High-frequency volatility in a time deformed framework, the role of volume, durations and jumps through intraday data

Antonio A. F. Santos

Faculty of Economics, Centre for Business and Economic Research (CeBER),

Monetary and Financial Research Group (GEMF), University of Coimbra, Portugal

Abstract

The analysis of volatility is paramount for decision-making in financial markets. Here is developed a strategy for increasing information to be used in the build of relevant volatility measures. With the increased information available through the new technologies most of the data analyses in finance are "big data" problems. The time deformed returns associated with the volume are used to estimate and forecast intraday high-frequency volatility. This kind of return can give a new perspective on volatility evolution, because allow estimation, forecasts and decisions to be considered at a varying speed when measured in calendar time, which is compatible with the reality in financial markets, periods of high and low activity are clearly identifiable. Through this strategy, other information elements can be extracted from the data, and not only the traditional fixed time-interval prices usually obtained. By extracting other information related to volatility evolution, for example, volume of trade and durations, better volatility estimates can be constructed. It allows the generalisation of models, namely the ones that include the possibility of jumps in the volatility, which without such increase of sample information present characteristics of underidentification, leading to inconsistent estimates and creating biased forecasts.

1 Introduction

The volatility in financial markets is an important element for countries, institutions (e.g. banks), and also for individuals. It is fully recognised by all, it is accepted that influences decisions, can decrease or increase the wealth, and the well being in general. However, it cannot be uniquely and unambiguously measured. The analysis of volatility has a long history in the academic world, but also in most financial institutions that influence or are influenced by the evolution of financial markets.

By now there is not a unique measure of volatility, but the main tendency since the 90's is to define "objective" measures of volatility obtained through data, which nowadays is a "big data" problem. It is the task of statistical and econometric models to extract the relevant information from such data. A substantial number of models and procedures have been proposed as able to define meaningful measures of volatility.

Most financial econometric models use mainly prices or returns, and almost all aim essentially to define a relevant volatility measure. The best known models for modelling volatility are the ones from the (G)ARCH family (Engle, 1982; Bollerslev, 1986) and the Stochastic Volatility (SV) (Taylor, 1986, 1994). Recently, some developments have emerged using model-free volatility measures and the most referenced is the Realised Volatility (RV) (Andersen et al., 1999, 2001, 2003, 2005; Andersen and Teräsvirta, 2009; Andersen et al., 2011; Barndorff-Nielsen and Shephard, 2002, 2004), and for an excellent survey see McAleer and Medeiros (2008).

Here we designated daily observations as high-frequency data, but with the development of information technologies, with trade made essentially through and by computers, and the dissemination of information available through those technologies, institutions and individuals can easily obtain ultra-high-frequency intraday data, with a frequency that can reach the millisecond.

The aim is to explore the availability of ultra-high-frequency data, and develop measures of volatility, and also relevant forecasts for intraday volatility. It is well established that daily volatility varies, and there is further evidence that intraday volatility as well (Andersen and Bollerslev, 1997, 1998). One of the more flexible models used to characterise the volatility evolution is the SV model.

The use of intraday returns was considered by Stroud and Johannes (2014) aiming to model intraday volatility through SV models. As new characteristics are found in intraday data, Stroud and Johannes (2014), following Eraker et al. (2003), Todorov (2011), and Todorov and Tauchen (2011), considered the possibility of jumps in the observation process associated with returns as well in the latent process associated with the volatility. Another important characteristic to be taken into account is the seasonal component that is found in intraday volatility (Andersen and Bollerslev, 1997, 1998). These were two components that were added to the model, even if jumps were already considered with daily returns (Chib et al., 2002), they were not associated with the latent process that drives the volatility. Another characteristic is the possibility of considering two states, which intends to characterise different speeds of volatility evolution.

Our approach has many resemblances with Stroud and Johannes (2014), in the sense that aims to characterise and model the intraday volatility, and considers the same type of models to perform such analysis. We address the problem by using a new type of return, in volume-domain instead of in time-domain. This is a type of return that is not commonly used in the literature, and hopefully it will incorporate a new perspective on volatility evolution. The use of such return overcomes some of the difficulties related to the seasonal component presented in intraday returns, and serves as a vehicle to incorporate information contained in the volume.

The assumption that volume of transactions may have an important role in the definition of prices or returns is not new (Clark, 1973; Epps, 1976; Epps and Epps, 1976; Tauchen and Pitts, 1983; Tauchen et al., 1996; Andersen, 1996). The dependence between trading volume and financial returns, mainly designed to characterise the high kurtosis of returns distributions was first investigated by Clark (1973), Epps (1976) and Epps and Epps (1976). Trading volume was used as a mixing variable, allowing unconditional distributions to present a fat tails characteristic. Clark (1973) and Lamoureux and Lastrapes (1990) used the trading volume as an exogenous variable. When returns and volume-trade are jointly determined, a different approach is needed. Tauchen and Pitts (1983) and Andersen (1996), both assume that the two processes, returns and volume, depend on a latent process which characterises an unobserved flow of information.

Even that theoretically Clark (1973) and Tauchen and Pitts (1983) established a link between volume and returns, where the volume drives return distributions, the inclusion of this component in a models like GARCH and SV has been difficult to justify. Attempts have been made as in Lamoureux and Lastrapes (1990), Watanabe (2000), Liesenfeld and Richard (2003), but essentially the volume process overpowers the other components. An essential characteristic, the volatility persistence on returns, is overshadowed by the mean dynamics associated with volume. This is perhaps one of the reasons that it has received a small amount of attention in the analysis of volatility evolution, and following Granger (2002), "Given its ready availability, until recently volume has been the missing variable in finance ... ".

Instead of considering directly the volume as an explanatory variable as did Lamoureux and Lastrapes (1990) and Watanabe (2000), we consider the influence of volume in an indirect way. This is possible by using intraday data. This kind of data is influenced by different characteristics associated with trading mechanisms in financial markets, and such kind of data can incorporate substantial amount of noise. With intraday data, it is not usual to use tick-by-tick data, but instead less frequent data as s-min returns, with $s = 1, 5, 10, 30$. Our main reference paper, Stroud and Johannes (2014), uses 5-min returns. Prices and returns are considered in time-domain, and are calculated using the nearest neighbour s-min price or return tag. For a fixed time period, the number of shares traded is random, and also the amount of information arriving to markets. As it does happen evenly, can be considered that the number of trades depend on relevant information that has arrived. More information arrive in the beginning of trading hours and near the closing, it is usual to find a seasonal pattern, where the volatility is higher at opening and closing of the markets.

Despite the volume has not received much attention within the finance literature, there are some research that present some results common with the ones addressed. An idea of a different referential from the one represented by the calendar time, and the consideration of information measures in different scales was considered by Maillet and Michel (1997) and Le Fol and Ludovic (1998), where a kind of time deformation is considered. This could be useful to analyse some stylised facts associated with intraday returns, and in this way allowing a better characterisation of such returns. Further research has been highlighting the role played by the volume (Gourieroux and Le Fol, 1997; Darolles et al., 2000, 2015, 2017), but unfortunately we think has not been yet incorporated in mainstream financial literature. If the former propose to deal with the characterisation of unconditional distribution for intraday returns, the latter, especially Darolles et al. (2017), deals with the dynamic features through conditional distributions. Our difference is that volume serves to define a new referential (time deformation), and we address the estimation and forecasting issues without discard the nonlinear features, presenting workable algorithms to estimate models' and forecast volatility evolution.

The seasonal pattern is perceived when the time period is fixed and the amount of information in each time interval is random. We can try to maintain fixed the amount of information, and as it is not directly observable, volume is used as a proxy. Instead of using a time-period, the volume is fixed (amount of information), and returns are calculated in this new referential, which will lead to the definition of volume-adapted return.

2 Deformed time observations

A natural referential to assume for a sequence of observations for a given variable is the calendar time, variables can be observed monthly, weekly, and daily. Until recently, for economic variables, daily observations were considered high-frequency data. Economic variables associated with human behaviour have an intrinsic time-tag from which is difficult to dissociate from, we assume that there is an annual budget, the salary is paid monthly, etc. In financial markets, the frequency mainly used is the daily one. Financial markets are opened during the day, and after they close is not possible to rebalance the portfolios, which means that closing prices are the benchmark.

If we zoom to a specific day, during trading time, it is not so obvious the use of calendar time as the main referential. During the day, markets move at different speeds, relevant information arrive in a random manner, and rebalancing the portfolio every 15 minutes in some situations can be too many and in others too few. Time deformed settings associated with evolution of prices in financial markets was already considered in Maillet and Michel (1997), Gourieroux and Le Fol (1997), Le Fol and Ludovic (1998), and called time deformation in a way that defines variables in a varying shrinking or stretching of the calendar time as a form to reveal certain characteristics that are manifest themselves randomly in time.

Following Gourieroux and Jasiak (2001), we consider two time scales, the calendar time t, assuming continuous non-negative values, and a discrete counterpart z, assuming non-negative values. The time changing process determines the mapping from calendar time to another referential,

$$Z : t \in \mathcal{R}^{+} \to Z_t \in \mathcal{N}$$

A variety of time scales can be considered, but in the case of financial markets, we use the general denomination of volume. When important information arrive to the market we can verify an increase in the activity, more people are selling and buying stocks, there is an increase in the volume, represented by the number of transactions, N, the total number of shares traded, V, or even the value associated with the total of shares traded, C. In this cases, the driving processes can be, $N(t)$, number of transactions before t, $V(t)$ total number of shares traded before t, or, $C(t)$, total market value traded before t.

3 Volume-adapted returns

For the definition of volume-adapted return, we use the notion of a subordinated stochastic process, which in the context of pricing financial assets as been considered by Clark (1973). The main idea is that prices evolve at different rates during identical intervals of time. To any transaction in financial markets is associated a given volume. When more information arrive to the markets, it is expected that there is an increase in volume, and prices adjust more rapidly, which can induce significant changes in volatility.

Consider a discrete stochastic process associated with the log-price of a given stock at t, $p(t)$, for $t = 1, 2, 3, \ldots$. Instead of indexing by $t = 0, 1, 2, \ldots$ representing a deterministic evolution of the time, the process can be indexed by another set of numbers, $t_1, t_2, \ldots$, that are themselves a realisation of a stochastic process, with positive increments, $t_1 \leq t_2 \leq t_3 \ldots$. This process can be used to characterise the observations for prices in an intraday framework. Considering tick-by-tick data, where observed prices are $p(t_i)$, with $0 = t_0 \leq t_1 \leq t_2 \leq \ldots \leq t_n = 1$, intraday returns are defined as $y(t_i) = p(t_i) - p(t_{i-1})$. As the t_i is a random variable from $T(t)$, a positive stochastic process, the observations for prices are unequally spaced in time. Here, $T(t)$ is called the directing process and $p(T(t))$ is the subordinated process.

When tick-by-tick observations are considered for the prices, for these are associated a given number of shares traded, which can be denoted by $V(t_i)$, where $V(t_i)$ is directly connected to $p(t_i)$. Using intraday data, in most econometric models, the time-period is fixed, for example, a 5-min period, and the prices considered are given by $p(t_i)$, where t_i corresponds to the value that is near to a 5-min tag. The novel approach consists in defining the volume-adapted returns through a double subordinated process. The log-prices will be indexed by $V(t_i)$, $p(V(t_i))$.

Considering $0 = t_0 \leq t_1 \leq t_2 \leq \ldots \leq t_n = 1$ in the time-domain, the partition $0 = v_0 \leq v_1 \leq v_2 \leq \ldots \leq v_n = V$ is defined in the volume-domain, where $v_i = V(t_i)$. The volume-adapted price and respective return are defined by $p(v_i)$ and $y(v_i) = \log(p(v_i)/p(v_{i-1}))$, where v_i is the volume index associated with a given target. The price and respective returns are defined in the volume-domain.

Usually the analyses associated with the volatility study use daily data or consider measures of daily volatility (through intraday data). The research on modelling intraday volatility is scarce. In the time-domain a time-period associated with the returns must be defined, and an usual trade-off between precision and measurement error must be considered.

One of the most used tag is the 5-min, as in Stroud and Johannes (2014).

However, different periods were considered from 1- to 30-min returns. A choice must be made, but that is unambiguous in the sense that price is observable for each time tag. The same happens with volume. Instead of time, volume of trade is considered. As we observe the time passing, we get also the increasing of volume, which indexes the volume-adapted prices and respective returns.

Different volumes can be considered to index returns calculations. More volume implies further information, but less observations per day. It makes sense to consider the same time period when different assets are considered, the same does not happen when the volume is considered. The number of shares in each transaction usually depends on the stock prices. To define a possible benchmark value for the volume, we established a link with returns observed in the time domain. We consider only observations in the normal period of transactions, 6.5 hours, which would give 78 observations associated with 5-min periods.

4 Modelling volatility

Giving the usual hypothesis related to evolution of prices in financial markets, namely, the weak efficiency hypothesis that future prices cannot be predictable from past evolution of the same, usually financial econometric models do not seek mean but instead variance dynamic. When daily volatility forecasts are asked for, if only daily data is available, the structure of a model is needed to produce meaningful forecasts. However, if intraday data is available, a daily forecast can be defined by aggregating the different information elements obtained intradaily.

The first utilization of such data to define measures of financial volatility was associated with nonparametric measures like the Realized Volatility (RV), an estimator of the integrated volatility

$$IV = \int_{t}^{t+1} \sigma^2(s)ds$$

defined assuming a standard diffusion process associated with the evolution of log-prices, $p(t)$.

A basic estimator to IV is given by the Realized Volatility. Assume for a given day n intraday returns y_{t_i} are observed with $0 = t_0 < t_1 < \ldots < t_n = n$,

$$RV = \sum_{i=1}^{n} y_{t_i}^2 \tag{1}$$

with a pure diffusion process for log-prices, $dp(t) = \sigma(t)\, W(t)$, and $W(t)$ a standard Winner process, $RV \overset{P}{\to} IV$.

The application of theory for calculation of a daily measure of volatility faces some difficulties due to the violation of a pure diffusion process hypothesis, and because the shrinking of time intervals is not possible without the introduction of additional noise with consequences for the statistical properties of the estimator. Using parametric models has been less frequent, but also encompasses modelling difficulties. The main idea is to use the information contained in intraday data as a way to ameliorate estimation and forecasting of financial volatility.

Recent research addresses the high-frequency volatility evolution, where the aim is to consider that volatility is varying intradaily. A nonparametric approach is not feasible, and a structure induced by a model is needed to obtain interpretable results. As with the daily frequency, ARCH and SV models have been the most used to perform the analyses. One recent reference and a benchmark for the results presented here is Stroud and Johannes (2014).

Apparently the extension of ARCH and SV models used with daily data should not be done to intraday data without considering new features presented in such kind of data. The tendency is to consider several complex extensions as a way of addressing new features. We take as example a model adopted in Stroud and Johannes (2014), a SV model that accounts for the presence of two volatility factors, a fast- and slow-moving, seasonal adjustments, news announcements, and jumps in returns and volatility. These characteristics must be extracted through a model and respective parameters with a unique series of 5-min returns.

A main question arises, is it possible through a unique series to distinguish so many effects? This question is related to the problem of model identification, which has been overlooked in the application of many econometric models. It has different interpretations depending if a classical or Bayesian view of Statistics is adopted, and it is easy to analyse and interpret within a linear regression model, but hard in nonlinear models.

State-space models have assumed a predominant role in the analysis of financial volatility. It is a well established fact that at some frequencies the volatility is time-varying whatever the reasonable measure we might use. The usual interpretation is that the volatility evolution depends on the stream of information that arrive to the market, which is not directly observable, and can be accommodated through a state-space representation.

The state-space models adequate for modelling volatility evolution have to assume nonlinear characteristics, which makes difficult estimation and forecasting, no analytical formula is available as linear models. Many algorithms associated with numerical methods have been developed to estimate and forecast through these models. Some algorithms are related to Markov chain Monte Carlo (MCMC) simulations that approximate the parameters' posterior distribution within a Bayesian estimation framework.

In the SV model, y_t represents the intraday volume-adapted for a given threshold volume (V), where here to alleviate the notation, t represents just the observation index. However, the return is defined trough $y(v_i)$ for a given threshold V. The element α_t represents the state that defines the variance of the process, for $t = 1, \ldots, n$. The model adopted has the usual nonlinear state-space form

$$y_t = \exp(\alpha_t/2)\,\sqrt{\lambda_t}\,\varepsilon_t \qquad (2)$$

$$\alpha_{t+1} = \mu + \phi\,(\alpha_t - \mu) + \sigma_\eta\,\eta_{t+1} \qquad (3)$$

where $(\varepsilon_t, \eta_{t+1}) \sim N(\mathbf{0}, \Sigma)$, with $\Sigma_{11} = 1$, $\Sigma_{12} = \Sigma_{21} = \rho\,\sigma_\eta$, and $\Sigma_{22} = \sigma_\eta^2$. To model the fat-tail characteristic, using the Stochastic Volatility with Student-t innovations (SVt), it is used the mixing variable that follows a gamma distribution, $\lambda_t \sim \mathcal{G}(\nu/2, \nu/2)$. The parameter vector is given by $(\mu, \phi, \sigma_\eta, \rho, \nu)$, where μ represents the mean level of the volatility, the parameter ϕ, with $|\phi| < 1$, the volatility persistence, σ_η the volatility of the volatility, ρ the parameter that characterises the leverage effect, i.e., with $corr(\varepsilon_t, \eta_{t+1}) = \rho < 0$, a give shock associated with an extreme negative return will increase the future expected volatility. Finally, ν characterises the thickness of the tails of the unconditional distribution of returns.

Univariate time-series are considered and the previous formulation encompasses the four most used versions of the SV model, the standard, the asymmetric

(ASV), the model with fat-tails (SVt), and the asymmetric with fat-tails (ASVt). The standard SV model is obtained making $\rho = 0$ and $\nu = \infty$. By considering the parameter vector $\theta = (\mu, \phi, \sigma_\eta, \rho, \nu)$, it is assumed that the leverage effect and the fat-tails characteristics are modelled simultaneously.

4.1 Model's estimation

Due the nature of the returns characterised in the section 2, where the volatility clustering characteristic is apparent, led us to think that when estimating the parameters as with daily returns, we find parameters that can be as interpretable as the ones obtained in that scenario. The key is the value of the persistence parameter and the ability of obtain meaningful forecasts that approximate the volatility measure. There is now a new feature that must be taken into account. There is a substantial increase in the number of observations that are used to estimate the parameters of the model.

Jacquier et al. (1994) began a series of important research on the best way to estimate the model, and for the most cited, see Shephard and Pitt (1997), Kim et al. (1998), Chib et al. (2002), Chib et al. (2006) and Omori et al. (2007). Different approaches have been proposed, which are associated with different algorithms to obtain samples for the vector of states. Here, we develop a gaussian approximation which gives high acceptance rates in the MCMC simulations used in the estimation process. Essentially, as in Chib and Greenberg (1994) independent samplers based on the gaussian approximations are used. The main aspect is that in the context of simulating the states in for the SV models, those approximations are very robust, and allow easily to generalise from simpler to more complex model. In the next section we only present the algorithm for the basic SV model, but the extensions are straightforward, and more importantly are easy to code.

4.2 Simulate the states

To estimate the model, the marginal posterior distribution of the parameters is approximated. This can be done by simulating from the distribution of $\alpha_{1:n}, \theta | y_{1:n}$, where $\alpha_{1:n} = (\alpha_1, \ldots, \alpha_n)$ and $y_{1:n} = (y_1, \ldots, y_n)$. Using Gibbs sampling the parameters are sampled conditional on the states, $\theta|\alpha_{1:n}, y_{1:n}$, and the states conditional on the parameters, $\alpha_{1:n}|\theta, y_{1:n}$.

The literature that addresses the problem of estimating the parameters of the SV model is very vast. Very sophisticated algorithms were developed to obtain efficient ways of estimating the parameters through the Bayesian estimation paradigm using MCMC sampling. We have found some conflicting results, and the development of very complex algorithms can be error prone in the implementation of a digital computer.

We develop a single-move sampler to simulate from $\alpha_t|\alpha_{\setminus t}, \theta, y_{1:n}$, where $\alpha_{\setminus t} = (\alpha_1, \ldots, \alpha_{t-1}, \alpha_{t+1}, \alpha_n)$. Based on a second order Taylor approximation to the target density gives a gaussian density as the approximating density. Here we present the results associated with the basic SV model, which allows the definition of an analytical formula for the mean and variance of the gaussian approximation, but it is straightforward if such components were obtained numerically, as they need to be in extensions of the SV considered.

Assuming that at iteration k the sampled elements are $\theta^{(k)} = (\mu^{(k)}, \phi^{(k)}, \sigma_\eta^{(k)})$ and $\alpha^{(k)} = (\alpha_1^{(k)}, \ldots, \alpha_n^{(k)})$, at iteration $k+1$ the algorithm proceeds as

1. Sample from $\alpha_t|\alpha_{t-1}^{(k+1)}, \alpha_{t+1}^{(k)}, y_t, \theta^{(k)};\ \ t = 1, \ldots, n.$
2. Sample from $\mu, \phi|\sigma_\eta^{(k)}, \alpha^{(k+1)}, y_{1:n}$.
3. Sample from $\sigma_\eta|\mu^{(k+1)}, \phi^{(k+1)}, \alpha^{(k+1)}$.

To obtain samples for the states in step 1, the algorithm proceeds as follows. The logarithm of the density function assumes the form

$$\ell(\alpha_t) \propto -\frac{\alpha_t}{2} - \frac{y_t^2}{2e^{\alpha_t}} - \frac{(\alpha_{t+1} - \mu - \phi(\alpha_t - \mu))^2}{2\sigma_\eta^2} - \frac{(\alpha_t - \mu - \phi(\alpha_{t-1} - \mu))^2}{2\sigma_\eta^2}\ , \tag{4}$$

for which the maximizer is defined as

$$\alpha_t^* = W\left(\frac{y_t^2\sigma_\eta^2\, e^{-\varphi}}{2(1+\phi^2)}\right) + \varphi, \text{ with } \varphi = \frac{\phi(\alpha_{t+1} + \alpha_{t-1})}{1+\phi^2} - \frac{4\phi\mu + \sigma_\eta^2}{2(1+\phi^2)} + \mu\ , \tag{5}$$

where W is the Lambert function. The second order Taylor approximation of $\ell(\alpha_t)$ around α_t^* is the log-kernel of a Gaussian density with mean α_t^* and variance

$$s_t^2 = -\frac{1}{\ell''(\alpha_t^*)} = \frac{2e^{\alpha_t^*}\sigma_\eta^2}{y_t^2\sigma_\eta^2 + 2(1+\phi^2)e^{\alpha_t^*}}\ . \tag{6}$$

This is the approximating density used to obtain samples for the vector of states. The approximation is very good and the acceptance rates are very high. However, even with chains that always move, sometimes they move slowly, and high levels of autocorrelation are obtained. Due to the simplicity of the sampler, several strategies may be considered to reduce the levels of autocorrelation and to define more efficient estimation procedures. The main point to highlight is that, with SV models, gaussian approximations are straightforward to implement.

4.3 Parameters' estimation

To estimate the parameter vector using Bayesian estimation methods, obtained using the marginal posterior distributions of the parameters, as these do not present an analytical tractable form, they are approximated through Markov chain Monte Carlo (MCMC) simulation techniques. The main task is to approximate the posterior distribution of $\theta|\alpha_{1:n}, y_{1:n}$, where $\alpha_{1:n} = (\alpha_1, \ldots, \alpha_n)$ and $y_{1:n} = (y_1, \ldots, y_n)$. The common approach is to simulate from the distribution of $\theta, \alpha_{1:n}|y_{1:n}$, and through marginalisation approximate the posterior distribution of the parameter vector. To simulate from the former, an iterative procedure of simulating from the conditional distributions is devised, simulate from the parameter vector given the vector of states, $\theta|\alpha_{1:n}, y_{1:n}$, and simulate from the vector of states given the parameter vector, $\alpha_{1:n}|\theta, y_{1:n}$. This kind of approach to estimate the SV models has appeared first in Jacquier et al. (1994), and further developments can be seen in Kim et al. (1998), Chib et al. (2002), Omori et al. (2007), Kastner and Frühwirth-Schnatter (2014), Djegnéné and McCausland (2015).

4.3.1 Sampling the parameter vector

It is the marginal posterior distribution of the parameters that is used in a Bayesian analysis to define a point estimate to the parameters. As the distribution does not have an analytical tractable form, it is approximated by simulating from $\theta|\alpha_{1:n}, y_{1:n}$, where $\theta = (\mu, \phi, \sigma_\eta^2)$ for the SV model and $\theta = (\mu, \phi, \sigma_\eta^2, \rho)$ for the ASV model. The critical element to simulate in this kind of models is the vector of states, mainly due to its dimension. Given the vector of states, it is relatively straightforward to simulate from the marginal posterior distribution of the parameters.

To define the posterior, a prior distribution for the parameter vector needs to be specified. It is common to assume the independence of the parameters within the prior and for the ASV model, $p(\theta) = p(\mu)p(\phi)p(\sigma_\eta^2)p(\rho)$ (Kim et al., 1998; Chib et al., 2002; Omori et al., 2007; Kastner and Frühwirth-Schnatter, 2014). For μ, ϕ and σ_η^2 taken individually, conditional on the observations, states, and remaining parameters, conjugate prior distributions can be defined, gaussian for μ and ϕ, and inverse-gamma to σ_η^2.

There is a well established set of results on the sensitivity of the marginal posterior distribution to the different forms that the prior distributions may assume. Naturally the influence of the priors depend on the information contained in the likelihood. With the amount of information usually available for the estimation of this kind of models, the sensitivity of the marginal posterior distribution to different forms of the priors can be small.

For the ASV the same priors are used, and to distinguish σ_η and ρ, the results presented in Jacquier et al. (2004) are considered. With $\psi = \rho\sigma_\eta$ and $\Omega = \sigma_\eta^2(1-\rho^2)$, estimated through a linear model, the original parameters are obtained as $\sigma_\eta^2 = \Omega + \psi^2$ and $\rho = \psi/\sigma_\eta$.

4.4 Model extensions

The SV model is the main model considered here. This model has been extensively applied using daily returns, and to intraday as been considered in Stroud and Johannes (2014). Several extensions of the basic SV have been introduced, but since Jacquier et al. (1994) the main focus has been in developing efficient estimation algorithms, which have been the object of the research presented in Shephard (1996) Shephard and Pitt (1997), Kim et al. (1998), Chib et al. (2002) Omori et al. (2007) Omori and Watanabe (2008) Kastner and Frühwirth-Schnatter (2014) and Djegnéné and McCausland (2015). However, it must be assumed that the series used carry sufficient information to distinguish the effect produced by all the parameters. There has been certain formulations that have implied some doubts on the ability of data and estimation methods being able to estimate in a consistent way all parameters.

In Stroud and Johannes (2014), extensions to the basic SV model have been taken to an extreme. The extensions include components associated with different levels of persistence (two-factor model), seasonal effects, news announcements, fat tails, asymmetric effects, and jumps in the observation and latent processes.

To our knowledge SV models were never applied to time deformed observations as with the ones that are considered here. On the other hand, we can extract from data other elements of information, variables that must be related with the volatility evolution of returns, here in a time deformed setting, durations between trades can be retrieved, and also the number of shares (or the mean) traded when the directing process in the time deformation is constituted by the number of trades.

We expect that the added sources of information allow some of the problems that we report for the estimation of wide extensions associated with the SV model as were considered in Stroud and Johannes (2014) can be avoided. Here we are going to consider the role played by the durations that can contribute for obtaining more consistent parameter estimates in complex models and in this manner be able to improve forecasts obtained through them.

We consider two simplified versions of the model applied in Stroud and Johannes (2014) trying to individualise two features of the model that potentially can cause estimation issues, the two-factor, and the presence of jumps in the volatility process. Consider the simple extension of the basic SV model

$$y_t = \sigma_y \exp(\alpha_t/2 + \gamma_t/2)\,\varepsilon_t \tag{7}$$

$$\alpha_{t+1} = \phi_1 \alpha_t + \sigma_1\,\eta_{1,t} \tag{8}$$

$$\gamma_{t+1} = \phi_2 \gamma_t + \sigma_2\,\eta_{2,t} \tag{9}$$

with $(\varepsilon_t, \eta_{1,t}, \eta_{2,t}) \sim N(0, I_3)$, the parameter σ_y represents the level of volatility, $|\phi_1| < 1$ and $|\phi_2| < 1$ represent the process persistence, and with $\phi_1 > \phi_2$, α_t represents the slow-moving latent factor whereas γ_t the fast-moving one. Parameters σ_1 and σ_2 represent the usual volatility of the volatility, here decomposed into two components. This kind of model was already referred in Shephard (1996), estimated by Liesenfeld and Richard (2003), analysed in Durham (2006), and applied to intraday data in Stroud and Johannes (2014). Even with strong priors and restrictions on the parameters ($\phi_1 > \phi_2$), consistent estimation of all parameters can be very tricky, essentially due to the non-identification of all parameters.

Another extension that has been considered is related to jumps, namely in the volatility process. Let us consider the simple extension of the basic SV model with

$$\begin{aligned} y_t &= \sigma_y \exp(\alpha_t/2)\,\varepsilon_t &(10)\\ \alpha_{t+1} &= \phi\,\alpha_t + J_t Z_t + \sigma_\eta \eta_t &(11) \end{aligned}$$

the parameter σ_y represents the level of volatility, $|\phi| < 1$ the volatility persistence, with $J_t \sim \mathcal{B}(p)$, $Z_t \sim N(\mu_z, \sigma_z^2)$, representing the existence of a jump, where $J_t = 0, 1$, with $P(J_t = 1) = p$, and a jump size Z_t that follows a normal distribution with mean μ_z and variance σ_z^2.

The model is not identifiable, the level of volatility is influenced by the parameters σ_y, μ_z, σ_z, and even by p, and a unique series cannot disentangle all these effects. Following the specification presented in Nakajima and Omori (2009), assuming $Z_t \sim N(\mu_z, \mu_z^2)$, can in a certain way soften the problem, but we have to recall that Nakajima and Omori (2009) considered the jumps only in the observation process. Finally, we have to note that the error term in the system equation ($J_t Z_t + \sigma_\eta \eta_t$) is given by a mixture, and following Celeux et al. (2000) saying "... we consider that almost the entirety of Markov chain Monte Carlo samplers implemented for mixture models has failed to converge! Moreover, we wish to stress that harm can result from the statistical interpretation of Markov chain Monte Carlo samples produced by placing constraints on the parameters", we are lead to think that care must be taken when jumps are included.

Our main goal is to analyse the evolution of volatility, but before we can do that, we must address the possible identification problems associated with the models proposed. Even within a Bayesian framework, with prior distributions for the parameters, and the so called identification restrictions, some doubts can be cast about the meaningfulness of some estimates for the parameters in certain models.

The idea of working with time deformed observations (prices and returns) is related to the fact that information do not arrive to the market in an evenly manner. On the other hand, if the aim is to analyse the volatility, other sources of information that not only the prices can be used to characterise the phenomenon, for example, the volume. In periods of time that a lot of information is arriving to the market, good or bad, these can trigger many adjustments on portfolios, which in turn will increase volatility, and this can be related to the rise in volume (number) of transactions. Using daily data, several attempts were considered, but until now, there is not a full acceptance and appreciation of the analysis that include the volume in volatility evolution characterisation.

Using intraday data we return to a kind of analysis that addresses the volatility evolution not in the calendar-time but in the trading-time. This setting accommodates some of the issues difficult to address in calendar-time, namely the seasonality, identification of all parameters in the models, and allow the introduction of other information sources as the volume and durations, that can help the characterisation of high-frequency volatility evolution.

Because of the identification issue, if we want to generalise the models and pick different characteristics of the returns distribution, the usual approach is to consider a new set of parameters to characterise the new features of interest. However, certainly there are limitations for which a given series can distinguish all kind of parameters. In a time deformed setting, new variables can be associated with the evolution of returns which allow the information disentanglement.

If there is lack of information in a series to distinguish the effect of all parameters, one possible solution is searching for other variables that may be included and are related to the main variable of interest. In the time deformed environment we can observe returns but also durations, and smaller ones imply increased activity with greater volatility and higher probability of jumps to happen. A first attempt is to model jointly returns and durations in a bivariate SV model, also including jumps,

$$\begin{aligned}
y_t &= \sigma_y \exp(\alpha_t/2)\,\varepsilon_t; \quad \varepsilon_t \sim N(0,1) \\
d_t &= \exp(-\alpha_t/2)\,\zeta_t; \quad \zeta_t \sim G(\lambda,1) \\
\alpha_{t+1} &= \phi\alpha_t + J_t\,Z_t + \sigma_\eta \eta_t; \quad \eta_t \sim N(0,1)
\end{aligned}$$

Table 1: Summary of the estimation using MCMC to approximate the posterior distribution of the parameters. Point estimate (Mean) and respective standard deviation (SD). The INEF is given by N/ESS, number of observations in the chain divided by the effective sample size

	With durations			Without durations		
Par.	Mean	SD	INEF.	Mean	SD	INEF
σ_y	0.100	0.002	24.69	0.125	0.018	226.01
ϕ	0.965	0.004	9.72	0.963	0.005	22.78
σ_η	0.186	0.008	24.47	0.188	0.018	125.15
δ	0.214	0.097	6.18	-0.126	0.186	45.52
p	0.039	0.020	2.39	0.040	0.028	83.76
λ	9.870	0.222	26.84			

where a new variable d_t is introduced representing the durations, and λ represents the durations level, $J_t \sim \mathcal{B}(p)$, $Z_t \sim N(\delta, \delta^2)$, representing the existence of a jump, where $J_t = 0, 1$, with $P(J_t = 1) = p$, and a jump size Z_t that follows a normal distribution with mean δ and variance δ^2. In this model the parameter vector is given by $\theta = (\sigma_y, \lambda, \phi, \sigma_\eta, \delta, p)$. Here the durations are included, which embodies information on the volatility evolution, but more importantly, it adds information into the model, which allow a more consistent parameters estimation, namely, the ones jump associated.

Two illustrate the differences we build a simulation example, simulating from the model with $\sigma_y = 0.1$, $\lambda = 10$, $\phi = 0.97$, $\sigma_\eta = 0.2$, $\delta = 0.25$, and $p = 0.02$. This values are in line with the ones found for real data, and even it gives a negative correlation around -0.3 between absolute returns and durations, smaller durations imply an increase in volatility. The model was estimated through MCMC using the STAN package (Stan Development Team. 2017. Stan Modeling Language Users Guide and Reference Manual, Version 2.17.0. http://mc-stan.org). The estimates are presented in Table 1, and we depict in Figure 1, the chains associated with the more problematic parameters, δ and p.

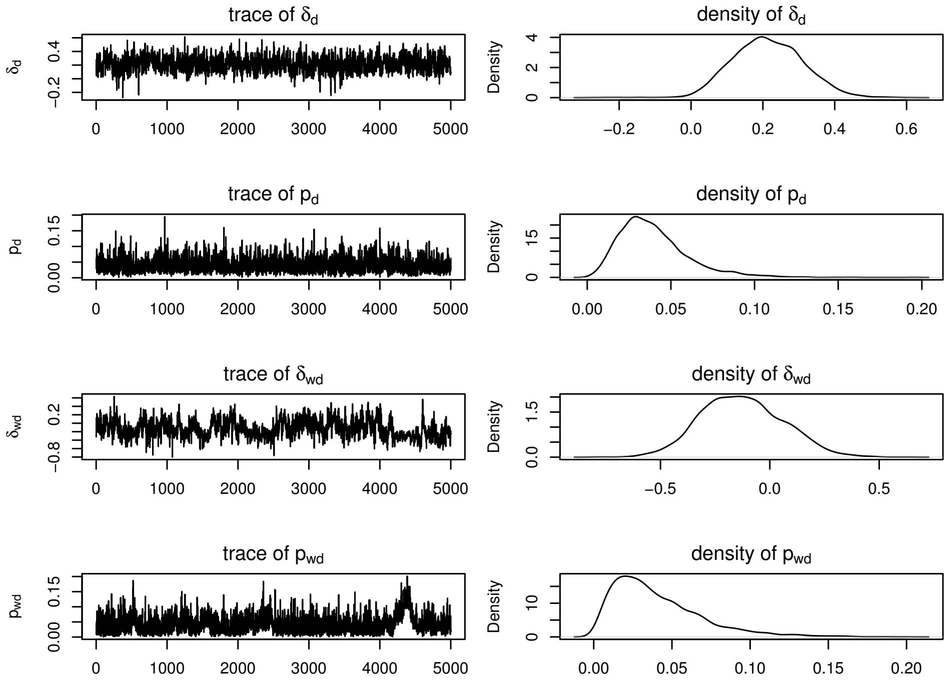

Figure 1: Trace and respective density associated with the chains used to approximate the parameters δ and p, with (d) and without (wd) durations in the respective model

As it can be established by the results presented with the introduction of durations allow an improvement of the estimation process, the estimated values are near the true ones and the mixing of in chains increase dramatically as it can be seen by the values of INEF. Without the increased information induced by durations the jump size is erroneously estimated.

5 Forecasting and particle filter

The SV model is a state space model where the evolution of the states defines the evolution of the volatility. Forecasts for the evolution of the states in this setting require the development of simulation techniques known as Sequential Monte Carlo (SMC), also referred as particle filter methods Andrieu et al. (2010); Carpenter et al. (1999); Del Moral et al. (2006); Doucet et al. (2000); Fearnhead et al. (2010); Godsill and Clapp (2001); Pitt and Shephard (1999). The aim is to update the filter distribution for the states when new information arrive.

Using a model that depends on a set of parameters, all forecasts are conditioned by the parameters. It is not realistic to assume that the parameters are know, and the parameters are estimated through Bayesian estimation methods. This constitutes an approximation, because even if model's uncertainty is not taken into account, it can be assumed that the parameters can vary over time.

The quantities of interest are the values of the states governing the evolution of the volatility, which are propagated to define the predictive density of the returns, defined here as $f(y_{t+1}|y_{1:t})$. However, essential to the definition of this distribution is the filter density associated with the states, $f(\alpha_t|y_{1:t})$. Bayes's rule allows us to assert that the posterior density $f(\alpha_t|y_{1:t})$ of states is related to the density $f(\alpha_t|y_{1:t-1})$ prior to y_t, and the density $f(y_t|\alpha_t)$ of y_t given α_t by $f(\alpha_t|y_{1:t}) \propto f(y_t|\alpha_t)f(\alpha_t|y_{1:t-1})$. The predictive density of y_{t+1} given $y_{1:t}$ is defined by $f(y_{t+1}|y_{1:t}) = \int f(y_{t+1}|\alpha_{t+1})f(\alpha_{t+1}|y_{1:t})\, d\alpha_{t+1}$.

Particle filters approximate the posterior density of interest, $f(\alpha_t|y_{1:t})$, through a set of m "particles" $\{\alpha_{t,1}, \ldots, \alpha_{t,m}\}$ and their respective weights $\{\pi_{t,1}, \ldots, \pi_{t,m}\}$, where $\pi_{t,j} \geq 0$ and $\sum_{j=1}^{m} \pi_{t,j} = 1$. This procedure must be implemented sequentially with the states evolving over time to accommodate new information that arrive. It is difficult to obtain samples from the target density, and an approximating density is used instead, afterwards the particles are resampled to better approximate the target density. This is known as the sample importance resampling (SIR) algorithm. A possible approximating density is given by $f(\alpha_t|\alpha_{t-1})$, however, Pitt and Shephard (1999, 2001) pointed out that as a density to approximate $f(\alpha_t|y_{1:t})$ is not generally efficient, because it constitutes a *blind* proposal that does not take into account the information contained in y_t.

5.1 Particle filter for the SV model

Through SMC with SIR the aim is to update sequentially the filter density for the states. The optimal importance density is given by $f(y_t|\alpha_t)f(\alpha_t|\alpha_{t-1}, y_t)$, which induces importance weights with zero variance. Usually it is not possible to obtain samples from this density, and an importance density $g(\alpha_t)$, different from the optimal density, is used to approximate the target density.

To approximate the filter densities associated with the SV model, Pitt and Shephard (1999) considered the same kind of approximations used to sample the states in a static MCMC setting. However, the approximations were based on a first order Taylor approximation, and it was demonstrated by Smith and Santos (2006) that they are not robust when information contained in more extreme observations need to be updated (also called very informative observations). In Smith and Santos (2006), a second order Taylor approximation for the likelihood combined with the predictive density for the states leads to improvements in the particle filter algorithm. As the auxiliary particle filter in Pitt and Shephard (1999), avoids *blind* proposals like the ones proposed in Gordon et al. (1993), takes into account the information in y_t, and defines a robust approximation for the target density, which also avoids the degeneracy of the weights.

Here we develop the aforementioned results using a robuster approximation for the importance density. The logarithm of the density $f(y_t|\alpha_t)f(\alpha_t|\alpha_{t-1})$, $\ell(\alpha_t)$, is concave on α_t, and to maximize the function in order to α_t, let us consider the first derivative equal to zero, $\ell'(\alpha_t) = 0$. Solving in order to α_t the solution is

$$\alpha_t^* = W\left(\frac{y_t^2 \sigma_\eta^2\, e^{-\gamma}}{2}\right) + \gamma, \text{ with } \gamma = \mu(1-\phi) + \phi\alpha_{t-1} - \frac{\sigma_\eta^2}{2}\ . \tag{12}$$

The second derivative is given by $\ell''(\alpha_t) = -(2e^{\alpha_t} + \sigma_\eta^2\ y_t^2)/(2\sigma_\eta^2\ e^{\alpha_t})$, which is strictly negative for all α_t, so α_t^* maximizes the function $\ell(\alpha_t)$ defining a global maximum. The second order Taylor expansion of $\ell(\alpha_t)$ around α_t^* defines the log-kernel of a gaussian density with mean $m_t = \alpha_t^*$ and variance

$$s_t^2 = \frac{2\sigma_\eta^2\ e^{m_t}}{2e^{m_t} + \sigma_\eta^2\ y_t^2}\ . \tag{13}$$

This gaussian density will be used as the importance density in the SIR algorithm.

In the procedures implemented, the estimates of interest were approximated using particles with equal weights, which means that a resampling step is performed. Assuming at $t-1$ a set of m particles $\alpha_{t-1}^m = \{\alpha_{t-1,1}, \ldots, \alpha_{t-1,m}\}$ with associated weights $1/m$, which approximate the density $f(\alpha_{t-1}|y_{1:t-1})$, the algorithm proceeds as follows

1. For each element of the set, $\alpha_{t,i}$, $i = 1, \ldots, m$, sample a value from a gaussian distribution with mean and variance defined by (12) and (13), respectively, obtaining the set $\{\alpha_{t,1}^*, \ldots, \alpha_{t,m}^*\}$.

2. Calculate the weights,

$$w_i = \frac{f(y_t|\alpha_{t,i}^*) f(\alpha_{t,i}^*|\alpha_{t-1,i})}{g(\alpha_{t,i}^*|m_t, s_t^2)}, \quad \pi_i = \frac{w_i}{\sum_{i=1}^m w_i}\ . \tag{14}$$

3. Resample from the set $\{\alpha_{t,1}^*, \ldots, \alpha_{t,m}^*\}$ using the set of weights $\{\pi_1, \ldots, \pi_m\}$ obtaining a sample $\{\alpha_{t|1:t,1}, \ldots, \alpha_{t|1:t,m}\}$, where to each particle a weight of $1/m$ is associated.

For the one step-ahead volatility forecast, having the approximation to the density $f(\alpha_t|y_{1:t})$, and due to the structure of the system equation in the SV model, AR(1) with gaussian noise, it is easy to sample from $f(\alpha_{t+1}|y_{1:t})$, the predictive density for the states.

6 Application results

The results in this article are associated with the analysis performed for three stocks traded in US markets, Apple (AAPL), General Electric (GE), and Exxon Mobile (XOM), using intraday data from March 2, 2017 to February 22, 2018, collected from the public available information on the site of BATS Exchange, now CBOE.

Table 2: Summary statistics of the intraday return distributions on three stocks, AAPL, GE and XOM. In the first column, in brackets following the designation of the stock is indicated the number of trades used to define the volume-adapted returns. It is given the variance (Var.), the kurtosis (Kurt.), and the autocorrelations associated with the absolute returns for lag $i = 1, 2, 3$ ($\rho_{|y_{t-i}|}$). The column MD represents the mean duration (in minutes) between each recorded observation, and the last column ($\rho_{|y|,d}$) represents the correlation between absolute returns and durations

Stocks	Var.	Kurt.	$\rho_{\|y_{t-1}\|}$	$\rho_{\|y_{t-2}\|}$	$\rho_{\|y_{t-3}\|}$	MD	$\rho_{\|y\|,d}$
AAPL (400)	0.012	7.529	0.252	0.240	0.245	4.935	-0.290
AAPL (1200)	0.033	7.452	0.198	0.206	0.180	15.110	-0.310
GE (250)	0.017	7.613	0.203	0.191	0.182	4.985	-0.205
GE (750)	0.047	7.672	0.196	0.177	0.136	15.199	-0.213
XOM (150)	0.007	11.932	0.265	0.214	0.216	4.751	-0.157
XOM (450)	0.018	6.472	0.210	0.192	0.167	14.466	-0.198

The data was collected from the four order books available, and the frequency available is the millisecond. In the process of recording the data some observations might be lost, but the recorded data is constituted by the price that the stock was traded, the number of shares traded, and the time (to the millisecond) of the

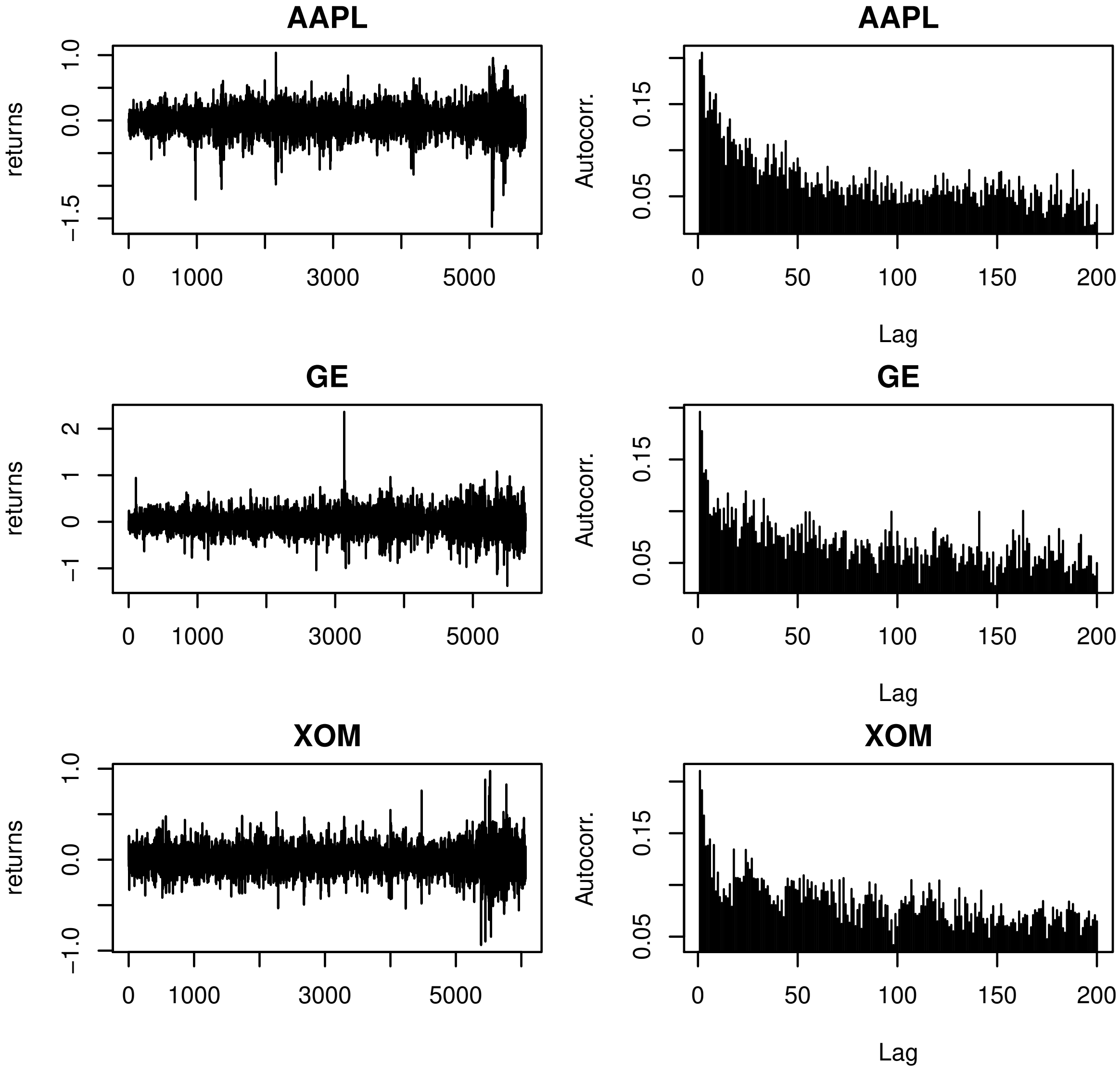

Figure 2: Evolution of volume-adapted returns for three stocks, AAPL, GE and XOM. The volume-returns were defined through the time deformation associated with a given target for the number of trades. The values were chosen to mimic the calendar time most used 1-min, 5-min, 15-min and 30-min. The referential presented in the figure correspond to 1200 for AAPL, 750 for GE and 450 for XOM, which correspond of an average of 15 minutes between recorded values. The autocorrelation for the absolute return is also given, indicating the persistence of volatility

Table 3: Summary of estimation results associated with the simulation used to approximate the posterior distribution of the parameters for four models, SV, SVt, ASV and ASVt, for the AAPL stock. The first column represents the mean of the posterior distribution (Mean), which represents the point estimate, the second column an estimate of the standard deviation (SD) of the estimator, the third is the inefficiency factor (INEF) calculated as the number of observations in the sample (N) and the effective sample size (ESS). The last two columns are related to the Geweke's statistics to assess the convergence of the chains

Par.	Mean	SD	INEF	Geweke-z	p-value
μ	-3.717	0.067	1.931	-1.228	0.220
ϕ	0.952	0.009	40.851	-0.074	0.941
σ_η	0.226	0.023	58.998	0.128	0.898
μ	-3.793	0.084	14.593	2.433	0.015
ϕ	0.970	0.008	89.401	-0.744	0.457
σ_η	0.170	0.023	120.670	0.844	0.399
ν	21.110	7.817	131.666	1.690	0.091
μ	-3.716	0.063	2.240	0.645	0.519
ϕ	0.949	0.009	32.969	0.903	0.366
σ_η	0.235	0.022	44.423	-0.550	0.582
ρ	-0.259	0.046	6.860	-1.618	0.106
μ	-3.774	0.074	8.647	0.382	0.702
ϕ	0.962	0.009	67.533	0.846	0.398
σ_η	0.196	0.025	99.102	-0.778	0.436
ρ	-0.298	0.054	14.283	-0.284	0.777
ν	28.251	10.403	55.686	-0.535	0.593

Table 4: Summary of estimation results associated with the simulation used to approximate the posterior distribution of the parameters for four models, SV, SVt, ASV and ASVt, for the GE stock. The first column represents the mean of the posterior distribution (Mean), which represents the point estimate, the second column an estimate of the standard deviation (SD) of the estimator, the third is the inefficiency factor (INEF) calculated as the number of observations in the sample (N) and the effective sample size (ESS). The last two columns are related to the Geweke's statistics to assess the convergence of the chains

Par.	Mean	SD	INEF	Geweke-z	p-value
μ	-3.375	0.060	3.260	-0.464	0.643
ϕ	0.931	0.013	42.860	-0.167	0.867
σ_η	0.283	0.030	52.476	0.050	0.960
μ	-3.506	0.095	11.633	-0.146	0.884
ϕ	0.977	0.007	116.961	-1.006	0.314
σ_η	0.142	0.023	171.839	0.897	0.370
ν	10.417	1.911	78.801	0.274	0.784
μ	-3.380	0.058	2.077	0.802	0.422
ϕ	0.926	0.013	30.393	0.006	0.995
σ_η	0.295	0.029	41.848	-0.189	0.850
ρ	-0.016	0.042	4.410	-0.671	0.502
μ	-3.490	0.083	9.877	-0.177	0.860
ϕ	0.967	0.009	98.020	0.485	0.628
σ_η	0.178	0.027	138.167	-0.480	0.631
ρ	-0.030	0.056	10.208	1.068	0.286
ν	12.846	3.175	75.655	-0.080	0.936

trade. For three stocks analysed, and for the sample period aforementioned the number of observations were 7 409 509, 4 585 078 and 2 888 670, respectively.

The aim is to show how such a rich dataset can be used to characterise the high-frequency volatility associated with the asset returns. In other contexts static environments were considered where the aim was to understand the fat tail characteristic of the financial returns, and the volume would be used as the scaling factor, giving rise to normal distributed returns when conveniently scaled. Here we assume that agents might need to take decision within very short periods of time, and that measures of volatility adapted to such periods are needed. We find that intraday returns also present fat tails, and as with other frequencies they also cluster, indicating volatility dynamics.

The pattern of volatility clustering can be perceived by the analysis of Figure 2, where returns and autocorrelations for the absolute values are depicted. The evolution of autocorrelation are in line with what is observed for daily returns where these procedures have been most applied. The correlations are small but highly persistent, indication a relationship, but naturally a nonlinear one.

As can be checked by the analysis of autocorrelations associated with the absolute returns, no seasonal component can be perceived. The adjustments is made through the durations. In time deformed setting the seasonal components are adjusted automatically, and in Figure 3, is depicted the evolution of durations. Through an autocorrelation analysis a seasonal pattern can be perceived, indicating that durations are smaller at opening and near closing of markets. A simple filter was applied to durations for obtaining a seasonal adjusted series, depicted in Figure 4.

As it is shown in Table 2, through the correlation between absolute returns and durations, in the same order of values as the ones of autocorrelations associated with absolute returns, with the definition and usage of time deformed returns, new information can be retrieved from intraday data, namely the durations, which coupled with the evolution of returns can give us a clearer approximation for the volatility evolution.

The aim is to estimate and forecast the volatility. If a used measure is the variance of returns, that is time-varying, dynamic models must be used and they

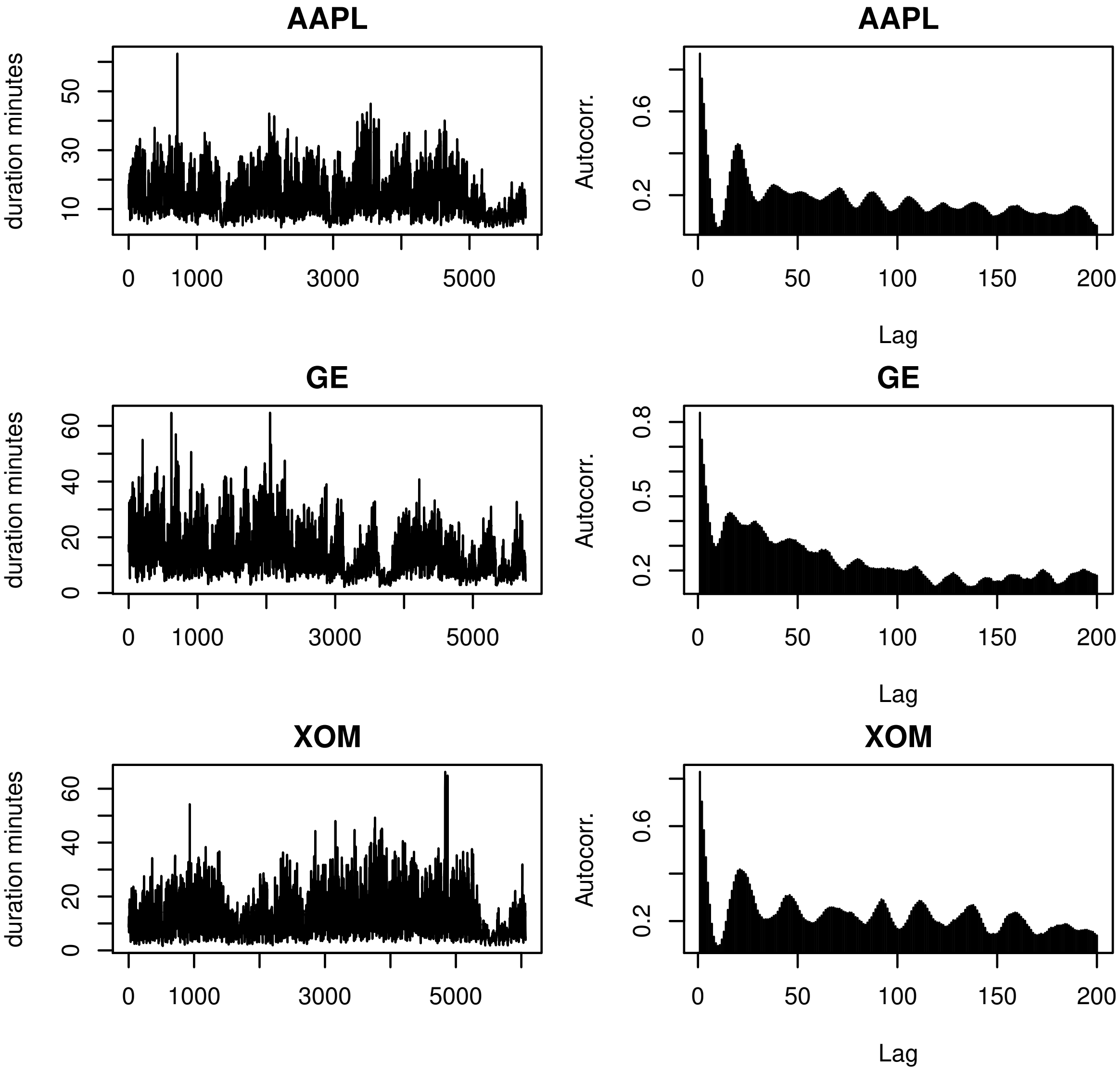

Figure 3: Durations and respective autocorrelations for the recorded observations associated with the returns depicted in Figure 2. To the evolution of the autocorrelations is clear the seasonal pattern that exists for the series

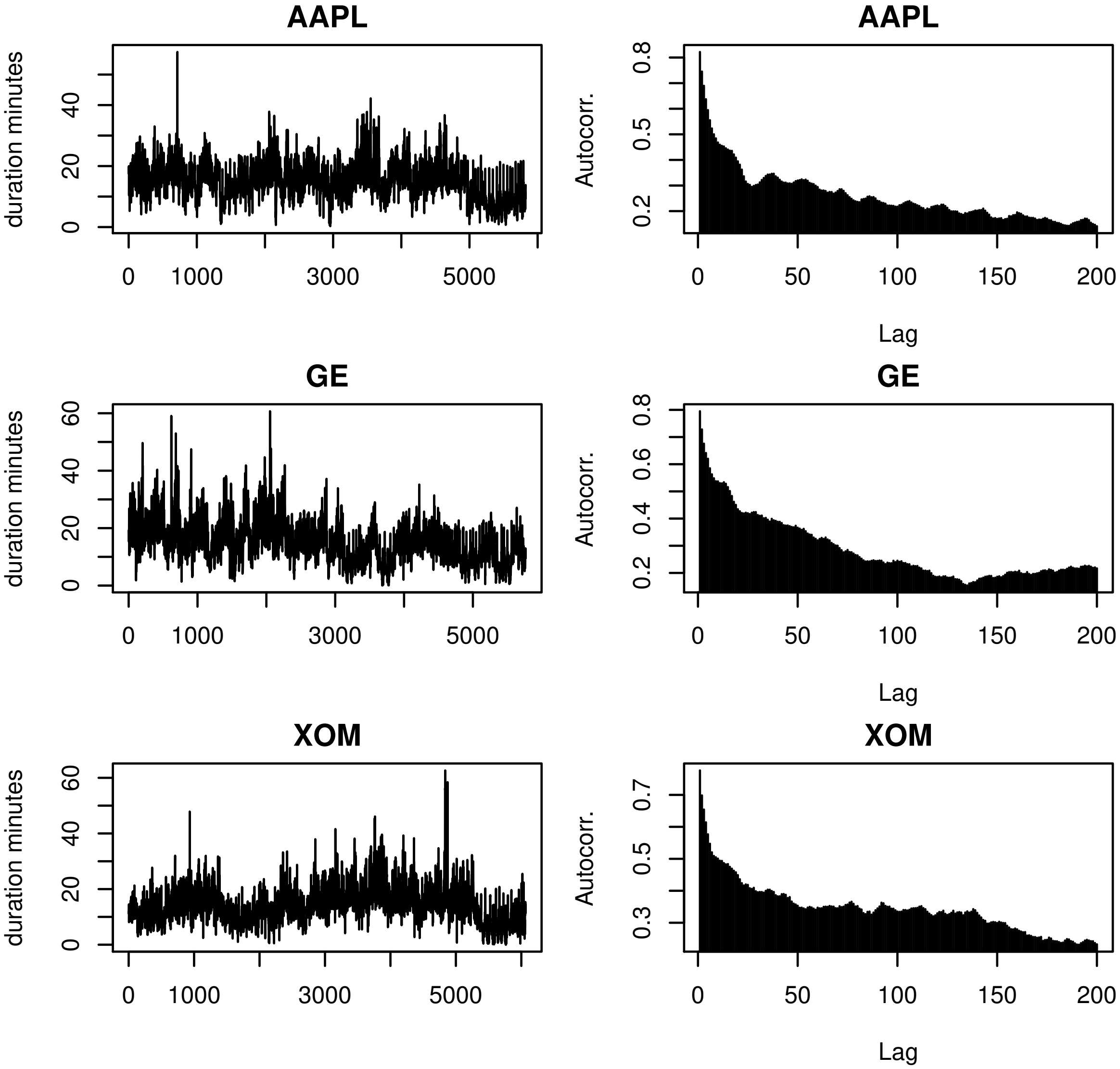

Figure 4: Durations and respective autocorrelations as in Figure 3 after passing a filter to remove part of the seasonality

assume nonlinear features. Forecasts based on a model usually need a set of parameters that link variables with the element that a forecast is built to. Parameters must be estimated through data and the quality of forecasts are inevitably linked with the one from parameters' estimates.

In nonlinear models used to characterise volatility evolution as are ARCH and SV, implies that large samples are needed to obtain reliable results. Less than 1 000 observations in the sample can result in estimates subjected to a high variability. In this context, using daily data, and to apply estimation procedures to these models, samples that correspond to 30 years of data have been used (around 8 000 observations). We can discuss the validity of such estimates, not in terms of sample variability, but instead asking if a unique structure associated with the model can accommodate the characterisation of a given market for such long period of time.

Intraday data is an element that can address problems associated with possible changes of structures, because we can obtain substantial amounts of data, necessary to estimate conveniently the parameters of high-dimensional nonlinear models, and that data being associated with a relatively short period of time, where no significant structural changes are expected.

Results associated with the estimation of different versions, SV, SVt, ASV and ASVt, are presented in Table 3 and 4. The characteristics most relevant when daily returns are used, can be found with intraday returns, especially the volatility clustering, indicated by a value for the persistence parameter higher than 0.9. In some cases the Student-t distribution is important for the observations equation, more with GE, and the leverage effect parameter is also important to model the AAPL returns.

To reinforce the argument present here that returns in a time deformed setting using trade and volume information can be important to model volatility evolution, apart from analysis and decisions at varying speeds when measured in calendar time, it allows other kind of information related to volatility that can be retrieve from the data, as are duration and volume of trade. For a series associated with AAPL returns (1200 trades - 15-min) the models as in the simulation above were estimated. The parameter λ had to be fixed, and the value $\lambda = 15$ was used.

Table 5: Summary of the estimation using MCMC to approximate the posterior distribution of the parameters for SV-Dur with jumps for AAPL stock. Point estimate (Mean) and respective standard deviation (SD). The INEF is given by N/ESS, number of observations in the chain divided by the effective sample size

	With durations			Without durations		
Par.	Mean	SD	INEF.	Mean	SD	INEF
σ_y	0.146	0.001	0.72	0.145	0.014	230.93
ϕ	0.933	0.005	1.77	0.951	0.009	132.22
σ_η	0.282	0.007	6.89	0.212	0.002	249.91
δ	0.396	0.067	3.02	0.221	0.250	111.25
p	0.041	0.011	1.04	0.034	0.020	26.28

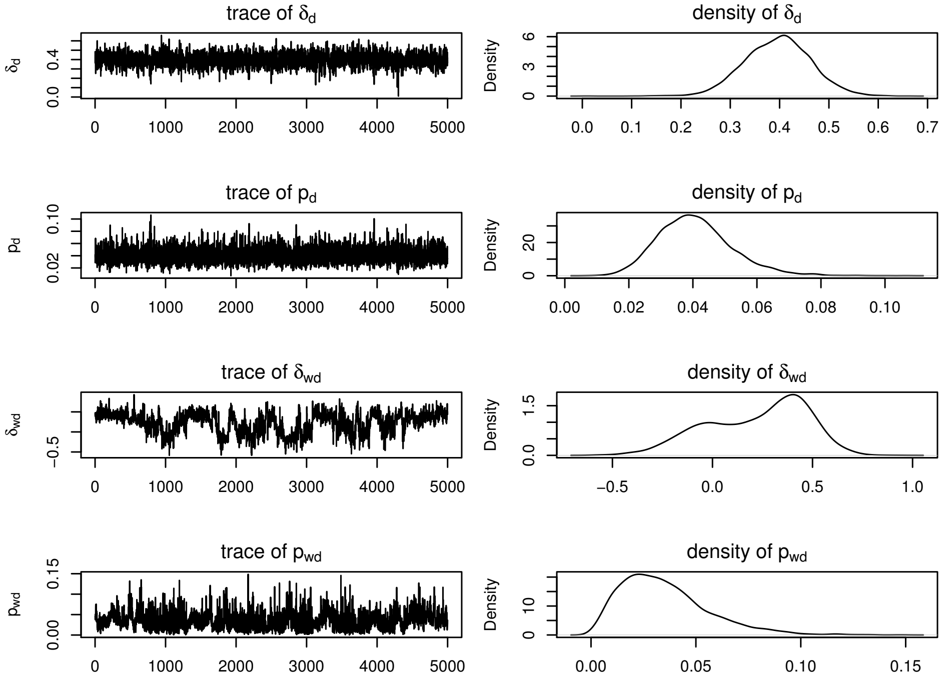

Figure 5: Trace and respective density associated with the chains used to approximate the parameters δ and p, with (d) and without (wd) durations in the models applied to AAPL returns

The results presented in Table 5, and depicted in Figure 5, show that it is problematic to estimate the parameters of a SV models with jumps related to the latent process, the chains have difficulties to converge, with posterior distributions presenting some degree of bimodality, which is critical to Bayesian estimation. On the other hand, by generalising the model including durations, mixing of chains increase dramatically, with reduction of standard deviations, which represent a most trustworthy scenario of estimation.

7 Concluding remarks

The results presented demonstrate that similar procedures used to estimate and forecast through SV models using daily observations can also be applied to intraday data. Assuming that the volatility is time-varying within a given day, the intraday volatility evolution is modelled. A major issue is how to define the appropriate frequency for the intraday returns. In this article is adopted a different approach by defining the returns in volume-domain, which can offer a different perspective for the volatility evolution. It incorporates the volume, which can turn the modelling of volatility evolution more flexible, taking into account the differences of activity on different days of negotiation, depending on the amount of relevant information that arrive to the market.

References

Andersen, T. G. (1996). Return volatility and trading volume: An information flow interpretation of stochastic volatility. *The Journal of Finance 51*(1), 169–204.

Andersen, T. G. and T. Bollerslev (1997). Intraday periodicity and volatility persistence in financial markets. *Journal of empirical finance 4*(2), 115–158.

Andersen, T. G. and T. Bollerslev (1998). Deutsche mark–dollar volatility: intraday activity patterns, macroeconomic announcements, and longer run dependencies. *the Journal of Finance 53*(1), 219–265.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens (2001). The distribution of realized stock return volatility. *Journal of Financial Economics 61*(1), 43–76.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica 71*(2), 579–625.

Andersen, T. G., T. Bollerslev, and S. Lange (1999). Forecasting financial market volatility: Sample frequency vis-a-vis forecast horizon. *Journal of Empirical Finance 6*(5), 457–477.

Andersen, T. G., T. Bollerslev, and N. Meddahi (2005). Correcting the errors: Volatility forecast evaluation using high-frequency data and realized volatilities. *Econometrica 73*(1), 279–296.

Andersen, T. G., T. Bollerslev, and N. Meddahi (2011). Realized volatility forecasting and market microstructure noise. *Journal of Econometrics 160*(1), 220–234.

Andersen, T. G. and T. Teräsvirta (2009). Realized volatility. In *Handbook of Financial Time Series*, pp. 555–575. Springer.

Andrieu, C., A. Doucet, and R. Holenstein (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(3), 269–342.

Barndorff-Nielsen, O. E. and N. Shephard (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*(2), 253–280.

Barndorff-Nielsen, O. E. and N. Shephard (2004). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica 72*(3), 885–925.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics 31*(3), 307–327.

Carpenter, J., P. Clifford, and P. Fearnhead (1999). Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation 146*(1), 2–7.

Celeux, G., M. Hurn, and C. P. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association 95*(451), 957–970.

Chib, S. and E. Greenberg (1994). Bayes inference in regression models with arma (p, q) errors. *Journal of Econometrics 64*(1), 183–206.

Chib, S., F. Nardari, and N. Shephard (2002). Markov chain monte carlo methods for stochastic volatility models. *Journal of Econometrics 108*(2), 281–316.

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

Clark, P. K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica: journal of the Econometric Society*, 135–155.

Darolles, S., C. Gourieroux, and G. Le Fol (2000). Intraday transaction price dynamics. *Annales d'Economie et de Statistique*, 207–238.

Darolles, S., G. Le Fol, and G. Mero (2015). Measuring the liquidity part of volume. *Journal of Banking & Finance 50*, 92–105.

Darolles, S., G. Le Fol, and G. Mero (2017). Mixture of distribution hypothesis: Analyzing daily liquidity frictions and information flows. *Journal of Econometrics 201*(2), 367–383.

Del Moral, P., A. Doucet, and A. Jasra (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(3), 411–436.

Djegnéné, B. and W. J. McCausland (2015). The hessian method for models with leverage-like effects. *Journal of Financial Econometrics 13*(3), 722–755.

Doucet, A., S. Godsill, and C. Andrieu (2000). On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing 10*(3), 197–208.

Durham, G. B. (2006). Monte carlo methods for estimating, smoothing, and filtering one-and two-factor stochastic volatility models. *Journal of Econometrics 133*(1), 273–305.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987–1007.

Epps, T. W. (1976). The stochastic dependence of security price changes and transaction volumes in a model with temporally dependent price changes. *Journal of the American Statistical Association 71*(356), 830–834.

Epps, T. W. and M. L. Epps (1976). The stochastic dependence of security price changes and transaction volumes: Implications for the mixture-of-distributions hypothesis. *Econometrica: Journal of the Econometric Society*, 305–321.

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

Fearnhead, P., D. Wyncoll, and J. Tawn (2010). A sequential smoothing algorithm with linear computational cost. *Biometrika 97*(2), 447–464.

Godsill, S. and T. Clapp (2001). Improvement strategies for monte carlo particle filters. In *Sequential Monte Carlo Methods in Practice*, pp. 139–158. Springer.

Gordon, N. J., D. J. Salmond, and A. F. Smith (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, Volume 140, pp. 107–113. IET.

Gourieroux, C. and J. Jasiak (2001). *Financial Econometrics.* Princeton.

Gourieroux, C. and G. Le Fol (1997). Volatilités et mesures de risque. *Journal de la Société de Statistique de Paris 138*(4), 1–32.

Granger, C. W. (2002). Some comments on risk. *Journal of Applied Econometrics 17*(5), 447–456.

Jacquier, E., N. G. Polson, and P. E. Rossi (1994). Bayesian analysis of stochastic volatility models. *Journal of Business & Economic Statistics 12*(4), 371–89.

Jacquier, E., N. G. Polson, and P. E. Rossi (2004). Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *Journal of Econometrics 122*(1), 185–212.

Kastner, G. and S. Frühwirth-Schnatter (2014). Ancillarity-sufficiency interweaving strategy (asis) for boosting mcmc estimation of stochastic volatility models. *Computational Statistics & Data Analysis 76*, 408–423.

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

Lamoureux, C. G. and W. D. Lastrapes (1990). Heteroskedasticity in stock return data: volume versus garch effects. *The Journal of Finance 45*(1), 221–229.

Le Fol, G. and M. Ludovic (1998). Time deformation: Definition and comparisons. *Journal of Computational Intelligence in Finance 6*(5), 19–33.

Liesenfeld, R. and J.-F. Richard (2003). Estimation of dynamic bivariate mixture models: comments on watanabe (2000). *Journal of Business & Economic Statistics 21*(4), 570–576.

Maillet, B. and T. Michel (1997). Mesures de temps, information et distribution des rendements intra-journaliers. *Journal de la société française de statistique 138*(4), 89–120.

McAleer, M. and M. C. Medeiros (2008). Realized volatility: A review. *Econometric Reviews 27*(1-3), 10–45.

Nakajima, J. and Y. Omori (2009). Leverage, heavy-tails and correlated jumps in stochastic volatility models. *Computational Statistics & Data Analysis 53*(6), 2335–2353.

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

Omori, Y. and T. Watanabe (2008). Block sampler and posterior mode estimation for asymmetric stochastic volatility models. *Computational Statistics & Data Analysis 52*(6), 2892–2910.

Pitt, M. K. and N. Shephard (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association 94*(446), 590–599.

Pitt, M. K. and N. Shephard (2001). Auxiliary variable based particle filters. In *Sequential Monte Carlo methods in practice*, pp. 273–293. Springer.

Shephard, N. (1996). *Statistical aspects of ARCH and stochastic volatility.* Springer.

Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-gaussian measurement time series. *Biometrika 84*(3), 653–667.

Smith, J. and A. A. F. Santos (2006). Second-order filter distribution approximations for financial time series with extreme outliers. *Journal of Business & Economic Statistics 24*(3), 329–337.

Stroud, J. R. and M. S. Johannes (2014). Bayesian modeling and forecasting of 24-hour high-frequency volatility. *Journal of the American Statistical Association 109*(508), 1368–1384.

Tauchen, G., H. Zhang, and M. Liu (1996). Volume, volatility, and leverage: A dynamic analysis. *Journal of Econometrics 74*(1), 177–208.

Tauchen, G. E. and M. Pitts (1983). The price variability-volume relationship on speculative markets. *Econometrica: Journal of the Econometric Society*, 485–505.

Taylor, S. J. (1986). *Modelling Financial Time Series.* Wiley: Chichester.

Taylor, S. J. (1994). Modeling stochastic volatility: A review and comparative study. *Mathematical finance 4*(2), 183–204.

Todorov, V. (2011). Econometric analysis of jump-driven stochastic volatility models. *Journal of Econometrics 160*(1), 12–21.

Todorov, V. and G. Tauchen (2011). Volatility jumps. *Journal of Business & Economic Statistics 29*(3), 356–371.

Watanabe, T. (2000). Bayesian analysis of dynamic bivariate mixture models: Can they explain the behavior of returns and trading volume? *Journal of Business & Economic Statistics 18*(2), 199–210.