MONASH University

MONASH BUSINESS SCHOOL

DEPARTMENT OF ECONOMETRICS AND BUSINESS STATISTICS

ISSN 1440-771X

WORKING PAPER SERIES

Bayesian Forecasting in the 21st Century: A Modern Review

Gael M. Martin, David T. Frazier and Ruben Loaiza-Maya, Florian Huber, Gary Koop, John Maheu, Didier Nibbering and Anastasios Panagiotelis

Working Paper No.01/23
JANUARY 2023

Bayesian Forecasting in the 21st Century: A Modern Review

Gael M. Martin, David T. Frazier, Worapree Maneesoonthorn, Rubén Loaiza-Maya,

Florian Huber, Gary Koop, John Maheu, Didier Nibbering and Anastasios Panagiotelis*

December 12, 2022

Abstract

The Bayesian statistical paradigm provides a principled and coherent approach to probabilistic forecasting. Uncertainty about all unknowns that characterize any forecasting problem – model, parameters, latent states – is factored into the forecast distribution, with forecasts conditioned *only* on what is known or observed. Allied with the elegance of the method, Bayesian forecasting is now underpinned by the burgeoning field of Bayesian computation, which enables Bayesian forecasts to be produced for virtually any problem, no matter how large, or complex. The current state of play in Bayesian forecasting is the subject of this review. The aim is to provide readers with an overview of modern approaches to the field, set in some historical context. Whilst our primary focus is on applications in the fields of economics and finance, and their allied disciplines, sufficient general details about implementation are provided to aid and inform all investigators.

Keywords: Bayesian prediction; macroeconomics; finance; marketing; electricity demand; Bayesian computational methods; loss-based Bayesian prediction

MSC2010 Subject Classification: 60G25, 62F15, 62M20

JEL Classifications: C01, C11, C53

*Authors: Gael M. Martin, Monash University, Australia (Corresponding author: gael.martin@monash.edu); David T. Frazier, Monash University, Australia; Worapreee (Ole) Maneesoonthorn, University of Melbourne, Australia; Rubén Loaiza-Maya, Monash University, Australia; Florian Huber; University of Salzburg, Austria; Gary Koop, University of Strathclyde, UK; John Maheu, McMaster University, Canada; Anastasios Panagiotelis; University of Sydney, Australia; Didier Nibbering, Monash University, Australia. Martin, Frazier and Maneesoonthorn have been supported by Australian Research Council Discovery Grant DP200101414. Frazier has also been supported by Australian Research Council Discovery Early Career Researcher Award DE200101070.

1 Introduction

1.1 Why Bayesian forecasting?

The Bayesian statistical paradigm uses the rules and language of probability to quantify uncertainty about all unknown aspects of phenomena that generate observed data. This core characteristic of the paradigm makes it particularly suitable for forecasting, with uncertainty about the unknown values of future observations automatically expressed in terms of a probability distribution. Moreover, Bayesian methods allow a user to seamlessly, and systematically, yield probabilistic forecasts that reflect uncertainty about all unknowns, and that condition *only* on known past events, or 'data'.

To paraphrase Geweke and Whiteman (2006), Bayesian methods prevail in forecasting because they explicitly employ the *principle of relevant conditioning*: the use of forecasts that condition only on observed data, with all uncertainty about the model, parameters, latent variables, etc., correctly accounted for in the production of said forecasts. This stands in stark contrast to forecasting approaches based on frequentist methods, which typically condition on a given model (or set of models), and on point estimates of model unknowns, plus adopt *ad hoc* methods (if any) to quantify uncertainty regarding unknown quantities. Indeed, the ability of Bayesian forecasters to incorporate the uncertainty associated with the production of forecasts, while utilizing all available information – both *a priori* and sample information − in a principled manner, led Granger *et al.* (1986) to conclude that:

> *"In terms of forecasting accuracy a good Bayesian will beat a non-Bayesian, who will do better than a bad Bayesian."*

Echoing these sentiments, in our opinion, the power of the Bayesian forecasting paradigm is a product of the paradigm's requirement that all elements of the statistical problem necessary to produce forecasts – future observations, past observations, parameters, latent variables, models – are treated as arguments of a joint probability distribution. The express probabilistic formulation of these elements, in turn, allows a Bayesian to invoke the standard rules of probability to produce a distribution for an unknown future value that is conditioned on the known past data, and is *marginal* of all other arguments that are inherently unknown.

While this ability to marginalize all unknowns through probability calculus is the hallmark of the Bayesian approach, the benefits of the paradigm, and what ultimately in our opinion defines a 'good Bayesian', is the attention to detail necessary to successfully implement Bayesian methods. In Bayesian forecasting, before we ever attempt to produce a forecast, we must first carefully enumerate all possible sources of uncertainty; and construct reasonable prior beliefs for these quantities, which often include (possibly several layers of) latent variables that have a specific and delicate interaction with the observed data; always taking great care to ensure that these prior beliefs do not conflict with the observed data. Then and only then can we 'turn the Bayesian crank' to produce the joint posterior distribution over all unknown quantities (including future values), and ultimately integrate out the quantities we are not interested in to obtain the (posterior) predictive distribution for the future values of our random variables of interest. The attention to detail necessary to produce Bayesian forecasts means that very

few (if any) *implicit* assumptions are maintained, and what explicit assumptions are maintained can often be rationalized/tested against the data.

Consistent with the internal coherence of the Bayesian statistical paradigm, the basic manner in which all Bayesian forecasting problems are *framed* is the same. What differs however, from case to case, is the way in which the problem is *solved* – i.e. the way in which the forecast distribution is *accessed.* To understand why this is so, it is sufficient to recognize that virtually all Bayesian quantities of interest, including forecast distributions, can be expressed as expectations of some sort. For most models that are used to predict empirically relevant data these expectations are not available in closed form. Hence, in any practical problem, implementation of Bayesian forecasting is both model- and data-dependent, and relies on advanced computational tools. Different forecasting problems – defined by different forms and 'sizes' of models and data sets – require, in turn, different approaches to computation. The evolution of the *practice* of Bayesian forecasting has, as a consequence, gone hand-in-hand with developments on the computational front; with increasingly large and complex models rendered amenable to a Bayesian forecasting approach via access to modern techniques of computation.

1.2 The purview of this review

In this review, we give a modern take on the current landscape of Bayesian forecasting. Whilst excellent textbook treatments of Bayesian forecasting are given in Geweke (2005) and West and Harrison (2006), and with Geweke and Whiteman (2006) reviewing specific aspects of Bayesian forecasting in a slew of practical settings, the field has advanced by leaps and bounds in the last twenty years. Therefore, we believe the time is ripe to consider a review of the subject that touches on many of the novel and exciting areas now being explored. The methodological advances we review have general applicability to all discipline areas; nevertheless, due to our own interests, expertise and experience – and to keep the scope of the paper manageable – we have chosen to focus primarily on applications in the economic sciences. Whilst the paper is not designed to be a treatise on Bayesian computation, sufficient details are provided to enable the practitioner to understand *why* numerical tools are needed in most forecasting settings, and *how* they are used.

The general structure of the paper is as follows. In Section 2 we provide a short tutorial on Bayesian forecasting. This begins with an outline of the Bayesian forecasting method, followed by an overview of the computational techniques used to *implement* the method. In Section 3 we then take the reader on a potted chronological tour of Bayesian forecasting, up to the present day. We begin by giving a snapshot of the forecasting problems tackled during the last few decades of the 20th century, and the computational solutions that were adopted then – most notably, Markov chain Monte Carlo (MCMC) algorithms. We then look at the types of 'intractable' forecasting problems that are increasingly encountered in the 21st century, and provide an overview of the new computational solutions that have been proposed to tackle such problems. We also outline very recent developments in which misspecification of the forecasting model is explicitly acknowledged, and conventional likelihood-based Bayesian forecasting eschewed as a consequence; with problem-specific measures of forecast accuracy (or *forecast loss*) used, instead, to drive the production of forecast distributions. Section 4 then provides the reader with more detailed reviews of modern Bayesian forecasting in the following four broad fields: macroeconomics, finance, marketing,

and electricity pricing and demand. Section 5 closes the paper with a brief summary of the current state of play.

Before proceeding further, we make a note about scope and language. To render the scope of the paper manageable we focus primarily on Bayesian forecasting in 'time series models' – i.e. models for random variables that are indexed by time – and on using such models to say something about the values that these random variables will assume in the future. These future values may be informed only by past observations on the variable, or may also depend on the known values of covariates, or regressors. We also follow the convention in the Bayesian literature by using the terms 'forecast' and 'prediction' (and all of their various grammatical derivations) synonymously and interchangeably in this case, for the sake of linguistic variety. The fundamental principles of Bayesian prediction apply equally to data indexed by something other than time. The term 'forecast' is not used in this case as it is a term reserved for temporal settings. The main exceptions to our focus on time series models, and forecasting *per se*, occur in Section 4.3, in which models for cross-sectional data are used to predict customer choice in marketing settings, and Section 4.4, in which models for electricity demand that have a spatial dimension are referenced.

2 A Tutorial on Bayesian Forecasting

2.1 The Bayesian forecasting method

For the sake of illustration, we assume a scalar random variable y_t, and define the $(T \times 1)$ vector of observations on y_t as $\mathbf{y}_{1:T} = (y_1, y_2, ..., y_T)'$. We assume (for the moment) that $\mathbf{y}_{1:T}$ has been generated from some *parametric* model with likelihood $p(\mathbf{y}_{1:T}|\boldsymbol{\theta})$, with $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)' \in \Theta \subseteq \mathbb{R}^p$ a p-dimensional vector of unknown parameters, and where we possess prior beliefs on $\boldsymbol{\theta}$ specified by $p(\boldsymbol{\theta})$. Using the same symbol $\mathbf{y}_{1:T}$ to denote both the vector of observed data and the T-dimensional vector random variable, we define the joint distribution over $\mathbf{y}_{1:T}$ and $\boldsymbol{\theta}$ as $p(\mathbf{y}_{1:T}, \boldsymbol{\theta})$. Application of the standard rules of probability to $p(\mathbf{y}_{1:T}, \boldsymbol{\theta})$ yields *Bayes theorem* (or *Bayes rule)*,

$$p(\boldsymbol{\theta}|\mathbf{y}_{1:T}) = \frac{p(\mathbf{y}_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y}_{1:T})}, \tag{1}$$

where $p(\mathbf{y}_{1:T}) = \int_\Theta p(\mathbf{y}_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$. Bayes theorem provides a representation for the *posterior* probability density function (pdf) for $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$, as proportional to the product of the likelihood function and the prior. The term $p(\mathbf{y}_{1:T})$ defines the marginal likelihood, and the scale factor $[p(\mathbf{y}_{1:T})]^{-1}$ in (1) ensures that $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ integrates to one.

Now, define y_{T+1} as the (one-step-ahead) future random variable, where we focus on one-step-ahead forecasting in Sections 2 and 3 merely to simplify the exposition. Assuming y_{T+1} to be a continuous random variable (again, for illustration), standard probability manipulations lead to the following expression for the *forecast* (or *predictive*) pdf for y_{T+1} :

$$p(y_{T+1}|\mathbf{y}_{1:T}) = \int_\Theta p(y_{T+1}|\boldsymbol{\theta}, \mathbf{y}_{1:T})p(\boldsymbol{\theta}|\mathbf{y}_{1:T})d\boldsymbol{\theta}. \tag{2}$$

When no confusion arises, we also refer to $p(y_{T+1}|\mathbf{y}_{1:T})$, albeit loosely, as the forecast (or predictive)

distribution, or simply as the 'predictive'.[1] The density $p(y_{T+1}|\mathbf{y}_{1:T})$ summarizes all uncertainty about y_{T+1}, conditional on the assumed model – which underpins the structure of both the conditional predictive, $p(y_{T+1}|\boldsymbol{\theta}, \mathbf{y}_{1:T})$, and the posterior itself – and the prior beliefs that inform $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$. Point and interval predictions of y_{T+1}, and indeed any other distributional summary, can be extracted from (2). In the case where the model itself is uncertain, and a finite set of parametric models, $\mathcal{M}_1$, $\mathcal{M}_2$,...,$\mathcal{M}_K$, is assumed to span the model space, a 'model-averaged' predictive, $p_{MA}(y_{T+1}|\mathbf{y}_{1:T})$, is produced as

$$p_{MA}(y_{T+1}|\mathbf{y}_{1:T}) = \sum_{k=1}^{K} p(y_{T+1}|\mathbf{y}_{1:T}, \mathcal{M}_k) p(\mathcal{M}_k|\mathbf{y}_{1:T}), \tag{3}$$

where $p(y_{T+1}|\mathbf{y}_{1:T}, \mathcal{M}_k)$ denotes the density in (2), but now conditioned explicitly on the kth model in the set. The kth posterior model probability, $p(\mathcal{M}_k|\mathbf{y}_{1:T})$, $k = 1, 2, ..., K$, is computed via a further application of Bayes theorem in which the (initial) joint distribution of interest is defined over both the model space and the space for the parameters of each of the K models. Standard manipulations lead to

$$p(\mathcal{M}_k|\mathbf{y}_{1:T}) \propto p(\mathbf{y}_{1:T}|\mathcal{M}_k) p(\mathcal{M}_k), \tag{4}$$

where

$$p(\mathbf{y}_{1:T}|\mathcal{M}_k) = \int_{\Theta_k} p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_k, \mathcal{M}_k) p(\boldsymbol{\theta}_k|\mathcal{M}_k) d\boldsymbol{\theta}_k, \tag{5}$$

for each $k = 1, 2, ..., K$, with $\boldsymbol{\theta}_k$ denoting the parameter set for the kth model.

As is clear, *analytical* evaluation of $p(y_{T+1}|\mathbf{y}_{1:T})$ in (2) requires, at the very least, a closed-form expression for $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$. Typically, however, such an expression is not available, with most posteriors being known only up to a constant of proportionality, as

$$p(\boldsymbol{\theta}|\mathbf{y}_{1:T}) \propto p(\mathbf{y}_{1:T}|\boldsymbol{\theta}) p(\boldsymbol{\theta}). \tag{6}$$

The main exceptions to this occur when $p(\mathbf{y}_{1:T}|\boldsymbol{\theta})$ is from the exponential family, and either a natural conjugate, or convenient noninformative prior is adopted; specifications which may be suitable for some simple (and low-dimensional) empirical problems, but are certainly not broadly applicable in practice. Analytical evaluation of $p_{MA}(y_{T+1}|\mathbf{y}_{1:T})$ in (3) also requires a closed-form expression for each $p(\mathbf{y}_{1:T}|\mathcal{M}_k)$ (with normalization of $p(\mathcal{M}_k|\mathbf{y}_{1:T})$ then straightforward); once again a rare thing beyond the exponential family (and standard prior) setting. Hence the need for numerical *computation* to implement Bayesian forecasting in virtually all realistic empirical problems.[2]

2.2 An overview of computation

The form of (2) makes it clear that the Bayesian predictive pdf, $p(y_{T+1}|\mathbf{y}_{1:T})$, is nothing more than the posterior expectation of the predictive *conditional* on $\boldsymbol{\theta}$. Hence, accessing $p(y_{T+1}|\mathbf{y}_{1:T})$ amounts to the

[1] We note that $p(y_{T+1}|\mathbf{y}_{1:T})$ is sometimes referred to as a 'posterior' predictive in the literature, given that it is produced by averaging the conditional predictive, $p(y_{T+1}|\boldsymbol{\theta}, \mathbf{y}_{1:T})$, with respect to the posterior density, $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$. We do not adopt this expression, leaving it to the context to make it clear as to whether the term 'predictive' is being used to refer to the distribution that is marginal of $\boldsymbol{\theta}$, $p(y_{T+1}|\mathbf{y}_{1:T})$, or that which is conditioned on $\boldsymbol{\theta}$, $p(y_{T+1}|\boldsymbol{\theta}, \mathbf{y}_{1:T})$. We also streamline the exposition by not using explicit notation for any observed covariates on which the model for y_t may depend, and on which the predictive for y_{T+1} would condition, unless this is essential.

[2] Numerous textbook illustrations of the material in this section can be found. In addition to the references Geweke (2005) and West and Harrison (2006) cited earlier, some examples are Zellner (1971), Koop (2003) and Robert (2007).

evaluation of an expectation. This insight is helpful, as it enables us to see many of the computational methods that are used to access $p(y_{T+1}|\mathbf{y}_{1:T})$ – in cases where it is not available in closed form – simply as different ways of numerically estimating an expectation.

It is convenient to group Bayesian computational methods into three categories: *1)* Deterministic integration (or quadrature) methods (Davis and Rabinowitz, 1975; Naylor and Smith, 1982); *2)* Exact simulation methods; and *3)* Approximate methods. Given that the production of $p(y_{T+1}|\mathbf{y}_{1:T})$ involves integration over $\boldsymbol{\theta}$, only in very low-dimensional models is *1)* a feasible computational method on its own, due to the well-known 'curse of dimensionality' that characterizes numerical quadrature. Hence, the computational methods in *2)* and *3)* are those most commonly adopted, and will be our focus here; noting that quadrature may play still a limited role *within* these alternative computational frameworks.

The methods in *2)* use simulation to produce M draws of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(i)}$, $i = 1, 2, ..., M$, from the posterior $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$, which, in turn, define M conditional predictives, $p(y_{T+1}|\boldsymbol{\theta}^{(i)}, \mathbf{y}_{1:T})$, $i = 1, 2, ..., M$, the mean of which is used to estimate (2). Alternatively, if it is easier to simulate from $p(y_{T+1}|\boldsymbol{\theta}^{(i)}, \mathbf{y}_{1:T})$ than to evaluate it at any point in the support of y_{T+1}, M draws of y_{T+1}, $y_{T+1}^{(i)}$, $i = 1, 2, ..., M$, are taken, one for each draw $\boldsymbol{\theta}^{(i)}$, and kernel density estimation methods used to produce an estimate of $p(y_{T+1}|\mathbf{y}_{1:T})$. Different simulation methods are distinguished by the way in which the posterior draws are produced. Methods in *2)* include Monte Carlo sampling (Metropolis and Ulam, 1949), importance sampling (IS) (Hammersley and Handscomb, 1964; Kloek and van Dijk, 1978; Geweke, 1989) and MCMC sampling – including Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) and Metropolis-Hastings (MH) algorithms (Metropolis *et al.*, 1953; Hastings, 1970) – with MCMC being by far the most common simulation method used to compute forecast distributions in practice. The term 'exact' arises from the fact that, under appropriate conditions (including convergence of the Markov chain to $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ in the case of the MCMC algorithms), such methods all produce a $\sqrt{M}$-consistent estimate of the ordinate $p(y_{T+1}|\mathbf{y}_{1:T})$, at any point in the support of the random variable y_{T+1}; this estimate can thus be rendered arbitrarily accurate, for large enough M.

We refer the reader to: Chib and Greenberg (1996) and Geyer (2011) for reviews of MCMC sampling; Casella and George (1992) and Chib and Greenberg (1995) for descriptions of the Gibbs and MH algorithms (respectively) that are useful for practitioners; and Andrieu *et al.* (2004), Robert and Casella (2011) and Martin *et al.* (2022b) for historical accounts of MCMC sampling. Geweke and Whiteman (2006) also serves as an excellent reference on the use of these computational methods in a forecasting context. Given the critical role played by MCMC methods in the production of Bayesian forecasts, the basic principles of the algorithms are also outlined below in Section 3.1; with more recent developments of both IS and MCMC – most notably sequential Monte Carlo (SMC) (Gordon *et al.*, 1993; Chopin and Papaspiliopoulos, 2020) and pseudo-marginal MCMC (Beaumont, 2003; Andrieu and Roberts, 2009; Andrieu *et al.*, 2011) – discussed briefly in Section 3.2.

The methods in *3)* replace $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ in the integrand of (2) with an *approximation* of some sort, and evaluate the resultant integral. In so doing, such methods do not aim to estimate $p(y_{T+1}|\mathbf{y}_{1:T})$ itself, but some representation of it, defined as the expectation of $p(y_{T+1}|\boldsymbol{\theta}, \mathbf{y}_{1:T})$ with respect to the relevant posterior approximation. The methods in *3)* have been based on the principles of approximate Bayesian computation (ABC) (Marin *et al.*, 2011; Sisson and Fan, 2011; Sisson *et al.*, 2019), Bayesian synthetic

likelihood (BSL) (Price *et al.*, 2018), variational Bayes (VB) (Blei *et al.*, 2017), and integrated nested Laplace approximation (INLA) (Rue *et al.*, 2009), and produce what are termed 'approximate' forecast, or predictive distributions. Suffice to say that the principle adopted for estimating the 'approximate predictive' so defined is typically one and the same: draws of $\boldsymbol{\theta}$ from the approximate posterior (however produced) are used to produce either a sample mean of conditional predictives, or M draws of y_{T+1} from $p(y_{T+1}|\boldsymbol{\theta}, \mathbf{y}_{1:T})$, with kernel density estimation then applied.

Production of (3) requires the computation of each model-specific predictive, plus the computation of each (5). The first set of K computations would proceed via the sorts of steps outlined above. Computation of the K marginal likelihoods could also be performed via one of the three broad methods listed above (in particular *2)* or *3)*); however, the fact that each (5) is a *prior*, rather than a *posterior* expectation does have implications for precise manner in which computation is implemented. (See Ardia *et al.*, 2012, and Llorente *et al.*, 2021, for details).

3 Bayesian Forecasting: A Chronological Tour

3.1 The late 20th century: The advent of MCMC

As is clear from the brief synopsis above, it is *simulation* that is key to computing forecast distributions when they are not available in closed form. While the use of simulation to compute statistical quantities of interest was known by the 1970s (Metropolis and Ulam, 1949; Metropolis *et al.*, 1953; Hammersley and Handscomb, 1964; Hastings, 1970), the technology required to perform simulation in a convenient and timely fashion was not yet available, and simulation-based computation thus remained largely out of reach. To quote Geweke and Whiteman (2006):

"In the beginning, there was diffuseness, conjugacy and analytical work!"

In the latter part of the 20th century, things changed. The increased speed and availability of desktop machines (Ceruzzi, 2003), allied with critical advances in simulation methodology, led to a proliferation of methods for accessing $p(y_{T+1}|\mathbf{y}_{1:T})$ via the simulation of draws from $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$. To this end, we give a brief outline of the pre-eminent posterior simulation algorithms of the 1990s (and into the early 2000s): Gibbs sampling (Section 3.1.1), MH-within-Gibbs sampling (Section 3.1.2), and (MH-within-) Gibbs sampling allied with data augmentation (Section 3.1.3); touching on the types of forecasting models that were able to be treated via such methods, including the ubiquitous state space models that underpin much modern Bayesian forecasting. To keep the exposition concise, we place all algorithmic details in Appendix A, and reference specific algorithms from Appendix A at suitable points in the text.

3.1.1 Gibbs sampling

As a general rule, if $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ does not have a closed-form representation, it is also not amenable to Monte Carlo sampling, as the latter requires that $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ can be decomposed into recognizable densities, from which computer simulation is feasible. IS (Kloek and van Dijk, 1978; Geweke, 1989), via use of an 'importance' or 'proposal' density, $q(\boldsymbol{\theta}|\mathbf{y}_{1:T})$, that matches $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ well and which *can* be drawn from,

is a possible solution in some cases. However, the algorithm can fail to produce representative draws from $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ when the dimension of $\boldsymbol{\theta}$ is large, due to the difficulty of finding a $q(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ that is a 'good match' to $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ in high dimensions.

In contrast, under certain conditions, a Gibbs sampler is able to produce a (dependent) set of draws from the *joint* posterior via iterative sampling from lower dimensional, and often standard, *conditional* posteriors. In other words, a Gibbs sampler takes advantage of the fact that, while joint and marginal posterior distributions are usually complex in form and unable to be simulated from directly, conditional posteriors are often standard and amenable to simulation. Given the satisfaction of the required convergence conditions (Geyer, 2011), draws $\boldsymbol{\theta}^{(i)}$, $i = 1, 2, ..., M$, produced via iterative sampling from the full conditionals converge in distribution to $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ as $M \to \infty$, and can be used to produce a $\sqrt{M}$-consistent estimate of the ordinates of $p(y_{T+1}|\mathbf{y}_{1:T})$ across the support of y_{T+1} in the manner described in Section 2.2. Decisions about how to partition, or 'block' $\boldsymbol{\theta}$ need to be made (Liu *et al.*, 1994; Roberts and Sahu, 1997), with a view to increasing the 'efficiency' of the chain which, in effect, amounts to ensuring an accurate estimate of $p(y_{T+1}|\mathbf{y}_{1:T})$ for an given number of draws, M. (See Algorithm 1 in Appendix A.1.)

Chib (1993) and McCulloch and Tsay (1994) are the earliest examples of using Gibbs algorithms for Bayesian estimation and prediction in time series settings. Whilst Chib (1993) looks expressly at an autoregressive (AR) structure in the error term of a linear regression model, the proposed Gibbs scheme also nests the steps required for sampling from an observed AR model. (See also Section 8.1, Koop, 2003.) The key points to note are that the conditional posteriors are shown to retain a closed form even when the specification allows for *i)* multiple lags of y_t, *ii)* an informative prior on $\boldsymbol{\theta}$, *iii)* a prior that imposes stationarity on the model, and *iv)* a Student t (rather than a Gaussian) distribution; with any one of these modifications *precluding* analytical treatment of $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ and $p(y_{T+1}|\mathbf{y}_{1:T})$. McCulloch and Tsay (1994) entertain even more complex versions of an AR model, to cater for random level shifts, outliers and missing values, and with an informative prior invoked for all unknowns. Once again, despite such model and prior specifications implying that closed forms for $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ and $p(y_{T+1}|\mathbf{y}_{1:T})$ do not exist, the conditional posteriors all retain a closed form and, hence, allow for a straightforward application of a Gibbs scheme.

As one would anticipate however, a 'pure' Gibbs algorithm based on a full set of standard conditionals is not always possible, with the more typical situation being one in which one or more of the conditionals – associated with any given partitioning of the parameter space – are not available in closed form. The following section describes how to adapt a Gibbs algorithm in cases where certain conditional components are not known in closed form, and, in so doing, illustrates a powerful simulation-based algorithm for accessing $p(y_{T+1}|\mathbf{y}_{1:T})$ in more complex settings.

3.1.2 MH-within-Gibbs sampling

The Gibbs sampler is only one example of an MCMC algorithm. The first such example – the 'Metropolis' algorithm – appeared in a paper that has assumed an important status in the history of statistics:

Metropolis *et al.* (1953)[3]. The 'Metropolis' algorithm was subsequently generalized by Hastings (1970), and it is this 'MH' version of the method that is typically referenced. For the purpose of this review, the key purpose of the MH algorithm is to enable sampling from non-standard conditionals within a Gibbs algorithm, in particular when the dimension of the conditionals precludes (say) the exclusive use of inverse cumulative distribution function (ICDF) sampling.[4]

Under regularity, a Markov chain that converges to $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ can be produced by embedding an MH algorithm (or MH *algorithms*) within an outer Gibbs loop. In short, an MH-within-Gibb algorithm proceeds by drawing from any non-standard conditional indirectly, via a 'candidate', or 'proposal' distribution that is deemed to be a good match to the inaccessible conditional, and accepting the draw with a given probability. Critically, the formula that defines the acceptance probability involves evaluation of the non-standard conditional only up to its integrating constant; hence the conditional need not be known in its entirety.[5] Again, under appropriate regularity, the draws $\boldsymbol{\theta}^{(i)}$, $i = 1, 2, ..., M$, from the MH-within-Gibbs algorithm converge in distribution to $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ as $M \to \infty$, and can be used to produce a $\sqrt{M}$-consistent estimate of the ordinates of $p(y_{T+1}|\mathbf{y}_{1:T})$. (See Algorithm 2 in Appendix A.2.)

As will become evident in the subsequent empirical review sections, MH-within-Gibbs algorithms remain the dominant form of method used to sample from posteriors – and to estimate predictive distributions – for time series models for which a convenient partitioning of the parameter space is available, and for which the conditional posteriors are known up to their integrating constants. Hence, we reserve further elaboration on the use of such algorithms in practice until the appropriate points in Section 4.

3.1.3 MCMC, data augmentation, and state space models

For many empirical problems in economics and related fields, a suitable model can be partitioned into two sets: static unknowns $\boldsymbol{\theta}$, which are fixed throughout time, and latent data, $\mathbf{z}_{1:T} = (z_1, z_2, ..., z_T)'$, that are a function of time. The latent states may be intrinsic to the model – as in a state space model – or may be auxiliary variables introduced purely for the purpose of facilitating posterior sampling. Application of a Gibbs-based MCMC scheme to the joint, or 'augmented' set of unknowns $\{\boldsymbol{\theta}, \mathbf{z}_{1:T}\}$ is often referred to as 'data augmentation', in the spirit of Tanner and Wong (1987), and such schemes have enabled the Bayesian analysis of large classes of time series models that would otherwise have been inaccessible.

We illustrate here the basic principles of the approach using a state space model governed by a measurement density for the observed scalar random variable, y_t, and a Markov transition density for a

[3]For example, Dongarra and Sullivan (2000) rank the 'Metropolis' algorithm proposed in Metropolis *et al.* as one of the 10 algorithms "with the greatest influence on the development and practice of science and engineering in the 20th century".

[4]Any non-standard probability distribution can, in principle, be drawn from using ICDF sampling. The term 'Griddy Gibbs' sampling was first used by Ritter and Tanner (1992) to refer to the use of ICDF sampling to draw from non-standard conditionals in a Gibbs scheme. Given that the method amounts to the use of numerical quadrature, it suffers from the curse of dimensionality, and is thus infeasible for drawing from anything other than very low-dimensional conditionals. See Bauwens and Lubrano (1998) for the application of the Griddy-Gibbs sampler to a generalized autoregressive conditionally heteroscedastic (GARCH) model for financial returns.

[5]Moreover, and in contrast to IS, the requirement to find a well-matched proposal distribution is facilitated by the dimension reduction invoked by the breaking down of the high-dimensional joint posterior into the lower dimensional conditionals, before any proposal distribution needs to be specified.

scalar state variable, z_t,

$$p(y_t|z_t, \boldsymbol{\theta}) \tag{7}$$

$$p(z_t|z_{t-1}, \boldsymbol{\theta}). \tag{8}$$

Using the generic notation in (7) and (8), the augmented posterior is

$$p(\boldsymbol{\theta}, \mathbf{z}_{1:T}|\mathbf{y}_{1:T}) \propto p(\mathbf{y}_{1:T}|\mathbf{z}_{1:T}, \boldsymbol{\theta})p(\mathbf{z}_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{9}$$

In certain cases, the model structure is such that a pure Gibbs scheme can be used to produce draws from $p(\boldsymbol{\theta}, \mathbf{z}_{1:T}|\mathbf{y}_{1:T})$ and, thus, from $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$; an insight obtained independently by Carter and Kohn (1994) and Frühwirth-Schnatter (1994) for the case of the linear Gaussian state space model, for example. However, implementation of such a scheme will, by definition, require both $p(\boldsymbol{\theta}|\mathbf{z}_{1:T}, \mathbf{y}_{1:T})$ and $p(\mathbf{z}_{1:T}|\boldsymbol{\theta}, \mathbf{y}_{1:T})$ to have recognizable forms. In more general cases, in which either the measurement or state equation has non-linear and/or non-Gaussian features, the resulting conditionals will not necessarily have a known closed form, which necessitates the addition of MH steps within the outer Gibbs loop. Such a treatment was the method of attack for large classes of models in the 1990s and 2000s. Relevant contributions here, which include specific treatments of the ubiquitous stochastic volatility (SV) model, are Polson *et al.* (1992), Jacquier *et al.* (1994), Shephard and Pitt (1997), Kim *et al.* (1998), Chib *et al.* (2002), Stroud *et al.* (2003), Chib *et al.* (2006), Strickland *et al.* (2006), Omori *et al.* (2007) and Strickland *et al.* (2008). The reviews of Fearnhead (2011) and Giordani *et al.* (2011) provide more detailed accounts and extensive referencing of this earlier literature.[6] (See also Appendix A.3.)

To conclude, and once again using the generic notation in (7) and (8), once draws have been produced from $p(\boldsymbol{\theta}, \mathbf{z}_{1:T}|\mathbf{y}_{1:T})$, the predictive pdf,

$$p(y_{T+1}|\mathbf{y}_{1:T}) = \int_{z_{T+1}} \int_{\mathbf{z}_{1:T}} \int_{\Theta} p(y_{T+1}|z_{T+1}, \boldsymbol{\theta}, \mathbf{y}_{1:T})p(z_{T+1}|z_T, \boldsymbol{\theta})p(\boldsymbol{\theta}, \mathbf{z}_{1:T}|\mathbf{y}_{1:T})d\boldsymbol{\theta}d\mathbf{z}_{1:T}dz_{T+1}, \tag{10}$$

can be estimated is the usual way, using subsequent draws from $p(z_{T+1}|z_T, \boldsymbol{\theta})$ and $p(y_{T+1}|z_{T+1}, \boldsymbol{\theta}, \mathbf{y}_{1:T})$, or by averaging the conditional predictives over all draws of z_{T+1} and $\boldsymbol{\theta}$.

3.2 The 21st Century: Intractable forecasting models

3.2.1 What do we mean by 'intractable'?

The MCMC methods that evolved during the late 20th century continue to serve as the 'bread and butter' of Bayesian forecasting, as will be made evident in Section 4. Nevertheless, more ambitious forecasting problems are now being tackled, and this has tested the mettle of some of the early algorithms! As a consequence, Bayesian forecasters have begun to exploit more modern computational techniques, and it is those techniques that we touch on briefly in this section.

It is convenient to characterize these newer computational developments as different types of solutions to so-called 'intractable' forecasting problems, by which we mean: *1)* Forecasts based on models with data

[6] We also note here the work of Chib and Greenberg (1994), in which the state space representation of an autoregressive moving average (ARMA(p,q)) model (Harvey, 1981) was exploited, and the principle of data augmentation invoked, in order to enable a MH-within-Gibbs scheme to be applied.

generating processes (DGPs) that cannot be readily expressed as a pdf, or probability mass function (pmf); *2)* Forecasts based on high-dimensional models, with a very large number of unknowns; *3)* Forecasts produced using extremely large data sets. Problems that feature problem *1)* are referred to as *doubly intractable* problems, as not only is $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ not available in its entirety (as is typical), but the DGP itself is also not able to be expressed analytically.

With reference to *1)*, the MCMC methods referenced so far entail the evaluation of the DGP as a pd(/m)f, either in the calculation of the acceptance probability in any MH sub-step, or in the specification of full conditionals in any 'pure' Gibbs step. Hence, they are infeasible when DGPs do not admit such a representation. Many such DGPs exist (see, for example, Martin *et al.*, 2022a, for a list of examples); however, particularly pertinent ones to mention here are continuous time models in finance with unknown transition densities (Gallant and Tauchen, 1996), α-stable models for financial returns (and/or their volatility) (Peters *et al.*, 2012; Martin *et al.*, 2019), and stochastic dynamic equilibrium models in economics (Calvet and Czellar, 2015). With regard to *2)*, whilst, in principle (and under appropriate regularity), a convergent MCMC chain can be constructed for any model, the exploration of a very high-dimensional parameter space via an MCMC algorithm *can* be prohibitively slow (Tavaré *et al.*, 1997; Rue *et al.*, 2009; Braun and McAuliffe, 2010; Lintusaari *et al.*, 2017; Betancourt, 2018; Johndrow *et al.*, 2019). Hence, in models with a very large number of unknowns – including those with multiple sets of high-dimensional latent variables – the production of an accurate MCMC-based estimate of $p(y_{T+1}|\mathbf{y}_{1:T})$ in a practical amount of time, may not be possible. Finally, regarding *3)* MCMC schemes require pointwise (i.e. for each y_t) evaluation of $p(\mathbf{y}_{1:T}|\boldsymbol{\theta})$ at each draw of $\boldsymbol{\theta}$, thereby inducing an $O(n)$ computational burden at each iteration in an MCMC chain. Such schemes can thus struggle when confronted with 'big data' (Bardenet *et al.*, 2017).

The methods in the following sections have been designed to solve one or more of these instances of intractability. The techniques in Section 3.2.2 do so whilst preserving the 'exact' nature of the estimate of $p(y_{T+1}|\mathbf{y}_{1:T})$, whilst those in Section 3.2.3 aim to produce an approximation of $p(y_{T+1}|\mathbf{y}_{1:T})$ only.

3.2.2 Exact computational solutions

The first two decades of the 21st century have witnessed a wealth of advances in both MCMC and IS-based algorithms. The goal of the newer MCMC algorithms – at their heart – is to explore the high mass region of the joint posterior more *efficiently*, in particular when the dimension of the parameter space is large. This, in turn, enables a more accurate estimate of $p(y_{T+1}|\mathbf{y}_{1:T})$ to be produced for a given computational budget. This goal has been achieved via a variety of means, which (in the spirit of Robert *et al.*, 2018, and Martin *et al.*, 2022b) can be summarized as: *i)* the use of more geometric information about the target posterior, most notably the use of 'Hamiltonian' up-dates (Neal, 2011b; Hoffman and Gelman, 2014); *ii)* the use of better MH candidate, or proposal distributions, including those that 'adapt' to previous draws (Nott and Kohn, 2005; Roberts and Rosenthal, 2009); *iii)* various types of combinations of multiple chains (Jacob *et al.*, 2011; Neal, 2011a; Neiswanger *et al.*, 2013; Glynn and Rhee, 2014; Huber, 2016; Jacob *et al.*, 2020); or *iv)* the use of ex-post variance reduction methods (Craiu and Meng, 2005; Douc and Robert, 2011; Owen, 2017; Baker *et al.*, 2019). We refer the reader to Green *et al.* (2015), Robert *et al.* (2018) and Dunson and Johndrow (2019) for detailed reviews of

modern developments in MCMC, and to Jahan *et al.* (2020) for an overview of the way in which certain of the newer methods manage the problem of scale – in terms of either the unknowns or the data, or both.

Whilst not designed expressly to deal with problems of scale, sequential Monte Carlo (SMC) methods – which exploit the principles of IS – have developed in parallel to the expansion of the MCMC stable. Developed initially for the sequential analysis of state space models, via methods of 'particle filtering' (Gordon *et al.*, 1993), SMC methods have evolved into a larger suite of methods used to perform both sequential and non-sequential tasks (Naesseth *et al.*, 2019; Chopin and Papaspiliopoulos, 2020). For the purpose of this review, the most pertinent development is the melding of particle filtering with MCMC in state space settings to produce a particle marginal MH (PMMH) algorithm (Andrieu *et al.*, 2011; Flury and Shephard, 2011; Pitt *et al.*, 2012; Doucet *et al.*, 2015; Deligiannidis *et al.*, 2018). Such algorithms tackle intractability type *1)* in the dichotomy of the previous section, by replacing an 'unavailable' likelihood function by an unbiased estimate – produced via the particle filter – in an MH algorithm which, under regularity, retains the posterior $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ as its invariant distribution. Given the increasingly important role played by PMMH, a brief algorithmic description of it is included in Algorithm 3 in Appendix A.4.[7]

3.2.3 Approximate computational solutions

In situations in which the dimension, or structure of the forecasting model, or the size of the data set, still precludes the use of an MCMC or a PMMH approach, an approximate method may be the only computational option. The cost of adopting such a solution is that these methods no longer directly target the exact predictive, $p(y_{T+1}|\mathbf{y}_{1:T})$; instead, an approximation of $p(y_{T+1}|\mathbf{y}_{1:T})$ becomes the goal.

The spirit of these methods is to approximate $p(y_{T+1}|\mathbf{y}_{1:T})$ via some feasible approximation to the posterior $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$. Denoting the posterior approximation generically by $g(\boldsymbol{\theta}|\mathbf{y}_{1:T})$, the resultant approximate predictive can be expressed as

$$g(y_{T+1}|\mathbf{y}_{1:T}) = \int_{\Theta} p(y_{T+1}|\boldsymbol{\theta}, \mathbf{y}_{1:T}) g(\boldsymbol{\theta}|\mathbf{y}_{1:T}) d\boldsymbol{\theta}, \tag{11}$$

in the case where there are only static unknowns. When the model features both static parameters and time-varying latent parameters, and exploiting the Markov property of the state process in (8), the approximate predictive can be represented as

$$g(y_{T+1}|\mathbf{y}_{1:T}) = \int_{z_{T+1}} \int_{z_T} \int_{\Theta} p(y_{T+1}|z_{T+1}, \boldsymbol{\theta}, \mathbf{y}_{1:T}) p(z_{T+1}|z_T, \boldsymbol{\theta}) p(z_T|\boldsymbol{\theta}, \mathbf{y}_{1:T}) g(\boldsymbol{\theta}|\mathbf{y}_{1:T}) d\boldsymbol{\theta} dz_T dz_{T+1}.$$

Given draws of $\boldsymbol{\theta}$ from $g(\boldsymbol{\theta}|\mathbf{y}_{1:T})$, and given an appropriate forward-filtering algorithm to draw from $p(z_T|\boldsymbol{\theta}, \mathbf{y}_{1:T})$ when needed, a simulation-based estimate of $g(y_{T+1}|\mathbf{y}_{1:T})$ can be produced in the usual way, either as a sample mean of the conditional predictives defined by the draws of $\boldsymbol{\theta}$ (and z_{T+1}), or by applying kernel density techniques to the draws of y_{T+1} from the conditional predictive.

[7] PMMH is actually a special case of the general pseudo-marginal MH technique (also sometimes denoted by the abbreviation 'PMMH'), in which a 'pseudo' likelihood, produced – in some manner or another – as an unbiased estimator of the true likelihood, is used within an MH algorithm. See, for example, the subsampling methods based on pseudo-marginal MCMC (Bardenet *et al.*, 2017; Quiroz *et al.*, 2018; Quiroz *et al.*, 2019) used expressly to improve the performance of MCMC in the case of a large-dimensional $\mathbf{y}_{1:T}$ (i.e. intractability type *3)*).

With reference to the taxonomy of intractable problems delineated in Section 3.2, the different methods of producing $g(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ (and, hence, $g(y_{T+1}|\mathbf{y}_{1:T})$) can be categorized according to whether they are being used to obviate *1)* or to tackle a problem of scale: *2)* and/or *3)*. Both ABC and BSL avoid the need to evaluate the DGP and, hence, are feasible methods in the doubly intractable settings of category *1)*. In brief, both methods require only *simulation*, not *evaluation*, of the DGP. The *approximation* of $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ arises, primarily, from the fact that both methods – in different ways – degrade the information in the full data set, $\mathbf{y}_{1:T}$, to the information contained in a set of summary statistics, $\eta(\mathbf{y}_{1:T})$. As such, the target becomes the so-called 'partial' posterior for $\boldsymbol{\theta}$, which conditions on $\eta(\mathbf{y}_{1:T})$, rather than $\mathbf{y}_{1:T}$. The quality of the approximation is thus dependent on the informativeness of the summaries, as well as on other forms of approximation invoked in the implementation of the methods. Vanilla versions of both algorithms are provided in Algorithms 4 (Appendix A.5) and 5 (Appendix A.6) respectively.

In contrast to ABC and BSL, VB and INLA still *target* the exact posterior $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$, but provide approximations that can be computationally convenient when the *scale* of the empirical problem is large in some sense (so problem *2)* and/or problem *3)*), often as a consequence of the specification of a high number of latent, or 'local', parameters in the model, in addition to the (usually) smaller set of 'global' parameters ($\boldsymbol{\theta}$ in our notation). Adopting the technique of the calculus of variations, VB produces an approximation of $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ that is 'closest' to $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ within a chosen *variational family*, whilst INLA applies a series of nested Laplace approximations (Laplace, 1774; Tierney and Kadane, 1986; Tierney *et al.*, 1989) to a high-dimensional latent Gaussian model to produce an approximation of $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$. Both VB and INLA exploit state-of-the-art optimization techniques, for the purpose of minimizing the 'distance' between $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ and the variational approximation in the case of VB, and for the purpose of producing the mode of the high-dimensional vector of latent states in the case of INLA. The basic principles of VB and INLA are provided in Appendices A.7 and A.8 respectively.

We refer the interested reader to Martin *et al.* (2022a) for an extensive review of all of these approximate Bayesian methods, as well as more complete coverage of the existing literature, including references to in-depth reviews of specific methods. Martin *et al.* also includes discussion of 'hybrid' methods that mix and match features of more than one computational technique, with the aim of tackling multiple instances of 'intractability' simultaneously.

Regardless of which approximation method is used, the hope is that the resulting approximate predictive $g(y_{T+1}|\mathbf{y}_{1:T})$ performs well relative to the inaccessible exact predictive, and that issue is addressed in certain work cited in the empirical reviews in Section 4.

3.3 The 21st Century: Misspecified forecasting models

3.3.1 The role of model specification in Bayesian forecasting

Inherent in the conventional Bayesian approach to forecasting is the assumption that the process that has generated the observed data tallies with the particular model that underpins the likelihood function. Bayesian model averaging (BMA) – and the resultant predictive in (3) – has evolved as a principled way of catering for *uncertainty* about the predictive model, and BMA remains a very important technique in the Bayesian arsenal. Nevertheless, underpinning BMA is still the assumption that the true process is spanned by the set of models over which one averages – i.e. that the so-called $\mathcal{M}$-closed view of the

world (Bernardo and Smith, 1994) prevails.

In response to these perceived limitations of the conventional approach, attention has recently been given to producing predictions that are 'fit for purpose', by focusing the Bayesian machinery on the *specific* goals of the predictive analysis at hand. In the following sections we briefly summarize three such approaches, all of which move beyond the conventional likelihood-based Bayesian up-date, and $\mathcal{M}$-closed paradigm: seeking to produce accurate predictions without recourse to the assumption of correct model specification.

3.3.2 Focused, or 'loss-based' Bayesian prediction

Loaiza-Maya *et al.* (2021) propose an approach to Bayesian prediction expressly designed for the context of misspecification. In brief, rather than a correct predictive model being assumed, a prior is placed over a class of *plausible* predictive models. The prior is then updated to a posterior via a sample criterion function that is constructed using a *scoring rule* (Gneiting and Raftery, 2007) that rewards the type of predictive accuracy (e.g. accurate prediction of extreme values) that is important for the particular empirical problem being tackled. With a criterion function that explicitly captures predictive accuracy replacing the likelihood function in the conventional Bayesian up-date, the explicit need for correct model specification is avoided.

Following Gneiting and Raftery (2007), and using generic notation, for $\mathcal{P}$ a convex class of predictive distributions on $(\Omega, \mathcal{F})$, the predictive accuracy of $P \in \mathcal{P}$ can be assessed using a scoring rule $S : \mathcal{P} \times \Omega \to \mathbb{R}$. If the value y eventuates, then the positively-oriented 'score' of the predictive P, is $S(P, y)$. The expected score under the true unknown predictive P_0 is defined as

$$\mathbb{S}(\cdot, P_0) := \int_{y \in \Omega} S(\cdot, y) dP_0(y). \tag{12}$$

A scoring rule is said to be proper relative to $\mathcal{P}$ if, for all $P, G \in \mathcal{P}$, $S(G, G) \geq S(P, G)$, and is strictly proper, relative to $\mathcal{P}$, if $S(G, G) = S(P, G) \iff P = G$. Scoring results are important mechanisms as they elicit truth telling within the forecasting exercise: if the true predictive P_0 *were* known, then in terms of forecasting accuracy as measured by the scoring rule $S(\cdot, \cdot)$ it would be optimal to use P_0.

Different scoring rules rewards different forms of predictive accuracy (see Gneiting and Raftery, 2007, Opschoor *et al.*, 2017, and Martin *et al.*, 2022c for expositions); hence the motivation to drive the up-date by the score that 'matters'. Since P_0 and the expected score $\mathbb{S}(\cdot, P_0)$ are unattainable in practice, an estimate based on $\mathbf{y}_{1:T}$ is used to define the sample criterion, $S_T(\boldsymbol{\theta}) := \sum_{t=0}^{T-1} S[p(y_{t+1}|\boldsymbol{\theta}, \mathbf{y}_{1:t}), y_{t+1}]$, where $p(y_{t+1}|\boldsymbol{\theta}, \mathbf{y}_{1:t})$ is the pdf associated with a given P. Adopting the *exponential* updating rule proposed by Bissiri *et al.* (2016) (see also Giummolè *et al.*, 2017, Holmes and Walker, 2017, Guedj, 2019, Lyddon *et al.*, 2019, and Syring and Martin, 2019), Loaiza-Maya *et al.* (2021) define the *generalized* (or *Gibbs*) posterior:

$$\pi_w(\boldsymbol{\theta}|\mathbf{y}_{1:T}) = \frac{\exp\left[wS_T(\boldsymbol{\theta})\right] \pi(\boldsymbol{\theta})}{\int_\Theta \exp\left[wS_T(\boldsymbol{\theta})\right] \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}, \tag{13}$$

for some learning rate $\omega \geq 0$, calibrated in a preliminary step. This posterior explicitly places high weight on – or *focuses* on – values of $\boldsymbol{\theta}$ that yield high predictive accuracy in the scoring rule $S(\cdot, \cdot)$. As

such, the process of building a Bayesian predictive as:

$$p_w(y_{T+1}|\mathbf{y}_{1:T}) = \int_\Theta p(y_{T+1}|\boldsymbol{\theta}, \mathbf{y}_{1:T})\pi_w(\boldsymbol{\theta}|\mathbf{y}_{1:T})d\boldsymbol{\theta}, \tag{14}$$

is termed *focused Bayesian prediction* (FBP) by the authors. By construction, when the predictive model, $p(y_{T+1}|\boldsymbol{\theta}, \mathbf{y}_{1:T})$, is misspecified, (14) will – out-of-sample – often outperform, in the chosen rule $S(\cdot, \cdot)$, the likelihood (or log-score)-based predictive in (2), and this is demonstrated in Loaiza-Maya *et al.* both theoretically and in extensive numerical illustrations.

Since a positively-oriented score can, equivalently, be viewed as the negative of a measure of predictive loss, FBP can also be referred to as 'loss-based' prediction. Such terminology is indeed adopted in Frazier *et al.* (2021), in which the principles delineated here are extended to high-dimensional models, and approximations to both $\pi_w(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ and $p_w(y_{T+1}|\mathbf{y}_{1:T})$ based on VB proposed, and validated.

3.3.3 Bayesian predictive combinations: Beyond BMA

The predictive distributions within the 'plausible class' referenced above may characterize a single dynamic structure depending on a vector of unknown parameters, $\boldsymbol{\theta}$, or may constitute weighted combinations of predictives from distinct models, in which case $\boldsymbol{\theta}$ comprises both the model-specific parameters and the weights. As such, FBP provides a coherent Bayesian method for estimating weighted combinations of predictives via predictive accuracy criteria, and without the need to assume that the true model is spanned by the set of constituent predictives – an assumption that underpins BMA, as we have noted.

A similar motivation underlies other contributions to the extensive Bayesian literature on estimating combinations of predictives that has now developed (and which rivals the large frequentist literature on forecast combinations that has also evolved[8]), with predictive performance – quantified by a range of user-specified loss measures – driving the posterior up-dating of the weights. Indeed, the Bayesian literature, having access as it does to powerful computational tools, has been able to invoke more complex weighting schemes than can be tackled via frequentist (optimization) methods. Notable contributions, including some also driven by the criterion of predictive calibration (Dawid, 1982; Dawid, 1985; Gneiting *et al.*, 2007), include Billio *et al.* (2013), Casarin *et al.* (2015a), Casarin *et al.* (2015b), Casarin *et al.* (2016), Pettenuzzo and Ravazzolo (2016), Aastveit *et al.* (2018), Bassetti *et al.* (2018), Baştürk *et al.* (2019) and Casarin *et al.* (2019). Once again adopting the language of Bernardo and Smith (2009), this literature seeks to move Bayesian predictive combinations beyond the $\mathcal{M}$-closed world of BMA to the $\mathcal{M}$-open world that accords with the reality of misspecification.

3.3.4 Bayesian predictive decision synthesis

A third approach that seeks to produce Bayesian predictions without relying explicitly on correct model specification is Bayesian predictive synthesis (BPS) (Johnson, 2017; McAlinn and West, 2019; McAlinn

[8]See Hall and Mitchell (2007), Ranjan and Gneiting (2010), Geweke and Amisano (2011) and Gneiting and Ranjan (2013) for early contributions to the frequentist forecast combination literature, and Wang *et al.* (2022) for a recent review. We note that whilst Geweke and Amisano (2011) is not explicitly Bayesian, in terms of estimating the optimal predictive combination, it provides important insights into the connection between the 'optimal linear pool' and BMA, and also uses Bayesian numerical methods in the production of some of the constituent forecast distributions.

et al., 2020; Aastveit *et al.*, 2022), recently expanded to Bayesian predictive *decision* synthesis (BPDS) by Tallman and West (2022). In particular, BPDS provides a sound decision-theoretic framework for constructing forecast combinations, and can be shown to encompass several commonly-suggested Bayesian forecasting approaches.

The starting point of BPDS is the production is a prior distribution over the m-dimensional unknown outcome $\mathbf{y}$ – implicitly indexed by $T+1$ in a time series forecasting application – and the information set $\mathcal{H}$, encoded via the J predictive models $\{h_j(\mathbf{y}|\mathbf{x}_j) : 1 \leq j \leq J\}$, where $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_J)$ denotes the collection of vectors of (possibly latent) dummy variables associated with a decision. The decision maker then constructs a predictive by integrating out $\mathbf{x}$ using a 'synthesis function' $\alpha(\mathbf{y}|\mathbf{x})$:

$$p(\mathbf{y}|\mathcal{H}) = \int_{\mathcal{X}} \alpha(\mathbf{y}|\mathbf{x}) \prod_{j=1}^{J} h_j(\mathbf{y}|\mathbf{x}_j) d\mathbf{x}_1 \ldots d\mathbf{x}_J.$$

The choice of the synthesis function $\alpha(\cdot|\mathbf{x})$ can be used to drive the analysis. For instance, in the case of forecast combinations, we can take $h_j(\mathbf{y}|\mathbf{x}) = p_j(\mathbf{y}|\mathbf{x}, \mathcal{M}_j)$, for some model $\mathcal{M}_j$, and then any set of synthesis functions $\alpha_j(\cdot|\mathbf{x})$ such that the combination density

$$p(\mathbf{y}|\mathcal{H}) = \int_{\mathcal{X}} \frac{\sum_{j=1}^{J} \omega_j \alpha_j(\mathbf{y}, \mathbf{x}_j \mid \mathbf{x}) \, p_j(\mathbf{y} \mid \mathbf{x}, \mathcal{M}_j)}{\sum_{k=1}^{J} \omega_k \alpha_k(\mathbf{y}, \mathbf{x}_k \mid \mathbf{x})} d\mathbf{x}_1 \ldots d\mathbf{x}_J.$$

is a valid density, for given weights $0 < \omega_j < 1$, $\sum_{j=1}^{J} \omega_j = 1$. Specific choices of $\alpha_j(\mathbf{y}, \mathbf{x}_j \mid \mathbf{x})$ then produce different forecast combination methods (see, Johnson, 2017, for a discussion); for example, in the case of McAlinn and West (2019) and McAlinn *et al.* (2020), the synthesis function is taken to be the density of a (possibly multivariate) dynamic linear factor model.

In an attempt to 'focus' the BPDS approach towards decisions that are tailored to a specific loss function, Tallman and West (2022) propose taking as their synthesis function, $\alpha_j(\mathbf{y}, \mathbf{x}_j|\mathbf{x}) = \exp\{\tau'(\mathbf{x}) S_j(\mathbf{y}, \mathbf{x}_j)\}$, where the score $S_j(\mathbf{y}, \mathbf{x}_j)$ is a k-dimensional vector that measures the utility one receives from realizing outcome $\mathbf{y}$ under decision $\mathbf{x}_j$, and $\tau(\mathbf{x})$ is a vector that weights the directional relevance of the scores $S_j(\mathbf{y}, \mathbf{x}_j)$.

While the BPS framework, as a whole, can set the tenor of the predictions towards dynamic forecast updates that produce predictions tailored towards a loss function of interest, via the choice of synthesis function $\alpha(\cdot|\cdot)$, BPS *is* ultimately tied to a 'likelihood-type' framework, or at least a log-loss function, due to the presence of the latent variables $\mathbf{x}$, which must be integrated out via assumed predictive models, $p_j(\mathbf{y}|\mathbf{x}, \mathcal{M}_j)$, and the production of these individual predictives using exact (i.e., likelihood-based) Bayesian methods.[9] While the BPDS approach can somewhat circumvent the reliance on the likelihood, due to its ability to focus on specific scores, this approach appears to be distinct from methods that entirely replace the likelihood function for the up-date. Therefore, a very interesting research path would involved combining the methods based on generalized posteriors discussed in Section 3.3.2 with the BPS framework.

[9] For an alternative combination approach based on dynamic weights and non-linear filtering, we refer to Billio *et al.* (2013), as cited earlier.

4 Selective Reviews of Bayesian Forecasting: Discipline-Specific Examples

Having established the necessary details regarding the production of Bayesian forecasts in general contexts, we now review how this general probabilistic mechanism is employed to produce Bayesian forecasts in several important empirical fields. In order to produce a comprehensive and up-to-date review of each area, a range of discipline experts have been invited to write the various sections, with the authorship flagged in the section headings. This means that the style of coverage differs somewhat across sections, as suits the topic, and as fits with the perspective of the authors; however, we have aimed to retain notation that – as far as possible – is consistent both across sections and with the notation used in the earlier parts of the paper, and in the technical appendix; and to ensure that the basic layout of all sections is the same. As noted earlier, other than in Section 4.3 – in which cross-sectional consumer choice data is modelled – and in Section 4.4, in which spatial models are briefly referenced, time series problems and forecasting are the primary focus.

4.1 Macroeconomics (Florian Huber and Gary Koop)

Central banks and other policy institutions routinely collect vast amounts of time series data on key macroeconomic outcomes. One stylized fact is that these data sets often display substantial co-movements and this calls for modeling all these series jointly to produce accurate point and density forecasts. This, however, leads to large-scale models that are prone to overfitting, ultimately resulting in weak out-of-sample forecasting performance. This helps explain the popularity of Bayesian methods for macroeconomic forecasting. They can easily handle many parameters and, through appropriate prior choice, deal effectively with questions related to model and specification uncertainty in macroeconomic models.

At a high level of generality, there are two modelling approaches used by macroeconomic forecasters. The first uses reduced-from models and imposes relatively little economic structure on the data. The second uses structural models such as dynamic stochastic general equilibrium (DSGE) models that are often estimated through Bayesian techniques; see, among many others, Smets and Wouters (2007), Adolfson *et al.* (2007) and Del Negro *et al.* (2016). However, reduced-form approaches have proved more popular and, in this section, our focus will be on them.

As stated above, macroeconomists are typically interested in modeling the joint evolution of a set of macroeconomic quantities. To set up a general framework for understanding the types of models used for forecasting, assume that an M-dimensional vector $\mathbf{y}_t$ is related to a K-dimensional vector of explanatory variables $\mathbf{x}_t$ through

$$\mathbf{y}_t = g(\mathbf{x}_t) + \boldsymbol{\varepsilon}_t, \tag{15}$$

where $g : \mathbb{R}^K \to \mathbb{R}^M$ is a function and $\boldsymbol{\varepsilon}_t$ is $\mathcal{N}(\mathbf{0}_M, \boldsymbol{\Sigma}_t)$.[10] This general specification nests most important reduced-form models commonly used in macroeconomics and can be used to explain the main issues that arise.

[10]Note that the Gaussianity assumption is not essential; mixtures of Gaussian distributions, for example, can be used to produce flexible error distributions if deemed necessary (see, for example, Clark *et al.*, 2022a, and Lenza and Primiceri, 2022).

For instance, if $\mathbf{x}_t = (\mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p})'$ contains p lags of $\mathbf{y}_t$, $g(\mathbf{x}_t) = \mathbf{A}\mathbf{x}_t$ is a linear function with $M \times K(= Mp)$ coefficient matrix $\mathbf{A}$, and $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}$ is constant over time we have a standard vector autoregressive (VAR) model. If we set $\mathbf{x}_t = \boldsymbol{f}_t$ with $\boldsymbol{f}_t$ denoting a set of $Q \ll M$ latent factors and $g(\boldsymbol{f}_t) = \boldsymbol{\Lambda}\boldsymbol{f}_t$ is linear with $\boldsymbol{\Lambda}$ being an $M \times Q$ matrix of factor loadings and $\boldsymbol{f}_t$ evolves according to some stochastic process (such as a VAR) we end up with a dynamic factor model (DFM, see Stock and Watson, 2011). Factor augmented VARs (Bernanke *et al.*, 2005) combine a VAR with a DFM. The dependent variables in the VAR part of the model are a subset of $\mathbf{y}_t$ plus a small number of factors.

Traditionally, VARs and factor models have been linear and homoskedastic. But there is a great amount of empirical evidence in most macroeconomic data sets of parameter change, both in the conditional mean and the conditional variance. This can be accommodated through particular choices for g and $\boldsymbol{\Sigma}_t$. For the latter, stochastic volatility processes have proved particularly popular. For the former, various parametric forms for g lead to time-varying parameter VARs (TVP-VARs) which assume that the coefficients of the VAR evolve according to a random walk. But it is also worth noting that there is an increasing literature which assumes g is unknown and uses Bayesian nonparametric methods to uncover its form (see, for example, Kalli and Griffin, 2018, Adrian *et al.*, 2021, and Huber *et al.*, 2020).

If we set $M = 1$ we obtain single-equation time series regressions which are particularly popular in inflation forecasting (e.g. based on the Phillips curve). If we additionally set $x_t = 1$ and allow for time varying parameters, we can obtain models such as the unobserved components stochastic volatility (UCSV) model of Stock and Watson (2007) which is commonly used to forecast inflation (for recent applications, see Chan *et al.*, 2013, Stock and Watson, 2016, and Huber and Pfarrhofer, 2021).

This general framework defines a class of likelihood functions. As per the outline in Section 2.1, Bayesian forecasting involves multiplying a chosen likelihood function by an appropriate prior to produce a posterior which can be used to produce the predictive density. The choice of prior and computational method used for posterior and predictive inference will be case specific and we will have more to say about some interesting cases below. But a few general comments are worth noting here. First, the choice of prior matters much more in models such as the large VAR, which have a large number of parameters relative to the number observations, than in models with fewer parameters such as the UCSV model or the DFM. Second, for linear homoskedastic models with conjugate priors analytical formulae for the posterior and the one-step-ahead predictive density are available. For all other cases, MCMC methods are available. These take the general form outlined in Section 3.1. However, as noted in Section 3.2, MCMC methods typically do not scale well and can be computationally slow in models involving large numbers of parameters (such as large VARs) or large numbers of latent states (such as TVP-VARs). Thus, the focus of many recent papers has been on developing either improved MCMC algorithms or approximate VB methods for speeding up computation. Thirdly, our discussion so far focuses on forecasting with a single model. In practice, it is common to find that forecasts improve if many models are combined. Thus, either BMA or, alternatively, the methods outlined in Section 3.3.3 are commonly used by macroeconomic forecasters.

With this general framework established, it is worthwhile to offer some additional detail about some of the most important 21st century developments and a discussion of how they have led to improvements in macroeconomic forecasting.

Large VARs Going back to early work such as Doan *et al.* (1984a), Bayesian VARs have been used successfully in a variety of macroeconomic forecasting applications. Recently, they have enjoyed even greater popularity due to the rise of the large VAR. The pioneering large VAR paper was Bańbura *et al.* (2010). Subsequently, dozens of papers have used large VARs for macroeconomic forecasting (see, among many others, Carriero *et al.*, 2009, Koop, 2013b, Carriero *et al.*, 2015, Giannone *et al.*, 2015, and Hauzenberger *et al.*, 2021b). Large VARs, involving dozens or even hundreds of dependent variables, have been found to forecast well and improve upon single-equation techniques and DFMs. Large VARs are heavily over-parameterized and, thus, Bayesian prior shrinkage has been essential in ensuring their forecasting success. We will discuss priors shortly, but at this point we highlight the fact that the use of large Bayesian VARs has been one of the major recent developments in macroeconomic forecasting.

Prior shrinkage in VARs Many different priors have been used with VARs. Traditionally, natural conjugate priors in the Minnesota tradition were used since these allowed for analytical posterior and one-step-ahead predictive inference. Definitions of these priors and discussions of their properties are available in standard sources such as Koop and Korobilis (2010) and Dieppe *et al.* (2016). These priors are subjective and require the user to select prior hyperparameters, most importantly those relating to the strength of prior shrinkage. In recent years, a range of alternative priors have been proposed which are more automatic, requiring fewer subjective prior choices by the researcher. For instance, Giannone *et al.* (2015) develop methods for estimating shrinkage parameters in conjugate priors, thus avoiding the need for their subjective elicitation. Chan (2022) also uses a conjugate prior and develops methods for selecting shrinkage parameters using a prior which relaxes some of the restrictive assumptions of the Minnesota prior. There are also a range of methods which automatically decide on the optimal degree of shrinkage for each VAR coefficient. These are the global-local shrinkage priors which are widely used with regressions and in machine learning applications and increasingly used with VARs.[11] Global-local shrinkage priors have the form

$$a_j \sim \mathcal{N}(0, \psi_j \lambda), \quad \psi_j \sim f_1, \quad \lambda \sim f_2,$$

where a_j is the j^{th} VAR coefficient, λ controls global shrinkage since it is common to all coefficients and ψ_j controls local shrinkage since it is specific to the j^{th} coefficient. The densities f_1 and f_2 are mixing densities and a large range of choices of them have been proposed. One choice leads to stochastic search variable selection, used with VARs in George *et al.* (2008), Koop (2013a) and Korobilis (2013) and many others. Other choices lead to the Dirichlet-Laplace prior used with VARs by Kastner and Huber (2021), or the normal-gamma and horseshoe priors used in Huber and Feldkircher (2019) and Cross *et al.* (2020), and there are many others. Since these priors are Gaussian at the first layer of the hierarchy, textbook MCMC algorithms for all the VAR parameters can be easily implemented.[12]

Adding stochastic volatility (SV) The other main development that has had a tremendous impact on applied macroeconomic forecasting in the 21st century is the development of models such as

[11] They are also used with DFMs to select the number of factors.

[12] In large VARs with global-local shrinkage priors, MCMC methods can nevertheless be very slow, with much faster VB methods developed in Gefang *et al.* (2022).

VARs that incorporate parameter change and nonlinearity. Put simply, the macroeconomic world is rarely linear and homoskedastic and models that relax these assumptions have been found to improve macroeconomic forecasting. These improvements lie not only in point forecasts, but more importantly in density forecasts. Given the increasing interest, by central banks and academics alike, in issues such as forecast uncertainty and tail risk, the fact that these new models produce more accurate predictive densities increases their value.

A popular specification for VARs with SV involves factorizing the error variance-covariance matrix as $\mathbf{\Sigma}_t = \mathbf{A}_0\mathbf{H}_t\mathbf{A}_0'$ with $\mathbf{A}_0$ being a lower triangular matrix with unit diagonals[13] and $\mathbf{H}_t = \text{diag}(e^{h_{1t}}, \ldots, e^{h_{Mt}})$ being a diagonal matrix with log-volatilities evolving according to simple stochastic processes such as a independent random walks or AR(1) processes. In an important contribution, Clark (2011) considers a VAR-SV and finds it to produce accurate point and density forecasts relative to homoskedastic models, with gains being particularly pronounced using forecast metrics involving the entire predictive density. Building on this insight, several other researchers have analyzed the role of heteroskedasticity in macroeconomic forecasting in VARs (see, for example, Clark and Ravazzolo, 2015, and Chiu *et al.*, 2017) and confirm the result that using SV pays off when the focus is on obtaining accurate density forecasts. However, a problem with the standard SV specification is that the computational burden relative to homoskedastic VARs is increased enormously. This makes it difficult to do Bayesian forecasting with large VARs with SV. As a remedy, Carriero *et al.* (2016) propose a simple common stochastic volatility (CSV) specification that assumes the shock variances to be driven by a single common volatility factor, maintaining conjugacy and thus leading to computationally efficient MCMC algorithms. They acknowledge that this model is simplistic but show that it yields much more accurate forecasts than homoskedastic VARs in a standard US macroeconomic forecasting application.

To gain more flexibility, researchers have developed algorithms that allow for estimating large VARs with M independent SV processes. Carriero *et al.* (2019) propose techniques that permit equation-by-equation estimation of such VARs and thus render estimation of larger models with SV feasible. Modified versions of this algorithm form the basis of several recent papers that combine large data sets with SV for macroeconomic forecasting (see, among others, Huber and Feldkircher, 2019, Chan, 2021, and Chan *et al.*, 2021).

Adding time variation in the VAR coefficients The previous discussion has emphasized that capturing changing error variances is key for obtaining precise forecasts. However, it may also be important to allow for structural change in the VAR coefficients themselves. One popular multivariate model that captures both changes in the VAR coefficients and error variances is the TVP-VAR-SV model proposed in Primiceri (2005), which assumes that the VAR coefficients $\boldsymbol{\beta}_t = \text{vec}(\boldsymbol{A}_t)$ are time-varying and evolve according to a multivariate random walk while $\mathbf{\Sigma}_t$ is a multivariate SV process. This model is a multivariate state space model which can be estimated using adaptations of the techniques outlined in Section 3.1.3. The innovations to the states govern the amount of time variation in the parameters. Various shrinkage priors (often based on the global-local shrinkage priors discussed above) have been proposed that allow for data-based selection of whether time variation in a corresponding coefficient is

[13] $\mathbf{A}_0$ can also be time varying.

necessary or not. These priors are typically elicited on the non-centered parameterization of the state space model (see Frühwirth-Schnatter and Wagner, 2010) and can help minimize overfitting concerns and produce improved forecasts.

D'Agostino *et al.* (2013) is an important early contribution to the macroeconomic forecasting literature using TVP models. This paper uses a small TVP-VAR with SV and shows that it produces more accurate point predictions, outperforming simpler univariate benchmarks and constant parameter VARs. One key shortcoming of this model, however, is that it only uses a small information set. This has led to several researchers proposing new methods that can be used in higher dimensions. Various approaches are possible, including models that restrict the TVP process (e.g. by imposing a factor structure which allows for time variation in a large number of parameters to be driven by a low number of factors, see Chan *et al.*, 2020). As mentioned above, shrinkage priors are used to keep the curse of dimensionality in check. These priors are typically used after transforming the model to allow for equation-by-equation estimation. Such approaches mean fairly high-dimensional TVP-VARs can be estimated without risk of over-fitting, and in a reasonable amount of time. MCMC-based forecasting with large TVP-VARs and regressions is also an active field of research and different shrinkage methods and advances in computation have led to improvements in the forecasting performance of TVP models (see, among many others, Huber *et al.*, 2021, and Hauzenberger *et al.*, 2021a). However, it is worth noting that if computation does become a concern, approximate methods (e.g. using the VB methods outlined in Section 3.2.3) can be used. Approaches which avoid the need for MCMC are developed in Koop and Korobilis (2013) and Koop and Korobilis (2018). The former paper proposes large approximate TVP-VARs based on forgetting factors whereas the latter one uses VB techniques to forecast inflation with large TVP regression models.

Bayesian nonparametric VARs Up to this point we have assumed that the conditional mean function g takes a known form. However, it could be that the functional form is unknown. Bayesian nonparametric techniques, such as Bayesian additive regression trees (BART, see Chipman *et al.*, 2010), Gaussian processes and kernel regressions (Adrian *et al.*, 2021) or infinite mixtures (Kalli and Griffin, 2018), allow the researcher to uncover such unknown functional forms and produce precise macroeconomic forecasts. In general, they have had great success, but they have been found to be particularly useful in studies that focus on the tails of predictive distributions or on the handling of outliers such as the ones experienced during the pandemic (see, for example, Huber *et al.*, 2020, and Clark *et al.*, 2022b).

Kalli and Griffin (2018) propose a nonparametric VAR that builds on an infinite mixture model with the mixture weights being driven by the lagged endogenous variables. They show, using US and UK data, that their model yields competitive forecasts, with accuracy gains in terms of point and density predictions increasing sharply for higher forecast horizons. Clark *et al.* (2022a) use BART-based VARs to perform tail forecasting of US output, unemployment and inflation in real time, finding that nonparametric techniques work well in the tails and for higher-order forecasts. With a particular focus on predictive accuracy during the pandemic, Huber *et al.* (2020) develop mixed frequency nonparametric VARs and show that these models yield substantially more precise nowcasts during the Covid-19 period.

Conclusions and further directions We have outlined how Bayesian methods have been used successfully for macroeconomic forecasting. Most of the discussion related to VARs, which are a class of models where Bayesian methods have proved particularly popular. But it is worth noting that empirically-relevant extensions (e.g. SV or TVP) can be added to other multivariate time series models such as DFMs or FAVARs, as can the VAR prior shrinkage methods (e.g. global-local shrinkage methods) we have discussed. It is also worth noting that we have focused on models that do not restrict the coefficients. However, restricted VARs are often used for forecasting. For instance, vector error correction models (which impose cointegrating restrictions) or multi-country VARs such as global VARs are restricted VARs.

We have also focused on forecasting as opposed to the closely related field of nowcasting. Mixed frequency VARs, which jointly model quickly-released, high-frequency variables (e.g. monthly variables such as surveys, employment and inflation) and slowly-released, low-frequency variables (e.g. quarterly variables such as GDP), have proved very popular with nowcasters. Bayesian methods are typically used with such models (see, for example, Schorfheide and Song, 2015, McCracken *et al.*, 2021, Huber *et al.*, 2020, and Koop *et al.*, 2020) and, in real-time nowcasting exercises they tend to perform well.

4.2 Finance (John Maheu, Worapree Maneesoonthorn and Gael Martin)

A pertinent question in financial analysis is whether the risks associated with financial assets – and the prices of those risks – are predictable in ways that are useful in applications such as portfolio allocation, risk management and derivative pricing. With risk factors typically being represented as latent distributional features of observable financial variables, it follows that two key goals in the statistical analysis of financial problems are: i) The accurate prediction of latent distributional features; and ii) The development of complex, non-linear state space models to underpin this prediction.

Both of these goals lend themselves naturally to a Bayesian treatment given, in turn, the automatic production of predictive *distributions* via the Bayesian paradigm, and the swathe of computational methods available to estimate complex models – most notably those with a latent variable structure. In particular, the growth in financial derivatives markets from the 1990s onwards has generated the need to model the underlying asset as a continuous time process, almost always augmented with continuous time processes for the asset volatility, and often via jump diffusions. Such models – whilst 'convenient' in the sense of allowing for closed-form solutions for derivative prices – are challenging from a statistical point of view, given that they typically need to be treated as a (discretized) non-linear state space model, and may require multiple sources of data to enable separate identification of model parameters and risk premia. Estimation of and forecasting with such models *is* nevertheless computationally feasible via Bayesian methods, with MCMC algorithms of one form or another forming the backbone of the early treatments (Eraker, 2001; Eraker *et al.*, 2003; Eraker, 2004; Forbes *et al.*, 2007; Johannes *et al.*, 2009).

We refer the reader to Jacquier and Polson (2011) and Johannes and Polson (2010) for comprehensive reviews of the application of Bayesian methods in finance up to the first decade of the 21st century. The coverage includes, in short, Bayesian approaches to: portfolio allocation, return predictability, asset pricing, volatility, covariance, 'beta' and 'value at risk' prediction, continuous time models (and discretized versions thereof), interest rate modelling, and derivative (e.g. option) pricing. Our goal in

the current review is to outline the more recent advances that have evolved over the last decade, in particular those that have exploited (in one way or another) new methodological advances, new sources of data, and modern computational techniques. In order, we shall briefly review: the use of diverse data sets, including derivative prices and high-frequency measures of financial quantities; the treatment of DGPs that are unavailable in closed form; the analysis of high-dimensional models; and the application of non-parametric modelling.

Multiple sources of financial data It is now a well-established fact that the constant volatility feature of a geometric diffusion process for a financial asset price is inconsistent with both the observed dynamics in return volatility and the excess kurtosis and skewness that characterizes the typical empirical return distribution; see Bollerslev *et al.* (1992) for an early review. The option pricing literature supports this finding, with certain empirical regularities, such as 'implied volatility smiles', seen as evidence that asset prices deviate from the geometric Brownian motion assumption that underlies the Black and Scholes (1973) option price (Bakshi *et al.*, 1997; Hafner and Herwartz, 2001; Lim *et al.*, 2005). Hence, the 21st century has seen the proliferation of many alternative specifications for asset prices, and associated theoretical derivative prices, most of which are nested in a general framework of (discretized) bivariate jump diffusion models for the asset itself and its volatility. Allied with these developments has been the growth in access to transaction-level 'high-frequency' data – in *both* the spot and options markets – which, in itself, has spawned new approaches to inference and forecasting in the financial sphere.

The Bayesian literature has brought to bear on this problem the power of computational methods – both established, and more recent – to enable the multivariate state space models that have emerged from this literature to be estimated, and probabilistic predictions of all dynamic variables – the return itself, volatility, random jumps (in either the return or the volatility, or both), and various risk premia – to be produced. With reference to the generic notation for a state space model in (7) and (8), Bayesian approaches over the last decade can be categorized according to the specification adopted for the (multivariate) measurement at time t, $\mathbf{y}_t$ and, hence, for the (multivariate) state, $\mathbf{z}_t$, being modelled and forecast. Some work exploits data from both the spot and options market to predict volatility and its risk premia (Maneesoonthorn *et al.*, 2012), and theoretical option prices (Yu *et al.*, 2011; Carverhill and Luo, 2022[14]); other work combines 'low-frequency' daily observations on returns with high-frequency measures of volatility and/or price jumps to predict (in some combination) returns, volatility, and the size and occurrence of price jumps (Jin and Maheu, 2013; Maneesoonthorn *et al.*, 2017; Frazier *et al.*, 2019); whilst further work combines daily returns with futures prices in predicting various financial quantities of interest (Fileccia and Sgarra, 2018; Gonzato and Sgarra, 2021).

Financial models that are 'unavailable' All but one of the papers cited in the previous paragraph share a common feature - namely, a DGP that can be expressed as a probability density (or mass) function. With reference to (9), it is the availability of a closed form for $p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}|\boldsymbol{\theta}) =$

[14]We note that whilst a time series model is constructed in the case of these two references, the (out-of-sample) prediction of option prices is across the cross section of strike prices and maturities. We also make note of Fulop and Li (2019) who exploit spot and options data to produce filtered estimates (as opposed to strictly out-of-sample predictions) of latent volatility and price jump intensity.

$p(\mathbf{y}_{1:T}|\mathbf{z}_{1:T}, \boldsymbol{\theta})p(\mathbf{z}_{1:T}|\boldsymbol{\theta})$, that renders feasible the MCMC methods used in the said works. In contrast, Frazier *et al.* (2019) adopt a process for the latent log-volatility that is driven by an α-stable innovation, such that $p(\mathbf{z}_{1:T}|\boldsymbol{\theta})$ is unavailable, and MCMC infeasible as a consequence. Instead, ABC is adopted for inference, and an approximate predictive of the form of (11) produced instead. In addition to providing theoretical validation of the approach, the authors demonstrate, in range of different simulation settings, that despite inaccuracy at the posterior level, the approximate predictive is always a very close match to the exact predictive. Related work in which an ABC method is used to conduct forecasting appears in Canale and Ruggiero (2016), Kon Kam King *et al.* (2019), Virbickaitė *et al.* (2020) and Pesonen *et al.* (2022). ABC treatment of a conditional likelihood for a time series of financial returns, $p(\mathbf{y}_{1:T}|\mathbf{z}_{1:T}, \boldsymbol{\theta})$, that is unavailable in closed form is also investigated in Creel and Kristensen (2015), Martin *et al.* (2019) and Chakraborty *et al.* (2022), with Chakraborty *et al.* (2022) proposing a modularized version of ABC. For other recent Bayesian treatments of intractable models of this sort that continue to exploit MCMC principles (with or without an ABC component), see Vankov *et al.* (2019) and Müller and Uhl (2021).[15]

Large financial models Thus far, we have reviewed Bayesian treatments of models for single financial assets. That is, the models may have specified multiple latent components, and potentially multiple measurements, but they still aim to explain (and forecast) quantities related to a *single* asset. Models for multiple assets are also critically important in financial applications, with the relationship between financial assets determining the extent to which diversification can be achieved, as well as how risks permeate the various sectors of the financial market. Indeed, Bayesian methods are particularly suitable for dealing with such multivariate models, since the dimensionality of $\mathbf{z}_{1:T}$ is typically much larger than that of $\mathbf{y}_{1:T}$ and, hence, challenging to deal with via any other means.

Chib *et al.* (2009) provide an early review of the Bayesian analysis of multivariate SV models, with all work up to this point utilizing traditional MCMC techniques, and the statistical and predictive analysis limited to relatively low-dimensional systems (up to ten assets). Subsequent work has focused on the development of more flexible multivariate distributions (Nakajima, 2017), and the use of sparse factor structures and shrinkage priors in constructing larger-dimensional models (Zhou *et al.*, 2014; Kastner *et al.*, 2017; Baştürk *et al.*, 2019). More recently, with the advances made in VB methods, inference and prediction in very large-dimensional financial models is now possible (Gunawan *et al.*, 2021; Chan and Yu, 2022; Frazier *et al.*, 2022; Quiroz *et al.*, 2022). There is also a growing interest in the prediction of *co-movements* of various sorts, with: Bernardi *et al.* (2015) predicting the interdependence between U.S. stocks with Bayesian time-varying quantile regressions; Geraci and Gnabo (2018) capturing and predicting the interconnectedness of financial institutions through Bayesian time-varying VARs; and Alexopoulos *et al.* (2022) modelling and predicting common jump factors in a large panel of financial returns.

Bayesian nonparametric modelling in finance As noted, simple parametric assumptions such as additive Gaussian innovations are inconsistent with the stylized features of financial data. Whilst

[15]The citation of Creel and Kristensen (2015), Martin *et al.* (2019), Vankov *et al.* (2019) and Müller and Uhl (2021) is relevant to this review, despite these references not having an *explicit* component on forecasting.

more suitable non-Gaussian/non-linear models can be built (as highlighted above), Bayesian nonparametric modelling allows for further flexibility via the incorporation of Dirichlet process mixture (DPM) structures. Such an approach has been shown to provide robustness to distributional assumptions and can improve point forecasts, but the main gain has been significant improvements in the accuracy of predictive densities, and of risk measures derived from those densities. The advancement of the literature in this direction has been aided by the stick-breaking representation (Sethuraman, 1994) and the introduction of the slice sampler (Walker, 2007; Kalli *et al.*, 2011).

Jensen and Maheu (2010) introduce an extension to a standard SV model to capture the unknown return innovation distribution via a DPM. The DPM specification has also been inserted into other popular models in finance, with: Jensen and Maheu (2014) adopting a DPM to jointly model the return and future log-volatility distribution; Delatola and Griffin (2013) capturing the so-called leverage effect; Ausín *et al.* (2014) applying a DPM to univariate GARCH models; and Kalli and Griffin (2015) using Bayesian nonparametric modelling to aggregate autoregressive processes to produce a SV model with long-range dependence. Extensions to multivariate financial models have also occurred: in a multivariate GARCH setting in Jensen and Maheu (2013); and in a Cholesky-type multivariate SV model in Zaharieva *et al.* (2020).

A potential drawback of the DPM model is that it neglects time dependence in the unknown distribution. An important extension of the DPM prior is the hierarchical Dirichlet process of Teh *et al.* (2006), which allows for the construction of a prior for an infinite hidden Markov model (IHMM), which allows for time dependence in a flexible manner. The introduction of the beam sampler of Van Gael *et al.* (2008), which extends the slice sampler, renders conventional posterior sampling methods for finite-state Markov switching models (Chib, 1996) feasible in the IHMM. The IHMM structure has been used to model the univariate GARCH distribution (Dufays, 2016), and the multivariate GARCH distribution (Li, 2022); and to provide a nonparametric model for realized measures, including realized covariance matrices (Jin and Maheu, 2016; Liu and Maheu, 2018; Jin *et al.*, 2019), with all papers documenting very large improvements in density forecast accuracy from the IHMM. Other applications of the IHMM include Shi and Song (2016), who use the IHMM to date and forecast speculative bubbles, and who also adopt a version with GARCH effects; Yang (2019), who studies the relationship between stock returns and real growth with a multivariate IHMM model; and more recently Jin *et al.* (2022), who employ the DPM prior in the infinite Markov pooling of predictive distributions, with forecasting applications to interest rates, realized covariances and asset returns. Other approaches to time dependence in Bayesian nonparametrics for finance include Griffin and Steel (2011), who introduce a time-dependent stick breaking process in a general setting and develop a SV model for returns. More recently, Sun *et al.* (2020) use a weighted DPM to forecast return distributions, while Zamenjani (2021) allows for lagged covariates to impact the weights in the DPM model through a probit stick-breaking process.

4.3 Marketing (Rubén Loaiza Maya and Didier Nibering)

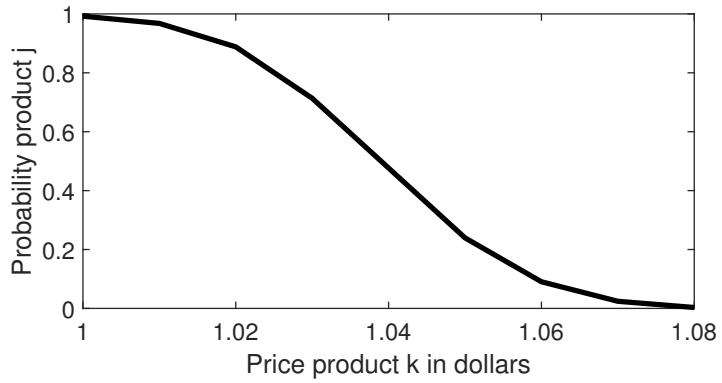
Bayesian methods are applied to a wide range of marketing problems; see Rossi and Allenby (2003) for a review of the early literature. More recently, these methods have been increasingly used for the purpose of prediction, for instance in customer choice behaviour (Toubia *et al.*, 2019; Araya *et al.*, 2022),

customer demand (Posch *et al.*, 2022), customer satisfaction (Mittal *et al.*, 2021), dynamic pricing (Bastani *et al.*, 2022), advertising effectiveness (Danaher *et al.*, 2020; Loaiza-Maya *et al.*, 2022) and recommender systems (Ansari *et al.*, 2018). Given the large variety of marketing applications, we focus in this section on the modelling of customer choice to illustrate the key principles of Bayesian prediction in marketing problems.

A common problem in marketing is that of setting the price level of a set of products so that total profits are maximized. To estimate these optimal prices, predictions of how customers will react to price changes are crucial. Predictions of customer choices under different marketing environments can be constructed by choice models. These models are estimated using data about the product choices of customers in the marketplace, a survey, an experiment, etc. (Rossi *et al.*, 2012).

Figure 1 shows a prediction of interest in this context: the predicted purchase probability of a customer for product j (y-axis) as a function of the price of product k (x-axis). The predicted purchase probability can be constructed for a customer for which only a few choices are observed, or for a new customer for which we do not observe choices in the data. Although this section, as noted earlier, focuses on cross-sectional data, choice models can also be applied to the forecasting of future choice probabilities by using panel data (Gilbride and Allenby, 2004; Terui *et al.*, 2011).

Figure 1: Predicted purchase probability of customer i



The two most popular models used to predict choice behaviour are the multinomial logit and multinomial probit models. The multinomial logit model imposes the independence of irrelevant alternatives (IIA) property (McFadden, 1989), which means that it cannot capture general substitution patterns among choice alternatives. The IIA property of this model can be relaxed under certain assumptions by extending the multinomial logit model to a nested logit model (Poirier, 1996; Lahiri and Gao, 2002) or a random parameter logit model (Train, 2009).

On the other hand, the multinomial probit model does not impose the IIA property, and as such is commonly used in the analysis of economic choice behaviour, where complementary and substitution effects are important. For instance, the multinomial probit model has been recently used in the analysis of car choices (Karmakar *et al.*, 2021), grocery brand choices (Miyazaki *et al.*, 2021), employment choices (Mishkin, 2021), and car parking choices (Paleti, 2018). This section presents a review of Bayesian prediction with the multinomial probit model.

Multinomial probit model specification The variable of interest is $y_i \in \{0, 1, 2, \ldots, J\}$, which indicates the choice made by individual i among a set of $J + 1$ choice alternatives. This choice is modeled to be conditional on a set of J latent utilities $\mathbf{z}_i = (z_{i1}, \ldots, z_{iJ})'$, so that

$$p(y_i|\mathbf{z}_i) = \begin{cases} I\left[z_{iy_i} = \max(\mathbf{z}_i)\right] & \text{if } \max(\mathbf{z}_i) > 0, \\ I[y_i = 0] & \text{if } \max(\mathbf{z}_i) \leq 0, \end{cases} \tag{16}$$

where $I[A]$ is one if statement A is true and zero otherwise. The base category $j = 0$ is one of the choice alternatives, which is selected *a priori*. The base category is observed whenever all the latent utilities are less than zero.

The utilities are expressed in terms of r predictors via a linear Gaussian model,

$$p(\mathbf{z}_i|X_i, \boldsymbol{\theta}) = \phi_J\left(\mathbf{z}_i; X_i\boldsymbol{\beta}, \Sigma\right), \tag{17}$$

where $\phi_J\left(\mathbf{z}; \boldsymbol{\mu}, C\right)$ denotes a J-variate normal density with mean $\boldsymbol{\mu}$ and covariance matrix C, X_i a $J \times r$ matrix of predictor values, $\boldsymbol{\beta}$ an $r-$dimensional vector of coefficients, and Σ a covariance matrix that captures complementary and substitution effects between the choice alternatives.

Combined, (16) and (17) give rise to the augmented likelihood function of the multinomial probit model

$$p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, X) = \prod_{i=1}^{n} p(y_i|\mathbf{z}_i)p(\mathbf{z}_i|X_i, \boldsymbol{\theta}), \tag{18}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \Sigma\}$, $\mathbf{y} = \{y_i\}_{i=1}^n$, $\mathbf{z} = \{\mathbf{z}_i\}_{i=1}^n$, and $X = \{X_i\}_{i=1}^n$, with n the total number of individuals. For a given prior distribution $p(\boldsymbol{\theta})$, the augmented posterior distribution of the model is given as

$$p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}, X) \propto p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, X)p(\boldsymbol{\theta}). \tag{19}$$

Albert and Chib (1993) were the first to propose the use of data augmentation (see Section 3.1.3 herein) for conducting Bayesian analysis of the multinomial probit model.

The predictive distribution Consider now an individual s, with predictor values X_s, whose choice behaviour we would like to predict. The predictive for individual s, can be written as

$$p(y_s|X_s, \mathbf{y}, X) = \int_\Theta \int_{\mathbf{z}_s} p(y_s|\mathbf{z}_s)p(\mathbf{z}_s|\boldsymbol{\theta}, X_s)d\mathbf{z}_s \int_{\mathbf{z}} p(\boldsymbol{\theta}, \boldsymbol{z}|\mathbf{y}, X)d\mathbf{z}\, d\boldsymbol{\theta}, \tag{20}$$

from which the predictive choice probabilities $Pr(y_s = j) = p(j|X_s, \mathbf{y}, X)$, such as those in Figure 1, can be constructed. The specification and computation of the predictive distribution in (20) poses three key challenges.

First, $p(y_s|\mathbf{z}_s)$ requires a choice of base category. This choice affects the prior predictive choice probabilities, and hence the (posterior) predictive choice probabilities can be sensitive to the choice of base category; see Burgette and Nordheim (2012). Burgette *et al.* (2021) propose a symmetric prior specification to address this problem. The parameters $\boldsymbol{\theta}$ are not identified under this prior, but this does not affect the predicted probabilities.

Second, the parameters $\boldsymbol{\theta}$ lack scale identification, as $p(y_i|\mathbf{z}_i) = p(y_i|c\mathbf{z}_i)$ for any positive scalar c. Different solutions have been proposed to fix the scale, all based on a constraint on the specification of

Σ. For instance, McCulloch *et al.* (2000) fix the first leading element of Σ to unity. This approach is sensitive to the ordering of the choice categories in the model. Burgette and Nordheim (2012) fix the trace of Σ, which is invariant to the way the choice categories enter the model.

Third, the computation of $p(y_s|X_s, \mathbf{y}, X)$ involves the evaluation of the integrals over the latent utilities in $\mathbf{z}_s$ and $\mathbf{z}$. Since no analytical solution for these integrals is available, they are solved with MCMC sampling steps. The latent utility of each choice category is sampled from a univariate truncated normal, conditional on the latent utilities for all the other choice alternatives, for each individual (McCulloch and Rossi, 1994). Conditional on the draws for the latent utilities, sampling $\boldsymbol{\beta}$ from its full conditional is straightforward. Generating from the conditional distribution of Σ is nonstandard as the scale restrictions on Σ have to be taken into account.

Scalable Bayesian prediction In addition to the challenges delineated above, it is difficult to scale $p(y_s|X_s, \mathbf{y}, X)$ to problems with large choice sets or a large number of observations. Recent advances in the computation of the predictive have focused on tackling the scalability issues in J and n, as we discuss below.

When considering a full covariance matrix specification for Σ, the total number of parameters increases quadratically with J. For problems with large choice sets and small samples, this implies that the ratio of total number of parameters to total number of observations is large, making it difficult to construct accurate predictions. Loaiza-Maya and Nibbering (2022b) propose a spherical transformation of the covariance matrix of the latent utilities that imposes a parsimonious factor structure and a trace restriction. As a result, the total number of parameters grows only linearly with J. The authors demonstrate that this parsimonious structure leads to improved predictive performance over full covariance matrix specifications.

Additionally, as noted above, the construction of the posterior predictive entails evaluation of the integral over the latent utilities $\boldsymbol{z}$. Although MCMC is able to solve this integral, it does so by generating the utility vector for each individual from a multivariate truncated normal, which is a computationally costly exercise (McCulloch and Rossi, 1994; Botev, 2017). This renders MCMC algorithms impractical for problems where a large n is considered.

VB can be employed to tackle problems with large n. Adapting the generic descriptions of VB in Section 3.2.3 and Appendix A.7, the application of VB in this setting considers the class of approximating densities $\mathcal{Q}$ with elements $q_\lambda(\boldsymbol{\theta}, \mathbf{z}) \in \mathcal{Q}$, indexed by the variational parameters $\boldsymbol{\lambda}$. The exact augmented posterior is approximated by $q_{\hat{\lambda}}(\boldsymbol{\theta}, \mathbf{z})$ with an optimal variational parameter equal to

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda} \in \Lambda}{\arg\min}\, \mathrm{KL}\left[q_\lambda(\boldsymbol{\theta}, \mathbf{z}) || p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}, X)\right], \tag{21}$$

where KL denotes the Kullback-Leibler divergence. Then the variational predictive is constructed as

$$\hat{p}_\lambda(y_s|X_s, \mathbf{y}, X) = \int_\Theta \int_{\mathbf{z}_s} p(y_s|\mathbf{z}_s) p(\mathbf{z}_s|\boldsymbol{\theta}, X_s) d\mathbf{z}_s \int_{\mathbf{z}} q_{\hat{\lambda}}(\boldsymbol{\theta}, \mathbf{z}) d\mathbf{z}\, d\boldsymbol{\theta}.$$

Calibration of the variational approximation requires a scale-identified expression for $p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, X)$. To achieve this, Girolami and Rogers (2006) consider an identity matrix covariance structure, while Fasano and Durante (2022) fix Σ at predetermined values. Loaiza-Maya and Nibbering (2022a) propose a

method for a multinomial probit model with a factor covariance structure. This method uses the hybrid variational approximation $q_\lambda(\boldsymbol{\theta}, \mathbf{z}) = q_\lambda(\boldsymbol{\theta})p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}, X)$ introduced by Loaiza-Maya *et al.* (2022).

4.4 Electricity Pricing and Demand (Anastasios Panagiotelis)

Forecasting in electricity markets is critical for efficient day-to-day operation of power grids, long-term planning of infrastructure and increasingly, at a disaggregated level, for the management of smart grids. The scope of this subsection will cover forecasting electricity prices, electricity load/demand and generation by source of power, primarily wind and solar. Hereafter these problems will collectively be referred to as 'electricity forecasting'. Motivations for electricity forecasting can be found in general reviews such as Weron (2014) for price forecasting, Lindberg *et al.* (2019) for load forecasting, Antonanzas *et al.* (2016) for solar power forecasting and Giebel and Kariniotakis (2017) for wind power forecasting. These reviews indicate that the majority of work in electricity forecasting does not employ a Bayesian approach; however notwithstanding this, Bayesian methods have found success in the field.

There are very few instances of Bayesian forecasting in electricity markets that predate the early 2000s, although we now cover some notable exceptions. Bunn (1980) consider the case of updating load forecasts in an online fashion by computing a Bayesian model average of load profiles of a cloudy and sunny day. Meanwhile, Bayesian VARs have been used by Gunel (1987), Beck and Solow (1994) and Joutz *et al.* (1995) to forecast energy demand, nuclear power generation and demand prices and consumption respectively. A Bayesian VAR shrinks autoregressive coefficients to either a random walk or white noise depending on whether data are stationary or non-stationary and was popularized in macroeconomics by Doan *et al.* (1984b) (see also Section 4.1). The performance of Bayesian VARs in early electricity forecasting applications is mixed; Beck and Solow (1994) find evidence in favour of Bayesian autoregression, Joutz *et al.* (1995) find that Bayesian VARs are effective for forecasting demand, but not price, while Gunel (1987) does not find any improvement at all from using Bayesian VARs rather than conventional ARIMA models.

With the advent and popularization of MCMC methods, Bayesian forecasting has begun to find greater success in the field of electricity forecasting. In the literature of roughly the past two decades, there are three common major motivations for using Bayesian forecasting, two of which have antecedents in the earlier literature. The first, is the use of 'Bayesian models'[16], which have now grown well beyond Bayesian VARs to include models with latent volatilities, models with a spatial dimension, and Bayesian neural networks. The second is the use of BMA for forecast combination. The third is the production of full probabilistic forecasts via Bayesian computation. These are now each discussed in turn.

Bayesian models The structure inherent in many electricity forecasting problems provides a motivation for the innovative use of priors to improve forecasting accuracy. Although the early literature cited before found somewhat ambiguous results when comparing Bayesian VARs to classical alternatives, more recent work finds evidence in favour of a Bayesian approach; see Raviv *et al.* (2015) for point forecasts and Gianfreda *et al.* (2020) for both point and density forecasts. An important aspect of this work is

[16]By a 'Bayesian model' we generally mean a model with a prior and likelihood estimated by Bayesian inference. Bayesian methods for finding tuning parameters such as the automatic relevance determination (see Hippert and Taylor, 2010, for an example in electricity forecasting), and Bayesian optimisation lie beyond the scope of this section.

the exploitation of the intraday nature of the data, since typically hourly prices are stacked in a VAR model. The intraday structure lends itself to priors that shrink parameters corresponding to consecutive hours of the day that are close to one another. An early application of this approach can be seen in Cottet and Smith (2003).

Since electricity data are increasingly available not only at a high temporal frequency but also at a high spatial resolution, there are further examples in the literature of using priors to exploit neighbourhood structure. Examples include Ohtsuka *et al.* (2010) who use spatial ARMA processes to predict electricity load in nine Japanese regions, and Gilanifar *et al.* (2019) who use spatio-temporal Gaussian processes to forecast residential-level electricity demand. Even where spatial information is unavailable, hierarchical models estimated using Bayesian methods have been used to produce disaggregate energy demand forecasts; examples can be found in Mori and Nakano (2014) and Wang *et al.* (2017) who use Gaussian processes, and Grillone *et al.* (2021) who use regression. Informative hierarchical priors have been used in instances where data sets are small in size, or unavailable; for example, Pezzulli *et al.* (2006) elicit priors for future trajectories of temperature in the winter using past observations, and Launay *et al.* (2015) elicit priors for the electricity demand of 'non-metered' households using data on 'metered' households.

While the aforementioned examples take a Bayesian approach to exploit the use of priors in novel ways, another strain of the Bayesian forecasting literature is based on estimating models with latent variables. Examples in electricity forecasting include a latent jump process for price spikes (Chan *et al.*, 2014) and SV models (Smith, 2010; Kostrzewski and Kostrzewska, 2019). Also, in recent years, Bayesian analysis of machine learning models has become increasingly popular. This includes neural network models (Brusaferri *et al.*, 2019; Ghayekhloo *et al.*, 2019; Capone *et al.*, 2020), where VB is typically used. Also, Bayesian regression trees (BART) have been applied to electricity forecasting by Nateghi *et al.* (2011) and Alipour *et al.* (2019), who find that they outperform non-Bayesian counterparts. Finally, there is an extensive literature on using Bayesian networks for forecasting in energy; see Adedipe *et al.* (2020) for a review of these methods in forecasting wind generation.

Bayesian model averaging (BMA) As noted earlier, the importance of forecast combination is widely appreciated in the forecasting literature. Whilst, as highlighted in Section 3.3.3, many different Bayesian approaches to forecast combination have now been explored, BMA remains a very important method in the sphere of electricity forecasting. As described in Section 2.1, BMA uses posterior model probabilities as combination weights. Whenever the choice of model is parameterized, the predictive density has an interpretation as a forecast combination. Examples include Smith (2000) who combines forecasts from regression models that include different predictor sets, and Panagiotelis and Smith (2008) who average over models with different combinations of skew and symmetric marginal distributions.

It is also common in the electricity forecasting literature to produce point forecasts from different models and then combine these using BMA as a post-processing step. This approach grew out of research combining ensembles of forecasts from numerical weather predictions (NWPs) (Raftery *et al.*, 2005; Sloughter *et al.*, 2010). In the NWP setting, forecasts are the outputs of deterministic physical models. Statistical models are then formed by assuming that for $k = 1, \ldots, K$, $p(y_t|a_k, b_k, f_k, \sigma^2, \mathcal{M}_k) \sim$

$N(a_k + b_k f_k, \sigma^2)$, where f_k is the k^{th} NWP and a_k, b_k and σ^2 are additional parameters. These statistical models are then combined using the usual BMA machinery described by (3). Uncertainty over a_k, b_k and σ^2 is integrated out in the usual way, and there are no additional parameters since the f_k are obtained deterministically. This approach has been used in energy forecasting by Coelho *et al.* (2006), who motivate forecasting rainfall as a input into forecasting generation from hydroelectric dams, and Du (2018) who use wind forecasts to predict generation from wind farms.

The work of Raftery *et al.* (2005) has been subsequently extended to the case where the forecasts f_k are not the outputs of deterministic physical models but are point forecasts from statistical models, each with their own unknown parameters. For example Nowotarski *et al.* (2014) adopt the approach of Raftery *et al.* (2005) but where the f_k are obtained from statistical time series models with parameters estimated using frequentist techniques. This approach is not fully Bayesian (despite being referred to as BMA in the literature), since although the model average integrates over the uncertainty in a_k, b_k and σ^2 it does not integrate over uncertainty in the parameters of the underlying time series models used to generate the point forecasts f_k.[17] This approach is found to perform poorly relative to frequentist forecast combination schemes. In a similar vein, Hassan *et al.* (2015) and Raza *et al.* (2017) combine electricity load forecasts from different neural networks.

Probabilistic forecasting A common motivation for taking a Bayesian approach is the ease with which the computational machinery of MCMC or approximate methods produce a full predictive density rather than only point forecasts. Key operational decisions in electricity forecasting depend on quantities other than the predicted mean; see Nowotarski and Weron (2018) and references therein for discussion. While the importance of probabilistic forecasting is often highlighted in Bayesian papers it is not always the case that forecasts are evaluated in a way that assesses the quality of the full predictive distribution[18]. For example, often probabilistic forecasts are summarized by prediction intervals, and the empirical coverage of these intervals used as a means of checking model quality; for an early example see Pezzulli *et al.* (2006), and more recently Wang *et al.* (2017) and Kostrzewski and Kostrzewska (2019), where the latter show that Bayesian methods compare favourably to non-Bayesian alternatives for forecasting electricity prices. Kostrzewski and Kostrzewska (2019) also evaluate quantile forecasts using the pinball loss, as do Yang *et al.* (2019) and Sun *et al.* (2019) both for forecasting residential-level load (net of solar PV generation in the latter case).

However, the use of scoring rules (Gneiting and Raftery, 2007) and, hence, the explicit recognition of the distributional form of the forecasts, *is* becoming increasingly popular as a means of evaluating predictive distribution in both Bayesian and non-Bayesian electricity forecasting. The CRPS is particularly amenable to Bayesian inference since it is usually approximated using a Monte Carlo sample from the predictive density. For an early example of its use in Bayesian electricity forecasting see Panagiotelis and Smith (2008); for later examples, see Bracale and De Falco (2015), Brusaferri *et al.* (2019) and Gianfreda *et al.* (2020). Other scoring rules are less commonly used in the Bayesian electricity forecasting literature although Ohtsuka *et al.* (2010), where the log score is used, is a notable exception.

[17] The same point does not apply when combining ensembles from NWPs since the forecasting models are deterministic.
[18] We note that in some cases, this is challenging; for example for long-run forecasts as in Da Silva *et al.* (2019).

5 In Summary

Bayesian forecasting is underpinned by a single core principle: uncertainty about the future value of a random variable is expressed using a probability distribution that is, ultimately, conditioned *only* on observable, or known, information about the random variable. Nothing, surely, could be more natural as a way of framing the forecasting exercise, and the Bayesian approach to forecasting is – arguably – one of the most compelling features of the paradigm.[19]

In large measure, the challenge has, potentially, been in the *implementation* of Bayesian forecasting: computing the expectation that defines the predictive distribution; most particularly when accessing (draws from) the posterior itself is difficult. And as models have become larger and more challenging, and as the data sets have grown 'bigger', this problem of accessing the exact posterior has only increased. However, as this review has demonstrated, the expansion of the forecasting problems being tackled has gone hand-in-hand with the development of new and improved computational methods designed expressly to access challenging posteriors, and in a reasonable computing time. Notably, when it comes to accurate forecasting, somewhat crude *approximations* of the posterior have been found to still yield accurate predictions; meaning that Bayesian forecasting remains viable for large and complex models for which approximate computation of posteriors is the only feasible approach.

The more fundamental problem of model mis-specification can also be managed, by moving away from the conventional likelihood-based Bayesian up-dating and allowing forecast accuracy itself – and its link to the future decisions that depend on that accuracy – to drive the up-dating. This, in turn, ensures that forecasts are 'fit for purpose', *despite* the inevitable mis-specification of the forecasting model. Allied with the computational power that now drives the Bayesian engine, this ability to generalize the paradigm beyond its traditional links with the likelihood principle is a potent, if not yet fully realized, force in forecasting.

References

Aastveit, K. A., Cross, J. L., and van Dijk, H. K. (2022). Quantifying time-varying forecast uncertainty and risk for the real price of oil. *Journal of Business & Economic Statistics*, 0(0):1–15.

Aastveit, K. A., Ravazzolo, F., and van Dijk, H. K. (2018). Combined density nowcasting in an uncertain economic environment. *Journal of Business & Economic Statistics*, 36(1):131–145.

Adedipe, T., Shafiee, M., and Zio, E. (2020). Bayesian network modelling for the wind energy industry: An overview. *Reliability Engineering & System Safety*, 202:107053.

Adolfson, M., Lindé, J., and Villani, M. (2007). Forecasting performance of an open economy DSGE model. *Econometric Reviews*, 26(2-4):289–328.

Adrian, T., Boyarchenko, N., and Giannone, D. (2021). Multimodality in macrofinancial dynamics. *International Economic Review*, 62(2):861–886.

[19] As seen in Section 4.3 – in which a non-temporal prediction problem was the focus – exactly the same coherent approach to what is 'known' and what is 'unknown' obtains in that setting.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679.

Alexopoulos, A., Dellaportas, P., and Papaspiliopoulos, O. (2022). Bayesian prediction of jumps in large panels of time series data. *Bayesian Analysis*, 17(2):651–683.

Alipour, P., Mukherjee, S., and Nateghi, R. (2019). Assessing climate sensitivity of peak electricity load for resilient power systems planning and operation: A study applied to the Texas region. *Energy*, 185:1143–1153.

Andrieu, C., Doucet, A., and Holenstein, R. (2011). Particle Markov chain Monte Carlo. *J. Royal Statist. Society Series B*, 72(2):269–342. With discussion.

Andrieu, C., Doucet, A., and Robert, C. (2004). Computational advances for and from Bayesian analysis. *Statist. Science*, 19(1):118–127.

Andrieu, C. and Roberts, G. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697–725.

Ansari, A., Li, Y., and Zhang, J. Z. (2018). Probabilistic topic model for hybrid recommender systems: A stochastic variational Bayesian approach. *Marketing Science*, 37(6):987–1008.

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de Pison, F. J., and Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. *Solar Energy*, 136:78–111.

Araya, S., Elberg, A., Noton, C., and Schwartz, D. (2022). Identifying food labeling effects on consumer behavior. *Marketing Science*.

Ardia, D., Baştürk, N., Hoogerheide, L., and van Dijk, H. K. (2012). A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood. *Computational Statistics and Data Analysis*, 56(11):3398–3414.

Ausín, M. C., Galeano, P., and Ghosh, P. (2014). A semiparametric Bayesian approach to the analysis of financial time series with applications to value at risk estimation. *European Journal of Operational Research*, 232(2):350–358.

Baker, J., Fearnhead, P., Fox, E., and Nemeth, C. (2019). Control variates for stochastic gradient MCMC. *Statist. Comp.*, 29:599–615.

Bakshi, G., Cao, C., and Chen, Z. (1997). Empirical performance of alternative option pricing models. *The Journal of finance*, 52(5):2003–2049.

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large Bayesian vector autoregressions. *Journal of Applied Econometrics*, 25(1):71–92.

Bardenet, R., Doucet, A., and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *J. Machine Learning Res.*, 18(1):1515–1557.

Bassetti, F., Casarin, R., and Ravazzolo, F. (2018). Bayesian nonparametric calibration and combination of predictive distributions. *J. American Statist. Assoc.*, 113(522):675–685.

Bastani, H., Simchi-Levi, D., and Zhu, R. (2022). Meta dynamic pricing: Transfer learning across experiments. *Management Science*, 68(3):1865–1881.

Baştürk, N., Borowska, A., Grassi, S., Hoogerheide, L., and van Dijk, H. K. (2019). Forecast density combinations of dynamic models and data driven portfolio strategies. *Journal of Econometrics*, 210(1):170–186.

Bauwens, L. and Lubrano, M. (1998). Bayesian inference on GARCH models using the Gibbs sampler. *The Econometrics Journal*, 1(1):23–46.

Baştürk, N., Borowska, A., Grassi, S., Hoogerheide, L., and van Dijk, H. (2019). Forecast density combinations of dynamic models and data driven portfolio strategies. *Journal of Econometrics*, 210(1):170–186.

Beaumont, M. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164:1139–1160.

Beck, R. and Solow, J. L. (1994). Forecasting nuclear power supply with Bayesian autoregression. *Energy Economics*, 16(3):185–192.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly journal of economics*, 120(1):387–422.

Bernardi, M., Gayraud, G., and Petrella, L. (2015). Bayesian tail risk interdependence using quantile regression. *Bayesian Analysis*, 10(3):553–603.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. John Wiley, New York.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. John Wiley & Sons.

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. *https://arxiv.org/abs/1701.02434v2*.

Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, 177(2):213–232.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *J. Royal Statist. Society Series B*, 78(5):1103–1130.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economies*, 81:637–659.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *J. American Statist. Assoc.*, 112(518):859–877.

Bollerslev, T., Chou, R., and Kroner, K. (1992). ARCH modeling in finance. a review of the theory and empirical evidence. *J. Econometrics*, 52:5–59.

Botev, Z. I. (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):125–148.

Bracale, A. and De Falco, P. (2015). An advanced Bayesian method for short-term probabilistic forecasting of the generation of wind power. *Energies*, 8(9):10293–10314.

Braun, M. and McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice. *J. American Statist. Assoc.*, 105(489):324–335.

Brusaferri, A., Matteucci, M., Portolani, P., and Vitali, A. (2019). A Bayesian deep learning-based method for probabilistic forecast of day-ahead electricity prices. *Applied Energy*, 250:1158–1175.

Bunn, D. W. (1980). Experimental study of a Bayesian method for daily electricity load forecasting. *Applied Mathematical Modelling*, 4(2):113–116.

Burgette, L. F. and Nordheim, E. V. (2012). The trace restriction: An alternative identification strategy for the Bayesian multinomial probit model. *Journal of Business & Economic Statistics*, 30(3):404–410.

Burgette, L. F., Puelz, D., and Hahn, P. R. (2021). A symmetric prior for multinomial probit models. *Bayesian Analysis*, 1(1):1–18.

Calvet, L. E. and Czellar, V. (2015). Accurate methods for approximate Bayesian computation filtering. *J. Finan. Econometrics*, 13(4):798–838.

Canale, A. and Ruggiero, M. (2016). Bayesian nonparametric forecasting of monotonic functional time series. *Electronic Journal of Statistics*, 10(2):3265–3286.

Capone, A., Helminger, C., and Hirche, S. (2020). Day-ahead scheduling of thermal storage systems using Bayesian neural networks. *IFAC-PapersOnLine*, 53(2):13281–13286.

Carriero, A., Clark, T. E., and Marcellino, M. (2015). Bayesian VARs: specification choices and forecast accuracy. *Journal of Applied Econometrics*, 30(1):46–73.

Carriero, A., Clark, T. E., and Marcellino, M. (2016). Common drifting volatility in large Bayesian VARs. *Journal of Business & Economic Statistics*, 34(3):375–390.

Carriero, A., Clark, T. E., and Marcellino, M. (2019). Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *Journal of Econometrics*, 212(1):137–154.

Carriero, A., Kapetanios, G., and Marcellino, M. (2009). Forecasting exchange rates with a large Bayesian VAR. *International Journal of Forecasting*, 25(2):400–417.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Carverhill, A. and Luo, D. (2022). A Bayesian analysis of time-varying jump risk in S&P 500 returns and options. *Journal of Financial Markets*, page 100786.

Casarin, R., Grassi, S., Ravazzollo, F., and van Dijk, H. (2019). Forecast density combinations with dynamic learning for large data sets in economics and finance. *Tinbergen Institue Discussion Paper 2019-025/III*.

Casarin, R., Grassi, S., Ravazzolo, F., and van Dijk, H. (2015a). Parallel sequential Monte Carlo for efficient density combination: The deco MATLAB toolbox. *Journal of Statistical Software, Articles*, 68(3):1–30.

Casarin, R., Leisen, F., Molina, G., and ter Horst, E. (2015b). A Bayesian beta Markov random field calibration of the term structure of implied risk neutral densities. *Bayesian Analysis*, 10(4):791–819.

Casarin, R., Mantoan, G., and Ravazzolo, F. (2016). Bayesian calibration of generalized pools of predictive distributions. *Econometrics*, 4(1):1–24.

Casella, G. and George, E. (1992). An introduction to Gibbs sampling. *American Statist.*, 46:167–174.

Ceruzzi, P. (2003). *A History of Modern Computing*. MIT Press, second edition.

Chakraborty, A., Nott, D. J., Drovandi, C., Frazier, D. T., and Sisson, S. A. (2022). Modularized Bayesian analyses and cutting feedback in likelihood-free inference. *arXiv preprint arXiv:2203.09782*.

Chan, J. (2022). Asymmetric conjugate priors for large Bayesian VARs. *Quantitative Economics*, 13:1145–1169.

Chan, J. C. (2021). Minnesota-type adaptive hierarchical priors for large Bayesian VARs. *International Journal of Forecasting*, 37(3):1212–1226.

Chan, J. C., Eisenstat, E., and Strachan, R. W. (2020). Reducing the state space dimension in a large TVP-VAR. *Journal of Econometrics*, 218(1):105–118.

Chan, J. C., Koop, G., and Potter, S. M. (2013). A new model of trend inflation. *Journal of Business & Economic Statistics*, 31(1):94–106.

Chan, J. C., Koop, G., and Yu, X. (2021). Large order-invariant Bayesian VARs with stochastic volatility. *arXiv preprint arXiv:2111.07225*.

Chan, J. C. and Yu, X. (2022). Fast and accurate variational inference for large Bayesian VARs with stochastic volatility. *Journal of Economic Dynamics and Control*, 143:104505.

Chan, J. S., Choy, S. B., and Lam, C. P. (2014). Modeling electricity price using a threshold conditional autoregressive geometric process jump model. *Communications in Statistics-Theory and Methods*, 43(10-12):2505–2515.

Chib, S. (1993). Bayes regression with autoregressive errors: A Gibbs sampling approach. *Journal of Econometrics*, 58(3):275–294.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, 75(1):79–97.

Chib, S. and Greenberg, E. (1994). Bayes inference for regression models with ARMA(p,q) errors. *J. Econometrics*, 64:183–206.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *American Statist.*, 49:327–335.

Chib, S. and Greenberg, E. (1996). Markov chain Monte Carlo simulation methods in econometrics. *Econometric Theory*, 12(3):409–431.

Chib, S., Nadari, F., and Shephard, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *J. Econometrics*, 108:281–316.

Chib, S., Nardari, F., and Shephard, N. (2006). Analysis of high dimensional multivariate stochastic volatility models. *Journal of Econometrics*, 134(2):341–371.

Chib, S., Omori, Y., and Asai, M. (2009). Multivariate stochastic volatility. In *Handbook of financial time series*, pages 365–400. Springer.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Chiu, C.-W. J., Mumtaz, H., and Pinter, G. (2017). Forecasting with VAR models: Fat tails and stochastic volatility. *International Journal of Forecasting*, 33(4):1124–1143.

Chopin, N. and Papaspiliopoulos (2020). *An introduction to sequential Monte Carlo*. Springer.

Clark, T., Huber, F., Koop, G., and Mercellino, M. (2022a). Tail forecasting with multivariate Bayesian additive regression trees. *Federal Reserve Bank of Cleveland Working Paper*, 21-08R.

Clark, T. E. (2011). Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. *Journal of Business & Economic Statistics*, 29(3):327–341.

Clark, T. E., Huber, F., Koop, G., and Marcellino, M. (2022b). Forecasting US inflation using Bayesian nonparametric models. *arXiv preprint arXiv:2202.13793*.

Clark, T. E. and Ravazzolo, F. (2015). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *Journal of Applied Econometrics*, 30(4):551–575.

Coelho, C., Stephenson, D., Doblas-Reyes, F., Balmaseda, M., Guetter, A., and Van Oldenborgh, G. (2006). A Bayesian approach for multi-model downscaling: Seasonal forecasting of regional rainfall and river flows in South America. *Meteorological Applications*, 13(1):73–82.

Cottet, R. and Smith, M. (2003). Bayesian modeling and forecasting of intraday electricity load. *Journal of the American Statistical Association*, 98(464):839–849.

Craiu, R. V. and Meng, X.-L. (2005). Multiprocess parallel antithetic coupling for backward and forward Markov chain Monte Carlo. *Ann. Statist.*, 33(2):661–697.

Creel, M. and Kristensen, D. (2015). ABC of SV: Limited information likelihood inference in stochastic volatility jump-diffusion models. *Journal of Empirical Finance*, 31:85–108.

Cross, J. L., Hou, C., and Poon, A. (2020). Macroeconomic forecasting with large Bayesian VARs: Global-local priors and the illusion of sparsity. *International Journal of Forecasting*, 36(3):899–915.

Da Silva, F. L., Oliveira, F. L. C., and Souza, R. C. (2019). A bottom-up Bayesian extension for long term electricity consumption forecasting. *Energy*, 167:198–210.

D'Agostino, A., Gambetti, L., and Giannone, D. (2013). Macroeconomic forecasting and structural change. *Journal of Applied Econometrics*, 28(1):82–101.

Danaher, P. J., Danaher, T. S., Smith, M. S., and Loaiza-Maya, R. (2020). Advertising effectiveness for multiple retailer-brands in a multimedia and multichannel environment. *Journal of Marketing Research*, 57(3):445–467.

Davis, P. and Rabinowitz, P. (1975). *Numerical Methods of Integration*. Academic Press, New York.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

Dawid, A. P. (1985). Calibration-based empirical probability. *The Annals of Statistics*, 13(4):1251–1274.

Del Negro, M., Hasegawa, R. B., and Schorfheide, F. (2016). Dynamic prediction pools: An investigation of financial frictions and forecasting performance. *Journal of Econometrics*, 192(2):391–405. Innovations in Multiple Time Series Analysis.

Delatola, E.-I. and Griffin, J. E. (2013). A Bayesian semiparametric model for volatility with a leverage effect. *Computational Statistics & Data Analysis*, 60:97–110.

Deligiannidis, G., Doucet, A., and Pitt, M. K. (2018). The correlated pseudomarginal method. *J. Royal Statist. Society Series B*, 80(5):839–870.

Dieppe, A., van Roye, B., and Legrand, R. (2016). The BEAR toolbox. *European Central Bank Working Paper*, 1934.

Doan, T., Litterman, R., and Sims, C. (1984a). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3(1):1–100.

Doan, T., Litterman, R., and Sims, C. (1984b). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3(1):1–100.

Dongarra, J. and Sullivan, F. (2000). Guest editors' introduction: The top 10 algorithms. *Computing in Science & Engineering*, 2(1):22–23.

Douc, R. and Robert, C. P. (2011). A vanilla Rao–Blackwellization of Metropolis–Hastings algorithms. *Ann. Statist.*, 39(1):261–277.

Doucet, A., Pitt, M. K., Deligiannidis, G., and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313.

Du, P. (2018). Ensemble machine learning-based wind forecasting to combine NWP output with data from weather station. *IEEE Transactions on Sustainable Energy*, 10(4):2133–2141.

Dufays, A. (2016). Infinite-state Markov-switching for dynamic volatility. *Journal of Financial Econometrics*, 14(2):418–460.

Dunson, D. and Johndrow, J. (2019). The Hastings algorithm at fifty. *Biometrika*, 107(1):1–23.

Eraker, B. (2001). MCMC analysis of diffusion models with application to finance. *Journal of Business & Economic Statistics*, 19(2):177–191.

Eraker, B. (2004). Do stock prices and volatility jump? reconciling evidence from spot and option prices. *The Journal of finance*, 59(3):1367–1403.

Eraker, B., Johannes, M., and Polson, N. (2003). The impact of jumps in volatility and returns. *The Journal of Finance*, 58(3):1269–1300.

Fasano, A. and Durante, D. (2022). A class of conjugate priors for multinomial probit models which includes the multivariate normal one. *Journal of Machine Learning Research*, 23(30):1–26.

Fearnhead, P. (2011). MCMC for state-space models. *Handbook of Markov Chain Monte Carlo*, pages 513–529. Chapman & Hall/CRC. Eds. Brooks, S., Gelman, A., Jones, G., Meng, X-L.

Fileccia, G. and Sgarra, C. (2018). A particle filtering approach to oil futures price calibration and forecasting. *Journal of Commodity Markets*, 9:21–34.

Flury, T. and Shephard, N. (2011). Bayesian inference based only on simulated likelihood: Particle filter analysis of dynamic economic models. *Econometric Theory*, 27(5):933–956.

Forbes, C. S., Martin, G. M., and Wright, J. (2007). Inference for a class of stochastic volatility models using option and spot prices: Application of a bivariate Kalman filter. *Econometric Reviews*, 26(2-4):387–418.

Frazier, D. T., Loaiza-Maya, R., and Martin, G. M. (2022). Variational Bayes in state space models: Inferential and predictive accuracy. *arXiv preprint arXiv:2106.12262*. Forthcoming, Journal of Computational and Graphical Statistics.

Frazier, D. T., Loaiza-Maya, R., Martin, G. M., and Koo, B. (2021). Loss-based variational Bayes prediction. *arXiv preprint arXiv:2104.14054*.

Frazier, D. T., Maneesoonthorn, W., Martin, G. M., and McCabe, B. P. (2019). Approximate Bayesian forecasting. *Intern. J. Forecasting*, 35(2):521–539.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *J. Time Ser. Anal.*, 15(2):183–202.

Frühwirth-Schnatter, S. (2004). Efficient Bayesian parameter estimation. *State space and unobserved component models: Theory and applications*, pages 123–151. CUP. Eds. Harvey, J., Koopman, S. and Shephard, N.

Frühwirth-Schnatter, S. and Wagner, H. (2010). Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics*, 154(1):85–100.

Fulop, A. and Li, J. (2019). Bayesian estimation of dynamic asset pricing models with informative observations. *Journal of Econometrics*, 209(1):114–138.

Gallant, A. R. and Tauchen, G. (1996). Which moments to match? *Econometric theory*, 12(4):657–681.

Gefang, D., Koop, G., and Poon, A. (2022). Forecasting using variational Bayesian inference in large vector autoregressions with hierarchical shrinkage. *International Journal of Forecasting*.

Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, 85(410):398–409.

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

George, E. I., Sun, D., and Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics*, 142(1):553–580.

Geraci, M. V. and Gnabo, J.-Y. (2018). Measuring interconnectedness between financial institutions with Bayesian time-varying vector autoregressions. *Journal of Financial and Quantitative Analysis*, 53(3):1371–1390.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1340.

Geweke, J. (2005). *Contemporary Bayesian econometrics and statistics*, volume 537. John Wiley & Sons.

Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, 164(1):130–141.

Geweke, J. and Whiteman, C. (2006). Bayesian forecasting. *Handbook of economic forecasting*, 1:3–80.

Geyer, C. J. (2011). Introduction to Markov chain Monte Carlo. *Handbook of Markov chain Monte Carlo*, pages 3–48. Chapman & Hall/CRC. Eds. Brooks, S., Gelman, A., Jones, G., Meng, X-L.

Ghayekhloo, M., Azimi, R., Ghofrani, M., Menhaj, M., and Shekari, E. (2019). A combination approach based on a novel data clustering method and Bayesian recurrent neural network for day-ahead price forecasting of electricity markets. *Electric Power Systems Research*, 168:184–199.

Gianfreda, A., Ravazzolo, F., and Rossini, L. (2020). Comparing the forecasting performances of linear models for electricity prices with high RES penetration. *International Journal of Forecasting*, 36(3):974–986.

Giannone, D., Lenza, M., and Primiceri, G. E. (2015). Prior selection for vector autoregressions. *The Review of Economics and Statistics*, 97(2):436–451.

Giebel, G. and Kariniotakis, G. (2017). Wind power forecasting—a review of the state of the art. *Renewable Energy Eorecasting*, pages 59–109.

Gilanifar, M., Wang, H., Ozguven, E. E., Zhou, Y., and Arghandeh, R. (2019). Bayesian spatiotemporal Gaussian process for short-term load forecasting using combined transportation and electricity data. *ACM Transactions on Cyber-Physical Systems*, 4(1):1–25.

Gilbride, T. J. and Allenby, G. M. (2004). A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Science*, 23(3):391–406.

Giordani, P., Pitt, M., and Kohn, R. (2011). Bayesian inference for time series state space models. *The Oxford Handbook of Bayesian Econometrics*, pages 61–124. OUP. Eds. Geweke, J., Koop, G. and van Dijk, H.

Girolami, M. and Rogers, S. (2006). Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, 18(8):1790–1817.

Giummolè, F., Mameli, V., Ruli, E., and Ventura, L. (2017). Objective Bayesian inference with proper scoring rules. *TEST*, 28(3):1–28.

Glynn, P. W. and Rhee, C.-H. (2014). Exact estimation for Markov chain equilibrium expectations. *J. Appl. Probab.*, 51(A):377–389.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electron. J. Statist.*, 7:1747–1782.

Gonzato, L. and Sgarra, C. (2021). Self-exciting jumps in the oil market: Bayesian estimation and dynamic hedging. *Energy Economics*, 99:105279.

Gordon, N., Salmond, J., and Smith, A. (1993). A novel approach to non-linear/non-Gaussian Bayesian state estimation. *IEEE Proceedings on Radar and Signal Processing*, 140(2):107–113.

Granger, C. W.*et al.* (1986). Forecasting accuracy of alternative techniques: a comparison of us macroeconomic forecasts: comment. *Journal of Business & Economic Statistics*, 4(1):16–17.

Green, P., Latuszynski, K., Pereyra, M., and Robert, C. (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statist. Comp.*, 25:835–862.

Griffin, J. E. and Steel, M. F. (2011). Stick-breaking autoregressive processes. *Journal of econometrics*, 162(2):383–396.

Grillone, B., Mor, G., Danov, S., Cipriano, J., Lazzari, F., and Sumper, A. (2021). Baseline energy use modeling and characterization in tertiary buildings using an interpretable Bayesian linear regression methodology. *Energies*, 14(17):5556.

Guedj, B. (2019). A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*.

Gunawan, D., Kohn, R., and Nott, D. (2021). Variational Bayes approximation of factor stochastic volatility models. *International Journal of Forecasting*, 37(4):1355–1375.

Gunel, I. (1987). Forecasting system energy demand. *Journal of Forecasting*, 6(2):137–156.

Hafner, C. M. and Herwartz, H. (2001). Option pricing under linear autoregressive dynamics, heteroskedasticity, and conditional leptokurtosis. *Journal of Empirical Finance*, 8(1):1–34.

Hall, S. G. and Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, 23(1):1–13.

Hammersley, J. and Handscomb, D. (1964). *Monte Carlo Methods*. John Wiley, New York.

Harvey, A. (1981). *The Econometric Analysis of Time Series*. John Wiley.

Hassan, S., Khosravi, A., and Jaafar, J. (2015). Examining performance of aggregation algorithms for neural network-based electricity demand forecasting. *International Journal of Electrical Power & Energy Systems*, 64:1098–1105.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 57(1):97–109.

Hauzenberger, N., Huber, F., Koop, G., and Onorante, L. (2021a). Fast and flexible Bayesian inference in time-varying parameter regression models. *Journal of Business & Economic Statistics*, 0(0):1–15.

Hauzenberger, N., Huber, F., and Onorante, L. (2021b). Combining shrinkage and sparsity in conjugate vector autoregressive models. *Journal of Applied Econometrics*, 36(3):304–327.

Hippert, H. S. and Taylor, J. W. (2010). An evaluation of Bayesian techniques for controlling model complexity and selecting inputs in a neural network for short-term load forecasting. *Neural networks*, 23(3):386–395.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Holmes, C. and Walker, S. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503.

Huber, F. and Feldkircher, M. (2019). Adaptive shrinkage in Bayesian vector autoregressive models. *Journal of Business & Economic Statistics*, 37(1):27–39.

Huber, F., Koop, G., and Onorante, L. (2021). Inducing sparsity and shrinkage in time-varying parameter models. *Journal of Business & Economic Statistics*, 39(3):669–683.

Huber, F., Koop, G., Onorante, L., Pfarrhofer, M., and Schreiner, J. (2020). Nowcasting in a pandemic using non-parametric mixed frequency VARs. *Journal of Econometrics*.

Huber, F. and Pfarrhofer, M. (2021). Dynamic shrinkage in time-varying parameter stochastic volatility in mean models. *Journal of Applied Econometrics*, 36(2):262–270.

Huber, M. L. (2016). *Perfect simulation*. Chapman & Hall/CRC.

Jacob, P., Robert, C., and Smith, M. (2011). Using parallel computation to improve independent Metropolis–Hastings based estimation. *J. Comput. Graph. Statist.*, 20(3):616–635.

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *J. Royal Statist. Society Series B*, 82:1–32. With discussion.

Jacquier, E. and Polson, N. (2011). Bayesian methods in finance. *The Oxford Handbook of Bayesian Econometrics*, pages 439–512. OUP. Eds. Geweke, J., Koop, G. and van Dijk, H.

Jacquier, R., Polson, N. G., and Rossi, P. E. (1994). Bayesian analysis of stochastic volatility models. *J. Business and Economic Statistics*, 12(4):371–389. With discussion.

Jahan, F., Ullah, I., and Mengersen, K. (2020). A review of Bayesian statistical approaches for Big Data. In Mengersen, K., Pudlo, P., and Robert, C., editors, *Case Studies in Applied Bayesian Science*, pages 17–44. Springer.

Jensen, M. J. and Maheu, J. M. (2010). Bayesian semiparametric stochastic volatility modeling. *Journal of Econometrics*, 157(2):306–316.

Jensen, M. J. and Maheu, J. M. (2013). Bayesian semiparametric multivariate GARCH modeling. *Journal of Econometrics*, 176(1):3–17.

Jensen, M. J. and Maheu, J. M. (2014). Estimating a semiparametric asymmetric stochastic volatility model with a Dirichlet process mixture. *Journal of Econometrics*, 178:523–538.

Jin, X. and Maheu, J. M. (2013). Modeling realized covariances and returns. *Journal of Financial Econometrics*, 11(2):335–369.

Jin, X. and Maheu, J. M. (2016). Bayesian semiparametric modeling of realized covariance matrices. *Journal of Econometrics*, 192(1):19–39.

Jin, X., Maheu, J. M., and Yang, Q. (2019). Bayesian parametric and semiparametric factor models for large realized covariance matrices. *Journal of Applied Econometrics*, 34:641–660.

Jin, X., Maheu, J. M., and Yang, Q. (2022). Infinite Markov pooling of predictive distributions. *Journal of Econometrics*, 228(2):302–321.

Johannes, M. and Polson, N. (2010). Chapter 13 - MCMC methods for continuous-time financial econometrics. In Ait-Sahalia, Y. and Hansen, L. P., editors, *Handbook of Financial Econometrics: Applications*, volume 2 of *Handbooks in Finance*, pages 1–72. Elsevier, San Diego.

Johannes, M. S., Polson, N. G., and Stroud, J. R. (2009). Optimal filtering of jump diffusions: Extracting latent states from asset prices. *The Review of Financial Studies*, 22(7):2759–2799.

Johndrow, J. E., Smith, A., Pillai, N., and Dunson, D. B. (2019). MCMC for imbalanced categorical data. *J. American Statist. Assoc.*, 114(527):1394–1403.

Johnson, M. C. (2017). *Bayesian predictive synthesis: Forecast calibration and combination.* PhD thesis, Duke University.

Joutz, F. L., Maddala, G. S., and Trost, R. P. (1995). An integrated Bayesian vector auto regression and error correction model for forecasting electricity consumption and prices. *Journal of Forecasting*, 14(3):287–310.

Kabisa, S., Dunson, D. B., and Morris, J. S. (2016). Online variational Bayes inference for high-dimensional correlated data. *J. Comput. Graph. Statist.*, 25(2):426–444.

Kalli, M. and Griffin, J. (2015). Flexible modeling of dependence in volatility processes. *Journal of Business & Economic Statistics*, 33(1):102–113.

Kalli, M. and Griffin, J. E. (2018). Bayesian nonparametric vector autoregressive models. *Journal of econometrics*, 203(2):267–282.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and computing*, 21(1):93–105.

Karmakar, B., Kwon, O., Mukherjee, G., and Siddarth, S. (2021). Understanding early adoption of hybrid cars via a new multinomial probit model with multiple network weights. Technical report.

Kastner, G., Frühwirth-Schnatter, S., and Lopes, H. F. (2017). Efficient Bayesian inference for multivariate factor stochastic volatility models. *Journal of Computational and Graphical Statistics*, 26(4):905–917.

Kastner, G. and Huber, F. (2021). Sparse Bayesian vector autoregressions in huge dimensions. *Journal of Forecasting*, 39:1142–1165.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65(3):361–393.

Kloek, T. and van Dijk, H. K. (1978). Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica*, 46(1):1–19.

Kon Kam King, G., Canale, A., and Ruggiero, M. (2019). Bayesian functional forecasting with locally-autoregressive dependent processes. *Bayesian Anal.*, 14(4):1121–1141.

Koop, G. (2013a). Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics*, 28:177–203.

Koop, G. and Korobilis, D. (2010). Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends® in Econometrics*, 3(4):267–358.

Koop, G. and Korobilis, D. (2013). Large time-varying parameter VARs. *Journal of Econometrics*, 177(2):185–198.

Koop, G. and Korobilis, D. (2018). Variational Bayes inference in high-dimensional time-varying parameter models. *SSRN 3246472*.

Koop, G., McIntyre, S., Mitchell, J., and Poon, A. (2020). Regional output growth in the united kingdom: More timely and higher frequency estimates from 1970. *Journal of Applied Econometrics*, 35(2):176–197.

Koop, G. M. (2003). *Bayesian Econometrics*. John Wiley & Sons Inc.

Koop, G. M. (2013b). Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics*, 28(2):177–203.

Korobilis, D. (2013). VAR forecasting using Bayesian variable selection. *Journal of Applied Econometrics*, 28:204–230.

Kostrzewski, M. and Kostrzewska, J. (2019). Probabilistic electricity price forecasting with Bayesian stochastic volatility models. *Energy Economics*, 80:610–620.

Lahiri, K. and Gao, J. (2002). Bayesian analysis of nested logit model by Markov chain Monte Carlo. *Journal of econometrics*, 111(1):103–133.

Laplace, P. (1774). Mémoire sur la probabilité des causes par les événemens. *Mémoires de l'Académie Royale des Sciences présentés par divers savants*, 6:621–656.

Launay, T., Philippe, A., and Lamarche, S. (2015). Construction of an informative hierarchical prior for a small sample with the help of historical data and application to electricity load forecasting. *Test*, 24(2):361–385.

Lenza, M. and Primiceri, G. E. (2022). How to estimate a vector autoregression after March 2020. *Journal of Applied Econometrics*, 37(4):688–699.

Li, C. (2022). A multivariate GARCH model with an infinite hidden Markov mixture. MPRA Paper No. 112792.

Lim, G.-C., Martin, G. M., and Martin, V. L. (2005). Parametric pricing of higher order moments in S&P500 options. *Journal of Applied Econometrics*, 20(3):377–404.

Lindberg, K., Seljom, P., Madsen, H., Fischer, D., and Korpås, M. (2019). Long-term electricity load forecasting: Current and future trends. *Utilities Policy*, 58:102–119.

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2017). Fundamentals and recent developments in approximate Bayesian computation. *Systematic biology*, 66(1):e66–e82.

Liu, J. and Maheu, J. M. (2018). Improving Markov switching models using realized variance. *Journal of Applied Econometrics*, 33:297–318.

Liu, J., Wong, W., and Kong, A. (1994). Covariance structure of the Gibbs sampler with application to the comparison of estimators and augmentation schemes. *Biometrika*, 81:27–40.

Llorente, F., Martino, L., Delgado, D., and Lopez-Santiago, J. (2021). Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *https://arXiv:2005.08334*.

Loaiza-Maya, R., Martin, G. M., and Frazier, D. T. (2021). Focused Bayesian prediction. *Journal of Applied Econometrics*, 36(5):517–543.

Loaiza-Maya, R. and Nibbering, D. (2022a). Fast variational Bayes methods for multinomial probit models. *arXiv preprint arXiv:2202.12495*. Forthcoming, Journal of Business & Economic Statistics.

Loaiza-Maya, R. and Nibbering, D. (2022b). Scalable Bayesian estimation in the multinomial probit model. *Journal of Business & Economic Statistics*, 40(4):1678–1690.

Loaiza-Maya, R., Smith, M. S., Nott, D. J., and Danaher, P. J. (2022). Fast and accurate variational inference for models with many latent variables. *Journal of Econometrics*, 230(2):339–362.

Lyddon, S., Holmes, C., and Walker, S. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478.

Maneesoonthorn, W., Forbes, C. S., and Martin, G. M. (2017). Inference on self-exciting jumps in prices and volatility using high-frequency measures. *Journal of Applied Econometrics*, 32(3):504–532.

Maneesoonthorn, W., Martin, G. M., Forbes, C. S., and Grose, S. D. (2012). Probabilistic forecasts of volatility and its risk premia. *Journal of Econometrics*, 171(2):217–236.

Marin, J., Pudlo, P., Robert, C., and Ryder, R. (2011). Approximate Bayesian computational methods. *Statist. Comp.*, 21(2):279–291.

Martin, G. M., Frazier, D. T., and Robert, C. P. (2022a). Approximating Bayes in the 21st century. *https://arXiv:2112.10342*. Forthcoming, Statistical Science.

Martin, G. M., Frazier, D. T., and Robert, C. P. (2022b). Computing Bayes: From then 'til now. *https://arxiv.org/abs/2208.00646*. Forthcoming, Statistical Science.

Martin, G. M., Loaiza-Maya, R., Maneesoonthorn, W., Frazier, D. T., and Ramírez-Hassan, A. (2022c). Optimal probabilistic forecasts: When do they work? *International Journal of Forecasting*, 38(1):384–406.

Martin, G. M., McCabe, B. P., Frazier, D. T., Maneesoonthorn, W., and Robert, C. P. (2019). Auxiliary likelihood-based approximate Bayesian computation in state space models. *J. Comput. Graph. Statist.*, 28(3):508–522.

Martino, S. and Riebler, A. (2019). Integrated nested Laplace approximations (INLA). *https://arXiv:1907.01248*.

McAlinn, K., Aastveit, K. A., Nakajima, J., and West, M. (2020). Multivariate Bayesian predictive synthesis in macroeconomic forecasting. *Journal of the American Statistical Association*, 115(531):1092–1110.

McAlinn, K. and West, M. (2019). Dynamic Bayesian predictive synthesis in time series forecasting. *Journal of econometrics*, 210(1):155–169.

McCracken, M., Owyang, M., and Sekhposyan, T. (2021). Real-time forecasting and scenario analysis with a large mixed-frequency Bayesian VAR. *International Journal of Central Banking*, 18(5):327–367.

McCulloch, R. and Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1-2):207–240.

McCulloch, R. E., Polson, N. G., and Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99(1):173–193.

McCulloch, R. E. and Tsay, R. S. (1994). Bayesian analysis of autoregressive time series via the Gibbs sampler. *Journal of Time Series Analysis*, 15(2):235–250.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica: Journal of the Econometric Society*, pages 995–1026.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *J. American Statist. Assoc.*, 44:335–341.

Mishkin, E. (2021). Gender and sibling dynamics in the intergenerational transmission of entrepreneurship. *Management Science*, 67(10):6116–6135.

Mittal, V., Han, K., Lee, J.-Y., and Sridhar, S. (2021). Improving business-to-business customer satisfaction programs: Assessment of asymmetry, heterogeneity, and financial impact. *Journal of Marketing Research*, 58(4):615–643.

Miyazaki, K., Hoshino, T., and Böckenholt, U. (2021). Dynamic two stage modeling for category-level and brand-level purchases using potential outcome approach with Bayes inference. *Journal of Business & Economic Statistics*, 39(3):622–635.

Mori, H. and Nakano, K. (2014). Application of Gaussian process to locational marginal pricing forecasting. *Procedia Computer Science*, 36:220–226.

Müller, G. and Uhl, S. (2021). Estimation of time-varying autoregressive stochastic volatility models with stable innovations. *Statistics and Computing*, 31(3):1–19.

Naesseth, C. A., Lindsten, F., Schön, T. B.,*et al.* (2019). Elements of sequential Monte Carlo. *Foundations and Trends in Machine Learning*, 12(3):307–392.

Nakajima, J. (2017). Bayesian analysis of multivariate stochastic volatility with skew return distribution. *Econometric Reviews*, 36(5):546–562.

Nateghi, R., Guikema, S. D., and Quiring, S. M. (2011). Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes. *Risk Analysis: An International Journal*, 31(12):1897–1906.

Naylor, J. and Smith, A. (1982). Application of a method for the efficient computation of posterior distributions. *Applied Statistics*, 31(3):214–225.

Neal, R. (2011a). MCMC using ensembles of states for problems with fast and slow variables such as Gaussian process regression. *https://arXiv:1101.0387*.

Neal, R. (2011b). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman & Hall/CRC. Eds. Brooks, S., Gelman, A., Jones, G., Meng, X-L.

Neiswanger, W., Wang, C., and Xing, E. (2013). Asymptotically exact, embarrassingly parallel MCMC. *https://arXiv:1311.4780*.

Nott, D. J. and Kohn, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika*, 92(4):747–763.

Nowotarski, J., Raviv, E., Trück, S., and Weron, R. (2014). An empirical comparison of alternative schemes for combining electricity spot price forecasts. *Energy Economics*, 46:395–412.

Nowotarski, J. and Weron, R. (2018). Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81:1548–1568.

Ohtsuka, Y., Oga, T., and Kakamu, K. (2010). Forecasting electricity demand in Japan: A Bayesian spatial autoregressive ARMA approach. *Computational Statistics & Data Analysis*, 54(11):2721–2735.

Omori, Y., Chib, S., Shephard, N., and Nakajima, J. (2007). Stochastic volatility with leverage: Fast and efficient likelihood inference. *Journal of Econometrics*, 140(2):425–449.

Opschoor, A., Van Dijk, D., and van der Wel, M. (2017). Combining density forecasts using focused scoring rules. *Journal of Applied Econometrics*, 32(7):1298–1313.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *American Statist.*, 64(2):140–153.

Owen, A. B. (2017). Statistically efficient thinning of a Markov chain sampler. *J. Comput. Graph. Statist.*, 26(3):738–744.

Paleti, R. (2018). Generalized multinomial probit model: Accommodating constrained random parameters. *Transportation Research Part B: Methodological*, 118:248–262.

Panagiotelis, A. and Smith, M. (2008). Bayesian density forecasting of intraday electricity prices using multivariate skew t distributions. *International Journal of Forecasting*, 24(4):710–727.

Pesonen, H., Simola, U., Köhn-Luque, A., Vuollekoski, H., Lai, X., Frigessi, A., Kaski, S., Frazier, D. T., Maneesoonthorn, W., Martin, G. M., and Corander, J. (2022). ABC of the future. *International Statistical Review*.

Peters, G. W., Sisson, S. A., and Fan, Y. (2012). Likelihood-free Bayesian inference for α-stable models. *Comput. Statist. Data Anal.*, 56(11):3743–3756.

Pettenuzzo, D. and Ravazzolo, F. (2016). Optimal portfolio choice under decision-based model combinations. *Journal of Applied Econometrics*, 31(7):1312–1332.

Pezzulli, S., Frederic, P., Majithia, S., Sabbagh, S., Black, E., Sutton, R., and Stephenson, D. (2006). The seasonal forecast of electricity demand: A hierarchical Bayesian model with climatological weather generator. *Applied Stochastic Models in Business and Industry*, 22(2):113–125.

Pitt, M. K., dos Santos Silva, R., Giordani, P., and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *J. Econometrics*, 171(2):134–151.

Poirier, D. J. (1996). A Bayesian analysis of nested logit models. *Journal of Econometrics*, 75(1):163–181.

Polson, N. G., Carlin, B. P., and Stoffer, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *J. American Statist. Assoc.*, 87(418):493–500.

Posch, K., Truden, C., Hungerländer, P., and Pilz, J. (2022). A Bayesian approach for predicting food and beverage sales in staff canteens and restaurants. *International Journal of Forecasting*, 38(1):321–338.

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2018). Bayesian synthetic likelihood. *J. Comput. Graph. Statist.*, 27(1):1–11.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72(3):821–852.

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2019). Speeding up MCMC by efficient data subsampling. *J. American Statist. Assoc.*, 114(526):831–843.

Quiroz, M., Nott, D. J., and Kohn, R. (2022). Gaussian variational approximation for high-dimensional state space models. *https://arXiv:1801.07873*. Forthcoming, Bayesian Analysis.

Quiroz, M., Tran, M.-N., Villani, M., and Kohn, R. (2018). Speeding up MCMC by delayed acceptance and data subsampling. *J. Comput. Graph. Statist.*, 27(1):12–22.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly weather review*, 133(5):1155–1174.

Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):71–91.

Raviv, E., Bouwman, K. E., and Van Dijk, D. (2015). Forecasting day-ahead electricity prices: Utilizing hourly prices. *Energy Economics*, 50:227–239.

Raza, M. Q., Nadarajah, M., and Ekanayake, C. (2017). Demand forecast of PV integrated bioclimatic buildings using ensemble framework. *Applied Energy*, 208:1626–1638.

Ritter, C. and Tanner, M. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *J. American Statist. Assoc.*, 87(419):861–868.

Robert, C. (2007). *The Bayesian Choice*. Springer-Verlag, New York.

Robert, C. and Casella, G. (2011). A history of Markov chain Monte Carlo—subjective recollections from incomplete data. *Statist. Science*, 26(1):102–115.

Robert, C. P., Elvira, V., Tawn, N., and Wu, C. (2018). Accelerating MCMC algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(5):e1435.

Roberts, G. and Sahu, S. (1997). Updating schemes, covariance structure, blocking and parametrisation for the Gibbs sampler. *J. Royal Statist. Society Series B*, 59(2):291–317.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Statist.*, 18(2):349–367.

Rossi, P., Allenby, G., and McCulloch, R. (2012). *Bayesian Statistics and Marketing*. Wiley Series in Probability and Statistics. Wiley.

Rossi, P. E. and Allenby, G. M. (2003). Bayesian statistics and marketing. *Marketing Science*, 22(3):304–328.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *J. Royal Statist. Society Series B*, 71(2):319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Application*, 4(1):395–421.

Schorfheide, F. and Song, D. (2015). Real-time forecasting with a mixed-frequency VAR. *Journal of Business & Economic Statistics*, 33(3):366–380.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, pages 639–650.

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

Shi, S. and Song, Y. (2016). Identifying speculative bubbles using an infinite hidden Markov model. *Journal of Financial Econometrics*, 14(1):159–184.

Sisson, S. and Fan, Y. (2011). Likelihood-free Markov chain Monte Carlo. *Handbook of Markov Chain Monte Carlo*, pages 313–333. Chapman & Hall/CRC. Eds. Brooks, S., Gelman, A., Jones, G., Meng, X-L.

Sisson, S. A., Fan, Y., and Beaumont, M. (2019). *Handbook of Approximate Bayesian Computation*. Chapman & Hall/CRC.

Sloughter, J. M., Gneiting, T., and Raftery, A. E. (2010). Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the american statistical association*, 105(489):25–35.

Smets, F. and Wouters, R. (2007). Shocks and frictions in US business cycles: A Bayesian DSGE approach. *American Economic Review*, 97(3):586–606.

Smith, M. (2000). Modeling and short-term forecasting of New South Wales electricity system load. *Journal of Business & Economic Statistics*, 18(4):465–478.

Smith, M. S. (2010). Bayesian inference for a periodic stochastic volatility model of intraday electricity prices. In *Statistical Modelling and Regression Structures*, pages 353–376. Springer.

Stock, J. H. and Watson, M. (2007). Why has US inflation become harder to forecast? *Journal of Money, Credit and Banking*, 39:3–33.

Stock, J. H. and Watson, M. (2011). Dynamic factor models. *Oxford Handbooks Online*.

Stock, J. H. and Watson, M. W. (2016). Core inflation and trend inflation. *Review of Economics and Statistics*, 98(4):770–784.

Strickland, C. M., Forbes, C. S., and Martin, G. M. (2006). Bayesian analysis of the stochastic conditional duration model. *Computational Statistics and Data Analysis*, 50(9):2247–2267.

Strickland, C. M., Martin, G. M., and Forbes, C. S. (2008). Parameterisation and efficient MCMC estimation of non-Gaussian state space models. *Computational Statistics and Data Analysis*, 52(6):2911–2930.

Stroud, J. R., Müller, P., and Polson, N. G. (2003). Nonlinear state-space models with state-dependent variances. *Journal of the American Statistical Association*, 98(462):377–386.

Sun, M., Zhang, T., Wang, Y., Strbac, G., and Kang, C. (2019). Using Bayesian deep learning to capture uncertainty for residential net load forecasting. *IEEE Transactions on Power Systems*, 35(1):188–201.

Sun, P., Kim, I., and Lee, K. (2020). Flexible weighted Dirichlet process mixture modelling and evaluation to address the problem of forecasting return distribution. *Journal of Nonparametric Statistics*, 32(4):989–1014.

Syring, N. and Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486.

Tallman, E. and West, M. (2022). Bayesian predictive decision synthesis. *arXiv preprint arXiv:2206.03815*.

Tanner, M. A. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *J. American Statist. Assoc.*, 82(398):528–550. With discussion.

Tavaré, S., Balding, D., Griffith, R., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145:505–518.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Terui, N., Ban, M., and Allenby, G. M. (2011). The effect of media advertising on brand consideration and choice. *Marketing Science*, 30(1):74–91.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, 22:1701–1762.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *J. American Statist. Assoc.*, 81(393):82–86.

Tierney, L., Kass, R., and Kadane, J. (1989). Fully exponential Laplace approximations to expectations and variances of non-positive functions. *J. American Statist. Assoc.*, 84(407):710–716.

Toubia, O., Iyengar, G., Bunnell, R., and Lemaire, A. (2019). Extracting features of entertainment products: A guided latent dirichlet allocation approach informed by the psychology of media consumption. *Journal of Marketing Research*, 56(1):18–36.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Van Gael, J., Saatci, Y., Teh, Y. W., and Ghahramani, Z. (2008). Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1088–1095. ACM.

Vankov, E. R., Guindani, M., and Ensor, K. B. (2019). Filtering and estimation for a class of stochastic volatility models with intractable likelihoods. *Bayesian Analysis*, 14(1):29–52.

Virbickaitė, A., Ausín, M. C., and Galeano, P. (2020). Copula stochastic volatility in oil returns: Approximate Bayesian computation with volatility prediction. *Energy Economics*, 92:104961.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation®*, 36(1):45–54.

Wand, M. P. (2017). Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *J. American Statist. Assoc.*, 112(517):137–168.

Wang, S., Sun, X., and Lall, U. (2017). A hierarchical Bayesian regression model for predicting summer residential electricity demand across the USA. *Energy*, 140:601–611.

Wang, X., Hyndman, R. J., Li, F., and Kang, Y. (2022). Forecast combinations: an over 50-year review.

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30(4):1030–1081.

West, M. and Harrison, J. (2006). *Bayesian forecasting and dynamic models*. Springer Science & Business Media.

Wood, S. (2019). Simplified integrated nested Laplace approximation. *Biometrika*, 107(1):223–230.

Yang, Q. (2019). Stock returns and real growth: A Bayesian nonparametric approach. *Journal of Empirical Finance*, 53:53–69.

Yang, Y., Li, W., Gulliver, T. A., and Li, S. (2019). Bayesian deep learning-based probabilistic load forecasting in smart grids. *IEEE Transactions on Industrial Informatics*, 16(7):4703–4713.

Yu, C. L., Li, H., and Wells, M. T. (2011). MCMC estimation of Levy jump models using stock and option prices. *Mathematical Finance*, 21(3):383–422.

Zaharieva, M. D., Trede, M., and Wilfling, B. (2020). Bayesian semiparametric multivariate stochastic volatility with application. *Econometric Reviews*, 39(9):947–970.

Zamenjani, A. S. (2021). Do financial variables help predict the conditional distribution of the market portfolio? *Journal of Empirical Finance*, 62:327–345.

Zellner, A. (1971). *An Introduction to Bayesian Econometrics*. John Wiley, New York.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. (2018). Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026.

Zhou, X., Nakajima, J., and West, M. (2014). Bayesian forecasting and portfolio decisions using dynamic dependent sparse factor models. *International Journal of Forecasting*, 30(4):963–980.

A Further Computational Details

A.1 Gibbs sampling

Under the required regularity conditions (see Tierney, 1994) the Gibbs sampler yields a Markov chain with invariant distribution, $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$, via a transition kernel that is defined as the product of full conditional posteriors associated with the joint. For the case of $\boldsymbol{\theta}$ partitioned into B mutually exclusive blocks, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_b, ..., \boldsymbol{\theta}_B)'$, the steps of the Gibbs algorithm are given in Algorithm 1.

Algorithm 1 Gibbs Sampling Algorithm

Specify an initial value $\boldsymbol{\theta}^{(0)}$ and partition the parameter set into B mutually exclusive blocks.
for $i = 1, \ldots, M$ **do**
  for $b = 1, \ldots, B$ **do**
    Draw $\boldsymbol{\theta}_b^{(i)} \sim p_b(\boldsymbol{\theta}_b | \boldsymbol{\theta}_1^{(i)}, \ldots, \boldsymbol{\theta}_{b-1}^{(i)}, \boldsymbol{\theta}_{b+1}^{(i-1)}, \ldots, \boldsymbol{\theta}_B^{(i-1)}, |\mathbf{y}_{1:T})$
  end for
end for
Return a sample of draws from $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$.

A.2 MH-within-Gibbs sampling

In Algorithm 2 we provide the generic steps of the so-called 'MH-within-Gibbs' algorithm, for the case of $\boldsymbol{\theta}$ partitioned into B mutually exclusive blocks, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_b, ..., \boldsymbol{\theta}_B)'$. The symbol p_b^* represents (the ordinate of) a kernel of the corresponding conditional $p_b(\cdot|\cdot)$.

Algorithm 2 MH-within-Gibbs Algorithm

Specify an initial value $\boldsymbol{\theta}^{(0)}$, a partition of the parameter set into B mutually exclusive blocks, and a proposal distribution $q_b(\boldsymbol{\theta}_b|\mathbf{y}_{1:T})$ for $b \in \{1, \ldots, B\}$.
for $i = 1, \ldots, M$ **do**
  for $b = 1, \ldots, B$ **do**
    Draw $\boldsymbol{\theta}_b^c \sim q_b(\boldsymbol{\theta}_b|\mathbf{y}_{1:T})$
    Compute the Metropolis-Hastings ratio:

$$r = \frac{p_b^*(\boldsymbol{\theta}_b^c|\boldsymbol{\theta}_1^{(i)}, \ldots, \boldsymbol{\theta}_{b-1}^{(i)}, \boldsymbol{\theta}_{b+1}^{(i-1)}, \ldots, \boldsymbol{\theta}_B^{(i-1)}, \mathbf{y}_{1:T}) \times q_b(\boldsymbol{\theta}_b^{(i-1)}|\mathbf{y}_{1:T})}{p_b^*(\boldsymbol{\theta}_b^{(i-1)}|\boldsymbol{\theta}_1^{(i)}, \ldots, \boldsymbol{\theta}_{b-1}^{(i)}, \boldsymbol{\theta}_{b+1}^{(i-1)}, \ldots, \boldsymbol{\theta}_B^{(i-1)}, \mathbf{y}_{1:T}) \times q_b(\boldsymbol{\theta}_b^c|\mathbf{y}_{1:T})}$$

    if $\mathcal{U}(0,1) < r$ **then**
      Set $\boldsymbol{\theta}_b^{(i)} = \boldsymbol{\theta}^c$
    else
      Set $\boldsymbol{\theta}_b^{(i)} = \boldsymbol{\theta}_b^{(i-1)}$
    end if
  end for
end for
Return a sample of draws from $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$.

The b^{th} candidate density $q_b(\boldsymbol{\theta}_b|\mathbf{y}_{1:T})$ may be chosen to deliberately target the form of the b^{th} conditional density, $p_b(\boldsymbol{\theta}_b|\boldsymbol{\theta}_1^{(i)}, ..., \boldsymbol{\theta}_{b-1}^{(i)}, \boldsymbol{\theta}_{b+1}^{(i-1)}, ..., \boldsymbol{\theta}_B^{(i-1)}, \mathbf{y}_{1:T})$, in which case the algorithm may be referred to as a 'tailored' algorithm; otherwise $q_b(\boldsymbol{\theta}_b|\mathbf{y}_{1:T})$ may be chosen in a more automated fashion, such as in a random-walk MH algorithm. The references cited in the text provide all details.

A.3 MH-within-Gibbs sampling in state space models

The application of MH-within-Gibbs sampling within a state space setting is qualitatively the same as described in Algorithm 2, except that the joint set of unknowns is augmented to ($\boldsymbol{\theta}$, $\mathbf{z}_{1:T}$), and decisions about partitioning need to be made for both $\boldsymbol{\theta}|\mathbf{z}_{1:T}$ and $\mathbf{z}_{1:T}|\boldsymbol{\theta}$. Decisions about the 'blocking' of $\mathbf{z}_{1:T}$ are particularly important, given both the dimension of $\mathbf{z}_{1:T}$ and the time-series dependence embedded in the state process, as are matters of parameterizing the state space model. We refer the reader to:

Shephard and Pitt (1997) and Strickland *et al.* (2006) for illustrations of state blocking in which the block sizes are selected randomly; and to Frühwirth-Schnatter (2004) and Strickland *et al.* (2008) for explorations of the impact of parameterization on the performance of the sampler.

A.4 PMMH in state space models

Early Bayesian treatments of non-linear state space models often exploited a linear Gaussian approximation at some point, for the purpose of defining candidate densities for (blocks of) $\mathbf{z}_{1:T}$ (e.g., Kim *et al.*, 1998; Stroud *et al.*, 2003; Strickland *et al.*, 2006), thereby enabling a Kalman filter-based 'forward filtering, backward sampling' algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994) to be used to produce a candidate draw of (any particular block of) $\mathbf{z}_{1:T}$, conditional on $\boldsymbol{\theta}$. As noted in Section 3.2.2 (and in the review, Giordani *et al.*, 2011), more recent approaches to such models have exploited PMMH principles instead. Algorithm 3 reproduces the algorithm in Andrieu *et al.* (2011), Section 2.4.2, adapted slightly to match the notation of the current paper. To simplify the exposition, the algorithm is presented for sampling the full vector $\boldsymbol{\theta}$. In practice the algorithm would be modified to cater for any blocking of $\boldsymbol{\theta}$.

Algorithm 3 PMMH Algorithm

Step 1: Initialization, $i = 0$

(a) Set $\boldsymbol{\theta}^{(0)}$ arbitrarily and

(b) Run an SMC algorithm targeting $p(\mathbf{z}_{1:T}|\mathbf{y}_{1:T}, \boldsymbol{\theta}^{(0)})$, sample $\mathbf{z}_{1:T}^{(0)} \sim \hat{p}(\mathbf{z}_{1:T}|\mathbf{y}_{1:T}, \boldsymbol{\theta}^{(0)})$ and let $\hat{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}^{(0)})$ denote the marginal likelihood estimate.

Step 2:
for $i = 1, \ldots, M$ **do**
(a) Draw $\boldsymbol{\theta}^c \sim q(\boldsymbol{\theta}|\mathbf{y}_{1:T}, \boldsymbol{\theta}^{(i-1)})$,

(b) Run an SMC algorithm targeting $p(\mathbf{z}_{1:T}|\mathbf{y}_{1:T}, \boldsymbol{\theta}^c)$, sample $\mathbf{z}_{1:T}^c \sim \hat{p}(\mathbf{z}_{1:T}|\mathbf{y}_{1:T}, \boldsymbol{\theta}^c)$ and let $\hat{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}^c)$ denote the marginal likelihood estimate.

(c) Compute the Metropolis-Hastings ratio:

$$r = \frac{\hat{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}^c)p(\boldsymbol{\theta}^c) \times q(\boldsymbol{\theta}^{(i-1)}|\mathbf{y}_{1:T}, \boldsymbol{\theta}^c)}{\hat{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}^{(i-1)})p(\boldsymbol{\theta}^{(i-1)}) \times q(\boldsymbol{\theta}^c|\mathbf{y}_{1:T}, \boldsymbol{\theta}^{(i-1)})}$$

if $\mathcal{U}(0,1) < r$ **then**
Set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^c$, $\mathbf{z}_{1:T}^{(i)} = \mathbf{z}_{1:T}^c$, $\hat{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}^{(i)}) = \hat{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}^c)$
else
Set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$, $\mathbf{z}_{1:T}^{(i)} = \mathbf{z}_{1:T}^{(i-1)}$, $\hat{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}^{(i)}) = \hat{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}^{(i-1)})$
end if
end for
Return a sample of draws from $p(\boldsymbol{\theta}|\mathbf{y})$.

A.5 ABC based on summary statistics

The simplest (accept/reject) form of the ABC algorithm, as based on a chosen vector of summaries, $\eta(\mathbf{y}_{1:T})$, proceeds via the steps in Algorithm 4, with the accepted draws of $\boldsymbol{\theta}$ used to produce an estimate of $p(\boldsymbol{\theta}|\eta(\mathbf{y}_{1:T}))$, via kernel density methods. That is, ABC targets only the *partial posterior*, $p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}_{1:T}))$. This partial posterior is equivalent to $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ if and only if $\eta(\mathbf{y}_{1:T})$ is sufficient for conducting inference on $\boldsymbol{\theta}$, and for $\varepsilon \to 0$. Clearly, the very problems for which ABC is required imply that sufficient statistics are not available, and the requirement that $\varepsilon \to 0$ is infeasible in practice; so inference via ABC is only ever *intrinsically* approximate.[20]

Algorithm 4 Accept/Reject ABC Algorithm Based on Summary Statistics

for $i = 1, \ldots, M$ **do**
  Simulate $\boldsymbol{\theta}^{(i)}$, $i = 1, 2, ..., M$, from $p(\boldsymbol{\theta})$, and artificial data $\mathbf{z}_{1:T}^{(i)}$ from $p(\cdot|\boldsymbol{\theta}^{(i)})$;
  Accept $\boldsymbol{\theta}^{(i)}$ if $d\{\eta(\mathbf{z}_{1:T}^{(i)}), \eta(\mathbf{y}_{1:T})\} \leq \varepsilon$, where $d\{\cdot, \cdot\}$ denotes a generic metric and $\varepsilon > 0$ a pre-specified tolerance parameter.
end for

A.6 BSL based on summary statistics

BSL mimics ABC in targeting a posterior for $\boldsymbol{\theta}$ that conditions on a vector of summaries $\eta(\mathbf{y}_{1:T})$, rather than the full data set $\mathbf{y}_{1:T}$; however the summaries play a different role in the algorithm. Once again with reference to the simplest version of the algorithm, the steps of the BSL-MCMC algorithm are as given in Algorithm 5. Note that, for a given $\boldsymbol{\theta}$, the draws $\mathbf{z}_{(j)1:T} \sim i.i.d.\ p(\cdot|\boldsymbol{\theta})$, $j = 1, \ldots, m$, are used to estimate $\mu(\boldsymbol{\theta})$ and $\Sigma(\boldsymbol{\theta})$ as $\mu_m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^{m} \eta(\mathbf{z}_{(j)1:T})$ and $\Sigma_m(\boldsymbol{\theta}) = \frac{1}{m-1}\sum_{j=1}^{m} (\eta(\mathbf{z}_{(j)1:T}) - \mu_m(\boldsymbol{\theta}))(\eta(\mathbf{z}_{(j)1:T}) - \mu_m(\boldsymbol{\theta}))'$. The M draws of $\boldsymbol{\theta}$ are used to produce an estimate of $p(\boldsymbol{\theta}|\eta(\mathbf{y}_{1:T}))$ via kernel density methods.

Algorithm 5 BSL-MCMC Algorithm

for $i = 1, \ldots, M$ **do**
  Draw $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)})$
  Produce $\mu_m(\boldsymbol{\theta})$ and $\Sigma_m(\boldsymbol{\theta})$ using $j = 1, \ldots, m$ independent model simulations at $\boldsymbol{\theta}^*$
  Compute the synthetic likelihood $L^* = \mathcal{N}[\eta(\mathbf{y}); \mu_m(\boldsymbol{\theta}^*), \Sigma_m(\boldsymbol{\theta}^*)]$ and $L^{(i-1)}$ defined in a corresponding manner
  Compute the Metropolis-Hastings ratio:

$$r = \frac{L^*\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^*)}{L^{(i-1)}\pi(\boldsymbol{\theta}^{(i-1)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)})}$$

  if $\mathcal{U}(0,1) < r$ **then**
    Set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^*$, $\mu_m(\boldsymbol{\theta}^{(i)}) = \mu_m(\boldsymbol{\theta}^*)$ and $\Sigma_m(\boldsymbol{\theta}^{(i)}) = \Sigma_m(\boldsymbol{\theta}^*)$
  else
    Set $\boldsymbol{\theta}^i = \boldsymbol{\theta}^{(i-1)}$, $\mu_m(\boldsymbol{\theta}^{(i)}) = \mu_m(\boldsymbol{\theta}^{(i-1)})$ and $\Sigma_m(\boldsymbol{\theta}^{(i)}) = \Sigma_m(\boldsymbol{\theta}^{(i-1)})$
  end if
end for

[20] The notation $\mathbf{z}_{1:T}$ used in this section and in Section A.6 below is not to be confused with the use of $\mathbf{z}_{1:T}$ to denote a vector of latent variables elsewhere in the paper.

A.7 Variational Bayes (VB)

VB seeks the best approximation to $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ over a 'variational family' of densities $\mathcal{Q}$, with generic element $q(\boldsymbol{\theta})$. Typically this proceeds by minimizing the Kullback-Leibler (KL) divergence between $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$, which produces the variational approximation as

$$q^*(\boldsymbol{\theta}) := \underset{q(\boldsymbol{\theta})\in\mathcal{Q}}{\arg\min}\, \mathrm{KL}\left[q(\boldsymbol{\theta})|p(\boldsymbol{\theta}|\mathbf{y}_{1:T})\right], \tag{22}$$

where

$$\mathrm{KL}\left[q(\boldsymbol{\theta})|p(\boldsymbol{\theta}|\mathbf{y}_{1:T})\right] = \mathbb{E}_q[\log(q(\boldsymbol{\theta}))] - \mathbb{E}_q[\log(p(\boldsymbol{\theta},\mathbf{y}_{1:T}))] + \log(p(\mathbf{y}_{1:T})) \tag{23}$$

and $p(\boldsymbol{\theta},\mathbf{y}_{1:T}) = p(\mathbf{y}_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Given that the unknown normalizing constant $\log(p(\mathbf{y}_{1:T}))$ in (23) does not depend on q, the (infeasible) optimization problem in (22), is replaced by the equivalent (and feasible) optimization problem:

$$q^*(\boldsymbol{\theta}) := \underset{q(\boldsymbol{\theta})\in\mathcal{Q}}{\arg\max}\, \left\{\mathbb{E}_q[\log(p(\boldsymbol{\theta},\mathbf{y}_{1:T}))] - \mathbb{E}_q[\log(q(\boldsymbol{\theta}))]\right\}, \tag{24}$$

with the so-called evidence lower bound (ELBO) defined as:

$$\mathrm{ELBO}[q(\boldsymbol{\theta})] := \mathbb{E}_q[\log(p(\boldsymbol{\theta},\mathbf{y}_{1:T}))] - \mathbb{E}_q[\log(q(\boldsymbol{\theta}))]. \tag{25}$$

The usefulness of VB is that, for certain models, $p(\mathbf{y}_{1:T}|\boldsymbol{\theta})$, and certain choices of $\mathcal{Q}$, the optimization problem in (24) can be solved efficiently using various numerical algorithms. Most notably, for problems in which $\boldsymbol{\theta}$, and possibly $\mathbf{y}_{1:T}$ also, are high-dimensional, the production of $q^*(\boldsymbol{\theta})$ is *much* faster (often orders of magnitude so) than producing an estimate of $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ (and any associated quantities, including predictives) via simulation (Braun and McAuliffe, 2010; Kabisa *et al.*, 2016; Wand, 2017; Koop and Korobilis, 2018). The relationship between (23) and (25), plus the fact that $KL[\cdot] \geq 0$, also means $\mathrm{ELBO}[q^*(\boldsymbol{\theta})]$ is a lower bound on the logarithm of the 'evidence', or marginal likelihood, $p(\mathbf{y}_{1:T})$; hence the abbreviation 'ELBO'.

Different VB methods are defined by both the choice of $\mathcal{Q}$ and the manner in which the optimization is implemented, and we refer the reader to Ormerod and Wand (2010), Blei *et al.* (2017), and Zhang *et al.* (2018), for reviews, including algorithmic details for specific VB methods.

A.8 Integrated nested Laplace approximation (INLA)

Rue *et al.* (2009) adapted the very early approximation method of Laplace (1774) to approximate posteriors (and associated quantities) in the latent Gaussian model class, which encompasses a large range of – potentially high-dimensional – models, including the non-Gaussian state space models that feature heavily in economics and finance. In brief, Rue *et al.* use a series of *nested* Laplace approximations, allied with low-dimensional numerical *integration*, denoting their method by *integrated nested Laplace approximation* (INLA) as a result. As with VB, INLA eschews simulation for optimization, exploiting bespoke numerical algorithms designed for the specific (albeit broad) model class. We refer the reader to Rue *et al.* (2009), Rue *et al.* (2017), Martino and Riebler (2019), and Wood (2019) for implementation details.