may be handled just as it was in the case of the AR(p) model, using a Metropolis within Gibbs step.

There are a number of variants on these approaches. Chib and Greenberg show that the data augmentation vector can be reduced to $\max(p, q+1)$ elements, with some increase in complexity. As an alternative to enforcing stationarity in the Metropolis within Gibbs step, the transformation of $\boldsymbol{\phi}$ to the corresponding vector of partial autocorrelations (see Barndorff-Nielsen and Schou (1973)) may be inverted and the Jacobian computed (see Monahan (1984)), thus transforming S_p to a unit hypercube. A similar treatment can restrict the roots of $1 - \sum_{s=1}^{q} \theta_s z^s$ to the exterior of the unit circle (see Marriott et al. (1996)).

There are no new essential complications introduced in extending any of these models or posterior simulators from univariate (ARMA) to multivariate (VARMA) models. On the other hand, VARMA models lead to large numbers of parameters as the number of variables increases, just as in the case of VAR models. The BVAR strategy of using shrinkage prior distributions appears not to have been applied in VARMA models. The approach has been, instead, to utilize exclusion restrictions for many parameters, the same strategy used in non-Bayesian approaches. In a Bayesian set-up, however, uncertainty about exclusion restrictions can be incorporated in posterior and predictive distributions. Ravishanker and Ray (1997a) do exactly this, in extending the model and methodology of Marriott et al. (1996) to VARMA models. Corresponding to each autoregressive coefficient ϕ_{ijs} there is a multiplicative Bernoulli random variable γ_{ijs}, indicating whether that coefficient is excluded, and similarly for each moving average coefficient θ_{ijs} there is a Bernoulli random variable δ_{ijs}:

$$y_{it} = \sum_{j=1}^{n} \sum_{s=1}^{p} \gamma_{ijs} \phi_{ijs} y_{j,t-s} + \sum_{j=1}^{n} \sum_{s=1}^{q} \theta_{ijs} \delta_{ijs} \varepsilon_{j,t-s} + \varepsilon_{it} \quad (i = 1, \ldots, n).$$

Prior probabilities on these random variables may be used to impose parsimony, both globally and also differentially at different lags and for different variables; independent Bernoulli prior distributions for the parameters γ_{ijs} and δ_{ijs}, embedded in a hierarchical prior with beta prior distributions for the probabilities, are the obvious alternatives to *ad hoc* non-Bayesian exclusion decisions, and are quite tractable. The conditional posterior distributions of the γ_{ijs} and δ_{ijs} are individually conditionally Bernoulli. This strategy is one of a family of similar approaches to exclusion restrictions in regression models (see George and McCulloch (1993) or Geweke (1996b)) and has also been employed in univariate ARMA models (see Barnett et al. (1996)). The posterior MCMC sampling algorithm for the parameters ϕ_{ijs} and δ_{ijs} also proceeds one parameter at a time; Ravishanker and Ray (1997a) report that this algorithm is computationally efficient in a three-variable VARMA model with $p = 3$, $q = 1$, applied to a data set with 75 quarterly observations.

5.3 Fractional integration

Fractional integration, also known as long memory, first drew the attention of economists because of the improved multi-step-ahead forecasts provided by even the simplest variants of these models as reported in Granger and Joyeux (1980) and Porter-Hudak (1982). In a fractionally integrated model $(1-L)^d y_t = u_t$, where

$$(1-L)^d = \sum_{j=0}^{\infty} \left(\begin{array}{c} d \\ j \end{array} \right) (-L)^j = \sum_{j=1}^{\infty} \frac{(-1)^j \Gamma(d-1)}{\Gamma(j-1)\Gamma(d-j-1)} L^j$$

and u_t is a stationary process whose autocovariance function decays geometrically. The fully parametric version of this model typically specifies

$$\phi(L)(1-L)^d (y_t - \mu) = \theta(L)\varepsilon_t, \tag{74}$$

with $\phi(L)$ and $\theta(L)$ being polynomials of specified finite order and ε_t being serially uncorrelated; most of the literature takes $\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$. Sowell (1992a, 1992b) first derived the likelihood function and implemented a maximum likelihood estimator. Koop et al. (1997) provided the first Bayesian treatment, employing a flat prior distribution for the parameters in $\phi(L)$ and $\theta(L)$, subject to invertibility restrictions. This study used importance sampling of the posterior distribution, with the prior distribution as the source distribution. The weighting function $w(\boldsymbol{\theta})$ is then just the likelihood function, evaluated using Sowell's computer code. The application in Koop et al. (1997) used quarterly US real GNP, 1947-1989, a standard data set for fractionally integrated models, and polynomials in $\phi(L)$ and $\theta(L)$ up to order 3. This study did not provide any evaluation of the efficiency of the prior density as the source distribution in the importance sampling algorithm; in typical situations this will be poor if there are a half-dozen or more dimensions of integration. In any event, the computing

times reported[3] indicate that subsequent more sophisticated algorithms are also much faster.

Much of the Bayesian treatment of fractionally integrated models originated with Ravishanker and coauthors, who applied these methods to forecasting. Pai and Ravishanker (1996) provided a thorough treatment of the univariate case based on a Metropolis random-walk algorithm. Their evaluation of the likelihood function differs from Sowell's. From the autocovariance function $r(s)$ corresponding to (74) given in Hosking (1981) the Levinson-Durbin algorithm provides the partial regression coefficients ϕ_j^k in

$$\mu_t = E(y_t \mid \mathbf{Y}_{t-1}) = \sum_{j=1}^{t-1} \phi_j^{t-1} y_{t-j}. \tag{75}$$

The likelihood function then follows from

$$y_t \mid \mathbf{Y}_{t-1} \sim N\left(\mu_t, \nu_t^2\right), \nu_t^2 = \left[r(0)/\sigma^2\right] \prod_{j=1}^{t-1} \left[1 - \left(\phi_j^j\right)^2\right]. \tag{76}$$

Pai and Ravishanker (1996) computed the maximum likelihood estimate as discussed in Haslett and Raftery (1989). The observed Fisher information matrix is the variance matrix used in the Metropolis random-walk algorithm, after integrating μ and σ^2 analytically from the posterior distribution. The study focused primarily on inference for the parameters; note that (75)-(76) provide the basis for sampling from the predictive distribution given the output of the posterior simulator.

A multivariate extension of (74), without cointegration, may be expressed

$$\Phi(L)\mathbf{D}(L)(\mathbf{y}_t - \boldsymbol{\mu}) = \Theta(L)\boldsymbol{\varepsilon}_t$$

in which $\mathbf{y}_t$ is $n \times 1$, $\mathbf{D}(L) = diag\left[(1-L)^{d_1}, \ldots, (1-L)^{d_n}\right]$, $\Phi(L)$ and $\Theta(L)$ are $n \times n$ matrix polynomials in L of specified order, and $\boldsymbol{\varepsilon}_t \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma})$. Ravishanker and Ray (1997, 2002) provided an exact Bayesian treatment and a forecasting application of this model. Their approach blends elements of Marriott et al. (1996) and Pai and Ravishanker (1996). It incorporates presample values of $\mathbf{z}_t = \mathbf{y}_t - \boldsymbol{\mu}$ and the pure fractionally integrated process $\mathbf{a}_t = \mathbf{D}(L)^{-1}\boldsymbol{\varepsilon}_t$ as latent variables. The autocovariance function $\mathbf{R}^a(s)$ of $\mathbf{a}_t$ is obtained recursively from

$$r^a(0)_{ij} = \sigma_{ij}\frac{\Gamma(1-d_i-d_j)}{\Gamma(1-d_i)\Gamma(1-d_j)}, \quad r^a(s)_{ij} = -\frac{1-d_i-s}{s-d_j} r^a(s-1)_{ij}.$$

The autocovariance function of $\mathbf{z}_t$ is then

$$\mathbf{R}^z(s) = \sum_{i=1}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{\Psi}_i \mathbf{R}^a(s+i-j)\boldsymbol{\Psi}_j'$$

[3] Contrast Koop et al. (1997, footnote 12) with Pai and Ravishanker (1996, p. 74).

where the coefficients $\mathbf{\Psi}_j$ are those in the moving average representation of the ARMA part of the process. Since these decay geometrically, truncation is not a serious issue. This provides the basis for a random walk Metropolis-within-Gibbs step constructed as in Pai and Ravishanker (1996). The other blocks in the Gibbs sampler are the pre-sample values of $\mathbf{z}_t$ and $\mathbf{a}_t$, plus $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$. The procedure requires on the order of n^3T^2 operations and storage of order n^2T^2; $T = 200$ and $n = 3$ requires a gigabyte of storage. If likelihood is computed conditional on all presample values being zero the problem is computationally much less demanding, but results differ substantially.

Ravishanker and Ray (2002) provide details of drawing from the predictive density, given the output of the posterior simulator. Since the presample values are a by-product of each iteration, the latent vectors $\mathbf{a}_t$ can be computed by means of $\mathbf{a}_t = -\sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{z}_{t-i} + \sum_{i=1}^{q} \mathbf{\Theta}_r \mathbf{a}_{t-r}$. Then sample $\mathbf{a}_t$ forward using the autocovariance function of the pure long-memory process, and finally apply the ARMA recursions to these values. The paper applies a simple version of the model ($n = 3$; $q = 0$; $p = 0$ or 1) to sea temperatures off the California coast. The coefficients of fractional integration are all about 0.4 when $p = 0$; $p = 1$ introduces the usual difficulties in distinguishing between long memory and slow geometric decay of the autocovariance function. There are substantial interactions in the off-diagonal elements of $\mathbf{\Phi}(L)$, but the study does not take up fractional cointegration.

5.4 Cointegration and error correction

Cointegration restricts the long-run behavior of multivariate time series that are otherwise nonstationary; see **OTHER CHAPTERS** for details. Error correction models (ECMs; **ANOTHER REFERENCE TO OTHER CHAPTERS**) provide a convenient representation of cointegration, and there is by now an enormous literature on inference in these models. By restricting the behavior of otherwise nonstationary time series, cointegration also has the promise of improving forecasts, especially at longer horizons. Coming hard on the heels of Bayesian vector autoregressions, ECMs were at first thought to be competitors of VARs:

> One could also compare these results with estimates which are obviously misspecified such as least squares on differences or Litterman's Bayesian Vector Autoregression which shrinks the parameter vector toward the first difference model which is itself misspecified for this system. The finding that such methods provided inferior forecasts would hardly be surprising. (Engle and Yoo (1987, p. 157))

Shoesmith (1995) carefully compared and combined the error correction specification and the prior distributions pioneered by Litterman, with illuminating results. He used the quarterly, six-lag VAR in Litterman (1980) for real GNP, the implicit GNP price deflator, real gross private domestic investment, the three-month treasury bill rate and the money supply (M1). Throughout

the exercise, Shoesmith repeatedly tested for lag length and the outcome consistently indicated six lags. The period 1959:1 through 1981:4 was the base estimation period, followed by 20 successive five-year experimental forecasts: the first was for 1982:1 through 1986:4; and the last was for 1986:4 through 1991:3 based on estimates using data from 1959:1 through 1986:3. Error correction specification tests were conducted using standard procedures (see Johansen (1988)). For all the samples used, these procedures identified the price deflator as I(2), all other variables as I(1), and two cointegrating vectors.

Shoesmith compared forecasts from Litterman's model with six other models. One, VAR/I1, was a VAR in I(1) series (i.e., first differences for the deflator and levels for all other variables) estimated by least squares, not incorporating any shrinkage or other prior. The second, ECM, was a conventional ECM, again with no shrinkage. The other four models all included the Minnesota prior. One of these models, BVAR/I1, differs from Litterman's model only in replacing the deflator with its first difference. Another, BECM, applies the Minnesota prior to the conventional ECM, with no shrinkage or other restrictions applied to the coefficients on the error correction terms. Yet another variant, BVAR/I0, applies the Minnesota prior to a VAR in I(0) variables (i.e., second differences for the deflator and first differences for all other variables). The final model, BECM/5Z, is identical to BECM except that five cointegrating relationships are specified, an intentional misreading of the outcome of the conventional procedure for determining the rank of the error correction matrix.

The paper offers an extensive comparison of root mean square forecasting errors for all of the variables. These are summarized here, by first forming the ratio of mean square error in each model to its counterpart in Litterman's model, and then averaging the ratios across the six variables.

Comparison of forecast RMSE in Shoesmith (1995)			
Horizon:	1 quarter	8 quarters	20 quarters
VAR/I1	1.33	1.00	1.14
ECM	1.28	0.89	0.91
BVAR/I1	0.97	0.96	0.85
BECM	0.89	0.72	0.45
BVAR/I0	0.95	0.87	0.59
BECM/5Z	0.99	1.02	0.88

The most notable feature of the results is the superiority of the BECM forecasts, which is realized at all forecasting horizons but becomes greater at more distant horizons. The ECM forecasts, by contrast, do not dominate those of either the original Litterman VAR or the BVAR/I1, contrary to the conjecture in Engle and Yoo (1987). The results show that most of the improvement comes from applying the Minnesota prior to a model that incorporates stationary time series: BVAR/I0 ranks second at all horizons, and the ECM without shrinkage performs poorly relative to BVAR/I0 at all horizons. In fact the VAR with the Minnesota prior and the error correction models are not competitors, but complementary methods of dealing with the profligate parameterization in multivariate time

series by shrinking toward reasonable models with fewer parameters. In the case of the ECM the shrinkage is a hard, but data driven, restriction, whereas in the Minnesota prior it is soft, allowing the data to override in cases where the more parsimoniously parameterized model is less applicable. The possibilities for employing both have hardly been exhausted. Shoesmith (1995) suggested that this may be a promising avenue for future research.

This experiment incorporated the Minnesota prior utilizing the mixed estimation methods described in Section 4.3, appropriate at the time to the investigation of the relative contributions of error correction and shrinkage in improving forecasts. More recent work has employed modern posterior simulators. A leading example is Villani (2001), which examined the inflation forecasting model of the central bank of Sweden. This model is expressed in error correction form

$$\Delta \mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{y}_{t-1} + \sum_{s=1}^{p} \boldsymbol{\Gamma}_s \Delta \mathbf{y}_{t-s} + \boldsymbol{\varepsilon}_t, \;\; \boldsymbol{\varepsilon}_t \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}). \tag{77}$$

It incorporates GDP, consumer prices and the three-month treasury rate, both Swedish and weighted averages of corresponding foreign series, as well as the trade-weighted exchange rate. Villani limits consideration to models in which $\boldsymbol{\beta}$ is 7×3, based on the bank's experience. He specifies four candidate coefficient vectors: for example, one based on purchasing power parity and another based on a Fisherian interpretation of the nominal interest rate given a stationary real rate. This forms the basis for competing models that utilize various combinations of these vectors in $\boldsymbol{\beta}$, as well as unknown cointegrating vectors. In the most restrictive formulations three vectors are specified and in the least restrictive all three are unknown. Villani specifies conventional uninformative priors for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, and conventional Minnesota priors for the parameters $\boldsymbol{\Gamma}_s$ of the short-run dynamics. The posterior distribution is sampled using a Gibbs sampler blocked in $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\{\boldsymbol{\Gamma}_s\}$ and $\boldsymbol{\Sigma}$.

The paper utilizes data from 1972:2 through 1993:3 for inference. Of all of the combinations of cointegrating vectors, Villani finds that the one in which all three are unrestricted is most favored. This is true using both likelihood ratio tests and an informal version (necessitated by the improper priors) of posterior odds ratios. This unrestricted specification ("$\boldsymbol{\beta}$ empirical" in the table below), as well as the most restricted one ("$\boldsymbol{\beta}$ specified"), are carried forward for the subsequent forecasting exercise. This exercise compares forecasts over the period 1994 - 1998, reporting forecast root mean square errors for the means of the predictive densities for price inflation ("Bayes ECM"). It also computes forecasts from the maximum likelihood estimates, treating these estimates as known coefficients ("ML unrestricted ECM"), and finds the forecast root mean square error. Finally, it constrains many of the coefficients to zero, using conventional stepwise deletion procedures in conjunction with maximum likelihood estimation, and again finds the forecast root mean square error. Taking averages of these root mean square errors over forecasting horizons of one to eight

quarters ahead yields the following comparison:

$\boldsymbol{\beta}$:	Specified	Empirical
Bayes ECM	0.485	0.488
ML unrestricted ECM	0.773	0.694
ML restricted ECM	0.675	0.532

The Bayesian ECM produces by far the lowest root mean square error of forecast, and results are about the same whether the restricted or unrestricted version of the cointegrating vectors are used. The forecasts based on restricted maximum likelihood estimates benefit from the additional restrictions imposed by stepwise deletion of coefficients, which is a crude from of shrinkage. In comparison with Shoesmith (1995), Villani (2001) has the further advantage of having used a full Monte Carlo simulation of the predictive density, whose mean is the Bayes estimate given a squared-error loss function.

These findings are supported by other studies that have made similar comparisons. An earlier literature on regional forecasting, of which the seminal paper is LeSage (1990), contains results that are broadly consistent but not directly comparable because of the differences in variables and data. Amisano and Serati (1999) utilized a three-variable VAR for Italian GDP, consumption and investment. Their approach was closer to mixed estimation than to full Bayesian inference. They employed not only a conventional Minnesota prior for the short-run dynamics, but also applied a shrinkage prior to the factor loading vector $\boldsymbol{\alpha}$ in (77). This combination produced a smaller root mean square error, for forecasts from one to twenty quarters ahead, than either a traditional VAR with a Minnesota prior, or an ECM that shrinks the short-run dynamics but not $\boldsymbol{\alpha}$.

5.5 Stochastic volatility

In classical linear processes, for example the vector autoregression (3), conditional means are time varying but conditional variances are not. By now it is well established that for many time series, including returns on financial assets, conditional variances in fact often vary greatly. Moreover, in the case of financial assets, conditional variances are fundamental to portfolio allocation. The ARCH family of models provides conditional variances that are functions of past realizations, likelihood functions that are relatively easy to evaluate, and a systematic basis for forecasting and solving the allocation problem. Stochastic volatility models provide an alternative approach, first motivated by autocorrelated information flows (see Tauchen and Pitts (1983)) and as discrete approximations to diffusion processes utilized in the continuous time asset pricing literature (see Hull and White (1987)). The canonical univariate model, introduced in Section 2.1.2, is

$$\begin{aligned} y_t &= \beta \exp\left(h_t/2\right)\varepsilon_t;\ \ h_t = \phi h_{t-1} + \sigma_\eta \eta_t; \\ h_1 &\sim N\left[0, \sigma_\eta^2/\left(1-\phi^2\right)\right];\ \ \left(\varepsilon_t, \eta_t\right)' \overset{iid}{\sim} N\left(\mathbf{0}, \mathbf{I}_2\right). \end{aligned} \tag{78}$$

Only the return y_t is observable. In the stochastic volatility model there are two shocks per time period, whereas in the ARCH family there is only one. As a consequence the stochastic volatility model can more readily generate extreme realizations of y_t. Such a realization will have an impact on the variance of future realizations if it arises because of an unusually large value of η_t, but not if it is due to large ε_t. Because h_t is a latent process not driven by past realizations of y_t, the likelihood function cannot be evaluated directly. Early applications like Taylor (1986) and Melino and Turnbull (1990) used method of moments rather than likelihood-based approaches.

Jacquier et al. (1994) were among the first to point out that the formulation of (78) in terms of latent variables is, by contrast, very natural in a Bayesian formulation that exploits a MCMC posterior simulator. The key insight is that conditional on the sequence of latent volatilities $\{h_t\}$, the likelihood function for (78) factors into a component for β and one for σ_η^2 and ϕ. Given an inverted gamma prior distribution for β^2 the posterior distribution of β^2 is also inverted gamma, and given an independent inverted gamma prior distribution for σ_η^2 and a truncated normal prior distribution for ϕ, the posterior distribution of $\left(\sigma_\eta^2, \phi\right)$ is the one discussed at the start of Section 5.2. Thus, the key step is sampling from the posterior distribution of $\{h_t\}$ conditional on $\{y_t^o\}$ and the parameters $\left(\beta, \sigma_\eta^2, \phi\right)$. Because $\{h_t\}$ is a first order Markov process, the conditional distribution of a single h_t given $\{h_s, s \neq t\}$, $\{y_t\}$ and $\left(\beta, \sigma_\eta^2, \phi\right)$ depends only on h_{t-1}, h_{t+1}, y_t and $\left(\beta, \sigma_\eta^2, \phi\right)$. The log-kernel of this distribution is

$$-\frac{\left(h_t - \mu_t\right)^2}{2\sigma_\eta^2/\left(1+\phi^2\right)} - \frac{h_t}{2} - \frac{y_t^2 \exp\left(-h_t\right)}{2\beta^2} \tag{79}$$

with

$$\mu_t = \frac{\phi\left(h_{t+1} + h_{t-1}\right)}{1+\phi^2} - \frac{\sigma_\eta^2}{2\left(1+\phi^2\right)}.$$

Since the kernel is non-standard, a Metropolis-within-Gibbs step can be used for the draw of each h_t. The candidate distribution in Jacquier et al. (1994) is inverted gamma, with parameters chosen to match the first two moments of the candidate density and the kernel.

There are many variants on this Metropolis-within-Gibbs step. Shephard and Pitt (1997) took a second-order Taylor series expansion of (79) about $h_t = \mu_t$, and then used a Gaussian proposal distribution with the corresponding mean and variance. Alternatively, one could find the mode of (79) and the second derivative at the mode to create a Gaussian proposal distribution. The practical limitation in all of these approaches is that sampling the latent variables h_t one at a time generates serial correlation in the MCMC algorithm: loosely speaking, the greater is $|\phi|$, the greater is the serial correlation in the Markov chain. An example in Shephard and Pitt (1997), using almost 1,000 daily exchange rate returns, showed a relative numerical efficiency (as defined in Section 3.1.3) for ϕ of about 0.001; the posterior mean of ϕ is 0.982. The Gaussian proposal

distribution is very effective, with a high acceptance rate. The difficulty is in the serial correlation in the draws of h_t from one iteration to the next.

Shephard and Pitt (1997) pointed out that there is no reason, in principle, why the latent variables h_t need to be drawn one at a time. The conditional posterior distribution of a subset $\{h_t, \ldots, h_{t+k}\}$ of $\{h_t\}$, conditional on $\{h_s, s < t, s > t + k\}$, $\{y_t\}$, and $\left(\beta, \sigma_\eta^2, \phi\right)$ depends only on h_{t-1}, h_{t+k+1}, $(y_t, \ldots, y_{t+k})$ and $\left(\beta, \sigma_\eta^2, \phi\right)$. Shephard and Pitt derived a multivariate Gaussian proposal distribution for $\{h_t, \ldots, h_{t+k}\}$ in the same way as the univariate proposal distribution for h_t. As all of the $\{h_t\}$ are blocked into subsets $\{h_t, \ldots, h_{t+k}\}$ that are fewer in number but larger in size the conditional correlation between the blocks diminishes, and this decreases the serial correlation in the MCMC algorithm. On the other hand, the increasing dimension of each block means that the Gaussian proposal distribution is less efficient, and the proportion of draws rejected in each Metropolis-Hastings step increases. Shephard and Pitt discussed methods for choosing the number of subsets that achieves an overall performance near the best attainable. In their exchange rate example 10 or 20 subsets of $\{h_t\}$, with 50 to 100 latent variables in each subset, provided the most efficient algorithm. The relative numerical efficiency of ϕ was about 0.020 for this choice.

Kim et al. (1998) provided yet another method for sampling from the posterior distribution. They began by noting that nothing is lost by working with $\log\left(y_t^2\right) = \log\left(\beta\right) + h_t + \log\varepsilon_t^2$. The disturbance term has a log-$\chi^2\left(1\right)$ distribution. This is intractable, but can be well-approximated by a mixture of seven normal distributions. Conditional on the corresponding seven latent states, most of the posterior distribution, including the latent variables $\{h_t\}$, is jointly Gaussian, and the $\{h_t\}$ can therefore be marginalized analytically. Each iteration of the resulting MCMC algorithm provides values of the parameter vector $\left(\beta, \sigma_\eta^2, \phi\right)$; given these values and the data, it is straightforward to draw $\{h_t\}$ from the Gaussian conditional posterior distribution. The algorithm is very efficient, there now being seven rather than T latent variables. The unique invariant distribution of the Markov chain is that of the posterior distribution based on the mixture approximation rather than the actual model. Conditional on the drawn values of the $\{h_t\}$ it is easy to evaluate the ratio of the true to the approximate posterior distribution. The approximate posterior distribution may thus be regarded as the source distribution in an importance sampling algorithm, and posterior moments can be computed by means of reweighting as discussed in Section 3.1.3.

Bos et al. (2000) provided an interesting application of stochastic volatility and competing models in a decision-theoretic prediction setting. The decision problem is hedging holdings of a foreign currency against fluctuations in the relevant exchange rate. The dollar value of a unit of foreign currency holdings in period t is the exchange rate S_t. If held to period $t+1$ the dollar value of these holdings will be S_{t+1}. Alternatively, at time t the unit of foreign currency may be exchanged for a contract for forward delivery of F_t dollars in period $t+1$. By covered interest parity, $F_t = S_t \exp\left(r_{t,t+1}^h - r_{t,t+1}^f\right)$, where $r_{t,\tau}^h$ and $r_{t,\tau}^f$ are the

risk-free home and foreign currency interest rates, respectively, each at time t with a maturity of τ periods. Bos et al. considered optimal hedging strategy in this context, corresponding to a CRRA utility function $U(W_t) = (W_t^\gamma - 1)/\gamma$. Initial wealth is $W_t = S_t$, and the fraction H_t is hedged by purchasing contracts for forward delivery of dollars. Taking advantage of the scale-invariance of $U(W_t)$, the decision problem is

$$\max_{H_t} \gamma^{-1} \langle E\{[(1-H_t)S_{t+1} + H_t F_t]/S_t \mid \Phi_t\}^\gamma - 1 \rangle .$$

Bos et al. took $\Phi_t = \{S_{t-j}\,(j \geq 0)\}$ and constrained $H_t \in [0,1]$. It is sufficient to model the continuously compounded exchange rate return $s_t = \log(S_t/S_{t-1})$, because

$$[(1-H_t)S_{t+1} + H_t F_t]/S_t = (1-H_t)\exp(s_{t+1}) + H_t \exp\left(r_t^h - r_t^f\right).$$

The study considered eight alternative models, all special cases of the state space model

$$\begin{aligned} s_t &= \mu_t + \varepsilon_t, \;\; \varepsilon_t \sim \left(0, \sigma_{\varepsilon,t}^2\right) \\ \mu_t &= \rho\mu_{t-1} + \eta_t, \;\; \eta_t \overset{iid}{\sim} N\left(0, \sigma_\eta^2\right). \end{aligned}$$

The two most competitive models are GARCH(1,1)-t,

$$\sigma_{\varepsilon,t}^2 = \omega + \delta\sigma_{\varepsilon,t-1}^2 + \alpha\varepsilon_{t-1}^2, \;\; \varepsilon_t \sim t\left[0, (\nu-2)\sigma_{\varepsilon,t}^2, \nu\right],$$

and the stochastic volatility model

$$\sigma_{\varepsilon,t}^2 = \mu_h + \phi\left(\sigma_{\varepsilon,t-1}^2 - \mu_h\right) + \zeta_t, \;\; \zeta_t \sim N\left(0, \sigma_\zeta^2\right).$$

After assigning similar proper priors to the models, the study used MCMC to simulate from the posterior distribution of each model. The algorithm for GARCH(1,1)-t copes with the Student-t distribution by data augmentation as proposed in Geweke (1993). Conditional on these latent variables the likelihood function has the same form as in the GARCH(1,1) model. It can be evaluated directly, and Metropolis-within-Gibbs steps are used for ν and the block of parameters $\left(\sigma_\varepsilon^2, \delta, \alpha\right)$. The Kim et al. (1998) algorithm is used for the stochastic volatility model.

Bos et al. applied these models to the overnight hedging problem for the dollar and Deutschmark. They used daily data from January 1, 1982 through December 31, 1997 for inference, and the period from January 1, 1998 through December 31, 1999 to evaluate optimal hedging performance using each model. The log-Bayes factor in favor of the stochastic volatility model is about 15. (The log-Bayes factors in favor of the stochastic volatility model, against the six models other than GARCH(1,1)-t considered, are all over 100.) Given the output of the posterior simulators, solving the optimal hedging problem is a simple and straightforward calculus problem, as described in Section 3.3.1. The

performance of any sequence of hedging decisions $\{H_t\}$ over the period $T+1, \ldots, T+F$ can be evaluated by the ex post realized utility

$$\sum_{t=T+1}^{T+F} U_t = \gamma^{-1} \sum_{t=T+1}^{T+F} \left[(1 - H_t) S_{t+1} + H_t F_t \right] / S_t.$$

The article undertook this exercise for all of the models considered as well as some benchmark *ad hoc* decision rules. In addition to the GARCH(1,1)-t and stochastic volatility models, the exercise included a benchmark model in which the exchange return s_t is Gaussian white noise. The best-performing *ad hoc* decision rule is the random walk strategy, which sets the hedge ratio to one (zero) if the foreign currency depreciated (appreciated) in the previous period. The comparisons are as follows:

Realized utility for alternative hedging strategies				
	White noise	GARCH-t	Stoch. vol.	RW hedge
Marginal likelihood	-4305.9	-4043.4	-4028.5	
$\sum U_t$ $(\gamma = -10)$	-2.24	-0.01	3.10	3.35
$\sum U_t$ $(\gamma = -2)$	0.23	7.42	7.69	6.73
$\sum U_t$ $(\gamma = 0)$	5.66	7.40	9.60	7.56

The stochastic volatility model leads to higher realized utility than does the GARCH-t model in all cases, and it outperforms the random walk hedge model except for the most risk-averse utility function. Hedging strategies based on the white noise model are always inferior. Model combination would place almost all weight on the stochastic volatility model, given the Bayes factors, and so the decision based on model combination, discussed in Sections 2.4.3 and 3.3.2, leads to the best outcome.

6 Practical experience with Bayesian forecasts

This section describes two long-term experiences with Bayesian forecasting: The Federal Reserve Bank of Minneapolis national forecasting project, and The Iowa Economic Forecast produced by The University of Iowa Institute for Economic Research. This is certainly not an exhaustive treatment of the production usage of Bayesian forecasting methods; we describe these experiences because they are well documented (Litterman, 1986; McNees, 1986; Whiteman, 1996) and because we have personal knowledge of each.

6.1 National BVAR forecasts: The Federal Reserve Bank of Minneapolis

Litterman's thesis work at the University of Minnesota ("the U") was coincident with his employment as a research assistant in the Research Department

at the Federal Reserve Bank of Minneapolis (the "Bank"). In 1978 and 1979, he wrote a computer program, "Predict" to carry out the calculations described in Section 4. At the same time, Thomas Doan, also a graduate student at the U and likewise a research assistant at the Bank, was writing code to carry out regression, ARIMA, and other calculations for staff economists. Thomas Turner, a staff economist at the Bank, had modified a program written by Christopher Sims, "Spectre", to incorporate regression calculations using complex arithmetic to facilitate frequency-domain treatment of serial correlation. By the summer of 1979, Doan had collected his own routines in a flexible shell and incorporated the features of Spectre and Predict (in most cases completely recoding their routines) to produce the program RATS (for "Regression Analysis of Time Series"). Indeed, Litterman (1979) indicates that some of the calculations for his paper were carried out in RATS. The program subsequently became a successful Doan-Litterman commercial venture, and did much to facilitate the adoption of BVAR methods throughout academics and business.

It was in fact Litterman himself who was responsible for the Bank's focus on BVAR forecasts. He had left Minnesota in 1979 to take a position as Assistant Professor of Economics at M.I.T., but was hired back to the Bank two years later. Based on work carried out while a graduate student and subsequently at M.I.T., in 1980 Litterman began issuing monthly forecasts using a six-variable BVAR of the type described in Section 4. The six variables were: real GNP, the GNP price deflator, real business fixed investment, the 3-month Treasury bill rate, the unemployment rate, and the money supply (M1). Upon his return to the Bank, the BVAR for these variables (described in Litterman, 1986) became known as the "Minneapolis Fed model."

In his description of five years of monthly experience forecasting with the BVAR model, Litterman (1986) notes that unlike his competition at the time–large, expensive commercial forecasts produced by the likes of Data Resources Inc. (DRI), Wharton Econometric Forecasting Associates (WEFA), and Chase–his forecasts were produced mechanically, without judgemental adjustment. The BVAR often produced forecasts very different from the commercial predictions, and Litterman notes that they were sometimes regarded by recipients (Litterman's mailing list of academics, which included one of us–Whiteman) as too "volatile" or "wild". Still, his procedure produced real time forecasts that were "at least competitive with the best forecasts commercially available." (Litterman, 1986, p. 35) McNees's (1986) independent assessment, which also involved comparisons with an even broader collection of competitors was that Litterman's BVAR was "generally the most accurate or among the most accurate" for real GNP, the unemployment rate, and investment. The BVAR price forecasts, on the other hand, were among the least accurate.

Subsequent study by Litterman resulted in the addition of an exchange rate measure and stock prices, and, at least experimentally, improved the performance of the model's price predictions. Other models were developed as well; Litterman (1984) describes a 46-variable monthly national forecasting model, while Amirizaheh and Todd (1984) describe a five-state model of the 9th Federal Reserve District (that of the Minneapolis Fed) involving 3 or 4 equations

per state. Moreover, the models were used regularly in Bank discussions, and reports based on them appeared regularly in the Minneapolis Fed *Quarterly Review* (e.g., Litterman, 1984a; Litterman, 1985).

In 1986, Litterman left the Bank to go to Goldman-Sachs. This required dissolution of the Doan-Litterman joint venture, and Doan subsequently formed Estima, Inc. to further develop and market RATS. It also meant that forecast production fell to staff economists whose research interests were not necessarily focused on the further development of BVARs (e.g., Miller and Roberds, 1987; Runkle, 1988; Miller and Runkle, 1989; Runkle, 1989; Runkle, 1990; Runkle, 1991). This, together with the pain associated with explaining the inevitable forecast errors, caused enthusiasm for the BVAR effort at the Bank to wane over the ensuing half dozen years, and the last *Quarterly Review* "outlook" article based on a BVAR forecast appeared in 1992 (Runkle, 1992). By the spring of 1993, the Bank's BVAR efforts were being overseen by a research assistant (albeit a quite capable one), and the authors of this paper were consulted by the leadership of the Bank's Research Department regarding what steps were required to ensure academic currency and reliability of the forecasting effort. The cost–our advice was to employ a staff economist whose research would be complementary to the production of forecasts–was regarded as too high given the configuration of economists in the department, and development of the forecasting model and procedures at the Bank effectively ceased.

Cutting-edge development of Bayesian forecasting models reappeared relatively soon within the Federal Reserve System. In 1995, Tao Zha, who had written a Minnesota thesis under the direction of Chris Sims, moved from the University of Saskatchewan to the Federal Reserve Bank of Atlanta, and began implementing the developments described in Sims and Zha (1998, 1999) to produce regular forecasts for internal briefing purposes. These efforts, which utilize the over-identified procedures described in Section 4.4, are described in Robertson and Tallman (1999a,b) and Zha (1998), but there is no continuous public record of forecasts comparable to Litterman's "Five Years of Experience".

6.2 Regional BVAR forecasts: economic conditions in Iowa

In 1990, Whiteman became Director of the Institute for Economic Research at the University of Iowa. Previously, the Institute had published forecasts of general economic conditions and had produced tax revenue forecasts for internal use of the state's Department of Management by judgmentally adjusting the product of a large commercial forecaster. These forecasts had not been especially accurate and were costing the state tens of thousands of dollars each year. As a consequence, an "Iowa Economic Forecast" model was constructed based on BVAR technology, and forecasts using it have been issued continuously each quarter since March 1990.

The Iowa model consists of four linked VARs. Three of these involve income, real income, and employment, and are treated using mixed estimation and the priors outlined in Litterman (1979) and Doan, Litterman, and Sims (1984). The fourth VAR, for predicting aggregate state tax revenue, is much smaller, and

fully Bayesian predictive densities are produced from it under a diffuse prior.

The income and employment VARs involve variables that were of interest to the Iowa Forecasting Council, a group of academic and business economists that met quarterly to advise the Governor on economic conditions. The nominal income VAR includes total nonfarm income and four of its components: wage and salary disbursements, property income, transfers, and farm income. These five variables together with their national analogues, four lags, of each, and a constant and seasonal dummy variables complete the specification of the model for the observables. The prior is Litterman's (1979) (recall specifications (61) and (62)), with a generalization of the "other's weight" that embodies the notion that national variables are much more likely to be helpful in predicting Iowa variables than the converse. Details can be found in Whiteman (1996) and Otrok and Whiteman (1998a). The real income VAR is constructed in parallel fashion after deflating each income variable by the GDP deflator.

The employment VAR is constructed similarly, using aggregate Iowa employment (nonfarm employment) together with the state's population and five components of employment: durable and nondurable goods manufacturing employment, and employment in services and wholesale and retail trade. National analogues of each are used for a total of 14 equations. Monthly data available from the U.S. Bureau of Labor Statistics and Iowa's Department of Workforce Development are aggregated to a quarterly basis. As in the income VAR, four lags, a constant, and seasonal dummies are included. The prior is very similar to the one employed in the income VARs.

The revenue VAR incorporates two variables: total personal income and total tax receipts (on a cash basis.) The small size was dictated by data availability at the time of the initial model construction: only seven years' of revenue data were available on a consistent accounting standard as of the beginning of 1990. Monthly data are aggregated to a quarterly basis; other variables include a constant and seasonal dummies. Until 1997, two lags were used; thereafter, four were employed. The prior is diffuse, as in (66).

Each quarter, the income and employment VARs are "estimated" (via mixed estimation), and (as in Litterman, 1979, and Doan, Litterman, and Sims, 1984) parameter estimates so obtained are used to produce forecasts using the chain rule of forecasting for horizons of 12 quarters. Measures of uncertainty at each horizon are calculated each quarter from a psuedo-real time forecasting experiment (recall the description of Litterman's (1979)) over the 40 quarters immediately prior to the end of the sample. Forecasts and uncertainty measures are published in the "Iowa Economic Forecast".

Production of the revenue forecasts involves normal-Wishart sampling. In particular, each quarter, the Wishart distribution is sampled repeatedly for innovation covariance matrices; using each such sampled covariance matrix, a conditionally Gaussian parameter vector and a sequence of Gaussian errors is drawn and used to seed a dynamic simulation of the VAR. These quarterly results are aggregated to annual figures and used to produce graphs of predictive densities and distribution functions. Additionally, asymmetric linear loss forecasts (see equation (29)) are produced. As noted above, this amounts to re-

porting quantiles of the predictive distribution. In the notation of (29), reports are for integer "loss factors" (ratios $(1-q)/q$); an example from July 2004 is given below:

Iowa Revenue Growth Forecasts				
Loss Factor	FY05	FY06	FY07	FY08
1	1.9	4.4	3.3	3.6
2	1.0	3.5	2.5	2.9
3	0.6	3.0	2.0	2.4
4	0.3	2.7	1.7	2.1
5	0.0	2.5	1.5	1.9

The forecasts produced by the income, employment, and revenue VARs are discussed by the Iowa Council of Economic Advisors (which replaced the Iowa Economic Forecast Council in 2004) and also the Revenue Estimating Conference (REC). The latter body consists of three individuals, of whom two are appointed by the Governor and the third is agreed to by the other two. It makes the official state revenue forecast using whatever information it chooses to consider. Regarding the use and interpretation of a predictive density forecast by state policymakers, one of the members of the REC during the 1990s, Director of the Department of Management, Gretchen Tegler remarked, "It lets the decision-maker choose how certain they want to be." (Cedar Rapids Gazette, 2004.) By law, the official estimate is binding in the sense that the governor cannot propose, and the legislature may not pass, expenditure bills that exceed 99% of revenue predicted to be available in the relevant fiscal year. The estimate is made by December 15 of each year, and conditions the Governor's "State of the State" address in early January, and the legislative session that runs from January to May.

Whiteman (1996) reports on five years of experience with the procedures. Although there are not competitive forecasts available, he compares forecasting results to historical data revisions and expectations of policy makers. During the period 1990-1994, personal income in the state ranged from about $50 billion to $60 billion. Root mean squared one-step ahead forecast errors relative to first releases of the data averaged $1 billion. The data themselves were only marginally more accurate: root mean squared revisions from first release to second release averaged $864 million. The revenue predictions made for the on-the-run fiscal year prior to the December REC meeting had root mean squared errors of 2%. Tegler's assessment: "If you are within 2 percent, you are phenomenal." (Cedar Rapids Gazette, 2004.) Subsequent difficulties in forecasting during fiscal years 2000 and 2001 (in the aftermath of a steep stock market decline and during an unusual national recession), which were widespread across the country in fact led to a reexamination of forecasting methods in the state in 2003-2004. The outcome of this was a reaffirmation of official faith in the approach, perhaps reflecting former State Comptroller Marvin Seldon's comment at the inception of BVAR use in Iowa revenue forecasting: "If you can find a revenue forecaster who can get you within 3 percent, keep him." (Seldon, 1990)

References

Aguilar, O. and M. West (2000), "Bayesian dynamic factor models and portfolio allocation", Journal of Business and Economic Statistics 18: 338-357.

Amirizadeh, H., and R. Todd (1984), "More growth ahead for ninth district states," Federal Reserve Bank of Minneapolis Quarterly Review 4:8-17.

Amisano, G. and M. Serati (1999), "Forecasting cointegrated series with BVAR models", Journal of Forecasting 18: 463-476.

Barnard, G.A. (1963), "New methods of quality control", Journal of the Royal Statistical Society Series A 126: 255-259.

Barndorff-Nielsen, O.E. and G. Schou (1973), "On the reparameterization of autoregressive models by partial autocorrelations", Journal of Multivariate Analysis 3: 408-419.

Barnett, G., R. Kohn and S. Sheather (1996), "Bayesian estimation of an autoregressive model using Markov chain Monte Carlo", Journal of Econometrics 74: 237-254.

Bates, J.M. and C.W.J. Granger (1969), "The combination of forecasts", Operations Research 20: 451-468.

Bayarri, M.J. and J.O. Berger (1998), "Quantifying surprise in the data and model verification" in: Berger, J.O., J.M. Bernardo, A.P. Dawid, D.V. Lindley and A.F.M. Smith, eds., Bayesian Statistics 6 (Oxford University Press, Oxford) 53-82.

Berger, J.O. and M. Delampady (1987), "Testing precise hypotheses", Statistical Science 2: 317-352.

Berger, J.O. and T. Selke (1987), "Testing a point null hypothesis: the irreconcilability of p values and evidence", Journal of the American Statistical Association 82: 112-122.

Bernardo, J.M. and A.F.M. Smith (1994), Bayesian Theory (Wiley, New York).

Bos, C.S., R.J. Mahieu and H.K. van Dijk (2000), "Daily exchange rate behaviour and hedging of currency risk", Journal of Applied Econometrics 15: 671-696.

Box, G.E.P. (1980), "Sampling and Bayes inference in scientific modeling and robustness", Journal of the Royal Statistical Society Series A 143: 383-430.

Box, G.E.P. and G.M. Jenkins (1976), Time Series Analysis, Forecasting and Control (Holden-Day, San Francisco).

Brav, A. (2000), "Inference in long-horizon event studies: A Bayesian approach with application to initial public offerings", The Journal of Finance 55: 1979-2016.

Carter, C.K. and R. Kohn (1994), "On Gibbs sampling for state-space models", Biometrika 81: 541-553.

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

Cedar Rapids Gazette (2004), "Rain or shine? Professor forecasts funding," Sunday February 1, 2004.

Chatfield, C. (1976), "Discussion on the paper by Professor Harrison and Mr. Stevens," Journal of the Royal Statistical Society. Series B (Methodological), Vol. 38, No. 3: 231-232.

Chatfield, C. (1993), "Calculating interval forecasts", Journal of Business and Economic Statistics 11: 121-135.

Chatfield, C. (1995), "Model uncertainty, data mining, and statistical inference", Journal of the Royal Statistical Society Series A 158: 419-468.

Chib, S. (1995), "Marginal likelihood from the Gibbs output", Journal of the American Statistical Association 90: 1313-1321.

Chib, S., and E. Greenberg (1994a), "Understanding the Metropolis-Hastings algorithm", The American Statistician 49: 327-335.

Chib, S., and E. Greenberg (1994b), "Bayes inference in regression models with ARMA(p,q) errors", Journal of Econometrics 64: 183-206.

Chib, S., and J. Jeliazkov (2001), "Marginal likelihood from the Metropolis-Hastings output", Journal of the American Statistical Association 96: 270-281.

Christoffersen, P.F. (1998), "Evaluating interval forecasts", International Economic Review 39: 841-862.

Chulani, S., B. Boehm and B. Steece (1999), "Bayesian analysis of empirical software engineering cost models", IEEE Transactions on Software Engineering 25: 573-583.

Clemen, R.T. (1989), "Combining forecasts – a review and annotated bibliography", International Journal of Forecasting 5: 559-583.

Cogley, T., Morozov, S., and T. Sargent (2004), "Bayesian fan charts for U.K. inflation: Forecasting and sources of uncertainty in an evolving monetary system," unpublished manuscript.

Dawid, A.P. (1984), "Statistical theory: The prequential approach", Journal of the Royal Statistical Society Series A 147: 278-292.

DeJong, D.N., B.F. Ingram and C.H. Whiteman (2000), "A Bayesian approach to dynamic macroeconomics", Journal of Econometrics 98: 203-223.

Diebold, F.X. (1998), Elements of Forecasting (South-Western College Publishing, Cincinnatti).

Doan, T., Litterman, R.B., and Sims, C.A. (1984), "Forecasting and conditional projection using realistic prior distributions," Econometric Reviews 3:1-100.

Draper, D. (1995), "Assessment and propagation of model uncertainty", Journal of the Royal Statistical Society Series B 57: 45-97.

Drèze, J.H. (1977), "Bayesian regression analysis using poly-t densities," Journal of Econometrics 6:329-354.

Drèze, J.H., and J.A. Morales (1976), "Bayesian full information analysis of simultaneous equations," Journal of the American Statistical Association 71:919-23.

Drèze, J.H., and J.F. Richard (1983), "Bayesian analysis of simultaneous equation systems," in Z. Griliches and M.D. Intrilligator (eds.), Handbook of Econometrics, Vol. I, Amsterdam: North-Holland.

Edwards, W., H. Lindman and L.J. Savage (1963), "Bayesian statistical inference for psychological research", Psychological Review 70: 193-242.

Engle, R.F. and B.S. Yoo (1987), "Forecasting and testing in cointegrated systems", Journal of Econometrics 35: 143-159.

Fair, R.C. (1980), "Estimating the expected predictive accuracy of econometric models," International Economic Review 21:355-378.

Foster, F.D., and C.H. Whiteman (2004), "Bayesian prediction, entropy, and option pricing in the U.S. soybean Mmarket, 1993-1997," University of Iowa manuscript.

Garcia-Ferer, A., R.A. Highfield, F. Palm and A. Zellner (1987), "Macroeconomic forecasting using pooled international data", Journal of Business and Economic Statistics 5: 53-67.

Geisel, M.S. (1975), "Bayesian comparison of simple macroeconomic models", in: S.E. Fienberg and A. Zellner, eds., Studies in Bayesian Econometrics and Statistics: In Honor of Leonard J. Savage (North-Holland, Amsterdam) 227-256.

Geisser, S. (1993), Predictive Inference: An Introduction (Chapman and Hall, London).

Gelfand, A.E. and D.K. Dey, "Bayesian model choice: Asymptotics and exact calculations", Journal of the Royal Statistical Society Series B 56: 501-514.

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

Gelman, A. (2003), "A Bayesian formulation of exploratory data analysis and goodness-of-fit testing", International Statistical Review 71: 369-382.

Gelman, A., J.B. Carlin, H.S. Stern and D.B. Rubin (1995), Bayesian Data Analysis (Chapman and Hall, London).

Gelman, A., X.L. Meng and H.S. Stern (1996), "Posterior predictive assessment of model fitness via realized discrepancies", Statistica Sinica 6: 733-760.

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

George, E. and R.E. McCulloch (1993), "Variable selection via Gibbs sampling", Journal of the American Statistical Association 99: 881-889.

Gerlach, R, C. Carter and R. Kohn (2000), "Efficient Bayesian inference for dynamic mixture models", Journal of the American Statistical Association 95: 819-828.

Geweke, J. (1988), "Antithetic acceleration of Monte Carlo integration in Bayesian inference", Journal of Econometrics 38: 73-90.

Geweke, J. (1989a), "Bayesian inference in econometric models using Monte Carlo integration", Econometrica 57: 1317-1340.

Geweke, J. (1989b), "Exact predictive densities in linear models with ARCH disturbances", Journal of Econometrics 40: 63-86.

Geweke, J. (1991), "Generic, algorithmic approaches to Monte Carlo integration in Bayesian inference", Contemporary Mathematics 115: 117-135.

Geweke, J. (1993), "Bayesian treatment of the independent Student-t linear model", Journal of Applied Econometrics 8: S19-S40.

Geweke, J. (1996a), "Monte Carlo simulation and numerical integration", in: H. Amman, D. Kendrick and J. Rust, eds., Handbook of Computational Economics (North-Holland, Amsterdam) 731-800.

Geweke, J. (1996b), "Variable selection and model comparison in regression", in: J.O. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith AFM, eds., Bayesian Statistics 5 (Oxford University Press, Oxford) 609-620.

Geweke, J. (1998), "Simulation methods for model criticism and robustness analysis", in: J.O. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith, eds., Bayesian Statistics 6 (Oxford University Press, Oxford) 275-299.

Geweke, J. (1999), "Using simulation methods for Bayesian econometric models: Inference, development and communication", Econometric Reviews 18: 1-126.

Geweke, J. (2000), "Bayesian communication: The BACC system", 2000 Proceedings of the Section on Bayesian Statistical Sciences - American Statistical Association 40-49.

Geweke, J. (2005), Contemporary Bayesian Econometrics and Statistics (Wiley, New York).

Geweke, J. and W. McCausland (2001), "Bayesian specification analysis in econometrics", American Journal of Agricultural Economics 83: 1181-1186.

Geweke, J. and G. Zhou (1996), "Measuring the pricing error of the arbitrage pricing theory", The Review of Financial Studies 9: 557-587.

Gilks, W.R., S. Richardson and D.J. Spiegelhaldter (1996), Markov Chain Monte Carlo in Practice (Chapman and Hall, London).

Good, I.J. (1956), "The surprise index for the multivariate normal distribution", Annals of Mathematical Statistics 27: 1130-1135.

Granger, C.W.J. (1986), "Comment" (on McNees, 1986), Journal of Business and Economic Statistics 4:16-17.

Granger, C.W.J. (1989), "Invited review: Combining forecasts – twenty years later", Journal of Forecasting 8: 167-173.

Granger, C.W.J. and R. Joyeux (1980), "An introduction to long memory time series models and fractional differencing", Journal of Time Series Analysis 1: 15-29.

Granger, C.W.J. and R. Ramanathan (1984), "Improved methods of combining forecasts", Journal of Forecasting 3: 197-204.

Greene, W.H. (2003), Econometric Analysis (Fifth Edition, Prentice-Hall, Upper Saddle River NJ).

Hammersly, J.M. and D.C. Handscomb (1964), Monte Carlo Methods (Methuen and Company, London).

Hammersly, J.M. and K.H. Morton (1956), "A new Monte Carlo technique: Antithetic variates", Proceedings of the Cambridge Philosophical Society 52: 449-474.

Harrison, P.J., and C.F. Stevens (1976), "Bayesian forecasting," Journal of the Royal Statistical Society. Series B (Methodological), Vol. 38, No. 3: 205-247.

Haslett, J. and A.E. Raftery (1989), "Space-time modeling with long-memory

dependence: Assessing Ireland's wind power resource", Applied Statistics 38: 1-50.

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

Heckerman, D. (1997), "Bayesian networks for data mining", Data Mining and Knowledge Discovery 1: 79-119.

Hildreth, C. (1963), "Bayesian statisticians and remote clients", Econometrica 31: 422-438.

Hoerl, A.E., and R.W. Kennard (1970), "Ridge regression: Biased estimtion for nonorthogonal problems," Technometrics 12:55-67.

Hoeting, J.A., D. Madigan, A.E. Raftery and C.T. Volinsky (1999), "Bayesian model averaging: A tutorial", Statistical Science 14: 382-401.

Hosking, J.R.M. (1981), "Fractional differencing", Biometrika 68: 165-176.

Huerta, G. and M. West (1999), "Priors and component structures in autoregressive time series models", Journal of the Royal Statistical Society Series B 61: 881-899.

Hull, J., and A. White (1987), "The pricing of options on assets with stochastic volatilities", Journal of Finance 42: 281-300.

Ingram, B.F. and C.H. Whiteman (1994), "Supplanting the Minnesota prior – forecasting macroeconomic time series using real business-cycle model priors", Journal of Monetary Economics 34: 497-510.

Iowa Economic Forecast, produced quarterly by the Institute for Economic Research in the Tippie College of Business at The University of Iowa. Available at www.biz.uiowa.edu/econ/econinst.

Jacquier, C., N.G. Polson and P.E. Rossi (1994), "Bayesian analysis of stochastic volatility models", Journal of Business and Economic Statistics 12: 371-389.

James, W. and C. Stein (1961), "Estimation with quadratic loss", in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability (University of California Press, Berkeley) 361-379.

Jeffreys, H. (1939), Theory of Probability (Clarendon Press, Oxford).

Johansen, S. (1988), "Statistical analysis of cointegration vectors", Journal of Economic Dynamics and Control 12: 231-254.

Kadiyala, K.R. and S. Karlsson (1993), "Forecasting with generalized Bayesian vector autoregressions", Journal of Forecasting 12: 365-378.

Kadiyala, K.R. and S. Karlsson (1997), "Numerical methods for estimation and inference in Bayesian VAR-models", Journal of Applied Econometrics 12: 99-132.

Kass, R.E. and A.E. Raftery (1996), "Bayes factors," Journal of the American Statistical Association 90: 773-795.

Kim, S., N. Shephard and S. Chib (1998), "Stochastic volatility: Likelihood inference and comparison with ARCH models", Review of Economic Studies 64: 361-393.

Kling, J.L. (1987), "Predicting the turning points of business and economic time series", Journal of Business 60: 201-238.

Kling, J.L. and D.A. Bessler (1989), "Calibration-based predictive distributions: An application of prequential analysis to interest rates, money, prices and output", Journal of Business 62: 477-499.

Kloek, T. and H.K. van Dijk (1978), "Bayesian estimates of equation system parameters: An application of integration by Monte Carlo", Econometrica 46: 1-20.

Koop, G. (2001), "Bayesian inference in models based on equilibrium search theory", Journal of Econometrics 102: 311-338.

Koop, G., E. Ley., J. Osiewalski and M.F.J. Steel (1997), "Bayesian analysis of long memory and persistence using ARFIMA models", Journal of Econometrics 76: 149-169.

Lancaster, T. (2004), An Introduction to Modern Bayesian Econometrics (Blackwell Publishing, Malden MA).

Leamer, E.E. (1972), "A class of informative priors and distributed lag analysis," Econometrica 40:1059-1081.

Leamer, E.E. (1978), Specification Searches (Wiley, New York).

Lesage, J.P. (1990), "A comparison of the forecasting ability of ECM and VAR models", The Review of Economics and Statistics 72: 664-671.

Lindley, D. and A.F.M. Smith (1972), "Bayes estimates for the linear model", Journal of the Royal Statistical Society Series B 34: 1-41.

Litterman, R.B. (1979) "Techniques of forecasting using vector autoregressions," Federal Reserve Bank of Minneapolis Working Paper 115.

Litterman, R.B. (1980), "A Bayesian procedure for forecasting with vector autoregressions", Working paper, Massachusetts Institute of Technology.

Litterman, R.B. (1984a), "Above average national growth in 1985 and 1986," Federal Reserve Bank of Minneapolis Quarterly Review.

Litterman, R.B. (1984b), "Forecasting and policy analysis with Bayesian vector autoregression models," Federal Reserve Bank of Minneapolis Quarterly Review.

Litterman, R.B. (1985), "How monetary policy in 1985 affects the outlook," Federal Reserve Bank of Minneapolis Quarterly Review.

Litterman, R.B. (1986), "Forecasting with Bayesian vector autoregressions - 5 years of experience", Journal of Business and Economic Statistics 4: 25-38.

Maddala, G.S. (1974), Econometrics (McGraw-Hill, New York).

Marriott, J, N. Ravishanker, A. Gelfand and J. Pai (1996), "Bayesian analysis of ARMA processes: Complete sampling-based inference under exact likelihoods" in: D.A. Barry, K.M. Chaloner and J. Geweke, eds., Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Arnold Zellner (Wiley, New York) 243-256.

McNees, S.K. (1975), "An evaluation of economic forecasts," New England Economic Review: 3-39.

McNees, S.K. (1986), "Forecasting accuracy of alternative techniques: A comparison of U.S. macroeconomic forecasts," Journal of Business and Economic Statistics 4:5-15.

Melino, A. and S. Turnbull (1990), "Pricing foreign currency options with stochastic volatility", Journal of Econometrics 45: 7-39.

Meng, X.L. (1994), "Posterior predictive p-values", Annals of Statistics 22: 1142-1160.

Meng, X.L. and W.H. Wong (1996), "Simulating ratios of normalizing constants via a simple identity: A theoretical exploration", Statistica Sinica 6: 831-860.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller (1953), "Equation of state calculations by fast computing machines", The Journal of Chemical Physics 21: 1087-1092.

Miller, P.J., and D.E. Runkle (1989), "The U.S. economy in 1989 and 1990: Walking a fine line," Federal Reserve Bank of Minneapolis Quarterly Review.

Min, C.K. and A. Zellner (1993), "Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates", Journal of Econometrics 56: 89-118.

Monahan, J.F. (1983), "Fully Bayesian analysis of ARMA time series models", Journal of Econometrics 21: 307-331.

Monahan, J.F. (1984), "A note on enforcing stationarity in autoregressive moving average models", Biometrika 71: 403-404.

Newbold, P. (1974), "The exact likelihood for a mixed autoregressive moving average process", Biometrika 61: 423-426.

Otrok, C., and C.H. Whiteman (1998a), "What to do when the crystal ball is cloudy: Conditional and unconditional forecasting in Iowa," Proceedings of the National Tax Association, 326-334.

Otrok, C., and C.H. Whiteman (1998b), "Bayesian leading indicators: Measuring and predicting economic conditions in Iowa," International Economic Review 39:997-1014.

Pai, J.S. and N. Ravishanker (1996), "Bayesian modelling of ARFIMA processes by Markov chain Monte Carlo methods", Journal of Forecasting 15: 63-82.

Palm, F.C. and A. Zellner (1992), "To combine or not to combine – Issues of combining forecasts", Journal of Forecasting 11: 687-701.

Peskun, P.H. (1973), "Optimum Monte-Carlo sampling using Markov chains", Biometrika 60: 607-612.

Petridis, V., A. Kehagias, L. Petrou, A. Bakirtzis A., S. Kiartzis, H. Panagiotou and N. Maslaris (2001), "A Bayesian multiple models combination method for time series prediction", Journal of Intelligent and Robotic Systems 31: 69-89.

Plackett, R.L. (1950), "Some theorems in least squares," Biometrica 37:149-157.

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

Porter-Hudak, S. (1982), Long-term memory modelling – a simplified spectral approach, Unpublished University of Wisconsin Ph.D. thesis.

RATS, computer program available from Estima, 1800 Sherman Ave., Suite 612, Evanston, IL 60201.

Ravishanker, N. and B.K. Ray (1997a), "Bayesian analysis of vector ARMA models using Gibbs sampling", Journal of Forecasting 16: 177-194.

Ravishanker, N. and B.K. Ray (1997b), "Bayesian analysis of vector ARFIMA process", Australian Journal of Statistics 39: 295-311.

Ravishanker, N. and B.K. Ray (2002), "Bayesian prediction for vector ARFIMA processes", International Journal of Forecasting 18: 207-214.

Ripley, R.D. (1987), Stochastic Simulation (Wiley, New York).

Roberds, W., and R. Todd (1987), "Forecasting and modelling the U.S. economy in 1986-1988," Federal Reserve Bank of Minneapolis Quarterly Review.

Roberts, H.V. (1965), "Probabilistic prediction", Journal of the American Statistical Association 60: 50-62.

Robertson, J.C., and E.W. Tallman (1999a), "Vector autoregressions: Forecasting and reality," Federal Reserve Bank of Atlanta Economic Review (First Quarter):4-18.

Robertson, J.C., and E.W. Tallman (1999b), "Improving forecasts of the Federal funds rate in a policy model," Journal of Business and Economic Statistics 19:324-30.

Rosenblatt, M. (1952), "Remarks on a multivariate transformation", Annals of Mathematical Statistics 23: 470-472.

Rothenberg, T.J. (1963), "A Bayesian analysis of simultaneous equation systems," Report 6315, Econometric Institute, Netherlands School of Economics, Rotterdam.

Rubin, D.B. (1984), "Bayesianly justifiable and relevant frequency calculations for the applied statistician", Annals of Statistics 12: 1151-1172.

Runkle, D.E. (1988), "Why no crunch from the crash?" Federal Reserve Bank of Minneapolis Quarterly Review.

Runkle, D.E. (1989), "The U.S. economy in 1990 and 1991: Continued expansion likely," Federal Reserve Bank of Minneapolis Quarterly Review.

Runkle, D.E. (1990), "Bad Nnews from a forecasting model of the U.S. economy," Federal Reserve Bank of Minneapolis Quarterly Review.

Runkle, D.E. (1991), "A bleak outlook for the U.S. economy," Federal Reserve Bank of Minneapolis Quarterly Review.

Runkle, D.E. (1992), "No relief in sight for the U.S. economy," Federal Reserve Bank of Minneapolis Quarterly Review.

Schotman, P. and H.K. van Dijk (1991), "A Bayesian analysis of the unit root in real exchange rates," Journal of Econometrics 49:195-238.

Seldon, M. (1990), personnal communication to Whiteman.

Shao, J. (1989), "Monte Carlo approximations in Bayesian decision theory", Journal of the American Statistical Association 84: 727-732.

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

Shiller, R.J. (1973), "A distributed lag estimator derived from smoothness priors," Econometrica 41:775-788.

Shoesmith, G.L. (1995), "Multiple cointegrating vectors, error correction, and forecasting with Litterman's model", International Journal of Forecasting 11: 557-567.

Sims, C.A. (1974), "Distributed lags," in M.D. Intrilligator and P.A. Kendrick, eds., Frontiers of Quantitative Economics, Volume II, 239-332. Amsterdam: North-Holland.

Sims, C.A. (1992), "A nine-variable probabilistic macroeconomic forecasting model," In J.H. Stock and M.W. Watson, eds., Business Cycles, Indicators, and Forecasting. Chicago: University of Chicago Press.

Sims, C.A., and T.A. Zha (1997), "Bayesian methods for dynamic multivariate models," International Economic Review 39:949-968.

Sims, C.A., and T.A. Zha (1999), "Error bands for impulse responses," Econometrica 67:1113-1155.

Smith, A.F.M. (1973), "A general Bayesian linear model", Journal of the Royal Statistical Society Series B 35: 67-75.

Smyth, D.J. (1983), "Short-run macroeconomic forecasting: the OECD performance", Journal of Forecasting 2: 37-49.

Sowell, F. (1992a), "Maximum likelihood estimation of stationary univariate fractionally integrated models", Journal of Econometrics 53: 165-188.

Sowell, F. (1992b), "Modeling long-run behavior with the fractional ARIMA model", Journal of Monetary Economics 29: 277-302.

Stein, C.M. (1974), "Multiple regression," in I. Olkin (ed.) Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling (Stanford University Press: Stanford).

Tanner, M.A. and H.W. Wong (1987), "The calculation of posterior distributions by data augmentation", Journal of the American Statistical Association 82: 528-540.

Tauchen, G. and M. Pitts (1983), "The price-variability-volume relationship on speculative markets", Econometrica 51: 485-505.

Tay, A.S. and K.F. Wallis (2000), "Density forecasting: A survey", Journal of Forecasting 19: 235-254.

Taylor, S. (1986), Modelling Financial Time Series (Wiley, New York).

Theil, H. (1963), "On the use of incomplete prior information in regression analysis," Journal of the American Statistical Association 58:401-414.

Thompson, P.A. (1984), Bayesian multiperiod prediction: Forecasting with graphics, Unpublished University of Wisconsin Ph.D. thesis.

Thompson, P.A. and R.B. Miller (1986), "Sampling the future: A Bayesian approach to forecasting from univariate time series models", Journal of Business and Economic Statistics 4: 427-436.

Tierney, L. (1994), "Markov chains for exploring posterior distributions", Annals of Statistics 22: 1701-1762.

Tobias, J.L. (2001), "Forecasting output growth rates and median output growth rates: A hierarchical Bayesian approach", Journal of Forecasting 20: 297-314.

Villani, M. (2001), "Bayesian prediction with cointegrated vector autoregressions", International Journal of Forecasting 17: 585-605.

Wecker, W. (1979), "Predicting the turning points of a time series", Journal of Business 52: 35-50.

Weiss, A.A. (1996), "Estimating time series models using the relevant cost function", Journal of Applied Econometrics 11: 539-560.

West, M. (1995), "Bayesian inference in cyclical component dynamic linear models", Journal of the American Statistical Association 90: 1301-1312.

West, M. and J. Harrison (1997), Bayesian Forecasting and Dynamic Models (Second Edition, Springer, New York).

Whiteman, C.H. (1996), "Bayesian prediction under asymmetric linear loss: Forecasting state tax revenues in Iowa," in W.O. Johnson, J.C. Lee, and A. Zellner, eds., Forecasting, Prediction and Modeling in Statistics and Econometrics: Bayesian and non-Bayesian Approaches. Springer-Verlag, New York.

Winkler, R.L. (1981), "Combining probability distributions from dependent information sources", Management Science 27: 479-488.

Zellner, A. (1971), An Introduction to Bayesian Inference in Econometrics (Wiley, New York).

Zellner, A. (1986), "Bayesian estimation and prediction using asymmetric loss functions", Journal of the American Statistical Association 81: 446-451.

Zellner A., and B. Chen (2001), "Bayesian modeling of economies and data requirements", Macroeconomic Dynamics 5: 673-700.

Zellner, A. and C. Hong (1989), "Forecasting international growth rates using Bayesian shrinkage and other procedures", Journal of Econometrics 40: 183-202.

Zellner A, C. Hong and G.M. Gulati (1990), "Turning points in economic time series, loss structures and Bayesian forecasting", in: S Geisser, J.S. Hodges, S.J. Press and A. Zellner, eds., Bayesian and likelihood methods in statistics and econometrics: Essays in honor of George A. Barnard (North-Holland, Amsterdam) 371-393.

Zellner, A., C. Hong and C.K. Min (1991), "Bayesian exponentially weighted autoregression, time-varying parameter, and pooling techniques", Journal of Econometrics 49: 275-304.

Zha, T.A. 1998. "A dynamic multivariate model for use in formulating policy." Federal Reserve Bank of Atlanta Economic Review 83 (First Quarter): 16–29.