



MONASH University

Australia

Department of Econometrics and Business Statistics

<http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/>

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

**Jason Ng, Catherine S. Forbes, Gael M. Martin and
Brendan P.M. McCabe**

September 2011

Working Paper 11/11

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

Jason Ng[†], Catherine S. Forbes[‡], Gael M. Martin[§] and Brendan P.M. McCabe[¶]

August 31, 2011

Abstract

The object of this paper is to produce non-parametric maximum likelihood estimates of forecast distributions in a general non-Gaussian, non-linear state space setting. The transition densities that define the evolution of the dynamic state process are represented in parametric form, but the conditional distribution of the non-Gaussian variable is estimated non-parametrically. The filtering and prediction distributions are estimated via a computationally efficient algorithm that exploits the functional relationship between the observed variable, the state variable and a measurement error with an invariant distribution. Simulation experiments are used to document the accuracy of the non-parametric method relative to both correctly and incorrectly specified parametric alternatives. In an empirical illustration, the method is used to produce sequential estimates of the forecast distribution of realized volatility on the S&P500 stock index during the recent financial crisis. A resampling technique for measuring sampling variation in the estimated forecast distributions is also demonstrated.

KEYWORDS: Probabilistic Forecasting; Non-Gaussian Time Series; Grid-based Filtering; Penalized Likelihood; Subsampling; Realized Volatility.

JEL CODES: C14, C22, C53.

*This research has been supported by Australian Research Council (ARC) Discovery Grant DP0985234 and ARC Future Fellowship FT0991045. The authors would like to thank participants at the National PhD Conference in Economics and Business (November, 2010, Australian National University); the 7th International Symposium on Econometric Theory and Applications (April, 2011, Monash University); the International Symposium of Forecasting (June, 2011, University of Prague); the 8th Workshop on Bayesian Nonparametrics (June, 2011, Veracruz, Mexico); the Australasian Econometric Society Meeting (July, 2011, University of Adelaide); and the Frontiers in Financial Econometrics Workshop (July, 2011, Queensland University of Technology) for constructive comments on earlier drafts of the paper.

[†]Department of Econometrics and Business Statistics, Monash University, Australia.

[‡]Department of Econometrics and Business Statistics, Monash University, Australia

[§]Corresponding author: Department of Econometrics and Business Statistics, Monash University, Australia. Email: gael.martin@monash.edu.

[¶]Management School, University of Liverpool, UK.

1 Introduction

The focus of this paper is on forecasting non-Gaussian time series variables. Such 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 (Freeland and McCabe, 2004; McCabe and Martin, 2005; Bu and McCabe, 2008; Czado *et al.*, 2009; McCabe, Martin and Harris, 2011; Bauwens *et al.*, 2004; Amisano and Giacomini, 2007). In addition, the need to forecast the probability of large financial losses has been the primary reason for the recent focus on distributional forecasting of portfolio returns (Diebold *et al.*, 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 (Giacomini and Komunjer, 2005; de Rossi and Harvey, 2009). The extraction of risk-neutral distributional forecasts of non-Gaussian asset returns from derivative prices (Ait-Sahalia and Lo, 1998; Bates, 2000; Lim *et al.*, 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

In the spirit of this evolving literature, we develop a new method for estimating the full forecast distribution of non-Gaussian time series variables. In contrast to the existing literature, in which the focus is almost exclusively on the specification of strict parametric models, 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 data generating process (DGP).

¹Discussion of the merits of probabilistic forecasting in general is provided by, amongst others, Dawid (1984), Tay and Wallis (2000), Gneiting *et al.* (2007) and Gneiting (2008).

The method is developed within the general framework of non-Gaussian, non-linear state space models, 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.²

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, Simpson's rule) can be used to estimate the relevant integrals. The ordinates of the (unknown) measurement density are estimated as unknown parameters using maximum likelihood (ML) methods, with this aspect drawing on the recent work of Berg *et al.* (2010) on discrete penalized likelihood (see also Scott *et al.*, 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 the Gaussian sum filter, an alternative method for avoiding a strict parametric specification for the measurement distribution (see Sorenson and Alspach, 1971; Kitagawa, 1994; Monteiro, 2010). 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 *et al.* 2008; Jensen and Maheu, 2010; Yau, Papaspiliopoulos, Roberts and Holmes, 2010).

Extensive simulation exercises are used to assess the predictive accuracy of the non-parametric method, against both correctly specified and misspecified parametric alternatives, and for a variety of DGPs. Assessment of the resulting forecast distributions is based on a range of comparative and evaluative methods including predictive score, probability integral transform and coverage methods. 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. Using the approach developed in McCabe, Martin and Harris (2011), resampling is used to cater for estimation uncertainty in the production of the probabilistic forecasts of volatility. (See also Rodriguez and Ruiz, 2009).

²Note that the terms 'forecast distribution' and 'predictive (or prediction) distribution' are used interchangeably in the paper.

The outline of the paper is as follows. In Section 2 we describe 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. In Section 3.1, we discuss the linear and non-linear models considered in the simulation experiments. In Section 3.2 we then outline the various tools used to compare and evaluate the predictive distributions obtained via the non-parametric and parametric methods, with simulation results then presented in Section 3.3. The empirical application is documented in Section 4, with details given of the subsampling method used to measure the impact of sampling variation on the estimated forecast distribution. Section 5 concludes.

2 Non-parametric Estimation of the Forecast Distribution

2.1 An Outline of the Basic Approach

Our non-parametric estimate of a forecast distribution is developed within the context of a general non-Gaussian, non-linear state space model for a scalar random variable y_t . Consider the system governed by a measurement equation for y_t and a transition equation for a scalar state variable x_t ,

$$y_t = h_t(x_t, \eta_t) \tag{1}$$

$$x_t = k_t(x_{t-1}, v_t), \tag{2}$$

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

$$G(x_t) = y_t - h_t(x_t, \eta_t)$$

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. For the sake of generality we focus on the case where y_t is continuous, with all distributions expressed using density functions as a consequence. 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, although the simple grid-based method emphasized here is clearly most suitable for reasonably low-dimensional problems. We also focus here on the case where $x_t^*(y_t, \eta_t)$ is analytically available, in addition to being unique, with adaptation of the method obviously required when neither of these conditions are met.

As is common, we assume that η_t is independent of x_t , in which case the probability density function (pdf) for η_t is simply $p(\eta_t|x_t) = p(\eta_t)$, for all $t = 1, 2, \dots, T$. We also assume time-series

independence for η_t ; 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 $p(\eta_t)$, we allow its distributional form to be unknown. An initial (parametric) distribution $p(x_1)$ is specified for the scalar state, with the transition densities resulting from (2) denoted by $p(x_t|x_{t-1})$, $t = 2, 3, \dots, T$. In the examples considered in the paper (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 , and k_t is such that the transition densities $p(x_t|x_{t-1})$ are available. To avoid unnecessary notation, we suppress the t subscript on the functions h and k from this point.

Given the model defined by (1) and (2), the one-step ahead forecast distribution for y_{T+1} , conditional on the observed data, $y_{1:T} = (y_1, y_2, \dots, y_T)'$, is

$$p(y_{T+1}|y_{1:T}) = \int p(y_{T+1}|x_{T+1}) p(x_{T+1}|y_{1:T}) dx_{T+1}, \quad (3)$$

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 $\{p(x_t|x_{t-1}), t = 2, 3, \dots, T\}$, has been suppressed. The primary aim of the paper 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), 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 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} .

Using standard prediction error decomposition, the likelihood function for the collection of all unknown fixed parameters $\boldsymbol{\theta}$, augmented, in the current context, by the unknown ordinates of $p(y_{T+1}|x_{T+1})$, is given by

$$L(\boldsymbol{\theta}) \propto p(y_1) \prod_{t=1}^{T-1} p(y_{t+1}|y_{1:t}), \quad (4)$$

where $y_{1:t} = (y_1, y_2, \dots, y_t)'$. The likelihood function thus requires the availability of the one-step

ahead prediction distributions,

$$p(y_{t+1}|y_{1:t}), \quad t = 1, 2, \dots, T - 1 \quad (5)$$

and the marginal distribution

$$p(y_1), \quad (6)$$

where both (5) and (6) are (suppressed) functions of θ . In the following section we outline a computationally efficient filtering algorithm for computing (5) and (6), needed for the specification of the likelihood function in (4). 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) is estimated, with sampling error able to be quantified in empirical settings using resampling methods, as illustrated in Section 4.3.

Crucially, the computational burden associated with evaluation of the likelihood function in (4) 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 is known to be geometric in T (see, for example, Kitagawa, 1994). 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, Maskell, Gordon and Clapp, 2002; Clements, Hurn and White, 2006).

2.2 A Grid-based Filter

The objective of filtering is to update knowledge of the system each time a new value of y_t is observed. Within the general state space model in (1) and (2), along with the initial distribution $p(x_1)$, filtering determines the distribution of the state vector, x_t , given a portion of the observed data, $y_{1:t}$, as represented by the filtered density $p(x_t|y_{1:t})$, for $t = 1, 2, \dots, T$. Therefore, filtering is a recursive procedure that is applied for each t , revising the filtered density, $p(x_t|y_{1:t})$, using the new observation y_{t+1} , to produce the updated density, $p(x_{t+1}|y_{1:t+1})$.

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. Our approach exploits the functional relationship between the observation y_t and the *i.i.d.* variable η_t , for given x_t , in (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.

2.2.1 Preliminaries

The δ -function³ may be represented as

$$\delta(z^* - z) = \begin{cases} \infty & \text{if } z^* = z \\ 0 & \text{if } z^* \neq z \end{cases}$$

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

$$\int_{-\infty}^{\infty} f(z) \delta(z^* - z) dz = f(z^*), \quad (7)$$

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

$$\int_{-\infty}^{\infty} f(z) \delta(G(z)) dz = \int_{-\infty}^{\infty} f(z) \left| \frac{\partial G(z)}{\partial z} \right|^{-1} \delta(z - z^*) dz, \quad (8)$$

resulting, via (7), in

$$\int_{-\infty}^{\infty} f(z) \delta(G(z)) dz = f(z^*) \left| \frac{\partial G(z)}{\partial z} \right|_{z=z^*}^{-1}, \quad (9)$$

where $\left| \frac{\partial G(z)}{\partial z} \right|_{z=z^*}$ denotes the modulus of the derivative of $G(z)$, evaluated at $z = z^*$. The transformation in (8) makes it explicit that the root of the argument of the δ function is $z = z^*$, and as a consequence of this result, we sometimes write

$$\delta(G(z)) = \left| \frac{\partial G(z)}{\partial z} \right|^{-1} \delta(z - z^*) \quad (10)$$

when considering the composite function $\delta(G(z))$ explicitly in terms of z . 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, we use the δ -function to express the transformation in (1) from the iid measurement error η_t to the observed data y_t , given x_t , so that

$$p(y_t|x_t) = \int_{-\infty}^{\infty} p(\eta) \delta(y_t - h(x_t, \eta)) d\eta, \quad (11)$$

³Strictly speaking, $\delta(x)$ is a generalized function, and is properly defined as a measure rather than as a function. However, we take advantage of the commonly used heuristic definition here as it is more convenient for the filtering manipulations that are to follow in the next section. See, for example, Hassani (2009).

where η is a variable of integration that traverses the support of $p(\eta)$. This result, along with the transformation of variables relation in (10), enables all integrals required to produce the likelihood function in (4) to be expressed in terms of the measurement error variable, η .

2.2.2 The initial filtered distribution: $p(x_1|y_1)$

Using the representation of the measurement density as an integral involving the δ -function in (11), it follows that the filtered density of the state variable at time $t = 1$ may be expressed as

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

We simplify the expression of the resulting filtered density in two ways. First, the numerator is written in terms of the state variable using (10). Second, the order of integration is reversed and (8) and (9) used in the denominator to obtain

$$p(x_1|y_1) = \frac{p(x_1) \int_{-\infty}^{\infty} p(\eta) \left| \frac{\partial h}{\partial x_1} \right|^{-1} \delta(x_1 - x_1^*(y_1, \eta)) d\eta}{\int_{-\infty}^{\infty} p(x_1^*(y_1, \eta)) p(\eta) \left| \frac{\partial h}{\partial x_1} \right|_{x_1=x_1^*(y_1, \eta)}^{-1} d\eta}, \quad (12)$$

where $x_1^*(y_1, \eta)$ is the (assumed unique) solution to $y_1 - h_1(x_1, \eta) = 0$ for any value η in the support of $p(\eta)$.

Next, to numerically evaluate the filtered distribution in (12) via rectangular integration, an evenly spaced grid $\{\eta^1, \eta^2, \dots, \eta^N\}$ is defined, with interval length m , resulting in the approximation for $p(x_1|y_1)$ given by

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

where $p(\eta^j)$ 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 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

$$p(x_1|y_1) = \sum_{j=1}^N W_1^j \delta(x_1 - x_1^{*j}), \quad (13)$$

and where

$$W_1^j = \frac{p(\eta^j) \left| \frac{\partial h}{\partial x_1} \right|_{x_1=x_1^{*j}}^{-1} p(x_1^{*j})}{\sum_{i=1}^N p(\eta^i) \left| \frac{\partial h}{\partial x_1} \right|_{x_1=x_1^{*i}}^{-1} p(x_1^{*i})}, \quad (14)$$

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

$$p(y_1) = m \sum_{i=1}^N p(\eta^i) \left| \frac{\partial h}{\partial x_1} \right|_{x_1=x_1^{*i}}^{-1} p(x_1^{*i}), \quad (15)$$

obtained from approximating the denominator in (12).

Having obtained the representation in (13) for time $t = 1$, we show that for any time $t = 2, 3, \dots, T$, an appropriate discrete distribution can be found to approximate the filtered distribution

$$p(x_t | y_{1:t}) = \sum_{j=1}^N W_t^j \delta(x_t - x_t^{*j}), \quad (16)$$

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

$$x_t^{*j} = x_t^*(y_t, \eta^j) \quad (17)$$

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

⁴Note that densities employing the Dirac delta notation should be interpreted carefully. In (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

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

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

2.2.3 The predictive distribution for the state: $p(x_{t+1}|y_{1:t})$

Assuming (16) holds in period t , it follows that the one-step ahead state prediction density is a mixture of transition densities, since

$$\begin{aligned}
p(x_{t+1}|y_{1:t}) &= \int p(x_{t+1}|x_t) p(x_t|y_{1:t}) dx_t \\
&= \int p(x_{t+1}|x_t) \sum_{j=1}^N W_t^j \delta(x_t - x_t^{*j}) dx_t \\
&= \sum_{j=1}^N \int W_t^j p(x_{t+1}|x_t) \delta(x_t - x_t^{*j}) dx_t \\
&= \sum_{j=1}^N W_t^j p(x_{t+1}|x_t^{*j}), \tag{18}
\end{aligned}$$

for $t = 1, 2, \dots, T$. The notation $p(x_{t+1}|x_t^{*j})$ denotes the transition density of $p(x_{t+1}|x_t)$, 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 $p(x_{t+1}|x_t)$ are available, no additional approximation is needed in moving from $p(x_t|y_{1:t})$ to $p(x_{t+1}|y_{1:t})$.

2.2.4 The one-step ahead predictive distribution for the observed: $p(y_{t+1}|y_{1:t})$

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

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

Utilizing (11) for $p(y_{t+1}|x_{t+1})$, the one-step ahead prediction density has representation

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

which, after integration with respect to x_{t+1} (and using (9) once again), yields

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

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

$$p(y_{t+1}|y_{1:t}) = m \sum_{i=1}^N p(\eta^i) \left| \frac{\partial h}{\partial x_{t+1}} \right|_{x_{t+1}=x_{t+1}^*(y_{t+1}, \eta^i)}^{-1} p(x_{t+1}^*(y_{t+1}, \eta^i)|y_{1:t}). \tag{19}$$

Noting that $p(x_{t+1}^*(y_{t+1}, \eta^i)|y_{1:t})$ in (19) denotes the one-step ahead predictive density from (18) evaluated at $x_{t+1} = x_{t+1}^*(y_{t+1}, \eta^i)$, it can be seen that $p(y_{t+1}|y_{1:t})$ is computed as an N^2 mixture of (specified) transition density functions as a consequence.

2.2.5 The updated filtered distribution: $p(x_{t+1}|y_{1:t+1})$

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

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

for $t = 1, 2, \dots, T - 1$, and where $x_{t+1}^{*j} = x_{t+1}^*(y_{t+1}, \eta^j)$ 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 (16), with density

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

where, for $j = 1, 2, \dots, N$,

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

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 $j = 1, 2, \dots, N$.

2.2.6 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 $x_t^{*j} = x_t^*(y_t, \eta^j)$ the unique zero of $y_t - h(x_t, \eta^j)$, for each $j = 1, 2, \dots, N$ and all $t = 1, 2, \dots, T$. Initialize the filter at period 1 with (13) and (14). For $t = 1, 2, \dots, T - 1$,

$$p(x_{t+1}|y_{1:t}) = \sum_{j=1}^N W_t^j p(x_{t+1}|x_t^{*j}), \quad (20)$$

$$p(y_{t+1}|y_{1:t}) = \sum_{i=1}^N M_{t+1}^i(y_{t+1}) p(x_{t+1}^{*i}(y_{t+1}, \eta^i)|y_{1:t}) \quad (21)$$

$$p(x_{t+1}|y_{1:t+1}) = \sum_{j=1}^N W_{t+1}^j \delta(x_{t+1} - x_{t+1}^{*j}), \quad (22)$$

with

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

and

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

The computational burden involved in the evaluation of the t th component of the likelihood function ($p(y_{t+1} | y_{1:t})$) 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 $p(\eta^j)$ (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.

Our approach has two key benefits. Firstly, 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 nonlinear state space model, 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 . Secondly, and the case of interest here, when the measurement error density, $p(\eta)$, is *unknown*, the mass associated with each of the grid points resulting from the rectangular integration procedure,

$$g^j = p(\eta^j) m, \tag{23}$$

for $j = 1, 2, \dots, N$, 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.

We complete this section by noting that the non-parametric filter could, in principle, be replaced by a filter in which the measurement error density at each grid point is represented as a K -mixture of normal distributions: $p(\eta^j) = \frac{1}{b} \sum_{k=1}^K g^k \phi\left(\frac{\eta^j - \eta^k}{b}\right)$, where η^k , $k = 1, 2, \dots, K$ are the K local grid points on which the normal mixtures are centred, while $\{\eta^j, j = 1, 2, \dots, N\}$ represent the grid-points in the support of η over which integration is performed.⁵ The parameter b is the standard deviation of each mixture density, assumed here to be constant across all

⁵Note that a mixture of non-normal parametric distributions is also possible.

the mixture densities and g^k is the unknown weight attached to the k^{th} mixture, which could again be estimated via ML. 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 $T(N^2K)$ (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 to us not pursuing this modification further in this paper. However, it is worth noting that this less flexible representation of $p(\eta)$ may produce some computational gains, relative to the non-parametric representation, in the high-dimensional case, given that the number of weights to be estimated, K , is independent of the dimension of η . We leave further exploration of this issue for later work, focussing here on novel and computationally feasible use of the non-parametric representation in the univariate (or low-dimensional) setting.

2.3 Penalized Log-likelihood Specification

The product of the elements $p(y_{t+1}|y_{1:t})$ in (21), for $t = 1, 2, \dots, T - 1$, along with the marginal distribution $p(y_1)$ in (15), defines the likelihood function in (4). 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, we augment the log likelihood with two components that (with reference to (23)) respectively: (i) impose smoothness on g^j as a function of j ; and (ii) penalize large values of $|\boldsymbol{\eta}'\mathbf{g} - \eta^j|$, 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

$$\ln L(\boldsymbol{\theta}) = \ln p(y_1) + \sum_{t=1}^{T-1} \ln p(y_{t+1}|y_{1:t}) - \omega \frac{1}{2} \mathbf{g}' \mathbf{H}(N, \lambda^2) \mathbf{g} - (1 - \omega) \mathbf{k}(c)' \mathbf{g}, \quad (24)$$

where

$$\mathbf{H}(N, \lambda^2) = N^3 \lambda^{-2} \boldsymbol{\Omega}' \mathbf{A} \boldsymbol{\Omega} + N^{-1} [\mathbf{e}\mathbf{e}' + \boldsymbol{\eta}\boldsymbol{\eta}'] \quad (25)$$

and $\mathbf{k}(c)$ is an $(N \times 1)$ vector with j^{th} element given by $k^j(c) = -\exp(c|\eta^j - \boldsymbol{\eta}'\mathbf{g}|)$. The matrix \mathbf{A} in (25) is an $(N - 2) \times (N - 2)$ tridiagonal matrix with $a_{jj} = 1/3$ (for $j = 1, \dots, N - 2$) and $a_{j,j+1} = a_{j+1,j} = 1/6$ ($j = 1, \dots, N - 3$); $\boldsymbol{\Omega}$ is an $(N - 2) \times N$ matrix with three nonzero elements $\Omega_{jj} = 1$, $\Omega_{j,j+1} = -2$, $\Omega_{j,j+2} = 1$ 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 (24) 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 (24) 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 $\sum_{i=1}^N g^j = 1$, $g^j \geq 0$, $j = 1, 2, \dots, N$, to produce ML estimates of the augmented θ . An estimate of the forecast distribution in (3) is subsequently produced using these estimated parameters.

3 Simulation Experiments

3.1 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 (in Section 3.1.1) 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 introduced into the measurement equation in Section 3.1.2, 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 (SCD) model; see Bauwens and Veredas (2004) and Strickland, Forbes and Martin (2006). The final model examined (in Section 3.1.3) is non-linear in both the measurement and state equations, with both Gaussian and non-Gaussian measurement errors considered. We refer to it as the realized volatility model, as the form of the model lends itself to the empirical investigation of this observable measure of latent volatility. It is, indeed, the model that underlies the empirical investigation of S&P500 volatility in Section 4.

3.1.1 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,

$$y_t = x_t + \eta_t \tag{26}$$

$$x_t = \alpha + \rho x_{t-1} + \sigma_v v_t, \tag{27}$$

where $\alpha = 0.1$, $\rho = 0.8$, $\sigma_v = 1.2$ and $v_t \sim N(0, 1)$. We entertain three different distributions for η_t : normal, 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 positively skewed skewed Student- t distribution. For the purpose of integration, the 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.

3.1.2 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, we specify an SCD model for y_t as

$$y_t = e^{x_t} \varepsilon_t \quad (28)$$

$$x_t = \alpha + \rho x_{t-1} + \sigma_v v_t, \quad (29)$$

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. We also assume 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 .⁶ Taking logarithms of (28), the measurement equation is transformed as

$$\ln(y_t) = x_t + b + \sigma_\eta \eta_t, \quad (30)$$

where $\varepsilon_t = \exp(b + \sigma_\eta \eta_t)$, $\eta_t \sim i.i.d.(0, 1)$, $b = E(\ln \varepsilon_t)$ and $\sigma_\eta^2 = var(\ln \varepsilon_t)$. We adopt three different distributions for ε_t : exponential, Weibull and gamma, with the associated expressions for b and σ_η^2 documented in Johnson *et al.* (1994). 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 log transformation of ε_t .

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 log duration. Given that this distributional assumption for η_t is incorrect, the approach based on the Kalman filter does not produce the correct

⁶Typically observed durations will 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 paper we use a t subscript 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.

forecast distribution, and we document the forecast accuracy of this (misspecified) approach in comparison with that of the non-parametric method.

3.1.3 Non-linear Model: Realized Volatility

As a second example of a non-linear state space specification, and one that is explored in Section 4, we consider the following bivariate jump diffusion process for the price of a financial asset, P_t , and its stochastic variance, V_t ,

$$\frac{dP_t}{P_t} = \mu_p dt + \sqrt{V_t} dB_t^p + dJ_t \quad (31)$$

$$dV_t = \kappa[\phi - V_t]dt + \sigma_v \sqrt{V_t} dB_t^v, \quad (32)$$

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 may 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 allowed to be correlated with a coefficient ρ . We assume, however, that dB_t^i and dJ_t are independent, for $i = \{p, v\}$. 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 variance process in (32), quadratic variation over the horizon $t - 1$ to t (assumed to be a day) is defined as

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

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

$$\mathcal{V}_{t-1,t} = \int_{t-1}^t V_s ds, \quad (33)$$

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, 2002, and Anderson *et al.*, 2003) that

$$RV_t = \sum_{t_i \in [t-1,t]}^B r_{t_i}^2 \xrightarrow{p} Q\mathcal{V}_{t-1,t}, \quad (34)$$

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 .⁷

⁷The result in (34) is based on the implicit assumption that microstructure noise effects are absent. See Martin *et al.* (2009) for a recent summary of modifications of (34) that cater for the presence of microstructure noise in the intraday prices.

We define the measurement equation as

$$\ln RV_t = \ln V_t + u_{RV_t}, \quad (35)$$

where the latent volatility evolves according to (32) and $u_{RV_t} = \ln RV_t - \ln V_t$ is the log realized volatility error. Based on the assumed DGP above, the error term in (35) will capture the effects 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 (33), 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.⁸ An Euler approximation of (32) is used to define the state equation,

$$V_t = \kappa\phi + (1 - \kappa)V_{t-1} + \sigma_v\sqrt{V_{t-1}}v_{t-1}, \quad (36)$$

where V_{t-1} = the point-in-time volatility on day $(t - 1)$ 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.

Using the generic notation of the paper, the model is thus

$$y_t = \ln x_t + \sigma_\eta\eta_t \quad (37)$$

$$x_t = \alpha + \rho x_{t-1} + \sigma_v\sqrt{x_{t-1}}v_{t-1} \quad (38)$$

where $y_t = \ln RV_t$, $x_t = V_t$, $\sigma_\eta = 0.12$, $\alpha = 0.005$, $\rho = 0.92$ and $\sigma_v = 0.04$. We assume that the state error, v_t , follows 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 is time-varying due to being dependent on the value of the previous state, as reflected in the inequality, $v_t > (-\alpha - \rho x_{t-1})/\sigma_v\sqrt{x_{t-1}}$. As in the linear model, we entertain three different distributions for η_t : 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 3.1.1, and the same supports adopted for the purpose of integration.

We adopt the extended Kalman filter (Anderson and Moore, 1979) as an alternative approach to estimating the forecast distribution. The extended filter deals with the non-linearity in the measurement and state equations (via Taylor series approximations) but assumes that both the measurement and state equation errors are Gaussian.

⁸We have chosen to define the measurement equation 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 (2002).

3.2 Comparison and Evaluation of Predictive Distributions

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 paper to compare the non-parametric and parametric estimates of the predictive distributions of the observed variables. Four proper scoring rules are adopted: logarithmic score (LS), quadratic score (QS), spherical score (SPHS) and the ranked probability score (RPS), given respectively by

$$LS = \ln p(y_{T+1}^o | y_{1:T}) \quad (39)$$

$$QS = 2p(y_{T+1}^o | y_{1:T}) - \int_{-\infty}^{\infty} [p(y_{T+1} | y_{1:T})]^2 dy_{T+1} \quad (40)$$

$$SPHS = p(y_{T+1}^o | y_{1:T}) / \left(\int_{-\infty}^{\infty} [p(y_{T+1} | y_{1:T})]^2 dy_{T+1} \right)^{1/2} \quad (41)$$

$$RPS = - \int_{-\infty}^{\infty} [P(y_{T+1} | y_{1:T}) - I(y_{T+1}^o \leq y_{T+1})]^2 dy_{T+1}, \quad (42)$$

where, in our context, the competing density forecasts, denoted generically by $p(y_{T+1} | y_{1:T})$, are produced by applying the non-parametric and (various) parametric methods to the state space models in Sections 3.1.1 to 3.1.3. As the scoring rule in (42) uses the forecast cumulative density functions rather than density forecasts, the former are analogously denoted by $P(y_{T+1} | y_{1:T})$. The symbol $I(\cdot)$ in (42) denotes the indicator function that takes a value of one if $y_{T+1}^o \leq y_{T+1}$ and zero otherwise, where y_{T+1}^o is *ex-post* the observed value of y_{T+1} .⁹

The *LS* in (39) is a simple ‘local’ scoring rule, returning a high value if y_{T+1}^o is in the high density region of $p(y_{T+1} | y_{1:T})$ 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* and *SPHS* - (40) and (41) respectively - combine a reward for a ‘well-calibrated’ prediction ($p(y_{T+1}^o | y_{1:T})$) with an implicit penalty ($\int_{-\infty}^{\infty} [p(y_{T+1} | y_{1:T})]^2 dy_{T+1}$) for misplaced ‘sharpness’, or certainty, in the prediction. The *RPS* in (42), on the other hand, is sensitive to distance, rewarding the assignment of high predictive mass near to the realized value of y_{T+1} . (See Gneiting and Raftery, 2007, Gneiting, Balabdaoui and Raftery, 2007, and Boero, Smith and Wallis, 2011, for recent expositions).

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

⁹The integrals with respect to the continuous random variable y_{T+1} in (40) to (42) are evaluated numerically.

appealing to a central limit theorem. Denote \overline{SD} as the average difference between the scores of the two competing predictive distributions, associated with a set of M (independently) replicated one-step ahead forecasts. Under the null hypothesis of no difference in the mean scores, the standardized test statistic, $z = \overline{SD}/\widehat{\sigma}_{SD}/\sqrt{M}$, has a limiting $N(0, 1)$ distribution, where $\widehat{\sigma}_{SD}/\sqrt{M}$ is the estimated standard deviation of \overline{SD} .

In contrast with the process of comparison, the evaluation of forecasts involves assessing the performance of a forecasting approach against an *absolute* standard. For example, the probability integral transform (PIT) method 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. Specifically, under the null hypothesis that the predictive distribution corresponds to the *true* data generating process, the PIT, defined as the cumulative predictive distribution evaluated at y_{T+1}^o ,

$$u_{T+1} = \int_{-\infty}^{y_{T+1}^o} p(y_{T+1}|y_{1:T}) dy_{T+1}, \quad (43)$$

is uniform $(0, 1)$ (Rosenblatt, 1952). Hence, the evaluation of $p(\cdot)$ is performed by assessing whether or not the probability integral transform over M replications, $\{u_{T+1}^i, \text{ for } i = 1, 2, \dots, M\}$, is $U(0, 1)$. Under $H_0 : u_{T+1} \sim i.i.d.U(0, 1)$, the joint distribution of the relative frequencies of the u_{T+1}^i is multinomial, and Pearson's goodness of fit statistic can be used to assess whether the empirical distribution (of the u_{T+1}^i) conforms with this theoretical distribution. 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} ,

$$\omega_{T+1} = \Phi^{-1}(u_{T+1}), \quad (44)$$

where $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal distribution function. A likelihood ratio (LR) test of $H_0 : \omega_{T+1} \sim i.i.d.N(0, 1)$, against the alternative that the $\{\omega_{T+1}^i, \text{ for } i = 1, 2, \dots, M\}$ have an autoregressive structure of order one, with Gaussian errors, is conducted. To supplement the LR results, the Jarque-Bera normality test is applied. All three tests have χ^2 null distributions, with 19, 3 and 2 degrees of freedom respectively. The degrees of freedom for the Pearson goodness of fit test corresponds to one less than the chosen number of bins (20) used in the construction of the test statistic.

The PIT-based tests are supplemented here by empirical coverage rates, calculated as the proportion of instances (over M replications) in which the realized value falls within the 95% highest predictive density (HPD) interval. If the (estimated) predictive distribution has a coverage rate higher (lower) than the nominal rate, it means that the distribution is too dispersed

(concentrated) relative to the true predictive distribution. We also calculate the proportion of samples with realizations that fall in the lower and upper 5% predictive tails. If the predictive has a tail coverage rate that is higher (lower) than the nominal rate, it means that extreme values are being over (under) predicted.

3.3 Simulation Results

All DGPs in the three broad models being investigated (as detailed in Sections (3.1.1), (3.1.2) and (3.1.3) respectively) were simulated over $M = 1000$ replications, with $T = 1000$. For both the linear and realized volatility (RV) models, $N = 11$ grid points were used in the support of the measurement error density, whilst $N = 21$ was used for the SCD model. All grid points are evenly spaced.¹⁰ The parameter values (other than the density ordinates defining the measurement error in the non-parametric case) are fixed in all simulation exercises and take on values recorded in the text. Table 1 records the distributional parameter values (if applicable) for the measurement error in each DGP, and the values of λ , c and ω in (24) 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 were chosen with reference to typical empirical data relevant to the model at hand.¹¹

Tables 2 to 4 record respectively all score, evaluation and coverage results. Results for the linear model, (26) and (27), the SCD model, (30) and (29) and the RV model, (37) and (38), are recorded in Panel A, B and C respectively of each table. With reference to Panel A in Table 2, the scores of the non-parametric estimate of $p(y_{T+1}|y_{1:T})$, 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. 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 $p(y_{T+1}|y_{1:T})$ are higher overall than for the parametric 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

¹⁰Some experimentation with different values of N was conducted, with results being reasonably 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 values of N that were not *too* large.

¹¹All numerical results in this and the following empirical section have been produced using the GAUSS programming language.

significantly out-perform the misspecified parametric estimates, for all four scoring measures.

Table 1

Constants, λ , c and ω , used in the penalized likelihood function in (24), in the simulation experiments for the linear, SCD and RV models, as detailed in Sections (3.1.1), (3.1.2) and (3.1.3) respectively.

	η_t	λ	c	ω
Linear Model	$N(0, 1)$	0.5	0.5	0.2
	<i>Student t</i> (0, 1, $\nu = 3$)	4.0	0.5	0.2
	<i>Skewed Student t</i> (0, 1, $\nu = 3, \gamma = 3$)	6.0	0.05	0.2
SCD Model	<i>Exponential</i> (1, 1)	1.0	1.0	0.4
	<i>Weibull</i> ($\gamma = 1.15, 1$)	1.0	1.0	0.4
	<i>Gamma</i> ($\zeta = 1.23, 1$)	1.0	1.0	0.4
RV model	$N(0, 1)$	1.0	0.5	0.2
	<i>Student t</i> (0, 1, $\nu = 3$)	8.0	0.05	0.4
	<i>Skewed Student t</i> (0, 1, $\nu = 3, \gamma = 3$)	4.0	0.5	0.2

Panel A of Table 3 records (for the linear model) the test statistics associated with the three PIT tests described in Section 3.2, namely, the Pearson test for the uniformity of $\{u_{T+1}^i, i = 1, 2, \dots, M\}$ in (43), the LR test of the normality (and independence) of $\{\omega_{T+1}^i, i = 1, 2, \dots, M\}$ in (44) and the Jarque-Bera test for the normality of $\{\omega_{T+1}^i, i = 1, 2, \dots, M\}$. 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 $p(y_{T+1}|y_{1:T})$ 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.

Table 2

Prediction comparison. Average scores for the non-parametric and parametric estimates of $p(y_{T+1}|y_{1:T})$ (Panels A and C) and $p(\ln y_{T+1}|\ln y_{1:T})$ (Panel B), for the respective DGPs, with z values for the difference in scores across the competing forecasts reported. In the table, ** represents statistical significance at the 5% level.

PANEL A: Estimated $p(y_{T+1} y_{1:T})$ for the linear model (Section 3.1.1)												
η_i :	Logarithmic Score			Quadratic Score			Spherical Score			Continuous Ranked Probability		
	<i>N</i>	<i>St</i>	<i>SkSt</i>	<i>N</i>	<i>St</i>	<i>SkSt</i>	<i>N</i>	<i>St</i>	<i>SkSt</i>	<i>N</i>	<i>St</i>	<i>SkSt</i>
Kalman filter	-1.9487	-1.9872	-2.0464	0.1665	0.1684	0.1615	0.4081	0.4104	0.4019	-0.9576	-0.9774	-1.0269
Non-parametric filter	-1.9512	-1.9695	-2.001	0.1662	0.1693	0.1652	0.4078	0.4113	0.4065	-0.9584	-0.9728	-1.0032
z-statistic	-1.2825	2.5027**	3.8688**	-0.7064	0.8918	2.4836**	-0.5760	0.7909	2.6586**	-0.5254	1.1732	3.4734**

PANEL B: Estimated $p(\ln y_{T+1} \ln y_{1:T})$ for the SCD model (Section 3.1.2)												
η_i :	Logarithmic Score			Quadratic Score			Spherical Score			Continuous Ranked Probability		
	<i>Exp</i>	<i>Wb</i>	<i>Gamma</i>	<i>Exp</i>	<i>Wb</i>	<i>Gamma</i>	<i>Exp</i>	<i>Wb</i>	<i>Gamma</i>	<i>Exp</i>	<i>Wb</i>	<i>Gamma</i>
Kalman filter	-1.7414	-1.6280	-1.6463	0.2086	0.2398	0.2303	0.4567	0.4898	0.4799	-0.7729	-0.6829	-0.7004
Non-parametric filter	-1.7114	-1.5958	-1.6115	0.2135	0.2470	0.2353	0.4621	0.4970	0.4851	-0.7643	-0.6712	-0.6914
z-statistic	3.2606**	3.1794**	3.6051**	2.1441**	2.9672**	2.2638**	2.2215**	2.9651**	2.3718**	2.5278**	2.9249**	3.3838**

PANEL C: Estimated $p(y_{T+1} y_{1:T})$ for the RV Model (Section 3.1.3)												
η_i :	Logarithmic Score			Quadratic Score			Spherical Score			Continuous Ranked Probability		
	<i>N</i>	<i>St</i>	<i>SkSt</i>	<i>N</i>	<i>St</i>	<i>SkSt</i>	<i>N</i>	<i>St</i>	<i>SkSt</i>	<i>N</i>	<i>St</i>	<i>SkSt</i>
Kalman filter	0.01035	0.1282	0.08348	1.2160	1.3567	1.3394	1.1013	1.1629	1.558	-0.13462	-0.1204	-0.1243
Non-parametric filter	0.02564	0.1422	0.1026	1.2232	1.3783	1.3729	1.1047	1.1712	1.1691	-0.13447	-0.1196	-0.1232
z-statistic	2.2647**	2.4188**	2.5861**	1.3683	2.0192**	2.7316**	1.4966	1.9258**	2.7818**	0.4573	1.4604	1.7499**

With reference to Panel A of Table 4, 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’) under coverage of the 95% interval.

Table 3

Prediction Evaluation. Pearson, LR and Jarque-Bera χ^2 test statistics, for the non-parametric and parametric estimates of $p(y_{T+1}|y_{1:T})$ (Panels A and C) and $p(\ln y_{T+1}|\ln y_{1:T})$ (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 $p(y_{T+1} y_{1:T})$ for Linear model						
	NP	KF	NP	KF	NP	KF
$\eta_t \sim N(0, 1)$	13.12	11.88	0.618	0.414	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 $p(\ln y_{T+1} \ln y_{1:T})$ for SCD model						
	NP	KF	NP	KF	NP	KF
$\eta_t \sim \exp(1, 1)$	20.68	44.68**	1.188	0.581	3.077	64.983**
$\eta_t \sim Wb(\gamma = 1.15, 1)$	9.96	48.64**	1.879	0.635	4.409	129.785**
$\eta_t \sim Gamma(\zeta = 1.23, 1)$	10.16	31.60**	3.933	2.554	1.131	77.524**
PANEL C: Estimated $p(y_{T+1} y_{1:T})$ for RV Model						
	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**

Considering now the score results for the SCD model, recorded in Panel B of Table 2, all four scores for the non-parametric estimate of $p(\ln y_{T+1}|\ln y_{1:T})$ 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 3, across all DGPs the non-parametric estimates of $p(\ln y_{T+1}|\ln y_{1:T})$ are

Table 4

Prediction Evaluation. Coverage rates (5% and 95%) for the non-parametric and parametric estimates of $p(y_{T+1}|y_{1:T})$ (Panels A and C) and $p(\ln y_{T+1}|\ln y_{1:T})$ (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 $p(y_{T+1} y_{1:T})$ for the linear model (Section 3.1.1)									
η_t :	<i>N</i>	<i>St</i>	<i>SkSt</i>	<i>N</i>	<i>St</i>	<i>SkSt</i>	<i>N</i>	<i>St</i>	<i>SkSt</i>
Kalman filter	4.8	4.5	5.5	5.0	5.3	6.4	94.9	93.3**	92.4**
Non-parametric filter	4.4	4.6	6.1	4.5	5.9	5.8	95.2	94.1	93.5
PANEL B: Estimated $p(\ln y_{T+1} \ln y_{1:T})$ for the SCD model (Section 3.1.2)									
η_t :	<i>Exp</i>	<i>Wb</i>	<i>Gamma</i>	<i>Exp</i>	<i>Wb</i>	<i>Gamma</i>	<i>Exp</i>	<i>Wb</i>	<i>Gamma</i>
Kalman filter	6.0	5.8	6.5	2.7**	2.8**	3.3**	94.9	94.9	95.4
Non-parametric filter	5.2	4.7	5.1	6.0	6.3	5.9	94.2	94.3	94.7
PANEL C: Estimated $p(y_{T+1} y_{1:T})$ for the RV Model (Section 3.1.3)									
η_t :	<i>N</i>	<i>St</i>	<i>SkSt</i>	<i>N</i>	<i>St</i>	<i>SkSt</i>	<i>N</i>	<i>St</i>	<i>Skst</i>
Kalman filter	6.1	5.6	6.0	5.2	3.0**	3.3**	93.4	95.6	94.4
Non-parametric filter	5.3	4.7	5.8	5.5	4.3	4.0	93.8	95.7	95.0

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 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 (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.

Finally, all scores (reported in Panel C of Table 2) for the non-parametric estimate of $p(y_{T+1}|y_{1:T})$ in the RV model are higher than those of the parametric estimate, under all DGPs. Despite the positive values of the relevant test statistics, 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 (correct) assumption of conditional Gaussianity. However, under the Student- t DGP, the non-parametric estimate is significantly more accurate than the (misspecified) parametric estimate, according to three of the four scores, and in all four cases under the skewed Student- t distribution.

The results in Panel C of Table 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 hypotheses 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 Panel C of Table 4, both forecast approaches have similar (and reasonable) coverage rates, apart from a significant undercoverage in the upper tail on the part of the misspecified parametric approach, under both the symmetric and (positively) skewed Student- t DGPs.

4 Empirical Illustration

4.1 Preliminary Analysis

In order to illustrate the non-parametric method, we produce and evaluate non-parametric estimates of the one-step ahead prediction distributions for realized volatility on the S&P500 market index, fitting the model described in (37) and (38). 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).¹²

The time series of the data is plotted in Panel A of Figure 1. 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, whereby 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 realized volatility caused by a sharp

¹²The authors would like to acknowledge the excellent research assistance of Chris Tse in producing the 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.

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 1 plots the histogram of log realized volatility, with the distinct skewness to the right reflecting the occurrence of the very extreme values of realized volatility itself.¹³ 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 log realized volatility the aim is to capture the impact of the jump variation in a computational simple way, rather than modelling price jumps explicitly.

4.2 Empirical results

We divide the S&P500 daily realized volatility data into two subsamples. The first subsample (2 January 1998 to 30 January 2007), containing 2245 observations, is reserved for estimation of the model parameters in (37) and (38), including the unknown ordinates of $p(\eta)$. The subsample used for forecast assessment comprises the remaining 400 realized volatility values, covering the period from 31 January 2007 to 29 August 2008, and as represented by the shaded area in Panel A of Figure 1. This subsample period 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 with each new daily observation. $N = 21$ grid points, equally spaced over the interval from -10 to 10, are used to represent the support of the measurement error density, $p(\eta)$. Values of the penalty parameters used in (24) are $\lambda = 4$, $c = 0.5$ and $\omega = 0.3$.¹⁴

We estimate the 400 one-step ahead predictive distributions for the level of realized volatility for the out-of-sample period. For each of the 400 prediction distributions, simulated draws

¹³A Jarque-Bera test applied to this log realized volatility series rejects the null hypothesis of Gaussianity at any conventional level of significance.

¹⁴Robustness of the results to different values of N (in a range from 21 to 51) for a fixed set of penalty values, and robustness of the results to different sets of penalty values ($1 \leq \lambda \leq 4$; $0.01 \leq c \leq 0.5$; $0.3 \leq \omega \leq 0.8$) was investigated. Differences in the estimated forecast distributions were negligible.

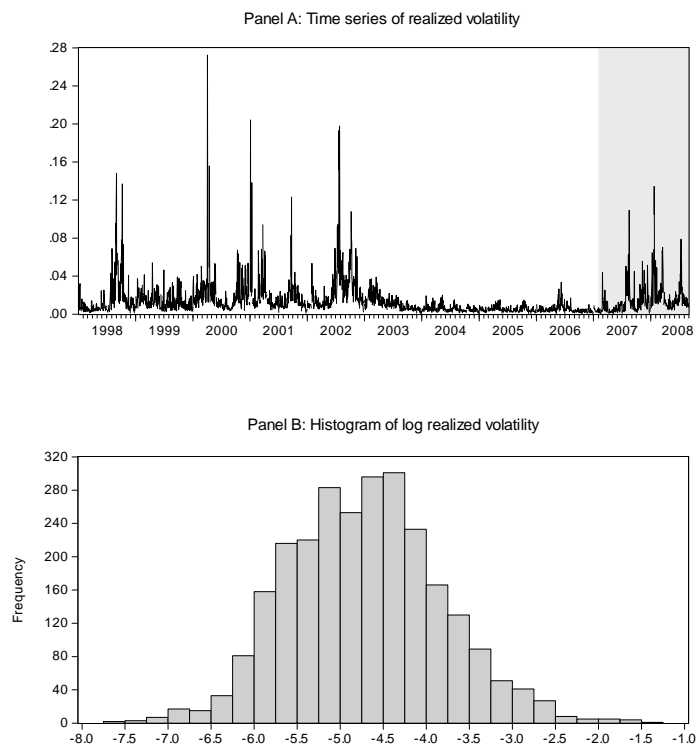


Figure 1: Time series of realized volatility and histogram of log realized volatility of S&P500 market index from 2 January 1998 to 29 August 2008.

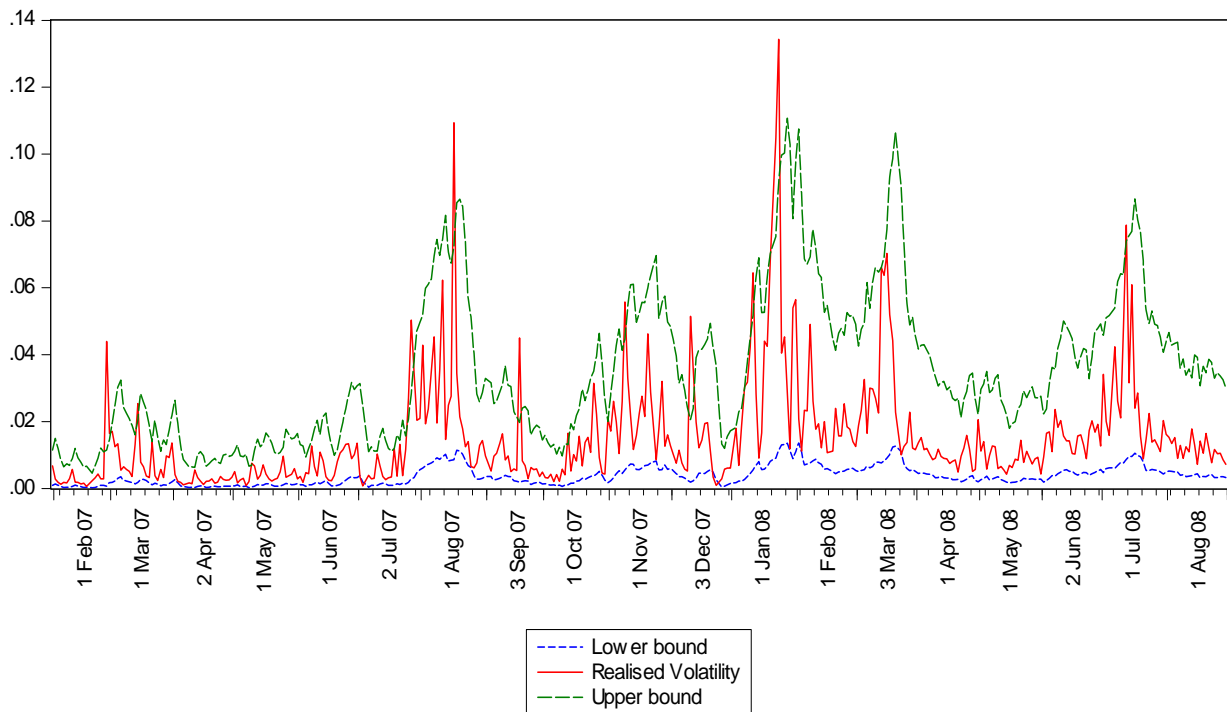


Figure 2: 95% one-step-ahead prediction intervals and the observed realized volatility, over the 400 day evaluation period (31 January 2007 to 29 August 2008)

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 2. 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 in (43). However, both the LR test of the normality (and independence) of $\{\omega_{T+1}^i, i = 1, 2, \dots, M\}$ in (44) and the Jarque-Bera test for the normality of $\{\omega_{T+1}^i, i = 1, 2, \dots, M\}$ in (44) 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 in Figure 3, indicates that too many realizations of volatility fall in the right tail of the forecast density, relative to the estimate thereof. This suggests that, despite the overall predictive accuracy evidenced, we are still unable to capture the *most* extreme values of volatility that occur on a few occasions during the evaluation period.

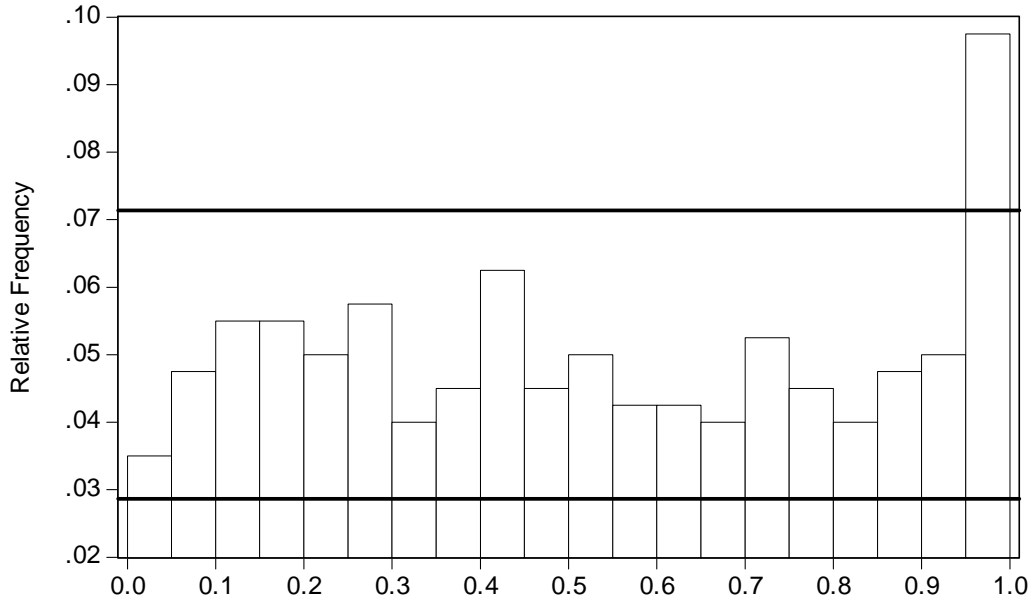


Figure 3: Histogram of the probability integral transform series, u , for the realized volatility model. The horizontal lines superimposed on the histogram are approximate 95% confidence intervals for the individual bin heights under the null that u is *i.i.d.* $U(0, 1)$.

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 quite high volatility day during the out-of-sample period. The technique mimicks 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.¹⁵ The steps of the procedure are as follows:

1. Obtain $T-b+1$ subsamples $Y_1 = (y_1, \dots, y_b)$, $Y_2 = (y_2, \dots, y_{b+1})$, ..., $Y_{T-b+1} = (y_{T-b+1}, \dots, y_T)$ from the set of empirical data, $y_{1:T} = (y_1, y_2, \dots, y_T)'$.
2. Use the proposed non-parametric ML method to produce an estimate of θ , $\hat{\theta}_{b,t}$, computed

¹⁵In related work, 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 paper. See also Pascual, Romo and Ruiz (2001, 2005).

from Y_t , for $t = 1, 2, \dots, T - b + 1$.

3. Use $\hat{\boldsymbol{\theta}}_{b,t}$ and the *observed* values, $y_{1:T}$, to compute the 1-step ahead forecast distribution $p\left(y_{T+1}|y_{1:T}, \hat{\boldsymbol{\theta}}_{b,t}\right)$.
4. Calculate (over an arbitrarily fine grid of values for y_{T+1}) the metric $d_{b,t} = \sqrt{T} \left\| p\left(y_{T+1}|y_{1:T}, \hat{\boldsymbol{\theta}}_{b,t}\right) - p\left(y_{T+1}|y_{1:T}, \hat{\boldsymbol{\theta}}\right) \right\|_1$, where $p\left(y_{T+1}|y_{1:T}, \hat{\boldsymbol{\theta}}\right)$ is the estimated forecast distribution based on the empirical data and $\hat{\boldsymbol{\theta}}$ is the empirical estimate of $\boldsymbol{\theta}$.
5. Find the 95th percentile of $\{d_{b,1}, \dots, d_{b,T-b+1}\}$, $d_b^{0.95}$, and the corresponding distribution $p_{0.95}(y_{T+1}|y_{1:T}, \cdot)$. Then, relative to the replicated distributions and in terms of the $\|\cdot\|_1$ distance from $p\left(y_{T+1}|y_{1:T}, \hat{\boldsymbol{\theta}}\right)$, the chances of seeing a distribution as or more ‘extreme’ than $p_{0.95}(y_{T+1}|y_{1:T}, \cdot)$ 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 $b \in \{b_{small}, \dots, b_{big}\}$ carry out Steps 1 to 5 above to compute $d_b^{0.95}$.
- b. For each b compute VI_b as the standard deviation of the $2k + 1$ consecutive values $\{d_{b-k}^{0.95}, \dots, d_{b+k}^{0.95}\}$ (for $k = 2$).
- c. Choose \hat{b} to minimise VI_b .¹⁶

Figure 4 shows the 10th, 50th and 95th percentile sub-sampled forecasts, along with the estimated empirical forecast, for the 17th March 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

¹⁶We have chosen to use $d_b^{0.95}$ as the percentile on which selection of b is based as we are interested, primarily, in ascertaining the changes in the forecast distributions that may occur at the extreme end of the scale (of the metric d).

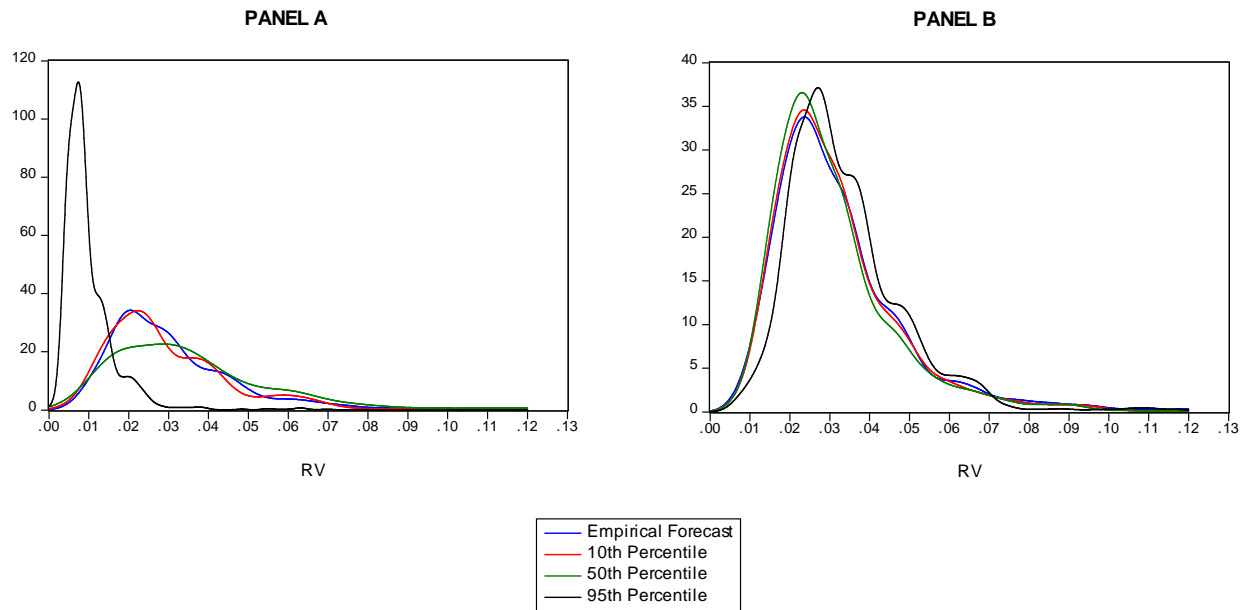


Figure 4: Plot of the 10th, 50th and 95th percentile bootstrap forecast against the empirical forecast for the 17th March 2008. Panel A shows the bootstrap forecasts estimated from the preceding 505 observations. Panel B shows the bootstrap forecasts estimated from the preceding 2528 observations.

distribution becomes much more concentrated around the empirical estimate, with the full suites of distributions leading to qualitatively similar conclusions regarding volatility on the given day.

5 Concluding Remarks

We have developed a new method for estimating the full forecast distribution of non-Gaussian time series variables in the context of a general non-Gaussian, non-linear state space model. A non-parametric filter is derived that exploits the functional relationship between the observed variable and the state and measurement 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 measurement 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 extension to higher dimensional problems also feasible, subject to the usual proviso that accompanies a grid-based method.

Using the proposed method, the predictive distributions for the observed and latent variables are produced for a range of linear and non-linear models, in a simulation setting. The non-parametric predictive distributions are compared against distributions produced via (misspecified) parametric approaches. Results show that the non-parametric method performs significantly better, overall, than (misspecified) parametric alternatives and is competitive with correctly specified parametric estimates. The new method is also applied to empirical data on the S&P500 index, with the non-parametric predictive distribution able to capture important distributional information about the future value of the realized volatility of the index. A subsampling method is used to highlight the effect that sampling variation can have on predictive conclusions, in small samples in particular.

We conclude by noting that despite our focus here on the non-parametric setting, our proposed algorithm is also directly applicable to models in which the measurement error distribution is specified parametrically. In that particular case, as long as the measurement error distribution is able to be simulated from and an appropriate transformation between each measurement and its error term is available, then the grid-based method may be replaced by an approach in which all relevant integrals are evaluated by Monte Carlo simulation, based on draws from the invariant distribution of the measurement error. The resulting alternative particle filter, unnecessary in scalar (or low) dimensional cases such as those explored in this paper, would be a powerful tool in high-dimensional settings, particularly as it avoids the degeneracy problems that are a feature of existing simulation-based filtering algorithms. This is currently the subject of investigation by the authors.

References

- [1] Ait-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] Amisano, G. and Giacomini, R. 2007. Comparing Density Forecasts via Weighted Likelihood Ratio Tests. *Journal of Business and Economic Statistics* 25, 177-190.
- [3] Anderson, B.D.O. and Moore, J.B. 1979. *Optimal Filtering*. Prentice.
- [4] Andersen, T.G., Bollerslev, T., Diebold, F.X. and Ebens, H. 2001. The Distribution of Realized Stock Return Volatility. *Journal of Financial Econometrics* 61, 43-76.
- [5] Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. 2003. Modeling and Forecasting Realized Volatility. *Econometrica* 71, 579-625.
- [6] 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 (2), 174-188.
- [7] Au, C. and Tam, J. 1999. Transforming Variables Using the Dirac Generalized Function. *The American Statistician* 53, 270-272.
- [8] Barndorff-Nielsen, O.E. and Shephard, N. 2002. 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.
- [9] Bates, D. 2000. Post-87 Crash Fears in the S&P 500 Futures Option Market. *Journal of Econometrics* 94, 181-238.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] Berkowitz, J. 2001. Testing Density Forecasts with Applications to Risk Management, *Journal of Business and Economic Statistics* 19, 465-474.

- [14] Boero, G., Smith, J. and Wallis, K.F. 2011. Scoring Rules and Survey Density Forecasts. *International Journal of Forecasting* 27, 379-393.
- [15] Broadie, M., Chernov, M. and Johannes, M. 2007. Model Specification and Risk Premia: Evidence from Futures Options. *The Journal of Finance* LXII: 1453 - 1490.
- [16] Brownless, 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
- [17] 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.
- [18] 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.
- [19] 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*.
- [20] Czado, C., Gneiting, T. and Held, L. 2009. Predictive Model Assessment for Count Data. In press, *Biometrics*.
- [21] 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.
- [22] De Rossi, G. and Harvey, A. 2009. Quantiles, expectiles and splines. *Journal of Econometrics* 152, 179-85.
- [23] 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.
- [24] Durham, G.B. 2007. Monte Carlo Methods for Estimating, Smoothing and Filtering One- and Two-Factor Stochastic Volatility Models. *Journal of Econometrics* 133, 273-305.
- [25] Durham, G.B. 2007. SV Mixture Models with Application to S&P 500 Index Returns. *Journal of Financial Econometrics* 85, 822-856.

- [26] Engle, R.F. and Gonzalez-Rivera, G. 1991. Semiparametric ARCH Models. *Journal of Business and Economic Statistics* 9, 345-359.
- [27] Eraker, B., Johannes, M.S. and Polson, N.G. 2003. The Impact of Jumps in Returns and Volatility, *Journal of Finance* 53, 1269-1300.
- [28] 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
- [29] Freeland, R and McCabe, B. 2004. Forecasting Discrete Valued Low Count Time Series. *International Journal of Forecasting* 20, 427-434.
- [30] 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.
- [31] Giacomini, R. and Komunjer, I. 2005. Evaluation and Combination of Conditional Quantile Forecasts. *JBES* 23, 416-431.
- [32] Gneiting, T. 2008. Editorial: Probabilistic Forecasting, *Journal of the Royal Statistical Society (A)* 171, 319-321.
- [33] Gneiting, T. and Raftery, A.E. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation, *Journal of the American Statistical Association* 102, 359-378.
- [34] Gneiting, T., Balabdaoui, F. and Raftery, A.E. 2007. Probabilistic Forecasts, Calibration and Sharpness. *Journal of Royal Statistical Society* 69, 243-268.
- [35] Hassani, S. 2009. *Mathematical Methods for Students of Physics and Related Fields* (2nd Edition).
- [36] Jensen, M.J. and Maheu, J.M. 2010. Bayesian Semiparametric Stochastic Volatility Modeling. *Journal of Econometrics* 157, 306-316.
- [37] Johnson, N.L., Kotz, S. and Balakrishnan, N. 1994. *Distributions in Statistics: Continuous Univariate Distributions*, Vol. 2. Wiley, New York
- [38] Kitagawa, G. 1987. Non-Gaussian State Space Modeling of Nonstationary Time Series. *Journal of the American Statistical Association* 76, 1032-1064.
- [39] 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.

- [40] Khuri, A.I. 2004. Application of Dirac's Delta Function in Statistics. *International Journal of Mathematical Education in Science and Technology* 35, 185-195.
- [41] 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.
- [42] McCabe, B and Martin, G. 2005. Bayesian Predictions of Low Count Time Series. *International Journal of Forecasting* 21, 315-330.
- [43] McCabe, B., Martin, G.M. and Harris, D. 2011. Efficient Probabilistic Forecasts for Counts. Forthcoming, *Journal of the Royal Statistical Society, Series B* 73, 253-272.
- [44] 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.
- [45] 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>.
- [46] 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.
- [47] Pascual, L., Romo, J. and Ruiz, E. 2005. Bootstrap prediction intervals for power-transformed time series. *International Journal of Forecasting* 21, 219-235.
- [48] Politis, D.N., Romano, J.P. and Wolf, M. 1999. *Subsampling*, Springer, New York.
- [49] Rodriguez, A. and Ruiz, E. 2009. Bootstrap Prediction Intervals in State-Space Models. *Journal of Time Series Analysis* 30, 167-178.
- [50] Rosenblatt, R.F. 1952. Remarks on a Multivariate Transformation. *Annals of Mathematical Statistics* 23, 470-472.
- [51] 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.
- [52] Sorenson H.W. and Alspach D.L. 1971. Recursive Bayesian Estimation Using Gaussian Sums. *Automatica* 7, 465-479

- [53] 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.
- [54] Tay, A and Wallis, K. 2000. Density forecasting: A Survey. *Journal of Forecasting* 19, 235-254.
- [55] 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.