Non-Parametric Estimation of Forecast Distributions in Non-linear, Non-Gaussian State Space Models

A thesis submitted for the degree of

Doctor of Philosophy

by

Jason Wei Jian Ng

B.Com.(Hons), Monash University, Australia

Department of Econometrics and Business Statistics

Faculty of Business and Economics

Monash University

Australia

January 2012

Contents

[1] All numerical results in this and the following empirical section have been produced using the GAUSS programming language.

List of Figures

List of Tables

Notice 1

Notice 2

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

Acknowledgments

I would like first and foremost, to express my sincere thanks to my supervisors Professor Gael Martin and Associate Professor Catherine Forbes. I could not be more grateful for the patient guidance, support, encouragement and direction that I have received from them over the last few years. I am also immensely grateful for the generosity of the Cochrane family in awarding me the Donald Cochrane Scholarship for the duration of my candidature.

I would like to give thanks for the support and invaluable assistance that I received, throughout many years, from the Department of Econometrics and Business Statistics at Monash University. In particular, I thank Professor Brett Inder and Dr. Katy Cornwell for the teaching opportunities at the undergraduate level. Thanks are also due to the administration team for its fantastic support over the years. Finally, I am most appreciative of the hand of friendship extended to me by the various academic staff in the Department.

I am thankful to my close circle of friends who have stood by me at all times. In particular, I would like to thank Ian Song, for his help and guidance, and my good friend, Kevin Liew, for being such an awesome brother to me.

I also wish to thank my parents and brother for their unconditional love,

encouragement and support in helping me achieve my goals.

Finally, I wish to thank God for His grace and provision in allowing me to reach this milestone in my life.

Abstract

Non-Gaussian time series variables are prevalent in the economic and finance spheres, with state space models often employed to analyze such variables and, ultimately, to produce forecasts. A review of the relevant literature reveals that existing methods are characterized by a reliance on (potentially incorrect) parametric assumptions and are often computationally expensive. The primary aim of this thesis is to develop a non-parametric approach to forecasting - within the state space framework - with computational ease an important focus. With a view to capturing all relevant information about the likely future values of the variable of interest, the approach is used to produce non-parametric estimates of the full forecast distribution over any time horizon.

Simulation experiments are used to document the accuracy of the non-parametric method relative to both correctly and incorrectly specified parametric alternatives, in a variety of relevant settings. Applying a range of methods for evaluating and comparing distributional forecasts, the non-parametric method is shown to perform significantly better, overall, than misspecified parametric alternatives while remaining competitive with correctly specified parametric estimators.

Focus is then given to the development of a new non-Gaussian state space model for observed realized volatility from which estimates of forecast distributions of future volatility are produced using the non-parametric method. In an empirical illustration, the non-parametric method is used to produce sequential estimates of the out-of-sample one-step-ahead forecast distribution of realized volatility on the S&P500 index during the recent financial crisis. A resampling technique for measuring sampling variation in an estimated forecast distribution is also demonstrated.

The proposed filtering algorithm is further extended to cater, in particular, for multi-step-ahead forecasting and multivariate systems. A simulation-based version of the algorithm is also illustrated, with the algorithm in this form seen to be a computationally efficient alternative to existing particle filtering algorithms.

Chapter 1

Introduction

1.1 Background

Non-Gaussian time series variables are prevalent in the economic and finance spheres, where deviations from the symmetric bell-shaped Gaussian distribution may arise for a variety of reasons, for example due to the positivity of the variable, to its inherent integer or binary nature, or to the prevalence of values that are far from, or unevenly distributed around, the mean. Against this backdrop, the challenge is to produce forecasts that are coherent - i.e. consistent with any restrictions on the values assumed by the variable - and that also encompass all important distributional information. Point forecasts, based on measures of central tendency, are common. However, they may not be coherent - evidenced, for example, by a real-valued conditional mean forecast of an integer-valued variable. Moreover, such measures convey none of the distributional information that is increasingly important for decision making (e.g. risk management), most notably as concerns the probability of occurrence of extreme outcomes. In contrast, an estimate of the full probability distribution, defined explicitly over all possible future values of the random variable is, by its very construction, coherent, as well as reflecting all of the important distributional features (including tail features) of the variable in question.

Such issues have informed recent work in which distributional forecasts have been produced for specific non-Gaussian data types (e.g. Freeland and McCabe, 2004a,b; McCabe and Martin, 2005; Jung and Tremayne, 2006; Bu and McCabe, 2008; Bu, Hadri and McCabe, 2008; Czado, Gneiting and Held, 2009; McCabe, Martin and Harris, 2011, for counts; Bauwens, Giot, Grammig and Veredas, 2004, for trade durations; Hong, Li and Zhou, 2004, for interest rates; Amisano and Giacomini, 2007, for inflation). The need to forecast the probability of large financial losses has also been the primary reason for the recent focus on distributional forecasting of portfolio returns (Diebold, Gunther and Tay, 1998; Berkowitz, 2001; Geweke and Amisano, 2010), with this literature, in turn, closely linked to that in which extreme quantiles (or Values at Risk) are the focus of the forecasting exercise (Danielsson and de Vries, 2000; Engle and Manganelli, 2004; Giacomini and Komunjer, 2005; de Rossi and Harvey, 2009). The extraction of risk-neutral distributional forecasts of non-Gaussian asset returns from derivative prices (Bakshi, Cao and Chen, 1997; Aït-Sahalia and Lo, 1998; Bates, 2000; Lim, Martin and Martin, 2005) is motivated by similar goals; i.e. that deviation from Gaussianity requires attention to be given to the prediction of higher order moments and to future distributional characteristics.[1]

[1] Discussion of the merits of probabilisitic forecasting in general is provided by, amongst others, Dawid (1984), Tay and Wallis (2000), Corradi and Swanson (2006), Gneiting and Raftery (2007), Gneiting, Balabdaoui and Raftery (2007) and Gneiting (2008).

The existing literature, in large part, uses parametric models to estimate forecast distributions, with the accuracy of the resultant predictions being dependent of the validity of the parametric assumptions. In contrast, certain analyses eschew strict parametric specifications, producing 'non-parametric' estimates of forecast distributions of one kind or another.[2] As an early example, Aït-Sahalia and Lo (1998) propose a non-parametric technique for estimating the state-price density (SPD) implicit in financial option prices. This technique involves using non-parametric kernel regression methods to estimate an option-pricing formula which is then used to derive the SPD. When applied to the S&P500 stock index, the non-parametric method is shown to be superior to a misspecified parametric approach based on the Black-Scholes pricing model. In a somewhat different context, McCabe, Martin and Harris (2011) propose a non-parametric approach for estimating predictive distributions of count time series modelled by the integer auto-regressive (INAR) class. The approach treats the arrivals, or innovations, process non-parametrically, with a parametric structure maintained for the dynamic component of the model. In addition to achieving asymptotic non-parametric efficiency, the non-parametric method is shown to perform better overall in finite samples than various misspecified parametric alternatives.

State space models are also used in the development of non-parametric methods for estimating predictive distributions. For example, Rodriguez and Ruiz

[2] We follow Li and Racine (2007) in using the term 'non-parametric' to refer to statistical techniques that do not require the specification of a functional form for an object being estimated; a forecast distribution in our case. We return to this matter of terminology at a later point.

(2009), building on earlier work of Pascual, Romo and Ruiz (2006), present a bootstrap-based approach for the estimation of prediction intervals in a linear state space setting. The bootstrap procedure is used to produce prediction intervals for the observed variable, via the Kalman filter recursions, but avoids the assumption of Gaussian measurement errors by using random draws from the empirical distributions of the measurement error. Although the bootstrap procedure focuses on the estimation of prediction intervals and not the predictive distribution of the observed variable *per se*, kernel smoothing techniques can be applied to the bootstrap draws of the future value of the observed variable to construct a non-parametric estimate of its predictive density.

Also within a state space framework, Durham (2007) uses a standard stochastic volatility (SV) model to obtain a semi-parametric estimate of the forecast distribution of financial returns, by using a mixture of normals for the conditional distribution of returns, allied with simulation-based inference. Monteiro (2010), on the other hand, represents the unknown measurement error density in a linear state space model as a Gaussian-sum. To avoid a geometric increase in the computational load imposed by the Gaussian-sum methodology, standard clustering algorithms are employed at each iteration of the filter. In addition to estimating the parameters of the model via maximum likelihood methods, Monteiro also estimates the unknown measurement error density itself.

In addition to time series work, the literature includes applications of non-parametric techniques to cross sectional data. Conditional density estimation was originally proposed in Rosenblatt (1969), and is especially important in problems where, for a given value of a vector of explanatory variables, the interest lies in estimating the conditional density of a response variable. Hall, Racine and Li (2004), for example, propose and apply a cross-validation technique to estimate the conditional probability density function of female labour force participation given certain female worker characteristics (e.g. age, educational attainment, number of children in different age groups). In another empirical example, Hansen (2004) applies a two-step conditional density estimator to estimate the conditional density of log-wages given the ages of the workers.

This brief literature survey provides a context for the focus of this thesis, as follows. First, in keeping with the large number of important economic and financial time series variables that exhibit non-Gaussianity, our primary interest is in forecasting non-Gaussian time series data *per se*. Second, with a view to capturing all uncertainty about the future realization of a time series variable - including extreme values - we focus on probabilistic (as opposed to point) forecasting. Third, in adopting a non-parametric approach, estimated forecast distributions that are not reliant on the correct specification of the true data generating process (DGP) are produced. Finally, and in contrast to the problem-specific approaches in the existing literature, we adopt a very general approach, via the non-linear, non-Gaussian state space framework that is applicable to many empirical settings.

1.2 Overview of the Thesis

The outline of the thesis is as follows. Chapter 2 begins with the introduction of a general state space model for a multivariate time series variable. The recursive filtering and prediction steps, used to produce the one-step-ahead predictive distributions for the observed variable, are outlined. The specific filters detailed are the Kalman filter, the extended Kalman filter, the unscented Kalman filter, the grid-based non-Gaussian filter of Kitagawa (1987), the Gaussian-sum filter and various particle filters. For fixed parameters of the model, the filtering algorithms can be used, at least in principle, to numerically evaluate the joint density associated with the observed data, marginal of the latent state variable. Treating this joint density as a function of the parameters, the resulting likelihood function is numerically maximized to produce maximum likelihood (ML) estimates of the unknown parameters in the model. The estimate of the one-step-ahead predictive distribution is subsequently produced by conditioning on the ML parameter estimates.

Chapter 3 contributes to the probabilistic forecasting literature by proposing a filtering algorithm - within the general non-linear, non-Gaussian state space framework - that provides an approximation to the true (but unavailable) filtering and predictive distributions. As will be shown, the proposed filter is a recursive algorithm, with the Dirac delta function used to recast all relevant filtered and predictive densities into integrals that are undertaken with respect to the invariant distribution of the measurement error. When the measurement error distribution is unknown, the method may be viewed as a non-parametric filtering algorithm, with ordinates of the unknown error density, at fixed grid locations, estimated within an ML procedure. The recursive filtering and prediction densities are then both used to define the likelihood function, and, ultimately, the out-of-sample predictive distribution for the non-Gaussian variable. Through this approach, the non-parametric filter produces distributional forecasts that are not reliant on the complete specification of the true DGP.

In order to assess the accuracy of the non-parametric forecasting method against other possible approaches, Chapter 4 presents the available tools used to compare and evaluate alternative forecast distributions produced from competing methods. Extensive simulation experiments are then conducted to produce non-parametric and parametric estimates of the forecast distributions in the context of the linear model and the non-linear stochastic conditional duration (SCD) model, with three different distributional assumptions for the true measurement error in each case. Parametric estimates of the forecast distributions are produced through the use of Kalman filter-based approaches, and are compared with the non-parametric estimates arising from the newly proposed method. The results show that the non-parametric estimator performs significantly better, overall, than do the (misspecified) parametric alternatives, while remaining competitive with a correctly specified parametric method.

Chapter 5 contributes to the financial volatility forecasting literature by proposing a realized volatility state space model, and by using the non-parametric methodology to estimate the full forecast distribution of realized volatility. A simulation exercise is undertaken to assess the forecasting performance of the non-parametric method, against the parametric alternatives, in this context. The non-parametric estimation method is then applied to the problem of estimating the forecast distribution of realized volatility for the S&P500 market index during the recent financial turmoil, with results showing that the non-parametric predictive distribution is able to capture important distributional information about the future value of the realized volatility of the index. A resampling method is also used to cater for estimation uncertainty in the production of the probabilistic forecasts of volatility.

Chapter 6 considers various extensions of the algorithm proposed in Chapter 3. In the first instance, the production of the multi-step-ahead forecast distribution is detailed. Second, two key assumptions that underpin the non-parametric filter in Chapter 3 are relaxed. Third, the non-parametric filter is illustrated in the multivariate setting. Lastly, despite the primary focus of the thesis being on a non-parametric setting in which the measurement error distribution is unknown, the filtering algorithm proposed in Chapter 3 is discussed here in the context of models in which the measurement error distribution is specified parametrically, and is able to be simulated from. This modification results in a simulation-based non-linear filter, in which all relevant integrals are evaluated by Monte Carlo simulation. This filter, in using simulation from the invariant measurement distribution, is shown to provide a simple and computationally efficient alternative to conventional particle filtering methods.

In the concluding chapter, the main contributions of the thesis are reiterated, namely the successful development and evaluation of a new method for producing non-parametric forecast distributions in non-linear, non-Gaussian state space models, and the application of the method to a new model developed for an empirically relevant problem. The novelty of the new filtering methodology and its computational simplicity are highlighted and its potential applicability to a wide range of settings summarized.

Chapter 2

State Space Models

2.1 Introduction

State space models provide a unified and flexible parametric framework for modelling and describing a wide range of time series data arising in a variety of disciplines. The paper by Kalman (1960) sparked the early development of the state space methodology in the field of engineering. Kalman illustrated that a broad class of problems could be expressed by a linear, Gaussian model, with the Markovian nature of the model allowing for the calculations needed for practical application of the model, to be set up in a simple recursive form convenient for computing. Further development of these ideas occurred subsequently in the engineering field, with contributions to the state space methodology from statisticians and econometricians occurring relatively infrequently until the early 1980's.

In more recent years, however, state space methods have been more widely adopted in the fields of statistics and econometrics, and are now routinely used in time series analysis and in other areas where longitudinal data play a role (e.g. Ansley and Kohn, 1985; Harvey and Durbin, 1986; Kitagawa 1987, 1989 and 1994; Carter and Kohn, 1994; Frühwirth-Schnatter, 1994 and 2004; De Jong and Shephard, 1995; Harvey and Chung, 2000; Durbin and Koopman, 2000 and 2001). They have been widely used in empirical finance, with leading examples being variants of the stochastic volatility model (see for example, Harvey, Ruiz and Shephard, 1994; Jacquier, Polson and Rossi, 1994; Shephard and Pitt, 1997; Anderson, 2001; Chernov, Gallant, Ghysels and Tauchen, 2003; Eraker, Johannes and Polson, 2003; Eraker, 2004; Broadie, Chernov and Johannes, 2007) and the stochastic conditional duration model (Bauwens and Veredas, 2004; Strickland, Forbes and Martin, 2006).

The increasingly widespread use of the state space approach can be attributed to the following four key characteristics. First, the state space approach lends itself to a structural analysis of the problem, whereby different components that make up the time series (e.g. trend, seasonal and cyclical) are modelled separately within the model (e.g. Durbin and Harvey, 1985; Harvey and Durbin, 1986). This is in contrast with the traditional approach of Box and Jenkins (1970), in which the model adopted depends solely on a reduced form, rather than on the structure of the system that is thought to have generated the data. Second, state space models are flexible, allowing for changes in the structure of the system over time. On the other hand, the autoregressive integrated moving average (ARIMA) models favoured by Box-Jenkins models are invariant through time, since these models require time series data to be stationary, or to be made stationary through either transformation or differencing of the data. Third, the state space framework is very general. For example, many commonly used statistical models, including ARIMA and multiple linear regression models, have a state space representation. Further, multivariate observed and explanatory variables can also be handled by simple extensions of univariate state space theory. Most importantly, appropriate specification of the measurement and/state equations can be used to cater for any data type, whether continuous or discrete, and whether defined on a restricted or an unrestricted support.

Key to the use of state state models, for the purposes of both inference and forecasting, is the need to manage the presence of the unobservable random states via filtering techniques. In the remainder of this chapter, the main concepts of the filtering literature are reviewed, with particular emphasis given to out-of-sample prediction and to flexibility in the specification of the measurement error distribution. Section 2.2 presents a general parametric state space model, with the associated inferential objects briefly discussed in Section 2.3. Section 2.4 presents the general filtering steps leading to the one-step-ahead predictive distributions that are needed for specifying the likelihood function, followed by the specific filters used, in various settings, to produce these predictive distributions. In an attempt to provide a coverage of the most commonly used filters in the relevant literature, we outline the following: the Kalman filter (Kalman, 1960, Kalman and Bucy, 1961), the extended Kalman filter (Anderson and Moore, 1979), the unscented Kalman filter (Julier, Uhlmann and Durrant-Whyte, 1995, 1996), the grid-based non-Gaussian filter (Kitagawa, 1987), the Gaussian-sum filter (Sorenson and Alspach, 1971; Monteiro, 2010) and representative particle filters. Section 2.5 discusses how the unknown static parameters in the state space model may be estimated using ML estimation, resulting in an estimate of the out-of-sample one-step-ahead predictive distribution. Section 2.6 concludes by discussing the limitations of the parametric forecasting approach, thereby motivating the proposed non-parametric approach in Chapter 3.

2.2 The General Parametric State Space Model

The general parametric state space model relates to a time series of (possibly vector valued) observations $\{y_t\ , t = 1, 2, ..., T\}$, with each observation y_t being a noisy measurement of an underlying (also possibly vector valued) *unobserved* state variable, x_t. This relationship is expressed through the measurement probability density function (pdf),

$$p\left(y_t|x_t, \theta\right), \tag{2.1}$$

for $t = 1, 2, ...T$, where y_t is a $(p \times 1)$ vector of observations, x_t is an $(m \times 1)$ vector of unobserved components and θ denotes a $(q \times 1)$ vector of unknown parameters. The evolution of the $(m \times 1)$ state variable, x_{t+1}, from the previous value x_t, is described by the state transition pdf,

$$p\left(x_{t+1}|x_t, \theta\right), \tag{2.2}$$

for $t = 1, 2, ..., T$, with an initial state pdf,

$$p\left(x_1|\theta\right), \tag{2.3}$$

also specified. It is noted that here the exposition focuses on both y_t and x_t as continuous random variables, with all distributions expressed using density functions as a consequence; however the framework is general and can be adapted, for example, to cater for the case where y_t and/or x_t is discrete.

The expressions in (2.1) and (2.2) are often defined via regression relationships, with the model formally expressed by the following measurement and state relationships:

$$\text{Measurement equation :} \qquad y_t = h_t\left(x_t, \eta_t\right) \tag{2.4}$$

$$\text{State equation :} \qquad x_{t+1} = k_t\left(x_t, v_t\right), \tag{2.5}$$

for $t = 1, 2, ..., T$. The measurement equation (2.4) expresses y_t as a function of the state variable and the measurement error η_t, while the state equation (2.5) expresses x_{t+1} as a function of its lagged value, x_t, and the state error v_t. Either or both of these functions may be dependent on (one or more element of) the vector of parameters θ, through the dependence on θ of the densities in (2.1) to (2.3). Each η_t is assumed to be an *i.i.d.* random variable, that is, any dynamic behaviour in y_t is captured completely by $h_t\left(\cdot,\cdot\right)$ and $k_t\left(\cdot,\cdot\right)$. As is also common, we assume that η_t is independent of x_t, in which case the pdf for η_t is simply $p\left(\eta_t|x_t\right) = p\left(\eta_t\right)$, for all $t = 1, 2, ..., T$.

2.3 Inference in State Space Models

Four types of inference may be conducted in the context of state space models: state filtering, state smoothing, out-of-sample forecasting (of both the observed and the latent), and estimation of the vector of unknown parameters, θ, with this section providing a brief overview of each. To aid the discussion, the notation

$$y_{1:t} = \{y_1, y_2, ..., y_t\},$$

for $t = 1, 2, ..., T$ is used to denote the subset of observations beginning with y_1, up to and including the t^{th} observation, y_t.

2.3.1 State Filtering

With reference to the general state space model in (2.3) to (2.5), filtering refers to the determination of the distribution of the state vector, x_t, given $y_{1:t}$ (i.e. given the observed data up to and including period t), as represented by

$$p\left(x_t|y_{1:t}, \theta\right), \tag{2.6}$$

for each $t = 1, 2, ..., T$. The objective of filtering is to update knowledge of the system each time a new value of y_t is observed. Therefore, filtering is a recursive procedure that is applied for each t, revising the filtered density, $p\left(x_t|y_{1:t}, \theta\right)$, using the new observation y_{t+1}, to produce the updated density, $p\left(x_{t+1}|y_{1:t+1}, \theta\right)$. Closed form expressions for (2.6) are available in only very few cases, with the methods detailed in Section 2.4 approximating these quantities in different ways. A filtering process, in revising each state density, produces (an approximation to) the one-step-ahead predictive densities, $p\left(y_{t+1}|y_{1:t}, \theta\right)$ for $t = 1, 2, ...T$, as a by-product. Further details of the general filtering (and up-dating) process, plus an outline of specific filtering algorithms will be given in Section 2.4.

2.3.2 State Smoothing

Smoothing refers to the determination of the distribution of the state vector, x_t, conditional upon *all* available observations, $y_{1:T}$, resulting in

$$p\left(x_t|y_{1:T},\theta\right), \tag{2.7}$$

for each $t = 1, 2, ..., T$. The difference between filtered distributions and smoothed distributions is that the smoothed distribution for x_t depends upon the entire set of observations, $y_{1:T}$, and not only on the portion of observations available at time t, $y_{1:t}$. Of course, for the special case of $t = T$, the filtered distribution is the same as the smoothed distribution. A smoothing procedure will typically take place after a filtering procedure has been completed, beginning at time $T-1$ and working backwards for $t = T-2, ..., 2, 1$ to update the filtered distributions to achieve conditioning on all of the observations.

2.3.3 Forecasting

Distributional forecasting of the measurement variable y_t refers to the determination of the distribution of a future observation, y_{T+s}, given all the observations, $y_{1:T}$, represented as

$$\begin{aligned} p\left(y_{T+s}|y_{1:T},\theta\right) &= \int p\left(y_{T+s},x_{T+s}|y_{1:T},\theta\right)\ dx_{T+s} \\ &= \int p\left(y_{T+s}|x_{T+s},\theta\right)p\left(x_{T+s}|y_{1:T},\theta\right)\ dx_{T+s}, \end{aligned} \tag{2.8}$$

where

$$\begin{aligned} p\left(x_{T+s}|y_{1:T},\theta\right) &= \int p\left(x_{T+s}|x_{T+s-1},\theta\right)p\left(x_{T+s-1}|y_{1:T},\theta\right)\ dx_{T+s-1} \\ &= \int ... \int p\left(x_{T+s}|x_{T+s-1},\theta\right)p\left(x_{T+s-1}|x_{T+s-2},\theta\right)... \\ &\quad p\left(x_{T+1}|x_{T},\theta\right)p\left(x_{T}|y_{1:T},\theta\right)\ dx_{T}dx_{T+1}...dx_{T+s-1}, \quad (2.9) \end{aligned}$$

and the term $s > 0$ is referred to as the forecast horizon. The density in (2.9) provides a distributional forecast of the latent variable x_t at future period $t = T+s$, a relevant output when the latent variable is of interest in its own right, in addition to being a necessary input into the forecast distribution of the observed in (2.8). Closed form expressions for (2.8) and (2.9) are typically unavailable, with the filtering (and up-dating) methods detailed in Section 2.4 in the main producing different *approximations* to these quantities only. When $s = 1$, (2.8) and (2.9) correspond to out-of-sample one-step-ahead predictive distributions. Primary focus is given to one-step-ahead prediction throughout this thesis, but with extension to the multi-step-ahead case outlined in Chapter 6.

2.3.4 Estimation

In the three previous sections, the vector of parameters, θ, is implictly assumed to be known. However, in the vast majority of applications, θ is unknown and has to be estimated. An ML approach to the estimation of θ is adopted in this thesis. Under this approach, the one-step-ahead prediction distributions, $p\left(y_{t+1}|y_{1:t},\theta\right)$ for $t = 1, 2, ..., T-1$, produced (or approximated) as a by-product of the filtering process, are used to construct the likelihood function, via the prediction error decomposition, as

$$L(\theta) \propto p(y_1|\theta) \prod_{t=1}^{T-1} p(y_{t+1}|y_{1:t}, \theta). \tag{2.10}$$

The ML estimate of θ, $\widehat{\theta}$, is obtained by maximizing (2.10) with respect to θ, typically via numerical optimization of the logarithm of $L(\theta)$. Bayesian methods are also commonly used in the state space framework, but will not be considered here.[1]

It is now evident that the ability to produce distributional forecasts, and to estimate the vector of parameters θ if unknown, lies in the ability to perform the filtering process that results in the one-step-ahead predictive distributions, $p(y_{t+1}|y_{1:t}, \theta)$. The next section will present the general steps required of a filtering procedure and the details of some specific filters that produce either exact (if available) or approximate representations of $p(y_{t+1}|y_{1:t}, \theta)$, and ultimately $p(y_{T+1}|y_{1:T}, \theta)$. As state smoothing is not required for forecasting and (ML) estimation purposes, the issue of smoothing will not be pursued further in the thesis. From this point on the explicit dependence of all distributions on θ is also suppressed in the exposition.

2.4 Filtering Methods in State Space Models

There are essentially three steps taken at each time t to produce the filtered distribution $p(x_t|y_{1:t})$, with a by-product of the process being the production of the one-step-ahead predictive distribution, $p(y_{t+1}|y_{1:t})$. Before these steps can

[1] A recent survey of Bayesian methods in state space models is provided in Giordani, Pitt and Kohn (2011).

be undertaken, the first filtered distribution, $p\left(x_1|y_{1:1}\right)$, must be obtained. This is done by initializing the filter with the assumed parametric distribution $p\left(x_1\right)$, updated having observed y_1 via Bayes theorem as

$$p\left(x_1|y_{1:1}\right) = \frac{p\left(y_1|x_1\right)p\left(x_1\right)}{p\left(y_1\right)}$$

where

$$p\left(y_1\right) = \int p\left(y_1|x_1\right)p\left(x_1\right)\ dx_1.$$

Then, for each $t = 1, 2, ..., T-1$, we undertake the following three filtering steps:

1. Produce the state predictive distribution as the integral of $p\left(x_{t+1}, x_t|y_{1:t}\right)$ with respect to x_t, with the joint density obtained from the product of the transition density and the filtered density. That is,

$$\begin{aligned} p\left(x_{t+1}|y_{1:t}\right) &= \int p\left(x_{t+1}, x_t|y_{1:t}\right)\ dx_t \\ &= \int p\left(x_{t+1}|x_t\right)p\left(x_t|y_{1:t}\right)\ dx_t. \end{aligned} \tag{2.11}$$

2. Given $p\left(x_{t+1}|y_{1:t}\right)$, and using knowledge of the measurement density in (2.1), produce the one-step-ahead predictive distribution for the measurement as

$$\begin{aligned} p\left(y_{t+1}|y_{1:t}\right) &= \int p\left(y_{t+1}, x_{t+1}|y_{1:t}\right)\ dx_{t+1} \\ &= \int p\left(y_{t+1}|x_{t+1}\right)p\left(x_{t+1}|y_{1:t}\right)\ dx_{t+1}. \end{aligned} \tag{2.12}$$

3. Produce the up-dated filtered state density, given the observation y_{t+1}, by applying Bayes theorem,

$$p\left(x_{t+1}|y_{1:t+1}\right) = \frac{p\left(y_{t+1}|x_{t+1}\right)p\left(x_{t+1}|y_{1:t}\right)}{p\left(y_{t+1}|y_{1:t}\right)}. \tag{2.13}$$

At period T, the out-of-sample one-step-ahead forecast distribution is produced by using Steps 1 and 2 above, without the revision Step 3.

This filtering process is often difficult to perform as it requires recursive integration. Apart from certain special cases, of which the linear, Gaussian case is the canonical example, the recursive integrals required to implement the filtering and up-dating process cannot be solved analytically, and some form of approximation is required as a consequence.

2.4.1 Kalman Filter

If the functions $h_t(\cdot,\cdot)$ and $k_t(\cdot,\cdot)$ in model (2.4) and (2.5), are linear, with both error terms additive and Gaussian, the state space model may be represented as

$$y_t = c_t + H_t x_t + \eta_t \tag{2.14}$$

$$x_{t+1} = d_t + K_t x_t + v_t \tag{2.15}$$

$$\begin{pmatrix} \eta_t \\ v_t \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} R_t & 0 \\ 0 & Q_t \end{pmatrix}\right), \tag{2.16}$$

for $t = 1, 2, ..., T$. As noted earlier, y_t is a $(p \times 1)$ vector of observations and x_t is an $(m \times 1)$ vector of unobserved components. The random components in the measurement and state equations, $\{\eta_t\}$ from (2.14) and $\{v_t\}$ from (2.15), respectively, represent mutually independent, zero-mean Gaussian random variables having variance-covariance matrices R_t and Q_t, respectively. For the present purpose, the system matrices H_t and K_t are assumed known, with H_t a $(p \times m)$ matrix and K_t an $(m \times m)$ matrix. We also assume that the dimension of the state variable x_t is the same as that of the state error term v_t, with modification of the filter available if $\dim(v_t) < \dim(x_t)$; see for example Durbin and Koopman (2001). The intercepts c_t in (2.14) and d_t in (2.15) are $(p \times 1)$ and $(m \times 1)$ vectors of known constants, respectively. In this specific case, θ could be comprised of q elements contained within c_t, d_t, H_t, K_t, R_t, Q_t, $a_{1|0}$ and $V_{1|0}$. The initial state vector, x_1 has a marginal distribution given by

$$x_1 \sim N\left(a_{1|0},\ V_{1|0}\right), \tag{2.17}$$

where $a_{1|0}$ and $V_{1|0}$ are also assumed known.

When the state space model is linear and Gaussian, as is the model in (2.14)-(2.17), all filtered and predictive distributions for the state are Gaussian, as are the predictive distributions for the observed, $p\left(y_{t+1}|y_t\right)$, for $t = 1, 2, ..., T$. Hence, only the first two moments of these distributions need to be calculated for each entire distribution to be determined. The Kalman filter provides the required recursions for the calculation of these moments. Denoting the first two moments respectively for each of (2.11) and (2.13) by

$$\begin{aligned} E\left(x_{t+1}|y_{1:t}\right) &= a_{t+1|t} \\ Var\left(x_{t+1}|y_{1:t}\right) &= V_{t+1|t}, \end{aligned}$$

and

$$\begin{aligned} E\left(x_{t+1}|y_{1:t+1}\right) &= a_{t+1|t+1} \\ Var\left(x_{t+1}|y_{1:t+1}\right) &= V_{t+1|t+1}, \end{aligned}$$

the Kalman filter equations corresponding to the state space model presented in (2.14)-(2.17) are

$$\begin{aligned} a_{t+1|t} &= d_t + K_t a_{t|t} && (2.18a) \\ V_{t+1|t} &= K_t V_{t|t} K_t' + Q_t && (2.18b) \\ a_{t+1|t+1} &= a_{t+1|t} + M_{t+1}\varepsilon_{t+1} && (2.18c) \\ V_{t+1|t+1} &= \left(I - M_{t+1}H_{t+1}\right) V_{t+1|t}, && (2.18d) \end{aligned}$$

where

$$\begin{aligned} M_{t+1} &= V_{t+1|t} H_{t+1}' F_{t+1}^{-1} && (2.18e) \\ F_{t+1} &= H_{t+1} V_{t+1|t} H_{t+1}' + R_{t+1} && (2.18f) \\ \varepsilon_{t+1} &= y_{t+1} - c_{t+1} - H_{t+1} a_{t+1|t}, && (2.18g) \end{aligned}$$

for $t = 1, 2, ..., T$. Therefore, the filtered distribution for the state x_{t+1} is a Gaussian density with its mean and covariance respectively given by $a_{t+1|t+1}$ in (2.18c) and $V_{t+1|t+1}$ in (2.18d), while the predictive state distribution is also a Gaussian density with its mean and covariance respectively given by $a_{t+1|t}$ in (2.18a) and $V_{t+1|t}$ in (2.18b). Resulting from the filter, each one-step-ahead predictive distribution corresponding to the measurement at time $t + 1$ is also Gaussian, with mean and covariance respectively given by

$$\begin{aligned} E\left(y_{t+1}|y_{1:t}\right) &= \mu_{t+1} = c_{t+1} + H_{t+1} a_{t+1|t} && (2.19) \\ Var\left(y_{t+1}|y_{1:t}\right) &= F_{t+1}, && (2.20) \end{aligned}$$

for each $t = 1, 2, ..., T,$ with F_{t+1} as given in (2.18f). Since $\varepsilon_{t+1} = y_{t+1} - \mu_{t+1}$, the corresponding predictive density is given by

$$p\left(y_{t+1}|y_{1:t}\right)=\left(2\pi\right)^{-\frac{p}{2}}\left|F_{t+1}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\varepsilon_{t+1}'F_{t+1}^{-1}\varepsilon_{t+1}\right)\right\}. \tag{2.21}$$

Note that the Kalman filter is initiated with $a_{1|1}$ and $V_{1|1}$ computed using (2.18c)-(2.18g).

2.4.2 Extended Kalman Filter

Even if the error sequences $\{\eta_t\}$ and $\{v_t\}$ are Gaussian, if the functions $h_t\left(\cdot,\cdot\right)$ and $k_t\left(\cdot,\cdot\right)$ in (2.4) and (2.5) are non-linear, exact solutions for the state predictive density in (2.11), the state filtered density in (2.13), and the one-step-ahead predictive density in (2.12), are not available in general. A simple approach to producing the densities in (2.11)-(2.13) is to approximate the model by linearizing the non-linear model about various points and then to perform filtering. The extended Kalman filter is based on this approach, whereby the non-linear state space model is linearized at each point t around the most relevant state estimate, taken to be the mean of either the previously filtered state density or the predictive state density, depending on the particular filtering step being undertaken at that point. Once a linearized model is obtained, and the assumption of Gaussian errors for both the measurement and state equations invoked, all filtered and predicted state distributions, along with the one-step-ahead predictive distributions, are produced using the appropriately modified version of the Kalman filter recursions. As a consequence, the extended Kalman filter provides approximate solutions for all filtered and predictive distributions associated with the model.

Following Anderson and Moore (1979) we demonstrate the extended Kalman filter in the context of the following representation of the state space model

$$y_t = h_t(x_t) + \eta_t \tag{2.22}$$

$$x_{t+1} = k_{1t}(x_t) + k_{2t}(x_t)\, v_t, \tag{2.23}$$

for $t = 1, 2, ..., T$. The functions $h_t(x_t)$, $k_{1t}(x_t)$ and $k_{2t}(x_t)$ are potentially non-linear functions of the state variable x_t, with $h_t(x_t)$ and $k_{1t}(x_t)$ assumed differentiable. The errors η_t and v_t are each assumed to be a zero mean, Gaussian random variable with a variance-covariance matrix represented by R_t and Q_t, respectively[2], with a joint distribution as per (2.16). As in the linear, Gaussian state space model, the initial state, x_1, is assumed to have a Gaussian distribution described by (2.17).

Referring to the state equation in (2.23), the non-linear functions $k_{1t}(x_t)$ and $k_{2t}(x_t)$, if sufficiently smooth, can be linearized by taking a Taylor series expansion for each function, about the (approximated) conditional mean of the filtered state density. In the spirit of the notation used in the Kalman filter, denote this approximated conditional mean by $\widehat{a}_{t|t}$. The functions $k_{1t}(x_t)$ and $k_{2t}(x_t)$ are each expanded about $\widehat{a}_{t|t}$, because having observed $y_{1:t}$, $\widehat{a}_{t|t}$ is, in some sense, the best available predictor of x_t. Taking a first-order expansion of

[2] Often in this setting, $Q_t \equiv 1$, as the scaling of v_t can be incorporated into the function $k_{2t}(x_t)$.

$k_{1t}(x_t)$ and a zero-order expansion of $k_{2t}(x_t)$ yields

$$\begin{aligned} k_{1t}(x_t) &\approx k_{1t}\left(\widehat{a}_{t|t}\right) + K_{1t}\left(x_t - \widehat{a}_{t|t}\right) && (2.24) \\ k_{2t}(x_t) &\approx k_{2t}\left(\widehat{a}_{t|t}\right) = K_{2t}, && (2.25) \end{aligned}$$

where

$$K_{1t} = \left.\frac{\partial k_{1t}(x)}{\partial x}\right|_{x=\widehat{a}_{t|t}}. \qquad (2.26)$$

Following the same reasoning, the non-linear function $h_t(x_t)$ in (2.22) is linearized by taking a first-order Taylor series expansion about the (approximate) conditional mean of the state predictive density, $\widehat{a}_{t|t-1}$, with the latter providing the most recent estimate of x_t used in determining the one-step-ahead prediction density $p(y_t|y_{1:t-1})$ at time $t-1$, in the filtering process. We then have

$$h_t(x_t) \approx h_t\left(\widehat{a}_{t|t-1}\right) + \Psi_t\left(x_t - \widehat{a}_{t|t-1}\right), \qquad (2.27)$$

where

$$\Psi_t = \left.\frac{\partial h_t(x)}{\partial x}\right|_{x=\widehat{a}_{t|t-1}}. \qquad (2.28)$$

Therefore, conditioning on the values of $\widehat{a}_{t|t-1}$ and $\widehat{a}_{t|t}$, the non-linear model in (2.22) and (2.23) is approximated by the linear, Gaussian state space model,

$$\begin{aligned} y_t &= \left(h_t\left(\widehat{a}_{t|t-1}\right) - \Psi_t\widehat{a}_{t|t-1}\right) + \Psi_t x_t + \eta_t && (2.29) \\ x_{t+1} &= \left(k_{1t}\left(\widehat{a}_{t|t}\right) - K_{1t}\widehat{a}_{t|t}\right) + K_{1t}x_t + K_{2t}v_t, && (2.30) \end{aligned}$$

and the Kalman filter recursions are then applied to the approximating model so defined. The so-called ‘extended’ Kalman filter equations for the non-linear model in (2.22) and (2.23) are then given by

$$\begin{aligned}
\widehat{a}_{t+1|t} &= k_{1t}\left(\widehat{a}_{t|t}\right) && \text{(2.31a)}\\
V_{t+1|t} &= K_{1t}V_{t|t}K_{1t}' + K_{2t}Q_tK_{2t}' && \text{(2.31b)}\\
\widehat{a}_{t+1|t+1} &= \widehat{a}_{t+1|t} + M_{t+1}\varepsilon_{t+1} && \text{(2.31c)}\\
V_{t+1|t+1} &= \left(I - M_{t+1}\Psi_{t+1}\right)V_{t+1|t}, && \text{(2.31d)}
\end{aligned}$$

where

$$\begin{aligned}
M_{t+1} &= V_{t+1|t}\Psi_{t+1}'F_{t+1}^{-1} && \text{(2.31e)}\\
F_{t+1} &= \Psi_{t+1}V_{t+1|t}\Psi_{t+1}' + R_{t+1} && \text{(2.31f)}\\
\varepsilon_{t+1} &= y_{t+1} - h_{t+1}\left(\widehat{a}_{t+1|t}\right), && \text{(2.31g)}
\end{aligned}$$

for $t = 1, 2, ..., T$. As is the case for the Kalman filter, the extended Kalman filter is initialized using (2.31c)-(2.31g) with $t = 0$.

The one-step-ahead predictive distribution for the measurement at each point t is also approximated as a Gaussian density as

$$p\left(y_{t+1}|y_{1:t}\right) \approx (2\pi)^{-\frac{p}{2}}\left|F_{t+1}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\varepsilon_{t+1}'F_{t+1}^{-1}\varepsilon_{t+1}\right)\right\}, \tag{2.32}$$

and, at time T, the out-of-sample one-step-ahead prediction distribution approximated as

$$p\left(y_{T+1}|y_{1:T}\right) \approx (2\pi)^{-\frac{p}{2}}\left|F_{T+1}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(y_{T+1} - \mu_{T+1}\right)'F_{T+1}^{-1}\left(y_{T+1} - \mu_{T+1}\right)\right\}, \tag{2.33}$$

where

$$E\left(y_{T+1}|y_{1:T}\right) = \mu_{T+1} = h_{T+1}\left(\widehat{a}_{T+1|T}\right) \tag{2.34}$$

$$Var\left(y_{T+1}|y_{1:T}\right) = F_{t+1}, \tag{2.35}$$

with F_{t+1} as given in (2.31f).

2.4.3 Unscented Kalman Filter

The extended Kalman filter, presented in Section 2.4.2, maintains the elegant and computationally efficient recursive updating form of the Kalman filter for producing predictive distributions in a non-linear state space model. However, the extended Kalman filter works on the assumption that a non-linear function of a random variable can be well approximated by a linear function of a Gaussian random variable, with the linear function obtained via a first-order Taylor series expansion of the original function around a single point while neglecting higher order terms. This assumption is dubious, especially for highly non-linear systems. The unscented Kalman filter (Julier *et al.*, 1995, 1996; Julier and Uhlmann, 1997) was proposed as an alternative to the extended Kalman filter that provides superior performance at a reduced level of computational complexity, with the added advantage of not requiring the computation of Jacobian matrices.

The unscented Kalman filter is based on the theory of unscented transformations, which is a method for calculating the moments of a random variable that has undergone a non-linear transformation. It involves selecting a set of points on the support of the random variable, called *sigma points*, according to a predetermined and deterministic criterion. These sigma points yield, in turn, a 'cloud' of transformed points through the non-linear function. One of the useful properties of the unscented transformation is that the mean of the transformed variable is calculated to a higher order of accuracy than is typically the case for the extended Kalman filter, while the variance-covariance of the transformed variable is calculated to the same order of accuracy as with the extended Kalman filter.

In the demonstration of the unscented filter we revert to the general representation for the (potentially) multi-dimensional y_t and x_t, as given by (2.4) and (2.5), but with simple additive forms of the error terms, η_t and v_t, invoked. Hence, the model is represented as

$$y_t = h_t(x_t) + \eta_t \tag{2.36}$$

$$x_{t+1} = k_t(x_t) + v_t \tag{2.37}$$

$$\begin{pmatrix} \eta_t \\ v_t \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} R_t & 0 \\ 0 & Q_t \end{pmatrix}\right), \tag{2.38}$$

for $t = 1, 2, ..., T$, where the functions $h_t(\cdot)$ and $k_t(\cdot)$ in (2.36) and (2.37) are non-linear in the state. The error terms η_t and v_t, are assumed to be Gaussian with a variance-covariance matrix represented by R_t and Q_t respectively, and with $p(x_1)$ given by (2.17). The unscented Kalman filter, like the extended Kalman filter, approximates all of the filtered and predictive state distributions, along with the one-step-ahead predictive distributions, with Gaussian distributions. The unscented filter then provides the required recursions for the calculation of the first two moments of these distributions, via the following steps.

1. Denote the (previously approximated) first two moments of $p\left(x_t|y_{1:t}\right)$ by $a_{t|t}$ and $V_{t|t}$. Approximate the first two moments of the state predictive distribution $p\left(x_{t+1}|y_{1:t}\right)$ as follows:

   (a) Form an $m \times (2m+1)$ matrix χ_t with columns containing the m-dimensional sigma vectors, $\chi_{i,t}$, for $i = 0, 1, ..., 2m$, defined according to

$$\chi_{0,t} = a_{t|t} \tag{2.39a}$$

$$\chi_{i,t} = a_{t|t} + \left(\sqrt{\left(m+\vartheta\right)V_{t|t}}\right)_i, \ i = 1, 2, ..., m \tag{2.39b}$$

$$\chi_{i,t} = a_{t|t} - \left(\sqrt{\left(m+\vartheta\right)V_{t|t}}\right)_i, \ i = m+1, ..., 2m \tag{2.39c}$$

where

$$\vartheta = \varphi^2\left(m+\tau\right) - m.$$

In the vector state case, i.e. with $dim(x_t) = m > 1$, the expression $\sqrt{\left(m+\vartheta\right)V_{t|t}}$ denotes the square root of the scaled covariance matrix $\left(m+\vartheta\right)V_{t|t}$ such that $\left(\sqrt{\left(m+\vartheta\right)V_{t|t}}\right)'\left(\sqrt{\left(m+\vartheta\right)V_{t|t}}\right) = \left(m+\vartheta\right)V_{t|t}$, with $\left(\sqrt{\left(m+\vartheta\right)V_{t|t}}\right)_i$ denoting the i^{th} column of $\sqrt{\left(m+\vartheta\right)V_{t|t}}$. The constants ϑ and φ are both scaling parameters determining the spread of the sigma points around $a_{t|t}$, with φ typically set to a small positive value. The constant τ is an additional scaling parameter which is usually set to 0 or $(3 - m)$. A useful heuristic is to select $m + \vartheta = 3$, so that if the transformed distribution were (scalar) Gaussian then the transformed sigma-points would achieve a weighted fourth moment equal to 3. (See Julier and Uhlmann, 2004, for details). Although the method may be applied to non-additive state and measurement equations, in that case the state variable would need to be augmented with the individual noise terms from each of the measurement equations and state, η_t and v_t, respectively, resulting in an increase in the number of sigma points required. For the case presented here, i.e. when the error terms are additive, the dimension of each sigma point is equal to m, the dimension of the state variable x_t, and the required number of sigma points used is lower than in the more general case (Wan and van der Merwe, 2001). If the state variable is a scalar and the error terms are additive, then $m = 1$ and only three sigma points are required within each transformation. (Note, however, that two separate transformations are required for each time t, given the presence of the two non-linear functions $h_t(\cdot)$ and $k_t(\cdot)$.)

(b) Propagate the sigma matrix in (2.39) through the non-linear state equation, by transforming each column vector according to $k_t(\cdot)$ in (2.37), i.e.

$$\chi^*_{i,t+1|t} = k_t\left(\chi_{i,t}\right), \tag{2.40}$$

and obtaining the matrix of transformed sigma vectors

$$\chi^*_{t+1|t} = \left(\chi^*_{0,t+1|t}, \chi^*_{1,t+1|t}, ..., \chi^*_{2m,t+1|t}\right).$$

The state predictive mean and covariance associated with $p\left(x_{t+1}|y_{1:t}\right)$, are then calculated using weighted functions of the transformed sigma vectors in (2.40), respectively, given by

$$\begin{aligned} a_{t+1|t} &= \sum_{i=0}^{2m} W_i^{(a)} \chi^*_{i,t+1|t} && (2.41) \\ V_{t+1|t} &= \sum_{i=0}^{2m} W_i^{(V)} \left[\chi^*_{i,t+1|t} - a_{t+1|t}\right] \left[\chi^*_{i,t+1|t} - a_{t+1|t}\right]' + Q_t. \end{aligned} \quad (2.42)$$

The weights, $W_i^{(a)}$ and $W_i^{(V)}$, for $i = 0, 1, 2, ...2m$, associated with the conditional mean $a_{t+1|t}$ and variance $V_{t+1|t}$ approximations respectively, are determined by

$$\begin{aligned} W_0^{(a)} &= \frac{\vartheta}{m+\vartheta} \\ W_0^{(V)} &= \frac{\vartheta}{m+\vartheta} + \left(1 - \varphi^2 - \varpi\right) \\ W_i^{(a)} &= W_i^{(V)} = \frac{1}{2\left(m+\vartheta\right)} \qquad \text{for } i = 1, 2, ..., 2m. \end{aligned} \quad (2.43)$$

Here the weight sets $\left\{W_i^{(a)}, i = 0, 1, ..., 2m\right\}$ and $\left\{W_i^{(V)}, i = 0, 1, ..., 2m\right\}$ must each sum to one. Note that when the scaling parameter $\vartheta = 0$, the weight associated with $\chi_{0,t}$ will be zero. In addition, although the individual weights may in general be non-negative, justification is also available for choosing the scaling parameters to ensure positive weights. The constant ϖ may be used to incorporate prior knowledge of the previously filtered distribution for $p\left(x_t|y_{1:t}\right)$, or the function $k_t\left(\cdot\right)$ in (2.37), and is set to 2 when the previously filtered distribution is assumed to be Gaussian and when no further information about $k_t\left(\cdot\right)$ is used (Julier and Uhlmann, 2004).

(c) The state predictive distribution $p\left(x_{t+1}|y_{1:t}\right)$ is then approximated by a Gaussian density, with its mean and covariance matrix given by (2.41) and (2.42), respectively.

2. The first two moments of $p\left(y_{t+1}|y_{1:t}\right)$ are obtained through the following steps:

(a) Determine a new $m \times (2m+1)$ matrix $\widetilde{\chi}_{t+1}$ of m-dimensional sigma vectors, $\widetilde{\chi}_{i,t+1}$, according to

$$\widetilde{\chi}_{0,t+1} = a_{t+1|t} \tag{2.44a}$$

$$\widetilde{\chi}_{i,t+1} = a_{t+1|t} + \left(\sqrt{\left(m+\widetilde{\vartheta}\right) V_{t+1|t}}\right)_i, \quad i = 1, 2, ..., m \tag{2.44b}$$

$$\widetilde{\chi}_{i,t+1} = a_{t+1|t} - \left(\sqrt{\left(m+\widetilde{\vartheta}\right) V_{t+1|t}}\right)_i, \quad i = m+1, ..., 2m \tag{2.44c}$$

where (analogously)

$$\widetilde{\vartheta} = \widetilde{\varphi}^2 \left(m + \widetilde{\tau}\right) - m.$$

Again here if $dim(x_t) = m > 1$, the expression $\left(\sqrt{\left(m + \widetilde{\vartheta}\right) V_{t+1|t}}\right)$ is the square root matrix of $\left(m + \widetilde{\vartheta}\right) V_{t+1|t}$, and $\left(\sqrt{\left(m + \widetilde{\vartheta}\right) V_{t+1|t}}\right)_i$ denotes the i^{th} column of $\sqrt{\left(m + \widetilde{\vartheta}\right) V_{t+1|t}}$.

(b) Propagate the sigma matrix in (2.44) through the non-linear measurement equation in (2.36), by transforming each column vector as

$$\widetilde{y}_{i,t+1} = h_{t+1}\left(\widetilde{\chi}_{i,t+1}\right) \tag{2.45}$$

and obtaining the corresponding matrix of transformed sigma vectors,

$$\widetilde{y}_{t+1} = \left(\widetilde{y}_{0,t+1}, \widetilde{y}_{1,t+1}, ..., \widetilde{y}_{2m,t+1}\right). \tag{2.46}$$

The approximate mean and covariance matrix, respectively, associated with the predictive distribution of $p\left(y_{t+1}|y_{1:t}\right)$, are then calculated as the weighted average of the transformed sigma vectors,

$$\mu_{t+1} = \sum_{i=0}^{2m} W_i^{(\mu)} \widetilde{y}_{i,t+1} \tag{2.47}$$

$$F_{t+1} = \sum_{i=0}^{2m} W_i^{(F)} \left[\widetilde{y}_{i,t+1} - \mu_{t+1}\right] \left[\widetilde{y}_{i,t+1} - \mu_{t+1}\right]' + R_{t+1}. \tag{2.48}$$

The weights, $W_i^{(\mu)}$ and $W_i^{(F)}$, for $i = 0, 2, ...2m$, associated with the conditional mean $E\left[y_{t+1}|y_{1:t}\right]$ and variance $Var\left(y_{t+1}|y_{1:t}\right)$, are determined by

$$
\begin{aligned}
W_0^{(\mu)} &= \frac{\widetilde{\vartheta}}{m+\widetilde{\vartheta}} \\
W_0^{(F)} &= \frac{\widetilde{\vartheta}}{m+\widetilde{\vartheta}} + \left(1-\widetilde{\varphi}^2-\widetilde{\varpi}\right) \\
W_i^{(\mu)} &= W_i^{(F)} = \frac{1}{2\left(m+\widetilde{\vartheta}\right)} \qquad \text{for } i=1,2,...,2m.
\end{aligned}
\tag{2.49}
$$

Here again each of the weight sets $\left\{W_i^{(\mu)}, i=1,2,...,2m\right\}$ and $\left\{W_i^{(F)}, i=1,2,...,2m\right\}$ must sum to one. Note that if the parameters satisfy $\widetilde{\vartheta}=\vartheta$, $\widetilde{\varphi}=\varphi$ and $\widetilde{\varpi}=\varpi$, then $W_i^{(\mu)}=W_i^{(a)}$ and $W_i^{(F)}=W_i^{(V)}$ for all $i=0,1,..,2m$. This is the case generally found in the literature (see, for example, Julier and Uhlmann, 2004, and Wan and van der Merwe, 2001); hence this restriction is imposed here.

(c) The predictive distribution $p\left(y_{t+1}|y_{1:t}\right)$ associated with the next observation is then approximated by a Gaussian density, with its mean and covariance matrix given by (2.47) and (2.48), respectively, with $W_i^{(\mu)}=W_i^{(a)}$ and $W_i^{(F)}=W_i^{(V)}$ for all $i=0,1,..,2m$.

3. Finally, with the observation of y_{t+1} revealed, the mean and covariance of the updated filtered distribution, $p\left(x_{t+1}|y_{1:t+1}\right)$, are respectively given by

$$a_{t+1|t+1} = a_{t+1|t} + M_{t+1}\left(y_{t+1}-\mu_{t+1}\right) \tag{2.50}$$

$$V_{t+1|t+1} = V_{t+1|t} - M_{t+1}F_{t+1}M'_{t+1}, \tag{2.51}$$

where

$$M_{t+1} = V_{t+1}^{xy} F_{t+1}^{-1},$$

with F_{t+1} as given in (2.48) and with V_{t+1}^{xy} denoting the cross-covariance matrix and defined as

$$V_{t+1}^{xy} = \sum_{i=0}^{2m} W_i^{(V)} \left[\chi_{i,t+1|t}^{*} - a_{t+1|t}\right] \left[\widetilde{y}_{i,t+1} - \mu_{t+1}\right]'.$$

The state filtered distribution, $p\left(x_{t+1}|y_{1:t+1}\right)$, is then approximated by a Gaussian density, with its mean $a_{t+1|t+1}$ and covariance matrix $V_{t+1|t+1}$ given by (2.50) and (2.51), respectively.

Finally, the out-of-sample one-step-ahead distribution is approximated by a Gaussian distribution, with its mean and variance-covariance given respectively by μ_{T+1} and F_{T+1} from Step 2 above. Note that the unscented Kalman filter is initiated with $a_{1|1}$ and $V_{1|1}$ computed using (2.50) and (2.51), with $x_1 \sim N\left(a_{1|0}, V_{1|0}\right)$ assumed known.

2.4.4 Grid-based Non-Gaussian Filter

The extended Kalman filter and the unscented Kalman filter both approximate the conditional densities in the recursive filtering and predictive algorithms as Gaussian densities, and therefore potentially miss out on other features of these distributions, such as skewness and leptokurtosis for example. In contrast, Kitagawa (1987) outlines a fully parametric grid-based non-Gaussian filter that realizes the recursive formulae for the predictive and filtered densities in (2.11) and (2.13), with a particular form of approximation. Specifically, the method is based on continuous piecewise linear approximations to the relevant integrands, with trapezoidal integration then used to produce the filtered and predictive state distributions. This grid-based filter involves having a set of grid points over the non-constant supports of the filtered and predictive distributions of the state variable. In this subsection, the generic form of the grid-based non-Gaussian filter is presented in the context of the context of scalar variables y_t and x_t.

Consider a state space model governed by the measurement density $p\left(y_t|x_t\right)$ and the transition density $p\left(x_{t+1}|x_t\right)$, where both x_t and y_t are scalar variables for each $t = 1, 2, ..., T$. The grid-based non-Gaussian filter is given by the following steps.

1. The state predictive distribution given by (2.11) is defined as an integral with respect to x_t. To evaluate the integral numerically, a finite set of grid points $\{x_t^i,\ i = 1, 2, ..., M\}$ is defined across the support of the distribution of x_t. The integral of the probability distribution over the (possibly infinite) support of the distribution of x_t, is then approximated as the sum of integrals over the M segments of the support defined by the choice of grid-points, yielding

$$\begin{aligned} p\left(x_{t+1}|y_{1:t}\right) &= \int_{-\infty}^{\infty} p\left(x_{t+1}|x_t\right) p\left(x_t|y_{1:t}\right)\ dx_t \\ &\approx \textstyle\sum_{i=1}^{M} \int_{x_t^{i-1}}^{x_t^i} p\left(x_{t+1}|x_t\right) p\left(x_t|y_{1:t}\right) dx_t. \end{aligned} \tag{2.52}$$

Each of the integrals in (2.52) is then evaluated via the trapezoidal rule, consistent with the integrand in each segment being approximated by a linear function. The resulting state predictive density $p\left(x_{t+1}|y_{1:t}\right)$ is then given by

$$\begin{aligned} p\left(x_{t+1}|y_{1:t}\right) &\approx \textstyle\sum_{i=1}^{M}\left\{p\left(x_{t+1}|x_{t}^{i-1}\right)p\left(x_{t}^{i-1}|y_{1:t}\right)\right. \\ &\qquad \left.+\,p\left(x_{t+1}|x_{t}^{i}\right)p\left(x_{t}^{i}|y_{1:t}\right)\right\}\left(x_{t}^{i}-x_{t}^{i-1}\right)/2. \qquad (2.53) \end{aligned}$$

It is noted that the set of grid points, $\{x_t^i, i = 1, 2, ..., M\}$, across the support of the distribution of x_t, has a t subscript to denote that the grid points must be modified to ensure efficient implementation of the numerical approximation. However, due to the computational problem associated with determining a set of grid-points for each distinct t, often the simpler, though less efficient, approach of using a large single set of common grid points for all t can be adopted, ensuring a very wide support for the (unknown) marginal distribution of x_t. The inefficiency of the algorithm is derived from the fact that many terms in (2.53) will have a negligible contribution to the numerical evaluation of the integral. (See Kitagawa, 1987).

2. Having produced $p\left(x_{t+1}|y_{1:t}\right)$, the one-step-ahead predictive distribution $p\left(y_{t+1}|y_{1:t}\right)$ is then approximated as the sum of integrals over the M segments of the support of the distribution of x_{t+1}, defined by the choice of grid-points, yielding

$$\begin{aligned} p\left(y_{t+1}|y_{1:t}\right) &= \int_{-\infty}^{\infty} p\left(y_{t+1}|x_{t+1}\right)p\left(x_{t+1}|y_{1:t}\right)\ dx_{t+1} \\ &\approx \textstyle\sum_{i=1}^{M}\displaystyle\int_{x_{t+1}^{i-1}}^{x_{t+1}^{i}} p\left(y_{t+1}|x_{t+1}\right)p\left(x_{t+1}|y_{1:t}\right)\ dx_{t+1}, \qquad (2.54) \end{aligned}$$

where $\left\{x_{t+1}^{i}, i=1,2,...,M\right\}$ are the grid points on the support. Using the trapezoidal rule again to evaluate each of the integrals in (2.54), the integrand within each integral is approximated by a linear function, resulting in

$$\begin{aligned} p\left(y_{t+1}|y_{1:t}\right) &\approx \textstyle\sum_{i=1}^{M}\left\{p\left(y_{t+1}|x_{t+1}^{i-1}\right)p\left(x_{t+1}^{i-1}|y_{1:t}\right)+\right. \\ &\left. p\left(y_{t+1}|x_{t+1}^{i}\right)p\left(x_{t+1}^{i}|y_{1:t}\right)\right\}\left(x_{t+1}^{i}-x_{t+1}^{i-1}\right)/2. \quad (2.55) \end{aligned}$$

The same issue as flagged above regarding the determination of the time-t dependent support applies here.

3. The filtered distribution of the state variable is up-dated upon the observation of y_{t+1} as

$$p\left(x_{t+1}|y_{1:t+1}\right)=\frac{p\left(y_{t+1}|x_{t+1}\right)p\left(x_{t+1}|y_{1:t}\right)}{p\left(y_{t+1}|y_{1:t}\right)},$$

where $p\left(y_{t+1}|x_{t+1}\right)$ is given by the model, $p\left(x_{t+1}|y_{1:t}\right)$ is the function defined in (2.53), and $p\left(y_{t+1}|y_{1:t}\right)$ is given by the sum in (2.55).

The out-of-sample one-step-ahead predictive distribution is subsequently approximated as

$$\begin{aligned} p\left(y_{T+1}|y_{1:T}\right) &\approx \textstyle\sum_{i=1}^{M}\left\{p\left(y_{T+1}|x_{T+1}^{i-1}\right)p\left(x_{T+1}^{i-1}|y_{1:T}\right)+\right. \\ &\left. p\left(y_{T+1}|x_{T+1}^{i}\right)p\left(x_{T+1}^{i}|y_{1:T}\right)\right\}\left(x_{T+1}^{i}-x_{T+1}^{i-1}\right)/2, \quad (2.56) \end{aligned}$$

with $\left\{x_{T+1}^{i}, i=1,2,...,M\right\}$ being the set of grid points on the support of the distribution of x_{T+1}.

2.4.5 Gaussian-sum Filter

The grid-based non-Gaussian filter in Section 2.4.4 is demonstrated for a univariate state space model. If the model were to have a multivariate state variable, the grid-based non-Gaussian filter would require the numerical evaluation of multiple integrals for the production of all relevant densities. The application of this filter is thus impractical in the case of a very high dimensional state, a problem that is exacerbated by the need to choose a number of grid points M that is sufficiently large to cover the time-varying supports of all x_t.

One potential solution to this computational burden, as suggested by Kitagawa (1989, 1994), is to use a Gaussian-sum filter (Sorenson and Alspach, 1971; Alspach and Sorenson, 1972; Anderson and Moore, 1979). The Gaussian-sum approach approximates each of the densities in the filtering process with a finite mixture of Gaussian densities, with such a mixture called a Gaussian sum. In this approach, any non-Gaussian measurement density $p\left(y_t|x_t\right)$ in (2.1) or state transition density $p\left(x_{t+1}|x_t\right)$ in (2.2) is approximated by an appropriately chosen Gaussian sum whose components all have means that are linear in the state variable, x_t, as well as variances that are constant. Owing to the linearity and Gaussianity of each component within these approximations, each of the filtered and predictive state densities is shown to result in a further Gaussian sum, with components whose means and variances are easily computed via repeated application of the Kalman filter. Although feasible for small sample sizes, as the number of terms in the Gaussian-sum representing each filtered state density grows exponentially with t, pruning algorithms are required for larger samples in order to keep the number of terms in the successive Gaussian sums manageable.

Monteiro (2010) has recently suggested that the Gaussian-sum filter can also be used as a flexible modelling approach, enabling a relaxation of the required distributional assumption for the measurement error in a linear state space model. The Gaussian-sum filter is illustrated with this particular context in mind.

Consider a *univariate* version of the linear state space model in (2.14) and (2.15), reproduced here for convenience,

$$y_t = c_t + H_t x_t + \eta_t \tag{2.57}$$

$$x_{t+1} = d_t + K_t x_t + v_t \tag{2.58}$$

$$v_t \sim N(0, Q_t) \tag{2.59}$$

for $t = 1, 2, ..., T$. In this model, both y_t and x_t are scalar variables for each $t = 1, 2, ..., T$, so that the expressions c_t, d_t, H_t, K_t and Q_t are now all scalars also. A parametric distribution is not specified for the measurement error η_t. Rather, Monteiro (2010) assumes that the measurement error density is a Gaussian sum given by the particular form

$$p(\eta_t) = \frac{1}{(T+1)} \sum_{t=0}^{T} \phi\left(\eta_t; \widehat{\eta}_t, b^2\right), \tag{2.60}$$

where the notation $\phi(x; \mu, \sigma^2)$ represents the Gaussian probability density for the variable x associated with mean μ and variance σ^2. As can be seen from (2.60), the density of the measurement error term in (2.57) is an equally weighted average of $T+1$ separate Gaussian kernels, one for each time point being centred on an estimate of the corresponding true (unobserved) error η_t, $\widehat{\eta}_t$. The constant scale parameter $b > 0$ associated with the Gaussian kernel is the so-called bandwidth of the kernel, and is treated as an unknown parameter to be estimated. The additional constraint $\widehat{\eta}_0 = -\sum \widehat{\eta}_t$ is imposed to ensure identifiability of the model. Following from (2.57) and (2.60), the measurement density for y_t given x_t is also a Gaussian sum, with

$$p\left(y_t|x_t\right) = \frac{1}{(T+1)} \sum_{t=0}^{T} \phi\left(y_t; c_t + H_t x_t + \widehat{\eta}_t, b^2\right). \qquad (2.61)$$

In the spirit of Sorenson and Alspach (1971), Monteiro (2010) shows that at any time step t, the structure of the filtered state density may be written as

$$p\left(x_t|y_{1:t}\right) = \sum_{l=0}^{L_t} \varsigma_t^{(l)} \phi\left(x_t; a_{t|t}^{(l)}, V_{t|t}^{(l)}\right), \qquad (2.62)$$

where $L_t = (T+1)^t - 1$, and where each of the $L_t + 1$ Gaussian components in $p\left(x_t|y_{1:t}\right)$ is associated with a mean $a_{t|t}^{(l)}$ and a variance $V_{t|t}^{(l)}$, each computed using the Kalman filter equations in (2.18) adapted for the univariate setting. Starting with the previously filtered state density in (2.62), one iteration of the Gaussian-sum filter is given by the following three steps:

1. Using the state filtered density in (2.62), the state predictive distribution

is

$$\begin{aligned} p\left(x_{t+1}|y_{1:t}\right) &= \int p\left(x_{t+1}|x_t\right) p\left(x_t|y_{1:t}\right)\ dx_t \\ &= \int \phi\left(x_{t+1}; d_t + K_t x_t, Q_t\right) \sum_{l=0}^{L_t} \varsigma_t^{(l)} \phi\left(x_t; a_{t|t}^{(l)}, V_{t|t}^{(l)}\right)\ dx_t \\ &= \sum_{l=0}^{L_t} \varsigma_t^{(l)} \int \phi\left(x_{t+1}; d_t + K_t x_t, Q_t\right) \phi\left(x_t; a_{t|t}^{(l)}, V_{t|t}^{(l)}\right)\ dx_t \\ &= \sum_{l=0}^{L_t} \varsigma_t^{(l)} \phi\left(x_{t+1}; a_{t+1|t}^{(l)}, V_{t+1|t}^{(l)}\right), \qquad (2.63) \end{aligned}$$

with the mean $a_{t+1|t}^{(l)}$ and variance $V_{t+1|t}^{(l)}$ of the l^{th} Gaussian component computed using (2.18a) and (2.18b) respectively, and given by

$$\begin{aligned} a_{t+1|t}^{(l)} &= d_t + K_t a_{t|t}^{(l)} \\ V_{t+1|t}^{(l)} &= K_t^2 V_{t|t}^{(l)} + Q_t. \end{aligned}$$

2. To produce the one-step-ahead predictive density $p\left(y_{t+1}|y_{1:t}\right)$, an expression for $p\left(y_{t+1}|x_{t+1}\right)$ is required. Although the specific form is given in (2.61), to ensure clarity and completeness in the presentation, the conditional density for the measurement y_{t+1} given the state x_{t+1} is denoted using the general form

$$p\left(y_{t+1}|x_{t+1}\right) = \sum_{j=0}^{J} \alpha_j \phi\left(y_{t+1}; c_{t+1}^j + H_{t+1}^j x_{t+1}, R_{t+1}^j\right), \qquad (2.64)$$

so that the Gaussian-sum is indexed with $j = 0, 1, ..., J$, with the j^{th} Gaussian component having a mean linear in the relevant state variable x_{t+1} and variance R_{t+1}^j free from x_{t+1}. In the setting of Monteiro (2010) given in (2.57)-(2.59) and (2.61), then $J = T$, $\alpha_j = \frac{1}{(T+1)}$, $c_{t+1}^j = \left(c_{t+1} + \widehat{\eta}_j\right)$,

$H^j_{t+1} = H_{t+1}$ and $R^j_{t+1} = b^2$ for each $t = 1, 2, ..., T-1$. Given the desired general representation of the one-step-ahead predictive distribution in (2.12), and referring to the mean and variance respectively of the one-step-ahead predictive density for the next observation given in equations (2.19) and (2.20), the one-step-ahead predictive density for y_{t+1} becomes

$$\begin{aligned} p\left(y_{t+1}|y_{1:t}\right) &= \int p\left(y_{t+1}|x_{t+1}\right) p\left(x_{t+1}|y_{1:t}\right) dx_{t+1} \\ &= \int \left\{ \sum_{j=0}^{J} \alpha_j \phi\left(y_{t+1}; c^j_{t+1} + H^j_{t+1} x_{t+1}, R^j_{t+1}\right) \right. \\ &\quad \left. \sum_{l=0}^{L_t} \varsigma^{(l)}_t \phi\left(x_{t+1}; a^{(l)}_{t+1|t}, V^{(l)}_{t+1|t}\right) \right\} dx_{t+1} \\ &= \sum_{l=0}^{L_t} \sum_{j=0}^{J} \varsigma^{(l)}_t \alpha_j \phi\left(y_{t+1}; \mu^{(l,j)}_{t+1}, F^{(l,j)}_{t+1}\right), \end{aligned} \tag{2.65}$$

with

$$\begin{aligned} \mu^{(l,j)}_{t+1} &= c^j_{t+1} + H^j_{t+1} a^{(l)}_{t+1|t} \\ F^{(l,j)}_{t+1} &= H^2_{t+1} V^{(l)}_{t+1|t} + R^j_{t+1} \end{aligned}$$

for $j = 0, 1, ..., J$ and $l = 0, 1, ..., L_t$.

3. Normalizing the product of (2.63) and (2.64), the updated filtered density is

$$\begin{aligned} p\left(x_{t+1}|y_{1:t+1}\right) &\propto \left\{ \sum_{j=0}^{J} \alpha_j \phi\left(y_{t+1}; c^j_{t+1} + H^j_{t+1} x_{t+1}, R^j_{t+1}\right) \right. \\ &\quad \left. \sum_{l=0}^{L_t} \varsigma^{(l)}_t \phi\left(x_{t+1}; a^{(l)}_{t+1|t}, V^{(l)}_{t+1|t}\right) \right\} \\ &= \sum_{l=0}^{L_t} \sum_{j=0}^{J} \varsigma^{(l,j)}_{t+1} \phi\left(x_{t+1}; a^{(l,j)}_{t+1|t+1}, V^{(l,j)}_{t+1|t+1}\right). \end{aligned} \tag{2.66}$$

with

$$\begin{aligned} a^{(l,j)}_{t+1|t+1} &= a^{(l)}_{t+1|t} + M^{(l,j)}_{t+1}\varepsilon^{(l,j)}_{t+1} \\ V^{(l,j)}_{t+1|t+1} &= \left(1 - M^{(l,j)}_{t+1}H_{t+1}\right)V^{(l)}_{t+1|t}, \end{aligned}$$

and where

$$\begin{aligned} M^{(l,j)}_{t+1} &= V^{(l)}_{t+1|t}H^{j}_{t+1}\left(F^{(l,j)}_{t+1}\right)^{-1} \\ F^{(l,j)}_{t+1} &= \left(H^{j}_{t+1}\right)^{2}V^{(l)}_{t+1|t} + R^{j}_{t+1} \\ \varepsilon^{(l,j)}_{t+1} &= y_{t+1} - c^{j}_{t+1} - H^{j}_{t+1}a^{(l)}_{t+1|t} \end{aligned}$$

and

$$\varsigma^{(l,j)}_{t+1} = \frac{\varsigma^{(l)}_{t}\alpha_j\phi\left(y_{t+1};\mu^{(l,j)}_{t+1|t},F^{(l,j)}_{t+1}\right)}{\sum\limits_{l=0}^{L}\sum\limits_{j=0}^{J}\varsigma^{(l)}_{t}\alpha_j\phi\left(y_{t+1};\mu^{(l,j)}_{t+1|t},F^{(l,l)}_{t+1}\right)},$$

for $j = 0,1,...,J$ and $l = 0,1,...,L_t$. That is, given that the one-step-ahead state prediction density given by (2.63) is a mixture of Gaussian densities, and that the measurement error density in (2.64) is also a mixture of Gaussian densities, then the updated state filtered density is the mixture of $(L_{t+1}+1) = (L_t+1)(J+1)$ Gaussian densities given in (2.66), with weights determined by $\varsigma^{(l,j)}_{t+1}$ and, the $(l,j)^{th}$ component mean $a^{(l,j)}_{t+1|t+1}$ and variance $V^{(l)}_{t+1|t+1}$ obtained from the appropriate Kalman filter equations. Hence, replacing the double index (l,j) in (2.66) with a suitable single index, the revised filtered density can be written in the required form,

$$p\left(x_{t+1}|y_{1:t+1}\right) = \sum_{l=0}^{L_{t+1}}\varsigma^{(l)}_{t+1}\phi\left(x_{t+1};a^{(l)}_{t+1|t+1},V^{(l)}_{t+1|t+1}\right).$$

A couple of comments are in order. First, note that the Gaussian-sum filter is initialized by

$$p(x_1) = \phi\left(x_1; a_{1|0}, V_{1|0}\right), \tag{2.67}$$

with $a_{1|0}$ and $V_{1|0}$ assumed known.[3] Therefore, comparing with (2.62) it can be seen that $(L_0 + 1) = 1$. In updating the filtered distribution from (2.62) to (2.66), the number of Gaussian mixtures has increased from $(L_t + 1)$ to $(L_{t+1} + 1) = (L_t + 1)(J + 1)$, and since $J = T$ in the Monteiro (2010) setting, the increase is from $(T + 1)^t$ to $(T + 1)^{t+1}$. Hence, it can now be seen that the number of mixture components in the filtered distributions indeed increases geometrically in T. To reduce the computational burden, Monteiro suggests using standard clustering algorithms to combine Gaussian components with similar mean and variance values, thereby keeping the Gaussian-sum structure but with the number of terms in the filtered state density at each iteration constrained to be small - of the order $L_t = \sqrt{T+1}$ for all $t = 1, 2, ...T$. Although this approach reduces the number of elements in each Gaussian sum, it does not remove the need for significant computational resources for the procedure to be implemented, particularly for large values of T, as a large clustering algorithm must be implemented at each iteration, and the geometric dependence on T still obtains.

Second, as the Gaussian-sum filter algorithm presented here requires esti-

[3] In fact, Monteiro applies a diffuse initial distribution on x_1, in the spirit of Ansley and Kohn (1985).

mates, $\widehat{\eta}_j$, of the unobserved measurement error, η_j, for each[4] $j = 1, 2, ..., T$, Monteiro suggests obtaining these from a Kalman smoothing procedure, such as that given by Kitagawa (1994). However, as these smoothed $\{\widehat{\eta}_t\}$ are obtained conditional on the ML estimate (say) of any unknown parameter vector and conversely, calculation of the likelihood function requires the $\{\widehat{\eta}_t\}$, Monteiro iterates the two procedures and thereby runs the Gaussian-sum filter many times until apparent convergence is achieved. It is the estimation of the bandwidth parameter b, in conjunction with the estimated measurement errors, $\widehat{\eta}_t$, that provides the non-parametric description to the measurement error, via the Gaussian sum in (2.60). However, the iteration required to produce these estimates, coupled with the fact that each iteration of the Gaussian-sum filter requires a clustering algorithm to manage the exponential growth in the number of mixture elements, means that the overall method is very computationally demanding, particularly for large sample sizes.

2.4.6 Particle Filters

The particle filter has become a very popular class of numerical method for conducting inference in non-linear, non-Gaussian state space models. Reviews of the topic include Doucet, de Freitas and Gordon (2001) and Cappe, Godsill and Moulines (2007). Particle filters are a form of simulation filter that approximate the conditional distributions, $p\left(x_t|y_{1:t}\right)$ and $p\left(x_{t+1}|y_{1:t}\right)$, with empirical distributions arising from a set of draws (or 'particles') obtained via recursive Monte

[4] Recall that the constraint $\widehat{\eta}_0 = -\sum \widehat{\eta}_t$ is imposed.

Carlo methods.

In this subsection, three alternative particle filters are reviewed. The first is the simplest version of the particle filter, referred to as the bootstrap filter, independently proposed by Gordon, Salmond and Smith (1993) and Kitagawa (1996). The bootstrap filter provides a straightforward introduction to more elaborate particle filtering methods. The second filter reviewed is the generic auxiliary particle filter, designed to reduce the degeneracy problem that the bootstrap filter may encounter. We then review a third method, referred to as the fully adapted auxiliary particle filter.

Bootstrap Filter

Given a set of random draws, or 'particles', $\{x_t^i,\ i = 1, 2, ..., P\}$ from $p\left(x_t|y_{1:t}\right)$, the bootstrap filter is an algorithm that propagates and updates these draws, to obtain a set of particles $\left\{x_{t+1}^i,\ i = 1, 2, ..., P\right\}$ from $p\left(x_{t+1}|y_{1:t+1}\right)$. The latter distribution is then approximated by the empirical distribution of a finite sample of particles, with the accuracy of the approximation increasing as the number of particles increases, due to the strong law of large numbers.

To initialize the bootstrap filter, an intial set of particles, $\{\widetilde{x}_1^i,\ i = 1, 2, ..., P\}$, is drawn independently from $p\left(x_1\right)$. Then, having observed y_1, the particles are resampled with replacement according to the probability mass function (pmf), $w_1^i = \frac{p\left(y_1|\widetilde{x}_1^i\right)}{\sum_{j=1}^{P} p\left(y_1|\widetilde{x}_1^j\right)}$, for $i = 1, 2, ..., P$, thereby producing a set of particles, $\{x_1^i,\ i = 1, 2, ..., P\}$, from $p\left(x_1|y_1\right)$. The bootstrap filter algorithm is then given by the following steps for each $t = 1, 2..., T$:

1. Draw $\widetilde{x}_{t+1}^{i}$ for $i = 1, 2, ..., P$ from $p(x_{t+1}|x_t^i)$. These particles $\widetilde{x}_{t+1}^{i}$ for $i = 1, 2, ..., P$ represent a sample from the state prediction distribution, $p(x_{t+1}|y_{1:t})$.

2. Approximate the one-step-ahead predictive density, $p(y_{t+1}|y_{1:t})$, by the average of the conditional distributions for the measurement, as

$$\begin{aligned} p(y_{t+1}|y_{1:t}) &= \int p(y_{t+1}|x_{t+1})\, p(x_{t+1}|y_{1:t})\ dx_{t+1} \\ &\approx \frac{1}{P}\sum_{i=1}^{P} p\left(y_{t+1}|\tilde{x}_{t+1}^{i}\right), \end{aligned} \tag{2.68}$$

where $\left\{\tilde{x}_{t+1}^{i},\ i = 1, 2, ..., P\right\}$ represent the draws of the predictive state from $p(x_{t+1}|y_{1:t})$ in Step 1.

3. Given the measurement y_{t+1}, construct a resampling pmf by assigning to each $\widetilde{x}_{t+1}^{i}$ a normalized weight

$$w_{t+1}^{i} = \frac{p\left(y_{t+1}|\widetilde{x}_{t+1}^{i}\right)}{\sum_{j=1}^{P} p\left(y_{t+1}|\widetilde{x}_{t+1}^{j}\right)}, \tag{2.69}$$

for $i = 1, 2, ..., P$. Resample x_{t+1}^{i} from the discrete distribution having pmf given by (2.69) for each $i = 1, 2, ..., P$. This completes the updating step, as the particles $\left\{x_{t+1}^{i},\ i = 1, 2, ..., P\right\}$ are used to approximate the density, $p(x_{t+1}|y_{1:t+1})$.

As can be seen, the only requirements for the bootstrap filter are that $p(x_1)$ and $p(x_{t+1}|x_t)$ are available for sampling, and that the conditional density $p(y_t|x_t)$ can be evaluated. However, one of the problems with the bootstrap filter is that the predictive states are drawn from $p(x_{t+1}|x_t^i)$ without accounting for the next time period's observation, y_{t+1}. This can lead to degeneracy, because many of the resulting weights for the predicted particles $\widetilde{x}_{t+1}^{i}$ in (2.69) become very small, so that many of the potential particles are assigned negligible (or no) weight. This problem is particularly prevalent if y_{t+1} is an outlier, or if the filtered density is diffuse relative to the likelihood component. For example, if y_{t+1} is exceptionally large, and given that the bootstrap filter simulates $\widetilde{x}_{t+1}^{i}$ without accounting for y_{t+1}, there are potentially many draws of $\widetilde{x}_{t+1}^{i}$ that are not in the high probability region associated with $p\left(y_{t+1}|x_{t+1}\right)$. The weights in (2.69) are thus unevenly distributed due to the large variation in $\left\{p\left(y_{t+1}|\widetilde{x}_{t+1}^{i}\right),\ i=1,2,...,P\right\}$. Pitt and Shephard (1999) address this problem by introducing the auxiliary particle filter.

On a related note, despite the fact that the resampling weights in (2.69) imply a filtered state density given by

$$\sum_{i=1}^{P} w_{t+1}^{i}\delta\left(x_{t+1}-x_{t+1}^{i}\right),$$

where $\delta\left(.\right)$ denotes the Dirac delta mass at point x_{t+1}^{i}, the actual form of the filtered density that feeds into the subsequent iteration of the filter is

$$\sum_{i=1}^{P} \frac{P_i}{P}\delta\left(x_{t+1}-x_{t+1}^{i}\right),$$

where P_i denotes the number of replicated values of x_{t+1}^{i} arising from the P resampled particles, so that $\sum_{i=1}^{P} P_i = P$.[5] When $w_{t+1}^{i} > 0$ is relatively small, it is nevertheless likely that the realized P_i will actually be zero. Hence, an

[5] The Dirac delta function $\delta\left(x_{t+1}-x_{t+1}^{i}\right)$ can be thought of as a very tall and thin spike with unit area located at the point x_{t+1}^{i}, corresponding to a degenerate distribution at $x_{t+1}=x_{t+1}^{i}$. The formal properties of the Dirac delta function are introduced later in Chapter 3.

added source of degeneracy is due to this type of Monte Carlo error. That is, even if a particular weight w_{t+1}^{i} is non-zero, there is a positive probability that the particle will not be propagated through to any of the future time periods, because if $P_i = 0$, then subsequent particle draws for x_s, $s > t+1$ will not include any having been conditioned upon $x_{t+1} = x_{t+1}^{i}$. This particular aspect is present in all of the particle filters considered in this thesis, apart from the new particle filter described in Section 6.6 of Chapter 6.

Auxiliary Particle Filter

The auxiliary particle filter (APF), introduced by Pitt and Shephard (1999), is a variant of the bootstrap filter. The APF ensures that the predicted particles, through the use of an auxiliary variable denoted by the index k, are more likely to match up with the observed data y_{t+1}, with more even weights for the predicted particles, $\widetilde{x}_{t+1}^{i}$, produced as a consequence. The APF operates by obtaining a sample of draws from the joint distribution $p\left(x_{t+1}, k|y_{1:t+1}\right)$ and then omitting the index k in the pair $\left(x_{t+1}, k\right)$ to produce a sample of particles $\left\{x_{t+1}^{i},\ i = 1, 2, ..., P\right\}$ from the marginalized filtered density $p\left(x_{t+1}|y_{1:t+1}\right)$. Crucially, the APF relies on the use of an importance density, $q\left(x_{t+1}, k|y_{1:t+1}\right)$, to draw the sample $\left(x_{t+1}^{i}, k^{i}\right)$ for $i = 1, 2, ..., P$.

In the generic APF setting (see Pitt and Shephard, 1999, or Arulampalam, Maskell, Gordon and Clapp, 2002), the importance density is defined to satisfy

$$q\left(x_{t+1}, k|y_{1:t+1}\right) \propto p\left(y_{t+1}|\mu_{t+1}^{k}\right) p\left(x_{t+1}|x_{t}^{k}\right), \tag{2.70}$$

where μ_{t+1}^{k} is some likely value associated with the density $p\left(x_{t+1}|x_{t}^{k}\right)$, and x_{t}^{k} is the k^{th} particle from the set of particles $\{x_t^i,\ i = 1, 2, ..., P\}$ that approximates the previous filtered distribution, $p(x_t|y_{1:t})$. The term μ_{t+1}^k could be, for example, the conditional mean $E\left(x_{t+1}|x_t^k\right)$, or a sampled value from $p\left(x_{t+1}|x_t^k\right)$.

The importance density in (2.70) is, in turn, decomposed into the product of a marginal and a conditional density, corresponding to the decomposition in (2.70),

$$q(x_{t+1}, k|y_{1:t+1}) = q(k|y_{1:t+1})\, q(x_{t+1}|k, y_{1:t+1}), \tag{2.71}$$

with

$$q(k|y_{1:t+1}) \propto p\left(y_{t+1}|\mu_{t+1}^k\right) \tag{2.72}$$

and

$$q(x_{t+1}|k, y_{1:t+1}) = p\left(x_{t+1}|x_t^k\right). \tag{2.73}$$

Therefore, a *single* pair of (x_{t+1}, k), denoted as $\left(x_{t+1}^i, k^i\right)$, can be sampled from the importance density in (2.71) in two preliminary steps:

1. With reference to (2.72), simulate the index k^i with probablity $\lambda_k \propto p\left(y_{t+1}|\mu_{t+1}^k\right)$, for each $k = 1, 2, ..., P$, with λ_k referred to as the first-stage weights.

2. Then, with reference to (2.73), draw x_{t+1}^i from the transition density $p\left(x_{t+1}|x_t^{k^i}\right)$. Unlike the bootstrap filter, where predictive states are drawn without accounting for y_{t+1}, the construction of the first-stage weights are informed by y_{t+1}, thereby allowing particles that give rise to high likelihoods, $p(y_{t+1}|x_{t+1})$, to be produced in this step.

The above two steps are repeated P times to generate $\left\{x_{t+1}^i, k^i,\, i = 1, 2, ..., P\right\}$, after which a re-weighting step is performed by assigning a second-stage weight, w_{t+1}^i, to the sample pair $\left\{x_{t+1}^i, k^i\right\}$, with

$$w_{t+1}^i \propto \frac{p\left(y_{t+1}|x_{t+1}^i\right)}{p\left(y_{t+1}|\mu_{t+1}^{k^i}\right)}. \tag{2.74}$$

The second-stage weights are likely to be less variable than those in (2.69), reducing the degeneracy problem associated with the bootstrap filter, since the initial particles are generated in Step 2 above using information from the data value y_{t+1} implied through the index in Step 1. A sample of x_{t+1} from $p\left(x_{t+1}|y_{1:t+1}\right)$ can then be produced from the discrete distribution defined by the weights in (2.74).

In summary, after initializing $\{x_1^i,\ i = 1, 2, ..., P\}$ as independent draws from $p\left(x_1\right)$, the generic APF algorithm is given by the following steps for each $t = 1, 2, ..., T$:

1. Obtain a sample of draws, $\left\{\widetilde{x}_{t+1}^i,\ i = 1, 2, ..., P\right\}$, that represent a sample from the state prediction distribution, $p\left(x_{t+1}|y_{1:t}\right)$, via the following steps:

   (a) Calculate μ_{t+1}^i for $i = 1, 2, ..., P$, where μ_{t+1}^i is a likely value of x_{t+1} from $p\left(x_{t+1}|x_t^i\right)$.

   (b) With reference to (2.72), construct a pmf for the auxiliary variable, k^i, by assigning the first-stage weights,

$$\lambda_{t+1}^i = \frac{p\left(y_{t+1}|\mu_{t+1}^i\right)}{\sum_{j=1}^P p\left(y_{t+1}|\mu_{t+1}^j\right)}, \tag{2.75}$$

   for $i = 1, 2, ..., P$.

(c) Sample the k^i from the discrete distribution, with pmf in (2.75), for $i = 1, 2, ..., P$.

(d) Sample $\widetilde{x}^i_{t+1}$ from the importance density $q\left(x_{t+1}|k^i, y_{1:t+1}\right) = p\left(x_{t+1}|x_t^{k^i}\right)$, for $i = 1, 2, ..., P$. These particles then represent a sample from $p\left(x_{t+1}|y_{1:t}\right)$.

2. Then approximate the one-step-ahead predictive density, $p\left(y_{t+1}|y_{1:t}\right)$, by the average of the conditional distributions for the measurement, as

$$\begin{aligned} p\left(y_{t+1}|y_{1:t}\right) &= \int p\left(y_{t+1}|x_{t+1}\right) p\left(x_{t+1}|y_{1:t}\right)\ dx_{t+1} \\ &\approx \frac{1}{P}\sum_{i=1}^{P} p\left(y_{t+1}|\tilde{x}^i_{t+1}\right), \end{aligned} \tag{2.76}$$

where $\left\{\tilde{x}^i_{t+1},\ i = 1, 2, ..., P\right\}$ represent the draws of the predictive state from $p\left(x_{t+1}|y_{1:t}\right)$ in Step 1.

3. Construct a resampling pmf by assigning to each $\widetilde{x}^i_{t+1}$, for $i = 1, 2, ..., P$, a second-stage normalized weight

$$w^i_{t+1} = \frac{\widetilde{w}^i_{t+1}}{\sum_{j=1}^{M} \widetilde{w}^j_{t+1}}, \tag{2.77}$$

where $\widetilde{w}^i_{t+1} = \frac{p\left(y_{t+1}|\tilde{x}^i_{t+1}\right)}{p\left(y_{t+1}|\mu^{k^i}_{t+1}\right)}$, for $i = 1, 2, ..., P$. Resample x^i_{t+1} from the discrete distribution having pmf given by (2.77) for $i = 1, 2, ..., P$. This completes the updating step, as the particles $\left\{x^i_{t+1},\ i = 1, 2, ..., P\right\}$ approximate the density, $p\left(x_{t+1}|y_{1:t+1}\right)$.

Fully Adapted Auxiliary Particle Filter

The generic APF above can usually reduce the variability of the second-stage weights in (2.77) relative to the resampling weights associated with the bootstrap filter; see Arulampalam *et al.* (2002). However, further improvement can be made over the generic scheme by tailoring the approach to the structure of the model under consideration. This approach is referred to as the fully adapted APF. The subsequent modifications to the APF involve changing Step 1(d) and the calculation of $\widetilde{w}_{t+1}^{i}$ in Step 3 in the algorithm above to:

- Sample $\widetilde{x}_{t+1}^{i}$ from the importance density $q\left(x_{t+1}|x_t^{k^i}, y_{t+1}, \mu_{t+1}^{k^i}\right)$, where the form of the density $q\left(x_{t+1}|\cdot\right)$ is dependent on the specific structure of the model considered.

- Calculate $\widetilde{w}_{t+1}^{i}$ for $i = 1, 2, ..., P$, using

  $$\widetilde{w}_{t+1}^{i} = \frac{p\left(y_{t+1}|\tilde{x}_{t+1}^{i}\right)}{q\left(y_{t+1}|x_{t+1}^{i}, \mu_{t+1}^{k^i}\right)},$$

  where the form of $q\left(y_{t+1}|\cdot\right)$ is dependent on the specific structure of the model considered.

Pitt and Shephard (1999) demonstrate how to implement this adaptation of the APF procedure in the case where the conditional measurement distribution $p\left(y_t|x_t\right)$ is log-concave and the state equation is linear and Gaussian. The density $q\left(y_{t+1}|x_{t+1}, \mu_{t+1}\right)$ can be obtained by taking a first-order Taylor series expansion of $\ln p\left(y_t|x_t\right)$ around μ_{t+1}, giving the approximation

$$\ln q\left(y_{t+1}|x_{t+1},\mu_{t+1}\right) \approx \ln p\left(y_{t+1}|\mu_{t+1}\right) + \left(x_{t+1}-\mu_{t+1}\right) \times \frac{\partial \ln p\left(y_{t+1}|\mu_{t+1}\right)}{\partial \mu_{t+1}}. \tag{2.78}$$

Given the assumed linear, Gaussian form of $p\left(x_{t+1}|x_t\right)$, the normal candidate prediction density,

$$q\left(x_{t+1}|x_t,y_{t+1},\mu_{t+1}\right) \propto q\left(y_{t+1}|x_{t+1},\mu_{t+1}\right) \times p\left(x_{t+1}|x_t\right), \tag{2.79}$$

is readily obtained via the product of the two normal kernels associated with $q\left(y_{t+1}|x_{t+1},\mu_{t+1}\right)$ and $p\left(x_{t+1}|x_t\right)$. Pitt and Shephard apply the adapted method to various models, including a stochastic volatility model and a model in which $p\left(y_t|x_t\right)$ is a discrete mixture of normals.

2.5 Maximum Likelihood Estimation of θ and $p\left(y_{T+1}|y_{1:T}\right)$

The various filtering algorithms presented in Section 2.4, used to produce the one-step-ahead predictive distribution $p\left(y_{t+1}|y_{1:t}\right)$ and, ultimately, the out-of-sample one-step-ahead predictive distribution, $p\left(y_{T+1}|y_{1:T}\right)$, are conditioned on a vector of parameters θ, that is assumed to be known. This section relaxes this assumption and assumes that θ is unknown, and has to be estimated using the sample observations, $y_{1:T}$.

As mentioned in Section 2.3.4, θ can be estimated via an ML approach by maximizing the logarithm of the likelihood function in (2.10) with respect to θ, to obtain the ML estimate, $\widehat{\theta}$. The (log) likelihood function only requires the availability of the marginal distribution $p(y_1)$, and the one-step-ahead predictive distributions, $p(y_{t+1}|y_{1:t})$ for $t = 1, 2, ...T - 1$, with the representations of these distributions having already been produced by the various filters presented in Section 2.4. However, as already highlighted, apart from the linear and Gaussian model, where $p(y_{t+1}|y_{1:t})$ has an exact analytical solution produced by the Kalman filter, the other filters applied to non-linear, non-Gaussian models produce approximations to $p(y_{t+1}|y_{1:t})$ only.[6] Hence, in the linear, Gaussian case, the likelihood function produced via the Kalman filter is exact, whilst in general non-linear, non-Gaussian models the other filters outlined above can be used to produce an approximation to the likelihood function only.

The (exact) log-likelihood function for the linear, Gaussian model in (2.14)-(2.17) is given by

$$\ln L(\theta|y) = -\frac{Tp}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=0}^{T-1}\ln|F_{t+1}| - \frac{1}{2}\sum_{t=0}^{T-1}\varepsilon_{t+1}' F_{t+1}^{-1}\varepsilon_{t+1},$$

where

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

and μ_{t+1} and F_{t+1} are the means and variance-covariance of the predictive distribution, given by (2.19) and (2.20) respectively. If certain regularity conditions are satisfied, then the ML estimate, $\widehat{\theta}$, based on a sample size T, is consistent and asymptotically normal. This result is discussed by Hamilton (1994) and in his discussion Hamilton gives a number of references to theoretical work on the

[6] Note that an exception to this statement applies if the state is discrete on a finite support. In this case all relevant integrals are defined exactly as finite sums. See Arulampalam *et al.* (2002) for an illustration.

subject. If the true distributions for the errors are non-Gaussian (i.e. model is misspecified), the estimation method is referred to as quasi-ML estimation.

The log-likelihood functions approximated by the extended Kalman filter for the non-linear model in (2.22) and (2.23), and by the unscented Kalman filter for the non-linear model in (2.36)-(2.38) have the same representation, given by

$$\ln L\left(\theta|y\right) \approx -\frac{Tp}{2}\ln\left(2\pi\right) - \frac{1}{2}\sum_{t=1}^{T-1}\ln\left|F_{t+1}\right| - \frac{1}{2}\sum_{t=1}^{T-1}\varepsilon_{t+1}' F_{t+1}^{-1}\varepsilon_{t+1},$$

where

$$\varepsilon_{t+1} = y_{t+1} - \mu_{t+1}.$$

However, here the mean and variance-covariance of the predictive distribution for the observation at time t, μ_{t+1} and F_{t+1}, are given by (2.34) and (2.35) respectively in the case of the extended Kalman filter, and by (2.47) and (2.48) respectively for the unscented Kalman filter case. Once again, this log-likelihood function has been specified under the assumption of Gaussianity for the error terms in the state space model as in (2.16) and with the initial state distribution given by (2.17). If the true distributions for the errors are non-Gaussian, an additional element of approximation error is involved in the specification of the likelihood function, i.e. over and above the error involved in the approximation of the relevant first and second moments.

The log-likelihood functions approximated by the grid-based non-Gaussian filter, Gaussian-sum filter and particle filters, are given by

$$\ln L\left(\theta|y\right) \approx \textstyle\sum_{t=0}^{T-1}\ln p\left(y_{t+1}|y_{1:t}\right),$$

with $p\left(y_{t+1}|y_{1:t}\right)$ defined in (2.55) for the grid-based non-Gaussian filter, (2.65) for the Gaussian-sum filter, (2.68) for the bootstrap filter and (2.76) for the auxiliary particle filter. In particular, see Pitt (2002) for discussion on the use of particle filters to estimate a likelihood function, and the associated properties of the estimator.

Finally, upon estimation of $\widehat{\theta}$, the ML estimate of the out-of-sample one-step-ahead predictive distribution is produced by conditioning the one-step-ahead forecast distribution at time T, on the estimated parameter vector $\widehat{\theta}$,

$$\widehat{p}\left(y_{T+1}|y_{1:T}\right) = p\left(y_{T+1}|y_{1:T}, \widehat{\theta}\right). \tag{2.80}$$

2.6 Limitations of the Parametric Forecasting Approach

In this chapter some of the existing filters that can be used to produce approximations of predictive distributions in state space models have been presented. In this section, we conclude with a discussion of the limitations of these approaches, with a brief overview given of how the non-parametric filter, to be proposed in Chapter 3, will overcome some of these limitations.

The first limitation is that most of the filters presented in Section 2.4 require parametric assumptions for the model. The Kalman filter requires the model to be linear in structure, with Gaussian error terms, in order to yield analytical solutions for the predictive distributions. The extended Kalman filter, whilst addressing the issue of non-linearity in the model, still assumes Gaussianity for the error terms. The grid-based non-Gaussian filter, in turn, allows for non-Gaussian errors in the state space model, but still requires the distributional assumptions for the measurement and state errors to be specified. The particle filters also require parametric specifications for both the measurement and state error terms. If these parametric specifications are incorrect, the out-of-sample predictive distributions produced by the parametric filters will also be incorrect. We note also that with regard to the Kalman filter in particular, the linear, Gaussian assumption implies Gaussianity for the marginal distribution of the data, which is not an accurate representation of the *non-Gaussian* empirical data that is the focus of this thesis.

The second limitation is that some of these filters are computationally burdensome. For example, the Gaussian-sum filter suffers from a curse of dimensionality due to the geometric increase in the number of Kalman filter components with each time step in the filter, making the evaluation of the likelihood difficult. Moreover, even when the number of Gaussian-sum terms is moderated, the computational demands are still large due to the need to establish a suitable clustering of representative Gaussian-sum terms. The particle filters, being simulation-based methods, are also computationally expensive.

In contrast, the non-parametric filter proposed in Chapter 3 is shown to overcome these two limitations. The non-parametric filter can be applied to non-linear, non-Gaussian models, with the measurement error treated non-parametrically, enabling a non-parametric estimate of the forecast distribution to be produced.

For reasonably low dimensions of the measurement and state variables, the non-parametric filter is also shown to have a relatively low computational burden, compared with both the Gaussian-sum and particle filters.

In addition to overcoming the above two broad limitations, the non-parametric filter proposed in Chapter 3 also avoids some of the limitations specific to the various filters. First, the grid-based non-Gaussian filter in Section 2.4.4 requires integration over the supports of the filtered and predictive distributions of x_t, which are dependent on the observed data up to time t and thus vary with t. In contrast, the non-parametric filter is also grid-based, but with integration occurring only over the invariant support of the measurement error density. Second, like the Gaussian-sum filter in Section 2.4.5, the non-parametric filter proposed in Chapter 3 does not assume a parametric specification for the measurement error. However, unlike the Gaussian-sum filter, the non-parametric filter does not suffer from the curse of dimensionality, having a computational burden that is linear, rather than exponential in T. Third, although initially conceived for the non-parametric case, it will become apparent that the proposed approach results in a general filtering algorithm useful for the parametric case also, as the properties of the Dirac delta function enable a switch of integrals so that numerical integration can occur with respect to the invariant distribution of the measurement error. A simulation-based approach that exploits these general features is explored in Chapter 6, with the resulting Monte Carlo filter seen to avoid the degeneracy problems inherent to existing particle filter methods.

Chapter 3

Non-Parametric Estimation of Forecast Distributions in Non-linear, Non-Gaussian State Space Models

3.1 Introduction

In the spirit of the evolving literature referenced in Chapter 1, in which *distributional* forecasts are produced for specific non-Gaussian data types, a new method for estimating the full forecast distribution of non-Gaussian time series variables is developed in this chapter. In contrast to much of the work cited in the first chapter, in which strict parametric models are used, a flexible *non-parametric* approach is to be adopted here, with a view to producing distributional forecasts that are not reliant on the complete specification of the true DGP. The method is developed within the general framework of non-linear, non-Gaussian state space models outlined in Chapter 2, but with the distribution for the observed non-Gaussian variable, conditional on the latent state(s), estimated non-parametrically. The estimated forecast distribution, defined by the relevant function of the non-parametric estimate of the conditional distribution, thereby serves as a flexible representation of the likely future values of the non-Gaussian variable, given its current and past values, and conditional on the (parametric) dynamic structure imposed by the state space form.[1]

The recursive filtering and prediction distributions used both to define the likelihood function and, ultimately, the predictive distribution for the non-Gaussian variable (and for the state also, when of inherent interest), are represented via the numerical solutions of integrals defined over the support of the independent and identically distributed (*i.i.d.*) measurement error - with this support readily approximated in empirical settings. Any standard deterministic integration technique (e.g. rectangular integration, trapezoidal rule, Simpson's rule) can be used to estimate the relevant integrals. The ordinates of the (unknown) measurement density are estimated as unknown parameters using a penalized ML method, with this aspect drawing on the recent work of Berg, Geweke and Rietz (2010) on discrete penalized likelihood (see also Scott, Tapia and Thompson, 1980; Engle and Gonzalez-Rivera, 1991). The relative computational simplicity of the proposed method - for reasonably low dimensions of the measurement and state variables - is in marked contrast with the high computational burden of Monteiro's (2010) Gaussian-sum filter reviewed in Section 2.4.5 - an alterna-

[1] The estimated forecast distribution produced by the non-parametric filter is referred to as a non-parametric estimate throughout this and the following chapters, because our proposed algorithm does not require the specification of a functional form for the forecast distribution. However, given that we do use a parametric structure for the state equation, plus invoke some parametric structure in the measurement equation, the term 'semi-parametric' could conceivably be suitable also.

tive method for allowing for flexibility in the specification of the measurement error. The modest computational burden of the proposed method also stands in contrast with the simulation-based estimation methods needed to implement flexible mixture modelling in the non-Gaussian state space realm (e.g. Durham, 2007; Caron, Davy, Doucet and Duflos, 2008; Jensen and Maheu, 2010; Yau, Papaspiliopoulos, Roberts and Holmes, 2011).

An outline of the rest of the chapter is as follows. Section 3.2 gives an outline of the basic approach, with its computational simplicity being highlighted. Section 3.3 describes the proposed recursive algorithm, with the Dirac delta function (δ-function) used to recast all filtering and predictive densities into integrals defined over the constant support of the measurement error. Section 3.4 constructs the penalized log-likelihood function which is maximized to produce ML estimates of the unknown parameters in the model. An estimate of the forecast distribution is then produced using the estimated parameters. Section 3.5 discusses a modification of the algorithm whereby the measurement error density at each grid point is represented as a mixture of normal distributions. Section 3.6 concludes.

3.2 An Outline of the Basic Approach

The non-parametric estimate of a forecast distribution is developed within the context of a general non-linear, non-Gaussian state space model for a scalar random variable y_t. Consider a univariate version of the general state space model in (2.4) and (2.5), governed by a measurement equation for the scalar y_t and a transition equation for a scalar state variable x_t, given as

$$y_t = h_t(x_t, \eta_t) \tag{3.1}$$

$$x_{t+1} = k_t(x_t, v_t), \tag{3.2}$$

respectively, for $t = 1, 2, ..., T$, where each η_t is assumed to be an *i.i.d.* random variable. The functions given by $h_t(\cdot, \cdot)$ are assumed to be differentiable with respect to each argument. Further, it is assumed that, for given values y_t and η_t, the function

$$G_t(x_t) = y_t - h_t(x_t, \eta_t) \tag{3.3}$$

is assumed to have a unique root at $x_t = x_t^*(y_t, \eta_t)$, as well as having a non-zero derivative at that root. As in the initial outline of the general model in Chapter 2, the focus will be on the case where y_t is continuous, with all distributions expressed using density functions as a consequence. However, with simple modifications the proposed methodology applies equally to the case of discrete measurements and/or states. Extension to the multivariate setting is also possible, as will be illustrated in Chapter 6, although the grid-based method emphasized here is clearly most suitable for reasonably low-dimensional problems. It is also assumed that $x_t^*(y_t, \eta_t)$ is analytically available, in addition to being unique, with adaptation of the method obviously required when neither of these conditions are met. These adaptations will also be explored in Chapter 6.

As is common, η_t is assumed to be independent of x_t, in which case the pdf for η_t is simply

$$p(\eta_t|x_t) = p(\eta_t)\,; \text{ for all } t = 1, 2, ..., T.$$

Time-series independence for η_t is also assumed; that is, any dynamic behaviour in y_t is captured completely by $h_t(\cdot,\cdot)$ and $k_t(\cdot,\cdot)$. However, rather than assume a potentially incorrect parametric specification for $p(\eta_t)$, its distributional form is allowed to be unknown. An initial (parametric) distribution $p(x_1)$ is specified for the scalar state, with the transition densities resulting from (3.2) denoted by $p(x_{t+1}|x_t)$, $t = 1, 2, 3, ..., T$. In the examples considered in the thesis (and as would be standard in many empirical problems), $h_t(\cdot,\cdot)$ and $k_t(\cdot,\cdot)$ are assumed to be known functions for all t, with k_t such that the transition densities $p(x_{t+1}|x_t)$ are available. To avoid unnecessary notation, the t subscript on the functions h and k is suppressed from this point, as generalization to time dependent functions is straightforward.

Given the model defined by (3.1) and (3.2), the one-step-ahead forecast distribution for y_{T+1}, conditional on the observed data, $y_{1:T} = (y_1, y_2, ..., y_T)'$ is

$$p(y_{T+1}|y_{1:T}) = \int p(y_{T+1}|x_{T+1})\, p(x_{T+1}|y_{1:T})\, dx_{T+1}, \tag{3.4}$$

where the explicit dependence of $p(y_{T+1}|y_{1:T})$ on any unknown fixed parameters that characterize $h(\cdot,\cdot)$, $p(x_1)$, or any of the transition densities $\{p(x_{t+1}|x_t)$, $t = 2, 3, ..., T\}$, has been suppressed. The primary aim of the approach is to incorporate, within an overarching ML inferential approach, non-parametric estimation of the conditional measurement distribution, $p(y_{T+1}|x_{T+1})$, which via (3.4), will yield a non-parametric estimate of the one-step-ahead forecast density, $p(y_{T+1}|y_{1:T})$. In cases where the state variable is also of inherent interest, a non-parametric estimate of the corresponding forecast density for the state, $p(x_{T+1}|y_{1:T})$, may be obtained. As outlined below, the non-parametric method is implemented by representing the unknown density, $p(y_{T+1}|x_{T+1})$, by its ordinates defined, in turn, for N grid points on the support of η_{T+1}. The nature of these grid-points is determined by the integration method used to estimate the integrals that define the relevant filtering/prediction algorithm. This approach introduces an additional N unknown parameters to be estimated (via ML) along with any other unknown parameters that characterize the model. Estimation is subject to the usual restrictions associated with probability distributions and to any restrictions to be imposed on the distribution as a consequence of the role played by x_{T+1}. A penalty function is used to impose smoothness on the estimated density of y_{T+1} given x_{T+1}.

As discussed in Section 2.3.4, the likelihood function for the vector of unknown static parameters θ, augmented in the current context by the unknown ordinates of $p(y_{T+1}|x_{T+1})$, requires the availability of the one-step-ahead prediction distributions,

$$p(y_{t+1}|y_{1:t}), \qquad t = 1, 2, ...T - 1 \tag{3.5}$$

and the marginal distribution

$$p(y_1), \tag{3.6}$$

where both (3.5) and (3.6) are (suppressed) functions of θ. In the following section, a computationally efficient filtering algorithm for computing (3.5) and

(3.6), needed for the specification of the likelihood function in (2.10), is outlined. The unknown parameters are estimated by maximizing the (penalized) likelihood function subject to the smoothness and coherence constraints noted above. Conditional on these estimates, the predictive density in (3.4) is estimated, with sampling error able to be quantified in empirical settings using resampling methods, as illustrated in Section 5.4.3.

Crucially, the computational burden associated with evaluation of the likelihood function is shown to be a linear function (only) of the sample size, T. This is in contrast with the computational burden associated with a kernel density representation of $p(\eta_t)$, such as the one used in the Gaussian-sum filter, which was shown to be geometric in T in Section 2.4.5. The computational simplicity of our method derives from the fact that given observed data for period t, the representation of the invariant measurement error density on a *common* grid implies a variable grid of values for the corresponding state variable, x_t. Hence, the computational requirements of evaluating the likelihood using our filter are equivalent to those that either assume or impose discretization on the state (see, for example, Arulampalam *et al.*, 2002; Clements, Hurn and White, 2006).

3.3 A Grid-based Filter

The filtering algorithm proposed here provides an approximation to the true filtering distributions that are in general not available in closed form, even when the measurement error distribution, $p(\eta)$, is known. This approach exploits the functional relationship between the observation y_t and the *i.i.d.* variable η_t, for given x_t, in (3.1). Utilizing this relationship, the filtering expressions are manipulated using properties of the δ-function, in such a way that all requisite integrals are undertaken with respect to the invariant distribution of η. When this measurement error distribution is unknown, the method may be viewed as a non-parametric filtering algorithm, with ordinates of the unknown error density $p(\eta)$, at fixed grid locations, estimated within an ML procedure.

3.3.1 Preliminaries

The δ-function[2] may be represented as

$$\delta\left(z^{*}-z\right)=\left\{\begin{array}{cl}\infty & \text{if } z^{*}=z \\ 0 & \text{if } z^{*}\neq z\end{array}\right.$$

where $\int_{-\infty}^{\infty}\delta\left(z^{*}-z\right)dz=1$ and

$$\int_{-\infty}^{\infty} f(z)\delta\left(z^{*}-z\right)dz=f(z^{*}), \tag{3.7}$$

for any continuous, real-valued function $f(\cdot)$. Note z^{*} is the root of the argument of the δ-function. Further, denoting by $\delta\left(G\left(z\right)\right)$ the δ-function composed with a differentiable function $G\left(z\right)$ having a unique zero at z^{*}, a transformation of variables yields

$$\int_{-\infty}^{\infty} f(z)\delta\left(G\left(z\right)\right)dz=\int_{-\infty}^{\infty} f(z)\left|\frac{\partial G\left(z\right)}{\partial z}\right|^{-1}\delta\left(z-z^{*}\right)dz, \tag{3.8}$$

[2] Strictly speaking, $\delta\left(x\right)$ is a generalized function, and is properly defined as a measure rather than as a function. However, the commonly used heuristic definition is taken advantage of here as it is more convenient for the filtering manipulations that are to follow in the next section. See, for example, Hassani (2009).

resulting, via (3.7), in

$$\int_{-\infty}^{\infty} f(z)\delta\left(G\left(z\right)\right)dz = f(z^{*})\left|\frac{\partial G\left(z\right)}{\partial z}\right|_{z=z^{*}}^{-1}, \tag{3.9}$$

where $\left|\frac{\partial G(z)}{\partial z}\right|_{z=z^{*}}$ denotes the modulus of the derivative of $G\left(z\right)$, evaluated at $z = z^{*}$. The transformation in (3.8) makes it explicit that the root of the argument of the δ-function is $z = z^{*}$, and as a consequence of this result, the δ-function satisfies the following relation

$$\delta\left(G\left(z\right)\right) = \left|\frac{\partial G\left(z\right)}{\partial z}\right|^{-1}\delta\left(z - z^{*}\right) \tag{3.10}$$

when considering the composite function $\delta\left(G\left(z\right)\right)$ explicitly in terms of z. In what follows, $G\left(x_t\right) = y_t - h\left(x_t, \eta_t\right)$ and, hence,

$$\left|\frac{\partial G\left(x_t\right)}{\partial x_t}\right| = \left|\frac{\partial h}{\partial x_t}\right|.$$

Further discussion of using these and other properties of the δ-function may be found in Au and Tam (1999) and Khuri (2004).

In the context of a state space model, the δ-function is used to express the transformation in (3.1) from the *i.i.d.* measurement error η_t to the observed data y_t, given x_t, so that

$$p\left(y_t|x_t\right) = \int_{-\infty}^{\infty} p\left(\eta\right)\delta\left(y_t - h(x_t, \eta)\right)d\eta, \tag{3.11}$$

where η is a variable of integration that traverses the support of $p(\eta)$. This result, along with the transformation of variables relation in (3.10), enables all integrals required to produce the likelihood function in (2.10) to be expressed in terms of the measurement error variable, η.

3.3.2 The Initial Filtered Distribution: $p\left(x_1|y_1\right)$

Using the representation of the measurement density as an integral involving the δ-function in (3.11), it follows that the filtered density of the state variable at time $t=1$ may be expressed as

$$\begin{aligned} p\left(x_1|y_1\right) &= \frac{p\left(x_1\right)p\left(y_1|x_1\right)}{p(y_1)} \\ &= \frac{p\left(x_1\right)\int_{-\infty}^{\infty}p\left(\eta\right)\ \delta\left(y_1-h(x_1,\eta)\right)d\eta}{\int_{-\infty}^{\infty}p\left(x_1\right)\left[\int_{-\infty}^{\infty}p\left(\eta\right)\ \delta\left(y_1-h(x_1,\eta)\right)d\eta\right]dx_1}. \end{aligned}$$

The expression of the resulting filtered density is then simplified in two ways. First, the numerator is written in terms of the state variable using (3.10). Second, the order of integration is reversed with (3.8) and (3.9) used in the denominator to obtain

$$p\left(x_1|y_1\right)=\frac{p\left(x_1\right)\int_{-\infty}^{\infty}p\left(\eta\right)\left|\frac{\partial h}{\partial x_1}\right|^{-1}\delta\left(x_1-x_1^*(y_1,\eta)\right)d\eta}{\int_{-\infty}^{\infty}p\left(x_1^*(y_1,\eta)\right)p\left(\eta\right)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^*(y_1,\eta)}^{-1}d\eta}, \tag{3.12}$$

where $x_1^*(y_1,\eta)$ is the (assumed unique) solution to $y_1-h(x_1,\eta)=0$ for any value η in the support of $p\left(\eta\right)$.

Next, to numerically evaluate the filtered distribution in (3.12) via rectangular integration, an evenly spaced grid $\left\{\eta^1,\eta^2,...,\eta^N\right\}$ is defined, with interval length m, resulting in the approximation for $p\left(x_1|y_1\right)$ given by

$$p\left(x_1|y_1\right)\approx\frac{p\left(x_1\right)\sum_{j=1}^{N}m\,p\left(\eta^j\right)\left|\frac{\partial h}{\partial x_1}\right|^{-1}\delta\left(x_1-x_1^{*j}\right)}{\sum_{i=1}^{N}m\,p\left(x_1^{*i}\right)p\left(\eta^i\right)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^{*i}}^{-1}},$$

where $p\left(\eta^j\right)$ is defined as the unknown density ordinate associated with the grid-point indexed by j.[3] Note that conveniently using the numerical integration

[3] In contrast to the grid-based non-Gaussian filter of Kitagawa (1987), discussed in Section

approach over the same grid $\left\{\eta^1, \eta^2, ..., \eta^N\right\}$ in the numerator as well as in the denominator serves to produce an implied state, $x_1^{*j} = x_1^*(y_1, \eta^j)$, associated with each η^j, such that the first filtered distribution has representation (up to numerical approximation error) as a discrete distribution, with density

$$p\left(x_1|y_1\right) = \sum_{j=1}^{N} W_1^j \delta\left(x_1 - x_1^{*j}\right), \tag{3.13}$$

and where

$$W_1^j = \frac{p\left(\eta^j\right)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^{*j}}^{-1} p\left(x_1^{*j}\right)}{\sum_{i=1}^{N} p\left(\eta^i\right)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^{*i}}^{-1} p\left(x_1^{*i}\right)}, \tag{3.14}$$

for $j = 1, 2, ..., N$.[4] Implicit in this approximation to the first filtered state density is the first likelihood contribution,

$$p\left(y_1\right) = m \sum_{i=1}^{N} p\left(\eta^i\right)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^{*i}}^{-1} p\left(x_1^{*i}\right), \tag{3.15}$$

2.4.4, in which M is used to denote the number of grid points, the notation N is used to denote the number of grid points in the proposed grid-based filter underlying the non-parametric method. The reason for this change in notation is to highlight the conceptual difference between Kitagawa's filter, where M is chosen with reference to the distribution of the state variable (x), and the new filtering method introduced in the thesis, in which N is chosen with reference to the time-invariant distribution of the measurement error (η). The distinction is important, as N is fixed for all t, whilst an efficient implementation of Kitagawa's method would require different values of M for every t.

[4] Note that densities employing the Dirac delta notation should be interpreted carefully. In (3.13), x_1 given y_1 has a discrete distribution with probability mass equal to W_1^j at $x_1 = x_1^{*j}$. It is referred to as a *density* because

$$\begin{aligned}\int_{-\infty}^{c} p\left(x_1|y_1\right) dx_1 &= \int_{-\infty}^{c} \sum_{j=1}^{N} W_1^j \delta\left(x_1 - x_1^{*j}\right) dx_1 \\ &= \sum_{j=1}^{N} W_1^j \int_{-\infty}^{c} \delta\left(x_1 - x_1^{*j}\right) dx_1 \\ &= \sum_{j=1}^{N^*(c)} W_1^j\end{aligned}$$

where $N^*\left(c\right)$ denotes the number of x_1^{*j} that are less than or equal to c.

obtained from approximating the denominator in (3.12).

Having obtained the representation in (3.13) for time $t = 1$, we show that for any time $t = 2, 3, ...T$, an appropriate discrete distribution can be found to approximate the filtered distribution

$$p\left(x_t|y_{1:t}\right) = \sum_{j=1}^{N} W_t^j \delta\left(x_t - x_t^{*j}\right), \tag{3.16}$$

where the recursively determined weights satisfy

$$\sum_{j=1}^{N} W_t^j = 1,$$

and each state grid location

$$x_t^{*j} = x_t^*(y_t, \eta^j) \tag{3.17}$$

is determined by the unique zero of $y_t - h(x_t, \eta^j)$, for $j = 1, 2, ...N$.

3.3.3 The Predictive Distribution for the State: $p\left(x_{t+1}|y_{1:t}\right)$

Assuming (3.16) holds in period t, it follows that the one-step-ahead state prediction density is a mixture of transition densities, since

$$\begin{aligned} p\left(x_{t+1}|y_{1:t}\right) &= \int p\left(x_{t+1}|x_t\right) p\left(x_t|y_{1:t}\right) dx_t \\ &= \int p\left(x_{t+1}|x_t\right) \sum_{j=1}^{N} W_t^j \delta\left(x_t - x_t^{*j}\right) dx_t \\ &= \sum_{j=1}^{N} \int W_t^j p\left(x_{t+1}|x_t\right) \delta\left(x_t - x_t^{*j}\right) dx_t \\ &= \sum_{j=1}^{N} W_t^j \; p\left(x_{t+1}|x_t^{*j}\right), \end{aligned} \tag{3.18}$$

for $t = 1, 2, ..., T$. The notation $p\left(x_{t+1}|x_t^{*j}\right)$ denotes the transition density of $p\left(x_{t+1}|x_t\right)$, viewed as a function of x_{t+1} and given the fixed value of $x_t = x_t^{*j}$. As it is assumed that the transition densities $p\left(x_{t+1}|x_t\right)$ are available, no additional approximation is needed in moving from $p\left(x_t|y_{1:t}\right)$ to $p\left(x_{t+1}|y_{1:t}\right)$. Note that the W_t^j values, for $j = 1, 2, ..., N$, are available from the previous iteration of the filter.

3.3.4 The One-step-ahead Predictive Distribution for the Observed: $p\left(y_{t+1}|y_{1:t}\right)$

Having obtained a representation for the filtered density for the future state variable, x_{t+1}, the corresponding predictive density for the next observation is given by

$$p\left(y_{t+1}|y_{1:t}\right) = \int_{-\infty}^{\infty} p\left(y_{t+1}|x_{t+1}\right) p\left(x_{t+1}|y_{1:t}\right) dx_{t+1}.$$

Utilizing (3.11) for $p\left(y_{t+1}|x_{t+1}\right)$, the one-step-ahead prediction density has representation

$$p\left(y_{t+1}|y_{1:t}\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p\left(\eta\right)\delta\left(y_{t+1} - h(x_{t+1},\eta)\right) d\eta\, p\left(x_{t+1}|y_{1:t}\right) dx_{t+1},$$

which, after integration with respect to x_{t+1}, (and using (3.9) once again), yields

$$p\left(y_{t+1}|y_{1:t}\right) = \int_{-\infty}^{\infty} p\left(\eta\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|^{-1}_{x_{t+1}=x^*_{t+1}(y_{t+1},\eta)} p(x^*_{t+1}(y_{t+1},\eta)|y_{1:t})d\eta. \quad (3.19)$$

Invoking again the pre-specified grid of values for η, we have (up to numerical approximation error),

$$p\left(y_{t+1}|y_{1:t}\right) = m\sum_{i=1}^{N} p\left(\eta^i\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|^{-1}_{x_{t+1}=x^*_{t+1}(y_{t+1},\eta^i)} p\left(x^*_{t+1}(y_{t+1},\eta^i)|y_{1:t}\right). \quad (3.20)$$

Noting that $p\left(x^*_{t+1}(y_{t+1},\eta^i)|y_{1:t}\right)$ in (3.20) denotes the one-step-ahead predictive density from (3.18) evaluated at $x_{t+1} = x^*_{t+1}(y_{t+1},\eta^i)$, it can be seen that $p\left(y_{t+1}|y_{1:t}\right)$ is computed as an N^2 mixture of (specified) transition density functions as a consequence.

3.3.5 The Updated Filtered Distribution: $p\left(x_{t+1}|y_{1:t+1}\right)$

Finally, the predictive distribution for the state at time $t+1$ is updated given the realization y_{t+1} as

$$\begin{aligned} p\left(x_{t+1}|y_{1:t+1}\right) &= \frac{p\left(y_{t+1}|x_{t+1}\right)p\left(x_{t+1}|y_{1:t}\right)}{p\left(y_{t+1}|y_{1:t}\right)} \\ &\approx \frac{m\sum_{j=1}^{N} p\left(\eta^j\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|^{-1}\delta\left(x_{t+1}-x^{*j}_{t+1}\right)\,p\left(x_{t+1}|y_{1:t}\right)}{m\sum_{i=1}^{N} p\left(\eta^i\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|^{-1}_{x_{t+1}=x^{*i}_{t+1}} p\left(x^{*i}_{t+1}|y_{1:t}\right)}, \end{aligned}$$

for $t = 1, 2, ..., T-1$, and where $x^{*j}_{t+1} = x^*_{t+1}(y_{t+1},\eta^j)$ is determined by the j^{th} grid point η^j and the observed y_{t+1}. Hence, the updated filtered distribution has representation (up to numerical approximation error) as a discrete distribution as in (3.16), with density

$$p\left(x_{t+1}|y_{1:t+1}\right) = \sum_{j=1}^{N} W^j_{t+1}\delta\left(x_{t+1}-x^{*j}_{t+1}\right),$$

where, for $j = 1, 2, ..., N$,

$$W^j_{t+1} = \frac{p\left(\eta^j\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|^{-1}_{x_{t+1}=x^{*j}_{t+1}} p\left(x^{*j}_{t+1}|y_{1:t}\right)}{\sum_{i=1}^{N} p\left(\eta^i\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|^{-1}_{x_{t+1}=x^{*i}_{t+1}} p\left(x^{*i}_{t+1}|y_{1:t}\right)}$$

denotes the probability associated with location x^{*j}_{t+1} given by the unique zero of $y_{t+1} - h(x_{t+1},\eta^j)$, for $j = 1, 2, ...N$.

3.3.6 Summary of the Algorithm for General t

While the derivation details the motivation behind the general filter, the actual algorithm is easily implemented using the following summary. Denote by $x_t^{*j} = x_t^*\left(y_t, \eta^j\right)$ the unique zero of $y_t - h\left(x_t, \eta^j\right)$, for each $j = 1, 2, ..., N$ and all $t = 1, 2, ..., T$. Initialize the filter at period 1 with (3.13) and (3.14). For $t = 1, 2, ..., T-1$

$$
\begin{aligned}
p(x_{t+1}|y_{1:t}) &= \sum_{j=1}^{N} W_t^j p\left(x_{t+1}|x_t^{*j}\right), && (3.21)\\
p\left(y_{t+1}|y_{1:t}\right) &= \sum_{i=1}^{N} M_{t+1}^i\ \left(y_{t+1}\right) p(x_{t+1}^*(y_{t+1}, \eta^i)|y_{1:t}), && (3.22)\\
p\left(x_{t+1}|y_{1:t+1}\right) &= \sum_{j=1}^{N} W_{t+1}^j \delta\left(x_{t+1} - x_{t+1}^{*j}\right). && (3.23)
\end{aligned}
$$

with

$$
M_{t+1}^i\left(y_{t+1}\right) = mp\left(\eta^i\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^*(y_{t+1}, \eta^i)}^{-1} \tag{3.24}
$$

and

$$
W_{t+1}^j = \frac{M_{t+1}^j\left(y_{t+1}\right) p\left(x_{t+1}^{*j}|y_{1:t}\right)}{\sum_{i=1}^{N} M_{t+1}^i\left(y_{t+1}\right) p\left(x_{t+1}^{*i}|y_{1:t}\right)}. \tag{3.25}
$$

The computational burden involved in the evaluation of the t^{th} component of the likelihood function $(p\left(y_{t+1}|y_{1:t}\right))$ is of order N^2 for all t, implying an overall computational burden that is linear in T. Note that, although the approximation renders the state filtered distribution discrete, the state prediction density is continuous, as is the prediction density for the observed variable. Conditional on known values for $p\left(\eta^j\right)$ (and all other parameters), for large enough N the filtering algorithm is exact, in the sense of recovering the true filtered and predictive distributions for the state, plus the true predictive distribution for the observed, at each time point.

This approach has three key benefits. First, establishing a grid of η^j values for the region of integration to a reasonable level of coverage need only be done *once* for the *i.i.d.* random variable η (and not for each t). This is in contrast, for example, with the approach of Kitagawa (1987) for the case of a fully parametric non-Gaussian non-linear state space model discussed in Section 2.4.4, in which numerical integration is performed over the non-constant effective supports of the filtered and predictive distributions of x_t, which are, in turn, determined by the observed data up to time point t. Second, and the case of interest here, when the measurement error density, $p(\eta)$, is *unknown*, the mass associated with at each of the grid points resulting from the rectangular integration procedure,

$$g^j = p\left(\eta^j\right) m, \tag{3.26}$$

for $j = 1, 2, ..., N$, may be estimated within an ML procedure. Since m is known, an estimate of $p(\eta)$ is obtained over the regular grid. Extensions of the algorithm incorporating alternative numerical integration methods, such as Simpson's rule, are straightforward but avoided here to keep the complexity to a minimum. Finally, note that while the state transition equation is often stated in the form $x_{t+1} = k_t(x_t, v_t)$ as in (3.2), with $p(v_t)$ independent of x_t, all that is required for the method is that the transition probability density functions $p(x_{t+1}|x_t)$ are available for each t.

3.4 Penalized Log-likelihood Specification

The product of the elements $p\left(y_{t+1}|y_{1:t}\right)$ in (3.22), for $t = 1, 2, ..., T-1$, along with the marginal distribution $p\left(y_1\right)$ in (3.15), defines the likelihood function in (2.10). Motivated by the prior belief that the true unknown distribution of η is a smooth function that declines in the tails, the logarithm of this likelihood function is penalized accordingly. Specifically, the log-likelihood is augmented with two components that (with reference to (3.26)) respectively: (i) impose smoothness on g^j as a function of j; and (ii) penalize large values of $|\boldsymbol{\eta}\prime\mathbf{g} - \eta^j|$, where $\mathbf{g}$ and $\boldsymbol{\eta}$ are the $(N \times 1)$ vectors containing the elements g^j and η^j. (See Berg *et al.*, 2010). The *penalized* log-likelihood function then becomes

$$\ln L\left(\theta\right) = \ln p\left(y_1\right) + \sum_{t=1}^{T-1} \ln p\left(y_{t+1}|y_{1:t}\right) - \omega\frac{1}{2}\mathbf{g}'\mathbf{H}\left(N, \lambda^2\right)\mathbf{g} - \left(1-\omega\right)\mathbf{k}\left(c\right)'\mathbf{g}, \tag{3.27}$$

where

$$\mathbf{H}\left(N, \lambda^2\right) = N^3\lambda^{-2}\boldsymbol{\Delta}'\mathbf{A}\boldsymbol{\Delta} + N^{-1}\left[\mathbf{e}\mathbf{e}' + \boldsymbol{\eta}\boldsymbol{\eta}'\right] \tag{3.28}$$

and $\mathbf{k}\left(c\right)$ is an $(N \times 1)$ vector with j^{th} element given by

$$k^j\left(c\right) = -\exp\left(c\left|\eta^j - \boldsymbol{\eta}'\mathbf{g}\right|\right).$$

The matrix $\mathbf{A}$ in (3.28) is an $(N-2) \times (N-2)$ tridiagonal matrix with $a_{jj} = 1/3$ (for $j = 1, ..., N-2$) and $a_{j,j+1} = a_{j+1,j} = 1/6$ (for $j = 1, ..., N-3$); $\boldsymbol{\Delta}$ is an $(N-2) \times N$ matrix with three nonzero elements $\Delta_{jj} = 1$, $\Delta_{j,j+1} = -2$, $\Delta_{j,j+2} = 1$ in each row j; $\mathbf{e}$ is an $(N \times 1)$ vector of ones; and N is the number of grid points. The first penalty component in (3.27) controls the smoothness of the estimated density function defined by the g^j, with smaller values of λ^2 corresponding to smoother densities. The second penalty term in (3.27) penalizes values of g^j associated with grid-points that are relatively far from the mean, with the value of c determining the size of the penalty. The constant $\omega \in (0, 1)$ weights the two types of penalty. The penalized log-likelihood function is then maximized, subject to $\sum_{i=1}^{N} g^j = 1$, $g^j \geq 0$, $j = 1, 2, ..., N$, to produce ML estimates of the augmented θ. An estimate of the forecast distribution in (3.4) is subsequently produced using these estimated parameters.

3.5 A Mixture-based Alternative

The non-parametric filter developed in Section 3.3 defines the measurement error density as unknown density ordinates associated with specified grid points η^j, where $j = 1, 2, ...N$ (see Equation (3.26)). The non-parametric filter could, in principle, be replaced by a filter in which the measurement error density is represented as a Gaussian sum

$$p(\eta) = \sum_{k=1}^{K} g^k \phi\left(\eta; \eta^k, b^2\right), \tag{3.29}$$

where $\left\{\eta^k, k = 1, 2, ..., K\right\}$ are the locations at which the normal mixtures are centred. The bandwidth parameter $b > 0$ is the standard deviation of each mixture density, assumed here to be constant across all the mixture densities and g^k is the unknown weight attached to the k^{th} mixture which would be estimated via ML, subject to

$$\sum_{k=1}^{K} g^k = 1.$$

Note that K could be small so that relatively few unknown mixture weights require estimation.

The derivation of the mixture-based algorithm is analogous to that of the original non-parametric filter given in Section 3.3, with the recursions in (3.21)-(3.23) remaining appropriate in this case but with the weights in (3.25) requiring modification to reflect the Gaussian sum specification of $p\left(\eta\right)$ in (3.29). Integration over the measurement error density is again undertaken numerically over the grid points $\{\eta^j,\ j=1,2,...,N\}$, selected to cover the support of η. Using the alternative representation for $p\left(\eta\right)$ in (3.29) in the filtering algorithm detailed in Section 3.3, it can be shown that for $t=1,2,...,T-1$, the expressions in (3.21), (3.22) and (3.23) still obtain, but with the values of $M_{t+1}^i(y_{t+1})$ now being given by

$$M_{t+1}^i(y_{t+1}) = m\sum_{k=1}^{K} g^k\phi\left(\eta^i;\eta^k,b^2\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|^{-1}_{x_{t+1}=x_{t+1}^*(y_{t+1},\eta^i)},$$

replacing those in (3.24) and feeding into the weights given in (3.25). The mixture-based non-parametric filter is initialized with (3.13), with the form of W_1^j now given by

$$W_1^j = \frac{\sum_{k=1}^{K} g^k\phi\left(\eta^j;\eta^k,b^2\right)\left|\frac{\partial h}{\partial x_1}\right|^{-1}_{x_1=x_1^{*j}} p\left(x_1^{*j}\right)}{\sum_{i=1}^{N}\sum_{k=1}^{K} g^k\phi\left(\eta^i;\eta^k,b^2\right)\left|\frac{\partial h}{\partial x_1}\right|^{-1}_{x_1=x_1^{*i}} p\left(x_1^{*i}\right)},$$

for $j=1,2,...,N$. In this case the corresponding first likelihood contribution is given by

$$p\left(y_1\right) = m\sum_{i=1}^{N}\sum_{k=1}^{K} g^k\phi\left(\eta^i;\eta^k,b^2\right)\left|\frac{\partial h}{\partial x_1}\right|^{-1}_{x_1=x_1^{*i}} p\left(x_1^{*i}\right).$$

The expression for $p\left(\eta\right)$ in (3.29) has a seemingly similar representation to that used by Monteiro (2010) in the Gaussian-sum filter discussed in Section 2.4.5 (see equation (2.60)). Indeed, the common features shared by the two representations are the use of the Gaussian distribution for the kernel/mixture and the use of the bandwidth as the standard deviation of each kernel/mixture component. There are, however, important differences between the two. First, the means of the Gaussian components of $p(\eta)$ in (2.60) are given by (ultimately, estimates of) the unobserved measurement disturbances η_t, while the Gaussian mixtures in (3.29) have means given by the set of local points, $\left\{\eta^k,\ k = 1, 2, ..., K\right\}$, that are spread across the invariant support of η. Second, each Gaussian component of the kernel density estimator in (2.60) has the same weight, while the weights attached to each Gaussian density in (3.29) are not constant and are able to be estimated freely using the data. Crucially, as described in Section 2.4.5 the Gaussian-sum filter involves a geometric increase in the number of components as the filter propagates through time, whilst the mixture-based approach suggested here maintains a fixed set of components over the full sample period. Hence, expressing the density of η as a mixture of normals in (3.29) still produces closed form representations for the relevant densities, whilst avoiding the curse of dimensionality that afflicts the Gaussian-sum filter. Nevertheless, insertion of this density for η into the filtering recursions (rather than the discrete non-parametric representation) would lead to an increase in the computational burden associated with evaluating the likelihood function from order TN^2 to order TN^2K (in the scalar case). This increase in computational requirement, along with the distinct decrease in the flexibility with which the unknown $p(\eta)$ is represented, has led us not to pursue this modification further. However, it is worth noting that this less flexible representation of $p(\eta)$ may produce some computational gains, relative to the full non-parametric represention, in the high-dimensional case, given that the number of weights to be estimated, K, is independent of the dimension of η. Further exploration of this issue is left for future work.

3.6 Summary

In this chapter a new method is developed for estimating the full forecast distribution of non-Gaussian time series variables in the context of a general non-linear, non-Gaussian state space model. A non-parametric filter is derived that exploits the functional relationship between the observed variable and the state and measurment error variables, expressed using Dirac's δ-function. This representation, along with a simple rectangular integration rule defined over the fixed support of the measurement error, allows the density of the measurement error to be estimated at N grid points using a penalized likelihood procedure. The approach enables predictive distributions to be produced with computational ease in any model in which the relationship between the measure and state is well understood, but the precise distributional form of the measurment error is unknown. The method is developed in the context of a model for a scalar measurement and state, as is suitable for many empirical problems, with extensions to multivariate problems to be demonstrated in Chapter 6. The method can also be modified when some of the assumptions invoked here are relaxed, with these modifications examined in Chapter 6.

Chapter 4

Forecast Distributions: Comparison and Evaluation

4.1 Introduction

Reviews of the forecast evaluation literature, such as Diebold and Lopez (1996), reveal that most attention had, in the past, been paid to evaluating point forecasts, with very little attention given to the evaluation of density forecasts. However, in more recent years, with the increased focus on the production and use of density forecasts, there has been a corresponding development of methods for assessing their accuracy, with key contributions here being Diebold *et al.* (1998); Corradi and Swanson (2006); Gneiting and Raftery (2007); Gneiting *et al.* (2007); and Geweke and Amisano (2010). In the context of this thesis, the aim is to use the available tools to ascertain whether the non-parametric approach developed in Chapter 3 produces distributional forecasts that are more accurate than those produced by (potentially misspecified) parametric alternatives.

Following Geweke and Amisano (2010), a distinction is drawn between the comparison and evaluation of probabilistic forecasts. Comparing forecasts involves measuring *relative* performance; that is, determining which approach is favoured over the other. Scoring rules are used in this thesis to compare the non-parametric and parametric estimates of the predictive distributions of the observed variables. The evaluation of forecasts, on the other hand, involves assessing the performance of a forecasting approach against an *absolute* standard. For example, the probability integral transform (PIT) method (which is adopted here) benchmarks the sequence of cumulative predictive distributions, produced from a single method and evaluated at *ex-post* values, against the distribution of independent and identically distributed uniform random variables that would result if the data *were* generated (in truth) by the assumed model.

The outline of this chapter is as follows: Section 4.2 details the scoring rules used in the comparison of competing forecast distributions. Section 4.3 presents the PIT method, along with the PIT-based tests used to evaluate the predictive distributions. Empirical coverage rates, used to supplement the PIT-based tests, are also described. Section 4.4 details the various DGPs used in simulation experiments in which the accuracy of the non-parametric and parametric estimates of the forecast distributions is assessed, with the simulation results presented in Section 4.4.2. Section 4.5 concludes.

4.2 Forecast Comparison

Scoring rules assess the quality of probabilistic forecasts through the assignment of numerical scores. These scores serve as summary measures of predictive performance and enable competing forecasting methods (non-parametric and parametric in our case) to be ranked. *Proper* scoring rules also encourage the assessor to make careful assessments and to be honest. (See Selten, 1998; Garthwaite, Kadane and O'Hagan, 2005; Gneiting and Raftery, 2007; Gneiting *et al.*, 2007; Boero, Smith and Wallis, 2011, for relevant expositions). Without loss of generality, the setting where scores reflect a reward for good forecasting performance is considered. Let $s\left(P, y_{T+1}^{o}\right)$ be the score that is assigned when the forecaster produces the predictive distribution P and the *ex-post* value, y_{T+1}^{o}, is observed. A scoring rule is proper if, for any observation y_{T+1}^{o} drawn at random from G, the expected value of $s\left(P, y_{T+1}^{o}\right)$ is maximized when $P = G$, and strictly proper if this maximum is unique.[1] Therefore, if a forecaster's personal belief tallies with the truth (G), there is an incentive for the forecaster to provide an honest assessment in reporting this belief, instead of other beliefs, in order to maximize the score.

Four proper scoring rules are adopted: logarithmic score (LS), quadratic score (QS), spherical score (SPHS) and the ranked probability score (RPS), given respectively by

[1] We note from Gneiting (2011) that scoring functions are generally taken to be negatively oriented, with a smaller score indicating a better forecast. However, in this thesis, the scoring functions are taken to be positively oriented, with a larger score therefore indicating a better forecast.

$$LS = \ln p\left(y^o_{T+1}|y_{1:T}\right) \tag{4.1}$$

$$QS = 2p\left(y^o_{T+1}|y_{1:T}\right) - \int_{-\infty}^{\infty} \left[p\left(y_{T+1}|y_{1:T}\right)\right]^2 dy_{T+1} \tag{4.2}$$

$$SPHS = p\left(y^o_{T+1}|y_{1:T}\right) / \left(\int_{-\infty}^{\infty} \left[p\left(y_{T+1}|y_{1:T}\right)\right]^2 dy_{T+1}\right)^{1/2} \tag{4.3}$$

$$RPS = -\int_{-\infty}^{\infty} \left[P(y_{T+1}|y_{1:T}) - I\left(y^o_{T+1} \leq y_{T+1}\right)\right]^2 dy_{T+1}, \tag{4.4}$$

where, in our context, the competing density forecasts, denoted generically by $p\left(y_{T+1}|y_{1:T}\right)$, are produced by applying the non-parametric and (various) parametric methods to the state space models considered later in Section 4.4.1. As the scoring rule in (4.4) uses the forecast cumulative distribution functions rather than density forecasts, the former are analogously denoted by $P\left(y_{T+1}|y_{1:T}\right)$. The symbol $I(\cdot)$ in (4.4) denotes the indicator function that takes a value of one if $y^o_{T+1} \leq y_{T+1}$ and zero otherwise. The integrals with respect to the continuous random variable y_{T+1} in (4.2) to (4.4) are evaluated numerically.

The LS in (4.1) is a simple 'local' scoring rule, returning a high value if y^o_{T+1} is in the high density region of $p\left(y_{T+1}|y_{1:T}\right)$, and a low value otherwise. In contrast, the other three rules depend not only on the ordinate of the predictive density at the realized value of y_{T+1}, but also on the shape of the entire predictive density. The QS in (4.2), for instance, is comprised of two components: a reward for a 'well-calibrated' prediction $(2p\left(y_{T+1}^{o}|y_{1:T}\right))$ and an implicit penalty $(-\int_{-\infty}^{\infty}\left[p\left(y_{T+1}|y_{1:T}\right)\right]^{2}dy_{T+1})$ for misplaced 'sharpness', or certainty, in the prediction. That is, for any given value for $p\left(y_{T+1}^{o}|y_{1:T}\right)$, the QS score is reduced according to the degree of concentration of $p\left(y_{T+1}|y_{1:T}\right)$ (measured by the magnitude of $\int_{-\infty}^{\infty}\left[p\left(y_{T+1}|y_{1:T}\right)\right]^{2}dy_{T+1}$). If $p\left(y_{T+1}|y_{1:T}\right)$ is very concentrated around y_{T+1}^{o} (or 'sharp' in its prediction of the true value) then the sharpness may produce a penalty, but only in association with a reward for correct calibration (i.e. $2p\left(y_{T+1}^{o}|y_{1:T}\right)$ is also large). However, if $p\left(y_{T+1}|y_{1:T}\right)$ is very concentrated elsewhere in the support of y_{T+1} (i.e. not at y_{T+1}^{o}), the sharpness will produce a true penalty, because $2p\left(y_{T+1}^{o}|y_{1:T}\right)$ will be low; the 'moral' here being that certainty is rewarded by the QS score only in conjunction with accuracy. The same principle applies in the calculation of the $SPHS$ in (4.3), with the denominator serving as the penalty term in this case. Neither the QS nor the $SPHS$ are, however, sensitive to the distance between the region of high predictive mass and y_{T+1}^{o}. The RPS in (4.4), on the other hand, is sensitive to distance, rewarding the assignment of high predictive mass *near* to the realized value of y_{T+1}.

In the spirit of Diebold and Mariano (1995), amongst others, we assess the significance of the difference between the average scores of the competing estimated predictive distributions by appealing to a central limit theorem. For each scoring rule, M independent replications of a time series, $\{y_1, y_2, ..., y_T\}^i$ for $i = 1, 2, ..., M$, were simulated, with the parametric and non-parametric estimates of the one-step-ahead forecast distribution produced for each. The scores for the two competing predictive distributions evaluated at the relevant 'observed' $\left(y_{T+1}^{o}\right)^{i}$, were calculated, for each $i = 1, 2, ..., M$. Denote $\overline{SD}$ as the average difference between the scores of the two competing predictive distributions, associated with the set of M (independently) replicated one-step-ahead forecasts. Under the null hypothesis of no difference in the mean scores, the standardized test statistic,

$$z = \frac{\overline{SD}}{\widehat{\sigma}_{SD}/\sqrt{M}}, \tag{4.5}$$

has a limiting $N(0,1)$ distribution, where $\widehat{\sigma}_{SD}/\sqrt{M}$ is the estimated standard deviation of $\overline{SD}$.

4.3 Forecast Evaluation

4.3.1 Probability Integral Transform

According to Diebold *et al.* (1998), regardless of the loss function, the correct predictive distribution (G from above) is superior to all alternatives. This suggests that any density forecast, $p\left(y_{T+1}|y_{1:T}\right)$, should simply be evaluated by testing whether or not $p\left(y_{T+1}|y_{1:T}\right) = g\left(y_{T+1}|y_{1:T}\right)$, where $g\left(\cdot\right)$ is the true predictive density associated with G. Specifically, under the null hypothesis that the predictive distribution corresponds to the *true* DGP, the PIT, defined as the cumulative predictive distribution evaluated at y_{T+1}^{o},

$$u_{T+1} = \int_{-\infty}^{y_{T+1}^{o}} p\left(y_{T+1}|y_{1:T}\right) dy_{T+1}, \tag{4.6}$$

is uniform $(0,1)$ (Rosenblatt, 1952). Hence, the evaluation of $p\left(\cdot\right)$ is performed by assessing whether or not the probability integral transform over M replications,

$\left\{u_{T+1}^{i},\ i=1,2,...,M\right\}$ is $U\left(0,1\right)$. Under $H_0 : u_{T+1} \sim i.i.d.\ U\left(0,1\right)$, the joint distribution of the relative frequencies (or height of the histogram) of the u_{T+1}^{i} is multinomial, with

$$p\left(m_j\right) = \left(\begin{array}{c} M \\ m_j \end{array}\right) p^{m_j}\left(1-p\right)^{M-m_j},$$

where m_j is the number of observations in the jth histogram bin and $p = \frac{1}{bin}$, with $bin = 20$ being the number of histogram bins. The Pearson goodness of fit statistic, used to assess whether the empirical distribution of the u_{T+1}^{i} conforms with this theoretical distribution, is then computed as

$$\chi^2 = \sum_{j=1}^{bin} \frac{\left(m_j - Mp\right)^2}{Mp}.$$

Under $H_0 : u_{T+1} \sim i.i.d.U\left(0,1\right)$, the Pearson's test statistic has an asymptotic χ^2 distribution with $bin-1$ degrees of freedom.

As the Pearson test requires large sample sizes to be reliable (Berkowitz, 2001), we supplement this test with one based on a quantile transformation of u_{T+1},

$$\omega_{T+1} = \Phi^{-1}\left(u_{T+1}\right), \tag{4.7}$$

where $\Phi^{-1}\left(\cdot\right)$ denotes the inverse of the standard normal distribution function. A likelihood ratio (LR) test of $H_0 : \omega_{T+1} \sim i.i.d.N\left(0,1\right)$, against the alternative that the $\left\{\omega_{T+1}^{i},\ \text{for } i=1,2,...,M\right\}$ have an autoregressive structure of order one, with Gaussian errors, is conducted.[2] To supplement the LR results, the

[2] The AR(1) alternative is a particular form of dependence, chosen here due to its simplicity. (See, for example, Berkowitz, 2001). This particular form of the LR test is a special case of an extended forecast evaluation method developed by De Raaig and Raunig (2005) who discuss the use of the LR test to determine the presence of higher order dependence in the transformed PIT values.

Jarque-Bera normality test is applied. Both tests have χ^2 null distributions, with 3 and 2 degrees of freedom respectively.

Finally, histogram plots of the u_{T+1} values can also be used to visually identify systematic errors in the estimated forecast distributions. For example, a humped-shaped histogram of u^i_{T+1} indicates that the estimated predictive distributions tend to underestimate the true probability of occurance in the middle of the support of y_{T+1}, and consequently tend to overestimate the true probability of occurence in one or both tails. Conversely, frequent underestimation of the probability of (both) extreme values of y_{T+1} results in a U-shaped histogram. Another possible example is a humped shape in the middle with peaks at both ends of the histogram. This pattern suggests that the true forecast densities tend to be more leptokurtic than the estimated forecast densities. By considering the apparent departures of the histogram of the u^i_{T+1} values from a uniform distribution in this way, the performance of the estimates of time varying forecast distributions may be evaluated. See Diebold *et al.* (1998) and Carney and Cunningham (2006) for further examples. This visual inspection of PIT plots is used in the empirical illustration in Chapter 5.

4.3.2 Empirical Coverage Rates

The PIT-based tests are supplemented here by empirical coverage rates that give a sense of how accurately an estimated predictive distribution has captured the *spread* of the true predictive distribution. The empirical coverage rate is calculated as the proportion of instances (over M replications) in which the realized value falls within the 95% highest predictive density (HPD) interval. The proportion of samples with realizations that fall in each of the lower and upper 5% predictive tails is also calculated. If the predictive density has a tail coverage rate that is higher (lower) than the nominal rate, it means that extreme values are being under (over) predicted.

4.4 Simulation Experiments

4.4.1 Alternative State Space Models

The non-parametric filter is applied to a range of state space models to produce the non-parametric ML estimates of forecast distributions, $p(y_{T+1}|y_{1:T})$, in a simulation setting. The first model considered (and outlined in Section 4.4.1) is a state space model in which both the measurement and state equations are linear, with both Gaussian and non-Gaussian measurement errors entertained for the true DGP. Non-linearity is then introduced into the measurement equation (in Section 4.4.1), and strictly positive (non-Gaussian) measurement errors assumed. This form of model has been used to characterize (amongst other things) the dynamic behaviour in financial trade durations and is known, in that context, as the stochastic conditional duration model; see Bauwens and Veredas (2004) and Strickland *et al.* (2006). In Chapter 5, in the context of an empirical investigation of S&P500 volatility, we introduce a third model, containing non-linearity in both the measurement and state equations and both Gaussian and non-Gaussian measurement errors invoked. As the motivation for this model derives from the empirical problem in question, we defer discussion of this third model - including the simulation results pertaining to it - until Section 5.3.1 of Chapter 5.

Linear Model

The linear model is the mainstay of the state space literature; hence, it is necessary to ascertain the performance of the non-parametric method in this relatively simple setting, prior to investigating its performance in more complex non-linear models. The comparator is the estimated forecast distribution produced via the application of the Kalman filter to a model in which the measurement error is assumed to be Gaussian. Clearly, when the Gaussian distributional assumption does not tally with the true DGP, the Kalman filter will not produce the correct forecast distribution. Our interest is in determining the extent to which the non-parametric method produces more accurate (distributional) forecasts than the misspecified Kalman filter-based approach.

The proposed linear state space model has the form,

$$y_t = x_t + \eta_t \tag{4.8}$$

$$x_{t+1} = \alpha + \rho x_t + \sigma_v v_t, \tag{4.9}$$

where $\alpha = 0.1$, $\rho = 0.8$, $\sigma_v = 1.2$ and $v_t \sim N(0,1)$. The non-parametric filter is initialized with $p(x_1)$ as the density associated with a $N(\mu, \tau^2)$ random variable, with $\mu = \frac{\alpha}{1-\rho}$ and $\tau^2 = \frac{\sigma_v^2}{1-\rho^2}$. The non-parametric filter is then implemented using the algorithm outlined in Section 3.3, with

$$\begin{aligned} x_t^{*j} &= y_t - \eta^j, \\ \left|\frac{\partial h}{\partial x_t}\right|_{x_t = x_t^{*j}}^{-1} &= 1, \end{aligned}$$

and

$$p\left(x_{t+1}|x_t^{*j}\right) = \left(2\pi\sigma_v^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{x_{t+1} - \left[\alpha + \rho x_t^{*j}\right]}{\sigma_v}\right)^2\right\},$$

for $t = 1, 2, ..., T$. We entertain three different distributions for η_t, including the Gaussian, Student-t and skewed Student-t (see Fernandez and Steel, 1998). The measurement error is standardized to have a mean of zero and variance equal to one ($\eta_t \sim i.i.d\,(0,1)$) and the degrees of freedom parameter is set to 3, implying very fat-tailed non-Gaussian distributions. The skewness parameter is also set to 3 (a value of 1 corresponding to symmetry), implying a heavy right tail associated with the skewed Student-t distribution. For the purpose of integration, the grid supports were set -4 to 4 in the Gaussian case, -6 to 6 in the (symmetric) Student-t case and -4 to 8 in the skewed Student-t case.

Non-linear Model: Stochastic Conditional Duration

The SCD specification models a sequence of trade durations and is based on the assumption that the dynamics in the durations are generated by a stochastic latent variable. Bauwens and Veredas (2004), for example, interpret the latent variable as one that captures the random flow of information into the market that is not directly observed. Denoting by x_t the duration between the trade at time t and the immediately preceding trade, an SCD model for y_t is specified as

$$\begin{aligned} y_t &= e^{x_t}\varepsilon_t && (4.10) \\ x_{t+1} &= \alpha + \rho x_t + \sigma_v v_t, && (4.11) \end{aligned}$$

where ε_t is assumed to be an $i.i.d.$ random variable defined on a positive support, with mean (and variance) equal to one. It is also assumed that $\alpha = 0.1$, $\rho = 0.9$, $\sigma_v = 0.3$ and $v_t \sim i.i.d.\ N(0,1)$, with ε_t and v_t independent for all t. Observed durations will typically exhibit a diurnal regularity that would be removed prior to implementation of the SCD model. Note also that for the purpose of retaining consistent notation throughout the thesis, a t subscript is used on the duration variable in the SCD model to denote sequential observations over time. These sequential durations are, of course, associated with irregularly spaced trades.

Taking logarithms of (4.10), the measurement equation is transformed as

$$\ln(y_t) = x_t + b + \sigma_\eta \eta_t, \tag{4.12}$$

where $\varepsilon_t = \exp(b + \sigma_\eta \eta_t)$, $\eta_t \sim i.i.d.\,(0,1)$, $b = E(\ln \varepsilon_t)$ and $\sigma_\eta^2 = Var(\ln \varepsilon_t)$. Three different distributions for ε_t are adopted: exponential, Weibull and gamma. The second column in Table 4.1 documents the form of the density for each of the three DGPs (exponential, Weibull and gamma, identified in the first column) that are adopted for ε_t in (4.10), while the third column in Table 4.1 documents the corresponding density for η_t in (4.12). The fourth and fifth columns in Table 4.1 provide (respectively) the associated values for b and σ_η^2 for each of the three distributional assumptions adopted for ε_t; see Johnson, Kotz and Balakrishnan

(1994).

The non-parametric filter is initialized with $p\left(x_1\right)$ as the density associated with a $N\left(\mu,\tau^2\right)$ variable, with again $\mu=\frac{\alpha}{1-\rho}$ and $\tau^2=\frac{\sigma_v^2}{1-\rho^2}$. A range of -7 to 3 for η_t is used in implementing the non-parametric approach, due to the negative skewness that results from the logarithmic transformation of ε_t. The filter is then implemented using the algorithm in Section 3.3, with

$$\begin{aligned}
x_t^{*j} &= x_t^{*j}(y_t,\eta^j) \\
&= \ln\left(y_t\right)-b-\sigma_\eta\eta^j, \\
\left|\frac{\partial h}{\partial x_t}\right|_{x_t=x_t^{*j}}^{-1} &= 1
\end{aligned}$$

and

$$p_t\left(x_{t+1}|x_t^{*j}\right)=\left(2\pi\sigma_v^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\frac{x_{t+1}-\left[\alpha+\rho x_t^{*j}\right]}{\sigma_v}\right)^2\right\}.$$

Table 4.1:
Density functions corresponding to specifications for ε_t in (4.10) and η_t in (4.12). Notation used is as follows: $b=E\left(\ln\varepsilon_t\right)$ and $\sigma_\eta^2=Var(\ln\varepsilon_t)$, with $\eta_t=\left(\ln\varepsilon_t-b\right)/\sigma_\eta\sim i.i.d.\left(0,1\right)$. In the table, $f\left(\eta_t\right)=b+\sigma_\eta\eta_t$, $\psi\left(\zeta\right)$ is the digamma function and $\psi'\left(\zeta\right)$ is the trigamma function.

	$p\left(\varepsilon_t\right)$	$p\left(\eta_t\right)$	b	σ_η^2
Exponential	$\exp\left(-\varepsilon_t\right)$	$\exp\left(-\exp\left(f\left(\eta_t\right)\right)\right)$ $\times\exp\left(f\left(\eta_t\right)\right)\sigma_\eta$	-0.5772	$\frac{\pi^2}{6}$
Weibull	$\gamma\left(\varepsilon_t\right)^{\gamma-1}\exp\left[-\left(\varepsilon_t\right)^\gamma\right]$	$\gamma\exp\left(-\exp\left(f\left(\eta_t\right)\gamma\right)\right.$ $\left.\times\exp\left(f\left(\eta_t\right)\gamma\right)\sigma_\eta\right.$	$\frac{-0.5772}{\gamma}$	$\frac{\pi^2}{6\gamma^2}$
Gamma	$\frac{\varepsilon_t^{\zeta-1}\exp(-\varepsilon_t)}{\Gamma(\zeta)}$	$\frac{\sigma_\eta}{\Gamma(\zeta)}\exp\left(-\exp\left(f\left(\eta_t\right)\right)\right)$ $\times\exp\left(\zeta f\left(\eta_t\right)\right)$	$\psi\left(\zeta\right)$	$\psi'\left(\zeta\right)$

The parametric comparator treats η_t as if it were *i.i.d.* $N(0,1)$ and uses the Kalman filter to produce the forecast density for the logarithmic duration. Given that this distributional assumption for η_t is incorrect, the approach based on the Kalman filter does not produce the correct forecast distribution, and the forecast accuracy of this (misspecified) approach in comparison with that of the non-parametric method is documented.

4.4.2 Simulation Results[3]

All DGPs in the two broad models being investigated (as detailed in Section (4.4.1)) are simulated over $M = 1000$ replications, with $T = 1000$. The parameter values (other than the density ordinates defining the measurement error in the non-parametric case) are fixed in the simulation exercise, taking on values recorded in Section 4.4.1. Table 4.2 records the values of λ, c and ω in (3.27) used to ensure smoothness of the estimate of the measurement error distribution. Values of the smoothing parameters were determined by a trial and error process. Other parameters values have been chosen with reference to typical empirical data relevant to the model at hand, with parameter values chosen for the SCD model being reasonably representative of the empirical estimates reported by Bauwens and Verdas (2004) and Strickland *et al.* (2006).

Prior to selecting particular values for N for use in the simulations, some experimentation with different numbers of grid-points was conducted, with a view to gauging the robustness of the predictive results to this choice parame-

[3] All numerical results in this and the following empirical section have been produced using the GAUSS programming language.

ter. Specifically, a single one-step-ahead predictive distribution was produced, based on data simulated from the two models, and under the three different measurement distributions for each model. Panel A of Figure 4.1 shows the plots of the non-parametric estimates of $p\left(y_{T+1}|y_{1:T}\right)$ for the linear model, with Gaussian, Student-t and skewed Student-t distributional assumptions respectively, and with N varying from 11 to 51. Panel B shows the corresponding estimates of $p\left(\ln y_{T+1}|\ln y_{1:T}\right)$ for the SCD model, with exponential, Weibull and gamma measurement errors, and with N varying from 21 to 61. It can be seen that estimates of the predictive distributions obtained using different values of N are almost indistinguishable from one another, leading us to choose the smallest values of N in the two ranges considered, namely $N = 11$ for the linear model and $N = 21$ for the SCD model, with all grid points evenly spaced.[4]

[4] It should be noted here that the support of η_t is wider in the case of the SCD model, than for the linear model, due to the skewness of $p\left(\eta_t\right)$ that results from the log-transformation of the exponential, Weibull and gamma probability distributions considered for ε_t in equation (4.10). Consequently, more grid points were found to be needed to capture the areas of non-negligible probability over this wider support.

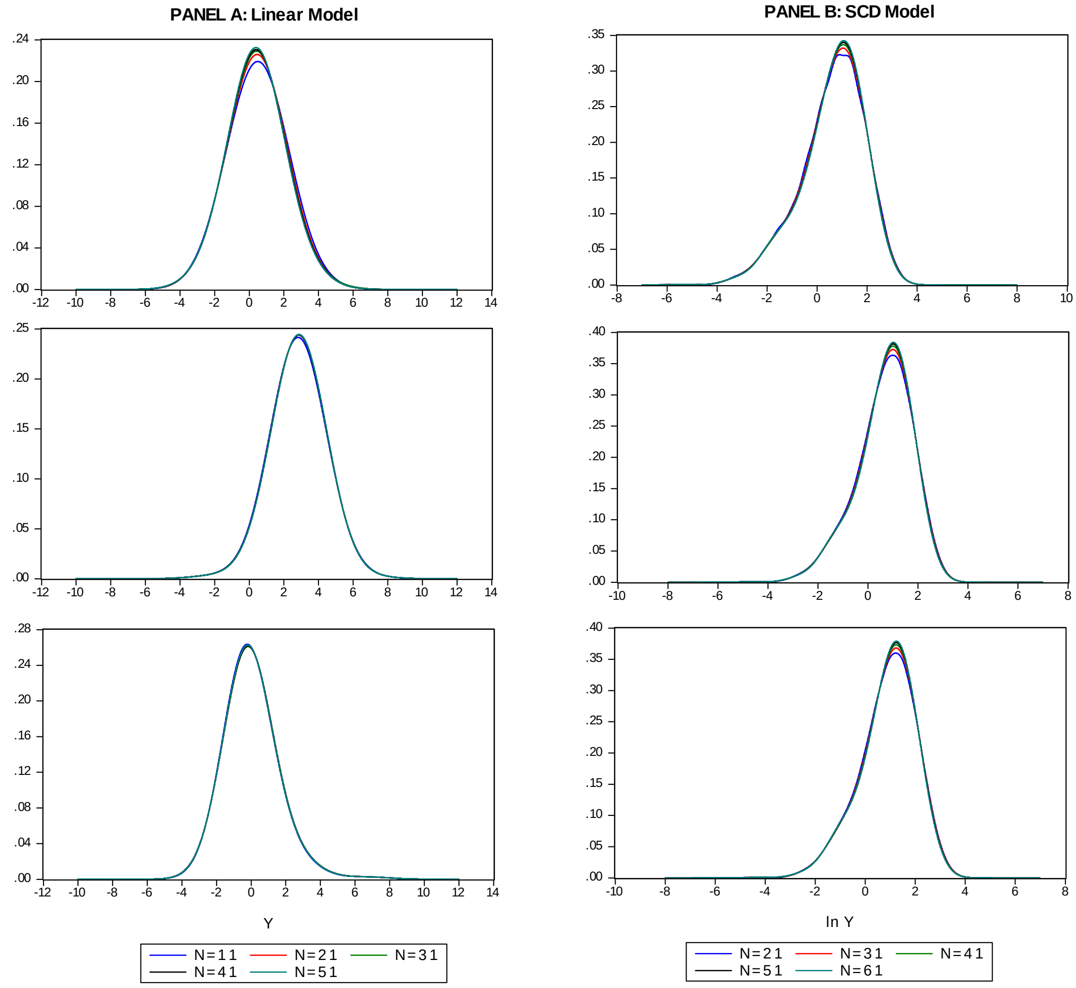


Figure 4.1: Estimated one-step-ahead predictive distribution of the linear and SCD models for varying number of grid points, N. Panel A shows (from top to bottom), $p\left(y_{T+1}|y_{1:T}\right)$, for the Gaussian, Student-t and skewed Student-t DGPs, with N ranging from 11 to 51. Panel B shows (from top to bottom), $p\left(\ln y_{T+1}|\ln y_{1:T}\right)$, for the exponential, Weibull and gamma DGPs, with N ranging from 21 to 61.

Table 4.2:
Constants, λ, c and ω, used in the penalized likelihood function in (3.27), in the simulation experiments for the linear and SCD models, as detailed in Section (4.4.1).

	η_t	λ	c	ω
Linear Model	$N(0,1)$	0.5	0.5	0.2
	Student $t(0,1,\nu=3)$	4.0	0.5	0.2
	Skewed Student $t(0,1,\nu=3,\gamma=3)$	6.0	0.05	0.2
SCD Model	*Exponential* $(1,1)$	1.0	1.0	0.4
	Weibull $(\gamma=1.15,1)$	1.0	1.0	0.4
	Gamma $(\zeta=1.23,1)$	1.0	1.0	0.4

Tables 4.3, 4.4 and 4.5 record respectively all score, evaluation and coverage results. Results for the linear model, (4.8) and (4.9), and the SCD model, (4.12) and (4.11), are recorded in Panel A and B respectively of each table. With reference to Panel A in Table 4.3, the scores of the non-parametric estimate of $p\left(y_{T+1}|y_{1:T}\right)$, under the Gaussian DGP, are seen to be lower overall than those of the parametric forecast, across all four measures. This is no surprise, given that the Kalman filter produces the correct forecast distribution in the linear, Gaussian case. However, the differences between the scores are insignificant at the 5% level, indicating that the non-parametric method does very well at recovering the true forecast distribution, despite not exploiting the distributional form of the measurment error. In the Student-t case - in which the Gaussian assumption underlying the Kalman filter-based distribution is incorrect - the scores of the non-parametric estimate of $p\left(y_{T+1}|y_{1:T}\right)$ are higher overall than for the para-

metric forecast, across all four measures. Once again, however, the differences are insignificant at the 5% level, except for the logarithmic score, according to which the non-parametric estimate significantly outperforms the misspecified parametric alternative. Under the *skewed* Student-t DGP, the non-parametric estimates significantly out-perform the misspecified parametric estimates, for all four scoring measures.

Panel A of Table 4.4 records (for the linear model) the test statistics associated with the three PIT tests described in Section 4.3, namely, the Pearson test for the uniformity of $\left\{u_{T+1}^{i},\ i=1,2,...,M\right\}$ in (4.6), the LR test of the normality (and independence) of $\left\{\omega_{T+1}^{i},\ i=1,2,...,M\right\}$ in (4.7) and the Jarque-Bera test for the normality of $\left\{\omega_{T+1}^{i},\ i=1,2,...,M\right\}$. For the (conditionally) Gaussian DGP, all test statistics - for both the non-parametric and parametric estimates - do not reject the null at the 5% level, indicating that both approaches produce accurate predictive distributions for this DGP. In contrast, in the Student-t and skewed Student-t cases, at least one of the LR and Jarque-Bera tests leads to rejection of the parametric estimates, indicating that the predictive distributions produced by the misspecified parametric approach under these two DGPs are inaccurate. The LR test of the non-parametric estimate of $p\left(y_{T+1}|y_{1:T}\right)$ in the skewed Student-t case leads to marginal rejection (at the 5% level), but the other two tests of the non-parametric estimate fail to reject the null hypothesis.

With reference to Panel A of Table 4.5, the lower and upper 5% coverage rates for both forecasting approaches, and under all three DGPs, are seen to be close to the nominal levels, indicating that both approaches are able to capture the tails of the true predictive distribution well enough, in the linear case, even under (parametric) misspecification. However, under misspecification, the parametric estimate has significant (although not 'substantial') undercoverage of the 95% interval.

Table 4.3:
Forecast comparison. Average scores for the parametric (Kalman filter based) and non-parametric estimates of $p\left(y_{T+1}|y_{1:T}\right)$ for the linear model (Panel A) and $p\left(\ln y_{T+1}|\ln y_{1:T}\right)$ for the SCD model (Panel B), for the respective DGPs, with z values (see (4.5)) for the difference in scores across the competing forecasts reported. In the table, ** represents statistical significance at the 5% level for a one-sided test.

PANEL A: Estimated $p\left(y_{T+1}|y_{1:T}\right)$ for the linear model (Section 4.4.1)

	Logarithmic Score			Quadratic Score		
η_t:	*N*	*St*	*SkSt*	*N*	*St*	*SkSt*
Kalman filter	-1.9487	-1.9872	-2.0464	0.1665	0.1684	0.1615
Non-parametric	-1.9512	-1.9695	-2.001	0.1662	0.1693	0.1652
z-statistic	-1.2825	2.5027**	3.8688**	-0.7064	0.8918	2.4836**
	Spherical Score			Continuous Ranked Probability Score		
η_t:	*N*	*St*	*SkSt*	*N*	*St*	*SkSt*
Kalman filter	0.4081	0.4104	0.4019	-0.9576	-0.9774	-1.0269
Non-parametric	0.4078	0.4113	0.4065	-0.9584	-0.9728	-1.0032
z-statistic	-0.5760	0.7909	2.6586**	-0.5254	1.1732	3.4734**

PANEL B: Estimated $p\left(\ln y_{T+1}|\ln y_{1:T}\right)$ for the SCD model (Section 4.4.1)

	Logarithmic Score			Quadratic Score		
η_t:	*Exp*	*Wb*	*Gamma*	*Exp*	*Wb*	*Gamma*
Kalman filter	-1.7414	-1.6280	-1.6463	0.2086	0.2398	0.2303
Non-parametric	-1.7114	-1.5958	-1.6115	0.2135	0.2470	0.2353
z-statistic	3.2606**	3.1794**	3.6051**	2.1441**	2.9672**	2.2638**
	Spherical Score			Continuous Ranked Probability Score		
η_t:	*Exp*	*Wb*	*Gamma*	*Exp*	*Wb*	*Gamma*
Kalman filter	0.4567	0.4898	0.4799	-0.7729	-0.6829	-0.7004
Non-parametric	0.4621	0.4970	0.4851	-0.7643	-0.6712	-0.6914
z-statistic	2.2215**	2.9651**	2.3718**	2.5278**	2.9249**	3.3838**

Table 4.4:

Forecast Evaluation. Pearson, LR and Jarque-Bera χ^2 test statistics, for the parametric and non-parametric estimates of $p\left(y_{T+1}|y_{1:T}\right)$ for the linear model (Panel A) and $p\left(\ln y_{T+1}|\ln y_{1:T}\right)$ for the SCD model (Panel B), for the respective DGPs. In the table, ** represents statistical significance at the 5% level. The critical values for the three tests are respectively 30.14, 7.82 and 5.99.

	Pearson		**LR**		**Jarque-Bera**	
PANEL A: Estimated $p\left(y_{T+1}\|y_{1:T}\right)$ for the Linear model (Section 4.4.1)						
	NP	KF	NP	KF	NP	KF
$\eta_t \sim N(0,1)$	13.12	11.88	1.646	0.081	0.826	0.0921
$\eta_t \sim St(0,1,\nu=3)$	13.44	11.56	3.228	3.648	3.251	37.619**
$\eta_t \sim SkSt(0,1,\nu=3,\gamma=3)$	12.48	21.40	9.053**	15.571**	1.6968	75.781**
PANEL B: Estimated $p\left(\ln y_{T+1}\|\ln y_{1:T}\right)$ for the SCD model (Section 4.4.1)						
	NP	KF	NP	KF	NP	KF
$\eta_t\sim \exp\left(1,1\right)$	20.68	44.68**	1.188	0.581	3.077	64.983**
$\eta_t\sim Wb\left(\gamma=1.15,1\right)$	9.96	48.64**	1.879	0.635	4.409	129.785**
$\eta_t\sim Gamma\left(\zeta=1.23,1\right)$	10.16	31.60**	3.933	2.554	1.131	77.524**

Considering now the score results for the SCD model, recorded in Panel B of Table 4.3, all four scores for the non-parametric estimate of $p\left(\ln y_{T+1}|\ln y_{1:T}\right)$ are seen to be significantly higher than the corresponding scores for the parametric estimate, for all three DGPs. With reference to Panel B of Table 4.4, across all DGPs the non-parametric estimates of $p\left(\ln y_{T+1}|\ln y_{1:T}\right)$ are assessed as being correct, as none of the null hypotheses for the three tests is rejected at the 5% level. The (misspecified) parametric estimate, on the other hand, is associated with rejection for all but one of the tests of fit. Whilst none of the 5% (lower tail) and 95% coverage rates recorded in Panel B of Table 4.5 (for either forecasting approach) is significantly different from the nominal level, the 5% (lower tail) coverage rates for the non-parametric estimate are closer to the nominal level than those of the parametric alternative, for all three DGPs. In addition, the 5% upper tail of the non-parametric forecast distribution has coverage that is not significantly different from the nominal level, whereas the estimate from the Kalman filter-based approach significantly underestimates the nominal level.

Table 4.5:
Forecast Evaluation. Coverage rates (5% and 95%) for the parametric (Kalman filter based) and non-parametric estimates of $p\left(y_{T+1}|y_{1:T}\right)$ for the linear model (Panel A) and $p\left(\ln y_{T+1}|\ln y_{1:T}\right)$ for the SCD model (Panel B), for the respective DGPs. In the table, ** represents significant difference from the nominal coverage, at the 5% significance level.

	5% lower tail			**5% upper tail**			**95% HPD**		
PANEL A: Estimated $p\left(y_{T+1}\|y_{1:T}\right)$ for the linear model (Section 4.4.1)									
η_t:	*N*	*St*	*SkSt*	*N*	*St*	*SkSt*	*N*	*St*	*SkSt*
Kalman filter	4.8	4.5	5.5	5.0	5.3	6.4	94.9	93.3**	92.4**
Non-parametric	4.4	4.6	6.1	4.5	5.9	5.8	95.2	94.1	93.5
PANEL B: Estimated $p\left(\ln y_{T+1}\|\ln y_{1:T}\right)$ for the SCD model (Section 4.4.1)									
η_t:	*Exp*	*Wb*	*Gamma*	*Exp*	*Wb*	*Gamma*	*Exp*	*Wb*	*Gamma*
Kalman filter	6.0	5.8	6.5	2.7**	2.8**	3.3**	94.9	94.9	95.4
Non-parametric	5.2	4.7	5.1	6.0	6.3	5.9	94.2	94.3	94.7

4.5 Conclusions

This chapter has addressed the issue of comparing and evaluating competing probabilistic forecasts. A clear distinction between the comparison and evaluation has been made, with four proper scoring rules presented and used to compare the probablistic forecasts produced by both the non-parametric method and various parametric alternatives. Evaluation of all probabilistic forecasts has been performed using PIT-based tests, supplemented with empirical coverage rates.

Simulated data is used in the assessment, with the parametric comparators all using a Kalman-filter based approach. Three distributions for the true measurement error are entertained for each of the linear and (non-linear) SCD models, with the parameter values chosen (for the SCD model in particular) with reference to typical empirical data. The simulation results show that the non-parametric method performs significantly better, overall, than (misspecified) parametric alternatives and is competitive with correctly specified parametric estimates.

Chapter 5

Non-Parametric Estimation of Forecast Distributions of Realized Volatility

5.1 Introduction

Financial market volatility is central to the theory and practice of asset and derivative pricing, asset allocation and risk management. For example, modern option pricing theory, beginning with Black and Scholes (1973), assigns volatility as an input into the determination of option prices. Volatility is also used in the calculation of Value-at-Risk (VaR), which is commonly used by financial risk managers to report the riskiness of an asset portfolio. Obtaining accurate forecasts of future volatility are critical for these, and other, financial applications, a task rendered challenging by the fact that volatility is a latent quantity and, hence, not directly observed.

The importance of volatility forecasts in practice has led to an enormous academic literature on the modelling and forecasting of returns volatility. A vast majority of the earlier studies relied on the autoregressive conditional heteroskedastic (ARCH) framework pioneered by Engle (1982), in which latent volatility is modelled as a time-varying function of lagged observed returns. The generalized ARCH (GARCH) model was subsequently introduced by Bollerslev (1986). Reviews of these two models can be found in Bollerslev, Chou and Kroner (1992), Bollerslev, Engle and Nelson (1994), Andersen and Bollerslev (1998a), Diebold (2004) and Bauwens, Laurent and Rombouts (2006), amongst others. Bollerslev (2010) also provides an extensive glossary of the plethora of models based on the ARCH and GARCH framework that have been developed over the years.

The ARCH and GARCH models are conditionally deterministic models; that is the volatility of returns, conditional on past information, is a non-stochastic function of observed past returns. Alternatively, volatility may be modelled as an intrinsically stochastic process, with models of this kind known as stochastic volatility models (Taylor, 1986, 1994; Ghysels, Harvey and Renault, 1996; Shephard, 1996). Whilst the additional flexibility associated with the stochastic specification of volatility may yield benefits from a modelling point of view, the introduction of an additional source of randomness creates obvious challenges for inference. A large literature, primarily simulation-based, has thus developed in which the latent volatility process is accommodated; see Broto and Ruiz (2004) and Shephard (2005) for recent reviews.

ARCH, GARCH and stochastic volatility models are parametric representations of latent volatility, with their usefulness depending upon the validity of the specific distributional assumptions invoked. In contrast, Andersen and

Bollerslev (1998b) first examined the use of high frequency data for providing model-free, non-parametric measures of the true unobserved volatility, referred to as realized volatility. This approach involves estimating volatility by the sum of squared intra-day high frequency returns, and is formally discussed by Andersen, Bollerslev, Diebold and Ebens (2001a), Andersen, Bollerslev, Diebold and Labys (2001b) and Barndorff-Nielsen and Shephard (2002a,b). Several other studies highlighting the advantage of using high-frequency data to measure volatility include Zhou (1996), Meddahi (2002) and Andersen, Bollerslev, Diebold and Labys (2003). Some surveys of this literature can be also found in Barndorff-Nielsen, Nicolato and Shephard (2002) and Andersen, Bollerslev and Diebold (2010).

The rationale behind the approach is that as the number of intraday returns used in the calculation of realized volatility approaches infinity, realized volatility converges to the true latent volatility factor. Hence, *ex-post* volatility becomes observable and can thus be modelled directly, rather than being treated as a latent variable. The non-parametric measure calculated for day t (say) may be used as a direct proxy for volatility on day t (or the subsequent day, $t+1$). However, increasingly, researchers have attempted to capture the stylized dynamic behaviour evident in the realized measures by fitting an observation-driven time series model of some sort to the observed measure and subsequently using the fitted model for forecasting. (See Andersen, Bollerslev and Diebold, 2007; Aït-Sahalia and Mancini, 2008; Martin, Reidy and Wright, 2009; Martens, van Dijk and de Pooter, 2009; and Liu and Maheu, 2009, for recent examples).

The importance of jump variation as a component of the observed measure has also been given increased attention of late, as the significance of large discrete jumps in asset prices has become more evident; e.g. Andersen *et al.* (2007); Bollerslev, Kretschmer, Pigorsch and Tauchen (2009); and Tauchen and Zhou (2011).

High frequency data are, however, often contaminated by microstructure effects, such as discrete clustering and bid-ask spreads, for example (Bai, Russell and Tiao, 2001; Andreou and Ghysels, 2002). One solution to this problem has been to sample over an intermediate frequency, with sampling frequencies ranging from 5 minute intervals (e.g. Andersen *et al.*, 2001a; Barndoff-Nielsen and Shephard, 2002a) to as long as 30 minutes (e.g. Andersen *et al.*, 2003) being common. Another solution has been to modify the raw measure itself, via one of a variety of methods, all of which attempt to retrieve consistency of the measure in the presence of microstructure noise. See *Journal of Econometrics*, 2011, Volume 160(1): Special Issue on Realized Volatility, for several recent contributions to this literature.

In practice of course, despite the asymptotic underpinnings of the realized volatility measures, the observed measures capture the theoretical variance quantity with error. Hence, the measures have also been exploited as observable quantities in a state space model, with filtering techniques employed to extract estimates of the latent variance; see Barndorff-Nielsen and Shephard (2002); Creal (2008); Jacquier and Miller (2010); and Maneesoonthorn, Martin, Forbes and Grose (2011) for illustrations. We pursue this approach in this chapter, with the key aspects of our approach summarized as follows. We specify a nonlinear state space model, referred to as the realized volatility (RV) model, via which we produce the one-step-ahead forecast distribution of realized volatility using the non-parametric filter developed in Section 3.3. The model makes use of a (discretized) version of a continuous time diffusion for the latent variance process, for which realized volatility is a noisy measure. Jumps in price (and/or volatility) are not measured explicitly, and the adjustment for microstructure noise is informal. Hence, the error term in the measurement equation will absorb these unmodelled effects, in addition to the effect of using a finite number (only) of high-frequency observations to construct the realize volatility measure. The non-parametric method will, in principle, capture the distributional features that arise from all of these factors.

Finally, we note that the focus on *probabilistic* forecasting of volatility *per se* is in marked contrast to the focus of related work whereby point forecasts of volatility are the key output. See Blair, Poon and Taylor (2001); Martens and Zein (2004); Pong, Shackleton, Taylor and Xu (2004); Koopman, Jungbacker and Hol (2005); Martin *et al.* (2009); and Busch, Christensen and Nielsen (2011) for examples.

The structure of this chapter is as follows. Section 5.2 begins with the model assumed to underlie spot price data for a given financial asset. Then the model used to estimate the one-step-ahead forecast distribution for RV, the observable measure of latent volatility for the given financial asset, is established. The non-parametric method is applied to the RV model to produce the one-step-ahead forecast distribution in a simulation experiment in Section 5.3, with the extended Kalman filter presented in Section 2.4.2 used as the comparator. Using the comparison and evaluation tools outlined in Chapter 4, the relative forecast accuracy of these two approaches is examined using simulated data, with the results documented in Section 5.3.1. The results of an empirical illustration of realized volatility for the S&P500 index from January 1998 to August 2008 are reported in Section 5.4, with a subsampling method used to measure the impact of sampling variation on the estimated forecast distribution.

5.2 Realized Volatility Model

We wish to specify a state space model for realized volatility, which will relate the observed measure to a dynamic model for the latent volatility process associated with a financial asset return. The following bivariate jump diffusion process is assumed for the price of a financial asset, P_t, and its stochastic variance, V_t,

$$\frac{dP_t}{P_t} = \mu_p dt + \sqrt{V_t} dB_t^p + dJ_t \tag{5.1}$$

$$dV_t = \kappa[\phi - V_t]dt + \sigma_v \sqrt{V_t} dB_t^v, \tag{5.2}$$

where $dJ_t = Z_t dN_t$, $Z_t \sim N(\mu_z, \sigma_z^2)$, and $P(dN_t = 1) = \delta_J dt$ and $P(dN_t = 0) = (1 - \delta_J)\, dt$. Under this specification, random jumps are allowed to occur in the asset price, at rate δ_J, and with a magnitude determined by a normal distribution. The pair of Brownian increments (dB_t^p, dB_t^v) are potentially correlated, having a contemporaneous correlation coefficient ρ. However, dB_t^i and dJ_t are assumed to be independent, for each of $i = \{p, v\}$. This model is referenced in the literature as the stochastic volatility with jumps (SVJ) model (Eraker *et al.,* 2003; Broadie *et al.,* 2007).

Given the assumed variance process in (5.2), quadratic variation over the horizon $t-1$ to t (assumed to be a day) is defined as

$$\mathcal{QV}_{t-1,t} = \int_{t-1}^{t} V_s ds + \sum_{t-1<s\leq t}^{N_t} Z_s^2.$$

That is, $\mathcal{QV}_{t-1,t}$ is equal to the sum of the *integrated variance* of the continuous sample path component of P_t,

$$\mathcal{V}_{t-1,t} = \int_{t-1}^{t} V_s ds, \tag{5.3}$$

and the sum of the $N_t - N_{t-1}$ squared jumps that occur on day t. Using the notation p_{t_i} to denote the i^{th} logarithmic price observed during day t, and $r_{t_i} = p_{t_i} - p_{t_{i-1}}$ as the i^{th} transaction return, it follows (see Barndorff-Nielsen and Shephard, 2002a; and Andersen *et al.*, 2003) that

$$RV_t = \sum_{t_i \in [t-1,t]}^{B} r_{t_i}^2 \xrightarrow{p} \mathcal{QV}_{t-1,t}, \tag{5.4}$$

where RV_t is referred as *realized variance* (or, in a slight abuse of terminology, *realized volatility*) and B is equal to the number of intraday returns on day t. The result in (5.4) is based on the implicit assumption that microstructure noise effects are absent.

The measurement equation is defined as

$$\ln RV_t = \ln V_t + u_{RV_t}, \tag{5.5}$$

where the latent volatility evolves according to (5.2) and $u_{RV_t} = \ln RV_t - \ln V_t$ is the logarithmic realized volatility error. The measurement equation has been defined in logarithmic form (for both RV_t and V_t) in order to (approximately) remove the dependence of the deviation of RV_t from V_t on the level of V_t. (See, for example, Barndorff-Nielsen and Shephard, 2002a). Based on the assumed DGP in (5.1) and (5.2), u_{RV_t} in (5.5) will capture the effect of ignoring the price jump variation contained in $\ln RV_t$; the error associated with using the point in time variance, V_t, as an estimate of the integrated variance in (5.3); and the error associated with the use of a finite value of B. If no adjustment is made to the realized variance measure to cater for the presence of microstructure noise, the error term will also capture this omitted effect. The non-parametric method will, in principle, capture the distributional features of u_{RV_t} that arise from all of these factors.

An Euler approximation of (5.2) is used to define the state equation,

$$V_{t+1} = \kappa\phi + (1-\kappa)\,V_t + \sigma_v\sqrt{V_t}v_t, \tag{5.6}$$

where V_t = the point-in-time volatility on day t and $v_t \sim i.i.d.N(0,1)$. The parameter ϕ is an annualized quantity, matching the annualized magnitude of the point in time volatility, V_t. The parameter κ is treated as a daily quantity, measuring the rate of mean reversion in the annualized V_t per day.

5.3 Simulation Experiment

Prior to applying the realized volatility model to empirical data, a simulation exercise such as that documented in Chapter 4, is undertaken here to assess the predictive performance of the non-parametric method, with the extended Kalman filter used as the comparator. Using the generic notation in Chapter 3, the model is thus

$$y_t = \ln x_t + \sigma_\eta \eta_t \tag{5.7}$$

$$x_{t+1} = \alpha + \rho x_t + \sigma_v \sqrt{x_t} v_t \tag{5.8}$$

where $y_t = \ln RV_t$ and $x_t = V_t$. Values of $\sigma_\eta = 0.12$, $\alpha = 0.005$, $\rho = 0.92$ and $\sigma_v = 0.04$ were chosen with reference to typical empirical results relevant to the RV model, including those based on the S&P500 data analyzed in Section 5.4. The state error, v_t, is assumed to follow a truncated normal distribution to ensure that volatility is non-negative (i.e. $x_t > 0$) in the implementation of the algorithm. The truncation value associated with x_{t+1} is dependent on the value of the previous state, as reflected in the inequality,

$$v_t > \frac{(-\alpha - \rho x_t)}{\sigma_v \sqrt{x_t}}.$$

The non-parametric filter is initialized with $p(x_1)$ as the density associated with a normal variate, $N(\mu, \tau^2)$, truncated with $x_1 > 0$, and with $\mu = \frac{\alpha}{1-\rho}$ and $\tau^2 = \mu\sigma_v^2$. The filter is then implemented using the algorithm in Section 3.3, with

$$\begin{aligned} x_t^{*j} &= \exp\left(y_t - \sigma_\eta \eta^j\right) \\ \left|\frac{\partial h}{\partial x_t}\right|_{x_t = x_t^{*j}}^{-1} &= x_t^{*j}, \end{aligned}$$

and

$$p\left(x_{t+1}|x_t^{*j}\right) = \frac{\left(2\pi\sigma_v^2 x_t^{*j}(y_t,\eta^j)\right)^{-\frac{1}{2}}}{1-\Phi\left(\frac{-\alpha-\rho x_t^{*j}}{\sigma_v\sqrt{x_t^{*j}}}\right)} \exp\left\{-\frac{1}{2}\left(\frac{x_{t+1}-\left[\alpha+\rho x_t^{*j}\right]}{\sigma_v\sqrt{x_t^{*j}}}\right)^2\right\}.$$

As in the linear model, three different distributions for η_t are entertained, including normal, Student-t and skewed Student-t. The measurement error is standardized to have a mean of zero and variance equal to one (i.e. $\eta_t \sim i.i.d\,(0,1)$), and with the same values assigned to the degrees of freedom and skewness parameters as detailed in Section 4.4.1, and the same supports adopted for the purpose of integration.

The extended Kalman filter presented in Section (2.4.2), with the assumption of additive Gaussian measurement errors, is adopted as the basis for an alternative approach to estimating the forecast distribution. Referring to the RV model in (5.7) and (5.8), and comparing these expressions with the non-linear state space model in (2.22) and (2.23), the non-linear functions $h_t\left(x_t\right)$, $k_{1t}\left(x_t\right)$ and $k_{2t}\left(x_t\right)$ in (2.22) and (2.23), for the RV model, are given by

$$h_t\left(x_t\right) = \ln x_t \tag{5.9a}$$

$$k_{1t}\left(x_t\right) = \kappa\phi + \left(1-\kappa\right)x_t \tag{5.9b}$$

$$k_{2t}\left(x_t\right) = \sqrt{x_t}. \tag{5.9c}$$

Referring to the Taylor series expansions about the filtered mean $\widehat{a}_{t|t}$, and predicted conditional mean $\widehat{a}_{t|t-1}$, in (2.24)-(2.27), we have, for the RV model,

$$\begin{aligned} K_{1t} &= 1-\kappa && \text{(5.10a)} \\ \Psi_t &= \frac{1}{\widehat{a}_{t|t-1}} && \text{(5.10b)} \\ K_{2t} &= \sqrt{\widehat{a}_{t|t}}. && \text{(5.10c)} \end{aligned}$$

Substituting (5.9) and (5.10) into the general linearized model of (2.29) and (2.30) produces a linear, Gaussian model approximation of the RV model as

$$\begin{aligned} y_t &= \left(\ln \widehat{a}_{t|t-1} - 1\right) + \frac{x_t}{\widehat{a}_{t|t-1}} + \sigma_\eta \eta_t \\ x_{t+1} &= \kappa\phi + (1-\kappa)\, x_t + \sigma_v \sqrt{\widehat{a}_{t|t}} \nu_t, \end{aligned}$$

with η_t and ν_t both assumed to be normally distributed.

Referring to (2.34) and (2.35), the out-of-sample one-step-ahead predictive distribution $p\left(y_{T+1}|y_{1:T}\right)$, is approximated as a normal distribution with its respective mean and variance given as

$$\begin{aligned} E\left(y_{T+1}|y_{1:T}\right) &= h\left(\widehat{a}_{T+1|T}\right) \\ &= \ln \widehat{a}_{T+1|T} \end{aligned}$$

$$\begin{aligned} Var\left(y_{T+1}|y_{1:T}\right) &= \Psi_{T+1}^2 V_{T+1|T} + \sigma_\eta^2 \\ &= \frac{V_{T+1|T}}{\widehat{a}_{T+1|T}^2} + \sigma_\eta^2. \end{aligned}$$

5.3.1 Simulation Results

The simulation experiment undertaken here is similar to that performed in Chapter 4, with the RV model being simulated over $M = 1000$ replications, with $T = 1000$. $N = 11$ grid points were used in the support of the measurement error density, with the grid points evenly spaced. As with the simulation exercise in Chapter 4, the parameter values for the model in (5.7) and (5.8) (other than the density ordinates defining the measurement error in the non-parametric case) are fixed in the simulation exercise and take on values recorded in Section 5.3. Table 5.1 also records the distributional parameter values (if applicable) for the measurement error in each DGP, and the values of λ, c and ω in (3.27) used to ensure smoothness of the estimate of the measurement error distribution.

Table 5.1:
Constants, λ, c and ω, used in the penalized likelihood function in (3.27), in the simulation experiment for the RV model, as detailed in Section (5.3).

	η_t	λ	c	ω
RV model	$N(0,1)$	1.0	0.5	0.2
	$Student\ t(0,1,\nu=3)$	8.0	0.05	0.4
	$Skewed\ Student\ t(0,1,\nu=3,\gamma=3)$	4.0	0.5	0.2

Tables 5.2 to 5.4 record respectively the score, evaluation and coverage results for the RV model. The scores reported in Table 5.2 for the non-parametric estimate of $p\left(y_{T+1}|y_{1:T}\right)$ in the RV model are higher than those of the parametric estimate, under all DGPs. Despite the positive values of the relevant test statis-

tics, in the Gaussian case three of the non-parametric scores are insignificantly higher than those of the corresponding parametric alternatives, indicating that the extended Kalman filter approach works reasonably well under the (correct) assumption of conditional Gaussianity for the measurement distribution. Our statistically significant result for the logarithmic score for the Gaussian case, however, may be due to model misspecification for the extended Kalman filter, resulting from the fact that the true state transition density in this case is a truncated Gaussian density. Of course, this model misspecification becomes more pronounced under the (truncated) Student-t DGP, where the non-parametric estimate is significantly more accurate than the (misspecified) parametric estimate, according to three of the four scores, as well as in all four cases under the (truncated) skewed Student-t DGP.

The results in Table 5.3 show that, as is the case for the SCD model, there is an overall tendency for the non-parametric approach to yield more accurate forecasts in the RV model, according to the tests of fit. Specifically, the null hypothesis is rejected at the 5% level in the non-parametric case in only one case out of nine (and marginally at that), whilst five rejections (out of nine cases) occur for the extended Kalman filter-based alternative. With reference to Table 5.4, both forecast approaches have similar (and reasonable) coverage rates, apart from a significant underestimation of the nominal coverage in the upper tail on the part of the misspecified parametric approach, under both the symmetric and (positively) skewed Student-t DGPs.

Table 5.2:
Forecast comparison. Average scores for the parametric (Kalman filter based) and non-parametric estimates of $p\left(y_{T+1}|y_{1:T}\right)$ for the respective DGPs, with z values (see (4.5)) for the difference in scores across the competing forecasts reported. In the table, ** represents statistical significance at the 5% level for a one-sided test.

Estimated $p\left(y_{T+1}|y_{1:T}\right)$ for the RV Model (Section 5.3)

	Logarithmic Score			Quadratic Score		
η_t:	N	St	$SkSt$	N	St	$SkSt$
Kalman filter	0.01035	0.1282	0.08348	1.2160	1.3567	1.3394
Non-parametric	0.02564	0.1422	0.1026	1.2232	1.3783	1.3729
z-statistic	2.2647**	2.4188**	2.5861**	1.3683	2.0192**	2.7316**
	Spherical Score			Continuous Ranked Probability Score		
η_t:	N	St	$SkSt$	N	St	$SkSt$
Kalman filter	1.1013	1.1629	1.1558	-0.13462	-0.1204	-0.1243
Non-parametric	1.1047	1.1712	1.1691	-0.13447	-0.1196	-0.1232
z-statistic	1.4966	1.9258**	2.7818**	0.4573	1.4604	1.7499**

Table 5.3:
Forecast Evaluation. Pearson, LR and Jarque-Bera χ^2 test statistics, for the non-parametric (NP) and parametric (KF) estimates of $p\left(y_{T+1}|y_{1:T}\right)$ for the respective DGPs. In the table, ** represents statistical significance at the 5% level. The critical values for the three tests are respectively 30.14, 7.82 and 5.99.

	Pearson		**LR**		**Jarque-Bera**	
	NP	KF	NP	KF	NP	KF
$\eta_t \sim N(0,1)$	21.28	37.32**	8.347**	13.284**	1.043	36.499**
$\eta_t \sim St(0,1,\nu=3)$	24.72	30.04	3.019	5.398	0.983	10.752**
$\eta_t \sim SkSt(0,1,\nu=3,\gamma=3)$	16.40	24.96	3.321	2.847	3.385	39.216**

Table 5.4:
Forecast Evaluation. Coverage rates (5% and 95%) for the parametric (Kalman filter based) and non-parametric estimates of $p\left(y_{T+1}|y_{1:T}\right)$ for the respective DGPs. In the table, ** represents significant difference from the nominal coverage, at the 5% significance level.

	5% lower tail			**5% upper tail**			**95% HPD**		
η_t:	N	St	$SkSt$	N	St	$SkSt$	N	St	$Skst$
Kalman filter	6.1	5.6	6.0	5.2	3.0**	3.3**	93.4	95.6	94.4
Non-parametric	5.3	4.7	5.8	5.5	4.3	4.0	93.8	95.7	95.0

As with the linear and SCD models, a sensitivity analysis is also performed to examine the effect of the number of grid points, N, on the estimate of the one-step-ahead predictive distribution, $p\left(y_{T+1}|y_{1:T}\right)$, in the RV model. Figure 5.1 plots the estimates of $p\left(y_{T+1}|y_{1:T}\right)$ for the RV model with Gaussian, Student-t and skewed Student-t error distributions, and with N varying from 11 to 51. It can be seen that estimates of the predictive distributions obtained from different values of N are very similar with one another. Hence, the results are considered to be robust to the number of grid points. As the number of grid points chosen corresponds to the number of unknown probabilities to be estimated, the computational requirements of the simulation experiment led to the use of a value of N at the lower end of the range considered.

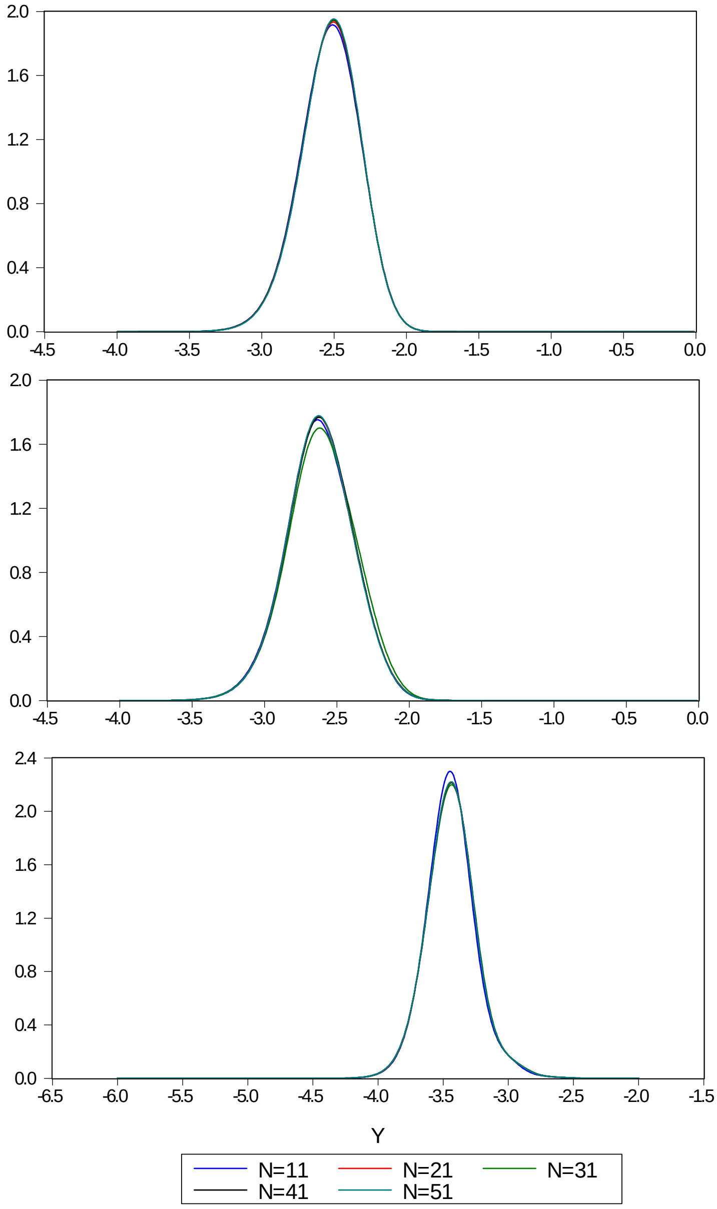


Figure 5.1: Estimated one-step-ahead predictive distribution of the RV model for varying number of grid points, N. The figure shows (from top to bottom), $p\left(y_{T+1}|y_{1:T}\right)$, for the Gaussian, Student-t and skewed Student-t DGPs, with N ranging from 11 to 51.

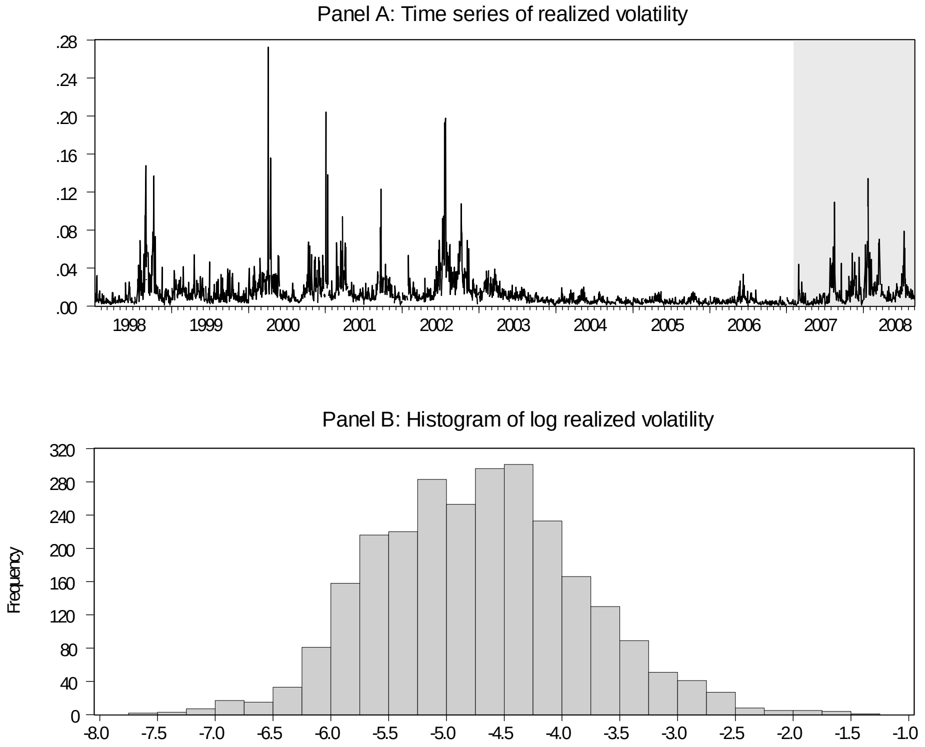
5.4 Empirical Illustration: Realized Volatility of the S&P500 Index

5.4.1 Preliminary Analysis

In order to illustrate the non-parametric method, non-parametric estimates of the one-step-ahead prediction distributions for realized volatility of the S&P500 market index are produced and evaluated, based on the model described in (5.7) and (5.8). The sample period extends from 2 January 1998 to 29 August 2008, providing a total of 2645 daily observations. All index data has been supplied by the Securities Industries Research Centre of Asia Pacific (SIRCA) on behalf of Reuters, with the raw index data having been cleaned using the methods of Brownlees and Gallo (2006).[1]

The time series of the data is plotted in Panel A of Figure 5.2. As is clear from that figure, there are several distinct periods in which volatility is seen to be significantly higher than during the remaining sample period. The first of these periods corresponds to the Asian currency crisis in 1998, when a financial crisis gripped much of Asia and raised fears of a worldwide economic slowdown. Realized volatility also reached high levels at the end of year 2000, following the burst of the 'Dot-com' bubble, and in year 2001 after the September 11th terrorist attacks in the United States. Year 2002 produced record values of re-

[1] The candidate would like to acknowledge the assistance of Chris Tse in producing the empirical realized variance series. The realized variation measure is based on fixed five minute sampling, with a 'nearest price' method used to construct artificial returns five minutes apart. Subsampling (or averaging) over the day is also used, in order to mitigate some of the effects of microstructure noise. See Martin *et al.* (2009) for details of such computations.



Panel A: Time series of realized volatility

Panel B: Histogram of log realized volatility

Figure 5.2: Time series of realized volatility and histogram of logarithmic realized volatility of S&P500 market index from 2 January 1998 to 29 August 2008.

alized volatility caused by a sharp drop in stock prices, generally viewed as a market correction to over-inflated prices following a decade-long ‘bull’ market. Also factoring in the speed of the fall in prices at this time were a series of large corporate collapses (e.g. Enron and WorldCom), prompting many corporations to revise earnings statements, and causing a general loss of investor confidence. The final period of high volatility in our sample corresponds to the year 2008, associated with the ‘global financial crisis’, triggered by the sub-prime mortgage defaults in the United States. During all of these periods, the peaks reached by the realized volatility values were between ten and twenty times larger than the average level over the full sample period. In contrast, there was a relatively long period of time, from 2003 to mid-2007, during which volatility was relatively stable and low. Panel B of Figure 5.2 plots the histogram of logarithmic realized volatility, with the distinct skewness to the right reflecting the occurrence of the very extreme values of realized volatility itself. A Jarque-Bera test applied to this logarithmic realized volatility series rejects the null hypothesis of Gaussianity at any conventional level of significance. These empirical characteristics are consistent with the existence of a jump diffusion model for the stock prices index, with realized volatility reflecting both diffusive and jump variation as a consequence. In using the non-parametric approach to estimate the forecast distribution for logarithmic realized volatility, the aim is to capture the impact of jump variation in a computationally simple way, rather than modelling price jumps explicitly.

5.4.2 Empirical Results

The S&P500 daily realized volatility data is divided into two subsamples. A portion of the sample (2 January 1998 to 30 January 2007), containing 2245 observations, is reserved for estimation of the model parameters in (5.7) and (5.8), including the unknown ordinates of $p(\eta)$. The sample used for forecast assessment comprises the remaining 400 realized volatility values, covering the period from 31 January 2007 to 29 August 2008, and is represented by the shaded area in Panel A of Figure 5.2. This second sample corresponds to the early period of the financial crisis, during which defaults on sub-prime mortgages began to impact on the viability of financial institutions and the availability of credit. The out-of-sample density forecasts are based on (parameter) estimates updated as the estimation window expands, incorporating each new daily observation within the second sample period. $N = 21$ grid points, equally spaced over the interval from -10 to 10, are used to represent the support of the measurement error density, $p(\eta)$, for all 400 forecast distributions. Values of the penalty parameters used in (3.27) are $\lambda = 4$, $c = 0.5$ and $\omega = 0.3$.[2]

For each of the 400 estimated forecast distributions, simulated draws of $\ln RV_{T+1}$ (in terms of which the measurement equation is specified) are exponentiated to produce future values of RV_{T+1}, with these values then used to produce a sequence of 95% prediction intervals for the evaluation period in Figure 5.3.

[2] Robustness of the empirical results to different sets of penalty values ($1 \leq \lambda \leq 4$; $0.01 \leq c \leq 0.5$; $0.3 \leq \omega \leq 0.8$) for a fixed N was investigated. Differences in the estimated forecast distributions were negligible.

The solid line represents the observed RV_t at each point t in the evaluation period, while the dotted lines represent the 2.5% and 97.5% predictive bounds. The empirical coverage for the evaluation period is 94.0%, insignificantly different from the nominal level of 95% and providing, thereby, extremely strong support for the overall accuracy of the non-parametric approach. Support is also provided via the Pearson test for uniformity of the probability integral transform series, u_{T+1} in (4.6). However, both the LR test of the normality (and independence) of $\left\{\omega_{T+1}^{i}, i = 1, 2, ..., M\right\}$ in (4.7) and the Jarque-Bera test for the normality of $\left\{\omega_{T+1}^{i}, i = 1, 2, ..., M\right\}$ in (4.7) lead to rejection, indicating that some aspect of the forecast distribution is not being adequately captured. Observation of the shape of the histogram of u_{T+1} in Figure 5.4 shows a relatively smooth and uniform shape, as is consistent with the support of the Pearson test, but that a few very extreme realizations are unable to be captured by the right tail of the estimated forecast density.

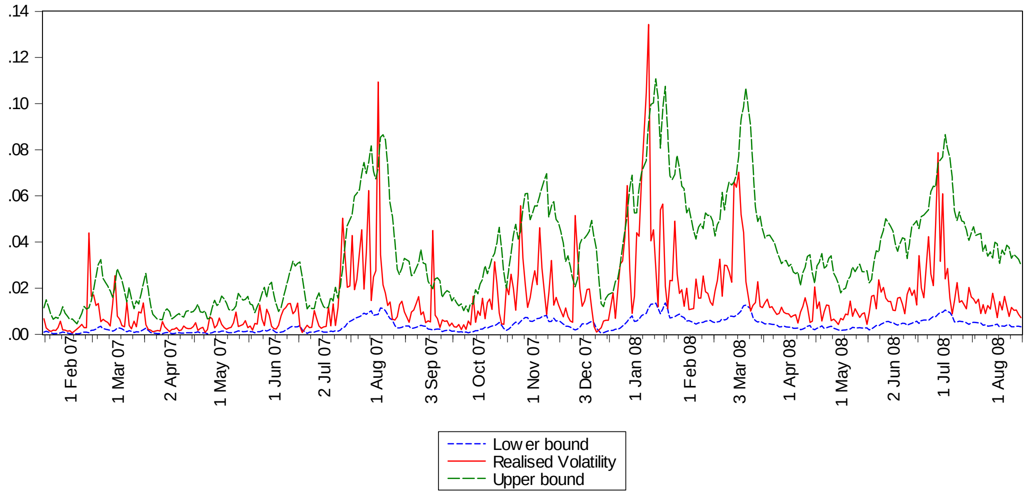


Figure 5.3: 95% one-step-ahead prediction intervals and the observed realized volatility, over the 400 day evaluation period (31 January 2007 to 29 August 2008).

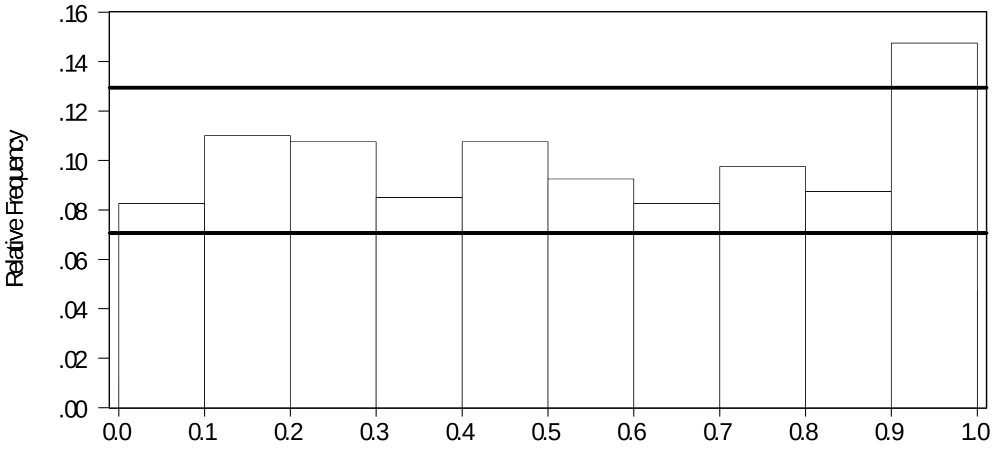


Figure 5.4: Histogram of the probability integral transform series, u, for the realized volatility model. The horizontal lines superimposed on the histogram are approximate 95% confidence intervals for the individual bin heights under the null that u is $i.i.d.$ $U\left(0,1\right)$.

5.4.3 Measuring Sampling Error

Finally, in the context of producing estimates of forecast distribution that are conditional on estimates of the fixed parameters, it is of interest to consider the issue of sampling error and the appropriate measurement thereof. In the spirit of McCabe *et al.* (2011), the subsampling approach of Politis, Romano and Wolf (1999) is used to quantify sampling variation in a single estimated one-step-ahead forecast distribution, for 17th March 2008, a day with quite high volatility during the out-of-sample period. The technique mimics the conventional prediction interval for a scalar point forecast, but ensures, at the same time, that the integration to unity property of the forecast distribution still holds.[3] The steps of the procedure are as follows:

1. Obtain $T-b+1$ subsamples $Y_1 = (y_1, \ldots, y_b)\,, Y_2 = (y_2, ..., y_{b+1}), ..., Y_{T-b+1} = (y_{T-b+1}, ..., y_T)$ from the set of empirical data, $y_{1:T} = (y_1, y_2, ..., y_T)'$.

2. Use the proposed non-parametric ML method to produce an estimate of θ, $\hat{\theta}_{b,t}$, computed from Y_t, for $t = 1, 2, ..., T - b + 1$.

3. Use $\hat{\theta}_{b,t}$ and the *observed* values, $y_{1:T}$, to compute the one-step-ahead forecast distribution $p\left(y_{T+1}|y_{1:T}, \hat{\theta}_{b,t}\right)$.

[3] As mentioned earlier in Chapter 1, Rodriguez and Ruiz (2009) present a bootstrap-based approach to estimating prediction intervals in a linear state space setting. Their method uses the Kalman filter recursions, but eschews the assumption of Gaussian innovations by using random draws from the empirical distributions of the innovations. It also factors sampling variation into the prediction intervals, but in a different way from that proposed by McCabe *et al.* (2011) and followed in this chapter. See also Pascual *et al.* (2001, 2006). Jung and Tremayne (2006) use a block bootstrapping technique to cater for parameter uncertainty in the estimation of forecast distributions for discrete count data, in an INAR setting.

4. Calculate (over an arbitrarily fine grid of values for y_{T+1}) the metric $d_{b,t} = \sqrt{T}\left\| p\left(y_{T+1}|y_{1:T},\hat{\theta}_{b,t}\right) - p\left(y_{T+1}|y_{1:T},\hat{\theta}\right)\right\|_1$, where $p\left(y_{T+1}|y_{1:T},\hat{\theta}\right)$ is the estimated forecast distribution based on the empirical data and $\hat{\theta}$ is the empirical estimate of θ.

5. Find the 95^{th} percentile of $\{d_{b,1},\ldots,d_{b,T-b+1}\}$, $d_b^{0.95}$, and the corresponding distribution $p_{0.95}\left(y_{T+1}|y_{1:T},\cdot\right)$. Then, relative to the replicated distributions and in terms of the $||.||_1$ distance from $p\left(y_{T+1}|y_{1:T},\hat{\theta}\right)$, the chances of seeing a distribution as or more 'extreme' than $p_{0.95}\left(y_{T+1}|y_{1:T},\cdot\right)$ is 5%.

The data-dependent method used to choose the size of the sub-samples, b (see Politis *et al.*, 1999, Chapter 9) is as follows:

a. For each $b \in \{b_{small},\ldots,b_{big}\}$ carry out Steps 1 to 5 above to compute $d_b^{0.95}$.[4]

b. For each b compute VI_b as the standard deviation of the $2k+1$ adjacent values $\left\{d_{b-k}^{0.95},\ldots,d_{b+k}^{0.95}\right\}$ (for $k=2$).

c. Choose $\hat{b}$ to minimise VI_b.[5]

Figure 5.5 shows the 10th, 50th and 95th percentile sub-sampled forecast distributions, along with the empirical forecast distribution, for the 17th March

[4] In the spirit of McCabe et al. (2011), we chose values for b_{small} and b_{big} that encompassed a value equivalent to half of the sample size being considered. For the smaller sample, with 505 observations, b_{small} and b_{big} were chosen to be 220 and 275 respectively, while b_{small} and b_{big} were chosen to be 900 and 1600 respectively for the larger sample with 2528 observations.

[5] $d_b^{0.95}$ has been chosen as the percentile on which selection of b is based, as the interest primarily rests in ascertaining the changes in the forecast distributions that may occur at the extreme end of the scale (of the metric d).

2008. Panel A shows the relevant results based on a sample size of 505 observations (approximately two trading years) with $\hat{b} = 255$. Panel B shows the results based on 2528 observations, with $\hat{b} = 1300$. As is clear, for the smaller sample size, there is a large amount of uncertainty in the predictive estimate, with that uncertainty serving to shift probability mass across the support of the predictive distribution. For example, the predictive distribution at the 50th percentile assigns a larger probability to extreme values of volatility, than does the actual empirical estimate. On the other hand, the predictive distribution at the 95th percentile assigns large probabilities to very *low* values of volatility. In other words, for the smaller sample size sampling variability has a substantial impact, serving to alter the qualitative nature of conclusions drawn about future volatility. For the larger sample size, the subsampled-based sampling distribution of the (estimated) forecast distribution becomes much more concentrated around the empirical estimate, with the full suite of distributions leading to qualitatively similar conclusions regarding volatility on the given day.

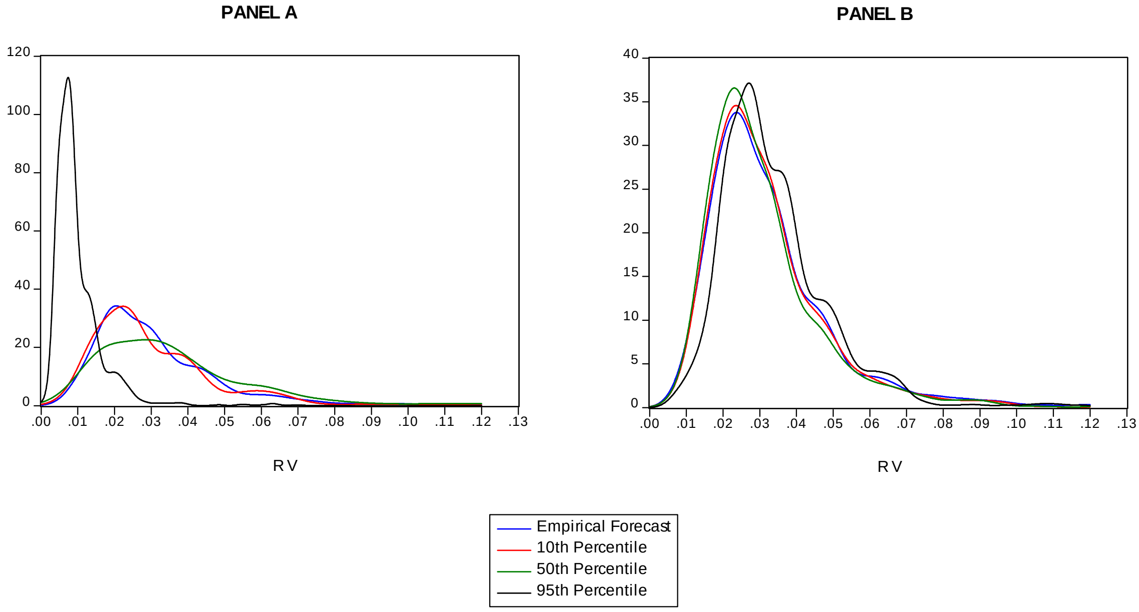


Figure 5.5: Plot of the 10th, 50th and 95th percentile bootstrap forecast distributions against the empirical forecast distribution for 17 March 2008. Panel A shows the one-step-ahead forecast distributions estimated from the preceding 505 observations. Panel B shows the one-step-ahead forecast distributions estimated from the preceding 2528 observations.

5.5 Conclusions

This chapter contributes to the current literature on realized volatility forecasting by developing a state space model to which the non-parametric filter is applied to obtain distributional forecasts. The non-parametric approach aims to capture the distributional properties of the measurement error which, in turn, contains the effects of all factors not explictly modelled. In principle, all unmodeled effects will be thereby reflected in the estimated forecast distributions.

The forecasting accuracy of the non-parametric approach and the extended Kalman filter-based parametric approach is assessed using simulated data. Three different DGPs are entertained for the simulation exercise. For all three cases, the non-parametric estimates of the forecast distribution are shown to be accurate, exhibiting the non-parametric filter's ability to perform well in this particular setting. The non-parametric estimate of the forecast distribution is also robust to the number of grid points on the measurement error density. When applied to the S&P 500 data, the proposed non-parametric methodology is shown to produce forecast distributions with excellent overall accuracy, only failing to fully capture the most extreme values of realized volatility that occured on a few occasions. A resampling method is used to highlight the effect that sampling variation can have on predictive conclusions, in small samples in particular.

Chapter 6

Extensions

6.1 Introduction

The non-parametric filter developed in Chapter 3 is a grid-based approach that yields a non-parametric estimate of the one-step-ahead forecast density, $p\left(y_{T+1}|y_{1:T}\right)$. The algorithm utilizes the properties of the δ-function to effectively switch the integration problem from x_t to η_t, where the latter has, by assumption, a support that is constant over t. The algorithm as outlined in Chapter 3 depends on the assumption that the function $G_t\left(x_t\right) = y_t - h_t\left(x_t, \eta_t\right)$ in (3.3) has a unique root at $x_t = x_t^*(y_t, \eta_t)$, and that this root is analytically available. The application of the algorithm has also been confined to univariate state space models thus far.

This chapter will expand the non-parametric methodology in four ways. First, with the emphasis having been on producing the out-of-sample one-step-ahead forecast distribution, Section 6.2 shows how the filter can be used to obtain a multi-step-ahead distribution. Second, the dual assumptions of the existence of a unique root, and the unique root being analytically available, will be examined. Specifically, Section 6.3 derives the grid-based filter in the case of multiple solutions to $G_t\left(x_t\right)$, while Section 6.4 discusses the issue of non-analytic roots. Third, with the non-parametric filter having been applied only to univariate state space models in prior chapters, the extension of the filter to a multivariate setting is demonstrated in Section 6.5. Finally, Section 6.6 explores the case where the grid-based filtering method is replaced by a Monte Carlo simulation-based method when the measurement error has a parametric specification that can be simulated from.

6.2 Multi-step-Ahead Forecast Distributions

In this section we demonstrate how the out-of-sample multi-step-ahead predictive distribution for the model in (3.1) and (3.2) can be computed using the non-parametric filter. The summary of the non-parametric filter for general t produced under the assumptions detailed in Chapter 3 is reproduced here for convenience:

$$\begin{aligned}
p(x_{t+1}|y_{1:t}) &= \sum_{j=1}^{N} W_t^j p\left(x_{t+1}|x_t^{*j}\right) \\
p\left(y_{t+1}|y_{1:t}\right) &= \sum_{i=1}^{N} M_{t+1}^i\ \left(y_{t+1}\right) p(x_{t+1}^*(y_{t+1},\eta^i)|y_{1:t}) \\
p\left(x_{t+1}|y_{1:t+1}\right) &= \sum_{j=1}^{N} W_{t+1}^j \delta\left(x_{t+1}-x_{t+1}^{*j}\right),
\end{aligned}$$

with

$$M_{t+1}^i\left(y_{t+1}\right) = mp\left(\eta^i\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^*(y_{t+1},\eta^i)}^{-1}$$

and

$$W_{t+1}^{j}=\frac{M_{t+1}^{j}\left(y_{t+1}\right)p\left(x_{t+1}^{*j}|y_{1:t}\right)}{\sum_{i=1}^{N}M_{t+1}^{i}\left(y_{t+1}\right)p\left(x_{t+1}^{*i}|y_{1:t}\right)}.$$

For multi-step-ahead forecast horizons, the prediction distribution for the observation in period $T+s$, with $s>1$, is

$$\begin{aligned}p\left(y_{T+s}|y_{1:T}\right) &= \int p\left(y_{T+s},x_{T+s}|y_{1:T}\right)\ dx_{T+s}\\ &= \int p\left(y_{T+s}|x_{T+s}\right)p\left(x_{T+s}|y_{1:T}\right)\ dx_{T+s},\end{aligned}$$

where the state distribution at a future period $T+s$, is obtained by repeated application of the transition recursion

$$p\left(x_{T+s}|y_{1:T}\right)=\int p\left(x_{T+s}|x_{T+s-1}\right)p\left(x_{T+s-1}|y_{1:T}\right)\ dx_{T+s-1}.$$

Given that

$$p(x_{T+1}|y_{1:T})=\sum_{j=1}^{N}W_{T}^{j}p\left(x_{T+1}|x_{T}^{*j}\right),$$

and with

$$W_{T+1}^{j}=\frac{M_{T+1}^{j}\left(y_{T+1}\right)p\left(x_{T+1}^{*j}|y_{1:T}\right)}{\sum_{i=1}^{N}M_{T+1}^{i}\left(y_{T+1}\right)p\left(x_{T+1}^{*i}|y_{1:T}\right)},$$

then, for example

$$\begin{aligned}p\left(x_{T+2}|y_{1:T}\right) &= \sum_{j=1}^{N}W_{T}^{j}\int p\left(x_{T+2}|x_{T+1}\right)p\left(x_{T+1}|x_{T}^{*j}\right)\ dx_{T+1},\\ p\left(x_{T+3}|y_{1:T}\right) &= \sum_{j=1}^{N}W_{T}^{j}\int p\left(x_{T+3}|x_{T+2}\right)\int p\left(x_{T+2}|x_{T+1}\right)p\left(x_{T+1}|x_{T}^{*j}\right)\ dx_{T+1}dx_{T+2},\end{aligned}$$

and so on. Once the state forecast distribution for period $T+s$ has been obtained, the corresponding prediction distribution for the observation in period $T+s$ is

$$\begin{aligned} p\left(y_{T+s}|y_{1:T}\right) &= \int p\left(y_{T+s}|x_{T+s}\right) p\left(x_{T+s}|y_{1:T}\right)\ dx_{T+s} \\ &= \int\int p\left(\eta\right)\delta\left(y_{T+s}-h(x_{T+s},\eta)\right)d\eta\ p\left(x_{T+s}|y_{1:T}\right)\ dx_{T+s}. \end{aligned} \tag{6.1}$$

If an analytical expression for $p\left(x_{T+s}|y_{1:T}\right)$ is available, (6.1) can first be integrated with respect to x_{T+s}, yielding

$$p\left(y_{T+s}|y_{1:T}\right) = \int_{-\infty}^{\infty} p\left(\eta\right)\left|\frac{\partial h}{\partial x_{T+s}}\right|^{-1}_{x_{T+s}=x^*_{T+s}(y_{T+s},\eta)} p\left(x^*_{T+s}\left(y_{T+s},\eta\right)|y_{1:T}\right)\ d\eta.$$

Subsequently, with numerical integration over the grid points $\left\{\eta^1,\eta^2,...,\eta^N\right\}$, we obtain

$$p\left(y_{T+s}|y_{1:T}\right) \approx m\sum_{i=1}^{N} p\left(\eta^i\right)\left|\frac{\partial h}{\partial x_{T+s}}\right|^{-1}_{x_{T+s}=x^*_{T+s}(y_{T+s},\eta^i)} p\left(x^*_{T+s}\left(y_{T+s},\eta^i\right)|y_{1:T}\right).$$

If, however, an analytical expression for $p\left(x_{T+s}|y_{1:T}\right)$ is not available, a simulation based method for obtaining $p\left(y_{T+s}|y_{1:T}\right)$ may be used. This can be implemented by drawing R replicates of x_{T+s}, obtained first by sampling x_T from the final filtered distribution $p\left(x_T|y_{1:T}\right)$, which assigns positive probabilities only to the values $x^{*j}_T = x^*_T\left(y_T,\eta^j\right)$ for $j=1,2,...,N$, and then via s sequential draws from the transition distribution given in (3.2). Each of the resulting R draws of x_{T+s}, denoted by $x^{(r)}_{T+s}$, could then be used to simulate an 'observed' value $y_{T+s}=y^{(r)}_{T+s}$ using (3.1) to produce

$$y^{(r)}_{T+s} = h\left(x^{(r)}_{T+s},\eta^{(r)}_{T+s}\right),$$

with $\eta_{T+s}^{(r)}$ drawn from the pmf implied by the non-parametric estimate of $p(\eta)$. The s-step-ahead forecast density, $p(y_{T+s}|y_{1:T})$, may then be obtained by applying a kernel density estimator to $y_{T+s}^{(r)}$, for $r = 1, 2, ..., R$.[1]

6.3 The Case of Multiple Roots of $G_t(x_t)$

In the development of the algorithm in Chapter 3, we invoke the assumption that for given values y_t and η_t, the function

$$G(x_t) = y_t - h(x_t, \eta_t) \tag{6.2}$$

has a unique root at $x_t = x_t^*(y_t, \eta_t)$, as well as having a non-zero derivative at that root. We now relax the unique root assumption and show how the non-parametric filter can be adapted to the case when (6.2) has multiple roots at $x_t = x_t^{*i}(y_t, \eta_t)$, for $i = 1, 2, ..., K$.

6.3.1 Preliminaries

The δ-function satisfies the following relation,

$$\int_{-\infty}^{\infty} f(z)\delta(z^* - z)\, dz = \sum_{i=1}^{K} f(z^{*i}), \tag{6.3}$$

for any continuous, real-valued function $f(.)$. Note z^{*i} for $i = 1, 2, ..., K$ are the roots of the argument of the δ-function, with the case of $K > 1$ now being explicitly allowed for. Further, denoting by $\delta(G(z))$ the composite function in

[1] One would expect the multi-step-ahead forecast distribution, being conditional only on current information up to time T, to become wider as s increases due to increased uncertainty of the forecasts. Therefore, a larger value of R may potentially be needed for a higher value of s, in order for a wider spread of possible values of y_{T+s} to be covered.

which $G(z)$ is a differentiable function with multiple roots at $z^{*i}, i = 1, 2, ..., K$, a transformation of variable yields

$$\int_{-\infty}^{\infty} f(z)\delta(G(z))\,dz = \int_{-\infty}^{\infty} f(z)\left|\frac{\partial G(z)}{\partial z}\right|^{-1} \delta(z - z^*)\,dz, \tag{6.4}$$

resulting, via (6.3), in

$$\int_{-\infty}^{\infty} f(z)\delta(G(z))\,dz = \sum_{i=1}^{K} f(z^{*i})\left|\frac{\partial G(z)}{\partial z}\right|_{z=z^{*i}}^{-1}, \tag{6.5}$$

where $\left|\frac{\partial G(z)}{\partial z}\right|_{z=z^{*i}}^{-1}$ denotes the modulus of the derivative of $G(z)$ evaluated at $z = z^{*i}$, for $i = 1, 2, ..., K$.

The Dirac delta function also satisfies the following relation,

$$\delta(G(z)) = \sum_{i=1}^{K} \left|\frac{\partial G(z)}{\partial z}\right|_{z=z^{*i}}^{-1} \delta\left(z - z^{*i}\right), \tag{6.6}$$

when considering the composite function $\delta(G(z))$ explicitly in terms of z. In what follows, $G(x_t) = y_t - h(x_t, \eta_t)$ and, hence,

$$\left|\frac{\partial G(x_t)}{\partial x_t}\right| = \left|\frac{\partial h}{\partial x_t}\right|,$$

and accordingly

$$\delta(y_t - h(x_t, \eta_t)) = \sum_{i=1}^{K} \left|\frac{\partial h}{\partial x_t}\right|_{x_t = x_t^{*i}(y_t, \eta_t)}^{-1} \delta\left(x_t - x_t^{*i}(y_t, \eta_t)\right).$$

6.3.2 The Initial Filtered Distribution: $p(x_1|y_1)$

Using the δ-function representation of the measurement density in (3.11), it follows that the filtered density of the state variable at time $t = 1$ may be expressed as

$$\begin{aligned} p\left(x_1|y_1\right) &= \frac{p\left(x_1\right) p\left(y_1|x_1\right)}{p(y_1)} \\ &= \frac{p\left(x_1\right) \int_{-\infty}^{\infty} p\left(\eta\right)\ \delta\left(y_1 - h(x_1,\eta)\right) d\eta}{\int_{-\infty}^{\infty} p\left(x_1\right) \left[\int_{-\infty}^{\infty} p\left(\eta\right)\ \delta\left(y_1 - h(x_1,\eta)\right) d\eta\right] dx_1}. \end{aligned}$$

The expression for $p\left(x_1|y_1\right)$ is then simplified by first re-writing $\delta\left(y_1 - h(x_1,\eta)\right)$ using (6.6). Then, the order of integration is reversed and (6.4) and (6.5) used in the denominator to obtain

$$p\left(x_1|y_1\right) = \frac{p\left(x_1\right) \int_{-\infty}^{\infty} p\left(\eta\right) \sum_{i=1}^{K} \left|\frac{\partial h}{\partial x_1}\right|_{x_1 = x_1^{*i}(y_1,\eta)}^{-1} \delta\left(x_1 - x_1^{*i}(y_1,\eta)\right) d\eta}{\int_{-\infty}^{\infty} p\left(\eta\right) \sum_{i=1}^{K} p\left(x_1^{*i}(y_1,\eta)\right) \left|\frac{\partial h}{\partial x_1}\right|_{x_1 = x_1^{*i}(y_1,\eta)}^{-1} d\eta}, \qquad (6.7)$$

where $x_1^{*i}(y_1,\eta)$ for $i = 1, 2, ..., K$ are the solutions to $y_1 - h(x_1,\eta) = 0$ for any value η in the support of $p\left(\eta\right)$.

Next, to numerically evaluate the filtered distribution in (6.7) via rectangular integration, an evenly spaced grid $\left\{\eta^1, \eta^2, ..., \eta^N\right\}$ is defined, with interval length m, resulting in the approximation for $p\left(x_1|y_1\right)$ given by

$$p\left(x_1|y_1\right) \approx \frac{p\left(x_1\right) \sum_{j=1}^{N} \sum_{i=1}^{K} m\, p\left(\eta^j\right) \left|\frac{\partial h}{\partial x_1}\right|_{x_1 = x_1^{*i}(y_1,\eta^j)}^{-1} \delta\left(x_1 - x_1^{*i}\left(y_1,\eta^j\right)\right)}{\sum_{j=1}^{N} \sum_{i=1}^{K} m\, p\left(x_1^{*i}(y_1,\eta^j)\right) p\left(\eta^j\right) \left|\frac{\partial h}{\partial x_1}\right|_{x_1 = x_1^{*i}(y_1,\eta^j)}^{-1}},$$

where $p\left(\eta^j\right)$ is defined as the unknown density ordinate associated with the grid-point indexed by j. Note that conveniently using the numerical integration approach in the numerator as well as in the denominator serves to produce multiple implied states, $x_1^{*ij} = x_1^{*i}(y_1,\eta^j)$, associated with each η^j, such that the first filtered distribution has representation (up to numerical approximation error) as a discrete distribution, with density

$$p\left(x_1|y_1\right)=\sum_{j=1}^{N}\sum_{i=1}^{K}W_1^{ij}\delta\left(x_1-x_1^{*ij}\right), \tag{6.8}$$

and where

$$W_1^{ij}=\frac{p\left(\eta^j\right)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^{*i}(y_1,\eta^j)}^{-1}p\left(x_1^{*i}\left(y_1,\eta^j\right)\right)}{\sum_{j=1}^{N}\sum_{i=1}^{K}p\left(\eta^j\right)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^{*i}(y_1,\eta^j)}^{-1}p\left(x_1^{*i}\left(y_1,\eta^j\right)\right)}, \tag{6.9}$$

for $i=1,2,...,K$ and $j=1,2,...,N$. Implicit in this approximation to the first filtered state density is the first likelihood contribution,

$$p\left(y_1\right)=m\sum_{j=1}^{N}\sum_{i=1}^{K}p\left(\eta^j\right)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^{*i}(y_1,\eta^j)}^{-1}p\left(x_1^{*i}\left(y_1,\eta^j\right)\right), \tag{6.10}$$

obtained from approximating the denominator in (6.7).

Having obtained the representation in (6.8) for time $t=1$, it will be shown that for any time $t=2,3,...T$, an appropriate discrete distribution can be found to approximate the filtered distribution

$$p\left(x_t|y_{1:t}\right)=\sum_{j=1}^{N}\sum_{i=1}^{K}W_t^{ij}\delta\left(x_t-x_t^{*ij}\right), \tag{6.11}$$

where the iteratively determined weights satisfy

$$\sum_{j=1}^{N}\sum_{i=1}^{K}W_t^{ij}=1,$$

and each state grid location

$$x_t^{*ij}=x_t^{*i}\left(y_t,\eta^j\right)$$

is determined by the i^{th} solution of $y_t-h_t\left(x_t,\eta^j\right)$, for $i=1,2,...,K$ and $j=1,2,...,N$.

6.3.3 The Predictive Distribution for the State: $p\left(x_{t+1}|y_{1:t}\right)$

Assuming (6.11) holds in period t, it follows that the one-step-ahead state prediction density is a mixture of transition densities, since

$$\begin{aligned} p\left(x_{t+1}|y_{1:t}\right) &= \int p\left(x_{t+1}|x_t\right) p\left(x_t|y_{1:t}\right) dx_t \\ &= \int p\left(x_{t+1}|x_t\right) \sum_{j=1}^{N}\sum_{i=1}^{K} W_t^{ij} \delta\left(x_t - x_t^{*ij}\right) dx_t \\ &= \sum_{j=1}^{N}\sum_{i=1}^{K} W_t^{ij} \int p\left(x_{t+1}|x_t\right) \delta\left(x_t - x_t^{*ij}\right) dx_t \\ &= \sum_{j=1}^{N}\sum_{i=1}^{K} W_t^{ij}\ p\left(x_{t+1}|x_t^{*ij}\right), \qquad (6.12) \end{aligned}$$

for $t = 1, 2, ..., T$. The notation $p\left(x_{t+1}|x_t^{*ij}\right)$ denotes the transition density of $p\left(x_{t+1}|x_t\right)$, viewed as a function of x_{t+1} and given the fixed value of $x_t = x_t^{*ij}$. As it is assumed that the transition densities $p\left(x_{t+1}|x_t\right)$ are available, no additional approximation is needed in moving from $p\left(x_t|y_{1:t}\right)$ to $p\left(x_{t+1}|y_{1:t}\right)$.

6.3.4 The One-step-ahead Predictive Distribution for the Observed: $p\left(y_{t+1}|y_{1:t}\right)$

Having obtained a representation for the filtered density for the future state variable, x_{t+1}, the corresponding predictive density for the next observation is given by

$$p\left(y_{t+1}|y_{1:t}\right) = \int_{-\infty}^{\infty} p\left(y_{t+1}|x_{t+1}\right) p\left(x_{t+1}|y_{1:t}\right) dx_{t+1}.$$

Utilizing (3.11) for $p\left(y_{t+1}|x_{t+1}\right)$, the one-step-ahead prediction density has representation

$$p\left(y_{t+1}|y_{1:t}\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p\left(\eta\right) \delta\left(y_{t+1} - h(x_{t+1}, \eta)\right) d\eta\, p\left(x_{t+1}|y_{1:t}\right) dx_{t+1},$$

which, after integration with respect to x_{t+1} (and using (6.5) once again), yields

$$p\left(y_{t+1}|y_{1:t}\right)=\int_{-\infty}^{\infty}p\left(\eta\right)\sum_{i=1}^{K}\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*i}(y_{t+1},\eta)}^{-1}p(x_{t+1}^{*i}(y_{t+1},\eta)|y_{1:t})d\eta.$$

Invoking again the pre-specified grid of values for η, the one-step-ahead prediction density (up to numerical approximation error) is,

$$p\left(y_{t+1}|y_{1:t}\right)=m\sum_{j=1}^{N}\sum_{i=1}^{K}p\left(\eta^{j}\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*i}(y_{t+1},\eta^{j})}^{-1}p\left(x_{t+1}^{*i}(y_{t+1},\eta^{j})|y_{1:t}\right). \tag{6.13}$$

Given that $p\left(x_{t+1}^{*i}(y_{t+1},\eta^{j})|y_{1:t}\right)$ in (6.13) denotes the one-step-ahead prediction density from (6.12) evaluated at $x_{t+1}=x_{t+1}^{*i}\left(y_{t+1},\eta^{j}\right)$, it can be seen that $p\left(y_{t+1}|y_{1:t}\right)$ is computed as an $(N^{2}\times K^{2})$ mixture of (specified) transition density functions as a consequence.

6.3.5 The Updated Filtered Distribution: $p\left(x_{t+1}|y_{1:t+1}\right)$

Finally, the predictive distribution for the state at time $t+1$ is updated given the realization y_{t+1} as

$$\begin{aligned}p\left(x_{t+1}|y_{1:t+1}\right) &= \frac{p\left(y_{t+1}|x_{t+1}\right)p\left(x_{t+1}|y_{1:t}\right)}{p\left(y_{t+1}|y_{1:t}\right)}\\ &\approx \frac{m\sum_{j=1}^{N}\sum_{i=1}^{K}p\left(\eta^{j}\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*i}(y_{t+1},\eta^{j})}^{-1}\delta\left(x_{t+1}-x_{t+1}^{*ij}\right)\,p\left(x_{t+1}|y_{1:t}\right)}{m\sum_{j=1}^{N}\sum_{i=1}^{K}p\left(\eta^{j}\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*i}(y_{t+1},\eta^{j})}^{-1}p\left(x_{t+1}^{*i}(y_{t+1},\eta^{j})|y_{1:t}\right)},\end{aligned}$$

for $t=1,2,...,T-1$, and where $x_{t+1}^{*ij}=x_{t+1}^{*i}(y_{t+1},\eta^{j})$ is determined by the i^{th} solution at the j^{th} grid point η^{j} and the observed y_{t+1}. Hence, the updated filtered distribution has representation (up to numerical approximation error) as a discrete distribution as in (6.11), with density

$$p\left(x_{t+1}|y_{1:t+1}\right)=\sum_{j=1}^{N}\sum_{i=1}^{K}W_{t+1}^{ij}\delta\left(x_{t+1}-x_{t+1}^{*ij}\right),$$

where, for $i=1,2,...,K$ and $j=1,2,...,N$,

$$W_{t+1}^{ij}=\frac{p\left(\eta^{j}\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*ij}}^{-1}p\left(x_{t+1}^{*ij}|y_{1:t}\right)}{\sum_{j=1}^{N}\sum_{i=1}^{K}p\left(\eta^{j}\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*ij}}^{-1}p\left(x_{t+1}^{*ij}|y_{1:t}\right)}$$

denotes the probability associated with location x_{t+1}^{*ij}.

6.3.6 Summary of the Algorithm for General t

The actual algorithm is easily implemented using the following summary. Denote by $x_t^{*ij}=x_t^{*i}\left(y_t,\eta^j\right)$ the K zeroes of $y_t-h\left(x_t,\eta^j\right)$, for $i=1,2,..,K$, for each $j=1,2,...,N$ and all $t=1,2,...,T$. Initialize the filter at period 1 with (6.8) and (6.9). For $t=1,2,...,T-1$

$$\begin{aligned}
p(x_{t+1}|y_{1:t}) &= \sum_{j=1}^{N}\sum_{i=1}^{K}W_t^{ij}\ p\left(x_{t+1}|x_t^{*ij}\right),\\
p\left(y_{t+1}|y_{1:t}\right) &= \sum_{j=1}^{N}\sum_{i=1}^{K}M_{t+1}^{ij}\left(y_{t+1}\right)\ p(x_{t+1}^{*i}(y_{t+1},\eta^j)|y_{1:t}),\\
p\left(x_{t+1}|y_{1:t+1}\right) &= \sum_{j=1}^{N}\sum_{i=1}^{K}W_{t+1}^{ij}\delta\left(x_{t+1}-x_{t+1}^{*ij}\right),
\end{aligned}$$

with

$$M_{t+1}^{ij}\left(y_{t+1}\right)=m\,p\left(\eta^j\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*i}\left(y_{t+1},\eta^j\right)}^{-1}$$

and

$$W_{t+1}^{ij}=\frac{M_{t+1}^{i\,j}\left(y_{t+1}\right)p\left(x_{t+1}^{*ij}|y_{1:t}\right)}{\sum_{j=1}^{N}\sum_{i=1}^{K}M_{t+1}^{i}\left(y_{t+1}\right)p\left(x_{t+1}^{*ij}|y_{1:t}\right)}.$$

The computational burden involved in the evaluation of the t^{th} component of the likelihood function $(p(y_{t+1}|y_{1:t}))$ is $(N^2 \times K^2)$ for all t, implying an overall computational burden that is linear in T. Nevertheless, this computational burden is heavier than the original case where the function $G(x_t)$ has a unique root. As with the unique root case, the state prediction density and the prediction density for the observed variable are continuous, despite the approximation rendering the state filtered distribution discrete. Conditional on known values of $p(\eta^j)$ (and all other parameters), for large enough N the filtering algorithm is exact, being able to produce the true filtered and predictive distributions for the state, and the true predictive distribution for the observed, at each time point.

6.4 The Case of Non-analytic Roots of $G_t(x_t)$

The linear, SCD and RV models examined in this thesis all have unique solutions, $x_t^*(y_t, \eta)$, that are analytically available. However, for some models an analytical solution for $x_t^*(y_t, \eta)$ may not be available. For example, consider the stochastic volatility model used in Stroud, Muller and Polson (2003). The equity price, P_t, is assumed to follow a geometric Brownian motion with volatility, V_t, modelled as a continuous time mean-reverting process,

$$d\ln P_t = (\mu - V_t/2)\,dt + \sqrt{V_t}dB_t^p, \qquad (6.14)$$

$$d\ln V_t = \kappa(\phi - \ln V_t)\,dt + \sigma_v dB_t^v, \qquad (6.15)$$

where B_t^p and dB_t^v are independent Brownian motions. The parameter κ governs the speed of mean reversion, ϕ is the long-run mean of log-volatility and σ_v is the volatility of volatility. Using an Euler discretization of (6.14) and (6.15) over the interval of time $\Delta t = 1$, and defining $y_t = \ln\left(\frac{P_{t+1}}{P_t}\right)$ and $x_t = \ln V_t$, the state space model becomes

$$\begin{aligned} y_t &= \left(\mu - \frac{1}{2}\exp(x_t)\right) + \exp\left(\frac{x_t}{2}\right)\eta_t & (6.16) \\ x_{t+1} &= x_t + \kappa(\phi - x_t) + \sigma_v v_t, & (6.17) \end{aligned}$$

where η_t and v_t are both $i.i.d.N(0,1)$. The measurement equation in (6.16) has a non-linear conditional mean, with the conditional variance of the additive error term being a non-linear function of x_t. A non-parametric treatment of the measurement error distribution for η_t via the method outlined in Chapter 3 is thus not immediately feasible in this case, due to the lack of an analytic solution of $y_t - \left(\mu - \frac{1}{2}\exp(x_t)\right) - \exp\left(\frac{x_t}{2}\right)\eta_t$ for x_t as a function of y_t and η_t.

In contrast, consider the discrete time stochastic volatility model in Harvey *et al.* (1994); see also Shephard and Pitt (1997) and Durbin and Koopman (2001). Denoting y_t once again as the logarithmic return and x_t as the logarithm of the variance of the return, where it is assumed that

$$\begin{aligned} y_t &= \exp\left(\frac{x_t}{2}\right)\eta_t & (6.18) \\ x_{t+1} &= \alpha + \rho x_t + v_t, & (6.19) \end{aligned}$$

with $\eta_t \sim i.i.d.\,(0,1)$ and $v_t \sim i.i.d.\,(0,\sigma_v^2)$. Solving the measurement equation in (6.18) for x_t returns the unique solution $x_t^*(y_t,\eta)$ as

$$x_t^* = 2\ln\left(\frac{y_t}{\eta_t}\right).$$

However, the constraint $\frac{y_t}{\eta_t} > 0$ has to be imposed in order for x_t^* to be defined. To overcome this, the measurement equation in (6.18) can be rewritten as

$$|y_t| = \exp\left(\frac{x_t}{2}\right)|\eta_t|,$$

allowing only the absolute values of y_t and η_t be used in the model. This rewriting of the model does not result in the loss of any information because, referring to (6.18), the sign of the observation y_t only reveals the sign of η_t, and contains no information about the latent state variable, x_t. Based on this representation of the measurement equation, the unique solution is given as $x_t^* = 2\ln\left(\left|\frac{y_t}{\eta_t}\right|\right)$, which by its contruction, will always be defined for $\eta_t > 0$.

In summary, the non-parametric filter is not directly applicable in models where an analytical solution for the state variable is unavailable. However, in certain cases, a re-parameterization of the model that admits a solution may be feasible, thereby extending the realm of models to which the filter can be applied.[2]

6.5 Multivariate State Space Models

The focus of the thesis up to this point has been on the non-parametric estimation of the forecast distribution of the scalar random variable y_t. However, the non-parametric approach can be extended to multivariate settings of the following form: (i) the dimension of the observed variable is larger than the dimension

[2] Of course numerical techniques could always, in principle, be used to solve $y_t - h(x_t, \eta_t) = 0$ for x_t, with this numerical solution producing an additional level of computational burden to the filter. We do not explore this possibility here.

of the state variable; (ii) the dimension of the state variable is larger than the dimension of the observed variable; (iii) the dimensions of the state and observed variables are the same. We illustrate here the non-parametric methodology for the simplest case only, namely case (iii), where the state space model is a square system. For illustration purposes, we consider the *bivariate* system governed by the measurement equation for each $y_t = (y_{1,t}, y_{2,t})'$,

$$\begin{aligned} y_t &= h(x_t, \eta_t) \\ &= \begin{bmatrix} h_1\left(x_{1,t}, x_{2,t}, \eta_{1,t}\right) \\ h_2\left(x_{1,t}, x_{2,t}, \eta_{2,t}\right) \end{bmatrix} \end{aligned}$$

where $\eta_t = \left(\eta_{1,t}, \eta_{2,t}\right)'$ is now a (2×1) vector for each $t = 1, 2, ..., T$, the transition probabilities for the state vector $x_{t+1} = (x_{1,t+1}, x_{2,t+1})'$ are given by

$$p(x_{t+1}|x_t) = p(x_{1,t+1}, x_{2,t+1}|x_{1,t}, x_{2,t}),$$

again for $t = 1, 2, ..., T$, and the initial state distribution is

$$p(x_1) = p(x_{1,1}, x_{2,1}).$$

As with the univariate case, the following assumptions are invoked for the multivariate setting:

1. η_t is assumed to be *i.i.d.*, with each having density given by $p(\eta_t) = p\left(\eta_{1,t}, \eta_{2,t}\right)$.

2. The functions given by $h(x_t, \eta_t)$ are assumed to be differentiable with respect to each argument.

3. For given values of y_t and η_t, the function $G\left(x_t\right) = y_t - h\left(x_t, \eta_t\right)$ is assumed to have a unique root at $x_t = x_t^*\left(y_t, \eta_t\right)$ which is analytically available, as well as having a non-zero derivative at that root.

Extending the δ-function representation of the distribution in (3.11) to multivariate systems, the measurement distribution is written as

$$p\left(y_t | x_t\right) = \int_{-\infty}^{\infty} p\left(\eta\right) \; \delta\left(y_t - h(x_t, \eta)\right) d\eta, \tag{6.20}$$

where η is a variable of integration that traverses the support of $p\left(\eta\right)$. Note that the form of the measurement density in (6.20) appears identical to the univariate case given in (3.11). However in this case, η, x_t and y_t are now all (2×1) variables, and the integral is over the two dimensional domain of η. Note that, as in the univariate case presented in Chapter 3, in the discussion that follows, the subscript indicating the time period t is not explicitly stated, so that, for example, $\left(\eta_1, \eta_2\right)'$ indicates the two-dimensional random variable, η. Further, the Dirac δ-function in (6.20) provides the same requisite properties as the univariate case, so that essentially the same steps may be used to derive the multivariate filter. For further information on the multivariate Dirac δ-function, see Khuri (2004).

6.5.1 The Initial Filtered Distribution: $p\left(x_1|y_1\right)$

Referring to the expression in (3.12) for the univariate state space model, the first filtered density for the bivariate state space model is expressed similarly as

$$p\left(x_1|y_1\right)=\frac{p\left(x_1\right)\int_{-\infty}^{\infty}p\left(\eta\right)\left|\frac{\partial h}{\partial x_1}\right|^{-1}\delta\left(x_1-x_1^*(y_1,\eta)\right)d\eta}{\int_{-\infty}^{\infty}p\left(x_1^*(y_1,\eta)\right)p\left(\eta\right)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^*(y_1,\eta)}^{-1}d\eta}, \tag{6.21}$$

where $x_1^*(y_1,\eta)$ is the (assumed unique) solution to $y_1-h(x_1,\eta)=0$ for any value η in the support of $p\left(\eta\right).$

Next, to numerically evaluate the filtered distribution in (6.21) via rectangular integration, evenly spaced grids $\left\{\eta_1^1,\eta_1^2,...,\eta_1^{N_1}\right\}$ and $\left\{\eta_2^1,\eta_2^2,...,\eta_2^{N_2}\right\}$ for the joint measurement error density are defined, with interval lengths m_1 and m_2, given by $m_1=\eta_1^j-\eta_1^{j-1}$, for $j=2,3,...,N_1$, and $m_2=\eta_2^k-\eta_2^{k-1}$, for $k=2,3,...,N_2$. The resulting approximation for $p\left(x_1|y_1\right)$ is then given by

$$p\left(x_1|y_1\right)\approx\frac{p\left(x_1\right)\sum_{j=1}^{N_1}\sum_{k=1}^{N_2}m_1m_2\,p\left(\eta_1^j,\eta_2^k\right)\left|\frac{\partial h}{\partial x_1}\right|^{-1}\delta\left(x_1-x_1^{*jk}\right)}{\sum_{j=1}^{N_1}\sum_{k=1}^{N_2}m_1m_2p\left(\eta_1^j,\eta_2^k\right)\ p\left(x_1^{*jk}\right)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^*\left(y_1,\eta_1^j,\eta_2^k\right)}^{-1}}, \tag{6.22}$$

with

$$\left|\frac{\partial h}{\partial x_1}\right|^{-1}=\left|\begin{array}{cc}\frac{\partial h_1}{\partial x_{1,1}} & \frac{\partial h_1}{\partial x_{2,1}}\\ \frac{\partial h_2}{\partial x_{1,1}} & \frac{\partial h_2}{\partial x_{2,1}}\end{array}\right|^{-1}, \tag{6.23}$$

where $p\left(\eta_1^j,\eta_2^k\right)$ is defined as the unknown density ordinate associated with the joint measurement error density at the bivariate grid-point indexed by j and k. The expression $x_{i,1}$ in (6.23) represents the i^{th} component of the x_1 variable, for $i=1,2$. Note that using the numerical integration approach in the numerator as well as in the denominator serves to produce an implied state, $x_1^{*jk}=x_1^*\left(y_1,\eta_1^j,\eta_2^k\right)$, associated with the bivariate grid-point $\left(\eta_1^j,\eta_2^k\right)'$, so that the first filtered distribution has representation (up to numerical approximation error) as a discrete distribution, with density

$$p\left(x_1|y_1\right)=\sum_{j=1}^{N_1}\sum_{k=1}^{N_2}W_1^{jk}\delta\left(x_1-x_1^{*jk}\right), \tag{6.24}$$

and where

$$W_1^{jk}=\frac{p\left(\eta_1^j,\eta_2^k\right)\,p\left(x_1^{*jk}\right)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^{*jk}}^{-1}}{\sum_{j=1}^{N_1}\sum_{k=1}^{N_2}p\left(\eta_1^j,\eta_2^k\right)\,p\left(x_1^{*jk}\right)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^{*jk}}^{-1}}, \tag{6.25}$$

for $j=1,2,...,N_1$ and $k=1,2,...,N_2$. Implicit in this approximation to the first filtered state density is the first likelihood contribution,

$$p\left(y_1\right)=m_1m_2\sum_{j=1}^{N_1}\sum_{k=1}^{N_2}p\left(\eta_1^j,\eta_2^k\right)\,p\left(x_1^{*jk}\right)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^{*jk}}^{-1}, \tag{6.26}$$

obtained from approximating the denominator in (6.21).

Having obtained the representation in (6.24) for time $t=1$, it will be shown that for any time $t=2,3,...T$, an appropriate discrete distribution can be found to approximate the filtered distribution as

$$p\left(x_t|y_t\right)=\sum_{j=1}^{N_1}\sum_{k=1}^{N_2}W_t^{jk}\delta\left(x_t-x_t^{*jk}\right), \tag{6.27}$$

where the iteratively determined weights satisfy

$$\sum_{j=1}^{N_1}\sum_{k=1}^{N_2}W_t^{jk}=1,$$

and each implied state

$$x_t^{*jk}=x_t^*\left(y_t,\eta_1^j,\eta_2^k\right), \tag{6.28}$$

arising from state grid location $\left(\eta_1^j,\eta_2^k\right)$ and observation y_t, is determined by the unique zero of $y_t-h(x_t,\eta)$.

6.5.2 The Predictive Distribution for the State: $p\left(x_{t+1}|y_{1:t}\right)$

Assuming (6.27) holds in period t, it follows that the one-step-ahead state prediction density is a mixture of transition densities, since

$$\begin{aligned}
p\left(x_{t+1}|y_{1:t}\right) &= \int p\left(x_{t+1}|x_t\right)p\left(x_t|y_{1:t}\right)dx_t \\
&= \int p\left(x_{t+1}|x_t\right)\sum_{j=1}^{N_1}\sum_{k=1}^{N_2}W_t^{jk}\delta\left(x_t - x_t^{*jk}\right)dx_t \\
&= \sum_{j=1}^{N_1}\sum_{k=1}^{N_2}W_t^{jk}\int p\left(x_{t+1}|x_t\right)\delta\left(x_t - x_t^{*jk}\right)dx_t \\
&= \sum_{j=1}^{N_1}\sum_{k=1}^{N_2}W_t^{jk}p\left(x_{t+1}|x_t^{*jk}\right), \qquad (6.29)
\end{aligned}$$

for $t = 1,2,...,T$. Note that the W_t^{jk} values, for $j = 1,2,...,N_1$ and $k = 1,2,...,N_2$, are available from the previous iteration of the filter.

6.5.3 The One-step-ahead Predictive Distribution for the Observed: $p\left(y_{t+1}|y_{1:t}\right)$

Referring to the expression in (3.19) for the univariate state space model, the one-step-ahead predictive distribution for the bivariate state space model is expressed similarly, using vector notation, as

$$p\left(y_{t+1}|y_{1:t}\right) = \int_{-\infty}^{\infty} p\left(\eta\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|^{-1}_{x_{t+1}=x^*_{t+1}(y_{t+1},\eta)} p(x^*_{t+1}\left(y_{t+1},\eta\right))|y_{1:t})d\eta.$$

Invoking again the pre-specified bivariate grid of values for $(\eta_1,\eta_2)'$, the one-step-ahead prediction density (up to numerical approximation error) is,

$$p\left(y_{t+1}|y_{1:t}\right) = m_1 m_2\sum_{j=1}^{N_1}\sum_{k=1}^{N_2} p\left(\eta_1^j,\eta_2^k\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|^{-1}_{x_{t+1}=x^*_{t+1}\left(y_{t+1},\eta_1^j,\eta_2^k\right)} p(x^*_{t+1}\left(y_{t+1},\eta_1^j,\eta_2^k\right))|y_{1:t}). \qquad (6.30)$$

Given that $p(x_{t+1}^{*}\left(y_{t+1},\eta_1^j,\eta_2^k\right))|y_{1:t})$ in (6.30) denotes the one-step-ahead predictive density from (6.29) evaluated at $x_{t+1}=x_{t+1}^{*}\left(y_{t+1},\eta_1^j,\eta_2^k\right)$, it can be seen that $p\left(y_{t+1}|y_{1:t}\right)$ is computed as an $\left(N_1\times N_2\right)^2$ mixture of (specified) transition density functions as a consequence.

6.5.4 The Updated Filtered Distribution: $p\left(x_{t+1}|y_{1:t+1}\right)$

Finally, the predictive distribution for the state at time $t+1$ is updated given the realization y_{t+1} as

$$\begin{aligned}
p\left(x_{t+1}|y_{1:t+1}\right) &= \frac{p\left(y_{t+1}|x_{t+1}\right)p\left(x_{t+1}|y_{1:t}\right)}{p\left(y_{t+1}|y_{1:t}\right)}\\
&\approx \frac{m_1m_2\sum_{j=1}^{N_1}\sum_{k=1}^{N_2} p\left(\eta_1^j,\eta_2^k\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|^{-1}\delta\left(x_{t+1}-x_{t+1}^{*jk}\right)\,p\left(x_{t+1}|y_{1:t}\right)}{m_1m_2\sum_{j=1}^{N_1}\sum_{k=1}^{N_2} p\left(\eta_1^j,\eta_2^k\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*jk}}^{-1} p(x_{t+1}^{*jk}|y_{1:t})},
\end{aligned}$$

for $t=1,2,...,T-1$, and where $x_{t+1}^{*jk}=x_{t+1}^{*}\left(y_{t+1},\eta_1^j,\eta_2^k\right)$ is determined by the $(j,k)^{th}$ grid point $\left(\eta^j,\eta^k\right)'$ and the observed y_{t+1}. Hence, the updated filtered distribution has representation (up to numerical approximation error) as a discrete distribution as in (3.16), with density

$$p\left(x_{t+1}|y_{1:t+1}\right)=\sum_{j=1}^{N_1}\sum_{k=1}^{N_2}W_{t+1}^{jk}\delta\left(x_{t+1}-x_{t+1}^{*jk}\right),$$

where, for $j=1,2,...,N_1$ and $k=1,2,...,N_2$,

$$W_{t+1}^{jk}=\frac{p\left(\eta_1^j,\eta_2^k\right)\,p\left(x_{t+1}^{*jk}|y_{1:t}\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*jk}}^{-1}}{\sum_{j=1}^{N_1}\sum_{k=1}^{N_2} p\left(\eta_1^j,\eta_2^k\right)\,p\left(x_{t+1}^{*jk}|y_{1:t}\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*jk}}^{-1}}$$

denotes the probability associated with location x_{t+1}^{*jk} given by the unique zero of $y_{t+1}-h(x_{t+1},\eta^j,\eta^k)$, for $j=1,2,...N_1$ and $k=1,2,...,N_2$.

6.5.5 Summary of the Algorithm for General t

The actual algorithm is implemented using the following summary. Denote by $x_t^{*jk} = x_t^*\left(y_t, \eta_1^j, \eta_2^k\right)$ the unique zero of $y_t - h(x_t, \eta)$, for each $j = 1, 2, ..., N_1$, $k = 1, 2, ..., N_2$ and all $t = 1, 2, ..., T$. Initialize the filter at period 1 with (6.24) and (6.25). For $t = 1, 2, ..., T-1$

$$\begin{aligned} p(x_{t+1}|y_{1:t}) &= \sum_{j=1}^{N_1}\sum_{k=1}^{N_2} W_t^{jk} p\left(x_{t+1}|x_t^{*jk}\right), \\ p\left(y_{t+1}|y_{1:t}\right) &= \sum_{j=1}^{N_1}\sum_{k=1}^{N_2} M_{t+1}^{jk}\left(y_{t+1}\right) p(x_{t+1}^{*jk}|y_{1:t}), \\ p\left(x_{t+1}|y_{1:t+1}\right) &= \sum_{j=1}^{N_1}\sum_{k=1}^{N_2} W_{t+1}^{jk}\delta\left(x_{t+1} - x_{t+1}^{*jk}\right), \end{aligned}$$

with

$$M_{t+1}^{jk}\left(y_{t+1}\right) = m_1 m_2\, p\left(\eta_1^j, \eta_2^k\right) \left|\frac{\partial h}{\partial x_{t+1}}\right|^{-1}_{x_{t+1} = x_{t+1}^*\left(y_{t+1}, \eta_1^j, \eta_2^k\right)}$$

and

$$W_{t+1}^{jk} = \frac{M_{t+1}^{jk}\left(y_{t+1}\right) p\left(x_{t+1}^{*jk}|y_{1:t}\right)}{\sum_{j=1}^{N_1}\sum_{k=1}^{N_2} M_{t+1}^{jk}\left(y_{t+1}\right) p\left(x_{t+1}^{*jk}|y_{1:t}\right)}.$$

The computational burden involved in the evaluation of the t^{th} component of the likelihood function $\left(p\left(y_{t+1}|y_{1:t}\right)\right)$ is of order $\left(N_1 \times N_2\right)^2$ for all t, an increase from the order of N^2 in the univariate state space model. That is, despite remaining linear in T the computational burden of the algorithm is exponential in the dimension of y_t (and x_t), highlighting the fact that the grid-based algorithm is most suitable for reasonably low-dimensional problems. For high-dimensional systems, evaluation of the relevant integrals via simulation techniques would be required. We outline a contribution towards that solution in the following section, but in a context in which a parametric model for the measurement error, η_t is specified. That is, we shift focus, for the sake of this final illustration, from the non-parametric aspect of the filter and explore the other main characteristic of the proposed algorithm - namely its dependence only on integration with respect to the invariant η.

6.6 Probabilistic Filtering using Monte Carlo Methods

6.6.1 Introduction

Chapter 3 introduced the grid-based filter (in the scalar case), whereby all the requisite integrals are evaluated with respect to the invariant distribution of η, and approximated via rectangular integration. If $p(\eta)$ is unknown, the resulting non-parametric filter allows for the measurement error density to be estimated via (penalized) maximum likelihood. On the other hand, if $p(\eta)$ were specified parametrically, and for given values of all unknown parameters, then in the limit, as the number of grid points on the support of the invariant distribution of η is increased, the grid-based filter produces the exact filtering and predictive distributions.

This section explores an alternative to the the grid-based method in which all relevant integrals are evaluated by Monte Carlo simulation, based on draws from the invariant distribution of η. This requires $p(\eta)$ to be specified parametrically and to be amenable to simulation. The idea is to replace the evenly-spaced grid on the support of η by a set of simulated draws $\eta^1, \eta^2, ..., \eta^N \sim p(\eta)$ wherever integration is required to produce the filtering and predictive distributions. We show that the resulting Monte Carlo filter avoids the degeneracy problems that are a feature of existing particle filtering algorithms and, hence, could be a powerful tool in high-dimensional state space models.

6.6.2 Derivation of the Monte Carlo Filter

In illustrating the probabilistic filtering algorithm, the same system in (3.1) and (3.2) is considered, along with the same assumptions regarding the function $G(x_t) = y_t - h(x_t, \eta_t)$, namely that it is differentiable, has a unique root at $x_t = x_t^*(y_t, \eta)$, and has a non-zero derivative at that root. It is further assumed that $p(\eta)$ is specified up to a small number of fixed parameters, and that draws of η can be obtained by simulation. Of course, despite the demonstration of the method in the scalar case, as highlighted above the primary motivation for the method occurs in the high-dimensional case, where the grid-based filter would be computationally infeasible.[3]

The Initial Filtered Distribution: $p(x_1|y_1)$

From (3.12), the δ-function representation of the initial filtered distribution expression is

$$p(x_1|y_1) = \frac{p(x_1)\int_{-\infty}^{\infty} p(\eta)\left|\frac{\partial h}{\partial x_1}\right|^{-1} \delta(x_1 - x_1^*(y_1,\eta))\, d\eta}{\int_{-\infty}^{\infty} p(x_1^*(y_1,\eta))\, p(\eta)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^*(y_1,\eta)}^{-1} d\eta}, \tag{6.31}$$

[3] Further to this point, simulation experiments have shown that in the one-dimensional case that has been the primary focus of this thesis, the simulation-based method produces an almost 10-fold increase in computational time relative to the grid-based algorithm, with no commensurate gain in accuracy.

where $x_1^*(y_1,\eta)$ is the assumed unique solution to $y_1 - h(x_1,\eta_1) = 0$ for any value of η in the support of $p(\eta)$. To evaluate the filtered distribution in (6.31) via Monte Carlo integration, a set of *i.i.d.* base particles $\left\{\eta^1,\eta^2,...,\eta^N\right\}$ is simulated from $p(\eta)$, resulting in the approximation for $p(x_1|y_1)$ given by

$$p(x_1|y_1) \approx \frac{p(x_1)\,N^{-1}\sum_{j=1}^N \left|\frac{\partial h}{\partial x_1}\right|^{-1}\delta\left(x_1 - x_1^{*j}\right)}{N^{-1}\sum_{i=1}^N p\left(x_1^{*i}\right)\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^{*i}}^{-1}}.$$

The expression $x_1^{*j} = x_1^*(y_1,\eta^j)$ is the implied state associated with the observed y_1 and each η^j, such that the first filtered distribution has representation (up to Monte Carlo approximation error) as a discrete distribution, with density

$$p(x_1|y_1) = \sum_{j=1}^N B_1^j \delta\left(x_1 - x_1^{*j}\right), \tag{6.32}$$

and where

$$B_1^j = \frac{\left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^{*j}}^{-1} p\left(x_1^{*j}\right)}{\sum_{i=1}^N \left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^{*i}}^{-1} p\left(x_1^{*i}\right)}, \tag{6.33}$$

for $j = 1,2,...,N$. Whilst the form of the filter in (6.32) looks similar to that of (3.13) in the grid-based filter, the locations x_1^{*j} in (6.32) are random due to being derived from the random set of particles $\{\eta^j,\ j = 1,2,...,N\}$. Therefore, the functional value of B_1^j is not the same as that of W_1^j in (3.14) under the grid-based filter, due to the absence of the $p(\eta^j)$ element. Implicit in this approximation to the first filtered state density is the first likelihood contribution,

$$p(y_1) = N^{-1}\sum_{i=1}^N \left|\frac{\partial h}{\partial x_1}\right|_{x_1=x_1^{*i}}^{-1} p\left(x_1^{*i}\right), \tag{6.34}$$

obtained from approximating the denominator in (6.31).

Having obtained the representation in (6.32) for time $t = 1$, it will be shown that for any time $t = 2, 3, ..., T$, an appropriate discrete distribution can be found to approximate the filtered distribution,

$$p\left(x_t | y_{1:t}\right) = \sum_{j=1}^{N} B_t^j \delta\left(x_t - x_t^{*j}\right), \tag{6.35}$$

where the iteratively determined weights satisfy $\sum_{j=1}^{N} B_t^j = 1$, and each state grid location

$$x_t^{*j} = x_t^*(y_t, \eta^j)$$

is determined by the unique zero of $y_t - h(x_t, \eta^j)$, for $j = 1, 2, ..., N$.

The Predictive Distribution for the State: $p\left(x_{t+1} | y_{1:t}\right)$

Assuming (6.35) holds in period t, it follows that the one-step-ahead state prediction density is a mixture of transition densities, since

$$\begin{aligned} p\left(x_{t+1} | y_{1:t}\right) &= \int p\left(x_{t+1} | x_t\right) p\left(x_t | y_{1:t}\right) dx_t \\ &= \int p\left(x_{t+1} | x_t\right) \sum_{j=1}^{N} B_t^j \delta\left(x_t - x_t^{*j}\right) dx_t \\ &= \sum_{j=1}^{N} B_t^j \int p\left(x_{t+1} | x_t\right) \delta\left(x_t - x_t^{*j}\right) dx_t \\ &= \sum_{j=1}^{N} B_t^j \; p\left(x_{t+1} | x_t^{*j}\right), \end{aligned} \tag{6.36}$$

for $t = 1, 2, ..., T$. The notation $p\left(x_{t+1} | x_t^{*j}\right)$ denotes the transition density of $p\left(x_{t+1} | x_t\right)$, viewed as a function of x_{t+1} and given the fixed value of $x_t = x_t^{*j}$. As it is assumed that the transition densities $p\left(x_{t+1} | x_t\right)$ are available, no additional approximation is needed in moving from $p\left(x_t | y_{1:t}\right)$ to $p\left(x_{t+1} | y_{1:t}\right)$.

The One-step-ahead Predictive Distribution for the Observed: $p\left(y_{t+1}|y_{1:t}\right)$

The one-step-ahead predictive distribution with its integral with respect to η is derived in (3.20) and represented as

$$p\left(y_{t+1}|y_{1:t}\right)=\int_{-\infty}^{\infty}p\left(\eta\right)\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*}(y_{t+1},\eta)}^{-1}p(x_{t+1}^{*}(y_{t+1},\eta)|y_{1:t})d\eta.$$

By invoking again the Monte Carlo sample of η values, the one-step-ahead prediction density (up to numerical approximation error) becomes,

$$p\left(y_{t+1}|y_{1:t}\right)=N^{-1}\sum_{i=1}^{N}\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*}(y_{t+1},\eta^{i})}^{-1}p\left(x_{t+1}^{*}(y_{t+1},\eta^{i})|y_{1:t}\right). \qquad (6.37)$$

Noting that $p\left(x_{t+1}^{*}(y_{t+1},\eta^{i})|y_{1:t}\right)$ in (6.37) denotes the one-step-ahead predictive density from (6.36) evaluated at $x_{t+1}=x_{t+1}^{*}(y_{t+1},\eta^{i})$, it can be seen that $p\left(y_{t+1}|y_{1:t}\right)$ is computed as an N^{2} mixture of (specified) transition density functions as a consequence.

The Updated Filtered Distribution: $p\left(x_{t+1}|y_{1:t+1}\right)$

Finally, the predictive distribution for the state at time $t+1$ is updated given the realization y_{t+1} as

$$\begin{aligned}p\left(x_{t+1}|y_{1:t+1}\right) &= \frac{p\left(y_{t+1}|x_{t+1}\right)p\left(x_{t+1}|y_{1:t}\right)}{p\left(y_{t+1}|y_{1:t}\right)}\\ &\approx \frac{\sum_{j=1}^{N}\left|\frac{\partial h}{\partial x_{t+1}}\right|^{-1}\delta\left(x_{t+1}-x_{t+1}^{*j}\right)\,p\left(x_{t+1}|y_{1:t}\right)}{\sum_{i=1}^{N}\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*i}}^{-1}p\left(x_{t+1}^{*i}|y_{1:t}\right)},\end{aligned}$$

for $t=1,2,...,T-1$, and where $x_{t+1}^{*j}=x_{t+1}^{*}(y_{t+1},\eta^{j})$ is determined by the simulated value η^{j} and the observed y_{t+1}. Hence, the updated filtered distribution has representation (up to numerical approximation error) as a discrete distribution as in (6.35), with density

$$p\left(x_{t+1}|y_{1:t+1}\right)=\sum_{j=1}^{N}B_{t+1}^{j}\delta\left(x_{t+1}-x_{t+1}^{*j}\right),$$

where, for $j=1,2,...,N$,

$$B_{t+1}^{j}=\frac{\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*j}}^{-1}p\left(x_{t+1}^{*j}|y_{1:t}\right)}{\sum_{i=1}^{N}\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*i}}^{-1}p\left(x_{t+1}^{*i}|y_{1:t}\right)}$$

denotes the probability associated with location x_{t+1}^{*j} given by the unique zero of $y_{t+1}-h(x_{t+1},\eta^{j})$, for $j=1,2,...,N$.

Summary of the Algorithm for General t

The actual algorithm is easily implemented using the following summary. Denote by $x_t^{*j}=x_t^{*}\left(y_t,\eta^j\right)$ the unique zero of $y_t-h\left(x_t,\eta^j\right)$, for each $j=1,2,...,N$ and all $t=1,2,...,T$. Initialize the filter at period 1 with (6.32) and (6.33), and with $\eta^1,\eta^2,\ldots,\eta^N\overset{iid}{\sim}p\left(\eta\right)$, with N sufficiently large. For $t=1,2,...,T-1$,

$$\begin{aligned}
p(x_{t+1}|y_{1:t}) &= \sum_{j=1}^{N}B_t^j\ p\left(x_{t+1}|x_t^{*j}\right),\\
p\left(y_{t+1}|y_{1:t}\right) &= N^{-1}\sum_{i=1}^{N}\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*}(y_{t+1},\eta^{i})}^{-1}p\left(x_{t+1}^{*}(y_{t+1},\eta^{i})|y_{1:t}\right),\\
p\left(x_{t+1}|y_{1:t+1}\right) &= \sum_{j=1}^{N}B_{t+1}^{j}\delta\left(x_{t+1}-x_{t+1}^{*j}\right),
\end{aligned}$$

with

$$B_{t+1}^{j}=\frac{\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*j}}^{-1}p\left(x_{t+1}^{*j}|y_{1:t}\right)}{\sum_{i=1}^{N}\left|\frac{\partial h}{\partial x_{t+1}}\right|_{x_{t+1}=x_{t+1}^{*i}}^{-1}p\left(x_{t+1}^{*i}|y_{1:t}\right)}.$$

A few points regarding the above algorithm are as follows. First, it is noted that the time subscript t has been again omitted from η as the distribution is assumed to be constant for all t. Further, although it is possible to use different simulated draws of η^j values, for $j = 1, 2, ..., N$, for each recursive iteration of the filter, it is also suitable to re-use the same set of simulated draws each time an integral is required to be evaluated. Additionally, the computational burden involved in the evaluation of the t^{th} component of the likelihood function is of order N^2 for all t, implying an overall computational burden that is linear in T.

Second, like the particle filtering methods outlined in Chapter 2, conditional on a parametric specification of $p(\eta)$ (and all other parameters), for large enough N, the Monte Carlo filtering algorithm is exact, in the sense of recovering the true filtered and predictive distributions for the state, plus the true predictive distribution for the observed, at each time point. However, in contrast to existing particle filters, which propagate draws of x_{t+1} values from the previous time period and subsequently resample those draws, the Monte Carlo filter initially derives the particles via simulated draws from the invariant distribution of the measurement error, with the resultant state particles produced as the appropriate zeros of the measurement equation. These particles are subsequently reweighted so that the distribution of the weighted particles defines an estimate of the desired filtered distribution. Crucially, owing to the fact that the implied particles at any given time period $t + 1$, x_{t+1}^{*i}, $i = 1, 2, ..., N$, are generated via the root $x_{t+1}^{*}(y_{t+1}, \eta^i)$ of the measurement equation $y_{t+1} = h\left(x_{t+1}, \eta_{t+1}\right)$, and hence are not propagated from the previous time period, the filter does not suffer from particle degeneracy. Further, at the subsequent reweighting stage, even if a particle, x_t^{*j}, from the previous period carries negligible weight, i.e. $B_t^j > 0$ but small, as no resampling is required such particles will impact upon the filtered density, $p(x_{t+1}|y_{1:t+1})$ via the predictive state density, $p(x_{t+1}|y_{1:t})$. Hence, not only is direct degeneracy from non-propagated particles avoided, but the added Monte Carlo error produced via the resampling weights that occurs in existing particle filters is not a feature of the new approach.

Finally, it is noted that the multivariate Monte Carlo filter could be implemented in higher dimensions, particularly in the square case discussed in Section 6.5. In addition, the Monte Carlo approach could be used in conjunction with alternative and flexible representations of $p\left(\eta\right)$, such as the mixture representation in (2.60), in order to produce new non-parametric (or semi-parametric) filters, which may provide a useful way to construct non-parametric filters in the multivariate setting.

6.7 Summary

In this chapter we have expanded the proposed non-parametric methodology in a number of different ways. First, its use in producing an estimate of a multi-step-ahead forecast distribution is demonstrated. Next, the issues related to the relaxation of two of the crucial assumptions that underlie the non-parametric filter in Chapter 3 are studied. In particular, the non-parametric approach is adapted to state space models in which the function $G(x_t)$ has multiple roots. Whilst the approach cannot be readily applied to models in which the roots of $G(x_t)$ are not analytically available, we give an example of the type of re-parameterization of a model that can be used in this case. Third, the non-parametric filter is shown to be applicable in a multivariate state space setting, with a square bivariate state space model used as an illustration. Given the usual curse of dimensionality that affects a grid-based method, the non-parametric filter is clearly well-suited for low-dimensional problems only. Finally, we demonstrate how the grid-based filter can be replaced by a Monte Carlo filter when the measurement error density is parametrically specified, potentially using a flexible parametric structure, with this alternative approach being particularly beneficial in the high-dimensional case. Most notably, this simulation-based filter is shown to avoid the degeneracy problems that adversely affect existing particle filters.

Chapter 7

Conclusions

The focus of this thesis is four-fold. First, the primary interest is in forecasting non-Gaussian time series data, with this interest motivated by the large number of important economic and financial time series variables that exhibit non-Gaussianity. Second, consistent with general developments in the recent forecasting literature, in which distributional (as opposed to point) forecasts are increasingly viewed as the principle object of interest, probabilistic forecasting is a focus. As such, the forecasting approach automatically adheres to the notion of coherence, as well as enabling uncertainty about the future value of the variable in question to be fully quantified. Third, estimated forecast distributions are produced without reliance upon the correct specification of all aspects of the true DGP. Finally, a very general approach is adopted, by producing a forecast methodology for the non-linear, non-Gaussian state space framework, a framework that has broad empirical applicability.

Following the introductory chapter where the main aims and motivations are exposited, Chapter 2 presents a general parametric state space model and briefly discusses the associated inferential objects. To perform inference and forecasting in state space models, the presence of the latent random states has to be managed via filtering techniques. Hence, the general filtering and updating steps needed to specify the one-step-ahead predictive distributions of which the likelihood function is comprised, are presented. The specific filters outlined are the Kalman filter, extended Kalman filter, unscented Kalman filter, grid-based non-Gaussian filter, Gaussian-sum filter, and representative particle filters. Apart from the Kalman filter, each of these methods attempt to deal, in one way or another, with the non-linear or non-Gaussian aspect of the general model that prohibits an exact filtering solution. When the static parameters in the model are unknown, they may be estimated using ML estimation, made possible via the various approximations from the filtering algorithms. As a result, an estimate of the out-of-sample one-step-ahead predictive distribution may be obtained.

Having highlighted some of the limitations of the parametric forecasting methods, in Chapter 3 a new non-parametric method is developed. A non-parametric filter that exploits the known functional relationship between the observed variable and the state and measurement error variables, but avoids a parametric specification for the distribution of the measurement error, is derived. The filtering computations are manipulated using properties of the Dirac δ-function in such a way that all requisite integrals are undertaken with respect to the invariant distribution of the measurement error. Rectangular integration is then used to numerically evaluate the relevant integrals, over a fixed grid of points.

This approach has two key advantages over the existing filters presented in Chapter 2. First, the relative computational simplicity of the proposed method - for reasonably low-dimensional systems - is in marked contrast with the high computational burden of the Gaussian-sum filter and the particle filters discussed in Chapter 2. Second, when the measurement error density is unknown, the ordinates of the density can be readily estimated using a penalized ML procedure. This is in contrast with the filters discussed in Chapter 2, many of which require parametric assumptions for the model, assumptions that may potentially be incorrect.

To assess the predictive accuracy of the non-parametric approach relative to parametric alternatives, Chapter 4 presents the tools used to compare and evaluate the forecast distributions produced by these competing approaches. The chapter distinguishes between the comparison and evaluation of probabilistic forecasts, detailing the various tools used for each category. Simulation exercises are then undertaken for the linear and (non-linear) SCD models, with the parametric competitors all based on the Kalman (or extended) filter. In the linear model, three different distributions for the true measurment error are considered: Gaussian, Student-t and skewed Student-t. The results show that the non-parametric filter is competitive with the correctly specified parametric estimates in the Gaussian case, and is significantly better in the skewed Student-t

case. In the non-linear SCD model, three different distributions for the true measurement error are also considered, with the non-parametric method performing significantly better than the (misspecified) parametric approach in all cases. Over and above the production of results pertaining to the forecast distributions in question, a contribution of Chapter 4 is the summary of the relevant forecast evaluation literature and the methods used to assess competing distributional forecasts.

Chapter 5 applies the non-parametric methodology to the RV model, with a view to producing distributional forecasts for realized volatility. A simulation exercise similar to those in Chapter 4 is undertaken, with Gaussian, Student-t and skewed Student-t measurement errors considered for the true DGPs. Simulation results are consistent with the findings documented in Chapter 4, with the results favoring the non-parametric approach overall. When applied to the S&P500 market index data for January 1998 to August 2008, the empirical results also provide strong support for the overall accuracy of the non-parametric approach. The chapter details a subsampling approach for quantifying the sampling variation in an estimated one-step-ahead forecast distribution. Results show that substantial sampling variability can arise in the case of forecasts produced using a small sample, leading to qualitatively different conclusions about future volatility. On the other hand, for sample sizes that are typical of financial applications, sampling variability is seen to be much less of a concern.

The key contribution of Chapter 5 to the current literature on realized volatility forecasting is the production of non-parametric distributional forecasts of realized volatility, in a state space setting. This is in direct contrast to the bulk of the existing literature, which focuses on point forecasts, constructed in turn from parametric observation-driven models for the observed quantity. The non-parametric approach aims to capture the distributional properties of the measurement error which, in turn, contains the effects of all factors that influence realized volatility but are not explicitly modelled. In principle, all unmodelled effects will be reflected in the estimated forecast distributions.

Lastly, Chapter 6 expands the proposed non-parametric methodology in four ways. First, extension of the methodology to the estimation of multi-step-ahead forecast distributions for the observation is demonstrated. Second, issues related to the relaxation of two assumptions underlying the non-parametric filter are explored. Third, the non-parametric methodology is shown to be applicable to a multivariate state space setting, with a square bivariate state space model used as an illustration. Finally, it is demonstrated how the grid-based filter can be replaced by a Monte Carlo filter when the measurement error distribution is parametrically specified, with this alternative approach being particularly useful in high-dimensional state space systems. These extensions of the proposed filtering algorithm serve to demonstrate further the versatility of the approach and its applicability to a wide range of empirically relevant problems.

Bibliography

[1] Aït-Sahalia, Y. and Lo, A.W. 1998. Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices. *Journal of Finance* 53, 499-547.

[2] Aït-Sahalia, Y. and Mancini, L. 2008. Out of Sample Forecasts of Quadratic Variation, *Journal of Econometrics* 147, 17-33.

[3] Alspach, D.L. and Sorenson, H.W. 1972. Nonlinear Bayesian Estimation Using Gaussian-sum Approximations. *IEEE Transactions on Automatic Control* AC-17, 439-448.

[4] Amisano, G. and Giacomini, R. 2007. Comparing Density Forecasts via Weighted Likelihood Ratio Tests. *Journal of Business and Economic Statistics* 25, 177-190.

[5] Andersen, T.G. and Bollerslev, T. and 1998a. ARCH and GARCH Models. *Encyclopedia of Statistical Sciences, Volume II.* (Eds. S. Kotz, C.B. Read and D.L. Banks), New York.

[6] Andersen, T.G. and Bollerslev, T. 1998b. Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts, *International Economic Review* 39, 885-905.

[7] Andersen, T.G., Bollerslev, T. and Diebold, F.X. 2007. Roughing It Up: Including Jump Components in the Measurement, Modeling and Forecasting of Return Volatility. *The Review of Economics and Statistics* 89, 701-720.

[8] Andersen, T.G., Bollerslev, T. and Diebold, F.X. 2010. Parametric and Nonparametric Volatility Measurement. *Handbook of Financial Econometrics* (Eds. Y. Aït-Sahalia and L.P. Hansen), North-Holland, Amsterdam.

[9] Andersen, T.G., Bollerslev, T., Diebold, F.X. and Ebens, H. 2001a. The Distribution of Realized Stock Return Volatility. *Journal of Financial Econometrics* 61, 43-76.

[10] Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. 2001b. The Distribution of Exchange Rate Volatility. *Journal of the American Statistical Association* 96, 42-55.

[11] Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. 2003. Modeling and Forecasting Realized Volatility. *Econometrica* 71, 579-625.

[12] Anderson, B.D.O. and Moore, J.B. 1979. *Optimal Filtering.* Prentice.

[13] Anderson, J. 2001. On the Normal Inverse Gaussian Stochastic Volatility Model. *Journal of Business and Economic Statistics* 19, 44-54.

[14] Andreou, E. and Ghysels, E. 2002. Rolling-Sampling Volatility Estimators: Some New Theoretical, Simulation and Empirical Results. *Journal of Business and Economic Statistics* 20, 363-376.

[15] Ansley, C.F. and Kohn, R. 1985. Estimation, Filtering and Smoothing in State Space Models with Incompletely Specified Initial Conditions. *Annals of Statistics* 13, 1286-1316.

[16] Arulampalam, M.S., Maskell, S., Gordon, N. and Clapp, T. 2002. A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking. *IEEE Transactions On Signal Processing* 50, 174-188.

[17] Au, C. and Tam, J. 1999. Transforming Variables Using the Dirac Generalized Function. *The American Statistician* 53, 270-272.

[18] Bai, X., Russell, J. and Tiao, G.C. 2001. Beyond Merton's Utopia: Effects of Non-Normality and Dependence on the Precision of Variance Estimates Using High Frequency Financial Data. Unpublished manuscript, University of Chicago.

[19] Bakshi, G., Cao, C. and Chen, Z. 1997. Empirical Performance of Alternative Option Pricing Models. *Journal of Finance* 52, 2003-2049.

[20] Barndorff-Nielsen, O.E. and Shephard, N. 2002a. Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 64, 253-280.

[21] Barndorff-Nielsen, O.E. and Shephard, N. 2002b. Estimating Quadratic Variation Using Realized Variance. *Journal of Applied Econometrics* 17, 457-477.

[22] Barndorff-Nielsen, O.E., Nicolato, E. and Shephard, N. 2002. Some Recent Developments in Stochastic Volatility Modelling. *Quantitative Finance* 2, 11-23.

[23] Bates, D. 2000. Post-87 Crash Fears in the S&P 500 Futures Option Market. *Journal of Econometrics* 94, 181-238.

[24] Bauwens, L. and Veredas, D. 2004. The Stochastic Conditional Duration Model: A Latent Variable Model for the Analysis of Financial Durations. *Journal of Econometrics* 199, 381-412.

[25] Bauwens, L., Giot, P., Grammig, J. and Veredas, D. 2004. A Comparison of Financial Duration Models via Density Forecasts. *International Journal of Forecasting* 20, 589-609.

[26] Bauwens, L., Laurent, S. and Rombouts, J.V.K. 2006. Multivariate GARCH Models: A Survey. *Journal of Applied Econometrics* 21, 71-109.

[27] Berg, J.E., Geweke, J. and Rietz, T.A. 2010. Memoirs of an Indifferent Trader: Estimating Forecast Distributions from Prediction Markets, *Quantitative Economics* 1, 163-186.

[28] Berkowitz, J. 2001. Testing Density Forecasts with Applications to Risk Management, *Journal of Business and Economic Statistics* 19, 465–474.

[29] Black, F. and Scholes, M. 1973. The Pricing of Options and Liabilities. *Journal of Political Economy* 81, 637-654.

[30] Blair, B.J., Poon, S.H. and Taylor, S.J. 2001. Forecasting S&P100 Volatility: the Incremental Information Content of Implied Volatilities and High Frequency Index Returns, *Journal of Econometrics* 105, 5-26.

[31] Boero, G., Smith, J. and Wallis, K.F. 2011. Scoring Rules and Survey Density Forecasts. *International Journal of Forecasting* 27, 379-393.

[32] Bollerslev, T. 1986. Generalized Autogressive Conditional Heteroskedasticity. *Journal of Econometrics* 31, 307-327.

[33] Bollerslev, T. 2010. Glossary to ARCH (GARCH). *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, Bollerslev, T., Russell, J.R., Watson, M.W. (eds). Oxford, United Kingdom.

[34] Bollerslev, T., Chou, R.Y. and Kroner, K.F. 1992. ARCH Modeling in Finance. *Journal of Econometrics* 52, 5-59.

[35] Bollerslev, T., Engle, R.F. and Nelson, D.B. 1994. ARCH Models. *Handbook of Econometrics,* Engle, R.F. and McFadden, D. (eds). Amsterdam.

[36] Bollerslev, T., Kretschmer, U., Pigorsch, C. and Tauchen, G. 2009. A Discrete-Time Model for Daily S&P500 Returns and Realized Variations: Jumps and Leverage Effects. *Journal of Econometrics* 150, 151-166.

[37] Box, G.E.P. and Jenkins, G.M. 1970. *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day.

[38] Broadie, M., Chernov, M. and Johannes, M. 2007. Model Specification and Risk Premia: Evidence from Futures Options. *The Journal of Finance* LXII, 1453 - 1490.

[39] Broto, C. and Ruiz, E. 2004. Estimation Methods for Stochastic Volatility Models: A Survey. *Journal of Economic Surveys* 18, 613-649.

[40] Brownlees, C.T. and Gallo, G.M. 2006. Financial Econometrics Analysis at Ultra-High Frequency: Data Handling Concerns, *Computational Statistics and Data Analysis* 51, 2232-2245.

[41] Bu, R. and McCabe, B.P.M. 2008. Model Selection, Estimation and Forecasting in INAR(p) Models: A Likelihood based Markov Chain Approach, *International Journal of Forecasting* 24, 151-162.

[42] Bu, R., Hadri, K. and McCabe, B.P.M. (2008). Maximum Likelihood Estimation of Higher-order Integer Valued Autoregressive Processes, *Journal of Time Series Analysis* 29, 973-994.

[43] Busch, T., Christensen, B.J. and Nielsen, M.O. 2011. The Role of Implied Volatility in Forecasting Future Realized Volatility and Jumps in Foreign Exchange, Stock and Bond Markets. *Journal of Econometrics* 160, 48-57.

[44] Cappe, O., Godsill, S.J. and Moulines, E. 2007. An Overview of Existing Methods and Recent Advances in Sequential Monte Carlo, *IEEE Proceedings* 95, 899-924.

[45] Carney, M. and Cunningham, P. 2006. Evaluating Density Forecasting Models. *Proceedings of 17th Irish Conference on Artificial Intelligence and Cognitive Science*, 121-130.

[46] Caron, F., Davy, M., Doucet, A. and Duflos, E. 2008. Bayesian Inference for Linear Dynamic Models with Dirichlet Process Mixtures. *IEEE Transactions on Signal Processing.* 56, 71-84.

[47] Carter, C. and Kohn, R. 1994. On Gibbs Sampling for State Space Models. *Biometrika* 81, 541-553.

[48] Chernov, M., Gallant, A.R., Ghysels, E. and Tauchen, G. 2003. Alternative Models for Stock Price Dynamics. *Journal of Econometrics* 116, 225-257.

[49] Clements, A., Hurn, S. and White, S. 2006. Estimating Stochastic Volatility Models Using a Discrete Non-Linear Filter. Working Paper No. 3. *National Centre for Econometric Research.* Available at http://www.ncer.edu.au/papers/documents/WPNo3.pdf.

[50] Corradi, V and Swanson, N, 2006. Predictive Density and Conditional Confidence Interval Accuracy Tests. *Journal of Econometrics* 135, 187–228.

[51] Creal, D.D. 2008. Analysis of Filtering and Smoothing Algorithms for Levy-driven Stochastic Volatility Models. *Computational Statistics and Data Analysis* 52, 2863-2876.

[52] Czado, C., Gneiting, T. and Held, L. 2009. Predictive Model Assessment for Count Data. In press, *Biometrics.*

[53] Danielsson, J. and de Vries, C. 2000. Value at Risk and Extreme Returns. *Annals of Economics and Statistics* 60, 239-270.

[54] Dawid, A.P. 1984. Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach. *Journal of the Royal Statistical Society, Series A* 147, No. 2, 278-292.

[55] De Jong, P. and Shephard, N. 1995. The Simulation Smoother for Time Series Models. *Biometrika* 82, 339-350.

[56] De Raaig, G. and Raunig, B. 2005. Evaluating Density Forecasts from Models of Stock Market Returns. *The European Journal of Finance* 11, 151-166.

[57] De Rossi, G. and Harvey, A. 2009. Quantiles, Expectiles and Splines. *Journal of Econometrics* 152, 179-85.

[58] Diebold, F.X. 2004. The Nobel Memorial Prize for Robert F. Engle. *Scandinavian Journal of Economics* 106, 165-185.

[59] Diebold, F.X. and Lopez, J.A. 1996. Forecast Evaluation and Combination. *Handbook of Statistics* 14, *Statistical Methods in Finance*, Maddala, G.S, Rao, C.R. (eds). North-Holland: Amsterdam, 241–268.

[60] Diebold, F.X. and Mariano, R.S. 1995. Comparing Predictive Accuracy. *Journal of Business and Economic Statistics* 13, 253-263.

[61] Diebold, F.X., Gunther, T.A. and Tay, A.S. 1998. Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review* 39, 863-883.

[62] Doucet, A., de Freitas and Gordon, N.J. 2001. *Sequential Monte Carlo Methods in Practice.* New York: Springer-Verlag.

[63] Durbin, J. and Harvey, A.C. 1985. *The Effects of Seat Belt Legislation on Road Casualties in Great Britain: Report on Assessment of Statistical Evidence.* Annexe to Compulsary Seat Belt Wearing Report, Department of Transport, London, HMSO.

[64] Durbin, J, and Koopman, S.J. 2000. Time Series Analysis of Non-Gaussian Observations Based on State Space Models from Classical and Bayesian Perspectives. *Journal of the Royal Statistical Society, Series B* 62, 3-56.

[65] Durbin, J. and Koopman, S.J. 2001. *Time Series Analysis by State Space Methods*. New York: Oxford University Press.

[66] Durham, G.B. 2007. SV Mixture Models with Application to S&P 500 Index Returns. *Journal of Financial Econometrics* 85, 822-856.

[67] Engle, R.F. 1982. Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation. *Econometrica* 50, 987-1008.

[68] Engle, R.F. and Gonzalez-Rivera, G. 1991. Semiparametric ARCH Models. *Journal of Business and Economic Statistics* 9, 345-359.

[69] Engle, R.F. and Manganelli, S. 2004. CAViar: Conditional Autoregressive Value at Risk by Regression Quantiles. *Journal of Business and Economic Statistics* 22, 367-381.

[70] Eraker, B. 2004. Do Stock Prices and Volatility Jump? Reconciling Evidence from Spot and Option Prices. *The Journal of Finance* 59, 1367-1403.

[71] Eraker, B., Johannes, M.S. and Polson, N.G. 2003. The Impact of Jumps in Returns and Volatility. *Journal of Finance* 58, 1269–1300.

[72] Fernandez, C. and Steel, M.F.J. 1998. On Bayesian Modelling of Fat Tails and Skewness. *Journal of the American Statistical Association* 93, 359-371.

[73] Freeland, R. and McCabe, B. 2004a. Forecasting Discrete Valued Low Count Time Series. *International Journal of Forecasting* 20, 427-434.

[74] Freeland, R. and McCabe, B. 2004b. Analysis of Low Count Time Series Data by Poisson Autoregression. *Journal of Time Series Analysis* 25, 701-722.

[75] Frühwirth-Schnatter, S. 1994. Data Augmentation and Dynamic Linear Models. *Journal of Time Series Analysis* 15, 183-202.

[76] Frühwirth-Schnatter, S. 2004. Efficient Bayesian Parameter Estimation for State Space Models Based on Reparameterisations. *State Space and Unobserved Component Models: Theory and Applications*, Cambridge University Press.

[77] Garthwaite, P.H., Kadane, J.B. and O'Hagan, A. 2005. Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association* 100, 680-700.

[78] Geweke, J. and Amisano, G. 2010. Comparing and Evaluating Bayesian Predictive Distributions of Asset Returns, *International Journal of Forecasting (Special Issue on Applied Bayesian Forecasting in Economics)* 26, 216-230.

[79] Ghysels, E., Harvey, A.C. and Renault, E., 1996. Stochastic Volatility. *Statistical Methods in Finance*, Rao, C.R., Maddala, G.S. (eds). North-Holland, Amsterdam.

[80] Giacomini, R. and Komunjer, I. 2005. Evaluation and Combination of Conditional Quantile Forecasts. *Journal of Business and Economic Statistics* 23, 416-431.

[81] Giordani, P., Pitt, M. and Kohn, R. 2011. Bayesian Inference for Time Series State Space Models. *The Oxford Handbook of Bayesian Econometrics.* Geweke, J., Koop, G. and van Dijk, H. (eds). Oxford University Press.

[82] Gneiting, T. 2008. Editorial: Probabilistic Forecasting. *Journal of the Royal Statistical Society, Series A* 171, 319-321.

[83] Gneiting, T. 2011. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association* 106, 746-762.

[84] Gneiting, T. and Raftery, A.E. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 102, 359-378.

[85] Gneiting, T., Balabdaoui, F. and Raftery, A.E. 2007. Probabilistic Forecasts, Calibration and Sharpness. *Journal of Royal Statistical Society, Series B* 69, 243-268.

[86] Gordon, N.J., Salmond, D.J. and Smith, A.F.M. 1993. A Novel Approach to Nonlinear and Non-Gaussian Bayesian State Estimation. *IEEE Proceedings F* 140, 107-113.

[87] Hall, P., Racine, J. and Li, Q. 2004. Cross-Validation and the Estimation of Conditional Probability Densities. *Journal of the American Statistical Association* 99, 1015-1026.

[88] Hamilton, J. 1994. *Time Series Analysis.* Princeton: Princeton University Press.

[89] Hansen, B.E. 2004. Nonparametric Conditional Density Estimation. Working paper. University of Wisconsin, United States of America. Available at http://www.ssc.wisc.edu/~bhansen/papers/ncde.pdf.

[90] Harvey, A.C. and Chung, C.H. 2000. Estimating the Underlying Change in Unemployment in the UK. *Journal of Royal Statistical Society* A, 163, 303-39.

[91] Harvey, A.C. and Durbin, J. 1986. The Effects of Seat Belt Legislation on British Road Casualties: A Case Study in Structural Time Series Modelling. *Journal of Royal Statistical Society* A, 149, 187-227.

[92] Harvey, A.C., Ruiz, E. and Shephard, N. 1994. Multivariate Stochastic Variance Models. *Review of Economic Studies* 61, 247-264.

[93] Hassani, S. 2009. *Mathematical Methods for Students of Physics and Related Fields (2nd Edition).* New York: Springer-Verlag.

[94] Hogg, R.V. and Craig, A.T. 1965. *Mathematical Statistics.* New York: Macmillan.

[95] Hong, Y., Li, H. and Zhou, F. 2004. Out-of-Sample Performance of Discrete Spot Interest Rate Models. *Journal of Business and Economic Statistics* 22, 457-473.

[96] Jacquier, E. and Miller, S. 2010. The Information Content of Realized Volatility. *Working Paper*, HEC, University of Montreal.

[97] Jacquier, E., Polson, N. and Rossi, P. 1994. Bayesian Analysis of Stochastic Volatility Models. *Journal of Business and Economic Statistics* 12, 69-87.

[98] Jensen, M.J. and Maheu, J.M. 2010. Bayesian Semiparametric Stochastic Volatility Modeling. *Journal of Econometrics* 157, 306-316.

[99] Julier, S.J. and Uhlmann, J.K. 1997. A New Extension of the Kalman Filter to Nonlinear Systems. *Proc. AeroSense: 11th Int. Symp. on Aerospace/Defence Sensing, Simulation and Controls*, 182-193.

[100] Julier, S.J. and Uhlmann, J.K. 2004. Unscented Filtering and Nonlinear Estimation. *Proceedings of the IEEE* 92, 401-422.

[101] Julier, S.J., Uhlmann, J.K. and Durrant-Whyte, H.F. 1995. A New Approach for Filtering Nonlinear Systems. *Proceedings of the American Control Conference*, 1628-1632.

[102] Julier, S.J., Uhlmann, J.K. and Durrant-Whyte, H.F. 1996. A New Approach for the Nonlinear Transformation of Means and Covariances in

Linear Filters. *IEEE Transactions on Automatic Control*, Accepted for publication as a Technical Note.

[103] Jung, R.C. and Tremayne, A. R. 2006. Coherent Forecasting in Integer Time Series Models. *International Journal of Forecasting* 22, 223-238.

[104] Johnson, N.L., Kotz, S. and Balakrishnan, N. 1994. *Distributions in Statistics: Continuous Univariate Distributions*, Vol. 2. Wiley, New York.

[105] Kalman, R.E. 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions ASME - Journal of Basic Engineering, Series D,* 82, 35-45.

[106] Kalman, R.E. and Bucy, R.S. 1961. New Results in Linear Filtering and Prediction Theory. *Transactions ASME - Journal of Basic Engineering, Series D*, 83, 95-108.

[107] Khuri, A.I. 2004. Application of Dirac's Delta Function in Statistics. *International Journal of Mathematical Eduction in Science and Technology* 35, 185-195.

[108] Kitagawa, G. 1987. Non-Gaussian State Space Modeling of Nonstationary Time Series. *Journal of the American Statistical Association* 76, 1032-1064.

[109] Kitagawa, G. 1989. Non-Gaussian Seasonal Adjustment. *Computers and Mathematics with Application* 18, 503-514.

[110] Kitagawa, G. 1994. The Two-Filter Formula for Smoothing and an Implementation of the Gaussian-sum Smoother. *Annals of the Institute of Statistical Mathematics* 46, 605-623.

[111] Kitagawa, G. 1996. Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *Journal of Computational and Graphical Statistics* 5, 1-25.

[112] Koopman, S.J., Jungbacker, B. and Hol, E. 2005. Forecasting Daily Variability of the S&P100 Stock Index Using Historical, Realized and Implied Volatility Measurements. *Journal of Empirical Finance* 12, 445-475.

[113] Li, Q. and Racine, J.S. 2007. *Nonparametric Econometrics.* Princeton University Press.

[114] Lim, G., Martin, G. and Martin, V. 2005. Parametric Pricing of Higher Order Moments in S&P500 Options. *Journal of Applied Econometrics* 20, 377-404.

[115] Liu, C. and Maheu, J. 2009. Forecasting Realized Volatility: A Bayesian Model-Averaging Approach, *Journal of Applied Econometrics* 24, 704-733.

[116] Maneesoonthorn, W., Martin, G.M., Forbes, C.S. and Grose, S. 2011. Probabilistic Forecasts of Volatility and its Risk Premia. *Working Paper*, Monash University. Available at http://www.buseco.monash.edu.au/ebs/pubs/wpapers/2010/wp22-10.pdf.

[117] Martens, M. and Zein, J. 2004. Predicting Financial Volatility: High-Frequency Time Series Forecasts Vis-a-Vis Implied Volatility. *Journal of Futures Markets* 24, 1005-1028.

[118] Martens, M., van Dijk, D. and de Pooter, M. 2009. Forecasting S&P500 Volatility: Long Memory, Level Shifts, Leverage Effects, Day-of-the-Week Seasonality, and Macroeconomic Announcements. *International Journal of Forecasting* 25, 282-303.

[119] Martin, G., Reidy, A. and Wright, J. 2009. Does the Option Market Produce Superior Forecasts of Noise-corrected Volatility Measures? *Journal of Applied Econometrics* 24, 77-104.

[120] McCabe, B. and Martin, G. 2005. Bayesian Predictions of Low Count Time Series. *International Journal of Forecasting* 21, 315-330.

[121] McCabe, B., Martin, G.M. and Harris, D. 2011. Efficient Probabilistic Forecasts for Counts. *Journal of the Royal Statistical Society, Series B* 73, 253-272.

[122] Meddahi, N. 2002. A Theoretical Comparison Between Integrated and Realized Volatility. *Journal of Applied Econometrics* 17, 479-508.

[123] Monteiro, A.A. 2010. A Semiparametric State Space Model. Working paper, *Statistics and Econometrics Series,* Universidad Carlos III de Madrid, Spain. Available at http://e-archivo.uc3m.es/bitstream/10016/9247/1/ws103418.pdf.

[124] Pascual, L., Romo, J. and Ruiz, E. 2001. Effects of Parameter Estimation on Prediction Densities: A Bootstrap Approach. *International Journal of Forecasting* 17, 83-103.

[125] Pascual, L., Romo, J. and Ruiz, E. 2006. Bootstrap Prediction for Returns and Volatilities in GARCH Models. *Computational Statistics & Data Analysis* 50, 2293-2312.

[126] Pitt, M.K. 2002. Smooth Particle Filters for Likelihood Evaluation and Maximisation. *The Warwick Economics Research Paper Series (TWERPS)* 651, University of Warwick, Department of Economics. Available at http://ideas.repec.org/p/wrk/warwec/651.html.

[127] Pitt, M.K. and Shephard, N. 1999. Filtering via Simulation: Auxiliary Particle Filters. *Journal of the American Statistical Association* 94, 590-599.

[128] Politis, D. N., Romano, J. P. and Wolf, M. 1999. Subsampling. New York: Springer.

[129] Pong, S., Shackleton, M.B., Taylor, S.J. and Xu, X. 2004. Forecasting Currency Volatility: A Comparison of Implied Volatilities and AR(FI)MA Models. *Journal of Banking and Finance* 28, 2541-2563.

[130] Rodriguez, A. and Ruiz, E. 2009. Bootstrap Prediction Intervals in State-Space Models. *Journal of Time Series Analysis* 30, 167-178.

[131] Rosenblatt, R.F. 1952. Remarks on a Multivariate Transformation. *Annals of Mathematical Statistics* 23, 470-472.

[132] Rosenblatt, M. 1969, Conditional Probability Density and Regression Estimates. In P. R. Krishnaiah (Ed.), *Multi-Variate Analysis* II, 25-31. Academic Press, New York.

[133] Scott, D.W., Tapia, R.A. and Thompson, J.R. 1980. Nonparametric Probability Density Estimation by Discrete Maximum Penalized-Likelihood Criteria. *Annals of Statistics* 8, 820-832.

[134] Selten, R. 1998. Axiomatic Characterization of the Quadratic Scoring Rule. *Experimental Economics* 1, 43-62.

[135] Shephard, N. 1996. Statistical Aspects of ARCH and Stochastic Volatility. *Time Series Models*, Cox, D.R., Hinkley, D.V., Barndorff-Nielsen, O.E. (eds). Chapman and Hall, London.

[136] Shephard, N. 2005. *Stochastic Volatility: Selected Readings (Advanced Texts in Econometrics)*. Oxford University Press.

[137] Shephard, N. and Pitt, M.K. 1997. Likelihood Analysis of Non-Gaussian Measurement Time Series. *Biometrika* 84, 653-667.

[138] Sorenson H.W. and Alspach D.L. 1971. Recursive Bayesian Estimation Using Gaussian-sums. *Automatica* 7, 465-479.

[139] Strickland, C.M., Forbes, C.S. and Martin, G.M. 2006. Bayesian Analysis of the Stochastic Conditional Duration Model. *Computational Statistics and Data Analysis (Special Issue on Statistical Signal Extraction and Filtering)* 50, 2247-2267.

[140] Stroud, J.R., Muller, P. and Polson, N.G. 2003. Nonlinear State-Space Models with State-Dependent Variances. *Journal of the American Statistical Association* 98, 377-386.

[141] Tauchen, G. and Zhou, H. 2011. Realized Jumps on Financial Markets and Predicting Credit Spreads. *Journal of Econometrics* 160, 102-118.

[142] Tay, A. and Wallis, K. 2000. Density Forecasting: A Survey. *Journal of Forecasting* 19, 235-254.

[143] Taylor, S.J. 1986. *Modeling Financial Time Series.* Wiley, Chichester.

[144] Taylor, S.J. 1994. Modeling Stochastic Volatility: A Review and Comparative Study. *Mathematical Finance* 4, 183-204.

[145] Wan, E.A. and van der Merwe, R. 2001. *Chapter 7: The Unscented Kalman Filter, In Kalman Filtering and Neural Networks.* (Eds. S. Haykin), Wiley Publishing.

[146] Yau, C., Papaspiliopoulos, O., Roberts, G.O. and Holmes, C. 2011. Bayesian Non-parametric Hidden Markov Models with Applications in Genomics. *Journal of the Royal Statistical Society, Series B* 73, 37-57.

[147] Zhou, B. 1996. High-Frequency Data and Volatility in Foreign Exchange Rates. *Journal of Business and Economic Statistics* 14, 45-52.